International Journal of Enhanced Research in Educational Development (IJERED)
ISSN: 2320-8708, Vol. 11 Issue 6, Nov. - Dec., 2023, Impact Factor: 7.326

Modeling and Performance Analysis of Server Vacations in Queuing Networks

Arvin[1], Vinod Bhatia[2]

[1,2] Department of Mathematics, Baba Mastnath University, Rohtak, Haryana, India

ABSTRACT

Waiting line dynamics may be studied and optimized in many different real-world contexts, and this is where queueing theory comes in. The M/M/1/N queuing system is a cornerstone of the original queueing concepts. The average number of customers (L), the average time a customer spends in the system (W), and the implications of different parameters like arrival rate, service rate, and system capacity on system behavior are all explored in this study through mathematical analysis, simulation studies, and practical insights. Multiple vacations are expected for the server. The system's performance metrics are then shown. We develop a cost model based on the performance analysis to establish the best possible service fee.

Keywords: Reneging, Vacation, Balk, Queueing, Cost model

INTRODUCTION

Analyzing and optimizing the performance of queuing systems is a critical endeavor in the fields of operations research, computer science, and industrial engineering. Among the myriad queuing models, the M/M/1/N queuing system stands as one of the fundamental and widely studied configurations. This system provides valuable insights into understanding the dynamics of waiting lines, service rates, and system utilization. In this extensive exploration, we delve into the intricacies of the M/M/1/N queuing system, breaking down its key components, elucidating its performance measures, and unraveling the implications of different parameters on its operation.

The M/M/1/N queuing system is a classical and highly simplified model used to describe and analyze waiting lines, often referred to as queues. The notation "M/M/1/N" provides insights into the essential characteristics of this queuing model. The first "M" signifies that the arrival process of customers or entities follows a Poisson distribution, which is memoryless and stochastic, commonly encountered in scenarios such as customer arrivals at a service center or packets arriving at a network router. The second "M" denotes the service process, which also follows a memoryless distribution, typically the exponential distribution. This implies that the time taken to serve a customer or process a request is random and independent of previous service times.

The "1" indicates that there is a single server in the system, implying that only one customer can be served at a time. This single-server configuration aligns with various real-world situations, such as a bank teller, a single-server helpdesk, or a single CPU in a computer system. The "N" represents the system's capacity or the greatest number of people who may be served in a particular period of time. When this threshold is reached, new users are denied access to the system, which is a crucial aspect of many practical applications, such as parking lots with limited spaces or a server with a finite number of processing units.

The M/M/1/N queuing system's performance metrics will make more sense after we've covered the groundwork. Lambda () is a crucial parameter since it reflects the typical rate at which new customers or other entities are added to the system. On the flip side, μ (mu) represents the typical pace at which the server handles requests from new users. By gauging the level of server activity, the utilization factor (ρ) may be utilized as a useful indicator of the system's efficacy ($\rho = \lambda/\mu$). When ρ approaches 1, it means that the server is highly used and almost constantly occupied. On the other hand, when ρ is very near to 0, it indicates that the server is underused and consumers almost never have to wait.
One of the most important performance indicators for an M/M/1/N queueing system is the average number of customers in the system (L), which measures the typical number of customers waiting to be served, including those who are currently receiving service. The total time a consumer spends in the system (W), which includes both waiting and service time, is another key indicator of its efficiency. These KPIs have an immediate effect on wait times and resource use, making them crucial for gauging system efficiency and customer satisfaction.

These efficiency metrics make sense only when applied to an M/M/1/N queuing system, and may be determined

theoretically with the help of queuing theory. Modeling, analyzing, and optimizing queueing systems are all possible thanks to queueing theory, a mathematical framework. Its usefulness extends well beyond contact center staffing optimization to include the design of efficient transportation systems because of the solid groundwork it offers for understanding and forecasting the behavior of queues.

Using Little's Law, a cornerstone of queuing theory, we can determine L and W for the M/M/1/N queuing system. When calculating the average number of consumers in a system, the arrival rate, and the average length of time a customer spends in the system, Little's Law indicates that L = W. Using this simple formula, we can establish a clear correlation between the arrival rate and dwell duration, two crucial performance indicators.

