DOI 10.1007/s10559-018-0005-4
Cybernetics and Systems Analysis, Vol. 54, No. 1, January, 2018

MODELS OF PERISHABLE QUEUEING-INVENTORY SYSTEMS WITH SERVER VACATIONS

V. S. Koroliuk,[1] A. Z. Melikov,[2] L. A. Ponomarenko,[3] and A. M. Rustamov[4] UDC 519.21:621.391

Abstract. *The model of perishable queueing-inventory system with server vacations is studied. Upon service completion, server takes vacation if there are no customers in the queue and it starts service at the end of the vacation if the number of customers in the system exceeds some threshold; otherwise, it takes new vacation. Exact and approximate methods are proposed to calculate the characteristics of the system.*

Keywords: *inventory control, perishable items, server vacation.*

INTRODUCTION

This paper continues the studies started in [1], where we proposed an efficient method to calculate the characteristics of perishable queueing-inventory system (PQIS) with server vacation and two-bin replenishment policy.

Here, we will analyze PQIS models that have the following features, unlike model [1]. First, server takes a vacation only if there are no customers in the system at the moment of the end of its operation; second, we will analyze here both models with a limited queue and models with unlimited queue of customers; and third, upon the end of vacation time, server returns to operating mode only when the number of customers in the system exceeds some threshold. Taking these features into account allows us to investigate one more wide class of real perishable queueing-inventory systems.

The main scientific result of this study is the development of the method of approximate calculation of the stationary distribution of three-dimensional high-dimensional Markov chain (MC). In this connection, we will specify some well-known approaches to the solution of this problem.

One of them uses the idea of aggregation of states of the chain. Note that using this idea for calculating the stationary distribution of an MC has a long history. In the well-known study [2] (see Ch. 6), it is shown that an aggregated process is Markov only if for the chosen partition of the phase space (PS) of the original chain the sum of transition probabilities from states of each class into any other class should be identical for all the states of the original class. In other words, the class of aggregated MC is rather narrow. Despite this, this idea was later used in some other studies. For example, the papers [3–5] suppose that aggregated states form classes such that transitions (inputs and outputs) between them are carried out only through one (fixed) state of each class of states. If such partition of the original PS is possible, then algorithms are proposed for calculation of stationary distribution of the chain.

The second approach is intended for calculation of stationary distribution of almost completely decomposable Markov chains. It is used, for example, in [6–8]. These studies suppose that original PS of the chain is divided into

[1]Institute of Mathematics, National Academy of Sciences of Ukraine, Kyiv, Ukraine, *vskorol@yahoo.com*. [2]Institute of Control Systems, Azerbaijan National Academy of Sciences, Baku, Azerbaijan, *agassi.melikov@rambler.ru*. [3]International Research and Training Center for Information Technologies and Systems, National Academy of Sciences of Ukraine and Ministry of Education and Science of Ukraine, Kyiv, Ukraine, *laponomarenko@ukr.net*. [4]Baku Engineering University, Baku, Azerbaijan, *anar.rustamov@gmail.com*. Translated from Kibernetika i Sistemnyi Analiz, No. 1, January–February, 2018, pp. 35–50. Original article submitted March 2, 2017.

classes that are almost not connected with each other, and with regard for this fact iterative algorithms are proposed to solve the problem. The third approach is intended for calculation of the stationary distribution of approximately aggregated Markov chains. It underlies the Takahashi method [9]. Note that the possibility of applying the last two approaches should be established for each specific chain. Checking whether the MC under study is actually almost completely decomposable or approximately aggregated is a challenge since it requires finding all of the eigenvalues of huge-dimensional matrices, which are often ill-conditioned.

It can be easily seen that practical application of the methods proposed in [3–9] demands ingenuity from the researcher in creating appropriate partition of the PS of the Markov chain under study. The complexity of this problem is especially aggravated in the analysis of multidimentional chains.

A detailed review of studies that use the above approaches can be found in [10–12].

The method we have used is based on the theory of phase integration of states of MC. The fundamentals of this theory were developed in the early 1970s in [13], where the idea was proposed (pp. 152–154) that a hierarchical version of the phase integration algorithm (PIA) should be developed to overcome the "curse of dimensionality" in calculating stationary distributions of a large-dimensional MC. This idea has not been implemented in practice for a long time, and only in the 2000s A. Z. Melikov and L. A. Ponomarenko under the guidance of Academician V. S. Koroliuk, the founder of this theory, started the development of such algorithms for two-dimensional Markov chains, which are successfully applied in specific models of teletraffic systems [14, 15]. Recently, this approach is applied by other authors as well, who also emphasize the high accuracy of the obtained results (see, for example, [16–18]).

PHYSICAL PQIS MODEL WITH SERVER VACATION

The system under study has a warehouse of limited size S and contains one server for service of customers. The input flow of customers is Poisson with intensity λ, and irrespective of the state of the server and system stock level, arriving customers are accepted in the queue and are serviced on FIFO basis. Service times are independent and equally distributed random variables (r.v.) with common exponential distribution function (d.f.) with mean value μ^{-1}, $\mu < \infty$.

For simplicity, we will assume that each customer requires a unit stock; in other words, after the end of service, warehouse inventory level decreases by one. Inventory level also decreases as a result of perishability, i.e., each stock unit becomes unsuitable irrespective of the other ones after random time, which has exponential d.f. with parameter γ, $\gamma > 0$. It is assumed that the stock that is already at distribution stage cannot perish.

Inventory replenishment in the system follows the two-bin policy, i.e., (s, S)-policy, where threshold s, $s< S$, is introduced, and if inventory level in the system is higher than this value, the system does not make replenishment orders. When the inventory level becomes equal to s, an order of size $S-s$ is made. To exclude repeated orders, it is assumed that $s< S/2$. Inventory is replenished with a time delay, i.e., lead time is a positive r.v. with exponential d.f., and its mean value depends on server status, i.e., if an order is placed during server vacation, average lead time is ν_0^{-1}; otherwise it is ν_1^{-1}.

In the paper, we consider models with finite and infinite queues. In a model with finite queue, we assume that arrived customer is lost with probability one if there are N, $1< N < \infty$, customers in the system at this instant of time. At the same time, in the model with infinite queue, any arrived customer is accepted in the queue.

We assume that if there is at least one customer in the system after the end of server operation and stock is available, the server immediately selects one of the customers for service. At the same time, the server takes a vacation irrespective of the inventory level if after the end of service and/or as a result of impatience of customers there are no customers in the system. Server vacation time is an r.v. with exponential d.f. with the mean value β^{-1}, $\beta< \infty$. After the end of vacation, the server turns to the operating mode irrespective of inventory level if there are no less than r customers in the queue, $r \geq 1$. If the stock level is positive at this instant of time, the server immediately begins service of customers; otherwise, the server takes vacation again, with the same vacation principle (multiple vacation).

