ANALYSIS OF UNRELIABLE MULTI-SERVER QUEUEING SYSTEM WITH BREAKDOWNS SPREAD AND QUARANTINE

Alexander Dudin[a]

Sergei Dudin[a,b]
Olga Dudina[a,b]

Konstantin Samouylov[b]

[a]Department of Applied Mathematics
and Computer Science
Belarusian State University,
4 Nezavisimosti Ave.,220030,
Minsk, Belarus

[b]Department of Applied Probability
and Informatics
RUDN University,
6 Miklukho-Maklaya st., 117198,
Moscow, Russia

Email:dudin@bsu.by

Email: dudins@bsu.by
dudina@bsu.by

Email:ksam@sci.pfu.edu.ru

KEYWORDS

Multi-server queueing system, breakdowns, quarantine, Markovian arrival flow

ABSTRACT

We consider an unreliable multi-server queue in which the rate of servers' breakdowns increases when the number of broken servers grows. To prevent quick degradation of the system, it is proposed to switch to a quarantine regime when the number of broken servers exceeds some threshold and to maintain this regime until the number of broken servers becomes less than another threshold. During the quarantine, service of customers is stopped, new breakdowns do not arrive while the broken servers continue recovering. Under the fixed values of the thresholds, behavior of the system is described by the multi-dimensional continuous time Markov chain. The steady state distribution of the chain and the key performance measures of the system are computed as the functions of the thresholds. Possibility of the optimal choice of the thresholds providing the minimal value of an economical criterion is numerically illustrated.

INTRODUCTION

Queueing theory provides powerful mathematical and algorithmic tool for analysis of a variety systems where a certain scarce resource is shared between the competitive users who generate requests at random moments. As an example of such a resource we can mention the channels and servers of a telecommunication system or operators and equipment of a contact center. Important feature of many queueing systems is the unreliability of the servers, i.e., possibility of their failure. Such systems are called unreliable queueing systems. As the first papers devoted to analysis of the multi-server unreliable queueing systems, the papers (Mitrani and Avi-Itzhak 1968) and (Neuts and Lucantoni 1979) deserve to be mentioned. In these papers, the multi-server unreliable queueing systems with identical servers, the stationary Poisson arrival processes of customers and breakdowns and the exponential distribution of customers service time and servers recovering time are considered. Behavior of the systems is described by the two-dimensional Markov chains where one component defines the number of customers in the system and the second component is the number of operable (non-broken) servers. Such Markov chains were analysed using the partial generating functions in (Mitrani and Avi-Itzhak 1968) and the matrix analytic approach in (Neuts and Lucantoni 1979).

Assumption about the stationary Poisson arrival processes (in such a process, the inter-arrival times are independent identical exponentially distributed random variables) is not realistic for queueing systems describing operation of modern telecommunication networks where information flows exhibit significant burstiness and correlation. As more realistic model of information flows, the model of the Markovian Arrival Process (MAP) was developed, see, e.g., (Chakravarthy 2001, Lucantoni 1991). Unreliable multi-server queueing systems with the MAP were considered, e.g., in (Klimenok et al. 2008, Dudin et al. 2015, Al-Begain et al. 2012).

Distinguishing feature of the model analysed in this paper is consideration of the breakdowns arrival process which, to the best of our knowledge, is not considered in the queueing literature. The most popular in the queueing literature assumption is that each operable server breaks down independently of other servers and the intensity of its breakdowns is constant, i.e., the total intensity of breakdowns is proportional to the number of *operable* servers and, therefore, it decreases

Proceedings 31st European Conference on Modelling and
Simulation ©ECMS Zita Zoltay Paprika, Péter Horák, Kata Váradi,
Péter Tamás Zwierczyk, Ágnes Vidovics-Dancs, János Péter Rádics (Editors)
ISBN: 978-0-9932440-4-9/ ISBN: 978-0-9932440-5-6 (CD)

when the number of the broken servers grows. In this paper, we make the opposite assumption, not considered in the queueing literature previously. We assume that the total intensity of breakdowns increases when the number of *broken* servers grows.

This assumption may be true in the situations when the breakdown consists of infecting the server in some database by a computer virus or the human operator in contact center by a virus. Because the servers may use common hardware, tables and indices of databases and software while the operators may be compactly located and infections spreads by airborne transmission, infection of one server may provoke the quick spread of infection to another servers. Another situation where the total intensity of breakdowns increases when the number of broken servers grows is as follows. Smooth operation of the servers in some manufacturing system is provided via implementation of certain preventive works. If these works are done by a limited team of repairmen along with recovering the broken servers, this team may have a lack of time for preventive works when the number of broken servers becomes large. In such a situation, each operable server experiences the increased load and this, along with the absence of enough preventive work, may cause more quick failure of the server.