As we explore the performance measures of the M/M/1/N queuing system, we will investigate various scenarios and parameter values, examining how changes in arrival rates, service rates, and system capacity impact the system's behavior. Additionally, we will delve into the concept of system stability, which determines whether the system can handle incoming customers without becoming overwhelmed and experiencing infinite waiting times.

The M/M/1/N queuing system is a fundamental and versatile model for analyzing waiting lines and system performance in a wide range of applications. Its simplicity and mathematical elegance make it a valuable tool for understanding the dynamics of queues and optimizing resource allocation in various real-world settings. In the following sections, we will explore the performance measures of this queuing system in depth, unveiling the mathematical foundations, practical implications, and managerial insights that it offers.

QUEUING SYSTEMS

Queuing systems, often referred to as waiting line systems or simply queues, are fundamental mathematical models used to analyze and understand the dynamics of waiting in various real-world scenarios. These systems are essential for assessing and optimizing the performance of processes involving the flow of entities, such as customers, tasks, data packets, or vehicles, where entities arrive, wait in line, are serviced, and eventually depart. Queuing theory, the mathematical discipline that studies these systems, plays a crucial role in multiple domains due to its ability to provide insights into resource allocation, efficiency, and service quality.

Queuing systems are used when customers, jobs, or data packets arrive at a service center and must wait in line to be processed—also known as waiting line systems or queues—are mathematical models used to examine and evaluate those circumstances. Understanding the key components of queuing systems is crucial for analyzing and optimizing various real-world processes. Let's delve into these components in detail:

Arrival Process (λ - Lambda)
The arrival process represents how entities enter the queuing system. It is characterized by the rate at which entities arrive, denoted as λ (lambda). Lambda represents the typical rate of occurrences per given interval of time. The processes of arrival might take on a variety of distributions, but the Poisson distribution is commonly used in queuing theory due to its memoryless and stochastic nature. Other distributions, such as exponential or normal, may also be used depending on the scenario.

Service Process (μ - Mu)
The service process defines how quickly entities are processed or served by the system. It is characterized by the service rate, denoted as μ (mu), which represents the average number of entities that can be serviced per unit of time. Service times often follow the exponential distribution due to its mathematical convenience. However, other distributions may be used to model service times when appropriate.

Queue
The queue represents the waiting area where entities wait to be serviced when the server(s) are busy. The quantity of entities in the queue at any particular time is known as the queue length. When several entities are waiting in a queue, the discipline rules for how each entity is handled may change. First-Come-First-Served (FCFS), Last-Come-First-Served (LCFS), Priority Queues, etc. are examples of common queue disciplines.

Server(s)
The server(s) perform the actual work or service on the entities. In single-server queuing systems (M/M/1), there is one server. In multi-server systems (M/M/c), there are multiple servers. Each server has a service capacity, μ, which represents the maximum rate at which it can process entities.

Capacity (N)
In several queuing systems, a constraint exists on the maximum number of entities that may be accommodated in the system concurrently, known as the system's capacity, symbolized as N. Once the total number of entities in the system exceeds a certain threshold, denoted as N, there are two possible outcomes for subsequent arrivals: they may either be

rejected and not allowed into the system, or they may be redirected to another system.

Arrival Rate (ρ - Rho)
The arrival rate, ρ (rho), is the ratio of the arrival rate (λ) to the service rate (μ), i.e., $\rho = \lambda/\mu$. It represents the level of utilization or the fraction of time the server is busy. The value of ρ is critical in determining system behavior. When $\rho < 1$, the system is stable and can handle the arrivals without becoming overwhelmed. If $\rho \geq 1$ the system is unstable, and queues will grow indefinitely.

Utilization Factor (U)
The utilization factor, U, is another term for ρ and represents the fraction of time the server is busy. $U = \rho$. U is used to assess how efficiently the server is being utilized. High utilization may lead to longer waiting times and reduced service quality.

Service Time Variability (σ - Sigma):
In some queuing systems, service times can vary. The variability of service times is characterized by the standard deviation, σ (sigma), and measures the spread or dispersion of service times. Higher service time variability can lead to increased waiting times and unpredictability in the system's performance.

