Analysis of Two-Way Communication Retrial Queuing Systems with Non-reliable Server, Impatient Customers to the Orbit and Blocking Using Simulation

Ádám Tóth(✉), János Sztrik, Tamás Bérczes, and Attila Kuki

University of Debrecen, Debrecen 4032, Hungary
{toth.adam,sztrik.janos,tamas.berczes,attila.kuki}@inf.unideb.hu

Abstract. The goal of this paper is to carry out a sensitivity analysis to examine the effect of different distributions of service time when blocking is applied with the help of retrial queueing systems having the property of two-way communication. This eventuates in outgoing calls (secondary customers) which are performed by the service unit after a random time in its idle state. Primary customers arrive from the finite-source according to an exponential distribution. This model does not contain queues so the service of an incoming request starts immediately if the server is functional and in an idle state. Impatience of the customers and server failures are characterized by this system which also follow an exponential distribution. The novelty of the investigation is to illustrate the effect of blocking with several figures obtained by simulation using various distributions of service time on the desired performance measures.

Keywords: Simulation · Blocking · Sensitivity analysis · Finite-source queueing system · Unreliable server · Retrial queue · Impatient customers

1 Introduction

The explosive growth of network traffic in recent years evokes the necessity of investigating communication networks to understand the behaviour of different systems. More and more communication sessions evolve partly almost every device becomes "smart" leading to higher bandwidth requirements not just in

The work of Ádám Tóth is supported by the ÚNKP-20-4 new national excellence program of the ministry for innovation and technology from the source of the national research, development and innovation fund. The research work of János Sztrik, Attila Kuki and Tamás Bérczes was supported by the construction EFOP-3.6.3-VEKOP-16-2017-00002. The project was supported by the European Union, co-financed by the European Social Fund.

V. M. Vishnevskiy et al. (Eds.): DCCN 2021, CCIS 1552, pp. 174–185, 2022.
https://doi.org/10.1007/978-3-030-97110-6_13

multinational companies but in our homes as well. So many unknown quantities may modify the performance of networking systems making them very complex and difficult to realize every aspect of their operation. Consequently, researchers dedicate their time to develop mathematical models describing telecommunication systems. With the help of retrial queueing systems arising real-life problems can be modelled in main telecommunication systems like telephone switching systems, call centers, or computer systems. These systems possess a virtual waiting room the so-called orbit where customers get into when the service unit is unavailable. Some examples are listed where queueing models are utilized: [1,5].

In this paper, the customer owns the impatience feature meaning that customers are able to decide to leave the system earlier without obtaining its service requisition. This is a natural occurrence of human behaviour and can be experienced in many fields of life like in healthcare applications, call centers, telecommunication networks so various works examine the effect of this phenomenon like in [11,13,15]. In these articles impatient request is portrayed: if the queue is sufficiently long balking customers choose to avoid entering the system, jockeying customers can alter queues if they encounter them may get served faster, and reneging customers leave the queue if they have waited a definite time for service.

Examining the available literature the considered models include service units that are assumed to be accessible all the time. This hypothesis does not reflect the reality as unexpected errors can take place like power outages, human negligence, or other sudden actions. Although devices are developing and become more reliable, unfortunate failures have a massive effect on the operation of the system modifying the performance measures significantly hence retrial queuing systems have been investigated in several papers recently for example in [4,8–10].

Two-way communication scheme gains ground ultimately due to its usefulness in many application fields modelling arising actual problems. One prime example is call-center where service units in an idle state may perform other activities besides satisfying the needs of incoming calls including selling, advertising, and promoting products. In other words, whenever the server is idle it may call for customers outside of the system after a random time. Utilization of such systems is always a key issue in that way many scientists are trying to optimize the service of different requests see for example [3,12,16,18].

The main focus of this paper is to carry out a sensitivity analysis inspecting the various distributions of service time of primary customers when blocking is applied on the main performance measures for instance the mean waiting time and the variance of an arbitrary, a successfully served and an impatient customer, the total utilization of the service unit, the probability of abandonment. Because giving exact formulas are difficult especially when one of the variables does not follow an exponential distribution, the obtained results are gathered by stochastic simulation based on SimPack [6] which contains the basic building blocks of the code. One of the main motivations is to develop simulation models in this way because it gives us the freedom to calculate any performance measure which we desire using various values of input parameters. The achieved results

indicate the relevance of the used distributions using various parameter settings and the effect of blocking illustrated by numerous figures concentrated on the interesting phenomena of these systems.

