



Contents lists available at SciVerse ScienceDirect

## Performance Evaluation

journal homepage: [www.elsevier.com/locate/peva](http://www.elsevier.com/locate/peva)

# A matrix geometric representation for the queue length distribution of multitype semi-Markovian queues

Benny Van Houdt\*

Performance Analysis of Telecommunication Systems, Department of Mathematics and Computer Science, University of Antwerp-IBBT, Middelheimlaan 1, B-2020 Antwerp, Belgium

## ARTICLE INFO

### Article history:

Received 2 February 2011

Accepted 9 January 2012

Available online xxxx

### Keywords:

Semi-Markovian queues

Queue length

Matrix geometric

Phase-type

Multitype queue

Correlated service and inter-arrival times

Quasi-Birth-Death Markov chain

Fluid queue

## ABSTRACT

In this paper we study a broad class of semi-Markovian queues introduced by Sengupta. This class contains many classical queues such as the GI/M/1 queue, SM/MAP/1 queue and others, as well as queues with correlated inter-arrival and service times. Queues belonging to this class are characterized by a set of matrices of size  $m$  and Sengupta showed that its waiting time distribution can be represented as a phase-type distribution of order  $m$ . For the special case of the SM/MAP/1 queue without correlated service and inter-arrival times the queue length distribution was also shown to be phase-type of order  $m$ , but no derivation for the queue length was provided in the general case.

This paper introduces an order  $m^2$  phase-type representation  $(\kappa, K)$  for the queue length distribution in the general case and proves that the order  $m^2$  of the distribution cannot be further reduced in general. A matrix geometric representation  $(\kappa, K, \nu)$  is also established for the number of type  $\tau \subseteq \{1, \dots, m\}$  customers in the system, where a customer is of type  $\tau$  if the phase in which it completes service belongs to  $\tau$ . We derive these results in both discrete and continuous time and also discuss the numerical procedure to compute  $(\kappa, K, \nu)$ . When the arrivals have a Markovian structure, the numerical procedure is reduced to solving a Quasi-Birth-Death (for the discrete time case) or fluid queue (for the continuous time case).

Finally, by combining a result of Sengupta and Ozawa, we provide a simple formula to compute the order  $m$  phase-type representation of the waiting time in a MAP/MAP/1 queue without correlated service and inter-arrival times, using the  $R$  matrix of a Quasi-Birth-Death Markov chain.

© 2012 Elsevier B.V. All rights reserved.

## 1. Introduction

In this paper we consider a broad class of semi-Markovian (SM) queues that have been used extensively to assess the performance of various communication (and manufacturing) systems (e.g., [1–3]). Characteristic of these queues is that they support correlated arrival processes, correlation between successive service times, as well as correlation between the service and inter-arrival times. Denote  $T_1 < T_2 < T_3 < \dots$ , with  $T_1 = 0$  as the customer arrival times,  $I_{n+1} = T_{n+1} - T_n$ , for  $n \geq 1$ , as the inter-arrival times and  $S_n$ , for  $n \geq 1$ , as the service time of the  $n$ -th customer. Let  $\{Y_n, n \geq 1\}$  be an irreducible aperiodic Markov chain with a finite state space  $\{1, \dots, m\}$ . Then, a single server queue is termed semi-Markovian [4] provided that

$$\begin{aligned} P[I_{n+1} \leq x, S_n \leq y, Y_{n+1} = j | Y_1, \dots, Y_n, S_1, \dots, S_{n-1}, I_1, \dots, I_n] \\ = P[I_{n+1} \leq x, S_n \leq y, Y_{n+1} = j | Y_n], \end{aligned} \quad (1)$$

\* Tel.: +32 3 265 3891; fax: +32 3 265 3777.

E-mail address: [benny.vanhoudt@ua.ac.be](mailto:benny.vanhoudt@ua.ac.be).

where the latter probability does not depend on  $n$ . In other words, given the state  $Y_n$  of the Markov chain, the service time of customer  $n$  and the inter-arrival time between customer  $n$  and  $n + 1$  are independent of all prior service times and inter-arrival times. Notice however that given  $Y_n$  the service time  $S_n$  and inter-arrival time  $I_{n+1}$  can be correlated.

Many traditional queues with independent inter-arrival times and service times belong to the above-mentioned class, such as the GI/PH/1 queue [5, Section 3], the SM/PH/1 queue [6, Section 4.2], the MAP/MAP/1 queue and the more general SM/MAP/1 queue [1], as well as queues with general and semi-Markovian service times. More importantly the class also contains various queues with correlated service and inter-arrival times (see Section 2), such as the MMAP[K]/PH[K]/1 [7,8], SM[K]/PH[K]/1 [9,10] multi-type queues, the D-MAP/PH/1 queue [2] and the M/SM/1 queue [11].

The main performance measures of this class of queues, such as their queue length and waiting time distribution, their transforms and moments, have only been obtained for special cases. For instance, in [11] the inter-arrival times  $I_{n+1}$  are assumed to be exponential, but can still be correlated with the semi-Markovian service times. In this case transforms for both distributions were obtained as well as recursive formulas for their moments. Sengupta [1] considered a very broad subclass by assuming that the service times  $S_n$  are phase-type (but still correlated). More specifically, denoting  $\bar{Y}_n$  as the phase in which the service of customer  $n$  is started and  $E_n$  as the phase in which customer  $n$  ends service, Sengupta demanded that

$$\begin{aligned} P[I_{n+1} \leq x, S_n \leq y, \bar{Y}_{n+1} = j, E_n = v | \bar{Y}_1, \dots, \bar{Y}_n, S_1, \dots, S_{n-1}, I_1, \dots, I_n] \\ = P[S_n \leq y, E_n = v | \bar{Y}_n] P[I_{n+1} \leq x, \bar{Y}_{n+1} = j | E_n = v], \end{aligned} \quad (2)$$

where the latter two probabilities are again independent of  $n$ . Notice,  $\bar{Y}_n$  assumes the role of  $Y_n$  in Eq. (1) and  $\bar{Y}_n$  determines  $\bar{Y}_{n+1}$ ,  $S_n$  and  $I_{n+1}$  as follows. First  $\bar{Y}_n$  determines the service time  $S_n$  and end phase  $E_n$ . Next, the inter-arrival time  $I_{n+1}$  and initial phase  $\bar{Y}_{n+1}$  are determined by  $E_n$ , meaning they are independent of  $S_n$  given  $E_n$ . The service and inter-arrival times can of course still be correlated via  $E_n$ . Furthermore, Sengupta also assumed that given that customer  $n$  starts his service in phase  $i$ , his service time is phase-type (PH) with characterization  $(e_i, S)$  for some  $m \times m$  matrix  $S$ , where  $e_i$  is a vector with a 1 in position  $i$  and 0 elsewhere.

In this paper we will consider the same subclass of semi-Markovian queues as Sengupta in both continuous and discrete time. In continuous time  $(e_i, S)$  is a continuous-time PH (CPH) distribution, that is, the probability that the service has a duration of length  $y$  or more is given by  $e_i \exp(Sy)e$  (with  $e$  a vector of ones), and the inter-arrival time can be discrete, continuous or a mixture of the two. In the discrete-time setting, time is slotted and the service time is a discrete-time PH (DPH) distribution such that all the service times are multiples of one time slot and the probability that the service lasts at least  $y$  time slots can be expressed by  $e_i S^{y-1} e$ . The inter-arrival times are also general, but discrete, meaning all the inter-arrival times are multiples of the length of a time slot as well. It is important to stress once more that the service and inter-arrival times are correlated.

Sengupta [1] showed (for the continuous time case), using the age process (see Sections 3 and 4) and the theory of Markov processes with a matrix exponential distribution [5], that the waiting time distribution in such a queuing system has a phase-type representation of order  $m$ . Moreover, for the special case of the SM/MAP/1 queue *without* dependencies between the service and inter-arrival times, the queue length distribution was also shown to be phase-type of order  $m$  (though the numerical procedure to compute it converges only linearly). However, no results on the queue length distribution were provided for the general case considered in [1].

In this paper we derive a phase-type representation  $(\kappa, K)$  of order  $m^2$  for the queue length distribution in the general case (in both discrete and continuous time) and show that in general this representation cannot be reduced in order. Additionally, we define customer  $n$  as type  $\tau$ , with  $\tau$  a non-empty subset of  $\{1, \dots, m\}$ , if its service end phase  $E_n$  is part of the set  $\tau$ . An order  $m^2$  matrix geometric representation  $(\kappa, K, v)$  for the type  $\tau$  queue length distribution is also established and setting  $\tau = \{1, \dots, m\}$  results in the above-mentioned phase-type representation  $(\kappa, K)$ . The derivation of these representations is based on the Markov chain that captures the age of the customer in service and relies on a simple observation made by Ozawa in [12]. Of course, for various subclasses such as the SM/MAP/1 queue without correlation between the service and inter-arrival times (and thus also the MAP/MAP/1 queue) this representation is redundant as a smaller, order  $m$  representation is known to exist. An example (with correlated inter-arrival and service times) for which the minimal order lies between  $m$  and  $m^2$  is provided as well. We also discuss the numerical issues related to the computation of this order  $m^2$  phase-type representation. Furthermore, we indicate that by combining some of the results of Sengupta [1] with those of Ozawa [12], the order  $m$  waiting time distribution of the traditional MAP/MAP/1 queue can be computed with hardly an effort from the  $R$ -matrix of the Quasi-Birth-Death Markov chain that describes the evolution of the queue length [13].

The results presented in this paper also resemble the ones obtained by Ozawa [12] for the class of queues that are defined by a general Quasi-Birth-Death (QBD) process. This class also supports queues with correlated service and arrival times and also includes the MAP/MAP/1 queues without such correlation. Actually, the latter queues seem to be the only ones that reside in the intersection of the queues considered in this paper and the ones considered by Ozawa. In [12] Ozawa derived an order  $m^2$  phase-type distribution for the sojourn time, while the order  $m$  phase-type representation for the queue length of such a queue is immediate from Neuts [6]. Hence, there seems to be some form of duality present between our results and the ones presented in [12]. The minimality of the order  $m^2$  representation was not proven by Ozawa for the general case. However, it is not hard to develop examples for which the order  $m^2$  sojourn time distribution is minimal.

In the next section we start by discussing a number of examples that fit within the subclass of semi-Markovian queues considered in this paper. In Section 3 we will present our main results for the discrete-time case, whereas Section 4 covers the somewhat more involved continuous-time setting. We conclude in Section 5 by providing some numerical examples.

