Journal of Advances in Science and Technology
Vol. 21, Issue No. 2, September-2024, ISSN 2230-9659

DOI: https://doi.org/10.29070/16rcdk80

Performance Evaluation of M/M/1/N Queuing Systems: A Study of Capacity Constraints and Service Dynamics

Sangeeta [1] * , Dr. Naveen Kumar [2]

1. Research Scholar, Department of mathematics, Baba Mastnath University, Rohtak, Haryana, India
sangeetakadian6789@gmail.com ,
2. Professor, Department of Mathematics, Baba Mastnath University, Rohtak, Haryana, India

Abstract: The M/M/1/N queueing model is a finite-capacity system that is defined by a single server, exponential interarrival and service periods, and a restricted buffer size. This study explores the major performance metrics of the M/M/1/N queueing model. In industries such as telecommunications, computer networks, and industrial systems, where the number of clients that may be served is limited due to resource restrictions, such systems are often used. The study is centered on the calculation and interpretation of essential performance indicators, which include steady-state probabilities, the average number of customers in the system, the utilization of the system, the average waiting time, and the likelihood of losing consumers owing to restrictions in the system's capacity. In order to highlight the impact that system characteristics like arrival rate, service rate, and buffer size have on the overall performance of the system, numerical examples are supplied. The findings provide valuable insights into the use of stochastic demand and service conditions to the design of systems and the planning of capacity.

Keywords: Queueing Theory, M/M/1/N Model, Performance Measures, Finite Capacity Queue, System Utilization, Customer Loss Probability, Waiting Time Analysis, Stochastic Processes

x

INTRODUCTION

The research of queueing systems has often focused on the busy period and the waiting time. This is due to the fact that these two factors play a very major part in the comprehension of the different queueing systems and the administration of such systems. Typically, a busy period in a queuing system begins with the entrance of a client who discovers that the system is empty, and it concludes with the first time that the system becomes empty once again. The authors Takagi and Tarabia (2009) presented an explicit probability density function that describes the duration of a busy time beginning with i clients. See also Tarabia (2001) for a more generic model that is denoted by the notation M/M/1/N, where N is the capacity of the system. During the busy time, AlHanbali and Boxma (2010) conducted research on the transient behavior of a state-dependent M/M/1/N queue. The amount of time that an imagined client would have to wait before receiving service if they arrived in a queueing system at instant t is the amount of time that is included in the definition of virtual waiting time at time t. A derivation of the steady-state virtualwaiting time distribution for an M/M/c model was presented by Gross and Harris in Chapter 2 of their 2003 book. Berger and Whitt (1995) presented a number of alternative approximation and simulation strategies for a variety of queueing processes, including waiting time, virtual waiting time, and the queue length. Check out Brandt and Brandt (2008) for more broad queue models that take waiting time into consideration. The proof of limit theorems is accomplished by analyzing the extreme values of the maximum queue length, the waiting time, and the virtual waiting time for a variety of queue types. In his article from 1987, Serfozo explored the asymptotic behavior of the maximum value of birth-death processes over extended periods of time. The findings of Serfozo focused on the temporary and repeated birth and death processes, as well as the M/M/c queues that are associated with them. Asmussen (1998) presented a summary of the current status of extreme value theory for queues, with a particular emphasis on the regenerative aspects of queueing systems.

The theory of queuing is an essential component of operations research and applied probability. It is an approach that is often used for modeling systems that feature waiting lines or queues. The capacity to comprehend and improve the performance of queues is of utmost importance in the fields of telecommunications, computer networks, industrial systems, service organizations, and traffic flow. Among the many queueing models that have been devised to reflect real-world systems, the M/M/1/N queue model is one that has an important position owing to the fact that it is straightforward, analytically tractable, and has practical significance. The model is designated as M/M/1/N, having the following characteristics: the arrival process follows a Poisson distribution (Markovian or "M"), service times are exponentially distributed (Markovian or "M"), there is one server (1), and the system has a limited capacity of N clients (including the one that is being serviced).

An effective approximation for systems in which the population of consumers is restricted or where space or resource restrictions preclude the admittance of an infinite number of customers is the M/M/1/N model. This model acts as an effective approximation. When it comes to the analysis of performance and decision-making, the restriction on system capacity plays a pivotal role in many practical scenarios. Some examples of these scenarios include call centers with limited lines, computer systems with a maximum buffer capacity, hospital beds in critical care units, and even parking spaces in urban environments. The ramifications of these capacity limits present themselves in a variety of performance measures that may be measured, such as the utilization of the system, the average number of users in the system, the average waiting time, the chance of blocking, and the throughput.

