Analysis of Two-Server Queueing Model with Phase-Type Service Time Distribution and Common Phases of Service

Chesoong Kim[1](✉), Alexander Dudin[2], Sergey Dudin[2], and Olga Dudina[2]

[1] Sangji University, Wonju, Kangwon 220-702, Korea
dowoo@sangji.ac.kr
[2] Belarusian State University, 4, Nezavisimosti Ave., 220030 Minsk, Belarus
dudin@bsu.by, dudin85@mail.ru, dudina_olga@email.com

Abstract. We consider two server queueing system with an infinite buffer. Customers arrive to the system according to the Markovian Arrival Process. Service time of a customer has a phase-type distribution. The servers use the same equipment (phases of PH) for customers processing. So, if service of a customer transits to the phase, at which another server is currently providing the service, the service of the customer is suspended until the phase will become available. Behavior of the system is described by the multi-dimensional Markov chain. The generator of this Markov chain is derived. Expressions for computation of the main performance measures are derived.

Keywords: Markovian arrival process · Phase-type service time distribution · Interacting service processes

1 Introduction

The simplest queueing models suggest that customer's service time is exponentially distributed. This allows to avoid the necessity of taking into account the elapsed service time when the Markovian process describing behavior of the queue should be constructed. It is quite obvious that in many real world applications of queueing theory assumption about the exponential distribution does not hold true and more general distributions of service time have to be considered. If service time has an arbitrary distribution, it is mandatory to take into account the elapsed or residual service times, e.g., by introducing supplementary variables or considering the embedded Markov chains. This may lead to huge analytical difficulties in analysis of the Markovian process describing dynamics of the queue, especially when multi-server queue is under study. To avoid such difficulties, so called phase type (PH) distribution, as natural extension of previously well-known Erlangian and hyper-exponential distribution, was offered, see, e.g., [1]. Good property of this distribution is its generality. Generally speaking, any distribution can be approximated, in sense of a weak convergence, by the

V. Vishnevsky and D. Kozyrev (Eds.): DCCN 2015, CCIS 601, pp. 19–29, 2016.
DOI: 10.1007/978-3-319-30843-2_3

PH type distribution, see, e.g., [3]. Random time having the PH type distribution can be interpreted as the sequence of phases duration of each of which has exponential distribution. The total number of existing phases of service is finite. However, implementation of the phases during the service of a customer may be repeated random number of times. So, in concrete realization of the random time having the PH type distribution the number of phases in a sequence is random.

Formal definition of PH type distribution is given in the next section. For purposes of this paper, we slightly rephrase this definition as follows. There is a virtual network with nodes (phases), say, $\{1,\ldots,M\}$. Random time having the PH type distribution with an irreducible representation $(\boldsymbol{\beta}, S)$ is the time during which some virtual customer stays in this network conditional on the fact that this virtual customer starts its staying in the network from the visit to the state m of the network with probability $\boldsymbol{\beta}_m,\ m=\overline{1,M}$, then it makes transitions inside of the set $\{1,\ldots,M\}$ with intensities given by the entries of the matrix S or leaves the network from any state m with intensity, which is the mth component of column vector $\mathbf{S}_0=-S\mathbf{e}$, where $\mathbf{e}$ denotes unit column vector. In brief, as it was already noted above, random time having the PH type distribution consists of the random number of *virtual* phases, duration of which is exponentially distributed. This allows to replace keeping track of the **continuous** elapsed or residual service time by the keeping track of the **discrete** current phase of the service what greatly simplifies the analysis.

So, phases of service in definition of PH type distribution may be just the virtual entities. However, in many real world situations, random time having PH type distribution may represent the sequence of *real* phases. E.g., processing time of the query in data base consists of implementation of a sequence of input/output operations alternating with the use of CPU. Processing of a car in service station consists of a sequence of technological operations. Security control in airport includes screening of a luggage and passengers with possible additional personal individual passenger inspection. If the service is provided by the single server, no collisions occur. But, if there are several servers operating in parallel, collisions may occur because the servers use some common resources, e.g. memory and CPU, tables and indices of data base, different equipment of a car service station, screening devices and cabins for the personal individual inspection, etc.