## 2. Definitions and examples

This section is mostly devoted to providing examples of well-known queuing systems that fit within the subclass of queues studied in this paper. The first two examples are queues *without* correlation between the service and inter-arrival time, these were also discussed in [1]. Recall from the previous section, and more specifically from Eq. (2), that the semi-Markovian queues considered in this paper are characterized by two sets of probabilities. The first set holds the probabilities

$$P[S_n \leq y, E_n = v \mid \bar{Y}_n = i],$$

that determine the probability that the service time is smaller than or equal to  $y$  and the service ends in phase  $v$ , given that the service started in phase  $i$ . The second set is formed by

$$P[I_{n+1} \leq x, \bar{Y}_{n+1} = j \mid E_n = v],$$

holding the probability that the inter-arrival time is smaller than or equal to  $x$  and the next customer starts service in phase  $j$ , provided that the current service ended in phase  $v$ . The first set of probabilities is denoted as  $V_{i,v}(y)$ , meaning  $V(y)$  are  $m \times m$  matrices for  $y \geq 0$ . Due to the assumption on the phase-type service characterized by  $(e_i, S)$ , we have in the continuous-time case

$$V(y) = \int_{z=0}^y \exp(Sz)S^*dz = (I_m - \exp(Sy))(-S)^{-1}S^*, \tag{3}$$

where  $I_m$  is the order  $m$  identity matrix,  $S^*$  is a diagonal matrix with  $S^*e = -Se$  and  $e$  is vector of ones. In discrete time on the other hand we find

$$V(y) = \sum_{k=0}^{y-1} S^k S^* = (I_m - S^y)(I - S)^{-1}S^*, \tag{4}$$

where  $S^*$  is a diagonal matrix with  $S^*e = (I - S)e$ . The second set of matrices is denoted as  $P_{v,j}(x)$ , where  $P(x)$  is also a square matrix of order  $m$ . Throughout the paper we assume that  $S + S^* \int_{x=0}^{\infty} dP(x)$  is irreducible and that  $P(0) = 0$ , meaning there are no batch arrivals. Before we provide a number of examples for which we will specify both  $V(y)$  (i.e.,  $S$ ) and  $P(x)$ , we remind the reader that customer  $n$  is said to be of type  $\tau$ , for  $\tau$  a non-empty subset of  $\{1, \dots, m\}$ , if its end service phase  $E_n$  is part of  $\tau$ . We also remark that this end phase, determined by  $V(y)$ , is not affected by  $S^*$  as it is a diagonal matrix.

*The GI/PH/1 queue.* Consider a queue where the arrivals form a renewal process with inter-arrival time distribution given by  $H(t)$  and assume the service is independent of the inter-arrival times and follows an order  $m$  PH distribution given by  $(\alpha, S)$ . This implies that  $P(x) = H(x)e\alpha$  and  $V(y)$  is determined by  $S$  as indicated in (3). This queue was studied by Sengupta in [5], where an order  $m$  representation for both the queue length and waiting time distribution was given.

*The SM/MAP/1 queue.* Consider the queue with semi-Markovian arrivals, i.e., the arrival process is a Markov renewal process, and Markovian services. Let the entries  $H_{i,j}(t)$  hold the probability of having an inter-arrival time smaller than or equal to  $t$ , while the state of the Markov renewal process changes from  $i$  to  $j$  (for  $i, j \in \{1, \dots, m_a\}$ ). Similarly let the size  $m_s$  matrices  $S_0$  and  $S_1$  characterize the Markovian service, meaning the service time of a customer starting in phase  $i$  is an order  $m_s$  PH distribution characterized by  $(e_i, S_0)$ , while  $(S_1)_{i,j}$  holds the probability that customer  $n+1$  starts service in phase  $j$  given that customer  $n$  ended his service in phase  $i$ . In this case a semi-Markovian queue is obtained by setting  $P(x) = H(x) \otimes S_1$ , while the matrix  $S$  in (3) is given by  $I_{m_a} \otimes S_0$ , with  $\otimes$  denoting the matrix Kronecker product. An order  $m = m_a m_s$  PH representation for both the queue length and waiting time distribution was provided by Sengupta in [1]. The popular MAP/MAP/1 queue clearly belongs to the set of SM/MAP/1 queues and we will provide a much faster way to compute Sengupta's order  $m$  representation for its waiting time distribution. Finally, Neuts studied the special case of the SM/PH/1 queue in [6, Section 4.2].

*The SMK/PHK/1 queue.* The SM[K] arrival process is a multi-type Markov renewal process characterized by the  $m_a \times m_a$  matrices  $H^{(k)}(t)$ , for  $k = 1, \dots, K$ . Entry  $H_{i,j}^{(k)}(t)$  holds the probability of having an inter-arrival time smaller than or equal to  $t$ , while the state of the Markov renewal process changes from  $i$  to  $j$  and the type of the arriving customer is  $k$ . The PH[K] service process indicates that type  $k$  customers follow an order  $m_s^{(k)}$  phase-type distribution with parameters  $(\alpha_k, S_k)$ , for  $k = 1, \dots, K$ . Notice, consecutive service times are correlated via the correlation between the customer types and as such there is also correlation between the service and inter-arrival times. To represent this queue as a semi-Markovian queue with  $m = m_a \sum_k m_s^{(k)}$ , it suffices to set

$$S = \begin{bmatrix} S_1 & 0 & \dots & 0 \\ 0 & S_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & S_K \end{bmatrix} \otimes I_{m_a},$$

and

$$P(x) = (e \otimes I_{m_a}) [\alpha_1 \otimes H^{(1)}(x) \cdots \alpha_K \otimes H^{(K)}(x)].$$

Examples 5–7 given in [1] are a special case of an SM[K]/PH[K]/1 queue in continuous time. HE [9,10] analyzed the SM[K]/PH[K]/1 queue in discrete and continuous time and provided an order  $m$  PH representation for the overall and per type waiting time distributions. In the discrete-time paper [9] no results were provided for the queue length, while in continuous time a recursive algorithm was provided that required the solution of a Sylvester matrix equation during each step in the special case of MMAP[K] arrivals. Thus, no phase-type representation for the overall (or per type) queue length distribution was given. In [7,8] the special case of the MMAP[K]/PH[K]/1 queue in discrete time was considered.

As a special case of our main result, we establish an order  $m^2$  PH representation for the overall queue length in an SM[K]/PH[K]/1 queue and obtain a matrix geometric representation for the type  $k$  queue length distribution by setting  $\tau = \{\sum_{i=1}^{k-1} m_s^{(i)} m_a + 1, \dots, \sum_{i=1}^k m_s^{(i)} m_a\}$ . Furthermore, for the continuous-time MMAP[K]/PH[K]/1 queue, we also introduce the first algorithm with quadratic convergence (by means of a reduction to a fluid queue) to compute the matrix  $T$  needed to obtain the above-mentioned phase-type representation of both the waiting time and queue length distribution.

*The D-MAP/PH/1 queue with correlated service and inter-arrival times.* The discrete-time Markovian arrival process (MAP) is characterized by the order  $m_a$  matrices  $D_0$  and  $D_1$ , while a customer starting service requires an order  $m_s$  DPH distributed amount of service characterized by  $(e_i, T)$ . Customer  $n + 1$  will start service in phase  $i$  according to the probability vector  $\alpha_i$  provided that the inter-arrival time between customer  $n$  and  $n + 1$  is equal to  $l$ , meaning the service time and inter-arrival time are clearly correlated. In [2] it was shown that this queue is equally general as assuming that the service time of customer  $n + 1$  is DPH with characterization  $(\alpha_i, T_i)$ , for some matrices  $T_i$ . This queue can be represented as a semi-Markovian queue by setting  $S = T \otimes I_{m_a}$ , while

$$P(x) = \sum_{l=1}^x e\alpha_l \otimes (D_0^{l-1} D_1),$$

for  $x = 1, 2, \dots$ . This queue was studied in [2] where the more general SM/PH/1 queue with correlated service and inter-arrival times was also discussed. This more general model is also a semi-Markovian queue (simply replace  $D_0^{l-1} D_1$  by  $H_l$  in the expression for  $P(x)$ ). Although the queue length distribution was computed using some recursive computations in [2], no phase-type representation was found (see also Section 3.3).

### 3. Semi-Markovian queue in discrete time

In this section we consider the discrete-time semi-Markovian queue, which implies that  $S$  is a substochastic matrix and  $P(x)$  is a step function with steps at  $x = 1, 2, \dots$ . Denote  $Q(x) = P(x) - P(x - 1)$ , i.e., the matrix holding the phase changes when the inter-arrival time equals  $x$ . We will derive an order  $m^2$  matrix geometric representation  $(\kappa, K, \nu)$  for the type  $\tau$  queue length distribution, i.e., we determine a stochastic vector  $\kappa$  of size  $m^2$  and a substochastic matrix  $K$  such that the probability of having  $i$  or more type  $\tau$  customers in the queue (provided that it is busy) equals  $\kappa K^{i-1} \nu$ , for  $i \geq 1$ . Furthermore, for  $\tau = \{1, \dots, m\}$ , meaning if we consider the overall queue length,  $\nu = e$  and the representation is a phase-type representation. We will also show that in general this order  $m^2$  representation cannot be reduced, though special cases are known for which a smaller representation exists, e.g., the SM/MAP/1 queue without correlation between the service and inter-arrival times.

To obtain the order  $m^2$  representation, we will rely on the discrete-time version of the age process used by Sengupta in [1] and we will make use of a simple lemma by Ozawa [12]. The age process of the discrete-time semi-Markovian queue is characterized by a GI/M/1-type Markov chain with transition matrix:

$$P = \begin{bmatrix} C_0 & A_0 & 0 & 0 & 0 & \dots \\ C_1 & A_1 & A_0 & 0 & 0 & \dots \\ C_2 & A_2 & A_1 & A_0 & 0 & \dots \\ C_3 & A_3 & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{5}$$

where  $C_i = \sum_{s>i} A_s$ . The  $m \times m$  matrices  $A_s$ , for  $s \geq 1$ , are said to decrease the *level* of the chain by  $s - 1$ , while  $A_0$  increases the level by one. Entry  $(i, j)$  of  $A_s$  is said to change the *phase* of the chain from  $i$  to  $j$ . The semi-Markovian queue will be observed by this chain whenever the server is busy and the level will represent the *age* of the customer in service, while the phase maintains the current phase of the DPH service process. The age  $a$  of a customer is defined as the number of time slots between the current time epoch  $t$  and its arrival time  $t - a$ . We assume that we observe the system just prior to time  $t$ , meaning the age of a customer, while the server is busy, is at least one.

