Simulation of M/G/1//N System with Collisions, 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 queuing system that features request collisions, primary server unreliability, and a backup server. When a new job arrives while the service facility is occupied, a collision occurs, sending both jobs to a virtual waiting room called the orbit. In the orbit, customers make repeated attempts to access the server after random intervals. If a server breakdown occurs, the customer at the server is sent to the orbit. The paper's novelty lies in the implementation of a backup facility for when the primary server is unavailable, and a sensitivity analysis using various service time distributions for primary customers.

We examined scenarios where key performance measures are visually represented to highlight observed disparities. Specifically, we analyzed two different scenarios, illustrating the most significant performance measures to emphasize the differences. These visual representations allowed us to identify critical performance bottlenecks and the effectiveness of the backup server. The findings provide valuable insights into the system's behavior under different conditions, helping to improve reliability and efficiency in similar queueing systems.

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

1 Introduction

In the contemporary context characterized by escalating traffic volumes and expanding user bases, the analysis of communication systems or the design of optimal configurations poses a formidable challenge. Given the pivotal role of information exchange across all spheres of life, it becomes imperative to develop or adapt mathematical and simulation models for telecommunication systems to align with these evolving dynamics. Retrial queues emerge as potent and apt tools for modeling real-world scenarios encountered in telecommunication systems, networks, mobile networks, call centers, and analogous domains. A plethora of scholarly works, exemplified by references such as [3,4], have been dedicated

V. M. Vishnevsky et al. (Eds.): DCCN 2024, LNCS 15460, pp. 144–155, 2025.
https://doi.org/10.1007/978-3-031-80853-1_12

to investigating various manifestations of retrial queuing systems characterized by retrial calls.

In certain contexts, researchers postulate the perpetual availability of service units, yet operational interruptions or unexpected events may occur, resulting in the rejection of incoming customers. Devices deployed across diverse industries are susceptible to malfunctions, rendering the presumption of their infallible operation overly sanguine and impractical. Likewise, within wireless communication environments, diverse factors can impinge upon transmission rates, precipitating interruptions during packet delivery. The inherent unreliability of retrial queuing systems significantly influences system functionality and performance metrics. Concurrently, halting production entirely is unviable, as it may engender delays in order fulfillment. Therefore, amidst such occurrences, machines or operators endowed with lower processing capacities may continue operating to sustain smoother functionality. Moreover, the authors investigate the viability of incorporating a backup server capable of delivering services at a diminished rate in instances of primary server unavailability. Numerous recent scholarly works have extensively examined retrial queuing systems featuring unreliable servers, as exemplified by references such as [6].

In service-oriented domains, service providers frequently encounter operational disruptions stemming from various factors, including database accessibility issues hindering the fulfillment of customer requests. In response to such disruptions, service providers commonly resort to contingency measures such as activating backup systems or eliciting additional information from customers to facilitate resolution. Several scholarly works extensively explore the dynamics of systems aimed at augmenting service provision through the integration of backup servers, as evidenced by references such as [1,7,11,13,14].

In technological contexts such as Ethernet networks or constrained communication sessions, the occurrence of job collisions is probable. Multiple entities within the source may initiate asynchronous attempts, causing signal interference and necessitating retransmissions. Hence, it is imperative to incorporate this phenomenon into investigations aimed at devising effective policies to mitigate conflicts and associated message delays. Publications addressing results related to collisions include [8–10,12].

The aim of our study is to conduct a sensitivity analysis, employing diverse service time distributions of the primary server, to assess the main performance metrics under scenarios involving a backup facility. During failure periods of the primary server, the service of the customers is traversed to the backup service facility and until restoration, new customers are permitted to reach the backup unit or the orbit if it is busy. Our investigation emphasize the effect of a backup service unit and the results are obtained through simulation using Simpack [5]. The simulation program is developed upon fundamental code elements enabling the computation of desired metrics across a range of input parameters. Graphical representations are provided to elucidate the impact of different parameters and distributions on the primary performance indicators.

