Queueing Models and
Service Management
Vol. 8, No. 1, page 21–49, 2025

QMSM

Analysis of a Dual Tandem Queue with Servers Floating Between the Stages

Sergei Dudin[1], Achyutha Krishnamoorhy[2], Alexander Dudin[1,*] and Olga Dudina[1]

[1]Belarusian State University

4, Nezavisimosti Ave., Minsk, 220030, Belarus

[2]Centre for Research in Mathematics

CMS College

Kottayam, Kerala, 686001, India

(*Received April 2024; accepted October 2024*)

Abstract: A tandem queueing system with a correlated arrival process and two multi-server stages is analyzed. The capacity of the buffer at stage 1 is infinite. The capacity of the buffer at stage 2 is finite. The total number of available servers is fixed. Servers are dynamically shared between the stages in such a way that any server cannot stay idle if at least one of the buffers is not idle. Servers can transit between the stages only at service completion epochs. Servers from stage 2 can transit to stage 1 even if the buffer at stage 2 is not empty, according to the control policy defined by two integer thresholds. If service by the server assigned to stage 2 is completed when the number of customers in the stage 1 buffer is not less than the first threshold and the number of customers in the stage 2 buffer is less than the second threshold, the released server is re-assigned to stage 1 and immediately starts service. If service by a server assigned to stage 1 is completed when the stage 2 buffer is full, the released server is re-assigned to stage 2 and immediately starts service. In the case of service completion at stage 2 during the epoch when the buffers at both stages are idle, the released server is re-assigned to stage 1 and waits for a new customer arrival at this stage. Customers' arrival is described by the Markov arrival process (MAP). Each customer has to receive service at both stages of the tandem or only at stage 1. The service times at both stages have a phase-type distribution with parameters depending on the stage. Under the fixed values of the thresholds, analysis of the stationary behavior of the tandem is implemented, including derivation of the ergodicity condition, computation of the stationary distribution of the number of customers at each stage, and derivation of expressions for the key performance indicators. Analysis is essentially based on the proper use of the notion of the generalized phase-type distribution. The results of numerical experiments illustrating the feasibility of the proposed algorithms and highlighting the dependence of the performance measures of the system on the parameters of the control policy are presented. The problem of the optimal choice of thresholds is briefly considered.

Keywords: Dynamic control by servers sharing, flexible servers, MAP, tandem queueing system.

* Corresponding author
Email: dudin@bsu.by

1. Introduction

1.1. Brief literature review

Tandem queueing systems are adequate mathematical models of many real-world systems, first of all, telecommunication and manufacturing systems in which processing of a customer requires service at more than one station (stage) in turn. This is why there are a lot of papers devoted to their analysis. A search in Google Scholar by the keywords ”tandem queue” made in the middle of September 2024 reveals more than 220 papers published since 2023. Thus, we do not try to give any survey of the existing research in tandem queues except for the tandem queues with moving (or flexible) servers one of which is the subject of research in this paper. Tandem queues with moving servers have been the focus of research at least since the early 1970s; see, e.g., [42, 43]. The relevant research and possible applications of the model are described, e.g., in [1, 2, 4, 7, 9, 23, 30, 45]. As motivation for consideration of the tandem queues with moving servers, it is mentioned in [2] that ”the use of a cross-trained workforce has become prevalent in the manufacturing and service industries”. Workers of many enterprises or department stores have multiple skills that allow them to implement different operations depending on the load. E.g., a store cashier can do the work of stocking goods when he/she is not busy with his/her main job. Therefore, the problem of choosing the optimal strategy of the cashier switching between the main job and the alternative work exists and deserves the interest of mathematicians.

A tandem with moving servers means the tandem queue, mainly dual, i.e., consisting of two sequential stages, in which there is a common pool of servers that can operate at both stages of a tandem. Tandems with a single server, which services customers at two stages according to some schedule, see, e.g., [30], essentially may be interpreted as polling systems in which a server sequentially polls the buffers. The literature on polling queues is huge; see, e.g., [52, 54, 55], and we will not touch here the tandems with a single flexible server like considered, e.g., in [59].

A tandem queue with two moving servers where a server operating at stage 1, who became idle, can help serve the customers at stage 2, was considered in [25]. The model with moving servers, where the server at either stage, while idle, can move to help the server at the other stage, is considered in [24]. In the dual tandem model considered in [46], each stage has a dedicated server, which can serve customers only at the assigned stage, and one flexible server that can operate at both stages. A tandem model with two servers, both flexible, was analyzed in [2]. Customers holding costs at both stages are taken into account. The arrival flow is defined by the stationary Poisson process. Exponential distribution of a customer service time has the rate dependent on the stage. Preemption of service is permitted. Collaborative (when the servers may collaborate to work on the same customer at the same time, thereby doubling the service rate at that stage), and non-collaborative versions are considered.

Several variants of the tandems with a finite intermediate buffer and two or more servers were considered in [3, 4, 5, 6, 7, 31, 32, 45] where the properties of the optimal strategy of servers moving between the stages are established, sometimes in the case of three available

servers.

All considered models of systems with more than one server at a stage assume the stationary Poisson flow of arriving customers and the exponential distribution of service times.

1.2. Contributions of the paper

The advantages of the tandem queueing model considered in the present paper over the known models are as follows:

1) Arrivals occur according to the MAP. More information about this arrival process can be found in [11, 12, 13, 16, 26, 39, 56]. Queueing systems with the MAP were recently considered, e.g., in [15, 22, 40, 49]. The MAP is an essentially better model of real-world flows in modern telecommunication and manufacturing systems than the stationary Poisson process because it allows to account for fluctuations in the instantaneous arrival rate, which are typical for many real-world systems. Tandem queues with the MAP and its generalization, a batch MAP, were previously considered in the literature; for the lists of the relevant papers, see, e.g., [19, 20, 33, 35, 56] and references therein. As motivation for consideration of the tandem queues with moving servers, it is mentioned in [2] that "The use of a cross-trained workforce has become prevalent in the manufacturing and service industries". Workers of many enterprises or department stores have multiple skills that allow them to implement different operations depending on the load. E.g., a store cashier can stock goods when he/she is not busy with his/her main job. As another simple example of the real system modeled by the tandem queue with moving servers, we can mention the system where unloading of containers for cargo transportation to a warehouse is performed. At the first stage, loaders extract containers from the vehicles and deliver them to an unloading zone having a finite capacity. The second stage corresponds to the delivering of a container from the unloading zone to a warehouse. If an unloader meets the unloading zone full, it can deliver the container directly to the warehouse. Therefore, considered tandem queues are the models of various real-world objects and the problem of choosing the optimal strategy of the server switching between the main job and the alternative work exists and deserves the interest of mathematicians.

2) The service time distribution at both stages is assumed to have a phase type (PH distribution) distinct on both stages. This distribution is essentially more general than the exponential distribution assumed in the majority of the relevant papers and allows for fitting not only the mean service rate, but the variance and the higher initial moments of service time of and the shape of the actual distribution of the service time in a real-world system.

3) The impatience of waiting customers (possibility of abandoning service in case of too long waiting), which is the inherent feature of many real-world systems, see, e.g., [14, 17, 34, 50, 51, 57] is taken into account. The impatience (at the second stage) for a tandem with flexible servers was recently taken into account in [7] and [9]. In [58], waiting customers are assumed to be impatient at both stages of a tandem. However, server transitions between the stages are not considered. We allow the impatience of waiting customers at both stages under more general assumptions about the arrival process and flexible servers. The analysis of the model is implemented exactly with the help of the tools of the well-known Quasi-

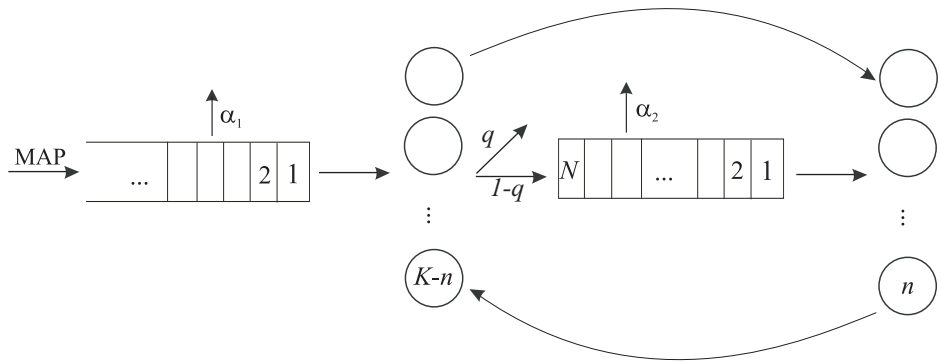

Figure 1. Structure of the tandem

Birth-and-Death processes or asymptotically quasi-Toeplitz Markov chains in cases where the customers at stage 1 are absolutely patient or impatient, respectively.

4) A new mechanism of server transitions between the stages is considered. This mechanism suggests that a server, which completes service at some stage when the buffer at this stage is empty, immediately transits to another stage and starts service if there are waiting customers. In addition to this, the threshold-type strategy of the distribution of available servers between the stages is fixed. This strategy assumes assigning a stage 2 server to stage 1 in the case of a long queue at stage 1 and a short queue at stage 2.

5) An effective way for simultaneous description of the states of the service underlying processes in multiple heterogeneous servers operating in parallel is presented and illustrated.

1.3. The structure of the presentation

The outline of the presentation is the following: An exact mathematical description of the tandem queue under study is given in Section 2. In Section 3, the behavior of the system is described by the continuous-time multidimensional Markov chain (MC). The generator of this chain is derived. Conditions for the ergodicity of this chain in the cases of patient and impatient customers are presented. The computation of the stationary distribution of the chain is briefly discussed. In Section 4, formulas for computing the values of the main performance measures of the tandem are presented. Section 5 contains the results of the numerical experiment. The dependencies of the main performance measures on the values of the thresholds are highlighted. An example of a solution to the optimization problem is presented. Section 6 summarizes the content of the paper and outlines possible generalizations of the model.

2. Description of the Model

Let us consider a tandem queueing system consisting of two stages. The scheme of the tandem under study is presented in Figure 1.

The tandem consists of two stages, each of which is modeled by the multi-server queueing system with a varying number of available servers. The arrivals of customers are defined by the MAP. This process is defined by the irreducible underlying Markov chain $\nu_t,\ t \geq 0,$

with the state space $\{1, 2, \ldots, W\}$ and by two matrices of size W. The matrix $D_1(D_0)$ defines rates of transitions of the chain ν_t that are accompanied (or not accompanied) by the arrival of a customer. The mean arrival rate is defined as $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$ where $\boldsymbol{\theta}$ is the invariant probability vector of the chain ν_t. This vector is the single solution to the equations $\boldsymbol{\theta}(D_0 + D_1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$. Here, $\mathbf{0}$ is the row vector consisting of 0s, and $\mathbf{e}$ is the column vector consisting of 1s.

