

# Delay analysis of a queue with re-sequencing buffer and Markov environment

Rostislav Razumchik<sup>1</sup> · Miklós Telek<sup>2</sup>

Received: 14 November 2014 / Revised: 16 March 2015 / Published online: 20 April 2015  
© Springer Science+Business Media New York 2015

**Abstract** There are simple service disciplines where the system time of a tagged customer depends only on the customers arriving in the system earlier (for example first-in-first-out (FIFO)) or later (for example LIFO) than the tagged one. In this paper we consider a single-server queueing system with two infinite queues in which the system time of a tagged customer may depend on both the customers arriving in the system earlier and later than the tagged one. New regular customers arrive in the system according to Markov arrival process (MAP) flow, occupy one place in the buffer and receive service in FIFO order. External re-sequencing signals also arrive at the system according to (different) MAP flow. Each re-sequencing signal transforms one regular customer into a delayed one by moving it to another queue (re-sequencing buffer), wherefrom it is served with lower priority than the regular ones. Service times of customers from both queues also have MAP distribution different from those which govern arrivals. Queueing system with memoryless ingredients (arrival, service, resequencing) has already been a subject of extensive research. In this paper we investigate how the essential analytical properties of scalar functions, which made the analysis of the memoryless system feasible, can be extended to the case of a Markov environment.

**Keywords** Delay analysis · Re-sequencing buffer · Matrix analytic methods · Kronecker expansion

---

✉ Rostislav Razumchik  
rrazumchik@gmail.com

Miklós Telek  
telek@hit.bme.hu

<sup>1</sup> Institute of Informatics Problems, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences and Peoples’ Friendship University of Russia, Moscow, Russia

<sup>2</sup> MTA-BME Information Systems Research Group and Technical University of Budapest, Budapest, Hungary

**Mathematics Subject Classification** 68M20 · 60K25**1 Introduction**

Since the introduction of the first matrix analytic methods [9] a wide range of Markov chain-based analysis approaches has been extended to Markov chains with regular matrix block structures. Currently, popular text books summarize basic matrix analytic approaches, for example [3, 8, 10], which establish the belief that the Markov chain-based analysis of stochastic models with memoryless components can be extended to the same models with modulating Markov environment. For example, the analysis of a queueing system with a Poisson arrival process can be extended to the analysis of the same system with a Markov arrival process (MAP). For simple queueing systems and straightforward performance measures this extension is more or less standard today. For important performance measures of complex queueing systems several standard queueing methodologies are not applicable when the system operates in a Markov environment [5], and it is still a challenge to find adequate matrix analytic methodology to handle these.

In this paper we consider the delay analysis of a rather complex queueing system operating in a Markov environment with customer re-sequencing. Regular customers arrive at the system and occupy one place in a buffer of infinite capacity. Re-sequencing signals arrive at the system and upon arrival each re-sequencing signal moves one regular customer from the buffer to another queue (re-sequencing buffer) of infinite capacity and leaves the system. There is one server which serves customers from both queues. Upon service completion one regular customer from the buffer goes to the server, or, only if there are no regular customers in the buffer, one customer from the re-sequencing buffer enters service. No service interruption is allowed. This model with memoryless components has already been solved in [12]. The basic idea of the solution method in [12] is rather standard: evaluate the joint stationary distribution of the number of customers in both queues (whose generating function has a closed form) and compute the delay distribution of a tagged customer which arrives at the system in steady state. Due to a geometric dependence of the delay distribution (expressed in terms of Laplace transform) of an arriving regular customer on the number of customers in both queues, delay distribution is obtained by appropriate substitution of Laplace transform functions into the parameters of the generating function describing the joint stationary distribution.

In the first phase of our research we very much shared the general belief of the extendibility of the analysis to Markov environments, but it turned out that the commonly applied approaches, including the one in [12], are not applicable in case of a Markov environment due to the presence of non-commuting matrices and matrix functions. In order to maintain our general belief we looked for an appropriate variant of the analysis approach which allows evaluation with non-commuting matrices.

One part of the proposed analysis approach is the application of Kronecker algebra [2, 13], which, to some extent, overcomes the limitations imposed by non-commuting matrices. Since the very beginning, the use of Kronecker algebra is quite common in matrix analytic methods [8, 10]. In this respect the only contribution of the paper is

a case study which demonstrates that multiple application of Kronecker transformations allows the analysis of more and more complex expressions with non-commuting matrices, which otherwise are not computable in closed form [6,7]. The other part of the proposed analysis approach is the variant of the queue analysis which allows matrix-based computations. In this respect the major difference is that our matrix-based analysis cannot be decomposed into stationary analysis of the number of customers and delay analysis of a tagged customer, but we have to compute both of these “ingredients” at once.

The rest of the paper is structured as follows. Section 2 briefly introduces the queueing model with re-sequencing and Sect. 3 summarizes the QBD-type analysis of the Markov chain which describes the joint stationary distribution of the number of customers in each buffer and states of the Markov environment. The analysis of the waiting time is presented in Sect. 4 and the most complex part of the analysis is deferred to Sect. 5. Finally, some results of numerical experiments carried out using obtained expressions are in Sect. 6.

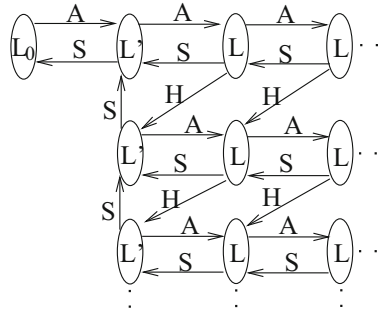
## 2 Model description

We consider a queueing system with two buffers: the regular buffer with high-priority customers and the re-sequencing buffer with low-priority customers. Arriving regular customers are added to the regular buffer of infinite capacity and wait for service. External re-sequencing signals also arrive at the system and each signal moves one customer at the head of the regular buffer (if any) to the re-sequencing buffer of infinite capacity and itself leaves the system. The service policy of customers in the regular and re-sequencing buffers is non-preemptive priority with the first-in-first-out (FIFO) discipline within each buffer, i.e. an arriving customer (also further referred to as high priority) occupies one place at the end of the regular buffer, and a re-sequenced customer (also further referred to as low priority) occupies one place at the end of the re-sequencing buffer. There is one server which serves customers from both buffers and the service process is the same for both high and low priority customers.

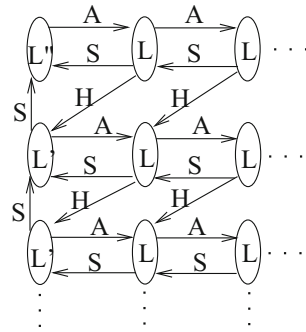
Customers arrive according to a MAP with generator matrices  $(\mathbf{A}_0, \mathbf{A}_1)$ , the service process is a MAP with  $(\mathbf{S}_0, \mathbf{S}_1)$  and re-sequencing signals arrive according to a MAP with  $(\mathbf{H}_0, \mathbf{H}_1)$ . Let  $\mathbf{A}_J = \mathbf{A}_0 + \mathbf{A}_1, \mathbf{S}_J = \mathbf{S}_0 + \mathbf{S}_1$  and  $\mathbf{H}_J = \mathbf{H}_0 + \mathbf{H}_1$ , denote the phase processes of the associated MAPs (see, for example, [8] for details). The block structure of the Markov chain representing the number of high- and low-priority customers in the system is depicted in Fig. 1. The block represents the set of states with the same number of high- and low-priority customers and with different phases of the MAPs. The letters on the figures describe

- arrival of a customer:  $\mathcal{A} = \mathbf{A}_1 \otimes I \otimes I,$
- service of a customer:  $\mathcal{S} = I \otimes \mathbf{S}_1 \otimes I,$
- re-sequencing of a customer:  $\mathcal{H} = I \otimes I \otimes \mathbf{H}_1,$
- phase change when re-sequencing is possible:  $\mathcal{L} = \mathbf{A}_0 \oplus \mathbf{S}_0 \oplus \mathbf{H}_0,$
- phase change when re-sequencing is not possible:  $\mathcal{L}' = \mathbf{A}_0 \oplus \mathbf{S}_0 \oplus \mathbf{H}_J,$
- phase change when re-sequencing is not possible and the service process is stopped:  $\mathcal{L}_0 = \mathbf{A}_0 \otimes I \oplus \mathbf{H}_J = \mathbf{A}_0 \otimes I \otimes I + I \otimes I \otimes \mathbf{H}_J,$

**Fig. 1** Block structure of the Markov chain representing the number of regular (high priority) and re-sequenced (low priority) customers



**Fig. 2** Block structure of the censored Markov chain representing the number of regular (high priority) and re-sequenced (low priority) customers



where  $I$  denotes the identity matrix of appropriate size. One has to note that the phase of the service process does not change when the system is empty and is equal to that phase in which the next service starts. The main goal of the analysis is to evaluate the stationary waiting time distribution of a regular customer arriving in the system.

### 3 Joint stationary distribution

Before deriving expressions for the stationary waiting time distribution, one has to obtain expressions for the joint stationary distribution of the number of customers in the regular buffer and the re-sequencing buffer, and the states of the regular and resequencing arrivals and service processes.

