Investigation of M/G/1//N System with Impatient Customers, Unreliable Primary and 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 finite-source retrial queueing system characterized by request collisions, primary server unreliability, and the inclusion of a backup server. In cases of collisions, when a new job arrives while the service facility is occupied, both jobs are sent to a virtual waiting area called the orbit. Customers in the orbit make further attempts to access the server after a random interval. During server breakdowns, the customer at the server is transferred to the orbit. The system consists of a backup facility when the primary server is unreachable to process requests while the main service unit is under repair. The novelty of this study lies in the implementation of the impatience of the customers and conducting a sensitivity analysis using various service time distributions for the primary customers. We analyzed two scenarios, presenting key performance metrics through visual representations to emphasize the observed differences.

Keywords: Simulation · Queueing system · Finite-source model · Backup server · Collisions · Unreliable operation · Impatience

1 Introduction

In the current era of increasing traffic volumes and growing user bases, analyzing communication systems and designing optimal configurations present significant challenges. Information exchange plays a crucial role in all aspects of life, making it essential to develop or adapt mathematical and simulation models for telecommunication systems to meet these evolving demands. Retrial queues are highly effective and well-suited for modeling real-world scenarios commonly found in telecommunications, networking, mobile systems, call centers, and related domains. Numerous scholarly works, such as those referenced in [3] and [4], have extensively investigated various aspects of retrial queuing systems characterized by retrial calls.

In some contexts, researchers often assume that service units are always available; however, operational disruptions or unforeseen events can arise, resulting in the rejection of incoming customers. Devices across various industries are

A. Dudin et al. (Eds.): ITMM 2024, CCIS 2472, pp. 134–145, 2025.
https://doi.org/10.1007/978-3-031-88307-1_11

prone to malfunctions, making the presumption of their infallible operation overly optimistic and impractical. Similarly, in wireless communication environments, diverse factors can affect transmission rates, causing interruptions during packet delivery. The inherent instability of retrial queuing systems plays a crucial role in influencing both system operations and performance outcomes. Additionally, halting production entirely is unfeasible, as it may cause delays in order fulfillment. As a result, in such situations, machines or operators with reduced processing capacities may remain active to ensure smoother system operations. Furthermore, the authors explore the feasibility of incorporating a backup server capable of providing services at a reduced rate when the primary server is unavailable. A multitude of recent studies has thoroughly investigated retrial queuing systems with unreliable servers, as evidenced by sources such as [1,6,8] and [13].

Waiting is a ubiquitous phenomenon in various aspects of life, often leading to dissatisfaction due to the time spent in queues. This dissatisfaction may lead to requests leaving the system prematurely without receiving service, a phenomenon known as impatience. Such behavior is observed in diverse domains including healthcare applications, call centers, and telecommunication networks. Impatience has been widely explored in numerous academic studies, identifying several distinct behaviors: balking occurs when customers decide not to enter the system due to long queues; jockeying involves customers switching between queues to expedite service; and reneging refers to customers abandoning the queue after waiting for a specific, often extended, duration. The impatience mechanism is a crucial aspect of the model, as it influences the overall system performance by potentially reducing the number of customers waiting in the system and affecting the service dynamics. Studies examining these behaviors include [7,9] and [12].

In technological settings like Ethernet networks or limited communication sessions, task collisions are highly probable. Asynchronous attempts by multiple entities within the source can cause signal interference, requiring retransmissions to resolve the conflicts. Therefore, it is essential to account for this phenomenon in research focused on devising effective strategies to minimize conflicts and reduce the resulting message delays. Studies discussing findings on collisions can be found in publications such as [10] and [11].

The objective of this study is to perform a sensitivity analysis using various service time distributions for the primary server, in order to evaluate the main performance metrics in scenarios that incorporate the feature of impatience of the customers. When the primary server fails, customer service is transferred to the backup facility. During this period, new customers are directed to the backup unit or to the orbit if the backup unit is busy. Our investigation focuses on the impact of the impatient feature, with results obtained through simulation using Simpack [5]. The simulation program is built using core code components designed to calculate the desired metrics over a variety of input parameters. Visual representations are included to demonstrate how different parameters and distributions influence key performance metrics.