Thus, the matrix  $A_0$  holds the probabilities that the server continues to serve the same customer (age increases by one), while  $A_s$  holds the probability that a service completion occurs and the next customer arrives  $s$  time units later (meaning the age at the next point of observation decreases by  $s - 1$ ). In other words,

$$\begin{aligned} A_0 &= S, \\ A_s &= S^* Q(s), \end{aligned}$$

for  $s = 1, 2, \dots$  with  $S^*$  the diagonal matrix such that  $S^*e = (I_m - S)e$ . The matrices  $C_s$ , for  $s \geq 0$  capture the case where the server becomes idle after the service completion (assuming arrivals occur after any possible service completions at time  $t$ ). As we only observe the queue when the system is busy, the age  $a$  during the next point of observation must be one.

Due to Neuts [6], this chain is positive recurrent if and only if  $\theta \sum_{s=1}^{\infty} sA_s e > 1$ , with  $\theta$  the invariant vector of  $A = \sum_{s=0}^{\infty} A_s$  (notice,  $A$  is irreducible as  $S + S^* \int_0^{\infty} dP(x) = S + S^* \sum_{s=1}^{\infty} Q(s)$  was assumed to be irreducible). Its stationary distribution  $\pi = (\pi(1), \pi(2), \dots)$ , with  $\pi(i)$  a  $1 \times m$  vector, has a matrix geometric form, that is,  $\pi(n) = \pi(1)R^{n-1}$ , with  $R$  the smallest non-negative solution to

$$R = \sum_{s=0}^{\infty} R^s A_s = S + \sum_{s=1}^{\infty} R^s S^* Q(s).$$

Notice, this  $R$ -matrix is the discrete-time analogue of the  $T$ -matrix of Sengupta [1]. Also remark that we cannot simply set up a Markov chain that has the number of customers in the queue as the level of the chain, except for special cases like the MAP/MAP/1 queue and others, because in general the phase at the end of the service of customer  $n$  influences the inter-arrival time between customer  $n$  and  $n + 1$ .

### 3.1. Order $m^2$ representation

To obtain the order  $m^2$  representation, we will make use of the following lemma by Ozawa [12] that can be proven by direct verification. Let

$$\xi = \begin{bmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_m^T \end{bmatrix},$$

and let  $a$  and  $b$  be two arbitrary  $1 \times m$  vectors, then

$$(a \otimes b)\xi = ab^T = ba^T,$$

where  $T$  denotes the transposed vector.

Let  $N_b(\tau)$  be the random variable representing the number of type  $\tau$  customers in the queue provided that it is busy (otherwise, the number is zero), that is,  $P[N_b(\tau) = n]$  is the probability that the queue holds  $n$  type  $\tau$  customers (including the possible type  $\tau$  customer in service) at an arbitrary point in time given that the server is busy.

**Theorem 1.** *The queue length distribution  $N_b(\tau)$  of the type  $\tau$  customers has a matrix geometric representation  $(\kappa, K, \nu)$  of order  $m^2$  given by*

$$\kappa = \xi^T (I_m \otimes \Delta(\theta)), \quad K = (I - K(\tau^c))^{-1} K(\tau) \tag{6}$$

and

$$\nu = (I - K(\tau^c))^{-1} ((I_m - S)^{-1} \Delta(e_\tau) (I_m - S)e \otimes e) \tag{7}$$

with

$$K(\zeta) = ((I - S)^{-1} \Delta(e_\zeta) \otimes I_m) \sum_{s \geq 1} (A_s \otimes G^s), \tag{8}$$

where  $\tau^c = \{1, \dots, m\} \setminus \tau$ ,  $G = \Delta^{-1}(\theta)R^T \Delta(\theta)$ ,  $\Delta(x)$  denotes a diagonal matrix such that  $\Delta(x)e = x^T$  and  $e_\zeta$  for  $\zeta$  a subset of  $\{1, \dots, m\}$ , is a vector with entry  $i$  equal to one if  $i \in \zeta$  and zero otherwise. Furthermore, the vector  $\kappa$  is stochastic and the matrix  $K$  is substochastic.

**Proof.** Define  $\bar{A}_s(\zeta) = (I_m - A_0)^{-1} \Delta(e_\zeta) A_s$  and  $\bar{C}_i(\zeta) = (I_m - A_0)^{-1} \Delta(e_\zeta) C_i$ , with  $\zeta$  a subset of  $\{1, \dots, m\}$  (the inverse  $(I_m - A_0)^{-1}$  exists as  $A_0 = S$  is a strictly substochastic matrix). Via the stationary probability vector  $\pi$  of the age process we can express  $P[N_b(\tau) = n]$  as

$$P[N_b(\tau) = n] = \sum_{i=1}^{\infty} \pi(i) P_{\tau,n}(i-1)e, \tag{9}$$

with  $P_{\tau,n}(i)$  an  $m \times m$  matrix defined as follows. Given that an arrival occurred in phase  $j$  at time 0, entry  $(j, j')$  of  $P_{\tau,n}(i)$  equals the probability that either (i) the arrival at time zero is of type  $\tau^c$  and  $n$  type  $\tau$  arrivals occur on the time epochs  $t = 1$  to  $i$ , while the phase is  $j'$  when the first arrival at time  $t > i$  occurs or (ii) if  $n > 0$ , the arrival at time zero is of type  $\tau$  and  $n - 1$  type  $\tau$  arrivals occur on the time epochs  $t = 1$  to  $i$ , while the phase is  $j'$  when the first arrival at time  $t > i$

occurs. Notice, we make use of the probabilities  $P_{\tau,n}(i-1)$  instead of  $P_{\tau,n}(i)$  as we are observing the system just prior to any possible arrivals or service completions, hence arrivals and service completions occurring at time  $t$  are not part of the system state at time  $t$ .

Using the matrix geometric form of  $\pi$  and the above-mentioned lemma of Ozawa with  $a^T = P_{\tau,n-1}(i)e$  and  $b = \pi(1)R^i$ , we find

$$\begin{aligned} P[N_b(\tau) = n] &= \sum_{i \geq 0} \pi(1)R^i P_{\tau,n}(i)e \\ &= (e^T \otimes \pi(1)) \sum_{i \geq 0} (P_{\tau,n}^T(i) \otimes R^i) \xi. \end{aligned} \tag{10}$$

Notice,  $\bar{A}_s(\tau)$  ( $\bar{A}_s(\tau^c)$ ) holds the probabilities that the customer arriving at time 0 is of type  $\tau$  ( $\tau^c$ ), while the next arrival occurs at time  $s$ . As such the matrices  $P_{\tau,n}(0) = 0$  for  $n > 1$ ,  $P_{\tau,1}(0) = \sum_{s>0} \bar{A}_s(\tau)$  and  $P_{\tau,0}(0) = \sum_{s>0} \bar{A}_s(\tau^c)$ , while

$$\begin{aligned} P_{\tau,0}(i) &= \sum_{s>i} \bar{A}_s(\tau^c) + \sum_{s=1}^i \bar{A}_s(\tau^c) P_{\tau,0}(i-s) \\ P_{\tau,1}(i) &= \sum_{s>i} \bar{A}_s(\tau) + \sum_{s=1}^i (\bar{A}_s(\tau) P_{\tau,0}(i-s) + \bar{A}_s(\tau^c) P_{\tau,1}(i-s)) \\ P_{\tau,n}(i) &= \sum_{s=1}^i (\bar{A}_s(\tau) P_{\tau,n-1}(i-s) + \bar{A}_s(\tau^c) P_{\tau,n}(i-s)) \end{aligned} \tag{11}$$

for  $i > 0$  and  $n \geq 1$ . This implies for  $n > 1$

$$P_{\tau,n}^T(i) \otimes R^i = \sum_{s=1}^i ((P_{\tau,n-1}^T(i-s) \otimes R^{i-s}) (\bar{A}_s^T(\tau) \otimes R^s) + (P_{\tau,n}^T(i-s) \otimes R^{i-s}) (\bar{A}_s^T(\tau^c) \otimes R^s)),$$

yielding

$$\sum_{i \geq 0} (P_{\tau,n}^T(i) \otimes R^i) = \left( \sum_{i \geq 0} (P_{\tau,n-1}^T(i) \otimes R^i) \right) \left( \sum_{s \geq 1} (\bar{A}_s^T(\tau) \otimes R^s) \right) + \left( \sum_{i \geq 0} (P_{\tau,n}^T(i) \otimes R^i) \right) \left( \sum_{s \geq 1} (\bar{A}_s^T(\tau^c) \otimes R^s) \right).$$

Thus, if we define

$$M(\zeta) = \sum_{s \geq 1} (\bar{A}_s^T(\zeta) \otimes R^s),$$

for  $\zeta$  subset of  $\{1, \dots, m\}$ , one finds for  $n > 1$

$$\sum_{i \geq 0} (P_{\tau,n}^T(i) \otimes R^i) = \left( \sum_{i \geq 0} (P_{\tau,n-1}^T(i) \otimes R^i) \right) M(\tau) (I - M(\tau^c))^{-1},$$

where the existence of  $(I - M(\tau^c))^{-1}$  will follow from the substochastic nature of  $K(\tau^c)$ .

