# Delay distribution of (im)patient customers in a discrete time D-MAP/PH/1 queue with age-dependent service times

# Delay distribution of (im)patient customers in a discrete time D-MAP/PH/1 queue with age dependent service times

B. Van Houdt, R.B. Lenin, C. Blondia
University of Antwerp, Department of Mathematics and Computer Science,
Performance Analysis of Telecommunication Systems Research Group,
Middelheimlaan, 1, B-2020 Antwerp - Belgium,
{benny.vanhoudt,lenin.bhavanandan,chris.blondia}@ua.ac.be

April 4, 2003

**Abstract**

This paper presents an algorithmic procedure to calculate the delay distribution of (im)patient customers in a discrete time D-MAP/PH/1 queue, where the service time distribution of a customer depends on his waiting time. We consider three different situations: impatient customers in the waiting room, impatient customers in the system, that is, if a customer has been in the waiting room, respectively, in the system for $a$ time units it leaves the waiting room, respectively, the system. In the third situation, all customers are patient—that is, they only leave the system after completing service. In all three situations the service time of a customer depends upon the time he has spent in the waiting room. As opposed to the general approach in many queueing systems, we calculate the delay distribution, using matrix analytic methods, without obtaining the steady state probabilities of the queue length. The trick used in this paper, which was also applied by Van Houdt and Blondia [J. Appl. Probab. Vol. 39, No 1, pp. 213–222 (2002)], is to keep track of the "age" of the customer in service, while remembering the D-MAP state immediately after the customer in service arrived. Possible extentions of this method to more general queues and numerical examples that demonstrate the strength of the algorithm are also included.

**Index Terms:** matrix analytic methods, D-MAP arrival process, phase-type services, (im)patient customers, age dependent service times.

## 1 Introduction

This paper introduces an algorithmic procedure to calculate the delay distribution of a first-come-first-serve queue with correlated arrivals (D-MAP, see Section 2), phase-type service times that might depend on the waiting time of a customer and customers who are either patient or impatient. As opposed to the general approach in many queueing systems, we calculate the delay distribution without obtaining the steady state probabilities of the queue length. The trick, which was also applied in [14], is to keep track of the "age" of the customer in service, while remembering the D-MAP state immediately after the customer in service arrived. As far as we know, there is currently no method available to compute the delay distribution of a queueing system that combines impatient customers with either correlated arrivals and/or age dependent service times. The same can be said as far as queueing systems with patient customers that have age dependent service times is concerned.

Queueing systems of this type have obvious applications in manufacturing, service industries and telecommunications. For instance, in service industries, items that have been stored for a certain amount of time might require additional or alternative processing. In telecommunication systems, packets belonging to real time services become worthless to a receiver if they do not arrive before a certain deadline, therefore, these packets can be modeled as impatient customers. A study of the telephone system based on a queue with (im)patient customers was performed in [15].

The paper is structured as follows. Section 2 introduces the three queueing systems of interest. In Section 3 we develop an algorithm to compute the waiting time distribution of the system where all customers are impatient in the system (thus, even if they reach their critical age while in the server, they immediately leave the queue). Section 4 indicates the necessary changes to Section 3 if we consider impatient customers in the waiting room instead of the system, thus, once a customer enters the server, he is no longer impatient. Patient customers are considered in Section 5, while numerical examples, for each of the three systems, are presented in Section 6. In Section 7, we briefly describe how to generalize the methods presented in Section 3 to 5, if we are dealing with an arrival process that distinguishes different customer types.

## 2 The D-MAP/PH/1 queue with (im)patient customers and age-dependent service times

The arrival process of the queueing system of interest is a discrete time Markov arrival process, commonly known as the D-MAP process [2, 3], that does not allow batch arrivals; therefore, it is a subclass of the D-BMAP arrival process, which allows batch arrivals. Formally, a D-MAP is characterized—similar to its continuous time variant the MAP process [10]—by two $m \times m$ matrices $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$, where $m$ is a positive integer. The $(j_1, j_2)^{th}$ entry of the matrix $\boldsymbol{D}_1$ represents the probability that a customer arrives and the underlying Markov chain makes a transition from state $j_1$ to state $j_2$. The matrix $\boldsymbol{D}_0$ covers the case when there is no arrival. Thus, if the D-MAP is in state $j_1$ at time $n$, then, with probability $(D_x)_{j_1,j_2}$, we have $x$ arrival(s) at time $n$ and the state at time $n+1$ equals $j_2$. The matrix $\boldsymbol{D}$, defined as

$$\boldsymbol{D} = \boldsymbol{D}_0 + \boldsymbol{D}_1$$

represents the stochastic $m \times m$ transition matrix of the underlying Markov chain of the arrival process. Let $\boldsymbol{\theta}$ be the stationary probability vector of $\boldsymbol{D}$, that is, $\boldsymbol{\theta}\boldsymbol{D} = \boldsymbol{\theta}$ and $\boldsymbol{\theta}\boldsymbol{e} = 1$,

where $\boldsymbol{e}$ is a column vector with all entries equal to one. The stationary arrival rate is given by $\lambda = \boldsymbol{\theta}\boldsymbol{D}_1\boldsymbol{e}$.

The service time of a customer depends upon his waiting time $w$ and has a common phase-type distribution function [12] with a matrix representation $(m_w, \boldsymbol{\alpha}'_w, \boldsymbol{T}_w)$, where $m_w$ is a positive integer, $\boldsymbol{\alpha}'_w$ is an $1 \times m_w$ nonnegative stochastic vector and $\boldsymbol{T}_w$ is an $m_w \times m_w$ substochastic matrix. The $i^{th}$ component of the vector $\boldsymbol{\alpha}'_w$ is the probability that a customer, who waited $w$ time units before entering the service system, starts his service in phase $i$—that is, the phase of the service equals $i$ at time $n+1$, if we denote $n$ as the time that the customer entered the service system (and if we observe the phase at time $n+1$ just prior to a possible phase change at time $n+1$). If $\boldsymbol{T}^F_w = \boldsymbol{e} - \boldsymbol{T}_w\boldsymbol{e}$, then the $i^{th}$ entry of the vector $\boldsymbol{T}^F_w$ denotes the probability that such a customer—being one that waited $w$ time units—completes his service provided that he is in the $i^{th}$ phase at the current time instant. The $(i_1, i_2)^{th}$ entry of $\boldsymbol{T}_w$, on the other hand, is the probability that such a customer continues his service in phase $i_2$ at the next time instant provided that he is in phase $i_1$ at the current time instant. Notice, the minimum service time of a customer is one time unit[1]. The mean service time of a customer who waited $w$ time units in the waiting room is given by $1/\mu_w = \boldsymbol{\alpha}'_w(\boldsymbol{I} - \boldsymbol{T}_w)^{-1}\boldsymbol{e}$. We assume that for some large $v$, $(m_w, \boldsymbol{\alpha}'_w, \boldsymbol{T}_w)$ equals $(m_v, \boldsymbol{\alpha}'_v, \boldsymbol{T}_v)$ for $w > v$. Notice, this assumption is without loss of generality if we consider impatient customers (the first two situations) by choosing $v = a$, where $a$ denotes the critical age of impatient customers.

