Methodology and Computing in Applied Probability (2020) 22: 401–432
https://doi.org/10.1007/s11009-019-09707-7

Analysis of Queueing System with Non-Preemptive Time Limited Service and Impatient Customers

Chesoong Kim[1] · Alexander Dudin[2,3] · Olga Dudina[2,3] · Valentina Klimenok[2,3]

Received: 9 March 2018 / Revised: 5 March 2019 /
Accepted: 6 March 2019 / Published online: 22 March 2019

Abstract
We consider a single-server queueing system with server vacations as the important component of the polling queueing model of a real-world system. Period of continuous operation of the server (the maximum server attendance time) is restricted, but the service of a customer cannot be interrupted when this period expires. Such features are inherent for many real-world systems with resource sharing. We assume that the customers arrival is described by the Markovian Arrival Process and service, vacation and maximum server attendance times have a phase-type distribution. We derive the stationary distributions of the system states and waiting time. Taking in mind the necessity of further application of the results to modeling the polling queueing systems, the distribution of the server visiting time is derived. Extensive numerical results are presented. They highlight that an account of the coefficient of variation of vacation and maximum attendance time is very important for exact evaluation of the key performance measures of the system, while only the results for the coefficient of variation equal to zero or one are known in the literature. Impact of the possible customers impatience, which is intuitively important because the *time-limited* service is considered, is confirmed by the results of the numerical experiments. Optimization problem of matching the durations of vacation and maximum attendance time is considered.

Keywords Queueing model · Time limited service · Vacation · Polling system · Phase-type distribution · Markovian arrival process

Mathematics Subject Classification (2010) 68M20 · 60K25 · 90B22

✉ Chesoong Kim
dowoo@sangji.ac.kr

[1] Department of Industrial Engineering, Sangji University, Wonju, Kangwon, 220-702, Korea

[2] Belarusian State University, 4, Nezavisimosti Ave., Minsk 220030, Belarus

[3] RUDN University, 6 Miklukho-Maklaya st., 117198 Moscow, Russia

1 Introduction

1.1 Practical Motivation

This research was started during the analysis of a large inter-banking processing center aiming to optimize operation of this center. The basic function of this center is to handle operations (transactions) between the cooperating entities (banks). The principle of operation of the center is a priori chosen as time division. Time is divided to the slots and during a slot transactions of only one entity can be processed. Entities are different with respect to the average number of the required transactions per unit of time and their importance for the system. This information has to be taken for decision support about the duration and frequency of slots provided to the entities. Mathematical modeling of the center is important both from the point of view of better performance of the center (e.g. in terms of the weighted average time for transaction implementation) and from the point of view of fair access of various entities to the center. The results of the modeling are useful to avoid underutilization of the center capacity during the slots assigned to some important entities and congestion during the other slots.

Previous attempts to optimize the work of this center based on the results of computer simulation or application of the simplest models of queueing networks did not lead to successful results. Due to existence of many choices of parameters of the center, simulation is extremely time consuming. The relevant queueing networks do not allow solution close to product form. Therefore, it was suggested to build and analyse queueing model more or less adequate to reality of the center operation.

From the point of view of queueing theory, it is evident that the work of the processing center should be described by the polling model. Theory of queueing systems with polling is quite well developed, see, e.g. the surveys (Boon et al. 2011; Hanbali et al. 2012; Vishnevsky and Semenova 2006). However, the direct application of the known results in the literature results was impossible due to the following reasons and imposed from the early beginning restrictions:

- The statistical analysis of the flows of transactions in the real center under investigation evidently showed the presence of positive correlation of successive inter-arrival times. As it is already well-known from the existing queueing literature, correlation in the arrival processes essentially deteriorates performance measures of queueing systems comparing to the corresponding systems with the stationary Poisson arrival process having the same arrival rate. But, the overwhelming majority of the relevant papers assume the stationary Poisson arrival process of customers. This assumption drastically simplifies the mathematical analysis because it reduces the dimension of the Markov process describing behavior of the system. However, this assumption does not hold true for the considered center.
- Essential restriction was to take into account the ban of interruption of a transaction. If the time slot assigned to a given entity expires during some transaction processing, this processing cannot be terminated ahead of the schedule. Only after transaction completion the center may switch to processing of transactions of the next entity. Such a ban is natural from the point of view of referential integrity of the information.
- Waiting time for a transaction is restricted.

Analysis of the existing literature has shown the lack of the models where all three listed restrictions are satisfied. Therefore, it was necessary to implement the analysis of the described system. It is clear that the analytical analysis of the whole system is not possible.

The well-known approach to analysis of a polling system consists of decomposition of the system into sub-systems each of which describes processing of transactions of one, tagged, entity in terms of the appropriate vacation queueing model. Time slot, during which the center provides service to these transactions, in what follows we call the server visiting time. When this times expires, the server switches to service of transactions from another entities. From the point of view of the tagged entity, the time, during which transactions of other entities are handled, may be considered as the server vacation. The analysis of the whole polling system can be successfully performed via the analysis of a set of vacation models describing the dynamics of the service of the transactions of the tagged entity with properly chosen distribution of the vacation time. However, in reality these distributions for different entities are not a priori known and depend on each other. The vacation time in the model of service of one entity is the sum of visiting times of other entities. But this problem is more or less easily solved heuristically by means of iterative computations where, at each step, an unknown distribution of a vacation time in the tagged vacation queue is rectified based on results of the computation of the distribution of visiting time of other queues by the server, see, e.g., Vishnevsky et al. (2012).

Thus, the important step in modeling the operation of the processing center is to elaborate the accurate vacation queueing model taking into account the listed above restrictions.

1.2 The Relevant Literature and Contributions of the Paper

In vacation queueing models, it is suggested that the server of the system can take a vacation during which the service of customers is temporarily suspended. As important references concerning the vacation queueing models the studies (Takagi 1990, 1991, 1997, 2000) by H. Takagi and the book Tian and Zhang (2006) can be mentioned. Vacation queueing models are very versatile with respect to the rules of beginning and ending vacations (correspondingly, ending and starting the service periods). The existing literature is very extensive. Therefore, for easier navigation in this literature, the classification of such models was elaborated, see, e.g., Takagi (1991). This classification is permanently developing due to appearance of new models of various real-world systems.

Basically, the most popular rules defining the duration of the service period are the following: (i) the exhaustive service that suggests that once the service period begins it will end only when the system will become empty; (ii) the decremented service that suggests that the service of customers is terminated when the number of customers in the system decreases to the predefined number; (iii) the gated service that suggests that the service period continues until all customers presenting in the system at this period beginning will be served; (iv) the limited service that suggests that the service period is restricted. This restriction may have two forms. One form assumes the limitation of the number of customers that can obtain service during one service period. Usually, this number is restricted from above, however, restriction from below might be considered as well, see, e.g., the recent paper Boxma et al. (2015). Another form assumes the limitation of the time during which the server can continuously provide service. We refer to the latter rule as time-limited service and to the time, during which the server can continuously provide service, as a service period or the maximum attendance time. To be short, further we abbreviate this time as MAT. Different combinations of the listed rules are considered in the literature as well. E.g., it is possible to combine the gated and time-limited service: a service period ends when the MAT expires or all customers presenting in the system at the service period beginning instant finish service, whichever occurs first. Another possible combination is a composition of the exhaustive and time-limited service: a service period ends when the MAT expires or the system becomes

empty, whichever occurs first. In this paper, we consider namely such a combination. This combination is very important from the point of view of potential real-world applications. In particular, it is effectively applied in the systems where a certain restricted resource is dynamically shared among several users. Limitation of the time of continuous service of the requests generated by some user helps to get more fair and timely access for various users, to avoid any types of monopolization of the resource. Exhaustive service allows to terminate an access for the user that currently does not need the resource.

The model considered in this paper has the following advantages over the existing in the literature.

- To take into account the existence of correlation of inter-arrival times, we consider the known in the literature Markovian Arrival Process (MAP) of arriving customers (transaction). For definition, properties and related literature, see, e.g., Chakravarthy (2001), Lucantoni (1991), and Vishnevski and Dudin (2017).
- Ideally from the mathematical point of view, durations of the maximum attendance time (MAT) defining restriction on the continuous time of the server operation without going to vacation, service and vacation times have to be random with the general distribution. However, practically all papers in this subject, probably except the papers where the authors intend to analyze the system only in some asymptotic conditions, these distributions are assumed to be exponential or degenerate. E.g., the distribution of the MAT is assumed to be exponential in de Haan et al. (2009), Hanbali et al. (2012), Katayama (2001, 2007), Katayama and Kobayashi (2007) adn Leung (1994). The constant MAT is considered in Frigui and Alfa (1998), Leung and Eisengerg (1990), Leung and Lucantoni (1994), and de Se Silva et al. (1995). The coefficient of variation of the exponential distribution is equal to 1 and the coefficient of variation of the degenerate distribution is equal to 0. If in the real-world system the MAT or service or vacation time has higher than 1 coefficient of variation, the hyper-exponential distribution can be recommended. If this coefficient is less than 1, Erlangian distribution can be applied. Both these distributions are the very special case of the so-called PH (phase-type) distribution, see, e.g. Neuts (1981). The possibility of approximation (in sense of a weak convergence) of an arbitrary distribution by the PH distribution is mentioned, e.g., in Asmussen (2003). By this reason, as a trade-off between the desire to analyze the model under the most general assumptions about the distribution of the MAT, service and vacation times and the possibility to get tractable results ready for computer realization, we suggest that these times in our model have the PH-type distribution. Very high importance of account of variation of the MAT and vacation times is demonstrated in our paper by the numerical examples. E.g., the probability of an arbitrary customer loss differs by two times in the systems with the same mean vacation time but the different coefficient of variation.
- When the time limited service is considered, it is quite often that the MAT expires during the service of a certain customer. Two options are possible in such a situation. The first option is that the service of this customer is preempted. The customer is lost or will be served after the vacation period completion. This option was considered, e.g., in de Haan et al. (2009), Hanbali et al. (2012), and Leung and Eisengerg (1990). The second option is that the service of this customer has to be continued until this customer receives complete service. This option was considered, e.g., in Katayama (2007), Katayama and Kobayashi (2007), and Leung (1994). Both options together were considered in Katayama (2001) and de Se Silva et al. (1995). The option with non-preemptive service is definitely more difficult for analysis in the case of non-exponential distribution of the service, vacation and MAT. This is because if one wishes

to describe the behavior of the system by the Markov process, during the MAT he/she must monitor simultaneously the elapsed (or residual) times for the MAT and service times. If the technique of the supplementary variables will be applied (with two continuous supplementary variables as the elapsed or residual times for the maximum attendance and service times at a given time moment), this will lead to the functional or integro-differential equations of the type solution for which is not known in the relevant literature. In our paper, we consider the more difficult option with non-preemptive service. The use of the PH distribution of the MAT and service times instead of the arbitrary distribution implies that it is possible to replace the account of two continuous supplementary variables by the account of two supplementary variables with the discrete state space. The corresponding multi-dimensional Markov process can be analyzed via the matrix analytic methods.

- In our paper, we assume that the customers waiting in the queue are impatient and may leave the system without service after some amount of waiting time. Queues with impatient customers are a popular subject of research. However, no one of the relevant papers cited above takes impatience into account. It is worth to note that we assume that the rate of customer's leaving the system without service depends on the state of the server (the server is on the vacation, provides service within the MAT or already after expiration of this time). Such a dependence, e.g. reflects the fact that the customers may leave the system more intensively when the MAT already expired, the server will switch-off soon and will return for service only after a vacation.
- Because consideration of the vacation queue is motivated by the further application of the obtained results for analysis of the polling system, we supplement the standard in the literature analysis of the stationary distributions of the queue length and waiting time with the analysis of the server visiting time defined as the time interval since the epoch of vacation completion till the moment of the next vacation beginning. Server visiting time is rarely analysed in the literature by the following reason. If service discipline is exhaustive, i.e., vacation starts only when the system becomes empty, visiting time coincides with busy period, distribution of which is well-known. If the service discipline is time limited with service interruption, visiting time is just the minimum of the busy period and MAT, therefore, its analysis is trivial. But in the case of the time limited discipline without service interruption the task of derivation of distribution of the visiting time becomes complicated. This task is solved in our paper. The corresponding results have the methodological value.
- The last but not the least, despite the fact that the model is quite complicated we provide the exact, not approximate, analytical and algorithmic analysis of the formulated model. Computer realization of the elaborated algorithms for computation of the stationary distributions of the system states, sojourn time of an arbitrary customer and server visiting time as well as the major performance measures of the system shows that the required computation time is very small, negligible comparing to the required for computer simulation of the model time.

The paper Frigui and Alfa (1998) deserves the special citing. In that paper, the model of $MAP/PH/1$ type with time-limited preemptive service is considered in discrete time settings. Analysis of the vacation models with time-limited service in discrete time is easier than the analysis of the corresponding models in continuous time. Time limit is not assumed to be having discrete PH distribution. Non-preemption of service, which also makes the analysis more complicated, is not allowed in Frigui and Alfa (1998).

1.3 The Outline of Presentation

The paper is organized as follows. In Section 2, the mathematical model under study is described in detail. In Section 3, the process of the system states is defined by a continuous-time multi-dimensional Markov chain. This chain belongs to the class of asymptotically Quasi-Toeplitz Markov chains. The generator of this chain is written down. In Section 4, the analysis of the Markov chain is presented. It is proved here that if the customers are impatient at least in one state of the server (the server is on the vacation, provides service within the MAT or already after expiration of this time), the stationary probabilities of the system states exist for any set of the system parameters. Expressions for computing key performance measures of the system are given in Section 5. The Laplace-Stieltjes transform of the waiting time distribution is derived in Section 6. The formula for computation of the mean waiting time is given there as well. The Laplace-Stieltjes transform of the distribution of the visiting time of the server in the case of customers patient during the MAT is derived in Section 7. The formula for computing the mean visiting time is given there. Results of numerical experiments are briefly described in Section 8. In particular, importance of consideration of more general, than the exponential, PH type distribution of the MAT and vacations time is illustrated. Section 9 concludes the paper.

