Asymptotic Analysis of the N-Model with Static Priority

Mariia Maltseva
Petrozavodsk State University
Petrozavodsk, Russia
masha.mariam.maltseva@mail.ru

Evsey Morozov
Institute of Applied Mathematical Research,
Karelian Research Centre RAS
Petrozavodsk State University
Petrozavodsk, Russia
emorozov@karelia.ru

***Abstract*—In this paper, we study the so-called N-model, which consists of two pools of servers. Pool 1 contains N_1 servers while pool 2 consists of one server, which can be inactive time to time. The 1st pool is fed by a Poisson input of the 1st class customers. Provided all servers of pool 1 are busy and the queue size exceeds a given threshold, a class-1 customer, with a given probability, jumps to server of pool 2, if this server is active. Under exponential assumptions, the number of customers in the 1st pool is a Markov birth-death process. The periods of activity/inactivity constitute initially a "transient" process, which converges to a stationary regime. In this research we deduce stationary distribution of this Markov process in an explicit form. Moreover, we find conditions, when the mentioned convergence of the alternating process of activity/inactivity of the server in the 2nd pool implies convergence of the birth-death process, describing the 1st pool, to stationary regime. Moreover, this convergence is demonstrated by simulation.**

I. Introduction

In this work, we develop stability analysis of the so-called N-model. This model has been proposed and analyzed by an extended fluid approach in the paper [1]. The authors also considered this model in previous work [2]. In general, this model belongs to a class of *Skills-Based* systems. This property can be expressed in particular by assignment of the servers among customers depending on their class or priority. Both the design and analysis of Skills-Based Routing is a complex and challenging problem. A classic multi-server system describes situation, when all servers are equally-skilled [3], [4]. The N-model we consider in this work is a two-pool system, which serves class-1 customers following Poisson input. These customers arrive in the 1st pool, and depending on the state of the queue may jump to pool 2 to be served there. At that, a class-1 customer meeting all servers in pool 1 busy (or when queue-size in pool 1 exceeds a threshold $C > 0$) jumps to pool 2 to be served as a class-(1,2) customer. A distinctive feature of this model is that server in pool 2 is unavailable time to time, and when the server is unavailable, class-1 customer can not jump to pool 2.

Another variation of this model has been considered in [5]. In the papers [1], [4], [6], [5], [2] can be found a motivation to introduce and study class of N-models. Interaction between servers, which exists in N-model, makes stability and performance investigation of these systems much more complicated. These models constitute a wide class of the systems with interacting servers. Also models under study belong to systems with *flexible servers* and, alternatively, can be called *cross-trained* servers [7], [8], [9], [10]. We note that flexible servers can be used, when there are a few different classes of customers arrive in the system. In this case, some servers can serve a fixed set of classes of customers, while remaining servers accept more classes of customers. In other setting, service capacity may be shared between servers to increase throughput. An important example of this setting is the *cognitive radio*, where a dynamic management is applied for using the best wireless channels to avoid congestion. Such a radio can detect unused frequency bands and switch between free channels without interruption of transmission.

We call regime of the server in pool 2 *stationary*, if the server is switching in inactive state (switch off) as soon as becomes idle. Alternatively, we call this regime *transient*, if server of pool 2 may stay active after departure of a served customer. Then convergence of the transient regime of the 2nd pool to stationary one can be described as follows: assume that, when the 2nd pool is idle, inactivity periods appear initially with a rate λ_2. In this case, server, being idle, can be also active. Assume further that $\lambda_2 \to \infty$. In this case, each time, when server (in pool 2) becomes idle after completion of service, is switches immediately in inactive state. Namely such a policy we call stationary regime of pool 2.

The contribution of this work is as follows. First of all, for exponential service times, that is for pure Markovian model, we construct Kolmogorov equations and derive the stationary distribution of the number of customers in the 1st pool with $N_1 \geq 1$ servers, *provided the 2nd pool (server) is in the stationary regime initially*. Another contribution of this work is that now we allow that class-1 customer jumps (in idle) pool 2 with a probability p only. (In previous works [5], [2], [13] we have considered the case $p = 1$ only.) Moreover, using approach from [11], we prove the following continuity property of the model: if the 2nd pool approach stationary regime (that is $\lambda_2 \to \infty$), then the distribution of the 1st pool approaches the stationary distribution corresponding to the system in which the 2nd pool is initially in the stationary regime. In this work we also verify the continuity property by simulation.