For useful information about the $MAP,$ its properties, particular cases (including the stationary Poisson process, renewal arrival process with PH distribution of inter-arrival times, Markov Modulated Poisson process, etc.), and computation of the main characteristics, see, e.g., [11, 12, 13, 16, 39, 56]. The problem of an adequate description of arriving flows in real-world systems by the MAP based on the observation of arrival epochs is already well-addressed in the literature; see, e.g., [10, 26, 44].

The capacity of the buffer at stage 1 is assumed to be infinite. The capacity of the buffer at stage 2 is finite and equal to N. After service at stage 1 of the tandem, a customer leaves the system permanently with probability $q,\ 0 \leq q \leq 1,$ and with the complementary probability $1 - q$ requires an additional service and proceeds to stage 2.

The total number K of available servers is fixed. Servers are dynamically shared between the stages. Normally, the server assigned to station r, $r = 1, 2,$ that finishes service of a customer immediately starts service of a customer from the rth buffer if it is not empty. If this buffer is empty, the server starts service at another stage. In the case of service completion at stage 2, during the epoch when the buffers at both stages are idle, the free server is re-assigned to stage 1 and waits for a new customer's arrival at this stage.

Re-assignment of the servers from stage 2 to state 1 is also implemented at service completion epochs when the number of requests in the buffer of stage 2 is relatively small compared to the buffer of stage 1. More exactly, such a re-assignment occurs according to the threshold strategy defined by two integer thresholds, n_1 and n_2. If service by the server assigned to stage 2 is completed when the number of customers in the buffer of stage 2 is less than the threshold $n_2,\ n_2 = \overline{1, N}$, while the number of customers in the buffer of stage 1 is greater or equal to the threshold $n_1,\ n_1 \geq 1$, then the released server is re-assigned to stage 1 and immediately starts service of a customer from the first buffer. Here, denotation like $n_2 = \overline{1, N}$ means that the variable n_2 admits the values from the set $\{1, 2, \ldots, N\}$.

If service by the server assigned to stage 1 is completed when the buffer of stage 2 is full and the customer who just obtained service decides to transit to stage 2, then the released server is re-assigned to stage 2 and immediately starts service of a customer at stage 2. Therefore, customer loss or stage 1 server blocking due to the overflow of the finite intermediate buffer is not possible in the considered model, in contrast to the standard tandem queues with the fixed numbers of servers at each stage.

The customer's service time at the stage r has a PH distribution defined by the underlying process $\eta_t^{(r)}$ having the set $\{1, 2, \ldots, M^{(r)}\}$ of transient phases and so-called irreducible representation $(\boldsymbol{\beta}^{(r)}, S^{(r)}),\ r = 1, 2$. The stochastic row vector $\boldsymbol{\beta}^{(r)}$ defines the distribution of the states of the process $\eta_t^{(r)}$ at the service beginning epoch at stage r. The sub-generator $S^{(r)}$ defines the transition rates of the process $\eta_t^{(r)}$ within the set of transient states. The col-

umn vector $\mathbf{S}_0^{(r)} = -S^{(r)}\mathbf{e}$ defines the transition rates of the process $\eta_t^{(r)}$ into the absorbing state. The transition to this state corresponds to the end of the customer's service. The mean service time of a customer is computed by $b_1^{(r)} = \boldsymbol{\beta}^{(r)}(-S^{(r)})^{-1}\mathbf{e},\ r = 1, 2.$

For useful information about the PH distribution, its properties, particular cases, characteristics, and suitability for approximation of an arbitrary distribution of a non-negative random variable, see, e.g., [16, 41] and [8]. The problem of estimation of parameters $(\boldsymbol{\beta}^{(r)}, S^{(r)})$ of the PH distribution based on the results of observation of values of service time in a real-world system has also been intensively addressed in the existing literature, see, e.g., [10, 26, 53].

We assume that the customers residing in the buffers are impatient. A customer waiting in the r-th buffer leaves the buffer after an exponentially distributed time with the parameter α_r, $r = 1, 2$, independently of other waiting customers. It permanently departs from the tandem.

Our aim is to analyze the stationary distribution of the states of the tandem (including the number of assigned servers and the number of customers in the buffer at each stage), derive expressions for the computation of its key performance indicators, and numerically solve the problem of the choice of the thresholds (n_1, n_2) providing the optimum to one of the possible criteria of the quality of operation of the tandem.

3. The Process of System States

3.1. The choice of the MC defining the behavior of the system

It is clear that a description of the state of the system has to include the following:

1) the number of customers in two buffers (or the total number of customers in buffers and the number of customers in buffers of stage 1 or stage 2);

2) the state of the underlying process of customer arrival;

3) the number of servers providing service at each stage (or the total number of busy servers and the number of servers providing service at stage 1 or stage 2);

4) information about the current phases of service provided by all busy servers at both stages.

As follows from this list of the mandatory components, there are many options for the choice of the multidimensional Markov process describing the behavior of the tandem. Thus, it is desirable to choose the best option. Taking into account the perspective of an algorithmic analysis of this process and its further computer implementation, the best choice assumes the minimal cardinality of the state space of the finite components of the process. This cardinality defines the size of the finite blocks of the infinite-size generator of the process. This, in turn, defines the feasibility of the computer realization of the computation of the stationary distribution of the introduced process.

We make the choice of the Markov process as follows:

Let

$i(t)$ be the total number of customers waiting in both buffers, $i(t) \geq 0$,

$n(t)$ be the number of customers in the buffer of stage 2, $n(t) = \overline{0, \min\{i(t), N\}}$,

$k(t)$ be the number of busy servers, $k(t) = \overline{0, K}$,

ν_t be the state of the underlying process of arrivals, $\nu_t = \overline{1, W}$,

$\boldsymbol{\eta}_t$ be the vector process defining the current state of underlying processes of service in $k(t)$ busy servers at the moment t, $t \geq 0$.

It is easy to see that, if $i(t) = 0$, then $n(t) = 0$. If $i(t) > 0$, then $k(t) = K$. If $k(t) = 0$, the process $\boldsymbol{\eta}_t$ is omitted.

Because the available servers are dynamically distributed among the stages and the PH distributions of service processes at various stages have distinct representations, the problem of properly defining the vector process $\boldsymbol{\eta}_t$ is quite difficult. To solve this problem, we use two effective tricks.

The first trick consists of the avoidance of permanent monitoring the association of any server with a certain stage of a tandem. This trick is realized here via the use of the so-called generalized phase type (GPH) distribution introduced in [35]. This distribution is defined by a continuous time MC having the space of transient states $\{1, 2, \ldots, M\}$ where $M = M^{(1)} + M^{(2)}$ and irreducible representation $(\boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2, S)$. Here, the vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are defined as

$$\boldsymbol{\beta}_1 = (\boldsymbol{\beta}^{(1)}, \mathbf{0}_{M^{(2)}}),\ \boldsymbol{\beta}_2 = (\mathbf{0}_{M^{(1)}}, \boldsymbol{\beta}^{(2)})$$

and the sub-generator S is a block-diagonal matrix with the diagonal blocks $S^{(r)},\ r = 1, 2$.

The main advantage of the use of the GPH distribution consists of the following: The choice of the initial phase of the service underlying process *depends* on the stage at which service begins. If service begins at stage r, this choice is implemented randomly with the probabilities given by the entries of the vector $\boldsymbol{\beta}_r,\ r = 1, 2$. After the service begins, transitions of the underlying process in the set of transient states are defined by the sub-generator S, *irrespective* of the stage of tandem. This allows us to avoid the separate consideration of the different variants of servers assignment to the stages of a tandem and the phases of the customer's service at these servers.

The second trick, which is effective in combination with the first one and was initially proposed in [47] and [48], consists of a description of service processes on the servers, not via tracking the phases of the underlying process on each server but via counting how many servers have each value of the service underlying process at an arbitrary moment. In the case of relatively small numbers M of transient states and large numbers K of servers, this significantly reduces (see [28]) the cardinality of the state space of the process $\boldsymbol{\eta}_t$ from $(M + 1)^K$ to $\frac{(K+M)!}{M!K!}$. E.g., if $M = 2$ and $K = 5$, then the cardinality reduces from 243 to 21. If $M = 2$ and $K = 10$, then the cardinality reduces from 59 049 to 66.

Therefore, we suggest that the process $\boldsymbol{\eta}_t$ is defined by the formula $\boldsymbol{\eta}_t = \{\eta_t^{(1)}, \eta_t^{(2)}, \ldots, \eta_t^{(M)}\}$, where $\eta_t^{(m)}$ is the number of customers at the phase m of the underlying process of the generalized PH service process.

The behavior of the considered tandem is completely described by the above-defined multidimensional continuous-time MC

$$\zeta_t = \{i(t), n(t), k(t), \nu_t, \boldsymbol{\eta}_t\},\ t \geq 0.$$

The state space of the process ζ_t is

$$\{(0,0,0,\nu),\ \nu=\overline{1,W}\}$$
$$\bigcup\{(0,0,k,\nu,\boldsymbol{\eta}),\ k=\overline{1,K},\ \nu=\overline{1,W},\boldsymbol{\eta}=\{\eta^{(1)},\eta^{(2)},\dots,\eta^{(M)}\},$$
$$\eta^{(m)}=\overline{0,k},\ m=\overline{1,M},\ \sum_{m=1}^{M}\eta^{(m)}=k\}$$
$$\bigcup\{(i,n,K,\nu,\boldsymbol{\eta}),\ i>0,\ n=\overline{0,\min\{i,N\}},\ \nu=\overline{1,W},\boldsymbol{\eta}=\{\eta^{(1)},\eta^{(2)},\dots,\eta^{(M)}\},$$
$$\eta^{(m)}=\overline{0,K},\ m=\overline{1,M},\ \sum_{m=1}^{M}\eta^{(m)}=K\}.$$

3.2. Generator of the MC

Let us enumerate the states of the MC $\zeta_t,\ t\geq 0$, in the direct lexicographic order of the components $(i(t),n(t),k(t),\nu_t)$ and the reverse lexicographic order of the components $(\eta_t^{(1)},\dots,\eta_t^{(M)})$. We call the set of states having the value i of the first component of the MC the *level* i of the chain ζ_t.

Let the entries of the matrix $Q_{i,j}$, except the diagonal entries of the matrix $Q_{i,i}$, define the rates of transition of the MC ζ_t from the states that belong to the level i to the states that belong to the level j. The diagonal entries of the matrix $Q_{i,i}$ are negative. Their modules are equal to the rates of the exit of the MC ζ_t from the states that belong to the level i.

