Simulating Retrial Queues with Finite Source, Two-Way Communication to the Orbit, Backup Server, and Impatient Customers

Ádám Tóth[1](✉), János Sztrik[1], and Avtandil Bardavelidze[2]

[1] University of Debrecen, University Square 1, Debrecen 4032, Hungary
{toth.adam,sztrik.janos}@inf.unideb.hu
[2] Akaki Tsereteli State University, Kutaisi, Georgia
avtandil.bardavelidzec@atsu.edu.ge

Abstract. This paper explores a retrial queuing system with two-way communication and an unreliable server that may encounter random breakdowns. The system is of the finite-source M/M/1//N type, where the idle server can initiate calls to customers in the orbit, termed as secondary customers. Both primary and secondary customer service times are characterized by exponential distributions, with rates denoted as μ_1 and μ_2, respectively. The novelty of this study lies in its investigation of various failure time distributions and their impact on critical performance metrics, such as the mean response time of a random customer, while utilizing a backup server with impatient customers. The backup server can be likened to a primary server operating at a reduced rate during maintenance intervals. To ensure a valid comparison, a fitting process equalizes the mean and variance across all distributions. The outcomes are visually depicted through the utilization of our self-made simulation program.

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

1 Introduction

Nowadays, the analysis of telecommunication systems and the creation of optimal designs for these schemes have become formidable endeavors due to the immense traffic and escalating number of users. Information exchange pervades every facet of contemporary life, underscoring the need to develop mathematical and simulation models for telecommunication systems or adapt existing ones to keep pace with these dynamic changes. Retrial queues stand out as potent and fitting tools for modeling real-world challenges that arise in telecommunications, networks, mobile networks, call centers, and similar systems. A plethora of literature, exemplified by works like [1,5,6,10], delves into the examination of various retrial queuing systems characterized by recurring calls.

A. Dudin et al. (Eds.): ITMM 2023/WRQ 2023, CCIS 2163, pp. 316–328, 2024.
https://doi.org/10.1007/978-3-031-65385-8_23

We are currently exploring a retrial queuing system endowed with two-way communication capabilities, a research area that has gained substantial prominence owing to its striking resemblance to certain real-world systems. This correspondence is particularly pronounced in the context of call centers, where service units often perform multitasking, engaging in activities such as sales, promotions, and product advertising alongside handling incoming calls. In our investigation, the primary server, following a random idle interval, calls customers in from the orbit, called secondary customers. The system's utilization of the service unit is under scrutiny and has undergone extensive examination in prior works, exemplified by studies like [4,9,13].

In various research scenarios, some assume that service units remain continuously available, but real-world events like failures or unexpected incidents can occur during their operation, resulting in the rejection of incoming customers. Devices used across different industries are prone to breakdowns, and relying on their uninterrupted operation is often overly optimistic and unrealistic. Likewise, in wireless communication, multiple factors can affect transmission rates, resulting in disruptions during packet transmission. The inherent lack of reliability in retrial queuing systems has a substantial impact on system operations and performance metrics.

Furthermore, ceasing production entirely is not a feasible choice, as it may result in delays in order fulfillment. Therefore, in the event of such failures, it becomes crucial to keep machines or operators with lower processing rates operational to ensure a continuous workflow. Additionally, the authors investigated the option of implementing a backup server that could provide services at a reduced rate when the primary server is inaccessible. This approach has attracted significant attention in recent research, with studies such as [8,12] being notable examples.

In the service sector, it's not uncommon for service providers to encounter disruptions for various reasons, including difficulties in accessing their databases to address customer requests. When such disruptions transpire, service providers frequently employ alternative measures, such as resorting to backup systems or gathering additional information from customers to meet their needs.

Numerous research papers explore the performance of systems with the objective of improving service by integrating a backup server, as demonstrated in studies such as [2,11,15,16]. These inquiries provide insights into strategies and approaches for sustaining service quality in challenging scenarios.

The primary aim of this investigation is to assess how the system's unreliable operation affects performance measures, such as the mean response time of a customer or service unit utilization, by comparing various failure time distributions while the customers may depart after a random long enough waiting. This study builds upon the authors' earlier research [17], where the system incorporated an unreliable server. In the current configuration, in the event of server unavailability, a backup server takes over the processing of incoming requests.