Before we proceed further, let us state the same thing in another way. The service time of a customer depends upon his waiting time $w$ and has a common phase-type distribution function with a matrix representation $(m_{ser}, \boldsymbol{\alpha}_w, \boldsymbol{T})$. Indeed, by setting $m_{ser} = \sum_{i=0}^{v} m_i$ and defining $\boldsymbol{\alpha}_w$ and $\boldsymbol{T}$ as

$$\begin{aligned} \boldsymbol{\alpha}_w &= [\mathbf{0}_{w,b}\boldsymbol{\alpha}'_w\mathbf{0}_{w,c}], \\ \boldsymbol{T} &= \begin{bmatrix} \boldsymbol{T}_0 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{T}_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{T}_v \end{bmatrix}, \end{aligned}$$

where $\mathbf{0}_{w,b}$, resp. $\mathbf{0}_{w,c}$, is a $1 \times \sum_{i=0}^{\min(w-1,v-1)} m_i$, resp. $1 \times \sum_{i=w+1}^{v} m_i$, vector filled with zeros, we see that both cases are equivalent. For many practical situations, one can often define a significantly smaller matrix $\boldsymbol{T}$ that is equivalent to the $\boldsymbol{T}$ matrix mentioned above. For example, if all the customers with a waiting time smaller than 100 time units have a deterministic service time of 4 time units and those with a longer waiting time (i.e., $w \geq 100$) are served in 2 time units, $v$ equals 100 and the $\boldsymbol{T}$ matrix defined above is a $402 \times 402$ matrix. However, it is sufficient to define $m_{ser} = 4$ and $\boldsymbol{T}$ as

$$\boldsymbol{T} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The vectors $\boldsymbol{\alpha}_w$ are equal to [1 0 0 0] for $w < 100$ and [0 0 1 0] for $w \geq 100$. It should be noted that such a reduction of $T$ is only possible in specific cases and not in the general case.

[1]The service time equals $l$ time units with probability $\boldsymbol{\alpha}'_w\boldsymbol{T}^{l-1}_w\boldsymbol{T}^F_w$

Finally, in the case of a simultaneous arrival and departure we assume that the departure occurs first. Also, if an arriving customer sees the server empty upon arrival, his service will start immediately (see [7]). In the forthcoming sections we describe the models using $(m_{ser}, \boldsymbol{\alpha}_w, \boldsymbol{T})$. Finally, we define $\boldsymbol{T}^F = \boldsymbol{T}\boldsymbol{e} - \boldsymbol{e}$ and $m_{tot} = m_{ser}m$. In each of the following three sections, we create a(n) (in)finite Markov chain (MC), indicate how to calculate the steady state vector of this MC and present a simple formula to find the delay distribution of a customer using the steady state vector. While constructing each of these MCs, we will always observe the system just prior to possible phase changes, arrivals or departures. Thus, if we refer to the system state at time $n$, such events happening at time $n$ are not yet taken into account by the system state.

## 3 Impatient Customers in the System

In this section, we consider the D-MAP/PH/1 queue with service times depending on the waiting time and impatient customers in the system. Thus, if a customer has spent a certain time, say $a$ time units, in the system (that is, waiting room and server) he immediately leaves the queue without starting/completing his service.

Consider a Markov chain (MC) with a finite number of states labeled $1, 2, \ldots, am_{tot} + m$. The set of states $\{1, \ldots, m\}$ is referred to as level zero of the MC, whereas the set of states $\{(i-1)m_{tot} + m + 1, \ldots, im_{tot} + m\}$ is referred to as level $i$ of the MC for $0 < i \leq a$. The states of level $i$, with $0 < i \leq a$, are labeled as $(s, j)$, where $1 \leq s \leq m_{ser}$ and $1 \leq j \leq m$. Let state $j$ of level zero of the MC correspond to the situation in which the queue and the server are empty, while the current state of the D-MAP is $j$. Let state $(s, j)$ of level $i$ of the MC correspond to the situation in which there is a customer in service, who arrived $i \leq a$ time units ago[2], while the service process is currently in phase $s$ and the D-MAP arrival process was in state $j$ at time $n - i + 1$, where $n$ is the current time instant. Recall, we observe the system just prior to possible phase changes, arrivals or departures.

The level of the Markov chain can never increase by more than one during a transition between time instant $n$ and $n + 1$. Let us explain. If the server is idle at time $n$, that is, the MC is at level zero at time $n$, we remain in level zero if the server remains idle, otherwise we make a transition to level one (because an arriving customer immediately enters the service system). For a busy server we have: either the customer in service, denoted as $c$, remains in the service system, thus, a transition is made from level $i$ to $i + 1$, or a new customer $c'$ enters the service system[3], the age at time $n$ of which can be at most one less than the age of customer $c$ at time $n$, meaning that the new level is at most $i$. As a result, the system can be described by a transition matrix $\boldsymbol{P}$ with the following structure:

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{B}_1 & \boldsymbol{B}_0 & \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{B}_2 & \boldsymbol{A}_1^1 & \boldsymbol{A}_0^1 & \boldsymbol{0} & \ldots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{B}_3 & \boldsymbol{A}_2^2 & \boldsymbol{A}_1^2 & \boldsymbol{A}_0^2 & \ldots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ \boldsymbol{B}_a & \boldsymbol{A}_{a-1}^{a-1} & \boldsymbol{A}_{a-1}^{a-1} & \boldsymbol{A}_{a-2}^{a-1} & \ldots & \boldsymbol{A}_1^{a-1} & \boldsymbol{A}_0^{a-1} \\ \boldsymbol{E} & \boldsymbol{C}_a & \boldsymbol{C}_{a-1} & \boldsymbol{C}_{a-2} & \ldots & \boldsymbol{C}_2 & \boldsymbol{C}_1 \end{bmatrix}, \tag{1}$$

[2]A customer with an age larger than $a$ can never be in the service system, because customers leave the system whenever their age equals $a$ (possibly making the server available for a new customer while leaving).

[3]The situation in which the server becomes idle corresponds to a transition to level zero.

where $\boldsymbol{A}_k^i$ and $\boldsymbol{C}_i$ are $m_{tot} \times m_{tot}$ matrices, $\boldsymbol{B}_i, i > 1$, and $\boldsymbol{E}$ are $m_{tot} \times m$ matrices, $\boldsymbol{B}_1$ is an $m \times m$ matrix and $\boldsymbol{B}_0$ is an $m \times m_{tot}$ matrix and hence $\boldsymbol{P}$ is a $(m + am_{tot}) \times (m + am_{tot})$ matrix.