Theorem 1. The infinitesimal generator Q of the MC $\zeta_t,\ t\geq 0$, has a block-tridiagonal structure

$$Q=\begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & O & \dots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & O & \dots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & O & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{1}$$

where the non-zero blocks $Q_{i,j},\ |i-j|\leq 1,$ are defined as follows:

$$Q_{0,0}=\begin{pmatrix} Q_{0,0}^{(0,0)} & Q_{0,0}^{(0,1)} & O & \dots & O & O \\ Q_{0,0}^{(1,0)} & Q_{0,0}^{(1,1)} & Q_{0,0}^{(1,2)} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & Q_{0,0}^{(K,K-1)} & Q_{0,0}^{(K,K)} \end{pmatrix}, \tag{2}$$

$$Q_{0,0}^{(0,0)}=D_0,$$

$$Q_{0,0}^{(k,k+1)}=D_1\otimes P_k(\boldsymbol{\beta}_1),\ k=\overline{0,K-1}, \tag{3}$$

$$Q_{0,0}^{(k,k)}=D_0\oplus(A_k+\Delta_k)+(1-q)I_W\otimes L_k^{(1)}P_{k-1}(\boldsymbol{\beta}_2),\ k=\overline{1,K}, \tag{4}$$

$$Q_{0,0}^{(k,k-1)}=I_W\otimes L_k^{(2)}+qI_W\otimes L_k^{(1)},\ k=\overline{1,K}, \tag{5}$$

$$Q_{0,1} = \begin{pmatrix} O_{W\times 2WT_K} \\ O_{WT_1\times 2WT_K} \\ O_{WT_2\times 2WT_K} \\ \cdots \\ O_{WT_{K-1}\times 2WT_K} \\ E_0^+ \otimes D_1 \otimes I_{T_K} \end{pmatrix}, \tag{6}$$

$$Q_{1,0} = \begin{pmatrix} O_{2WT_K\times W} & O_{2WT_K\times WT_1} & O_{2WT_K\times WT_2} & \cdots & O_{2WT_K\times WT_{K-1}} & X \end{pmatrix}, \tag{7}$$

where

$$X = \begin{pmatrix} \alpha_1 I_{WT_K} + I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_1) + qI_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1) \\ \alpha_2 I_{WT_K} + I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_2) + qI_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_2) \end{pmatrix}, \tag{8}$$

$$\begin{aligned} Q_{i,i} =& I_{i+1} \otimes (D_0 \oplus (A_K + \Delta_K)) + (1-q)\bar{I}_i \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_2) \\ & - (\alpha_2 C_i + \alpha_1(iI_{i+1} - C_i)) \otimes I_{WT_K} + (1-q)I_i^+ \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1),\ 0 < i < N, \end{aligned} \tag{9}$$

$$\begin{aligned} Q_{i,i} =& I_{N+1} \otimes (D_0 \oplus (A_K + \Delta_K)) + (1-q)\bar{I}_N \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_2) \\ & - (\alpha_2 C_N + \alpha_1(iI_{N+1} - C_N)) \otimes I_{WT_K} + (1-q)I_N^+ \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1),\ i \geq N, \end{aligned} \tag{10}$$

$$\begin{aligned} Q_{i,i-1} =& E_i^- \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_2) + \hat{E}_i^- \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_1) \\ & + \alpha_2 C_i E_i^- \otimes I_{WT_K} + \alpha_1(iI_{i+1} - C_i)\tilde{E}_i^- \otimes I_{WT_K} + q\tilde{E}_i^- \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1) \\ & + q\bar{E}_i^- \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_2), 1 < i \leq \min\{n_1 - 1, N\}, \end{aligned} \tag{11}$$

$$\begin{aligned} Q_{i,i-1} =& I^- \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_2) + \hat{I} \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_1) + \alpha_2 C_N I^- \otimes I_{WT_K} \\ & + \alpha_1(iI_{N+1} - C_N) \otimes I_{WT_K} + qI_{(N+1)W} \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1),\ N < i < n_1, \end{aligned} \tag{12}$$

$$\begin{aligned} Q_{i,i-1} =& \tilde{Z}_i^- \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_2) + Z_i^- \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_1) + \alpha_2 C_i E_i^- \otimes I_{WT_K} \\ & + \alpha_1(iI_{i+1} - C_i)\tilde{E}_i^- \otimes I_{WT_K} + q\tilde{E}_i^- \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1) \\ & + q\bar{E}_i^- \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_2),\ N \geq i \geq n_1, \end{aligned} \tag{13}$$

$$\begin{aligned} Q_{i,i-1} =& \tilde{J}_i^- \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_2) + J_i^- \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_1) + \alpha_2 C_N I^- \otimes I_{WT_K} \\ & + \alpha_1(iI_{N+1} - C_N) \otimes I_{WT_K} + qI_{(N+1)W} \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1),\ i > \max\{n_1 - 1, N\}, \end{aligned} \tag{14}$$

$$Q_{i,i+1} = E_i^+ \otimes D_1 \otimes I_{T_K},\ 1 < i < N, \tag{15}$$

$$Q_{i,i+1} = I_{N+1} \otimes D_1 \otimes I_{T_K},\ i \geq N. \tag{16}$$

Here,

- the column vector $\mathbf{S}_0$ is defined as $\mathbf{S}_0 = -S\mathbf{e}$;
- the matrix $L_n^{(r)} = L_n^{(r)}(\mathbf{S}_0)$ defines the transition rates of the process $\{\eta_t^{(1)},\ \ldots, \eta_t^{(M)}\}$ at the epoch when service in one of n busy servers is completed at stage r, $r = 1, 2,\ n = \overline{1, K}$;
- the matrix $A_n = A_n(S)$ contains the transition rates of the process $\{\eta_t^{(1)},\ \ldots, \eta_t^{(M)}\}$ at the epoch of the change of the phase of service in one of n busy servers, $n = \overline{1, K}$;
- the matrix $P_n(\boldsymbol{\beta}_r)$ defines the transition rates of the process $\{\eta_t^{(1)},\ \ldots, \eta_t^{(M)}\}$ when a customer starts service at stage r and n servers are busy, $r,\ r = 1, 2,\ n = \overline{0, K-1}$;
- the diagonal elements of the diagonal matrix Δ_n define the rates of the exit of the process $\{\eta_t^{(1)},\ \ldots, \eta_t^{(M)}\}$ from its states. This matrix Δ_n is computed by the formula

$$\Delta_n = -\text{diag}\{A_n\mathbf{e} + L_n^{(1)}\mathbf{e} + L_n^{(2)}\mathbf{e}\},\ n = \overline{1, K};$$

- $\otimes$ and $\oplus$ are symbols of Kronecker product and sum of matrices; see their definition and properties in [27];
- E_i^+, $i = \overline{0, N-1}$, is the matrix of size $(i+1) \times (i+2)$ with all zero entries except the entries $(E_i^+)_{l,l}$, $l = \overline{0, i}$, which are equal to 1;
- E_i^-, $i = \overline{2, N}$, is the matrix of size $(i+1) \times i$ with all zero entries except the entries $(E_i^-)_{l,l-1}$, $l = \overline{1, i}$, which are equal to 1;
- $\tilde{E}_i^-$, $i = \overline{2, N}$, is the matrix of size $(i+1) \times i$ with all zero entries except the entries $(\tilde{E}_i^-)_{l,l}$, $l = \overline{0, i}$, which are equal to 1;
- $\hat{E}_i^-$, $i = \overline{2, N}$, is the matrix of size $(i+1) \times i$ with all zero entries except the entry $(\hat{E}_i^-)_{0,0}$ which is equal to 1;
- $\bar{E}_i^-$, $i = \overline{2, N}$, is the matrix of size $(i+1) \times i$ with all zero entries except the entry $(\bar{E}_i^-)_{i,i-1}$ which is equal to 1;
- C_i, $i = \overline{1, N}$, is the diagonal matrix of size $i+1$ with the diagonal entries $\{0, 1, \ldots, i\}$;
- $\bar{I}_i$, $i = \overline{1, N}$, is square the matrix of size $i+1$ with all zero entries except the entry $(\bar{I}_i)_{i,i}$ which is equal to 1;
- I_i^+, $i = \overline{1, N}$, is the square matrix of size $i+1$ with all zero entries except the entries $(I_i^+)_{l,l+1}$, $l = \overline{0, i-1}$, which are equal to 1;
- I^- is the square matrix of size $N+1$ with all zero entries except the entries $(I^-)_{l,l-1}$, $l = \overline{1, N}$, which are equal to 1;

- $\hat{I}$ is the square matrix of size $N+1$ with all zero entries except the entry $(\hat{I})_{0,0}$ which is equal to 1;

- Z_i^-, $i = \overline{n_1, N}$, is the matrix of size $(i+1) \times i$ with all zero entries except the entries $(Z_i^-)_{l,l}$, $l = \overline{0, \min\{i-n_1, n_2-1\}}$, which are equal to 1;

- $\tilde{Z}_i^-$, $i = \overline{n_1, N}$, is the matrix of size $(i+1) \times i$ with all zero entries except the entries $(\tilde{Z}_i^-)_{l,l-1}$, $l = \overline{\min\{i-n_1, n_2-1\}+1, i}$, which are equal to 1;

- J_i^-, $i \geq \max\{n_1, N+1\}$, is the square matrix of size $N+1$ with all zero entries except the entries $(J_i^-)_{l,l}$, $l = \overline{0, \min\{i-n_1, n_2-1\}}$, which are equal to 1;

- $\tilde{J}_i^-$, $i \geq \max\{n_1, N+1\}$, is the square matrix of size $N+1$ with all zero entries except the entries $(\tilde{J}_i^-)_{l,l-1}$, $l = \overline{\min\{i-n_1, n_2-1\}+1, N}$, which are equal to 1;

- the number T_n is equal to the cardinality of the state space of the process $\{\eta_t^{(1)}, \ldots, \eta_t^{(M)}\}$ when n servers provide service. It is calculated as

$$T_0 = 1, \; T_n = \frac{(n+M-1)!}{n!(M-1)!}, \; n = \overline{1, K}.$$

The recursive algorithms for the calculation of matrices $P_n(\boldsymbol{\beta}_r)$, $n = \overline{0, K-1}$, $L_n^{(r)}$, and A_n, $n = \overline{1, K}$, $r = 1, 2$, can be found in [37].

Proof of the theorem is implemented by means of careful analysis of all possible transitions of the MC ζ_t during an interval of infinitesimal length. Because the probabilistic meaning of the matrices $P_n(\boldsymbol{\beta}_r)$, $n = \overline{0, K-1}$, $L_n^{(r)}$, and A_n, $n = \overline{1, K}$, $r = 1, 2$, is explained above, the formulas for the blocks of the generator are almost self-explanatory. Thus, we present only a brief proof.