The paper is organized as follows. We describe the model in Section II. Section III contains the main mentioned above theoretical results: solution of Kolmogorov equations and the proof of convergence queue-size distribution in pool 1 to distribution for the model with initially stationary pool 2.

Section IV contains simulation results.

II. Description of the model

We consider a two-pool queueing model with infinite-capacity buffers. The 1st pool contains N_1 servers, while the 2nd pool contains only one server (we often will not distinguish 2nd pool and 2nd server). The 2nd pool uses multiple vacations policy: when idle, it becomes inactive time to time. It is assumed that the 1st pool is fed by a Poisson input with rate λ_1. Arriving class-1 customers can be served by the servers of both pools, and it reflects a flexibility of the servers. Service times of class-1 customers are exponential with rate μ_1. If server 2 is active at a moment t, then it inspects the state of the 1st pool and, if the queue size of the 1st pool $Q_1(t)$ exceeds a given threshold $C \geq 0$, then a waiting class-1 customer may jump to the server of pool 2 with a probability p, becoming a class-(1, 2) customer. Service time of such a customer is exponential with rate μ_{12}. But if the 2nd pool is active at the instant t and $Q_1(t) < C$, then the 2nd pool starts an inactive period (vacation). The inactivity periods have exponential duration with rate μ_2. As we mentioned above such a regime of the 2nd pool is called stationary. To obtain an analytic solution, we consider only the case when service rate of class-(1,2) customers equals the rate of the inactivity period lengths of pool 2, that is $\mu_1 = \mu_{12}$.

It is assumed that service times of class-i customers $\{S_k^{(i)},\, k \geq 1\}$ are independent, exponential with rate

$$\mu_i = 1/\mathsf{E}S^{(i)} \in (0,\, \infty),\, i = 1, (1,2).$$

(In what follows, we omit the serial index to denote a generic element of an i.i.d sequence.) All sequences are assumed to be independent.

We denote $Q_1(t)$, $X_1(t)$, $Z_1(t)$ the number of customer waiting in the queue, the number of busy servers and the total number of customers in pool 1, respectively, at instant t^-. We note, that it is does not matter for stability analysis, which waiting class-1 customer jumps to pool 2, when server of pool 2 is active and $Q_1(t) \geq C$.

III. Theoretical results

In this section, we derive the stationary distribution of the number of customers at pool 1.

For this purpose we compose Kolmogorov equations for the stationary probabilities of the state of the 1st queue, considering that the 2nd pool is in stationary regime initially. Introduce traffic intensities

for $k = 1, \ldots, N_1$

$$\rho_k = \frac{\lambda_1}{k\mu_1},$$

and

$$\rho_{N_1+C+1} = \frac{\lambda_1}{N_1\mu_1 + p\mu_2}.$$

It is easy to check, that the following balance relations for stationary distribution of the process $\{Z_1(t)\}$ hold true:

for $k = 0, \ldots, N_1 - 1$ the following equations hold

$$\lambda_1 \mathsf{P}_k = (k+1)\mu_1 \mathsf{P}_{k+1},$$

whence it follows, that

$$\mathsf{P}_{k+1} = \prod_1^{k+1} \rho_i \mathsf{P}_0. \tag{1}$$

For $k = 0, \ldots, C-1$

$$\lambda_1 \mathsf{P}_{N_1+k} = N_1\mu_1 \mathsf{P}_{N_1+k+1},$$

implying

$$\mathsf{P}_{N_1+k+1} = \rho_{N_1}^{k+1} \prod_1^{N_1} \rho_i \mathsf{P}_0. \tag{2}$$

Also for $k \geq 0$

$$\lambda_1 \mathsf{P}_{N_1+C+k} = (N_1\mu_1 + p\mu_2)\mathsf{P}_{N_1+C+k+1},$$

and we obtain

$$\mathsf{P}_{N_1+C+k+1} = [\rho_{N_1+C+1}]^{k+1}[\rho_{N_1}]^C \prod_1^{N_1} \rho_i \mathsf{P}_0. \tag{3}$$