2 The Mathematical Model

We consider a single-server queueing system with an infinite buffer. The input flow is described by the MAP. Customer's arrival in the MAP is directed by an underlying irreducible continuous-time Markov chain ν_t, $t \geq 0$, with a finite state space $\{0, ..., W\}$. The sojourn time of the chain ν_t, $t \geq 0$, in the state ν has an exponential distribution with the parameter λ_ν, $\nu = \overline{0, W}$. Here and throughout this paper the notation of type $\nu = \overline{0, W}$ means that ν takes values from the set $\{0, ..., W\}$. After this sojourn time expires, with probability $p_k(\nu, \nu')$, the process ν_t jumps to the state ν', and k customers, $k = 0, 1$, arrive into the system. The rates of jumps of the underlying Markov chain from one state into another with generation of k customers are combined into the matrices D_k, $k = 0, 1$, of size $(W + 1) \times (W + 1)$. The matrix $D(1) = D_0 + D_1$ is the infinitesimal generator of the process ν_t, $t \geq 0$. The invariant probability vector (vector of stationary distribution) $\boldsymbol{\theta}$ of this process is computed as the unique solution to the equations $\boldsymbol{\theta} D(1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$. Here and throughout this paper $\mathbf{0}$ is a zero row vector and $\mathbf{e}$ is a column vector of appropriate size consisting of ones. In the case when the size of a vector is not clear from context, it is indicated as a lower index, e.g. $\mathbf{e}_{\overline{W}}$ denotes the unit column vector of size $\overline{W} = W + 1$. The fundamental rate λ of the MAP is defined as $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$ and gives the expected number of arrivals per unit of time in the stationary mode. The variance v of intervals between customer arrivals is calculated as $v = 2\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - \lambda^{-2}$, the squared coefficient c_{var} of variation is equal to $c_{var} = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$, while the coefficient c_{cor} of correlation of successive intervals between arrivals is given by $c_{cor} = (\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1}D_1(-D_0)^{-1}\mathbf{e} - \lambda^{-2})/v$.

For more information about the MAP, its special cases, properties and related research see Lucantoni (1991) and the survey paper Chakravarthy (2001). Usefulness of the MAP in modeling customers flows in telecommunication systems is mentioned in Heyman and Lucantoni (2003) and Klemm et al. (2003). Methods for constructing the MAP based on the traces of the customer flows in real-world systems are available in the literature. As the recent paper, Buchholz and Kriege (2017) can be mentioned.

The state of the server alternates between the busy (service) and the vacation periods. Therefore, an arriving customer never starts service immediately upon arrival. This customer is placed to the buffer and is then picked up for the service according to the First In - First Out discipline. The vacation period starts after completion of the service period. The length of the vacation period has the PH distribution with an irreducible representation $(\boldsymbol{\gamma}, \Gamma)$. This means the following. The vacation is governed by the underlying process $\xi_t,\ t \geq 0$, which is a continuous time Markov chain with state space $\{1, \ldots, R, R+1\}$. The initial state of the process $\xi_t,\ t \geq 0$, at the epoch of starting the vacation is determined within the set $\{1, \ldots, R\}$ of transient states by the probabilistic row-vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_R)$. The rates of the process $\xi_t,\ t \geq 0$, transitions within the set $\{1, \ldots, R\}$, which do not lead to the vacation period completion, are defined by the square irreducible matrix Γ of size R. The rates of transitions to the absorbing state $R+1$, which lead to vacation completion, are given by the entries of the column-vector $\boldsymbol{\Gamma}_0 = -\Gamma\mathbf{e}$. The distribution function of vacation time has the form $1 - \boldsymbol{\gamma} e^{\Gamma x}\mathbf{e}$. The Laplace-Stieltjes transform of this distribution function is $\boldsymbol{\gamma}(sI - \Gamma)^{-1}\boldsymbol{\Gamma}_0,\ Re\ s > 0$. The average length of the vacation time is given by $v_1 = \boldsymbol{\gamma}\,(-\Gamma)^{-1}\,\mathbf{e}$.

If at the vacation completion instant the system is empty, a new vacation starts immediately. The new vacation period also has the PH distribution with an irreducible representation $(\boldsymbol{\gamma}, \Gamma)$. If the system is not empty, the service period starts. The service period (maximum server attendance time) has the PH distribution with an irreducible representation $(\boldsymbol{\tau}, T)$. The underlying process $\chi_t,\ t \geq 0$, of the service period is a continuous-time Markov chain with the state space $\{1, \ldots, K\}$. The average duration of the service period is defined by formula $\tau_1 = \boldsymbol{\tau}(-T)^{-1}\mathbf{e}$. Simultaneously with the beginning of a service period, the service time of the first customer in the service period starts. The service time has the PH distribution with an irreducible representation $(\boldsymbol{\beta}, S)$. An underlying process of the service time is $\eta_t,\ t \geq 0$, with finite state space $\{1, \ldots, M\}$. The average service time is defined by formula $b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$. Service of customers is stopped and the vacation period starts if during the service completion instant the system is idle. Alternatively, customers service should be finished if the maximum service attendance time expires. However, in contrast to the standard T-limited service, see Tian and Zhang (2006), here we assume that the currently provided service is not preempted. A new vacation period will start only when this service will be completed.

Customers staying in the buffer are impatient. Each customer leaves the system independently of other customers if its waiting time exceeds an exponentially distributed time. The parameter of the exponential distribution is equal to α_r where $r = 0$ if a vacation period is in a progress, $r = 1$ if the server provides service while the maximum service attendance time is not expired, and $r = 2$ if the server provides service but the maximum service attendance time is finished. We suggest that $\alpha_r > 0$ at least for one value of $r,\ r = 0, 1, 2$. It is worth to mention that we cannot consider here more general, PH, distribution of patience time because the number of customers in the system is unlimited and the corresponding Markov chain describing behavior of the system does not belong to the known class of Asymptotically Quasi-Toeplitz Markov chains which are used for the system analysis in the next section.

Our aim is to analyze stationary behavior of the described queueing model.

3 The Process of the System States

Let

- i_t be the number of customers in the system, $i_t \geq 0$,

- r_t be the current state of the server: $r_t = 0$ if the vacation period is in a progress, $r_t = 1$ if the server provides service while the maximum service attendance time is not expired, and $r_t = 2$ if the server provides service but the MAT is already finished,
- ν_t be the state of the underlying process of the MAP, $\nu_t = \overline{0, W}$,
- m_t be the current phase of the underlying process ξ_t of the vacation time if $r_t = 0$; $m_t = (\chi_t, \eta_t)$ (the pair of the current phases of the underlying processes of the MAT and service time) if $r_t = 1$; and $m_t = \eta_t$ if $r_t = 2$,

at the moment $t,\ t \geq 0$.

It is easy to see that the state space of the multi-dimensional process $\zeta_t = \{i_t,\ r_t,\ \nu_t,\ m_t\},\ t \geq 0$, is

$$\Omega = \{(i,\ 0, \nu, \xi),\ i \geq 0,\ 0 \leq \nu \leq W,\ 1 \leq \xi \leq R)\}$$

$$\bigcup \{(i,\ 1, \nu, \chi, \eta),\ i \geq 1,\ 0 \leq \nu \leq W,\ 1 \leq \chi \leq K,\ 1 \leq \eta \leq M\}$$

$$\bigcup \{(i,\ 2, \nu, \eta),\ i \geq 1,\ 0 \leq \nu \leq W,\ 1 \leq \eta \leq M\}$$

and this process is an irreducible continuous-time Markov chain with one component, i_t, having infinite state space and finite other components.

To analyse the behavior and properties of the Markov chain ζ_t, we have to compute the infinitesimal generator of this chain. Let us denote this generator as $\mathbf{Q}$. The diagonal entries of the generator are negative. Modulus of each diagonal entry defines the rate of departure of the Markov chain from the corresponding state. The non-diagonal entries are non-negative and define the rates of the transitions of the Markov chain within its state space.

To simplify the structure of the generator $\mathbf{Q}$, let us enumerate the states of the Markov chain ξ_t in the lexicographic order and compose all the states of the chain having value (i, r) of the first two components to a *sub-level* (i, r). The sub-level $(i, 0)$ contains $\overline{W}R$ states, the sub-level $(i, 1)$ contains $\overline{W}KM$ states and the sub-level $(i, 2)$ contains $\overline{W}M$ states. Then, we compose sub-levels $(i, r),\ r = 0, 1, 2$, to the *level* i.

Lemma 1 *The generator* $\mathbf{Q}$ *has a block-tridiagonal structure:*

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

where non-zero blocks $\mathbf{Q}_{i,j}$ *defining the rates of the transition from the level* i *to the level* $j,\ j = \max\{0, i-1\}, i, i+1$, *are defined as follows:*

$$\mathbf{Q}_{0,0} = D_0 \oplus (\Gamma + \mathbf{\Gamma}_0 \boldsymbol{\gamma}), \qquad \mathbf{Q}_{1,0} = \begin{pmatrix} \alpha_0 I_{\overline{W}R} \\ I_{\overline{W}} \otimes \boldsymbol{e}_K \otimes \boldsymbol{S}_0 \otimes \boldsymbol{\gamma} \\ I_{\overline{W}} \otimes \boldsymbol{S}_0 \otimes \boldsymbol{\gamma} \end{pmatrix},\ \mathbf{Q}_{0,1} = \begin{pmatrix} D_1 \otimes I_R & | O & | O \end{pmatrix},$$

$$\mathbf{Q}_{i,i-1} = \begin{pmatrix} O & O & O \\ O & I_{\overline{W}} \otimes I_K \otimes \boldsymbol{S}_0 \otimes \boldsymbol{\beta} & O \\ I_{\overline{W}} \otimes \boldsymbol{S}_0 \otimes \boldsymbol{\gamma} & O & O \end{pmatrix} +$$

$$+ diag\{i\alpha_0 I_{\overline{W}R}, (i-1)\alpha_1 I_{\overline{W}KM}, (i-1)\alpha_2 I_{\overline{W}M}\},\ i \geq 2,$$

$$\mathbf{Q}_{i,i} = \begin{pmatrix} D_0 \oplus \Gamma & I_{\overline{W}} \otimes \mathbf{\Gamma}_0 \otimes \boldsymbol{\tau} \otimes \boldsymbol{\beta} & O \\ O & D_0 \oplus T \oplus S & I_{\overline{W}} \otimes \boldsymbol{T}_0 \otimes I_M \\ O & O & D_0 \oplus S \end{pmatrix} -$$

$$-diag\{i\alpha_0 I_{\overline{W}R}, (i-1)\alpha_1 I_{\overline{W}KM}, (i-1)\alpha_2 I_{\overline{W}M}\},\ i \geq 1,$$

$$\mathbf{Q}_{i,i+1} = \begin{pmatrix} D_1 \otimes I_R & O & O \\ O & D_1 \otimes I_K \otimes I_M & O \\ O & O & D_1 \otimes I_M \end{pmatrix},\ i \geq 1.$$

Here, I is the identity matrix, and O is a zero matrix of appropriate dimension, diag$\{\dots\}$ means a diagonal matrix with the diagonal blocks listed in the brackets, $\otimes, \oplus$ are the symbols of the Kronecker product and sum of matrices correspondingly, see Graham (1981).

Proof of Lemma 1 consists of analysis of the Markov chain $\xi_t,\ t \geq 0$, transitions during an infinitesimal interval of time and further combining the corresponding transition rates into the matrix blocks. The block structure with three block rows and three block columns of the matrices $\mathbf{Q}_{i,i}, \mathbf{Q}_{i,i+1},\ i \geq 1, \mathbf{Q}_{i,i-1},\ i \geq 2$, corresponds to possible transitions of the component r_t of the Markov chain ζ_t from the states 0,1,2 to the states 0,1,2, respectively. E.g., the blocks $(\mathbf{Q}_{i,i+1})_{r,r'}$ contain the rates of the transitions from the sub-level $(i, r),\ i \geq 1$, to the sub-level $(i+1, r')$. The Kronecker product and sum of matrices are very useful here for compact description of the rates of joint transition of several independent Markov processes. The boundary blocks $\mathbf{Q}_{0,0}, \mathbf{Q}_{0,1}, \mathbf{Q}_{1,0}$ have less sub-blocks because the process r_t may have only state 0 when the system is empty.

4 Analysis of the Markov Chain

The following statement is true.

Theorem 1 *Let $\alpha_r > 0$ at least for one value of $r,\ r = 0, 1, 2$. The Markov chain ζ_t is ergodic for any set of the system parameters.*

Proof Let $A_r,\ r = 0, 1, 2$, be the diagonal matrix with the diagonal entries defined by the moduli of the diagonal entries of the matrix $D_0 \oplus \Gamma$, if $r = 0$, $D_0 \oplus T \oplus S$, if $r = 1$ and $D_0 \oplus S$, if $r = 2$.