Block tridiagonal form (1) of the generator Q is clear because, due to the properties of the MAP flow and PH distribution, as well as the exponential distribution, the total number of customers in buffers during a very short interval of time can increase or decrease only by one.

The blocks $Q_{0,0}^{(k,k')}$ of the matrix $Q_{0,0}$ define (except the diagonal entries of the blocks $Q_{0,0}^{(k,k)}$) the rates of transitions of the MC ζ_t which are related to the change of the number of busy servers from k to k' when both buffers are empty. This number, during a short interval of time, can remain the same, decrease, or increase by one. This explains the block tridiagonal form (2) of the matrix $Q_{0,0}$.

Form (3) of the block $Q_{0,0}^{(k,k+1)}$ is clear because the increase in the number of busy servers can occur only due to a new customer arrival (the transition intensities are given by the matrix D_1). This customer starts service at stage 1. The initial state of the underlying process of this service is defined by the matrix $P_k(\boldsymbol{\beta}_1)$. Because the arrival of a new customer and the installation of the initial state occur at the same moment, as follows from the properties of the Kronecker product of matrices, we obtain formula (3).

The blocks $Q_{0,0}^{(k,k)}$ have form (4) because the exits from the states of the MC ζ_t and transitions without the change of the number of customers in buffers and the number of serviced customers can occur only if: (i) the underlying process of arrivals makes an exit or a transition without generation of a customer; or (ii) the underlying process of the GPH makes an exit or a transition without service completion; or (iii) a customer completes service at stage 1, transits to stage 2, and initiation of the state of the underlying process at stage 2 occurs. Taking into account that

$$D_0 \otimes I_{T_k} + I_W \otimes (A_k + \Delta_k) = D_0 \oplus (A_k + \Delta_k),$$

and the rates corresponding to option (iii) are given by the matrix $(1-q)I_W \otimes L_k^{(1)} P_{k-1}(\boldsymbol{\beta}_2)$, $k = \overline{1, K}$, we obtain formula (4).

Form (5) of the block $Q_{0,0}^{(k,k-1)}$ is clear because the decrease in the number of busy servers (under the empty buffers) can occur only due to customer service completion at stage 2 (the transition intensities are given by the matrix $L_k^{(2)}$), or service completion at stage 1 of a customer that does not need service at stage 2. As a result, we obtain formula (5).

The form of the block $Q_{0,0}$ is completely explained.

The transition from level 0 to level 1 can occur only if a new customer arrives at the moment when both buffers are idle but all servers are busy. Therefore, the matrix $Q_{0,1}$ is a block column of form (6) with all zero entries (because appearance of a customer in a buffer is impossible if the number of busy servers is less than K) except the last block, corresponding namely to the situation when all servers are busy, a new customer arrives, the total number of customers in the buffers increases from 0 to 1, and namely the number of customers in the first buffer increases. Thus, we obtain the form $E_0^+ \otimes D_1 \otimes I_{T_K}$ of the last matrix block in block column (6).

The transition from level 1 to level 0 can occur only when all servers are busy and one of them finishes service or one customer residing in the buffers departs from the system due to impatience. Thus, the first K blocks of the block row vector $Q_{1,0}$, corresponding to the busyness of $0, 1, \ldots, K-1$ servers are equal to zero matrices, and only the last block, X, in (7) is not a zero matrix. Form (8) of the block X is explained as follows: The existence of two block rows in (8) is caused by the existence of two options for placing a single customer in two buffers. The first (according to the lexicographic order) option is that the second buffer is empty, which implies that the first one is not empty. In this case, disappearance of a customer from this buffer can occur if: (i) this customer exits from the buffer due to impatience (with rate α_1); (ii) service is finished at stage 2 (with rates defined by the matrix $L_K^{(2)}$) and the customer from the first buffer starts service at stage 1 (probabilities of the choice of the initial state of the service underlying process are given by the matrix $P_{K-1}(\boldsymbol{\beta}_1)$); (iii) service is finished at stage 1, the customer abandons service at stage 2 and server of stage 1 starts new service. As a result of these considerations, we obtain the first block row in (8) as $\alpha_1 I_{WT_K} + I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_1) + qI_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1)$. The form of the second block row is explained analogously, taking into account that this row corresponds to the location of the single waiting customer in the second buffer.

Formula (9) for the matrix $Q_{i,i}$ in the case $0 < i < N$ (the buffer of stage 2 cannot be full) is explained as follows: The negative diagonal entries of this matrix define, up to the sign, the rates of the exit of the MC ζ_t from its states. These entries are given by the corresponding diagonal entries of the matrix $I_{i+1} \otimes (D_0 \oplus (A_K + \Delta_K))$ and the exit rates due to the customers impatience in the first buffer. The latter rates are given by the diagonal entries of the diagonal matrix $(\alpha_2 C_i + \alpha_1 (iI_{i+1} - C_i)) \otimes I_{WT_K}$. The non-diagonal entries of the matrix $Q_{i,i}$ define the rates of transitions in the chain without changing the number of customers in the buffers. They are given by the non-diagonal entries of the matrix $I_{i+1} \otimes (D_0 \oplus (A_K + \Delta_K))$ and by the matrix $(1-q)\bar{I}_i \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_2) + (1-q) I_i^+ \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1)$. The first of the summands here corresponds to the scenario when the server of stage 1 completes service, the buffer of this stage is empty, and the released server continues service for this customer at stage 2. The second summand corresponds to the scenario when the server of stage 1 completes service and the buffer of this stage is not empty. This server starts service for the next customer at stage 1, while the customer, whose service is completed, joins the buffer at stage 2. Formula (10) is explained analogously, taking into account that for $i \geq N$ the buffer at stage 2 can be full and the state space of the second component of the MC is $\{0, \ldots, N\}$, not $\{0, \ldots, i\}$ as in the derivation of formula (9).

There are four different forms (11)–(14) of the block $Q_{i,i-1}$ depending on relations between i, n_1, and N. Let us comment on form (14) for $i > \max\{n_1 - 1, N\}$. All summands in the right-hand size of (14) give the rates of decreasing the total number i of customers in the buffers under the occurrence of the different events.

The summand $\tilde{J}_i^- \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_2)$ corresponds to service completion by the server at stage 2, after which the server takes for service the customer from the buffer of stage 2. The summand $J_i^- \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_1)$ corresponds to service completion by the server at stage 2, after which the server takes for service the customer from the buffer of stage 1. The summand $\alpha_2 C_N I^- \otimes I_{WT_K}$ gives the rates of customer departure from the buffer at stage 2 due to impatience. The summand $\alpha_1 (iI_{N+1} - C_N) \otimes I_{WT_K}$ gives the rates of customer departure from the buffer at stage 1 due to impatience. The summand $qI_{(N+1)W} \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1)$ reflects the case when service is completed at stage 1, the served customer abandons service at stage 2, and the released served starts service for a customer from the buffer of stage 1.

Formulas (15) and (16) for the block $Q_{i,i+1}$ correspond to a new customer arrival and its joining the buffer of stage 1. When $1 < i < N$, this arrival causes the expansion of the set space of the number of customers at the buffer of stage 2, while without the increase of this number. When $i \geq N$, a new arrival increases the number of customers in the buffer at stage 1 without any effect on stage 2.

Theorem 1 is proven.

3.3. Conditions for the ergodicity of the MC

Let us investigate the problem of the existence of the stationary distribution of the MC ζ_t having the generator Q. Let us distinguish two cases: when customers are absolutely patient

at stage 1 ($\alpha_1 = 0$) and when they are impatient there ($\alpha_1 > 0$).

Theorem 2. If customers are absolutely patient at stage 1, i.e., $\alpha_1 = 0$, then the necessary and sufficient condition for the existence of the stationary distribution (ergodicity condition) of the MC ζ_t is the fulfillment of the inequality

$$\mathbf{u}Q^{-}\mathbf{e} > \mathbf{u}Q^{+}\mathbf{e}$$

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

$$\mathbf{u}(Q^{-} + Q^{0} + Q^{+}) = \mathbf{0},\ \mathbf{ue} = 1$$

where

$$\begin{aligned} Q^{-} =& \tilde{J}_N^{-} \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_2) + J_N^{-} \otimes I_W \otimes L_K^{(2)} P_{K-1}(\boldsymbol{\beta}_1) \\ &+ \alpha_2 C_N I^{-} \otimes I_{WT_K} + qI_{(N+1)W} \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1), \end{aligned}$$

$$\begin{aligned} Q^{0} =& I_{N+1} \otimes (D_0 \oplus (A_K + \Delta_K)) + (1-q)\bar{I}_N \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_2) \\ &- \alpha_2 C_N \otimes I_{WT_K} + (1-q) I_N^{+} \otimes I_W \otimes L_K^{(1)} P_{K-1}(\boldsymbol{\beta}_1), \end{aligned}$$

$$Q^{+} = I_{N+1} \otimes D_1 \otimes I_{T_K}.$$

Proof easily follows from [38] because in the case of patient customers, the MC ζ_t is the quasi-Toeplitz MC (or Quasi-Birth-and-Death process) with many boundary states.

Vector $\mathbf{u}$ defines the joint stationary distribution of the number $n(t)$ of customers in the second buffer, the underlying process of arrivals ν_t, and the underlying process of service $\boldsymbol{\eta}_t$ during periods of time when the system is overloaded, i.e., the queue in the first buffer is huge. The size of this vector is equal to $(N+1)WT_N$ and it is easily calculated as the solution of a finite system of linear algebraic equations. When this vector is computed, checking whether an ergodicity condition is met is easy.

Theorem 3. If customers are impatient at stage 1, i.e., $\alpha_1 > 0$, then the stationary distribution of the MC ζ_t exists for any set of the system parameters.

Proof. If $\alpha_1 > 0$, is possible to prove that the MC ζ_t belongs to the class of asymptotically quasi-Toeplitz Markov chains ($AQTMC$), defined in [38].

According to the definition of $AQTMC$ given in [38], the generator Q of the MC ζ_t has to be such that the following matrices exist:

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

where $\mathcal{Q}_i = -I \circ Q_{i,i}$, $i \geq 0$, where $\circ$ is the symbol of the Hadamard product of matrices; see, e.g., [29].

Having explicit expressions for the blocks of the generator Q, it is not difficult to check that the limiting matrices Y_0, Y_1, and Y_2 indeed exist and are equal to I, O, and O, respectively. Thus, the MC ζ_t belongs to the class of $AQTMC$.