Customers are impatient only while in the queue, i.e., a customer in the server does not leave the system without service. The degree of impatience of customers in the queue generally depends on server's status, i.e., in case of server vacation admissible values of queueing time are independent and equally distributed r.v. having exponential d.f. with the mean value α_0^{-1}; if the server is in operating mode, then these r.v. have exponential d.f. as well but with the mean value α_1^{-1}. Generally speaking, $\alpha_0 \neq \alpha_1$.

The problem is to find joint distribution of the inventory level of the system, the number of customers in it, and server's status. Its solution will allow calculating averaged characteristics of the system under study: average inventory level (S_{av}); average inventory damage rate (Γ_{av}); average order rate (RR); probability of server vacation (P_{vac}); probability of customer loss (PL); and average rate of customer loss from the queue because of their impatience (RL_{av}). Finding these characteristics will allow performing the value analysis for the system.

MATHEMATICAL MODEL OF PQIS WITH SERVER VACATION

Based on the form of distribution of the r.v. used in the model statement, we determine that the stochastic process being introduced is a three-dimensional Markov chain (3-D MC). Indeed, system operation at an arbitrary instant of time is described by a stochastic process whose state is specified by three-dimensional vector $\boldsymbol{n} = (n_1, n_2, \theta)$, where the first and second components are the current inventory level and number of customers in the system, respectively, and the third binary component means server's status, i.e.,

$$\theta = \begin{cases} 0 & \text{in case of server vacation,} \\ 1 & \text{in case of server operating mode.} \end{cases}$$

First, we will consider a model with finite queue, i.e., assume that the maximum number of customers in the system (including customers in the server) is N, $N < \infty$. In this case, we will denote the PS of the respective chain by E. It is defined as follows:

$$E = E_0 \cup E_1, \; E_0 \cap E_1 = \varnothing, \tag{1}$$

where $E_0 = \{\boldsymbol{n} : n_1 = 0, 1, \dots, S;\; n_2 = 0, 1, \dots, N; \theta = 0\}$ and $E_1 = \{\boldsymbol{n} : n_1 = 0, 1, \dots, S;\; n_2 = 1, 2, \dots, N; \theta = 1\}$.

From formula (1) it follows that geometrically PS of the model is specified by points with integer coordinates that belong to parallelepipeds of unit height and with bases being rectangles with side length N and S.

To determine this chain, it is necessary to find its generating matrix (Q-matrix). Elements of this matrix are the intensities of transitions between its states. Denote the intensity of transition from state $\boldsymbol{n}$ into state $\boldsymbol{n}'$ by $q(\boldsymbol{n}, \boldsymbol{n}')$, $\boldsymbol{n}, \boldsymbol{n}' \in S$.

As is seen from the description of the system under study, transitions between states of the PS E are related to the following events: (i) arrival of customers; (ii) end of the process of their service; (iii) abandoning the queue due to customers impatience; (iv) end of inventory lifetime; (v) stock replenishment; (vi) server vacation; and (vii) server's return from vacation.

To construct Q-matrix with regard for the mechanism of changes in inventory level and number of customers in the system, as well as server vacation and return scheme, it is expedient to distinguish the following cases in determining the initial state $\boldsymbol{n} \in E$: (i) $\boldsymbol{n} \in E_0$; (ii) $\boldsymbol{n} \in E_1$.

First, we will consider the case $\boldsymbol{n} \in E_0$. In this case, outputs from this state because of events of types (ii) and (vi) are impossible. The intensities of the output from this state under the other events are defined as follows. If some customer (type (i) events) arrives, it is queued provided that $n_2 < N$; in other words, transition from this state into state $\boldsymbol{n} + \boldsymbol{e}_2 \in E_0$ is executed. The intensity of such transition is λ. Hereinafter, $\boldsymbol{e}_i$ means the ith unit vector of the three-dimensional Euclidean space, $i = 1, 2, 3$.

If some customer leaves the queue without service (type (iii) events), then transition from this state into state $\boldsymbol{n} - \boldsymbol{e}_2 \in E_0$ takes place; the intensity of such transition is $n_2 \alpha_0$. Upon termination of inventory lifetime (type (iv) event) in the state $\boldsymbol{n} \in E_0$, $n_1 > 0$, transition into state $\boldsymbol{n} - \boldsymbol{e}_1 \in E_0$ is performed; the intensity of such transition is $n_1 \gamma$. At the instant of time of an order arrival from a higher-level warehouse (type (v) event) in state $\boldsymbol{n} \in E_0$, $n_1 \le s$, transition from this state into state $\boldsymbol{n} + (S - s)\boldsymbol{e}_1 \in E_0$ takes place; the intensity of such transition is ν_0. If in state $\boldsymbol{n} \in E_0$, $n_1 > 0$, $n_2 \ge r$, server returns from vacation (type (vii) event), then transition from this state into state $\boldsymbol{n} + \boldsymbol{e}_3 \in E_1$ takes place. Hence, for the cases of the initial state $\boldsymbol{n} \in E_0$, elements of the Q-matrix are defined as follows:

$$q(\boldsymbol{n},\boldsymbol{n}')=\begin{cases}\lambda & \text{if } \boldsymbol{n}'=\boldsymbol{n}+\boldsymbol{e}_2,\\ \beta & \text{if } n_2\ge r,\ \boldsymbol{n}'=\boldsymbol{n}+\boldsymbol{e}_3,\\ n_1\gamma & \text{if } \boldsymbol{n}'=\boldsymbol{n}-\boldsymbol{e}_1,\\ n_2\alpha_0 & \text{if } \boldsymbol{n}'=\boldsymbol{n}-\boldsymbol{e}_2,\\ \nu_0 & \text{if } n_1\le s,\ \boldsymbol{n}'=\boldsymbol{n}+(S-s)\boldsymbol{e}_1.\end{cases} \tag{2}$$

Now, let us consider the case $\boldsymbol{n}\in E_1$. In this case, outputs from this state because of type (vii) events are impossible, and transition intensities for the above type (i), (iii), and (v) events are defined similarly to relations (2). At the same time, since the inventory at customer distribution stage cannot perish, after the end of inventory lifetime (type (iv) event) in state $\boldsymbol{n}\in E_1$, $n_1>1$, transition into state $\boldsymbol{n}-\boldsymbol{e}_1\in E_1$ is carried out, with intensity $(n_1-1)\gamma$. Since it is impossible to service customers if $n_1=0$, to determine the next state after the end of service of a customer (type (ii) event) in states $\boldsymbol{n}\in E_1$, $n_1>0$, it is necessary to distinguish the following cases: (i) $n_2=1$; (ii) $n_2\ge 2$. In case (i), transition into state $\boldsymbol{n}-\boldsymbol{e}_1-\boldsymbol{e}_2-\boldsymbol{e}_3\in E_0$ is carried out, and in case (ii), the next state will be $\boldsymbol{n}-\boldsymbol{e}_1-\boldsymbol{e}_2\in E_1$. In both cases, the transition intensities are μ. Note that in type (i) cases server goes on vacation (type (vi) events); the server takes vacation also from the unique state $(0,1,1)\in E_1$ because of customer's impatience, i.e., transition into new state $(0,0,0)\in E_0$ is executed with intensity α_1. Hence, for cases of the initial state $\boldsymbol{n}\in E_1$, elements of the Q-matrix are defined as follows:

$$q(\boldsymbol{n},\boldsymbol{n}')=\begin{cases}\lambda & \text{if } \boldsymbol{n}'=\boldsymbol{n}+\boldsymbol{e}_2,\\ \mu & \text{if } n_2>1,\ \boldsymbol{n}'=\boldsymbol{n}-\boldsymbol{e}_1-\boldsymbol{e}_2\\ & \text{or } n_2=1,\ \boldsymbol{n}'=\boldsymbol{n}-\boldsymbol{e}_1-\boldsymbol{e}_2-\boldsymbol{e}_3,\\ (n_1-1)\gamma & \text{if } \boldsymbol{n}'=\boldsymbol{n}-\boldsymbol{e}_1,\\ (n_2-1)\alpha_1 & \text{if } n_1>0,\ n_2>1,\ \boldsymbol{n}'=\boldsymbol{n}-\boldsymbol{e}_2,\\ n_2\alpha_1 & \text{if } n_1=0,\ n_2>1,\ \boldsymbol{n}'=\boldsymbol{n}-\boldsymbol{e}_2\\ & \text{or } n_1=0,\ n_2=1,\ n'=\boldsymbol{n}-\boldsymbol{e}_2-\boldsymbol{e}_3,\\ \nu_1 & \text{if } n_1\ge s,\ \boldsymbol{n}'=\boldsymbol{n}+(S-s)\boldsymbol{e}_1.\end{cases} \tag{3}$$

Thus, the mathematical model of the system under study is a three-dimensional Markov chain with PS (1), and elements of its generating matrix are defined from relations (2) and (3).

CALCULATING THE CHARACTERISTICS OF PQIS MODEL WITH SERVER VACATION

Let $p(\boldsymbol{n})$ denote the stationary probability of state $\boldsymbol{n}\in E$ (stationary probabilities of states of the constructed finite-dimensional 3-D MC exist since this chain is irreducible). These quantities satisfy the system of equilibrium equations (SEE), which is composed from relations (2) and (3). We omit here the explicit form of this SEE because it is awkward and its composition is evident.

As we have already mentioned, the following quantities are characteristics of the PQIS under study: average warehouse inventory level; average damage rate; average order intensity, probability of server vacation; probability of customer loss; and average intensity of customer loss from the queue because of their impatience. These characteristics are defined in terms of state probabilities of the described 3-D MC with application of the approach proposed in [1].

Average warehouse stock level can be calculated as follows:

$$S_{av}=\sum_{k=1}^{S}k\sum_{\boldsymbol{n}\in E}p(\boldsymbol{n})\delta(n_1,k), \tag{4}$$

where $\delta(i,j)$ are Kronecker deltas.

Since inventory at distribution stage cannot perish, we get:

$$\Gamma_{av}=\gamma\left(\sum_{k=1}^{S}k\sum_{\boldsymbol{n}\in E_0}p(\boldsymbol{n})\delta(n_1,k)+\sum_{k=2}^{S}(k-1)\sum_{\boldsymbol{n}\in E_1}p(\boldsymbol{n})\delta(n_1,k)\right). \tag{5}$$

The probability of server vacation is calculated as follows:

$$P_{vac} = \sum_{\boldsymbol{n}\in E_0} p(\boldsymbol{n}). \tag{6}$$

Using the total probability formula, we find that the customer loss probability (PL) is

$$PL = P_{vac} PL_v + (1 - P_{vac}) PL_s, \tag{7}$$

where PL_v is the probability of customer loss during server vacation and PL_s is the probability of customer loss when the server is in operating mode. These quantities consist of two terms: probabilities of customer loss during its arrival because of buffer overflow and probability of customer loss from the queue because of its impatience. In other words, we have:

$$PL_v = \sum_{\boldsymbol{n}\in E_0} p(\boldsymbol{n})(\delta(n_2, N) + (1-\delta(n_2, N))P_0(n_1, n_2)), \tag{8}$$

where $P_0(n_1, n_2)$ is the probability that in state $(n_1, n_2, 0)$ a customer is lost because of impatience;

$$PL_s = \sum_{\boldsymbol{n}\in E_1} p(\boldsymbol{n})(\delta(n_2, N) + (1-\delta(n_2, N))P_1(n_1, n_2)), \tag{9}$$

where $P_1(n_1, n_2)$ is the probability that in state $(n_1, n_2, 1)$ a customer is lost because of impatience.

The quantities $P_k(n_1, n_2)$, $k = 0, 1$, in formulas (8) and (9) are calculated as follows:

$$P_0(n_1, n_2) = \frac{n_2\alpha_0}{n_2\alpha_0 + \lambda I(n_2 < N) + n_1\gamma},$$

$$P_1(n_1, n_2) = \begin{cases} \dfrac{n_2\alpha_1}{n_2\alpha_1 + \lambda} & \text{if } n_1 = 0,\ 1 < n_2 < N, \\[2ex] \dfrac{(n_2-1)\alpha_1}{(n_2-1)\alpha_1 + \lambda + (n_1-1)\gamma I(n_1 > 1) + \mu} & \text{if } n_1 > 0,\ 1 < n_2 < N, \end{cases}$$

where $I(A)$ is an indicator function of event A.

Average rate of customer loss from the queue because of impatience is defined as follows:

$$RL_{av} = \alpha_0 L_v + \alpha_1 L_s, \tag{10}$$

where L_v and L_s are average number of customers in the queue during server vacation and in operating mode, respectively, i.e.,

$$L_v = \sum_{k=1}^{N} k \sum_{\boldsymbol{n}\in E_0} p(\boldsymbol{n})\delta(n_2, k);\ L_s = \sum_{k=2}^{N} (k-1) \sum_{\boldsymbol{n}\in E_1} p(\boldsymbol{n})\delta(n_2, k).$$

Unlike [1], we introduce here a new system's characteristics: average intensity of orders for inventory replenishment (RR). As is described above, orders are formed in the following cases: (i) if the system is in states of type $\boldsymbol{n} \in E_0$, $n_1 = s+1$, and inventory perishes, and (ii) if the system is in states of types $\boldsymbol{n} \in E_1$, $n_1 = s+1$, and inventory level decreases as a result of perish or issue on request. Hence, this characteristics is defined as follows:

$$RR = \gamma(s+1) \sum_{\boldsymbol{n}\in E_0} p(\boldsymbol{n})\delta(n_1, s+1) + (\mu + s\gamma) \sum_{\boldsymbol{n}\in E_1} p(\boldsymbol{n})\delta(n_1, s+1). \tag{11}$$

Due to the complicated structure of the Q-matrix, applying well-known numerical methods to solve the SEE of the model under study for high-dimensional PS (1) involves huge computing difficulties. In this connection, below we will propose an alternative solution of this problem, based on PIA [1]. Let us briefly remind the special features of application of this algorithm to our system.