Using (11), one similarly finds

$$\sum_{i \geq 0} (P_{\tau,1}^T(i) \otimes R^i) = \left( \sum_{i \geq 0} (P_{\tau,0}^T(i) \otimes R^i) \right) M(\tau) (I - M(\tau^c))^{-1} + \sum_{i \geq 0} (\bar{C}_i^T(\tau) \otimes R^i) (I - M(\tau^c))^{-1},$$

and

$$\sum_{i \geq 0} (P_{\tau,0}^T(i) \otimes R^i) = \sum_{i \geq 0} (\bar{C}_i^T(\tau^c) \otimes R^i) (I - M(\tau^c))^{-1}.$$

This allows us to express the probabilities  $P[N_b(\tau) = n]$  via (10) in a matrix geometric form of order  $m^2$  as

$$P[N_b(\tau) = 0] = \left( \sum_{i \geq 0} ((\bar{C}_i(\tau^c)e)^T \otimes \pi(1)R^i) \right) (I - M(\tau^c))^{-1} \xi,$$

$$P[N_b(\tau) = n] = \alpha (I - M(\tau^c))^{-1} (M(\tau) (I - M(\tau^c))^{-1})^{n-1} \xi,$$

for  $n > 0$  with

$$\alpha = \sum_{i \geq 0} ((\bar{C}_i(\tau^c)e)^T \otimes \pi(1)R^i) (I - M(\tau^c))^{-1} M(\tau) + \sum_{i \geq 0} ((\bar{C}_i(\tau)e)^T \otimes \pi(1)R^i).$$

By expanding  $\bar{C}_i(\zeta)$  and switching the order of the sums, we can rewrite  $\sum_{i \geq 0} ((\bar{C}_i(\zeta)e)^T \otimes \pi(1)R^i)$ , with  $\zeta$  a subset of  $\{1, \dots, m\}$  as

$$\sum_{s \geq 1} (e^T \bar{A}_s^T(\zeta) \otimes \pi(1)) \sum_{i=0}^{s-1} (I_m \otimes R^i) = (e^T \otimes \pi(1)(I_m - R)^{-1}) \sum_{s \geq 1} (\bar{A}_s^T(\zeta) \otimes (I_m - R^s))$$

which leads to

$$\sum_{i \geq 0} ((\bar{C}_i(\zeta)e)^T \otimes \pi(1)R^i) = (e^T \otimes \pi(1)(I_m - R)^{-1}) (((I_m - A_0)^{-1} \Delta(e_\zeta)(I_m - A_0)^T \otimes I_m) - M(\zeta)).$$

If we define  $S^T(\zeta) = ((I_m - A_0)^{-1} \Delta(e_\zeta)(I_m - A_0)^T)$ ,  $\alpha$  can be expressed as

$$\alpha = (e^T \otimes \pi(1)(I_m - R)^{-1}) ((S^T(\tau^c) \otimes I_m) - M(\tau^c))(I - M(\tau^c))^{-1} M(\tau) + (S^T(\tau) \otimes I_m) - M(\tau),$$

which can be further simplified to

$$\alpha = (e^T \otimes \pi(1)(I_m - R)^{-1}) (S^T(\tau) \otimes I_m) (I - (I - M(\tau^c))^{-1} M(\tau)),$$

by noting that  $e^T S^T(\tau) + e^T S^T(\tau^c) = e^T$ . Due to the form of  $C_i$ ,  $\pi(1)(I_m - R)^{-1}$  is readily recognized as the unique stochastic invariant vector of  $A = \sum_{s \geq 0} A_s$ , which we denoted earlier on as  $\theta$ . Thus, by further simplifying  $\alpha$ , one finds that  $P[N_b(\tau) = n]$  can be written as

$$P[N_b(\tau) = n] = (e^T S^T(\tau) \otimes \theta) (I - M(\tau^c))^{-1} (I - M) M^{n-1} \xi$$

with  $M = M(\tau)(I - M(\tau^c))^{-1}$ .

Moreover,  $P[N_b(\tau) \geq n] = (e^T S^T(\tau) \otimes \theta) (I - M(\tau^c))^{-1} M^{n-1} \xi$ , meaning  $N_b(\tau)$  has a matrix geometric representation  $((e^T S^T(\tau) \otimes \theta) (I - M(\tau^c))^{-1}, M, \xi)$  of order  $m^2$ . If  $\theta > 0$ , which holds due to the irreducibility assumption on  $A$ ,  $P[N_b(\tau) \geq n]$  can be rewritten as

$$P[N_b(\tau) \geq n] = (e^T S^T(\tau) \otimes \theta) (I_m \otimes \Delta^{-1}(\theta)) (I_m \otimes \Delta(\theta)) (I - M(\tau^c))^{-1} (I_m \otimes \Delta^{-1}(\theta)) \\ \times ((I_m \otimes \Delta(\theta)) M (I_m \otimes \Delta^{-1}(\theta)))^{n-1} (I_m \otimes \Delta(\theta)) \xi = v^T (K^T)^{n-1} \kappa^T.$$

This proves the theorem provided that  $\kappa$  is stochastic and  $K = (I - K(\tau^c))^{-1} K(\tau)$  is strictly substochastic. The matrix  $G = \Delta^{-1}(\theta) R^T \Delta(\theta)$  is recognized as the  $G$ -matrix of the Ramaswami dual of the GI/M/1-type Markov chain characterized by  $P$  [14]. As  $P$  is positive recurrent, its dual process is a transient M/G/1-type Markov chain and therefore  $G$  is strictly substochastic [15]. As a result  $K(\tau) + K(\tau^c)$  is strictly substochastic due to  $\sum_{s \geq 1} (I_m - A_0)^{-1} A_s e = e$ . Hence,  $K(\tau)$  and  $K(\tau^c)$  are strictly substochastic, the inverse  $(I - K(\tau^c))^{-1}$  exists and  $K$  is strictly substochastic. The vector  $\kappa$  is clearly stochastic.  $\square$

**Corollary 1.** The overall queue length distribution  $N_b$  has a phase-type representation  $(\kappa, K)$  of order  $m^2$  given by

$$\kappa = \xi^T (I_m \otimes \Delta(\theta)), \quad K = \sum_{s \geq 1} (\bar{A}_s \otimes G^s), \tag{12}$$

where  $\bar{A}_s = (I_m - A_0)^{-1} A_s$ ,  $G = \Delta^{-1}(\theta) R^T \Delta(\theta)$  and  $\Delta(x)$  denotes a diagonal matrix such that  $\Delta(x)e = x^T$ .

**Proof.** The proof is immediate from the previous theorem by noting that  $\tau = \{1, \dots, m\}$ , such that  $K(\tau^c) = 0$  and  $v = e$ .  $\square$

### 3.2. Redundancy of the representation

In this section we provide an example of a semi-Markovian queue with  $m = 2$  such that its order  $m^2 = 4$  phase-type representation  $(\kappa, K)$  for the overall queue length distribution cannot be represented by a phase-type (or matrix geometric) distribution with an order below four. This implies that the  $m^2$  order cannot be reduced in general. Examples with  $m > 2$  can be constructed in a similar manner.

We consider a queue with 2 types of customers, both customer types require a geometric amount of service. Type 1 customers have a mean service time of  $1/(1-s)$  time slots, while the type 2 customers require a mean service of  $1/(1-r)$  time slots. The arrival process is periodic in the sense that at times  $3t$ , for  $t = 0, 1, 2, \dots$  there is a type 1 arrival and at times  $3t + 1$ , for  $t = 0, 1, \dots$  we have a type 2 arrival. There is no need to consider a process with periodic arrivals, one can also easily generate examples for which the arrival process is aperiodic, neither does the type of the customer need to alternate between type 1 and type 2. Thus, the queue under consideration has

$$S = \begin{bmatrix} s & 0 \\ 0 & r \end{bmatrix}, \quad Q(1) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad Q(2) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

As  $A_2$  is of rank 1, the positive recurrent GI/M/1-type Markov chain is actually a Quasi-Birth-Death Markov chain and its  $G$  matrix is equal to

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

As a result,  $R = A_0(I_m - A_1 - A_0G)^{-1}$  is the rational matrix

$$R = \frac{1}{(1-r)(1-s)} \begin{bmatrix} s & s(1-s) \\ r^2 & r(1-s) \end{bmatrix}$$

and the phase-type representation  $(\kappa, K)$  is found as

$$\kappa = \left( \frac{1-r}{2-r-s}, 0, 0, \frac{1-s}{2-r-s} \right),$$

and

$$K = \frac{1}{(1-r)(1-s)} \begin{bmatrix} 0 & 0 & s & r^2 \frac{1-s}{1-r} \\ 0 & 0 & s(1-r) & r(1-s) \\ \frac{s(r^2 + s - r^2s)}{(1-r)(1-s)} & \frac{r^2(r + s - rs)}{(1-r)^2} & 0 & 0 \\ \frac{s(r + s - rs)}{1-s} & \frac{r^2}{1-r} & 0 & 0 \end{bmatrix}.$$

The matrix  $K$  is periodic due to the periodicity of the arrival process. Examples where  $K$  is aperiodic are also easy to find. In order to prove that a DPH representation  $(\beta, T)$  of order  $n$  is minimal, one first computes  $m_k = \beta(I_n - T)^{-k}e$  for  $k = 0, \dots, 2n - 2$  and checks whether the Hankel matrix defined by these  $2n - 1$  values has a determinant different from zero [16–18]. Actually, the results in [16–18] are for continuous-time phase-type (CPH) distributions, but can be used directly in discrete time by remarking that any order  $n$  DPH characterized by  $(\beta, T)$  can be transformed into a CPH characterized by  $(\beta, T - I_n)$  that has  $m_k$  as its  $k$ -th moment. A smaller order representation will exist for the DPH if and only if it exists for the CPH. We also remark that in general the minimal order does not coincide with the number of distinct eigenvalues of  $T$  as examples can be given where the minimal order is smaller.

As  $K$  is in rational form, we can compute  $m_0$  to  $m_6$  in rational form and perform an exact computation of the determinant of the corresponding Hankel matrix, which is a rational function of  $r$  and  $s$ . If we now fix  $r = 1/4$  (this value was chosen arbitrarily), we obtain a rational function of  $s$  for the determinant. This function has three real zeros:  $s = 0, 1/4$  and  $1$ . Thus, for all  $s$  values different from these three the queue length distribution  $N_b$  does not have a phase-type (or matrix geometric) representation with an order below 4. For  $s = 1$ , the system is unstable, while for  $s = 1/4$ , the service time of all customers is geometric with mean  $4/3$ , meaning the service time and inter-arrival times are independent and an order 2 phase-type representation is known to exist as the queue reduces to an SM/M/1 queue. The last case  $s = 0$  is rather interesting: type 1 customers require a deterministic service and type 2 a geometric one. Also, there is still correlation between the service and inter-arrival times and it turns out that the rank of its Hankel matrix is 3, with  $m < 3 < m^2$ . Hence, examples exist for which the minimal order lies between  $m$  and  $m^2$ . Even for any  $0 < r < 1$  and  $s = 0$ , the minimal phase-type representation of the queue length distribution is of order 3 and can be represented by  $\kappa = ((1-r)/(2-r), 0, 1/(1-r))$  and

$$K = \begin{bmatrix} 0 & 0 & \frac{r^2}{(1-r)^2} \\ 0 & 0 & \frac{r}{1-r} \\ 0 & \frac{r^2}{(1-r)^2} & 0 \end{bmatrix}.$$

### 3.3. Computation of the queue length distribution

#### 3.3.1. General case

To compute the representation  $(\kappa, K, \nu)$ , one first computes  $\theta$ , the invariant vector of  $\sum_s A_s$ , from which  $\kappa$  is obtained via Eq. (6). Next, the matrices  $A_s^{(r)}$  that characterize the Ramaswami dual of  $P$  are computed as

$$A_s^{(r)} = \Delta^{-1}(\theta)A_s^T \Delta(\theta),$$

for  $s = 0, 1, \dots$  These matrices characterize a transient M/G/1-type Markov chain and its  $G$  matrix is the smallest non-negative solution to

$$G = \sum_{s \geq 0} A_s^{(r)} G^s.$$



This non-linear equation can be solved using the cyclic reduction algorithm [19] which converges quadratically. The matrices  $K(\tau)$  and  $K(\tau^c)$  can be computed from  $G$  using Eq. (8), from which  $K$  and  $\nu$  can be readily obtained. The probabilities  $P[N_b(\tau) \geq n]$  can now be computed as  $\kappa K^{n-1} \nu$ .

