The Simulation of Finite-Source Retrial Queues with Two-Way Communication to the Orbit, Incorporating a Backup Server

Ádám Tóth(✉) and János Sztrik

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

Abstract. This paper investigates a two-way communication retrial queuing system with a server that may experience random breakdowns. The system is a finite-source M/M/1//N type, and the idle server can make calls to the customers in the orbit, also known as secondary customers. The service time of the primary and secondary customers follow independent exponential distributions with rates of μ_1 and μ_2, respectively. The novelty of this study is to analyze the impact of various distributions of failure time on the key performance measures using a backup server, such as the mean response time of an arbitrary customer. One could think of a backup server as a primary server that operates at a reduced rate during periods of repair. To ensure a valid comparison, a fitting process is conducted so that the mean and variance of every distribution are equal. The self-developed simulation program provides graphical illustrations of the results.

Keywords: Simulation · Queueing system · Finite-source model · Sensitivity analysis · Backup server · Unreliable operation · Outgoing calls

1 Introduction

Nowadays, due to the growth of traffic and the increasing number of users, analyzing communication systems or designing optimal patterns for these schemes is a challenging task. Information exchange is essential in every aspect of life, and it is crucial to develop mathematical and simulation models of telecommunication systems or modify the existing ones to keep pace with these changes. Retrial queues are effective and appropriate tools for modeling real-life problems that arise in telecommunication systems, networks, mobile networks, call centers, and similar systems. Numerous papers and books have been dedicated to studying a variety of retrial queuing systems with repeated calls like in [4,5].

We are investigating a retrial queuing system with two-way communication capabilities, which has become a popular research topic due to its resemblance to certain real-life systems. This is particularly relevant in call centers, where

V. M. Vishnevskiy et al. (Eds.): DCCN 2023, LNCS 14123, pp. 44–55, 2024.
https://doi.org/10.1007/978-3-031-50482-2_4

service units may engage in additional activities such as sales, promotion, and product advertising while attending to incoming calls. In our study, the primary server calls in customers from the orbit, known as secondary customers, when it becomes idle after a random period of time. The utilization of the service unit is monitored and has been extensively studied in previous works, for example in [3,10].

In some scenarios, it is assumed by researchers that service units are available continuously, but failures or sudden events may happen during operation resulting in the rejection of incoming customers. Devices used in different industries are subject to breakdowns, and considering their reliable operation is quite an optimistic and unrealistic approach. Similarly, in wireless communication, various elements can affect the transmission rate, and interruptions may occur during packet transmission. The unreliable nature of retrial queuing systems greatly affects the system's operation and performance measures. At the same time, completely stopping production is not feasible as it can lead to delays in fulfilling the orders. Hence, during such failures, machines or operators with lower processing rates can continue to work to ensure a smoother operation. Additionally, the authors examined the possibility of having a backup server available to provide service at a reduced rate in cases where the main server is unavailable. Many recent papers have extensively studied retrial queuing systems with unreliable servers, [7,9] are just a few examples.

In service sectors, it is not uncommon for service providers to experience breakdowns due to various reasons, including the inability to access their database to address customer requests. When such breakdowns occur, service providers often resort to alternative measures such as accessing backup systems or gathering additional information from the customers to provide the required solutions. Here are some papers which thoroughly investigate the behaviour of systems that tries to enhance the service by adding a backup server like [1,8,11,12] or [15].

The main objective of this study is to investigate the impact of the unreliable operation of a system by comparing various failure time distributions on performance measures such as the mean response time of a customer or the service unit utilization. This paper is a continuation of the previous work by the authors [13], where the system had an unreliable server, but now, if the server is unavailable a backup server takes its place to serve incoming requests. To obtain the desired performance measures, a simulation model was developed using SimPack [6], a set of C/C++ libraries and executable programs for computer simulation. Simulation is an excellent alternative to deriving exact formulas, particularly when it is problematic or almost impossible. The user can apply as many distributions as needed to approximate performance measures. In this paper, we present a sensitivity analysis of various failure time distributions on the main performance measures. We illustrate the results through graphical representations of interesting phenomena related to sensitivity problems.

