



Stability Analysis and Simulation of an N -Model with Two Interacting Pools

Mariia Maltseva²  and Evsey Morozov^{1,2} 

¹ Institute of Applied Mathematical Research, Karelian Research Centre of the RAS,
Petrozavodsk, Russia,
emorozov@karelia.ru

² Petrozavodsk State University,
Petrozavodsk, Russia

Abstract. We consider the so-called N -model which contains two pools of servers and two classes of external customers following independent Poisson inputs. Service times are class-dependent and, in each pool, are i.i.d. Pool 1 consists of N_1 servers and pool 2 consists of one server. When all servers of pool 1 are occupied, and there are waiting customers in the queue of pool 1, then a class-1 customer jumps to server of pool 2, becoming a class-(1,2) customer. We consider a non-preemptive service priority: a class-(1,2) customer starts service in the server of pool 2, when an ongoing customer service, if any, is completed. The purpose of the research is to deduce explicit stationary distribution of number of customers at the 1st pool, and to verify stability conditions of the model. Moreover, we simulate a model with one server in the 1st pool, $N_1 = 1$, in which class-1 customers jump to pool 2 provided the queue at pool 1 exceeds a positive threshold C . In this setting we verify by simulation that (i) for each fixed C , the stationary idle probability P_0 of server 1 attains minimum when the 2nd server is always busy with class-2 customers (saturated regime) and (ii) P_0 decreases as the threshold C increases.

Keywords: two-pool N -model, stability, stationary distribution, non-preemptive priority, monotonicity, simulation

1 Introduction

This work continues the study of the so-called N -model considered earlier in [4], and which, in turn, is a variation of the model studied in [8]. This model belongs to a class of network models with so-called Skills-Based Routing which describes the routing of customers. This routing can be dynamic, depending on the state of the system. In the opposite case, this routing is predefined in advance, say by assignment of the servers among customers depending on their priority. The latter routing is called static. The design of Skills-Based Routing is an actual and challenging problem. A multi-server pool is equivalent to multiple equally-skilled single servers, see for instance, [3,9].

For instance, N -model means that in the two-pool system, each server of the 1st pool serves only the customers of class-1 customers, while each server of the 2nd pool can serve both class-1 and class-2 customers. In this model, there are two independent Poisson inputs of class- i customers, where class- i customers join the queue or occupy the available server of pool $i = 1, 2$. 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 with preemptive-resume priority over class-2 customers, if any. As it is mentioned above, such a model is called N -model with static priority in [8]. This model has been also studied in [5] where stability conditions of each pool and both pools are found. However, stability conditions of the 2nd pool and the entire system contain an unknown parameter which in general depends on distributions of service times in both pools (and also on the input rates and threshold C). Motivation of this N -model can be found in [8,9,10,5,4]. In particular, servers of pool 1 can be treated as *beneficiary*, while the server of pool 2 can be called *donor*. These models constitute a wide class of the systems with interacting servers. This feature makes analytic investigation of such systems highly difficult. Also models under study belong to systems with flexible servers. Alternatively, to describe them one can use term *cross-trained* servers [1,2,6,7]. In these systems, some servers can serve a limited set of classes of customers, while remaining servers accept a broader set of classes of customers.

In some situation, service capacity may be transferred between servers for optimization of service process. Also one of possible applications of the model is the so-called *cognitive radio*, where a dynamic management is applied for using the best wireless channels in its neighborhood to avoid congestion. Such a radio can detect currently unused frequency bands and switch between such free channels without interruption of data transmission.

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 always busy by class-2 customers*. We call this system *saturated*. Moreover, using approach from [11], we prove the following intuitive continuity property of the model: if the queue size in pool 2 increases (in probability) then the stationary distribution of the 1st pool approaches the corresponding stationary distribution of the 1st pool with initially saturated pool 2. Evidently, each class-1 customer, before jumping to pool 2, must wait a time until class-2 customer, being served, departs server. (We recall that class-1 customers have non-preemptive priority.) It indicates that the stationary idle probability P_0 of the 1st pool (that is the probability that all N_1 servers are idle) must attained minimum in the saturated system. In this work we verify this property by simulation for the system with threshold $C > 0$, and it is another contribution of the research. (In previous paper [4] we studied the system in which $C = 0$.) Finally, we verify by simulation that the probability P_0 decreases as the threshold C increases.

