Annales Mathematicae et Informaticae
58 (2023) pp. 160–169
DOI: 10.33039/ami.2023.08.015
URL: https://ami.uni-eszterhazy.hu

Analysis of retrial queueing systems with two-way communication and impatient customers using simulation

Ádám Tóth, János Sztrik

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

Abstract. The aim of this research is to examine a finite-source retrial queueing system with two-way communication. The primary customers, who arrive from a finite-source following an exponential distribution, either receive service immediately if the service unit is available, or are redirected to the orbit and try again to reach the server after a random period. The system is unique in that when the server becomes idle, an outgoing call (secondary customer) is performed from the orbit or the source with varying parameters. Both primary and secondary customers have been serviced according to an exponential distribution but with different rates. Customers exhibit an impatience characteristic, which may lead to their departure before receiving service if they spend a certain amount of time waiting for the service unit. This investigation conduct a sensitivity analysis on the system's performance measures by utilizing different distributions of the customers' retrial time in two separate cases. The findings of the analysis have been presented graphically for comparison purposes.

Keywords: Finite-source queuing system, retrial queues, two-way communication, sensitivity analysis, simulation

AMS Subject Classification: 60K25

1. Introduction

Two-way communication is a popular research topic because it can be effectively modeled using retrial queueing systems in many real-life situations. Call centers are a prime example of this, where agents engage in various activities such as selling, advertising, and promoting products when not handling customer calls. Utilization

Submitted: July 25, 2023
Accepted: August 9, 2023
Published online: September 13, 2023

is one of the most crucial metrics in call center operations, and optimizing the efficiency of service units or agents is always a critical concern, see for example [1, 3, 7, 9, 13]. The distinctive feature of two-way communication is based on the occurrence of calls both inside and outside the system when the server is idle. Two types of outgoing calls can be identified:

- One type of outgoing call is when the server contacts a customer from the source for service, which is referred to as a primary outgoing call,
- Another type of outgoing call occurs when the server contacts a customer from the orbit, which is known as a secondary outgoing call.

In our model, we consider outgoing calls that can be made to either the source or the orbit. Existing literature on queueing systems reveals different schemes, where some models assume infinite queue size, causing incoming customers to wait until they receive service, while in others, customers leave the system immediately upon arrival if the service unit is fully occupied. However, in reality, there are scenarios where customers do not leave the system but wait in a virtual waiting room, known as an orbit, and attempt to connect with a server after some random time. Retrial queues are a suitable modeling tool for systems that involve an orbit. Queuing systems that utilize retrial queues are commonly used to model various problems arising in telecommunications systems, such as call centers, telephone switching systems, and computer networks like in [2, 6, 8]. Previously, scholars have explored examples of retrial queueing systems with two-way communication and infinite sources, some of which are listed below: [12, 14, 15].

Dragieva and Phung-Duc [5] examined a scenario in which secondary outgoing calls return to the source after service, while this study is a natural extension of [10], which considered a more realistic scenario. In this case, rather than returning secondary outgoing customers to the source, they are sent back to the orbit where they can retry their request for servicing the original incoming call. Investigating finite-source retrial models with two-way communication is motivated by real-life situations in which customers cannot receive immediate service upon arrival and must go to another location before attempting to check the system again or wait for the server to call for them when idle.

Some queuing models make the assumption that a consumer must wait in line indefinitely before being serviced. When a customer enters and discovers the service area is occupied, some additional models—known as loss models—have the customer leave and lose it forever. However, there are countless situations in real life where customers choose to give up the attempt to be served after an arbitrary amount of time rather than waiting. In this scenario, the client waits in a virtual waiting area called an orbit before making another attempt to contact the server. Retrial queues can be used to represent models that have an orbit.

The originality of this work lies in the sensitivity analysis that was conducted to examine how different retrial time distributions affect the key performance metrics. Our stochastic simulation program, which is based on SimPack produces the results.

To support discrete event simulation, continuous simulation, and combined (multi-model) simulation, this is a collection of C/C++ libraries and executable programs. Any sort of queueing system and simulation model can be freely modeled, and any performance metric can be calculated using any random number generator for the specified random variable. The comparison of the operating modes and various distributions will be shown through graphical representations.

