World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Variability of queue length and waiting times in a system with general service times

Przemysław Włodarski *

Department of Signal Processing and Multimedia Engineering, West Pomeranian University of Technology in Szczecin, 70-313 Szczecin, Poland.

World Journal of Advanced Engineering Technology and Sciences, 2025, 17(02), 454-461

Publication history: Received 17 October 2025; revised on 22 November 2025; accepted on 24 November 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.17.2.1520

Abstract

An essential issue in telecommunication network design and performance analysis involves determining how variability in arrivals and service affects system congestion. In particular, focus is placed on queue length, which corresponds to the number of waiting packets, and on delay, which is the main factor influencing flow performance. To ensure a given level quality of service, it is important to know these values before transmission or processing. The article demonstrates the impact of observation time on the estimated mean number of packets in the system, as well as on the mean packet waiting time in the queue, for a queueing system with a general service time. The results obtained for two different service time distributions, with constant and variable coefficients of variation, were compared with each other and with theoretical values.

Keywords: Queueing system; M/G/1 queue; General service time

1. Introduction

Queueing analysis plays a crucial role in optimizing the performance of computer networks. Network traffic is inherently stochastic, meaning that packets arrive randomly and require variable processing or transmission times. Queueing theory provides a structured framework for modeling these dynamics, enabling quantitative evaluation of network performance. By modeling network nodes, such as routers, switches, and servers, by using queueing systems, it becomes possible to estimate key metrics: queue length, representing the number of packets waiting in the system, and waiting time, which is the time packets spend in the queue before being processed. Accurate predictions of these metrics help prevent bottlenecks, optimize resource usage, and reduce the probability of packet loss. Using an M/G/1 queueing model and experiments based on workloads, it can be shown that the model can predict how increasing request arrival rates affect system performance and can potentially forecast the impact of adding more workloads in the future [1]. This queueing model can also support the development of traffic management strategies, including load balancing across multiple servers or network paths, prioritization of packets based on application requirements and admission control mechanisms [2]. By analyzing network behavior under varying traffic conditions, telecommunication operators can implement policies that maintain stability and efficiency.

The M/G/1 queueing system is a fundamental model in queueing theory, representing a single-server queue with Markovian (Poisson) arrivals and a general service time distribution. Unlike M/M/1 queues, where service times are exponentially distributed and memoryless, the M/G/1 model allows service times to follow any arbitrary distribution. This generality introduces an important aspect of queueing dynamics: service time variability, which significantly impacts system performance [3]. Variants of M/G/1, often extended with vacations, retrials or priority disciplines, are used to model nodes with low-power and sleep modes and channel access with variable service times due to scheduling and retransmissions. These models capture energy-saving policies and delays that are not well represented by exponential service times [4,5]. For full-duplex, wireless sensor networks the M/G/1-based analysis is used to separate

* Corresponding author: Przemysław Włodarski

queuing delay from MAC channel access delay, enabling tractable analytic expressions for mean delay and higher moments for realistic service distributions. That separation supports protocol design (e.g., backoff settings) and the evaluation of quality of service (QoS) targets for latency-sensitive traffic [6]. Recent work has also focused on data-driven and statistical fitting of M/G/1 and related single-server models to the captured network traces [7]. Maximum likelihood estimators (MLE) for arrival or service parameters make the M/G/1 useful for performance diagnosis and capacity planning in cloud and edge applications where service distributions should be taken from real data rather than assumed. Another study extends classical retrial queue models by applying an M/G/1 system in which customers leave to an 'orbit' while their service continues, enabling the derivation of performance metrics and the comparison of different service time distributions [8].