2 System Model

We analyze a finite-source retrial queuing system of type $M/G/1//N$, illustrated in Fig. 1, which incorporates an unreliable primary service unit, collision events, and a backup service unit. This model features a finite-source, where each of the N individuals generates requests to the system following an exponential distribution with parameter λ. Arrival times adhere to an exponential distribution with a mean of $\lambda * N$. When queues are absent, arriving jobs are processed immediately according to one of several distributions—gamma, hypo—exponential, hyper-exponential, Pareto, or lognormal-each defined by unique parameters but maintaining identical mean and variance values (η).

In cases of server busyness, an arriving customer causes a collision with the customer currently being serviced, resulting in both customers being transferred to the orbit. Jobs in the orbit make additional attempts to access the server after a random time following an exponential distribution with parameter σ. Moreover, the server experiences random breakdowns, with failure times governed by exponential random variables. The failure time is characterized by parameter γ_0 when the server is busy and γ_1 when it is idle.

When the primary service unit fails, repairs commence immediately, with the repair time following an exponential distribution characterized by the parameter γ_2. If the server fails while busy, the customer is promptly moved to the orbit. During the primary server's unavailability, all customers in the source continue to generate requests, which are then directed to the backup server. The backup server functions at a slower rate, modeled by an exponentially distributed random variable with parameter μ, and is assumed to be fully reliable. It operates only when the primary server is out of service. If the backup server is occupied, incoming requests are redirected to the orbit, and no collisions occur at the backup service unit.

Each primary customer in the system is characterized by an impatience property, which reflects their potential decision to leave the system if not served within a certain time frame. This decision to abandon the system is made after a random time period, which follows an exponential distribution with rate parameter τ.

The model assumes that all random variables are entirely independent in its formulation.

3 Simulation Results

3.1 First Scenario

We employed a statistical module class featuring an advanced analysis tool to estimate the mean and variance of observed variables using the batch mean method. This approach aggregates n consecutive observations from a steady-state simulation to produce a sequence of nearly independent samples. Renowned for its reliability, the batch mean method is widely used to construct confidence

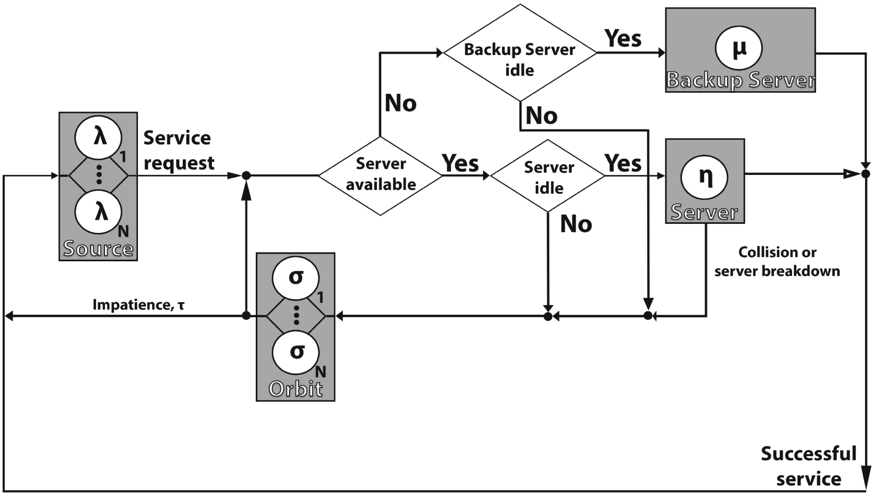

Fig. 1. System model

intervals for the steady-state mean of a process. To achieve approximate independence among sample averages, sufficiently large batch sizes are required. Further details on the batch mean method can be found in [2]. Our simulations were conducted with a 99.9% confidence level, terminating once the relative half-width of the confidence interval was reduced to 0.00001.

Table 1. Numerical values of model parameters

N	γ_0	γ_1	γ_2	σ	μ	τ
100	0.1	0.1	1	0.05	0.1	0.01

In this section, we aimed to determine service time parameters for each distribution to ensure they have identical mean values and variances. Four different distributions were examined to assess their influence on performance metrics. The hyper-exponential distribution was selected specifically to achieve a squared coefficient of variation greater than one. Table 2 outlines the input parameters for the various distributions, while Table 1 lists the values of other relevant parameters.