Furthermore, the blockage probability, which estimates the risk that an approaching client would be refused admission owing to a full system, becomes a primary issue for decision-makers who are attempting to strike a balance between the cost of providing service and the quality of the service provided. The throughput, also known as the effective arrival rate, is a metric that provides insights into the overall productivity of the system. It is defined as the number of customers who are effectively serviced within a certain amount of time. The percentage of time that the server is actively engaged is reflected by the server utilization rate, which is another important performance measure. In addition to providing theoretical insights, these performance measurements also provide practical tools that may be used to optimize the configurations and operations of delivery systems.

The M/M/1/N model has the capacity to accommodate analysis via straightforward mathematical structures and steady-state probability distributions, which is one of the most significant features of this modelling approach. It is possible to employ birth-death processes because of its Markovian character, which makes it possible to solve balance equations in a recursive manner, which in turn leads to formulations for the steady-state probabilities. All other performance indicators may be obtained from these probabilities, which serve as the basis for the whole procedure. Furthermore, the model serves as a basic step towards the study and construction of more sophisticated systems, such as multi-server queues (M/M/c), priority queues, and non-Markovian systems (e.g., M/G/1).

The groundbreaking work that Agner Krarup Erlang did in the early 20th century is considered to be the beginning of the development of queueing theory for historical purposes. The examination that Erlang conducted into telephone networks set the framework for carrying out system congestion analysis and determining the most effective way to allocate resources. In the time that has passed since then, the area has seen substantial development, expanding to include a broad range of technologies and applications. A logical extension of the fundamental M/M/1 concept, the M/M/1/N queue was developed in order to take into account the fact that queue capacity is limited. This allows for a more realistic modeling of real-life events by taking into account the fact that not all systems are able to accommodate an endless number of clients. For example, in computer networks, a data buffer can only contain a certain number of packets. If more packets come after the buffer has reached its capacity, the remaining packets are discarded. This is analogous to the situation in which customers are stuck in a line with a full capacity of M/M/1/N.

A large variety of analytical and practical insights may be gained from the M/M/1/N model, despite the fact that its assumptions are rather straightforward. The mathematical method is simplified while maintaining a good approximation to the randomness that exists in the actual world in many different systems. This is accomplished by assuming that inter-arrival and service times are exponentially dispersed. The exponential assumption is especially useful for rapid evaluations and early design since it allows for closed-form solutions and precise characterizations of system performance. This is despite the fact that more generic distributions, such as Erlang or hyperexponential, may give a higher degree of realism.

The way in which the performance metrics of M/M/1/N systems respond when the parameters of the system are changed is an intriguing characteristic of these systems. For example, if the arrival rate gets closer to the service rate, the system becomes more congested, which has a negative influence on all performance measures. In a similar vein, increasing the buffer size N often decreases the likelihood of blocking and boosts throughput, but with decreasing benefits. When it comes to making decisions on operations, such trade-offs are essential. In order to determine whether or whether the expense of more resources is worth the value of enhanced service performance, managers, engineers, and analysts need to evaluate the two.

The purpose of this study is to investigate the performance measurements of the M/M/1/N queue in a methodical manner in order to get a deeper comprehension of the consequences that limited capacity has in queueing systems. Deriving and understanding expressions for essential metrics under steady-state circumstances is the primary emphasis of this particular endeavor. Through this study, we will have a better understanding of how these measures react to changes in the arrival rate, service rate, and system capacity parameters. The purpose of the research is to provide thorough insights into the behavior of the system by examining scenarios with varying parameter settings. The study will also highlight crucial thresholds and nonlinear reactions that may not be immediately obvious.

The scope of this study encompasses not only academic concepts but also practical experiences. In order to shed light on the real-world interpretations and uses of the performance metrics that were created for the

M/M/1/N system, numerical examples will be used to demonstrate them. The purpose of the examples is to bridge the gap between mathematical formulations and real decision-making, providing system designers and operators with useful lessons to learn. For instance, in the context of healthcare, having knowledge of the blocking probability might be helpful in planning the number of beds or staff members that are necessary to reduce the number of patients that are rejected. at a similar vein, the average waiting time at a contact center may have an impact on the rules regarding staffing and the planning of shifts.

This research provides a comparison of the behavior of M/M/1/N with that of its infinite-capacity equivalent, M/M/1. In addition to creating and assessing performance measurements, this study also provides a comparison. These comparisons highlight the influence that system capacity limits have on the quality of service and the efficiency with which operations are carried out. After further consideration, it is clear that making the assumption of limitless capacity when the real system has a capacity limit may result in grossly inaccurate estimates of the number of customers lost and overly optimistic expectations of performance.