In order to analyze these complex variants, multiple mathematical techniques are employed. Laplace-Stieltjes (LST) transforms and supplementary-variable methods are widely used to derive performance measures (e.g., waiting times distributions, queue length distributions) in systems with general service times. The second technique is the asymptotic analysis that helps in systems with large or growing populations: for example, finite-source M/G/1 retrial queues with server breakdowns [9]. There are also matrix-geometric methods that are used in modified M/G/1 systems (e.g., with vacations or interruptions), enabling tractable computations of steady-state distributions [10,11,12]. Another research study analyze M/G/1 queues with heavy-tailed service time distributions, which often have no closed-form Laplace transforms. It proposes a method to approximate these transforms, enabling algorithms to compute the steady-state waiting time distribution to a desired accuracy [13].

2. M/G/1 queueing model

The single server M/G/1 queue is a very common model used in performance analysis of telecommunication systems. It assumes exponential interarrival times with the rate λ in pkts/s (packets per second) and general service time distribution with service rate μ also in pkts/s. In order to obtain performance metrics, the frequently used Pollaczek-Khinchine (P-K) transform approach is applied, which provides closed-form expressions for generating functions or Laplace-Stieltjes transforms (LST) of the steady-state distributions [3].

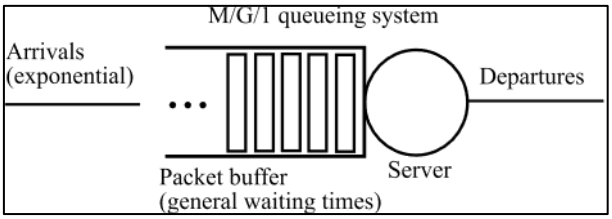

Figure 1 Diagram of the M/G/1 queueing system**.**

Probability density function of interarrivals has the following form:

$$f_a(t;\lambda) = \lambda e^{-\lambda t} \tag{1}$$

and the cumulative distribution function is defined as:

$$F_a(t;\lambda) = \int_0^t f_a(x;\lambda)dx = 1 - e^{-\lambda t} \tag{2}$$

Using the P-K formula, one can derive expressions for the mean waiting time in the queueing system:

$$W_q = \frac{\lambda \mathbb{E}(S^2)}{2(1-\rho)} = \frac{\rho[1+\mathbb{V}(S)\mu^2]}{2\mu(1-\rho)} \tag{3}$$

where $E(S^2)$ and $V(S)$ are the second moment and variance of the service time distribution, respectively, and ρ = λ/μ is the traffic load, which should be less than one in order for the M/G/1 queue to be stable. From Little's law, one can easily obtain the mean number of packets in the queue:

$$L_q = \lambda W_q = \frac{\lambda^2 \mathbb{E}(S^2)}{2(1-\rho)}. \tag{4}$$

The mean number of packets in the system, i.e. all packets in the queue plus the one in the server, can be derived from the formula for mean number of packets in the queue in (4):

$$L = L_q + \rho = \rho + \frac{\lambda^2 \mathbb{E}(S^2)}{2(1-\rho)} = \frac{\rho}{1-\rho}\left[1 - \frac{\rho}{2}(1 - \mu^2 \mathbb{V}(S))\right]. \tag{5}$$

3. Simulation results

The aim of the experiment is to examine the impact of the observation duration of a queueing system on the stability of the estimated queue length and queueing delay times. Simulations were carried out for different observation times (5s, 10s, 20s and 500s) and for various values of system load, ranging from 0.1 to 0.9 in steps of 0.1. Two service-time distributions were considered: exponential and uniform. The arrival rate was constant for all simulations and equal to λ = 10 pkts/s. The focus was on two fundamental metrics of the M/G/1 system, namely: the mean number of packets in the system and the mean packet delay in the queue. For both metrics, the values obtained in the experiment were compared with the analytical values given by formulas (5) and (3). Furthermore, for the mean number of packets in the system, the absolute error (AE) and the relative absolute error (RAE) were calculated for traffic loads ranging from 0.1 to 0.9.