If  $m$  is large, the construction of  $K(\tau)$  and  $K(\tau^c)$  can be avoided by noting that they are a sum of Kronecker products and the required multiplications can be performed using the shuffle algorithm [20] (after rewriting  $(I - K(\tau^c))^{-1}$  as an infinite sum). It might also be beneficial to compute the probabilities  $P[N_b(\tau) = n]$  directly from Eq. (9) by first computing the necessary  $P_{\tau,n}(i)$  matrices recursively using the relation in (11) (unless the spectral radius of  $R$  is close to one). This approach was used to compute the overall queue length distribution  $N_b$  in [2] for the discrete-time D-MAP/PH/1 queue with correlated service and inter-arrival times. In the continuous-time setting such a direct approach seems less attractive due to the numerical integrations involved. Furthermore, if we are only interested in the first few moments of the queue length distribution, we can compute these from  $(\kappa, K, \nu)$  directly, avoiding the need to compute the entire distribution.

Next, we consider a special case for which we can compute  $G$  using a Quasi-Birth-Death process with blocks of size  $2m$ , which results in an even better time and memory complexity.

### 3.3.2. Markovian inter-arrival times

Assume  $Q(u)$  can be written as  $Q(u) = \bar{D}_0^{u-1} \bar{D}_1$  for some  $m \times m$  matrices  $\bar{D}_0$  and  $\bar{D}_1$ , such that  $(\bar{D}_0, \bar{D}_1)$  characterizes a discrete-time MAP process. Notice, apart from  $\bar{D}_0$  and  $\bar{D}_1$ , the service process influences the arrival process as well via the matrix  $S$ , meaning in general the arrival process is *not* the MAP characterized by  $(\bar{D}_0, \bar{D}_1)$ . In this case we can define a Quasi-Birth-Death process characterized by the size  $2m$  matrices  $\hat{A}_0, \hat{A}_1$  and  $\hat{A}_2$  (where  $\hat{A}_0$  captures the upward transitions and  $\hat{A}_2$  the downward transitions)

$$\hat{A}_0 = \begin{bmatrix} 0 & 0 \\ 0 & S \end{bmatrix}, \quad \hat{A}_1 = \begin{bmatrix} 0 & \bar{D}_1 \\ 0 & S^* \bar{D}_1 \end{bmatrix}, \quad \hat{A}_2 = \begin{bmatrix} \bar{D}_0 & 0 \\ S^* \bar{D}_0 & 0 \end{bmatrix}.$$

Let the  $2m \times 2m$  matrix  $\hat{R}$  be the smallest non-negative solution to

$$\hat{R} = \hat{A}_0 + \hat{R} \hat{A}_1 + \hat{R}^2 \hat{A}_2,$$

which can be computed using the cyclic (or logarithmic) reduction algorithm with quadratic convergence [19]. Looking at the probabilistic interpretation of the matrices  $R$  and  $\hat{R}$  (see [6]), we find that  $R$  is identical to the size  $m$  lower right corner of  $\hat{R}$ . Thus, having computed  $\hat{R}$ , we can retrieve  $R$ , compute  $G$  and determine  $K(\zeta)$  for  $\zeta = \tau$  and  $\tau^c$  via

$$K(\zeta) = ((I_m - S)^{-1} \Delta(e_\zeta) S^* \otimes I_m) (I - (\bar{D}_0 \otimes G))^{-1} (\bar{D}_1 \otimes G).$$

The idea of the Quasi-Birth-Death reduction introduced above is a further generalization of the method first developed in [8] for the MMAP[K]/PH[K]/1 and later generalized to the SM[K]/PH[K]/1 queue in [9], where the reductions in computation time were illustrated by various examples.

## 4. Semi-Markovian queue in continuous time

Let us now consider the continuous-time case. As indicated in Section 2, the continuous-time semi-Markovian queues considered in this paper are characterized by an order  $m$  rate matrix  $S$  (i.e., the diagonal entries of  $S$  are negative, the remaining elements are non-negative and the row sums are non-positive) and a set of matrices  $P(u)$ , for  $u > 0$ . The matrix  $S$  describes the evolution of the phase while a customer remains in service, while entry  $(i, j)$  of  $P(u)$  held the probability of having an inter-arrival time smaller than or equal to  $u$ , while a customer ended his service in phase  $i$  and the next customer starts service in phase  $j$ . Recall that  $S^*$  was defined as a diagonal matrix such that  $S^* e = -S e$ , i.e., it contains the rates at which a service completion occurs and define  $A(u) = S^* P(u)$  as the rate of having a service completion followed by an inter-arrival time smaller than or equal to  $u$ . Denote  $dA(u)$  as the rate of having a service completion and an inter-arrival time between  $u$  and  $u + du$ .

As in [1], we consider the age process that observes the queue during the busy periods and that keeps track of the age of the customer in service and the current phase of the server. Thus, the age of the customer in service increases linearly while the phase evolves according to  $S$  until a service completion occurs that causes the chain to jump down by  $u$  according to  $dA(u)$ . In other words, this age process is a Markov process with a matrix-exponential steady state distribution, provided that it is positive recurrent [5]. Notice, such a process is similar to the GI/M/1-type Markov chains introduced by Neuts [6], but the *level* is a continuous variable that takes values in  $[0, \infty)$ . This irreducible Markov process is positive recurrent if and only if  $\theta S^* \int_0^\infty u dP(u) e > 1$ , where  $\theta$  is the unique invariant vector of  $S + S^* \int_0^\infty dP(u)$  [1].

Let  $\pi_i(x)$  denote the density of having a customer of age  $x$  in service in phase  $i$  at an arbitrary moment in time provided that the server is busy. Due to Sengupta,  $\pi(x)$  has a matrix exponential form, meaning  $\pi(x) = \pi(0) \exp(Tx)$  for some size  $m$  matrix  $T$  and  $\pi(0) = -\theta T$ . The matrix  $T$  is the minimal solution to the non-linear integral equation

$$T = S + \int_0^\infty \exp(Tu) dA(u). \tag{13}$$

4.1. An order  $m^2$  representation

**Theorem 2.** The type  $\tau \subseteq \{1, \dots, m\}$  queue length distribution  $N_b(\tau)$  has a matrix geometric representation  $(\kappa, K, \nu)$  of order  $m^2$  given by

$$\kappa = \xi^T(I_m \otimes \Delta(\theta)), \quad K = (I - K(\tau^c))^{-1}K(\tau) \tag{14}$$

and

$$\nu = (I - K(\tau^c))^{-1}((-S)^{-1}\Delta(e_\tau)(-S)e \otimes e) \tag{15}$$

with

$$K(\zeta) = ((-S)^{-1}\Delta(e_\zeta) \otimes I_m) \int_{x=0}^{\infty} (dA(x) \otimes \exp(Qx)), \tag{16}$$

where  $\tau^c = \{1, \dots, m\} \setminus \tau$ ,  $Q = \Delta^{-1}(\theta)T^T \Delta(\theta)$ ,  $\Delta(x)$  denotes a diagonal matrix such that  $\Delta(x)e = x^T$  and  $e_\zeta$  for  $\zeta$  a subset of  $\{1, \dots, m\}$ , is a vector with entry  $i$  equal to one if  $i \in \zeta$  and zero otherwise. Furthermore, the vector  $\kappa$  is stochastic and the matrix  $K$  is substochastic.

**Proof.** The proof is analogue to the discrete-time case and as such presented in a more compact form. By making use of the age process, the probability of having  $n$  type  $\tau$  customers in the queue at an arbitrary busy time epoch is given by

$$P[N_b(\tau) = n] = \int_{x=0}^{\infty} \pi(x)P_{\tau, n-1}(x)edx,$$

where the  $(i, j)$ -th element of  $P_{\tau, n}(x)$  holds the following probability: given an arrival in phase  $i$  at time 0, entry  $(i, j)$  holds the probability that either (i) the arrival at time zero is of type  $\tau^c$  and  $n$  type  $\tau$  arrivals occur in the interval  $(0, x]$ , while the phase is  $j$  when the first arrival at time  $t > x$  occurs or (ii) if  $n > 0$ , the arrival at time zero is of type  $\tau$  and  $n - 1$  type  $\tau$  arrivals occur in the interval  $(0, x]$ , while the phase is  $j$  when the first arrival at time  $t > x$  occurs. Expanding the matrix exponential  $\exp(Tx)$  and applying the aforementioned Lemma of Ozawa [12] with  $a^T = P_{\tau, n-1}(x)e$  and  $b = \pi(0)T^k$  results in

$$P[N_b(\tau) = n] = (e^T \otimes \pi(0)) \int_{x=0}^{\infty} (P_{\tau, n-1}^T(x) \otimes \exp(Tx)) dx\xi.$$

By definition  $P_n(0) = 0$  for  $n > 0$  and

$$P_{\tau, 0}(x) = \int_{u=x}^{\infty} d\bar{A}_{\tau^c}(u) + \int_{u=0}^x d\bar{A}_{\tau^c}(u)P_{\tau, 0}(x-u)$$

$$P_{\tau, n}(x) = 1_{[n=1]} \left( \int_{u=x}^{\infty} d\bar{A}_{\tau}(u) \right) + \int_{u=0}^x (d\bar{A}_{\tau}(u)P_{\tau, n-1}(x-u) + d\bar{A}_{\tau^c}(u)P_{\tau, n}(x-u))$$

for  $n > 1$ , where  $d\bar{A}_{\zeta}(x) = (-S)^{-1}\Delta(e_\zeta)dA(x)$  and  $1_{[n=1]}$  equals 1 for  $n = 1$  and zero otherwise. This allows us to rewrite  $\int_{x=0}^{\infty} (P_n^T(x) \otimes \exp(Tx)) dx$  which, similar to the discrete-time case, yields

$$P[N_b = n] = \alpha(I - M(\tau^c))^{-1}M^{n-1}\xi,$$

where  $M(\zeta) = \int_{x=0}^{\infty} (d\bar{A}_{\zeta}^T(x) \otimes \exp(Tx))$ ,  $M = M(\tau)(I - M(\tau^c))^{-1}$  and

$$\alpha = \left( \int_{x=0}^{\infty} \left( \int_{u=x}^{\infty} (d\bar{A}_{\tau^c}(u)e)^T \right) \otimes \pi(0) \exp(Tx)dx \right) (I - M(\tau^c))^{-1}M(\tau)$$

$$+ \int_{x=0}^{\infty} \left( \int_{u=x}^{\infty} (d\bar{A}_{\tau}(u)e)^T \right) \otimes \pi(0) \exp(Tx)dx.$$

By switching the order of the integrations, noting that  $-\pi(0)T^{-1}$  is the unique stochastic invariant vector of  $\int_{u=0}^{\infty} dA(u) + S$ , which we denoted as  $\theta$ , and remarking that  $\int_{x=0}^{\infty} (d\bar{A}_{\tau}(x) + d\bar{A}_{\tau^c}(x))e = e$ ,  $\alpha$  can be simplified to  $\alpha = (e^T \otimes \theta)(S^T(\tau) \otimes I_m)(I - M)$ , where  $S^T(\tau) = ((-S)^{-1}\Delta(e_\tau)(-S))^T$ .