#### 3.1 Censored process

To simplify the analysis and obtain a Markov chain with a regular structure we censor the Markov chain in Fig. 1 for the cases when the server is busy. The structure of the censored Markov chain is depicted in Fig. 2. The transitions of the upper left block of the censored chain are obtained as

$$\mathcal{L}'' = \mathcal{L}' - S\mathcal{L}_0^{-1}A = (\mathbf{A}_0 \oplus \mathbf{S}_0 \oplus \mathbf{H}_J) - (I \otimes \mathbf{S}_1 \otimes I)(\mathbf{A}_0 \otimes I \oplus \mathbf{H}_J)^{-1}(\mathbf{A}_1 \otimes I \otimes I).$$

### 3.2 QBD representation of the censored process

Following, for example, the discussion of Section 13.1 in [8], we can represent the censored Markov chain as a QBD process where the levels are composed of the set of states where the number of regular customers is the same (these states form the columns of blocks in Fig. 2). The generator  $\mathbb{Q}$  of the censored process can be represented in hyper-block tridiagonal form, where the hyper-block refers to the set of (infinitely many) states on the same level.

$$\mathbb{Q} = \begin{pmatrix} \mathbb{L}' & \mathbb{F} & 0 & 0 & 0 & \dots \\ \mathbb{B} & \mathbb{L} & \mathbb{F} & 0 & 0 & \dots \\ 0 & \mathbb{B} & \mathbb{L} & \mathbb{F} & 0 & \dots \\ 0 & 0 & \mathbb{B} & \mathbb{L} & \mathbb{F} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Due to the fact that the number of states within each level is infinite, matrices  $\mathbb{L}'$ ,  $\mathbb{L}$ ,  $\mathbb{B}$ ,  $\mathbb{F}$  have infinite rows and columns which are associated with the blocks in Fig. 2.

$$\mathbb{L}' = \begin{pmatrix} \mathcal{L}'' & 0 & 0 & \dots \\ \mathcal{S} & \mathcal{L}' & 0 & \dots \\ 0 & \mathcal{S} & \mathcal{L}' & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad \mathbb{L} = \begin{pmatrix} \mathcal{L} & 0 & 0 & \dots \\ 0 & \mathcal{L} & 0 & \dots \\ 0 & 0 & \mathcal{L} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad \mathbb{F} = \begin{pmatrix} \mathcal{A} & 0 & 0 & \dots \\ 0 & \mathcal{A} & 0 & \dots \\ 0 & 0 & \mathcal{A} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

$$\mathbb{B} = \begin{pmatrix} \mathcal{S} & \mathcal{H} & 0 & 0 & \dots \\ 0 & \mathcal{S} & \mathcal{H} & 0 & \dots \\ 0 & 0 & \mathcal{S} & \mathcal{H} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

### 3.3 Condition for stability

The number of customers increases in the system due to arrivals, whose average rate is  $\lambda$ , and decreases due to service, whose average rate is  $\mu$ . Since the considered system is work conserving and the re-sequencing signal does not change the number of customer in the system, the condition for stability is  $\mu > \lambda$ . Here  $\lambda$  is computed from  $(\mathbf{A}_0, \mathbf{A}_1)$  as  $\lambda = \pi_A \mathbf{A}_1 \mathbf{1}$ , where  $\pi_A$  is the solution of  $\pi_A \mathbf{A}_J = 0$ ,  $\pi_A \mathbf{1} = 1$  and  $\mathbf{1}$  is the column vector of ones of the appropriate size. The value of  $\mu$  can be computed similarly from  $(\mathbf{S}_0, \mathbf{S}_1)$ . Throughout the paper it is assumed that the queue is stable.

### 3.4 QBD analysis of the process

In the censored Markov chain we denote the stationary probability vector of the set of states with  $i$  regular and  $j$  delayed customers by  $\pi_{ij}$  ( $i, j \geq 0$ ) and compose the following row vectors

$$\mathbf{p}_i = (\pi_{i,0}, \pi_{i,1}, \pi_{i,2}, \pi_{i,3}, \dots), \quad i \geq 0,$$

$$\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \dots).$$

Considering the hyper-block structure of  $\mathbb{Q}$ , the linear infinite system of equations  $\mathbf{p}\mathbb{Q} = \mathbf{0}, \mathbf{p}\mathbf{1} = 1$  have the following hyper-block structure

$$\mathbf{p}_0\mathbb{L}' + \mathbf{p}_1\mathbb{B} = \mathbf{0}, \tag{1}$$

$$\mathbf{p}_{i-1}\mathbb{F} + \mathbf{p}_i\mathbb{L} + \mathbf{p}_{i+1}\mathbb{B} = \mathbf{0}, \quad \forall i \geq 1, \tag{2}$$

$$\sum_{k=0}^{\infty} \mathbf{p}_k\mathbf{1} = 1. \tag{3}$$

The solution of Eqs. (1–3) has a matrix geometric structure  $\mathbf{p}_i = \mathbf{p}_{i-1}\mathbb{R}, i \geq 1$ , where the matrix  $\mathbb{R}$  is the minimal non-negative solution of the equation

$$\mathbb{F} + \mathbb{R}\mathbb{L} + \mathbb{R}^2\mathbb{B} = \mathbf{0}, \tag{4}$$

where  $\mathbf{0}$  denotes the zero matrix [8]. Due to the level-independent behaviour of the  $\mathbb{Q}$  matrix,  $\mathbb{R}$  has the following upper-diagonal block-Toeplitz form

$$\mathbb{R} = \begin{pmatrix} \mathbf{R}_0 & \mathbf{R}_1 & \mathbf{R}_2 & \mathbf{R}_3 & \mathbf{R}_4 & \dots \\ 0 & \mathbf{R}_0 & \mathbf{R}_1 & \mathbf{R}_2 & \mathbf{R}_3 & \dots \\ 0 & 0 & \mathbf{R}_0 & \mathbf{R}_1 & \mathbf{R}_2 & \dots \\ 0 & 0 & 0 & \mathbf{R}_0 & \mathbf{R}_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Based on the block structure of  $\mathbb{R}$ , the hyper-block level equation  $\mathbf{p}_i = \mathbf{p}_{i-1}\mathbb{R}$  can be written as

$$\pi_{ij} = \sum_{k=0}^j \pi_{i-1,k} \mathbf{R}_{j-k}, \quad i \geq 1, \quad j \geq 0. \tag{5}$$

The explicit computation of the  $\mathbf{R}_j$  matrices is possible, in general, utilizing the structure of matrices  $\mathbb{L}, \mathbb{B}, \mathbb{F}$  and  $\mathbb{R}$ , and we can write equations for the  $(i, j)$ th block of matrix  $\mathbb{F} + \mathbb{R}\mathbb{L} + \mathbb{R}^2\mathbb{B}$ . It can be verified by direct calculation that Eq. (4) is equivalent to the following system of equations:

$$\mathbf{A} + \mathbf{R}_0\mathcal{L} + \mathbf{R}_0^2\mathcal{S} = \mathbf{0}, \tag{6}$$

$$\mathbf{R}_1\mathcal{L} + \mathbf{R}_0^2\mathcal{H} + \sum_{i=0}^1 \mathbf{R}_i\mathbf{R}_{1-i}\mathcal{S} = \mathbf{0}, \tag{7}$$

$$\mathbf{R}_j\mathcal{L} + \sum_{i=0}^{j-1} \mathbf{R}_i\mathbf{R}_{j-i-1}\mathcal{H} + \sum_{i=0}^j \mathbf{R}_i\mathbf{R}_{j-i}\mathcal{S} = \mathbf{0}, \quad j \geq 2. \tag{8}$$

The elements of the main diagonal of the matrix  $\mathbb{F} + \mathbb{R}\mathbb{L} + \mathbb{R}^2\mathbb{B}$  are equal to the left part of (6), and the elements of the first upper diagonal are equal to the left part of (7). Finally, the  $j$ -th upper diagonals contain elements which are equal to the left part of (8) for the corresponding value of  $j$ .

Let us introduce the matrix generating function

$$\bar{\mathbf{R}}(z) = \sum_{i=0}^{\infty} z^i \mathbf{R}_i, \quad |z| < 1.$$

Multiplying the left-hand and right-hand sides of (6) by  $z^0$ , (7) by  $z^1$ , and (8) by  $z^j$ , and summing up over all values of  $j \geq 0$ , we obtain

$$\mathcal{A} + \mathbf{R}_0\mathcal{L} + \mathbf{R}_0^2\mathcal{S} + \sum_{j=1}^{\infty} z^j \left( \mathbf{R}_j\mathcal{L} + \sum_{i=0}^{j-1} \mathbf{R}_i\mathbf{R}_{j-i-1}\mathcal{H} + \sum_{i=0}^j \mathbf{R}_i\mathbf{R}_{j-i}\mathcal{S} \right) = \mathbf{0}.$$

Therefore  $\bar{\mathbf{R}}(z)$  is the minimal non-negative solution of the quadratic matrix equation

$$\mathcal{A} + \bar{\mathbf{R}}(z)\mathcal{L} + \bar{\mathbf{R}}^2(z)(z\mathcal{H} + \mathcal{S}) = \mathbf{0}. \tag{9}$$

Throughout the paper we rely on the assumption that efficient numerical methods are available for the solution of quadratic matrix equations [1,4] and consequently we consider the matrices defined by quadratic matrix equations to be known. From (5), for  $\hat{\pi}_i(z) = \sum_{j=0}^{\infty} \pi_{ij}z^j, i \geq 1$ , we have

$$\hat{\pi}_i(z) = \sum_{j=0}^{\infty} \pi_{ij}z^j = \sum_{j=0}^{\infty} z^j \sum_{k=0}^j \pi_{i-1,k} \mathbf{R}_{j-k} = \hat{\pi}_{i-1}(z)\bar{\mathbf{R}}(z). \tag{10}$$