2 System Model

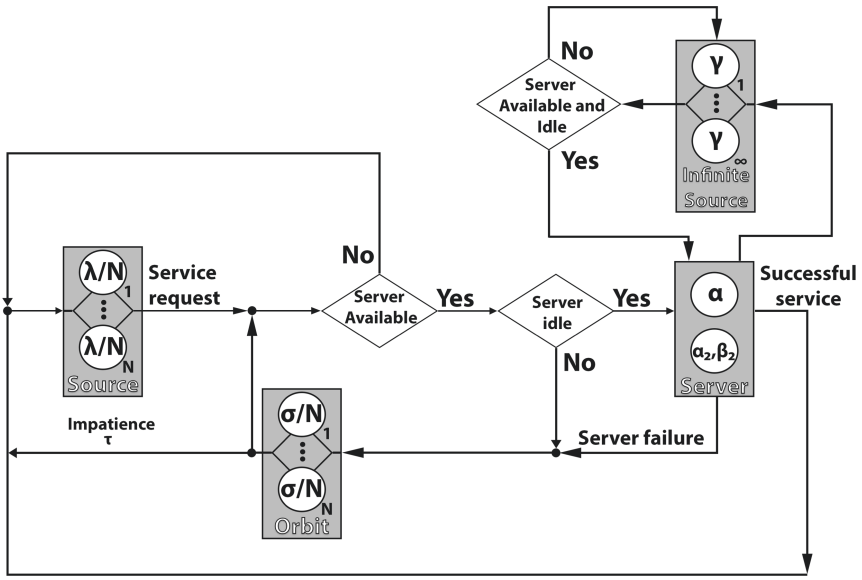

Fig. 1. System model

The regarded system is a retrial queueing system of type $M/G/1//N$ with impatient customers and an unreliable server that is capable of producing outgoing calls. N denotes the number of sources where each individual generates requests according to an exponential distribution with rate λ/N so the distribution of inter-request time is exponential with parameter λ/N (Fig. 1). There are no queues in our model in this way whenever an incoming customer finds the server in a busy state, it will be forwarded to the orbit. Otherwise, the service of an incoming customer starts instantly that follows gamma, hypo-exponential, hyper-exponential, Pareto, and lognormal distribution with different parameters but with the same mean value. During its residence in the orbit, a customer may launch an attempt to reach the service unit after an exponentially distributed time with parameter σ/N. Call generation can not occur until the end of the successful service of the individual in the source. We suppose that the service unit breaks down after an exponentially distributed time interval with parameter γ_0 when it is busy and with parameter γ_1 when idle. The repair time is also an exponentially distributed random variable with parameter γ_2 which starts

instantly after a failure takes place. During a faulty period, requests can not enter the system because of blocking. Customers have impatient characteristics therefore they may decide to leave the system after waiting an exponential time in the orbit with rate τ. As mentioned earlier an idle server may perform an outgoing call towards the customers (secondary) from an infinite source after an exponentially distributed time with parameter γ. The service of secondary customers is a gamma-distributed random variable with parameters α_2 and β_2. At the time the secondary request is arriving, if the server is busy or non-operational then it will be cancelled and returns without entering the system. In the case of breakdown:

- The service of a primary request is interrupted and it is forwarded immediately towards the orbit.
- The service of a secondary request is also interrupted but it departs the system.

3 Simulation

As mentioned earlier results are obtained by a self-developed simulation program and a statistical package [7] was integrated into our code to determine the performance measures. The method of batch means is used where the useful run is divided into N batches thus $n = M - K/N$ observations are carried out in every batch. K represents the warm-up period observations at the beginning of the simulation which is rejected. M represents the length of the simulation. We just simply calculate the sample average of the whole run after the warm-up period. To have a valid estimation, batches should be long enough and the sample averages of the batches should be approximately independent. In the following articles you can find more information about this process [2,14]. The simulations are performed with a confidence level of 99.9%. The relative half-width of the confidence interval required to stop the simulation run is 0.00001. The size of a batch used to detect the initial transient duration is 1000.

Table 1 display the used values of input parameters in our scenarios.

