Informatics and Control Problems 42 Issue 1 (2022)
journal homepage: www.icp.az

On the problem of the switching of the fast server

E.B. Bayramov[1*], E.V. Mehbaliyeva[2]

[1]*Datarockets (Red Panda Technology Inc.), Toronto, Canada*
[2]*Sumgait State University, Sumgait, Azerbaijan*

ARTICLE INFO

Article history:
Received 04.05.2022
Received in revised form 12.05.2022
Accepted 19.05.2022
Available online 25.05.2022

Keywords:
System with heterogeneous servers
Randomized *N*-policy
Numerical analysis

ABSTRACT

A Markovian model of a system with two heterogeneous servers is investigated, in which the slow server is always on, and the fast server can turn on when the job queue length reaches a certain value. Models with unlimited and limited size of the buffer of waiting jobs are considered. A numerical method for calculating the probability distribution of states and characteristics of the systems under study is developed. The results of numerical experiments are given.

1. Introduction

Queues with Heterogeneous Servers (QHS) are common in the modeling of computer and telecommunication networks, because it is necessary to use heterogeneous computers (servers) in the process of their expansion. Apart from computer and telecommunication networks, different speed servers are found in call centers with multiple operators who have different qualifications, as well as in production systems, where the process of servicing jobs involves not machines, but humans.

In [1], devoted to the study of QHS models, a Markovian infinite-queue system is studied, in which one of the free servers is assigned to service jobs of equal probability (this scheme is called randomized access). An analysis of the available literature has shown that the vast majority of studies investigate QHS models, which adopt randomized access schemes [1-3] and ordered access schemes [4-8]. An overview of these studies can be found in [9, 10].

Important problems in QHS are determining the optimal access strategies and selecting the optimal server switching schemes depending on the current state of the system. In order to optimize QHS, the FSF (Fast Server First) access scheme is often used, i.e., the server that has the highest speed among the free servers is always selected for servicing at the moment a job arrives. This is because the probability of loss in the M/M/k/k system has a minimum value when using the FSF access scheme [11]. However, the FSF access scheme is not always optimal or even suboptimal for QHS with queues. For instance, it is proved in [12] that for QHS with unlimited queueing, the *N*-policy is optimal for minimizing the average number of jobs in the system. In accordance with this policy, in two server QHSs, the fast server always runs if there is at least one job in the system, and

[*] Corresponding author.
E-mail addresses: elvin.bayramov@protonmail.com (E.B. Bayramov), esmira.mehbaliyeva@mail.ru (E.V. Mehbaliyeva).

www.icp.az/2022/1-08.pdf https://doi.org/10.54381/icp.2022.1.08

the slow server is only started when the queue length reaches a certain (threshold) value N.

Note that a generalization of the classical N-policy was introduced in [13], i.e., a randomized N-policy for the slow server switching on was defined. According to this policy, the slow server turns on with a certain probability and remains idle with an additional probability. The study proposed a numerical method to find the stationary probabilities of states of the system, also calculating the characteristics of a system with unlimited and limited size of the buffer of waiting jobs.

In this study, we propose a QHS model with an alternative randomized N-policy, i.e., here, in contrast to [13], it is assumed that the slow server is always on, and the fast server turns on with a certain probability when the queue length reaches a value N, and with an additional probability it remains idle. The use of this scheme is relevant in systems where the startup of the fast server involves heavy economic (technical or technological) losses.

2. Method for calculating the state probabilities and characteristics of the systems under study

We consider a QHS with an unlimited queue that contains two servers: a fast server (F-server) and a slow server (S-server). The following assumptions are made here.

(i) The system receives a Poisson flow of identical jobs at a rate λ, and this value does not depend on the state of the servers;

(ii) The service times of both servers are random variables with exponential d.f.; the average service times of F-server and S-server are μ_F^{-1} and μ_S^{-1} respectively, with $\mu_F > \mu_S$.