Traditionally in analysis of multi-server queues with PH type service time distribution, it is assumed that service processes in the servers are independent. To the best of our knowledge, the systems with interference of phases of service in different servers are not considered in literature. In this paper, we start research of such systems from a relatively simple model where there are only two servers, state spaces of the virtual networks, in terms of which PH type service time distribution is interpreted, coincide and if some phase of the service by a server is required while this phase is busy by another server this phase of service is postponed until it will be released by another server. In some sense, in this model we somehow unite the problems considered in two different popular branches of operations research, queueing theory and scheduling theory, which consider the similar objects (service systems) but with emphasis to different aspects of the problem. Scheduling theory addresses the aspect how the required phases of service should be ordered for different jobs to provide the shortest common processing time and completely ignores the stochastic aspects (arrival of jobs at random moments and random duration of implementation of a service at each phase). Queueing theory accounts these stochastic aspects, but is not focused on proper ordering the phases of the service.

The mathematical model is described in detail in Sect. 2. Behavior of the system is described by the multidimensional continuous time Markov chain in Sect. 3. The infinitesimal generator of this Markov chain as the block structured matrix is presented here. This Markov chain belongs to the class of continuous-time asymptotically quasi-Toeplitz Markov chains. Using this fact, it is shown that, due to impatience of customers, this Markov chain is ergodic for any set of the system parameters. The problem of computation of the stationary distribution of this Markov chain is touched. In Sect. 4, formulas for computation of the main performance measures of the system are presented. Section 5 concludes the paper.

2 Mathematical Model

Queueing system with two servers and a buffer of infinite capacity is considered. The structure of the system under study is presented in Fig. 1.

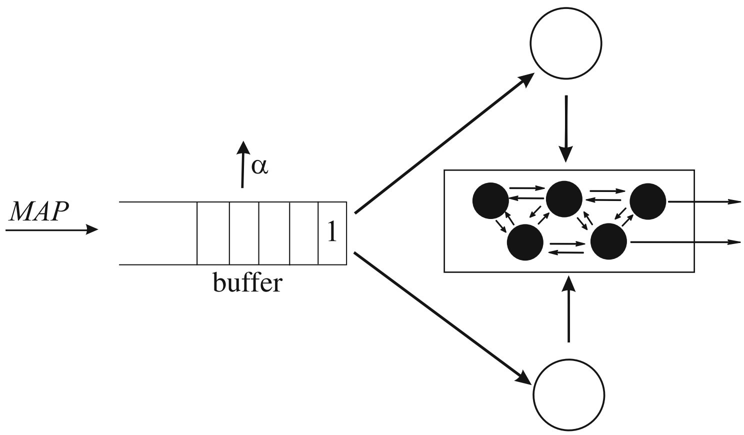

Fig. 1. Queueing system under study

Customers arrive at the system according to the Markovian arrival process (MAP). Arrivals in the MAP are directed by an irreducible continuous time Markov chain ν_t, $t \geq 0$, with the finite state space $\{0, 1, ..., W\}$. The sojourn time of the Markov chain ν_t, $t \geq 0$, in the state ν has an exponential distribution with the parameter λ_ν, $\nu = \overline{0, W}$. Here, notation such as $\nu = \overline{0, W}$ means that ν assumes values from the set $\{0, 1, ..., W\}$. After this sojourn time expires, with probability $p_k(\nu, \nu')$ the process ν_t transits to the state ν' and k customers, $k = 0, 1$, arrive at the system. The intensities of transitions from one state to another, that are accompanied by the arrival of k customers, are combined to the square matrices D_k, $k = 0, 1$, of size $W + 1$. The matrix generating function of these matrices is $D(z) = D_0 + D_1 z$, $|z| \leq 1$. The matrix $D(1)$ is an infinitesimal generator of the process ν_t, $t \geq 0$. The stationary distribution vector $\boldsymbol{\theta}$ of this process satisfies the system of equations $\boldsymbol{\theta} D(1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$. Here and throughout this paper, $\mathbf{0}$ is a zero row vector. In case if the dimension of a vector is not clear from the context, it is indicated as a lower index.