It is clear that the increase of the total intensity of breakdowns when the number of broken servers grows may eventually lead to the full degradation of the system if a certain additional (to the routine repair works) mechanism for reducing the number of broken servers will be not applied. To propose such a mechanism, it is worth to note that an effective way to struggle against the spread of infection in medicine is to impose the quarantine (ward closure) when the number of infected persons becomes large. When the regime of quarantine is established in some entity (school class, children group in kindergarten, hospital, etc), this entity stops its routine operation and members of the entity are isolated in the maximally possible extent until the level of infection drops to admissible level. By analogy, to prevent quick spread of breakdowns (up to the complete collapse of the system), we assume that the considered queueing system may switch to quarantine regime. This regime suggests that all servers stop operation, new customers may arrive but they are not allowed to enter service and are stored to the buffer. New breakdowns do not arrive. The broken servers are repaired. Thus, in this paper, we consider the unreliable multi-server queueing system with possibility to use the quarantine regime. The strategy of using this regime is as follows. The system switches from the operable mode to quarantine regime when the number of broken servers exceeds some predefined threshold, say, M_1. The system switches back to the operable mode when the number of broken servers drops below other threshold, say M_2, where $M_2 < M_1$. The final goal of the study is to provide the way for choosing the optimal values of the thresholds M_1 and M_2 providing the best quality of the system operation.

MATHEMATICAL MODEL

We consider an N-server queueing system with an infinite buffer and Markovian arrival process of customers. The structure of the system under study is presented in Figure 1.

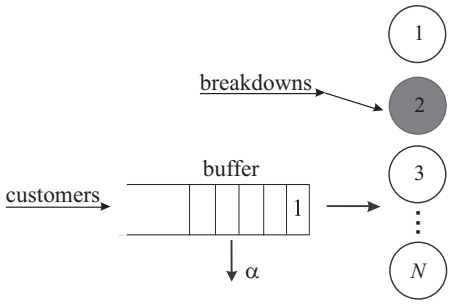

Figure 1: Queueing system under study

Customers arrive at the system according to the MAP. The advantage of the MAP pattern of the arrival process comparing to the popular in the queueing literature model of the stationary Poisson process (which is a very particular case of the MAP) is that the MAP allows to take into account correlation of the successive inter-arrival times and burstyness typical for modern telecommunication networks. Arrivals in the MAP are directed by an irreducible continuous-time Markov chain w_t, $t \geq 0$, with the finite state space $\{0, 1, ..., W\}$. The intensities of transitions that are accompanied by the arrival of k customers, are combined to the square matrices D_k, $k = 0, 1$, of size $W+1$. Formulas for computation of the average intensity λ (fundamental rate) of the MAP, the squared coefficient of the variation and the coefficient of correlation of intervals between successive arrivals can be found, e.g., in (Chakravarthy 2001).

The service time of a customer by a server has an exponential distribution with the parameter μ, $\mu > 0$. The servers can break down. The intensity of breakdowns arrival depends on the current number of broken servers. When the number of broken servers is k, the intensity of breakdowns arrival is equal to γ_k, $k = \overline{0, N}$. By default, we assume that the intensity γ_k does not decrease when k grows, i.e. $\gamma_k \leq \gamma_{k+1}$, $k = \overline{0, N-1}$. Arrival of a breakdown implies the failure of an arbitrary non-broken server. If the broken server provided service to a customer, this customer moves to a free server if it is available, or joins the queue from the head, otherwise.

Each broken server is recovering, independently of another broken servers, during the exponentially distributed time with the intensity β. After recovering, the server immediately resumes service.

Because we assume that the intensity of breakdowns γ_k, generally speaking, increases when the number of broken servers k grows, to prevent the quick spread of servers failures (and possible complete crash of the system), we suggest that when the number of broken servers becomes large the system can pass to the regime of quarantine. This means that all servers stop ser-

vice, customers, which are getting service, move to the head of the queue in the random order. During the quarantine, new customers can arrive and move to the buffer while the arrivals of the new breakdowns are ignored. The strategy of the quarantine beginning and ending is defined by the thresholds, M_1 and M_2, such as $1 \le M_1 \le N,\ 0 \le M_2 < M_1$. If the number of broken servers reaches the level M_1, the quarantine starts. It finishes when the number of broken servers falls to the level M_2. When the quarantine ends, all non-broken servers immediately resume service.

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

Our goal is to analyse the stationary behavior of the described system under the fixed parameters of the system and the thresholds M_1 and M_2 and to illustrate the effect of these thresholds. It is clear that the problem of the optimal choosing the thresholds is not the trivial one. If M_1 is large, the system benefits from late interruption of service but then it suffers from many servers being broken and not available for service provisioning. If M_1 is small, the situation when many servers are broken is avoided, however, the systems wastes time due to the frequent use of quarantines. If M_2 is small, the quarantine is long and performance of the servers is lost. If M_2 is large, the system weakly uses an opportunity to temporarily ignore new breakdowns arrival and repair the broken servers. Therefore, likely the next quarantine will be required very soon. Thus, to provide the best conditions for customers service, the careful quantitative analysis of the model is required.

