manuscript No.
(will be inserted by the editor)

Analysis of generalized QBD queues with matrix-geometrically distributed batch arrivals and services

Gábor Horváth

the date of receipt and acceptance should be inserted later

Abstract In a QBD (quasi birth-death) queue, the level forward and level backward transitions of a QBD-type Markov chain are interpreted as customer arrivals and services. In the generalized QBD queue considered in this paper arrivals and services can occur in matrix-geometrically distributed batches. This paper presents the queue length and sojourn time analysis of generalized QBD queues. It is shown that, if the number of phases is N, the number of customers in the system is order-N matrix-geometrically distributed, and the sojourn time is order-N^2 matrix-exponentially distributed, just like in the case of classical QBD queues without batches. Furthermore, phase-type representations are provided for both distributions. In the special case of the arrival and service processes being independent, further simplifications make it possible to obtain a more compact, order-N representation for the sojourn time distribution.

Keywords Matrix-analytic methods · age process · batch arrival · batch service · queue length analysis · sojourn time analysis

Mathematics Subject Classification (2000) 60K25 · 68M20

1 Introduction

Several solution procedures exist for the stationary analysis of Markov chains with a regular structure. Over the last three decades, matrix analytic methods have been developed for the efficient solution of M/G/1-type Markov

G. Horváth
Budapest University of Technology and Economics, Dept. of Networked Systems and Services
MTA-BME Information Systems Research Group
Magyar tudósok krt 2., 1117 Budapest, Hungary
Tel.: +36-1-4633254
Fax: +36-1-4633263
E-mail: ghorvath@hit.bme.hu

chains (where the generator matrix has an upper block Hessenberg form, [Neuts(1989)]), for the GI/M/1-type Markov chains (where the generator matrix has a lower block Hessenberg form, [Neuts(1981)]), and for the GI/G/1-type Markov chains (which have a dense generator having a block Toeplitz structure, [Gail et al(1997)Gail, Hantler, and Taylor]).

Quasi birth-death processes (QBDs), Markov chains with a regular block-tridiagonal structure proved especially successful in queueing theory. There are many books available on QBDs (the first extensive one is [Neuts(1981)]), and thousands of papers were published where QBDs are applied to solve practical problems.

QBD queues are FCFS (first-come first-served) queues, which are closely related to QBD processes: the level forward and level backward transitions of a QBD process are interpreted as customer arrivals and services in the corresponding QBD queue. Since the stationary distribution of a QBD process is matrix-geometric, the number of customers in a QBD queue is order N matrix-geometrically distributed as well, given that the number of phases is N. The sojourn time of QBD queues has been studied in [Ozawa(2006)], where an order N^2 matrix-exponential distribution is derived for the sojourn time.

In this paper an extension of the QBD queue is analyzed, where batch arrivals and services are both allowed. We show that if the batch sizes have a matrix-geometric form, the number of customers in the system is order-N matrix-geometrically distributed and the sojourn time is order-N^2 matrix-exponentially distributed, just like in the case without batches. Furthermore, we show that phase-type representations exist for both distributions. The special case with independent arrival and service processes is also investigated. For this case, we were able to obtain a more compact, order-N matrix-exponential distribution for the sojourn time.

A system somewhat similar to the one studied in this paper has been considered in [Éltető and Telek(2008)], but it is not as general as ours: batch services are not allowed, the batch size can not depend on the phase of the background process, and the arrival and service processes are assumed to be independent as well. The representation for the queue length distribution obtained there is not minimal, and the sojourn time is not analyzed either.

The system considered in [Jafari and Sohraby(2001)] is of greater relevance, since the Markov chain studied there is identical to the one investigated by this paper. However, this paper still has several contributions. While the steady-state solution is derived with the tools of system theory (and the invariant subspace approach) in [Jafari and Sohraby(2001)], we provide a purely matrix-analytic solution here[1]. Furthermore, the sojourn time analysis and the derivation of phase type representations, which are among the objectives of this paper, are not provided in [Jafari and Sohraby(2001)].

[1] The matrix-analytic and the invariant subspace approach coexist for ordinary QBDs and other more advanced queueing models.

2 Model description

QBD queues are First-Come-First-Served queues with a continuous time Markov chain $\{\mathcal{J}(t), t > 0\}$ in the background. Some marked transitions of this Markov chain lead to an arrival of a customer increasing the number of customers in the system (denoted by $\{\mathcal{X}(t), t > 0\}$) by one. Some other transitions are accompanied by a service of a customer (decreasing $\mathcal{X}(t)$ by one), and the rest of the transitions are internal in the sense that they do not change the length of the queue. The generator of the two-dimensional Markov process $\{\mathcal{X}(t), \mathcal{J}(t)\}$ has a QBD (block-tridiagonal) structure.

In this paper we study a more general system where customers arrive and are served in batches. The generator of the Markov chain $\{\mathcal{X}(t), \mathcal{J}(t)\}$ denoted by $\mathbf{Q}$ is dense, we have

$$\mathbf{Q} = \begin{bmatrix} \mathbf{L_0} & \mathbf{F_1} & \mathbf{F_2} & \mathbf{F_3} & \mathbf{F_4} & \cdots \\ \bar{\mathbf{B}}_\mathbf{1} & \mathbf{L} & \mathbf{F_1} & \mathbf{F_2} & \mathbf{F_3} & \cdots \\ \bar{\mathbf{B}}_\mathbf{2} & \mathbf{B_1} & \mathbf{L} & \mathbf{F_1} & \mathbf{F_2} & \cdots \\ \bar{\mathbf{B}}_\mathbf{3} & \mathbf{B_2} & \mathbf{B_1} & \mathbf{L} & \mathbf{F_1} & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{1}$$

where all matrix blocks are of size N.

However, the matrices corresponding to level forward and level backward transitions, $\mathbf{F_k}$ and $\mathbf{B_k}$ (for $k \geq 1$) are not arbitrary, they are defined by a matrix-geometric form

$$\mathbf{F_k} = \mathbf{F}\mathbf{X}_A^{k-1}\mathbf{Y}_A, \tag{2}$$

$$\mathbf{B_k} = \mathbf{B}\mathbf{X}_S^{k-1}\mathbf{Y}_S, \tag{3}$$

and the matrices at the boundary are

$$\bar{\mathbf{B}}_\mathbf{k} = \sum_{i=k}^{\infty} \mathbf{B_i}. \tag{4}$$

In fact, generator $\mathbf{Q}$ represents a GI/G/1-type Markov chain, however the matrix-geometric definition of the matrix blocks enables us to develop analysis procedures that are more efficient than those available for the general GI/G/1-type structure.

Throughout the paper the *batch* arrivals and the *individual* arrivals of the batch are distinguished, they have different meanings. Matrix $\mathbf{F}$ holds the transition rates leading to an arrival of a batch. At the batch arrival instant the first individual customer of the batch joins the queue immediately. The probability that the batch does not end yet and a new individual customer enters the queue is determined by sub-stochastic matrix $\mathbf{X}_A$ (note that the arrival of each individual customer in the batch can change the phase of the background process as well). The probabilities that the batch ends, with the corresponding phase transitions, are given by matrix $\mathbf{Y}_A$. I.e., $\mathbf{X}_A\mathbb{1} + \mathbf{Y}_A\mathbb{1} = \mathbb{1}$ holds. The batch and the individual service events are interpreted similarly.

Generalized QBD queues with matrix-geometric batch arrivals and services can represent a wide range of systems, some examples are listed below.

QBD queues without batches Setting $\mathbf{X}_A = \mathbf{0}, \mathbf{Y}_A = \mathbf{I}$ and $\mathbf{X}_S = \mathbf{0}, \mathbf{Y}_S = \mathbf{I}$ leads to an ordinary QBD queue without batches.

MAP/MAP/1 queues with phase-type distributed batch arrivals and services Assume that the batch arrivals and batch services are generated by Markovian Arrival Processes (MAPs), defined by matrices $(\mathbf{D_0}, \mathbf{D_1})$ and $(\mathbf{S_0}, \mathbf{S_1})$ for the arrivals and services, respectively. Let the size of the arrival and service batches be discrete phase-type (DPH) distributed, with parameters $(\alpha_A, \mathbf{A}_A, a_A = \mathbb{1} - \mathbf{A}_A\mathbb{1})$ and $(\alpha_S, \mathbf{A}_S, a_S = \mathbb{1} - \mathbf{A}_S\mathbb{1})$ ($\mathbb{1}$ denotes a column vector of ones). If the service discipline is FCFS, we get a generalized QBD queue with batches, where the parameters of the system are

$$\begin{aligned}
\mathbf{L_0} &= (\mathbf{D_0} \otimes \mathbf{I}) \otimes (\mathbf{I} \otimes \mathbf{I}),\\
\mathbf{L} &= (\mathbf{D_0} \otimes \mathbf{I}) \otimes (\mathbf{I} \otimes \mathbf{I}) + (\mathbf{I} \otimes \mathbf{I}) \otimes (\mathbf{S_0} \otimes \mathbf{I}),\\
\mathbf{F} &= (\mathbf{D_1} \otimes \mathbf{I}) \otimes (\mathbf{I} \otimes \mathbf{I}),\\
\mathbf{X}_A &= (\mathbf{I} \otimes \mathbf{A}_A) \otimes (\mathbf{I} \otimes \mathbf{I}),\\
\mathbf{Y}_A &= (\mathbf{I} \otimes a_A\alpha_A) \otimes (\mathbf{I} \otimes \mathbf{I}),\\
\mathbf{B} &= (\mathbf{I} \otimes \mathbf{I}) \otimes (\mathbf{S_1} \otimes \mathbf{I}),\\
\mathbf{X}_S &= (\mathbf{I} \otimes \mathbf{I}) \otimes (\mathbf{I} \otimes \mathbf{A}_S),\\
\mathbf{Y}_S &= (\mathbf{I} \otimes \mathbf{I}) \otimes (\mathbf{I} \otimes a_S\alpha_S),
\end{aligned}$$

since the background process has to keep track of 1) the phase of the arrival MAP, 2) the phase of the PH providing the size of the arrival batch, 3) the phase of the service MAP, 4) the phase of the PH determining the size of the service batch.

MM CPP/GE/1 The MM CPP/GE/c queueing model with negative customers proved to be useful in the analysis of a large number of telecommunication systems. The queue length and the sojourn time analysis of the basic variant of this queue have been published in [Chakka and Harrison(2001)] and [Harrison and Zatschler(2004)], respectively. If $c = 1$ (single server) and there are no negative customers the system belongs to the model class presented in this paper. The "MM" in the notation of the queueing system means that the arrivals and services are Markov modulated, with generator matrix denoted by $\mathbf{Q}$. The arrival and service times are exponentially distributed. The diagonal matrices of the arrival and service rates in various phases of the background process are denoted by $\mathbf{\Lambda}$ and $\mathbf{M}$. Batch arrivals and batch services are both allowed. $\mathbf{\Theta}$ and $\mathbf{\Phi}$ are the diagonal matrices of the parameters of the geometrically distributed batch sizes corresponding to the arrivals and services,

respectively. With these notations, the parameters of the generalized QBD queue with batches are

$$\begin{aligned} \mathbf{L_0} &= \mathbf{Q} - \mathbf{\Lambda}, & \mathbf{L} &= \mathbf{Q} - \mathbf{\Lambda} - \mathbf{M}, & & \\ \mathbf{F} &= \mathbf{\Lambda}, & \mathbf{X}_A &= \mathbf{\Theta}, & \mathbf{Y}_A &= \mathbf{I} - \mathbf{\Theta}, \\ \mathbf{B} &= \mathbf{M}, & \mathbf{X}_S &= \mathbf{\Phi}, & \mathbf{Y}_S &= \mathbf{I} - \mathbf{\Phi}. \end{aligned}$$

3 Analysis of the number of customers in the system

According to the results available for M/G/1 and GI/M/1 type queues (see [Neuts(1989)] and [Neuts(1981)]), the stability condition is that the upward drift must be less than the downward drift, thus $\theta \sum_{k=1}^{\infty} k\mathbf{F_k}\mathbb{1} < \theta \sum_{k=1}^{\infty} k\mathbf{B_k}\mathbb{1}$ must hold, where θ is the stationary phase probability vector of the background process. For the particular structure of the studied system this translates to

$$\theta\mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbb{1} < \theta\mathbf{B}(\mathbf{I} - \mathbf{X}_S)^{-1}\mathbb{1}, \tag{5}$$

where vector θ is the solution to $\theta(\mathbf{B}(\mathbf{I} - \mathbf{X}_S)^{-1}\mathbf{Y}_S + \mathbf{L} + \mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A) = 0$, $\theta\mathbb{1} = 1$ ($\mathbb{1}$ denotes the column vector of ones and $\mathbf{I}$ denotes the identity matrix of appropriate size).

Throughout the paper the stability condition is assumed to hold.