2 System Model

The system under consideration (in Fig. 1) is a retrial queuing system with an unreliable server and a finite-source. The source contains N customers, each generating primary customer requests with a rate of λ, such that inter-arrival times are exponentially distributed with a parameter of λ. Note that our model does not include waiting queues, so incoming customers occupy the server only when it is available and not busy. The service time of primary customers follows an exponential distribution with a parameter of μ_1. Following a successful service, the customer returns to the source. However, if an arriving customer (either from the source or orbit) encounters the server in a busy or failed state, the request is forwarded to the orbit. In the orbit, the customer may make an attempt to get its service requirement after an exponentially distributed random time with a parameter of σ. The system is assumed to have an unreliable server that can break down according to different distributions such as gamma, hypo-exponential, hyper-exponential, Pareto, and lognormal, each with different parameters but the same mean value. The repair process begins immediately after the server fails, and the repair time is exponentially distributed with parameter γ_2. If the server is busy and fails, the customer is immediately transferred to the orbit. All customers in the source can generate requests even if the service unit is unavailable, but these requests are directed to the backup server, which serves at a reduced rate (this is also an exponentially distributed random variable with parameter μ_3) when the main server is unavailable. The backup server is assumed to be reliable and works only if the main server is down. In the case of a busy backup server, the incoming requests are placed in the orbit. However, when the server is idle, it can initiate an outgoing call to the customers in the orbit after a random time, which is exponentially distributed with rate τ. The service time of these secondary customers follows an exponential distribution with parameters μ_2. The assumption made during model creation is that all random variables are completely independent of one another.

3 Simulation Results

We utilized a statistical module class providing a statistical analysis tool that enables us to quantitatively estimate the mean and variance values of observed variables using the batch mean method. The method aggregates n successive observations of a steady-state simulation to generate a sequence of independent samples. The batch mean method is a common technique used to establish confidence intervals for the steady-state mean of a process. To ensure the sample averages are approximately independent, large batches are required. More information on the batch mean method can be found in [2]. We conducted simulations with a 99.9% confidence level, and the simulation run was halted when the relative half-width of the confidence interval reached 0.00001.

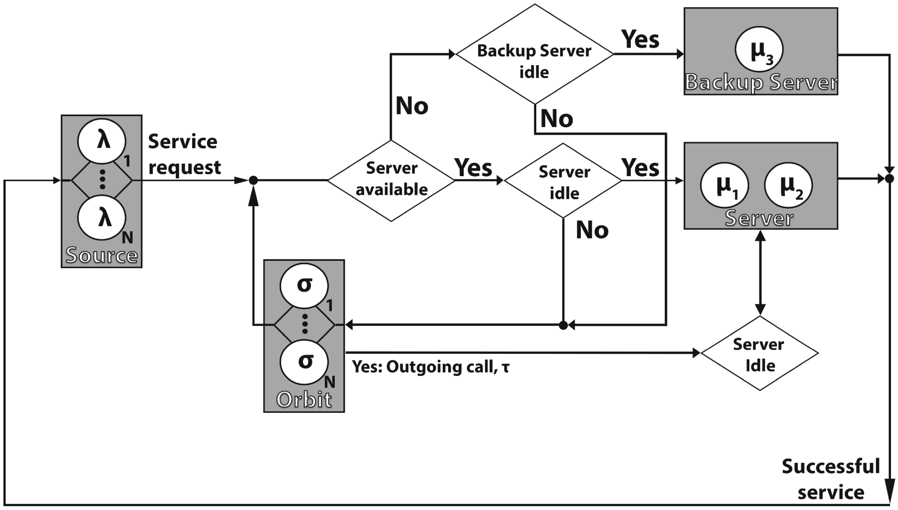

Fig. 1. The system model

3.1 First Scenario

In this section, we aimed to set the parameters of failure time for each distribution in such a way that the mean value and variance would be equal. The fitting process used for this purpose can be found in the following paper [14]. Four different distributions were considered in order to investigate their impact on performance measures. The hyper-exponential distribution was chosen to ensure that the squared coefficient of variation is greater than one. Table 2 presents the input parameters of the various distributions, while Table 1 shows the values of other applied parameters.