By means of normalization condition $\sum_{k=0}^{\infty} \mathsf{P}_k = 1$, we obtain

$$\begin{aligned} 1 &= \mathsf{P}_0 + \mathsf{P}_0 \sum_{l=1}^{N_1} \prod_{i=1}^{l} \rho_i + \mathsf{P}_0 \sum_{l=1}^{C} \rho_{N_1}{}^{l} \prod_{i=1}^{N_1} \rho_i \\ &+ \mathsf{P}_0 [\rho_{N_1}]^C \sum_{k=1}^{\infty} [\rho_{N_1+C+1}]^k \prod_{i=1}^{N_1} \rho_i. \end{aligned} \tag{4}$$

It gives the following explicit expression for P_0:

$$\begin{aligned} \mathsf{P}_0 &= \Big[1 + \sum_{l=1}^{N_1} \prod_{i=1}^{l} \rho_i + \prod_{i=1}^{N_1} \rho_i \frac{\rho_{N_1}(1 - [\rho_{N_1}]^C)}{1 - \rho_{N_1}} \\ &+ \prod_{i=1}^{N_1} \rho_i [\rho_{N_1}]^C \frac{\rho_{N_1+C+1}}{1 - \rho_{N_1+C+1}} \Big]^{-1}, \end{aligned} \tag{5}$$

where, recall,

$$\rho_{N_1+C+1} = \frac{\lambda_1}{N_1\mu_1 + p\mu_2}. \tag{6}$$

Recall, that $\mathsf{E}Q_1$ is the mean stationary number of customers in the queue of the pool 1.

$$\begin{aligned} \mathsf{E}Q_1 &= \sum_{k=N_1+1}^{\infty} (k - N_1)\mathsf{P}_k \\ &= \prod_{i=1}^{N_1} \rho_i P_0 \sum_{k=N_1+1}^{N_1+C} (k - N_1)\Big[\rho_{N_1}\Big]^{k-N_1} \\ &+ \rho_{N_1}^C \prod_{i=1}^{N_1} \rho_i P_0 \sum_{k=N_1+C+1}^{\infty} (k - N_1)[\rho_{N_1+C+1}]^{k-N_1-C} \\ &= \prod_{i=1}^{N_1} \rho_i P_0 [\frac{\rho_{N_1}}{(1-\rho_{N_1})^2}(1 - \rho_{N_1}^C - C\rho_{N_1}^C + C\rho_{N_1}^{C+1}) \\ &+ \frac{\rho_{N_1}^C \rho_{N_1+C+1}(C - C\rho_{N_1+C+1} + 1)}{(1 - \rho_{N_1+C+1})^2}\Big]. \end{aligned} \tag{7}$$

Recall, that $\mathsf{E}X_1$ is the mean stationary number of busy servers in the 1st pool (assumed, that pool 2 is in a stationary regime).

$$\begin{aligned}\mathsf{E}X_1 &= \sum_{k=1}^{N_1} k\mathsf{P}_k + N_1 P(k \geq N_1+1, k \leq N_1 + C)\\ &+ N_1 P(k \geq N_1 + C + 1)\\ &= P_0\Big[\sum_{k=1}^{N_1} k \prod_{i=1}^{k}\rho_i + N_1\rho_{N_1}\frac{1-\rho_{N_1}{}^C}{1-\rho_{N_1}}\prod_{i=1}^{N_1}\rho_i\\ &+ N_1\rho_{N_1}{}^C\frac{\rho_{N_1+C+1}}{1-\rho_{N_1+C+1}}\prod_{i=1}^{N_1}\rho_i\Big]. \qquad (8)\end{aligned}$$

Next we study the model with the server of pool 2, which is in a "transient "regime. It means, that intervals between starting points of inactivity periods of the server are exponentially distributed with rate λ_2. Thus the stationary distribution $\{\mathsf{P}_k\}$ we found above (see (1)-(3), (5)) formally relates to the case, when $\lambda_2 = \infty$, which we call *stationary multiple vacation regime* of the 2nd pool, or stationary regime, for short. Now we prove the following convergence property of the process $\{Z_1(t)\}$: the distribution of the process $\{Z_1(t)\}$ converges, as $\lambda_2 \to \infty$, to the stationary distribution $\{\mathsf{P}_k\}$ (which corresponds to initially stationary pool 2).

Note, that it has been proved in [12], that the 1st pool is stationary, if the following sufficient condition holds:

$$\frac{\mu_1 N_1 + p\mu_2 - \lambda_1}{\lambda_1} > 0. \qquad (9)$$

Condition 9 provides stability of the 1st pool apart from the threshold C.