(iii) The S-server is always on, and the F-server can turn on only when the queue length is not less than a certain threshold value $N, 0 < N < \infty$. In this case, the scheme of F-server switching on and off is defined as follows. If at the time a job arrives, the queue length is at least N, then the F-server turns on with probability $\alpha, 0 < \alpha < 1$ and one job from the queue is selected for service in that server; with additional probability $1 - \alpha$ it remains idle. When the F-server finishes serving a job, it selects one job from the queue if the queue length at that moment is greater than the value of N; otherwise, the F-server goes into hibernation mode.

(iv) In a limited-queue system, an incoming job is lost with probability one if at that moment the buffer is full and the F-server is on; if at that moment the buffer is full and the F-server is off, then either the F-server turns on with probability α or the incoming job is lost with additional probability $1 - \alpha$.

The task is to find the joint distribution of the number of jobs in the system and the F-server status, and to develop methods to calculate its characteristics.

3. Method for calculating the state probabilities and characteristics of the systems under study

Let us first consider a QHS model with unlimited queueing. The state of this system at an arbitrary time instant is determined by a two-dimensional vector (n, k), where n is the number of jobs in the system, k is the state of the F-server, i.e., $k = 0$, if the F-server is off, and $k = 1$ otherwise. Then the system operation is described by a two-dimensional Markov chain (MC) with the following state space:

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

where $E_k = \{(n, k): n = 0,1,2, \dots\}, k = 0,1$

The elements of the generating matrix of this MC are denoted by $q\big((n, k), (n', k')\big)$, i.e., $q\big((n, k), (n', k')\big)$ denotes the rate of the transition from the state (n, k) to the state (n', k').

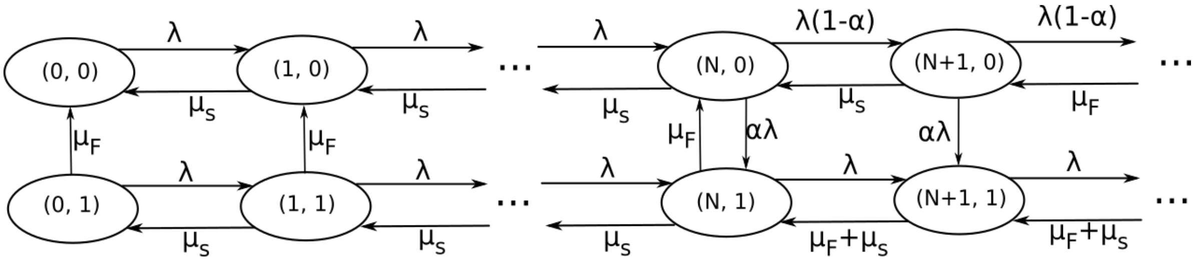


Fig. 1. The graph of transitions between states of the system

Based on the described mechanism of the system, we conclude that these quantities are determined as follows (see Fig. 1):

case $(n,0) \in E_0$:

$$q\big((n,0),(n',k')\big) = \begin{cases} \lambda, \text{if } n < N, n' = n+1, k' = 0, \\ \lambda(1-\alpha), \text{if } n \geq N, n' = n+1, k' = 0, \\ \lambda\alpha, \text{if } n \geq N, n' = n, k' = 0\,, \\ \mu_S, \text{if } n > 0, n' = n-1, k' = 0; \end{cases} \tag{2}$$

case $(n,1) \in E_1$:

$$q\big((n,1),(n',k')\big) = \begin{cases} \lambda, \text{if } n' = n+1, k' = 1, \\ \mu_F, \text{if } 0 < n \leq N, n' = n, k' = 0, \\ \mu_F + \mu_S, \text{if } n > N, n' = n-1, k' = 1, \\ \mu_S, \text{if } n \leq N, n' = n-1, k' = 1. \end{cases} \tag{3}$$

We denote the stationary probability of the state $(n,k) \in E$ by $p(n,k)$. The condition for the existence of the stationary mode is $\lambda < \mu_F + \mu_S$.