The paper is organized as follows. We describe the model in Section 2. Section 3 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 saturated pool 2. Section 4 contains simulation results.

2 Description of the model

We study a Markovian queueing model containing two pools of servers with infinite-capacity buffers. The 1st pool consists of N_1 servers working in parallel, while the 2nd pool contains only one server. Thus the first pool is a N_1 -server queueing system, while the second pool is a single-server system. It is worth mentioning that even in this relatively simple setting the analytic solution is available only in a special case, when service rate of class-(1,2) customers equals the service rate of class-2 customers.

The choice of such configuration of the pools is also caused by the following reason: in this case we are able to solve the corresponding Kolmogorov equations to find stationary distribution of the state of the first pool explicitly.

It is assumed that pool i is fed by a Poisson input with rate λ_i , $i = 1, 2$. Class-1 customers can be served in both pools, while class-2 customers can be accommodated by pool 2 only. Moreover, provided the number of waiting in the queue (in pool 1) class-1 customers exceeds a given threshold C , an arbitrary class-1 customer waiting in the 1st queue, can jump to pool 2, becoming class-(1,2) customer, where it has non-preemptive priority over class-2 customers. That is, the jumped customer starts service after the customer being served (if any) leaves the server of pool 2. We stress that at most one class-(1,2) customer can be in pool 2 simultaneously.

We assume that the service times of class- i customers $\{S_k^{(i)}, k \geq 1\}$ are independent, exponential with rate $\mu_i = 1/ES^{(i)} \in (0, \infty)$, $i = 1, 2, (1, 2)$. (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 $Q_i(t)$ the summary number of customers (including waiting in the queue) in pool i at instant t^- , $i = 1, 2$. We assume an arbitrary *work-conserving service discipline* in each pool, in particular, an *arbitrary waiting class-1 customer* may jump to the server of pool 2, provided $Q_1(t) > N_1 + C$ and pool 2 is idle, because performance analysis and stability/instability does not depend on the order of customers, which (in each class) are stochastically undistinguishable.

3 Theoretical results

In this section, we deduce explicit formulas for the 1st pool stationary probabilities. Then we formulate and verify a condition implying convergence of the queue-size distribution in original pool 1 with increasing class-2 customers to

stationary distribution of the 1st pool with initially saturated (by class-2 customers) pool 2.

A key assumption is that $\mu_{12} = \mu_2$, in which case class-2 and class-(1,2) customers are indistinguishable. As a result the one-dimensional process $Q_1(t)$, $t \geq 0$, turns out to be Markov, provided $Q_2(t) = \infty$.

Recall that the 1st pool consists of N_1 servers and the 2nd pool contains only one server. It is easy to verify that service rate of class-1 customers, at instant t , is defined as

$$\begin{aligned} \mu_1(t) &= \sum_{k=0}^{N_1} \mathbb{P}(Q_2(t) > 0, Q_1(t) = k)(\mu_1 k + \mu_2) \\ &\quad + \mathbb{P}(Q_2(t) > 0, Q_1(t) > N_1)(\mu_1 N_1 + \mu_2) \\ &\quad + \sum_{k=0}^{N_1} \mathbb{P}(Q_2(t) = 0, Q_1(t) = k)\mu_1 k. \end{aligned} \quad (1)$$

Our *main assumption* is that the 2nd pool approaches saturated regime, that is the queue size in the 2nd pool increases unlimitedly in distribution:

$$Q_2(t) \Rightarrow \infty, \quad t \rightarrow \infty. \quad (2)$$

Then it easily follows that, as $t \rightarrow \infty$,

$$\sum_{k=0}^{N_1} \mathbb{P}(Q_2(t) = 0, Q_1(t) = k) \rightarrow 0, \quad (3)$$

$$\mathbb{P}(Q_2(t) > 0, Q_1(t) > N_1) \rightarrow \mathbb{P}(Q_1 > N_1) =: P_{>N_1}, \quad (4)$$

$$\mathbb{P}(Q_2(t) > 0, Q_1(t) = k) \rightarrow \mathbb{P}(Q_1 = k) =: P_k, \quad k \geq 0. \quad (5)$$