At the first hierarchy level, the following aggregation function is introduced:

$$U(\boldsymbol{n}) = \langle \theta \rangle \text{ if } \boldsymbol{n} \in E_\theta, \tag{12}$$

where $\langle \theta \rangle$ is a lumped state that includes all the states from class E_θ, $\theta = 0, 1$. Denote $\Omega = \{\langle \theta \rangle : \theta = 0, 1\}$. Then according to PIA the state probabilities of the original model are defined as follows:

$$p(\boldsymbol{n}) \approx \rho_\theta(n_1, n_2)\pi(\langle \theta \rangle), \tag{13}$$

where $\rho_\theta(n_1, n_2)$ is the probability of state (n_1, n_2) inside the split model with the state space E_θ and $\pi(\langle \theta \rangle)$ is the probability of lumped state $\langle \theta \rangle \in \Omega$.

At the second hierarchy level, PIA is applied again to each class E_k, $k = 0, 1$. For correct application of the method, customers' arrival intensity is assumed to significantly exceed the inventory perish rate, i.e., $\lambda >> \gamma$. This assumption corresponds to the operation mode of real PQIS [1].

Under this assumption, the following decomposition is considered in the state space E_0:

$$E_0 = \bigcup_{i=1}^{S} E_0^i, \quad E_0^i \bigcap E_0^j = \varnothing \text{ if } i \neq j, \tag{14}$$

where $E_0^i = \{(n_1, n_2) \in E_0 : n_1 = i\}$, $i = 1, \dots, S$, i.e., the class of states E_0^i includes the states from E_0 at which the inventory level is equal to i irrespective of the number of customers in the queue.

At this hierarchy level, on the basis of decomposition (14) the following integration function is determined in the state space E_0:

$$U_0((n_1, n_2)) = \langle n_1 \rangle \text{ if } (n_1, n_2) \in E_0^{n_1}, \tag{15}$$

where $\langle n_1 \rangle$ is the lumped state that includes all the states from class $E_0^{n_1}$. Denote $\Omega_0 = \{\langle n_1 \rangle : i = 0, 1, \dots, S\}$.

According to PIA, we have

$$\rho_0(n_1, n_2) \approx \rho_0^{n_1}(n_2)\pi_0(\langle n_1 \rangle), \tag{16}$$

where $\rho_0^{n_1}(n_2)$ is the probability of state (n_1, n_2) inside the split model with the state space $E_0^{n_1}$ and $\pi_0(\langle n_1 \rangle)$ is the probability of lumped state $\langle n_1 \rangle \in \Omega_0$.

Since the first component in the class of states E_0^i is constant (it is equal to i), the microstate $(i, n_2) \in E_0^i$ is specified by the second component n_2, $n_2 = 0, 1, \dots, N$. Denote the intensities of transitions between states n_2 and n_2' of the split model with PS E_0^i by $q_0(n_2, n_2')$. These parameters are defined as follows:

$$q_0(n_2, n_2') = \begin{cases} \lambda & \text{if } n_2' = n_2 + 1, \\ n_2\alpha_0 & \text{if } n_2' = n_2 - 1, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

From relations (17) it follows that the probabilities of states in all the split models with PS E_0^i, $i = 0, 1, \dots, S$, coincide with the probabilities of states of the Erlang model $M / M / N / 0$ with loading λ / α_0 erl., i.e.,

$$\rho_0^i(n_2) = \frac{\sigma_0(n_2)}{\sum_{j=0}^{N} \sigma_0(j)}, \quad n_2 = 0, 1, \dots, N. \tag{18}$$

Hereinafter, the following notation is introduced: $\sigma_k(j) = \dfrac{(\lambda / \alpha_k)^j}{j!}$, $k = 0, 1$.

Remark 1. Since the probabilities $\rho_0^i(n_2)$ do not depend on parameter i, in what follows we will omit the superscript in these quantities.

From (2) and (18) we obtain that the intensities of transitions $q_0(\langle i\rangle, \langle j\rangle)$ from the lumped state $\langle i\rangle \in \Omega_0$ into another lumped state $\langle j\rangle \in \Omega_0$ are defined as follows:

$$q_0(\langle i\rangle, \langle j\rangle) = \begin{cases} i\gamma & \text{if } j = i-1, \\ \nu_0 & \text{if } i \le s,\ j = i+S-s, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Thus, we obtain the following expressions from relations (19) and the results from [1] to calculate the probabilities of lumped states $\pi_0(\langle n_1\rangle), \langle n_1\rangle \in \Omega_0$:

$$\pi_0(\langle n_1\rangle) = \begin{cases} a_{n_1}(0)\pi_0(\langle s+1\rangle) & \text{if } 0 \le n_1 \le s, \\ b_{n_1}(0)\pi_0(\langle s+1\rangle) & \text{if } s+1 \le n_1 \le S-s, \\ c_{n_1}(0)\pi_0(\langle s+1\rangle) & \text{if } S-s+1 \le n_1 \le S, \end{cases} \tag{20}$$

where $a_{n_1}(0) = \prod_{i=n_1+1}^{s+1} \frac{i\gamma}{\nu_0 + (i-1)\gamma}$; $b_{n_1}(0) = \frac{s+1}{n_1}$; and $c_{n_1}(0) = \frac{\nu_0}{n_1\gamma} \sum_{i=n_1-S+s}^{s} a_i(0)$.

The probability $\pi_0(\langle s+1\rangle)$ can be calculated from the normalization condition, i.e.,

$$\pi_0(\langle s+1\rangle) = \left(\sum_{i=0}^{s} a_i(0) + \sum_{i=s+1}^{S-s} b_i(0) + \sum_{i=S-s+1}^{S} c_i(0) \right)^{-1}.$$

Let us analyze the split model with PS E_1. Similarly to (14), we consider the decomposition

$$E_1 = \bigcup_{i=0}^{S} E_1^i,\ E_1^i \bigcap E_1^j = \varnothing,\ i \ne j, \tag{21}$$

where $E_1^i = \{(n_1, n_2) \in E_0 : n_1 = 0\}$, $i = 0, 1, \dots, S$. Similarly to (14), on the basis of decomposition (21) in the state space E_1 the following integration function is defined here:

$$U_1((n_1, n_2)) = \langle n_1\rangle \text{ if } (n_1, n_2) \in E_1^{n_1}, \tag{22}$$

where $\langle n_1\rangle$ is the lumped state that includes all the states from class $E_1^{n_1}$. Denote $\Omega_1 = \{\langle n_1\rangle : i = 0, 1, \dots, S\}$.