2 System Model

We examine a finite-source retrial queueing system characterized by type $M/G/1//N$ representing Fig. 1, incorporating an unreliable primary service unit, occurrences of collisions, and a backup service unit. This model features a finite source, with each of the N individuals generating requests to the system according to an exponential distribution with parameter λ. Arrival times follow an exponential distribution with a mean of $\lambda * N$. With no queues present, service for arriving jobs commences immediately following a gamma, hypo-exponential, hyper-exponential, Pareto, or lognormal distribution, each with distinct parameters but equivalent mean and variance values (η). In instances of server congestion, an arriving customer triggers a collision with the customer currently under service, resulting in both being transferred to the orbit. Jobs residing in the orbit initiate further attempts to access the server after an exponentially distributed random time with parameter σ. Additionally, random breakdowns occur, with failure times represented by exponential random variables. The failure time has a parameter of γ_0 when the server is occupied and γ_1 when idle.

Upon the failure of the service unit, the repair process commences immediately, with the duration of the repair following an exponential distribution characterized by parameter γ_2.

In the event of a busy server experiencing a failure, the customer is promptly transitioned to the orbit. Despite the unavailability of the service unit, all customers in the source retain the capability to generate requests, albeit directed towards the backup server, which operates at a reduced rate characterized by an exponentially distributed random variable with parameter μ during periods of primary server unavailability. The backup server is assumed to be reliable and operates solely in the absence of the primary server. When the backup server is occupied, incoming requests are directed to the orbit. The phenomenon of collision does not occur in front of the backup service unit. The model assumes complete independence among all random variables during its formulation.

3 Simulation Results

3.1 First Scenario

We employed a statistical module class equipped with a statistical analysis tool to quantitatively estimate the mean and variance values of observed variables using the batch mean method. This method involves aggregating n successive 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 for the steady-state mean of a process. To ensure that the sample averages are approximately independent, large batches are necessary. Further details on the batch mean method can be found in [2]. Our simulations were conducted with a confidence level of 99.9%, and the simulation run was terminated once the relative half-width of the confidence interval reached 0.00001.

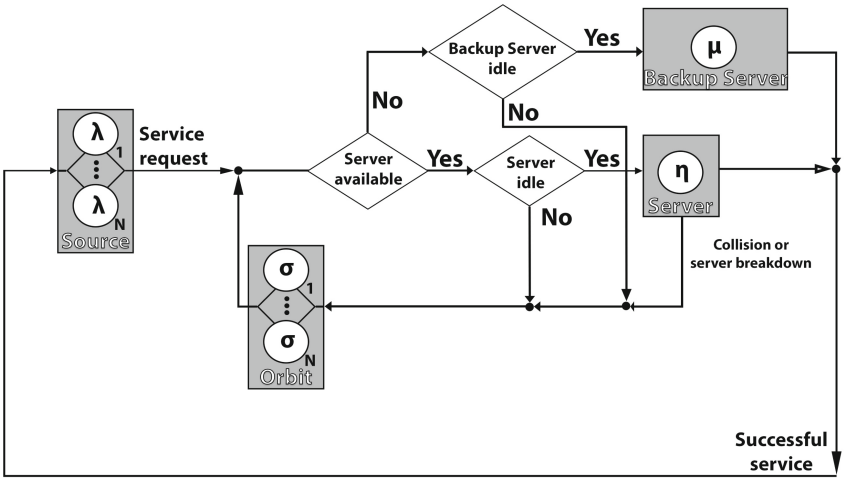

Fig. 1. System model

Table 1. Numerical values of model parameters

N	γ_0	γ_1	γ_2	σ	μ
100	0.1	0.1	1	0.05	0.6

In this section, our objective was to determine the service time parameters for each distribution in a manner that ensures equal mean values and variances. Four distinct distributions were examined to assess their influence on performance metrics. Specifically, the hyper-exponential distribution was selected to ensure a squared coefficient of variation greater than one. The input parameters of the various distributions are presented in Table 2, while Table 1 provides the values of other relevant parameters.