Let $\mathbf{R}_i$ be the diagonal matrix with the diagonal entries given by the moduli of the diagonal entries of the matrix $\mathbf{Q}_{i,i}$. It can be verified that the matrix $\mathbf{R}_i$ is defined by the formula

$$\mathbf{R}_i = \text{diag}\{A_0, A_1, A_2\} + \text{diag}\{i\alpha_0 I_{\bar{W}R}, (i-1)\alpha_1 I_{\bar{W}KM}, (i-1)\alpha_2 I_{\bar{W}M}\}.$$

It can be checked that the following limits exist:

$$\mathbf{Y}_0 = \lim_{i\to\infty} \mathbf{R}_i^{-1}\mathbf{Q}_{i,i-1},\ \mathbf{Y}_1 = \lim_{i\to\infty} \mathbf{R}_i^{-1}\mathbf{Q}_{i,i} + I,\ \mathbf{Y}_2 = \lim_{i\to\infty} \mathbf{R}_i^{-1}\mathbf{Q}_{i,i+1}$$

and the matrix $\mathbf{Y} = \mathbf{Y}_0 + \mathbf{Y}_1 + \mathbf{Y}_2$ is stochastic.

This implies that all conditions of the definition of asymptotically Quasi-Toeplitz Markov Chain ($AQTMC$) given in Klimenok and Dudin (2006) are fulfilled and the Markov chain ξ_t belongs to the class of $AQTMC$. This gives us an opportunity to derive the ergodicity and non-ergodicity conditions for this Markov chain and compute its steady-state distribution.

It is easy to make sure that, if $\alpha_r > 0$ at least for one value of $r,\ r = 0, 1, 2$, then the matrix $\mathbf{Y}$ is reducible. In such a case, according to Klimenok and Dudin (2006) the matrix $\mathbf{Y}$ has to be transformed into the canonical normal form, for details see Gantmakher (1967). Let this normal form contains m irreducible stochastic blocks, say, $\mathbf{Y}^{(l)},\ l = \overline{1, m}$. Then, as

follows from Klimenok and Dudin (2006), the sufficient condition for the ergodicity of the Markov chain ζ_t is the simultaneous fulfilment of the inequalities

$$\mathbf{y}^{(l)}\mathbf{Y}_0^{(l)}\mathbf{e} > \mathbf{y}^{(l)}\mathbf{Y}_2^{(l)}\mathbf{e},\ l = \overline{1,m}, \tag{2}$$

where the vectors $\mathbf{y}^{(l)}$ are defined as solutions of the systems of linear algebraic equations

$$\mathbf{y}^{(l)} = \mathbf{y}^{(l)}\mathbf{Y}^{(l)},\ \mathbf{y}^{(l)}\mathbf{e} = 1,\ l = \overline{1,m},$$

and $\mathbf{Y}_0^{(l)}$, $\mathbf{Y}_2^{(l)}$ are blocks of the matrices $\mathbf{Y}_0$ and $\mathbf{Y}_2$ corresponding to the block $\mathbf{Y}^{(l)}$ in the canonical normal form of the matrix $\mathbf{Y}$.

It can be shown, that, if exactly m' rates among α_r, $r = 0, 1, 2$, are positive, $m' = 1, 2, 3$, then the number m of irreducible stochastic blocks $\mathbf{Y}^{(l)}$ in the canonical normal form of the matrix $\mathbf{Y}$ is equal to m' and $\mathbf{Y}^{(l)} = \mathbf{Y}_0^{(l)} = I,\ l = \overline{1,m'}$. Correspondingly, $\mathbf{Y}_2^{(l)} = O,\ l = \overline{1,m'}$. Thus, for any stochastic vector $\mathbf{y}^{(l)}$ inequalities (2) take the form $1 > 0$ what is always true. Theorem 1 is proved. □

Remark 1 The statement of Theorem 1 may seem, at first sight, a bit strange. The status r of the server can have three possible values, $r = 0, 1, 2$, see above. However, if $\alpha_r > 0$ at least for one value of r, the system is always ergodic. The explanation of this phenomenon stems from the intuitive consideration that an ergodicity condition for any queueing system is defined as the condition of its ability to reduce the number of customers in the system in the situation when the system is overloaded. Because the mean duration of the time when the server has status r is strictly positive, positive value of the rate α_r of departure of the customers from the system under this status of the server implies departure of the huge number of customers from the overloaded buffer.

We suggest that conditions of Theorem 1 are fulfilled. Then the stationary distribution of the Markov chain ζ_t exists. Denote the stationary state probabilities of the chain as

$$\boldsymbol{\pi}(i,0,\nu,\xi) = \lim_{t\to\infty} P\{i_t = i,\ r_t = 0,\ \nu_t = \nu,\ \xi_t = \xi\},\ i \geq 0,\ \nu = \overline{0,W}, \xi = \overline{1,R},$$

$$\boldsymbol{\pi}(i,1,\nu,\chi,\eta) = \lim_{t\to\infty} P\{i_t = i,\ r_t = 1,\ \nu_t = \nu,\ \chi_t = \chi,\ \eta_t = \eta\},$$

$$i \geq 1,\ \nu = \overline{0,W},\ \chi = \overline{1,K},\ \eta = \overline{1,M},$$

$$\boldsymbol{\pi}(i,2,\nu,\eta) = \lim_{t\to\infty} P\{i_t = i,\ r_t = 2,\ \nu_t = \nu,\ \eta_t = \eta\},\ i \geq 1,\ \nu = \overline{0,W},\ \eta = \overline{1,M}.$$

Let $\boldsymbol{\pi}(i,r)$ be the row vector of probabilities of the states belonging to the sub-level (i,r) and $\boldsymbol{\pi}_i$ be the row vector of probabilities of the states belonging to the level i :

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1), \boldsymbol{\pi}(i,2)),\ i \geq 1,\ \boldsymbol{\pi}_0 = \boldsymbol{\pi}(0,0).$$

Computation of stationary probability vectors for asymptotically quasi-Toeplitz Markov chains is a pretty difficult task. Fortunately, the corresponding effective numerically stable algorithms are elaborated in Klimenok and Dudin (2006) for the case when the generator $\mathbf{Q}$ of the Markov chain ζ_t has the block upper-Hessenbergian form and in Dudina et al. (2013) when this generator has more simple block tridiagonal form. We use for computations the algorithm from Dudina et al. (2013).

5 Performance Measures of the System

As soon as the vectors $\boldsymbol{\pi}_i$, $i \geq 0$, have been computed, we are able to calculate various performance measures of the system.

– The average number of customers in the system $L = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i \mathbf{e}$.
– The average number of customers in the queue

$$L_q = \sum_{i=1}^{\infty} i\boldsymbol{\pi}(i,0)\mathbf{e}_{\bar{W}R} + \sum_{i=2}^{\infty}(i-1)\boldsymbol{\pi}(i,1)\mathbf{e}_{\bar{W}KM} + \sum_{i=2}^{\infty}(i-1)\boldsymbol{\pi}(i,2)\mathbf{e}_{\bar{W}M}.$$

– The fraction of time when the server has the vacation (has status 0) $F^{(0)} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0)\mathbf{e}$.
– The fraction of time when the server provides the service within the MAT (has status 1) $F^{(1)} = \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,1)\mathbf{e}$.
– The fraction of time when the MAT is over but the server does not finish service (has status 2) $F^{(2)} = \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,2)\mathbf{e}$.
– The average number of customers in the system conditional that the vacation is in a progress

$$L^{(0)} = (F^{(0)})^{-1} \sum_{i=1}^{\infty} i\boldsymbol{\pi}(i,0)\mathbf{e}.$$

– The average number of customers in the system conditional that the server has status r

$$L^{(r)} = (F^{(r)})^{-1} \sum_{i=1}^{\infty} i\boldsymbol{\pi}(i,r)\mathbf{e},\ r = 1,2.$$

– The probability that an arbitrary arriving customer meets the server being on the vacation

$$P_0 = \lambda^{-1} \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0)(D_1 \otimes I_R)\mathbf{e}.$$

– The probability that an arbitrary arriving customer meets the server having status r

$$P_r = \lambda^{-1} \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,r)(D_1 \otimes I_{K^{2-r}M})\mathbf{e},\ r = 1,2.$$

– The rate $\lambda^{(out)}$ of the flow of customers, which successfully received service in the system

$$\lambda^{(out)} = \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,1)(\mathbf{e}_{\bar{W}K} \otimes \mathbf{S}_0) + \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,2)(\mathbf{e}_{\bar{W}} \otimes \mathbf{S}_0).$$

– The probability of an arbitrary customer loss from the system (due to impatience)

$$P^{(loss)} = \lambda^{-1}\left[\sum_{i=1}^{\infty} i\alpha_0\boldsymbol{\pi}(i,0)\mathbf{e} + \sum_{i=1}^{\infty}(i-1)\alpha_1\boldsymbol{\pi}(i,1)\mathbf{e} + \sum_{i=1}^{\infty}(i-1)\alpha_2\boldsymbol{\pi}(i,2)\mathbf{e}\right]$$

or

$$P^{(loss)} = 1 - \frac{\lambda^{(out)}}{\lambda}.$$

– The rate J of server's switching on (average number of server's switching on per unit time)

$$J = \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,0)(\mathbf{e}_{\bar{W}} \otimes \Gamma_0).$$

6 Waiting Time Distribution

Let $Z(x)$ be the distribution function of the waiting time of an arbitrary customer and $z(s) = \int_0^\infty e^{-sx} dZ(x)$, $Re\ s > 0$, be its Laplace-Stieltjes transform. Let also $\mathbf{z}_i^{(r)}(s)$ be the column vectors of LSTs of the waiting time of an arbitrary customer conditional it arrives when there are i customers in the system, the status of the server is r and the states of the underlying Markov processes of the vacation time (if $r = 0$), the service time and the MAT (if $r = 1$), or the service time (if $r = 2$) are fixed.

Theorem 2 *The Laplace-Stieltjes transform $z(s)$ can be computed as follows:*

$$z(s) = \frac{1}{\lambda}\Bigg[\sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0)(D_1 \mathbf{e}_{\bar{W}} \otimes I_R)\mathbf{z}_i^{(0)}(s) + \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,1)(D_1 \mathbf{e}_{\bar{W}} \otimes I_{KM})\mathbf{z}_i^{(1)}(s) + \\ + \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,2)(D_1 \mathbf{e}_{\bar{W}} \otimes I_M)\mathbf{z}_i^{(2)}(s) \Bigg] \tag{3}$$

where the column vectors $\mathbf{z}_i^{(r)}(s)$, $r = 0, 1, 2$, constitute the column vectors $\mathbf{z}_i(s)$ of dimension $R + KM + M$, $\mathbf{z}_i(s) = \left(\mathbf{z}_i^{(0)}(s),\ \mathbf{z}_i^{(1)}(s),\ \mathbf{z}_i^{(2)}(s) \right)'$, $i \geq 0$, that are computed recursively by

$$\mathbf{z}_0(s) = \left(\mathbf{z}_0^{(0)}(s),\ \mathbf{e}_{KM},\ \mathbf{e}_M \right)', \tag{4}$$

$$\mathbf{z}_i(s) = \mathcal{M}_i(s)(\mathbf{g}_i(s) + \mathcal{N}_i(s)\mathbf{z}_{i-1}(s)),\ i \geq 1, \tag{5}$$

where

$$\mathbf{z}_0^{(0)}(s) = \mathcal{A}_0^{(0)}(s)(\Gamma_0 + \alpha_0 \mathbf{e}),$$

$$\mathbf{g}_i(s) = \left(\alpha_0 \mathcal{A}_i^{(0)}(s)\mathbf{e},\ \alpha_1 \mathcal{A}_i^{(1)}(s)\mathbf{e},\ \alpha_2 \mathcal{A}_i^{(2)}(s)\mathbf{e} \right),$$

$$\mathcal{M}_i(s) = \begin{pmatrix} I & \mathcal{A}_i^{(0)}(s)\Gamma_0(\boldsymbol{\tau} \otimes \boldsymbol{\beta}) & \mathcal{A}_i^{(0)}(s)\Gamma_0(\boldsymbol{\tau} \otimes \boldsymbol{\beta})\mathcal{A}_i^{(1)}(s)(\boldsymbol{T}_0 \otimes I_M) \\ O & I & \mathcal{A}_i^{(1)}(s)(\boldsymbol{T}_0 \otimes I_M) \\ O & O & I \end{pmatrix},$$

$$\mathcal{N}_i(s) = \begin{pmatrix} i\alpha_0 \mathcal{A}_i^{(0)}(s) & O & O \\ O & \mathcal{A}_i^{(1)}(s)(I_K \otimes \boldsymbol{S}_0 \boldsymbol{\beta} + (i-1)\alpha_1 I_{KM}) & O \\ \mathcal{A}_i^{(2)}(s)\boldsymbol{S}_0 \boldsymbol{\gamma} & O & (i-1)\alpha_2 \mathcal{A}_i^{(2)}(s) \end{pmatrix},$$

$$\mathcal{A}_i^{(0)}(s) = (sI + (i+1)\alpha_0 I - \Gamma)^{-1},$$

$$\mathcal{A}_i^{(1)}(s) = (sI + i\alpha_1 I - T \oplus S)^{-1},\ \ \mathcal{A}_i^{(2)}(s) = (sI + i\alpha_2 I - S)^{-1}.$$

Here $\boldsymbol{a}'$ denotes the transpose of the vector $\boldsymbol{a}$.

Proof To derive an expression for the LST $z(s)$, we use the method of collective marks, see, e.g., Kesten and Runnenburg (1956) and van Dantzig (1955). Let us tag an arbitrary arriving customer and monitor its stay in the system. According to the idea of the method of collective marks, $z(s)$ has the meaning of the probability that no catastrophe from some virtual stationary Poisson flow of catastrophes with the rate s arrives during the waiting time of the tagged customer. A catastrophe does not have any physical meaning and does not have any impact on the behavior of the queueing system. The notion of the catastrophe is used just to give the probabilistic interpretation for the LST. Analogously, components of the vector $\mathbf{z}_i^{(r)}(s)$ have the meaning of the probability that no catastrophe occurs during

the waiting time of the tagged customer conditional it arrives when there are i customers in the system, the status of the server is r and the states of the underlying Markov processes of the vacation time (if $r = 0$), the service time and the MAT (if $r = 1$), or the service time (if $r = 2$) have the corresponding values.