3.1 Scenario 1

We distinguished different scenarios where the values of service times of incoming customers are different to check how the various distribution modify the operation of the system. First, the squared coefficient of variation is greater than one, and to have a valid comparison we chose the parameters that the mean and variance would be the same in every case. For this, a fitting process was performed and [17] contains detailed info about these mechanisms (Table 2).

Figure 2 displays the probability ($P(i)$) that exactly i customer is located in the system. The figure shows that there is a significant disparity among the used distributions in the average number of requests in the system. Looking carefully at the obtained curves we could state that each of them corresponds to Gaussian distribution.

Table 1. Numerical values of model parameters

N	γ_0	γ_1	σ/N	γ	α_2	β_2	τ
100	0.05	0.5	0.01	0.8	1	1	0.001

Table 2. Parameters of service time of primary customers

Distribution	**Gamma**	**Hyper-exponential**	**Pareto**	**Lognormal**
Parameters	$\alpha = 0.037$ $\beta = 0.015$	$p = 0.482$ $\lambda_1 = 0.385$ $\lambda_2 = 0.416$	$\alpha = 2.018$ $k = 1.261$	$m = -0.751$ $\sigma = 1.826$
Mean	2.5			
Variance	169			
Squared coefficient of variation	27.04			

Figure 2 demonstrates the mean waiting time of an arbitrary customer in the function of arrival intensity when the service time of the customer follows a gamma distribution. The results prove what we expected aforehand when blocking is applied lower mean waiting time is obtained especially besides higher arrival intensity. The seen ratio is true for the other used distributions as well.

After noticing the effect of blocking, the next Figure (Fig. 4) shows the comparison of mean waiting time of an arbitrary customer besides the used distributions. With increasing arrival intensity, the mean waiting time increases and then, after reaching a certain value, starts to decrease. This tendency is valid for every curve regardless of the distribution. Although having the same first two moments, maximum property characteristic of a finite-source retrial queueing system arises even with the appearance of blocking (at Fig. 3). The other noteworthy thing about the figure is that the difference between the values obtained using the different distributions is significant especially in the case of Pareto distribution.

The variance of waiting time of a successfully served customer is depicted in Fig. 5 versus arrival intensity. Interestingly the differences are significant among the used distributions in spite of the selected parameters having the same first two moments. This is especially remarkable if we compare the values at gamma distribution with the values at Pareto distribution. This performance measure starts to escalate rapidly and after λ/N reaches 0.1 variance stagnates around a certain value.

3.2 Scenario 2

In this part after observing the results of the previous scenario, we were intrigued to see the effect of another parameter setting on the performance measures. In scenario 1 the squared coefficient of variation was greater than one so this time the parameters are chosen in order the squared coefficient of variation would be less than one. Because of this, the hyper-exponential distribution can

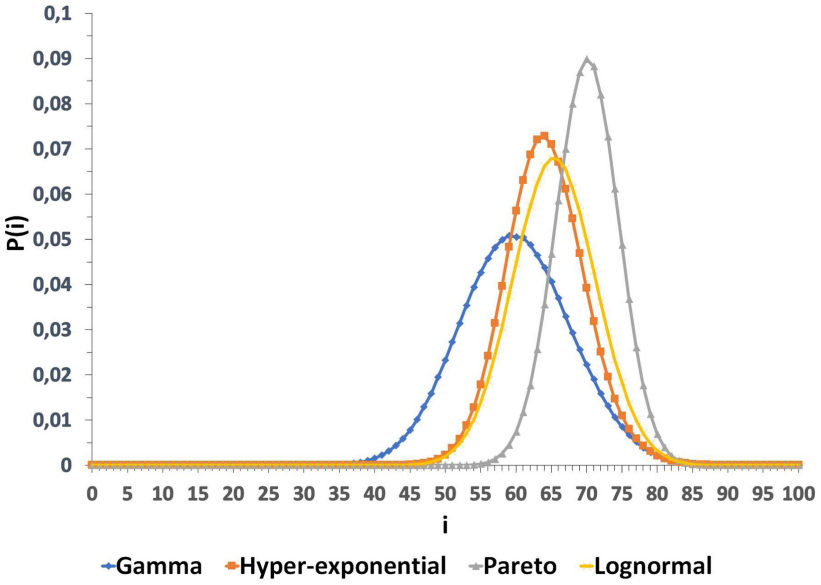

Fig. 2. Distribution of the number of customers in the system, $\lambda/N = 0.01$