To acquire the desired performance metrics, we developed a simulation model utilizing SimPack [7], which encompasses a collection of C/C++ libraries and executable programs tailored for computer simulation. Simulation serves as an excellent alternative for approximating performance metrics when deriving precise formulas proves problematic or nearly impossible. This paper introduces a sensitivity analysis of different failure time distributions' impact on key performance measures. We elucidate these findings by means of graphical representations that highlight intriguing facets of sensitivity-related issues.

2 System Model

The system is a retrial queuing system characterized by an unreliable server and a finite source of customers which is shown in Fig. 1. Within the source, there exist N customers, each generating primary requests at an exponential rate denoted by λ. Consequently, the inter-arrival times adhere to an exponential distribution parameterized by λ. Notably, our model does not contain waiting queues; thus, incoming customers can only occupy the server when it is available and idle. The service time for primary customers follows an exponential distribution with a parameter of μ_1. Following the successful completion of a service, the customer returns to the source. However, if an incoming customer (whether from the source or orbit) encounters a server in a busy or failed state, its request is redirected to the orbit. While within the orbit, a customer may attempt to fulfill its service requirement after an exponentially distributed random time with a parameter of σ.

The system assumes the presence of an unreliable server prone to failures, which can occur according to different distributions-such as gamma, hypo-exponential, hyper-exponential, Pareto, and lognormal. Each distribution comes with distinct parameters while sharing the same mean value. The repair process initiates immediately upon the server's failure, with the repair time following an exponential distribution characterized by parameter γ_2. If the server is busy and subsequently fails, the customer is promptly transferred to the orbit. Regardless of the service unit's availability, all customers within the source can generate requests. However, these requests are directed to the backup server, which operates at a reduced rate-an exponentially distributed random variable with parameter μ_3-when the primary server is unavailable. Importantly, the backup server is assumed to be reliable and functions solely during periods of primary server unavailability. In cases where the backup server is busy, incoming requests are placed into the orbit. Yet, during idle periods, the main server can initiate outgoing calls to customers within the orbit after a random time interval, characterized by an exponential distribution with a rate of ν. The service time for these secondary customers follows an exponential distribution with parameters μ_2. Customers in the orbit, after waiting an exponentially distributed time with parameter τ, may choose to leave the system without getting their service.

Throughout the model's creation, the fundamental assumption is maintained that all random variables remain entirely independent of each other.

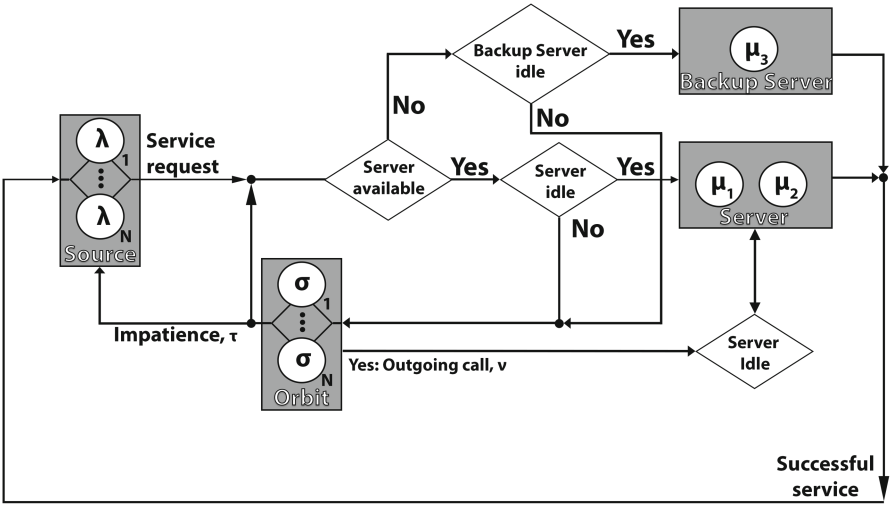

Fig. 1. System model

3 Simulation Results

We employed a statistical module class that incorporates a statistical analysis tool, enabling us to quantitatively estimate both the mean and variance values of observed variables via the batch mean method. This method involves aggregating n consecutive observations from a steady-state simulation to generate a sequence of independent samples. The batch mean method is a widely utilized technique for establishing confidence intervals concerning the steady-state mean of a process. It is important to note that, in order to ensure that the sample averages exhibit approximate independence, the use of sizable batches is imperative. Further details on the batch mean method can be found in [3,14]. In our simulations, we conducted operations with a confidence level of 99.9%, and the simulation run concluded when the relative half-width of the confidence interval reached the threshold of 0.00001.