Three queue length related stationary distributions are studied in this section, the joint probability of the number of customers and the phase of the background process

- at random time instants, denoted by $\pi = [\pi_i,\ i \geq 0]$ (formally, $(\pi_i)_j = \lim_{t\to\infty} P(\mathcal{X}(t) = i, \mathcal{J}(t) = j)$),
- right after individual arrival instants, denoted by $x = [x_i,\ i \geq 1]$,
- right after individual service instants, denoted by $y = [y_i,\ i \geq 0]$,

where π_k, x_k and y_k are size N row vectors.

3.1 The distribution of the number of customers in the system

According to the following theorem, π, x and y are matrix-geometrically distributed.

Theorem 1 *The stationary solutions of the Markov chain* (1) *at random time instants, right after individual arrival instants and right after individual service instants are matrix-geometric, i.e.,*

$$\pi_k = c_\pi x_1 \hat{\mathbf{R}}^{k-1}\mathbf{V}, \tag{6}$$

$$x_k = x_1 \hat{\mathbf{R}}^{k-1}, \tag{7}$$

$$y_k = c_y x_1 \hat{\mathbf{R}}^{k-1}\mathbf{H}, \tag{8}$$

for $k \geq 1$, where size N square matrices $\mathbf{R}, \mathbf{V}$ and $\mathbf{H}$ are the minimal non-negative solutions to matrix equations

$$\hat{\mathbf{R}} = \mathbf{X}_A + \mathbf{V}\mathbf{F}, \tag{9}$$

$$\mathbf{0} = \mathbf{H}\mathbf{Y}_S + \mathbf{V}\mathbf{L} + \mathbf{Y}_A, \tag{10}$$

$$\mathbf{H} = \hat{\mathbf{R}}\mathbf{V}\mathbf{B} + \hat{\mathbf{R}}\mathbf{H}\mathbf{X}_S, \tag{11}$$

and c_π, c_y are normalization constants.

Proof Due to the matrix-geometric nature of the batch sizes it is possible to define a GI/M/1-type discrete time Markov chain (DTMC) for the queue length observed by the individual arrivals right after they join the queue. The transition probability matrix of the DTMC is

$$\hat{\mathbf{Q}} = \begin{bmatrix} \hat{\mathbf{A}}'_{\mathbf{1}} & \hat{\mathbf{A}}_{\mathbf{0}} & & & \\ \hat{\mathbf{A}}'_{\mathbf{2}} & \hat{\mathbf{A}}_{\mathbf{1}} & \hat{\mathbf{A}}_{\mathbf{0}} & & \\ \hat{\mathbf{A}}'_{\mathbf{3}} & \hat{\mathbf{A}}_{\mathbf{2}} & \hat{\mathbf{A}}_{\mathbf{1}} & \hat{\mathbf{A}}_{\mathbf{0}} & \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix} \tag{12}$$

where the blocks corresponding to the regular part are

$$\hat{\mathbf{A}}_{\mathbf{0}} = \mathbf{X}_A + \mathbf{Y}_A\hat{\mathbf{P}}_{\mathbf{0}}\mathbf{F}, \qquad \hat{\mathbf{A}}_{\mathbf{k}} = \mathbf{Y}_A\hat{\mathbf{P}}_{\mathbf{k}}\mathbf{F}, \quad \text{for } k > 0. \tag{13}$$

Entry i, j of matrix $\hat{\mathbf{P}}_{\mathbf{k}}$ is the mean time spent in the state where the background process is in phase j and k customers are served between two batch arrival instants, given that the phase was i initially.

According to (13), the DTMC moves one level forward if the next customer of the batch (which is *not* the first in its batch) arrives, or if a batch arrival is completed and no customers are served till the first customer of the next batch arrives. The DTMC moves k levels backwards if $k+1$ customers are served between two batch arrivals.

The probabilities that k customers are served in time t before the next arrival with the corresponding phase transitions are given by matrix $\hat{\mathbf{P}}(k,t)$ and are characterized by differential equations

$$\begin{aligned} \frac{d}{dt}\hat{\mathbf{P}}(k,t) &= \hat{\mathbf{P}}(k,t)\mathbf{L} + \sum_{i=1}^{k} \hat{\mathbf{P}}(k-i,t)\mathbf{B}\mathbf{X}_S^{i-1}\mathbf{Y}_S, \quad \text{for } k > 0, \\ \frac{d}{dt}\hat{\mathbf{P}}(0,t) &= \hat{\mathbf{P}}(0,t)\mathbf{L}, \end{aligned} \tag{14}$$

from which $\hat{\mathbf{P}}_{\mathbf{k}} = \int_{t=0}^{\infty} \hat{\mathbf{P}}(k,t)\,dt$ is obtained.

The stationary distribution of GI/M/1 type DTMCs is known to be matrix-geometric [Neuts(1981)], i.e., $x_k = x_1\hat{\mathbf{R}}^{k-1}$ where matrix $\hat{\mathbf{R}}$ is the minimal non-negative solution of

$$\hat{\mathbf{R}} = \sum_{k=0}^{\infty} \hat{\mathbf{R}}^k\hat{\mathbf{A}}_{\mathbf{k}} = \mathbf{X}_A + \underbrace{\sum_{k=0}^{\infty} \hat{\mathbf{R}}^k\mathbf{Y}_A\hat{\mathbf{P}}_{\mathbf{k}}}_{\mathbf{V}}\mathbf{F}, \tag{15}$$

that is equal to (9) if the under-braced term is denoted by $\mathbf{V}$.

Let us now derive equations for matrix $\mathbf{V}$. The integral of (14) between 0 and ∞ gives

$$-\mathbf{I}\delta_k = \hat{\mathbf{P}}_\mathbf{k}\mathbf{L} + \sum_{i=1}^{k} \hat{\mathbf{P}}_{\mathbf{k-i}}\mathbf{B}\mathbf{X}_S^{i-1}\mathbf{Y}_S, \tag{16}$$

where δ_k is the Kronecker delta, i.e. $\delta_0 = 1$ and $\delta_k = 0$ for $k \neq 0$. Multiplying both sides of (16) by $\hat{\mathbf{R}}^k\mathbf{Y}_A$ from the left, summing from 0 to ∞ and swapping the summations leads to

$$-\mathbf{Y}_A = \mathbf{V}\mathbf{L} + \underbrace{\sum_{i=1}^{\infty} \hat{\mathbf{R}}^i\mathbf{V}\mathbf{B}\mathbf{X}_S^{i-1}}_{\mathbf{H}}\mathbf{Y}_S, \tag{17}$$

which equals (10). Finally, matrix $\mathbf{H}$, defined by the infinite sum above, is the solution of the discrete Sylvester equation (11) (also called Stein equation, see [Antoulas(2005)], Section 6.1.7).

Right now we have proven (7), and the matrix equations for $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$.

The stationary distribution at random time instant π_k is proportional to the time spent at level k between two batch arrival instants. Conditioning on the queue length distribution at arrivals and noting that the queue length can only decrease between arrival instants, for $k > 0$ we have that

$$\pi_k = c_\pi \sum_{i=k}^{\infty} x_i\mathbf{Y}_A\hat{\mathbf{P}}_{\mathbf{i-k}} = c_\pi\, x_1\hat{\mathbf{R}}^{k-1}\sum_{i=k}^{\infty}\hat{\mathbf{R}}^{i-k}\mathbf{Y}_A\hat{\mathbf{P}}_{\mathbf{i-k}} = c_\pi\, x_1\mathbf{R}^{k-1}\mathbf{V}, \tag{18}$$

which proves (6) (c_π is a normalization constant).

Finally, when a batch service is initiated at level i, the $(i-k)$th individual service of the batch leaves k customers in the system, thus

$$y_k = c'\sum_{i=k+1}^{\infty}\pi_i\mathbf{B}\mathbf{X}_S^{i-k-1} = c'\,c_\pi x_1\hat{\mathbf{R}}^{k-1}\sum_{i=k+1}^{\infty}\hat{\mathbf{R}}^{i-k}\mathbf{V}\mathbf{B}\mathbf{X}_S^{i-k-1} = \underbrace{c'\,c_\pi}_{c_y}x_1\hat{\mathbf{R}}^{k-1}\mathbf{H} \tag{19}$$

holds for $k > 0$, where c' is a normalization constant. □

The next theorem provides the missing components of the matrix-geometric solutions.

Theorem 2 *The initial vectors x_1 and normalization constants c_π, c_y are*

$$x_1 = \pi_0\mathbf{F}/\lambda, \tag{20}$$

$$c_\pi = \lambda, \tag{21}$$

$$c_y = \lambda/\mu. \tag{22}$$

Vector π_0 is the solution to the linear equation

$$\begin{aligned} 0 &= \pi_0(\mathbf{L_0} + \mathbf{F}(\mathbf{VB} + \mathbf{HX}_S)(\mathbf{I} - \mathbf{X}_S)^{-1}\mathbf{Y}_S), \\ 1 &= \pi_0(\mathbf{I} + \mathbf{F}(\mathbf{I} - \hat{\mathbf{R}})^{-1}\mathbf{V})\mathbb{1}, \end{aligned} \tag{23}$$

and vector y_0 is calculated by

$$y_0 = \frac{1}{\mu}\pi_0\mathbf{F}(\mathbf{VB} + \mathbf{HX}_S)(\mathbf{I} - \mathbf{X}_S)^{-1}. \tag{24}$$

λ and μ are the mean arrival and service rates, given by

$$\lambda = \pi_0\mathbf{F}(\mathbf{I} - \hat{\mathbf{R}})^{-1}\mathbb{1}, \tag{25}$$

$$\mu = \pi_0\mathbf{F}\left((\mathbf{VB} + \mathbf{HX}_S)(\mathbf{I} - \mathbf{X}_S)^{-1} + (\mathbf{I} - \hat{\mathbf{R}})^{-1}\mathbf{H}\right)\mathbb{1}. \tag{26}$$

Proof Level 1 can be observed only by the first arrival of the batch in an empty system, implying $x_1 = \pi_0\mathbf{F}/\lambda$, that provides (20). The normalization condition for x_k gives the expression for λ in (25). Now we show that $c_\pi = \lambda$. To this end λ, the mean arrival rate is expressed from π_i as well:

$$\begin{aligned} \lambda &= \sum_{i=0}^{\infty}\pi_i\mathbf{F}\sum_{k=0}^{\infty}(k+1)\mathbf{X}_A^k\mathbf{Y}_A\mathbb{1} \\ &= \pi_0(\mathbf{I} + (c_\pi/\lambda)\mathbf{F}(\mathbf{I} - \hat{\mathbf{R}})^{-1}\mathbf{V})\mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-2}\mathbf{Y}_A\mathbb{1} \\ &= \pi_0(\mathbf{I} + (c_\pi/\lambda)\mathbf{F}(\mathbf{I} - \hat{\mathbf{R}})^{-1}\mathbf{V})\mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbb{1} \\ &= \pi_0\mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbb{1} + (c_\pi/\lambda)\pi_0\mathbf{F}(\mathbf{I} - \hat{\mathbf{R}})^{-1}\mathbf{VF}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbb{1} \\ &= \pi_0\mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbb{1} + (c_\pi/\lambda)\pi_0\mathbf{F}(\mathbf{I} - \hat{\mathbf{R}})^{-1}\hat{\mathbf{R}}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbb{1} \\ &\quad - (c_\pi/\lambda)\pi_0\mathbf{F}(\mathbf{I} - \hat{\mathbf{R}})^{-1}\mathbf{X}_A(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbb{1} \\ &= \pi_0\mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbb{1} + (c_\pi/\lambda)\pi_0\mathbf{F}(\mathbf{I} - \hat{\mathbf{R}})^{-1}\mathbb{1} - (c_\pi/\lambda)\pi_0\mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbb{1}, \end{aligned}$$

which is clearly satisfied if $c_\pi = \lambda$, proving (21). During the transformations we utilized that $\mathbf{VF} = \hat{\mathbf{R}} - \mathbf{X}_A$ (based on (9)), and that $\hat{\mathbf{R}}(\mathbf{I} - \hat{\mathbf{R}})^{-1} = (\mathbf{I} - \hat{\mathbf{R}})^{-1} - \mathbf{I}$ and $\mathbf{X}_A(\mathbf{I} - \mathbf{X}_A)^{-1} = (\mathbf{I} - \mathbf{X}_A)^{-1} - \mathbf{I}$ hold.