Next, let us derive an expression for each of the matrices $\boldsymbol{A}_k^i$, $\boldsymbol{B}_i$, $\boldsymbol{C}_i$ and $\boldsymbol{E}$. Assume that the MC is in state $1 \leq j_1 \leq m$ at level zero at time $n$, that is, the D-MAP is in state $j_1$ at time $n$ and the server is idle. Then, anyone of two events can occur: (i) with probability $(\boldsymbol{D}_0)_{j_1,j_2}$, there is no arrival at time $n$ and the D-MAP makes a transition to state $j_2$. (ii) with probability $(\boldsymbol{\alpha}_0)_{s_2}(\boldsymbol{D}_1)_{j_1,j_2}$, a customer arrives at time $n$ (to the empty system) and starts his service in phase $s_2$ (for some $1 \leq s_2 \leq m_{ser}$), while the D-MAP is in state $j_2$ at time $n+1$. An arrival implies a transition to level one (because the customer has been in the service system for one time unit at time $n+1$), otherwise the MC remains at level zero, hence,

$$\boldsymbol{B}_1 = \boldsymbol{D}_0, \tag{2}$$

$$\boldsymbol{B}_0 = \boldsymbol{\alpha}_0 \otimes \boldsymbol{D}_1, \tag{3}$$

where $\otimes$ denotes the Kronecker product between matrices. Suppose that the MC is in state $(s_1, j_1)$ at level $i$, with $0 < i < a$, at time $n$. Then, we get a transition to level zero if the customer in service completes his service (with probability $(\boldsymbol{T}^F)_{s_1}$) and there is no arrival at time instant $n-i+1, n-i+2, \ldots, n$. Hence,

$$\boldsymbol{B}_{i+1} = \boldsymbol{T}^F \otimes \boldsymbol{D}_0^i, \tag{4}$$

for $0 < i < a$. A transition to state $(s_2, j_1)$ of level $i+1$ occurs if the customer remains in the service system (with probability $(\boldsymbol{T})_{s_1,s_2}$). Notice, in this case the state of the D-MAP remains the same, therefore,

$$\boldsymbol{A}_0^i = \boldsymbol{A}_0 = \boldsymbol{T} \otimes \boldsymbol{I}_m, \tag{5}$$

where $\boldsymbol{I}_m$ denotes the $m \times m$ unity matrix. Finally, a transition to level $i-l$, with $0 \leq l < i$, occurs if the customer in service completes his service (with probability $(\boldsymbol{T}^F)_{s_1}$) and there is no arrival until time $n+1-(i-l)$, that is, there is no arrival at time $n-i+1, \ldots, n-i+l$ and at time $n+1-i+l$ we have an arrival[4]. The age at time $n$ of the customer who arrived at time $n+1-i+l$ clearly equals $i-l-1$. The phase at the start of the service depends upon this age. Transitions from level $i$ to $i-l$ are governed by the $\boldsymbol{A}_{l+1}^i$, hence, for $1 \leq i < a$ and $0 \leq l < i$,

$$\boldsymbol{A}_{l+1}^i = \boldsymbol{T}^F \boldsymbol{\alpha}_{i-l-1} \otimes (\boldsymbol{D}_0^l \boldsymbol{D}_1), \tag{6}$$

for hereon, any matrix to the power 0 is taken to be the identity matrix of appropriate dimension. Next, consider the case where the MC is at level $a$ at time $n$. No matter whether the customer in service completes his service at time $n$, this customer leaves the system. Thus, as far as the transition probabilities are concerned, we could regard level $a$ as if the service completion probability vector $\boldsymbol{T}^F$ is identical to $\boldsymbol{e}$. As a result, a transition is made to level $a-i$, for $0 \leq i < a$, if there are no arrivals until time $n+1-(a-i)$, that is, there is no arrival at time $n-a+1, \ldots, n-a+i$ and at time $n+1-a+i$ we have an arrival. The

[4]Indeed, we are at level $i-l$ at time $n+1$ if the customer in service has an age $i-l$ at time $n+1$, thus, he arrived at time $n+1-i+l$.

age at time $n$ of the arriving customer is obviously equal to $a-i-1$. The phase at the start of the service depends upon this age. Transitions from level $a$ to $a-i$ are governed by the $\boldsymbol{C}_{i+1}$ matrix, hence, for $0 \leq i < a$,

$$\boldsymbol{C}_{i+1} = \boldsymbol{e}\boldsymbol{\alpha}_{a-i-1} \otimes (\boldsymbol{D}_0^i \boldsymbol{D}_1). \tag{7}$$

Finally, in order to get from level $a$ to level zero, there should not be any arrival at time $n-a+1, n-a+2, \ldots, n$, therefore,

$$\boldsymbol{E} = \boldsymbol{e} \otimes \boldsymbol{D}_0^a. \tag{8}$$

This concludes the description of the transition matrices.

The steady state vector $\boldsymbol{\pi}$ of $\boldsymbol{P}$, that is, $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$ and $\boldsymbol{\pi}\boldsymbol{e} = 1$, is found using the Latouche-Jacobs-Gaver (LJG) algorithm [9]. This algorithm calculates the solution of $\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}$, where $\boldsymbol{Q}$ is an infinitesimal generator matrix for a bidimensional Markov process with a lower block-Hessenberg form. By rewriting $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$ as $\boldsymbol{\pi}(\boldsymbol{P} - \boldsymbol{I}) = \boldsymbol{0}$, we see that ($\boldsymbol{P}$ - $\boldsymbol{I}$) is such a generator matrix. Applying the LJG algorithm results in the following algorithm, the time and memory complexity of which are $O(m_{tot}^3 a^2)$ and $O(m_{tot}^2 a)$:

**Algorithm:**

- INPUT: the matrices $\boldsymbol{D}_0$, $\boldsymbol{D}_1$, the phase-type distribution function characterized by $(m_{ser}, \boldsymbol{\alpha}_w, \boldsymbol{T})$ and the critical age $a$.
- STEP 1a: Calculate the matrices $\boldsymbol{C}_i$ and $\boldsymbol{E}$ by means of equations (7) and (8). Define $\boldsymbol{G}_a = \boldsymbol{C}_1 - \boldsymbol{I}$, $\boldsymbol{\Pi}_{a,0} = -(\boldsymbol{G}_a)^{-1}\boldsymbol{E}$ and for $k = 1, \ldots, a-1$, $\boldsymbol{\Pi}_{a,k} = -(\boldsymbol{G}_a)^{-1}\boldsymbol{C}_{a-k+1}$.
- STEP 1b: Calculate the matrix $\boldsymbol{A}_0$ by means of equation (5). For $s = a-1, a-2, \ldots, 1$, calculate the matrices $\boldsymbol{B}_{s+1}$ and $\boldsymbol{A}_k^s$ by means of equations (4) and (6) and define $\boldsymbol{G}_s = \boldsymbol{A}_1^s - \boldsymbol{I} + \boldsymbol{A}_0\boldsymbol{\Pi}_{s+1,s}$, $\boldsymbol{\Pi}_{s,0} = -(\boldsymbol{G}_s)^{-1}\left[\boldsymbol{B}_{s+1} + \boldsymbol{A}_0\boldsymbol{\Pi}_{s+1,0}\right]$ and for $k = 1, \ldots, s-1$, $\boldsymbol{\Pi}_{s,k} = -(\boldsymbol{G}_s)^{-1}\left[\boldsymbol{A}_{s-k+1}^s + \boldsymbol{A}_0\boldsymbol{\Pi}_{s+1,k}\right]$.
- STEP 1c: Calculate the matrices $\boldsymbol{B}_0$ and $\boldsymbol{B}_1$ by means of equations (2) and (3). Define $\boldsymbol{G}_0 = \boldsymbol{B}_1 - \boldsymbol{I} + \boldsymbol{B}_0\boldsymbol{\Pi}_{1,0}$.
- STEP 2: Calculate de steady state vector $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_a)$ by $\boldsymbol{\pi}_0\boldsymbol{G}_0 = \boldsymbol{0}$, $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_0\boldsymbol{B}_0(-\boldsymbol{G}_1)^{-1}$ and for $s = 2, \ldots, a$, $\boldsymbol{\pi}_s = \boldsymbol{\pi}_{s-1}\boldsymbol{A}_0(-\boldsymbol{G}_s)^{-1}$.