The characteristics of this QHS model are the average number of jobs in the system (L_S), the average dwell time of a job in the system (W_S) and the rate of switching of the F-server (RF).

The average number of jobs in the system is determined as the mathematical expectation of the corresponding random variable, i.e.,

$$L_S = \sum_{n=1}^{\infty} n \sum_{k=0}^{1} p\,(n,k) \tag{4}$$

The average dwell time of jobs in the system is calculated using Little's formula:

$$W_S = \frac{1}{\lambda} L_S\,. \tag{5}$$

Since the F-server turns on with probability α, then if at the time a job arrives, the number of jobs in the system is not less than the value N, then the rate of switching of the F-server is calculated as follows:

$$RF = \lambda\alpha \sum_{n=N}^{\infty} p\,(n,0)\,. \tag{6}$$

To find the stationary probabilities of states $p(n,k)$, the method of generating functions can be used. However, as shown in [13], its use is associated with certain methodological and technical difficulties because of the complex structure of the generating matrix of the MC under investigation. Therefore, the approximate method developed in [13] is used to solve the problem under consideration.

For the correct application of this method in the system under study, we assume that $\alpha \ll 1-\alpha$, i.e., $\alpha \ll 0.5$. Since the speed of the F-server is greater than the speed of the S-server, when the condition $\alpha \ll 0.5$ is satisfied, the system under study remains in the states from classes E_0 and E_1 for a long time, i.e., transitions between these classes occur rarely (see Fig. 1).

Based on this fact, all states within each class E_k are combined into one lumped state $\langle k \rangle$, and thus the set of lumped states $\Omega = \{\langle k \rangle : k = 0,1\}$ is determined.

The approximate values of the probabilities of states $\not{p}(n,k), (n,k) \in E$ of the original model are determined as

$$\not{p}(n,k) = \rho_k(n)\pi(\langle k \rangle), \tag{7}$$

where $\rho_k(n)$ is the probability of the state (n,k) within a split model with the state space E_k, and $\pi(\langle k \rangle)$ is the probability of the lumped state $\langle k \rangle \in \Omega$.

From relations (2) we conclude that if the condition $\nu_S < (1-\alpha)^{-1}$, where $\nu_S = \lambda/\mu_S$, is satisfied, the probabilities of states of the split model with the state space E_0 are as follows:

$$\rho_0(n) = \begin{cases} \nu_S^n \rho_0(0), \text{if } 0 \le n \le N\,, \\ (1-\alpha)^{-N}\big((1-\alpha)\nu_S\big)^n \rho_0(0), \text{if } n > N\,, \end{cases} \tag{8}$$

where $\rho_0(0)$ is determined from the normalization condition, i.e., $\rho_0(0) = \left(\sum_{n=0}^{N-1} \nu_S^n + \frac{\nu_S^N}{1-(1-\alpha)\nu_S}\right)^{-1}$

If the condition $\nu_{FS} < \alpha$ is fulfilled, where $\nu_{FS} = \frac{\lambda}{\mu_F + \mu_S}$, then from (3) we obtain that the probabilities of states of the split model with the state space E_1 are found as follows:

$$\rho_1(n) = \begin{cases} \nu_S^n \rho_1(0), \text{if } 0 \le n \le N, \\ \left(\frac{\nu_S}{\nu_{FS}}\right)^N \nu_{FS}^n \rho_0(0), \text{if } n > N, \end{cases} \tag{9}$$

where $\rho_1(0)$ is determined from the normalization condition, i.e., $\rho_1(0) = \left(\sum_{n=0}^{N-1} \nu_S^n + \frac{\nu_S^N}{1-\nu_{FS}}\right)^{-1}$.