3.1 First Scenario

In Table 1 the used values of input parameters are presented. The parameters of the failure time are presented in the following table (Table 2). To ensure a valid comparison, parameters are selected to have the same mean and variance values. The simulation program was executed with various parameter values, and this paper will highlight the most significant results. As indicated in the table, the squared coefficient of variation is greater than one in this scenario, enabling the examination of the impact of specific random variables. Additionally, we present results with a different set of parameters when the squared coefficient of variation is less than one.

Table 1. Used numerical values of model parameters

N	λ	γ_2	σ	μ_1	μ_2	ν	μ_3	τ
100	0.01	1	0.01	1	1.2	0.02	0.1	0.001

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			

The steady-state distribution, corresponding to different failure time distributions, is visually represented in Fig. 2. In this graph, the X-axis is labeled as i, which denotes the number of customers present in the system, while the Y-axis is labeled as $P(i)$, indicating the probability of precisely i customers being in the system. A closer examination of the curves reveals that all of them closely resemble the normal distribution. Notably, the Pareto distribution seems to exhibit a lower number of customers in the system. Nevertheless, when comparing the different distributions examined in our study, no significant disparities emerge.

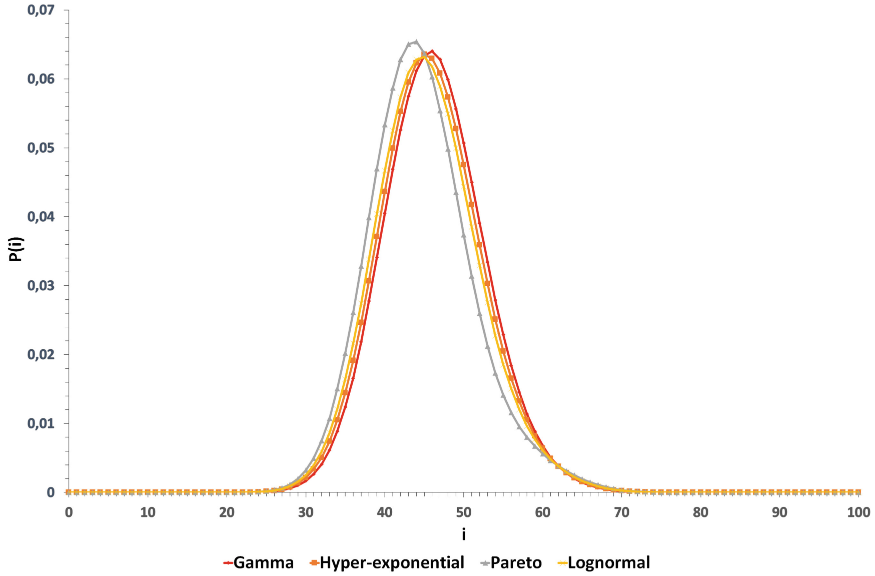

Fig. 2. Comparison of steady-state distributions

Figure 3 provides an illustration of the correlation between the mean response time of customers and the arrival intensity. On the contrary to the patterns observed in Fig. 2, the highest mean response time is associated with the Pareto distribution. However, the distinctions among the other distribution types become more pronounced. Remarkably, the gamma distribution stands out by yielding the lowest mean response time. A noteworthy phenomenon is that, as the arrival intensity increases, the mean response time initially experiences an uptrend, but subsequently, it starts to decrease after reaching a specific threshold. This behavior is a distinctive characteristic of retrial queuing systems with a finite source, and it tends to manifest when appropriate parameter configurations are applied.