Table 2. Parameters of service time of primary customers

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

Figure 2 depicts the correlation between the mean response time of customers and the arrival intensity. The Pareto distribution exhibits the highest mean

response time, while the distinctions among the other distribution types become more apparent. Notably, the gamma distribution stands out by yielding the lowest mean response time. An intriguing observation is that, as the arrival intensity increases, the mean response time initially rises but subsequently decreases after reaching a specific threshold. This behavior is a characteristic feature of retrial queuing systems with a finite source, and it tends to manifest under appropriate parameter configurations.

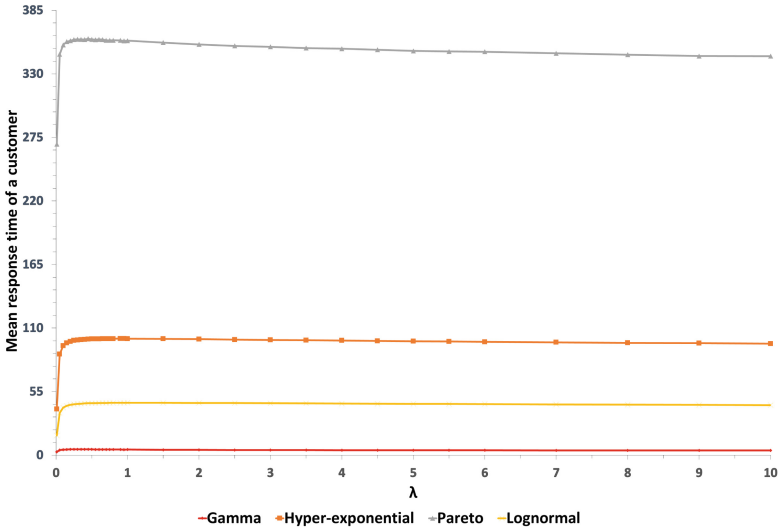

Fig. 2. Mean response time vs. arrival intensity

Figure 3 illustrates the utilization of the service unit in relation to the arrival rate of incoming customers. Despite possessing identical mean and variance values, notable distinctions are observed among different distributions. As the arrival rate escalates, the utilization of the service unit correspondingly rises. Specifically, the utilization rate is lower with the gamma distribution compared to other distributions, particularly evident with the hyper-exponential distribution. Interestingly, in the case of Pareto distribution the tendency is reversed as the utilization of the primary service unit starts to decrease besides increasing arrival intensity.

Figure 4 illustrates the utilization of the backup service unit as a function of arrival intensity comparing the used distributions with each other. Noticing the huge differences in the previous figures this time the results are quite close to each other. Upon closer inspection, it is evident that the utilization of the backup service unit is approximately 20%, indicating that the backup service unit is occupied by customers for one-fifth of the total simulation time. As the arrival intensity increases, the utilization of the backup service units also rises.