To prove this property, we use a condition obtained in [11], which is formulated below for the birth-and-death process $\{Z_1(t), t \geq 0\}$ with birth (input) rates $\lambda(k)$ and death (service) rates $\mu(k)$, where k is the current state of the process Z_1. In case, when $C = 0$, we obtain, that birth and death rates as follows:

$$\begin{aligned}\lambda(k) &= \lambda_1,\\ \mu(k) &= \mu_1 k, \ \ k \leq N_1,\\ \mu(k) &= \mu_1 N_1 + p\mu_2, \ \ k > N_1.\end{aligned}$$

We must verify the following condition from [11]:

$$\inf_{k\geq 0}\Big(\lambda(k) + \mu(k+1) - \frac{d_{k-1}}{d_k}\mu(k) - \frac{d_{k+1}}{d_k}\lambda(k+1) \quad > 0, \quad (10)$$

where constants d_k must be positive. We take the following constants:

$$\begin{aligned}d_k &= 1, k = -1, \cdots, N_1 - 1,\\ d_{N_1} &= 1 + \epsilon = \delta,\\ d_{N_1+k} &= \delta^{k+1}, \ \ k \geq 1,\end{aligned}$$

where $\epsilon > 0$ will be selected below.

For $k = 0\ , \cdots\ , N_1{-}2$ we obtain, that condition (10) indeed holds:

$$\lambda_1 + \mu_1(k+1) - \mu_1 k - \lambda_1 = \mu_1 > 0.$$

For k=$N_1 - 1$, we have

$$\lambda_1 + \mu_1 N_1 - \mu_1(N_1 - 1) - (1+\epsilon)\lambda_1 = \mu_1 - \epsilon\lambda_1 > 0,$$

if we take $\epsilon < \mu_1/\lambda_1$.

For k=N_1 it follows, that

$$\begin{aligned}&\lambda_1 + \mu_1 N_1 + p\mu_2 - \frac{1}{1+\epsilon}\mu_1 N_1 - (1+\epsilon)\lambda_1\\ &= \frac{-\epsilon^2\lambda_1 + \epsilon(\mu_1 N_1 + \mu_2 - \lambda_1) + \mu_2}{1+\epsilon} > 0,\end{aligned}$$

if in turn, we select $\epsilon < \epsilon_1$, where

$$\epsilon_1 = \frac{\mu_1 N_1 + p\mu_2 - \lambda_1 + \sqrt{(\mu_1 N_1 + p\mu_2 - \lambda_1)^2 + 4p\lambda_1\mu_2}}{2\lambda_1}$$

is a positive root of a quadratic function

$$-\epsilon^2\lambda_1 + \epsilon(\mu_1 N_1 + \mu_2 - \lambda_1) + \mu_2 = 0.$$

Finally, for $k \geq N_1 + 1$, we have

$$\lambda_1 + \mu_1 N_1 + p\mu_2 - \frac{1}{1+\epsilon}(\mu_1 N_1 + p\mu_2) - (1+\epsilon)\lambda_1 > 0,$$

if

$$\epsilon < \frac{1}{\lambda_1}(\mu_1 N_1 + p\mu_2 - \lambda_1).$$

It is clear, that

$$\epsilon_1 > \frac{1}{\lambda_1}(\mu_1 N_1 + p\mu_2 - \lambda_1).$$

Taking into account all restrictions to ϵ, we obtain, that condition (10) holds if ϵ satisfies the following constraints:

$$0 < \epsilon < \min\Big(\frac{\mu_1}{\lambda_1}, \frac{\mu_1 N_1 + p\mu_2 - \lambda_1}{\lambda_1} \quad . \qquad (11)$$

It remain to note that $\epsilon > 0$, satisfying (11) exists by condition (9).

IV. Simulation

In this section we demonstrate convergence of $\mathsf{E}Q_1$ and $\mathsf{E}X_1$ in the model with the 2nd pool, which is in a transient regime to the corresponding values in the model with initially stationary server of the 2nd pool.

We denote by $\hat{\mathsf{E}}Q_1$ and $\hat{\mathsf{E}}X_1$ the sample mean estimates of the mean queue size $\mathsf{E}Q_1$ and the mean number of busy servers $\mathsf{E}X_1$ obtained in formulas (7) and (8), respectively.