The system model

This section introduces the finite-source retrial queueing paradigm with a single server under consideration (see Figure 1). The source contains a total of N re-

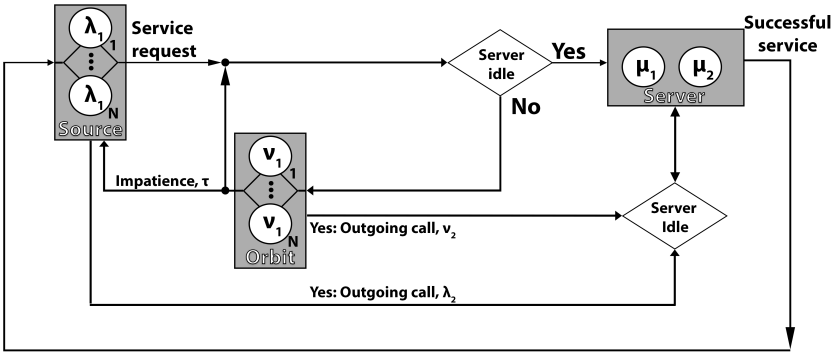

Figure 1. System model.

quests, each of which can produce a primary incoming call to the server. The inter-request times are determined by exponentially distributed random variables with the parameter λ_1. When the server is idle, an incoming customer's service starts immediately and follows an exponential distribution with parameter μ_1. After receiving satisfactory service, clients return to the original provider. Customers who arrive and find the service unit busy will not be lost; instead, they are transported to orbit. These are the secondary arriving jobs from the orbit that might make another attempt to contact the service unit following an arbitrary waiting period. Gamma, hyper-exponential, Pareto, and lognormal distributions are all used to describe this period's distribution, albeit they all have the same mean value. But the idle server may also request calls from the orbit and the source. We distinguish between two categories of outgoing calls:

- After an exponentially varying amount of time, the service unit may request a primary outgoing call from the source to be served with parameter λ_2,
- After an exponentially distributed period, the service unit may make a call (secondary outgoing call) from orbit with parameter ν_2.

The outgoing customers' service time is distributed exponentially with the parameter μ_2. When an incoming call is received from the orbit, there are two distinct scenarios:

- Operation mode number 1: After the outgoing service is complete, the call is returned back to the orbit to have its incoming call served because it has an unmet incoming request,
- Operation mode number 2: Here, the call also has an incoming request that hasn't been fulfilled, but as soon as the outgoing service is complete, the service unit fulfills the incoming request. A two-phase service will result from this, with the outgoing call being processed first and then the incoming one. When both service phases have been completed, the call goes back to the source.

Every primary customer has an impatience trait, and in our investigated model a primary customer eventually departs the system after waiting in the orbit for some time without obtaining the proper service, which is also an exponentially distributed random variable with rate τ. The arrivals of primary incoming calls, retrial intervals for secondary incoming calls, service times for incoming and outgoing calls, and the amount of time needed to make outgoing calls are all considered to be independent of one another.

Utilizing this model, our goal is to perform a sensitivity analysis on the main performance measures using several distributions of retrial times. A number of system properties are compared between various operating modes as well. We developed a simulation program to get the results, which will be shown in a series of graphs.

2. Simulation results

2.1. First scenario

We utilized SimPack as the foundation of our program and incorporated the necessary functionalities. To estimate the desired performance measures, we employed a statistical package that utilizes the popular batch means method. The simulation period is divided into a set of batches, with $s = R - M/T$ observations conducted in each batch, where M represents the discarded warm-up period observations and R is the simulation length. Once the initial phase is complete, the average of the entire simulation is computed. It is crucial for the batches to be of sufficient length and for each batch average to be independent for meaningful results. For additional information about the process used, please refer to the following papers: [4, 11].

The input parameters used in the simulations are presented in Table 1. A relative half-width of 0.00001 and a confidence level of 99.9% were employed to halt the simulation sequence. To ensure the accuracy of the results, the size of a batch during the initial transient period was set to 1000 and cannot be too small.

Table 1. Numerical values of model parameters.

N	μ_1	μ_2	λ_2	ν_2	τ
10	1	1	0.2	0.2	0.01;0.05;0.1

Table 2 lists the retrial time parameters of the customers, which were selected to have the same mean and variance value for a valid comparison. Various parameter values were tested in the simulation program, and the most significant results will be discussed in this paper. As demonstrated in the table, the squared coefficient of variation exceeds one in this case, enabling an investigation into the influence of specific random variables. Additionally, we will present outcomes with a distinct set of parameters when the squared coefficient of variation is less than one.