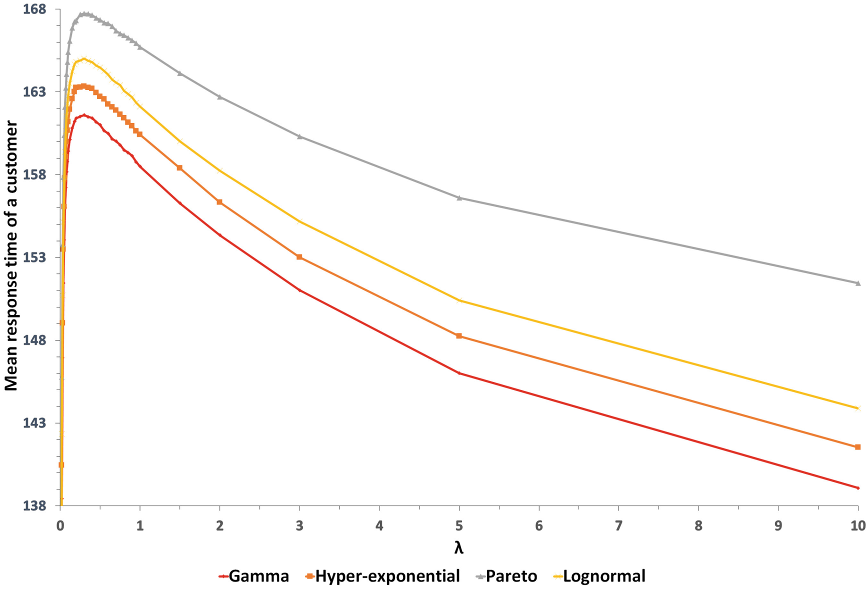

Fig. 3. Mean response time vs. arrival intensity

The variance of the response time is presented in the function of the arrival intensity of the incoming customers in Fig. 4. Looking at the results, it can be said that there are differences in this indicator as well, considering the used failure distributions. Similar trends to the previous chart are observed, with the smallest values occurring in the gamma distribution and the largest values in the Pareto distribution. However, at higher arrival intensity values, we find the smallest numbers in the Pareto distribution, which is an interesting development and requires further experiments and runs to explain this change.

The utilization of the backup service unit is shown in Fig. 5 besides the arrival of the incoming primary customers. In choosing the parameters, we aimed to simulate an environment where as many interruptions or failures as possible

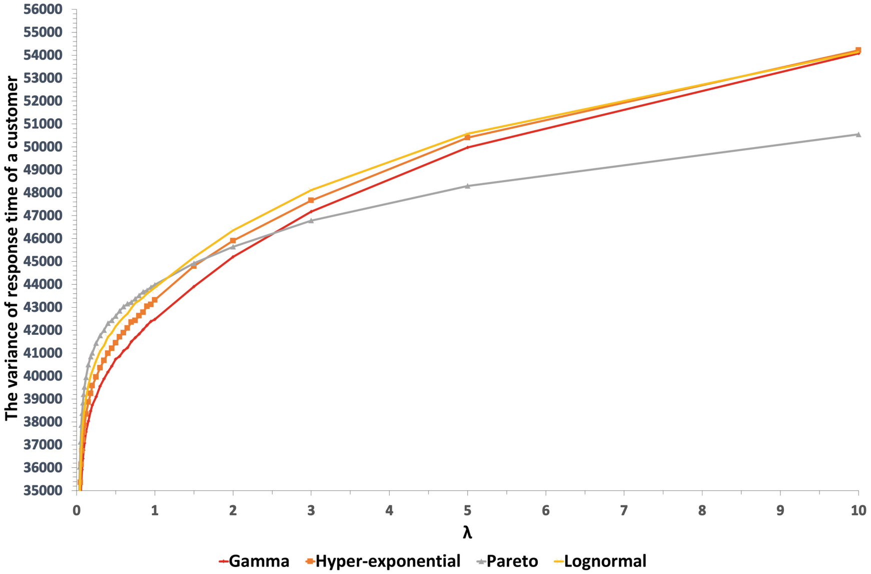

Fig. 4. Variance of the response time vs. arrival intensity

occur. Thus, in the chart, the utilization of the backup server unit becomes significant, and it is evident that this server is busy for most of the time. There are no significant differences among the used failure distributions, but with the Pareto distribution, the utilization is higher compared to the other distributions.

Figure 6 demonstrates the development of the probability of abandonment of a primary customer besides increasing arrival intensity. This metric indicates the likelihood of any given primary customer exiting the system during the orbit, signifying that the request does not meet its specified service requirement (impatient customers). As λ increases, the value of this performance measure also starts to increase, and this holds true for every utilized distribution, but the discrepancy among them is relatively significant. In the case of the gamma distribution, the inclination to exit the system earlier is much lower than in the others, especially when compared to the Pareto distribution.

