






Analysis of Retrial Queueing System with Two-Way Communication in Different Scenarios Using Simulation

János Sztrik  and Ádám Tóth  

University of Debrecen, University Square 1, Debrecen 4032, Hungary
{sztrik.janos,toth.adam}@inf.unideb.hu

Abstract. The purpose of this study is to investigate a finite-source retrial queueing system with two-way communication. Customers, who arrive from a finite source according to an exponential distribution, are referred to as primary customers. If the service unit is available, these customers will receive service immediately, but if not, they are redirected to the orbit and attempt to reach the server again after a random amount of time. The system is unique in that when the server becomes idle, an outgoing call, also known as a secondary customer, is made to the orbit and source with varying parameters. Both primary and secondary customers receive service following an exponential distribution, but with differing rates. This investigation aims to conduct a sensitivity analysis on the performance measures by using different distributions of the customers' retrial time in two separate cases. The results of the comparison will be displayed graphically.

Keywords: Finite-source queueing system · Retrial queues · Two-way communication · Sensitivity analysis · Simulation

1 Introduction

The topic of two-way communication is widely popular due to its ability to be modeled using retrial queueing systems in a variety of real-life scenarios. A prime example is the operation of call centers, where during idle periods, agents engage in activities such as selling, advertising, and promoting products in addition to handling customer calls. One of the most important measures is utilization, and how to optimize the efficiency of the service units or agents which is always a key issue, see for example [1, 4, 9, 13, 17]. The characteristic of two-way communication relies on performing calls inside and outside of the system when the server becomes idle. There are two types of outgoing calls that are distinguished:

- One where the server calls a customer from the source for service known as a primary outgoing call,
- And another where the server calls a customer from the orbit referred to as a secondary outgoing call.

In our model, outgoing calls can be made to either the source or the orbit. Exploring the available literature reveals many queueing schemes: in some the incoming customer waits until it is served because the queue size is infinite. In others, the arriving customer at the time of arrival can leave the system observing that the service units are fully occupied. However, in real life, there are various situations where customers do not leave the system, stay close to the service units and try to reach a server again after some random time. In this case, this customer will stay in a so-called virtual waiting room called orbit before launching another attempt to reach a server again. Systems containing an orbit can be modeled easily with retrial queues. Queueing systems with retrial queues are useful tools for modeling various problems that arise in telecommunication systems, such as call centers, telephone switching systems, and computer networks like in [2, 8, 11]. In the past, researchers investigated infinite source retrial queueing systems with two-way communication, and here are some examples: [3, 7, 12, 16, 18, 19].

Dragieva and Phung-Duc [6] have investigated the scenario when a secondary outgoing call returns to the source after the service. This paper is the natural continuation of [14] where a more realistic scenario was considered. Instead of sending back the secondary outgoing customers to the source, they will be sent back to the orbit where the call has the opportunity to retry his request for servicing the original incoming call. The motivation for investigating finite source retrial models with two-way communication is based on real-life scenarios in which customers are unable to receive service immediately upon arrival and must go to another location before checking the system again, or the server, once idle, calls for customers.

The uniqueness of this research lies in conducting a sensitivity analysis to assess the impact of various distributions of retrial time on multiple performance measures. The results were generated using a stochastic simulation program based on the SimPack framework ([10]), which is a collection of C/C++ libraries and programs for computer simulation to support various types of simulations, including discrete event simulation, continuous simulation, and multi-model simulation. It provides the flexibility to model any queueing system and perform simulations with custom random number generators to calculate any desired performance measures. The input parameters are presented in a table, and the results of the comparison between different operation modes and distributions are shown through graphical illustrations.

2 System Model

In this section, the considered finite-source retrial queueing model with one server is introduced, which is represented in Fig. 1. Altogether N requests are located in the source, and each of them is capable of generating a primary incoming call toward the server, and the inter-request times are exponentially distributed random variables with parameter λ_1 . In the case of an idle server, the service of an incoming customer begins instantaneously that follows an exponential distribution with parameter μ_1 . After the successful service, the customers go back

to the source. When the incoming customer finds the service unit busy, those customers will not be lost and they are transmitted to the orbit. These will be the secondary incoming jobs from the orbit that may retry to reach the service unit after a random waiting time. The distribution of this period follows different distributions including gamma, hyper-exponential, Pareto, and lognormal, with varying parameters, but all with the same mean value. However, the idle server can also initiate outgoing calls from both the source and the orbit. There are two types of outgoing calls that are distinguished:

- The service unit may call a job from the source to receive service (known as a primary outgoing call) after an exponentially distributed period with rate λ_2 ,
- Or the service unit may initiate a call from the orbit (referred to as a secondary outgoing call) after an exponentially distributed period with rate ν_2 .