Figure 2 depicts the correlation between arrival intensity and the mean response time of customers who are successfully served. Successfully served customers refer to those who remain in the system until receiving service, unaffected by impatience. Among the distributions, the Pareto distribution exhibits the highest mean response time, while the distinctions between the remaining distributions become more evident. Of particular note, the gamma distribution yields the shortest mean response time.

Table 2. Parameters of service time of primary customers

Distribution	Gamma	Hyper-exponential	Pareto	Lognormal
Parameters	$\alpha = 0.011$ $\beta = 0.011$	$p = 0.494$ $\lambda_1 = 0.989$ $\lambda_2 = 1.011$	$\alpha = 2.005$ $k = 0.501$	$m = -2.257$ $\sigma = 2.125$
Mean	1			
Variance	90.25			
Squared coefficient of variation	90.25			

An intriguing observation is that as arrival intensity grows, the mean response time initially increases but then begins to decline after surpassing a certain threshold. This behavior is a hallmark of retrial queuing systems with a finite source and typically emerges under specific parameter settings.

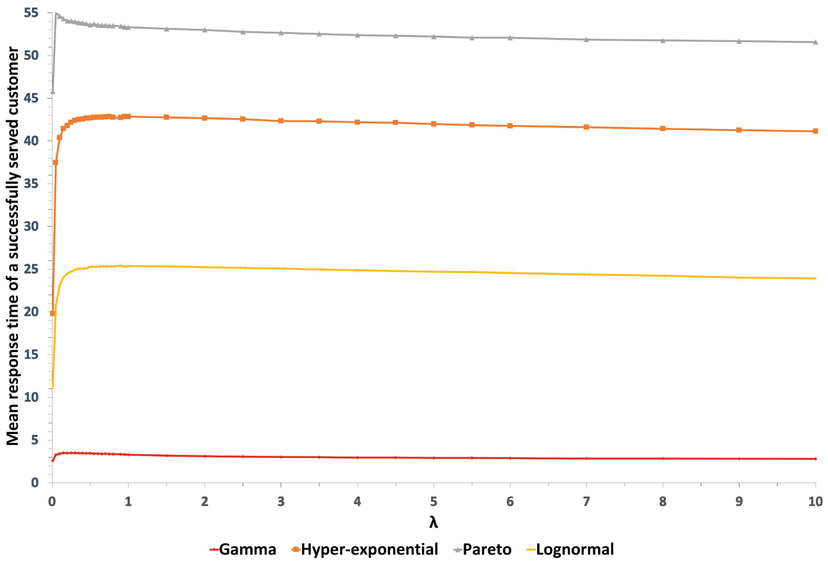

Fig. 2. Mean response time vs. arrival intensity

Figure 3 shows the mean response time of those customers who are served by the backup service unit in relation to the arrival rate of incoming customers. Inspecting closely the obtained results the same tendency develops what we observed in the previous figure. The Pareto distribution exhibits the highest values, while the gamma distribution produces the lowest values. Because the service rate of the backup unit is smaller than the rate of the primary service unit the spent time in the system of the customer served by the backup unit is higher on average.