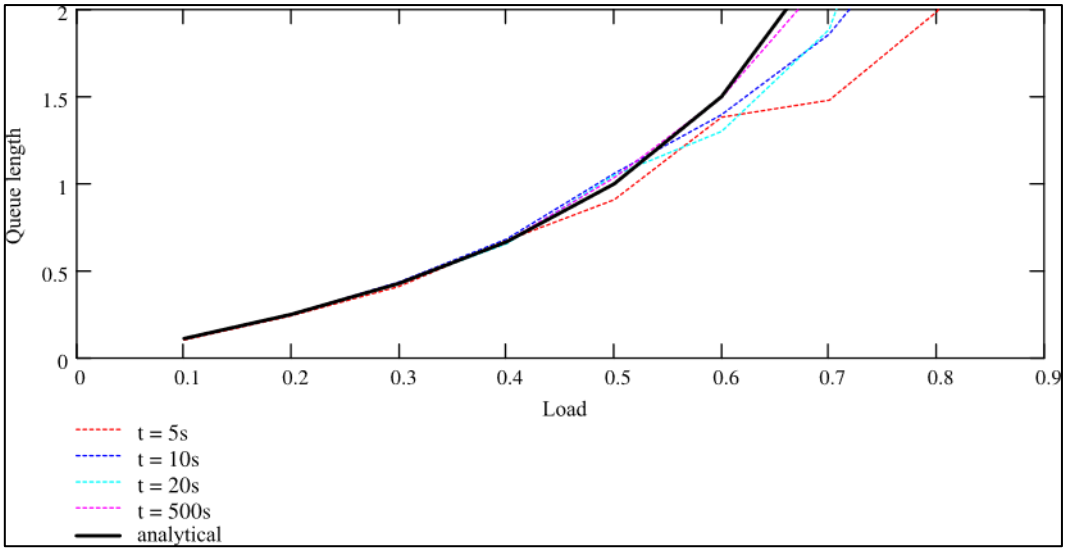

Figure 2 Mean queue lengths for different observation times and analytical curve, exponential service distribution

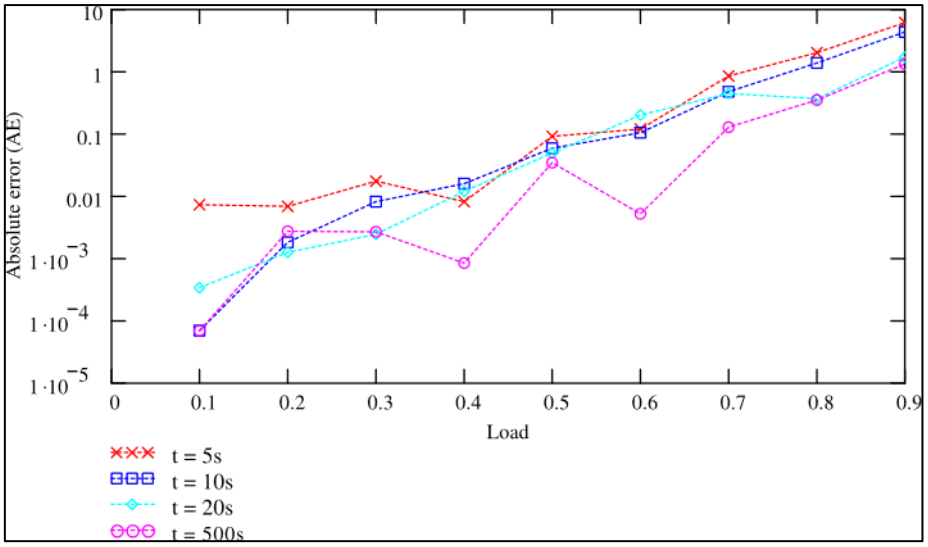

Figure 3 Absolute errors for mean queue lengths and different observation times, exponential service distribution