Similarly to the definition of  $\mathbb{R}$ , we can define the matrix  $\mathbb{G}$  of the hyper-block QBD process, which is the minimal non-negative solution of  $\mathbb{B} + \mathbb{L}\mathbb{G} + \mathbb{F}\mathbb{G}^2 = \mathbf{0}$ . The level independent block structure of the hyper-block QBD process ensures an upper-diagonal block-Toeplitz form for the matrix  $\mathbb{G}$  as well.

$$\mathbb{G} = \begin{pmatrix} \mathbf{G}_0 & \mathbf{G}_1 & \mathbf{G}_2 & \mathbf{G}_3 & \mathbf{G}_4 & \dots \\ 0 & \mathbf{G}_0 & \mathbf{G}_1 & \mathbf{G}_2 & \mathbf{G}_3 & \dots \\ 0 & 0 & \mathbf{G}_0 & \mathbf{G}_1 & \mathbf{G}_2 & \dots \\ 0 & 0 & 0 & \mathbf{G}_0 & \mathbf{G}_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The equation for the  $(i, j)$ th block of the matrix  $\mathbb{B} + \mathbb{L}\mathbb{G} + \mathbb{F}\mathbb{G}^2$  gives

$$\mathcal{S} + \mathcal{L}\mathbf{G}_0 + \mathcal{A}\mathbf{G}_0^2 = \mathbf{0}, \tag{11}$$

$$\mathcal{H} + \mathcal{L}\mathbf{G}_1 + \mathcal{A} \sum_{i=0}^1 \mathbf{G}_i\mathbf{G}_{1-i} = \mathbf{0}, \tag{12}$$

$$\mathcal{L}\mathbf{G}_j + \mathcal{A} \sum_{i=0}^j \mathbf{G}_i \mathbf{G}_{j-i} = \mathbf{0}, \quad j \geq 2. \tag{13}$$

Introducing the matrix generating function

$$\overline{\mathbf{G}}(z) = \sum_{i=0}^{\infty} z^i \mathbf{G}_i, \quad |z| < 1,$$

multiplying the  $j$ -th equation of (11–13) with  $z^j$  and summing up over all values of  $j$  leads to

$$(S + z\mathcal{H}) + \mathcal{L}\overline{\mathbf{G}}(z) + \mathcal{A}\overline{\mathbf{G}}^2(z) = \mathbf{0}. \tag{14}$$

### 3.5 Censored process on level 0

From the hyper-block level description we can obtain the generator of the censored process on level 0 as  $\mathbb{L}' + \mathbb{F}\mathbb{G}$ , which has the following M/G/1-type block level structure

$$\mathbb{L}' + \mathbb{F}\mathbb{G} = \begin{pmatrix} \mathcal{L}'' + \mathcal{A}\mathbf{G}_0 & \mathcal{A}\mathbf{G}_1 & \mathcal{A}\mathbf{G}_2 & \mathcal{A}\mathbf{G}_3 & \mathcal{A}\mathbf{G}_4 & \dots \\ S & \mathcal{L}' + \mathcal{A}\mathbf{G}_0 & \mathcal{A}\mathbf{G}_1 & \mathcal{A}\mathbf{G}_2 & \mathcal{A}\mathbf{G}_3 & \dots \\ 0 & S & \mathcal{L}' + \mathcal{A}\mathbf{G}_0 & \mathcal{A}\mathbf{G}_1 & \mathcal{A}\mathbf{G}_2 & \dots \\ 0 & 0 & S & \mathcal{L}' + \mathcal{A}\mathbf{G}_0 & \mathcal{A}\mathbf{G}_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The fundamental matrix of this M/G/1-type Markov chain,  $\check{\mathbf{G}}$ , is the minimal non-negative solution of the matrix equation

$$S + \mathcal{L}'\check{\mathbf{G}} + \mathcal{A} \sum_{i=0}^{\infty} \mathbf{G}_i \check{\mathbf{G}}^{i+1} = \mathbf{0}, \tag{15}$$

for which efficient numerical analysis is available [1] as well.

### 3.6 Obtaining $\pi_{00}$

Based on the fundamental matrix of the censored process on level 0 we can compute the generator matrix of the censored process on level 0 and block 0:

$$\mathbf{Q}_{00} = \mathcal{L}'' + \mathcal{A} \sum_{i=0}^{\infty} \mathbf{G}_i \check{\mathbf{G}}^i = \mathcal{L}' - \underbrace{S\mathcal{L}_0^{-1}\mathcal{A}} + \mathcal{A} \sum_{i=0}^{\infty} \mathbf{G}_i \check{\mathbf{G}}^i, \tag{16}$$

and  $\pi_{00}$  is the properly normalized solution of  $\pi_{00}\mathbf{Q}_{00} = 0$ . We note that the under-braced term comes from censoring out the block with an idle server. To compute  $\pi_{00}$  we need to get back to the original non-censored Markov chain in Fig. 1.



Let  $\pi_{\text{idle}}$  be the stationary distribution of the block of states representing an idle server (the left most block in Fig. 1). The stationary utilization of the system is  $\rho = \lambda/\mu$ , from which  $\pi_{\text{idle}}\mathbf{1} = 1 - \rho$ .

We compute the stationary distribution of  $\pi_{\text{idle}}$  in two steps. First we consider the non-censored Markov chain in Fig. 1 and analyse the behaviour of this Markov chain restricted to the block of states representing an idle server (the left most block in Fig. 1) and to the block of states representing a busy server and idle queues (the block next to the left most one in Fig. 1). According to the state transitions between these two blocks in Fig. 1 and (16), the generator of the process restricted to these two blocks is

$$\mathbf{Q}_{nq} = \begin{pmatrix} \mathcal{L}_0 & \mathcal{A} \\ \mathcal{S} & \mathcal{L}' + \mathcal{A} \sum_{i=0}^{\infty} \mathbf{G}_i \check{\mathbf{G}}^i \end{pmatrix} = \begin{pmatrix} \mathcal{L}_0 & \mathcal{A} \\ \mathcal{S} & \mathbf{T}_0 \end{pmatrix},$$

where we introduced the notation  $\mathbf{T}_0 = \mathcal{L}' + \mathcal{A} \sum_{i=0}^{\infty} \mathbf{G}_i \check{\mathbf{G}}^i$ . Further censoring this process to the idle block we have

$$\mathbf{Q}_{\text{idle}} = \mathcal{L}_0 - \mathcal{A}\mathbf{T}_0^{-1}\mathcal{S},$$

and  $\pi_{\text{idle}}$  is obtained as the solution of the linear system of equations

$$\pi_{\text{idle}}\mathbf{Q}_{\text{idle}} = 0 \quad \text{and} \quad \pi_{\text{idle}}\mathbf{1} = 1 - \rho.$$

Finally,  $\pi_{00}$  is obtained from

$$\pi_{00} = -\pi_{\text{idle}}\mathcal{A}\mathbf{T}_0^{-1}.$$

### 3.7 Computing $\hat{\pi}_0(z)$

The components of the vector  $\mathbf{p}_0 = (\pi_{0,0}, \pi_{0,1}, \pi_{0,2}, \pi_{0,3}, \dots)$ , can be computed from Ramaswami’s recursive formula (see, for example, [8]). Specifically,

$$\pi_{0,m} = -\left(\sum_{i=0}^{m-1} \pi_{0,i}\mathbf{T}_{m-i}\right)\mathbf{T}_0^{-1}, \quad m \geq 1, \tag{17}$$

where  $\mathbf{T}_0$  is defined above and

$$\mathbf{T}_m = \mathcal{A} \sum_{k=m}^{\infty} \mathbf{G}_k \check{\mathbf{G}}^{k-m}, \quad m \geq 1.$$

Though Ramaswami’s formula gives a way to compute  $\mathbf{p}_0$ , we are interested in an expression for its components  $\pi_{0,m}$ ,  $m \geq 0$ , in terms of the generating function

$$\hat{\pi}_0(z) = \sum_{m=0}^{\infty} \pi_{0,m}z^m, \quad 0 < z < 1.$$

In order to obtain an equation for  $\hat{\pi}_0(z)$  one has to write the equation  $\mathbf{p}_0(\mathbb{L}' + \mathbb{F}\mathbb{G}) = 0$  in block form:

$$\pi_{0,0}(\mathcal{L}'' + \mathcal{A}\mathbf{G}_0) + \pi_{0,1}\mathcal{S} = 0, \quad (18)$$

$$\pi_{0,0}\mathcal{A}\mathbf{G}_1 + \pi_{0,1}(\mathcal{L}' + \mathcal{A}\mathbf{G}_0) + \pi_{0,2}\mathcal{S} = 0, \quad (19)$$

$$\sum_{k=0}^j \pi_{0,k}\mathcal{A}\mathbf{G}_{j-k} + \pi_{0,j}\mathcal{L}' + \pi_{0,j+1}\mathcal{S} = 0, \quad j \geq 2. \quad (20)$$

By multiplying the  $k$ -th equation with  $z^k$  and summing up over all values of  $k$  one obtains

$$0 = \hat{\pi}_0(z) \left( \mathcal{A}\overline{\mathbf{G}}(z) + \mathcal{L}' + \frac{1}{z}\mathcal{S} \right) + \pi_{0,0} \left( \mathcal{L}'' - \mathcal{L}' - \frac{1}{z}\mathcal{S} \right) \quad (21)$$

and

$$\hat{\pi}_0(z) = \pi_{0,0} \left( \mathcal{L}' - \mathcal{L}'' + \frac{1}{z}\mathcal{S} \right) \left( \mathcal{A}\overline{\mathbf{G}}(z) + \mathcal{L}' + \frac{1}{z}\mathcal{S} \right)^{-1}. \quad (22)$$

We note that the function  $\hat{\pi}_0(z)$  is undefined at  $z = 1$ , and  $\hat{\pi}_0(1)$ , whenever used, should be read as  $\lim_{z \rightarrow 1} \hat{\pi}_0(z)$ .