It is established in [38] that the sufficient condition for the ergodicity of the $MC\ \zeta_t$ is the fulfillment of inequality

$$\mathbf{w}Y_0\mathbf{e} > \mathbf{w}Y_2\mathbf{e}$$

where the row-vector **w** satisfies the system:

$$\mathbf{w}(Y_0 + Y_1 + Y_2) = \mathbf{w},\ \mathbf{we} = 1.$$

Taking into account that $Y_0 = I, Y_1 = O$, and $Y_2 = O$, the presented inequality admits the trivial form $1 > 0$. Therefore, the $MC\ \zeta_t$ is ergodic for any choice of the system parameters. Theorem 3 is proven.

3.4. Outline of computation of stationary distribution of the MC

Let the conditions of existence of the stationary distribution of the $MC\ \zeta_t$ be fulfilled. Then, the following limits exist:

$$\pi(0,0,0,\nu) = \lim_{t\to\infty} P\{i(t) = 0, n(t) = 0, k(t) = 0, \nu_t = \nu\},\ \nu = \overline{1,W},$$

$$\pi(0,0,k,\nu,\boldsymbol{\eta}) = \lim_{t\to\infty} P\{i(t) = 0, n(t) = 0, k(t) = k, \nu_t = \nu, \boldsymbol{\eta}_t = \boldsymbol{\eta}\},$$
$$k = \overline{1,K},\ \nu = \overline{1,W},$$

$$\pi(i,n,K,\nu,\boldsymbol{\eta}) = \lim_{t\to\infty} P\{i(t) = i, n(t) = n, k(t) = K, \nu_t = \nu, \boldsymbol{\eta}_t = \boldsymbol{\eta}\},$$
$$i \geq 1,\ n = \overline{0, \min\{i, N\}},\ \nu = \overline{1,W},$$

called the stationary probabilities of the system states.

Corresponding to the above-defined ordering of the states of the $MC\ \zeta_t$, let us combine these stationary probabilities into the row vectors $\boldsymbol{\pi}(0,0,0)$ of size W, $\boldsymbol{\pi}(0,0,k)$ of size $WT_k,\ k = \overline{1,K}$, and $\boldsymbol{\pi}(i,n,K)$ of size $WT_K,\ i \geq 1$.

Introduce also the vectors $\boldsymbol{\pi}(0,0)$ of size $W\sum_{k=0}^{K} T_k$, and $\boldsymbol{\pi}(i,n)$ of size $WT_K,\ i \geq 1$.

Finally, introduce the vectors $\boldsymbol{\pi}_0 = \boldsymbol{\pi}(0,0)$ and $\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \dots, \boldsymbol{\pi}(i,N)),\ i \geq 1$.

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

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots)Q = \mathbf{0},\ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots)\mathbf{e} = 1.$$

This system has an infinite size.

In the case of absolutely patient customers, the solution of this system, up to the probabilities $\boldsymbol{\pi}_i,\ i = \overline{0,N}$, of the boundary levels, has the matrix geometric form, while these probabilities of boundary levels are computed as a solution of a finite system of linear algebraic equations. More details about the algorithms for solving this system and their realization can be found in [16, 41].

In the case of impatient customers, such systems are less investigated in the existing literature. However, because we have shown above that in this case the $MC\,\zeta_t$ belongs to the class of $AQTMC,$ the effective and numerically stable algorithms given in [16, 18, 34, 38] can be applied for solving this system.

Therefore, the problem of the computation of the stationary distribution of the analyzed tandem system with the fixed values of control parameters (n_1, n_2) can be considered solved.

4. Performance Measures

Having computed the vectors $\boldsymbol{\pi}_i,\ i \geq 0,$ we can calculate the values of the basic performance characteristics of the system.

The average number N_{buf} of customers in the buffers is computed as

$$N_{buf} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}.$$

The average number $N_{buf}^{(1)}$ of customers in the first buffer is computed as

$$N_{buf}^{(1)} = \sum_{i=1}^{\infty} \sum_{n=0}^{\min\{i-1,N\}} (i-n)\boldsymbol{\pi}(i,n)\mathbf{e}.$$

The average number $N_{buf}^{(2)}$ of customers in the second buffer is computed as

$$N_{buf}^{(2)} = \sum_{i=1}^{\infty} \sum_{n=1}^{\min\{i,N\}} n\boldsymbol{\pi}(i,n)\mathbf{e} = N_{buf} - N_{buf}^{(1)}.$$

The average number N_{serv} of busy servers is computed as

$$N_{serv} = K\sum_{i=1}^{\infty} \boldsymbol{\pi}_i\mathbf{e} + \sum_{k=1}^{K} k\boldsymbol{\pi}(0,0,k)\mathbf{e}.$$

Let

$$\boldsymbol{\gamma}_1 = (\underbrace{1,\ 1,\ldots,1}_{M^{(1)}},\underbrace{0,0,\ldots,0}_{M^{(2)}}),\ \boldsymbol{\gamma}_2 = (\underbrace{0,0,\ldots,0}_{M^{(1)}},\underbrace{1,\ 1,\ldots,1}_{M^{(2)}}).$$

Introduce the matrix $L_k(\boldsymbol{\gamma}_r)$ that is computed in the same way as the matrix $L_k(\mathbf{S}_0)$ defined above, but the column vector $\boldsymbol{\gamma}_r$ is used instead of the column vector $\mathbf{S}_0,\ k = \overline{1,K},\ r = 1,2.$

The average number $N_{serv}^{(1)}$ of busy servers at stage 1 is computed as

$$N_{serv}^{(1)} = \sum_{i=1}^{\infty} \sum_{n=0}^{\min\{i,N\}} \boldsymbol{\pi}(i,n,K)(I_W \otimes L_K(\boldsymbol{\gamma}_1))\mathbf{e} + \sum_{k=1}^{K} \boldsymbol{\pi}(0,0,k)(I_W \otimes L_k(\boldsymbol{\gamma}_1))\mathbf{e}.$$

The average number $N_{serv}^{(2)}$ of busy servers at stage 2 is computed as

$$N_{serv}^{(2)} = \sum_{i=1}^{\infty} \sum_{n=0}^{\min\{i,N\}} \boldsymbol{\pi}(i,n,K)(I_W \otimes L_K(\boldsymbol{\gamma}_2))\mathbf{e} + \sum_{k=1}^{K} \boldsymbol{\pi}(0,0,k)(I_W \otimes L_k(\boldsymbol{\gamma}_2))\mathbf{e}.$$

The average number N_{tandem} of customers in the tandem is computed as

$$N_{tandem} = N_{serv} + N_{buf}.$$

The average output rate from stage 1 of the tandem $\lambda_{out}^{(1)}$ is computed as

$$\lambda_{out}^{(1)} = \sum_{i=1}^{\infty} \sum_{n=0}^{\min\{i,N\}} \boldsymbol{\pi}(i,n,K)(I_W \otimes L_K^{(1)})\mathbf{e} + \sum_{k=1}^{K} \boldsymbol{\pi}(0,0,k)(I_W \otimes L_k^{(1)})\mathbf{e} = \frac{N_{serv}^{(1)}}{b_1^{(1)}}.$$

The average output rate from stage 2 of the tandem $\lambda_{out}^{(2)}$ is computed as

$$\lambda_{out}^{(2)} = \sum_{i=1}^{\infty} \sum_{n=0}^{\min\{i,N\}} \boldsymbol{\pi}(i,n,K)(I_W \otimes L_K^{(2)})\mathbf{e} + \sum_{k=1}^{K} \boldsymbol{\pi}(0,0,k)(I_W \otimes L_k^{(2)})\mathbf{e} = \frac{N_{serv}^{(2)}}{b_1^{(2)}}.$$

The probability $P_{loss}^{(1)}$ of loss of a customer at stage 1 of the tandem is computed as

$$P_{loss}^{(1)} = 1 - \frac{\lambda_{out}^{(1)}}{\lambda} = \frac{\alpha_1 N_{buf}^{(1)}}{\lambda}.$$

The probability $P_{loss}^{(2)}$ of loss of a customer at stage 2 of the tandem is computed as

$$P_{loss}^{(2)} = \frac{\alpha_2 N_{buf}^{(2)}}{\lambda} = \frac{(1-q)\lambda_{out}^{(1)} - \lambda_{out}^{(2)}}{\lambda}.$$

The probability $P_{loss-arr}^{(2)}$ of loss of a customer that arrives at stage 2 of the tandem at this stage is computed as

$$P_{loss-arr}^{(2)} = \frac{\alpha_2 N_{buf}^{(2)}}{(1-q)\lambda_{out}^{(1)}} = 1 - \frac{\lambda_{out}^{(2)}}{(1-q)\lambda_{out}^{(1)}}.$$

The probability P_{loss} of loss of an arbitrary customer in the tandem is computed as

$$P_{loss} = P_{loss}^{(1)} + P_{loss}^{(2)} = 1 - \frac{q\lambda_{out}^{(1)} + \lambda_{out}^{(2)}}{\lambda}.$$

The average intensity $\varphi^{(1-2)}$ of servers transition from stage 1 to stage 2 of the tandem is computed as

$$\varphi^{(1-2)} = \sum_{i=1}^{N} \boldsymbol{\pi}(i,i,K)(I_W \otimes L_K^{(1)})\mathbf{e} + (1-q)\sum_{i=N+1}^{\infty} \boldsymbol{\pi}(i,N,K)(I_W \otimes L_K^{(1)})\mathbf{e}$$
$$+ (1-q)\sum_{k=1}^{K} \boldsymbol{\pi}(0,0,k)(I_W \otimes L_k^{(1)})\mathbf{e}.$$

The average intensity $\varphi^{(2-1)}$ of servers transition from stage 2 to stage 1 of the tandem is computed as

$$\varphi^{(2-1)} = \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,0,K)(I_W \otimes L_K^{(2)})\mathbf{e} + \sum_{k=1}^{K} \boldsymbol{\pi}(0,0,k)(I_W \otimes L_k^{(2)})\mathbf{e}$$
$$+ \sum_{i=n_1+1}^{\infty} \sum_{n=1}^{\min\{i-n_1,n_2-1\}} \boldsymbol{\pi}(i,n,K)(I_W \otimes L_K^{(2)})\mathbf{e}.$$

Remark. We presented two different formulas for the computation of both loss probabilities $P_{loss}^{(1)}$ and $P_{loss}^{(2)}$. It is evident also that in the stationary operation of the tandem the intensity $\varphi^{(1-2)}$ is equal to the intensity $\varphi^{(2-1)}$. The fact of existence of different formulas for computation of loss probabilities and transition intensities was used to control the accuracy of the computation of the stationary distribution of the system states.

5. Numerical Example

In this section, we present a numerical example, the goals of which are to highlight the dependence of the main performance characteristics of the system on the control parameters n_1 and n_2 and illustrate the possibility of optimizing the quality of a tandem operation via the proper choice of these parameters.