The implementation of this algorithm requires the storage of a single $m_{tot} \times (m + m_{tot}a)$ matrix $\boldsymbol{W}$ using the following method. First, during step 1a, we store $[\boldsymbol{\Pi}_{a,0}\ \boldsymbol{\Pi}_{a,1}\ \ldots\ \boldsymbol{\Pi}_{a,a-1}\ (-\boldsymbol{G}_a)^{-1}]$ in $\boldsymbol{W}$. During step 1b, we overwrite $\boldsymbol{\Pi}_{s+1,s}$ by$(-\boldsymbol{G}_s)^{-1}$ and $\boldsymbol{\Pi}_{s+1,k}$ by $\boldsymbol{\Pi}_{s,k}$. Finally, during step 1c, we store $\boldsymbol{G}_0$ in $\boldsymbol{\Pi}_{1,0}$[5]. Thus, at the start of step 2 we have $\boldsymbol{W} = [\boldsymbol{G}_0\ (-\boldsymbol{G}_1)^{-1}\ (-\boldsymbol{G}_2)^{-1}\ \ldots\ (-\boldsymbol{G}_a)^{-1}]$, which suffices to obtain $\boldsymbol{\pi}$ in step 2. Other algorithms, having the same time and memory complexity as the LJG algorithm, to obtain $\boldsymbol{\pi}$ can be found in the literature [4, 8]. Morover, the LJG algorithm is, as far as we know, as efficient as any known algorithm to solve finite level independent MCs of the $GI/M/1$ type. Such a MC has a transition matrix similar to $\boldsymbol{P}$ but with $\boldsymbol{A}_k^i = \boldsymbol{A}_k^{a-1}$ for any $k$ and $i$. For instance, if we consider impatient customers with a service time that is independent of the waiting time $w$, that is, $\boldsymbol{\alpha}_w = \boldsymbol{\alpha}_0$ for any $w$, we obtain such an MC.

[5]The matrix $\boldsymbol{G}_0$ is a $m \times m$ matrix, thus, it requires only the first $m$ rows of $\boldsymbol{W}$.

Denote $P[X = i]$ as the probability that a customer completes his service $i$ time units after entering the system. Then, by noticing that this probability equals the expected number of customers who complete their service at age $i$ at an arbitrary time instant, divided by the expected number of customers who leave the system at an arbitrary time instant, we have

$$P[X = i] = \frac{1}{\lambda} \sum_{s=1}^{m_{ser}} (\boldsymbol{T}^F)_s \sum_{j=1}^{m} (\boldsymbol{\pi}_i)_{(s,j)}, \tag{9}$$

for $0 < i \leq a$. $P_{out} = 1 - \sum_{i=1}^{a} P[X = i]$ equals the probability that a customer leaves the system before starting/completing his service.

## 4 Impatient Customers in the Waiting Room

This queueing system is same as the one discussed in Section 3 except for the following: If a customer enters the service system before reaching the critical age $a$, he remains in the service system until his service is completed. Thus, the customers are no longer impatient once they are being served. If we define the same MC as in the previous section for this queueing system, we get a transition probability matrix $\boldsymbol{P'}$ that is either finite or infinite depending on whether the service time distribution of a customer is bounded. For unbounded service times we get an infinite $\boldsymbol{P'}$ matrix, because the age of the customer in service is unbounded. For bounded service times, where we denote $s_{max}$ as the maximum service time, we get a $\boldsymbol{P'}$ matrix of dimension $m + m_{tot}(a + s_{max} - 1)$, because the maximum age of a customer in service is $a + s_{max} - 1$ (a customer enters the service system with an age of at most $a - 1$).

In order to find $\boldsymbol{\pi'}$, the steady state vector of $\boldsymbol{P'}$, we define a new MC with a transition matrix $\tilde{\boldsymbol{P}}$ of dimension $m + m_{tot}a$ as follows. Instead of observing the MC at each time instance $n$, we only observe the chain when the server is either occupied by a customer of age *at most* $a$ or it is idle. Therefore, the transition probability matrix $\tilde{\boldsymbol{P}}$ is identical to $\boldsymbol{P}$ except for the last $m_{tot}$ rows – the rows corresponding to level $a$. Thus, $\tilde{\boldsymbol{P}}$ can be written as

$$\tilde{\boldsymbol{P}} = \begin{bmatrix} \boldsymbol{B}_1 & \boldsymbol{B}_0 & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{B}_2 & \boldsymbol{A}_1^1 & \boldsymbol{A}_0^1 & \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{B}_3 & \boldsymbol{A}_2^2 & \boldsymbol{A}_1^2 & \boldsymbol{A}_0^2 & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ \boldsymbol{B}_a & \boldsymbol{A}_{a-1}^{a-1} & \boldsymbol{A}_{a-1}^{a-1} & \boldsymbol{A}_{a-2}^{a-1} & \dots & \boldsymbol{A}_1^{a-1} & \boldsymbol{A}_0^{a-1} \\ \tilde{\boldsymbol{E}} & \tilde{\boldsymbol{C}}_a & \tilde{\boldsymbol{C}}_{a-1} & \tilde{\boldsymbol{C}}_{a-2} & \dots & \tilde{\boldsymbol{C}}_2 & \tilde{\boldsymbol{C}}_1 \end{bmatrix}. \tag{10}$$