Next, the number of customers at service instants is investigated. The system is left empty if the service batch is longer than the number of customers present in the system, hence

$$\begin{aligned} y_0 &= c'\sum_{i=1}^{\infty}\pi_i\sum_{j=i}^{\infty}\mathbf{B}\mathbf{X}_S^{j-1} = c'\pi_0\mathbf{F}\sum_{i=1}^{\infty}\hat{\mathbf{R}}^{i-1}\mathbf{VB}\mathbf{X}_S^{i-1}(\mathbf{I} - \mathbf{X}_S)^{-1} \\ &= c'\pi_0\mathbf{F}(\mathbf{VB} + \sum_{i=2}^{\infty}\hat{\mathbf{R}}^{i-1}\mathbf{VB}\mathbf{X}_S^{i-1})(\mathbf{I} - \mathbf{X}_S)^{-1} \\ &= c'\pi_0\mathbf{F}(\mathbf{VB} + \mathbf{HX}_S)(\mathbf{I} - \mathbf{X}_S)^{-1}. \end{aligned} \tag{27}$$

The normalization condition for y_k provides the constant c' as

$$\begin{aligned} 1/c' &= y_0\mathbb{1} + \sum_{k=1}^{\infty} y_k\mathbb{1} \\ &= \pi_0\mathbf{F}(\mathbf{VB}+\mathbf{HX}_S)(\mathbf{I}-\mathbf{X}_S)^{-1}\mathbb{1} + \pi_0\mathbf{F}(\mathbf{I}-\hat{\mathbf{R}})^{-1}\mathbf{H}\mathbb{1}. \end{aligned} \tag{28}$$

It remains to show that $1/c'$ equals the mean service rate. Expressing and transforming the mean service rate leads to

$$\begin{aligned} \mu &= \sum_{i=1}^{\infty}\pi_i\mathbf{B}\sum_{k=0}^{\infty}(k+1)\mathbf{X}_S^k\mathbf{Y}_S\mathbb{1} \\ &= \pi_0\mathbf{F}(\mathbf{I}-\hat{\mathbf{R}})^{-1}\mathbf{VB}(\mathbf{I}-\mathbf{X}_S)^{-1}\mathbb{1} \\ &= \pi_0\mathbf{FVB}(\mathbf{I}-\mathbf{X}_S)^{-1}\mathbb{1} + \pi_0\mathbf{F}(\mathbf{I}-\hat{\mathbf{R}})^{-1}\hat{\mathbf{R}}\mathbf{VB}(\mathbf{I}-\mathbf{X}_S)^{-1}\mathbb{1} \\ &= \pi_0\mathbf{FVB}(\mathbf{I}-\mathbf{X}_S)^{-1}\mathbb{1} + \pi_0\mathbf{F}(\mathbf{I}-\hat{\mathbf{R}})^{-1}(\mathbf{H}-\hat{\mathbf{R}}\mathbf{HX}_S)(\mathbf{I}-\mathbf{X}_S)^{-1}\mathbb{1} \\ &= \pi_0\mathbf{FVB}(\mathbf{I}-\mathbf{X}_S)^{-1}\mathbb{1} + \pi_0\mathbf{F}(\mathbf{I}-\hat{\mathbf{R}})^{-1}\mathbf{H}\mathbb{1} \\ &\quad + \pi_0\mathbf{FH}(\mathbf{I}-\mathbf{X}_S)^{-1}\mathbb{1} - \pi_0\mathbf{FH}\mathbb{1} \\ &= \pi_0\mathbf{F}(\mathbf{VB}+\mathbf{HX}_S)(\mathbf{I}-\mathbf{X}_S)^{-1}\mathbb{1} + \pi_0\mathbf{F}(\mathbf{I}-\hat{\mathbf{R}})^{-1}\mathbf{H}\mathbb{1}. \end{aligned} \tag{29}$$

In the manipulations we exploited $\hat{\mathbf{R}}\mathbf{VB} = \mathbf{H} - \hat{\mathbf{R}}\mathbf{HX}_S$ (based on (11)), and that $\hat{\mathbf{R}}(\mathbf{I}-\hat{\mathbf{R}})^{-1} = (\mathbf{I}-\hat{\mathbf{R}})^{-1} - \mathbf{I}$ and $\mathbf{X}_S(\mathbf{I}-\mathbf{X}_S)^{-1} = (\mathbf{I}-\mathbf{X}_S)^{-1} - \mathbf{I}$ hold. Since $1/c' = \mu$ and $c_y = c'\,c_\pi$ (see (19)), (22) is proven.

Finally, (23) is derived from the first equilibrium equation for generator (1) and the normalization condition for π_k:

$$\begin{aligned} 0 &= \pi_0\mathbf{L_0} + \sum_{i=1}^{\infty}\pi_i\bar{\mathbf{B}}_\mathbf{i} = \pi_0\mathbf{L_0} + \pi_0\mathbf{F}\sum_{i=1}^{\infty}\hat{\mathbf{R}}^{i-1}\mathbf{V}\sum_{j=i}^{\infty}\mathbf{B}\mathbf{X}_S^{j-1}\mathbf{Y}_S \\ &= \pi_0\mathbf{L_0} + \pi_0\mathbf{F}\left(\mathbf{VB} + \sum_{i=2}^{\infty}\hat{\mathbf{R}}^{i-1}\mathbf{VB}\mathbf{X}_S^{i-1}\right)(\mathbf{I}-\mathbf{X}_S)^{-1}\mathbf{Y}_S \\ &= \pi_0\mathbf{L_0} + \pi_0\mathbf{F}\left(\mathbf{VB}+\mathbf{HX}_S\right)(\mathbf{I}-\mathbf{X}_S)^{-1}\mathbf{Y}_S. \end{aligned} \tag{30}$$

□

Corollary 1 *The stationary solutions of the Markov chain* (1) *at random time instants, right after individual arrival instants and right after individual service instants are*

$$\pi_k = \pi_0\mathbf{F}\hat{\mathbf{R}}^{k-1}\mathbf{V}, \tag{31}$$

$$x_k = \pi_0\mathbf{F}\hat{\mathbf{R}}^{k-1}/\lambda, \tag{32}$$

$$y_k = \pi_0\mathbf{F}\hat{\mathbf{R}}^{k-1}\mathbf{H}/\mu, \tag{33}$$

for $k > 1$.

Remark 1 The fundamental matrix is denoted by $\hat{\mathbf{R}}$ instead of $\mathbf{R}$ because it corresponds to the arrival instants. The same notation was used in [Ozawa(2006)] and in [Horváth et al(2014)Horváth, Van Houdt, and Telek] as well. With an appropriate similarity transformation it is possible to transform $\pi_i = \pi_1 \hat{\mathbf{R}}^{i-1}\mathbf{V}$ into a purely matrix-geometric form $\pi_i = \pi_1' \mathbf{R}^{i-1}$, but in this case such a matrix $\mathbf{R}$ could have negative entries, which is not beneficial from the numerical stability point of view. Therefore it is better to stick with $\pi_i = \pi_1 \hat{\mathbf{R}}^{i-1}\mathbf{V}$, where the non-negativity of $\hat{\mathbf{R}}$ and $\mathbf{V}$ are guaranteed.

3.2 Efficient algorithms to obtain matrices $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$

From the numerical point of view the most critical question is how to obtain matrices $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$. This section provides three different solutions: a functional iteration based, a Newton iteration based solution, and a way to reduce the problem to the solution of a matrix-quadratic equation.

3.2.1 Basic procedures for $\hat{\mathbf{R}}, \mathbf{V}$ *and* $\mathbf{H}$

The most straight forward algorithm is based on a functional iteration, see Algorithm 1. (Note that this algorithm is similar to the one in Figure 8.1 in [Latouche and Ramaswami(1999)] if there are no batches.) In each iteration, the computationally most demanding step is the solution of a discrete Sylvester equation to obtain $\mathbf{H}^{(n+1)}$. One of the fastest and widely used direct method for solving such equations is the Hessenberg-Schur method [Golub et al(1979)Golub, Nash, and Van Loan] which has a computational complexity of $\mathcal{O}(N^3)$. Unfortunately the functional iteration in Algorithm 1 suffers from linear (slow) convergence speed.

Algorithm 1 Linearly convergent algorithm to obtain $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$

$\mathbf{V}^{(0)} \leftarrow \mathbf{Y}_A(-\mathbf{L})^{-1}$
$\hat{\mathbf{R}}^{(0)} \leftarrow \mathbf{X}_A + \mathbf{V}^{(0)}\mathbf{F}$
$n \leftarrow 0$
repeat
  $\mathbf{H}^{(n+1)} \leftarrow$ solution of $\hat{\mathbf{R}}^{(n)}\mathbf{H}^{(n+1)}\mathbf{X}_S - \mathbf{H}^{(n+1)} = -\hat{\mathbf{R}}^{(n)}\mathbf{V}^{(n)}\mathbf{B}$
  $\mathbf{V}^{(n+1)} \leftarrow (\mathbf{H}^{(n+1)}\mathbf{Y}_S + \mathbf{Y}_A)(-\mathbf{L})^{-1}$
  $\hat{\mathbf{R}}^{(n+1)} \leftarrow \mathbf{V}^{(n+1)}\mathbf{F} + \mathbf{X}_A$
  $n \leftarrow n+1$
until $||\hat{\mathbf{R}}^{(n)} - \hat{\mathbf{R}}^{(n-1)}|| < \epsilon$
return $\hat{\mathbf{R}}^{(n)}, \mathbf{V}^{(n)}, \mathbf{H}^{(n)}$

We also present a quadratically convergent algorithm based on the Newton iteration. Inserting $\hat{\mathbf{R}} = \mathbf{X}_A + \mathbf{V}\mathbf{F}$ (from (9)) and $\mathbf{V} = (\mathbf{Y}_A + \mathbf{H}\mathbf{Y}_S)(-\mathbf{L})^{-1}$ (from (10)) into (11) provides a matrix equation for $\mathbf{H}$ that does not depend on $\mathbf{V}$ and $\hat{\mathbf{R}}$, yielding

$$\mathbf{0} = (\mathbf{H}\mathbf{Z}_{\mathbf{s_1}} + \mathbf{Z}_{\mathbf{a_2}})(\mathbf{H}\mathbf{Z}_{\mathbf{s_2}} + \mathbf{Z}_{\mathbf{a_1}}) - \mathbf{H} := \mathcal{M}(\mathbf{H}), \tag{34}$$

where the matrix coefficients are

$$\mathbf{Z_{s_1}} = \mathbf{Y}_S(-\mathbf{L})^{-1}\mathbf{F}, \qquad \mathbf{Z_{s_2}} = \mathbf{X}_S + \mathbf{Y}_S(-\mathbf{L})^{-1}\mathbf{B},$$
$$\mathbf{Z_{a_1}} = \mathbf{Y}_A(-\mathbf{L})^{-1}\mathbf{B}, \qquad \mathbf{Z_{a_2}} = \mathbf{X}_A + \mathbf{Y}_A(-\mathbf{L})^{-1}\mathbf{F}.$$

The steps of the Newton iteration are $\mathbf{H}^{(n+1)} = \mathbf{H}^{(n)} + \mathbf{\Delta}^{(n)}$, where the update $\mathbf{\Delta}^{(n)}$ is the solution of

$$\mathcal{M}'|_{\mathbf{H}^{(n)}}(\mathbf{\Delta}^{(n)}) = -\mathcal{M}(\mathbf{H}^{(n)}), \tag{35}$$

where $\mathcal{M}'|_{\mathbf{H}^{(n)}}$ is the Fréchet derivative of operator $\mathcal{M}$ at $\mathbf{H}^{(n)}$, which, in our case is

$$\mathcal{M}'|_{\mathbf{H}^{(n)}} : \mathbf{\Delta}^{(n)} \to \mathbf{\Delta}^{(n)}\mathbf{Z_{s_1}}(\mathbf{H}\mathbf{Z_{s_2}} + \mathbf{Z_{a_1}}) + (\mathbf{H}\mathbf{Z_{s_1}} + \mathbf{Z_{a_2}})\mathbf{\Delta}^{(n)}\mathbf{Z_{s_2}} - \mathbf{\Delta}^{(n)}.$$

Thus, in step n the solution of the discrete Sylvester equation

$$(\mathbf{H}^{(n)}\mathbf{Z_{s_1}} + \mathbf{Z_{a_2}})\mathbf{\Delta}^{(n)}\mathbf{Z_{s_2}}(\mathbf{I} - \mathbf{Z_{s_1}}(\mathbf{H}^{(n)}\mathbf{Z_{s_2}} + \mathbf{Z_{a_1}}))^{-1} - \mathbf{\Delta}^{(n)} =$$
$$-\left((\mathbf{H}^{(n)}\mathbf{Z_{s_1}} + \mathbf{Z_{a_2}})(\mathbf{H}^{(n)}\mathbf{Z_{s_2}} + \mathbf{Z_{a_1}}) - \mathbf{H}^{(n)}\right)(\mathbf{I} - \mathbf{Z_{s_1}}(\mathbf{H}^{(n)}\mathbf{Z_{s_2}} + \mathbf{Z_{a_1}}))^{-1}$$

provides the update $\mathbf{\Delta}^{(n)}$ (see Algorithm 2).

The computationally most demanding steps in this algorithm are the solution of a discrete Sylvester equation providing $\mathbf{\Delta}^{(n)}$ ($\mathcal{O}(N^3)$ steps), and the calculation of a matrix inverse for $\mathbf{T_3}$ (the related Gauss-Jordan elimination needs $\mathcal{O}(N^3)$ steps).