The service time of the outgoing customers is exponentially distributed with parameter μ_2 . Two scenarios are distinguished when an outgoing call comes from the orbit:

- Case 1: The call has an unserved incoming request so that call is sent back to the orbit after the outgoing service is finished to have its incoming call be served,
- Case 2: Here, the call has also got an unserved incoming request but after the outgoing service is done the service unit serves the incoming request right away. This will result in a two-phase service, first the outgoing call then the incoming one is executed. The call returns to the source after both service phases are finished.

It is assumed that the arrivals of primary incoming calls, the intervals between retries of secondary incoming calls, the service times of both incoming and outgoing calls, and the time it takes to make outgoing calls are all mutually independent.

3 Simulation Results

3.1 First Scenario

SimPack is used to obtain the results as a basic block of our program and it was extended with the desired features. We used a statistical package that can estimate the desired measures. It utilizes the batch means method which is a quite popular method. In brief, the running period is divided into a number of batches (totaling T). In each batch, $s = R - M/T$ observations are conducted. M represents the discarded warm-up period observations that occur at the beginning of the simulation, and R is the length of the simulation. After the initial phase, the average of the entire run is calculated. To obtain meaningful results, the batches should be of sufficient length and the average of each batch should be independent. More detailed information about the used process you can find in these papers: [5, 15].

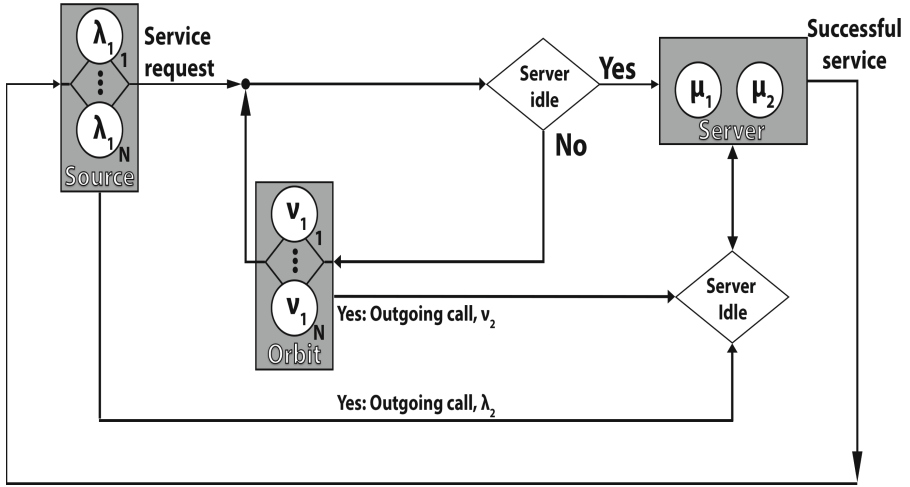


Fig. 1. System model

Throughout the simulations, a confidence level of 99.9% is used, and a relative half-width of the confidence interval of 0.00001 is employed to halt the actual simulation sequence. The size of a batch during the initial transient period cannot be too small, so it is set to 1000. The values of the input parameters used are presented in Table 1.

Table 1. Numerical values of model parameters

N	μ_1	μ_2	λ_2	ν_2
10	1	1	0.2	0.2

The next table (Table 2) contains the parameters of the retrial time of the customers, to achieve a valid comparison parameters are chosen according to having the same mean and variance value. The simulation program was run using various parameter values and the most noteworthy results will be presented in this paper. As shown in the table, the squared coefficient of variation is greater than one in this scenario, allowing for the examination of the impact of specific random variables. We also aim to present results with a different set of parameters when the squared coefficient of variation is less than one.