Taking into account the probabilistic meaning of the conditional LSTs $\mathbf{z}_i^{(r)}(s)$ and the formula of total probability, the expressions (4) and (5) for the LSTs $\mathbf{z}_i^{(r)}(s)$, $r = 0, 1, 2$, can be derived. In the derivation of these expressions for the column vectors $\mathbf{z}_i(s)$ it is taken into account that the number of customers in the system can decrease by one not only due to customer's service completion. The decrease can be also caused by the escape of some waiting customer from the system due to impatience. If the departing customer is one of the customers, which arrived to the system earlier than the tagged customer, this departure causes the reduction of the queue length before the tagged customer. But the departing customer can be the tagged customer itself. In this case, the customer leaves the system permanently. Its waiting time is finished and the probability of no catastrophe arrival during the residual waiting time is equal to 1. This explains the presence of the vectors $\mathbf{g}_i(s)$ in the right hand sides of relations (5). □

Corollary 1 *The average waiting time of an arbitrary customer, Z_1, can be computed as follows:*

$$Z_1 = -z'(s)|_{s=0} = \frac{1}{\lambda}\Bigg[\sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0)(D_1\mathbf{e}_{\bar{W}} \otimes I_R)\tilde{\mathbf{Z}}_i^{(0)}$$
$$+ \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,1)(D_1\mathbf{e}_{\bar{W}} \otimes I_{KM})\tilde{\mathbf{Z}}_i^{(1)} + \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,2)(D_1\mathbf{e}_{\bar{W}} \otimes I_M)\tilde{\mathbf{Z}}_i^{(2)} \Bigg]$$

where the column vectors $\tilde{\mathbf{Z}}_i^{(r)} = -(\mathbf{z}_i^{(r)}(s))'|_{s=0}$, $i \geq 0$, $r = 0, 1, 2$, are defined in the following way. These column vectors constitute the column vectors

$$\tilde{\mathbf{Z}}_i = \left(\tilde{\mathbf{Z}}_i^{(0)},\ \tilde{\mathbf{Z}}_i^{(1)},\ \tilde{\mathbf{Z}}_i^{(2)} \right)^T,$$

which are recursively computed by

$$\tilde{\mathbf{Z}}_0 = \left(\tilde{\mathbf{Z}}_0^{(0)},\ \boldsymbol{0}^T,\ \boldsymbol{0}^T \right)^T,$$

$$\tilde{\mathbf{Z}}_i = \tilde{\mathcal{M}}_i(\mathbf{g}_i(0) + \mathcal{N}_i(0)z_{i-1}(0)) + \mathcal{M}_i(0)(\tilde{\mathbf{g}}_i + \tilde{\mathcal{N}}_i z_{i-1}(0) + \mathcal{N}_i(0)\tilde{\mathbf{Z}}_{i-1}),\ i \geq 1,$$

where

$$\tilde{\mathbf{Z}}_0^{(0)} = \tilde{\mathcal{A}}_0^{(0)}(\Gamma_0 + \alpha_0\mathbf{e}),$$

$$\tilde{\mathbf{g}}_i = \left(\alpha_0\tilde{\mathcal{A}}_i^{(0)}\mathbf{e},\ \alpha_1\tilde{\mathcal{A}}_i^{(1)}\mathbf{e},\ \alpha_2\tilde{\mathcal{A}}_i^{(2)}\mathbf{e} \right)^T,$$

$$\tilde{\mathcal{M}}_i(0) = \begin{pmatrix} I & \tilde{\mathcal{A}}_i^{(0)}(0)\Gamma_0(\boldsymbol{\tau} \otimes \boldsymbol{\beta}) & \mathcal{L}_i \\ O & I & \tilde{\mathcal{A}}_i^{(1)}(0)(\boldsymbol{T}_0 \otimes I_M) \\ O & O & I \end{pmatrix},$$

$$\mathcal{L}_i = (\tilde{\mathcal{A}}_i^{(0)}(0)\Gamma_0(\boldsymbol{\tau} \otimes \boldsymbol{\beta})\mathcal{A}_i^{(1)}(0) + \mathcal{A}_i^{(0)}(0)\Gamma_0(\boldsymbol{\tau} \otimes \boldsymbol{\beta})\tilde{\mathcal{A}}_i^{(1)}(0))(\boldsymbol{T}_0 \otimes I_M),$$

$$\tilde{\mathcal{N}}_i = \begin{pmatrix} i\alpha_0\tilde{\mathcal{A}}_i^{(0)} & O & O \\ O & \tilde{\mathcal{A}}_i^{(1)}(I_K \otimes \boldsymbol{S}_0\boldsymbol{\beta} + (i-1)\alpha_1 I_{KM}) & O \\ \tilde{\mathcal{A}}_i^{(2)}\boldsymbol{S}_0\boldsymbol{\gamma} & O & (i-1)\alpha_2\tilde{\mathcal{A}}_i^{(2)} \end{pmatrix},$$

$$\tilde{\mathcal{A}}_i^{(0)} = ((i+1)\alpha_0 I - \Gamma)^{-2}, \tilde{\mathcal{A}}_i^{(1)} = (i\alpha_1 I - T \oplus S)^{-2},\ \ \tilde{\mathcal{A}}_i^{(2)} = (i\alpha_2 I - S)^{-2}.$$

The presented above results concern the waiting time of an *arbitrary* customer, including a customer which is lost due to impatience. Let us consider now the waiting time distribution of an *arbitrary customer, which is not lost in the system*, and let $V(x)$ be the distribution function of waiting time of such a customer and $v(s) = \int\limits_0^\infty e^{-sx} dV(x)$, $Re\ s > 0$, be its LST. Let us stress that we assume that $V(x)$ is not the distribution function of the waiting time of a customer conditional it is not lost in the system. $V(x)$ is the probability that an arbitrary customer is not lost in the system and its waiting time is less than x. Denote by $\mathbf{v}_i^{(r)}(s)$ the column vectors of the LSTs of the waiting time of an arbitrary customer, which is not lost in the system, conditional it arrives to the system when there are i customers in the system, the status of the server is r and the states of the underlying Markov processes are fixed.

Theorem 3 *The Laplace-Stieltjes transform $v(s)$ can be computed as follows:*

$$v(s) = \frac{1}{\lambda}\Bigg[\sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0)(D_1 \mathbf{e}_{\bar{W}} \otimes I_R)\boldsymbol{v}_i^{(0)}(s) + \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,1)(D_1 \mathbf{e}_{\bar{W}} \otimes I_{KM})\boldsymbol{v}_i^{(1)}(s) + \\ + \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,2)(D_1 \mathbf{e}_{\bar{W}} \otimes I_M)\boldsymbol{v}_i^{(2)}(s) \Bigg]$$

where the column vectors $\boldsymbol{v}_i^{(r)}(s)$, $r = 0, 1, 2$, constitute the column vectors $\boldsymbol{v}_i(s)$ of dimension $R + KM + M$

$$\mathbf{v}_i(s) = \left(\mathbf{v}_i^{(0)}(s),\ \mathbf{v}_i^{(1)}(s),\ \mathbf{v}_i^{(2)}(s) \right)^T,\ i \geq 0,$$

that are recursively computed by

$$\boldsymbol{v}_0(s) = \begin{pmatrix} (sI - \Gamma)^{-1}\Gamma_0 \\ \mathbf{e}_{(K+1)M} \end{pmatrix},\quad \boldsymbol{v}_i(s) = \mathcal{M}_i(s)\mathcal{N}_i(s)\boldsymbol{v}_{i-1}(s)),\ i \geq 1.$$

Corollary 2 *The average waiting time V_1 of an arbitrary customer, which is not lost in the system, can be computed as follows:*

$$V_1 = -v'(s)|_{s=0} = \frac{1}{\lambda}\Bigg[\sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0)(D_1 \mathbf{e}_{\bar{W}} \otimes I_R)\tilde{\mathbf{V}}_i^{(0)} \\ + \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,1)(D_1 \mathbf{e}_{\bar{W}} \otimes I_{KM})\tilde{\mathbf{V}}_i^{(1)} + \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,2)(D_1 \mathbf{e}_{\bar{W}} \otimes I_M)\tilde{\mathbf{V}}_i^{(2)} \Bigg]$$

where the column vectors $\tilde{\mathbf{V}}_i^{(r)} = -(\boldsymbol{v}_i^{(r)}(s))'|_{s=0}$, $i \geq 0$, $r = 0, 1, 2$, are defined in the following way. These column vectors constitute the column vectors

$$\tilde{\mathbf{V}}_i = \left(\tilde{\mathbf{V}}_i^{(0)},\ \tilde{\mathbf{V}}_i^{(1)},\ \tilde{\mathbf{V}}_i^{(2)} \right)^T,$$

which are recursively computed by

$$\tilde{\mathbf{V}}_0 = \left(\tilde{\mathbf{V}}_0^{(0)},\ \boldsymbol{0}^T,\ \boldsymbol{0}^T \right)^T,$$

$$\tilde{\boldsymbol{V}}_i = \tilde{\mathcal{M}}_i\mathcal{N}_i(0)\boldsymbol{v}_{i-1}(0) + \mathcal{M}_i(0)(\tilde{\mathcal{N}}_i\boldsymbol{v}_{i-1}(0) + \mathcal{N}_i(0)\tilde{\boldsymbol{V}}_{i-1}),\ i \geq 1,$$

where

$$\tilde{\mathbf{V}}_0^{(0)} = \tilde{\mathcal{A}}_0^{(0)}\alpha_0 \mathbf{e}.$$

Corollary 3 *The probability $P^{(loss)}$ that an arbitrary customer will be lost is computed as follows:*

$$P^{(loss)} = 1 - \frac{1}{\lambda}\Big[\sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0)(D_1 \mathbf{e}_{\bar{W}} \otimes I_R)\boldsymbol{v}_i^{(0)}(0)$$

$$+ \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,1)(D_1 \mathbf{e}_{\bar{W}} \otimes I_{KM})\boldsymbol{v}_i^{(1)}(0) + \sum_{i=1}^{\infty} \boldsymbol{\pi}(i,2)(D_1 \mathbf{e}_{\bar{W}} \otimes I_M)\boldsymbol{v}_i^{(2)}(0) \Big].$$

Proof According to the definition and probabilistic meaning of the LST $v(s)$, the value $v(0)$ is a probability that an arbitrary customer will not be lost during its waiting time. Because an arbitrary customer may be lost due to impatience only during its waiting time, the statement of Corollary 3 immediately follows from Theorem 3.

Note, that in Section 5 we got another two expressions for the probability $P^{(loss)}$. Availability of the three different formulas is helpful for the control of analytical derivations and computer implementation. □

7 Visiting Time Distribution and Perspectives of Application to the Analysis of Polling System

Mention, that the vacation time for the tagged vacation queue consists of a sequence of phases representing the visiting times to other buffers alternating with possible switching times between the queues. Thus, we can conclude that: (i) assumption made in our paper that the vacation time has the PH distribution ideally fits to possible application of results of the analysis of a vacation queue to the analysis of a polling system; (ii) the presented above analysis has to be complemented by the analysis of a visiting time in the considered queue. The visiting time in the vacation queueing model under consideration is the time interval since the epoch of vacation completion till the moment of the next vacation beginning. Let us remind that in this model the next vacation begins when the system becomes empty or the MATs expires and, then, service of a customer being in service at the moment of the MATs expiration finishes, whichever occurs earlier. Account of the residual service time after the MAT completion essentially complicates analysis comparing to the discipline with service termination at the MAT completion epoch because it is necessary to simultaneously monitor two non-exponentially distributed random variables: the residual service time and residual MAT. Assumption that the full service time and full MAT have PH distribution helps in implementation of this analysis. If al least one of these distributions is arbitrary, the analysis does not seem to be feasible.

Let $\boldsymbol{\kappa}_j$ be the row vector of size $\bar{W}$ the ν-th component of which is equal to the probability that the state of the underlying Markov process of customers arrival is equal to ν and j customers present in the system at an arbitrary vacation completion moment, $\nu = \overline{0, W}$, $j \geq 0$.

Lemma 2 *The vectors $\boldsymbol{\kappa}_j$, $j \geq 0$, are computed by the formula*

$$\boldsymbol{\kappa}_j = \frac{\boldsymbol{\pi}(j,0)(I_{\bar{W}} \otimes \boldsymbol{\Gamma}_0)}{\sum_{k=0}^{\infty} \boldsymbol{\pi}(k,0)(\boldsymbol{e}_{\bar{W}} \otimes \boldsymbol{\Gamma}_0)}, \quad j \geq 0.$$

Proof of Lemma 2 is straightforward because j customers can present in the system at an arbitrary vacation completion moment if j customers present in the system at an arbitrary moment when the server is on vacation and vacation time expires.