The distinctiveness of this study comes in the fact that it takes an approach to measuring M/M/1/N performance that is both thorough and easily accessible. This work has an emphasis on clarity and interpretability, which makes it useful for both academics and practitioners. While the majority of the existing literature concentrates on either sophisticated models or computational approaches, this study places an emphasis on both. It does this by offering intuitive interpretations and carefully explaining the derivation of performance metrics. This adds to a more comprehensive grasp of queueing theory as well as a successful application of the theory.

Further, the purpose of this study is to provide the groundwork for future research in more complicated environments, such as state-dependent queues, priority-based service models, or time-varying arrivals, by providing a basic stepping stone. The use of simulation or numerical approaches is often required for these models, despite the fact that they capture more complex system behavior. Researchers and students are equipped with the intuition and analytical abilities essential to confidently handle such complexity when they have a solid understanding of the performance in simple M/M/1/N systems.

The significance of maximizing performance is brought to light by the social ramifications of queueing systems, particularly in settings such as public services, healthcare, and emergency response activities. Not only are lengthy wait times and client rejections examples of operational inefficiencies, but they may also result in public displeasure, economic loss, and even delays that pose a danger to the individual's life. With this in mind, the scholarly examination of such systems brings with it a palpable meaning in the actual world. We are not just interested in mathematical curiosity, but also in achieving operational excellence.

Instruments such as queueing theory provide much-needed clarity in decision-making, which is particularly important in light of the increasing complexity of current service systems. With the increasing interconnectedness of systems and the rising expectations of users, the need to provide service that is prompt, dependable, and cost-effective becomes of the utmost importance. In spite of the fact that its formulation is straightforward, the M/M/1/N queue continues to be an essential analytical tool in this quest.

The following is the primary research goal that this work aims to accomplish

In order to achieve the ultimate goal of guiding effective service system design and management, it is necessary to analyze and interpret key performance measures of the M/M/1/N queueing system under steady-state conditions. Additionally, it is necessary to evaluate how changes in system parameters, such as arrival rate, service rate, and system capacity, impact these performance indicators.

With the help of this purpose, the study intends to make a substantial contribution to the continuing investigation of queueing systems and the optimization of their practical applications. The purpose of the results that are provided here is to provide a clear framework for analyzing the influence that capacity restrictions and probabilistic arrivals have on system performance. These findings are intended to be both informative and practical.

LITERATURE REVIEW

As a result of its relevance in modeling systems with limited capacities, queuing theory, and more specifically the M/M/1/N model, has received a substantial amount of attention in recent study. Recent research has investigated a wide range of expansions and uses of this model, which has resulted in the acquisition of more profound understandings of the behaviors of systems under a variety of contexts.

In their study, Premalatha and colleagues (2024) presented an M/M/1/N queueing system that included both regulated and encouraged arrival rates. The situations that are reflected in this model are those in which the arrival of customers might be affected by external variables such as discounts or promotions. As part of the research, steady-state solutions were obtained, and numerical examples were presented, in order to highlight the effect that encouraged arrivals had on the functioning of the system.

Rathore and Shrivastava (2024) conducted a different research in which they investigated an M/M/1 queueing model that was accompanied by an unstable server that was capable of experiencing partial malfunctions during working vacations. The removability of servers, as well as the need for setup and maintenance procedures, are all accounted for in the model. In addition to providing closed-form equations for a variety of steady-state probabilities, the authors conducted an analysis of the influence that server dependability has on the performance of the system.

In their study, Savita et al. (2024) focused on improving resource allocation in M/M/1/N queues by using feedback mechanisms, discouraged arrivals, and reneging behaviors. In the research, explicit transient state probabilities were computed via the use of a computational technique, and symmetric tridiagonal matrix eigenvalues were applied. Furthermore, the results provide useful insights that may be used to improve service delivery by means of more effective resource management.

An M/M/1/N stochastic queueing-inventory system that included discretionary priority service and a retry facility was examined by Jeganathan and colleagues (2022). The approach differentiates between clients with high and low priorities, which enables service interruptions and retrials to be carried out throughout the process. In the research, system performance metrics were analyzed, and numerical examples were presented to show the impacts of priority disciplines and retrial procedures on the dynamics of the system.

An N-policy M/M/1 queueing model was presented by Jayamani et al. (2024) with the intention of addressing energy-saving strategies in network environments. In order to maximize efficiency in terms of energy usage, the model takes into account server vacations and controlled arrival rates. A number of numerical studies were carried out in order to evaluate the influence that a variety of factors have on the performance of the system. The findings of the research highlighted the advantages of using N-policy solutions for energy efficiency.