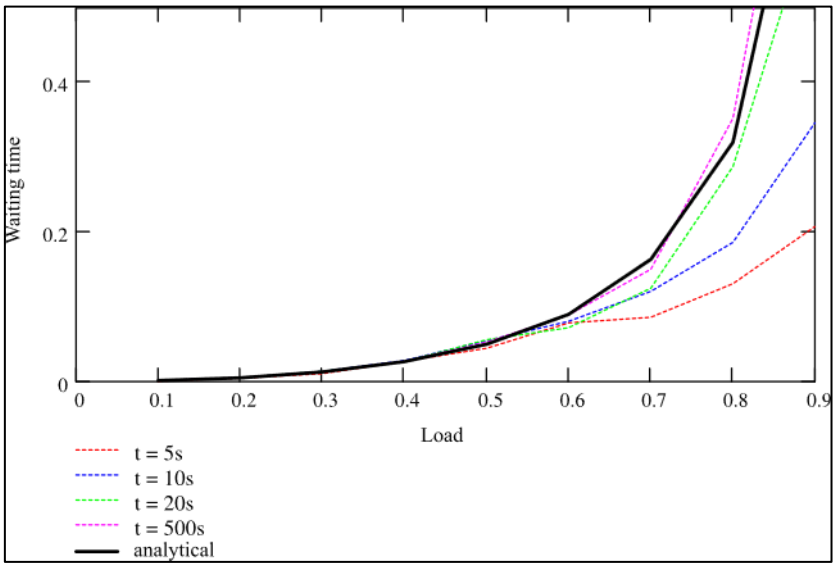

Figure 4 Mean waiting times for different observation times and analytical curve, exponential service distribution

Table 1 Errors of mean queue lengths for different times of observation and all range of analyzed load values, exponential service times (coefficient of variation: 1)

Time	Absolute error	Relative absolute error [%]	Root mean square error
5s	1.031	20.7	2.067
10s	0.712	13.4	1.455
20s	0.311	7.6	0.575
500s	0.205	3.8	0.434

The second service time distribution tested is the uniform distribution, whose probability density function and cumulative distribution function have the following form:

$$f_u(t;a,b) = \begin{cases} \dfrac{1}{b-a}, & \text{for } a \le t \le b \\ 0, & \text{for } t < a \vee t > b \end{cases}, \tag{6}$$

$$F_u(t;a,b) = \begin{cases} 0, & \text{for } t < a \\ \dfrac{t-a}{b-a}, & \text{for } a \le t \le b \\ 1, & \text{for } t > b \end{cases}. \tag{7}$$

For values a > 0 of the uniform distribution parameters, the coefficient of variation changes with the traffic load, as shown in Fig. 1. In the experiment, the value a = 0.008 was used due to the high variability of the coefficient of variation, but for comparison the relationships for smaller values of a (0.004 and 0.001) were also presented.

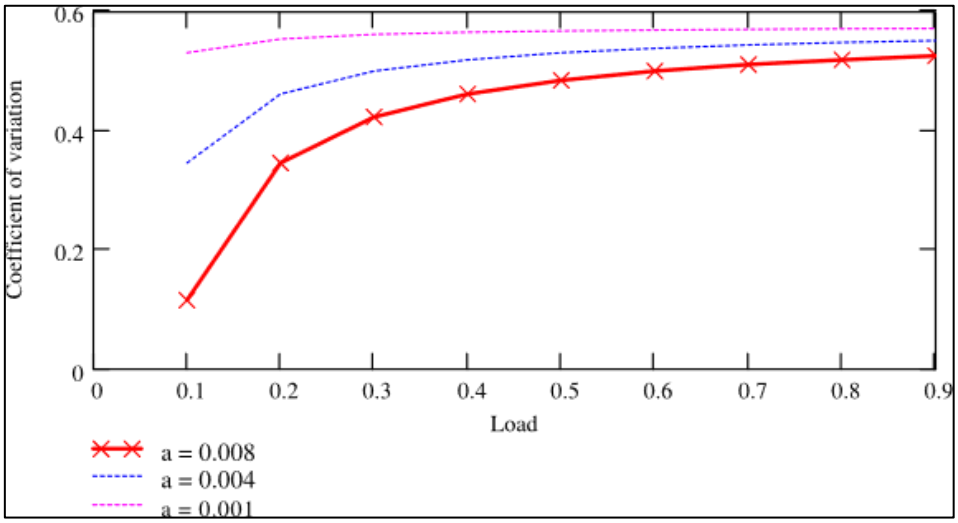

Figure 5 Coefficient of variation for different values of a in the uniform distribution