Theorem 4 *If customers are patient when the MAT is not finished (i.e., $\alpha_1 = 0$), the Laplace-Stieltjes transform $\psi(u)$, $Re\ u > 0$, of the visiting time can be computed as follows:*

$$\psi(u) = \sum_{k=0}^{\infty} \boldsymbol{\kappa}_k (I_{\bar{W}} \otimes \boldsymbol{\tau} \otimes \boldsymbol{\beta}) \boldsymbol{\psi}(u,k) \tag{6}$$

where the column vectors $\boldsymbol{\psi}(u,k)$, $k \geq 0$, of size $\bar{W}KM$ are defined by the formula

$$\boldsymbol{\psi}(u,k) = \mathbf{F}^k(u)\boldsymbol{e}_{\bar{W}KM} +$$

$$+ (I - \mathbf{F}^k(u))(uI - (D_0 + D_1) \oplus T \oplus (S + \boldsymbol{S}_0\boldsymbol{\beta}))^{-1}(\boldsymbol{e}_{\bar{W}} \otimes \boldsymbol{T}_0 \otimes (uI - S)^{-1}\boldsymbol{S}_0) \tag{7}$$

and the matrix $\mathbf{F}(u)$ is defined as the minimal non-negative solution to the quadratic matrix equation:

$$I_{\bar{W}K} \otimes \boldsymbol{S}_0\boldsymbol{\beta} - (uI - D_0 \oplus T \oplus S)\mathbf{F}(u) + (D_1 \otimes I_{KM})\mathbf{F}^2(u) = O. \tag{8}$$

Proof To derive an expression for the *LST* $\psi(u)$, we again use the method of collective marks. According to the idea of the method of collective marks, $\psi(u)$ has the meaning of the probability that no catastrophe from a virtual stationary Poisson flow of catastrophes with the rate u arrives during the visiting time. Formula (6) evidently follows from the formula of total probability if we take into account the probabilistic meaning of the vector $\boldsymbol{\psi}(u,k)$. The entries of this column vector of size $\bar{W}KM$ define the probability that no catastrophe arrives during the rest of the visiting time conditional on the fact that, at the given moment, k customers present in the system, the server provides service, the MAT is not expired and the states of underlying Markov chains of arrival, MAT and service processes have the corresponding values. Therefore, to finish the proof, we need to derive formulas (6) and (7).

Using the probabilistic meaning of the vectors $\boldsymbol{\psi}(u,k)$, $k \geq 1$, and the law of total probability, it is not difficult to derive the following recursive equations:

$$\boldsymbol{\psi}(u,k) = \int_0^{\infty} e^{-(uI - D_0 \oplus T \oplus S)t}[D_1 \otimes I_{KM}\boldsymbol{\psi}(u,k+1) + I_{\bar{W}K} \otimes (\mathbf{S}_0\boldsymbol{\beta})\boldsymbol{\psi}(u,k-1)$$

$$+ \mathbf{e}_{\bar{W}} \otimes \mathbf{T}_0 \otimes (uI - S)^{-1}\mathbf{S}_0]dt,\ k \geq 2. \tag{9}$$

The matrix

$$\mathbf{H}(u) = -(uI - D_0 \oplus T \oplus S)$$

is a sub-generator with strict domination of the diagonal entries. Consequently, it is non-singular, the real parts of its eigenvalues are negative and the following relation is true:

$$\int_0^{\infty} e^{\mathbf{H}(u)t}dt = (-\mathbf{H}(u))^{-1}.$$

Therefore, by introducing, for brevity, the following notation:

$$\mathbf{r}(u) = \mathbf{e}_{\bar{W}} \otimes \mathbf{T}_0 \otimes (uI - S)^{-1}\mathbf{S}_0,\ \hat{\mathbf{D}}_1 = D_1 \otimes I_{KM},\ \mathbf{C}_1 = I_{\bar{W}K} \otimes (\mathbf{S}_0\boldsymbol{\beta}),$$

system (9) can be rewritten in the form:

$$\mathbf{H}(u)\boldsymbol{\psi}(u,k) + \hat{\mathbf{D}}_1\boldsymbol{\psi}(u,k+1) + \mathbf{C}_1\boldsymbol{\psi}(u,k-1) + \mathbf{r}(u) = \mathbf{0},\ k \geq 2. \tag{10}$$

Analogously, it is possible to derive the equation

$$\mathbf{H}(u)\boldsymbol{\psi}(u,1)+\hat{\mathbf{D}}_1\boldsymbol{\psi}(u,2)+\mathbf{c}_2+\mathbf{r}(u)=\mathbf{0} \tag{11}$$

where $\mathbf{c}_2=\mathbf{e}_{\bar{W}}\otimes\mathbf{S}_0$.

On noting that $\mathbf{c}_2=\mathbf{C}_1\mathbf{e}_{\bar{W}KM}$ and setting

$$\boldsymbol{\psi}(u,0)=\mathbf{e}_{\bar{W}KM}$$

we combine Eqs. 10 and 11 into the following inhomogeneous system of the vector difference equations of the second order for the vectors $\boldsymbol{\psi}(u,k),\ k\geq 1$:

$$\mathbf{H}(u)\boldsymbol{\psi}(u,k)+\hat{\mathbf{D}}_1\boldsymbol{\psi}(u,k+1)+\mathbf{C}_1\boldsymbol{\psi}(u,k-1)+\mathbf{r}(u)=\mathbf{0},\ k\geq 1. \tag{12}$$

By analogy with the known way for solving the scalar counterpart of such equations, we will try to find solution to system (12) in the following form:

$$\boldsymbol{\psi}(u,k)=\mathbf{F}^k(u)\mathbf{x}(u)+\mathbf{A}(u)\mathbf{r}(u),\ k\geq 1, \tag{13}$$

where $\mathbf{F}(u)$ and $\mathbf{A}(u)$ are still unknown matrices and $\mathbf{x}(u)$ is an unknown vector. By substituting (13) into (12), after performing some transformations, we get the following system of equations for $k\geq 1$:

$$(\mathbf{H}(u)\mathbf{F}(u)+\hat{\mathbf{D}}_1\mathbf{F}^2(u)+\mathbf{C}_1)\mathbf{F}^{k-1}(u)\mathbf{x}(u)+((\mathbf{H}(u)+\hat{\mathbf{D}}_1+\mathbf{C}_1)\mathbf{A}(u)+I)\mathbf{r}(u)=\mathbf{0}. \tag{14}$$

It is easy to see that the expression $\mathbf{H}(u)\mathbf{F}(u)+\hat{\mathbf{D}}_1\mathbf{F}^2(u)+\mathbf{C}_1$ is equal to zero matrix because it is assumed in the theorem statement that the matrix $\mathbf{F}(u)$ is defined as the minimal non-negative solution to quadratic matrix equation (8).

Therefore, Eq. 14 reduce to the form

$$((\mathbf{H}(u)+\hat{\mathbf{D}}_1+\mathbf{C}_1)\mathbf{A}(u)+I)\mathbf{r}(u)=O.$$

In particular, this equation becomes identity if

$$(\mathbf{H}(u)+\hat{\mathbf{D}}_1+\mathbf{C}_1)\mathbf{A}(u)+I=O.$$

In this case, the still unknown matrix $\mathbf{A}(u)$ is defined by the formula

$$\mathbf{A}(u)=-(\mathbf{H}(u)+\hat{\mathbf{D}}_1+\mathbf{C}_1)^{-1}. \tag{15}$$

Note that the inverse matrix exists because the inverted matrix is a sub-generator with strict domination of diagonal entries. Thus, only the vector $\mathbf{x}(u)$ remains unknown. To derive an expression for this vector, we substitute the vectors $\boldsymbol{\psi}(u,k)$ of form (13) to equation of the system (12) with $k=1$. Taking into account equation (8), we reduce this equation to the following one:

$$\mathbf{C}_1(\mathbf{e}_{\bar{W}KM}-\mathbf{A}(u)\mathbf{r}(u)-\mathbf{x}(u))=\mathbf{0},$$

which is fulfilled if the vector $\mathbf{x}(u)$ is chosen as

$$\mathbf{x}(u)=\mathbf{e}_{\bar{W}KM}-\mathbf{A}(u)\mathbf{r}(u). \tag{16}$$

By substituting (15) and (16) to (13) we get (7). Theorem 7 is proved. □

Remark 2 It is easy to understand that the entries of the matrix $\mathbf{F}(u)$ define the probability that a catastrophe does not arrive, the MAT does not expire and the components $\{\nu_t,\chi_t,\eta_t\}$ of the Markov chain ζ_t make the corresponding transitions at the time interval during which the number of customers in the system decreases by 1. This observation makes formula (7) more transparent. The first summand in the right hand side of Eq. 7 is a vector, components of which define the probabilities that no catastrophe arrives during the rest of the visiting time conditional of the fact that, at the given moment, k customers present in the system,

the server provides service, the MAT is not expired, the states of the processes $\{\nu_t, \chi_t, \eta_t\}$ have the corresponding values and the visiting time will be finished because the system will become empty. The second summand in the right hand side of Eq. 7 is a vector, components of which define analogous probabilities when the visiting time will be finished after the maximum attendance time expiration and the finish of the residual service time.

Remark 3 Using results by M. Neuts from the book Neuts (1981), it is possible to show that, for any nonnegative u, the minimal non-negative solution to quadratic matrix equation (7) exists and its maximal eigenvalue is less than 1. Thus, $\mathbf{F}^k(u)$ tends to zero matrix when k approaches infinity. Then, it follows from Eq. 7 that

$$\boldsymbol{\psi}(u,\infty) = \lim_{k\to\infty} \boldsymbol{\psi}(u,k) = \mathbf{A}(u)\mathbf{r}(u)$$

what coincides with the formula

$$\boldsymbol{\psi}(u,\infty) = (uI - (D_0 + D_1) \oplus T \oplus (S + \mathbf{S}_0\boldsymbol{\beta}))^{-1}\mathbf{e}_{\bar{W}} \otimes \mathbf{T}_0 \otimes (uI - S)^{-1}\mathbf{S}_0$$

which can be derived in the direct way based on the obvious consideration that if k is huge, the visiting time will be finished by expiration of the MAT and the residual service time, but not by emptying the system. This consideration essentially simplifies the derivation because it in not necessary to monitor the number of customers which receive service before the MAT expires. It is necessary only to monitor the state of the service underlying process to compute the Laplace-Stieltjes transform of the residual service time after the MAT expires.

Remark 4 As it was discussed above, the LST of the distribution of the visiting time given by formula (6) does not separate the visiting times finished via emptying the system or via the MAT expiration. Let now $\psi^{empty}(u)$ be the LST of the distribution of the visiting time that is finished by emptying the system and $\psi^{expire}(u)$ be the LST transform of the distribution of the visiting time that is finished after the MAT expiration and the finish of the residual service time. It can be verified that these LST transforms are given by the formulas

$$\psi^{empty}(u) = \sum_{k=0}^{\infty} \boldsymbol{\kappa}_k (I_{\bar{W}} \otimes \boldsymbol{\tau} \otimes \boldsymbol{\beta})\mathbf{F}^k(u)\mathbf{e}_{\bar{W}KM},$$

$$\psi^{expire}(u) = \sum_{k=1}^{\infty} \boldsymbol{\kappa}_k (I_{\bar{W}} \otimes \boldsymbol{\tau} \otimes \boldsymbol{\beta})(I - \mathbf{F}^k(u))(uI - (D_0 + D_1) \oplus T \oplus (S + \mathbf{S}_0\boldsymbol{\beta}))^{-1} \times$$
$$\times (\mathbf{e}_{\bar{W}} \otimes \mathbf{T}_0 \otimes (uI - S)^{-1}\mathbf{S}_0).$$

Corollary 4 *The probabilities P^{empty} and P^{expire} that an arbitrary visiting time finishes due to emptying the system and due to the MAT expiration, correspondingly, are computed as*

$$P^{empty} = \psi^{empty}(0), \quad P^{expire} = \psi^{expire}(0).$$

Note that formula (6) defines the LST transform of the distribution of an arbitrary visiting time, including possible zero visiting time when the system is empty upon vacation completion and the server immediately takes one more vacation. The LST transform $\psi^{non-zero}(u)$ of the distribution of the visiting time conditional on the fact that zero visiting times are not accounted is given by formula

$$\psi^{non-zero}(u) = \frac{\sum_{k=1}^{\infty} \boldsymbol{\kappa}_k (I_{\bar{W}} \otimes \boldsymbol{\tau} \otimes \boldsymbol{\beta})\boldsymbol{\psi}(u,k)}{1 - \boldsymbol{\kappa}_0 \boldsymbol{e}_{\bar{W}}}.$$

Using (7), it is possible to verify that the evident from probabilistic considerations relation $\boldsymbol{\psi}(0,k) = \boldsymbol{e}_{\bar{W}KM}$ *holds for all* $k,\ k \geq 1$.

Corollary 5 *The average visiting time* Ψ_1 *is given by the formula*

$$\Psi_1 = -\psi'(0) = -\sum_{k=1}^{\infty} \boldsymbol{\kappa}_k (I_{\bar{W}} \otimes \boldsymbol{\tau} \otimes \boldsymbol{\beta}) \boldsymbol{\psi}'(0,k)$$

where the derivatives $\boldsymbol{\psi}'(0,k)$ *are given by the formula*

$$\boldsymbol{\psi}'(0,k) = (I - \mathbf{F}^k(0))[\boldsymbol{e}_{\bar{W}} \otimes (T^{-1}\boldsymbol{e}_K) \otimes \boldsymbol{e}_M + (T \oplus (S + \boldsymbol{S}_0\boldsymbol{\beta}))^{-1}(\boldsymbol{T}_0 \otimes (-S)^{-1}\boldsymbol{e}_M))].$$

Note that Ψ_1 *is the average visiting time including the visiting times which are equal to zero (because the system is empty at the vacation completion moment). The average visiting time* $\hat{\Psi}_1$ *of visits having non-zero length is obviously computed by*

$$\hat{\Psi}_1 = \frac{\Psi_1}{1 - \boldsymbol{\kappa}_0 \boldsymbol{e}}.$$

In application of the considered vacation model to the analysis of a polling model, the service processes of customers from the different buffers can be modeled by this vacation model. The dependence between these processes stems from the fact that the vacation time in the tagged buffer indeed is the sum of server's visiting times to another buffers. Having computed the mean values of these visiting times, one can assume the distribution of the vacation time in the tagged buffer as the generalized Erlangian distribution. This distribution can be considered as the particular case of the PH distribution of the service time which consists of a fixed number of sequential phases having an exponential distribution with the mean value equal to the average visiting time in the corresponding system.