PROCESS OF SYSTEM STATES

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

$$\xi_t = \{i_t, r_t, k_t, w_t\},\ t \ge 0,$$

where, during the epoch t,
- i_t is the number of customers in the system, $i_t \ge 0$;
- r_t is an indicator that specifies the state of the system: $r_t = 0$ corresponds to the operable state and $r_t = 1$ corresponds to the quarantine regime;
- k_t is the number of broken servers, $k_t = \overline{0, M_1}$;
- w_t is the state of the underlying process of the MAP, $w_t = \overline{0, W}$.

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

$$\left(\{i, 0, k, w\},\ k = \overline{0, M_1 - 1}\right) \bigcup$$

$$\left(\{i, 1, k, w\},\ k = \overline{M_2 + 1, M_1}\right), i \ge 0,\ w = \overline{0, W}.$$

To simplify the analysis of the Markov chain ξ_t, let us enumerate the states of this chain in the direct lexicographic order of the components r, k, w and refer to the set of the states of the Markov chain having values (i, r) of the first two components as the macro-state (i, r). Let Q be the generator of the Markov chain ξ_t. It consists of the blocks $Q_{i,j},\ i, j \ge 0$, each of which contains four blocks $Q_{i,j}^{(r,r')},\ r, r' = 0, 1$, defining the intensities of the transition from the macro-state (i, r) to the macro-state (j, r').

Analysing all possible transitions of the Markov chain ξ_t during an interval of an infinitesimal length and rewriting the intensities of these transitions in the block matrix form we obtain the following result.

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

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

The matrix $Q_{i,i}$, $i \ge 0$, has the following form:

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

where

$$Q_{i,i}^{(0,0)} = I_{M_1} \otimes D_0 - (\alpha A_i + \beta C_1 (I_{M_1} - E_1^-) + G(I_{M_1} - E^+) + \mu N_i) \otimes I_{\bar{W}},$$

$$Q_{i,i}^{(0,1)} = G\hat{I} \otimes I_{\bar{W}},\ \ Q_{i,i}^{(1,0)} = \beta\tilde{I} \otimes I_{\bar{W}},$$

$$Q_{i,i}^{(1,1)} = I_{M_1 - M_2} \otimes D_0 - i\alpha I_{(M_1 - M_2)\bar{W}} - \beta C_2 (I_{M_1 - M_2} - E_2^-) \otimes I_{\bar{W}},\ i \ge 0,$$

$$Q_{i,i-1} = \mathrm{diag}\{(\mu N_i + \alpha A_i) \otimes I_{\bar{W}}, i\alpha I_{(M_1 - M_2)\bar{W}}\},\ i \ge 1,$$

$$Q^+ = I_{2M_1 - M_2} \otimes D_1.$$

Here,
- I is the identity matrix, and O is a zero matrix of appropriate dimension;
- $\otimes$ indicates the symbol of Kronecker product of matrices, see (Graham 1981);
- $\bar{W} = W + 1$;
- $\mathrm{diag}\{\dots\}$ denotes the diagonal matrix with the diagonal blocks listed in the brackets;
- $A_i = \mathrm{diag}\{\max\{i - N, 0\}, \max\{i - (N-1), 0\}, \dots, \max\{i - (N - M_1 - 1), 0\}\}$;
- $N_i = \mathrm{diag}\{\min\{i, N\}, \max\{\min\{i, N\} - 1, 0\}, \dots, \max\{\min\{i, N\} - M_1 - 1, 0\}\}$;
- $C_1 = \mathrm{diag}\{0, 1, \dots, M_1 - 1\}$;
- $C_2 = \mathrm{diag}\{M_2 + 1, M_2 + 2, \dots, M_1\}$;
- $G = \mathrm{diag}\{\gamma_0, \gamma_1, \dots, \gamma_{M_1 - 1}\}$;
- E^+ is the square matrix of size M_1 with all zero entries except the entries $(E^+)_{l,l+1}$, $l = \overline{0, M_1 - 2}$, which are equal to 1;
- E_1^- is the square matrix of size M_1 with all zero entries except the entries $(E_1^-)_{l,l-1}$, $l = \overline{1, M_1 - 1}$, which are equal to 1;
- E_2^- is the square matrix of size $M_1 - M_2$ with all zero entries except the entries $(E_2^-)_{l,l-1}$, $l = \overline{1, M_1 - M_2}$, which are equal to 1;

• $\hat{I}$ is the matrix of size $M_1 \times (M_1 - M_2)$ with all zero entries except the entry $(\hat{I})_{M_1-1,M_1-M_2-1}$ which is equal to 1;
• $\tilde{I}$ is the matrix of size $(M_1 - M_2) \times M_1$ with all zero entries except the entry $(\hat{I})_{0,M_2}$ which is equal to $M_2 + 1$.