Let us derive an expression for $\tilde{\boldsymbol{E}}$ and $\tilde{\boldsymbol{C}}_i$ for $i = 1, \dots, a$. First, consider a transition from state $(s_1, j_1)$ at level $a$ to state $j_2$ at level 0. Denote by $c$ the customer of age $a$ who is in the server at time $n$. This customer will remain in the server for $k$ more time units (with probability $(\boldsymbol{T}^k \boldsymbol{T}^F)_{s_1}$, recall we observe the system just prior to the possible phase changes and departure epochs) for some $k \geq 0$. Thus, customer $c$ leaves the system at time $n + k$. Therefore, the next time instant that we observe is time $n + k + 1$. At time instant $n-a+1, \dots, n-a+k$ either there is an arrival or no arrival—if there is an arrival at time $n-a+l$ for some $l = 1, \dots, k$, the arrived customer leaves the waiting room at the time instant $n+l$ as his age is $a$ by then—and there is no arrival at the time $n - a + k + 1, n - a + k + 2, \dots, n + k$ (a customer arriving at any of these time instants would have an age of at most $a - 1$ at time $n + k$ and therefore would enter the service system at time $n + k$, which cannot be if we

want to make a transition to level zero). This event happens with a probability $(\boldsymbol{D}^k\boldsymbol{D}_0^a)_{j_1,j_2}$. Hence, the matrix $\tilde{\boldsymbol{E}}$ corresponding to the transition from level $a$ to level 0 is given by

$$\tilde{\boldsymbol{E}} = \sum_{k=0}^{\infty}(\boldsymbol{T}^k\boldsymbol{T}^F)\otimes(\boldsymbol{D}^k\boldsymbol{D}_0^a) = \left(\sum_{k=0}^{\infty}(\boldsymbol{T}^k\boldsymbol{T}^F)\otimes\boldsymbol{D}^k\right)\boldsymbol{D}_0^a. \tag{11}$$

Now, consider the transition from state $(s_1, j_1)$ at level $a$ to state $(s_2, j_2)$ at level $a-i$, for $0 \le i < a$. Again, denote by $c$ the customer of age $a$ who is in service at time $n$. This customer will remain in the service system for $k$ more time units for some $k \ge 0$. Thus, customer $c$ leaves the system at time $n+k$. Therefore, the next time instant that we observe is time $n+k+1$. At time instant $n-a+1,\ldots,n-a+k$ either there is an arrival or no arrival, there is no arrival at time $n-a+k+1,\ldots,n-a+k+i$ and an arrival occurs at time $n-a+k+i+1$. The age at time $n+k$ of this arrived customer is clearly $a-i-1$, therefore, the vector $\boldsymbol{\alpha}_{a-i-1}$ determines $s_2$. The above mentioned event happens with a probability $(\boldsymbol{D}^k\boldsymbol{D}_0^i\boldsymbol{D}_1)_{j_1,j_2}$. As a result, transition from level $a$ to $a-i$, for $0 \le i < a$, governed by the matrix $\tilde{\boldsymbol{C}}_{i+1}$, is found as

$$\tilde{\boldsymbol{C}}_{i+1} = \sum_{k=0}^{\infty}(\boldsymbol{T}^k\boldsymbol{T}^F)\otimes(\boldsymbol{\alpha}_{a-i-1}\otimes(\boldsymbol{D}^k\boldsymbol{D}_0^i\boldsymbol{D}_1)) = \left(\sum_{k=0}^{\infty}(\boldsymbol{T}^k\boldsymbol{T}^F)\otimes\boldsymbol{D}^k\right)(\boldsymbol{\alpha}_{a-i-1}\otimes\boldsymbol{D}_0^i\boldsymbol{D}_1). \tag{12}$$

The sum occuring in both equation (11) and (12) is finite if the service time distribution is bounded, otherwise we approximate it by the first $k'$ terms if $\sum_{k\ge k'}^{\infty}(\boldsymbol{T}^k\boldsymbol{T}^F)\otimes\boldsymbol{D}^k < \epsilon$, for some $\epsilon$ small, e.g., $10^{-20}$. Such a $k'$ exists because $\boldsymbol{T}$ is a substochastic matrix.

The steady state vector of $\tilde{\boldsymbol{P}}$ is found by the LJG algorithm of Section 3 if we replace $\boldsymbol{E}$ and $\boldsymbol{C}_i$ by $\tilde{\boldsymbol{E}}$ and $\tilde{\boldsymbol{C}}_i$. Having found $\tilde{\boldsymbol{\pi}} = [\tilde{\boldsymbol{\pi}}_0\tilde{\boldsymbol{\pi}}_1\ldots\tilde{\boldsymbol{\pi}}_a]$, the steady state vector of $\tilde{\boldsymbol{P}}$, we obtain $\boldsymbol{\pi}' = [\boldsymbol{\pi}'_0\boldsymbol{\pi}'_1\ldots]$, the steady state vector of $\boldsymbol{P}'$, as $\boldsymbol{\pi}'_i = \tilde{\boldsymbol{\pi}}_i/c$, for $0 \le i \le a$, and

$$\boldsymbol{\pi}'_{a+i} = \tilde{\boldsymbol{\pi}}_a(\boldsymbol{T}^i\otimes\boldsymbol{I})/c, \tag{13}$$

where $i > 0$, $\boldsymbol{I}$ is the unity matrix of dimension $m$ and $c = \sum_{i=0}^{a-1}\tilde{\boldsymbol{\pi}}_i\boldsymbol{e} + \tilde{\boldsymbol{\pi}}_a((\boldsymbol{I}-\boldsymbol{T})^{-1}\otimes\boldsymbol{I})\boldsymbol{e}$ is the normalization factor. Notice, if the service time distribution is bounded, $\boldsymbol{T}^i$ becomes zero for $i$ sufficiently large, otherwise it decreases exponentially to zero.

Denote $P[X'=i]$ as the probability that a customer completes his service $i$ time units after entering the system. Then, similar to Section 3, we have

$$P[X'=i] = \frac{1}{\lambda}\sum_{s=1}^{m_{ser}}(\boldsymbol{T}^F)_s\sum_{j=1}^{m}(\boldsymbol{\pi}'_i)_{(s,j)}, \tag{14}$$

for $i > 0$. $P'_{out} = 1 - \sum_{i>0}P[X'=i]$ equals the probability that a customer leaves the system before starting his service (once started, a customer always completes its service).

## 5 Patient Customers in the System

As opposed to the previous two sections, all customers are now patient, that is, they wait in the waiting room until the server becomes available and do not leave the server until their service is completed. We consider the same MC as in Section 3. Since there is no upper

bound on the time that a customer spends in the system, the transition probability matrix $\bar{\boldsymbol{P}}$ of this MC is an infinite matrix with the following structure:

$$\bar{\boldsymbol{P}} = \begin{bmatrix} \boldsymbol{B}_1 & \boldsymbol{B}_0 & \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots \\ \boldsymbol{B}_2 & \boldsymbol{A}_1^1 & \boldsymbol{A}_0^1 & \dots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots \\ \boldsymbol{B}_3 & \boldsymbol{A}_2^2 & \boldsymbol{A}_1^2 & \dots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots \\ \boldsymbol{B}_v & \boldsymbol{A}_{v-1}^{v-1} & \boldsymbol{A}_{v-2}^{v-1} & \dots & \boldsymbol{A}_1^{v-1} & \boldsymbol{A}_0^{v-1} & \boldsymbol{0} & \boldsymbol{0} & \dots \\ \boldsymbol{B}_{v+1} & \boldsymbol{A}_v^v & \boldsymbol{A}_{v-1}^v & \dots & \boldsymbol{A}_2^v & \boldsymbol{A}_1^v & \boldsymbol{A}_0 & \boldsymbol{0} & \ddots \\ \boldsymbol{B}_{v+2} & \boldsymbol{A}_{v+1}^{v+1} & \boldsymbol{A}_v^{v+1} & \dots & \boldsymbol{A}_3^{v+1} & \boldsymbol{A}_2^{v+1} & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \ddots \\ \boldsymbol{B}_{v+3} & \boldsymbol{A}_{v+2}^{v+2} & \boldsymbol{A}_{v+1}^{v+2} & \dots & \boldsymbol{A}_4^{v+2} & \boldsymbol{A}_3^{v+2} & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \ddots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & & \end{bmatrix}, \tag{15}$$