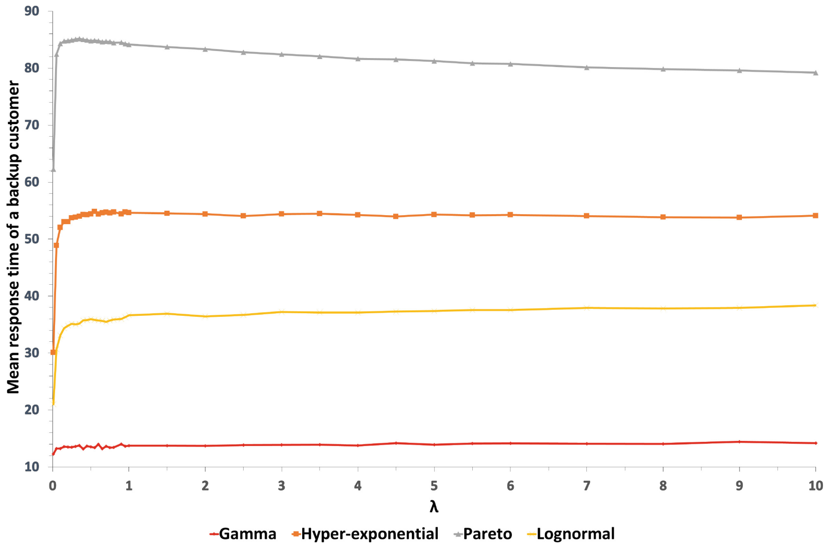

Fig. 3. Mean response time of a backup customer vs. arrival intensity

Figure 4 depicts the utilization of the backup service unit as a function of arrival intensity, comparing the different distributions. Unlike the significant differences observed in the previous figures, the results here are relatively close to each other. A closer look reveals that the backup service unit is utilized about 50% of the time, meaning it is occupied by customers for half of the total simulation period. As the arrival intensity increases, the utilization of the backup service unit also rises. However, once the arrival intensity reaches a certain threshold (around 1 in this case), the utilization plateaus.

Figure 5 illustrates the mean number of customer retrials as the arrival intensity increases. There are notable differences between the distributions used, with service times following a Pareto distribution showing particularly high retrial rates. In contrast, with a gamma distribution, requests typically do not attempt to reengage with the service unit, while other distributions exhibit a significantly higher number of collisions. The results also clearly show that after reaching a certain arrival intensity, the number of retrials stabilizes and does not continue to increase.

3.2 Second Scenario

Building on the results from the previous section, we next aimed to explore how changing the service time parameters would influence performance metrics. This time, we carefully chose parameters so that the squared coefficient of variation remained below one. Since the squared coefficient of variation for a hypo-exponential distribution is consistently less than one, we substituted the hyper-exponential distribution with its hypo-exponential counterpart. Distributions

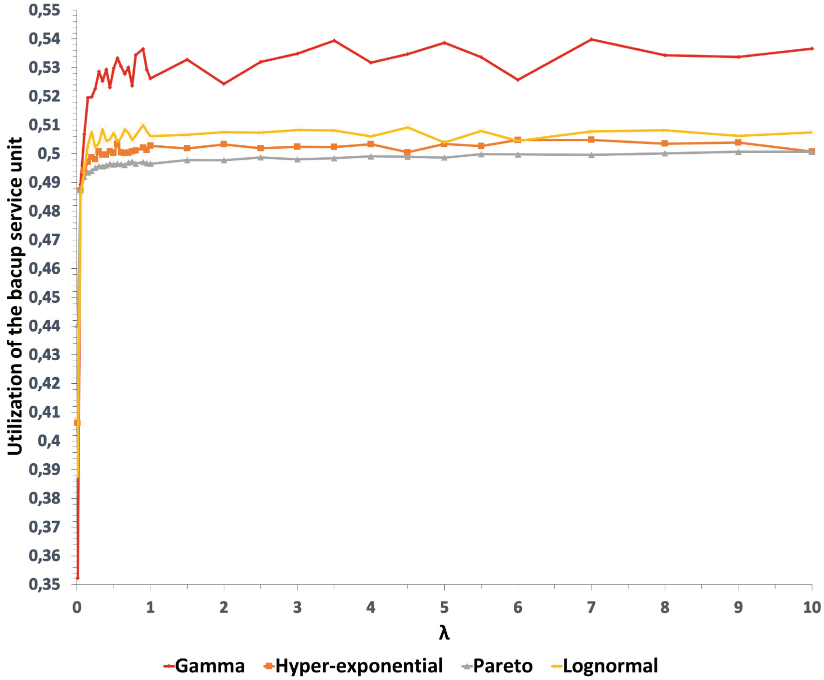

Fig. 4. The utilization of the primary service unit vs. arrival intensity

with squared coefficient of variation values below one tend to be more regular, meaning that values are clustered more closely around the mean, which is often the case for hypo-exponential distributions. In queuing theory, squared coefficient of variation below one typically leads to less fluctuation in waiting times, contributing to more predictable performance metrics. With these updated service time parameters, we revisited the same performance figures to assess the impact of these adjustments, as shown in Table 3. All other parameters were kept constant, as specified in Table 1.