Let us consider a tandem queue with $K = 5$ servers and $N = 10$ places in the buffer of stage 2. We assume that the process of customer arrivals is defined by the MAP determined by the matrices

$$D_0 = \begin{pmatrix} -15 & 0 \\ 0 & -1 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 14.5 & 0.5 \\ 0.05 & 0.95 \end{pmatrix}.$$

The vector $\boldsymbol{\theta}$ of the invariant distribution of the underlying process of the MAP is defined by $\boldsymbol{\theta} = (0.090909, 0.909091)$. The mean arrival rate is $\lambda = 2.27273$. The coefficient of correlation of the lengths of the successive inter-arrival times of this MAP is equal to 0.313281, the squared coefficient of variation of the inter-arrival times is 3.15978.

The service process of a customer at stage 1 of the tandem is defined by the vector $\boldsymbol{\beta}^{(1)} = (1,0)$ and matrix

$$S^{(1)} = \begin{pmatrix} -5 & 5 \\ 0 & -5 \end{pmatrix}.$$

The mean service time at stage 1 of the tandem is equal to $b_1^{(1)} = 0.4$.

After the service at stage 1, a customer leaves the system with the probability $q = 0.4$ and with the complimentary probability transits for service to stage 2.

The service time of a customer at stage 2 has a PH distribution with the vector $\boldsymbol{\beta}^{(2)} = (0.5, 0.5)$ and sub-generator

$$S^{(2)} = \begin{pmatrix} -5 & 0 \\ 0 & -1 \end{pmatrix}.$$

The mean service time at stage 2 of the tandem is equal to $b_1^{(2)} = 0.6$.

We assume that the impatience rates of customers in the buffers are equal to $\alpha_1 = 0.08$ and $\alpha_2 = 0.02$.

Let us vary the parameter n_1 over the interval $[1, 15]$ with step 1, and the parameter n_2 over the interval $[1, N + 1]$ also with step 1.

Figures 2 and 3 illustrate the dependence of the average numbers $N_{buf}^{(1)}$ and $N_{buf}^{(2)}$ of customers in the first and second buffers on the parameters n_1 and n_2.

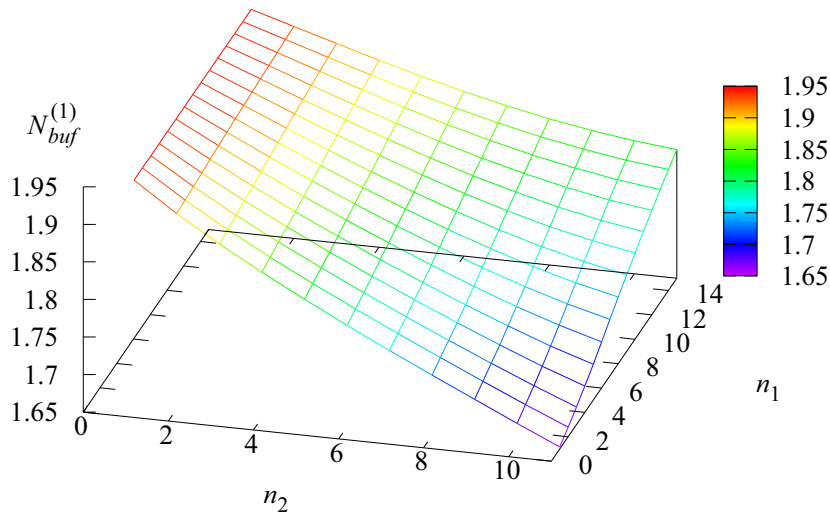

Figure 2. Dependence of the average number $N_{buf}^{(1)}$ of customers in the first buffer on the parameters n_1 and n_2

The value $N_{buf}^{(1)}$ has a maximum when n_1 is large and n_2 is small, and a minimum when n_2 is large and n_1 is small. This is clear because, in these cases, the rate of server redistribution to stage 1 is minimal and maximal, respectively. A smaller number of assigned servers at some stage implies a larger number of waiting customers at this stage. Correspondingly, the value $N_{buf}^{(2)}$ has a minimum when n_1 is large and n_2 is small and a maximum when n_2 is large and n_1 is small. This is explained by the higher rate of server redistribution to stage 2 in the first case and the lowest rate in the second case.

Figure 4 illustrates the dependence of the probability $P_{loss}^{(1)}$ of loss of a customer at stage 1 of the tandem on the parameters n_1 and n_2.

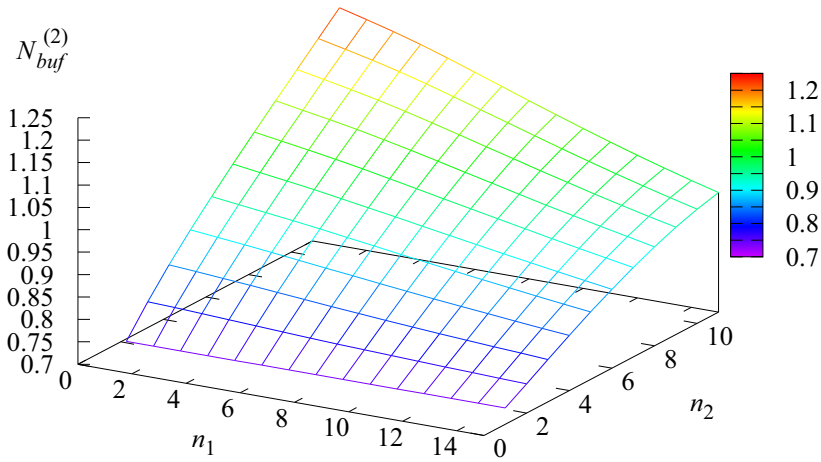

Figure 3. Dependence of the average number $N_{buf}^{(2)}$ of customers in the second buffer on the parameters n_1 and n_2

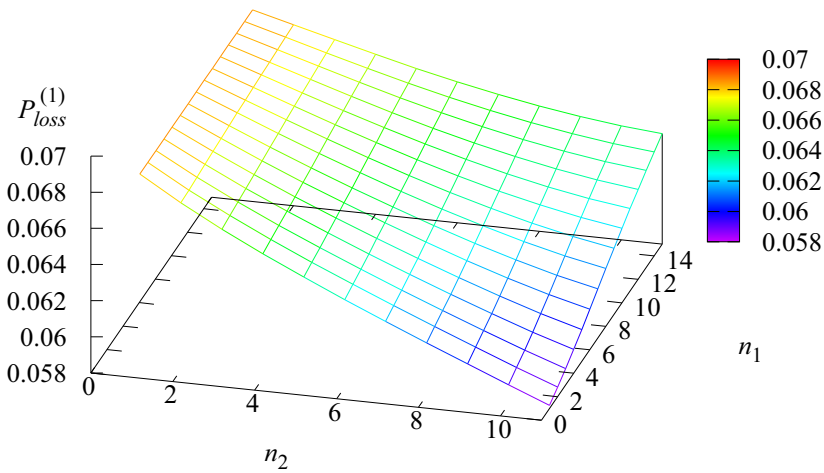

Figure 4. Dependence of the probability $P_{loss}^{(1)}$ of loss of a customer at stage 1 of the tandem on the parameters n_1 and n_2

Figures 5 and 6 illustrate the dependence of the loss probabilities $P_{loss}^{(2)}$ and $P_{loss-arr}^{(2)}$ of an arbitrary customer and a customer, which arrives at stage 2, during waiting at stage 2 of the tandem on the parameters n_1 and n_2.

The behavior of the surfaces given by Figures 4 and 5 matches the form of the surfaces given by Figures 2 and 3, correspondingly because the probabilities $P_{loss}^{(r)}$, $r = 1, 2$, linearly depend on the values $N_{buf}^{(2)}$, $r = 1, 2$, respectably. Figure 6 is similar to 5. The values given at Figure 6 are a bit larger because, with probability q, an arbitrary customer, which finishes service at stage 1, abandons stage 2 and is not lost there.

Figure 7 illustrates the dependence of the probability P_{loss} of loss of an arbitrary customer on the parameters n_1 and n_2. The form of the surface in this figure matched the surfaces drawn in Figures 4 and 5 because $P_{loss} = P_{loss}^{(1)} + P_{loss}^{(2)}$.

Figure 8 illustrates the dependence of the average intensity $\varphi^{(1-2)}$ of servers transition from stage 1 to stage 2 of the tandem on the parameters n_1 and n_2. Note that this intensity

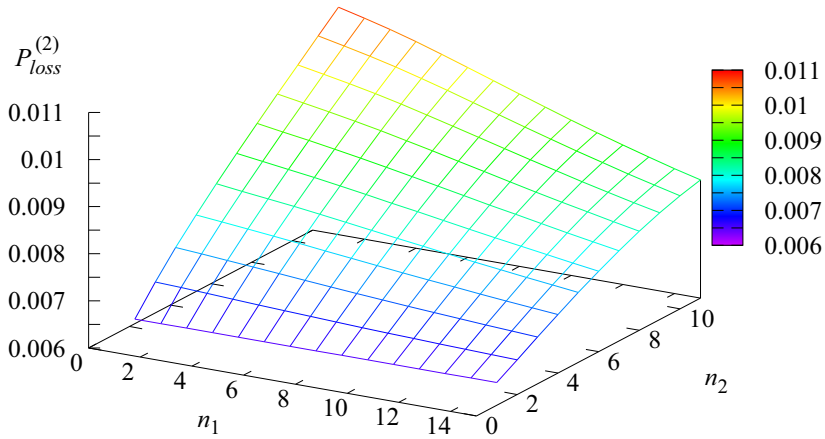

Figure 5. Dependence of the probability $P^{(2)}_{loss}$ of loss of a customer at stage 2 of the tandem on the parameters n_1 and n_2

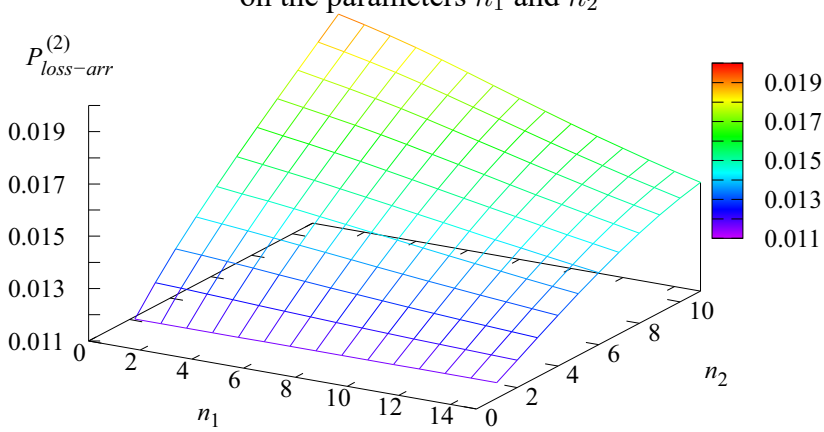

Figure 6. Dependence of the probability $P^{(2)}_{loss-arr}$ of loss of an arriving to stage 2 of the tandem customer on the parameters n_1 and n_2

is equal to the average intensity $\varphi^{(2-1)}$ of servers transition from stage 2 to stage 1.