This implies that  $N_b(\tau)$  has a matrix geometric representation  $((e^T \otimes \theta)(I - M(\tau^c))^{-1}, M, \xi)$  of order  $m^2$ , which can be transformed into the matrix geometric representation  $(\kappa, K, \nu)$  as  $\theta > 0$  (due to the irreducibility assumption).  $K(\tau) + K(\tau^c)$  is strictly substochastic as  $Q = \Delta(\theta)^{-1}T^T \Delta(\theta)$  was shown to be the generator of a transient Markov chain by Sengupta and  $\int_{x=0}^{\infty} (d\bar{A}_{\tau}(x) + d\bar{A}_{\tau^c}(x))e = e$ . This suffices to show that the inverse  $(I - K(\tau^c))^{-1}$  exists and that  $K$  is substochastic as well. The vector  $\kappa$  is clearly stochastic.  $\square$

**Corollary 2.** The overall queue length distribution  $N_b$  has a phase-type representation  $(\kappa, K)$  of order  $m^2$  given by

$$\kappa = \xi^T (I_m \otimes \Delta(\theta)), \quad K = \int_{x=0}^{\infty} (d\bar{A}(x) \otimes \exp(Qx)), \quad (17)$$

where  $d\bar{A}(u) = (-S)^{-1}dA(u)$  and  $Q = \Delta(\theta)^{-1}T^T \Delta(\theta)$ .

#### 4.2. Redundancy of the representation

Proving that examples exist for which the queue length distribution has a minimal order of  $m^2$  is more difficult in continuous time. Mostly because we wish to find an explicit expression for the phase-type representation  $(\kappa, K)$  and its first few moments, such that an exact evaluation of the determinant of its Hankel matrix can be performed. Thus, we need to specify the  $S$  and  $P(u)$  matrices such that  $T$  and its matrix exponential can be expressed explicitly, where  $T$  was a solution to the integral equation (13). As opposed to the discrete-time setting, where examples can be constructed such that  $R$  is known explicitly, there is no continuous-time analog for which  $T$  is known explicitly.

To construct such a rational  $T$ , we make use of a queue somewhat similar to the one considered in Section 3.2, that is, we define  $m = 2$ ,  $\tau = \{1, 2\}$ ,  $P(1) = Q(1)$  and  $P(2) = Q(1) + Q(2)$  and set  $dP(u) = 0$ , for all  $u \neq 1, 2$ . Notice, even for the continuous-time case  $P(u)$  may be chosen as a discrete distribution, which implies that the Stieltjes integration in (13) is a simple summation. Denote the  $2 \times 2$  rate matrix  $S$  as

$$S = \begin{bmatrix} -x_1 & x_2 \\ x_3 & -x_4 \end{bmatrix}.$$

Next we determine  $S$  such that  $T$  is the rational matrix below

$$T = \begin{bmatrix} -1 & 3/4 \\ 1/3 & -1 \end{bmatrix},$$

with matrix exponential

$$\exp(T) = \begin{bmatrix} \frac{1}{2}e^{-\frac{1}{2}} + \frac{1}{2}e^{-\frac{3}{2}} & \frac{3}{4}e^{-\frac{1}{2}} - \frac{3}{4}e^{-\frac{3}{2}} \\ \frac{1}{3}e^{-\frac{1}{2}} - \frac{1}{3}e^{-\frac{3}{2}} & \frac{1}{2}e^{-\frac{1}{2}} + \frac{1}{2}e^{-\frac{3}{2}} \end{bmatrix}.$$

Because of (13), it suffices to solve the linear system with 4 equations and 4 unknowns

$$T = S + \exp(T) \begin{bmatrix} 0 & x_1 - x_2 \\ 0 & 0 \end{bmatrix} + \exp(2T) \begin{bmatrix} 0 & 0 \\ x_4 - x_3 & 0 \end{bmatrix},$$

because of the form of the matrices  $P(u)$ . As  $T$ ,  $\exp(T)$  and its square  $\exp(2T)$  are known explicitly, the solution for  $x_1, x_2, x_3$  and  $x_4$  can be determined as

$$\begin{aligned} x_1 &= \frac{16q^6 + 8q^5 + 24q^4 + 5q^3 + 21q^2 + 5q + 5}{8q^2(2q^4 + q^3 + 2q^2 + 1)} \\ x_2 &= \frac{12q^7 - 8q^6 + 6q^5 - 19q^4 + 6q^3 - 16q^2 - 5}{8q^2(2q^5 - q^4 + q^3 - 2q^2 + q - 1)} \\ x_3 &= \frac{8q^8 + 4q^7 - 8q^6 - 9q^5 - 9q^4 - 12q^2 - 5q - 5}{12q^2(2q^6 + q^5 - q^3 - q^2 - 1)} \\ x_4 &= \frac{24q^6 + 14q^5 + 26q^4 + 5q^3 + 17q^2 + 5q + 5}{12q^2(2q^4 + q^3 + 2q^2 + 1)}, \end{aligned}$$

with  $q = e^{1/2}$  and  $S$  turns out to be a well-defined rate matrix. Using the expression for the exponential of  $T$  we can compute the phase-type representation  $(\kappa, K)$  in explicit form, as well as its corresponding Hankel matrix, allowing us to conclude that the order 4 representation is indeed minimal.

#### 4.3. Computation of the queue length distribution

##### 4.3.1. General case

The main step in computing the representation  $(\kappa, K, \nu)$  via (14) to (16) is to determine the matrix  $Q = \Delta(\theta)^{-1}T^T \Delta(\theta)$ , where  $T^T$  is the transposed matrix of  $T$ . As indicated in [5], the matrix  $T$  can be computed by setting  $T_0 = S$  and letting

$$T_{n+1} = S + \int_0^{\infty} \exp(T_n u) dA(u), \quad (18)$$

for  $n \geq 0$  until  $|T_{n+1} - T_n|$  is below some predefined parameter  $\epsilon$  (e.g.,  $\epsilon = 10^{-10}$ ). One could also compute  $Q$  directly by defining a dual process. Indeed, if we extend the Ramaswami dual [14] to M/G/1- and GI/M/1-type Markov chains with a continuous level, the dual of our age process becomes an M/G/1-type Markov chain with a continuous level as introduced by Takine [21]. As opposed to the discrete-time case, there is however no gain in doing so, as the iterative algorithm to compute  $Q$  is such that  $Q_n$ , the matrix obtained after  $n$  steps, can be written as  $\Delta(\theta)^{-1} T_n^T \Delta(\theta)$ . Before we move on to the Markovian inter-arrival times, we briefly discuss two other special cases in which numerical integration can be avoided.

*Discrete inter-arrival times.* Assume  $P(u)$  is a step-function with steps occurring at  $t_1, t_2, \dots$  and define  $Q(t_1) = P(t_1)$  and  $Q(t_i) = P(t_i) - P(t_{i-1})$  for  $i > 1$ , then (18) reduces to

$$T_{n+1} = S + \sum_{i=1}^{\infty} \exp(T_n t_i) Q(t_i),$$

as in the example in Section 4.2.

*Some well-known functions  $P_{i,j}(u)$ .* As explained in [1,5] the integration in (18) can be avoided if

$$\int_0^{\infty} \exp(T_n u) dP_{i,j}(u)$$

can be expressed in terms of  $T_n$  for all  $i, j \in \{1, \dots, m\}$  with  $P_{i,j}(u)$  the  $(i, j)$ -th entry of  $P(u)$ . Sengupta lists various examples for  $P_{i,j}(u)$  such as the uniform distribution on  $[a, b]$ , the gamma distribution with parameters  $(n, \alpha)$ , etc. The numerical integration in (16) can be avoided in a similar manner.

#### 4.3.2. Markovian inter-arrival times

*Avoiding numerical integration* Assume  $dP(u)$  can be written as  $dP(u) = \exp(\bar{D}_0 u) \bar{D}_1 du$  for some  $m \times m$  matrices  $\bar{D}_0$  and  $\bar{D}_1$ , such that  $(\bar{D}_0, \bar{D}_1)$  characterizes a MAP process. Notice, the actual arrival process of the semi-Markovian queue is not the MAP characterized by  $(\bar{D}_0, \bar{D}_1)$  as the service process affects the arrival process as well. In this case, we can express  $T_{n+1}$  as

$$T_{n+1} = S + \int_0^{\infty} \exp(T_n u) S^* \exp(\bar{D}_0 u) du \bar{D}_1,$$

and by applying integration by parts we find that  $T_{n+1}$  can be written as  $S + X_n \bar{D}_1$ , where  $X_n$  is the solution to the linear system

$$X_n \bar{D}_0 + T_n X_n = -S^*.$$

The equation for  $X_n$  is a Sylvester matrix equation that can be solved in  $O(m^3)$  time [22]. This approach is a generalization of the method used in [10] to compute  $T$  for the MMAP[K]/PH[K]/1 queue if we define  $\bar{D}_0 = (I \otimes D_0)$  and

$$\bar{D}_1 = [\alpha_1 \otimes D_1 \quad \dots \quad \alpha_K \otimes D_K],$$

where  $(D_0, D_1, \dots, D_K)$  characterizes the MMAP[K] arrival process and the service of a type  $k$  customer is phase-type  $(\alpha_k, S_k)$  (see also Section 2(c)). Due to  $\exp(A) \otimes \exp(B) = \exp(A \oplus B)$ , we can also simplify (16) to

$$K(\zeta) = -((-S)^{-1} \Delta(e_{\zeta}) S^* \otimes I_m) (\bar{D}_0 \oplus Q)^{-1} (\bar{D}_1 \otimes I_m).$$