Proof. The entries of the blocks Q^+ define the intensities of the transition of the Markov chain $\xi_t = \{i_t, r_t, k_t, w_t\}$, $t \geq 0$, that lead to increase of the component i_t (the number of customers in the system) by one. This can happen only if a customer arrives to the system, i.e., the underlying process w_t of the MAP arrival flow makes a transition with generation of a new customer. The intensities of such transition are defined by the entries of the matrix D_1. Arrival of a customer does not change the regime of the system operation (component r_t) and the number of broken servers (component k_t). Therefore, the matrix Q^+ has the form $Q^+ = \mathrm{diag}\{I_{M_1} \otimes D_1, I_{M_1-M_2} \otimes D_1\} = I_{2M_1-M_2} \otimes D_1$. Note, that operation of Kronecker product of matrices ($\otimes$) is very useful for describing the intensity or the probability of simultaneous transition of several independent Markovian components.

The entries of the blocks $Q_{i,i-1}$, $i \geq 1$, define the intensities of the transition of the Markov chain ξ_t, $t \geq 0$, that lead to the decrease in the number of customers in the system by one. This can happen if 1) a customer completes service or 2) a customer leaves the system due to impatience. If the system is in the operable state, both the options 1) and 2) are possible. The intensities of service completions depend on the number of customers in the system and the number of broken servers and are defined by the corresponding entries of the matrix μN_i. The matrix N_i defines the number of busy servers for each possible number of broken servers. The intensities of customers abandonment also depend on the number of customers in the system and the number of broken servers and are defined by the corresponding entries of the matrix αA_i.

If the system in the quarantine regime, only option 2) is possible. In this case, all i customers stay in the buffer, therefore, the intensity of a customer abandonment is $i\alpha$. A customer departure from the system does not change the regime of the system operation (component r_t), the number of broken servers (component k_t), and the state of the MAP underlying process (component w_t). Therefore, the matrix $Q_{i,i-1}$ has the form $Q_{i,i-1} = \mathrm{diag}\{(\mu N_i + \alpha A_i) \otimes I_{\bar{W}}, i\alpha I_{(M_1-M_2)\bar{W}}\}$, $i \geq 1$.

The non-diagonal entries of the blocks $Q_{i,i}$, $i \geq 0$, define the intensities of the transition of the chain $\xi_t, t \geq 0$, that do not lead to the change of the number of customers in the system. The entries of the matrix $Q_{i,i}^{(0,0)}$ define such intensities when the system is in the operable regime. The events that lead to such transitions are the following:

1) A breakdown arrives when the number of broken server is less then $M_1 - 1$. In this case, the number of broken servers increases by one and the component w_t does not change. The intensities of these transitions are given by the entries of the matrix $GE^+ \otimes I_{\bar{W}}$.

2) The underlying process of the MAP transits to another state without generation of a customer. In this case, the number of broken servers does not change. Thus, the intensities of these transitions are defined as the non-diagonal entries of the matrix $I_{M_1} \otimes D_0$.

3) The repairing of one broken server is finished. In this case, the number of broken servers decreases by one and the component w_t does not change. The intensities of these transitions are given by the entries of the matrix $\beta C_1 E_1^- \otimes I_{\bar{W}}$.

The entries of the matrix $Q_{i,i}^{(0,1)}$ define the intensities of the transitions that do not lead to the change the number of customers in the system, but lead to the transition to the quarantine regime. These transitions are possible only when the system is in the operable regime, the number of broken servers is $M_1 - 1$ and a new breakdown arrives. The intensities of these transitions are defined by the entries of the matrix $Q_{i,i}^{(0,1)} = G\hat{I} \otimes I_{\bar{W}}$.

The entries of the matrix $Q_{i,i}^{(1,1)}$ define the intensities of the transitions that also do not lead to the termination of the use of the quarantine regime and to the change the number of customers in the system. The events that lead to such transitions are the following: the repair completion of a broken server when the number of broken servers is greater than $M_2 + 1$ (the intensities are defined by the entries of the matrix $\beta C_2 E_2^- \otimes I_{\bar{W}}$) and the transition of the underlying MAP process without generation of a customer (the intensities are defined by the non-diagonal entries of the matrix $I_{M_1-M_2} \otimes D_0$).

The entries of the matrix $Q_{i,i}^{(1,0)}$ define the intensities of the transitions that do not lead to the change the number of customers in the system, but lead to the transition from the quarantine regime to the operable regime. These transitions are possible only if the number of broken servers is $M_2 + 1$ and one server is repaired. The intensities of these transitions are defined by the entries of the matrix $Q_{i,i}^{(1,0)} = \beta\tilde{I} \otimes I_{\bar{W}}$.