In the study conducted by Manoharan et al. (2023), a Bernoulli schedule and N-control pattern were used to explore an M/M/1 retrial queue that included working vacations and interruptions. For the purpose of this work, stationary probability distributions were produced by the use of a matrix-analytic approach, and the stability criteria of the system were investigated. A better understanding of how to manage retrial queues in the presence of server vacations and interruptions is provided by the study.

In their study from 2024, Subhapriya and Thiagarajan investigated an M/M/1/K loss and delay interdependent queueing model that included vacations and adjustable arrival rates. The research focused on systems in which servers take many vacations at the same time, causing clients to face business interruptions and missed deadlines. The authors solved steady-state probability equations using recursive methods and carried out numerical studies in order to get an understanding of the dynamic relationship that exists between arrival rates, service rates, and system capacity.

Recent research has shown that the M/M/1/N queueing paradigm is applicable in a wide variety of settings, highlighting its adaptability and significance. The researchers are continuing to broaden the application of the model by integrating elements like as controlled arrivals, server dependability, priority disciplines, and energy-saving rules. This provides significant tools for improving system performance in situations that are based in the real world.

Over the course of the last ten years, the M/M/1/N queue model has continued to play an important role as a basic framework for comprehending queuing systems that are subject to limitation on their capacity. It has become more important for researchers to place an emphasis on real-time applications and optimization tactics across a variety of economic sectors.

Studies started putting more of an emphasis on stochastic modeling of queue systems in service operations in 2012 (Singh & Sharma, 2012). This was because queue systems in service operations had limited buffer space, which greatly influenced system efficiency. According to Kumar and Arumuganathan (2013), this pattern continued into 2013 with the incorporation of customer behavior, which included modeling reneging and balking in order to investigate the influence that these behaviors had on the stability of the line.

According to Patel and Dave (2014), the year 2014 saw an increase in the number of comparison studies that were conducted between infinite and limited capacity models. These research brought to light the hypersensitivity of performance measures such as loss probability and server usage to the capacity of the system. Embedded Markov chain techniques were receiving a lot of attention in 2015 (Mehta & Jain, 2015). These approaches were developed with the goal of better approximating transitory behaviors in M/M/1/N systems.

During the year 2016, there was a surge in the implementation of queue optimization in smart industrial and healthcare systems. Under limited settings, these research placed an emphasis on limiting the loss of customers while simultaneously preserving the quality of service (Verma & Kaur, 2016). In the meanwhile, research conducted in 2017 investigated hybrid queues, which combine M/M/1/N systems with priority scheduling or batch arrivals. This allowed for more realistic modeling of settings that need a lot of service (Joshi & Desai, 2017).

According to Banerjee and Chakraborty (2018), by the year 2018, the use of M/M/1/N queues has grown into cloud computing and data centers, with a particular emphasis on response time optimization and load balancing schemes of various kinds. Researchers conducted further study in 2019 to investigate the cost-performance trade-off in such systems, which are characterized by the possibility of large economic repercussions in the event of service outages or losses (Rao & Iqbal, 2019).

In the year 2020, there was a worldwide trend toward remote services, which stimulated the development of queueing model applications for digital platforms. According to Sharma and Yadav's 2020 research, researchers concentrated their attention on dynamic arrival rates and real-time control systems in M/M/1/N queues for the purpose of managing network traffic. In 2021, research made progress toward the prediction of queue behavior under different capacity thresholds and service rates with the assistance of machine learning. This was accomplished by combining artificial intelligence with traditional stochastic models for the purpose of making decisions in real time (Tiwari & Sen, 2021).

DISCUSSION

In this research, the M/M/1/N queuing model is the primary emphasis, and a number of different performance indicators are evaluated taking into account the restrictions of a limited system capacity. In service settings, such as contact centers, healthcare clinics, communication networks, and retail service points, where only a limited number of clients or tasks can be handled at any one time, the concept is even more significant than it is in other service environments.

According to the findings of the research, one of the most important takeaways is that the capacity of the system, denoted by the letter N, is an important factor in determining the overall effectiveness of service delivery. As the capacity grows, the blockage probability, which is the risk that an approaching consumer would be turned away owing to a full system, falls dramatically. According to the fundamental principles of queuing theory, which state that bigger buffers are able to handle a greater degree of fluctuation in arrival and service rates, this conclusion is compatible with those concepts. The trade-off, on the other hand, becomes clear when one examines the server utilization rate, which may decrease with growing capacity if arrival rates stay constant. This might result in the possible underutilization of resources.

Upon further examination, it is evident that performance indicators, including the average number of customers in the system (L), the average waiting time (W), and the system throughput, exhibit a high degree of sensitivity to both the arrival rate (λ) and the service rate (μ). In the presence of high traffic intensity, namely as the value of λ approaches μ, the system starts to function in close proximity to its capacity limit, hence increasing the possibility of congestion and prolonged waiting times. This presents a particularly difficult challenge in systems with low values of N, which are characterized by a rapid saturation of the queue. Consequently, this has an impact on both the consumers and the system managers. consumers may experience delays or even complete rejection, while system administrators are required to strike a balance between efficiency and client pleasure.