### 3.8 Distribution immediately after customer arrival

As MAP arrival do not see time averages (that is, the PASTA property does not hold) one has to calculate stationary probabilities  $\tilde{\pi}_{ij}$  that after a customer arrival there are  $i$  ( $i \geq 1$ ) customers in the regular buffer and  $j$  ( $j \geq 0$ ) in the re-sequencing buffer. Following the same argument as in [11], we can write

$$\tilde{\pi}_{ij} = \frac{1}{\lambda} \pi_{i-1,j} \mathcal{A}, \quad i \geq 1, \quad j \geq 0, \quad \text{and} \quad \tilde{\pi}_{00} = \frac{1}{\lambda} \pi_{\text{idle}} \mathcal{A}.$$

### 3.9 Analysis with Kronecker expansion

In the sequel we need to evaluate various infinite summations of matrix expressions with non-commuting matrices. The analysis of those expressions is based on the technique which we refer to as Kronecker expansion [2, 13] and is based on the identity  $\text{vec}(ABC) = (C^T \otimes A)\text{vec}(B)$ . In this identity  $\text{vec}$  denotes the column stacking vector operator which transforms a matrix of size  $n \times m$  into a vector of size  $nm \times 1$ . We are going to utilize the property of the  $\text{vec}$  operator that  $\text{vec}(A) = A$  for a matrix  $A$  of size  $n \times 1$ . We note that the above identity has several seemingly different forms, for example  $\text{vec}(AB) = (I^T \otimes A)\text{vec}(B) = (B^T \otimes A)\text{vec}(I) = (B^T \otimes I)\text{vec}(A)$ .

Indeed, depending on the complexity of the obtained matrix expressions, we need to apply the Kronecker expansion multiple times, which generates larger and larger matrix expressions. This is a price we need to pay for generalizing the memoryless models to a Markov-modulated environment.

### 4 Stationary waiting time distribution

We analyse the stationary waiting time ( $W$ ), starting from the instant when a regular customer arrives at the system up to the instant when it enters service. Its distribution will be evaluated in terms of the Laplace transform  $\omega(s) = E(e^{-sW})$ . A regular customer may arrive at the server from the regular buffer or from the re-sequencing buffer, and thus the stationary waiting time distribution can be computed as

$$\begin{aligned} \omega(s) &= E(e^{-sW}) = \omega_H(s) + \omega_L(s) \\ &= E(e^{-sW} I_{\{\text{served from regular buffer}\}}) \\ &\quad + E\left(e^{-sW} I_{\{\text{served from re-sequencing buffer}\}}\right), \end{aligned}$$

where  $I_{\{a\}}$  is the indicator of event  $a$ .

#### 4.1 Stationary waiting time distribution of customer that receives service from regular buffer

When a customer is served from the regular buffer the waiting time is

$$\begin{aligned} \omega_H(s) &= E(e^{-sW} I_{\{\text{served from regular buffer}\}}) \\ &= \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \tilde{\pi}_{ij} \left( (sI - \mathcal{L}_A)^{-1} (\mathcal{S} + \mathcal{H}) \right)^{i-1} (sI - \mathcal{L}_A)^{-1} \mathcal{S} \mathbf{1} \\ &= \frac{1}{\lambda} \sum_{i=1}^{\infty} \underbrace{\sum_{j=0}^{\infty} \pi_{i-1,j}}_{\hat{\pi}_{i-1}(1)} \underbrace{\mathcal{A} \left( (sI - \mathcal{L}_A)^{-1} (\mathcal{S} + \mathcal{H}) \right)^{i-1}}_{\mathbf{U}(s)} \underbrace{(sI - \mathcal{L}_A)^{-1} \mathcal{S} \mathbf{1}}_{v(s)} \\ &= \frac{1}{\lambda} \sum_{i=0}^{\infty} \hat{\pi}_i(1) \mathcal{A} \mathbf{U}(s)^i v(s) = \frac{1}{\lambda} (v(s)^T \otimes \mathbf{1}) \underbrace{\sum_{i=0}^{\infty} \left( \mathbf{U}(s)^{iT} \otimes \hat{\pi}_i(1) \right)}_{\mathbf{K}(s)} \text{vec}(\mathcal{A}), \end{aligned}$$

where  $\mathcal{L}_A = \mathbf{A}_J \oplus \mathbf{S}_0 \oplus \mathbf{H}_0$ ,  $\text{vec}$  is the column stacking vector operator and we used the identity  $\text{vec}(ABC) = (C^T \otimes A)\text{vec}(B)$ . For  $\mathbf{K}(s)$  we have

$$\begin{aligned} \mathbf{K}(s) &= \sum_{i=0}^{\infty} \left( \mathbf{U}(s)^{iT} \otimes \hat{\pi}_i(1) \right) = (I \otimes \hat{\pi}_0(1)) + \sum_{i=1}^{\infty} \left( \mathbf{U}(s)^{iT} \otimes \hat{\pi}_i(1) \right) \\ &= (I \otimes \hat{\pi}_0(1)) + \sum_{i=1}^{\infty} \left( \mathbf{U}(s)^{i-1T} \mathbf{U}(s)^T \otimes \hat{\pi}_{i-1}(1) \bar{\mathbf{R}}(1) \right) \\ &= (I \otimes \hat{\pi}_0(1)) + \mathbf{K}(s) \left( \mathbf{U}(s)^T \otimes \bar{\mathbf{R}}(1) \right) \end{aligned}$$

and

$$\mathbf{K}(s) = (I \otimes \hat{\pi}_0(1)) \left( I - \mathbf{U}(s)^T \otimes \bar{\mathbf{R}}(1) \right)^{-1}. \quad (23)$$

In this final expression for  $\mathbf{K}(s)$  the values of  $\bar{\mathbf{R}}(1)$  and  $\hat{\pi}_0(1)$  are obtained from (9) and (22), respectively. Note that without the Kronecker expansion the expression for  $\omega_H(s)$  could not be obtained in closed form because of non-commutativity of the matrices  $\mathcal{A}$  and  $\mathbf{U}(s)$ . In the scalar case these matrices are reduced to scalars and  $\omega_H(s)$  is computed at once in terms of the generating function. In the Markov environment considered one had to search a representation in the form  $\mathbf{K}(s) = K_1(s) + \mathbf{K}(s)K_2(s)$ . Here  $K_1(s)$  is the 0-th element of the infinite sum. The second term  $\mathbf{K}(s)K_2(s)$  is, on one hand, the sum from 1 to  $\infty$  and, on the other hand, it is the product of the original infinite sum from 0 to  $\infty$  with some matrix  $K_2(s)$ .

#### 4.2 Stationary waiting time distribution of the customer that receives service from re-sequencing buffer

For  $j \leq i$ , let  $\mathbb{F}(t, i, j, k)$  be the matrix (according to the initial and final phases of the MAPs  $(\mathbf{A}_0, \mathbf{A}_1)$ ,  $(\mathbf{S}_0, \mathbf{S}_1)$  and  $(\mathbf{H}_0, \mathbf{H}_1)$ ) of the probabilities that  $k$  customers arrive,  $i - j$  customers are served and  $j$  are moved to the re-sequencing buffer in time  $t$ , when the initial number of customers in the regular buffer is larger than  $i$ . For the Laplace transform  $\tilde{\mathbb{F}}(s, i, j, k) = \int_t e^{-st} \mathbb{F}(t, i, j, k) dt$  we have

$$\tilde{\mathbb{F}}(s, 0, 0, 0) = (sI - \mathcal{L})^{-1} = \mathcal{L}(s), \quad (24)$$

and otherwise

$$\begin{aligned} \tilde{\mathbb{F}}(s, i, j, k) &= I_{\{i > j\}} \mathcal{L}(s) \mathcal{S} \tilde{\mathbb{F}}(s, i - 1, j, k) \\ &\quad + I_{\{j > 0\}} \mathcal{L}(s) \mathcal{H} \tilde{\mathbb{F}}(s, i - 1, j - 1, k) \\ &\quad + I_{\{k > 0\}} \mathcal{L}(s) \mathcal{A} \tilde{\mathbb{F}}(s, i, j, k - 1), \end{aligned} \quad (25)$$

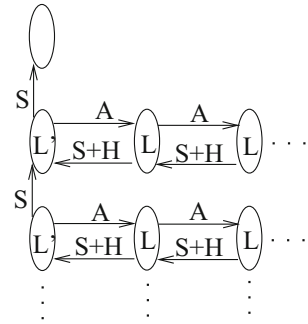
where  $\mathcal{L}(s)$  is defined in (24). The case that the tagged customer moves to the re-sequencing buffer is described by  $\tilde{\mathbb{F}}(s, i, j, k) \mathcal{H}$ .

After getting to the re-sequencing buffer, the customer needs to wait for the service of the high-priority customers (in the regular buffer) and the low-priority customers which were in the re-sequencing buffer. The Markov chain representing the waiting time in the re-sequencing buffer is depicted in Fig. 3.