*Reduction to fluid queues.* The iterative solution above that relies on solving a Sylvester matrix equation during each step has linear convergence only and might require many iterations. To obtain a solution method with quadratic convergence, we can construct a fluid queue [23], by replacing the immediate downward jumps by intervals of the appropriate length during which the level decreases linearly. In other words, we obtain a fluid queue with  $m$  phases in which the fluid increases (our original  $m$  states) and  $m$  phases in which the fluid decreases (the  $m$  artificial states that are added) such that  $F_{++} = S$ ,  $F_{+-} = S^*$ ,  $F_{-+} = \bar{D}_1$  and  $F_{--} = \bar{D}_0$ , where  $F_{++}, F_{+-}, F_{-+}$  and  $F_{--}$  holds the rates at which the phase changes while going up, from an up to a down phase, from a down to an up phase and while going down, respectively. Notice,  $F$  defined as

$$F = \begin{bmatrix} F_{++} & F_{+-} \\ F_{-+} & F_{--} \end{bmatrix},$$

is the rate matrix of a continuous-time Markov chain. If we take the expression for the steady state of a fluid queue [23] and observe the queue only when the level increases, one finds that its steady state  $\pi(x)$  has a matrix exponential form  $\pi(x) = \pi(0) \exp(Tx)$ , with  $T = F_{++} + \Psi F_{-+}$  where  $\Psi$  is the minimal non-negative solution to an algebraic Riccati equation [23]. Thus, to compute  $T$ , it suffices to determine  $\Psi$  and this can be done by constructing the following discrete-time Quasi-Birth-Death Markov chain [24] (where  $\hat{A}_0$  captures the upward and  $\hat{A}_2$  the downward transitions):

$$\hat{A}_0 = \begin{bmatrix} 1 & & \\ 2 & -I & \\ & & 0 \end{bmatrix}, \quad \hat{A}_1 = \begin{bmatrix} 1 & & \\ 2 & P_{++} & \\ & & 0 \end{bmatrix}, \quad \hat{A}_2 = \begin{bmatrix} 0 & 1 \\ & 2 & P_{+-} \\ 0 & & P_{--} \end{bmatrix},$$

where  $P$  is the transition matrix of a discrete-time Markov chain obtained from  $F$  by means of a uniformization and is partitioned in the same manner as  $F$  (i.e.,  $P = I + F/\mu$  with  $\mu = \max_i |F_{ii}|$ ). Let the  $2m \times 2m$  matrix  $\hat{G}$  be the smallest non-negative solution to

$$\hat{G} = \hat{A}_2 + \hat{A}_1 \hat{G} + \hat{A}_0 \hat{G}^2,$$

then  $\Psi$  is found as the upper right corner of  $\hat{G}$  [24] as

$$\hat{G} = \begin{bmatrix} 0 & \Psi \\ 0 & P_{--} + P_{-+} \Psi \end{bmatrix}.$$

In conclusion,  $T$  can be computed via  $\Psi$  for which we can use any algorithm with quadratic convergence such as cyclic or logarithmic reduction [19], the runtime of which can even be further improved by exploiting the internal structure of the matrices  $\hat{A}_0$ ,  $\hat{A}_1$  and  $\hat{A}_2$  (e.g., see [25]). Alternatively, the algebraic Riccati equation can also be solved with quadratic convergence using the Newton iteration [26], which requires the solution of a Sylvester matrix equation [22] of the form  $AX + XB = C$  during each iteration and therefore has a similar time complexity as the Quasi-Birth-Death approach discussed above.

#### 4.4. Special case: overall queue length of the MAP/MAP/1 queue

For the special case of the continuous-time MAP/MAP/1 queue with  $(D_0, D_1)$  characterizing the arrival process and  $(S_0, S_1)$  the service process we have  $S = I \otimes S_0$  and  $dA(u) = \exp(D_0 u) D_1 \otimes S_1$ . Hence,

$$d\bar{A}^T(u) = (D_1 \otimes (-S_0)^{-1} S_1^T \exp(D_0^T u) \otimes I).$$

and

$$M = -((D_1 \otimes (-S_0)^{-1} S_1^T \otimes I)((D_0^T \otimes I) \oplus T)^{-1}.$$

By remarking that  $P_0(x) = \int_{u=x}^{\infty} d\bar{A}(u) = (\exp(D_0 x) \otimes I)((-D_0)^{-1} D_1 \otimes (-S_0)^{-1} S_1)$ , we have

$$\alpha = -(e^T \otimes \pi(0))((D_0^T \otimes I) \oplus T)^{-1}.$$

Recall,  $\pi(0)$  equals  $-\theta T$  and  $\theta$  is the unique stationary vector of  $((-D_0)^{-1} D_1 \otimes S_1) + (I \otimes S_0)$  for the MAP/MAP/1 queue. This order  $m^2$  representation is clearly redundant as an order  $m$  representation can be obtained directly from the  $R$  matrix of the Quasi-Birth-Death Markov chain where the level represents the number of customers in the queue, that is,  $R$  is the smallest non-negative solution to

$$0 = (D_1 \otimes I) + R(D_0 \oplus S_0) + R^2(I \otimes S_1),$$

which can be computed using an algorithm with quadratic convergence (e.g., by cyclic or logarithmic reduction [19]). This also raises the question whether  $T$  could be computed from  $R$ .

To express  $T$  via  $R$  for the MAP/MAP/1 queue, we first note that for the more general SM/MAP/1 queue Sengupta [1, Theorem 6 and Eq. (15)] showed that

$$T = (I \otimes S_0) + \tilde{R}(I \otimes S_1),$$

where  $\tilde{R}$  is the  $R$ -matrix of the discrete-time GI/M/1-type Markov chain obtained by observing the queue length at arrival times only. Using Theorem 1 of Ozawa [12], we find that

$$\tilde{R} = (-U)^{-1}(D_1 \otimes I),$$

with  $U = (D_0 \oplus S_0) + R(I \otimes S_1) = (D_0 \oplus S_0) + (D_1 \otimes I)G$ . This allows us to conclude

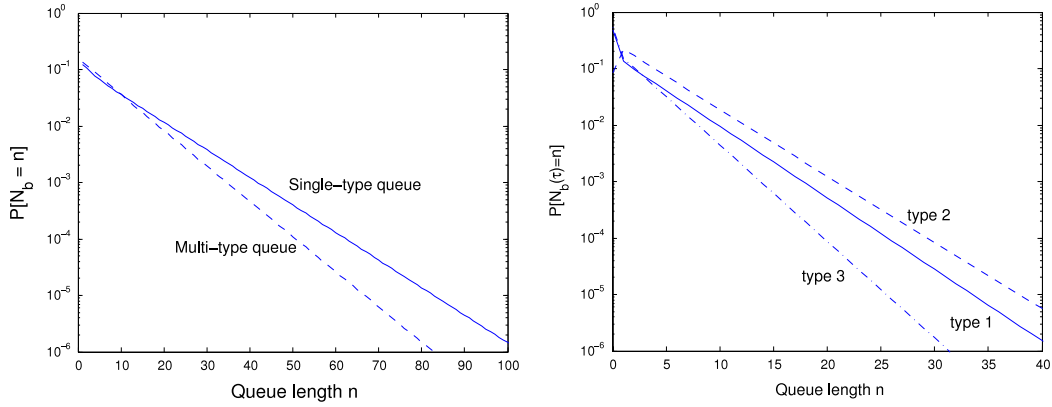
$$T = (I \otimes S_0) + (-U)^{-1}(D_1 \otimes S_1),$$

where  $U$  can be expressed via  $R$  as indicated above. Using either this new relation or the fluid queue reduction presented in the previous section, we can significantly outperform existing methods [27] to compute the order  $m$  phase-type distribution for the waiting time in a MAP/MAP/1 queue.

## 5. Numerical examples

We conclude by presenting a number of numerical examples. We restrict ourselves to two continuous-time examples as these are slightly more challenging. Discrete-time examples can be generated as well and can be solved even faster as we can rely on algorithms with quadratic convergence even in the general case.

We start with an example of an MMAP[K]/PH[K]/1 queue with three types of customers to validate our results with the more involved method in [10]. The first type of customers require an Erlang-2 amount of service with rate parameter  $\lambda = 1$ ,



**Fig. 1.** Overall (left) and per type (right) queue length distribution for an MMAP[K]/PH[K]/1 queue with and without correlation between the service and inter-arrival time.

the second follow a size 3 Coxian distribution with  $(\lambda_1, \lambda_2, \lambda_3) = (1/2, 1/3, 1/2)$  and the service time of the third class is exponential with rate  $\lambda = 1/5$ . The MMAP[K] arrival process is characterized by

$$D_0 = \begin{bmatrix} -104/500 & 4/500 \\ 6/700 & -106/700 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 1/10 & 0 \\ 0 & 0 \end{bmatrix},$$

$$D_2 = \begin{bmatrix} 1/10 & 0 \\ 0 & 1/14 \end{bmatrix}, \quad D_3 = \begin{bmatrix} 0 & 0 \\ 0 & 1/14 \end{bmatrix}.$$

Hence, this queue has periods with arrivals of type-1 and type-2 followed by periods with arrivals of type-2 and type-3 with a lower arrival rate. Furthermore, 30% of the customers are type-1, 50% type-2 and 20% type-3. As  $m = 12$  in this particular example,  $K$  is a size 144 matrix. The queue length distribution  $N_b$  is depicted on the left in Fig. 1 and is labeled *Multi-type queue*. These results are in perfect agreement with the method presented in [10] and implemented in [27]. The computation time required was less than 0.25 seconds using the Sylvester equation approach discussed in Section 4.3.2. On this particular example, the fluid queue approach of Section 4.3.2 further reduced the computation time of the matrix  $T$  by a factor 3.

When we ignore the correlation between the customer types, the queue is reduced to an MAP/PH/1 queue where the arrival process is the two-state MAP characterized by  $(D_0, D_1 + D_2 + D_3)$  and the phase-type service is of order 6. In Fig. 1 we show the impact of neglecting the correlation in the customer types, where we labeled the results as *Single-type queue*. We clearly see that the queue length distribution is highly affected by the correlation between the service and inter-arrival times, even the mean queue length increases by more than 20%. The queue length increases because the MAP state with the highest arrival rate produces type-1 and type-2 customers (with equal probability), while the state with the lower rate created type-2 and type-3 customers. The mean service time of the type-3 customers is however 2.5 times as high as the type-1 customers, meaning during the periods where the arrival rate is higher, the average amount of work per customer is less. Neglecting this results in longer queue lengths.