Algorithm 2 Quadratically convergent algorithm to obtain $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$

$\mathbf{H}^{(0)} \leftarrow \mathbf{0}$
$\mathbf{Z_{s_1}} \leftarrow \mathbf{Y}_S(-\mathbf{L})^{-1}\mathbf{F}, \qquad \mathbf{Z_{s_2}} \leftarrow \mathbf{X}_S + \mathbf{Y}_S(-\mathbf{L})^{-1}\mathbf{B}$
$\mathbf{Z_{a_1}} \leftarrow \mathbf{Y}_A(-\mathbf{L})^{-1}\mathbf{B}, \qquad \mathbf{Z_{a_2}} \leftarrow \mathbf{X}_A + \mathbf{Y}_A(-\mathbf{L})^{-1}\mathbf{F}$
$n \leftarrow 0$
repeat
  $\mathbf{T_1} \leftarrow \mathbf{H}^{(n)}\mathbf{Z_{s_1}} + \mathbf{Z_{a_2}}$
  $\mathbf{T_2} \leftarrow \mathbf{H}^{(n)}\mathbf{Z_{s_2}} + \mathbf{Z_{a_1}}$
  $\mathbf{T_3} \leftarrow (\mathbf{I} - \mathbf{Z_{s_1}}\mathbf{T_2})^{-1}$
  $\mathbf{\Delta}^{(n)} \leftarrow$ solution of $\mathbf{T_1}\mathbf{\Delta}^{(n)}\mathbf{Z_{s_2}}\mathbf{T_3} - \mathbf{\Delta}^{(n)} = -\mathbf{T_1}\mathbf{T_2}\mathbf{T_3} + \mathbf{H}^{(n)}\mathbf{T_3}$
  $\mathbf{H}^{(n+1)} \leftarrow \mathbf{H}^{(n)} + \mathbf{\Delta}^{(n)}$
  $n \leftarrow n + 1$
until $||\mathbf{H}^{(n)} - \mathbf{H}^{(n-1)}|| < \epsilon$
$\mathbf{V} \leftarrow (\mathbf{H}^{(n)}\mathbf{Y}_S + \mathbf{Y}_A)(-\mathbf{L})^{-1}$
$\hat{\mathbf{R}} \leftarrow \mathbf{X}_A + \mathbf{V}\mathbf{F}$
return $\hat{\mathbf{R}}, \mathbf{V}, \mathbf{H}^{(n)}$

3.2.2 Obtaining the matrices by the solution of a matrix-quadratic equation

Observe that the behavior of the queue can be characterized by a QBD as well, where the block size is $3N$. The blocks of this QBD are:

$$\tilde{\mathbf{F}} = \begin{bmatrix} \mathbf{0} & \mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{L}} = \begin{bmatrix} \mathbf{L} & \mathbf{0} & \mathbf{0} \\ \mathbf{Y}_A & -\mathbf{I} & \mathbf{0} \\ \mathbf{Y}_S & \mathbf{0} & -\mathbf{I} \end{bmatrix}, \quad \tilde{\mathbf{B}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{B} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_S \end{bmatrix}. \tag{36}$$

The first group of the phases corresponds to the local transitions of the background process. Whenever a batch of customers arrives, a transition to the second group of phases occurs. The role of the second group of phases is to increase the queue length gradually according to the size of the arriving batch, and the role of the third state group is to decrease it according to the size of the batch service.

By construction, censoring the stationary probabilities of this QBD (denoted by $\tilde{\pi}_k$) to the first phase group gives the stationary distribution of the original system, while censoring to the second and third phase groups provide the queue length distribution at individual arrivals and services, respectively, thus we have $\tilde{\pi}_k = \tilde{\pi}_0 \tilde{\mathbf{R}}^k = \left[\pi_k/c_1 \; x_k/c_2 \; y_k/c_3\right]$ for $k \geq 1$, where c_1, c_2 and c_3 are normalizing constants.

Moreover, as expected, there is a strong relationship between matrix $\tilde{\mathbf{R}}$ and matrices $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$.

Theorem 3 *Matrices* $\tilde{\mathbf{R}}$ *and* $\hat{\mathbf{R}}, \mathbf{V}, \mathbf{H}$ *are related as*

$$\left[\mathbf{V}\ \mathbf{I}\ \mathbf{H}\right] \tilde{\mathbf{R}}^i = \hat{\mathbf{R}}^i \left[\mathbf{V}\ \mathbf{I}\ \mathbf{H}\right], \qquad i \geq 0. \tag{37}$$

Proof First we show that

$$\tilde{\mathbf{R}} = \begin{bmatrix} \mathbf{F} \\ \mathbf{X}_A \\ \mathbf{0} \end{bmatrix} \left[\mathbf{V}\ \mathbf{I}\ \mathbf{H}\right] \tag{38}$$

holds, by substituting it to the matrix-quadratic equation $\mathbf{0} = \tilde{\mathbf{F}} + \tilde{\mathbf{R}}\tilde{\mathbf{L}} + \tilde{\mathbf{R}}^2\tilde{\mathbf{B}}$. Exploiting that $\tilde{\mathbf{F}} = \begin{bmatrix} \mathbf{F} \\ \mathbf{X}_A \\ \mathbf{0} \end{bmatrix} \left[\mathbf{0}\ \mathbf{I}\ \mathbf{0}\right]$ and that $\tilde{\mathbf{R}}^2 = \begin{bmatrix} \mathbf{F} \\ \mathbf{X}_A \\ \mathbf{0} \end{bmatrix} \hat{\mathbf{R}} \left[\mathbf{V}\ \mathbf{I}\ \mathbf{H}\right]$ we get

$$\begin{aligned} \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} &= \begin{bmatrix} \mathbf{F} \\ \mathbf{X}_A \\ \mathbf{0} \end{bmatrix} \left[\mathbf{0}\ \mathbf{I}\ \mathbf{0}\right] + \begin{bmatrix} \mathbf{F} \\ \mathbf{X}_A \\ \mathbf{0} \end{bmatrix} \left[\mathbf{V}\mathbf{L} + \mathbf{Y}_A + \mathbf{H}\mathbf{X}_S\ -\mathbf{I}\ -\mathbf{H}\right] \\ &\quad + \begin{bmatrix} \mathbf{F} \\ \mathbf{X}_A \\ \mathbf{0} \end{bmatrix} \left[\mathbf{0}\ \mathbf{0}\ \hat{\mathbf{R}}\mathbf{V}\mathbf{B} + \hat{\mathbf{R}}\mathbf{H}\mathbf{X}_S\right], \end{aligned} \tag{39}$$

which is satisfied since $\mathbf{V}\mathbf{L} + \mathbf{Y}_A + \mathbf{H}\mathbf{X}_S = \mathbf{0}$ holds due to (10) and $\hat{\mathbf{R}}\mathbf{V}\mathbf{B} + \hat{\mathbf{R}}\mathbf{H}\mathbf{X}_S - \mathbf{H} = \mathbf{0}$ holds due to (11).

As $\begin{bmatrix}\mathbf{V}\ \mathbf{I}\ \mathbf{H}\end{bmatrix}\begin{bmatrix}\mathbf{F}\\ \mathbf{X}_A\\ \mathbf{0}\end{bmatrix} = \hat{\mathbf{R}}$ according to (9), it is easy to see that

$$\tilde{\mathbf{R}}^i = \begin{bmatrix}\mathbf{F}\\ \mathbf{X}_A\\ \mathbf{0}\end{bmatrix}\hat{\mathbf{R}}^{i-1}\begin{bmatrix}\mathbf{V}\ \mathbf{I}\ \mathbf{H}\end{bmatrix}, \tag{40}$$

which, pre-multiplied by $\begin{bmatrix}\mathbf{V}\ \mathbf{I}\ \mathbf{H}\end{bmatrix}$ establishes (37). □

Theorem 4 *Matrices* $\mathbf{V}$ *and* $\mathbf{H}$ *can be obtained from equation*

$$\begin{bmatrix}\mathbf{V}\ \mathbf{I}\ \mathbf{H}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\ \mathbf{I}\ \mathbf{0}\end{bmatrix}(-\tilde{\mathbf{U}})^{-1}, \tag{41}$$

where $\tilde{\mathbf{U}} = \tilde{\mathbf{L}} + \tilde{\mathbf{R}}\tilde{\mathbf{B}}$.

Proof Applying Theorem 3 for $i = 1$ gives $\begin{bmatrix}\mathbf{V}\ \mathbf{I}\ \mathbf{H}\end{bmatrix}\tilde{\mathbf{R}} = \hat{\mathbf{R}}\begin{bmatrix}\mathbf{V}\ \mathbf{I}\ \mathbf{H}\end{bmatrix}$. Multiplying both sides by $\tilde{\mathbf{B}}$ from the right and adding $\begin{bmatrix}\mathbf{V}\ \mathbf{I}\ \mathbf{H}\end{bmatrix}\tilde{\mathbf{L}}$ leads to

$$\begin{bmatrix}\mathbf{V}\ \mathbf{I}\ \mathbf{H}\end{bmatrix}\tilde{\mathbf{U}} = \big[\underbrace{\mathbf{VL} + \mathbf{Y}_A + \mathbf{HY}_S}_{\mathbf{0}}\quad -\mathbf{I}\quad \underbrace{\hat{\mathbf{R}}\mathbf{VB} + \hat{\mathbf{R}}\mathbf{HX}_S - \mathbf{H}}_{\mathbf{0}}\big]. \tag{42}$$

Matrix $\tilde{\mathbf{U}}$ is the infinitesimal generator of the Markov process restricted to level n before the first visit to level $n-1$ in the QBD defined by (36) (see [Latouche and Ramaswami(1999)]). If the QBD is stable, $\tilde{\mathbf{U}}$ is a transient generator, hence it is invertible, providing the theorem. □

Theorem 4 enables the reduction of the problem of obtaining $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$ to the solution of a matrix-quadratic equation involving size $3N$ matrices (Algorithm 3). The availability of mature, efficient solution algorithms for matrix-quadratic equations may compensate the slightly increased complexity due to the larger matrices.

Algorithm 3 Obtaining $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$ by solving a matrix-quadratic equation

$\tilde{\mathbf{R}} \leftarrow$ solution of $\mathbf{0} = \tilde{\mathbf{F}} + \tilde{\mathbf{R}}\tilde{\mathbf{L}} + \tilde{\mathbf{R}}^2\tilde{\mathbf{B}}$
$\tilde{\mathbf{U}} \leftarrow \tilde{\mathbf{L}} + \tilde{\mathbf{R}}\tilde{\mathbf{B}}$
$\mathbf{Z} \leftarrow \begin{bmatrix}\mathbf{0}\ \mathbf{I}\ \mathbf{0}\end{bmatrix}(-\tilde{\mathbf{U}})^{-1}$
$\mathbf{V} \leftarrow \mathbf{Z}[1:N, 1:N]$
$\mathbf{H} \leftarrow \mathbf{Z}[1:N, 2N+1:3N]$
$\hat{\mathbf{R}} \leftarrow \mathbf{X}_A + \mathbf{VF}$
return $\hat{\mathbf{R}}, \mathbf{V}, \mathbf{H}$

3.3 Phase-type representation of the queue length distribution

The next theorem, providing a discrete phase-type distribution for the number of customers, is inspired by [Sengupta(1990a)].

Theorem 5 *Assuming that vector* $\xi = \pi_0 \mathbf{F}(\mathbf{I} - \hat{\mathbf{R}})^{-1}$ *is strictly positive, the number of customers in the system is (discrete) phase-type distributed with parameters* $(\tau, \mathbf{T})$*, thus*

$$p_k = \tau \mathbf{T}^{k-1}(\mathbf{I} - \mathbf{T})\mathbb{1}, \qquad k \geq 1, \tag{43}$$

$$p_0 = 1 - \tau\mathbb{1}, \tag{44}$$

where $p_k = \pi_k \mathbb{1}$*. The initial probability vector* τ *and the sub-stochastic transition probability matrix* $\mathbf{T}$ *are given by*

$$\tau = \mathbb{1}^T \mathbf{V}^T \Delta, \tag{45}$$

$$\mathbf{T} = \Delta^{-1} \hat{\mathbf{R}}^T \Delta, \tag{46}$$

with $\Delta = diag\langle \xi \rangle$*.*

Proof Transposing $p_k = \pi_k \mathbb{1}$ and inserting $\Delta\Delta^{-1}$ terms leads to

$$p_k = \pi_0 \mathbf{F} \hat{\mathbf{R}}^{k-1} \mathbf{V} \mathbb{1} = \underbrace{\mathbb{1}^T \mathbf{V}^T \Delta}_{\tau} (\underbrace{\Delta^{-1} \hat{\mathbf{R}}^T \Delta}_{\mathbf{T}})^{k-1} \underbrace{\Delta^{-1} \mathbf{F}^T \pi_0^T}_{t_0}. \tag{47}$$

Observe that the entries of $\tau, \mathbf{T}$ and t_0 are all non-negative, since $\hat{\mathbf{R}}, \mathbf{V}, \pi_0$ and Δ are all non-negative.