8 Numerical Results

As it was mentioned in Introduction, advantages of our results over the existing in the literature are more general assumptions about the arrival process, distribution of vacation, service and MATs, account of impatience of customers and analysis of the server's visiting time. It is obvious that these generalizations are valuable from the mathematical point of view. However, their usefulness for adequate modeling of real-world systems can be shown only via the computer experiments. The first part of this section is devoted to analysis of degree of importance of these generalizations. We separately illustrate the importance of generalizations of the existing models listed in Introduction.

Account of Correlation in the Arrival Process High importance of account of correlation in the arrival process and huge errors in the prediction of performance measures of the system if the existing correlated arrival process is approximated by the stationary Poisson process is already known in the queueing literature, see, e.g. Dudin et al. (2015) and Kim et al. (2014). Results of our computations confirm this importance for the model under study as well. Here, we omit these results to reduce the size of the paper.

Experiment 1. Account of the Coefficient of Variation of the Service Time In further experiments we fix the basic MAP arrival process. Let this MAP be defined by the matrices

$$D_0 = \begin{pmatrix} -1.352 & 0.0 \\ 0.0 & -0.043875 \end{pmatrix},\ D_1 = \begin{pmatrix} 1.343 & 0.009 \\ 0.04443 & 0.019445 \end{pmatrix}. \tag{17}$$

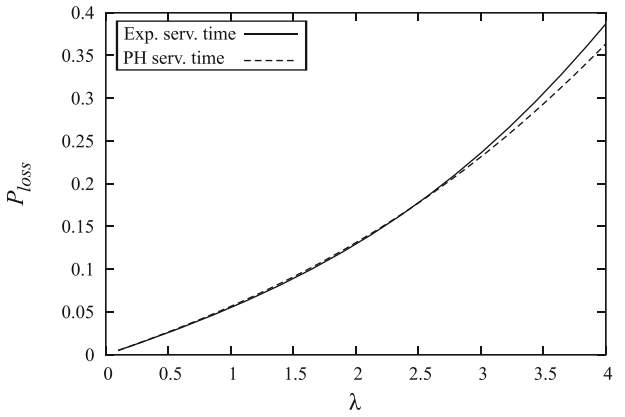

Fig. 1 Dependence of the probability of an arbitrary customer loss $P^{(loss)}$ on λ

This arrival process has the average arrival rate $\lambda = 1$, the coefficient of correlation of two successive intervals between arrivals $c_{cor} = 0.2$, and the squared coefficient of variation of the intervals between customer arrivals $c_{var} = 12.34$.

In the experiments we will show the dependence of some performance measures of the system on the average arrival rate λ. The MAP having a fixed value λ of the average arrival rate is defined by the matrices D_0 and D_1 given by Eq. 17 entries of which are multiplied by λ.

The distribution of the vacation time is assumed to be exponential with the rate 0.2. The distribution of the MAT is assumed to be exponential with the rate 0.5. The rates of customers impatience during the vacation period, the MAT and the residual service time after the MAT expiration are $\alpha_0 = 0.05$, $\alpha_1 = 0$, $\alpha_2 = 0.08$, correspondingly.

In experiment 1, we clarify the importance of account of variation of the service time. To this end, we consider two distributions of the service time with the same average service time. The first distribution is the exponential with the rate 10. The coefficient of variation of this distribution is equal to 1. The second distribution is the hyper-exponential distribution.

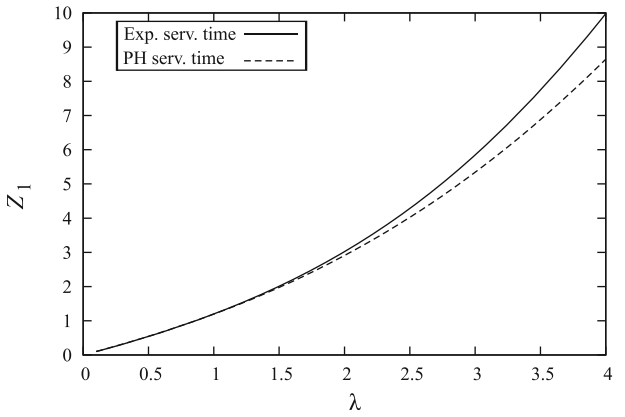

Fig. 2 Dependence of the average waiting time of an arbitrary customer Z_1 on λ

It is defined by the vector $\boldsymbol{\beta} = (\frac{27}{29}, \frac{2}{29})$ and the diagonal matrix S having the diagonal entries -30 and -1. The coefficient of variation of this distribution is equal to 3.6. The average service time for both distributions is equal to 0.1.

Figures 1, 2 and 3 illustrate the dependence of the probability $P^{(loss)}$ of an arbitrary customer loss from the system (due to impatience), the average waiting time of an arbitrary customer Z_1 and the average visiting time (conditional that the visit does not have a duration equal to 0) $\hat{\Psi}_1$ on the average arrival intensity λ for the exponential and hyper-exponential distributions of the service time.

The following conclusions follow from Figs. 1, 2 and 3:

1) Influence of the variation of the service time is not very essential. The difference between the values of Z_1 for two considered distributions of service time is only about 14 percent for $\lambda = 4$ and decreases when λ decreases.
2) For large λ, the values of $P^{(loss)}$ and Z_1 are smaller for the service time with the hyper-exponential distribution having a higher variation. This may be explained that with high probability $(\frac{27}{29})$ the service time is very short (with rate 30) and customers quickly get service and only with the small probability $(\frac{2}{29})$ an arbitrary customer has a long waiting time.
3) If the service time has the hyper-exponential distribution and we approximate this distribution by the exponential one, we *overestimate* the values of $P^{(loss)}$ and Z_1. This is quite good in practical applications because it is much worse when the approximation gives too optimistic prediction of the values of $P^{(loss)}$ and Z_1 than when the approximation is a bit pessimistic.
4) For large λ, the value of the average duration $\hat{\Psi}_1$ of visiting period is larger for the service time with the hyper-exponential distribution. This may be explained by the fact that the visiting period includes the residual service time after the MAT expires. This time is longer for the hyper-exponential distribution.

Experiment 2. Account of the Coefficient of Variation of the MAT Let the arrival process, impatience rates and distribution of the vacation time be the same as in Experiment 1. Let the service time distribution be exponential with the rate 10. We consider two distributions of the MAT with the mean value 2. The first distribution is the exponential with the rate 0.5. The coefficient of variation of this distribution is equal to 1. The second distribution is

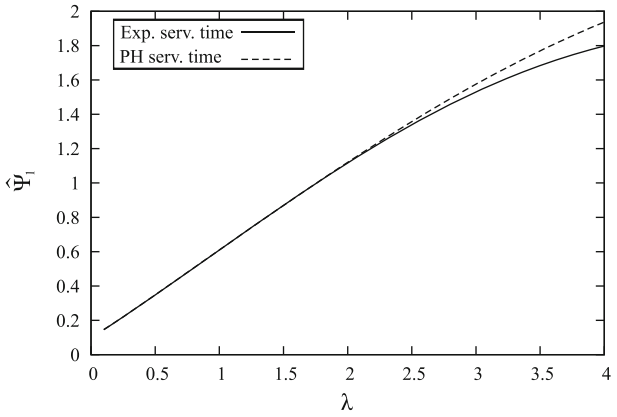

Fig. 3 Dependence of the average duration $\hat{\Psi}_1$ of visiting period on λ

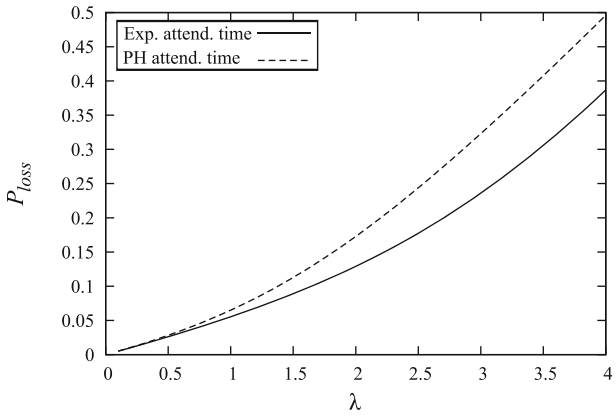

Fig. 4 Dependence of the probability of an arbitrary customer loss $P^{(loss)}$ on λ

the hyper-exponential distribution. It is defined by the vector $\boldsymbol{\theta} = (\frac{8}{9}, \frac{1}{9})$ and the diagonal matrix S having the diagonal entries -1 and -0.05. The coefficient of variation of this distribution is equal to 3.16.

Figures 4, 5 and 6 illustrate the dependence of the probability $P^{(loss)}$ of an arbitrary customer loss (due to impatience), the average waiting time of an arbitrary customer Z_1 and the average visiting time $\hat{\Psi}_1$ on the arrival intensity λ for the exponential and hyper-exponential distributions of the MAT.

The following conclusions follow from Figs. 4–6:

1) Influence of the coefficient of variation of the MAT is quite essential, especially for large values of λ.
2) If the MAT has the hyper-exponential distribution and we approximate this distribution by the exponential one, we *underestimate* the values of $P^{(loss)}$ and Z_1. This is bad in practical applications because this approximation gives too optimistic prediction of the values of $P^{(loss)}$ and Z_1.

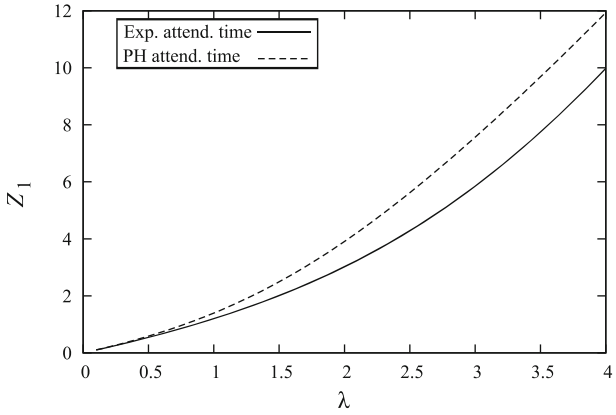

Fig. 5 Dependence of the average waiting time of an arbitrary customer Z_1 on λ

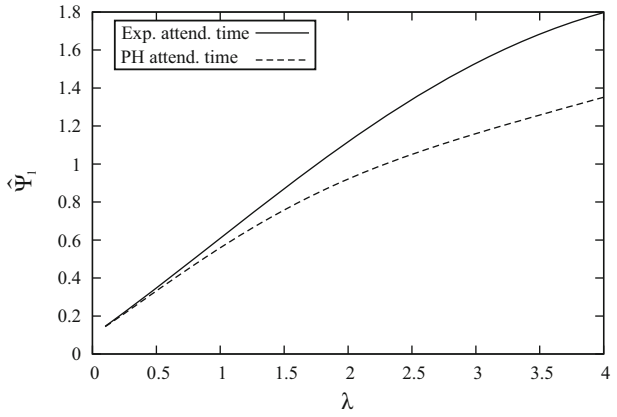

Fig. 6 Dependence of the average duration $\hat{\Psi}_1$ of visiting period on λ

3) Worse values of $P^{(loss)}$ and Z_1 in the case of the hyper-exponential distribution of the MAT are explained by Fig. 6. In the case of the hyper-exponential distribution, the mean time, during which the server continuously provides the service, is less than in the case of the exponential distribution. In turn, this evidently stems from the analysis of the parameters of the hyper-exponential distribution. With the high probability, $\frac{8}{9}$, the duration of the MAT is quite short. With the complimentary probability, $\frac{1}{9}$, the duration of the MAT is pretty long. But in the latter case the server visiting time may finish earlier than the MAT expires because the system becomes empty.

Experiment 3. Account of the Coefficient of Variation of the Vacation Time Let the arrival process, impatience rates and distribution of the MAT be the same as in Experiment 1 and the service time distribution be exponential with the rate 10. We consider two distributions of the vacation time with the mean value 5. The first distribution is the exponential with the rate 0.2. The coefficient of variation of this distribution is equal to 1. The second one is the hyper-exponential distribution. It is defined by the vector $\boldsymbol{\theta} = (\frac{16}{19}, \frac{3}{19})$ and the diagonal

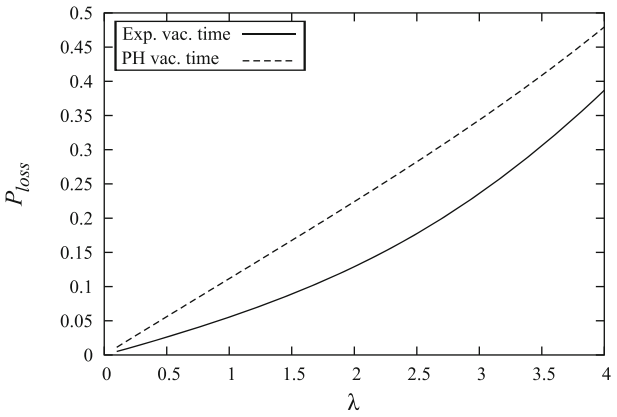

Fig. 7 Dependence of the probability of an arbitrary customer loss $P^{(loss)}$ on λ

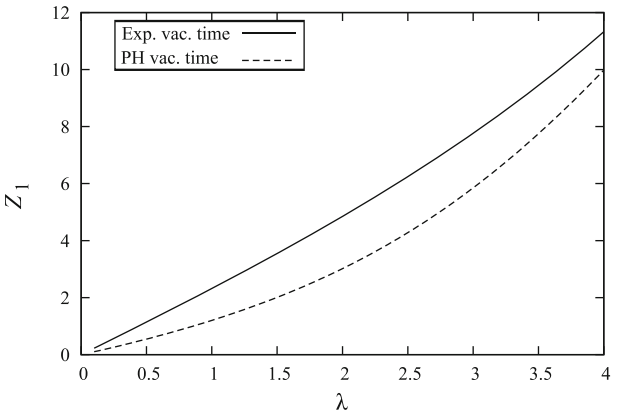

Fig. 8 Dependence of the average waiting time of an arbitrary customer Z_1 on λ

matrix S having the diagonal entries -0.8 and -0.04. The coefficient of variation of this distribution is equal to 2.65.