In particular,

$$\sum_{k=0}^{N_1} \mathbb{P}(Q_2(t) > 0, Q_1(t) = k) \rightarrow \sum_{k=0}^{N_1} P_k. \quad (6)$$

Now relations (3)-(6) imply that, as $t \rightarrow \infty$,

$$\mu_1(t) \rightarrow \sum_{k=0}^{N_1} P_k(\mu_1 k + \mu_2) + P_{>N_1}(\mu_1 N_1 + \mu_2) =: \mu. \quad (7)$$

Now we construct Kolmogorov equations for the stationary probabilities of the state of the 1st queue, provided the 2nd queue is overloaded. Introduce traffic intensities

$$\rho_1 = \frac{\lambda_1}{\mu_1},$$

and, for $k = 2, \dots, N_1$,

$$\rho_k = \frac{\lambda}{k\mu_1 + \mu_2}.$$

It is easy to check that the following balance relations for stationary distribution of the process $\{Q_1(t)\}$ hold true:

$$\lambda_1 P_0 = \mu_1 P_1,$$

and, for $k = 1, \dots, N_1 - 1$,

$$\lambda_1 P_k = ((k+1)\mu_1 + \mu_2) P_{k+1},$$

implying

$$P_{k+1} = \prod_1^{k+1} \rho_i P_0. \quad (8)$$

At the same time,

$$\lambda_1 P_{N_1} = (N_1 \mu_1 + \mu_2) P_{N_1+1},$$

and we obtain

$$P_{N_1+1} = \prod_1^{N_1} \rho_i \rho_{N_1} P_0. \quad (9)$$

Further analysis shows that

$$\lambda_1 P_{N_1+k} = (N_1 \mu_1 + \mu_2) P_{N_1+k+1},$$

and finally we obtain

$$P_{N_1+k+1} = \prod_1^{N_1} \rho_i [\rho_{N_1}]^{k+1} P_0. \quad (10)$$

Using normalization condition $\sum_{k=0}^{\infty} P_k = 1$, we obtain

$$1 = P_0 + P_0 \sum_{l=1}^{N_1} \prod_{i=1}^l \rho_i + P_0 \sum_{k=1}^{\infty} \prod_{i=1}^{N_1} \rho_i [\rho_{N_1}]^k. \quad (11)$$

It gives the following explicit expression for P_0 :

$$P_0 = \frac{1}{1 + \sum_{l=1}^{N_1} \prod_{i=1}^l \rho_i + \prod_{i=1}^{N_1} \rho_i \frac{\rho_{N_1}}{1 - \rho_{N_1}}}, \quad (12)$$

where, recall,

$$\rho_{N_1} = \frac{\lambda_1}{N_1 \mu_1 + \mu_2}. \quad (13)$$

Thus, the stationary probabilities P_k in (7) satisfy relations (8)-(13).

Our next goal is to show the following intuitive continuity result, which however needs to be strongly proved below. It is expected that, under main assumption (2), the distribution of the process $\{Q_1(t)\}$ converges to stationary distribution which corresponds to initially overloaded pool 2, by class-2 customers, that is $Q(0) = \infty$ with probability 1.

First of all we note that it has been proved in [5] that the 1st pool is stationary if the following sufficient condition holds:

$$\frac{\mu_1 N_1 + \mu_2 - \lambda_1}{\lambda_1} > 0. \quad (14)$$

Note, that this condition guarantees stability of the 1st pool regardless of the threshold C , when class-1 customers may jump to pool 2. The strong proof of the mentioned continuity property is based on a condition from [11] which is formulated below for the birth-and-death process $\{Q_1(t), t \geq 0\}$ with birth (input) rates $\lambda(k)$ and death (service) rates $\mu(k)$, where k is the current state of the process Q_1 . It is easy to see that in our setting

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

The condition we must verify is as follows [11]:

$$\inf_{k \geq 0} \left(\lambda(k) + \mu(k+1) - \frac{d_{k-1}}{d_k} \mu(k) - \frac{d_{k+1}}{d_k} \lambda(k+1) \right) > 0, \quad (15)$$