Table 2. Parameters of the retrial time of the customers

Distribution	Gamma	Hyper-exponential	Pareto	Lognormal
Parameters	$\alpha = 0.02$ $\beta = 0.2$	$p = 0.489$ $\lambda_1 = 9.798$ $\lambda_2 = 10.202$	$\alpha = 2.01$ $k = 0.05$	$m = -4.258$ $\sigma = 1.978$
Mean	0.1			
Variance	0.49			
Squared coefficient of variation	49			

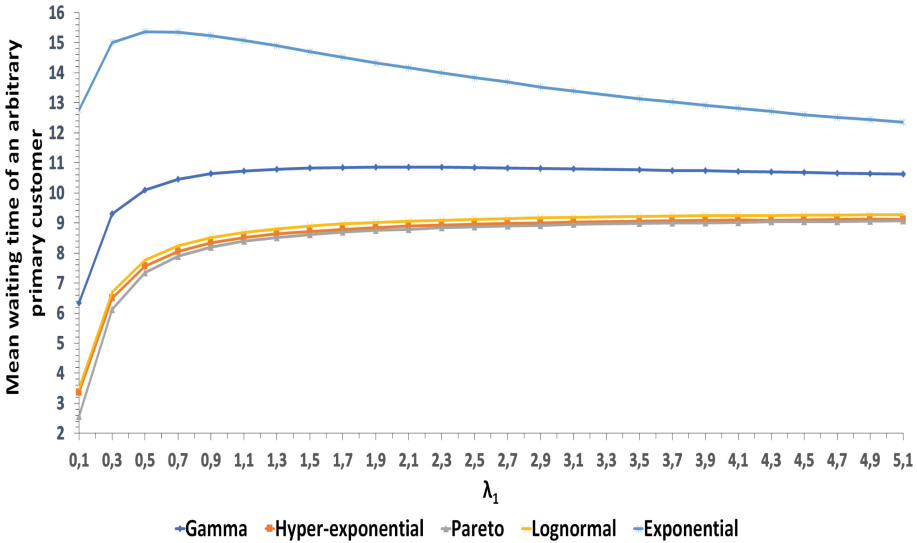


Fig. 2. Mean waiting time of an arbitrary primary customer vs. arrival intensity

The mean waiting time of calls for Case 2 is depicted as a function of the incoming generation rate on Figs. 2 and 3, and comparisons between the different cases are made. Figure 2 demonstrates five cases, the four different distributions, and the exponential case. In the scenario of exponential distribution, the maximum feature is observable, which is a general characteristic of retrial queues when the parameters are set appropriately. Among the distributions applied, the cases of gamma and exponential distribution result in a higher mean waiting time .

On Fig. 3 the comparison of the scenarios is shown using gamma distributed retrial time. The label “No outgoing” indicates that there are only incoming calls in the system, representing a typical finite source retrial system. This figure demonstrates the expected behavior of Cases 1 and 2, showing that Case 2 has a lower mean waiting time, but the lowest values are observed when there are no outgoing calls.

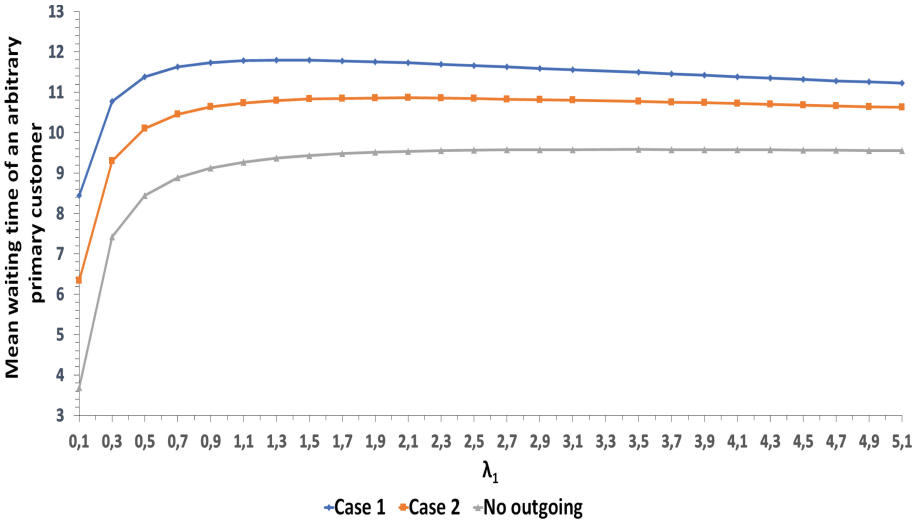


Fig. 3. Comparison of the mean waiting times of the different scenarios

The usage of the service unit is shown in relation to the arrival rate of incoming customers in Fig. 4. Despite having the same mean and variance, there are substantial differences between the different distributions. As the arrival rate increases, the utilization of the service unit also increases. The utilization rate is lower with the gamma and exponential distributions compared to the other distributions, particularly with the Pareto distribution.