Figures 7, 8 and 9 illustrate the dependence of the probability $P^{(loss)}$ of an arbitrary customer loss, the average waiting time Z_1 and the average visiting time $\hat{\Psi}_1$ on the arrival intensity λ for the exponential and hyper-exponential distributions of the vacation time.

The following conclusions stem from Figs. 7–9:

1) Again, if the vacation time has the hyper-exponential distribution but this distribution is approximated by the exponential distribution with the same expectation, this leads to the *underestimation* of the loss probability and the average waiting time what is bad in the practical applications.
2) Worse values of $P^{(loss)}$ and Z_1 in the case of the hyper-exponential distribution of the vacation time are explained by Fig. 9. In the case of the hyper-exponential distribution of the vacation time, the mean time, during which the server continuously provides the service, is less than in the case of the exponential distribution. Again, this evidently stems from analysis of the parameters of the hyper-exponential distribution. With the

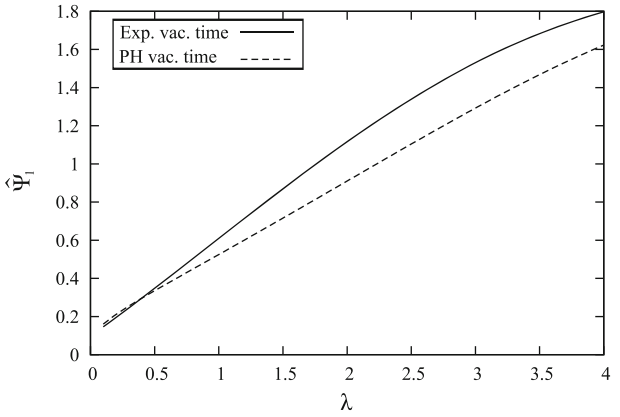

Fig. 9 Dependence of the average duration $\hat{\Psi}_1$ of visiting period on λ

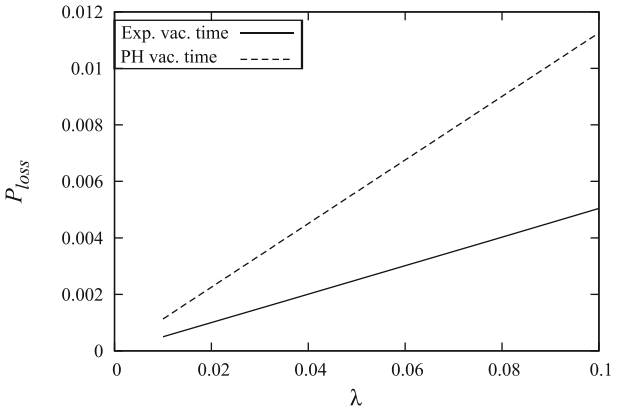

Fig. 10 Dependence of the probability of an arbitrary customer loss $P^{(loss)}$ on small λ

probability $\frac{3}{19}$, the duration of the vacation time is quite long. Therefore, many customers are lost due to impatience and often the visiting time finishes not because the MAT expires, but because the system becomes empty.

Figure 10 illustrates the same dependencies as Fig. 7, but only for small rates λ. It shows that the loss probability in the case of the hyper-exponential distribution of the vacation time is twice larger than in the case of the exponential distribution.

Experiment 4. Account of the Coefficients of Variation of the MAT and Vacation Times In this experiment, we compare the values of $P^{(loss)}$, Z_1 and $\hat{\Psi}_1$ in the cases when both distributions of the MAT and vacation times are exponential and both are hyper-exponential. The parameters of the corresponding distributions are the same as in Experiments 3 and 4.

Figures 11–13 illustrate the dependence of the values of $P^{(loss)}$, Z_1 and $\hat{\Psi}_1$ on the arrival intensity λ for these cases.

Comparing Figs. 7 and 11, 8 and 12, 9 and 13, it is easy to conclude that if both distributions of the MAT and vacation times are hyper-exponential then the approximation

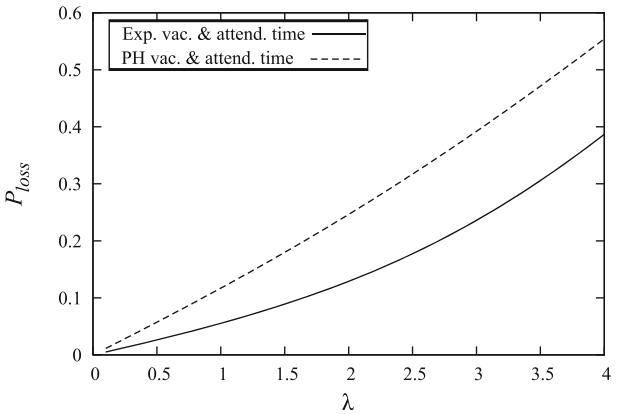

Fig. 11 Dependence of the probability of an arbitrary customer loss $P^{(loss)}$ on λ

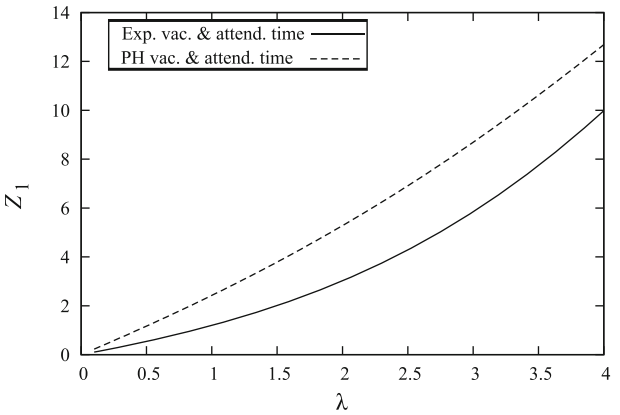

Fig. 12 Dependence of the average waiting time of an arbitrary customer Z_1 on λ

of the major performance measures of the system by their values in the case of the exponential distributions is worse comparing the case when only one MAT or vacation time is approximated via the exponentially distributed random variable. Consideration of the PH distribution instead of the exponential distribution in the model under study has not only theoretical importance but is also very valuable for exact prediction of performance measures of the real-world systems described by the queueing model under study.

Experiment 5. Account of Customers Impatience As it was mentioned in Introduction, it is obvious that the impatience phenomenon is vital in the context of the systems with the time-limited service because service of customers may be interrupted due to termination of the working period. In principle, even several vacations can occur during sojourn time of a customer and the customer may decide not to wait in the queue during a long time. Thus, effect of customers impatience must be taken into account. The impatience can be related, e.g., with the psychological factors if the customer is a human or the obsolescence of information if the customer is the information unit, etc. This subsection contains figures

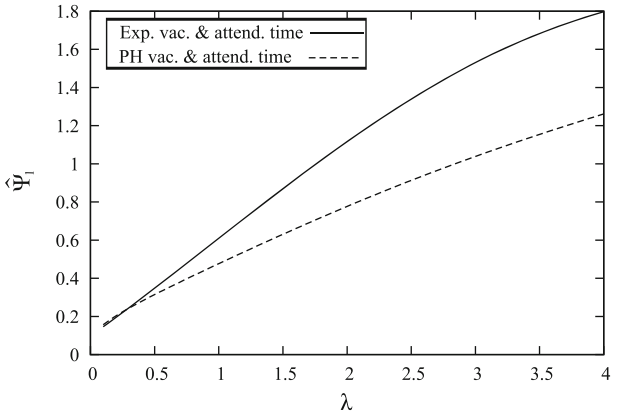

Fig. 13 Dependence of the average duration $\hat{\Psi}_1$ of visiting period on λ

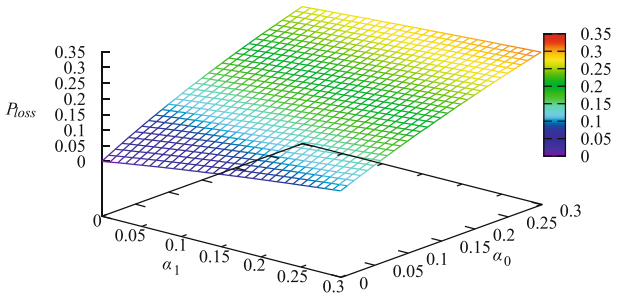

Fig. 14 Dependence of the probability of an arbitrary customer loss $P^{(loss)}$ on intensities α_0 and α_1

illustrating the dependence of performance measures of the system on the intensities of impatience.

The arrival process is defined by the matrices D_0 and D_1 given by formula (17) all entries of which are multiplied by 4. This arrival process has the average arrival rate $\lambda = 4$.

The vacation, service, and MATs have the Erlangian distribution of order 2 with the intensities of the phases equal to 1, 40, and 0.5, correspondingly. Under the fixed above parameters of the system, in particular because the service time (as well as the residual service time) has small expectation and the coefficient of variation, the value of the intensity α_2 of impatience of each customer during the interval when the server provides the residual service to a customer when the MAT expired has a very small impact. Let us fix this intensity equal to 0.08.

Figures 14 and 15 show the dependencies of the probability of an arbitrary customer loss $P^{(loss)}$ and the average waiting time of an arbitrary customer Z_1 on the intensities α_0 of impatience of each customer during the vacation time and α_1 of impatience during the MAT. It is evidently seen from these Figures that impatience of customer has the significant effect and must be taken into account.

Experiment 6. Optimization Problem Besides the transparent possible application of the analyzed vacation model to performance evaluation and capacity planning of polling systems, this vacation model can be applied, e.g., in the following situation. Some company provides the service to customers using some leased equipment, e.g., an information transmission channel. According to conditions of leasing, the maximum time of continuous using the equipment is limited. After this time expires, during a certain time the equipment is not

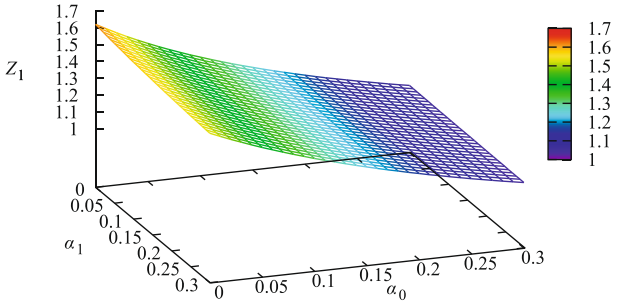

Fig. 15 Dependence of the average waiting time of an arbitrary customer Z_1 on intensities α_0 and α_1

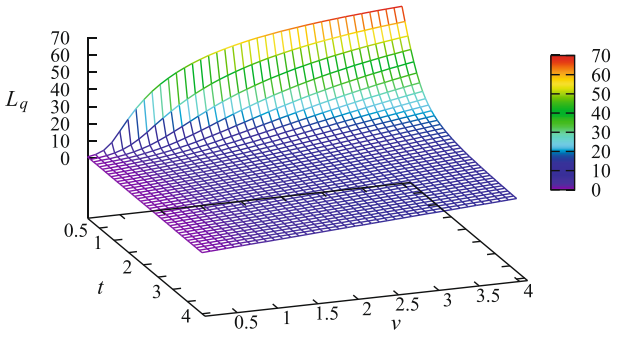

Fig. 16 Dependence of the average number of customers in the queue L_q on the parameters v_1 and t_1

available to this company. An interruption of current service at the moment of the maximum time expiration is not allowed. There are several tariff plans. These plans are distinguished by the mean duration of the maximum time of using the equipment and mean duration of time when the equipment is not available. Plans with longer maximum time of continuous using of the equipment are more expensive. Company's managers should optimally choose a tariff plan taking into account the price of a plan, the profit gained by service of customers and the possibility of the potential users loss due to long waiting for the service. The goal of the example presented in this subsection is to show some dependencies of the key performance measures of the system on the average vacation period and MAT.

Let us again the MAP arrival process is defined by the matrices given by formula (17) all entries of which are multiplied by 4.

The PH distribution of customer's service process is characterized by the vector $\boldsymbol{\beta} = (1,\ 0)$ and the matrix $S = \begin{pmatrix} -40 & 40 \\ 0 & -40 \end{pmatrix}$. The mean service time b_1 in this service process is equal to 0.05, the coefficient of variation $c_{var}^{(1)}$ is equal to 0.5.

The PH distribution of the vacation time is characterized by the vector $\boldsymbol{\gamma} = (1,\ 0)$ and the matrix $\Gamma = \begin{pmatrix} -\frac{2}{v_1} & \frac{2}{v_1} \\ 0 & -\frac{2}{v_1} \end{pmatrix}$. The mean vacation time is equal to v_1, the coefficient of variation is equal to 0.5.