To prove the theorem, we first show that $t_0 = (\mathbf{I} - \mathbf{T})\mathbb{1}$ as

$$(\mathbf{I} - \mathbf{T})\mathbb{1} = \Delta^{-1}(\mathbf{I} - \hat{\mathbf{R}}^T)\Delta\mathbb{1} = \Delta^{-1}(\mathbf{I} - \hat{\mathbf{R}}^T)\xi^T = \Delta^{-1}\mathbf{F}^T\pi_0^T = t_0. \tag{48}$$

The non-negativity of t_0 implies that the row sums of $\mathbf{T}$ are less then or equal to 1, thus $\mathbf{T}$ is a proper sub-stochastic matrix.

For vector τ we have that

$$\tau\mathbb{1} = \mathbb{1}^T \mathbf{V}^T \Delta\mathbb{1} = \mathbb{1}^T \mathbf{V}^T \xi^T = \pi_0 \mathbf{F}(\mathbf{I} - \hat{\mathbf{R}})^{-1} \mathbf{V}\mathbb{1} = 1 - \pi_0\mathbb{1}, \tag{49}$$

which is clearly less than 1, thus τ is a proper initial vector for a discrete phase-type distribution. □

4 Stationary analysis of the sojourn time of customers

Ozawa showed in [Ozawa(2006)] that the sojourn time in a QBD queue has a matrix-exponential distribution of order N^2.

In this section we prove that the sojourn time is matrix-exponentially distributed in our more general system with matrix-geometric batch arrivals and batch services as well.

If the queue length is k when a new customer arrives, the sojourn time of the newly arrived customer is the time needed by the system to serve $k + 1$ individual customers. Thus, to obtain the distribution of the sojourn time, the following two ingredients are needed:

– The distribution of the queue length after arrival instants,
– and the distribution of the time taken by the system to serve k individual customers.

4.1 The queue length distribution right after arrival instants

Observe that vectors $x_i, i \geq 1$ derived in Section 3 can not be used directly for the sojourn time analysis, since they correspond to the distribution of the queue length and the phase of the background process right after the individual arrivals. However, the service of the customers can start only after the entire batch has arrived. I.e., the vectors representing the joint probability that there are i customers after an individual arrival and the phase of the background process right after the entire batch arrived is

$$x'_i = x_i(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A = \frac{1}{\lambda}\pi_0 \mathbf{F}\hat{\mathbf{R}}^{i-1}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A, \quad i \geq 1. \tag{50}$$

4.2 The behavior of the service process

Let us denote the probability that exactly k individual customers are served till time t by matrix $\mathbf{N}(k,t)$ (the entries of the matrix correspond to the phase transitions between time 0 and t). $\mathbf{N}(k,t)$ is determined by a set of differential equations, similar to the one in [Latouche and Ramaswami(1999)] (Section 3.6) as

$$\frac{\partial}{\partial t}\mathbf{N}(0,t) = \mathbf{N}(0,t)(\mathbf{L} + \mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A), \tag{51}$$

$$\begin{aligned}\frac{\partial}{\partial t}\mathbf{N}(k,t) = {} & \mathbf{N}(k,t)(\mathbf{L} + \mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A) \\ & + \sum_{i=1}^{k} \mathbf{N}(k-i,t)\mathbf{B}\mathbf{X}_S^{i-1}\mathbf{Y}_S, \qquad k = 1, \ldots, \infty.\end{aligned} \tag{52}$$

i.e., transitions not accompanied by service events are characterized by rates $\mathbf{L}+\mathbf{F}(\mathbf{I}-\mathbf{X}_A)^{-1}\mathbf{Y}_A$, while transitions accompanied by the service of i customers are given by $\mathbf{B}\mathbf{X}_S^{i-1}\mathbf{Y}_S$.

4.3 The distribution of the sojourn time

Theorem 6 *The distribution of the sojourn time is given by*

$$P(\mathcal{V} < t) = 1 - (\mathbb{1}^T \otimes \hat{\eta})e^{\mathbf{M}t} vec\langle(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A\rangle, \tag{53}$$

where matrix $\mathbf{M}$ *is equal to*

$$\mathbf{M} = ((\mathbf{L} + \mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A)^T \otimes \mathbf{I}) + (\mathbf{Y}_S^T \otimes \mathbf{I})(\mathbf{I} - \mathbf{X}_S^T \otimes \hat{\mathbf{R}})^{-1}(\mathbf{B}^T \otimes \hat{\mathbf{R}}) \tag{54}$$

and vector $\hat{\eta}$ is the stationary phase distribution at arrivals

$$\hat{\eta} = \pi_0 \mathbf{F}(\mathbf{I} - \hat{\mathbf{R}})^{-1}/\lambda, \tag{55}$$

and vec$\langle\rangle$ denotes the column-stacking operator.

Proof The probability that the sojourn time of an arriving customer is greater than t equals the probability that the number of customers served up to time t is less than the number of customers the arriving customer found in the system (including itself). Hence we have

$$\begin{aligned} P(\mathcal{V} > t) &= \sum_{n=1}^{\infty} x_n' \sum_{k=0}^{n-1} \mathbf{N}(k,t)\mathbb{1} \\ &= \sum_{n=1}^{\infty} x_1 \hat{\mathbf{R}}^{n-1}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A \sum_{k=0}^{n-1} \mathbf{N}(k,t)\mathbb{1} \\ &= \underbrace{\frac{1}{\lambda}\pi_0 \mathbf{F} \sum_{n=1}^{\infty} \hat{\mathbf{R}}^{n-1}}_{\hat{\eta}} \underbrace{\sum_{k=0}^{\infty} \hat{\mathbf{R}}^{k}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A \mathbf{N}(k,t)}_{\mathbf{W}(t)} \mathbb{1}. \end{aligned} \tag{56}$$

thus, $P(\mathcal{V} > t) = \hat{\eta}\mathbf{W}(t)\mathbb{1}$.

To obtain differential equations for $\mathbf{W}(t)$ we have to multiply (51) and (52) by $\hat{\mathbf{R}}^k(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A$ from the left, sum up (52) from 1 to ∞ with regards to k, and add (51) to it. We get

$$\frac{d}{dt}\mathbf{W}(t) = \mathbf{W}(t)(\mathbf{L} + \mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A) + \sum_{i=1}^{\infty} \hat{\mathbf{R}}^i \mathbf{W}(t)\mathbf{B}\mathbf{X}_S^{i-1}\mathbf{Y}_S. \tag{57}$$

Making use of the vec$\langle\rangle$ operator and utilizing that vec$\langle AXB\rangle = (B^T \otimes A)vec\langle X\rangle$ (see [Steeb(1997)]) yields

$$\begin{aligned} \frac{d}{dt}\mathrm{vec}\langle\mathbf{W}(t)\rangle &= ((\mathbf{L} + \mathbf{F}(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A)^T \otimes \mathbf{I})\mathrm{vec}\langle\mathbf{W}(t)\rangle \\ &\quad + \left(\sum_{i=1}^{\infty} \mathbf{Y}_S^T \mathbf{X}_S^{T\,i-1}\mathbf{B}^T \otimes \hat{\mathbf{R}}^i\right)\mathrm{vec}\langle\mathbf{W}(t)\rangle \\ &= \mathbf{M}\,\mathrm{vec}\langle\mathbf{W}(t)\rangle. \end{aligned} \tag{58}$$

Since $\mathbf{N}(k,0)$, the number of customers served in time 0 equals $\mathbf{I}$ if $k = 0$ and 0 if $k > 0$, $\mathbf{W}(0)$ is given by

$$\mathrm{vec}\langle\mathbf{W}(0)\rangle = \mathrm{vec}\left\langle \sum_{k=0}^{\infty} \hat{\mathbf{R}}^k(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A\mathbf{N}(k,0)\right\rangle = \mathrm{vec}\langle(\mathbf{I} - \mathbf{X}_A)^{-1}\mathbf{Y}_A\rangle,$$

from which the closed form solution for $\mathrm{vec}\langle \mathbf{W}(t)\rangle$ is

$$\mathrm{vec}\langle \mathbf{W}(t)\rangle = e^{\mathbf{M}\,t}\mathrm{vec}\langle (\mathbf{I}-\mathbf{X}_A)^{-1}\mathbf{Y}_A\rangle. \tag{59}$$

Finally, the distribution of the sojourn time is given by

$$P(\mathcal{V}<t) = 1-\hat{\eta}\mathbf{W}(t)\mathbb{1} = 1-(\mathbb{1}^T\otimes\hat{\eta})e^{\mathbf{M}\,t}\mathrm{vec}\langle (\mathbf{I}-\mathbf{X}_A)^{-1}\mathbf{Y}_A\rangle, \tag{60}$$

thus the sojourn time distribution is matrix exponential of order N^2. □

4.4 Phase-type representation of the sojourn time distribution

The following theorem is the generalization of Corollary 1 in [Ozawa(2006)], which corresponds to ordinary QBD queues.

Theorem 7 *Assuming that vector $\hat{\eta}$ (see (55)) is strictly positive, the sojourn time of the customers is phase-type distributed with parameters $(\kappa,\mathbf{K})$, thus*

$$P(\mathcal{V}<t) = 1-\kappa e^{\mathbf{K}t}\mathbb{1}. \tag{61}$$

The initial probability vector κ and the transient generator matrix $\mathbf{K}$ are

$$\kappa = vec^T\langle (\mathbf{I}-\mathbf{X}_A)^{-1}\mathbf{Y}_A\rangle(\mathbf{I}\otimes\Delta), \tag{62}$$

$$\begin{aligned}\mathbf{K} = {}& ((\mathbf{L}+\mathbf{F}(\mathbf{I}-\mathbf{X}_A)^{-1}\mathbf{Y}_A)\otimes\mathbf{I})\\ &+(\mathbf{B}\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)(\mathbf{I}-\mathbf{X}_S\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)^{-1}(\mathbf{Y}_S\otimes\mathbf{I}),\end{aligned} \tag{63}$$

with $\Delta = diag\langle\hat{\eta}\rangle$.

Proof Let us transpose $P(\mathcal{V}>t)$ based on (53), and insert $(\mathbf{I}\otimes\Delta)(\mathbf{I}\otimes\Delta^{-1})$ at some places:

$$\begin{aligned}P(\mathcal{V}>t) = {}& \mathrm{vec}^T\langle (\mathbf{I}-\mathbf{X}_A)^{-1}\mathbf{Y}_A\rangle(\mathbf{I}\otimes\Delta)\\ &\times e^{(((\mathbf{L}+\mathbf{F}(\mathbf{I}-\mathbf{X}_A)^{-1}\mathbf{Y}_A)\otimes\mathbf{I})+(\mathbf{B}\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)(\mathbf{I}-\mathbf{X}_S\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)^{-1}(\mathbf{Y}_S\otimes\mathbf{I}))t}\\ &\times(\mathbf{I}\otimes\Delta^{-1})(\mathbb{1}\otimes\hat{\eta}^T).\end{aligned} \tag{64}$$

The first term equals κ, the exponent in the second term is matrix $\mathbf{K}$, while the third term simplifies to $\mathbb{1}$ due to the definition of Δ.

Vector κ is non-negative, and for the sum of the entries we have

$$\begin{aligned}\kappa\mathbb{1} &= \mathrm{vec}^T\langle (\mathbf{I}-\mathbf{X}_A)^{-1}\mathbf{Y}_A\rangle(\mathbb{1}\otimes\hat{\eta}^T)\\ &= (\mathbb{1}^T\otimes\hat{\eta})\mathrm{vec}\langle (\mathbf{I}-\mathbf{X}_A)^{-1}\mathbf{Y}_A\rangle\\ &= \hat{\eta}(\mathbf{I}-\mathbf{X}_A)^{-1}\mathbf{Y}_A\mathbb{1} = \hat{\eta}\mathbb{1} = 1,\end{aligned}$$

thus κ is a proper initial probability vector for a phase-type distribution.