The average intensity λ (fundamental rate) of the MAP is defined by $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$.

The MAP arrival process was introduced as a versatile Markovian point process ($VMPP$) by M.F. Neuts in the 70th. The original development of the $VMPP$ contained extensive notations; however these notations were greatly simplified in [4] and ever since this process bears the name Markovian arrival process. The class of MAPs includes many input flows considered previously, such as stationary Poisson (M), Erlangian (E_k), hyper-Markovian (HM), phase-type (PH), Markov modulated Poisson process ($MMPP$). Generally speaking, the MAP is correlated, so it is ideal to model correlated and or bursty traffic in the modern telecommunication networks, see, e.g., [5–7].

The service time of a customer for the server has a PH (phase-type) distribution with an irreducible representation $(\boldsymbol{\beta}, S)$. This service time traditionally is interpreted as the time until the underlying Markov process $m_t, t \geq 0$, with a finite state space $\{1, \ldots, M, M+1\}$ reaches the single absorbing state $M + 1$ condition on the fact that the initial state of this process is selected among the states $\{1, \ldots, M\}$ according to the probabilistic row vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$. The transition rates of the process m_t within the set $\{1, \ldots, M\}$ are defined by the sub-generator S, and the transition rates into the absorbing state (what leads to service completion) are given by the entries of the column vector $\mathbf{S}_0 = -S\mathbf{e}$. The mean service time is calculated as $b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$.

But here we assume that the service processes of customers by two servers are not independent. The state space of the underlying Markov processes for both the servers is the same. If there is no collision, these processes make transitions according to definition given above, independently of each other. However if the service of a customer by the l-th server, $l = 1, 2$ is at some phase, say, m while the underlying service process of other customer by l'th, $l' = 1, 2$, $l' \neq l$, server should transit to this phase m, the l'-th server is stopped and the service is blocked until the l-th server finishes the mth phase of the service. The servers of the system are assumed to be identical and are enumerated in arbitrary order. However, if two servers need the same phase of the service, they are assumed being enumerated in order of occupation of the phase. Number 1 is appointed to the server that currently occupies the phase under the conflict. Number 2 is appointed to the server that currently waits for releasing this phase.

The customers from the buffer are impatient, i.e., the customer leaves the buffer and the system after an exponentially distributed waiting time described by the parameter α, $0 < \alpha < \infty$.

3 Process of System States

It is easy to see that the dynamics of the system under study is described by the following regular irreducible multi-dimensional continuous-time Markov chain

$$\xi_t = \{i_t, r_t, \nu_t, n_t, m_t\},\ t \geq 0,$$

where, during the epoch t, $t \geq 0$,

- i_t is the number of customers in the system, $i_t \geq 0$;
- r_t is an indicator that indicates whether some server is blocked or not: $r_t = 0$ corresponds to the case when a server isn't blocked and $r_t = 1$ otherwise;
- ν_t is the state of the underlying process of the MAP, $\nu_t = \overline{0, W}$;
- n_t is the state of PH service process at the first server, $n_t = \overline{1, M}$.
- m_t is the state of PH service process at the second server, $m_t = \overline{1, M}$, $m_t \neq n_t$.