Table 1. Used numerical values of model parameters

N	λ	γ_2	σ	μ_1	μ_2	ν	μ_3
100	0.01	1	0.01	1	1.2	0.02	0.1; 0.6

The steady-state distribution for different failure time distributions is presented in Fig. 2. On the X-axes i represents the number of customers located in the system, and on the Y-axes $P(i)$ denotes the probability that exactly i customer is found in the system. Upon closer examination of the curves, it can be observed that all of them resemble the normal distribution. Although the Pareto distribution appears to have more customers in the system, there are no significant differences among the various distributions tested. Including a backup server results in a lower mean number of customers in the system in comparison with the paper of [13].

Table 2. Parameters of failure time

Distribution	Gamma	Hyper-exponential	Pareto	Lognormal
Parameters	$\alpha = 0.6$ $\beta = 0.5$	$p = 0.25$ $\lambda_1 = 0.41667$ $\lambda_2 = 1.25$	$\alpha = 2.2649$ $k = 0.67018$	$m = -0.3081$ $\sigma = 0.99037$
Mean	1.2			
Variance	2.4			
Squared coefficient of variation	1.6666666667			

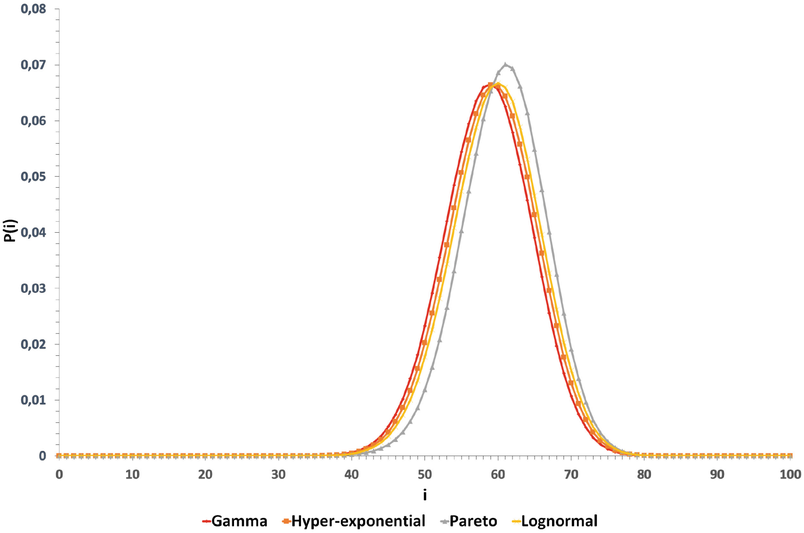

Fig. 2. Comparison of steady-state distributions

Figure 3 illustrates the relationship between the mean response time of customers and the arrival intensity. Consistent with the observations from Fig. 2, the highest mean response time is observed with the Pareto distribution. However, the differences among the other distributions are more noticeable. The gamma distribution yields the lowest mean response time. Interestingly, as the arrival intensity increases, the mean response time initially increases, but then starts to decrease after a certain point. This is a unique feature of retrial queuing systems with a finite source, and is a general characteristic when suitable parameter settings are used.