Figure 5 demonstrates the comparison of the utilization of the service unit beside various scenarios. Here, it is clear that in cases where there is an outgoing call the server utilization is significantly higher which is true for Case 1 and Case 2 as well. Compared with Fig. 3, it is an optimization problem because there is not a single case where the average waiting time is the lowest and the occupancy rate is the highest. It can be stated that Case 2 can be an optimal solution in terms of mean waiting time and utilization. As λ_1 increases, the value of this performance measure begins to raise and after a certain value, the utilization becomes constant.

3.2 Second Scenario

Having seen the results of the first scenario, we also wondered if different parameter values were used for different distributions. We did not change the mean, but the squared coefficient of variation became less than 1 for the second scenario. Those parameters of each distribution are shown in Table 3, and all the other parameters remain the same (see Table 2). In order to conduct a sensitivity analysis, a hypo-exponential distribution is used instead of a hyper-exponential distribution.

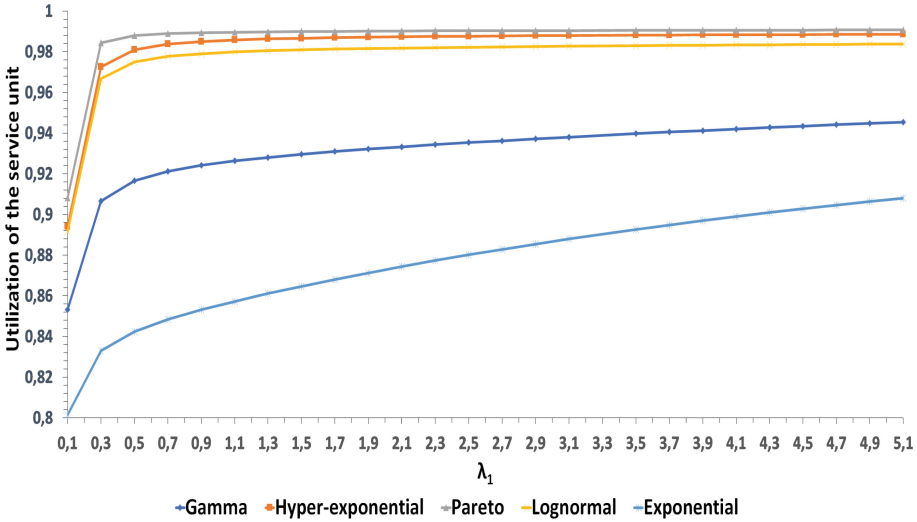


Fig. 4. The utilization of the service unit vs. arrival intensity

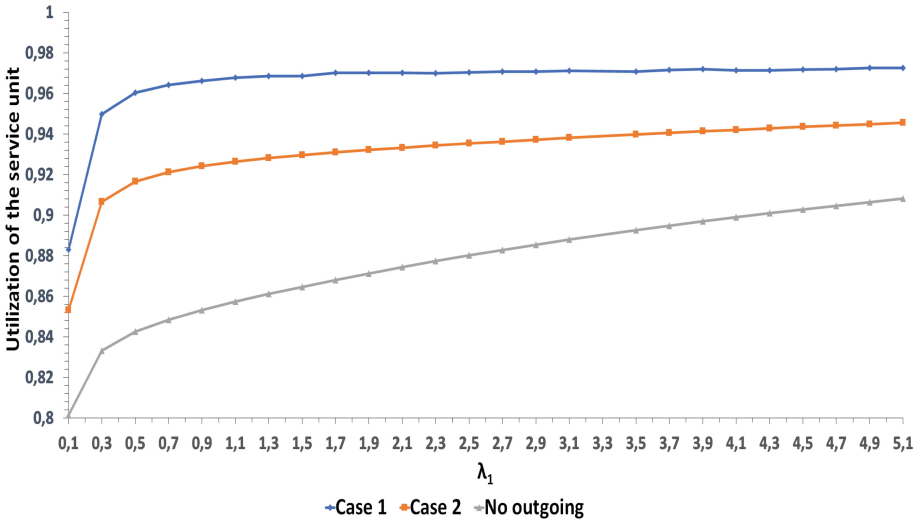


Fig. 5. Comparison of the utilization of the service unit of the different scenarios