The Markov chain ξ_t, $t \geq 0$, has the following state space:

$$\left(\{0,0,\nu\}\right)\bigcup\left(\{1,0,\nu,n\},\ n=\overline{1,M}\right)\bigcup$$

$$\left(\{i,0,\nu,n,m\},\ i\geq 2,\ n=\overline{1,M},\ m=\overline{1,M},\ m\neq n\right)\bigcup$$

$$\left(\{i,1,\nu,n\},\ i\geq 2,\ n=\overline{1,M}\right),\ \nu=\overline{0,W}.$$

For further use throughout this paper, we introduce the following notation:

- I is the identity matrix, and O is a zero matrix of appropriate dimension;
- $\otimes$ and $\oplus$ indicate the symbols of Kronecker product and sum of matrices (see [8]), respectively;
- $\bar{W} = W + 1$;
- I_{l_1,l_2}, $l_1, l_2 = \overline{1, M}$, $l_1 \neq l_2$, is the matrix of size $(M-1) \times (M-1)$ with all zero entries except the entries $(I_{l_1,l_2})_{k,k}$, $k = \overline{0, M-2}$, $k \neq l_2 - 2$, in the case $(l_1 < l_2)$ and $(I_{l_1,l_2})_{k,k}$, $k = \overline{0, M-2}$, $k \neq l_2 - 1$, in the case $(l_1 > l_2)$ which are equal to 1;
- $\tilde{S}_l$, $l = \overline{1, M}$, is the square matrix of size $M - 1$ that is obtained from matrix S by removing the $l-1$-th column and the l-th row;

- $\mathbf{e}_{l_1,l_2}$, $l_1, l_2 = \overline{1, M}$, $l_1 \neq l_2$, is the column vector of size $(M-1)$ with all zero entries except the entry $(\mathbf{e}_{l_1,l_2})_{l_2-1}$ in the case $(l_1 > l_2)$ and $(\mathbf{e}_{l_1,l_2})_{l_2-2}$, in the case $(l_1 < l_2)$ which are equal to 1;
- $\mathbf{c}_{l_1,l_2}$, $l_1, l_2 = \overline{1, M}$, $l_1 \neq l_2$, is the row vector of size $(M-1)$ with all zero entries except the entry $(\mathbf{c}_{l_1,l_2})_{l_1-2}$ in the case $(l_1 > l_2)$ and $(\mathbf{c}_{l_1,l_2})_{l_1-1}$, in the case $(l_1 < l_2)$ which are equal to 1;
- $\boldsymbol{\beta}_l$, $l = \overline{1, M}$, - the row vector that obtained from the vector $\boldsymbol{\beta}$ by deleting $l-1$-th component;
- $\mathbf{a}_l$, $l = \overline{1, M}$, is the column vector of size $M-1$ that is obtained from the $l-1$-th column of the matrix S by removing the $l-1$-th entry;
- I_l^+, $l = \overline{1, M}$, is the matrix of size $(M-1) \times M$ which obtained from the identity matrix of size $M-1$ by adding the zero column in position $l-1$;
- $\mathbf{S}_0^l$, $l = \overline{1, M}$ is a column vector of size $M-1$ which is obtained from the vector $\mathbf{S}_0$ by removing the $l-1$-th component.
- $\tilde{\mathbf{a}}_l$, $l = \overline{1, M}$, is a row vector of size M with all zero components except the component $(\tilde{\mathbf{a}}_l)_{l-1}$ which is equal to 1;
- B_l, $l = \overline{1, M}$, is the matrix of size $(M-1) \times M(M-1)$ which obtained from the matrix $\text{diag}\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M\}$ by deleting the $l-1$-th row;
- C_l, $l = \overline{1, M}$, is the matrix of size $(M-1) \times M$ which obtained from the matrix $\text{diag}\{\beta_1, \ldots, \beta_M\}$ by deleting the $l-1$-th row;

Let us enumerate the states of the Markov chain ξ_t, $t \geq 0$, in the direct lexicographic order of the components r, k, ν, ζ, η and refer to the set of the states of the chain having values (i, r) of the first two components of the Markov chain as a macro-state (i, r).