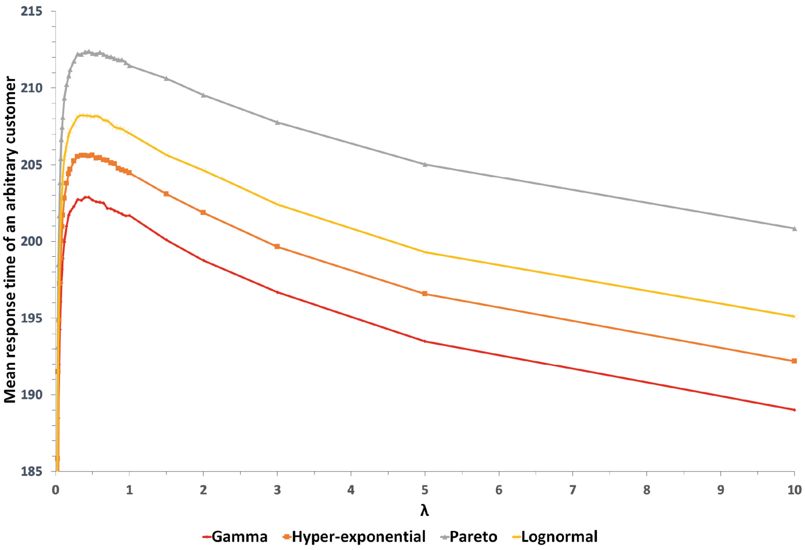

Fig. 3. Mean response time vs. arrival intensity

Figure 4 highlights the effect of using a backup server besides increasing arrival intensity. Under the "No backup server" it is meant when the normal server is either reliable or not but a backup server is not used at all. Actually, the expected behaviour occurs, the "No backup server, reliable normal server" case would be the ideal situation but in reality sudden acts or breakdowns can happen at any time. Comparing the obtained results when the service intensity of the backup is the lowest then the time spent in the system of the customers is the highest. From this figure we experience the advantage of using a backup server hence the customers spend less time in the system while they receive their service demand properly.

The final figure in this section depicts the utilization of the normal and backup service unit beside the arrival intensity using gamma distributed failure time. The red and orange curves represent the total time spent by the clients at the backup server. Upon careful examination of the figure, naturally, the utilization of the backup service unit decreases when the intensity of the service of the backup server increases. For the other distributions, the same tendency is observed, in this way, those ones are not depicted in this paper. As the arrival intensity increases, the utilization of the service units also increases. However, after reaching a certain arrival value (in this case it is around 5), utilization becomes basically stagnant.

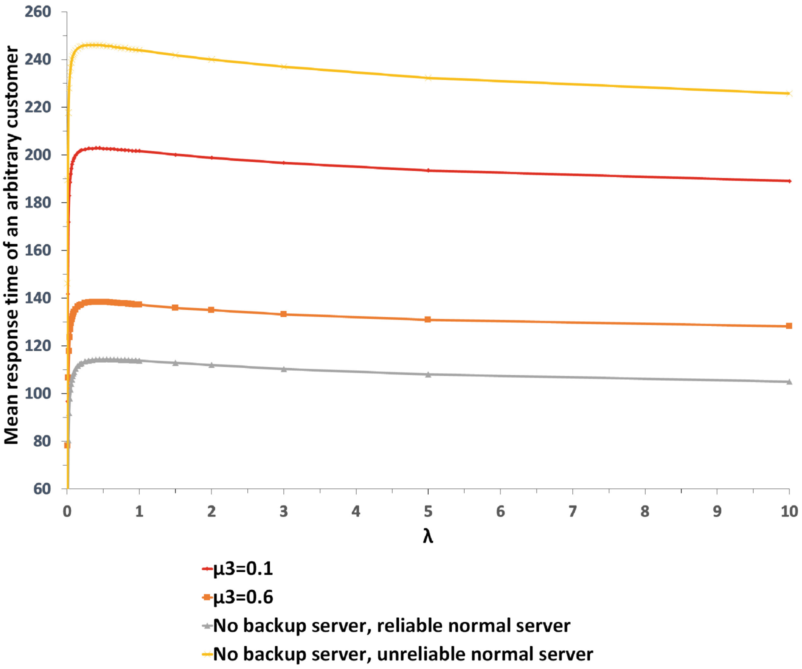

Fig. 4. Mean response time vs. arrival intensity

3.2 Second Scenario

We were curious about how the performance measurements are altered with the modification of the failure time parameters after observing the results of the previous section. The parameters were now selected to ensure that the squared coefficient of variation was below one. Because the squared coefficient of variation for a hypo-exponential distribution is always smaller than one, we replace the hyper-exponential distribution with hypo-exponential distribution. By utilizing the new failure time parameters, we will review the same figures as in the previous section to check the effect of newly chosen parameters, which is shown in Table 3. The other parameters remain unchanged (see Table 1) (Fig. 5).