Intensities $\varphi^{(1-2)}$ and $\varphi^{(2-1)}$ have a maximum under a small value of n_1 and a large value of n_2 because, for such values of n_1 and n_2, re-assigning of a server from stage 2 to stage 1 (and back) is more frequent.

Having highlighted the dependencies of various performance measures on the control parameters n_1 and of n_2, it is possible to formulate and solve different optimization problems.

To illustrate this, let us assume that the quality of the system's operation is described by the following economical cost criterion:

$$E = E(n_1, n_2) = a_1 q \lambda^{(1)}_{out} + a_2 \lambda^{(2)}_{out} - c_1 \lambda P^{(1)}_{loss} - c_2 \lambda P^{(2)}_{loss} - 2d\varphi^{(1-2)},$$

where a_1 is the profit gained by the system for service of a customer who needs only one stage service, a_2 is the profit gained by the system for service of a customer at both stages of the tandem, c_1 and c_2 are the charges paid by the system for loss of a customer at stages 1 and 2, respectively, and d is the charge paid by the system for one server transition between

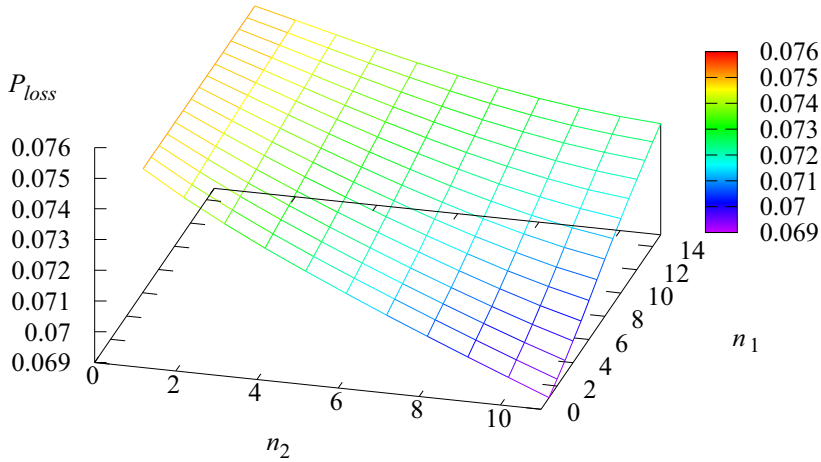

Figure 7. Dependence of the probability P_{loss} of loss of a customer on the parameters n_1 and n_2

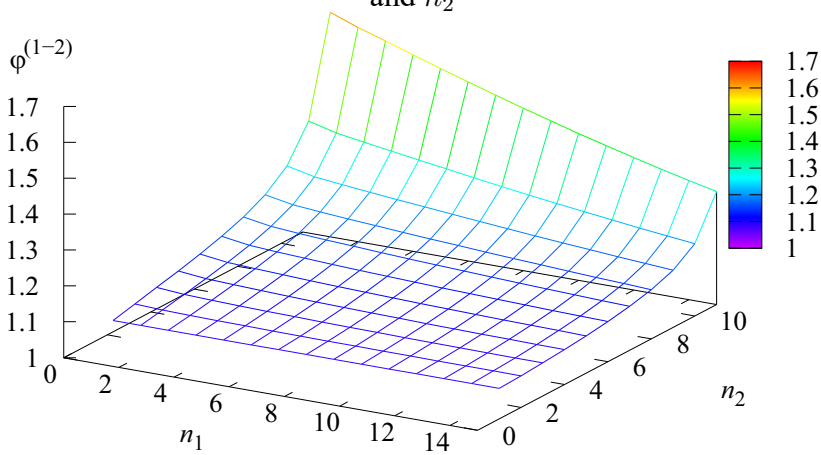

Figure 8. Dependence of the average intensity $\varphi^{(1-2)}$ of servers transitions from stage 1 to stage 2 of the tandem on the parameters n_1 and n_2

stages.

The value $E(n_1, n_2)$ has the meaning of the revenue of the system gained during a unit of time.

Let us fix the following values of the cost coefficients:

$$a_1 = 10,\ a_2 = 20,\ c_1 = 20,\ c_2 = 30,\ d = 0.5.$$

The dependence of the cost criterion E on the parameters n_1 and n_2, under the fixed total number of servers K and other parameters of the tandem, is illustrated in Figure 9.

The optimal (maximal) value of the cost criterion E is $E(2, 9) = 29.503$. Thus, an arbitrary server that finishes service at stage 2 has to move to stage 1 if the number of customers in the buffer of stage 2 is less than 9, while the number of customers in the buffer of stage 1 is greater than 2.

Having the possibility to compute the value of the cost criterion for any fixed set of the system parameters, various variants of analysis of the sensitivity of the optimal set of the

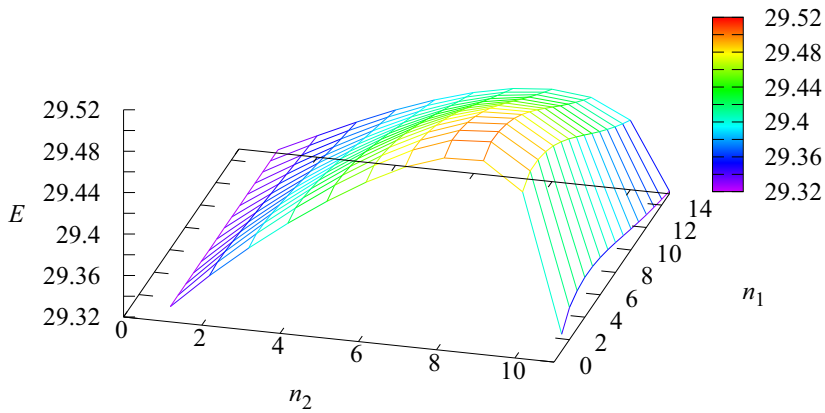

Figure 9. Dependence of the value of the cost criterion $E(n_1, n_2)$ on the parameters n_1 and n_2

thresholds n_1 and n_2 with respect to variation of these parameters and cost coefficients can be implemented. E.g., let us vary in the considered example the value d of the charge paid for one server transition between stages. It was written above that for $d = 0.5$ the optimal set of the thresholds is $(2, 9)$ and the value of the cost criterion is 29.503. For $d = 1$, the optimal set of the thresholds is $(4, 7)$ and the value of the cost criterion is 28.6311. For $d = 2$ the optimal set of the thresholds given above is $(9, 5)$ and the value of the cost criterion is 27.2931. When $d = 3$, the optimal set of the thresholds is $(11, 4)$ and the value of the cost criterion is 25.8838. When $d = 4$, the optimal set of the thresholds is $(13, 3)$, and the value of the cost criterion is 24.4882. The difference $n_1 - n_2$ between the optimal values of the thresholds varies in these examples as: $-7, -3, +4, +7, +10$.

6. Conclusion

We have analyzed the dual tandem queueing system, in which servers may transit between the stages when a server finishes service and the buffer of a stage, where this server was operating, is empty. When both buffers are empty, the server is assigned to stage 1. The server of stage 2 is reassigned at the service completion moment to stage 1 if the queue length at the buffer of stage 1 is greater or equal to the threshold n_1 while the queue length at the buffer of stage 2 is less than the threshold n_2. Aiming to formulate and solve an optimization problem consisting of the optimal choice of the thresholds, we fix an arbitrary pair of the thresholds n_1 and n_2 and analyze the stationary distribution of the multidimensional Markov process describing the behavior of a tandem. This is done via application of the notion of the generalized PH distribution, description of the service process in busy servers via the number of servers providing service at all phases of the underlying process of GPH, and use of the results for Quasi-Birth-and Death processes (if customers are absolutely patient at stage 1) and for $AQTMC$s (if customers are impatient at stage 1). Expressions for the main performance characteristics of the model are presented. A numerical example is given. In this example, 3D plots illustrate the dependence of some performance characteristics on the thresholds n_1 and n_2. The possibility of applying the obtained result to solving optimization

problems is illustrated.

The presented results, in particular the offered effective way of simultaneous tracking the underlying processes of PH type service at multiple servers floating between the stages, give a background for the investigation of various generalizations of the considered model.

As the possible variants of generalization that sound interesting from the point of view of possible practical applications, the following ones can be mentioned and implemented:

1) Buffers at both stages are finite. Or the first buffer is finite, while the second one is infinite.

2) The buffer at stage 1 is absent, and customers who met all servers at stage 1 busy retry for service later on.

3) Customers are heterogeneous and can have different priorities.

4) There is a cross-traffic of customers arriving directly at stage 2.

5) There are feedbacks of customers at both or one of the stages.

6) There are dedicated servers at both or some stages that, in contrast to the flexible servers, cannot move to another stage; see, e.g., [60].

7) Servers may be unreliable with MAP or $MMAP$ flows of breakdowns and PH type distributions of repair times.

8) Disasters can occur that temporarily destroy one of both stages of the tandem.

9) A vacation of stages is assumed when the corresponding buffer is empty.

10) Servers may be of several types, and the parameters of the service time of a customer may depend on the type of server and stage where it is used.

11) A semi-open queueing network can be considered instead of a dual tandem (see, e.g., [36]).

12) Other kinds of control strategies are applied. E.g., the following threshold-type strategy defined by the thresholds j_1 and j_2 can be considered. Let i_r be the current number of customers in the buffer of stage r, $r = 1, 2$. The server is reassigned from stage r to stage r' if, at a service completion epoch, the following inequality holds good: $i_{r'} \geq i_r + j_r$, $r = 1, 2, r' = 1, 2, r' \neq r$. This means that the servers transit between the stages when the difference between queue lengths is beyond a preassigned interval.

References

[1] Ahn, H. S., Duenyas, I., & Zhang, R. Q. (1999). Optimal stochastic scheduling of a two-stage tandem queue with parallel servers. *Advances in Applied Probability*, 31(4), 1095-1117.

[2] Ahn, H. S., Duenyas, I., & Lewis, M. E. (2002). Optimal control of a two-stage tandem queuing system with flexible servers. *Probability in the Engineering and Informational Sciences*, 16(4), 453-469.