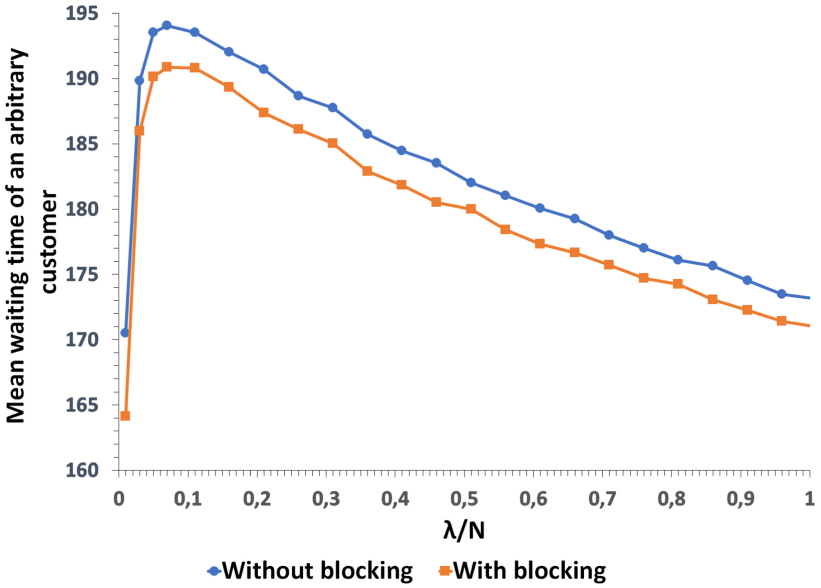

Fig. 3. The effect of blocking on the mean waiting of an arbitrary customer besides service time of gamma distribution

