Study on the Queuing Model

Dr. Fouzia Sajjad
Dawood University of Engineering and Technology, Karachi (Pakistan)
Corresponding author Email: fouziaabid2005@gmail.com

Abstract: Queuing are the most frequently encountered problems in everyday life. For example, queue at a cafeteria, library, bank, etc. Common to all of these cases are the arrivals of objects requiring service and the attendant delays when the service mechanism is busy. Waiting lines cannot be eliminated completely, but suitable techniques can be used to reduce the waiting time of an object in the system. A long waiting line may result in loss of customers to an organization. Waiting time can be reduced by providing additional service facilities, but it may result in an increase in the idle time of the service mechanism.

[Sajjad, F.. **Study on the Queuing Model**. *The International Journal of Interpretation, Observation and Analysis*, 2026; Volume 1, Issue 1 (January-March): Page Number: 40-46. **ISSN 2349-0713, Peer-reviewed (online/offline), Refereed, Indexed and International Journal (Since 2013), Global Impact Factor: 6.489**
Keywords: Queuing Model, System Parameter, The M/M/s Impatient model

1. Introduction

Queuing models are used to predict the performance of service systems when there is uncertainty in arrival and service times. In this note we will explain queuing terminology and discuss some simple queuing models. This note uses the terminology and conventions of QMacros, a spreadsheet add-in to analyze queuing systems.

The simplest possible (single stage) queuing systems have the following components: customers, servers, and a waiting area (queue), see figure 1. An arriving customer is placed in the queue until a server is available. To model such a system we need to specify:

the characteristics of the arrival process;
the characteristics of the service process; and
how (in what order) waiting customers are dispatched to available servers.

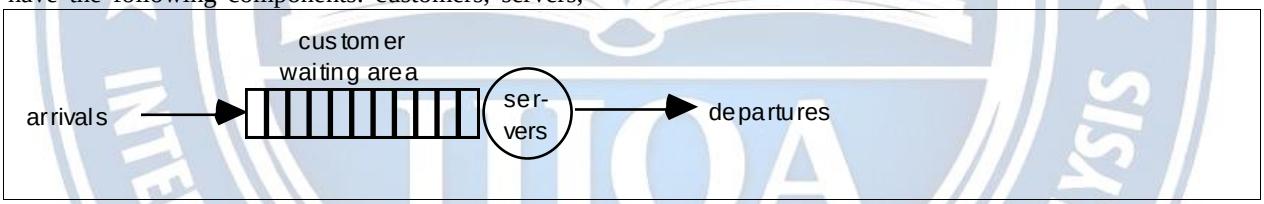

Figure 1. Single stage queuing system.

In this note we will always assume that customers are served in the order in which they arrive in the system (First-Come-First-Served or FCFS). For the characteristics of the arrival and service processes we will make various assumptions, and in general, queuing models are classified according to the specific assumptions made.

In section 2 below we will discuss probability basics, and in section 4 we will go over various performance measures for queuing systems. This material is somewhat technical in nature, but it is necessary for a precise understanding of what queuing models can and cannot do.

In section 3 some basic queuing models will be discussed in detail:

M/M/s – a multi-server model with Poisson arrivals and Exponential service times;

G/G/s – a multi-server model with General arrival process and General distribution of service times;

M/M/s/N – a multi-server model with Poisson arrivals, Exponential service times and a finite facility size so that no more than *N* customers can be present at any time;

M/M/s Impatient – a multi-server model with Poisson arrivals, Exponential service times and Impatient customers prone to balking or reneging.

Finally, section 4 gives an overview of basic queuing system parameters and performance measures.

Paper received: 15.12.2025 **Reviewed: 20.12.2025** **Accepted: 28.12.2025** **Published: 03.01.2026**

2. Probability Basics

2.1 The Exponential Distribution

Clearly, it is impossible to always predict in advance precisely how much service time a customer requires. Hence the service time of a customer is assumed to follow some probability distribution. The exponential distribution is the most frequently used distribution in queuing models. Quite often, we assume that the service time of a customer is independent of the service time of all other customers and follows an exponential distribution. In addition, as we will see in the next subsection, the time between two consecutive arrivals to a queuing system is also frequently assumed to follow an exponential distribution.