where the matrices $\boldsymbol{A}_k^i$ and $\boldsymbol{B}_i$ were defined in Section 3. Remember, in Section 2 we assumed that the service time distribution becomes identical for all customers with an age of $v$ or higher (for some $v$), that is, the vectors $\boldsymbol{\alpha}_{v+i} = \boldsymbol{\alpha}_v$, for $i > 0$. Therefore, looking at Equation 6, the matrices $\boldsymbol{A}_k^i$, for $i - k \geq v$, can be written as $\boldsymbol{A}_k$ because the vector $\boldsymbol{\alpha}_{i-k}$ is equal to $\boldsymbol{\alpha}_v$.

$\bar{\boldsymbol{\pi}}$, the steady state vector of $\bar{\boldsymbol{P}}$, can be calculated as follows. We reblock $\bar{\boldsymbol{P}}$ by gathering level $0, 1, \dots, v$ into a single level, thus,

$$\bar{\boldsymbol{P}} = \begin{bmatrix} \boldsymbol{F}_1 & \boldsymbol{F}_2 & \boldsymbol{0} & \dots \\ \boldsymbol{H}_1 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \dots \\ \boldsymbol{H}_2 & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \ddots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix}, \tag{16}$$

where $\boldsymbol{F}_1$ is the $(m + vm_{tot}) \times (m + vm_{tot})$ matrix in the upper left corner of Equation (15), $\boldsymbol{F}_2$ is a $(m + vm_{tot}) \times m_{tot}$ matrix with all its entries equal to zero, except for the last $m_{tot}$ rows which equal $\boldsymbol{A}_0$. Finally, $\boldsymbol{H}_i$, for $i > 0$, is given by

$$\boldsymbol{H}_i = \begin{bmatrix} \boldsymbol{B}_{v+i+1} & \boldsymbol{A}_{v+i}^{v+i} & \boldsymbol{A}_{v+i-1}^{v+i} & \dots & \boldsymbol{A}_{i+1}^{v+i} \end{bmatrix}. \tag{17}$$

Thus, $\bar{\boldsymbol{P}}$ can be seen as an infinite GI/M/1 Type MC with a generalized initial condition [12]. Due to [14, Theorem 1], the steady state vector $\bar{\boldsymbol{\pi}}$ of $\bar{\boldsymbol{P}}$ exists if and only if $\lambda/\mu_v < 1$, where $\lambda$ and $\mu_v$ where defined in Section 2. Moreover, $\bar{\boldsymbol{\pi}}_{v+1+i} = \bar{\boldsymbol{\pi}}_{v+1}\boldsymbol{R}^i$, for $i > 0$, where $\boldsymbol{R}$, an $m_{tot} \times m_{tot}$ matrix, is the smallest nonnegative solution of the following equation:

$$\boldsymbol{R} = \sum_{i=0}^{\infty} \boldsymbol{R}^i \boldsymbol{A}_i. \tag{18}$$

This equation can be solved by means of an iterative scheme [11, 13, 1]. The first $v + 2$ components $[\bar{\boldsymbol{\pi}}_0 \dots \bar{\boldsymbol{\pi}}_{v+1}]$ of the vector $\bar{\boldsymbol{\pi}}$ are then found by solving the boundary condition

$$[\bar{\boldsymbol{\pi}}_0 \dots \bar{\boldsymbol{\pi}}_{v+1}] = [\bar{\boldsymbol{\pi}}_0 \dots \bar{\boldsymbol{\pi}}_{v+1}] \begin{bmatrix} \boldsymbol{F}_1 & \boldsymbol{F}_2 \\ \bar{\boldsymbol{H}} & \bar{\boldsymbol{A}} \end{bmatrix}, \tag{19}$$

where $\bar{\boldsymbol{H}} = \sum_{l \geq 1} \boldsymbol{R}^{l-1}\boldsymbol{H}_l$ and $\bar{\boldsymbol{A}} = \sum_{l \geq 1} \boldsymbol{R}^{l-1}\boldsymbol{A}_l$ (see [12, 11]). The matrix appearing in Equation (19) has the same structure as $\bar{\boldsymbol{P}}$, thus, we can use the LJG algorithm to solve the

boundary condition. The vector $[\bar{\boldsymbol{\pi}}_0 \dots \bar{\boldsymbol{\pi}}_{v+1}]$ is normalized as

$$\bar{\boldsymbol{\pi}}_0 \boldsymbol{e} + \sum_{i=1}^{v} \bar{\boldsymbol{\pi}}_i \boldsymbol{e} + \bar{\boldsymbol{\pi}}_{v+1} (\boldsymbol{I} - \boldsymbol{R})^{-1} \boldsymbol{e} = 1. \tag{20}$$

Denote $P[\bar{X} = i]$ as the probability that a customer completes his service $i$ time units after entering the system. Then, similar to Sections 3 and 4, we have

$$P[\bar{X} = i] = \frac{1}{\lambda} \sum_{s=1}^{m_{ser}} (\boldsymbol{T}^F)_s \sum_{j=1}^{m} (\bar{\boldsymbol{\pi}}_i)_{(s,j)}, \tag{21}$$

for $i > 0$.

## 6 Numerical Examples

A fairly arbitrary example that demonstrates the strength of our approach is presented in this section. We consider a D-MAP arrival process with two states. While in state one, resp. two, an arrival occurs at the current time instant with a probability 0.1, resp. 0.4. The state of the D-MAP changes with a probability of 0.01. Thus,

$$\boldsymbol{D}_0 = \begin{bmatrix} 0.891 & 0.009 \\ 0.006 & 0.594 \end{bmatrix}, \quad \boldsymbol{D}_1 = \begin{bmatrix} 0.099 & 0.001 \\ 0.004 & 0.396 \end{bmatrix}. \tag{22}$$

The service time of a customer is as follows. If the customer waited $w$ time units, with $w \leq 100$, then his service time equals either two or three time units, each with a probability 0.5. If his waiting time $w$ is more than 100, he might need some additional processing, that is, with a probability 0.2, 0.05 and 0.005 he requires an additional geometrically distributed service time with a mean $1.5, 5$ and 50, respectively. Hence,

$$\boldsymbol{T} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.05 & 0.005 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4/5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 49/50 \end{bmatrix}, \tag{23}$$

the vectors $\boldsymbol{\alpha}_w = [0.5\ 0.5\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$, for $w \leq 100$ and $\boldsymbol{\alpha}_w = [0\ 0\ 0\ 0.5\ 0.5\ 0\ 0\ 0\ 0]$, for $w > 100$. We consider the following three systems, requiring the algorithms developed in Section 3, 4 and 5: impatient customers in the system, impatient customers in the waiting room and patient customers. The critical age $a$ for the first two systems varies between 10 and 500. In case of patient customers it suffices to set $v = 101$.