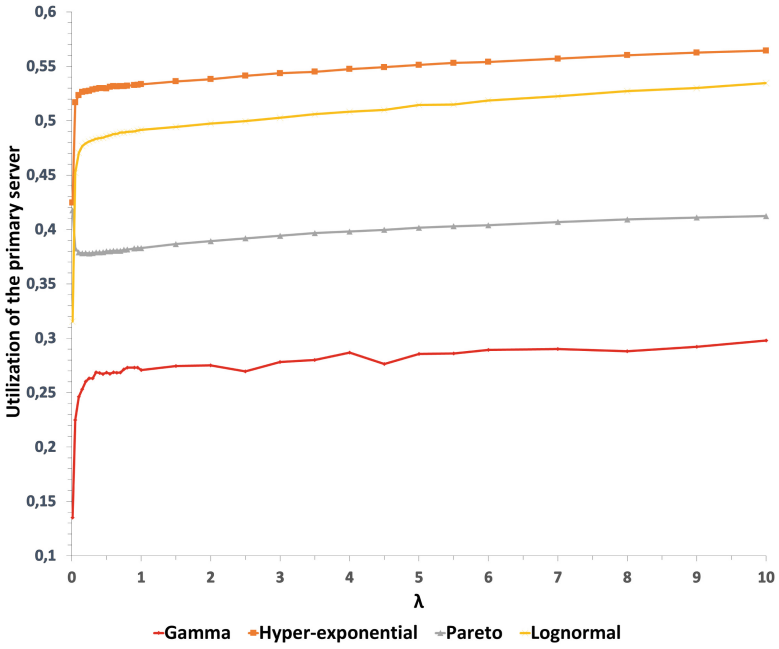

Fig. 3. Comparison of utilization

However, after reaching a certain arrival intensity (approximately 1 in this case), utilization becomes essentially stagnant.

Figure 5 highlights the mean number of retrials of a customer while the arrival intensity of the customers is increasing. Significant differences arise between the distributions used, with particularly high values observed for service times following a Pareto distribution. While with a gamma distribution, requests on average do not retry to engage with the service unit, other distributions show a considerably higher number of collisions. The results also clearly indicate that upon reaching a certain arrival intensity, the number of retries does not increase but rather remains at a constant value.

3.2 Second Scenario

We were curious about how the performance measurements would change with the modification of the service time parameters, following the results from the previous section. This time, the parameters were selected to ensure that the squared coefficient of variation was below one. Since the squared coefficient of variation for a hypo-exponential distribution is always less than one, we replaced the hyper-exponential distribution with the hypo-exponential distribution. Using these new service time parameters, we will review the same figures as in the

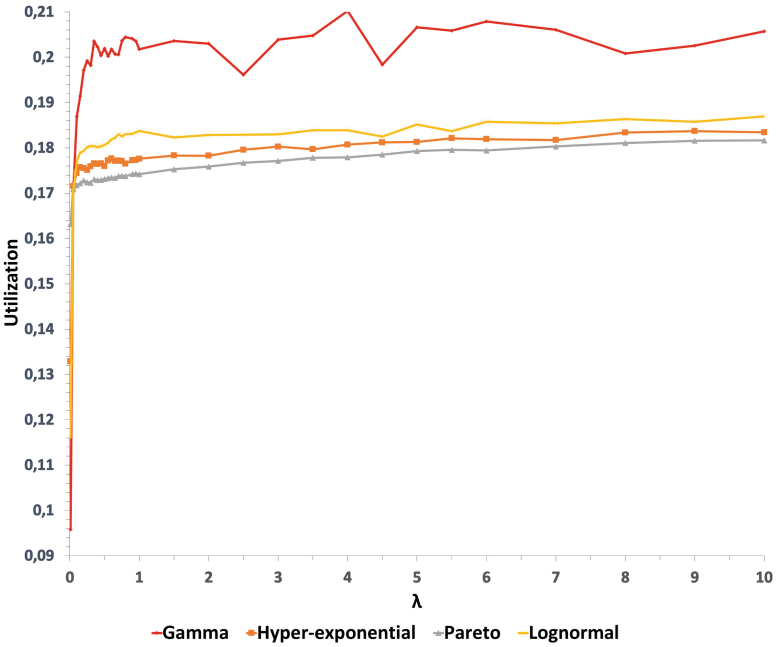

Fig. 4. Comparison of utilization of the backup service unit

previous section to observe the impact of the newly chosen parameters, which are shown in Table 3. The other parameters remain unchanged (see Table 1).

Table 3. Parameters of service time of primary customers

Distribution	Gamma	Hypo-exponential	Pareto	Lognormal
Parameters	$\alpha = 1.8$	$\mu_1 = 1.5$	$\alpha = 2.673$	$m = -0.22$
	$\beta = 1.8$	$\mu_2 = 3$	$k = 0.626$	$\sigma = 0.665$
Mean	1			
Variance	0.555			
Squared coefficient of variation	0.555			

To elucidate the differences between the two scenarios, we first examine the mean response time of a customer, as depicted in Fig. 6. The resulting curves are notably closer to each other, exhibiting less significant differences, except for the Pareto distribution, which continues to yield substantially higher values compared to the other distributions. As shown in Fig. 2, the mean response time reaches a maximum value, a common phenomenon in retrial queuing systems