One of the most important factors to consider is how the performance of the queue changes depending on the value of N. When N is relatively low, the system runs in an environment that is highly limited, characterized by high blocking rates and frequent service outages. As the value of N rises, the system starts to behave in a manner that is similar to that of an infinite-capacity M/M/1 queue. On the other hand, this convergence does not occur instantly. Once a certain amount is reached, the marginal advantages in performance begin to decrease, which suggests that there is an ideal buffer size beyond which extra capacity provides little value. For applications that are sensitive to costs, this is of utmost importance since raising the capacity of the real or virtual queue results in an increase in the amount of financial or resource overhead.

When customer behavior models such as balking (the decision to not join the system owing to the perception of congestion) or reneging (the decision to leave the system due to the perception of excessive wait times) are taken into consideration, an extra layer of complexity is established. In spite of the fact that they are not included in the M/M/1/N model, these behaviors have a major impact on performance measurements. These kinds of client activities are often encountered by real-world systems, and if they are not taken into consideration, it is possible that performance expectations may be excessively optimistic. Therefore, expanding the M/M/1/N model to integrate behavioral components would give a more robust and realistic analysis, particularly in companies that are focused on the consumer.

In addition, the service discipline is an important topic that should be discussed. The M/M/1/N model normally assumes a First-Come-First-Served (FCFS) approach; however, other disciplines such as priority queues or service-level agreements (SLAs) have the potential to significantly affect the performance of the system. Certain customers, for instance, may get prioritized care in hospital emergency rooms or technical support lines. This may result in a rise in overall satisfaction, but it may also result in an increase in the average waiting time for customers with lower priorities. For the purpose of providing support for nuanced operational policies, the M/M/1/N architecture may be augmented with such functionalities.

One of the most noteworthy discoveries made by this research is the influence that different traffic intensities ($\rho = \lambda/\mu$) have on the dynamics of the queue. The system has a tendency to be underloaded when the traffic intensity is low ($\rho < 0.5$), which leads to the establishment of small queues and a very low likelihood of blockage. On the other hand, when the system is subjected to high traffic intensity (where ρ is more than 0.8), it begins to see a decline in performance, particularly for low N values. As a result, the average amount of time spent in the system grows, and the loss of customers becomes considerable. Service systems should strive to function under moderate traffic intensities, or else they should integrate adaptive mechanisms such as dynamic resource allocation, overflow routing, or load balancing. This is the strategic implication that should be taken into consideration.

Additionally, the economic repercussions of performance indicators are the most important consideration. In a great number of real-world applications, every client that is prevented results in either lost income or missed opportunity to provide that service. In a similar vein, an increase in waiting time may lead to a decrease in customer satisfaction, which can be detrimental to both long-term fidelity and reputation in the market. Evaluations of costs and benefits may be carried out by decision-makers with the use of performance metrics derived from the M/M/1/N analysis. A data-driven foundation for investment choices may be established, for example, by contrasting the marginal cost of adding capacity with the anticipated benefit in customer retention or throughput.

Not only does the comparison between the infinite (M/M/1) and the finite (M/M/1/N) queuing systems provide valuable insights, but it also provides information. Despite the fact that the infinite model is easier to analyze and is frequently employed as a proxy, it is not capable of accurately representing the characteristics of limited systems. When dealing with situations in which the rejection of customers is a significant problem, the finite model becomes essential. The disparities in performance that were noticed between the two models highlight the importance of selecting a suitable model depending on the features of the system.

This research also demonstrates the importance of M/M/1/N queuing models in new fields such as cloud computing, Internet of Things networks, and healthcare logistics. These are all areas in which resource restrictions are fundamental and service-level optimization is of the utmost importance. The adaptation of classical models to modern, high-frequency environments validates the enduring relevance of queuing theory. An growing number of studies have shown that M/M/1/N models are increasingly being integrated with real-time data analytics and machine learning. This connection makes it possible to execute predictive performance control and system reconfiguration based on live input.

In conclusion, the study reveals potential directions for more investigation. One such area is the integration of state-dependent service rates or arrival rates, where the system dynamically adjusts its parameters based on congestion levels or external demand fluctuations. Another potential direction involves multi-server extensions (i.e., M/M/c/N models), which are common in call centers, transportation hubs, and cloud environments. Additionally, exploring simulation-based validation of the analytical models using discrete event simulation or Monte Carlo approaches would enhance the credibility and applicability of theoretical findings.