The diagonal entries of the matrix $Q_{i,i}$ are negative and the modulus of each entry defines the total intensity of leaving the corresponding state of the chain $\xi_t, t \geq 0$. The modules of the diagonal entries of the matrix $I_{M_1} \otimes D_0 - (\alpha A_i + \beta C_1 + G + \mu N_i) \otimes I_{\bar{W}}$ define the intensities of leaving the corresponding state of the Markov chain when the system is in operable mode. Thus, the matrix $Q_{i,i}^{(0,0)}$ is defined by the formula $Q_{i,i}^{(0,0)} = I_{M_1} \otimes D_0 - (\alpha A_i + \beta C_1 (I_{M_1} - E_1^-) + G(I_{M_1} - E^+) + \mu N_i) \otimes I_{\bar{W}}$. If the system is in the quarantine regime, the intensities of leaving the corresponding state of the Markov chain ξ_t are defined by the modulus of the diagonal entries of the matrix $I_{M_1-M_2} \otimes D_0 - i\alpha I_{(M_1-M_2)\bar{W}} - \beta C_2 \otimes I_{\bar{W}}$. Thus, the matrix $Q_{i,i}^{(1,1)}$ has the form $Q_{i,i}^{(1,1)} = I_{M_1-M_2} \otimes D_0 - i\alpha I_{(M_1-M_2)\bar{W}} - \beta C_2 (I_{M_1-M_2} - E_2^-) \otimes I_{\bar{W}}$.

Because the probability that the number of customers decreases or increases by more than one during the interval of an infinitesimal length is negligible, the

blocks $Q_{i,j}$ are zero matrices for all i, j when $|i-j| > 1$. Therefore, the generator Q has the block-tridiagonal structure. Theorem 1 is proved.

Corollary 1. The Markov chain $\xi_t,\ t \geq 0$, belongs to the class of continuous-time asymptotically quasi-Toeplitz Markov chains ($AQTMC$), see (Klimenok and Dudin 2006).

Proof directly follows from comparison of the properties of the blocks $Q_{i,j}$ of the block-tridiagonal generator Q for large values of i with the required properties of the generator listed in the definition of the $AQTMC$.

As the customers staying in the buffer are impatient ($\alpha > 0$), based on the results for $AQTMC$ it can be shown that the stationary probabilities of the system states $p(i, r, k, w)$, $i \geq 0$, $r = \overline{0, 1}$, $k = \overline{0, M_1 - 1}$, $w = \overline{0, W}$, exist for all possible values of the system parameters. Let us form the row vectors $\mathbf{p}_i$ of these probabilities enumerated in the lexicographic order of the components r, k, w. To this end, we sequentially form the row vectors

$$\mathbf{p}(i,0,k) = (p(i,0,k,0), p(i,0,k,1), \dots, p(i,0,k,W)),$$
$$k = \overline{0, M_1 - 1},$$
$$\mathbf{p}(i,0) = (\mathbf{p}(i,0,0), \mathbf{p}(i,0,1), \dots, \mathbf{p}(i,0,M_1-1)),$$
$$\mathbf{p}(i,1,k) = (p(i,1,k,0), p(i,1,k,1), \dots, p(i,1,k,W)),$$
$$k = \overline{M_2+1, M_1},$$
$$\mathbf{p}(i,1) = (\mathbf{p}(i,1,M_2{+}1), \mathbf{p}(i,1,M_2{+}2), \dots, \mathbf{p}(i,1,M_1)),$$
$$\mathbf{p}_i = (\mathbf{p}(i,0), \mathbf{p}(i,1)),\ i \geq 0.$$

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

$$(\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_i, \dots)Q = \mathbf{0},\ \ (\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_i, \dots)\mathbf{e} = 1$$

where Q is the infinitesimal generator of the Markov chain $\xi_t,\ t \geq 0$, $\mathbf{0}$ is a zero row vector, and $\mathbf{e}$ denotes a unit column vector. Due to the existing dependence of the blocks $Q_{i,i-1}$ and $Q_{i,i}$ on i, the problem of solving this infinite system of linear algebraic equations for the components of the vectors $\mathbf{p}_i$, $i \geq 0$, is far of easy. We omit the details of derivations and just mention that, to solve this system, we used the numerically stable algorithm that takes into account that the matrix Q has a block-tridiagonal structure, see (Dudina et al. 2013).

PERFORMANCE MEASURES

As soon as the vectors $\mathbf{p}_i,\ i \geq 0$, have been calculated, we are able to find various performance measures of the system.

The average number L of customers in the system is computed by $L = \sum_{i=1}^{\infty} i\mathbf{p}_i\mathbf{e}$.

Note that some formulas for performance measures contain the infinite sums. However, computation of these sums do not create difficulties. It is well known that if the Markov chain is ergodic the stationary probability vectors $\mathbf{p}_i$ converge in norm to a zero vector when i approaches infinity. Therefore, the computation of an sum may be terminated if the norm of the summand becomes less than a preassigned value ϵ.