Uniting the inequalities $\nu_S < (1-\alpha)^{-1}$ and $\mu_F > \mu_S$, we find the following necessary condition under which the proposed method can be correctly applied to the model under study:

$$\lambda < max\{\mu_S + \mu_F, (1-\alpha)\mu_S\}\,. \tag{10}$$

Suppose that $q_{kk'}, \langle k \rangle, \langle k' \rangle \in \Omega$ denotes the rate of transition from the lumped state $\langle k \rangle$ into the lumped state $\langle k' \rangle$. The specified transition rates are calculated as follows:

$$q_{01} = \lambda\alpha \sum_{n=N}^{\infty} \rho_0\,(n) = \lambda\alpha\left(1 - \sum_{n=0}^{N-1} \rho_0\,(n)\right);\ q_{10} = \mu_F \sum_{n=0}^{N} \rho_1\,(n)\,. \tag{11}$$

Therefore, from relations (11) we have

$$\pi(\langle 0 \rangle) = \frac{q_{10}}{q_{01} + q_{10}}, \pi(\langle 1 \rangle) = 1 - \pi(\langle 0 \rangle)\,. \tag{12}$$

Thus, if ergodicity condition (10) is satisfied, taking into account relations (7)-(9) and (12), approximate values of state probabilities are calculated. After standard transformations the approximate value of the average number of jobs in the system is determined:

$$\begin{gathered} L_S \approx \sum_{k=0}^{1} \pi(\langle k \rangle) \sum_{k=1}^{\infty} n\rho_k(n) = \pi(\langle 0 \rangle)\rho_0(0)\left(\sum_{n=1}^{N-1} n\nu_S^n + (1-\alpha)^{-N} G\big((1-\alpha)\nu_S\big)\right) + \\ +\pi(\langle 1 \rangle)\rho_1(0)\left(\sum_{n=1}^{N-1} n\nu_S^n + \left(\frac{\nu_S}{\nu_{FS}}\right)^N G(\nu_{FS})\right) \end{gathered} \tag{13}$$

where $(x) = \frac{x^N(N - x(N-1))}{(1-x)^2}$.

The approximate value of the F-server's rate of swtiching is calculated as follows:

$$RS \approx \lambda\alpha\pi(\langle 0 \rangle) \sum_{n=N}^{\infty} \rho_0(n) = \lambda\alpha\pi(\langle 0 \rangle)(1 - \sum_{n=0}^{N-1} \rho_0(n)). \tag{14}$$

Note that the method of balance equations can also be used to calculate the probabilities of states and characteristics of the QHS model with a limited buffer size. At the same time, the approximate method described above can be used for this purpose.

For example, suppose that the total capacity of the system is M, where $M > N$ (i.e., the buffer size is $M-1$). When using the approximate method, in the second line of formulas (8) and (9) the upper bound of the parameter n (infinity) is replaced by a finite value $X(t)$. The formulas for finding the values of $\rho_k(0), k=0{,}1$ are changed accordingly, i.e., we have:

$$\rho_0(0)=\left(\sum_{n=0}^{N} v_S^n+(1-\alpha)^{-N}\sum_{n=N+1}^{M}\left((1-\alpha)v_S\right)^n\right)^{-1};\ \rho_1(0)=\left(\sum_{n=0}^{N} v_S^n+\left(\frac{v_S}{v_{FS}}\right)^N\sum_{n=N+1}^{M} v_{FS}^n\right)^{-1}. \quad (15)$$

Taking into account formulas (15), the rates of transitions between lumped states are calculated using formulas (11), and the probabilities of the states themselves are calculated from formulas (12). Further, from formula (4) and (6) we conclude that in this model QHS the approximate values of the system characteristics are calculated as follows:

$$L_S \approx \sum_{k=0}^{1}\pi(\langle k\rangle)\sum_{n=1}^{M} n\rho_k(n);\quad RS \approx \lambda\alpha\pi(\langle 0\rangle)\sum_{n=N}^{M}\rho_0(n).$$