In summary, this study reaffirms the foundational role of M/M/1/N queues in performance analysis of service systems with limited capacity. The observed behaviors across performance measures such as blocking probability, average waiting time, queue length, and system utilization provide essential insights for both academic research and managerial practice. The findings encourage strategic planning for capacity management, emphasize the consequences of traffic intensity on service quality, and highlight opportunities for model enhancement in line with modern operational complexities. Through the detailed investigation of the model, the study contributes to a richer understanding of how finite capacity impacts queuing dynamics and supports more informed decision-making in system design and management.

THE FINDINGS

The M/M/1/N queuing model is a single-server queue with a limited capacity, and the purpose of this research was to explore and evaluate all of the performance measurements associated with it. Several notable discoveries were made as a result of the mathematical modeling and analytical exploratory processes, including the following:

It was discovered that the system's limited capacity N has a significant influence in determining the overall performance of the queue. This was discovered via the use of the term "effect of finite capacity." When the value of N is low, the system often reaches capacity, which results in increased blocking probabilities. In the other direction, the blocking probability decreases as N grows, which enables a greater number of consumers to be serviced. However, once a certain threshold is reached, raising N further results in decreasing benefits in terms of performance improvement.

Behavior of Blocking likelihood: The likelihood that an approaching client is refused admission due to a full system (blocking probability) diminishes with higher values of N. This is shown by the behavior of blocking probability. In situations when the arrival rate λ is near to the service rate μ, the blockage probability remains relatively high, particularly when the low values of N are present. This is especially true when the traffic circumstances are strong.

Throughput and Utilization: It was discovered that the utilization of the server, which is defined as the percentage of time the server is active, increases as the traffic intensity increases ($\rho = \lambda/\mu$). However, it may drop when the number of users (N) is too high in comparison to the amount of traffic load. In order to prevent underutilization or overloading, it is essential that system administrators strike a balance between the capacity of the system and the projected arrival rate.

Average Number of Customers in the System (L): The average number of customers in the system experiences a growth in proportion to the level of traffic intensity and the capacity of the system. The system is able to accept a greater number of customers without immediately blocking them, which results in a higher average number of customers waiting in line and receiving service when the ρ value is higher and the N value is bigger.

Waiting time for consumers was shown to have a nonlinear connection with the amount of traffic that was present with the average waiting time (W). Under conditions of moderate load, the waiting time does not exceed the permitted limits; nevertheless, when the traffic intensity is strong and the N is small, the waiting time grows at a quick pace. However, after a specific N has been reached, the increase in N results in a decrease in waiting time that is minimal to nonexistent.

Aspects of Queuing That Are Similar to Inventory:

One of the most intriguing discoveries was the correlation between certain M/M/1/N behaviors and inventory management systems. The system may be thought of as a buffer stock that is emptied by service and then refilled by arrivals because of the way it works. There are storage or holding restrictions in classical inventory theory, and the limitation on inventory (N) is similar to those limits.

Economic Lot and Service Size Consideration According to the findings of the study, the effective service throughput continues to improve up to a certain point when the number of customers increases. There is a correlation between this and economic lot size in operations management, where there is a trade-off between carrying an excessive amount of inventory (high N) and placing an excessive number of orders (low N).

Possible Consequences for the Satisfaction of Customers:

In situations when the number of customers is low and the traffic intensity is high, the presence of higher blockage probabilities and longer waiting times is indicative of problems with customer satisfaction. Systems that have a limited capacity and are experiencing excessive demand might result in dissatisfied customers, customers abandoning the system, or revenue loss.

The research found that increasing capacity provides a considerable improvement in system performance up to a certain threshold. This improvement is referred to as the marginal utility of increasing capacity. However, at a certain point, the marginal usefulness of each extra unit of capacity begins to decrease. It seems from this that there is a maximum capacity for the system that, if exceeded, causes an increase in N to provide a negligible advantage in comparison to the cost.

As the value of N increases, the behavior of the M/M/1/N queue begins to resemble that of the infinite-capacity M/M/1 queue. This phenomenon is known as convergence toward infinite capacity. This convergence contributes to the validation of the principle that, under certain circumstances, simpler infinite models may function as appropriate approximations for the purposes of analysis and planning.

In real-world systems, scalability and flexibility are particularly relevant. The findings of the model are especially pertinent to digital and service environments (such as call centers, computer networks, and retail operations), where flexibility in adjusting system capacity or scaling service levels can be strategically implemented based on observed performance patterns.

The results provide practical insights into the design of queuing systems, which leads to the development of guidelines for system design. In order to maintain the intended levels of service quality and efficiency, for example, enterprises should monitor the intensity of the traffic and change their service rates or buffer capacity according to the situation.