Table 2. Parameters of retrial time.

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			

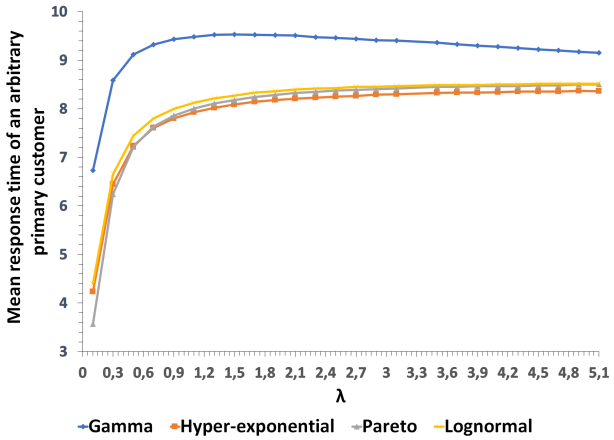

Figure 2. Mean response time of an arbitrary primary customer vs. arrival intensity.

The mean response time of an arbitrary primary customer is depicted in Figure 2 in the function of the arrival intensity. By comparing the impact of various distributions with the same first two moments, a noticeable discrepancy is observed.

Customers spend relatively more time in the orbit when gamma distribution is employed, but more or less the same amount of time when other distribution is utilized. Additionally, the system's intriguing maximum property is evident even as the arrival intensity increases, which is a characteristic of a finite-source retrial queueing system.

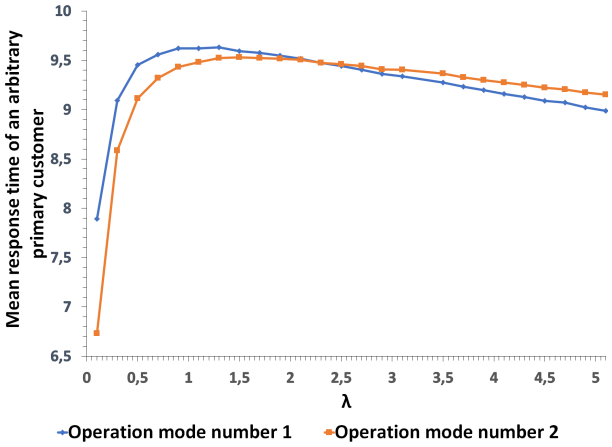

Figure 3. Comparison of the mean response times using different operation modes.

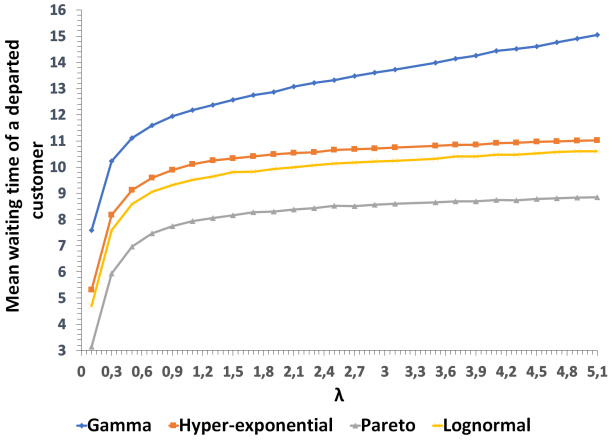

Figure 4. Mean waiting time of a departed customer using the different distributions of retrial times.

Figure 3 illustrates the impact of two operation modes on the mean response time of a customer under gamma distribution of retrial times, as the arrival intensity increases. It is interesting to observe that in case of lower λ values Operation mode

number 2 performs better resulting in lower mean response time and this trend changes after λ is greater than 2. However, the differences among operation modes are not that significant in this parameter setting.

Figure 4 demonstrates the comparison of mean waiting time of a departed customer beside various scenarios. Under this performance measure we refer to those mean waiting times of customers who exit from the system due to impatience. Despite having the same mean and variance, there are significant differences between the applied distributions, with increasing gaps observed as the arrival intensity increases. The mean waiting time of an impatient customer also increases with the arrival intensity, and the Pareto distribution consistently results in lower mean waiting times compared to the other distributions, particularly in comparison to the gamma distribution.