Note that unlike the split model with PS E_0, the state probabilities in the split model with PS E_1^0 and split models E_1^i, $i = 1, \dots, S$, do not coincide here. The state probabilities of the split model with PS E_1^0 can be calculated as follows:

$$\rho_1^0(n_2) = \frac{\sigma_1(n_2-1)}{n_2 \sum_{j=1}^{N} \frac{\sigma_1(j-1)}{j}},\ n_2 = 1, 2, \dots, N. \tag{23}$$

The state probabilities inside all the split models with PS E_1^i, $i = 1, \dots, S$, can be calculated similarly (do not depend on the subscript i, $i = 1, \dots, S$), i.e.,

$$\rho_1^i(n_2) = \frac{\sigma_1(n_2-1)}{\sum_{j=0}^{N-1} \sigma_1(j)},\ n_2 = 1, 2, \dots, N. \tag{24}$$

Taking into account (23) and (24), we find that the intensities of transitions $q_1(\langle i\rangle, \langle j\rangle)$ from the lumped state $\langle i\rangle \in \Omega_1$ into the lumped state $\langle j\rangle \in \Omega_1$ can be calculated as follows in this case:

$$q_1(\langle i\rangle, \langle j\rangle) = \begin{cases} (i-1)\gamma + \mu(1-\rho_1^1(1)) & \text{if } j = i-1, \\ \nu_1 & \text{if } i \le s,\ j = i+S-s, \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

Then we obtain the following expressions from relations (25) and the results of [1] to calculate the probabilities of lumped states $\pi_1(\langle n_1\rangle), \langle n_1\rangle \in \Omega_1$:

$$\pi_1(\langle n_1\rangle)=\begin{cases} a_{n_1}(1)\pi_1(\langle s+1\rangle) & \text{if } 0\le n_1\le s, \\ b_{n_1}(1)\pi_1(\langle s+1\rangle) & \text{if } s+1\le n_1\le S-s, \\ c_{n_1}(1)\pi_1(\langle s+1\rangle) & \text{if } S-s+1\le n_1\le S, \end{cases} \tag{26}$$

where

$$a_{n_1}(1)=\prod_{i=n_1+1}^{s+1}\frac{\Lambda_i}{\nu_1+\Lambda_{i-1}},\ b_{n_1}(1)=\frac{\Lambda_{s+1}}{\Lambda_{n_1}},\ c_{n_1}(1)=\frac{\nu_1}{\Lambda_{n_1}}\sum_{i=n_1-S+s}^{s}a_i(1),$$

$$\Lambda_i=\begin{cases} 0, & i=0, \\ (i-1)\gamma+\mu(1-\rho_1^1(1)), & 1\le i\le S. \end{cases}$$

The probability $\pi_1(\langle s+1\rangle)$ can be calculated from the respective normalization condition, i.e.,

$$\pi_1(\langle s+1\rangle)=\left(\sum_{i=0}^{s}a_i(1)+\sum_{i=s+1}^{S-s}b_i(1)+\sum_{i=S-s+1}^{S}c_i(1)\right)^{-1}.$$

Denote the intensities of transitions between classes E_k, $k=0,1$, by $q(\langle k\rangle,\langle k'\rangle)$. These parameters are defined as follows:

$$q(\langle k\rangle,\langle k'\rangle)=\begin{cases} \beta\sum_{i=r}^{N}\rho_0(i) & \text{if } k=0,\ k'=1, \\ \mu\rho_1^1(1)(1-\pi_1(\langle 0\rangle))+\alpha_1\rho_1^0(1)\pi_1(\langle 0\rangle) & \text{if } k=1,\ k'=0. \end{cases} \tag{27}$$

The required probabilities $\pi(\langle k\rangle)$, $\langle k\rangle\in\Omega$, can be easily calculated from relations (27), i.e.,

$$\pi(\langle 0\rangle)=\frac{q(\langle 1\rangle,\langle 0\rangle)}{q(\langle 0\rangle,\langle 1\rangle)+q(\langle 1\rangle,\langle 0\rangle)},\ \pi(\langle 1\rangle)=1-\pi(\langle 0\rangle).$$

The stationary probabilities of states of the original chain are defined as follows [1]:

$$p(n_1,n_2,\theta)\approx\rho_\theta(n_2)\pi_\theta(\langle n_1\rangle)\pi(\langle\theta\rangle). \tag{28}$$

Finally, after certain transformations we obtain the following formulas for approximate calculation of the characteristics of the PQIS model under study:

$$S_{av}\approx\sum_{i=0}^{1}\pi(\langle i\rangle)\sum_{k=1}^{S}k\pi_i(\langle k\rangle); \tag{29}$$

$$\Gamma_{av}\approx\gamma\sum_{i=0}^{1}\pi(\langle i\rangle)\sum_{k=i+1}^{S}(k-i)\pi_i(\langle k\rangle); \tag{30}$$

$$P_{vac}\approx\pi(\langle 0\rangle); \tag{31}$$

$$PL_v\approx\pi(\langle 0\rangle)\left(\rho_0(N)+\sum_{k=0}^{S}\pi_0(\langle k\rangle)\sum_{i=1}^{N-1}\rho_0(i)P_0(k,i)\right); \tag{32}$$

$$PL_s\approx\pi(\langle 1\rangle)\left(\rho_1^0(N)\pi_1(\langle 0\rangle)+\rho_1^1(N)(1-\pi_1(\langle 0\rangle))+\sum_{k=0}^{S}\pi_1(\langle k\rangle)\sum_{i=2}^{N-1}\rho_1^k(i)P_1(k,i)\right); \tag{33}$$

$$RR \approx \gamma(s+1)\pi(\langle 0\rangle)\pi_0(\langle s+1\rangle)+(\mu+s\gamma)\pi(\langle 1\rangle)\pi_1(\langle s+1\rangle); \quad (34)$$

$$RL_{av} \approx \alpha_0 L_v + \alpha_1 L_s, \quad (35)$$

where

$$L_v = \pi(\langle 0\rangle)\sum_{k=1}^{N} k\rho_0(k);$$

$$L_s = \pi(\langle 1\rangle)\left(\sum_{k=2}^{N}(k-1)\rho_1^1(k)(1-\pi_1(\langle 0\rangle))+\sum_{k=1}^{N}k\rho_1^0(k)\pi_1(\langle 0\rangle)\right).$$

Let us consider the model with infinite queue of spending customers, i.e., assume that $N=\infty$. In this case, PS of the model E is also specified by means of (1), but the classes of states E_0 and E_1 are infinite-dimensional sets, i.e., $E_0=\{\boldsymbol{n}: n_1=0,1,\dots,S;\, n_2=0,1,\dots;\, \theta=0\}$ and $E_1=\{\boldsymbol{n}: n_1=0,1,\dots,S;\, n_2=1,2,\dots;\, \theta=1\}$.

Elements of the Q-matrix of this model are defined similarly to relations (2). As to model's characteristics, losses of customers because of buffer overflow are impossible here; however, losses from the queue because of impatience of customers are possible. Other characteristics are defined from formulas (4)–(8), (10), and (11), but it is necessary to take into account that $N=\infty$.