In simulation we apply the number of arrivals $n = 100000$ and

$$\lambda_2{}^{(j)} = (j+1)\lambda_2{}^{(0)}, 0 \leq j \leq 13,$$

where $\lambda_2{}^{(0)} = 0.5$. To obtain smoothed trajectory, in each case we perform 100 runs. We use "R studio" software to run the simulative model.

Recall, that this property was proved for case with $C = 0$. And we demonstrate it for the system with exponential service time and the following parameters

$$\lambda_1 = 18, \mu_1 = 10, \mu_{12} = \mu_2 = 5, p = 0.1, N_1 = 2$$

(see Fig. 1 and Fig. 2).

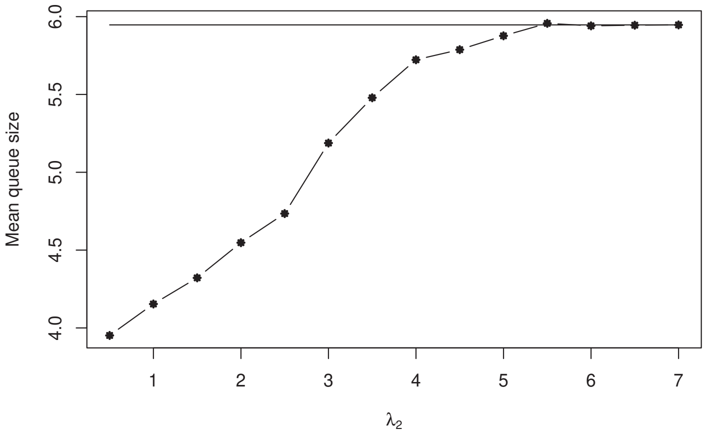

Fig. 1. Convergence of $\hat{\mathsf{E}}Q_1$ to the theoretical value $\mathsf{E}Q_1 = 5.947, C = 0, p = 0.1$

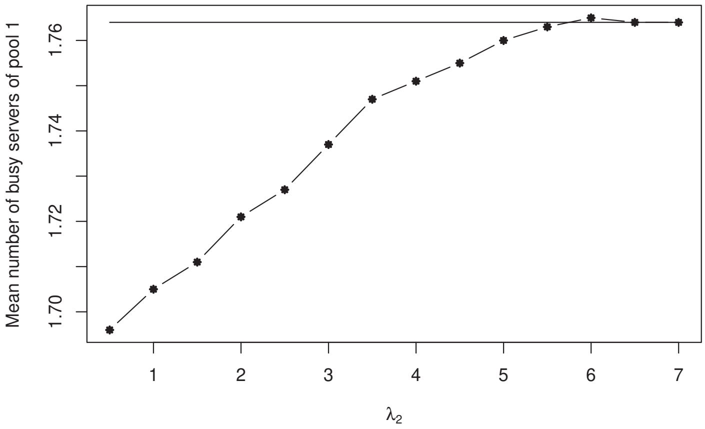

Fig. 2. Convergence of $\hat{\mathsf{E}}X_1$ to the theoretical value $\mathsf{E}X_1 = 1.764, C = 0, p = 0.1$

Also we demonstrate convergence of $\hat{\mathsf{E}}Q_1$ to $\mathsf{E}Q_1$ and $\hat{\mathsf{E}}X_1$ to $\mathsf{E}X_1$ for the system with exponential service time and the following parameters

$$\lambda_1 = 18, \mu_1 = 10, \mu_{12} = \mu_2 = 5, p = 0.1, N_1 = 2$$

with $C = 1$ (see Fig. 3 and Fig. 4).

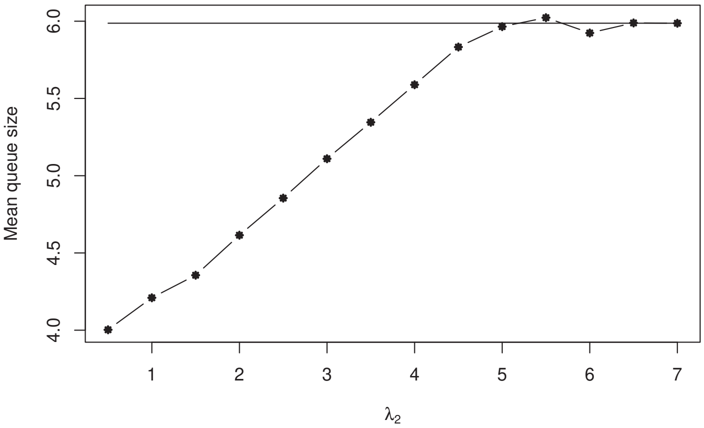

Fig. 3. Convergence of $\hat{\mathsf{E}}Q_1$ to the theoretical value $\mathsf{E}Q_1 = 5.987, C = 1, p = 0.1$