The average number N_{serv} of busy servers is computed by $N_{serv} = \sum_{i=1}^{\infty} \sum_{k=0}^{M_1-1} \max\{\min\{i, N\} - k, 0\}\mathbf{p}(i,0,k)\mathbf{e}$.

The average number N_{buffer} of customers in the buffer is computed by $N_{buffer} = \sum_{i=1}^{\infty}(\sum_{k=0}^{M_1-1} \max\{i-(N-k), 0\}\mathbf{p}(i,0,k)\mathbf{e} + \sum_{k=M_2+1}^{M_1} i\mathbf{p}(i,1,k)\mathbf{e})$.

The average number N_{broken} of broken servers is computed by $N_{broken} = \sum_{i=0}^{\infty}(\sum_{k=1}^{M_1-1} k\mathbf{p}(i,0,k)\mathbf{e} + \sum_{k=M_2+1}^{M_1} k\mathbf{p}(i,1,k)\mathbf{e})$.

The average intensity λ_{out} of flow of customers who receive service is computed by $\lambda_{out} = \mu N_{serv}$.

The probability P_{loss} that an arbitrary customer will be lost is computed by

$$P_{loss} = 1 - \frac{\lambda_{out}}{\lambda} = \frac{\alpha N_{buffer}}{\lambda}.$$

The average intensity γ_{quar} of the quarantine beginning can be found as

$$\gamma_{quar} = \gamma_{M_1-1} \sum_{i=0}^{\infty} \mathbf{p}(i,0,M_1-1)\mathbf{e}.$$

The average duration T_{quar} of the quarantine is computed by $T_{quar} = \beta^{-1} \sum_{k=M_2+1}^{M_1} \frac{1}{k}$.

The probability P_{quar} that at an arbitrary moment system is in the quarantine regime is calculated as

$$P_{quar} = \sum_{i=0}^{\infty} \mathbf{p}(i,1)\mathbf{e} = \gamma_{quar} T_{quar}.$$

NUMERICAL EXAMPLE

The goal of the numerical example is to demonstrate the feasibility of the proposed results and illustrate possible way for optimizing the system operation by means of the optimal choosing the thresholds M_1 and M_2 for starting and finishing the regime of quarantine, correspondingly.

Let consider the system with $N = 20$ servers. We assume that the MAP arrival flow of customers is defined by the matrices

$$D_0 = \begin{pmatrix} -12.168 & 0 \\ 0 & -0.395 \end{pmatrix},\ D_1 = \begin{pmatrix} 12.087 & 0.081 \\ 0.22 & 0.175 \end{pmatrix}.$$

This arrival flow has the average intensity of customers arrival $\lambda = 9$, the coefficient of correlation of successive inter-arrival intervals is $c_{cor} = 0.2$ and the coefficient of variation of such intervals is $c_{var} = 12.34$.

The rest of the system parameters are chosen as follows: the intensities γ_k, $k = \overline{0, M_1 - 1}$, of breakdowns arrival are defined as $\gamma_0 = 0.00001$, $\gamma_1 = 0.0007$, $\gamma_2 =$

0.0015, $\gamma_l = \gamma_{l-1} + 0.0015l$, $l = \overline{3, M_1 - 1}$; the service intensity $\mu = 0.8$; the intensity of customers impatience $\alpha = 0.2$; the recovering intensity $\beta = 0.0005$.

Let us vary the threshold M_1 over the interval $[1, N]$ and the threshold M_2 over the interval $[0, M_1 - 1]$.

Figure 2 illustrates the dependence of the average number N_{broken} of broken servers on the thresholds M_1 and M_2.

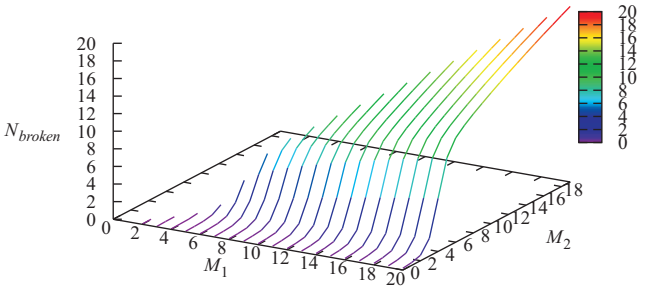

Figure 2: Dependence of the average number N_{broken} of broken servers on the thresholds M_1 and M_2

It is evidently seen in Figure 2 that the grows of the threshold M_1 (more late switching to the quarantine regime) causes the significant increase of the average number N_{broken} of broken servers. This increase becomes more essential when the threshold M_2 increases (the work in the quarantine regime becomes shorter).

Figure 3 illustrates the dependence of the average intensity γ_{quar} of the quarantine beginning on the thresholds M_1 and M_2.