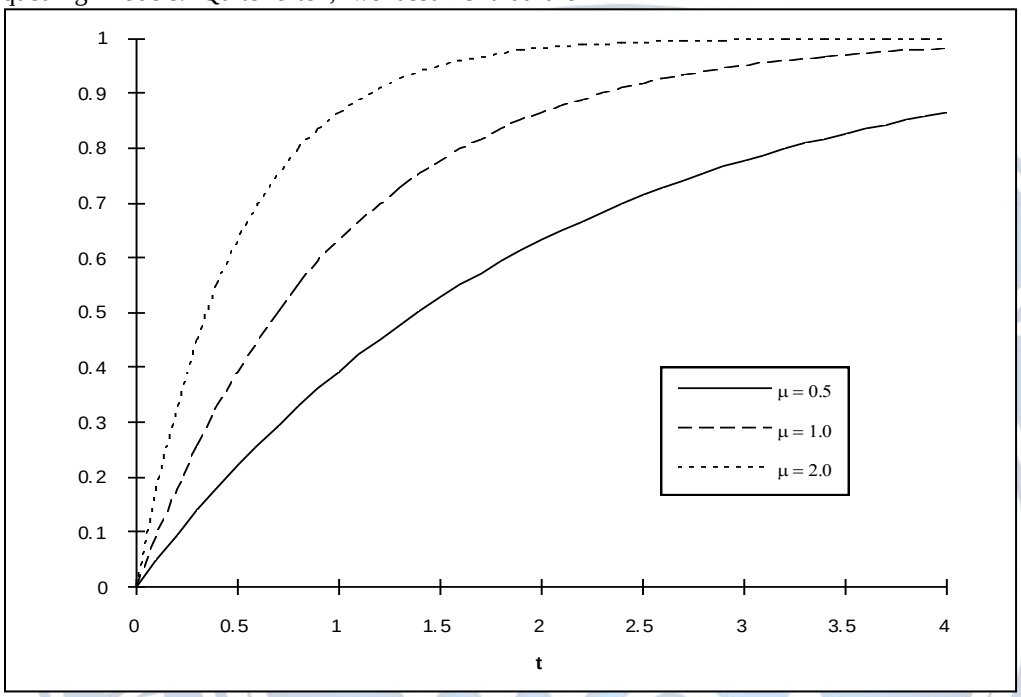

Figure 2. Exponential Cumulative Distribution Functions for □ = 0.5, 1.0, 2.0.

Let S be (the random variable describing) the service time of an arbitrary customer. Then S is said to follow an exponential distribution if for all $t \geq 0$:

$$\Pr\{S \leq t\} = 1 - e^{-\mu t}$$

The parameter is called the **service rate** and gives the average number of customers that a single server can process in the long run if she or he never runs out of customers to serve. This distribution is plotted for several values of in figure 2.

Some useful and/or interesting facts about the exponential distribution are:

- mean service time = E[S] = 1/□,
- the standard deviation of S = □[S] = 1/□,
- the coefficient of variation of S = cv[S] = □[S] / E[S] = 1.

In addition, the exponential distribution is what is called *memoryless*. This means that the distribution of the *remaining* service time of a customer who is currently in service follows the same exponential distribution as the service time of a different customer who starts service now. Formally, for all t, $s \geq 0$ we have

$$\Pr\{S \leq t + s | S > s\} = 1 - e^{-\mu t}.$$

Note that the right hand side of this equation does not depend on s at all.

It is this property that makes the exponential distribution easy to work with in queuing models: it is not necessary to keep track of how long a customer has already been in service since his remaining service time always follows the same distribution. In the commonly used shorthand notation for queuing models, the exponential distribution is represented by an "M" for Memoryless.

2.2 The Poisson Process

Arrivals to a service system usually occur in a random, unpredictable fashion. Even if a good forecast of the total number of arrivals is available,

Paper received: 15.12.2025 Reviewed: 20.12.2025 Accepted: 28.12.2025 Published: 03.01.2026

there is often still considerable uncertainty about the precise timing of arrivals. Consider, e.g., the checkout counter at a large hotel. Management has a good estimate of the number of guests that will check out that day, but there is considerable uncertainty about the number that will check out in the 7-7:30am time interval, and how the checkouts will bunch together during this interval. In a queuing model we therefore need some assumptions about the arrival process. A common assumption in many queuing models is that customers arrive according to a so-called Poisson Process. Formally, this process can be defined as follows.