In order to observe the similarities and differences between the two scenarios, we go through the same figures but with the modified parameter setting. Initially, we examine the mean waiting time of the calls as a function of the incoming call generation rate for Case 2 in Fig.6. In this scenario, the plotted curves are perfectly coincident, we got back almost exactly the same results to the hundredth of a percent using all 4 distributions, except the exponential case.

Table 3. Parameters of the retrieval time of the customers

Distribution	Gamma	Hypo-exponential	Pareto	Lognormal
Parameters	$\alpha = 1.4706$ $\beta = 14.706$	$\mu_1 = 50$ $\mu_2 = 12.5$	$\alpha = 2.572$ $k = 0.061$	$m = -2.562$ $\sigma = 0.72$
Mean	0.1			
Variance	0.0068			
Squared coefficient of variation	0.68			

The same trend is observed in this scenario, as the intensity of incoming demand increases, the average waiting time starts to decrease after a certain value that is a characteristic of the retrieval queues with appropriate parameters set.

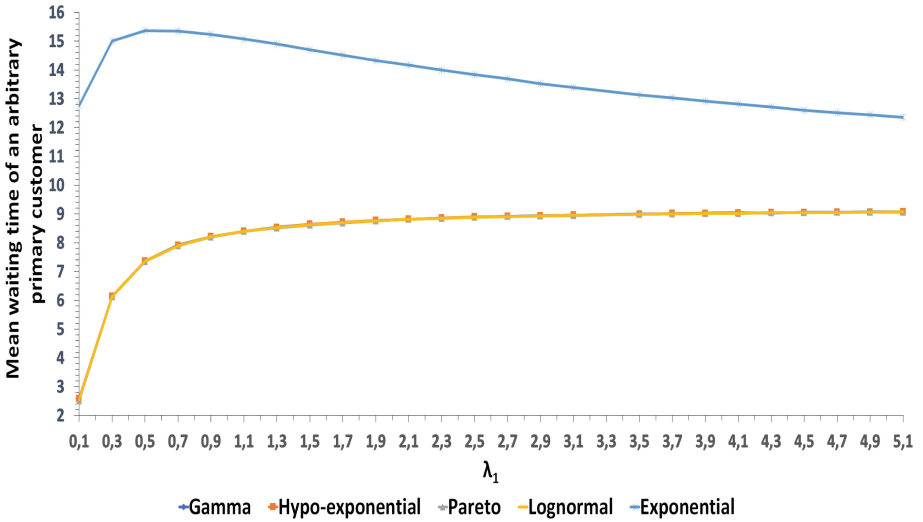


Fig. 6. Mean waiting time of an arbitrary primary customer vs. arrival intensity

Secondly, the comparison of the scenarios is exhibited using gamma distributed retrieval time (Fig. 7). With this parameter setting, it is observed that the average waiting times are actually the same not only between the distributions but also between the different operating modes. In fact, external calls are not relevant here as the curves are essentially the same.

The next figure (Fig. 8) shows the utilization of the service unit besides increasing arrival intensity. From the results of the previous two graphs, it is perhaps not surprising that, in addition to the average latency, the server utilization is the same for all the distributions used except the exponential one where the utilization is much less compared with the others.