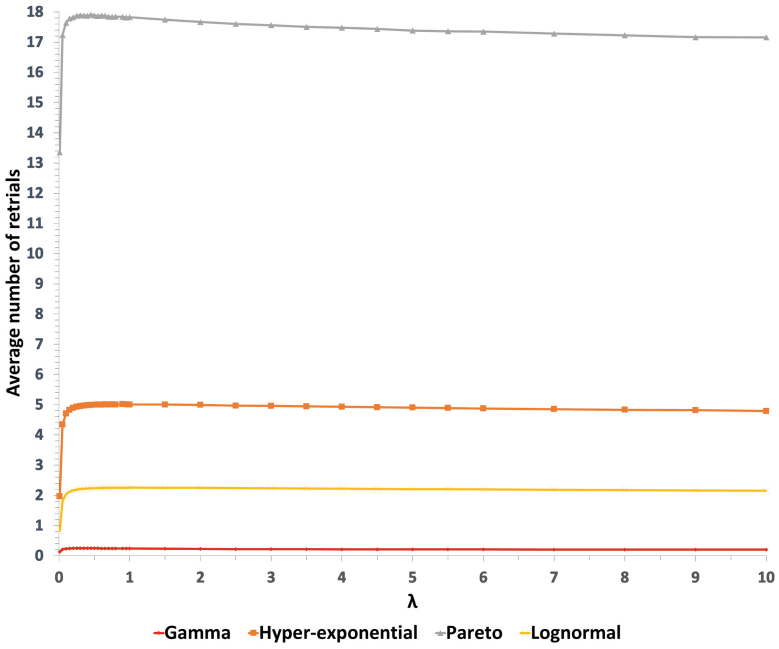

Fig. 5. The mean number of retrials of a customer

with a finite number of customers in the source. The same tendency is observed: after reaching a certain arrival intensity, the mean response time peaks and subsequently decreases as the arrival intensity continues to increase.

The next figure (Fig. 7) in this section illustrates the utilization of the primary service unit by customers. Close examination of the figure reveals that the Pareto distribution results in lower utilization values, indicating fewer customers under service and greater number of collisions. For the other distributions, the values are relatively close to each other. As the arrival intensity increases, the utilization of the primary service unit initially decreases, but beyond an arrival intensity of 0.01, it gradually increases, a trend consistent across all investigated cases.

Regarding the average number of retries, which is shown on Fig. 9 a similar trend is observed as in the previous scenario. The highest values are found with Pareto-distributed service times, while the lowest values are with gamma-distributed service times, though the differences are significantly smaller. Another notable observation, upon closely examining the figure, is that the number of retries is higher for all distributions in the scenario run with the previous parameter settings.