These key components collectively define the structure and behavior of queuing systems. Analyzing and optimizing queuing systems involves understanding how changes in these components impact system performance measures such as the average number of entities in the system (L), the average time entities spend in the system (W), and the probability of certain events occurring, such as entities having to wait in the queue. Queuing theory provides the mathematical tools to model and analyze these systems, helping businesses and organizations make informed decisions about resource allocation and process design.

REVIEW OF LITERATURE

Olarewaju, Adeoye&Udokang, Anietie (2020) Managers who interact with clients in waiting areas sometimes struggle to assess the effectiveness of their service center and its associated metrics. The performance indicators include the number of customers serviced at a given moment and the likelihood that n consumers are using the system. The results of applying the M/M/1 models for performance measurement to the banking industry show how these models may be used to evaluate the effectiveness of any firm. Thilaka, B. et al., (2020) We look at a single-server queue with Poisson arrivals and exponential response times, subject to a single-workweek vacation/shutdown/catastrophe policy. Each phase, "active" or "working vacation," has its own set of service hours. If there are users currently logged onto the system, a catastrophic event is considered to have wiped them out, rendering the system inoperable for an arbitrary amount of time. The probability of the shutdown time, the maintenance condition and the system size during the active phase and the working vacation phase has been explicitly expressed. Hanumantha Raoet al. Numerical examples are used to examine the impact of different parameters on the system's performance metrics.

Majid, Shakiretet et al took into account a working vacation policy with an unlimited capacity M/M/1 queue model, where arrivals balk with a probability and may renege due to impatience. As a result of our effort, we now have clear expressions for the queue sizes that occur throughout the server's typical busy time and during its working vacation period. As performance indicators, the expected size of the system, the proportion of client loss as a result of reneging and balking, and the typical wait time for a customer to be served are all generated. Awasthi et al. The queue length and waiting time are derived using stochastic decomposition structures. Numerical examples have been provided to show how changing the parameters can affect the system's efficiency. Seth, Sunny & Som, Bhupender In this study, we create a Markovianqueueing system with encouraged arrivals and a single server's finite capacity. The model is recursively solved in steady state. The appropriate metrics for success are extracted. The model's economic viability is analyzed by creating a cost model. Numerical and arbitrary simulations of the model are investigated. Som et al. "encouraged arrivals" was coined to describe the state of a system following the introduction of price reductions and promotional offerings by businesses. The concept of "encouraged arrivals" is a relatively recent contribution to the study of customer behavior in queues.

JAIN, N.K. et al., (2014) In this study, we create a multi-master, single-network, and negative-balanced (M/M/1/N) feedback queuing system. Customer behavior known as "reverse balking" occurs when a new customer meets a big system size and is more likely to join the system than if the system size is small. Customer behavior like this is common in the investing industry and other fields. In the literature on queueing systems, a "feedback customer" is one who expresses dissatisfaction after receiving only a subpar level of service. The model's steady-state solution is derived, and key performance metrics are obtained. Additionally, a sensitivity analysis of the model's parameters is conducted.

Kumar, Rakesh& Sharma, Sumeet (2012) Recent years have seen extensive use of the idea of reneging by queueing modelers. Potential clients are lost when a consumer reneges. With this drawback in mind, a new queuing model was

created to address the issue of retaining clients who had previously reneged. An M/M/1/N queuing approach for recovering abandoned patrons is presented here. We have calculated certain performance metrics and derived the model's steady-state solution. The estimated system size is affected by the possibility of client retention. The effect of customer retention on the predicted system size was examined through a comparative examination of many queuing models. Lastly, certain models of queuing have been derived as special examples of this concept.