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

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

where $\epsilon > 0$ will be selected below. For $k = 0, \dots, N_1 - 2$ we obtain, that condition (15) indeed holds:

$$\lambda_1 + \mu_1(k+1) + \mu_2 - \mu_1 k - \mu_2 - \lambda_1 = \mu_1 > 0.$$

For $k=N_1 - 1$, we have

$$\lambda_1 + \mu_1 N_1 + \mu_2 - \mu_1(N_1 - 1) - \mu_2 - (1 + \epsilon)\lambda_1 = \mu_1 - \epsilon\lambda_1 > 0,$$

if we take $\epsilon < \mu_1/\lambda_1$. For $k=N_1$ it follows that

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

if in turn, we select $\epsilon < (\mu_1 N_1 + \mu_2 - \lambda_1)/\lambda_1$. Finally, for $k \geq N_1 + 1$, we have

$$\lambda_1(1 + \epsilon)^2 + (1 + \epsilon)^2(\mu_1 N_1 + \mu_2) - (1 + \epsilon)(\mu_1 N_1 + \mu_2) > 0,$$

if

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

Collecting all requirements to ϵ , we conclude that condition (15) is satisfied if we select ϵ in such a way that the following inequality holds:

$$0 < \epsilon < \min\left(\frac{\mu_1}{\lambda_1}, \frac{\mu_1 N_1 + \mu_2 - \lambda_1}{\lambda_1}\right). \quad (16)$$

It remain to note that $\epsilon > 0$, satisfying (16) exists by condition (14).

Denote $\mathbf{E}X_1$ the mean stationary number of busy servers in the 1st pool (provided the 2nd pool is overloaded). It has been proved in [8], that if

$$\frac{\lambda_1 - \mu_1 \mathbf{E}X_1}{\mu_{12}} + \frac{\lambda_2}{\mu_2} < 1, \quad (17)$$

then the system is stable. Using obtained above explicit expressions (11), (12) for stationary distribution of the 1st pool, we calculate that

$$\begin{aligned} \mathbf{E}X_1 &= \sum_{k=1}^{N_1} P_k k + N_1 P_{>N_1} \\ &= P_0 \sum_{k=1}^{N_1} k \prod_{i=1}^k \rho_i + N_1 P_0 \sum_{k=1}^{\infty} \rho_{N_1}^k \prod_{i=1}^{N_1} \rho_i \\ &= P_0 \left(\sum_{k=1}^{N_1} \prod_{i=1}^k \rho_i + N_1 \frac{\rho_{N_1}}{1 - \rho_{N_1}} \prod_{i=1}^{N_1} \rho_i \right). \end{aligned}$$

In particular, if $N_1 = 1$, then we find the required parameter $\mathbf{E}X_1$ in (17):

$$\mathbf{E}X_1 = P_0 \left(\rho_1 + \rho_1 \frac{\rho_{N_1}^2}{1 - \rho_{N_1}} \right) = \frac{\lambda_1(\mu_1 + \mu_2)}{\mu_1^2 + \mu_1 \mu_2 + \lambda_1 \mu_2}.$$

Note that, in this single-server case, the stationary number of busy servers equals the stationary busy probability of the 1st pool (server) in the saturated system, that is $\mathbf{E}X_1 = P_b = 1 - P_0$.

In the next section we check by simulation the monotone decrease of the stationary idle probability P_0 when threshold C increases.

4 Simulation results

In the following experiments, we demonstrate a monotonicity property of the estimate \hat{P}_0 of the stationary idle probability in original system with fixed value

of the threshold $C > 0$. We emphasize that these experiments have been started in previous paper [4] where we shown, for the system with threshold $C = 1$, that the estimate \hat{P}_0 of the stationary idle probability attains minimum in the saturated system.

Let $\hat{P}_0^{(o)}$ be the estimate of stationary idle probability P_0 (of the 1st server) in the saturated system. Denote the difference $\hat{d} = \hat{P}_0 - \hat{P}_0^{(o)}$. We show that the minimum of estimate \hat{P}_0 is $\hat{P}_0^{(o)}$ (when the 2nd server is permanently busy), implying $\hat{d} \geq 0$, provided the number of observations is large enough. For the system with exponential service time and the following parameters

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

the property is confirmed for $C = 5$ and $C = 10$, see Fig. 1 and Fig. 2, that illustrate convergence of \hat{d} to 0.008 and 0.001, respectively.