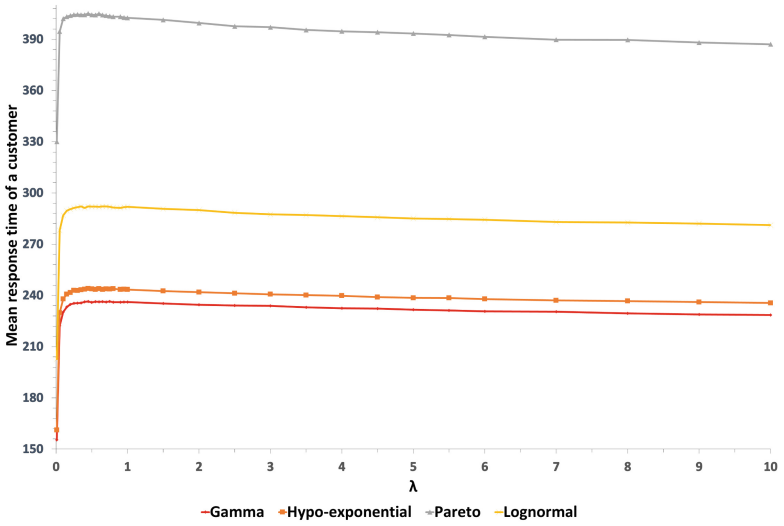

Fig. 6. Mean response time vs. arrival intensity

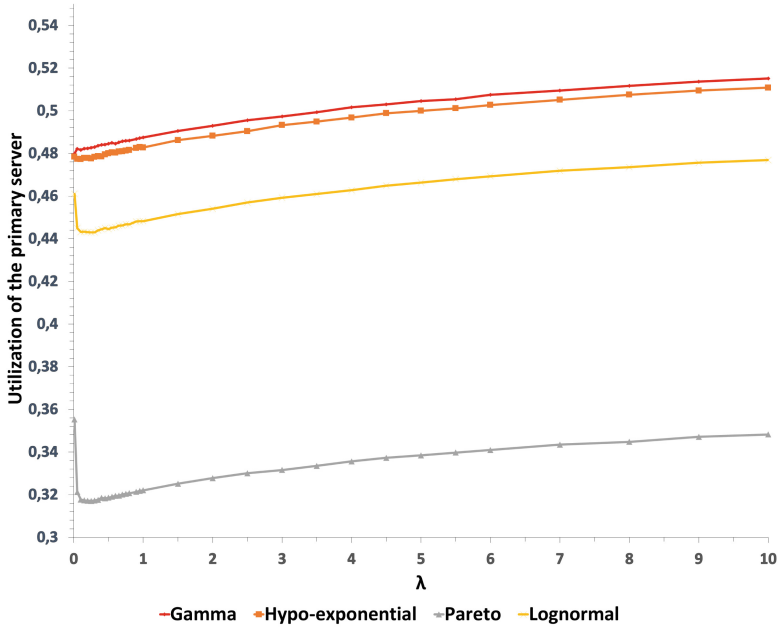

Fig. 7. Comparison of utilization

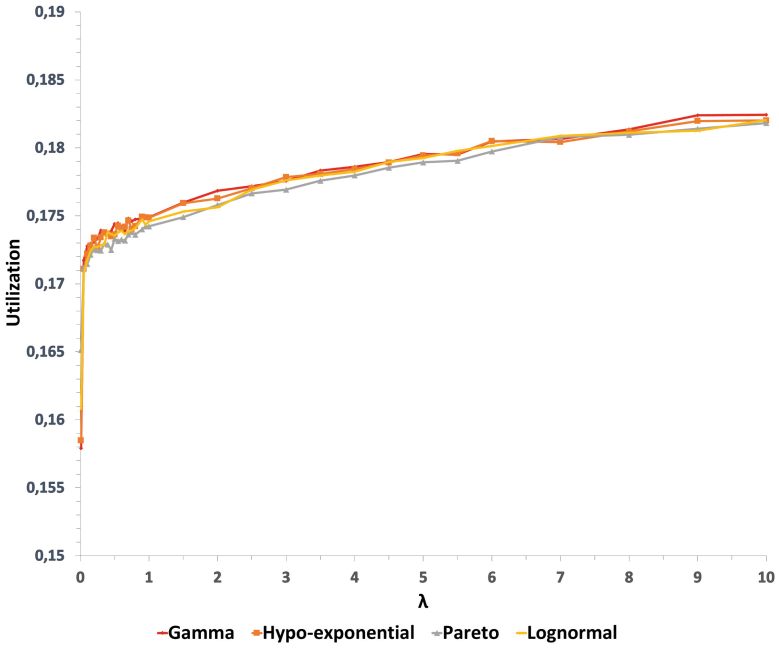

Fig. 8. Comparison of utilization of the backup service unit

Figure 8 illustrates 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 are quite similar. Closer inspection reveals that the utilization of the backup service unit is approximately 18%, indicating that it is occupied by customers for one-fifth of the total simulation time. The same trend arises so as the arrival intensity increases, the utilization of the backup service unit also rises. However, after reaching a certain arrival intensity (approximately 1 in this case), utilization becomes essentially stagnant.