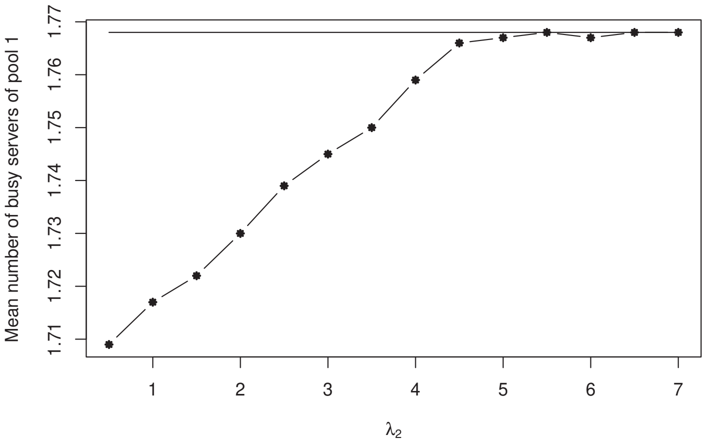

Fig. 4. Convergence of $\hat{\mathsf{E}}X_1$ to the theoretical value $\mathsf{E}X_1 = 1.768, C = 1, p = 0.1$

Fig. 5 and Fig. 6 demonstrate monotone decrease of $\mathsf{E}\,\hat{}\,Q_1$ and $\hat{\mathsf{E}}X_1$ as probability of a jump p increases for the system with

$$\lambda_1 = 18, \lambda_2 = 7, \mu_1 = 10, \mu_{12} = \mu_2 = 5, N_1 = 2, C = 0$$

and number of arrivals $n = 100000$. Fig. 7 and Fig. 8 illustrate the same property of $\hat{\mathsf{E}}Q_1$ and $\hat{\mathsf{E}}X_1$ for the same system with $C = 1$.