Table 3. Parameters of service time of primary customers

Distribution	Gamma	Hypo-exponential	Pareto	Lognormal
Parameters	$\alpha = 1.522$	$\mu_1 = 4.5454$	$\alpha = 2.588$	$m = -0.252$
	$\beta = 1.522$	$\mu_2 = 1.282$	$k = 0.614$	$\sigma = 0.71$
Mean	1			
Variance	0.6568			
Squared coefficient of variation	0.6568			

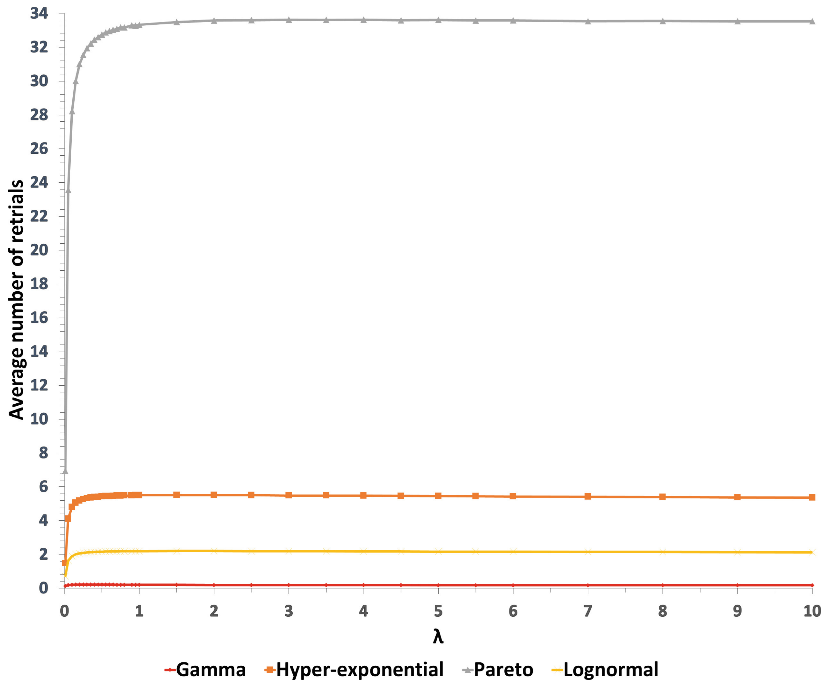

Fig. 5. The mean number of retrials vs. arrival intensity

To highlight the differences between the two scenarios, we begin by examining the mean response time of a customer, as shown in Fig. 6. The curves are noticeably closer to each other, with less pronounced differences, except for the Pareto distribution, which still produces significantly higher values compared to the other distributions. Similar to Fig. 2, the mean response time reaches a maximum, a common occurrence in retrial queuing systems with a finite customer pool. The same pattern is observed: after reaching a certain arrival intensity, the mean response time peaks and then gradually decreases as arrival intensity continues to rise.

Figure 7 illustrates the mean response time for customers served by the backup service unit. Similar to the previous scenario, the results follow the same trend observed in the prior figure, with the resulting curves notably close to each other and displaying minimal differences. Since the service rate of the backup unit is lower than that of the primary unit, customers served by the backup unit spend, on average, more time in the system.

Figure 8 illustrates the utilization of the backup service unit as a function of arrival intensity across different distributions. Unlike the substantial differences observed in previous figures, the results here are fairly similar. Closer inspection reveals that the backup service unit's utilization is around 50%, indicating it is