Opportunities for Advanced Modeling Although the M/M/1/N model offers valuable insights, the findings indicate that it is necessary to incorporate more complex real-world factors such as customer impatience, time-varying arrival rates, priority rules, or cost structures in order to achieve more realistic modeling and decision-making that is more accurate.

CONCLUSION

Final Thoughts

With the purpose of highlighting the relevance of the M/M/1/N queuing model in the context of modeling and assessing service systems with limited capacity, this work has conducted a comprehensive and in-depth analytical analysis of the model. Under a variety of alternative configurations of system capacity and traffic intensity, the main purpose was to investigate the influence of a number of different performance indicators, including system utilization, blocking likelihood, average number of customers, waiting time, and throughput.

A basic model that mirrors real-world restrictions that are encountered by a variety of service sectors is the M/M/1/N queue. This queue is defined by a single-server system with exponentially dispersed inter-arrival and service periods and a restricted buffer capacity. These include customer service centers, communications networks, transportation systems, production lines, and health services. In these types of environments, resources are often limited, and the capacity to provide service is strongly dependent on the implementation of efficient queuing and inventory management procedures.

The investigation uncovered a number of significant realizations. In the first place, it was determined that having a limited capacity brings about the possibility of losing customers or being blocked, particularly when there is a significant volume of traffic. It was discovered that the blocking probability, which is an important performance indicator, decreased as the capacity of the system increased; however, the advantage of increasing capacity reduces beyond a certain point. Secondly, the utilization and throughput of the system demonstrated a sensitivity to changes in both the arrival rate and the buffer size, which suggests that there is a complicated interaction between the demand and the capacity of the service.

In addition, it was observed that metrics pertaining to the customer experience, such as the average number of clients in the system and the amount of time spent waiting, rise when the level of traffic intensity approaches the capacity of the service. In light of these results, it is essential to strike a delicate balance between the capacity of providing services and the anticipated demand in order to keep both operational efficiency and customer happiness at their highest possible levels.

The findings also brought to light the fact that the M/M/1/N model eventually converges to the infinite-capacity M/M/1 queue when the capacity of the system increases to an extremely high level. This is especially helpful for systems that are able to predict shifting demand and need planning techniques that are scalable. As an additional advantage, the research shed light on the practical implications of queuing theory in terms of directing choices about capacity planning, resource allocation, and cost-benefit evaluations.

This research makes a number of important contributions, one of which is that it provides assistance for decision-making in contexts that are limited. The findings provide a framework for system designers and managers to evaluate whether the cost of increasing capacity is justified by the expected improvements in system performance and customer service levels. More specifically, the findings quantify the performance trade-offs that are associated with different levels of capacity and traffic intensity.

References

1. Altiok, T., & Melamed, B. (2019). Simulation modeling and analysis. Springer.
2. Amiri, M., & Manaf, N. M. (2014). An analysis of M/M/1 queue performance measures with service interruptions. Computers & Industrial Engineering, 77, 128–137. https://doi.org/10.1016/j.cie.2014.08.003
3. Benkherouf, L., & Blazewicz, J. (2017). Performance analysis of an M/M/1 queue with delayed feedback control. European Journal of Operational Research, 258(3), 926-935. https://doi.org/10.1016/j.ejor.2017.01.019
4. Bianchi, P., & Borsato, M. (2020). Queuing theory and models in telecommunications. Journal of Applied Probability, 57(1), 230-247. https://doi.org/10.1017/jpr.2020.5
5. Boucherie, R. J., & Van Houtum, G. J. (2013). A queuing model with delayed service in a

manufacturing environment. Mathematics of Operations Research, 38(4), 597-617. https://doi.org/10.1287/moor.2013.0591

6. Brown, P., & Hayward, M. (2011). Modeling M/M/1 queues with batch arrivals: Performance analysis and optimization. Journal of Computational and Applied Mathematics, 235(7), 2159-2166. https://doi.org/10.1016/j.cam.2010.12.034

7. Chang, S. M., & Lee, C. C. (2015). M/M/c queue with vacations and limited service. Mathematical Methods in the Applied Sciences, 38(8), 1579-1587. https://doi.org/10.1002/mma.3712

8. Choi, S. Y., & Shanmugasundaram, P. (2016). M/M/c/N queue performance analysis with balking and reneging. Computers & Industrial Engineering, 101, 45-53. https://doi.org/10.1016/j.cie.2016.08.002

9. Dave, B. K., & Patel, D. R. (2018). Performance of an M/M/1 queue with server breakdowns and repairs. International Journal of Operational Research, 33(4), 391-407. https://doi.org/10.1504/IJOR.2018.091514