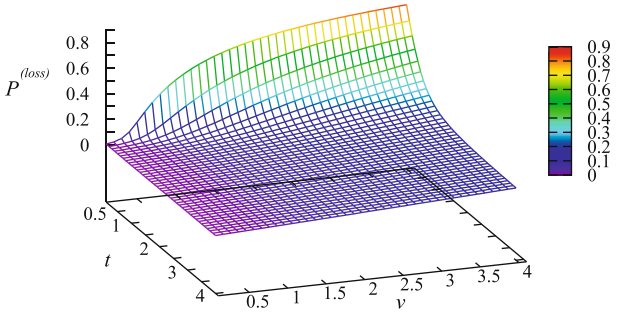

Fig. 17 Dependence of the probability of an arbitrary customer loss from the system (due to impatience) $P^{(loss)}$ on the parameters v_1 and t_1

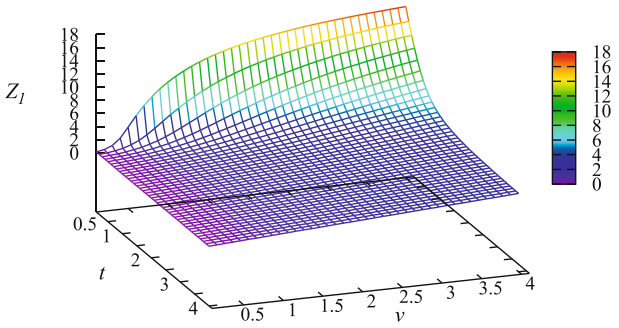

Fig. 18 Dependence of the average waiting time of an arbitrary customer Z_1 on the parameters v_1 and t_1

The PH distribution of the MAT is characterized by the vector $\boldsymbol{\tau} = (1,\ 0)$ and the matrix $T = \begin{pmatrix} -\frac{2}{t_1} & \frac{2}{t_1} \\ 0 & -\frac{2}{t_1} \end{pmatrix}$. The mean MAT is equal to t_1, the coefficient of variation is equal to 0.5.

The rates of customers impatience during the vacation period, the MAT and the residual service time after the MAT expiration are $\alpha_0 = 0.05,\ \alpha_1 = 0,\ \alpha_2 = 0.08$, correspondingly.

Let us vary parameters v_1 and t_1 over the interval $[0.1, 4]$ with step 0.1 and show the dependence of various performance measures of the system on v_1 and t_1. Figures 16, 17 and 18 illustrate the behavior of the average number of customers in the queue L_q, the probability of an arbitrary customer loss $P^{(loss)}$ and the average waiting time of an arbitrary customer Z_1.

The qualitative behavior of these system's performance measures is clear. All these measures are fairly small when the mean vacation time v_1 is small and the mean MAT t_1 is large. Figures 16–18 evaluate this behavior more exactly, quantitatively. It is evidently seen that L_q and $P^{(loss)}$ are close to zero when v_1 is small and t_1 is large. When v_1 is large, about 4, and t_1 is small, about 0.1, the values of L_q and $P^{(loss)}$ become very large, about 70 and 0.9, respectively. It is worth to mention that the surfaces for the average number of customers in the queue L_q and the average waiting time of an arbitrary customer Z_1 look very similar. This is easily explained by the fact established by means of more extensive numerical results that the famous Little formula is valid for this system in the form

$$\lambda Z_1 = L_q.$$

Figure 19 illustrates the behavior of the rate J of the server's switching on (average number of server's switching on per unit time) depending on the values of v_1 and t_1. The

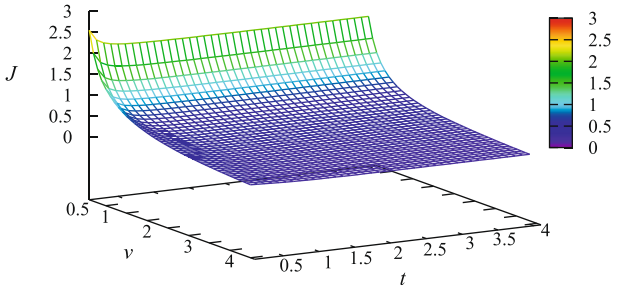

Fig. 19 Dependence of the rate J of the server's switching on on the parameters v_1 and t_1

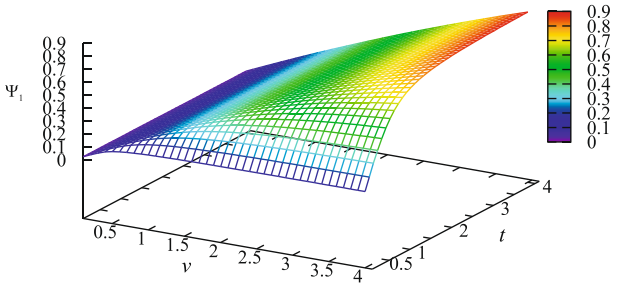

Fig. 20 Dependence of the average visiting time Ψ_1 on the parameters v_1 and t_1

value of J is more sensitive with respect to the parameter v_1. The largest value of J is achieved when both v_1 and t_1 are small.

Figures 20 and 21 illustrate the behavior of the average visiting time Ψ_1 and the average visiting time $\hat{\Psi}_1$ of visits having non-zero length.

The values Ψ_1 and $\hat{\Psi}_1$ are pretty small when t_1 is small. They grow when v_1 and t_1 increase. This is clear because the increase of v_1 implies larger number of customers in the system at the visit beginning epoch and larger chances that the visit will be finished due to the maximal attendance time expiration, not due to exhausting the queue. The increase of t_1 obviously leads to the increase of the average visit times when the visit is finished after the maximal attendance time expiration and the finish of the residual service time.

After we got information about the quantitative behavior of the main performance measures of the system, we can formulate a cost criterion for evaluation of quality of the system operation and consider optimization problem. Usually the cost criterion is defined as a profit gained from a system operation, which should be maximized, or losses of the system which should be minimized. We choose the latter option and fix the criterion in the form

$$N = N(v_1, t_1) = a\lambda P^{(loss)} + bJ + ct_1$$

where the cost coefficient a is the penalty cost for one customer loss, b is the fee for each switching off the server operation and c is the average cost paid per unit of time according to the tariff plan with the value of the average MAT t_1. In the presented above example, we fix the values of the cost coefficient as follows: $a = 2$, $b = 1$, $c = 0.2$.

The surface representing dependence of the cost criterion $N(t_1, v_1)$ on the controlled parameters v_1 and t_1 has the form given in Fig. 22.

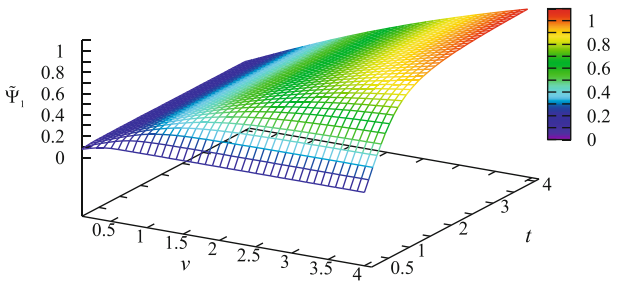

Fig. 21 Dependence of the average visiting time $\hat{\Psi}_1$ for the visits having non-zero length on the parameters v_1 and t_1

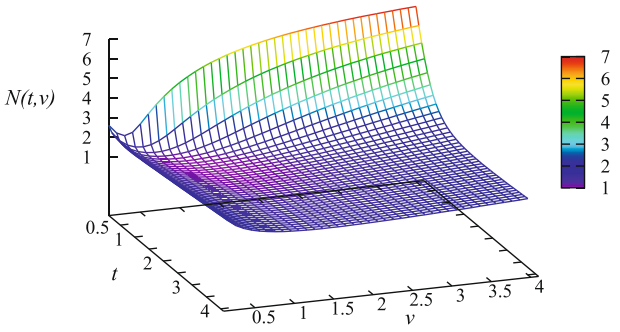

Fig. 22 Dependence of the cost criterion $N(t, v)$ on the parameters v_1 and t_1

The minimal value N^* of this cost criterion is achieved when $v_1 = 1.1$ and $t_1 = 0.9$. This minimal value is equal to 1.0184. The value of the loss probability $P^{(loss)}$ for this optimal choice of v_1 and t_1 is equal to 0.040634.

9 Conclusion

In this paper, a single-server queueing system with vacations and restriction on the continuous time of the server operation is considered. The arrival flow is described by the MAP, the distributions of the vacation time, the service and the MAT are of the phase-type. Service is not preemptive: if the MAT expires, currently provided service has to be performed completely. Customers are impatient. The individual rate of customer's leaving the system without service depends on the state of the server (the server is on the vacation, the MAT is not finished, the MAT is expired).

Condition for existence of the stationary distribution of the system states is proved, the stationary distributions of the system states, waiting times and the server visiting time are obtained. Numerical results highlight the importance of account of the coefficient of variation of vacation and MAT and show the potential applicability of the results to optimization of operation of the system with vacations and time-limited service.

Acknowledgments This research has been prepared with the support by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A3A03000523) and the support by RUDN University Program 5-100.

References

Asmussen S (2003) Applied probability and queues. Springer, New York

Boon MAA, Van der Mei RD, Winands EMM (2011) Applications of polling systems. Surv Oper Res Manag Sci 16:67–82

Boxma O, Claeys D, Gulikers L, Kella O (2015) A queueing system with vacations after N services. Nav Res Logist 62:646–658

Buchholz P, Kriege J (2017) Fitting correlated arrival and service times and related queueing performance. Queu Syst 85(3-4):337–359

Chakravarthy SR (2001) The batch Markovian arrival process: a review and future work. In: Krishnamoorthy A, Raju N, Ramaswami V (eds) Advances in probability theory and stochastic processes. Notable Publications Inc., New Jersey, pp 21–29

de Haan R, Boucherie RJ, van Ommeren RJ (2009) A polling model with an autonomous server. Queu Syst 62:279–308
de Se Silva E, Gail HR, Muntz RR (1995) Polling systems with server timeouts and their application to token passing networks. IEEE/ACM Trans Network 3:560–575
Dudin AN, Piscopo R, Manzo R (2015) Queue with group admission of customers. Comput Oper Res 61:89–99
Dudina O, Kim CS, Dudin S (2013) Retrial queueing system with Markovian arrival flow and phase type service time distribution. Comput Ind Eng 66:360–373
Frigui I, Alfa A-S (1998) Analysis of a time-limited polling system. Comput Commun 21:558–571
Gantmakher FR (1967) The matrix theory. Science, Moscow
Graham A (1981) Kronecker products and matrix calculus with applications. Ellis Horwood, Cichester
Hanbali AA, de Haan R, Boucherie RJ, van Ommeren J-KS (2012) Time-limited polling systems with batch arrivals and phase-type service times. Ann Oper Res 198:57–82
Heyman DP, Lucantoni D (2003) Modelling multiple IP traffic streams with rate limits. IEEE/ACM Trans Network 11:948–958
Katayama T (2001) Waiting time analysis for a queueing system with time-limited service and exponential timer. Nav Res Logist 48:638–651
Katayama T (2007) Analysis of a time-limited service priority queueing system with exponential timer and server vacations. Queu Syst 57:169–178
Katayama T, Kobayashi K (2007) Analysis of a nonpreemptive priority queue with exponential timer and server vacations. Perform Eval 64:495–506
Kesten H, Runnenburg JTh (1956) Priority in waiting line problems. CWI, Amsterdam
Kim CS, Klimenok V, Dudin A (2014) Optimization of guard channel policy in cellular mobile networks with account of Retrials. Comput Oper Res 43:181–190
Klemm A, Lindermann C, Lohmann M (2003) Modelling IP traffic using the batch Markovian arrival process. Perform Eval 54:149–173
Klimenok VI, Dudin AN (2006) Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. Queu Syst 54:245–259
Leung KK (1994) Cyclic-service systems with nonpreemptive, time-limited service. IEEE Trans Commun 42:2521–2524
Leung KK, Eisengerg M (1990) Queue with vacations and gated time-limited service. IEEE Trans Commun 38:1454–1462
Leung KK, Lucantoni D (1994) Two vacation models for token-ring networks where service is controlled by timers. Perform Eval 20:165–184
Lucantoni D (1991) New results on the single server queue with a batch Markovian arrival process. Commun Statist-Stoch Models 7:1–46
Neuts M (1981) Matrix-geometric solutions in stochastic models. The Johns Hopkins University Press, Baltimore
Takagi H (1990) Queueing analysis of polling models: and update. In: Takagi H (ed) Stochastic analysis of computer and communication systems, North-Holland, pp 267–318
Takagi H (1991) Queueing analysis: a foundation of performance evaluation. North-Holland
Takagi H (1997) Queueing analysis of polling models: progress in 1990-1994. In: Dshalalow JH (ed) Frontiers in queueing. CRC, Boca Raton, pp 119–146
Takagi H (2000) Analysis and applications of polling models. Performance evaluation: origins and directions. In: Harvig G, Lindeman C, Reiser M (eds) Lecture notes in computer science, pp 423–442
Tian N, Zhang ZG (2006) Vacation queueing models: theory and applications. Springer, New York
van Dantzig D (1955) Chaines de Markof dans les ensembles abstraits et applications aux processus avec regions absorbantes et au probleme des boucles. Ann de l'Inst H Pioncare 14:145–199
Vishnevski VM, Dudin AN (2017) Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks. Autom Remote Control 78:1361–1403
Vishnevsky V, Semenova O (2006) Mathematical methods to study the polling systems. Autom Remote Control 67:173–220
Vishnevsky V, Dudin A, Klimenok V, Semenova O, Shpilev S (2012) Approximate method to study $M/G/1$-type polling system with adaptive polling mechanism. Quality Technol Quant Manag 9:211–228