3.2 Second Scenario

Upon analyzing the outcomes from the previous section, our keen interest was focused on understanding how modifications to the failure time parameters would impact the performance measures. In this scenario, the parameters were selected to ensure that the squared coefficient of variation remains below one. Instead of employing a hyper-exponential distribution, we opt for a hypo-exponential distribution. This choice is motivated by the fact that, in the case of a hypo-exponential distribution, the squared coefficient of variation is always less than one. The identical performance measures will be visually presented as earlier,

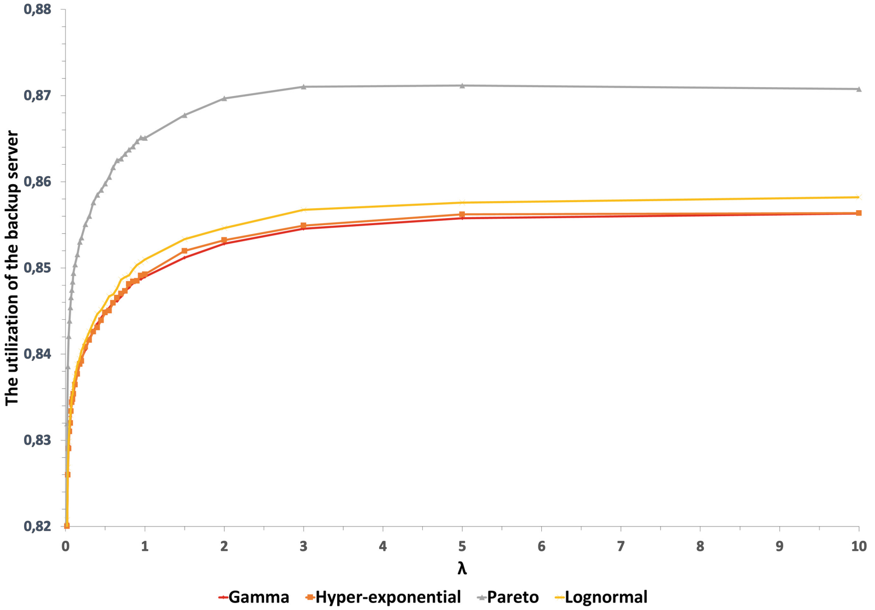

Fig. 5. Utilization of the backup server vs. arrival intensity

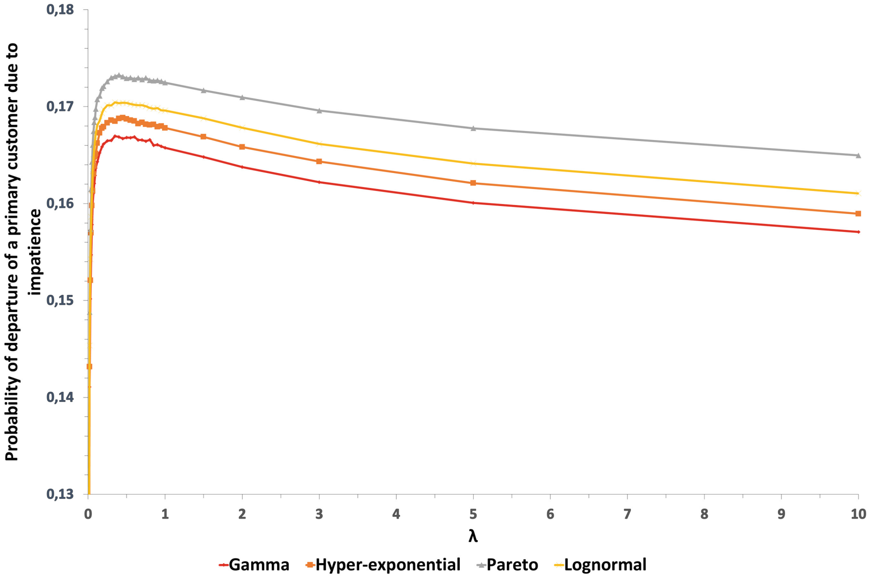

Fig. 6. The probability of the departure of a primary customer vs. arrival intensity

but with the incorporation of the new failure time parameters, as indicated in Table 2. The remaining parameters remain unchanged, as depicted in Table 1 (Table 3).

Table 3. Parameters of failure time

Distribution	Gamma	Hypo-exponential	Pareto	Lognormal
Parameters	$\alpha = 1.3846$ $\beta = 1.1538$	$\mu_1 = 1$ $\mu_2 = 5$	$\alpha = 2.5442$ $k = 0.7283$	$m = -0.08948$ $\sigma = 0.7373$
Mean	1.2			
Variance	1.04			
Squared coefficient of variation	0.72222222			

We will examine the same figures but with the updated parameter setting. Initially, Fig. 7 is related to the distribution of the number of customers in the system. Upon closer analysis of the curves, the obtained values are much more similar. Concerning the shape of the curves, they align with a normal distribution. Nevertheless, there isn't much disparity observed. As evident, the curves are nearly identical. The mean number of customers is slightly higher compared to the previous scenario.