Kumar, Rakesh & Sharma, Sumeet Recently, queueing modelers have made extensive use of the idea of customers backing out of their purchases. Reneging consumers cost businesses money since they don't end up buying from them in the future. A novel queuing model concerned with customer retention has been designed to account for the loss of these consumers as a result of cancellation. This model suggests that in many circumstances, a reneging client may be persuaded to remain in the line until his service is complete by using certain persuasion mechanisms. Thus, a client who cancels their order has a certain chance of being kept in the line (let's call it q) and an equal chance of leaving the queue without being served (p = 1-q). Retaining existing clients is known as customer retention. The inter-arrival and service times in our single-server, finite-capacity, customer-retention queueing system are assumed to follow a negative-exponential distribution. It is assumed that the revocation intervals follow an exponential distribution. We have solved the model and found the steady state solution. Srinivas and dattabanik analyzed Several indicators of performance have been calculated. The model's sensitivity analysis has been completed. Average system size is impacted by retention likelihood, which has been investigated. The average size of the system rises predictably with the retention probability, as shown by the numerical data. We have developed and explored a few special examples of the model. Abdelkader, Yousry& Al-Wohaibi, Maram (2011) The novel performance metrics for Markovianqueueing systems with a single server are the subject of this study. Moments of order statistics are crucial to the calculation of these quantities. We give the anticipated value and the variance of both the maximum (minimum) number of consumers in the system and the minimum (maximal) waiting time. To demonstrate the concept and the viability of the suggested methods, an application to an M/M/1 model is provided.

SYSTEM MODEL

Here, we think about an M/M/1/N queueing system where customers may balk, cancel, or take a break from waiting on servers. The following are the presumptions made by the system model:

a) One at a time, according to a Poisson process with a rate λ, customers enter the system. Upon arrival, a client has a b_n chance of joining the queue if n customers are ahead of him (n = 0, 1,..., N-1) where N is the maximum number of consumers in the system, or 1 - b_n likelihood of rejecting the line if n customers are ahead of him and

$$0 \le b_{n+1} \le b_n < 1,\ 1 \le n \le N-1,$$

$$b_0 = 1,\ \text{and}\ b_n = 0,\ n \ge N$$

b) Once a client has joined the line, they must wait until time T has passed before service can begin. If it hasn't started by then, he'll grow frustrated and leave the line without being served. T is a random variable in this case, with a density function of

$$d(t) = \alpha e^{-\alpha t},\ t \ge 0,\ \alpha > 0$$

where T represents a constant time interval α. Due to the random nature of clients arriving and leaving while waiting for service, the average cancellation rate may be written as (n-i)α. In light of this, we may write down as a function of the typical cancellation rate of customers

$$r(n) = (n - i)\,\alpha,\ i \le n \le N,\ i = 0, 1,$$

$$r(n) = 0,\ n > N$$

c) Customers are accommodated according to the principle of "first come, first served" (FCFS). Once work has begun, it is always finished. It is assumed that the service times follow an exponential distribution with the density function shown below.

$$s(t) = \mu e^{-\mu t},\quad t \ge 0,\ \mu > 0$$

where μ is the service rate.

d) When no users are logged in, the server will take a vacation for a variable amount of time V. The server will instantly begin another vacation if he returns from one and finds no customers waiting for his services. Assuming V follows an exponential distribution, the density function would look like this:

$$v(t) = \eta e - \eta t,\ t \ge 0,\ \eta > 0$$ where η is the vacation rate of a server.

PERFORMANCE MEASURES AND COST MODEL

Performance Measures

The busy probability of the server (PB), the server's likelihood of taking a vacation (PV), the anticipated number of customers in line (E(Nq)), and the anticipated number of consumers in the system may all be deduced from the steady-state probability (E(N)).

$$P_B = \sum_{n=1}^{N} p_1(n), \tag{1}$$

$$P_V = \sum_{n=0}^{N} p_0(n) = 1 - P_B, \tag{2}$$

$$E(N_q) = \sum_{i=0}^{1} \sum_{n=0}^{N} (n-i) p_i(n), \tag{3}$$

$$E(N) = \sum_{n=1}^{N} n p_1(n) + \sum_{n=0}^{N} n p_0(n). \tag{4}$$

Cost Model

We construct a model of predicted costs in which the variable of interest is the service rate. Our goal is to regulate the service charge such that the per-unit operating expenses of the system

C1 ≡ cost per unit time when the server is busy,

C2 ≡ cost per unit time when the server is on vacation,

C3 ≡ cost per unit time when a customer joins in the queue and waits for service,

C4 ≡ cost per unit time when a customer balks or reneges.

Using the above-described parameters for each kind of expenditure, we can calculate the projected total cost per unit time as

$$F(\mu) = C1PB + C2PV + C3E(Nq) + C4L.R$$

The server's expenses are indicated by the first two lines. The customer's wait time has an associated cost, shown by the third item C3E(Nq). The customer loss cost is included in the last line item, C4L.R.