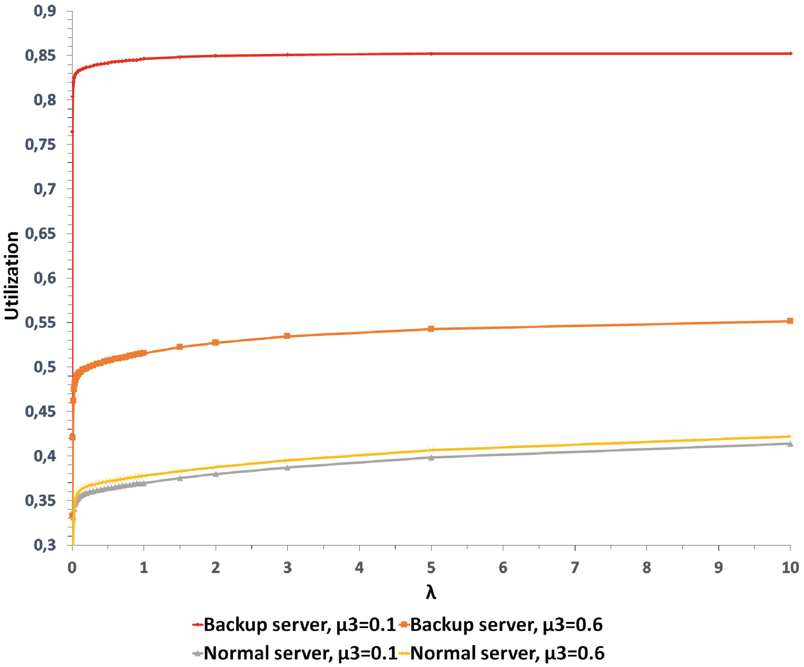

Fig. 5. Comparison of utilization

Table 3. Parameters of failure time

Distribution	Gamma	Hypo-exponential	Pareto	Lognormal
Parameters	$\alpha = 1.3846$ $\beta = 1.1538$	$\mu_1 = 1$ $\mu_2 = 5$	$\alpha = 2.5442$ $k = 0.7283$	$m = -0.0894$ $\sigma = 0.7373$
Mean	1.2			
Variance	1.04			
Squared coefficient of variation	0.722222			

Figure 6 displays the steady-state distributions with a squared coefficient variation of less than one. The curves overlap closely, indicating that regardless of the chosen distribution of failure time, the average number of customers in the system remains the same. In comparison to Fig. 2, the average values are a little bit greater and the curve of Pareto is closer to the other curves.

Figure 7 illustrates the development of the mean response time of an arbitrary customer as the arrival intensity increases. Upon closer inspection, it is evident that the curves are much closer to each other compared to Fig. 3, although there are minor differences among the chosen distributions. Similarly to Fig. 3, the highest values are observed in the case of Pareto distribution. When the squared coefficient of variation is less than one for each distribution, the mean waiting times are higher compared to the previous section (Fig. 7).

