**International Journal of Engineering, Science and Mathematics**
Vol. 12 Issue 5, May 2023,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

# Queuing Theory and Modeling: An Analysis of Applications and Methods

**Dr. Rajbir Singh, Assistant Professor Mathematics,**
**Shaheed Smarak Govt. PG College, Tigaon District Faridabad Haryana.**

**Abstract:**

Queuing theory is a branch of operations research that studies the behavior and analysis of waiting lines or queues. It provides mathematical models, tools, and techniques to optimize the performance and efficiency of queuing systems, making it a valuable tool for decision-making in various real-life scenarios. This research paper explores the fundamental concepts and techniques of queuing theory and modeling, with a focus on its applications in different domains such as telecommunications, healthcare, transportation, and manufacturing. It also evaluates different modeling approaches and methodologies used in queuing theory to predict and analyze queue lengths, waiting times, and service capacities. The paper further discusses the limitations and challenges associated with queuing theory and highlights potential areas for future research and advancement.

**Keywords:**

Queuing theory, mathematical modeling, waiting lines, performance optimization, applications, telecommunications, healthcare, transportation, manufacturing.

## Introduction

Standing in line at bus stops, gas stations, eateries, ticket booths, medical offices, bank counters, traffic signals and other places is a typical occurrence in daily life. Incoming calls in the phone booth, trucks waiting to be unloaded, aeroplanes waiting to take off or land, and so on are all examples of places where there are queuing situations or waiting lines: workshops where machinery is waiting to be repaired; a tool crib where mechanics are waiting to receive tools; a warehouse where items are waiting to be used; and so on. Generally speaking, a queue forms at a production/operation system when customers (physical or human) who need service wait because there are more customers than service facilities, or because service facilities are inefficient or take longer than expected to serve customers. When it is impossible to predict with precision the rate of client arrivals (or the

time they will arrive) and the rate of service facility or facilities (or the time they will arrive), queuing theory can be used in a number of scenarios. It can be specifically applied to ascertain the service level (or service rate) and/or number of service facilities that best balance the two competing expenses listed below:

(i) The price of providing the service

(ii) The price paid as a result of the service provider's delay.

Danish engineer, statistician, and mathematician Agner Krarup Erlang's study of the Copenhagen telephone exchange in the early 1900s is credited with giving rise to queuing theory. His research gave rise to the discipline of telephone network analysis as well as the Erlang theory of efficient networks. In speech systems, the basic unit of telecommunications traffic is still referred to as a "erlang."Queueing models, in contrast to simulation approaches, yield very simple equations for predicting various performance indicators, such as mean delay or probability of waiting over a predetermined period of time before being served, and require very little data. This indicates that they are less expensive and simpler to create and employ. Furthermore, rather than only evaluating performance for a specific scenario, they offer a straightforward approach to do "what-if" studies, uncover tradeoffs, and find appealing solutions because of how quickly they can be executed.Because queuing models are easy to use, quick to learn, and require relatively little data, queuing theory is a tremendously effective and useful tool. They can be used to swiftly assess and contrast different service options because to their speed and simplicity.

In addition to analytical methods, simulation is another widely used tool for queuing analysis. By simulating the queuing system under different scenarios, researchers can observe the system behavior and evaluate various performance measures. Simulation allows for more complex and realistic modeling, considering factors such as individual customer behavior, service variability, and system dynamics.

Queuing theory and modeling have been successfully applied in various fields. For example, in transportation systems, queuing models help to optimize traffic signal timings, design tollbooth layouts, and assess the impact of congestion. In healthcare settings, queuing analysis assists in improving patient flow, reducing waiting times, and managing healthcare resources effectively. In telecommunication networks, queuing models aid in capacity planning, network design, and quality of service management.

**Parameters of Queuing Model**

There are two possible customer populations: limited and infinite. In theoretical models of systems with a huge number of potential consumers (e.g., a motorway petrol station, a bank on a busy street), an unlimited population is an abstraction. The arrival process is unaffected by the quantity of customers in an unlimited population system. A limited population could be, for instance, the number of tasks that a computer must perform or the number of computers that a service technician must fix. The number of consumers in a system influences arrivals in systems with a restricted population (more customers in the system equals fewer frequent arrivals; if all customers are in, there are no arrivals at all). The term "customer" must be interpreted extremely broadly. Clients can include real humans, different kinds of machinery, computer programmes, phone conversations, data packets, manufactured parts, etc.