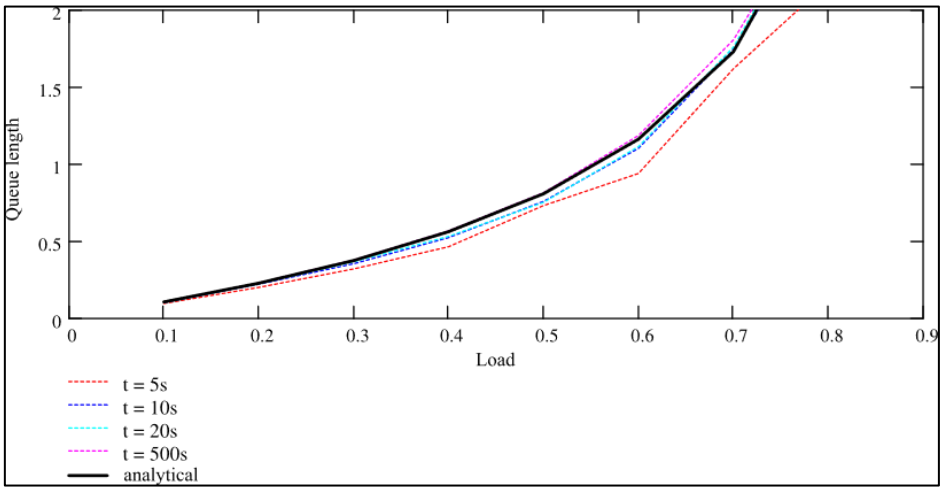

Figure 6 Mean queue lengths for different observation times and analytical curve, uniform service distribution

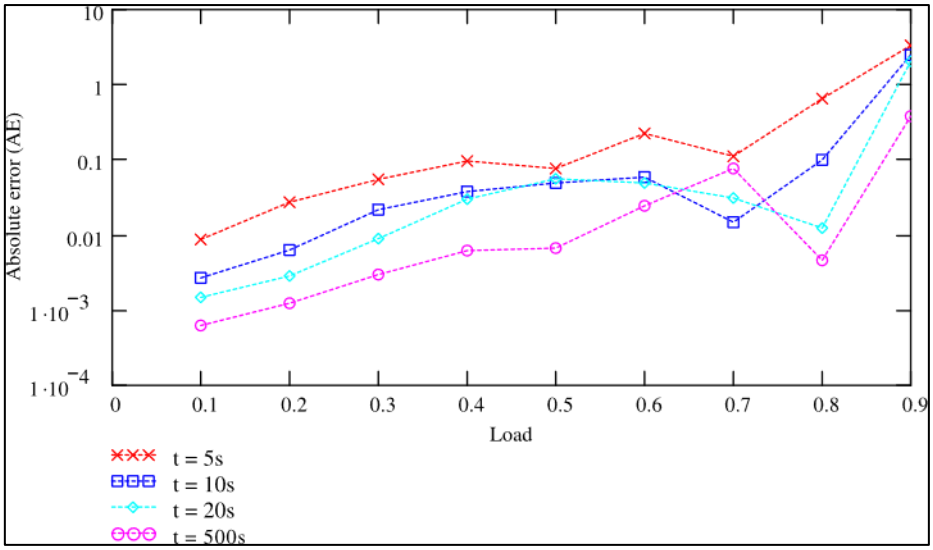

Figure 7 Absolute errors for mean queue lengths and different observation times, uniform service distribution