2.2. Second scenario

After observing the results of the first scenario, we became curious about the effects of using different parameter values for each distribution while keeping the mean constant. For the second scenario, we reduced the squared coefficient of variation to less than 1 for each distribution, as shown in Table 3, while all other parameters remained the same as in Table 2. To conduct a sensitivity analysis, we replaced the hyper-exponential distribution with a hypo-exponential distribution.

Table 3. Parameters of retrial time.

Distribution	**Gamma**	**Hypo-exponential**	**Pareto**	**Lognormal**
Parameters	$\alpha = 1.6$ $\beta = 16$	$\mu_1 = 13.333$ $\mu_2 = 40$	$\alpha = 2.612$ $k = 0.062$	$m = -2.545$ $\sigma = 0.697$
Mean	0.1			
Variance	0.00625			
Squared coefficient of variation	0.625			

Figure 5 illustrates how the mean response time of an arbitrary primary customer changes with increasing arrival intensity in a scenario where the mean value remains constant but the variance value is significantly reduced. The difference in average response time among the distributions is not very significant, it can be stated that they overlap each other totally. This indicates that variance has a considerable impact on performance measures, as larger variance values can result in greater disparities in performance measures.

Figure 6 compares the mean waiting time of a departed customer across varying arrival intensities. As expected from the previous figure, the differences in the obtained values are relatively small, which is true for every utilized distribution. Therefore, it can be concluded that in this parameter setting, the performance measures do not exhibit significant differences among the distributions. Naturally, the mean waiting time of a departed customer increases with the increment of the arrival intensity of the primary customer.

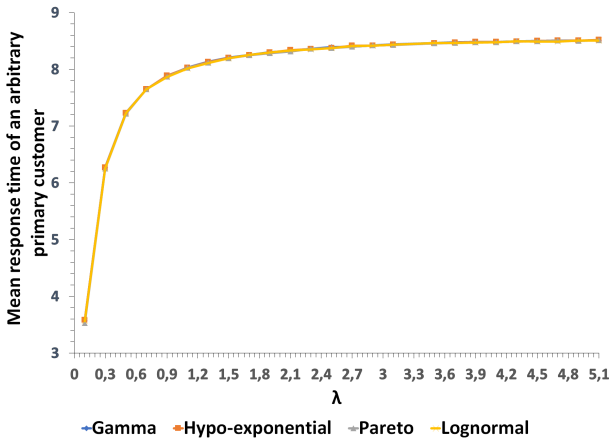

Figure 5. Mean response time of an arbitrary primary customer vs. arrival intensity.

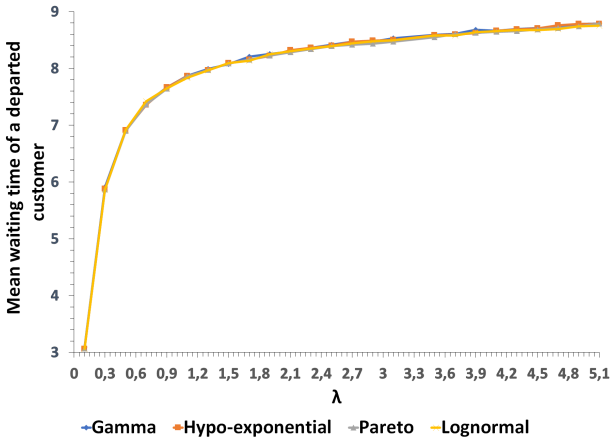

Figure 6. Mean waiting time of a departed customer using the different distributions of retrial times.

Finally, Figure 7 depicts the impact of increasing arrival intensity on the mean response times of the different operation modes. The values obtained in this scenario are very close to each other compared to the previous scenario. In this scenario, looking at the graphs, it can be said that there are no discrepancies in terms of comparing different distributions of retrial times or different operation modes.

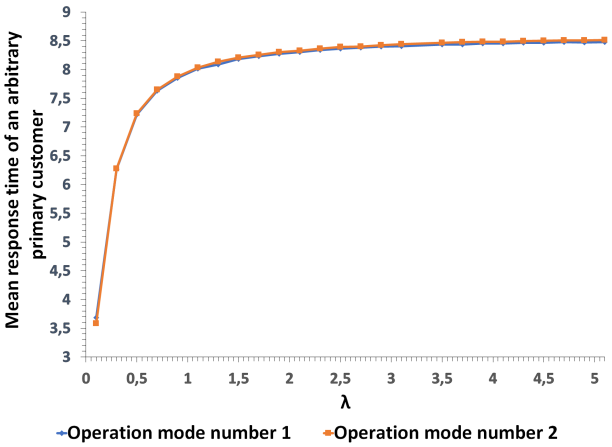

Figure 7. Comparison of the mean response times using different operation modes.