Before investigating matrix $\mathbf{K}$ we show that $\Delta^{-1}\hat{\mathbf{R}}^T\Delta$ is sub-stochastic, which follows from the fact that all entries are non-negative, and the row sums are less than or equal to one according to

$$\mathbb{1} - \Delta^{-1}\hat{\mathbf{R}}^T\Delta\mathbb{1} = \Delta^{-1}(\mathbf{I} - \hat{\mathbf{R}})^T\Delta\mathbb{1} = \Delta^{-1}(\mathbf{I} - \hat{\mathbf{R}})^T\hat{\eta}^T = \Delta^{-1}\mathbf{F}^T\pi_0^T/\lambda \geq 0.$$

As for matrix $\mathbf{K}$, negative entries can appear only in the diagonal due to term $\mathbf{L}\otimes\mathbf{I}$ of the definition (see (63)). Since $(\mathbf{L}+\mathbf{F})\mathbb{1} = -\mathbf{B}\mathbb{1}$, $\mathbf{Y}_S\mathbb{1} = \mathbb{1} - \mathbf{X}_S\mathbb{1}$ and $(\Delta^{-1}\hat{\mathbf{R}}^T\Delta)\mathbb{1} \leq \mathbb{1}$, the row sums of $\mathbf{K}$ are upper bounded as

$$\begin{aligned}
&\mathbf{K}\mathbb{1} = -\mathbf{B}\mathbb{1}\otimes\mathbb{1} + (\mathbf{B}\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)(\mathbf{I} - \mathbf{X}_S\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)^{-1}(\mathbf{Y}_S\mathbb{1}\otimes\mathbb{1})\\
&= -\mathbf{B}\mathbb{1}\otimes\mathbb{1} + (\mathbf{B}\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)(\mathbf{I} - \mathbf{X}_S\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)^{-1}(\mathbf{I} - \mathbf{X}_S\otimes\mathbf{I})\mathbb{1}\\
&\leq -\mathbf{B}\mathbb{1}\otimes\mathbb{1} + (\mathbf{B}\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)(\mathbf{I} - \mathbf{X}_S\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)^{-1}(\mathbf{I} - \mathbf{X}_S\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)\mathbb{1}\\
&\leq -\mathbf{B}\mathbb{1}\otimes\mathbb{1} + (\mathbf{B}\otimes\Delta^{-1}\hat{\mathbf{R}}^T\Delta)\mathbb{1}\\
&\leq 0,
\end{aligned}$$

thus $(\kappa, \mathbf{K})$ define a Markovian phase-type representation. □

Remark 2 Theorem 5 assumes that ξ is strictly positive, and Theorem 7 assumes that $\hat{\eta}$ is strictly positive. The same assumption is made in [Ozawa(2006)] as well. We have to add, however, that this restriction can be relaxed if the pseudo-inverse of Δ is used in the formulas instead of the inverse. Since Δ is a diagonal matrix, this means that all non-zero entries of the diagonal have to be inverted and the zero entries have to be kept. Unfortunately, we have no proof for this generalization yet.

5 The case of independent arrivals and services

In this section a special case of the general model is considered where the arrival and service processes are independent. This special case is essentially a continuous time BMAP/BMAP/1 queue with matrix-geometric batch sizes. The batch arrivals are generated by a MAP characterized by matrices $\check{\mathbf{D}}_\mathbf{0}, \check{\mathbf{D}}_\mathbf{1}$, and the matrix-geometric parameters of the batch sizes are given by $\check{\mathbf{X}}_\mathbf{A}, \check{\mathbf{Y}}_\mathbf{A}$, Hence, the matrices defining the BMAP are

$$\mathbf{D_0} = \check{\mathbf{D}}_\mathbf{0}, \quad \mathbf{D_k} = \check{\mathbf{D}}_\mathbf{1}\check{\mathbf{X}}_\mathbf{A}^{k-1}\check{\mathbf{Y}}_\mathbf{A}, \quad k \geq 1. \tag{65}$$

Similarly, the parameters of the MAP characterizing the batch services are $\check{\mathbf{S}}_\mathbf{0}, \check{\mathbf{S}}_\mathbf{1}, \check{\mathbf{X}}_\mathbf{S}$ and $\check{\mathbf{Y}}_\mathbf{S}$, thus the matrices of the corresponding BMAP are

$$\mathbf{S_0} = \check{\mathbf{S}}_\mathbf{0}, \quad \mathbf{S_k} = \check{\mathbf{S}}_\mathbf{1}\check{\mathbf{X}}_\mathbf{S}^{k-1}\check{\mathbf{Y}}_\mathbf{S}, \quad k \geq 1. \tag{66}$$

The matrix parameters of the corresponding QBD queue are obtained by the appropriate Kronecker operations, giving

$$\begin{aligned}
\mathbf{B} &= \mathbf{I}\otimes\check{\mathbf{S}}_\mathbf{1}, \qquad \mathbf{L} = \check{\mathbf{D}}_\mathbf{0}\oplus\check{\mathbf{S}}_\mathbf{0}, \quad \mathbf{F} = \check{\mathbf{D}}_\mathbf{1}\otimes\mathbf{I}, \quad \mathbf{L_0} = \check{\mathbf{D}}_\mathbf{0}\otimes\mathbf{I},\\
\mathbf{X}_A &= \check{\mathbf{X}}_\mathbf{A}\otimes\mathbf{I}, \quad \mathbf{Y}_A = \check{\mathbf{Y}}_\mathbf{A}\otimes\mathbf{I}, \quad \mathbf{X}_S = \mathbf{I}\otimes\check{\mathbf{X}}_\mathbf{S}, \quad \mathbf{Y}_S = \mathbf{I}\otimes\check{\mathbf{Y}}_\mathbf{S}.
\end{aligned}$$

Based on the results of the previous sections it is possible to obtain order N PH representation for the number of customers in the system and order N^2 PH representation for the sojourn time. The main contribution of this section is that in the special case introduced above a more compact order N representation exists for the sojourn time distribution.

Note that this observation is in line with the results available for the classical case without batches. The sojourn time distribution is of order N^2 in the general case (see [Ozawa(2006)]), and it is of order N if the arrival and service processes are independent, thus in the case of a MAP/MAP/1 queue (see [Sengupta(1990b)] and [He(2012)]). The order N representation is obtained by using the same technique as in [Sengupta(1990b)] and [He(2012)], based on the age process.

5.1 Analysis of the age process

The age process $\{\mathcal{A}(t), t > 0\}$ keeps track of the age of the current customer in the server. It increases by a slope of one during the service periods and it has a downward jump right after the service, the size of the jump equals the inter-arrival time of the next customer.

Here we investigate a two-dimensional Markov process $\{\mathcal{A}(t), \mathcal{Z}(t)\}$, where $\mathcal{A}(t)$ is the age of the customer in the server and $\mathcal{Z}(t)$ is the phase of the system at time t. However, the interpretation of the phases is different from $\mathcal{J}(t)$ used in the previous sections. $\mathcal{Z}(t)$ follows the phase of the service BMAP at time t, and the phase of the arrival BMAP right after the arrival of the individual customer at the head of the queue. The joint density is denoted by $\alpha_i(x) = \frac{d}{dx}P(\mathcal{A}(t) < x, \mathcal{Z}(t) = i)$, and the corresponding vector quantity is $\alpha(x) = [\alpha_i(x)]$.

Before stating the main theorem providing the distribution of $\alpha(x)$, it is beneficial to introduce some matrices in order to shorten the formulas. Hence,

$$\mathbf{D_0'} = (\check{\mathbf{D}}_\mathbf{0} \otimes \mathbf{I}) + (\check{\mathbf{D}}_\mathbf{1} \otimes \mathbf{I})(\mathbf{I} - \check{\mathbf{X}}_\mathbf{A} \otimes \check{\mathbf{X}}_\mathbf{S})^{-1}(\check{\mathbf{Y}}_\mathbf{A} \otimes \check{\mathbf{X}}_\mathbf{S}), \tag{67}$$

$$\mathbf{D_1'} = (\check{\mathbf{D}}_\mathbf{1} \otimes \mathbf{I})(\mathbf{I} - \check{\mathbf{X}}_\mathbf{A} \otimes \check{\mathbf{X}}_\mathbf{S})^{-1}(\mathbf{I} \otimes \check{\mathbf{Y}}_\mathbf{S}), \tag{68}$$

$$\mathbf{S_0'} = (\mathbf{I} \otimes \check{\mathbf{S}}_\mathbf{0}) + (\mathbf{I} \otimes \check{\mathbf{S}}_\mathbf{1})(\mathbf{I} - \check{\mathbf{X}}_\mathbf{A} \otimes \check{\mathbf{X}}_\mathbf{S})^{-1}(\check{\mathbf{X}}_\mathbf{A} \otimes \check{\mathbf{Y}}_\mathbf{S}), \tag{69}$$

$$\mathbf{S_1'} = (\mathbf{I} \otimes \check{\mathbf{S}}_\mathbf{1})(\mathbf{I} - \check{\mathbf{X}}_\mathbf{A} \otimes \check{\mathbf{X}}_\mathbf{S})^{-1}(\check{\mathbf{Y}}_\mathbf{A} \otimes \mathbf{I}). \tag{70}$$

Theorem 8 *Vector $\alpha(x)$ is matrix-exponentially distributed as*

$$\alpha(x) = \alpha(0)e^{\mathbf{T}x}, \tag{71}$$

where matrices $\mathbf{T}$ and $\mathbf{X}$ are the minimal solutions to the matrix equations

$$\begin{aligned} \mathbf{T} &= \mathbf{S_0'} + \mathbf{X}\mathbf{D_1'}, \\ \mathbf{0} &= \mathbf{T}\mathbf{X} + \mathbf{X}\mathbf{D_0'} + \mathbf{S_1'}, \end{aligned} \tag{72}$$

and vector $\alpha(0)$ is the solution to the linear system

$$\alpha(0) = \alpha(0)\mathbf{X}\left((-\check{\mathbf{D}}_\mathbf{0})^{-1}\check{\mathbf{D}}_\mathbf{1} \otimes (\mathbf{I} - \check{\mathbf{X}}_\mathbf{S})^{-1}\check{\mathbf{Y}}_\mathbf{S}\right), \quad \alpha(0)(-\mathbf{T})^{-1}\mathbb{1} = 1. \tag{73}$$

Proof The main characteristics of the age process are as follows.

$\mathcal{A}(t)$ increases at a slope of one until a batch service occurs. At a batch service instant several customers leave the system immediately. If the size of the service batch (N_S, the number of customers that can be served in the batch) is less than the size of the (remaining) arrival batch located at the head of the queue (N_A), then, after the departure of N_S customers, the customer at the head of the queue belongs to the same batch as the departing customers. Consequently, the age process continues to increase at a slope of one, there is no downward jump.

The age process has a downward jump only if $N_A \leq N_S$, when the size of the remaining batch of arrivals waiting at the head of the queue is not greater than the number of customers the server can serve. In this case, the server starts to serve further (younger) customers from the queue, that have a lower age.

Hence, due to the matrix-geometric nature of the batch sizes the probability density that the duration of the increasing period is t with the corresponding phase transitions is $e^{\mathbf{S}'_{\mathbf{0}}t}\mathbf{S}'_{\mathbf{1}}$ where $\mathbf{S}'_{\mathbf{0}}$ and $\mathbf{S}'_{\mathbf{1}}$ are defined by (69) and (70).

The lengths of the downward jumps are not trivial to characterize either. When $N_A \leq N_S$, the server serves further (potentially many) arrival batches from the queue up to N_S, or up to the point when the queue gets empty. Again, due to the matrix-geometric batch sizes the density of the length of the jump is $e^{\mathbf{D}'_{\mathbf{0}}t}\mathbf{D}'_{\mathbf{1}}$, with (67) and (68).

Based on these considerations the differential equation describing the evolution of $\{\mathcal{A}(t), \mathcal{Z}(t)\}$ is as follows:

$$\begin{aligned}\alpha_i(t,x) = \alpha_i(t-\Delta, x-\Delta)(1-\mathbf{S}'_{\mathbf{0}ii}\Delta) + \sum_{j\neq i}\alpha_j(t-\Delta,x-\Delta)\mathbf{S}'_{\mathbf{0}ji}\Delta \\ + \int_{u=0}^{\infty}\sum_{\forall j}\alpha_j(t-\Delta,x+u)\Delta\left[\mathbf{S}'_{\mathbf{1}}e^{\mathbf{D}'_{\mathbf{0}}u}\mathbf{D}'_{\mathbf{1}}\right]_{ji}du,\end{aligned} \tag{74}$$

that, letting $t\to\infty, \Delta\to 0$ and expressing in vector form equals

$$\frac{d}{dx}\alpha(x) = \alpha(x)\mathbf{S}'_{\mathbf{0}} + \int_{u=0}^{\infty}\alpha(x+u)\mathbf{S}'_{\mathbf{1}}e^{\mathbf{D}'_{\mathbf{0}}u}\mathbf{D}'_{\mathbf{1}}\,du. \tag{75}$$

According to [Sengupta(1990b)] the solution for $\alpha(x)$ is matrix-exponential. Inserting (71) into the differential equation leads to

$$\mathbf{T} = \mathbf{S}'_{\mathbf{0}} + \underbrace{\int_{u=0}^{\infty}e^{\mathbf{T}u}\mathbf{S}'_{\mathbf{1}}e^{\mathbf{D}'_{\mathbf{0}}u}\,du}_{\mathbf{X}}\mathbf{D}'_{\mathbf{1}}, \tag{76}$$

which provides (72) since the value of the integral (denoted by $\mathbf{X}$) can be obtained as the solution of a Sylvester equation.