To calculate the approximate values of stationary probabilities of states of this model, it is also possible to use the method described above. Since it is presented in detail for a model with finite PS, below we will only present the final form of the necessary formulas.

In this case, the probabilities of states inside all the split models with PS E_0^i, $i=0,1,\dots,S$, can be calculated as respective probabilities of states of the model $M/M/\infty$ with loading λ/α_0 erl., i.e.,

$$\rho_0(i)=\sigma_0(i)e^{-\lambda/\alpha_0},\; i=0,1,\dots. \quad (36)$$

The probabilities of lumped states $\pi_0(\langle n_1\rangle)$, $\langle n_1\rangle\in\Omega_0$, in this case can also be calculated by means of relations (20).

After certain transformations we obtain that the probabilities of states of the split model with PS E_1^0 and E_1^i, $i=1,\dots,S$, can be calculated as follows in this case:

$$\rho_1^0(j)=\frac{\sigma_1(j-1)}{j}\frac{\lambda/\alpha_1}{e^{\lambda/\alpha_1}-1},\; j=1,2,\dots, \quad (37)$$

$$\rho_1^i(j)=\sigma_1(j-1)e^{-\lambda/\alpha_1},\; j=1,2,\dots. \quad (38)$$

Remark 2. When deriving formula (37), we took into account that $\sum_{j=1}^{\infty}\frac{x^{j-1}}{j!}=\frac{e^x-1}{x}$.

The probabilities of lumped states $\pi_1(\langle n_1\rangle)$, $\langle n_1\rangle\in\Omega_1$, in this case can also be calculated by means of relations (26), but it is necessary to take into account that $\rho_1^1(1)$ is defined from formula (38).

When calculating the probability of lumped states $\pi(\langle k\rangle)$, $\langle k\rangle\in\Omega$, it is necessary to take into account that $q(\langle 0\rangle,\langle 1\rangle)$ (see formulas (27)) is defined as follows in this case:

$$q(\langle 0\rangle,\langle 1\rangle)=\beta\left(1-\sum_{i=0}^{r-1}\rho_0(i)\right).$$

Average warehouse inventory level, average perishability rate in the system, probability of server vacation, and average intensity of inventory replenishment orders in this model can be calculated by formulas (29)–(32), respectively. In this model, the probabilities of customer loss during server vacation and its operating mode are defined as follows:

$$PL_v \approx \pi(\langle 0\rangle) e^{-\lambda/\alpha_0} \sum_{k=0}^{S} \pi_0(\langle k\rangle) \sum_{i=1}^{\infty} \sigma_0(i) P_0(k,i); \quad (39)$$

$$PL_s \approx \pi(\langle 1\rangle)\left(\frac{\lambda/\alpha_1}{e^{\lambda/\alpha_1}-1} \pi_1(\langle 0\rangle) \sum_{i=2}^{\infty} \frac{\sigma_1(i-1)}{i} P_1(0,i) + e^{-\lambda/\alpha_1} \sum_{k=1}^{S} \pi_1(\langle k\rangle) \sum_{i=2}^{\infty} \sigma_1(i-1) P_1(k,i) \right). \quad (40)$$

After certain transformations we obtain that average number of customers in the queue during server vacation and its operating mode are calculated from the following formulas:

$$L_v \approx \pi(\langle 0\rangle) \frac{\lambda}{\alpha_0}; \; L_s \approx \pi(\langle 1\rangle) \frac{\lambda}{\alpha_1} e^{\lambda/\alpha_1} \left(1 - \pi_1(\langle 0\rangle)\left(1 - \frac{1}{e^{\lambda/\alpha_1}-1}\right)\right). \quad (41)$$

Formulas (35) and (41) are used to calculate average intensity of the loss of customers from the queue because of their impatience for different server's states.

NUMERICAL RESULTS

The developed algorithms allow us to analyze the behavior of the characteristics of systems under study with respect to variations in their loading and structural parameters. Because of the space restriction, we will present the results for a system with limited queue. To make the presentation more specific, we will analyze the behavior of the characteristics of the systems under study with respect to variation in the critical inventory level (s) and threshold value of the queue length at which the server returning from vacation returns to operating mode (r).

First, let us consider the dependences of system's characteristics on parameter s. We choose system's input parameters as follows [19]:

$$S = 50, \; N = 200, \; \lambda = 15, \; \mu = 4, \; \nu_0 = 4, \; \nu_1 = 6, \; \beta = 1.3, \; \gamma = 1, \; \alpha_0 = 2, \; \alpha_1 = 1.$$

Note that input parameters satisfy the above conditions for correct application of the developed asymptotic method, i.e., $\lambda >> \max\{\beta, \gamma, \mu\}$.

Here, we present the results for two different values of parameter r, i.e., in the first series of experiments we accept that $r = 5$ and in the second that $r = 20$. The results are presented in Figs. 1–5, where the notation × and ○ correspond to the first and second series of experiments, respectively.

As one would expect, function S_{av} is nondecreasing with respect to the increase in the critical inventory level (see Fig. 1). At the same time, the arrangement of graphs of function S_{av} for different values of parameter r (see Fig. 1) is surprising here at first sight, i.e., from these graphs we can see that the less this parameter, the greater the value of the function. In other words, it was to be expected that the earlier the server returns to operating mode after vacation (i.e., the server turns into operating mode for small values of the length of customers queue), the more intensively the customers queue is serviced and hence the less the average inventory level should be. However, the opposite pattern is observed here. This is because for the selected input data (which underlie all the subsequent reasoning) the probability of server vacation (P_{vac}) for $r = 20$ is almost equal to one (see Fig. 2). This is because in this case $L_s \approx 0$, i.e., for $r = 20$ the server is almost always on vacation, i.e., does not service customers; therefore, average inventory level in the second series of experiments appears smaller than in the first series of experiments (in this series of experiments $L_s \approx 1$).

As the critical inventory level increases, function Γ_{av} also grows (see Fig. 3) and the values of this function in the first series of experiments appear a bit larger than in the second series of experiments (sometimes almost coincide with them) since average inventory level in the first series of experiments appears larger than in the second series of experiments (see Fig. 1). These facts completely correspond to our expectations.

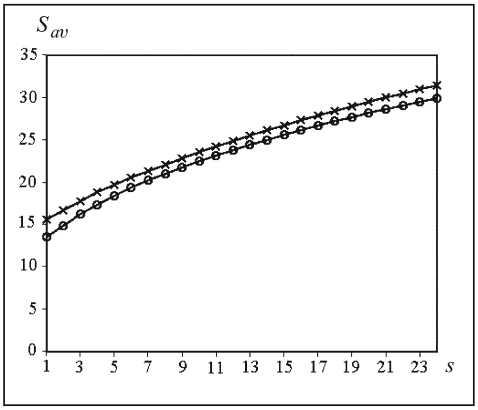

Fig. 1. Dependence of S_{av} on parameter *s*.