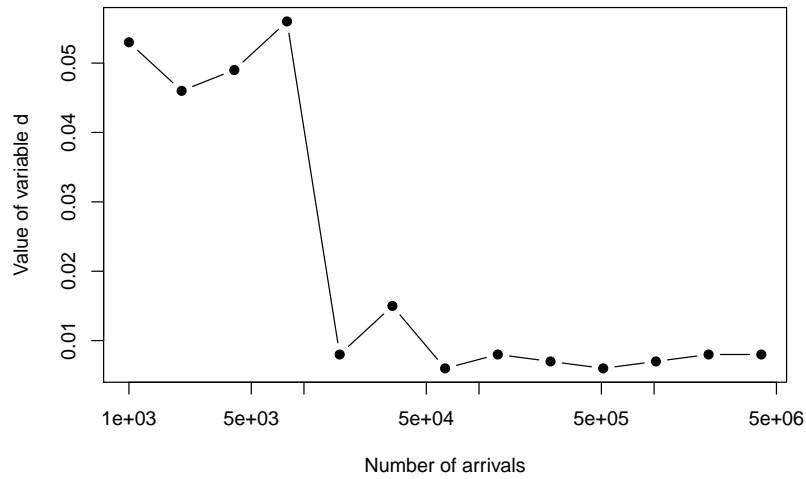


Fig. 1. Estimate \hat{d} for exponential service times, $C=5$

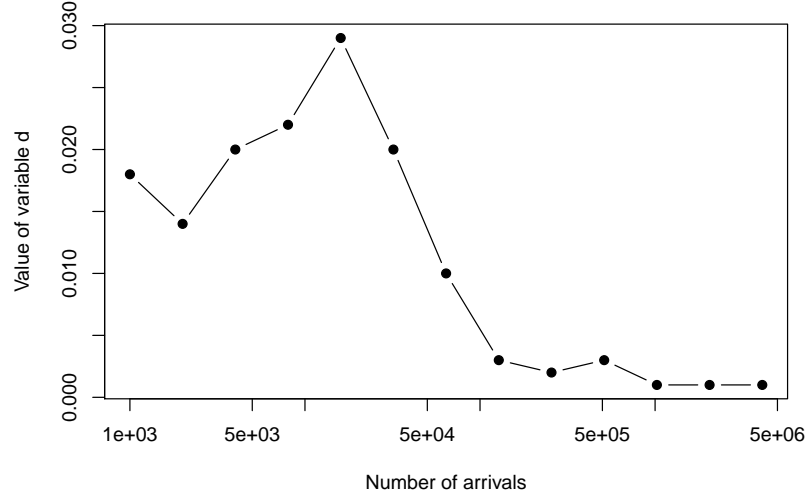


Fig. 2. Estimate \hat{d} for exponential service times, $C=10$

This property is also confirmed for the system with Pareto service time,

$$F(x) = 1 - (1/x)^{k_i}, \quad x \geq 1, \quad i = 1, 2, \quad (12),$$

with parameters

$$\lambda_1 = 0.5, \quad \lambda_2 = 0.4, \quad k_1 = 3, \quad k_{12} = k_2 = 5.$$

Fig. 3 illustrates convergence of \hat{d} to 0.001 and Fig. 4 illustrates convergence of \hat{d} to 0.002.