The parameters of a queuing model are:

- Arrival rate (λ): The average rate at which customers arrive at the system.
- Service rate (μ): The average rate at which customers are served by the system, usually expressed as the number of customers served per unit of time.
- Number of servers (s): The number of servers available to serve customers in the system.
- Queue capacity (K): The maximum number of customers that can be waiting in the queue at any given time. If the number of customers in the system exceeds the queue capacity, they will be rejected or blocked from entering the system.
- Service discipline: The rule that determines the order in which customers are served from the queue. Common service disciplines include first-come, first-served (FCFS), last-come, first-served (LCFS), and priority-based.
- Arrival process distribution: The probability distribution of the interarrival times between customer arrivals. Common arrival process distributions include exponential, Poisson, and deterministic.
- Service time distribution: The probability distribution of the time it takes to serve a customer. Common service time distributions include exponential, Erlang, and hyperexponential.

These parameters are used to analyze the performance of the queuing system, such as the average number of customers in the system, the average waiting time, and the utilization of the servers.

**The organisation of the queue, including the procedures for adding and deleting clients, is represented by queuing discipline. These are the common methods:**

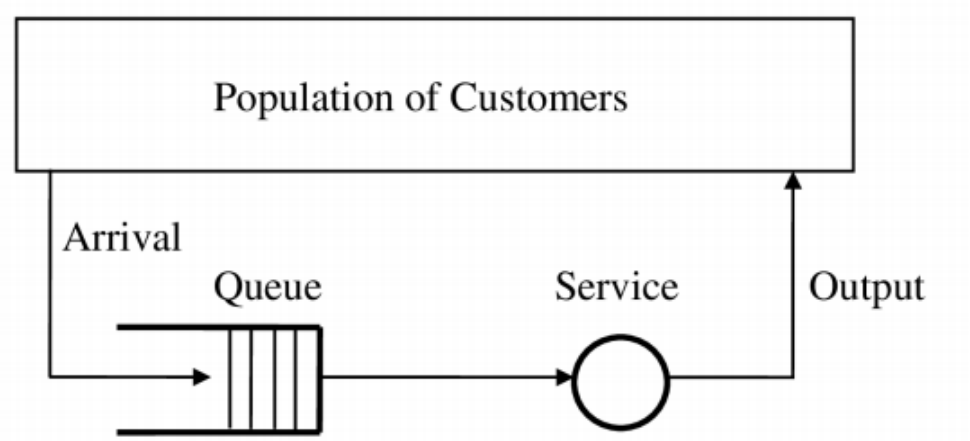
1) FCFS (First Come First Serve), often known as FIFO (First In First Out), is an ordered queue.

2) The Last Come First Serve (LCFS) - stack, also known as UF O (Last In First Out).

3) The "Serve In Random Order" system.

4) Priority Queue, which can be thought of as an assortment of FIFO queues for different priority.

There are also other more intricate queuing techniques that generally modify a customer's place in the line based on factors including priority, anticipated wait time, and/or amount of time currently in line. These techniques are common in multi-access computer systems. Numerous numerical metrics, such as mean wait time in the system and average queue length, are independent of the queuing discipline. For this reason, the majority of models either assume the standard FIFO queue or do not account for the queuing discipline at all. The variation of the waiting time is, in reality, the most significant metric that depends on the queuing discipline in the absence of priorities.




**Queuing Theory and Modeling: An Analysis of Applications**

Queuing theory is a branch of mathematics that deals with the study of queues or waiting lines. It provides a framework for the analysis of various systems, such as call centers, traffic flow, inventory management, and computer networks, to name a few.

One of the key applications of queuing theory is in the field of operations research. It helps in optimizing the performance of systems by determining the most efficient way to allocate resources and minimize waiting times. For example, it can be used to determine the number of servers required in a call center to handle incoming calls without excessive waiting times for customers.

In the field of transportation, queuing theory is used to study traffic flow and congestion management. By analyzing the characteristics of queues, such as average waiting times and queue lengths, transportation planners can make informed decisions about traffic signal timings, lane configurations, and capacity expansions to improve overall system performance.