RESULTS AND DISCUSSIONS

To illustrate the effect of the model's parameters on the ideal service rate, the optimal anticipated cost of the system F (), and other performance metrics, we provide some numerical examples in this section. We restrict the system to no more than three users by setting N = 3, $b_n = 1/(n + 1)$, C1 = 15, C2 = 12, C3 = 18, and C4 = 12.

First, we make a decision on the waiting time rate (α= 0.1), the vacation time rate (η = 0.1), and the client arrival rate λ. Table 1 provides an overview of the numerical findings. The information in Table 1 shows that: As rises, there are two primary tendencies that emerge: (i) While its minimal projected cost F, the ideal service rate declines, then climbs somewhat with growing. (μ∗) substantially increases with increasing ; and (ii) the server's PB busy probability and the anticipated number of consumers waiting in line. In contrast to the chance of the server taking a vacation, E(Nq), the expected number of consumers in the system, E(N), and the average rate of customer loss L.R. both increase with growing. This is due to the fact that when the system's size increases, it also client base. Consequently, as PB, E(Nq), and L.R. all rise, the optimum cost rises as well.

Table 1 The case for α = 0.1 and η = 0.1

λ	0.4	0.5	0.6	0.7	0.8	0.9
μ∗	0.2640	0.2588	0.2575	0.2576	0.2601	0.2631
F (μ∗)	33.9795	37.1881	40.0082	42.5360	44.8405	46.9682
PB	0.3728	0.4301	0.4757	0.5133	0.5448	0.5716
PV	0.6271	0.5705	0.5245	0.4864	0.4549	0.4288
E(Nq)	0.9583	1.0682	1.1585	1.2322	1.2951	1.3472
E(N)	1.3309	1.4985	1.6342	1.7461	1.8396	1.9191
L.R.	0.3016	0.3890	0.4779	0.5672	0.6582	0.7495

We then choose α = 0.1, λ = 0.5, and a variety of numbers η. Table 2 provides a summary of the numerical findings. As may be seen in Table 2, (i) as is increased η by a little amount, the lowest estimated cost η the ideal service rate μ∗ rises dramatically. F () declines as increases; (ii) PB, E(Nq), E(N), and L.R. all decrease as rises η; and (iii) PV grows as increases η. This is because as becomes larger, the average server downtime, measured as 1/η, lowers η. Since PB, E(Nq), and L.R. are all dropping, the ideal cost is as well.

Table 2The case for λ = 0.5, α = 0.1

H	0.03	0.05	0.1	0.15	0.2	0.25
μ∗	0.0441	0.1047	0.2592	0.4353	0.6441	0.8972
F (μ∗)	41.8105	40.3542	37.1881	34.6792	32.6312	30.9190
PB	0.7192	0.5796	0.4305	0.3440	0.2820	0.2319
PV	0.2809	0.4202	0.5698	0.6552	0.7179	0.7683
E(Nq)	1.2242	1.1852	1.0682	0.9690	0.8871	0.8181
E(N)	1.9435	1.7651	1.4979	1.3132	1.1686	1.0490
L.R.	0.4679	0.4392	0.3889	0.3505	0.3185	0.2918

In the end, we settle on λ = 0.5, η= 0.1, and modify α. The numerical findings are summarized in Table 3. A look at Table 3 reveals: (i) the best service rate μ∗ shifts somewhat as rises α, but maintains the same minimal projected cost For every given value of, the following is true: (i) α as increases, F (μ∗) reduces, (ii) α as increases, E(Nq) and E(N) decrease, (iii) α as grows, L.R. and PV increase, and (iv) α as increases, PB decreases. This is due to the fact that as α rises, the average waiting time of irate consumers falls. The optimal cost decreases as the average rate of customer attrition (L.R.) increases, while the expected number of customers waiting (E(Nq)) decreases and the chance of the system being busy (PB) decreases.