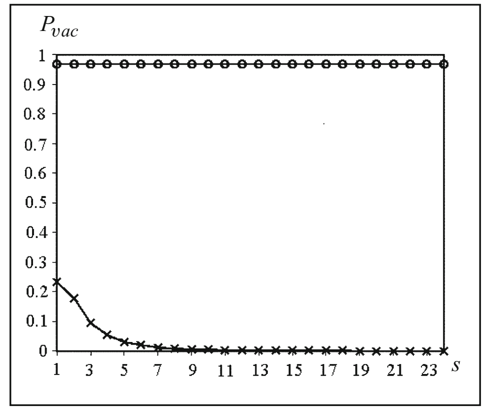

Fig. 2. Dependence of P_{vac} on parameter *s*.

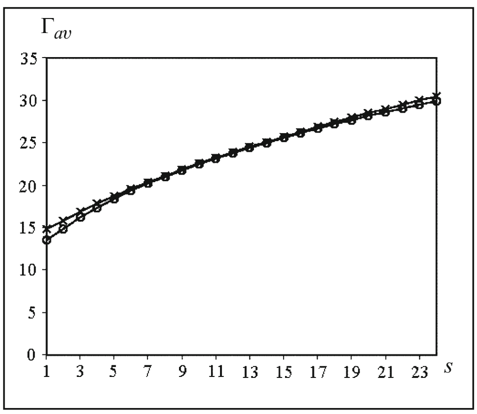

Fig. 3. Dependence of Γ_{av} on parameter *s*.

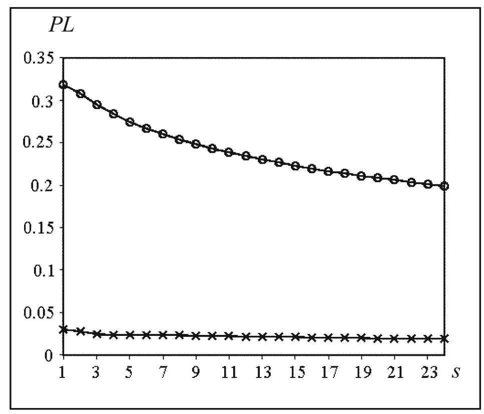

Fig. 4. Dependence of *PL* on parameter *s*.

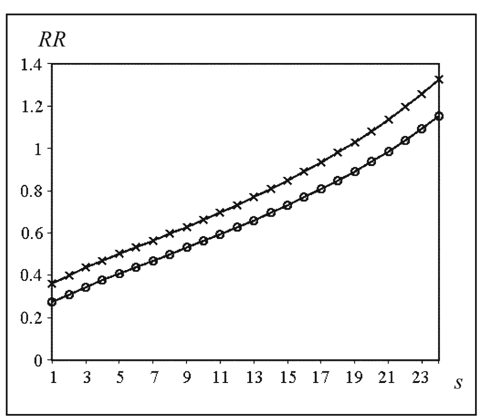

Fig. 5. Dependence of *RR* on parameter *s*.

Function *PL* is decreasing with respect to the increase in the critical inventory level, and its values in the first series of experiments appear substantially smaller than in the second series of experiments (see Fig. 4). This is because for $r = 20$ the relation $P_{vac} \approx 1$ takes place (see also formulas (7)).

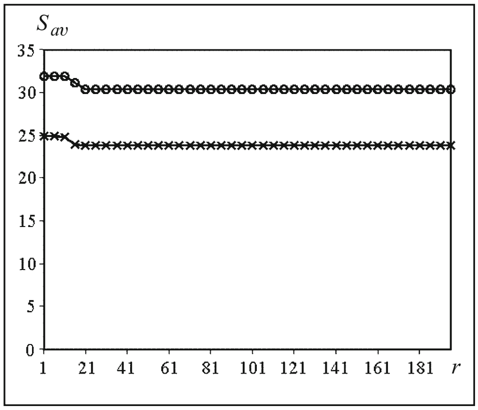

Fig. 6. Dependence of S_{av} on parameter r.

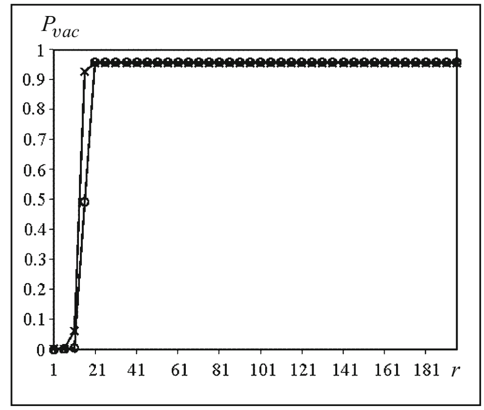

Fig. 7. Dependence of P_{vac} on parameter r.

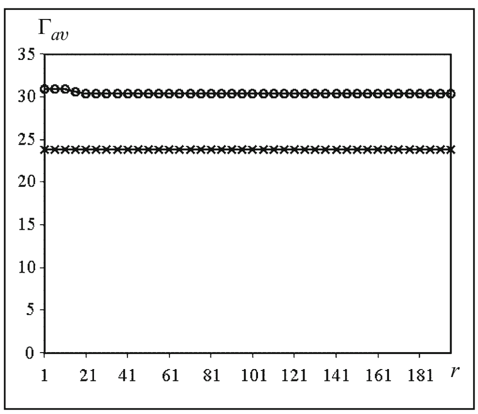

Fig. 8. Dependence of Γ_{av} on parameter r.

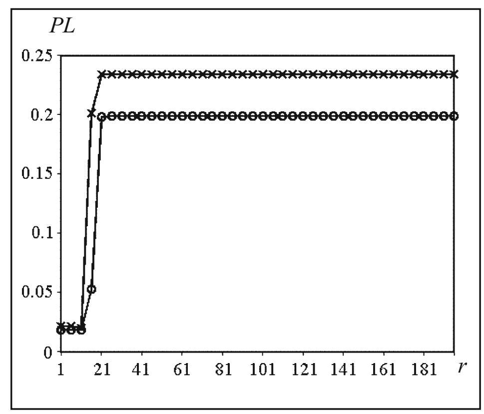

Fig. 9. Dependence of PL on parameter r.

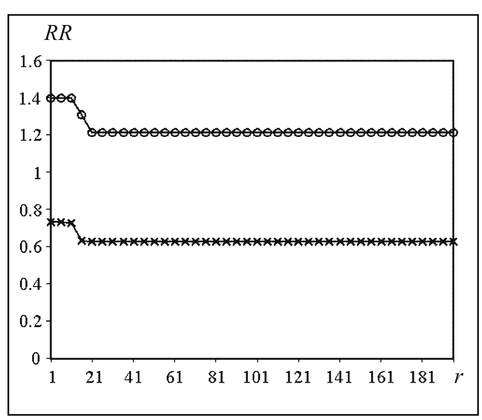

Fig. 10. Dependence of RR on parameter r.