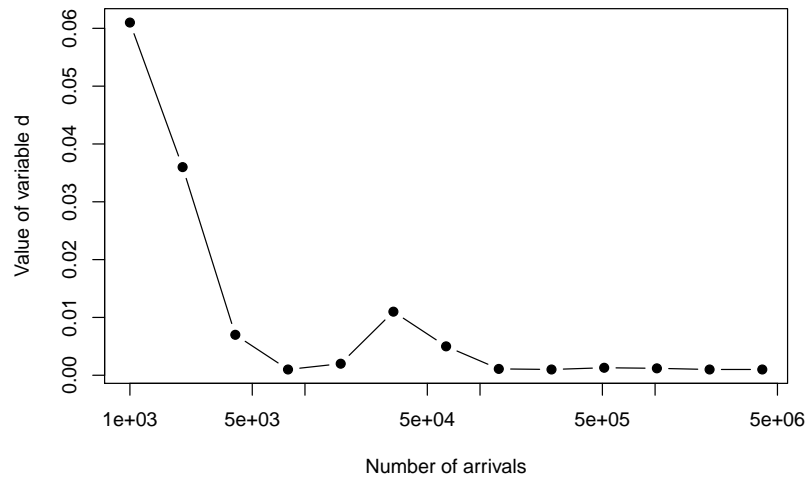


Fig. 3. Estimate \hat{d} for Pareto service times, C=5

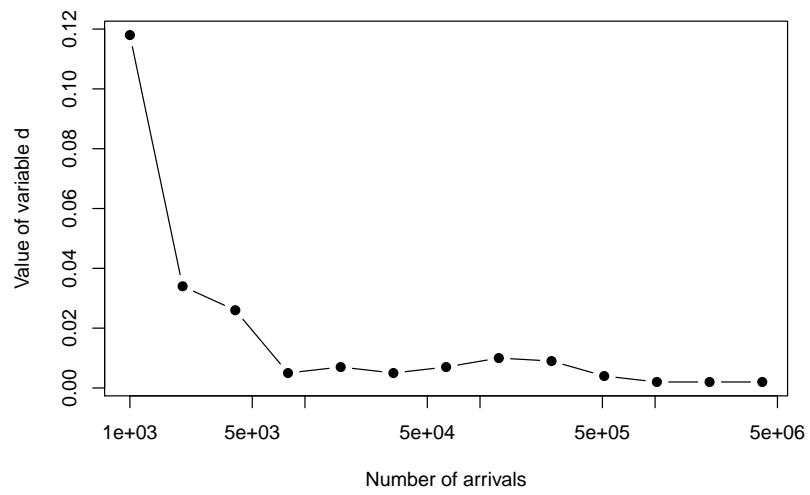


Fig. 4. Estimate \hat{d} for Pareto service times, C=10

Finally, we demonstrate that increasing C implies decreasing estimate \hat{P}_0 . We demonstrate it for the original model and model with the 2nd server overloaded by class-2 customers. Intuitively, this property is clear, because, when C increases, then class-1 customers in general wait more time in the 1st queue before jumping to pool 2. This property for 100000 arrivals, is confirmed for i) exponential service time with parameters $\lambda_1 = 7$, $\lambda_2 = 1$, $\mu_1 = 10$, $\mu_{12} = 5$, $\mu_2 = 5$ (Fig. 5); and ii) for Pareto service time with parameters $\lambda_1 = 0.5$, $\lambda_2 = 0.4$, $k_1 = 3$, $k_{12} = k_2 = 5$ (Fig. 6).

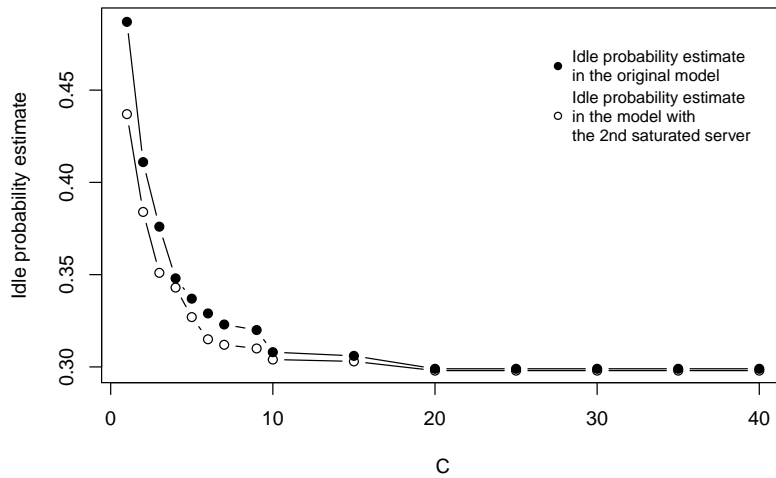


Fig. 5. The monotonicity of \hat{P}_0 and $\hat{P}_0^{(o)}$ for exponential service times as C increases