Here we have a new characteristic, the probability of job loss (PB). The exact value of this quantity is determined as follows:

$$PB=(1-\alpha)p(M,0)+p(M,1)\,.$$

The average dwell time of jobs in this system is calculated as follows:

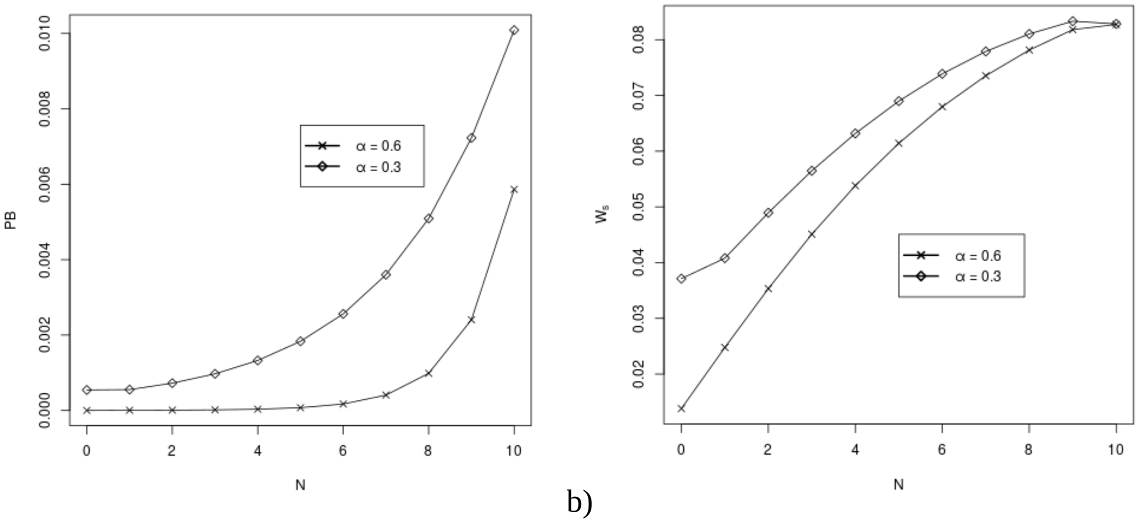
$$W_s=\frac{1}{\lambda(1-PB)}L_{s\,.}$$

From (7) we conclude that the approximate value of this quantity is determined as follows:

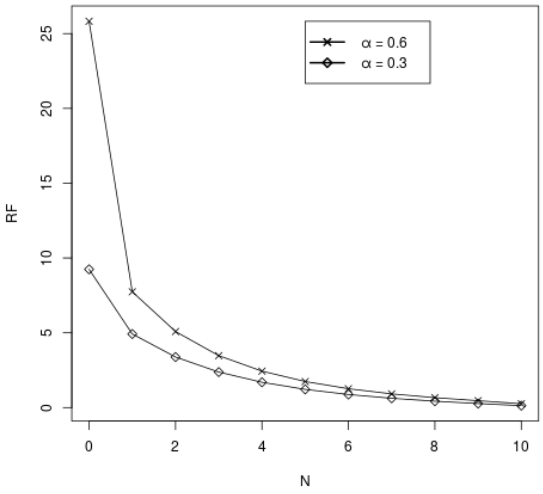
$$PB \approx (1-\alpha)\rho_0(M)\pi(\langle 0\rangle)+\rho_1(M)\pi(\langle 1\rangle).$$

4. Numerical results

Here we study the relationships between the characteristics of the system and the values of the threshold parameter N and the probability of the fast server's switching on and conduct a comparative analysis of two schemes: the slow server's switching on [13] and the fast server's switching on, which is proposed in this study. We also solve the problems of finding the optimal value of the threshold parameter N and conduct a comparative analysis of the results of optimization problems for the above two server switching schemes.



a) b)