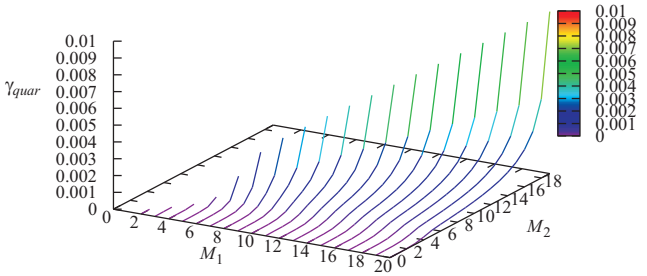

Figure 3: Dependence of the average intensity γ_{quar} on the thresholds M_1 and M_2

The intensity γ_{quar} of the quarantine beginning essentially increases when M_2 grows (what causes short duration of each quarantine while a frequent switching the system to the quarantine regime).

Figure 4 illustrates the dependence of the probability P_{quar} that an arbitrary moment system is in the quarantine regime on the thresholds M_1 and M_2.

The essential increase of P_{quar} when M_2 grows has the same explanation as the one given to Figure 3. More flat shape of the dependence of P_{quar} on the thresholds M_1 and M_2 in comparison to the shape of the dependence of γ_{quar} for large values of M_1 and M_2 is easily explained by the fact that the frequent switching to the quarantine regime (illustrated by Figure 3) is accompanied by the shorter duration of each period of using the quarantine regime.

Figure 5 illustrates the dependence of the loss probability P_{loss} on the thresholds M_1 and M_2.

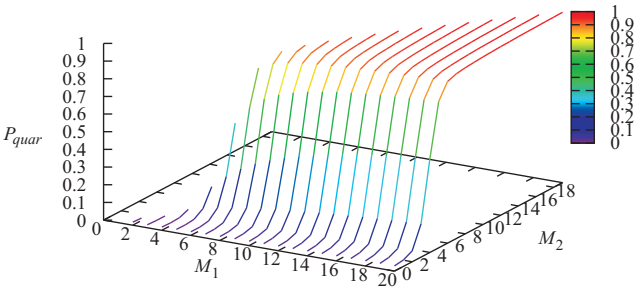

Figure 4: Dependence of the probability P^{quar} on the thresholds M_1 and M_2

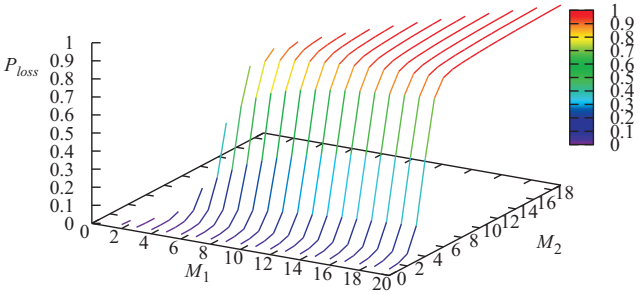

Figure 5: Dependence of the loss probability P_{loss} on the thresholds M_1 and M_2

The shape of the surface at Figure 5 well agrees with the shape of the surface at Figure 4. The large probability P_{quar} implies that during an essential share of time service is not provided. This causes large average queue length N_{buffer} and large value of the loss probability P_{loss} which is proportional to N_{buffer}. The minimal value of the loss probability $P_{loss} = 0.0182787$ is reached for $M_1 = 2$ and $M_2 = 1$.

Depending on application area of the considered queueing model, the quality of the system operation can be characterised in terms of different cost criteria. E.g., let us consider the following cost criterion:

$$E(M_1, M_2) = a\lambda_{out} - b\gamma_{quar} - cN_{broken}$$

where a is the profit obtained by the system from service of one customer, b is the charge paid by the system for the transition to the quarantine regime and c is the charge paid by the system for repair of one broken server.

Figure 6 illustrates the dependence of this cost criterion on the thresholds M_1 and M_2 when the cost coefficients are fixed as follows: $a = 1$, $b = 100$ and $c = 2$.

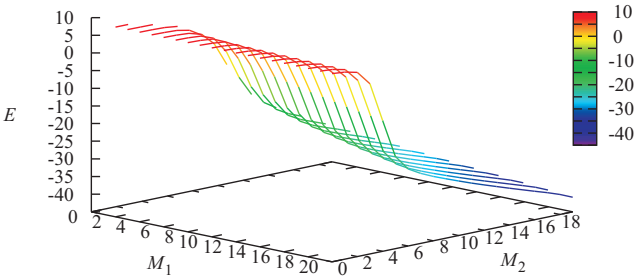

Figure 6: Dependence of the value $E(M_1, M_2)$ of the cost criterion on the thresholds M_1 and M_2

The optimal (maximal) value of the cost criterion is equal to $E^*(M_1, M_2) = 8.7419$ and is reached when $M_1 = 2$ and $M_2 = 0$, i.e., the quarantine starts when the number of broken servers reaches the value $M_2 = 2$ and the system resumes the work when all broken servers will be repaired.