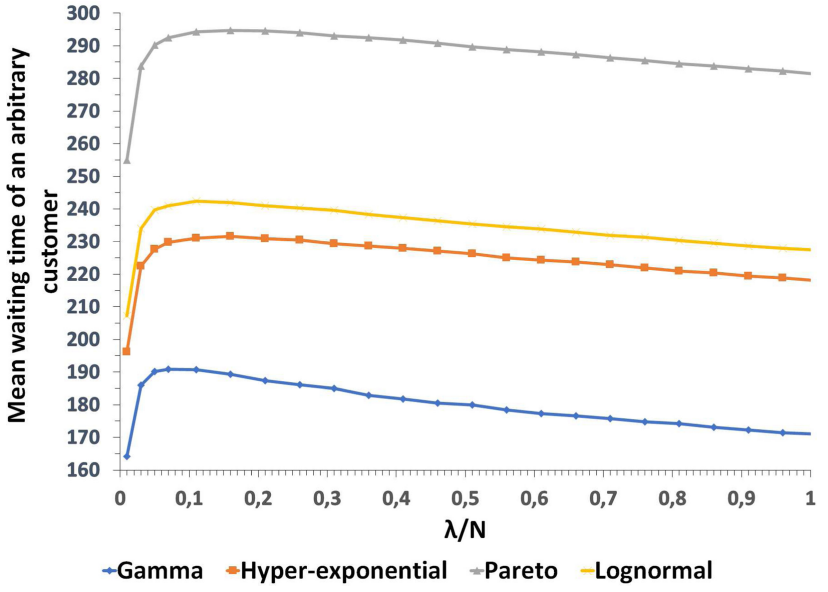

Fig. 4. The mean waiting time of an arbitrary customer

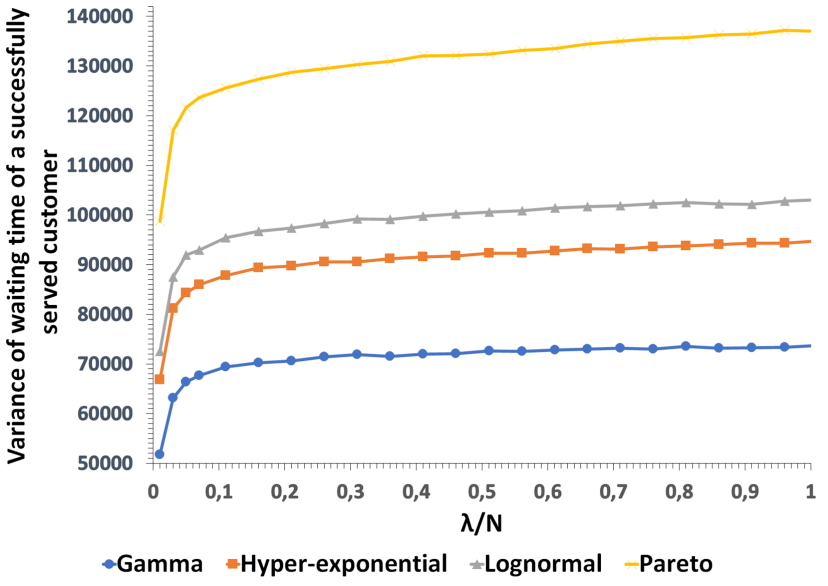

Fig. 5. The variance of waiting time of a successfully served customer

not be used that's why we replaced it with the hypo-exponential distribution. Table 3 contains the exact values of the parameters of the service time of primary customers in the case of this scenario, the other parameters remain unchanged which is shown in Table 1.

Table 3. Parameters of service time of primary customers

Distribution	**Gamma**	**Hypo-exponential**	**Pareto**	**Lognormal**
Parameters	$\alpha = 1.8$	$\mu_1 = 0.6$	$\alpha = 0.69$	$m = 2.67$
	$\beta = 0.72$	$\mu_2 = 1.2$	$k = 0.66$	$\sigma = 1.57$
Mean	2.5			
Variance	1.04			
Squared coefficient of variation	0.72222222			

The first figure (Fig. 6) shows the effect of blocking on the average waiting time of an arbitrary request as a function of the arrival intensity. With the other parameter setting, we saw that the average waiting time is lower in the blocking case, which of course applies here as well. Perhaps the only difference is that the curves are a little closer together in this scenario. However, a system with finite-source the maximum property characteristics appears even in the blocking case. It is worth mentioning that for the other distributions the difference is similar between the two cases.

How the increasing arrival intensity of the customers has an influence on the mean waiting time is illustrated in Fig. 7. Here, the mean and variance are the same again but compared to Fig. 4 the results indicate a completely different tendency. The obtained curves almost overlap each other, a minor difference can be observed at Pareto distribution but it is not significant. Similarly, after a while, the mean waiting time starts to decrease as in the previous scenario which is a characteristic of finite-source retrial queuing systems. Although the article presents results for one parameter setting, the interesting results described here were obtained are true for other settings.

After taking a closer look at the mean waiting time of an arbitrary customer, Fig. 8 demonstrates the variance of waiting time of a successfully served customer. In Scenario 1 the results in the previous scenario were significantly different from each other but here, with this parameter setting the curves are almost totally identical even for Pareto distribution. Another interesting thing about the figure is that, except for the Pareto distribution, the values obtained in this scenario are significantly higher than in the previous section.

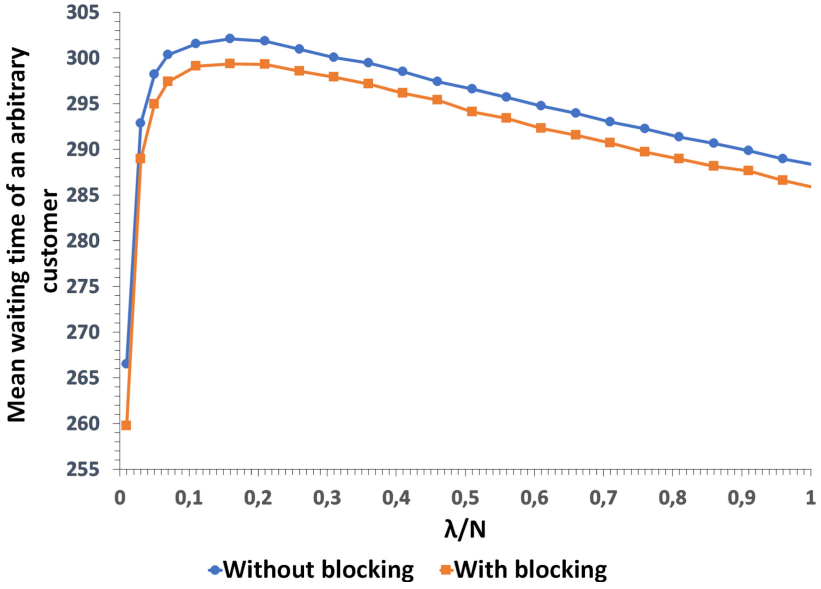

Fig. 6. The effect of blocking on the mean waiting of an arbitrary customer besides service time of gamma distribution