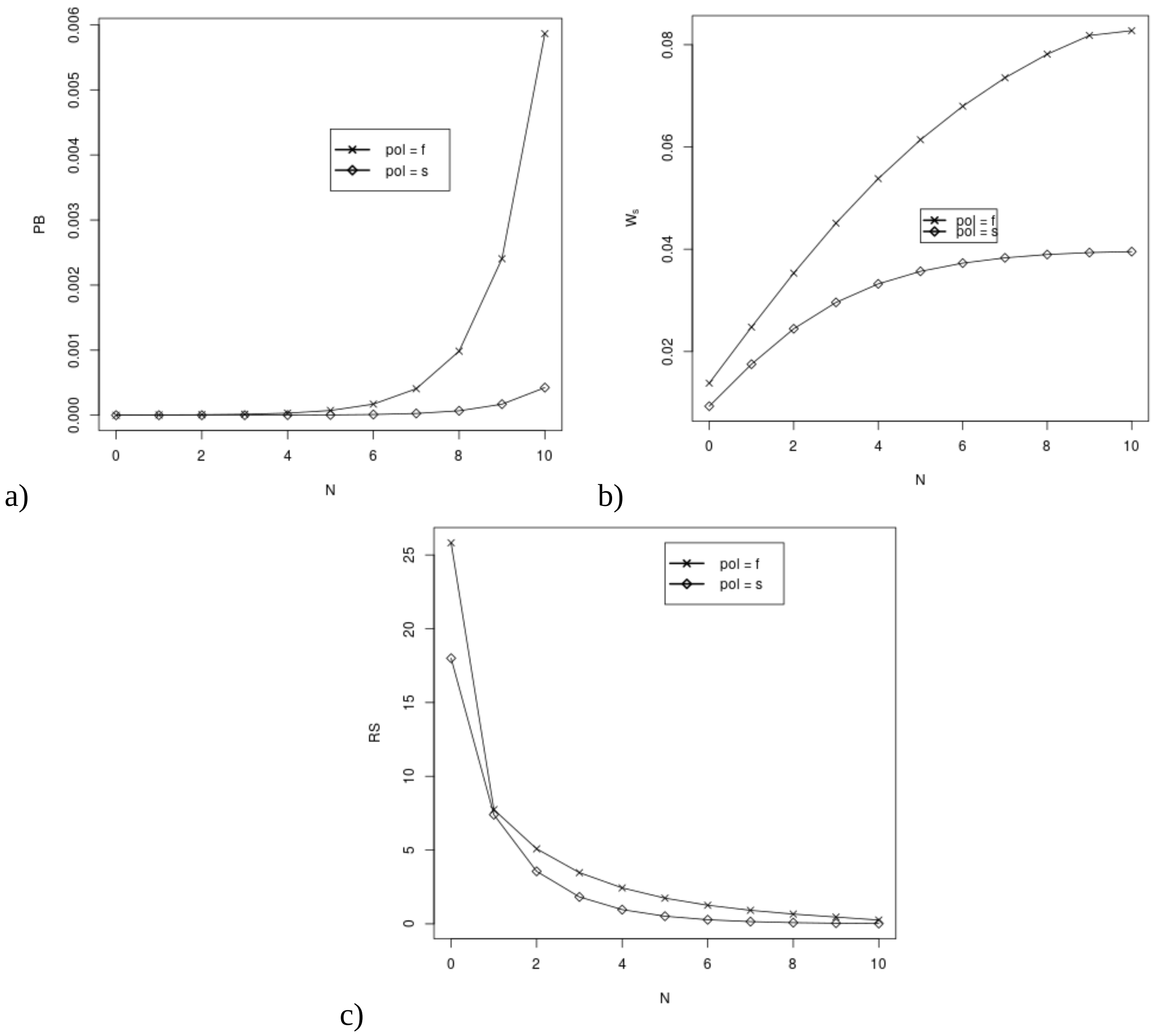
c)