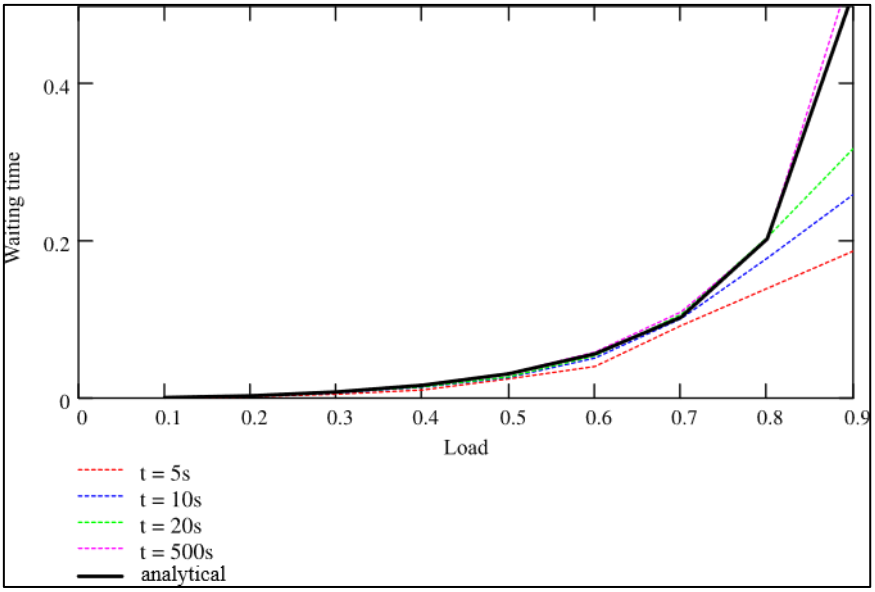

Figure 8 Absolute errors for mean queue lengths and different observation times, uniform service distribution

Table 2 Errors of mean queue lengths for different times of observation and all range of analyzed load values, uniform service times (mean coefficient of variation: 0.432).

Time	Absolute error	Relative absolute error [%]	Root mean square error
5s	0.508	18.295	1.075
10s	0.308	8.236	0.786
20s	0.247	6.346	0.643
500s	0.056	1.862	0.123

4. Conclusions

The M/G/1 single-server queue with Poisson arrivals and a general service-time distribution is a powerful tool for modeling components of telecommunication networks. Its flexibility to represent arbitrary service-time variability makes it especially useful for burst, non-exponential distributions, which occurs in packet processing at network devices. The M/G/1 encapsulates phenomena that are important for the design and analysis of computer systems or communication links. In many real situations the arrival process is less complex than the service mechanism and is assumed to have exponential interarrival times. Service times often correspond to packet transmission durations or file download sizes, which typically have non-exponential distributions. In a real network, the observation time and the accuracy of calculations of metrics related to the performance of the queuing system, such as the mean number of packets in the system and the mean queueing delay, are very important. As the results obtained for two distributions with different coefficients of variation show, the value of this coefficient is correlated with the values of the tested queuing system metrics.

In a real network, the observation time and the accuracy of calculations of metrics related to the performance of the queuing system, such as the mean number of packets in the system and the mean waiting time in the queue, are very important. As the results obtained for two distributions with different coefficients of variation show, the value of this coefficient is correlated with the values of the tested queuing system metrics – the longer time of observation, the greater the convergence to the analytical values. This is particularly evident when analyzing deviations from the analytical values (Table 1 and Table 2). However, this difference is not as noticeable in the case of the mean waiting time in the queue. Although the mean coefficient of variation for the uniform distribution over the entire analyzed range (from 0.1 to 0.9) is 0.432, significantly lower than that for the exponential distribution, Figures 4 and 8 do not differ much from each other, suggesting that, in this case, the coefficient of variation has little impact on the mean waiting times in the queue.