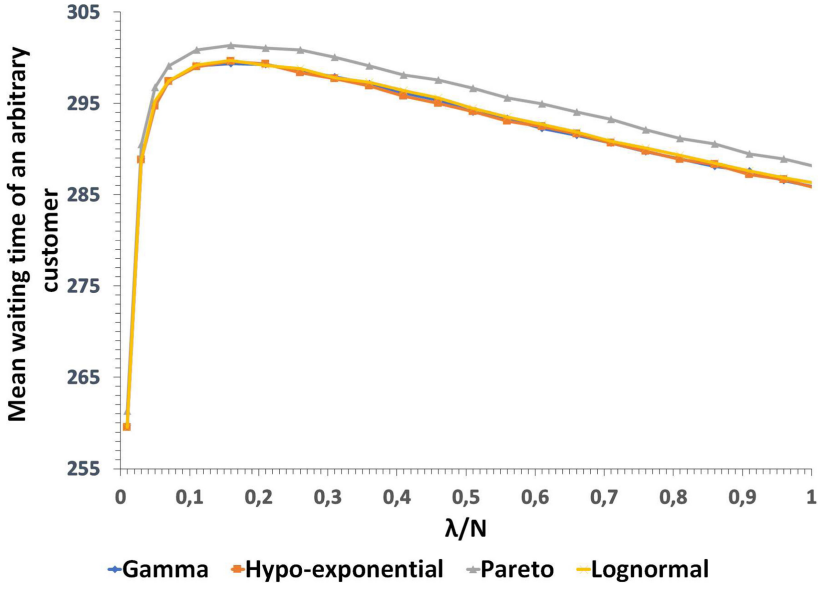

Fig. 7. The mean waiting time of an arbitrary customer

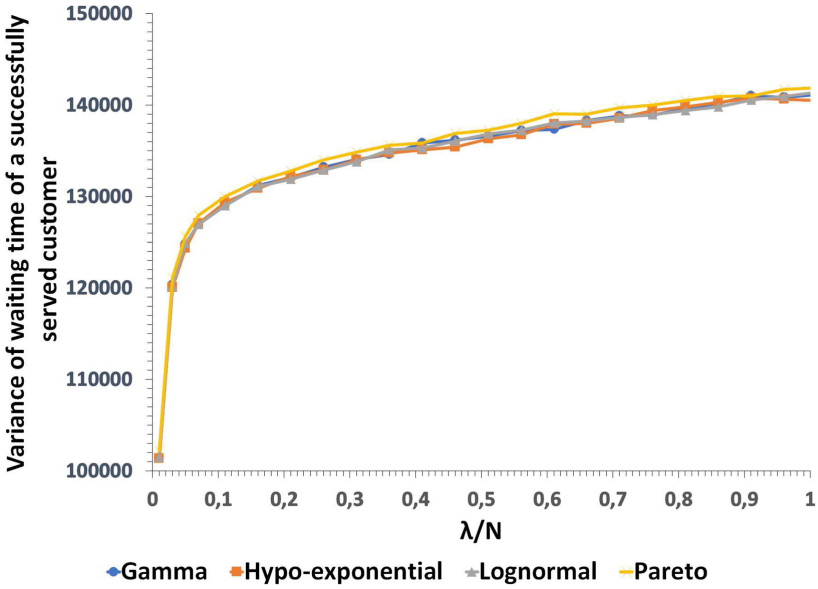

Fig. 8. The variance of waiting time of a successfully served customer

4 Conclusion

We introduced a retrial queueing system of type $M/G/1//N$ with impatient customers in the orbit and with an unreliable server having a two-way communication feature from an infinite source when blocking is implemented. Results are obtained by stochastic simulation and it is shown that the stationary probability distribution of the number of customers in the orbit tends to correspond to the Gaussian distribution despite the used distribution of service time of the primary customers. We investigated different scenarios for example when the squared coefficient of variation is greater than one the obtained values of mean waiting time of an arbitrary, successfully served customer significantly differ from each other even though the parameters are chosen that the mean and variance would be equal in case of every distribution. Results also revealed the effect of blocking which lowers the value of mean waiting time and the number of customers in the system. In our second scenario when the squared coefficient of variation is less than one interestingly the curves almost overlap each other minor disparity turns up examining all the desired performance measures. In the future, the authors intend to continue their research work, analyzing other features of the system like collisions, outgoing calls toward the customers from the orbit, or carrying out sensitivity analysis on other random variables.