The matrix Laplace transform of the time to reduce the number of high-priority customers by one,  $\tilde{\mathbf{G}}(s)$ , is the solution of

$$s\tilde{\mathbf{G}}(s) = (\mathcal{S} + \mathcal{H}) + \mathcal{L}\tilde{\mathbf{G}}(s) + \mathcal{A}\tilde{\mathbf{G}}(s)^2,$$

**Fig. 3** Block structure of the Markov chain representing the waiting time in the re-sequencing buffer



and the Laplace transform of the time to reduce the number of low-priority customers (in the re-sequencing buffer) by one is the solution of

$$s\widehat{\mathbf{G}}(s) = \mathcal{S} + \mathcal{L}'\widehat{\mathbf{G}}(s) + \mathcal{A}\widetilde{\mathbf{G}}(s)\widehat{\mathbf{G}}(s).$$

Hence, if the number of customers in the regular buffer is  $j \geq 0$  and in the re-sequencing buffer is  $k \geq 0$  at the arrival of the tagged customer in the re-sequencing buffer, then the subsequent waiting time is  $\widetilde{\mathbf{G}}(s)^j \widehat{\mathbf{G}}(s)^{k+1}$ .

Based on the previously computed matrix Laplace transforms, the waiting time of the customer which enters service from the re-sequencing buffer can be computed as

$$\begin{aligned} \omega_L(s) &= E \left( e^{-sW} I_{\{\text{served from re-sequencing buffer}\}} \right) \\ &= \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \tilde{\pi}_{ij} \sum_{\ell=0}^{i-1} \sum_{k=0}^{\infty} \tilde{\mathbb{F}}(s, i-1, \ell, k) \mathcal{H} \widetilde{\mathbf{G}}(s)^k \widehat{\mathbf{G}}(s)^{j+\ell+1} \mathbf{1} \\ &= \frac{1}{\lambda} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \pi_{i,j} \mathcal{A} \sum_{\ell=0}^i \sum_{k=0}^{\infty} \tilde{\mathbb{F}}(s, i, \ell, k) \mathcal{H} \widetilde{\mathbf{G}}(s)^k \widehat{\mathbf{G}}(s)^{j+\ell+1} \mathbf{1}. \end{aligned} \tag{26}$$

The main part of the analysis of  $\omega_L(s)$  is deferred to the next section. But in the course of the subsequent derivations we will make use of several quantities which are better introduced if one firstly considers terms of  $\omega_L(s)$  with  $i = 0$ . Thus we represent  $\omega_L(s)$  as

$$\begin{aligned} \omega_L(s) &= \omega_L^{i>0}(s) + \omega_L^{i=0}(s) \\ &= \frac{1}{\lambda} \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \pi_{i,j} \mathcal{A} \sum_{\ell=0}^i \sum_{k=0}^{\infty} \tilde{\mathbb{F}}(s, i, \ell, k) \mathcal{H} \widetilde{\mathbf{G}}(s)^k \widehat{\mathbf{G}}(s)^{j+\ell+1} \mathbf{1} \\ &\quad + \frac{1}{\lambda} \sum_{j=0}^{\infty} \pi_{0,j} \mathcal{A} \sum_{k=0}^{\infty} \underbrace{\tilde{\mathbb{F}}(s, 0, 0, k)}_{(\mathcal{L}(s), \mathcal{A})^k \mathcal{L}(s)} \mathcal{H} \widetilde{\mathbf{G}}(s)^k \widehat{\mathbf{G}}(s)^{j+1} \mathbf{1} \end{aligned} \tag{27}$$

and in the next subsection derive expression for  $\omega_L^{i=0}(s)$ .

### 4.3 Computation of $\omega_L^{i=0}(s)$

Applying the identity  $vec(ABC) = (C^T \otimes A)vec(B)$ , to  $\omega_L^{i=0}(s)$  one obtains

$$\begin{aligned} \omega_L^{i=0}(s) &= \frac{1}{\lambda} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \pi_{0,j} \mathcal{A}(\mathcal{L}(s) \mathcal{A})^k \mathcal{L}(s) \mathcal{H} \tilde{\mathbf{G}}(s)^k \hat{\mathbf{G}}(s)^{j+1} \mathbf{1} \\ &= \frac{1}{\lambda} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \left( \mathbf{1}^T \hat{\mathbf{G}}(s)^{j+1} \tilde{\mathbf{G}}(s)^k \otimes \pi_{0,j} \mathcal{A}(\mathcal{L}(s) \mathcal{A})^k \right) vec(\mathcal{L}(s) \mathcal{H}) \\ &= \frac{1}{\lambda} \left( \mathbf{1}^T \hat{\mathbf{G}}(s)^T \otimes \mathbf{1} \right) \\ &\quad \times \underbrace{\sum_{j=0}^{\infty} \left( \hat{\mathbf{G}}(s)^{jT} \otimes \pi_{0,j} \right)}_{\Theta(s)} \underbrace{(I \otimes \mathcal{A}) \sum_{k=0}^{\infty} \left( \tilde{\mathbf{G}}(s)^k \otimes (\mathcal{L}(s) \mathcal{A})^k \right)}_{(I - \tilde{\mathbf{G}}(s)^T \otimes \mathcal{L}(s) \mathcal{A})^{-1}} vec(\mathcal{L}(s) \mathcal{H}) \\ &= \frac{1}{\lambda} \left( \mathbf{1}^T \hat{\mathbf{G}}(s)^T \otimes \mathbf{1} \right) \Theta(s) (I \otimes \mathcal{A}) \left( I - \tilde{\mathbf{G}}(s)^T \otimes \mathcal{L}(s) \mathcal{A} \right)^{-1} vec(\mathcal{L}(s) \mathcal{H}). \end{aligned}$$

The only unknown in this expression is  $\Theta(s)$ . In order to compute it we revisit (18–20). Kronecker multiply the  $j$ -th equation with  $\hat{\mathbf{G}}(s)^{j+1T}$  from the left and sum up over all values of  $j$ . This gives

$$\begin{aligned} 0 &= \left( \hat{\mathbf{G}}(s)^T \otimes \pi_{0,0} (\mathcal{L}'' - \mathcal{L}') \right) + \underbrace{\sum_{j=0}^{\infty} \left( \hat{\mathbf{G}}(s)^{jT} \otimes \pi_{0,j} \right)}_{\Theta(s)} \left( \hat{\mathbf{G}}(s)^T \otimes \mathcal{L}' \right) \\ &\quad + \underbrace{\sum_{j=0}^{\infty} \left( \hat{\mathbf{G}}(s)^{j+1T} \otimes \pi_{0,j+1} \right)}_{\Theta(s) - (I \otimes \pi_{0,0})} (I \otimes \mathcal{S}) + \sum_{j=0}^{\infty} \left( \hat{\mathbf{G}}(s)^{j+1T} \otimes \sum_{k=0}^j \pi_{0,k} \mathcal{A} \mathbf{G}_{j-k} \right). \end{aligned}$$

The last term can be rewritten as

$$\begin{aligned} &\sum_{j=0}^{\infty} \left( \hat{\mathbf{G}}(s)^{j+1T} \otimes \sum_{k=0}^j \pi_{0,k} \mathcal{A} \mathbf{G}_{j-k} \right) \\ &= \underbrace{\sum_{k=0}^{\infty} \left( \hat{\mathbf{G}}(s)^k \otimes \pi_{0,k} \right)}_{\Theta(s)} \left( \hat{\mathbf{G}}(s)^T \otimes \mathcal{A} \right) \underbrace{\sum_{j=k}^{\infty} \left( \hat{\mathbf{G}}(s)^{j-kT} \otimes \mathbf{G}_{j-k} \right)}_{\Phi(s) = \sum_{j=0}^{\infty} \left( \hat{\mathbf{G}}(s)^{jT} \otimes \mathbf{G}_j \right)}. \end{aligned}$$

Putting this altogether, one obtains the equation for  $\Theta(\mathbf{s})$  in the form

$$0 = \left(\widehat{\mathbf{G}}(s)^T \otimes \pi_{0,0}(\mathcal{L}'' - \mathcal{L}')\right) - (I \otimes \pi_{0,0}\mathcal{S}) + \Theta(\mathbf{s}) \left[ \left(\widehat{\mathbf{G}}(s)^T \otimes \mathcal{L}'\right) + (I \otimes \mathcal{S}) + \left(\widehat{\mathbf{G}}(s)^T \otimes \mathcal{A}\right) \Phi(\mathbf{s}) \right], \tag{28}$$

where the only unknown is  $\Phi(\mathbf{s})$ . The computation of  $\Phi(\mathbf{s})$  follows the same pattern as that of  $\Theta(\mathbf{s})$ . Revisit (11–13) and Kronecker multiply the  $j$ -th equation with  $\widehat{\mathbf{G}}(s)^{jT}$  from the left and sum up over all values of  $j$ . This leads to the equation

$$\begin{aligned} \mathbf{0} &= (I \otimes \mathcal{S}) + \left(\widehat{\mathbf{G}}(s)^T \otimes \mathcal{H}\right) + (I \otimes \mathcal{L}) \sum_{j=0}^{\infty} \left(\widehat{\mathbf{G}}(s)^{jT} \otimes \mathbf{G}_j\right) \\ &\quad + (I \otimes \mathcal{A}) \sum_{j=0}^{\infty} \left(\widehat{\mathbf{G}}(s)^{jT} \otimes \sum_{i=0}^j \mathbf{G}_i \mathbf{G}_{j-i}\right) \\ &= (I \otimes \mathcal{S}) + \left(\widehat{\mathbf{G}}(s)^T \otimes \mathcal{H}\right) + (I \otimes \mathcal{L}) \Phi(\mathbf{s}) + (I \otimes \mathcal{A}) \Phi(\mathbf{s})^2, \end{aligned}$$

which is a quadratic matrix equation for  $\Phi(\mathbf{s})$ .