Function RR is increasing with respect to the increase in the critical inventory level; and its values in the first series of experiments appear a bit larger than in the second series of experiments (see Fig. 5). This is because the larger the value of critical inventory level, the more often system's inventory level appears equal to $s+1$, and increase in r reduces the values of this function (see comments to formula (11)).

Let us now consider the dependences of system's characteristics on parameter r. We will also present the results for two different values of parameter s, i.e., in the first series of experiments it is accepted that $s = 24$ and in the second that $s = 12$. The results are shown in Figs. 6–10, where the notation × and ○ correspond to the first and second series of experiments, respectively.

In the both series of experiments, function S_{av} is almost piecewise constant with respect to the increase in parameter r (see Fig. 6), and in the first series of experiments (for $r \le 16$) we have $S_{av} \approx 32$ and for $r > 16$ the value of this function is approximately equal to 30; in the second series of experiments for $r \le 16$ we have $S_{av} \approx 24$ and for $r > 16$ the value of this function is approximately equal to 23. Such result can be explained as follows. The values of this function are mainly influenced by two factors: probability of the server being in operating mode (or on vacation) and inventory perish intensity. For the selected input data, the probability of server vacation for $r > 16$ is very close to one (see Fig. 7). This means that for large values of parameter r inventory level almost does not decrease as a result of their distribution to customers, and inventory perish intensity, as one would expect, almost does not depend on parameter r (see Fig. 8). The fact that the values of function S_{av} in the first series of experiments is greater than in the second series of experiments has quite logical explanation, i.e., the higher the critical inventory level, the higher the average inventory level.

Here, function PL is nondecreasing with respect to the increase in parameter r and its values in the first series of experiments appear a bit smaller than in the second series of experiments (see Fig. 9). Function RR is nonincreasing with respect to the increase in parameter r and its values in the first series of experiments appear almost twice larger than in the second series of experiments (see Fig. 10). This fact is explained in the comment to Fig. 5.

At the end of this section, we will note that three infinite series (see formulas (39) and (40)) appear in the model with unlimited queue of customers in the formula for calculating the customer loss probability; therefore, explicit formulas to calculate their sum cannot be found. These series converge, since the majorizing series $\sum_{i=1}^{\infty} \sigma_k(i)$, $k = 0, 1$, converge. Therefore, we use here the series tail cuttoff method [20], i.e., the upper bounds of each sum are replaced with sufficiently large (finite) quantities; then they are gradually increased, and this procedure continues until the values of the respective sums almost stop varying.

CONCLUSIONS

We have analyzed the model of a queueing-inventory system with one server and perishable stock in which impatient spending customers can generate queues of limited or unlimited length. If there is no queue of customers, the server goes on multiple vacation with exponential vacation time, i.e., the server returning from vacation takes vacation again if the number of customers in the queue at this instant of time is less than certain threshold value. We have analyzed models that use two-bin replenishment policy and the lead time is assumed to be a positive r.v. with exponential d.f. We have developed an exact and an approximate methods to find the characteristics of the models under study and used them to carry out numerical experiments.

The proposed approximate method allows investigating PQIS models with other replenishment policies and retrial customers, with delayed and working vacations of the server, etc., and solving the problems of their optimization. Such problems can be a subject of further study.

REFERENCES

1. V. S. Koroliuk, A. Z. Melikov, L. A. Ponomarenko, and A. M. Rustamov, "Asymptotic analysis of the system with server vacation and perishable inventory," Cybern. Syst. Analysis, Vol. 53, No. 4, 543–553 (2017).
2. J. G. Kemeny and J. L. Snell, Finite Markov Chains, Princeton University Press, Princeton (1960).
3. G. Guardabassi and S. Rinaldi, "The problems in Markov chains: A topological approach," Operations Research, Vol. 18, Issue 2, 324–333 (1970).
4. B. N. Feinberg and S. S. Chiu, "A method to calculate steady-state distributions of large Markov chains by aggregating states," Operations Research, Vol. 35, Issue 2, 282–290 (1987).

5. D. S. Kim and R. L. Smith, “An exact aggregation/disaggregation algorithm for large scale Markov chain,” Naval Research Logistics, Vol. 42, No. 7, 1115–1128 (1995).
6. P. J. Courtois, Decomposability: Queueing and Computer System Applications, Academic Press, New York–London–San Francisco (1977).
7. J. R. Koury, D. F. McAllister, and W. J. Stewart, “Iterative methods for computing stationary distribution of nearly completely decomposable Markov chains,” SIAM J. on Algebraic and Discrete Methods, Vol. 5, No. 2, 164–186 (1984).
8. M. Haviv, “Aggregation/disaggregation methods for computing the stationary distributions of a Markov chain,” SIAM J. on Numerical Analysis, Vol. 24, No. 4, 952–966 (1987).
9. W. J. Stewart, Introduction to the Numerical Solution of Markov Chains, Princeton University Press, Princeton (1994).
10. D. F. Rogers and R. Plante, “Estimating equilibrium probabilities for band diagonal Markov chains using aggregation and disaggregation techniques,” Computers & Operations Research, Vol. 20, No. 8, 857–877 (1993).
11. P. O. Dianne, Iterative Methods for Finding the Stationary Vector for Markov Chain, IMA Preprint Series, No. 932 (1992).
12. U. Sumita and M. Rieders, “Application of the replacement process approach for computing the ergodic probability vector of large scale row-continuous Markov chains,” J. of the Operations Research Society of Japan, Vol. 33, No. 4, 279–307 (1990).
13. V. S. Koroliuk and A. F. Turbin, Semi-Markov Processes and Their Applications [in Russian], Naukova Dumka, Kyiv (1976).
14. L. Ponomarenko, C. S. Kim, and A. Melikov, Performance Analysis and Optimization of Multi-Traffic on Communication Networks, Springer, Berlin–Heidelberg (2010).
15. A. Melikov and L. Ponomarenko, Multi-Dimensional Queueing Models in Telecommunication Networks, Springer, Heidelberg (2014).
16. C. Liang and H. Luh, “Optimal services for content delivery based on business priority,” J. of the Chinese Institute of Engineers, Vol. 36, No. 4, 422–440 (2013).
17. C. Liang and H. Luh, “Efficient method for solving a two-dimensional Markov chain model for call centers,” Industrial Management & Data Systems, Vol. 115, No. 5, 901–922 (2015).
18. M. E. Haddad and F. Belarbi, “Approximate analysis of an unreliable $M / M / c$ retrial queue with phase merging algorithm,” New Trends in Mathematical Sciences, Vol. 4, No. 3, 9–21 (2016).
19. B. Jajaraman, B. Sivakumar, and G. Arivarignan, “A perishable inventory system with postponed demands and multiple server vacations,” in: Modeling and Simulation in Engineering, Hindawi Publishing Corporation, Vol. 2012, Article ID 620960.
20. A. Z. Melikov, L. A. Ponomarenko, and S. A. Bagirova, “Markov models of queueing-inventory systems with variable order size,” Cybern. Syst. Analysis, Vol. 53, No. 3, 373–386 (2017).