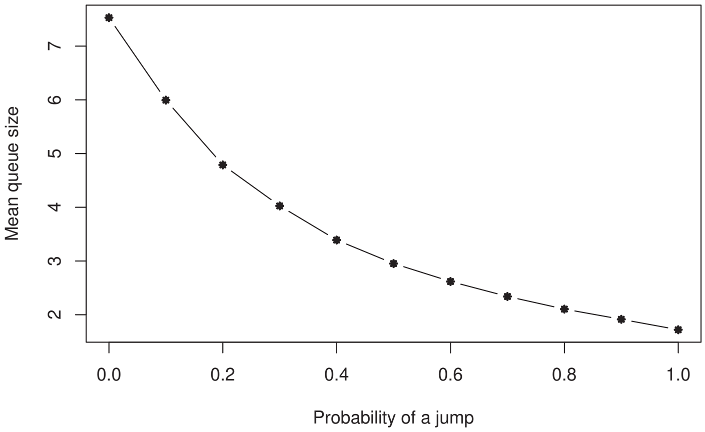

Fig. 5. Monotone decrease of $\hat{\mathsf{E}}Q_1$, $C = 0$

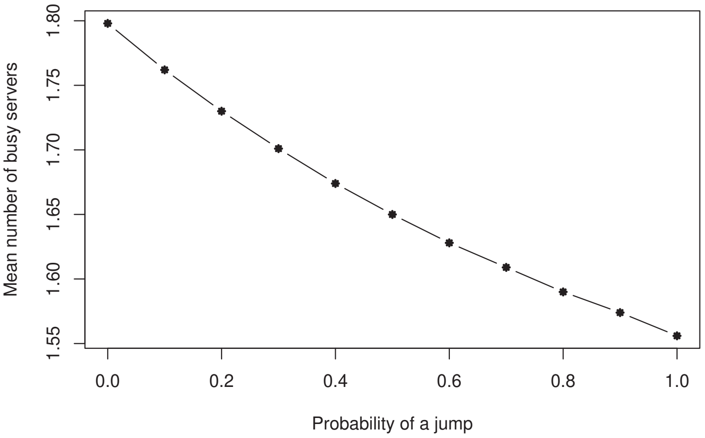

Fig. 6. Monotone decrease of $\hat{\mathsf{E}}X_1$, $C = 0$

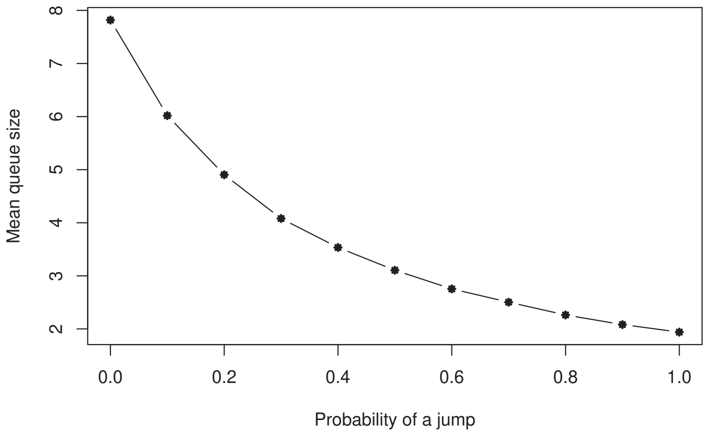

Fig. 7. Monotone decrease of $\hat{\mathsf{E}}Q_1$, $C = 1$

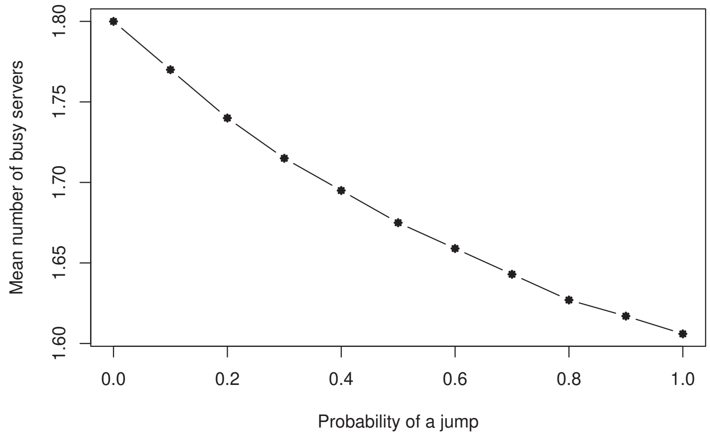

Fig. 8. Monotone decrease of $\hat{\mathsf{E}}X_1$, $C = 1$

As expected, when threshold C is fixed and probability of a jump p increases, both $\hat{\mathsf{E}}Q_1$ and $\hat{\mathsf{E}}X_1$ decreases. It happens, because waiting class-1 customers jump to the 2nd pool with larger probability.

Fig. 9 and Fig. 10 demonstrate monotone increase of $\mathsf{E}\,\hat{}\,Q_1$ and $\hat{\mathsf{E}}X_1$ as threshold C increases for the system with

$$\lambda_1 = 18, \lambda_2 = 7, \mu_1 = 10, \mu_{12} = \mu_2 = 5, N_1 = 2, p = 1$$

and number of arrivals $n = 100000$.