CONCLUSIONS

The novel unreliable queueing model, in which the intensity of servers breakdowns increases when the number of broken servers grows, is analysed. To smooth the negative effect of the increase in this intensity, it is proposed to switch the system to quarantine regime during which new breakdowns cannot occur. Switching to quarantine regime is performed when the number of broken servers reaches the threshold value M_1. Switch back to operable regime is performed when the number of broken servers drops below the threshold value M_2. The problem of computation of performance measures of the model as function of the thresholds (M_1, M_2) is solved. This allows us to make managerial decisions providing the best quality of the system operation. Results may be extended to the systems with retrials and more complicated arrival flows.

ACKNOWLEDGMENT

The publication was financially supported by the Ministry of Education and Science of the Russian Federation (the Agreement number 02.a03.21.0008).

REFERENCES

Al-Begain, K., Dudin, A., Klimenok, V., Dudin, S. 2012. "Generalised survivability analysis of systems with propagated failures", *Computers and Mathematics with Applications*, 64, 3777-3791.

Chakravarthy, S. 2001. "The batch Markovian arrival process: a review and future work", *Advances in Probability Theory and Stochastic Processes*, Notable Publications Inc., New Jersey, 21-29.

Dudin, A., Jacob, V., Krishnamoorthy, A. 2015. "A multi-server queueing system with service interruption, partial protection and repetition of service", *Annals of Operations Research*, 233, 101-121.

Dudina, O., Kim, Ch., Dudin, S. 2013. "Retrial queueing system with Markovian arrival flow and phase type service time distribution", *Computers and Industrial Engineering*, 66, 360-373.

Graham, A. 1981. "Kronecker products and matrix calculus with applications", Ellis Horwood, Cichester.

Klimenok, V., Orlovsky, D., Kim, Ch. 2008. "The $BMAP/PH/N$ retrial queue with Markovian flow of breakdowns", *European Journal of Operational Research*, 189, 1057-1072.

Klimenok, V.I. and Dudin, A.N. 2006. "Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory", *Queueing Systems*, 54, 245-259.

Lucantoni, D. 1991. "New results on the single server queue with a batch Markovian arrival process", *Communication in Statistics-Stochastic Models*, 7, 1-46.

Mitrani, I. and Avi-Itzhak, B. 1968. "A many-server queue with server interruptions", *Operations Research*, 163, 28-638.

Neuts, M. and Lucantoni, D. 1979. "A Markovian queue with N servers subject to breakdowns and repair", *Management Science*, 25, 49-861.

AUTHOR BIOGRAPHIES

ALEXANDER DUDIN has got PhD degree in Probability Theory and Mathematical Statistics in 1982 from Vilnius University and Doctor of Science degree in 1992 from Tomsk University. He is Head of Laboratory of Applied Probabilistic Analysis in Belarusian State University, Professor of the Probability Theory and Mathematical Statistics Department. He works also part time at the Peoples' Friendship University of Russia. He is author of 350 publications including more than 80 papers in top level Journals. Field of scientific interests are: Random Processes in Queueing Systems and Applications of Queueing Theory to Telecommunication. His email address is dudin@bsu.by.

SERGEI DUDIN was graduated from Belarusian State University in 2007. In 2010, he got PhD degree in Belarusian State University in System Analysis, Control and Information Processing and works currently as leading scientific researcher of Research Laboratory of Applied Probabilistic Analysis in Belarusian State University. He works also part time at the Peoples' Friendship University of Russia. His main fields of interests are queueing systems with session arrivals and controlled tandem models. His email address is dudins@bsu.by.

OLGA DUDINA was graduated from Belarusian State University in 2007. In 2010, she got PhD degree in Belarusian State University in Probability Theory and Mathematical Statistics and works currently as leading scientific researcher of Research Laboratory of Applied Probabilistic Analysis in Belarusian State University. She works also part time at the Peoples' Friendship University of Russia. Her main fields of interests are tandem queueing models with correlated arrival flows, non-markovian queueing systems. Her email address is dudina@bsu.by.

KONSTANTIN SAMOUYLOV received his Ph.D. from the Moscow State University and a Doctor of Sciences degree from the Moscow Technical University of Communications and Informatics. During 1985-1996 he held several positions at the Faculty of Science of the Peoples' Friendship University of Russia where he became a head of Telecommunication System Department in 1996. Since 2014 he is a head of the Department of Applied Informatics and Probability Theory. His current research interests are probability theory and theory of queuing systems, performance analysis of 4G/5G networks, teletraffic of triple play networks, and signaling networks planning. He is the author of more than 100 scientific and technical papers and three books. His email address is ksam@sci.pfu.edu.ru.