References

[1] Arzuaga, E.; Kaeli, D. An M/G/1 queue model for multiple applications on storage area networks. In Proceedings of the Proceedings for the 11th Workshop on Computer Architecture and Evaluation (2008) using Commercial Workloads CAECW, 2008, Vol. 11, pp. 25–32. https://doi.org/10.1109/ICCPCT61902.2024.10672928.

[2] Hyytiä, E.; Righter, R.; Virtamo, J. Admission Control to M/G/1 Subject to General Class-Specific Admission and Rejection Costs. In Proceedings of the 2020 32nd International Teletraffic Congress (ITC 32), 2020, pp. 123–128. https://doi.org/10.1109/ITC3249928.2020.00023.

[3] Daigle, J.N. Queueing Theory with Applications to Packet Telecommunication; Springer Science + Business Media, Inc.,: Boston, MA, 2005. https://doi.org/10.1007/b99875.

[4] Alouf, S.; Altman, E.; Azad, A.P. M/G/1 queue with repeated inhomogeneous vacations applied to ieee 802.16e power saving. In Proceedings of the Proceedings of the 2008 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, New York, NY, USA, 2008; SIGMETRICS '08, pp. 451–452. https://doi.org/10.1145/1375457.1375516.

[5] Lv, S.L.; Lyu, Y.; Sun, X.C. The M/G/1 queue system with random vacation. Journal of Industrial and Production Engineering 2019, 36, 229–236. https://doi.org/10.1080/21681015.2019.1646326.

[6] Sun, Y.; Zuo, H.; Chen, Y.; Wang, B. Queueing and channel access delay analysis in in-band full-duplex wireless networks. International Journal of Distributed Sensor Networks 2019, 15, 1550147719844458. https://doi.org/10.1177/1550147719844458.

[7] Dieleman, N.; Heidergott, B.; Peng, Y. Data-Driven Fitting of the M/G/1 Queue. In Proceedings of the 2019 16th International Conference on Service Systems and Service Management (ICSSSM), 2019, pp. 1–5. https://doi.org/10.1109/ICSSSM.2019.8887609.

[8] Hanukov, G.; Barron, Y.; Yechiali, U. An M/G/1 Queue with Repeated Orbit While in Service. Mathematics 2024, 12. https://doi.org/10.3390/math12233722.

[9] Nazarov, A.; Sztrik, J.; Kvach, A.; Tóth, Á. Asymptotic Analysis of Finite-Source M/GI/1 Retrial Queueing Systems with Collisions and Server Subject to Breakdowns and Repairs. Methodology and Computing in Applied Probability 2022, 24, 1503–1518. https://doi.org/10.1007/s11009-021-09870-w.

[10] Joshi, P.K.; Gupta, S.; Rajeshwari, K.N. Matrix geometric method for the analysis of M/M/1 model under repair. Advances and Applications in Mathematical Sciences 2021, 20, 1239–1248.

[11] Majid, S.; Manoharan, P. Analysis of the M/M/1 queue with single working vacation and vacation interruption. International Journal of Mathematics Trends and Technology (IJMTT) 2017, 47, 31–39. https://doi.org/10.14445/22315373/IJMTT-V47P505.

[12] Pérez, J.F.; Van Houdt, B. The M/G/1-type Markov chain with restricted transitions and its application to queues with batch arrivals. Probability in the Engineering and Informational Sciences 2011, 25, 487–517. https://doi.org/10.1017/S0269964811000155.

[13] Shortle, J.; H. Brill, P.; J. Fischer, M.; Gross, D.; M. Bevilacqua Masi, D. An Algorithm to Compute the Waiting Time Distribution for the M/G/1 Queue. INFORMS J. Comput. 2004, 16, 152–161. https://doi.org/10.1287/IJOC.1030.0045.