The per type queue length distribution is shown on the right in Fig. 1 and the results are in perfect agreement with the more elaborate method presented in [10] and implemented in [27]. As indicated in Section 2, the type  $k$  queue length distribution, for  $k = 1, 2$  and  $3$ , is obtained by properly defining  $\tau$  (i.e., for type 1 we have  $\tau = \{1, \dots, 4\}$ , for type 2 we set  $\tau = \{5, \dots, 10\}$  and for type 3 we have  $\tau = \{11, 12\}$ ).

In the second example we consider a queue that does not belong to the SM[K]/PH[K]/1 class with  $m = 2$ . The matrices  $S$  and  $P(\infty) = \int dP(u)$  are given by

$$S = \begin{bmatrix} -1/14 - 1/(40r) & 1/(40r) \\ 1/(20r) & -1/27 - 1/(20r) \end{bmatrix},$$

and

$$P(\infty) = \begin{bmatrix} 1 - 1/(5r) & 1/(5r) \\ 1/(6r) & 1 - 1/(6r) \end{bmatrix},$$

with  $r = 100$  or  $1000$ . Moreover, when the service of a customer ends in phase 1 (2), the inter-arrival time is uniform between 10 and 25 (between 20 and 35) and the phase after the arrival changes according to  $P(\infty)$ . Notice, this queue tends to have fairly long periods with uniform inter-arrival times between 10 and 25 and exponential services with mean length 14, typically followed by periods of uniform inter-arrival times between 20 and 35 and exponential service times with mean 27. Thus, it has periods with a load close to one and periods with a substantially lower load, such that the overall load is 86.6%.

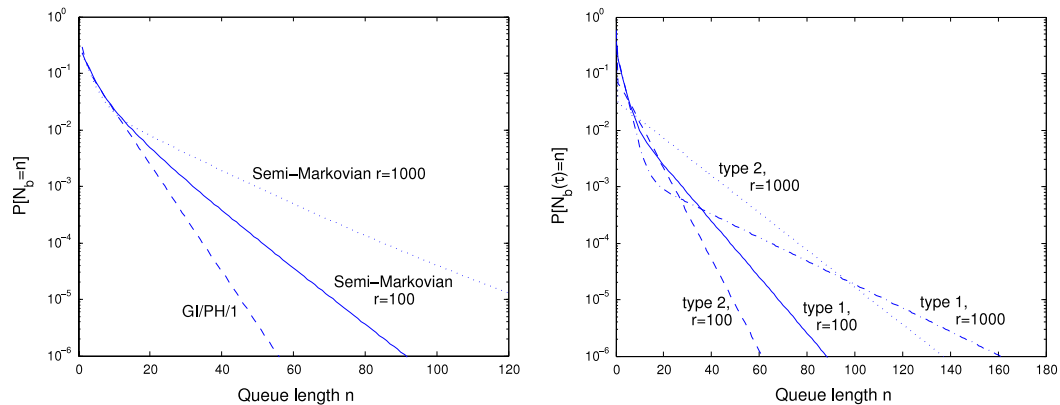


Fig. 2. Overall (left) and per type (right) queue length distribution for a semi-Markovian queue and its corresponding GI/PH/1 queue.

The queue length distribution can be computed in a fraction of a second and the results for the overall queue length are shown on the left in Fig. 2 for  $r = 100$  and  $r = 1000$  (labeled *Semi-Markovian queue*). To compute  $T$  we made use of the method presented in Section 4.3 (Well-known functions). If we were to neglect all the types of correlation we end up with a GI/PH/1 queue (for which the value of  $r$  is irrelevant) that does not exhibit the behavior above and is therefore far too optimistic with respect to the queue length distribution (see Fig. 2, labeled *GI/PH/1 queue*). On the right Fig. 2 shows the per type queue length distribution for  $r = 100$  and  $r = 1000$ . Recall, the type 1 customers arrive in low load periods, while the type 2 customers arrive in high load periods, where the mean length of these periods increases with  $r$ . As expected the mean queue length of the type 1 customers decreases with  $r$ , while the mean type 2 queue length increases. Somewhat less expected, the tail of the type 1 customers decays *slower* than the tail of the type 2 customers. This can be understood by remarking that the first type 1 customers that arrive during a low load period still have to face the queue created during the high load period, while the arrival rate in the low load period is higher than in the high load period. Indeed, the point of intersection between the type 1 and 2 queue length distributions decreases rapidly as  $r$  increases.

## References

- [1] B. Sengupta, The semi-Markovian queue: theory and applications, *Stochastic Models* 6 (3) (1990) 383–413.
- [2] J. Lambert, B. Van Houdt, C. Blondia, Queues with correlated service and inter-arrival times and their application to optical buffers, *Stochastic Models* 22 (2) (2006) 233–251.
- [3] J. Lambert, B. Van Houdt, C. Blondia, Queues in DOCSIS cable modem networks, *Computers & Operations Research* 35 (8) (2008) 2482–2496.
- [4] J.H.A. Smit de, The single server semi-Markov queue, *Stochastic Processes and Their Applications* 22 (1) (1986) 37–50.
- [5] B. Sengupta, Markov processes whose steady state distribution is matrix exponential with an application to the GI/PH/1 queue, *Advances in Applied Probability* 21 (1989) 159–180.
- [6] M. Neuts, *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach*, John Hopkins University Press, 1981.
- [7] B. Van Houdt, C. Blondia, The delay distribution of a type  $k$  customer in a first come first served MMAP[K]/PH[K]/1 queue, *Journal of Applied Probability* 39 (1) (2002) 213–222.
- [8] B. Van Houdt, C. Blondia, The waiting time distribution of a type  $k$  customer in a MMAP[K]/PH[K]/c ( $c = 1, 2$ ) queue using QBDs, *Stochastic Models* 20 (1) (2004) 55–69.
- [9] Q. HE, Age process, workload process, sojourn times, and waiting times in a discrete-time SM[K]/PH[K]/1/FCFS queue, *Queueing Systems* 49 (2005) 363–403.
- [10] Q. HE, Analysis of a continuous time SM[K]/PH[K]/1/FCFS queue: age process, sojourn times, and queue lengths, Working paper 04-01, Department of Industrial Engineering, Dalhousie University (2004).
- [11] I. Adan, V. Kulkarni, Single-server queue with Markov-dependent inter-arrival and service times, *Queueing Systems and its Applications* 45 (2003) 113–134.
- [12] T. Ozawa, Sojourn time distributions in the queue defined by a general QBD process, *Queueing Systems and its Applications* 53 (4) (2006) 203–211.
- [13] A. Horváth, G. Horváth, M. Telek, A joint moments based analysis of networks of MAP/MAP/1 queues, in: *QEST '08: Proceedings of the 2008 Fifth International Conference on Quantitative Evaluation of Systems*, IEEE Computer Society, 2008, pp. 125–134. <http://dx.doi.org/10.1109/QEST.2008.26>.
- [14] V. Ramaswami, A duality theorem for the matrix paradigms in queueing theory, *Stochastic Models* 6 (1) (1990) 151–161.
- [15] M. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, Inc., New York and Basel, 1989.
- [16] Q. HE, H. Zhang, On matrix exponential distributions, *Advances in Applied Probability* 39 (1) (2007) 271–292.
- [17] W.B. Gragg, A. Lindquist, On the partial realization problem, *Linear Algebra and its Applications* 50 (1983) 277–319.
- [18] A. van de Liefvoort, The moment problem for continuous distributions, Tech. Rep., University of Missouri (1990).
- [19] D. A. Bini, B. Meini, S. Steffé, B. Van Houdt, Structured Markov chains solver: algorithms, in: *SMCtools Workshop*, ACM Press, Pisa, Italy, 2006.
- [20] P. Fernandes, B. Plateau, W.J. Stewart, Efficient descriptor-vector multiplications in stochastic automata networks, *Journal of the ACM* 45 (3) (1998) 381–414. <http://doi.acm.org/10.1145/278298.278303>.
- [21] T. Takine, A continuous version of matrix-analytic methods with skip-free to the left property, *Stochastic Models* 12 (4) (1996) 673–682.
- [22] G.H. Golub, S. Nash, C. Van Loan, A Hessenberg–Schur method for the problem  $AX + XB = C$ , *IEEE Transactions on Automatic Control* 24 (1979) 909–913.
- [23] G. Latouche, Structured Markov chains in applied probability and numerical analysis, in: *Markov Anniversary Meeting, 2006*, pp. 69–78.
- [24] V. Ramaswami, Matrix analytic methods for stochastic fluid flows, in: *Teletraffic Engineering in a Competitive World – Proc. of the 16th International Teletraffic Congress (ITC 16)*, Elsevier Science B.V., 1999, pp. 1019–1030.

- [25] C.-H. Guo, Efficient methods for solving a nonsymmetric algebraic Riccati equation arising in stochastic fluid models, *Journal of Computational and Applied Mathematics* 192 (2006) 353–373.
- [26] C.-H. Guo, Nonsymmetric algebraic Riccati equations and Wiener–Hopf factorization for M-matrices, *SIAM Journal on Matrix Analysis and Applications* 23 (2001) 225–242.
- [27] J. F. Pérez, J. Van Velthoven, B. Van Houdt, Q-MAM: a tool for solving infinite queues using matrix-analytic methods, in: *ValueTools '08: Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools, 2008*, pp. 1–9.  
<http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4368>.



**Benny Van Houdt** received his M.Sc. degree in mathematics and computer science, and a Ph.D. in science from the University of Antwerp (Belgium) in July 1997, and May 2001, respectively. From August 1997 until September 2001 he held an Assistant position at the University of Antwerp, while from October 2001 until September 2007 he has been a postdoctoral fellow of the FWO-Flanders. In 2007, he became a professor at the Mathematics and Computer Science Department of the University of Antwerp, where he is a leading member of the PATS research group. His main research interest goes to the performance evaluation and stochastic modelling of communication networks. Other areas of interest include manufacturing, operating systems, tool development, etc. He has published various papers, containing both theoretical and practical contributions, in a variety of international journals (e.g., *IEEE JSAC*, *Performance Evaluation*, *Journal of Applied Probability*, *Stochastic Models*, *Queueing Systems*, etc.) and in conference proceedings (e.g., *ACM Sigmetrics*, *Networking*, *Globecom*, *Opticomm*, *ITC*, etc.).