### 5 Computation of $\omega_{\mathbf{L}}(s)$

In the following we split the expression (26) for  $\omega_{\mathbf{L}}(s)$  into the following two terms:

$$\omega_{\mathbf{L}}(s) = \omega_{\mathbf{L}}^{k=0}(s) + \omega_{\mathbf{L}}^{k>0}(s),$$

where  $\omega_{\mathbf{L}}^{k=0}(s)$  includes only terms of (26) with  $k = 0$  and  $\omega_{\mathbf{L}}^{k>0}(s)$  all other terms, and further obtain expressions for each of them individually.

#### 5.1 Analysis of $\omega_{\mathbf{L}}^{k=0}(s)$

In order to compute  $\omega_{\mathbf{L}}^{k=0}(s)$  we perform Kronecker expansion, applying the relation  $vec(ABC) = (C^T \otimes A)vec(B)$  two times. We have

$$\begin{aligned} \omega_{\mathbf{L}}^{k=0}(s) &= \frac{1}{\lambda} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \pi_{i,j} \mathcal{A} \underbrace{\sum_{\ell=0}^i \mathbb{F}(s, i, \ell, 0) \mathcal{H} \widehat{\mathbf{G}}(s)^\ell \widehat{\mathbf{G}}(s)^{j+1} \mathbf{1}}_{\widehat{\mathcal{F}}_{k=0}(s,i)} \\ &= \frac{1}{\lambda} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \pi_{i,j} \mathcal{A} \widehat{\mathcal{F}}_{k=0}(s, i) \widehat{\mathbf{G}}(s)^{j+1} \mathbf{1} \\ &= \frac{1}{\lambda} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left( \mathbf{1}^T \widehat{\mathbf{G}}(s)^{j+1T} \otimes \pi_{i,j} \mathcal{A} \right) vec(\widehat{\mathcal{F}}_{k=0}(s, i)) \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{\lambda} \left( \mathbf{1}^T \widehat{\mathbf{G}}(s)^T \otimes \mathbf{1} \right) \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i,j} \right) \left( I \otimes \mathcal{A} \right) \text{vec}(\widehat{\mathcal{F}}_{k=0}(s, i)) \\
 &= \frac{1}{\lambda} \left( \mathbf{1}^T \widehat{\mathbf{G}}(s)^T \otimes \mathbf{1} \right) \underbrace{\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left[ \text{vec}(\widehat{\mathcal{F}}_{k=0}(s, i))^T \otimes \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i,j} \right) \right]}_{\mathbf{M}(s)} \text{vec} \left( I \otimes \mathcal{A} \right).
 \end{aligned}$$

Now we focus on the analysis of  $\mathbf{M}(s)$ . Just as in the case of  $\mathbf{K}(s)$  in Sect. 4.1, we will show that the unknown matrix  $\mathbf{M}(s)$  can be expressed in the form  $\mathbf{M}(s) = M_1(s) + \mathbf{M}(s)M_2(s)$ . Extracting the term with  $i = 0$  from the sum one can write

$$\begin{aligned}
 \mathbf{M}(s) &= \left[ \underbrace{\text{vec}(\widehat{\mathcal{F}}_{k=0}(s, 0))^T}_{\mathcal{L}(s)\mathcal{H}} \otimes \underbrace{\sum_{j=0}^{\infty} \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{0,j} \right)}_{\Theta(s)} \right] \\
 &\quad + \sum_{i=1}^{\infty} \left[ \text{vec}(\widehat{\mathcal{F}}_{k=0}(s, i))^T \otimes \sum_{j=0}^{\infty} \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i,j} \right) \right]. \tag{29}
 \end{aligned}$$

Here the only two unknown quantities are inside the sum over  $i$ . Firstly we obtain an expression for  $\text{vec}(\widehat{\mathcal{F}}_{k=0}(s, i))^T$ . Revisiting the definition of  $\widehat{\mathcal{F}}_{k=0}(s, i)$  and applying (25) when  $i > 0$ , we obtain

$$\begin{aligned}
 \widehat{\mathcal{F}}_{k=0}(s, i) &= \sum_{\ell=0}^i \widetilde{\mathbb{F}}(s, i, \ell, 0) \mathcal{H} \widehat{\mathbf{G}}(s)^\ell \\
 &= \sum_{\ell=1}^{i-1} \widetilde{\mathbb{F}}(s, i, \ell, 0) \mathcal{H} \widehat{\mathbf{G}}(s)^\ell + \widetilde{\mathbb{F}}(s, i, 0, 0) \mathcal{H} + \widetilde{\mathbb{F}}(s, i, i, 0) \mathcal{H} \widehat{\mathbf{G}}(s)^i \\
 &= \sum_{\ell=1}^{i-1} \mathcal{L}(s) \mathcal{S} \widetilde{\mathbb{F}}(s, i-1, \ell, 0) \mathcal{H} \widehat{\mathbf{G}}(s)^\ell \\
 &\quad + \sum_{\ell=1}^{i-1} \mathcal{L}(s) \mathcal{H} \widetilde{\mathbb{F}}(s, i-1, \ell-1, 0) \mathcal{H} \widehat{\mathbf{G}}(s)^\ell \\
 &\quad + \mathcal{L}(s) \mathcal{S} \widetilde{\mathbb{F}}(s, i-1, 0, 0) \mathcal{H} + \mathcal{L}(s) \mathcal{H} \widetilde{\mathbb{F}}(s, i-1, i-1, 0) \mathcal{H} \widehat{\mathbf{G}}(s)^i \\
 &= \mathcal{L}(s) \mathcal{S} \sum_{\ell=0}^{i-1} \widetilde{\mathbb{F}}(s, i-1, \ell, 0) \mathcal{H} \widehat{\mathbf{G}}(s)^\ell \\
 &\quad + \mathcal{L}(s) \mathcal{H} \sum_{\ell=0}^{i-1} \widetilde{\mathbb{F}}(s, i-1, \ell, 0) \mathcal{H} \widehat{\mathbf{G}}(s)^\ell \widehat{\mathbf{G}}(s),
 \end{aligned}$$

or, equivalently, in terms of  $\widehat{\mathcal{F}}_{k=0}(s, i)$ :

$$\widehat{\mathcal{F}}_{k=0}(s, i) = \mathcal{L}(s) \mathcal{S} \widehat{\mathcal{F}}_{k=0}(s, i-1) + \mathcal{L}(s) \mathcal{H} \widehat{\mathcal{F}}_{k=0}(s, i-1) \widehat{\mathbf{G}}(s), \quad i \geq 1. \tag{30}$$



By applying the *vec* operator to (30), one finds the following expression for  $vec(\widehat{\mathcal{F}}_{k=0}(s, i))^T, i \geq 1$ :

$$vec(\widehat{\mathcal{F}}_{k=0}(s, i))^T = vec(\widehat{\mathcal{F}}_{k=0}(s, i - 1))^T \left[ \left( I \otimes \mathcal{L}(s)\mathcal{S} \right) + \left( \widehat{\mathbf{G}}(s)^T \otimes \mathcal{L}(s)\mathcal{H} \right) \right]^T. \tag{31}$$

Now we obtain an expression for the second unknown quantity in (29) which is in the sum over *i* on the right-hand side of the first Kronecker product. With respect to (5) it can be rewritten in the form

$$\begin{aligned} \sum_{j=0}^{\infty} \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i,j} \right) &= \sum_{j=0}^{\infty} \sum_{m=0}^j \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i-1,m} \mathbf{R}_{j-m} \right) \\ &= \sum_{j=0}^{\infty} \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i-1,j} \right) \underbrace{\sum_{n=0}^{\infty} \left( \widehat{\mathbf{G}}(s)^n \otimes \mathbf{R}_n \right)}_{\Psi(s)} \\ &= \sum_{j=0}^{\infty} \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i-1,j} \right) \Psi(s), \end{aligned} \tag{32}$$

where we redefined the running indexes in the second step. Substitution of (31) and (32) into (29) yields the sought-for expression for  $\mathbf{M}(s)$ :

$$\begin{aligned} \mathbf{M}(s) &= \left[ vec(\mathcal{L}(s)\mathcal{H})^T \otimes \Theta(s) \right] \\ &\quad + \sum_{i=1}^{\infty} \left[ vec(\widehat{\mathcal{F}}_{k=0}(s, i - 1))^T \otimes \sum_{j=0}^{\infty} \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i-1,j} \right) \right] \\ &\quad \times \left\{ \left[ \left( I \otimes \mathcal{L}(s)\mathcal{S} \right) + \left( \widehat{\mathbf{G}}(s) \otimes \mathcal{L}(s)\mathcal{H} \right) \right]^T \otimes \Psi(s) \right\} \\ &= \left[ vec(\mathcal{L}(s)\mathcal{H})^T \otimes \Theta(s) \right] \\ &\quad + \mathbf{M}(s) \left\{ \left[ \left( I \otimes \mathcal{L}(s)\mathcal{S} \right) + \left( \widehat{\mathbf{G}}(s)^T \otimes \mathcal{L}(s)\mathcal{H} \right) \right]^T \otimes \Psi(s) \right\}. \end{aligned}$$

The expression for  $\Psi(s)$  can be obtained from (6) to (8) completely in the same way as is it done for  $\Phi(s)$ , and thus is omitted.