Let Q be the generator of the Markov chain ξ_t, $t \geq 0$, consisting of the blocks $Q_{i,j}$, which, in turn, consist of the matrices $(Q_{i,j})_{r,r'}$ of the transition rates of this chain from the macro-state (i, r) to the macro-state (j, r'), $r, r' = 0, 1$. The diagonal entries of the matrices $Q_{i,i}$ are negative, and the modulus of the diagonal entry of the blocks $(Q_{i,i})_{r,r}$ defines the total intensity of leaving the corresponding state of the Markov chain ξ_t, $t \geq 0$.

Analysing all transitions of the Markov chain $\xi_t, t \geq 0$, during an interval of an infinitesimal length and rewriting the intensities of these transitions in the block matrix form we obtain the following result.

Theorem 1. *The infinitesimal generator $Q = (Q_{i,j})_{i,j \geq 0}$ of the Markov chain ξ_t, $t \geq 0$, has a block-tridiagonal structure:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \ldots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & \ldots \\ O & Q_{2,1} & Q_{2,2} & Q^+ & \ldots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}.$$

The non-zero blocks $Q_{i,j}$, $i, j \geq 0$, have the following form:

$$Q_{0,0} = D_0,$$

$$Q_{1,1} = D_0 \oplus S,$$

$$Q_{i,i} = \begin{pmatrix} Q_{i,i}^{(0,0)} & Q_{i,i}^{(0,1)} \\ Q_{i,i}^{(1,0)} & Q_{i,i}^{(1,1)} \end{pmatrix},\ i > 1,$$

$$Q_{i,i}^{(0,0)} = D_0 \otimes I_{M(M-1)} + I_{\bar{W}} \otimes (\mathcal{S} + \mathrm{diag}\{\tilde{S}_1, \dots, \tilde{S}_M\}) - (i-2)\alpha I_{\bar{W}M(M-1)},\ i > 1,$$

$$\mathcal{S} = \begin{pmatrix} S_{1,1}I_{M-1} & S_{1,2}I_{1,2} & \dots & S_{1,M}I_{1,M} \\ S_{2,1}I_{2,1} & S_{2,2}I_{M-1} & \dots & S_{2,M}I_{2,M} \\ \vdots & \vdots & \dots & \vdots \\ S_{M,1}I_{M,1} & S_{M,2}I_{M,2} & \dots & S_{M,M}I_{M-1} \end{pmatrix},$$

$$Q_{i,i}^{(0,1)} = I_{\bar{W}} \otimes \left(\begin{pmatrix} \mathbf{0}^T & S_{1,2}\mathbf{e}_{1,2} & \dots & S_{1,M}\mathbf{e}_{1,M} \\ S_{2,1}\mathbf{e}_{2,1} & \mathbf{0}^T & \dots & S_{2,M}\mathbf{e}_{2,M} \\ \vdots & \vdots & \dots & \vdots \\ S_{M,1}\mathbf{e}_{M,1} & S_{M,2}\mathbf{e}_{M,2} & \dots & \mathbf{0}^T \end{pmatrix} + \mathrm{diag}\{\mathbf{a}_1, \dots, \mathbf{a}_M\} \right),\ i > 1,$$

$$Q_{i,i}^{(1,0)} = I_{\bar{W}} \otimes \begin{pmatrix} \mathbf{0} & S_{1,2}\mathbf{c}_{1,2} & \dots & S_{1,M}\mathbf{c}_{1,M} \\ S_{2,1}\mathbf{c}_{1,M} & \mathbf{0} & \dots & S_{2,M}\mathbf{c}_{1,M} \\ \vdots & \vdots & \dots & \vdots \\ S_{M,1}\mathbf{c}_{1,M} & S_{M,2}\mathbf{c}_{1,M} & \dots & \mathbf{0} \end{pmatrix},$$

$$Q_{i,i}^{(1,1)} = D_0 \oplus \mathrm{diag}\{S_{1,1}, \dots, S_{M,M}\} - (i-2)\alpha I_{\bar{W}M},\ i > 1,$$