To express $\alpha(0)$, we have to observe that the age process can be 0 if the size of the service batch N_S is greater than or equal to the number of customers

present in the system. Hence

$$\begin{aligned}\alpha(0) &= \int_{x=0}^{\infty} \alpha(x)\mathbf{S}_1' e^{\mathbf{D}_0' x}\, dx \left((-\check{\mathbf{D}}_0)^{-1}\check{\mathbf{D}}_1 \otimes (\mathbf{I}-\check{\mathbf{X}}_\mathbf{S})^{-1}\check{\mathbf{Y}}_\mathbf{S}\right) \\ &= \alpha(0)\mathbf{X}\left((-\check{\mathbf{D}}_0)^{-1}\check{\mathbf{D}}_1 \otimes (\mathbf{I}-\check{\mathbf{X}}_\mathbf{S})^{-1}\check{\mathbf{Y}}_\mathbf{S}\right),\end{aligned} \tag{77}$$

where $(\mathbf{I}-\check{\mathbf{X}}_\mathbf{S})^{-1}\check{\mathbf{Y}}_\mathbf{S}$ corresponds to the phase transition of the service process when the batch is closed (there are no customers to serve any more), and $(-\check{\mathbf{D}}_0)^{-1}\check{\mathbf{D}}_1$ provides the phase transitions of the arrival process till a new busy period is initiated, where the age process is defined again.

Equation $\alpha(0)(-\mathbf{T})^{-1}\mathbb{1} = 1$ in (73) comes from the normalization condition.

Since $\mathbf{X}$ is a stochastic matrix (it contains the phase transition probabilities over the non-zero periods of the age process), matrices $(-\check{\mathbf{D}}_0)^{-1}\check{\mathbf{D}}_1$ and $(\mathbf{I}-\check{\mathbf{X}}_\mathbf{S})^{-1}\check{\mathbf{Y}}_\mathbf{S}$ are both stochastic as well, (73) defines a fully determined system of equations. □

The straight forward procedure to obtain matrix $\mathbf{T}$ numerically is to apply a functional iteration as shown in Algorithm 4.

Algorithm 4 Functional iteration to obtain $\mathbf{T}$ and $\mathbf{X}$

$\mathbf{T}^{(0)} \leftarrow \mathbf{S}_0'$
$n \leftarrow 0$
repeat
  $\mathbf{X}^{(n+1)} \leftarrow$ solution of $\mathbf{T}^{(n)}\mathbf{X}^{(n+1)} + \mathbf{X}^{(n+1)}\mathbf{D}_0' + \mathbf{S}_1' = \mathbf{0}$
  $\mathbf{T}^{(n+1)} \leftarrow \mathbf{S}_0' + \mathbf{X}^{(n+1)}\mathbf{D}_1'$
  $n \leftarrow n+1$
until $||\mathbf{T}^{(n)} - \mathbf{T}^{(n-1)}|| < \epsilon$
return $\mathbf{T}^{(n)}, \mathbf{X}^{(n)}$

Remark 3 The functional iteration for $\mathbf{T}$ suffers from linear convergence. Based on the intuition learned from [Horváth et al(2014)Horváth, Van Houdt, and Telek] we found that matrix $\mathbf{T}$ can be expressed from $\hat{\mathbf{R}}$ (for which quadratically convergent algorithms exist) as

$$\mathbf{T} = (\mathbf{I}\otimes\check{\mathbf{S}}_0) + \sum_{i=1}^{\infty}\hat{\mathbf{R}}^i(\mathbf{I}\otimes\check{\mathbf{S}}_1)(\mathbf{I}\otimes\check{\mathbf{X}}_\mathbf{S})^{i-1}(\mathbf{I}\otimes\check{\mathbf{Y}}_\mathbf{S}), \tag{78}$$

where the sum is the solution of a discrete Sylvester equation, but unfortunately we can not prove this relation yet.

5.2 The distribution of the sojourn time

In many queueing models the sojourn time distribution can be easily derived from the distribution of the age process, since the sojourn time of a customer is equal to its age at the departure instant. In our model, however, there are batch arrivals and services, making this approach a bit more difficult to apply.

Theorem 9 *The probability density function of the sojourn time, $w(x)$, is given by*

$$\begin{aligned} w(x) &= \frac{1}{c}\alpha(x) \\ &\times \left((\mathbf{I}\otimes\check{\mathbf{S}}_\mathbf{1})(\mathbf{I}-\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})^{-1}+\mathbf{X}(\check{\mathbf{D}}_\mathbf{1}\otimes\mathbf{I})(\mathbf{I}-\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})^{-1}(\mathbf{I}\otimes\check{\mathbf{X}}_\mathbf{S})\right)\mathbb{1}, \end{aligned} \tag{79}$$

where c is a normalization constant

$$\begin{aligned} c &= \alpha(0)(-\mathbf{T})^{-1} \\ &\times \left((\mathbf{I}\otimes\check{\mathbf{S}}_\mathbf{1})(\mathbf{I}-\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})^{-1}+\mathbf{X}(\check{\mathbf{D}}_\mathbf{1}\otimes\mathbf{I})(\mathbf{I}-\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})^{-1}(\mathbf{I}\otimes\check{\mathbf{X}}_\mathbf{S})\right)\mathbb{1}. \end{aligned} \tag{80}$$

Proof Two cases are distinguished leading to sojourn time x.

1. If $\mathcal{A}(t)=x$, and a batch service starts at time t, all customers belonging to the (remaining) arrival batch at the head of the queue which are taken by the service batch have a sojourn time of x.
2. When $\mathcal{A}(t)=x+u$ $(u>0)$, and a batch service starts at time t serving all customers from the batch at the head of the queue, further younger arrivals are served as well. If a customer arriving u time later than the one at the head of the queue gets served by the batch service as well, its sojourn time will be x.

   Taking into account the transitions leading to case 1 we get

$$\alpha(x)(\mathbf{I}\otimes\check{\mathbf{S}}_\mathbf{1})\sum_{k=0}^{\infty}(\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})^k(k+1)(\underbrace{\check{\mathbf{Y}}_\mathbf{A}\otimes\mathbf{I}}+\underbrace{\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{Y}}_\mathbf{S}})\mathbb{1}, \tag{81}$$

where the first under-braced term corresponds to the case when all $k+1$ customers of the remaining arrival batch at the head of the queue are served (and further customers might be served as well, but they will have a sojourn time other than x). The second under-braced term belongs to the reverse situation: the size of the service batch was $k+1$, but it was not enough to serve the arrival batch at the head of the queue. Some manipulations (exploiting that $\check{\mathbf{Y}}_\mathbf{A}\mathbb{1}=(\mathbf{I}-\check{\mathbf{X}}_\mathbf{A})\mathbb{1}$ and $\check{\mathbf{Y}}_\mathbf{S}\mathbb{1}=(\mathbf{I}-\check{\mathbf{X}}_\mathbf{S})\mathbb{1}$) lead to

$$\begin{aligned} &= \alpha(x)(\mathbf{I}\otimes\check{\mathbf{S}}_\mathbf{1})(\mathbf{I}-\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})^{-2}(\mathbb{1}-(\check{\mathbf{X}}_\mathbf{A}\otimes\mathbf{I})\mathbb{1}+(\check{\mathbf{X}}_\mathbf{A}\otimes\mathbf{I})\mathbb{1}-(\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})\mathbb{1}) \\ &= \alpha(x)(\mathbf{I}\otimes\check{\mathbf{S}}_\mathbf{1})(\mathbf{I}-\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})^{-2}(\mathbf{I}-\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})\mathbb{1}, \end{aligned}$$

which is the first term of (79).

For case 2 we have

$$\begin{aligned} &\int_{u=0}^{\infty}\alpha(x+u)\mathbf{S}'_\mathbf{1}e^{\mathbf{D}'_\mathbf{0}u}(\check{\mathbf{D}}_\mathbf{1}\otimes\mathbf{I})\left(\sum_{k=0}^{\infty}(\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})^k(\check{\mathbf{Y}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})\cdot(k+1)\right. \\ &\left.+\sum_{k=0}^{\infty}(\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})^k(\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{Y}}_\mathbf{S})\cdot k+\sum_{k=0}^{\infty}(\check{\mathbf{X}}_\mathbf{A}\otimes\check{\mathbf{X}}_\mathbf{S})^k(\check{\mathbf{Y}}_\mathbf{A}\otimes\check{\mathbf{Y}}_\mathbf{S})\cdot k\right)\mathbb{1}\,du. \end{aligned} \tag{82}$$

The first sum represents the case when the batch arriving u time later than the one at the head of the queue gets fully served. The kth term of the sum

corresponds to $k+1$ served customers. In case of the second and third sum of (82) the service batch ends before serving the entire arrival batch, leading to k services in the kth term of the sums. Manipulating (82) gives

$$\begin{aligned} &= \alpha(x)\mathbf{X}(\check{\mathbf{D}}_\mathbf{1} \otimes \mathbf{I})\Bigg(\sum_{k=0}^{\infty} (\check{\mathbf{X}}_\mathbf{A} \otimes \check{\mathbf{X}}_\mathbf{S})^k (k+1)(\check{\mathbf{Y}}_\mathbf{A} \otimes \check{\mathbf{X}}_\mathbf{S} + \mathbf{I} \otimes \check{\mathbf{Y}}_\mathbf{S})\mathbb{1} \\ &\quad - \sum_{k=0}^{\infty} (\check{\mathbf{X}}_\mathbf{A} \otimes \check{\mathbf{X}}_\mathbf{S})^k (\mathbf{I} \otimes \check{\mathbf{Y}}_\mathbf{S})\mathbb{1} \Bigg) \\ &= \alpha(x)\mathbf{X}(\check{\mathbf{D}}_\mathbf{1} \otimes \mathbf{I})\Big((\mathbf{I} - \check{\mathbf{X}}_\mathbf{A} \otimes \check{\mathbf{X}}_\mathbf{S})^{-1}\mathbb{1} - (\mathbf{I} - \check{\mathbf{X}}_\mathbf{A} \otimes \check{\mathbf{X}}_\mathbf{S})^{-1}(\mathbf{I} \otimes \check{\mathbf{Y}}_\mathbf{S})\mathbb{1} \Big), \end{aligned} \tag{83}$$

which equals the second term of (79). □

6 Summary of the analysis methods

This section gives a short summary on the analysis procedures proposed in this paper.

Given the 8 matrices characterizing the model ($\mathbf{L_0}, \mathbf{L}, \mathbf{F}, \mathbf{X}_A, \mathbf{Y}_A, \mathbf{B}, \mathbf{X}_S$ and $\mathbf{Y}_S$), the three main steps to obtain the distribution of the number of customers in the system are as follows.

1. The computation of matrices $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$ by Algorithm 1, 2 or 3.
2. Obtaining vector π_0 by solving the linear system (23).
3. The matrix-geometric distribution of the number of customers is given by (31).

The distribution of the sojourn time of the individual customers is determined by the following three steps.

1. The computation of $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$ by Algorithm 1, 2 or 3, and the computation of π_0 by (23).
2. The distribution of the sojourn time is provided by Theorem 6.
3. Theorem 7 gives the phase-type representation for the sojourn time, if needed.

If the arrival and service processes are independent, the smaller representation for the sojourn time distribution is given by the steps below.

1. Calculate matrix $\mathbf{T}$ by Algorithm 4, or from matrix $\hat{\mathbf{R}}$ by (78).
2. Obtain the initial vector of the age process $\alpha(0)$ by the solution of (73).
3. The density of the sojourn time is provided by Theorem 9.

7 Numerical examples

We have implemented the presented procedures in Matlab environment[2]. To solve the matrix-quadratic equations the cyclic reduction based algorithm of

[2] The implementation can be downloaded from `http://www.hit.bme.hu/~ghorvath/software`

Algorithm 1:	0.092636 s
Algorithm 2:	0.003452 s
Algorithm 3:	0.002932 s

Table 1 Execution times of different algorithms to determine matrices $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$

	1^{st} moment	2^{nd} moment	3^{rd} moment
Number of customers:	9.1588	183.418	5505.36
Sojourn time:	1.9302	7.4727	43.4

Table 2 Moments of the number of customers and the sojourn times

the SMCSolver tool [Bini et al(2006)Bini, Meini, Steffé, and Van Houdt] was used. The Sylvester equations were solved by the built-in `lyap` function of Matlab, which is based on the Hessenberg–Schur algorithm. All execution time related results are obtained on an average PC with an Intel Core i7-2600 CPU clocked at 3.4 GHz and having 4 GB of RAM.