Figure 1 presents the waiting time distribution for the following cases: impatient customers in the system with the critical age $a$ equal to $50, 200, 300, 400$ and 500 (full lines labeled as S($a$)), impatient customers in the waiting room for the same 5 values for $a$ (dotted lines labeled W($a$)) and patient customers (dashed line labeled P). All the curves more or less

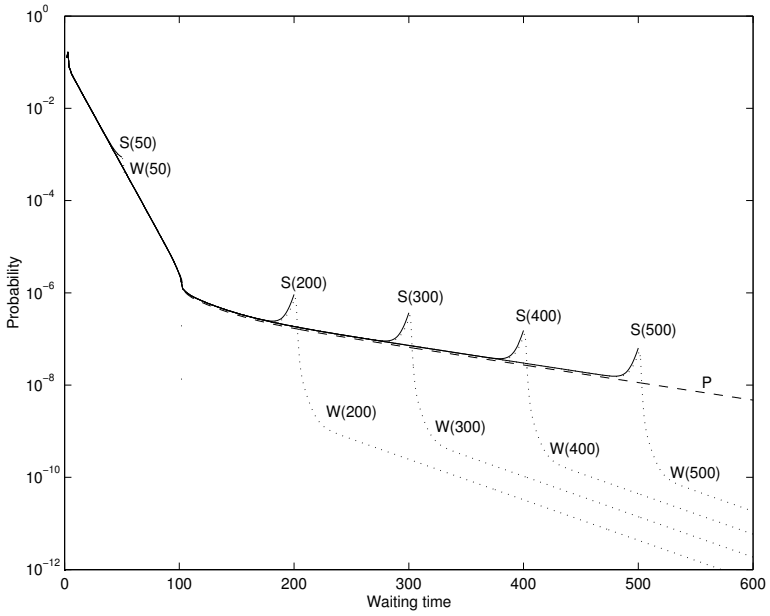


**Figure 1:** *Delay distribution of (im)patient customers*

coincide with the patient customer case as long as the waiting time $w$ is well below the critical age $a$. When $w$ approaches $a$ we get an increase, this is caused by the fact that the system is overloaded during the time periods in which the D-MAP resides in state 2. The W($a$) curves are, as opposed to the S($a$) curves, infinite, because there is no upper bound on the service time of a customer (except for $a = 50$). The difference between W($a$) and S($a$) seems very small, nevertheless the number of customers that leave the system without starting/completing service differs significantly (see Figure 2). Indeed, for $a \leq 100$, it is much more efficient to finish the service of the customer in progress (because the service time is nearly deterministic), whereas for $a > 100$, we get the reverse effect, because the service time of those customers has a large variation (thus, it is better to drop the customers who require a large service time).

The computation times for the curves in Figure 1 are below one minute, e.g., for S(500) we needed 58.6371 seconds using an AMD K7 ATHLON 1.4 GHz processor with 512 Mb of memory, the system with patient customers required 42.3194 seconds. In general, the computation time can be further reduced by removing the possible transient states from the MCs of interest. For instance, in the example above, the states $(s, j)$, for $s > 3$, at level $i$, for $i \leq 100$, are transient (because a customer entering the server with an age below 100 can never be in phase 4 to 9). Similarly, the states $(s, j)$, for $s \leq 3$, at level $i$, for $i > 103$, are also transient. Removing the transient states from the transition matrix reduces its dimension, while maintaining its structure (one simply eliminates the rows and columns that correspond to these states). It is, however, not necessary to remove these states, because their corresponding entries in the steady state vector will equal zero when applying the LJG algorithm. In some particular cases removing transient states could result in a dramatic reduction of the state space. For instance, assume that the interarrival times of the customers are independent and identically distributed according to some distribution $G$. Denote $g_i$, for $i \geq 1$, as the probability that the interarrival time equals $i$. Thus, this GI arrival process is

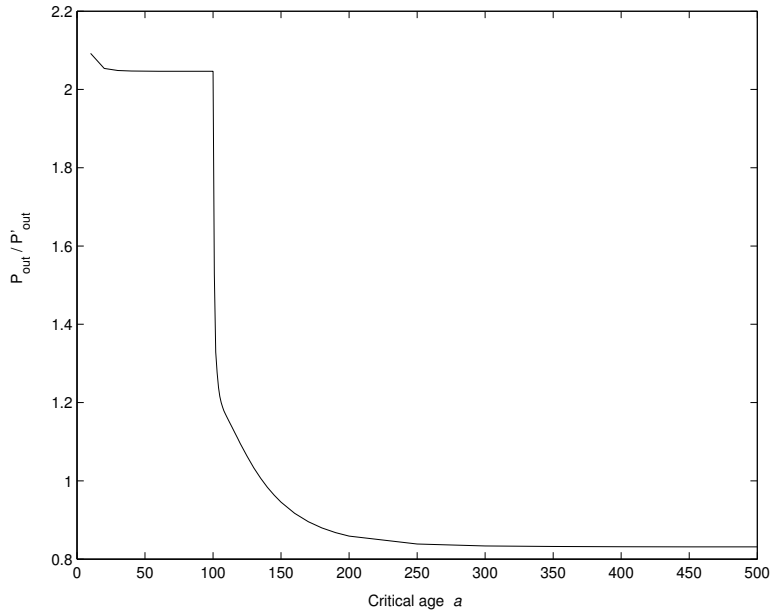


**Figure 2:** *Ratio of lost impatient customers customers in the system over the waiting room* ($P_{out}/P'_{out}$)

described by a D-MAP as follows:

$$\boldsymbol{D}_0 = \begin{bmatrix} 0 & g_2 & g_3 & g_4 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \boldsymbol{D}_1 = \begin{bmatrix} g_1 & 0 & 0 & 0 & \dots \\ 1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{24}$$

This D-MAP is always in the first state after an arrival occured, therefore, all the states $(s, j)$ at level $i$, for $i > 0$, are transient if $j \neq 1$ (because we remember the state of the D-MAP immediately after an arrival occured). Thus, for this particular arrival process, the transition matrices, e.g., $\boldsymbol{A}_k^i$, used in Section 3 to 5, can be reduced to $m_{ser} \times m_{ser}$ matrices (i.e., it seems as if the D-MAP has only one state). In general, if the $j_1$-th entry of the vector $\boldsymbol{e}\boldsymbol{D}_1$ equals zero, the states $(s, j_1)$ at level $i > 0$ are transient for any $s$.