$$Q_{0,1} = D_1 \otimes \boldsymbol{\beta},$$

$$Q_{1,2} = \begin{pmatrix} Q_{1,2}^{(0,0)} & Q_{1,2}^{(0,1)} \end{pmatrix},$$

$$Q_{1,2}^{(0,0)} = D_1 \otimes \mathrm{diag}\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_M\},$$

$$Q_{1,2}^{(0,1)} = D_1 \otimes \mathrm{diag}\{\beta_1, \dots, \beta_M\},$$

$$Q^+ = \begin{pmatrix} D_1 \otimes I_{M(M-1)} & O \\ O & D_1 \otimes I_M \end{pmatrix},$$

$$Q_{1,0} = I_{\bar{W}} \otimes S_0,\ Q_{2,1} = \begin{pmatrix} Q_{2,1}^{(0,0)} \\ Q_{2,1}^{(1,0)} \end{pmatrix},\ i > 1,$$

$$Q_{2,1}^{(0,0)} = I_{\bar{W}} \otimes \left(\begin{pmatrix} \mathbf{S_0}_1 I_1^+ \\ \vdots \\ (\mathbf{S_0})_M I_M^+ \end{pmatrix} + \mathrm{diag}\{(\mathbf{S}_0)^l,\ l = \overline{1, M}\} \right),$$

$$Q_{2,1}^{(1,0)} = I_{\bar{W}} \otimes \begin{pmatrix} (\mathbf{S_0})_1 \tilde{\mathbf{a}}_1 \\ \vdots \\ (\mathbf{S_0})_M \tilde{\mathbf{a}}_M \end{pmatrix},$$

$$Q_{i,i-1} = \begin{pmatrix} Q_{i,i-1}^{(0,0)} & Q_{i,i-1}^{(0,1)} \\ Q_{i,i-1}^{(1,0)} & Q_{i,i-1}^{(1,1)} \end{pmatrix},\ i > 2,$$

$$Q_{i,i-1}^{(0,0)} = I_{\bar{W}} \otimes \left(\begin{pmatrix} \mathbf{S_0}_1 B_1 \\ \vdots \\ (\mathbf{S_0})_M B_M \end{pmatrix} + \mathrm{diag}\{(\mathbf{S}_0)^l \boldsymbol{\beta}_l,\ l = \overline{1, M}\} \right) + (i-2)\alpha I_{\bar{W}M(M-1)},$$

$$Q_{i,i-1}^{(0,1)} = I_{\bar{W}} \otimes \left(\begin{pmatrix} (\mathbf{S_0})_1 C_1 \\ \vdots \\ (\mathbf{S_0})_M C_M \end{pmatrix} + \mathrm{diag}\{\mathbf{S}_0^l \beta_l,\ l = \overline{1, M}\} \right),$$

$$Q_{i,i-1}^{(1,0)} = I_{\bar{W}} \otimes \mathrm{diag}\{(\mathbf{S_0})_l \boldsymbol{\beta}_l,\ l = \overline{1, M}\},$$

$$Q_{i,i-1}^{(1,1)} = I_{\bar{W}} \otimes \mathrm{diag}\{(\mathbf{S_0})_l \beta_l,\ l = \overline{1, M}\} + (i-2)\alpha I_{\bar{W}M}.$$

Corollary 1. The Markov chain $\xi_t,\ t \geq 0$, belongs to the class of continuous-time asymptotically quasi-Toeplitz Markov chains ($AQTMC$), for definition and relevant information see paper [9].