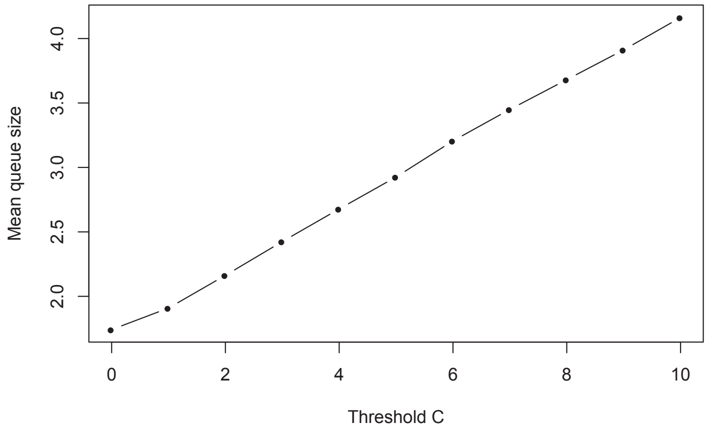

Fig. 9. Monotone increase of $\hat{\mathsf{E}}Q_1$

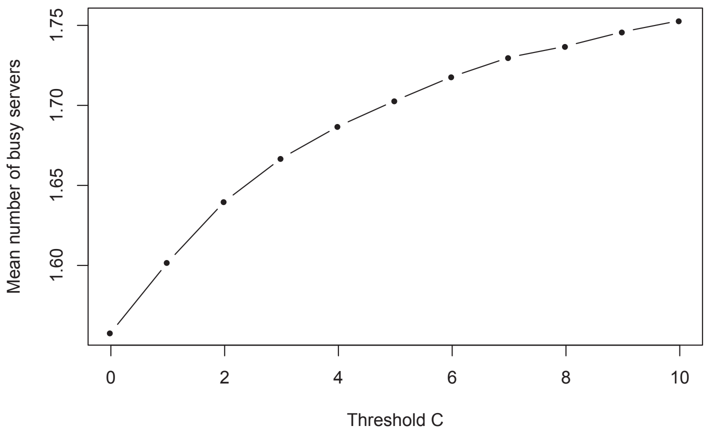

Fig. 10. Monotone increase of $\hat{\mathsf{E}}X_1$

V. Conclusion

In this paper, we study the N-model consisting of two pools where the 1st pool is a classic queueing system and the server of the 2nd pool uses multiple vacations policy. When queue size in the 1st pool exceeds a threshold C, a waiting customer may jump to the 2nd pool, if it is active at this instant. We derive the stationary distribution of the number of customers of the 1st pool. Moreover, we find condition implying convergence of the basic birth-death process in the 1st pool to a stationary distribution, when the 2nd pool approaches a multiple vacation policy. Theoretical results are illustrated by a few numerical examples obtained by simulation.

Acknowledgement

The research is in part supported by the Russian Foundation for Basic Research, projects 18-07-00147, 18-07-00156 and 19-07-00303. The authors are thankful to Alexander Rumyantsev for useful comments.

References

[1] T. Tezcan, "Stability analysis of N-model systems under a static priority rule", *Queueing System*, vol. 73, 2013, pp.235-259.

[2] E. Morozov, M. Maltseva, B. Steyaert, "Verification of the stability of a two-server queueing system with static priority", *2018 22nd Conference of Open Innovations Association (FRUCT)*, pp. 166-172, 2018.

[3] O. Garnet,A. Mandelbaum, "An introduction to Skills-Based Routing and its operational complexities", http://iew3.technion.ac.il/serveng/Lectures/SBR.pdf, 2000.

[4] W. Whitt, "Blocking when service is required from several facilities simultaneously", *ATT Techincal Journal*, vol. 64, issue 8, 1985, pp. 1807-1856.

[5] E. Morozov, "Stability of a two-pool n-model with preemptive-resume priority", *Distributed Computer and Communication Networks, Springer Internation Publishing*, pp. 399-409, 2018.

[6] D. Wong, N. Paciorek, T. Walsh, J. DiCelie, M. Young, B. Peet, "Concordia: An infrastructure for collaborating mobile agents. International Workshop on Mobile Agents MA 1997: Mobile Agents", *LNCS*, Springer, vol. 1219, 1997, pp. 86-97.

[7] S.R. Agnihothri, A.K. Mishra, D.E. Simmons, "Workforce cross-training decisions in field service systems with two job types", *Journal of the Operational Research Society*, vol. 54, issue 4, 2003, pp. 410-418.

[8] M. Ahghari, B. Balcioglu, “Benefits of cross-training in a skill-based routing contact center with priority queues and impatient customers”, *IIE Transactions* , vol. 41, 2009, pp. 524-536.

[9] E. Tekin, W.J. Hopp, M.P. Van Oyen, “Pooling strategies for call center agent cross-training”, *IIE Transactions*, vol. 41, no. 6, 2009, pp. 546-561.

[10] D. Terekhov, J.C. Beck, “An extended queueing control model for facilities with front room and back room operations and mixed-skilled workers”, *European Journal of Operational Research*, vol. 198, issue 1, 2009, pp. 223-231.

[11] A. Zeifman, “On the ergodicity of nonhomogeneous birth and death processes”, *Journal of Mathematical Sciences*, 1994.

[12] E. Morozov, “Stability of a two-pool n-model with preemptive-resume priority”, *Springer International Publishing*, 2018.

[13] M. Maltseva, E. Morozov, “Stability analysis and simulation of an N-model with two interacting pools”, *CEUR Workshop Proceedings*, 2018.