Let T_i be the arrival time of the i^{th} customer. Arrivals are said to follow a Poisson Process if the successive inter-arrival times $T_1 - T_0$, $T_2 - T_1$, …, $T_i - T_{i-1}$, …, are independent and all follow the same exponential distribution, i.e., for all $t \geq 0$ and $i = 1, 2, \ldots$ we have

$$\Pr\{T_i - T_{i-1} \leq t\} = 1 - e^{-\lambda t}.$$

Here the parameter λ is called the **arrival rate**, and it gives the average or expected number of arrivals per time unit.

An alternative way of looking at the arrival process is to count arrivals. Let $N(t)$ = the number of arrivals up to time t. For every t, $N(t)$ is a random variable that simply counts the number of arrivals. Then the (random!) number of arrivals between time s and time t is given by $N(t) - N(s)$. Of course, $N(t) \geq N(s)$ whenever $t \geq s$, so that $N(t)$ and $N(s)$ are not independent. A typical realization of a Poisson process is depicted in figure 3.

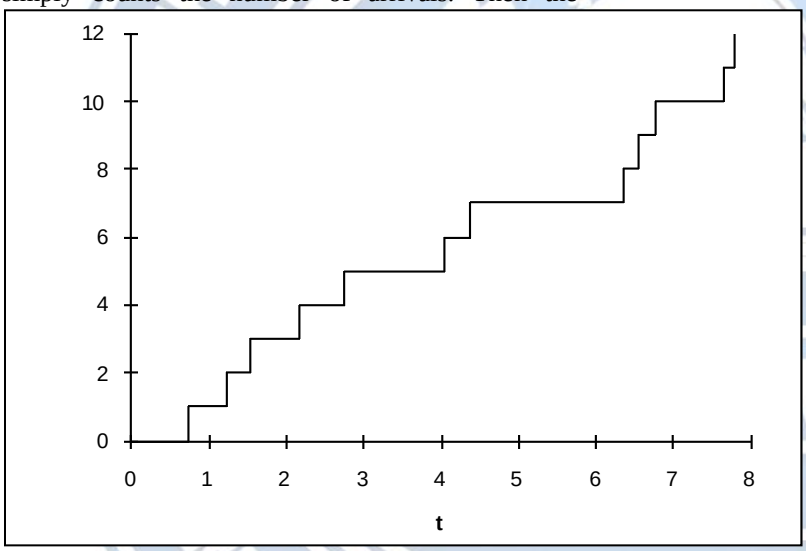

Figure 3. A realization of a Poisson Process.

The name Poisson Process can be explained with the following property: if the arrivals follow a Poisson process, then one can show that for every $s \geq t$, $N(t) - N(s)$ (= the number of arrivals between time s and time t) has a Poisson distribution with mean $\lambda(t-s)$, i.e., for $k = 0, 1, \ldots$:

$$\Pr\{N(t) - N(s) = k\} = \frac{(\lambda(t-s))^k}{k!} e^{\lambda(t-s)}.$$

In addition, the numbers of arrivals in any collection of non-overlapping time intervals are independent.

3. Descriptions of Four Basic Queuing Models

In this section we will describe four simple queuing models.

3.1 The M/M/s model

In this model arrivals follow a Poisson process, the service times are i.i.d. (independent and identically distributed) and follow an exponential distribution. There are s servers ($s \geq 1$). In the M/M/s model there is no balking or reneging, so all arrivals eventually receive service. This model is easy to analyze, and software packages usually give exact values for the various performance measures.

3.2 The G/G/s model

This model is a generalization of the standard M/M/s model. Arrivals follow a General arrival process with i.i.d. inter-arrival times, and the service times are i.i.d. and follow a General distribution. There are s servers ($s \geq 1$). There is no blocking or reneging, so all arrivals eventually receive service. This model is harder to analyze: to obtained exact values for most

Paper received: 15.12.2025 Reviewed: 20.12.2025 Accepted: 28.12.2025 Published: 03.01.2026

performance measures the analysis needs to be tailored to the specific arrival process and service time distribution. In practice, one often fits standard distributions by matching the means and standard deviations of the inter-arrival time distribution and the service time distribution. This results in approximate values for the various performance measures.

3.3 The M/M/s/N model

This model is like the M/M/s model, except that there can be no more than *N* customers present in the system (waiting in queue or in service) at the same time. An arriving customer who finds all *N* available positions occupied is blocked, i.e., this customer does not enter the system and departs without receiving service. This arrival is assumed lost to the system.