Proof. It can be verified that the limits Y_0, Y_1 and Y_2

$$Y_0 = \lim_{i\to\infty} R_i^{-1} Q_{i,i-1},\ Y_1 = \lim_{i\to\infty} R_i^{-1} Q_{i,i} + I,\ Y_2 = \lim_{i\to\infty} R_i^{-1} Q_{i,i+1}$$

exist and the matrix $Y_0 + Y_1 + Y_2$ is stochastic where the matrix R_i is a diagonal matrix with the diagonal entries which are defined as the moduli of the corresponding diagonal entries of the matrix $Q_{i,i},\ i \geq 0$. According to the definition given in [9] this means that the Markov chain $\xi_t,\ t \geq 0$, belongs to the class of $AQTMC$.

Let us analyze the properties of this Markov chain. This analysis includes derivation of conditions which should be imposed on the system parameters to guarantee existence of the stationary distribution of the states of the chain (ergodicity condition) and a procedure for computation of the stationary probabilities of the states.

As follows from [9], a sufficient condition for the existence of a stationary distribution of $AQTMC$ $\xi_t,\ t \geq 0$, is expressed in terms of the matrices Y_0, Y_1 and Y_2 defined above. This sufficient condition of ergodicity of Markov chain $\xi_t,\ t \geq 0$, is fulfillment of inequality

$$\mathbf{y} Y_0 \mathbf{e} > \mathbf{y} Y_2 \mathbf{e} \tag{1}$$

where the vector $\mathbf{y}$ is the unique solution to the system

$$\mathbf{y}(Y_0 + Y_1 + Y_2) = \mathbf{y},\ \mathbf{ye} = 1.$$

It is easy to verify that for the considered Markov chain the matrices Y_0, Y_1 and Y_2 have the following form:

$$Y_0 = I,\ Y_1 = O,\ Y_2 = O.$$

It is easy to see that here the ergodicity condition (1) is transformed to inequality $1 > 0$ which is true for all possible values of the system parameters.

Thus, the following limits (stationary probabilities) exist for any set of the system parameters:

$$\pi(i,r,\nu,n,m) = \lim_{t\to\infty} P\{i_t = i,\ r_t = r,\ \nu_t = \nu,\ n_t = n, m_t = m\},$$

$$i \geq 0,\ r = \overline{0,1},\ \nu = \overline{0,W},\ n = \overline{1,M},\ m = \overline{1,M}.$$

Then let us form the row vectors $\boldsymbol{\pi}_i$ of the stationary probabilities as follows:

$$\boldsymbol{\pi}_0 = (\pi(0,0,0), \pi(0,0,1), \ldots, \pi(0,0,W)),$$

$$\boldsymbol{\pi}_1 = (\boldsymbol{\pi}(1,0,0), \boldsymbol{\pi}(1,0,1), \ldots, \boldsymbol{\pi}(1,0,W)),$$

where

$$\boldsymbol{\pi}(1,0,\nu) = (\pi(1,0,\nu,1), \pi(1,0,\nu,2), \ldots, \pi(1,0,\nu,M)),\ \nu = \overline{0,W}.$$

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1)),\ i \geq 2,$$

where

$$\boldsymbol{\pi}(i,r) = (\boldsymbol{\pi}(i,r,0), \boldsymbol{\pi}(i,r,1), \ldots, \boldsymbol{\pi}(i,r,W)),\ r = 0,1,\quad i \geq 2,$$

$$\boldsymbol{\pi}(i,0,\nu) = (\boldsymbol{\pi}(i,0,\nu,1), \boldsymbol{\pi}(i,0,\nu,2), \ldots, \boldsymbol{\pi}(i,0,\nu,M)),\ \nu = \overline{0,W},$$

$$\boldsymbol{\pi}(i,0,\nu,n) = (\pi(i,0,\nu,n,1), \pi(i,0,\nu,n,2), \ldots, \pi(i,0,\nu,n,M)),\ n = \overline{1,M},$$

$$\boldsymbol{\pi}(i,1,\nu) = (\pi(i,1,\nu,1), \pi(i,1,\nu,2), \ldots, \pi(i,1,\nu,M)),\ \nu = \overline{0,W}.$$

It is well known that the probability vectors $\boldsymbol{\pi}_i$, $i \geq 0$, satisfy the following system of linear algebraic equations:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)Q = \mathbf{0},\quad (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)\mathbf{e} = 1. \tag{2}$$

To solve system (2), we use the numerically stable algorithm for computation of the probability vectors $\boldsymbol{\pi}_i$, $i \geq 0$, developed in [10] which effectively uses information about the asymptotic behavior of the Markov chain ξ_t, $t \geq 0$, and the sparse structure of the generator Q.