### 5.2 Analysis of $\omega_L^{k>0}(s)$

In this subsection we tackle the most complex case with the highest number of infinite summations and non-commuting matrices. For  $\omega_L^{k>0}(s)$  one has to apply Kronecker expansion multiple times. Firstly we recall that the definition of  $\omega_L^{k>0}(s)$  is

$$\omega_L^{k>0}(s) = \frac{1}{\lambda} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \pi_{i,j} \mathcal{A} \underbrace{\sum_{\ell=0}^i \sum_{k=1}^{\infty} \tilde{\mathbb{F}}(s, i, \ell, k) \mathcal{H} \tilde{\mathbf{G}}(s)^k \hat{\mathbf{G}}(s)^\ell \hat{\mathbf{G}}(s)^{j+1}}_{\mathcal{F}(s,i)} \mathbf{1}$$

and now consider term  $\mathcal{F}(s, i)$ . Applying the *vec* operator to  $\mathcal{F}(s, i)$  according to the following Kronecker expansion

$$\begin{aligned} \text{vec}(ABCD) &= (D^T \otimes A)\text{vec}(BC) = (\text{vec}(BC)^T \otimes (D^T \otimes A))\text{vec}(I) \\ &= (\text{vec}(I)^T \otimes I \otimes I)(C \otimes B^T \otimes D^T \otimes A)\text{vec}(I), \end{aligned}$$

one gets

$$\begin{aligned} &\text{vec}(\mathcal{F}(s, i)) \\ &= (\text{vec}(I)^T \otimes I \otimes I) \underbrace{\sum_{\ell=0}^i \sum_{k=1}^{\infty} \left( \tilde{\mathbf{G}}(s)^k \otimes \mathcal{H}^T \otimes \hat{\mathbf{G}}(s)^{\ell T} \otimes \tilde{\mathbb{F}}(s, i, \ell, k) \right)}_{\mathcal{F}^\otimes(s,i)} \text{vec}(I) \\ &= (\text{vec}(I)^T \otimes I \otimes I) \mathcal{F}^\otimes(s, i) \text{vec}(I). \end{aligned}$$

Considering the expression for  $\mathcal{F}(s, i)$  and using (25), when  $i > 0$  and  $k > 0$ , we get

$$\begin{aligned} \mathcal{F}(s, i) &= \sum_{\ell=0}^i \sum_{k=1}^{\infty} \tilde{\mathbb{F}}(s, i, \ell, k) \mathcal{H} \tilde{\mathbf{G}}(s)^k \hat{\mathbf{G}}(s)^\ell \\ &= \sum_{\ell=0}^{i-1} \sum_{k=1}^{\infty} \mathcal{L}(s) \mathcal{S} \tilde{\mathbb{F}}(s, i-1, \ell, k) \mathcal{H} \tilde{\mathbf{G}}(s)^k \hat{\mathbf{G}}(s)^\ell \\ &\quad + \sum_{\ell=0}^{i-1} \sum_{k=1}^{\infty} \mathcal{L}(s) \mathcal{H} \tilde{\mathbb{F}}(s, i-1, \ell, k) \mathcal{H} \tilde{\mathbf{G}}(s)^k \hat{\mathbf{G}}(s)^{\ell+1} \\ &\quad + \sum_{\ell=0}^i \sum_{k=0}^{\infty} \mathcal{L}(s) \mathcal{A} \tilde{\mathbb{F}}(s, i, \ell, k) \mathcal{H} \tilde{\mathbf{G}}(s)^{k+1} \hat{\mathbf{G}}(s)^\ell. \end{aligned} \tag{33}$$

Having such an expression for  $\mathcal{F}(s, i)$  we can now write out a relation for the term  $\mathcal{F}^\otimes(s, i)$  in the form

$$\begin{aligned} \mathcal{F}^\otimes(s, i) &= \underbrace{\left[ \left( I \otimes I \otimes I \otimes \mathcal{L}(s) \mathcal{S} \right) + \left( I \otimes I \otimes \hat{\mathbf{G}}(s)^T \otimes \mathcal{L}(s) \mathcal{H} \right) \right]}_{\mathbf{L}(s)} \mathcal{F}^\otimes(s, i-1) \\ &\quad + \underbrace{\left( \tilde{\mathbf{G}}(s) \otimes I \otimes I \otimes \mathcal{L}(s) \mathcal{A} \right)}_{\mathbf{K}(s)} \left( \mathcal{F}^\otimes(s, i) + \hat{\mathcal{F}}_{k=0}^\otimes(s, i) \right) \\ &= [I - \mathbf{K}(s)]^{-1} [\mathbf{L}(s) \mathcal{F}^\otimes(s, i-1) + \mathbf{K}(s) \hat{\mathcal{F}}_{k=0}^\otimes(s, i)], \end{aligned} \tag{34}$$

where we introduced the notation

$$\widehat{\mathcal{F}}_{k=0}^{\otimes}(s, i) = \sum_{\ell=0}^i \left( I \otimes \mathcal{H}^T \otimes \widehat{\mathbf{G}}(s)^{\ell T} \otimes \widetilde{\mathbb{F}}(s, i, \ell, 0) \right), \quad i \geq 0.$$

From (25) it follows that

$$\begin{aligned} \mathcal{F}^{\otimes}(s, 0) &= \sum_{k=1}^{\infty} \left( \widetilde{\mathbf{G}}(s)^k \otimes \mathcal{H}^T \otimes I \otimes (\mathcal{L}(s)\mathcal{A})^k \mathcal{L}(s) \right) \\ &= \left[ I - \left( \widetilde{\mathbf{G}}(s) \otimes I \otimes I \otimes \mathcal{L}(s)\mathcal{A} \right) \right]^{-1} \left( \widetilde{\mathbf{G}}(s) \otimes \mathcal{H}^T \otimes I \otimes \mathcal{L}(s)\mathcal{A}\mathcal{L}(s) \right) \end{aligned}$$

and  $\widehat{\mathcal{F}}_{k=0}^{\otimes}(s, 0) = I \otimes \mathcal{H}^T \otimes I \otimes \mathcal{L}(s)$ . For  $i \geq 1$ , from (30) we have

$$\widehat{\mathcal{F}}_{k=0}^{\otimes}(s, i) = \mathbf{L}(s) \mathcal{F}_{k=0}^{\otimes}(s, i - 1), \quad i \geq 1.$$

Now we go back to  $\omega_{\mathbf{L}}^{k>0}(s)$  and apply the *vec* operator multiple times in the following way:

$$\begin{aligned} \omega_{\mathbf{L}}^{k>0}(s) &= \frac{1}{\lambda} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \pi_{i,j} \mathcal{A} \mathcal{F}(s, i) \widehat{\mathbf{G}}(s)^{j+1} \mathbf{1} \\ &= \frac{1}{\lambda} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left( \mathbf{1}^T \widehat{\mathbf{G}}(s)^{j+1 T} \otimes \pi_{i,j} \mathcal{A} \right) \text{vec} \left( \mathcal{F}(s, i) \right) \\ &= \frac{1}{\lambda} \left( \mathbf{1}^T \widehat{\mathbf{G}}(s)^T \otimes \mathbf{1} \right) \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left( \widehat{\mathbf{G}}(s)^{j T} \otimes \pi_{i,j} \right) \left( I \otimes \mathcal{A} \right) \text{vec} \left( \mathcal{F}(s, i) \right) \\ &= \frac{1}{\lambda} \left( \mathbf{1}^T \widehat{\mathbf{G}}(s)^T \otimes \mathbf{1} \right) \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left[ \text{vec} \left( \mathcal{F}(s, i) \right)^T \otimes \left( \widehat{\mathbf{G}}(s)^{j T} \otimes \pi_{i,j} \right) \right] \text{vec} \left( I \otimes \mathcal{A} \right) \\ &= \frac{1}{\lambda} \left( \mathbf{1}^T \widehat{\mathbf{G}}(s)^T \otimes \mathbf{1} \right) \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left[ \text{vec}(I)^T \mathcal{F}^{\otimes}(s, i)^T \left( \text{vec}(I) \otimes I \otimes I \right) \right. \\ &\quad \left. \otimes \left( \widehat{\mathbf{G}}(s)^{j T} \otimes \pi_{i,j} \right) \right] \text{vec} \left( I \otimes \mathcal{A} \right) \\ &= \frac{1}{\lambda} \left( \mathbf{1}^T \widehat{\mathbf{G}}(s)^T \otimes \mathbf{1} \right) \underbrace{\left[ \text{vec}(I)^T \otimes I \right] \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left[ \mathcal{F}^{\otimes}(s, i)^T \otimes \left( \widehat{\mathbf{G}}(s)^{j T} \otimes \pi_{i,j} \right) \right]}_{\mathbf{N}(s)} \\ &\quad \times \left[ \text{vec}(I) \otimes I \otimes I \otimes I \right] \text{vec} \left( I \otimes \mathcal{A} \right). \end{aligned}$$