Fig. 2. Characteristics of a limited buffer system vs. parameter N

Fig. 2 shows the relationship between the characteristics of the system with a limited buffer and the threshold parameter N for different values of the probability of the fast server switching on. The values of initial parameters of the system are chosen as follows: $\lambda = 30, \mu_F = 55, \mu_S = 40, M = 10$. From these graphs we conclude that the probability of job loss (PB) increases with the increase in the value of the threshold parameter N (see Fig. 2, a). This is to be expected, because as the value of the parameter N grows, the length of the job queue increases, i.e., the probability that the buffer will be completely filled increases; at the same time, this characteristic decreases as the value of the probability of the fast server switching on increases. Average dwell time of jobs in the system (W_S) also grows with the growth of the value of the threshold parameter N (see Fig. 2, b), because the job queue length increases with it; here this characteristic decreases as the probability of the fast server switching on increases, because the rate of servicing grows with it. As expected, with the growth of the threshold parameter N the rate of switching of the F-server decreases (see Fig. 2, c), and with the growth of the probability of the fast server switching on this characteristic decreases, as at the same time the probability of the F-server switching on decreases.

We have investigated similar dependences for the system with an unlimited buffer for the same values of the load parameters of the system. Note that the behavior of the characteristics Ws and RF completely coincides with the curves shown in Fig. 2, b and Fig. 2, c. At the same time, the values of the characteristic Ws remain almost unchanged, and the values of the characteristic RF for the model with an unlimited buffer are almost 2 times smaller than for the model with a limited buffer.

We now consider comparisons of the system characteristics for different server switching schemes. As mentioned in the introduction, in QHS with unlimited queueing, the optimal policy for minimizing the average number of jobs in the system is that the fast server always runs if there is at least one job in the system, and the slow server only turns on when the queue length reaches the threshold value.



a) b)

c)

Fig. 3. Characteristics of a limited buffer system vs. parameter *N* for different server switching schemes

Here, for brevity, we present only the corresponding graphs for the system with a limited buffer, see Fig. 3, where it is assumed that $\lambda = 30, \mu_F = 55, \mu_S = 40, \alpha = 0.6$. In these graphs, pol=s corresponds to a policy where the fast server is always on, and pol=f indicates results for a policy where the slow server is always on. On the ordinate axis, RS (Rate of Switching) refers to the server's rate of switching. As can be seen from these graphs, for a system with a limited buffer, the policy where the fast server is always on and the slow server is always on/off depending on the length of the job queue is also the optimal in all three characteristics. We can notice that the server's rate of switching is higher with the policy described in this article. This is because when the majority of jobs are handled by the slow server, the queue grows faster and the chance that it will reach threshold N is higher than if the fast server were on.

In addition to the operational characteristics, we have also studied the behavior of Total Cost (TC) in the system, associated with the dwell of jobs in the system, their losses, as well as the switching on and operation of servers in various schemes.

Total penalty costs for both schemes are determined as follows:

$$TC = c_s W_s + \lambda c_l PB + c_{ox}\pi(\langle 1 \rangle) + c_x RX, \qquad (16)$$

where c_s is the penalty cost per unit time of one job dwell in the system; c_l is the penalty cost for the loss of one job; c_{0x} is penalty cost per unit time of x-server's operation; c_x is the penalty cost of one-time switching on of x-server, $x \in \{F, S\}$.

We studied models with an unlimited and a limited buffer. For the model with an unlimited buffer, it is assumed in formula (16) that PB=0.