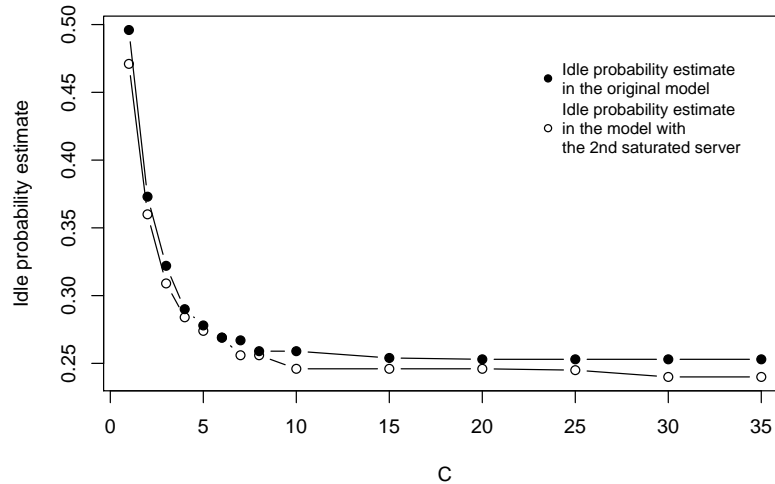


Fig. 6. The monotonicity of \hat{P}_0 and $\hat{P}_0^{(o)}$ for Pareto service times as C increases

ACKNOWLEDGEMENTS

The study was carried out under state order to the Karelian Research Centre of the Russian Academy of Sciences (Institute of Applied Mathematical Research KRC RAS). The research is partly supported by Russian Foundation for Basic Research, projects 18-07-00147, 18-07-00156.

References

1. Agnihotri, S.R., Mishra, A.K., Simmons, D.E.: Workforce cross-training decisions in field service systems with two job types. *Journal of the Operational Research Society* **54**(4), 410–418 (Apr 2003). <https://doi.org/10.1057/palgrave.jors.2601535>
2. Ahghari, M., Balcioglu, B.: Benefits of cross-training in a skill-based routing contact center with priority queues and impatient customers. *IIE Transactions* **41**(6), 524–536 (2009). <https://doi.org/10.1080/07408170802432975>
3. Garnet, O., Mandelbaum, A.: An introduction to skills-based routing and its operational complexities (2000), <http://iew3.technion.ac.il/serveng/Lectures/SBR.pdf>
4. Morozov, E., Maltseva, M., Steyaert, B.: Verification of the stability of a two-server queueing system with static priority. In: 2018 22nd Conference of Open Innovations Association (FRUCT). pp. 166–172 (May 2018). <https://doi.org/10.23919/FRUCT.2018.8468271>

5. Morozov, E.: Stability of a two-pool n-model with preemptive-resume priority. In: Vishnevskiy, V.M., Kozyrev, D.V. (eds.) *Distributed Computer and Communication Networks*. pp. 399–409. Springer International Publishing, Cham (2018)
6. Tekin, E., Hopp, W.J., Oyen, M.P.V.: Pooling strategies for call center agent cross-training. *IIE Transactions* **41**(6), 546–561 (2009). <https://doi.org/10.1080/07408170802512586>
7. Terekhov, D., Beck, J.C.: An extended queueing control model for facilities with front room and back room operations and mixed-skilled workers. *European Journal of Operational Research* **198**(1), 223 – 231 (2009). <https://doi.org/https://doi.org/10.1016/j.ejor.2008.08.013>
8. Tezcan, T.: Stability analysis of n-model systems under a static priority rule. *Queueing Systems* **73**(3), 235–259 (Mar 2013). <https://doi.org/10.1007/s11134-012-9304-z>
9. Whitt, W.: Blocking when service is required from several facilities simultaneously. *AT T Technical Journal* **64**(8), 1807–1856 (Oct 1985). <https://doi.org/10.1002/j.1538-7305.1985.tb00038.x>
10. Wong, D., Paciorek, N., Walsh, T., DiCelie, J., Young, M., Peet, B.: Concordia: An infrastructure for collaborating mobile agents. In: Rothermel, K., Popescu-Zeletin, R. (eds.) *Mobile Agents*. pp. 86–97. Springer Berlin Heidelberg, Berlin, Heidelberg (1997)
11. Zeifman, A.I.: On the ergodicity of nonhomogeneous birth and death processes. *Journal of Mathematical Sciences* **72**(1), 2893–2899 (Oct 1994). <https://doi.org/10.1007/BF01249905>