[3] Andradóttir, S., Ayhan, H., & Down, D. G. (2001). Server assignment policies for maximizing the steady-state throughput of finite queueing systems. *Management Science*, 47(10), 1421-1439.

[4] Andradóttir, S., & Ayhan, H. (2005). Throughput maximization for tandem lines with two stations and flexible servers. *Operations Research*, 53(3), 516-531.

[5] Andradottir, S., Ayhan, H., & Down, D. G. (2007). Dynamic assignment of dedicated and flexible servers in tandem lines. *Probability in the Engineering and Informational Sciences*, 21(4), 497-538.

[6] Andradottir, S., Ayhan, H., & Kırkızlar, E. (2012). Flexible servers in tandem lines with setup costs. *Queueing Systems*, 70, 165-186.

[7] Andradottir, S., & Ayhan, H. (2023). Optimal server assignment in queues with flexible servers and abandonments. *Stochastic Systems*, 13(3), 360-376.

[8] Asmussen, S. (2003). Applied probability and queues. New York: Springer.

[9] Ayhan, H. (2022). Server assignment policies in queues with customer abandonments. *Queueing Systems*, 100(3-4), 393-395.

[10] Buchholz, P., Kriege, J., & Felko, I. (2014). Input modeling with phase-type distributions and Markov models: theory and applications. Springer.

[11] Chakravarthy, S. R. (2001). The batch Markovian arrival process: a review and future work. *Advances in probability theory and stochastic processes*, 1(1), 21-49.

[12] Chakravarthy, S. R. (2022). Introduction to Matrix-Analytic Methods in Queues 1: Analytical and Simulation Approach - Basics. ISTE Ltd, London and John Wiley and Sons, New York.

[13] Chakravarthy, S. R. (2022). Introduction to Matrix-Analytic Methods in Queues 2: Analytical and Simulation Approach - Queues and Simulation. ISTE Ltd, London and John Wiley and Sons, New York.

[14] De Kok, A. G., & Tijms, H. C. (1985). A queueing system with impatient customers. *Journal of Applied Probabilit*y, 22(3), 688-696.

[15] Dudin, A., Krishnamoorthy, A., Dudin, S., & Dudina, O. (2024). Queueing system with control by admission of retrial requests depending on the number of busy servers and state of the underlying process of Markov arrival process of primary requests. *Annals of Operations Research*, 335(1), 135-150.

[16] Dudin, A. N., Klimenok, V. I., & Vishnevsky, V. M. (2020). The theory of queuing systems with correlated flows. Springer Nature, Cham.

[17] Dudin, S., Kim, C., & Dudina, O. (2013). $MMAP|M|N$ queueing system with impatient heterogeneous customers as a model of a contact center. *Computers & Operations Research*, 40(7), 1790-1803.

[18] Dudin, S., Dudin, A., Kostyukova, O., & Dudina, O. (2020). Effective algorithm for computation of the stationary distribution of multi-dimensional level-dependent Markov chains with upper block-Hessenberg structure of the generator. *Journal of Computational and Applied Mathematics*, 366, 112425.

[19] Dudin, S. A., Dudin, A. N., Dudina, O. S., & Chakravarthy, S. R. (2023). Analysis of a tandem queuing system with blocking and group service in the second node. *International Journal of Systems Science: Operations & Logistics*, 10(1), 2235270.

[20] Dudin, S. A., Dudina, O. S., & Dudin, A. N. (2024). Analysis of tandem queue with multi-server stages and group service at the second stage. *Axioms*, 13(4), 214.

[21] Dudin, S., Dudin, A., Manzo, R., & Rarità, L. (2024). Analysis of Semi-open Queueing Network with Correlated Arrival Process and Multi-server Nodes. *Operations Research Forum*, 5(4), 99.

[22] Dudina O. S. (2024) Analytical modelling of systems with a ticket queue. *Journal of the Belarusian State University. Mathematics and Informatics*, 2, 40–53. Russian.

[23] Farrar, T. M. (1993). Optimal use of an extra server in a two station tandem queueing network. *IEEE Transactions on Automatic Control*, 38(8), 1296-1299

[24] Grassmann, W., & Tavakoli, J. (2005). Two-stations queueing networks with moving servers, blocking, and customer loss. *The Electronic Journal of Linear Algebra*, 13, 72-89.

[25] Grassmann, W. K., & Tavakoli, J. (2002). A tandem queuewith a movable server: An eigenvalue approach. *SIAM Journal on Matrix analysis and Applications*, 24(2), 465-474.

[26] Gonzalez Bernal, M., Lillo, R. E., & Ramirez-Cobo, P. (2024). Call center data modeling: a queueing science approach based on Markovian arrival process. *Quality Technology & Quantitative Management*, 1-28. DOI:10.1080/16843703.2024.2371715.

[27] Graham, A. (2018). Kronecker products and matrix calculus with applications. Courier Dover Publications.

[28] He, Q. M., & Alfa, A. S. (2018). Space reduction for a class of multidimensional Markov chains: A summary and some applications. *INFORMS Journal on Computing*, 30(1), 1-10.

[29] Horn, R. A., & Johnson, C. R. (2012). Matrix analysis. Cambridge university press.

[30] Iravani, S. M., Posner, M. J. M., & Buzacott, J. A. (1997). A two-stage tandem queue attended by a moving server with holding and switching costs. *Queueing systems*, 26, 203-228.

[31] Işsık, T., Andradóttir, S., & Ayhan, H. (2016). Optimal control of queueing systems with non-collaborating servers. *Queueing Systems*, 84, 79-110.

[32] Isık, T., Andradóttir, S., & Ayhan, H. (2022). Dynamic Control of Non-Collaborative Workers When Reassignment Is Costly. *Production and Operations Management*, 31(3), 1332-1352.

[33] Kim, C. S., Park, S. H., Dudin, A., Klimenok, V., & Tsarenkov, G. (2010). Investigation of the $BMAP/G/1 \rightarrow /PH/1/M$ tandem queue with retrials and losses. *Applied Mathematical Modelling*, 34(10), 2926-2940.

[34] Kim, C., Dudin, S., Taramin, O., & Baek, J. (2013). Queueing system MAP| PH| N| N+ R with impatient heterogeneous customers as a model of call center. *Applied Mathematical Modelling*, 37(3), 958-976.

[35] Kim, C., Dudin, A., Dudina, O., & Dudin, S. (2014). Tandem queueing system with infinite and finite intermediate buffers and generalized phase-type service time distribution. *European Journal of Operational Research*, 235(1), 170-179.

[36] Kim, J., Dudin, A., Dudin, S., & Kim, C. (2018). Analysis of a semi-open queueing network with Markovian arrival process. *Performance Evaluation*, 120, 1-19.

[37] Kim, C., Dudin, A., Dudin, S., & Dudina, O. (2021). Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users. *IEEE Access*, 9, 106933-106946.

[38] Klimenok, V., & Dudin, A. (2006). Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Systems*, 54, 245-259.

[39] Lucantoni, D. (1991). New results on the single server queue with a batch Markovian arrival process. *Commun.-Stat.-Stoch. Model.*, 7, 1-46.

[40] Melikov, A., Chakravarthy, S. R., & Aliyeva, S. (2023). A retrial queueing model with feedback. *Queueing Models and Service Management*, 6(1), 63-95.

[41] Neuts, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models*. The Johns Hopkins University Press, Baltimore.

[42] Nair, S. S. (1970). Semi-Markov analysis of two queues in series attended by a single server. *Bull. Soc. Math. Belgique*, 22, 355-367.

[43] Nair, S. S. (1973). Two queues in series attended by a single server. *Bull. Soc. Math*, 25, 160-176.

[44] Okamura, H., & Dohi, T. (2015). mapfit: An R-based tool for PH/MAP parameter estimation. In *Quantitative Evaluation of Systems: 12th International Conference, QEST 2015, Madrid, Spain, September 1-3, 2015, Proceedings 12* (pp. 105-112). Springer International Publishing.

[45] Pandelis, D. G. (2008). Optimal control of flexible servers in two tandem queues with operating costs. *Probability in the Engineering and Informational Sciences*, 22(1), 107-131.

[46] Papachristos, I., & Pandelis, D. G. (2018). Optimal dynamic allocation of collaborative servers in two station tandem systems. *IEEE Transactions on Automatic Control*, 64(4), 1640-1647.

[47] Ramaswami, V. (1985). Independent Markov processes in parallel. *Stochastic Models*, *1*(3), 419-432.

[48] Ramaswami, V., & Lucantoni, D. M. (1985). Algorithms for the multi-server queue with phase type service. *Stochastic Models*, 1(3), 393-417.

[49] Sindhu, S., & Krishnamoorthy, A. (2023). MAP/Ek/1 Queue with Working Vacation Providing Main Service Only in Normal Mode of Service. *Queueing Models and Service Management*, 6(2), 1-27.

[50] Sharma, S., Kumar, R., Soodan, B. S., & Singh, P. (2023). Queuing models with customers' impatience: a survey. *International Journal of Mathematics in Operational Research*, 26(4), 523-547.

[51] Stanford, R. E. (1990). On queues with impatience. *Advances in applied probability*, 22(3), 768-769.

[52] Takagi, H. (1988). Queuing analysis of polling models. *ACM Computing Surveys (CSUR)*, 20(1), 5-28.

[53] Telek, M., & Horváth, G. (2007). A minimal representation of Markov arrival processes and a moments matching method. *Performance Evaluation*, 64(9-12), 1153-1168.

[54] Vishnevskii, V. M., & Semenova, O. V. (2006). Mathematical methods to study the polling systems. *Automation and Remote Control*, 67, 173-220.

[55] Vishnevsky, V., & Semenova, O. (2021). Polling systems and their application to telecommunication networks. *Mathematics*, 9(2), 117.

[56] Vishnevskii, V. M., & Dudin, A. N. (2017). Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks. *Automation and Remote Control*, 78, 1361-1403.

[57] Wang, K., Li, N., & Jiang, Z. (2010). Queueing system with impatient customers: A review. In *Proceedings of 2010 IEEE international conference on service operations and logistics, and informatics* (pp. 82-87). IEEE.

[58] Wang, J., Abouee-Mehrizi, H., Baron, O., & Berman, O. (2019). Tandem queues with impatient customers. *Performance Evaluation*, 135, 102011.

[59] Zabala, L., Doncel, J., & Ferro, A. (2023). Optimality of a Network Monitoring Agent and Validation in a Real Probe. *Mathematics*, 11(3), 610.

[60] Zou, A. A., & Down, D. G. (2018). Asymptotically maximal throughput in tandem systems with flexible and dedicated servers. Asia-Pacific Journal of Operational Research, 35(05), 1850038 1-15.