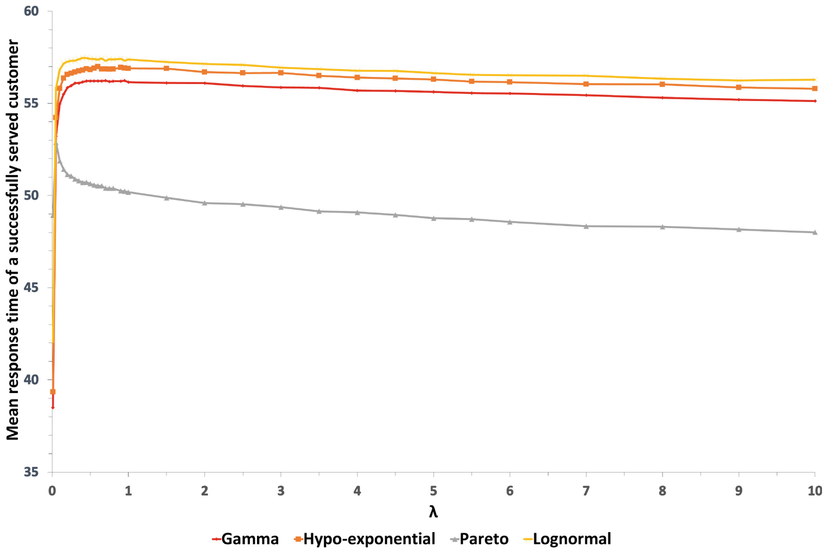

Fig. 6. Mean response time vs. arrival intensity

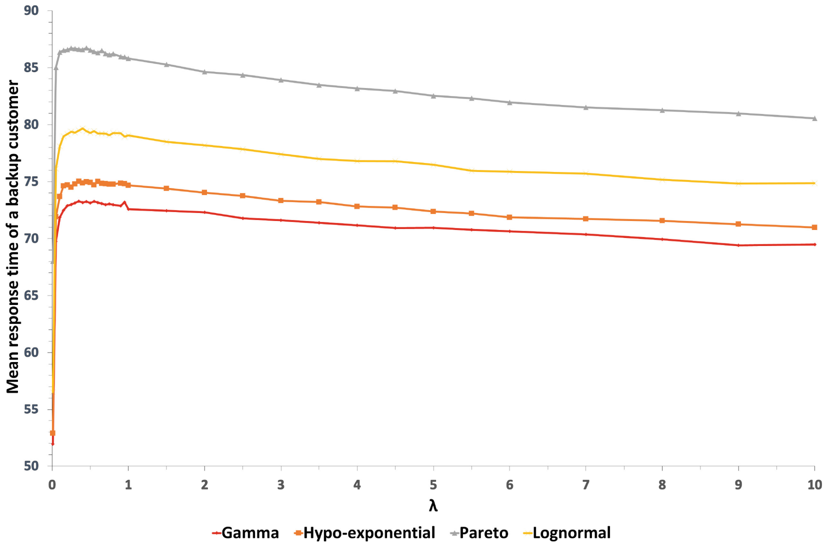

Fig. 7. Mean response time of a backup customer vs. arrival intensity

occupied by customers for about half of the total simulation time. A similar trend is evident: as arrival intensity increases, the backup service unit's utilization also rises. Once the arrival intensity reaches a specific threshold (around 1 in this instance), the utilization stabilizes.