3. Conclusion

We present a study on a two-way communication finite-source retrial queueing system that utilizes different retrial time distributions. Our investigation includes various scenarios with different parameters, focusing on the mean response time of an arbitrary primary customer and the mean waiting time of a departed customer. Through simulations and graphical figures, we demonstrate that choosing an appropriate distribution is critical when the squared coefficient of variation is greater than one. The figures also display the impact of outgoing calls and suggest that Operation mode number 2 (keeping customers waiting inside the bank) may result in smaller waiting and response times than Operation mode number 1.

For instance, in a banking setting, outgoing calls may be used to allocate signatures both inside and outside the bank while customers wait for their transactions. It is more advantageous for the bank to keep the customer waiting inside (Operation mode number 2) rather than turning them away or serving their initial request after obtaining the signature (Operation mode number 1).

Future research may explore other types of two-way communication finite-source retrial queuing systems or consider adding a backup service unit.

References

[1] S. Aguir, F. Karaesmen, O. Z. Akşin, F. Chauvet: *The impact of retrials on call center performance*, OR Spectrum 26.3 (2004), pp. 353–376.

[2] Z. Aksin, M. Armony, V. Mehrotra: *The modern call center: A multi-disciplinary perspective on operations management research*, Production and operations management 16.6 (2007), pp. 665–688.

[3] J. Artalejo, A. G. Corral: *Retrial Queueing Systems: A Computational Approach*, Springer, 2008.

[4] E. J. Chen, W. D. Kelton: *A Procedure for Generating Batch-Means Confidence Intervals for Simulation: Checking Independence and Normality*, SIMULATION 83.10 (2007), pp. 683–694.

[5] V. Dragieva, T. Phung-Duc: *Two-Way Communication $M/M/1//N$ Retrial Queue*, in: International Conference on Analytical and Stochastic Modeling Techniques and Applications, Springer, 2017, pp. 81–94.

[6] G. Falin, J. Artalejo: *A finite source retrial queue*, European Journal of Operational Research 108 (1998), pp. 409–424.

[7] D. Fiems, T. Phung-Duc: *Light-traffic analysis of random access systems without collisions*, Annals of Operations Research 277.2 (2019), pp. 311–327, doi: 10.1007/s10479-017-2636-7.

[8] A. Gómez-Corral, T. Phung-Duc: *Retrial queues and related models*, Annals of Operations Research 247.1 (2016), pp. 1–2, issn: 1572-9338, doi: 10.1007/s10479-016-2305-2.

[9] J. Kim, B. Kim: *A survey of retrial queueing systems*, Annals of Operations Research 247.1 (2016), pp. 3–36, issn: 1572-9338, doi: 10.1007/s10479-015-2038-7.

[10] A. Kuki, J. Sztrik, Á. Tóth, T. Bérczes: *A Contribution to Modeling Two-Way Communication with Retrial Queueing Systems*, in: Information Technologies and Mathematical Modelling. Queueing Theory and Applications, Springer, 2018, pp. 236–247, doi: 10.1007/978-3-319-97595-5_19.

[11] A. M. Law, W. D. Kelton: *Simulation Modeling and Analysis*, McGraw-Hill Education, 1991, isbn: 0-07-100803-9.

[12] A. Nazarov, T. Phung-Duc, S. Paul: *Heavy outgoing call asymptotics for MMP P/M/1/1 retrial queue with two-way communication*, in: Information Technologies and Mathematical Modelling. Queueing Theory and Applications, ed. by A. Dudin, A. Nazarov, A. Kirpichnikov, vol. 800, Cham: Springer International Publishing, 2017, pp. 28–41, doi: 10.1007/978-3-319-68069-9_3.

[13] S. Pustova: *Investigation of call centers as retrial queuing systems*, Cybernetics and Systems Analysis 46.3 (2010), pp. 494–499.

[14] H. Sakurai, T. Phung-Duc: *Scaling limits for single server retrial queues with two-way communication*, Ann. Oper. Res. 247.1 (2016), pp. 229–256.

[15] H. Sakurai, T. Phung-Duc: *Two-way communication retrial queues with multiple types of outgoing calls*, Top 23.2 (2015), pp. 466–492.