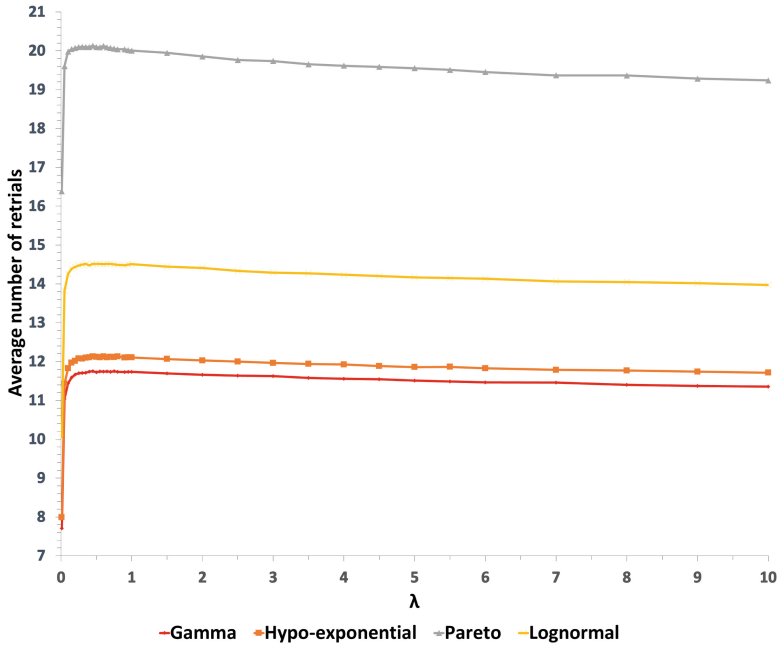

Fig. 9. The mean number of retrials of a customer

4 Conclusion

We conducted simulations of a retrial queuing system following the $M/G/1//N$ model, incorporating an unreliable primary server and a backup service unit. Our program was utilized to perform a sensitivity analysis on various performance metrics, including the mean response of times of the customers. From a multitude of parameter configurations, the most relevant measures were selected and graphically depicted. Notably, when the squared coefficient of variation exceeds one, significant deviations are observed among distributions across multiple aspects of the investigated metrics. When the squared coefficient of variation is less than one, differences among the utilized distributions tend to be less significant, and the curves almost overlap each other except for Pareto distribution. In summary, the obtained results display the influence of various distributions of service time on the performance measures like the mean response time of a customer or the utilization of the primary service unit. In future studies, the authors intend to further explore the impact of server blocking, impatience of the customers in alternative models and conduct sensitivity analyses for other variables, such as 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. 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
8. Krishnamoorthy, A., Pramod, P.K., Chakravarthy, S.R.: Queues with interruptions: a survey. TOP **22**(1), 290–320 (2012). https://doi.org/10.1007/s11750-012-0256-6
9. Kvach, A., Nazarov, A.: Sojourn time analysis of finite source Markov retrial queuing system with collision. In: Dudin, A., Nazarov, A., Yakupov, R. (eds.) ITMM 2015. CCIS, vol. 564, pp. 64–72. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25861-4_6
10. Kvach, A., Nazarov, A.: Numerical research of a closed retrial queueing system M/GI/1//N with collision of the customers. In: Proceedings of Tomsk State University. A Series of Physics and Mathematics. Tomsk. Materials of the III All-Russian Scientific Conference, vol. 297, pp. 65–70. TSU Publishing House (2015). (in Russian)
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. Nazarov, A., Kvach, A., Yampolsky, V.: Asymptotic analysis of closed Markov retrial queuing system with collision. In: Dudin, A., Nazarov, A., Yakupov, R., Gortsev, A. (eds.) ITMM 2014. CCIS, vol. 487, pp. 334–341. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13671-4_38
13. 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
14. Won, Y., Ban, J., Min, J., Hur, J., Oh, S., Lee, J.: Efficient index lookup for de-duplication backup system, pp. 383–384 (2008). https://doi.org/10.1109/MASCOT.2008.4770594