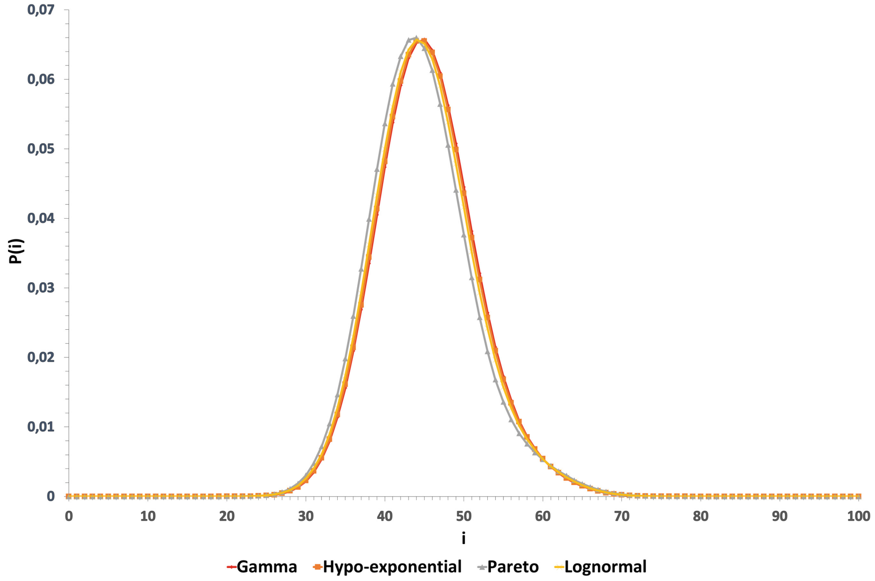

Fig. 7. Comparison of steady-state distributions

Figure 8 illustrates the evolution of the mean response time for a successfully served customer as the arrival intensity increases. In this situation, the mean value remains constant, but the variance is substantially reduced. The difference in the average mean response time among the distributions is not very pronounced, except for Pareto, where the values are notably higher. Therefore, it appears that variance has a noteworthy impact on performance measures, with larger values potentially leading to greater disparities in performance measures.

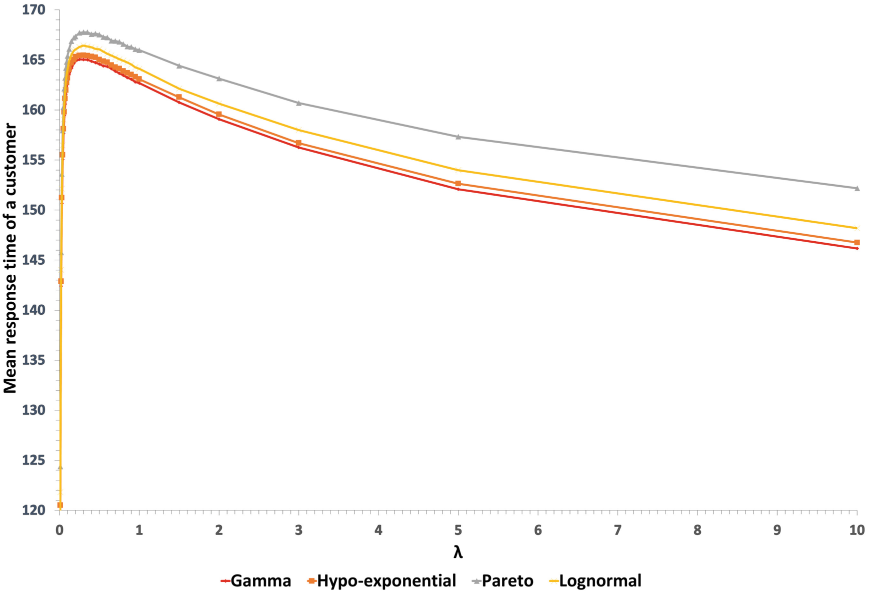

Fig. 8. Mean response time vs. arrival intensity

In the next, in Fig. 9 the variance of the response time is presented with the increment of the arrival intensity of the incoming customers. In comparison to the previous scenario, perhaps the difference is most evident in this figure with the use of newly employed parameters. In practice, the lines overlap completely, with prominent values only occurring in the log-normal distribution for higher arrival intensity values. What may be worth mentioning is that the values obtained in this scenario are smaller compared to the previous one.