3.4 The M/M/s Impatient model

The M/M/s Impatient model is also similar to the M/M/s model, except that now each customer has a patience time. The patience time can of course differ from one customer to another, and is typically assumed to be a random variable. There are two approaches to modeling impatient customers: balking and reneging. Under the balking approach, If a customer's wait in queue (before service starts) would be longer than the patience time, he or she balks, i.e., leaves without entering the system and without receiving service. Of course balking can occur only if the customer can determine upon arrival how long the wait would be. If this is not the case, the reneging approach can be used: a customer may join the queue initially, but when the wait reaches his or her patience time, the customer reneges, i.e., leaves the queue without receiving service. Hence under both approaches the set of customers that are ultimately served is the same. There is a difference in the average number of customers in the queue however: under the balking approach, customers who balk don't spend any time in the queue, under the reneging approach, customers who renege spend their patience time in the queue. Models with impatient customers are not very common in queuing software packages, but exact values for the performance measures are fairly easy to obtain for deterministic or exponential patience times.

4. System Parameters and Performance Measures

In this section we will define some frequently used parameters and performance measures for queuing systems. Throughout, we will make the assumption that the system is in "steady state", i.e., it has operated for a long time with the same values for all the parameters. Since customers arrive in a random fashion and service times are random, a customer's experience in the system (such as the waiting time experienced by the customer) are also random. Basically, the steady state assumption makes sure that as we take many observations of (say) customer waiting time, the frequency distribution converges to a fixed limiting distribution. It is this limiting distribution that the queuing models calculate. We can then interpret a performance measures in two ways: as an expected value for an arbitrarily chosen single customer, or as the average of observations made on a large number of customers (say all those arriving during a given long time interval). It is however wise to realize that each customer will have a different experience in the system.

In the absence of "steady state", the frequency distribution of (say) customer waiting time will not have a fixed limit. This means that the expected waiting time of a customer arriving at 10am may be quite different from the expected waiting time of a customer arriving at 11am. The queuing models that deal with this complication are much more difficult to analyze and are beyond the scope of this note.

WARNING: Time units play a role in measuring many of the performance measures and parameters. It is important that when you use a queuing model the same time unit is used throughout. If you specify an arrival rate in terms of customers/hr and the mean service time in minutes you will get at best meaningless answers and most likely a lot of trouble!

4.1 System Parameters

A summary of parameters is given in table 1.

Paper received: 15.12.2025 Reviewed: 20.12.2025 Accepted: 28.12.2025 Published: 03.01.2026

Parameter	Description	Models
Arrival Rate	Average number of customers arriving per time unit	All models
Service Rate	Average number of customers that a single server processes per time unit if he or she is never idle	All models
Number of Servers	Number of parallel, identical servers in the system	All models
Number of Positions	Total number of customer positions (waiting plus in service) in the system	M/M/s/N
(Mean) Patience Time	Average amount of time that a customer is willing to wait before service starts	M/M/s Impatient
Coefficient of Variation of the Inter-arrival Times	Measures the variability of the time between consecutive arrivals. $\text{cv}(A) = \frac{\text{std.dev.}(A)}{E[A]}$	G/G/s
Coefficient of Variation of the Service Times	Measures the variability of the service time distribution. $\text{cv}(S) = \frac{\text{std.dev.}(S)}{E[S]}$	G/G/s

Table 1. Summary of parameters.

4.2 Performance Measures

4.2.1 Load Factor

The system load factor is a measure of the relative load on the system. Formally, it is defined as

$$\text{Load Factor} = \frac{\text{Amount of work arriving at the system per time unit}}{\text{Amount of work that can be processed by the system per time unit}}.$$

EXAMPLE: In an M/M/s, G/G/s, M/M/s Impatient, or M/M/s/N system with arrival rate □ = 5 customers per hour, service rate of = 2 per hour, and 3 servers, the load factor = 5 / (2 □□3) = 0.8333.

4.2.2 Fraction Not Served

In some models, some unlucky arrivals never receive service, in the M/M/s/N because of blocking, in the M/M/s Impatient model because of balking or reneging. The probability of not receiving service gives the probability that an arbitrarily chosen arrival will be one of the unfortunate ones. Alternatively, it gives the fraction of all arrivals that never receive service.