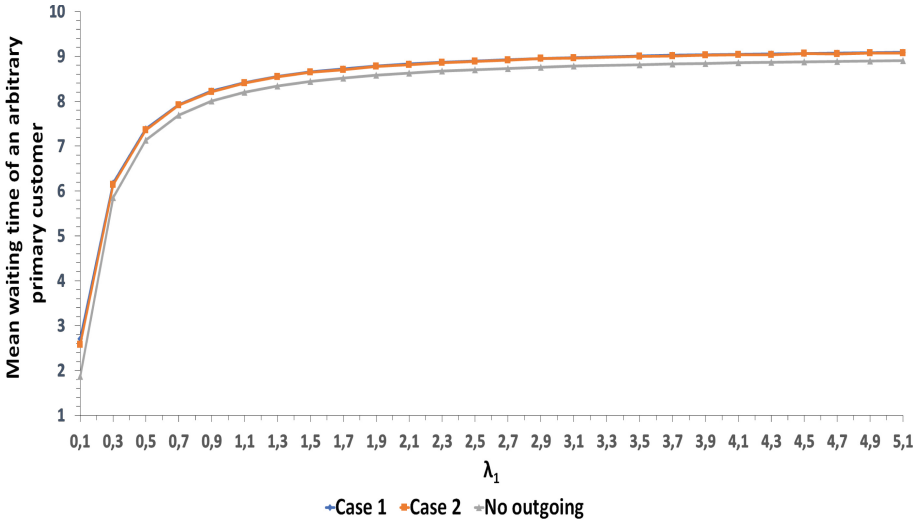


Fig. 7. Comparison of the mean waiting times of the different scenarios

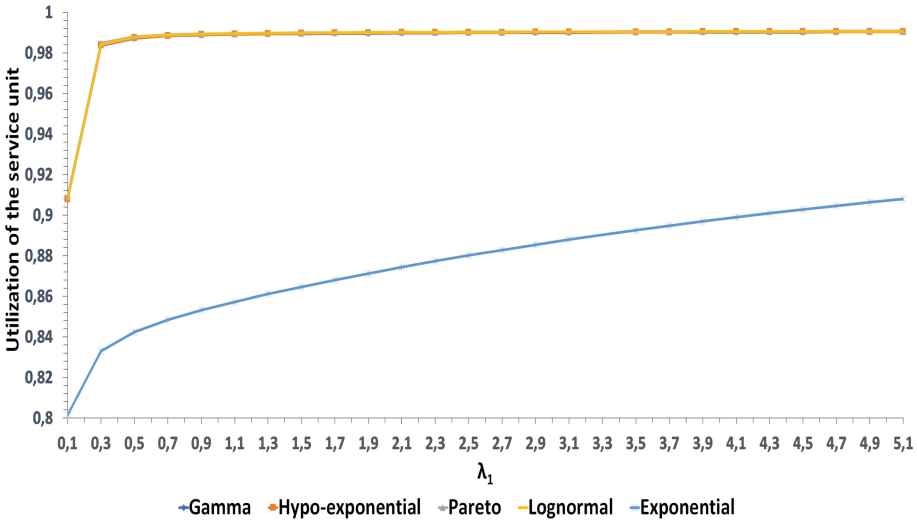


Fig. 8. The utilization of the service unit vs. arrival intensity

To emphasize the effect of this parameter setting, Fig. 9 illustrates the utilization of the service unit using different operation modes. It can be seen from the plotted curves that there is a difference between the cases studied at very low arrival intensities. When there are no external calls, the server utilization is significantly lower, but from $\lambda=0,3$ the results are practically the same. This is not surprising after studying the previous figures.

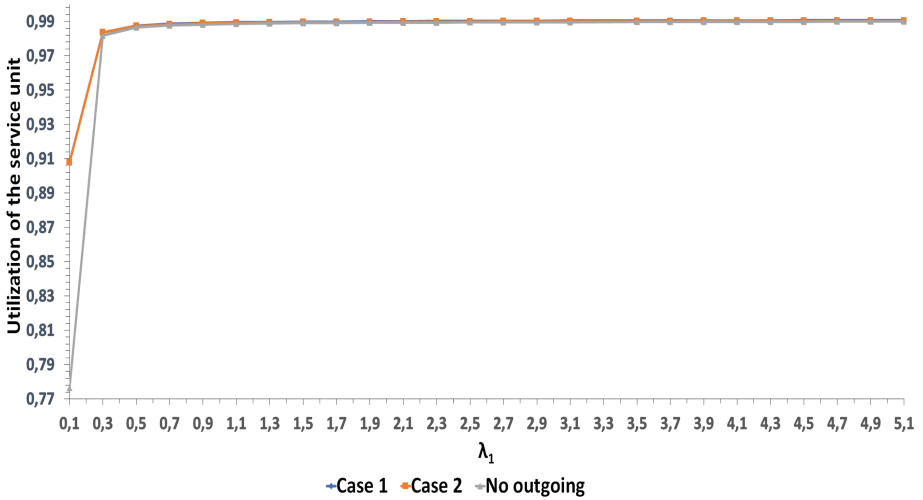


Fig. 9. Comparison of the utilization of the service unit of the different scenarios