The only unknown quantity in the expression for  $\omega_{\mathbf{L}}^{k>0}(s)$  is  $\mathbf{N}(s)$ . Its form can be found from (32) and (34). We have

$$\begin{aligned} \mathbf{N}(s) &= \left[ \mathcal{F}^{\otimes}(s, 0)^T \otimes \underbrace{\sum_{j=0}^{\infty} \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{0,j} \right)}_{\boldsymbol{\Theta}(s)} \right] \\ &\quad + \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \left[ \mathcal{F}^{\otimes}(s, i)^T \otimes \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i,j} \right) \right] \\ &= \left[ \mathcal{F}^{\otimes}(s, 0)^T \otimes \boldsymbol{\Theta}(s) \right] \\ &\quad + \underbrace{\sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \left[ \widehat{\mathcal{F}}_{k=0}^{\otimes}(s, i)^T \otimes \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i,j} \right) \right]}_{\mathbf{Z}(s)} \left( \mathbf{K}(s)^T [I - \mathbf{K}(s)]^{-1T} \otimes I \right) \\ &\quad + \underbrace{\sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \left[ \mathcal{F}^{\otimes}(s, i-1)^T \mathbf{L}(s)^T [I - \mathbf{K}(s)]^{-1T} \otimes \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i-1,j} \right) \boldsymbol{\Psi}(s) \right]}_{\mathbf{N}(s) \left( \mathbf{L}(s)^T [I - \mathbf{K}(s)]^{-1T} \otimes \boldsymbol{\Psi}(s) \right)}, \end{aligned}$$

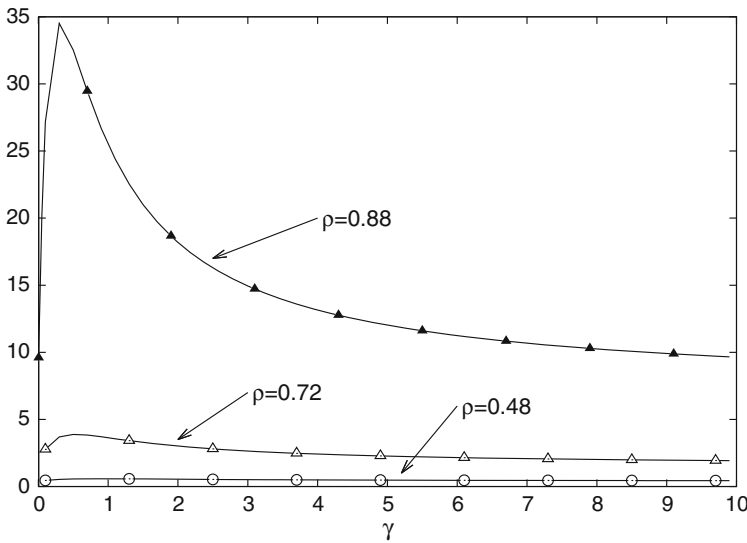
and for  $\mathbf{Z}(s)$ , using properties of Kronecker product, one obtains the relation

$$\begin{aligned} \mathbf{Z}(s) &= \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \left[ \widehat{\mathcal{F}}_{k=0}^{\otimes}(s, i-1)^T \mathbf{L}(s)^T \otimes \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i-1,j} \right) \boldsymbol{\Psi}(s) \right] \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \left[ \widehat{\mathcal{F}}_{k=0}^{\otimes}(s, i)^T \otimes \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{i,j} \right) \right] \left( \mathbf{L}(s)^T \otimes \boldsymbol{\Psi}(s) \right) \\ &= \left[ \left( \widehat{\mathcal{F}}_{k=0}^{\otimes}(s, 0)^T \otimes \sum_{j=0}^{\infty} \left( \widehat{\mathbf{G}}(s)^{jT} \otimes \pi_{0,j} \right) \right) + \mathbf{Z}(s) \right] \left( \mathbf{L}(s)^T \otimes \boldsymbol{\Psi}(s) \right) \\ &= \left[ \left( \left( I \otimes \mathcal{H}^T \otimes I \otimes \mathcal{L}(s) \right)^T \otimes \boldsymbol{\Theta}(s) \right) + \mathbf{Z}(s) \right] \left( \mathbf{L}(s)^T \otimes \boldsymbol{\Psi}(s) \right). \end{aligned}$$

The latter relation allows computation of  $\mathbf{Z}(s)$  and subsequently  $\mathbf{N}(s)$  and  $\omega_{\mathbf{L}}^{k>0}(s)$ .

### 6 Numerical example

In this section we present a simple numerical example, where only the service time depends on the Markov environment. Regular customers and re-sequencing signals arrive according to Poisson processes with rates  $\lambda$  and  $\gamma$ , respectively. The service process has a phase-type distribution with the following representation  $(\boldsymbol{\beta}, \mathbf{B})$ :



**Fig. 4** Behaviour of variance of customer’s waiting time as a function of re-sequencing intensity  $\gamma$  for different values of load

$$\beta = (0.5 \ 0.5), \quad \mathbf{B} = \begin{pmatrix} -4 & 2 \\ 1 & -4 \end{pmatrix}.$$

The service rate is  $\mu = -1/\beta\mathbf{B}^{-1}\mathbf{1} = 2.5$ . Using the notation introduced in Sect. 2, one has

$$\mathbf{A}_0 = (-\lambda), \quad \mathbf{A}_1 = (\lambda), \quad \mathbf{H}_0 = (-\gamma), \quad \mathbf{H}_1 = (\gamma), \quad \mathbf{S}_0 = \mathbf{B}, \quad \mathbf{S}_1 = -\mathbf{B}\mathbf{1}\beta.$$

Substituting this into the expression for  $\omega(s)$ , one can compute moments of the waiting time by differentiation. As the mean waiting time of an arbitrary customer can be easily computed from Little’s law, one is mainly interested in higher moments, for example variance. Even for the simple case considered the expression for the variance is very cumbersome and thus we just state numerical results without providing the expression itself. Figure 4 plots variance of the waiting time as a function of the re-sequencing rate,  $\gamma$ , for three values of load  $\rho = \lambda/\mu = 0.48, \rho = 0.72$  and  $\rho = 0.88$ .

It is worth noticing that in each case considered there exists a point with maximal variance inside the plotted range, which is in line with intuitive expectations. If the re-sequencing rate,  $\gamma$ , is low then a very small number of customers are re-sequenced and the variance of the waiting time is close to the variance of the waiting time in the ordinary  $M/PH/1$  queue (with the same arrival and service process). If the re-sequencing rate is large, almost every arriving customer is re-sequenced and thus again the variance of the waiting time almost coincides with the variance in the  $M/PH/1$  queue. In between these extreme cases some customers get re-sequenced, some do not and thus the variance of the waiting time increases.

## 7 Conclusion

The paper considers a queueing model with a re-sequencing buffer. The delay analysis of this model with memoryless arrival and service processes is provided in [12] with the use of generating functions. We extended the delay analysis to Markov-modulated arrival and service processes. This extension required the introduction of a completely new methodology due to the presence of non-commuting matrices. The main elements of the proposed new methodology are the replacement of the generating functions by utilizing the space inhomogeneity of the model (based on the relation of infinite summations from 0 to infinity with infinite summations from 1 to infinity), and the use of the Kronecker expansion. The price to pay for the use of the Kronecker expansion is the multiplicative increase of the matrix dimensions, which might result in inhibitive computational complexity for “large” models.

**Acknowledgments** R. Razumchik thanks the support of Russian Foundation for Basic Research (Grant 13-07-00223). M. Telek thanks the support of OTKA Grant K101150.

## References

1. Bini, D.A., Latouche, G., Meini, B.: Numerical Methods for Structured Markov Chains. Oxford University Press, New York, NY (2005)
2. Graham, A.: Kronecker Products and Matrix Calculus: With Applications. Wiley, New York, NY (1982)
3. He, Q.-M.: Fundamentals of Matrix-Analytic Methods. Springer, New York (2013)
4. Higham, N.J., Kim, H.-M.: Numerical analysis of a quadratic matrix equation. *IMA J. Numer. Anal.* **20**(4), 499–519 (2000)
5. Horváth, G.: Efficient analysis of the queue length moments of the MMAP/MAP/1 preemptive priority queue. *Perform. Eval.* **69**(12), 684–700 (2012)
6. Horváth, G., Van Houdt, B.: Departure process analysis of the multi-type MMAP[K]/PH[K]/1 FCFS queue. *Perform. Eval.* **70**(6), 423–439 (2013)
7. Horváth, G., Van Houdt, B., Telek, M.: Commuting matrices in the queue length and sojourn time analysis of MAP/MAP/1 queues. *Stoch. Models* **30**(4), 554–575 (2014)
8. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. Society for Industrial and Applied Mathematics, Philadelphia (1999)
9. Neuts, M.F.: Matrix Geometric Solutions in Stochastic Models. Johns Hopkins University Press, Baltimore (1981)
10. Neuts, M.F.: Structured Stochastic Matrices of M/G/1 Type and Their Applications. Marcel Dekker, New York (1989)
11. Ozawa, T.: Sojourn time distributions in the queue defined by a general QBD process. *Queueing Syst. Theory Appl.* **53**(4), 203–211 (2006)
12. Pechinkin, A.V., Razumchik, R.V.: On temporal characteristics in an exponential queueing system with negative claims and a bunker for ousted claims. *Autom. Remote Control* **72**(12), 2492–2504 (2011)
13. Steeb, W.H., Hardy, Y.: Matrix Calculus and Kronecker Product: A Practical Approach to Linear and Multilinear Algebra. World Scientific, River Edge, NJ (2011)