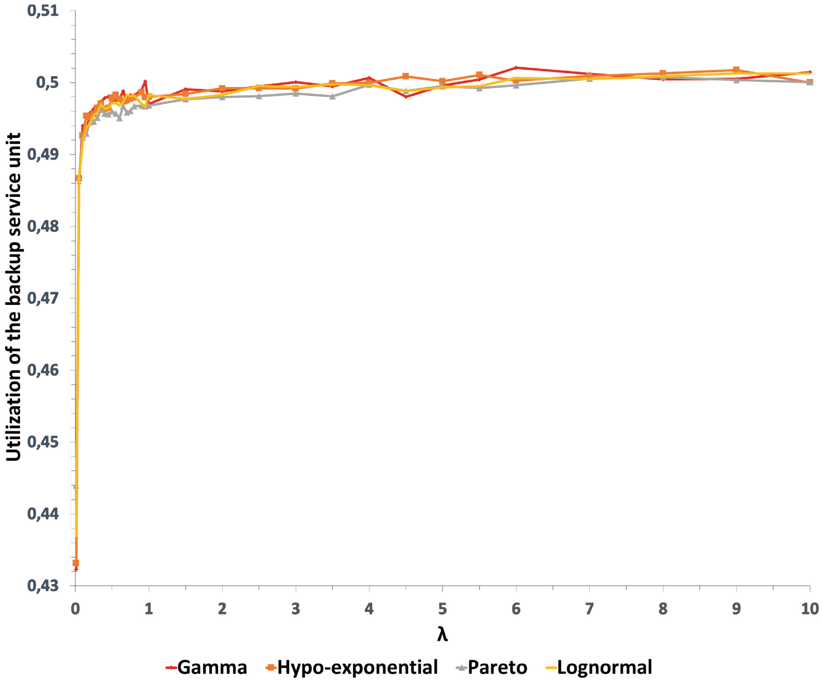

Fig. 8. The utilization of the primary service unit vs. arrival intensity

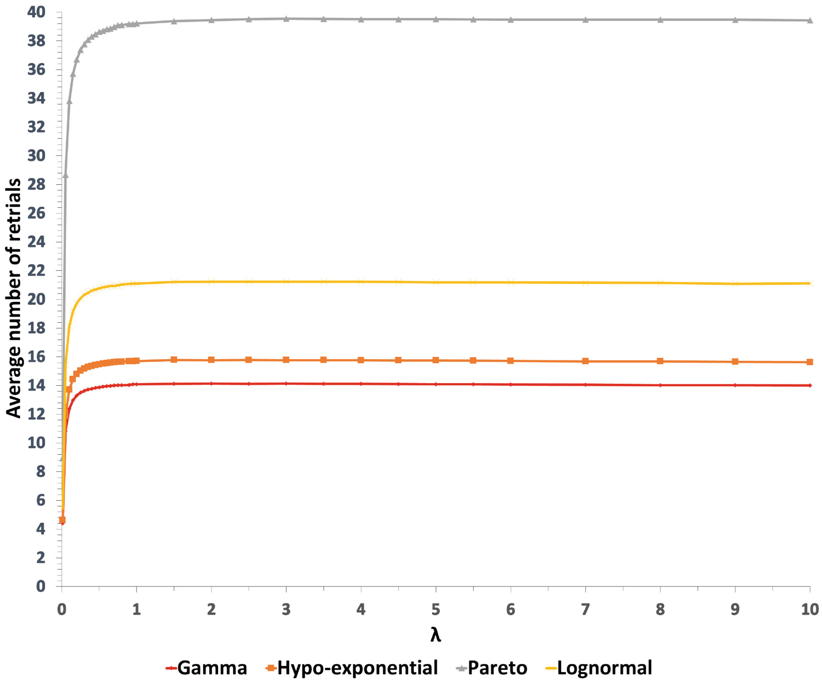

Fig. 9. The mean number of retrials vs. arrival intensity

Figure 9, which illustrates the average number of retries, shows a trend similar to the previous scenario. The highest retrial rates are seen with Pareto-distributed service times, while the lowest occur with gamma-distributed service times, although the differences are considerably smaller. Another noteworthy observation, upon closer inspection of the figure, is that the number of retries is higher for all distributions in the scenario using the previous parameter settings.

4 Conclusion

We simulated a retrial queuing system modeled as $M/G/1//N$, featuring an unreliable primary server and a backup service unit. The program enabled a sensitivity analysis of key performance metrics, such as the mean response time of successfully served customers. The results revealed significant differences in performance measures when the squared coefficient of variation exceeded one, highlighting the influence of the chosen distribution, while only minor deviations were observed when it was below one. The curves also demonstrated how customer impatience contributes to reducing the average response time for primary customers. Future research will focus on examining the effects of server blocking, incorporating two-way communication, exploring alternative impatience behaviors in different models, and performing sensitivity analyses on other variables, including failure rates.

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.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
4. 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
5. Fishwick, P.A.: SimPack: getting started with simulation programming in C and C++. In: In 1992 Winter Simulation Conference, pp. 154–162 (1992)
6. 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)
7. 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, no. 4 (2012)

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. 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
10. Kvach, A., Nazarov, A.: Sojourn time analysis of finite source Markov retrial queuing system with collision, chap. 8, pp. 64–72. Springer, Cham (2015)
11. Nazarov, A., Kvach, A., Yampolsky, V.: Asymptotic analysis of closed markov retrial queuing system with collision, chap. 1, pp. 334–341. Springer, Cham (2014)
12. 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. Industr. Manage. Optim. **12**, 851–878 (2015). https://doi.org/10.3934/jimo.2016.12.851
13. Satheesh R.K., Praba S.K.: A multi-server with backup system employs decision strategies to enhance its service. Res. Square 1–31 (2023). https://doi.org/10.21203/rs.3.rs-2498761/v1