4.2.3 Thruput

The thruput is the average number of customers that complete service per time unit. This number is always less than or equal to the lesser of the average number of arrivals per time unit and the total processing capacity of the system in a time unit.

4.2.4 Server Utilization

This is the fraction of time that the average server spends serving customers. If customers are distributed randomly or evenly to the servers, it is also equal to the probability that a server is busy at an arbitrary point in time.

4.2.5 Average Number in System

This gives the time average number of customers in the system (counting both customers waiting for service and customers being served). In the M/M/s/N model, blocked customers are not included in this measure. In the M/M/s Impatient model with balking, balking customers are not included in this measure. In the M/M/s Impatient model with reneging, reneging customers are included.

4.2.6. Average Number in Queue

This gives the average number of customers waiting for service (hence customers being served are *not* included). In the M/M/s/N model, blocked customers are not included in this measure. In the M/M/s Impatient model with balking, balking customers are not included in this measure. In the M/M/s Impatient model with reneging, reneging customers are included.

4.2.7 (Average) Time in System

This gives the expected amount of time that an arbitrary customer (who ultimately gets served) spends in the system. Alternatively, it gives the average amount of time that those customers who ultimately get served spend in the system (waiting for service and being served). In the M/M/s/N model, blocked customers are assumed not to spend time in the system. In the M/M/s Impatient model with balking, balking customers are assumed not to spend any time in the system. In the M/M/s Impatient model with reneging, reneging customers are assumed to spend their patience time in the queue, and no time in service.

Paper received: 15.12.2025 Reviewed: 20.12.2025 Accepted: 28.12.2025 Published: 03.01.2026

4.2.8 (Average) Wait in Queue

This gives the expected amount of time that an arbitrary customer spends in the queue before service begins given that the customer is ultimately served. Alternatively, it gives the average amount of time that customers that are ultimately served spend in the queue. In the M/M/s/N model, blocked customers are assumed not to spend time in the system. In the M/M/s Impatient model with balking, balking customers are assumed not to spend any time waiting in the queue. In the M/M/s Impatient model with reneging, reneging customers are assumed to spend their patience time in the queue.

4.2.9 Distribution of the Wait in Queue

The average wait in queue tells only a part of the story. Managers of queuing systems often need to worry about the extremes. In particular, measures of the form Pr{Wait in Queue ≤ *t*} are of interest. Which value(s) of *t* are considered important depends on the situation, of course. For the M/M/s/N model and the M/M/s Impatient model with balking, waiting time probabilities are given for served customers only. For the M/M/s Impatient model with reneging, waiting time probabilities are given for all customers.

EXAMPLE: Assume that 90% of all arrivals get served, and the other 10% are blocked or balk in an M/M/s/N system or an M/M/s Impatient system with balking. We will say Pr{Wait in Queue ≤ 1 minute} = 0.2 if 20% of the 90% who get served wait 1 minute or less. This means that an arriving customer has a probability of 0.1 of being blocked or reneging, a probability of 0.2*0.9 = 0.18 of waiting 1 minute or less before service starts, and a probability of 0.8*0.9 = 0.72 of waiting more than 1 minute before service starts.

EXAMPLE (continued): Assume that 90% of all arrivals get served and the other 10% renege in an M/M/s Impatient system with reneging. We will say Pr{Wait in Queue ≤ 1 minute} = 0.2 if 20% of all customers spend 1 minute or less in the queue before they either get served or they renege. The remaining 80% of all customers spend more than 1 minute in the queue.

4.3 Stability

Even if all the parameters of the queuing model satisfy the steady state assumption, there can be a problem with some of the performance measures. This happens when the Load Factor is 1 or larger in an M/M/s or G/G/s models. The problem is then that there is not enough service capacity to keep pace with the stream of entering customers. Thus (under steady state assumptions for the parameters) the number of customers in the queue will continue to grow without bound. This is not a very realistic model, but the performance measures can still be interpreted, and we will assign them consistent values, i.e., Average Nr. in Queue and Average Nr. in System are infinity, Average Wait in Queue and Average Time in System are infinity, the probability of not waiting is 0, the Fraction Not Served is 0 (even though the average waiting time grows infinitely large, every arriving customer is eventually served!), the probability of waiting less than (any) fixed amount of time is 1, the Thruput is equal to (number of servers) □□(service rate), and Utilization is 1.