4 Conclusion

In this research, we introduce a finite-source retrial queuing system with a two-way communication scheme that uses different distributions for retrial times. We examine multiple scenarios with varying parameters to carry out a sensitivity analysis and focus on the mean waiting time of customers and the utilization of the service unit. The results are obtained through simulations and demonstrated through graphical figures. The figures reveal that slight differences exist among the performance measures when the squared coefficient of variation is greater than one, highlighting the importance of choosing an appropriate distribution. The curves also show the effect of outgoing calls and suggest that in Case 2, the waiting time and utilization are better than in Case 1.

Taking the example of a bank, outgoing calls may be made for signature allocation, both outside and inside the bank as customers wait for transactions. It is more beneficial for the bank to keep a customer waiting inside the bank (Case 2) rather than sending them away or serving their initial request after the signature (Case 1).

Moving forward, we plan to continue our research by exploring other types of finite-source retrial queuing systems with two-way communication or adding a backup service unit.

References

1. Aguir, S., Karaesmen, F., Akşin, O.Z., Chauvet, F.: The impact of retrials on call center performance. *OR Spectrum* **26**(3), 353–376 (2004)
2. Aksin, Z., Armony, M., Mehrotra, V.: The modern call center: A multi-disciplinary perspective on operations management research. *Prod. Oper. Manag.* **16**(6), 665–688 (2007)

3. Artalejo, J.R.: New results in retrial queueing systems with breakdown of the servers. *Statistica Neerlandica* **48**(1), 23–36 (1994). <https://doi.org/10.1111/j.1467-9574.1994.tb01429.x>, <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1994.tb01429.x>
4. Artalejo, J., Corral, A.G.: *Retrial Queueing Systems: A Computational Approach*. Springer (2008). <https://doi.org/10.1007/978-3-540-78725-9>
5. 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)
6. 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
7. Dragieva, V.I.: Steady state analysis of the M/G/1//N queue with orbit of blocked customers. *Ann. Oper. Res.* **247**(1), 121–140 (2016)
8. Falin, G., Artalejo, J.: A finite source retrial queue. *Eur. J. Oper. Res.* **108**, 409–424 (1998)
9. Fiems, D., Phung-Duc, T.: Light-traffic analysis of random access systems without collisions. *Ann. Oper. Res.* **277**(2), 311–327 (2019)
10. Fishwick, P.A.: Simpack: Getting started with simulation programming in c and c++. In: *1992 Winter Simulation Conference*, pp. 154–162 (1992)
11. Gómez-Corral, A., Phung-Duc, T.: Retrial queues and related models. *Ann. Oper. Res.* **247**(1), 1–2 (2016). <https://doi.org/10.1007/s10479-016-2305-2>
12. Jinting, W.: Reliability analysis M/G/1 queues with general retrial times and server breakdowns. *Progress Natural Sci.* **16**(5), 464–473 (2006). <https://doi.org/10.1080/10020070612330021>, <https://www.tandfonline.com/doi/abs/10.1080/10020070612330021>
13. Kim, J., Kim, B.: A survey of retrial queueing systems. *Ann. Oper. Res.* **247**(1), 3–36 (2016)
14. 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
15. Law, A.M., Kelton, W.D.: *Simulation Modeling and Analysis*. McGraw-Hill Education (1991)
16. Nazarov, A., Phung-Duc, T., Paul, S.: Heavy outgoing call asymptotics for MMP P/M/1/1 retrial queue with two-way communication. In: Dudin, A., Nazarov, A., Kirpichnikov, A. (eds.) *Information Technologies and Mathematical Modelling. Queueing Theory and Applications*, vol. 800, pp. 28–41. Springer International Publishing, Cham (2017)
17. Pustova, S.: Investigation of call centers as retrial queueing systems. *Cybern. Syst. Anal.* **46**(3), 494–499 (2010)
18. Sakurai, H., Phung-Duc, T.: Two-way communication retrial queues with multiple types of outgoing calls. *TOP* **23**(2), 466–492 (2015)
19. Sakurai, H., Phung-Duc, T.: Scaling limits for single server retrial queues with two-way communication. *Ann. Oper. Res.* **247**(1), 229–256 (2016)