## 7 Extentions to other Queueing Systems

It is possible to combine the techniques presented in Sections 3 to 5 with those used in [14], to obtain the delay distribution of (im)patient customers in a discrete time MMAP[K]/ PH[K]/1 queue with age dependent service times. The discrete time Markov arrival process with marked transitions (MMAP[K], see [6, 5]) distinguishes customers into $K$ different types. The MMAP[K] is sometimes defined such that it allows batch arrivals to occur, but we do not consider batch arrivals. An MMAP[K] that does not allow batches to occur, is characterized

by a set[6] of $m \times m$ matrices $\{\boldsymbol{D}_k \mid 0 \leq k \leq K\}$, with $m$ a positive integer. The $(j_1, j_2)^{th}$ entry of the matrix $\boldsymbol{D}_k$, for $k > 0$, represents the probability that a customer of type $k$ arrives and the underlying Markov chain makes a transition from state $j_1$ to state $j_2$. The matrix $\boldsymbol{D}_0$ covers the case when there are no arrivals. Similar to Section 2, the matrix $\boldsymbol{D}$, defined as $\boldsymbol{D} = \sum_{k=0}^{K} \boldsymbol{D}_k$, represents the stochastic $m \times m$ transition matrix of the underlying Markov chain of the arrival process. Let $\boldsymbol{\theta}$ be the stationary probability vector of $\boldsymbol{D}$, that is, $\boldsymbol{\theta}\boldsymbol{D} = \boldsymbol{\theta}$ and $\boldsymbol{\theta}\boldsymbol{e} = 1$, where $\boldsymbol{e}$ is a column vector with all entries equal to one. The stationary arrival rate of type $k$ customers is given by $\lambda_k = \boldsymbol{\theta}\boldsymbol{D}_k\boldsymbol{e}$.

The service time of a type $k$ customer who waited $w$ time units in the waiting room of such a queue has a common phase-type distribution function with a matrix representation $(m_k, \boldsymbol{\alpha}_{k,w}, \boldsymbol{T}_k)$, where $m_k$ is a positive integer, $\boldsymbol{\alpha}_{k,w}$ is an $1 \times m_k$ nonnegative stochastic vector and $\boldsymbol{T}_k$ is an $m_k \times m_k$ substochastic matrix. The mean service time of a type $k$ customer who waited $w$ time units in the waiting room, equals $1/\mu_{k,w} = \boldsymbol{\alpha}_{k,w}(\boldsymbol{I} - \boldsymbol{T}_k)^{-1}\boldsymbol{e}$. Thus, the service time of a customer depends on his type and age. Define $m_{tot}^K$ as $m \sum_{k=1}^{K} m_k$.

The following systems can be solved by extending the methods discussed in Sections 3 to 5 in combination with [14]. Firstly, assume that all customers are patient and that $\boldsymbol{\alpha}_{k,w} = \boldsymbol{\alpha}_{k,v}$, for $w > v$ for some $v$ large. Thus, the service times become age independent from a certain age, but still depend upon the customer type. We can obtain the delay distribution using methods similar to Section 5. In order to do so, we have to add the type of the customer in service to the couple $(s, j)$. Thus, the states of level $i$, for $i > 0$, are now labeled as $(k, s, j)$, where $1 \leq k \leq K$, $1 \leq s \leq m_k$ and $1 \leq j \leq m$. The states of level zero remain the same. The $m_{tot}^K \times m_{tot}^K$ matrices $\boldsymbol{A}_i^k$ are found by replacing the $\boldsymbol{\alpha}_k$ vectors in [14, Section 2.1] by the appropriate $\boldsymbol{\alpha}_{k,w}$ vectors. The $\boldsymbol{B}_i$ are identical to those in [14, Section 2.1]. Secondly, assume that all the customers are impatient, that is, they leave the system, resp. waiting room, when they reach the critical age $a$. Notice, $a$ does not depend on the type of a customer. These two cases, i.e., impatient in the system or waiting room, can be treated similar to Sections 3 and 4 after adding the type of the customer $k$ to the couple $(s, j)$. Thus, the transition matrices $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ are now $(am_{tot}^K + m) \times (am_{tot}^K + m)$ matrices. Finally, it is also possible to consider a system with impatient customers where the critical age $a$ depends upon the type $k$ of a customer or to combine patient with impatient customers (where some types are patient and others are impatient with a critical age $a_k$). The MCs required to obtain the delay distribution of these queueing systems differ significantly from the ones presented in this paper, but are nevertheless based on the same ideas.

## Acknowledgements

The first author is a postdoctoral fellow of the FWO-Flanders. The second and third author were supported by the the FWO-Flanders project G.0315.01 entitled *Computational Methods for performance evaluation and simulation of complex technical systems.*

## References

[1] A.S. Alfa, B. Sengupta, T. Takine, and J. Xue. A new algorithm for computing the rate matrix of GI/M/1 type Markov chains. In *Proc. of the 4th Int. Conf. on Matrix Analytic*

[6] As opposed to the D-MAP, which is characterized by two matrices $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$. For $K = 1$, a MMAP[K] reduces to a D-MAP, hence, an MMAP[K] is a generalization of the D-MAP.

*Methods*, pages 1–16, Adelaide, Australia, 2002.

[2] C. Blondia. A discrete-time batch markovian arrival process as B-ISDN traffic model. *Belgian Journal of Operations Research, Statistics and Computer Science*, 32(3,4), 1993.

[3] C. Blondia and O. Casals. Statistical multiplexing of VBR sources: A matrix-analytical approach. *Performance Evaluation*, 16:5–20, 1992.

[4] J.Y. Le Boudec. An efficient solution method for Markov models of ATM links with loss priorities. *IEEE JSAC*, 9(3), April 1991.

[5] Q. He. The versatility of the MMAP[K] and the MMAP[K]/G[K]/1 queue. *Queueing Systems*, 38:397–418, 2001.

[6] Q. He and M.F. Neuts. Markov chains with marked transitions. *Stochastic Processes and their Applications*, 74:37–52, 1998.

[7] J.J. Hunter. *Mathematical Techniques of Applied Probability, vol II, discrete time models: techniques and applications*. Academic Press, New York, 1983.

[8] A.E. Kamal. Efficient solution of multiple server queues with applications to the modeling of ATM concentrators. In *Proc. of IEEE Infocom*, pages 248–254, San Francisco, CA, 1996.

[9] G. Latouche, P.A. Jacobs, and D.P. Gaver. Finite markov chain models skip-free in one direction. *Naval Research Logistics Quarterly*, 31:571–588, 1984.

[10] D.M. Lucantoni, K.S. Meier-Hellstern, and M.F. Neuts. A single server queue with server vacations and a class of non-renewal arrival processes. *Adv. Appl. Prob.*, 22:676–705, 1990.

[11] M.F. Neuts. Markov chains with applications in queueing theory, which have a matrix geometric invariant probability vector. *Adv. Appl. Prob.*, 10:185–212, 1978.

[12] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach*. John Hopkins University Press, 1981.

[13] V. Ramaswami. Nonlinear matrix equations in applied probability - solution techniques and open problems. *SIAM review*, 30(2):256–263, June 1988.

[14] B. Van Houdt and C. Blondia. The delay distribution of a type k customer in a first come first served MMAP[K]/PH[K]/1 queue. *J. of Appl. Probab.*, 39(1):213–222, 2002.

[15] Y.Q. Zhao and A.S. Alfa. Performance analysis of a telephone system with both patient and impatient customers. *Telecommunication Systems*, 4:201–215, 1995.