Table 3 The case for λ = 0.5, η = 0.1

A	0.05	0.1	0.2	0.3	0.4	0.5
μ*	0.2640	0.2590	0.2478	0.2417	0.2492	0.2811
F (μ*)	39.8796	37.1882	33.4390	30.9772	29.2291	27.9161
PB	0.4853	0.4295	0.3563	0.3047	0.2565	0.2049
PV	0.5152	0.5706	0.6439	0.6951	0.7439	0.7955
E(Nq)	1.2205	1.0684	0.8569	0.7195	0.6241	0.5548
E(N)	1.7048	1.4979	1.2131	1.0239	0.8805	0.7596
L.R.	0.3721	0.3885	0.4120	0.4261	0.4362	0.4422

CONCLUSIONS

The M/M/1/N queuing system continues to be a valuable tool for modeling and analyzing real-world waiting line scenarios. Its mathematical elegance, combined with its practical relevance, positions it as an indispensable asset for businesses and organizations striving to provide efficient and responsive services in an increasingly dynamic and competitive landscape. As we move forward, further research in queuing theory will undoubtedly refine our understanding of queuing systems, opening new avenues for innovation and improved system performance.

REFERENCES

[1]. Olarewaju, Adeoye&Udokang, Anietie. (2020). M/M/1 MODEL FOR PERFORMANCE MEASURE IN MANAGERIAL DECISION MAKING.
[2]. Thilaka, B. &Balasubramanian, Poorani&Udayabaskaran, Swaminathan. (2020). Steady state performance measures of an M/M/1 queue with single working vacation subject to catastrophe.
[3]. Hanumantha Rao, Sama&Vemuri, Vasanta& Kumar, K.. (2020). ENCOURAGED OR DISCOURSED ARRIVALS OF AN M=M=1=N QUEUEING SYSTEM WITH MODIFIED RENEGING. Advances in Mathematics: Scientific Journal. 9. 6641-6647. 10.37418/amsj.9.9.21.
[4]. Majid, Shakir&Manoharan, P. & Ashok, A. (2019). Analysis of an M/M/1 Queueing System with Working Vacation and Impatient Customers.
[5]. Awasthi, Bhavtosh & Sharma, Seema. (2018). Performance Analysis of MarkovianQueueing Model in presence of Encouraged Arrivals. 10.13140/RG.2.2.24117.60645.
[6]. Awasthi, Bhavtosh. (2018). Performance Analysis of M/M/1/K Finite Capacity Queueing Model with Reverse Balking and Reverse Reneging. Journal of Computer and Mathematical Sciences. 9. 850-855. 10.29055/jcms/821.
[7]. Seth, Sunny &Som, Bhupender. (2017). An M/M/1/N Queuing System with Encouraged Arrivals. Global Journal of Pure and Applied Mathematics. 13. 3443-3453.
[8]. Som, Bhupender. (2015). An M/M/1/N Queuing system with reverse reneging.
[9]. JAIN, N.K. & Kumar, Rakesh&Som, Bhupender. (2014). An M/M/1/N Queuing System with Reverse Balking. 4. 17-20.
[10]. Som, Bhupender. (2014). An M/M/1/N Queuing System with Reverse Balking. American Journal of Operational Research. 4. 17-20.
[11]. Kumar, Rakesh& Sharma, Sumeet. (2012). M/M/1/N Queuing System with Retention of Reneged Customers. Pakistan Journal of Statistics and Operation Research. 8. 719-735. 10.18187/pjsor.v8i4.408.
[12]. Kumar, Rakesh& Sharma, Sumeet. (2012). M/M/1/N Queuing System with Retention of Reneged Customers. Pakistan Journal of Statistics and Operation Research. 8. 10.1234/pjsor.v8i4.408.
[13]. Abdelkader, Yousry& Al-Wohaibi, Maram. (2011). Computing the Performance Measures in Queueing Models via the Method of Order Statistics. Journal of Applied Mathematics. 790253. 10.1155/2011/790253.
[14]. Srinivas, V. &Rao, Subba& Kale, B.. (2011). Estimation of Measures in M/M/1 Queue. Communications in Statistics—Theory and Methods. 40. 3327-3336. 10.1080/03610926.2010.498653.
[15]. DattaBanik, Abhijit. (2010). Analysis of single working vacation in GI/M/1/N and GI/M/1/ queueing system. International Journal of Operational Research. 7. 10.1504/IJOR.2010.032111.