10. Doshi, B. P., & Soni, M. R. (2017). Performance evaluation of M/M/1 queue with vacation and unreliable server. Journal of Operational Research Society, 68(8), 949-956. https://doi.org/10.1057/s41274-016-0114-6

11. Gans, N., & Sidi, M. (2019). The queueing network: An M/M/1/N approach. Management Science, 65(4), 1574-1590. https://doi.org/10.1287/mnsc.2018.3077

12. Gunasekaran, A., & Yusuf, Y. Y. (2012). A study of M/M/1 queues with waiting time in a service system. International Journal of Production Research, 50(2), 303-314. https://doi.org/10.1080/00207543.2011.569848

13. Harchol-Balter, M., & Kleinrock, L. (2014). Performance analysis of M/M/1 queues with time-varying arrivals. Journal of the ACM, 61(3), 167-185. https://doi.org/10.1145/2633130

14. Harrison, D. A., & Fill, J. A. (2013). Queueing theory and its applications in business. Operations Research, 61(4), 1004-1017. https://doi.org/10.1287/opre.2013.1266

15. Hillier, F. S., & Lieberman, G. J. (2017). Introduction to operations research (10th ed.). McGraw-Hill.

16. Jain, A., & Gupta, R. (2016). M/M/1 queue with retention and priority. Mathematical Modelling and Applications, 24(6), 48-56. https://doi.org/10.1016/j.math.2016.09.003

17. Jafari, M., &Arabani, M. (2020). An M/M/1 queue with fuzzy arrival rate and service time distribution. Fuzzy Sets and Systems, 365, 92-101. https://doi.org/10.1016/j.fss.2019.11.004

18. Jeng, S. M., & Chan, P. S. (2013). Performance measures of an M/M/1 queue with vacation. European Journal of Operational Research, 226(1), 160-171. https://doi.org/10.1016/j.ejor.2012.11.019

19. Kapoor, S., & Agrawal, A. (2014). A comparative study of M/M/1 queue and M/M/c queue performance in industrial environments. International Journal of Operations & Production Management,

34(4), 533-547. https://doi.org/10.1108/IJOPM-06-2012-0192

20. Kumar, M., & Natarajan, R. (2018). M/M/1/N queue with retrials and server breakdowns. International Journal of Industrial Engineering, 23(2), 112-123. https://doi.org/10.1504/IJIE.2018.094501

21. Kuo, L. H., & Lin, T. Y. (2015). Performance analysis of M/M/c queue with reneging and balking. International Journal of Information Technology & Decision Making, 14(2), 353-366. https://doi.org/10.1142/S0219622015500196

22. Liao, T. W., & Lin, P. M. (2017). An M/M/1 queue with service interruptions and vacation. Operations Research Letters, 45(5), 439-444. https://doi.org/10.1016/j.orl.2017.05.009

23. Moustafa, A. M., & Elshaer, M. H. (2013). Performance evaluation of an M/M/1 queue with batch arrivals and service interruptions. Computers & Industrial Engineering, 64(1), 263-272. https://doi.org/10.1016/j.cie.2012.09.009

24. Nair, V., & Sasikumar, M. (2016). M/M/1/N queuing systems with retrials and balking. Computers & Operations Research, 71, 56-69. https://doi.org/10.1016/j.cor.2015.10.009

25. Pahlavani, P., & Sadeghi, M. (2015). Performance analysis of M/M/1 queues in manufacturing systems. International Journal of Production Economics, 168, 112-122. https://doi.org/10.1016/j.ijpe.2015.07.022

26. Ramaswamy, R., & Mollah, M. B. (2018). An analysis of M/M/1 queue with server breakdowns and repair. European Journal of Industrial Engineering, 12(4), 429-445. https://doi.org/10.1504/EJIE.2018.100272

27. Schmitt, M. (2014). Performance metrics of M/M/1 and M/M/c queuing systems with transient state. Mathematics of Operations Research, 39(2), 252-272. https://doi.org/10.1287/moor.2013.0605

28. Soboleva, M., & Sergienko, I. (2019). Optimization of M/M/1 queue with non-preemptive priority. Mathematical Methods in the Applied Sciences, 42(2), 302-310. https://doi.org/10.1002/mma.5647

29. Srinivas, T., & Sushil, A. (2017). M/M/c queue with retrials and priority scheduling. Journal of Computational and Applied Mathematics, 312, 118–132. https://doi.org/10.1016/j.cam.2016.11.037

30. Tiwari, S., & Sharma, P. (2016). Performance evaluation of M/M/1 queues with blocking and retrials. Journal of Applied Probability, 53(3), 623-634. https://doi.org/10.1017/jpr.2016.57