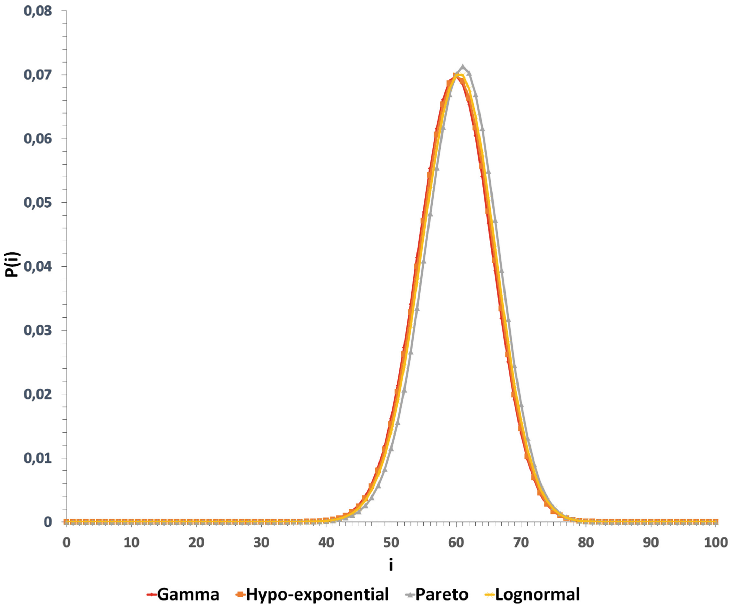

Fig. 6. Comparison of steady-state distributions

The final figure in this section illustrates the utilization of the primary and backup service units as a function of arrival intensity, using gamma distributed failure time. The utilization of the backup service unit is represented by the red and orange curves. Upon careful examination of the figure, it is observed that as the intensity of the service provided by the backup server increases, the utilization of the backup service unit decreases. This tendency is also observed for other distributions, although they are not depicted in this paper. With an increase in arrival intensity, the utilization of the service units also increases. However, after reaching a certain arrival value (in this case, approximately 5), the utilization reaches a plateau. In this figure, the same direction can be observed as in the previous section, the obtained values are basically equal to the previous scenario (Fig. 8).