4.4 Relationships among Performance Measures

Some simple general relationships among the performance measures and parameters are:

1) (Average Nr. in Queue) + (Average Nr. in Service) = (Average Nr. in System),

2) (Average Nr. in Service) = (Nr. of Servers) □ Utilization,

3) (Time in Queue) + 1 / (Service Rate) = (Time in System),

(Average Service Time) = 1 / (Service Rate),

Thruput = (Arrival Rate) □ (1 – (Fraction Not Served))

A somewhat deeper relationship is *Little's Law*, which comes in several versions:

6) (Average Nr. in System) = Thruput □ (Time in System)

7) (Average Nr. in Queue) = Thruput □ (Time in Queue)

8) (Average Nr. in Service) = (Thruput □ Average Service Time)

Some of these relationships (in particular Little's Law) hold in much more general settings than the simple queuing models discussed here.

References:

[1] P.K. Agrawal, A. Jain and M. Jain, M/M/1 Queueing Model with Working Vacation and Two Type of Server Breakdown,Journal of Physics: Conference Series Vol.1849(1), pp. 012021(2021).

[2] A. Aissani, S. Taleb, T. Kernane, G. Saidi and D. Hamadouche, An M/G/1 retrial queue with working vacation,Advances in Systems Science,,pp.443-452 (2014).

[3] S.I. Ammar, Transient solution of an M/M/1 vacation queue with a waiting server and impatient customers, Journal of the Egyptian Mathematical Society, Vol 25(3),pp.337-342(2017).

[4] D. Arivudainambi, P. Godhandaraman, and P. Rajadurai. Performance analysis of a single server

Paper received: 15.12.2025 Reviewed: 20.12.2025 Accepted: 28.12.2025 Published: 03.01.2026

retrial queue with working vacation Opsearch, Vol 51(3),pp. 434-462 (2014).
[5] Y. Baba, Analysis of a GI/M/1 queue with multiple working vacations, Operations Research Letters, Vol 33(2), pp.201-209 (2005).
[6] Y. Baba, The M[X]/M/1 queue with multiple working vacation,(2012).
[7] A.D. Banik, U.C. Gupta, and S.S Pathak, On the GI/M/1/N queue with multiple working vacations—analytic analysis and computation, Applied Mathematical Modelling, Vol 31(9), pp.1701-1710 (2007).
[8] A.A. Bouchentouf, M. Cherfaoui, and M. Boualem, Performance and economic analysis of a single server feedback queueing model with vacation and impatient customers, Opsearch, Vol 56(1),pp.300-323 (2019).
[9] A.A. Bouchentouf, A. Guendouzib, and A. Kandoucib, Performance and economic study of heterogeneous M/M/2/N feedback queue with working vacation and impatient customers, ProbStat Forum, Vol 12,pp. 15–35 (2019).
[10] S.R. Chakravarthy, and R. Kulshrestha, A queueing model with server breakdowns, repairs, vacations, and backup server, Operations Research Perspectives, Vol 7, pp.100131 (2020).
[11] H.Y. Chen, J.H. Li, and N.S. Tian, The GI/M/1 queue with phase-type working vacations and vacation interruption, Journal of Applied Mathematics and Computing, Vol 30(1),pp. 121-141(2009).
[12] N.H. Do, T. Van Do, and A. Melikov, Equilibrium customer behavior in the M/M/1 retrial queue with working vacations and a constant retrial rate. Operational Research, Vol 20(2), pp.627-646 (2020).
[13] B.T. Doshi, Queueing systems with vacations—a survey, Queueing systems, Vol 1(1), pp.29-66 (1986).
[14] S. Gao, and Z. Liu, An M/G/1 queue with single working vacation and vacation interruption under Bernoulli schedule, Applied Mathematical Modelling, Vol 37(3), pp.1564-1579 (2013).
[15] S. Gao, and Y. Yao, An M[X]/G/1 queue with randomized working vacations and at most J vacations, International Journal of Computer Mathematics, Vol 91(3),pp. 368-383 (2014).
[16] S. Gao, J. Wang, and W.W. Li. An M/G/1 retrial queue with general retrial times, working vacations and vacation interruption, Asia-Pacific Journal of Operational Research, Vol 31(02),pp. 1440006 (2014).

Paper received: 15.12.2025 Reviewed: 20.12.2025 Accepted: 28.12.2025 Published: 03.01.2026