Figure 10 depicts the comparison of the utilization of the backup server as a function of the arrival intensity. As expected, considering the results from the previous scenario, the differences in the obtained values are relatively close to each other, even in the case of Pareto distribution. It can be concluded that with this parameter setting, the distinctions among the distributions are not prominent. Regardless of the distribution, the utilization is nearly the same, meaning that the backup server is occupied approximately 87% of the simulation time.

Finally, Fig. 11 illustrates the variations in the abandonment probability with the increase in arrival intensity. The values are more closely aligned compared to

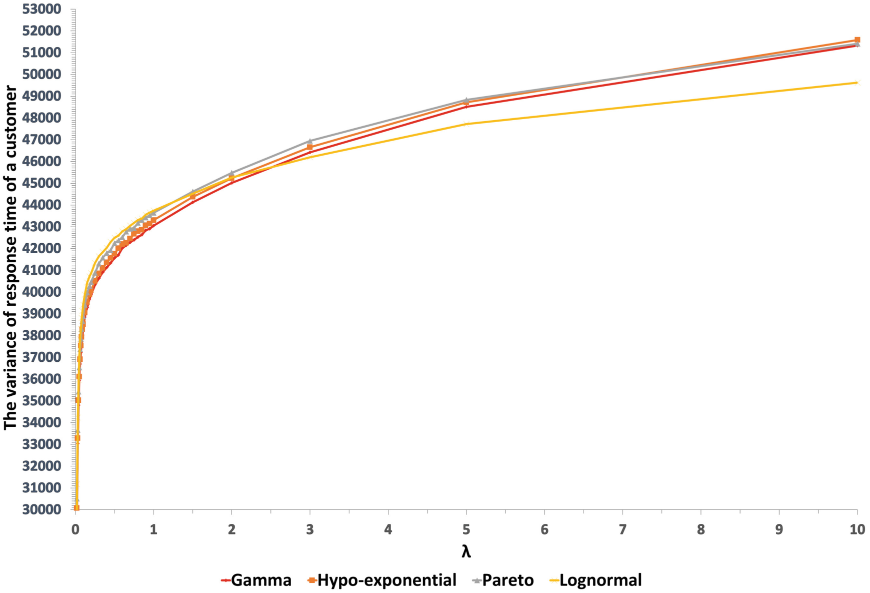

Fig. 9. Variance of the response time vs. arrival intensity

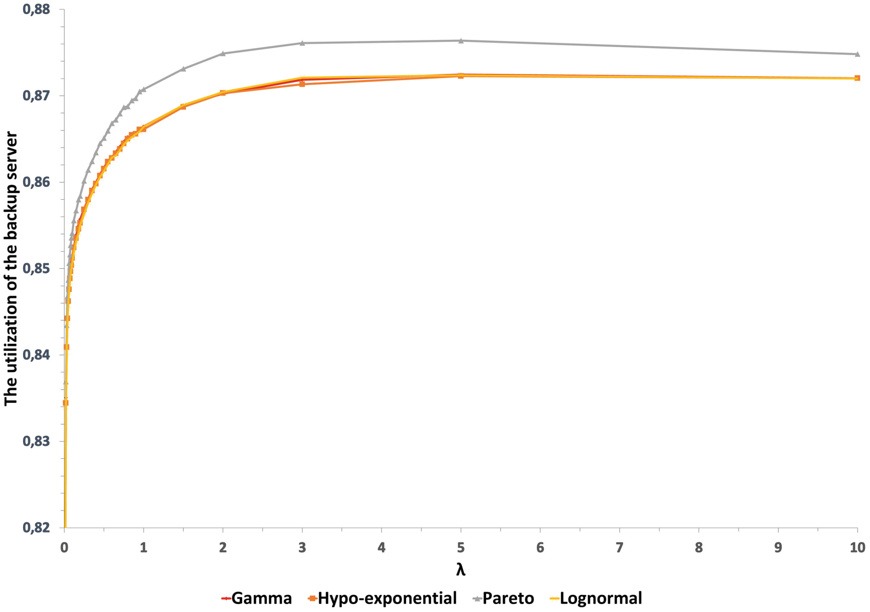

Fig. 10. Utilization of the backup server vs. arrival intensity

the alternative scenario. However, the highest values are observed in the case of the Pareto distribution; otherwise, the difference is minimal. The values obtained for one distribution do not stand out compared to the others; in each case, approximately 17% of incoming requests decide to abandon the system without being served.