7.1 Example with dependent arrival and service processes

The arrival and service related matrices used in the first numerical example are as follows:

$$\mathbf{F} = \begin{bmatrix} 2 & 1 & 0 \\ 5 & 0 & 0 \\ 1 & 3 & 0 \end{bmatrix}, \quad \mathbf{X}_A = \begin{bmatrix} 0.04 & 0.2 & 0.03 \\ 0.09 & 0 & 0.02 \\ 0.07 & 0.05 & 0.01 \end{bmatrix}, \quad \mathbf{Y}_A = \begin{bmatrix} 0.13 & 0.25 & 0.35 \\ 0.59 & 0.2 & 0.1 \\ 0.7 & 0.17 & 0 \end{bmatrix},$$
$$\mathbf{B} = \begin{bmatrix} 6 & 1 & 0 \\ 0 & 4 & 1 \\ 2 & 0 & 0 \end{bmatrix}, \quad \mathbf{X}_S = \begin{bmatrix} 0.12 & 0.01 & 0.02 \\ 0 & 0.07 & 0.04 \\ 0.1 & 0.03 & 0.08 \end{bmatrix}, \quad \mathbf{Y}_S = \begin{bmatrix} 0.5 & 0.2 & 0.15 \\ 0.89 & 0 & 0 \\ 0.79 & 0 & 0 \end{bmatrix},$$

and the internal transitions are given by

$$\mathbf{L} = \begin{bmatrix} -15 & 3 & 2 \\ 0 & -14 & 4 \\ 3 & 1 & -10 \end{bmatrix}, \quad \mathbf{L_0} = \begin{bmatrix} -6 & 0 & 3 \\ 4 & -12 & 3 \\ 1 & 1 & -6 \end{bmatrix}.$$

With these parameters the downward drift is 5.5797, and the upward one is 4.789, thus the system is stable according to (5).

As the first step, matrices $\hat{\mathbf{R}}, \mathbf{V}$ and $\mathbf{H}$ need to be determined, that are necessary both for the queue length and the sojourn time analysis. The execution times (the average of 10 executions) of all three algorithms are depicted in Table 1. Among the tree algorithms the QBD based (cyclic reduction) was slightly faster than the Newton iteration.

Table 2 presents the moments of the number of customers in the system and the moments of the sojourn time of individual customers, finally, the corresponding distributions are depicted in Figure 1.

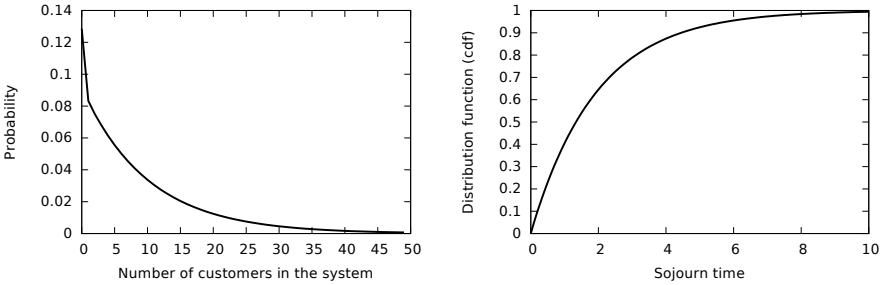

Fig. 1 Distribution of the number of customers in the system and the sojourn time

7.2 Example with independent arrival and service processes

To investigate how scalable the solution method is, the arrival and service processes are constructed such that their size and their drift can be adjusted arbitrarily. The structure of the matrices characterizing the arrival process are

$$\check{\mathbf{D}}_\mathbf{0} = \begin{bmatrix} \bullet & K\nu_A & & & \\ \gamma_A & \bullet & (K-1)\nu_A & & \\ & \ddots & & \ddots & \ddots \\ & & (K-1)\gamma_A & \bullet & \nu_A \\ & & & K\gamma_A & \bullet \end{bmatrix}, \quad \check{\mathbf{D}}_\mathbf{1} = \begin{bmatrix} 0 & & & & \\ & r_A/K & & & \\ & & 2r_A/K & & \\ & & & \ddots & \\ & & & & r_A \end{bmatrix},$$

and for the batch sizes $\check{\mathbf{X}}_\mathbf{A}$ and $\check{\mathbf{Y}}_\mathbf{A}$ are diagonal matrices such that $\check{\mathbf{X}}_\mathbf{A} = x_A\mathbf{I}$ and $\check{\mathbf{Y}}_\mathbf{A} = (1-x_A)\mathbf{I}$. In this particular example the fixed parameters are $\nu_A = 0.5, \gamma_A = 1.0, x_A = 0.35$, while K and r_A are changing. The diagonal entries denoted by $\bullet$ are determined uniquely such that the row sums $\check{\mathbf{D}}_\mathbf{0} + \check{\mathbf{D}}_\mathbf{1}$ are zeros.

The matrix parameters of the service time have a similar structure, with $\nu_S = 0.33, \gamma_S = 0.2, x_S = 0.55$.

In the first experiment we investigate how long it takes to determine the parameters of the steady state distribution (vector π_0 and matrices $\hat{\mathbf{R}}$ and $\mathbf{V}$) as the function of the drift. The downward drift is set to 5, while the upward drift is changed in the stability region between 0.1 and 4.9 by the appropriate setting of r_A. The number of phases of the arrival and service processes are both $K = 5$. According to the results (see Figure 2), the functional iteration (Algorithm 1) performs the worst (as expected), while the two quadratically convergent algorithms (Algorithm 2 and 3) are more than one magnitude faster. The QBD based (cyclic reduction) method seems to be especially insensitive to the utilization of the queue.

In the second experiment the upward drift is set to 3, and the effect of the number of phases (K) is investigated. Figure 3 shows the superiority of the quadratically convergent algorithms again. Note that at the last point, at $K = 20$ the matrices are huge, the algorithms operate on 400×400 matrices, and Algorithm 3, which is based on the reduction of the problem to a QBD,

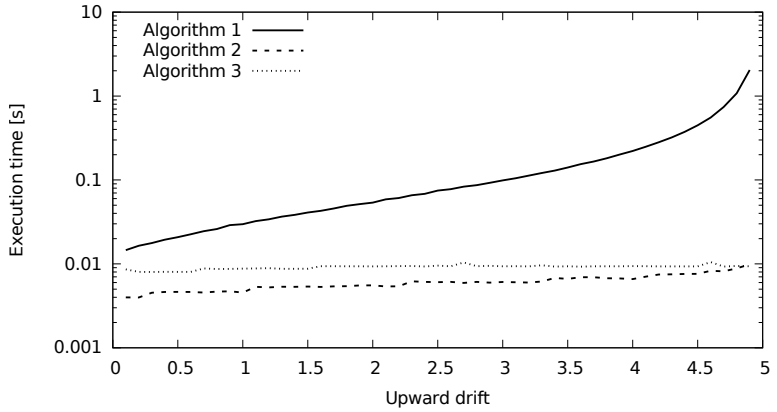

Fig. 2 Execution time vs. the utilization of the queue

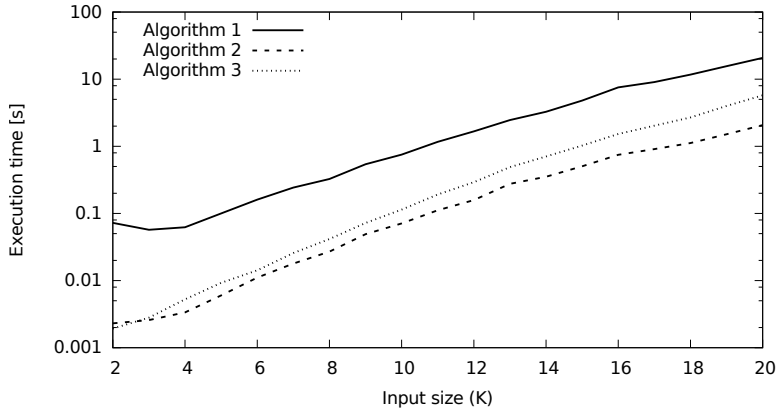

Fig. 3 Execution time vs. the input size

operates on 1200×1200 matrices. Despite of the large size of the input, the execution time of the analysis is still reasonable.

In the final experiment the speed of the sojourn time analysis is studied as the function of the number of phases (K). Three solution methods are involved in the comparison. The first one is the one developed in Section 4, which produces an order N^2 representation. In the other two cases we exploit the independence of the arrival and service processes in order to obtain a much smaller, order N representation. In one of these methods matrix $\mathbf{T}$ is obtained by Algorithm 4, and in the other one it is obtained by (78) with Algorithm 2. According to Figure 4 it is obvious that it is well worth to exploit the independence. The largest model the general method could solve corresponds to $K = 7$, where this general method returned a size 2401 representation, which is both slow to obtain and too large to work with. The memory of the computer was too small to compute cases for $K > 7$.

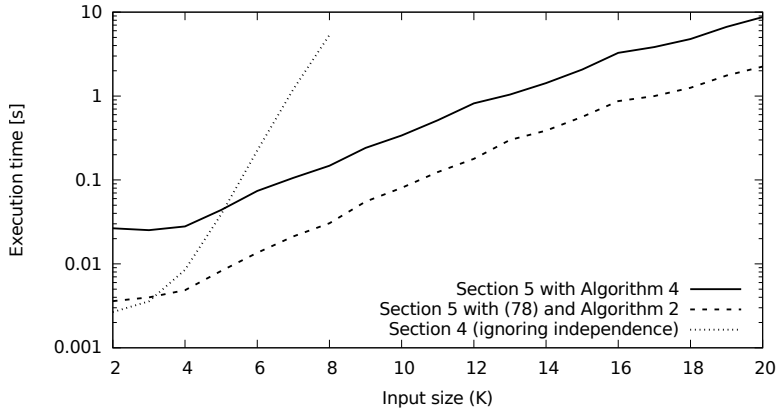

Fig. 4 Execution time for the sojourn time parameters vs. the input size

Acknowledgements This work was supported by the Hungarian research project OTKA K101150, and by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

[Antoulas(2005)] Antoulas AC (2005) Approximation of large-scale dynamical systems, vol 6. Siam

[Bini et al(2006)Bini, Meini, Steffé, and Van Houdt] Bini D, Meini B, Steffé S, Van Houdt B (2006) Structured Markov chains solver: software tools. In: Proceeding from the 2006 workshop on Tools for solving structured Markov chains, ACM, p 14

[Chakka and Harrison(2001)] Chakka R, Harrison P (2001) The MMCPP/GE/c queue. Queueing Systems 38(3):307–326

[Éltető and Telek(2008)] Éltető T, Telek M (2008) Numerical analysis of M/G/1 type queueing systems with phase type transition structure. Journal of Computational and Applied Mathematics 212(2):331–340

[Gail et al(1997)Gail, Hantler, and Taylor] Gail H, Hantler S, Taylor B (1997) Non-skip-free M/G/1 and G/M/1 type Markov chains. Advances in Applied Probability pp 733–758

[Golub et al(1979)Golub, Nash, and Van Loan] Golub GH, Nash S, Van Loan C (1979) A Hessenberg-Schur method for the problem AX+XB=C. Automatic Control, IEEE Transactions on 24(6):909–913

[Harrison and Zatschler(2004)] Harrison PG, Zatschler H (2004) Sojourn time distributions in modulated G-queues with batch processing. In: Quantitative Evaluation of Systems, 2004. Proceedings. First International Conference on the, pp 90–99

[He(2012)] He Q (2012) Analysis of a continuous time SM[K]/PH[K]/1/FCFS queue: Age process, sojourn times, and queue lengths. Journal of Systems Science and Complexity 25(1):133–155

[Horváth et al(2014)Horváth, Van Houdt, and Telek] Horváth G, Van Houdt B, Telek M (2014) Commuting matrices in the queue length and sojourn time analysis of MAP/MAP/1 queues. Stochastic Models 30(4):554–575

[Jafari and Sohraby(2001)] Jafari R, Sohraby K (2001) Combined M/G/1-G/M/1 type structured chains: a simple algorithmic solution and applications. In: INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings, vol 2, pp 1065–1074

[Latouche and Ramaswami(1999)] Latouche G, Ramaswami V (1999) Introduction to matrix analytic methods in stochastic modeling. ASA-SIAM, Philadelphia

[Neuts(1981)] Neuts MF (1981) Matrix-geometric solutions in stochastic models: an algorithmic approach. Courier Corporation
[Neuts(1989)] Neuts MF (1989) Structured stochastic matrices of M/G/1 type and their applications. Dekker
[Ozawa(2006)] Ozawa T (2006) Sojourn time distributions in the queue defined by a general QBD process. Queueing Systems 53(4):203–211
[Sengupta(1990a)] Sengupta B (1990a) Phase-type representations for matrix-geometric solutions. Stochastic Models 6(1):163–167
[Sengupta(1990b)] Sengupta B (1990b) The semi-Markovian queue: theory and applications. Stochastic Models 6(3):383–413
[Steeb(1997)] Steeb WH (1997) Matrix calculus and Kronecker product with applications and C++ programs. World Scientific