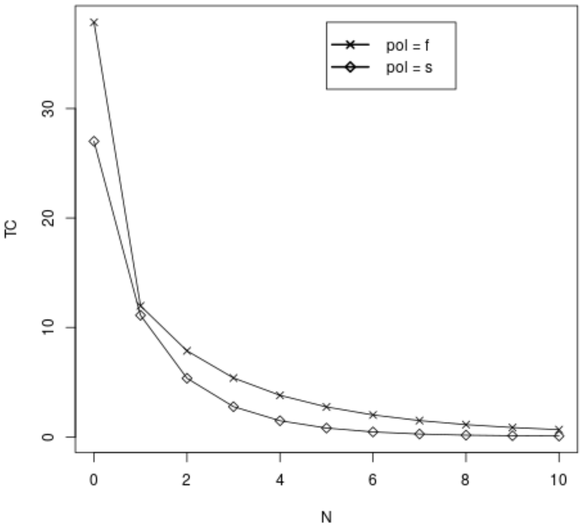


Fig. 4. TC vs. parameter N

For the sake of brevity, Fig. 4 shows results only for a limited-buffer model, where it is assumed that $c_s = 1, c_l = 1, c_{0x} = 2, c_x = 2.5$.

From Fig. 4 we conclude that the TC function is monotonically decreasing with respect to the parameter N; the policy in which the fast server is always running and the slow server switches on/off depending on the job queue length is optimal here as well.

5. Conclusion

In this study, we propose a randomized N-policy for the switching on of the fast server, according to which when the queue length reaches the value N, the fast server either switches on with a certain probability, or remains idle with an additional probability. An ergodicity condition for a model with an unlimited buffer size has been obtained. An approximate method for calculating the probability distribution of states and characteristics of the system under study has been developed, and the problem of optimization of a limited-queue system has been solved.

References

[1] H.Gumbel, Waiting Lines with Heterogeneous Servers, Operations Research. 8 (1960) pp.504-511.

[2] D.D. Yao, The Arrangement of Servers in an Ordered Entry System, Operations Research. 35 (1987) pp. 759-763.

[3] V.S. Singh, Markovian Queues with Three Servers, IIE Transactions. 3 (1971) pp. 45-48.

[4] E.A. Elsayed, Multichannel Queuing Systems with Ordered Entry and Finite Source, Computers & Operations Research. 10 (1983) pp. 213-222.

[5] D. Fakinos, The Generalized M/G/k Blocking System with Heterogeneous Servers, Journal of Operations Research Society. 33 (1982) pp.801-809.

[6] B. Pourbabai, D. Sonderman, Server Utilization Factors in Queuing Loss Systems with Ordered Entry and Heterogeneous Servers, Journal of Applied Probability. 23 (1986) pp. 236-242.

[7] W.M. Nawijn, On a Two-Server Finite Queuing System with Ordered Entry and Deterministic Arrivals, European Journal of Operations Research. 18 (1984) pp. 388-395.

[8] H.O. Isguder, U.U. Kocer, Analysis of GI/M/n/n Queuing System with Ordered Entry and no waiting line, Applied Mathematical Modelling. 38 (2014) pp. 1024-1032.

[9] A.Z. Melikov, E.V. Mekhbaliyeva, Analysis and optimization of system with heterogeneous servers and jump priorities, Journal of Computer and Systems Sciences International. 58 (2019) pp.718-735.

[10] A.Z. Melikov, L.A. Ponomarenko, E.V. Mekhbaliyeva, Analysis of models of systems with heterogeneous servers, Cybernetics and System Analysis. 56 (2020) pp.89-99.

[11] G. Nath, E. Enns, Optimal Service Rates in the Multi-Server Loss System with Heterogeneous Servers, Journal of Applied Probability.18 (1981) pp.776-781.

[12] R.L. Larsen, A.K. Agrawala, Control of Heterogeneous Two-Server Exponential Queuing System, IEEE Transactions on Software Engineering. 9 (1983) pp.522-526.

[13] А.З. Меликов, Э.В. Мехбалыева, Численное исследование системы с гетерогенными серверами и рандомизированной N-политикой, Вестник Томского Университета. Управление, вычислительная техника и информатика. No.53 (2020) pp.25-38. [In Russian: A.Z. Melikov, E.V. Mehbaliyeva, Numerical study of a system with heterogeneous servers and randomized N-policy, Vestnik Tomskogo Universiteta. Upravleniye, Vychislitelnaya Tekhnika i Informatika].