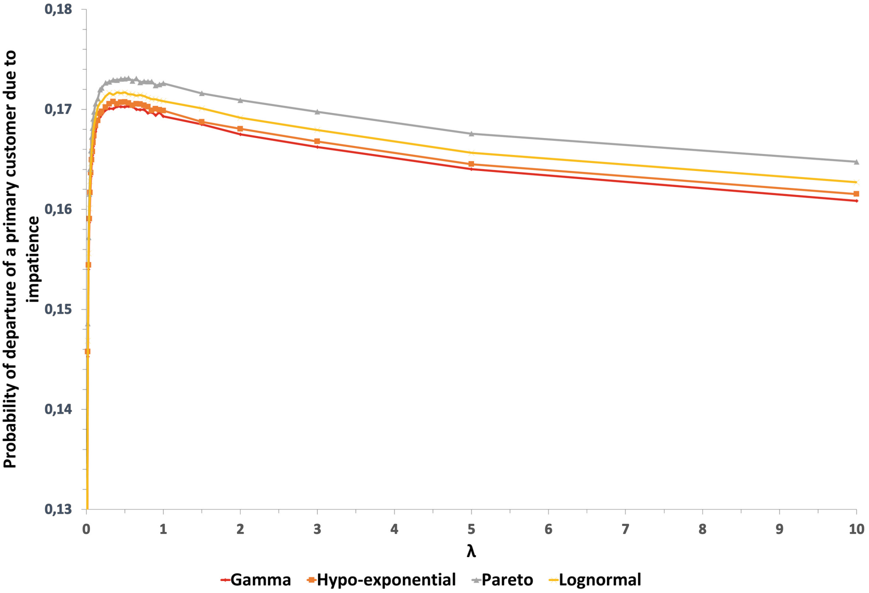

Fig. 11. The probability of the departure of a primary customer vs. arrival intensity

4 Conclusion

We introduced a retrial queuing system characterized by a finite source and two-way communication with impatient customers. Within this system, a primary server exhibits unreliability, and during periods of malfunction, a secondary service unit takes over. Furthermore, we conducted a sensitivity analysis utilizing a range of random number generators to investigate how different distributions of failure time impact performance metrics, such as the mean response time of any given customer. It's worth noting that when the squared coefficient of variation exceeds one, we observed variations in the mean response time among the values. Results also suggest that there is minimal difference among the measured values when the squared coefficient of variation is below one. The authors intend to further their research, delving into the observed phenomenon with greater scrutiny and enhancing their model by incorporating additional elements such as collisions and conducting additional sensitivity analyses on various random variables.

References

1. Artalejo, J., Gomez-Corral, A.: Retrial Queueing Systems: A Computational Approach. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78725-9
2. Chakravarthy, S.R., Shruti, Kulshrestha, R.: A queueing model with server breakdowns, repairs, vacations, and backup server. Oper. Res. Pers. **7**, 100131 (2020). https://doi.org/10.1016/j.orp.2019.100131. https://www.sciencedirect.com/science/article/pii/S2214716019302076
3. 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)
4. 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
5. 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
6. Fiems, D., Phung-Duc, T.: Light-traffic analysis of random access systems without collisions. Ann. Oper. Res. **277**, 1–17 (2017)
7. Fishwick, P.A.: SimPack: getting started with simulation programming in C and C++. In: 1992 Winter Simulation Conference, pp. 154–162 (1992)
8. 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)
9. 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
10. Kim, J., Kim, B.: A survey of retrial queueing systems. Ann. Oper. Res. **247**(1), 3–36 (2016). https://doi.org/10.1007/s10479-015-2038-7
11. Klimenok, V., Dudin, A., Semenova, O.: Unreliable retrial queueing system with a backup server, pp. 308–322 (2021). https://doi.org/10.1007/978-3-030-92507-9_25
12. 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
13. 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
14. Law, A.M., Kelton, W.D.: Simulation Modeling and Analysis. McGraw-Hill Education, New York (1991)
15. 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
16. 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
17. 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