References

1. Artalejo, J., Corral, A.G.: Retrial Queueing Systems: A Computational Approach. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78725-9
2. Chen, E.J., Kelton, W.D.: A procedure for generating batch-means confidence intervals for simulation: checking independence and normality. Simulation **83**(10), 683–694 (2007)
3. Dragieva, V., Phung-Duc, T.: Two-way communication M/M/1//N retrial queue. In: Thomas, N., Forshaw, M. (eds.) ASMTA 2017. LNCS, vol. 10378, pp. 81–94. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61428-1_6
4. Dragieva, V.I.: Number of retrials in a finite source retrial queue with unreliable server. Asia-Pac. J. Oper. Res. **31**(2), 23 (2014). https://doi.org/10.1142/S0217595914400053
5. Fiems, D., Phung-Duc, T.: Light-traffic analysis of random access systems without collisions. Ann. Oper. Res. **277**(2), 311–327 (2017). https://doi.org/10.1007/s10479-017-2636-7
6. Fishwick, P.A.: Simpack: getting started with simulation programming in C and C++. In: In 1992 Winter Simulation Conference, pp. 154–162 (1992)
7. Francini, A., Neri, F.: A comparison of methodologies for the stationary analysis of data gathered in the simulation of telecommunication networks. In: Proceedings of MASCOTS 1996 - 4th International Workshop on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, pp. 116–122, February 1996
8. Gharbi, N., Dutheillet, C.: An algorithmic approach for analysis of finite-source retrial systems with unreliable servers. Comput. Math. Appl. **62**(6), 2535–2546 (2011)
9. Gharbi, N., Ioualalen, M.: GSPN analysis of retrial systems with servers breakdowns and repairs. Appl. Math. Comput. **174**(2), 1151–1168 (2006). https://doi.org/10.1016/j.amc.2005.06.005
10. Gharbi, N., Nemmouchi, B., Mokdad, L., Ben-Othman, J.: The impact of breakdowns disciplines and repeated attempts on performances of small cell networks. J. Comput. Sci. **5**(4), 633–644 (2014)
11. Gupta, N.: Article: a view of queue analysis with customer behaviour and priorities. In: IJCA Proceedings on National Workshop-Cum-Conference on Recent Trends in Mathematics and Computing 2011 RTMC(4), May 2012
12. Kuki, A., Sztrik, J., Tóth, Á., Bérczes, T.: A contribution to modeling two-way communication with retrial queueing systems. In: Dudin, A., Nazarov, A., Moiseev, A. (eds.) ITMM/WRQ -2018. CCIS, vol. 912, pp. 236–247. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97595-5_19
13. Kumar, R., Jain, N., Som, B.: Optimization of an $M/M/1/N$ feedback queue with retention of reneged customers. Oper. Res. Decis. **24**, 45–58 (2014). https://doi.org/10.5277/ord140303
14. Law, A.M., Kelton, W.D.: Simulation Modeling and Analysis. McGraw-Hill Education, New York (1991)
15. Panda, G., Goswami, V., Datta Banik, A., Guha, D.: Equilibrium balking strategies in renewal input queue with Bernoulli-schedule controlled vacation and vacation interruption. J. Ind. Manag. Optim. **12**, 851–878 (2015). https://doi.org/10.3934/jimo.2016.12.851
16. Pustova, S.: Investigation of call centers as retrial queuing systems. Cybern. Syst. Anal. **46**(3), 494–499 (2010)

17. Sztrik, J., Tóth, Á., Pintér, Á., Bács, Z.: Simulation of finite-source retrial queues with two-way communications to the orbit. In: Dudin, A., Nazarov, A., Moiseev, A. (eds.) ITMM 2019. CCIS, vol. 1109, pp. 270–284. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33388-1_22
18. Wolf, T.: System and method for improving call center communications. US Patent App. 15/604,068, 30 November 2017