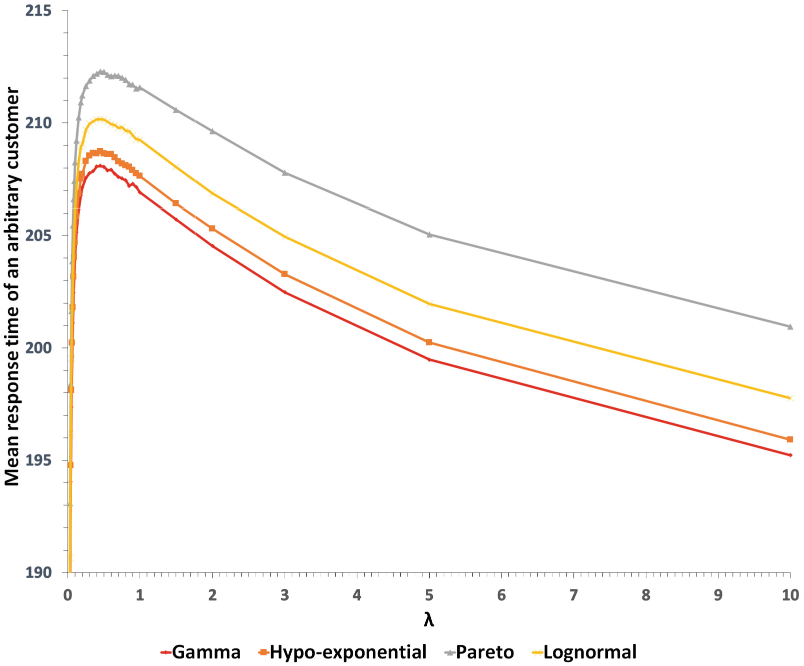

Fig. 7. Mean response time vs. arrival intensity

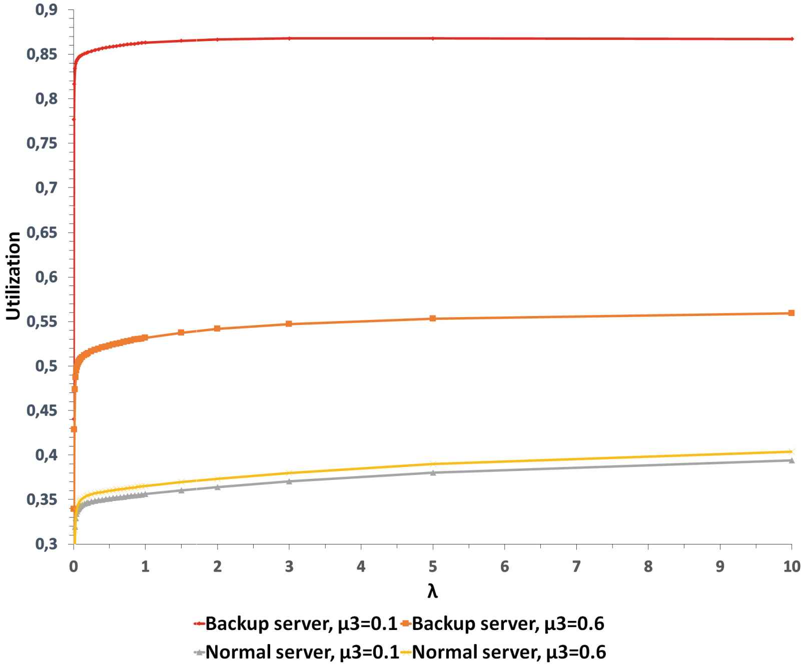

Fig. 8. Comparison of utilization

4 Conclusion

We present a retrial queuing system with finite source and two-way communication, where there is a primary server that is unreliable, and there is also a secondary service unit replacing it in the faulty periods. Moreover, we performed a sensitivity analysis using various random number generators to explore the effect of different distributions of failure time on the performance measures like the mean response time of an arbitrary customer or the utilization of the service units. We observe that when the squared coefficient of variation is greater than one, the mean response time of a customer exhibits some disparity among the values, but the influence is negligible when it is less than one. Using a backup service unit may significantly decrease the time spent in the system of the customers, especially in those scenarios where the primary service unit is under common breakdowns, or the channel is not reliable or the customers are moving rapidly.

In the future, we plan to make further modifications to the system, such as considering additional distributions and incorporating features like vacation.

References

1. Chakravarthy, S.R., Shruti, Kulshrestha, R.: A queueing model with server breakdowns, repairs, vacations, and backup server. Oper. Res. Perspect. **7**, 100131 (2020). https://doi.org/10.1016/j.orp.2019.100131. https://www.sciencedirect.com/science/article/pii/S2214716019302076
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**, 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: 1992 Winter Simulation Conference, pp. 154–162 (1992)
7. 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)
8. Klimenok, V., Dudin, A., Semenova, O.: Unreliable retrial queueing system with a backup server. In: Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V. (eds.) DCCN 2021. LNCS, vol. 13144, pp. 308–322. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-92507-9_25
9. Krishnamoorthy, A., Pramod, P.K., Chakravarthy, S.R.: Queues with interruptions: a survey. TOP **22**(1), 290–320 (2014). https://doi.org/10.1007/s11750-012-0256-6

10. 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
11. Liu, Y., Zhong, Q., Chang, L., Xia, Z., He, D., Cheng, C.: A secure data backup scheme using multi-factor authentication. IET Inf. Secur. **11**(5), 250–255 (2017). https://doi.org/10.1049/iet-ifs.2016.0103. https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-ifs.2016.0103
12. Satheesh, R.K., Praba, S.K.: A multi-server with backup system employs decision strategies to enhance its service. Research Square, pp. 1–31 (2023). https://doi.org/10.21203/rs.3.rs-2498761/v1
13. Sztrik, J., Tóth, Á., Pintér, Á., Bács, Z.: The effect of operation time of the server on the performance of finite-source retrial queues with two-way communications to the orbit. J. Math. Sci. **267**, 196–204 (2022). https://doi.org/10.1007/s10958-022-06124-z
14. Toth, A., Sztrik, J., Kuki, A., Berczes, T., Effosinin, D.: Reliability analysis of finite-source retrial queues with outgoing calls using simulation. In: 2019 International Conference on Information and Digital Technologies (IDT), June 2019, pp. 504–511 (2019). https://doi.org/10.1109/DT.2019.8813419
15. Won, Y., Ban, J., Min, J., Hur, J., Oh, S., Lee, J.: Efficient index lookup for de-duplication backup system, pp. 383–384, September 2008. https://doi.org/10.1109/MASCOT.2008.4770594