4 Performance Measures

The average number N_{buffer} of customers in the buffer is computed by

$$N_{buffer} = \sum_{i=3}^{\infty} (i-2)\boldsymbol{\pi}_i \mathbf{e}.$$

The average number N_{busy} of busy servers at an arbitrary moment is computed by

$$N_{busy} = \sum_{i=1}^{\infty} \min\{i,2\}\boldsymbol{\pi}_i \mathbf{e}.$$

The probability $P_{blocked}$ that a server is blocked at an arbitrary moment is computed by

$$P_{blocked} = \sum_{i=2}^{\infty} \boldsymbol{\pi}(i, 1)\mathbf{e}.$$

The probability P_{loss} that an arbitrary customer will be lost (due to impatience) is computed by

$$P_{loss} = \frac{\alpha N_{buffer}}{\lambda}.$$

The average intensity λ_{out} of flow of customers who receive service is computed by

$$\lambda_{out} = \lambda(1 - P_{loss}).$$

5 Conclusion

Two server queueing model with an infinite buffer and Markovian arrival process is analysed. Service times by two servers have phase type distribution with coinciding state spaces of underlying Markov chains. The phases of service times at two servers are implemented independently if the underlying Markov chains of services currently have different states (phases). If the required phases of service coincide, service by one of the servers is postponed until the phase will be released by the competitive server. Generator of multi-dimensional Markov chain describing behavior of the system is written down. Formulas for computation of the key performance measure of the system in terms of stationary probabilities of the Markov chain are presented. It is planned to extend the results to the cases when the number of servers is more than two, when the state spaces of underlying processes of service coincide partially, when the vectors and subgenerators defining irreducible representations of service times may we selected in such a way as to minimize possibility of conflicts, when the number of active servers can be dynamically changed, etc.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant No.2014R1A1A4A01007517).

References

1. Neuts, M.F.: Matrix-geometric Solutions in Stochastic Models. The Johns Hopkins University Press, Baltimore (1981)
2. Bocharov, P.P., D'Apice, C., Pechinkin, A.V., Salerno, S.: Queueing Theory. VSP, Utrecht-Boston (2004)
3. Asmussen, S.: Applied Probability and Queues. Springer, New York (2003)
4. Lucantoni, D.: New results on the single server queue with a batch Markovian arrival process. Commun. Stat. Stoch. Models **7**, 1–46 (1991)

5. Heyman, D.P., Lucantoni, D.: Modelling multiple IP traffic streams with rate limits. IEEE/ACM Trans. Netw. **11**, 948–958 (2003)
6. Klemm, A., Lindermann, C., Lohmann, M.: Modelling IP traffic using the batch Markovian arrival process. Perform. Eval. **54**, 149–173 (2003)
7. Chakravarthy, S.R.: The batch Markovian arrival process: a review and future work. In: Krishnamoorthy, A., Raju, N., Ramaswami, V. (eds.) Advances in Probability Theory and Stochastic Processes, pp. 21–29. Notable Publications Inc., New Jersey (2001)
8. Graham, A.: Kronecker Products and Matrix Calculus with Applications. Ellis Horwood, Cichester (1981)
9. Klimenok, V.I., Dudin, A.N.: Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. Queueing Syst. **54**, 245–259 (2006)
10. Dudina, O., Kim, C., Dudin, S.: Retrial queueing system with Markovian arrival flow and phase-type service time distribution. Comput. Ind. Eng. **66**, 360–373 (2013)