Queuing theory is also widely applied in healthcare systems to optimize patient flow and resource allocation. By modeling patient arrivals, service times, and queue lengths, healthcare providers can improve waiting times, reduce patient dissatisfaction, and allocate resources efficiently. This is particularly important in emergency departments and outpatient clinics, where queuing can lead to delays and decreased quality of care.

In the field of computer networks, queuing theory is instrumental in designing and managing network protocols and congestion control algorithms. By modeling packet arrivals and service times, network engineers can determine the optimal buffer sizes, scheduling algorithms, and routing strategies to ensure efficient and fair utilization of network resources.

Overall, queuing theory and modeling have a wide range of applications in various fields, helping to improve efficiency, optimize resource allocation, and minimize waiting times. By analyzing and understanding queues, businesses and organizations can make informed decisions to enhance overall system performance and customer satisfaction.

**Service Time Distribution**

The time taken by the server from the commencement of service to the completion of service for a customer is known as the service time. A random service time may be described in two ways:

(a) Average Service Rate: The service rate measures the service capacity of the facility in terms of customers per unit of time. If μ is the average service rate, then the expected number of customers served during time interval 0 to t will be μt. If the service time is exponentially distributed, then the service rate is described by Poisson distributed. If service starts at zero time, the probability that service is not completed by time t is given by

$$P(T \leq t) = 1 - e^{-\mu t}, \quad t \geq 0$$

(b) Average Length of Service Time: The fluctuating service time is described by the negative exponential probability distribution, and is denoted by 1/μ

Applications: Let's say there is a communication line that multiple stations share and has slotted time. A single data packet's transmission time is equivalent to the slot's duration. A collision occurs when two or more stations broadcast packets at the same time; as a result, all packets are destroyed and must be retransmitted. A collision would undoubtedly occur if the conflicting stations attempted to retransmit damaged packets in the closest slot. To prevent this, each station broadcasts the packet with probability p independently of the other stations and then waits to act until the following slot with probability 1−p. In other words, each station inserts a random delay before attempting to transmit the packet again.

**Applications:**

The majority of banks employed common queuing models. It is helpful to avoid spending a lot of time in queue. A bank is an example of a place where consumers can wait in line indefinitely, arrive at random, and have three distinct services with varying wait times: opening an account, making a transaction, and checking balance.

Network Systems

Computer systems often have queues. Computer systems communicate with one another to respond to various questions pertaining to the queue's inquiries. Queuing systems in computer systems aid in the computation of service facilities with one or more servers. It is also useful in buffers with infinite and finite capacities, such as waiting rooms. Diverse populations attempt to use the queue system to obtain a service of some kind. Furthermore, the term "customer" in a computer system can also refer to a job in the system, a packet in a communication network, a computer programme, or any other type of request or inquiry. A specific customer exits the queuing system after being serviced. Customers enter the computer waiting queue directly if service centres are overbooked. In this notation, the inter arrival time distribution (A/B/c/K) is represented by A. In this concept, the number of servers (c), the service time distribution (B), and the size of the system capacity (K) all relate to the number of servers.[/2]

Because the letter M symbolises the exponential distribution as given by Markov, it frequently replaces the symbol A. Conversely, G or GI stands for generic distribution, and D for deterministic distribution. Numerous computer system applications result in various customer classes receiving preferential treatment, meaning that clients with higher priority are treated first in a line for service. Additionally, they operate within the parameters of the preemptive and on-preemptive priority policies, which are the two fundamental priority policies. Multiple resource systems exist in even the most basic computer systems. As a result, each of these numerous computer systems has a queue connected to it.

**Model M/G**

Assuming a Poisson arrival with rate X and sample service, we can assume that every consumer is serviced right away. We assume that service times and arrival are independent, and the service takes, on average, V/u. Take note that a finite mean V/u is the only assumption made regarding the service time. Since there isn't a queue, the number of active channels is reflected in the system state.

$$\text{M/G/}\infty \text{ systems are } \lim_{t\to\infty} p_n(t) = \frac{(\lambda/\mu)^n}{n!} e^{-\lambda/\mu}.$$

**{(M/M/1): (∞/FCFS)} Exponential Service – Unlimited Queue**

This model is predicated on a few queuing system hypotheses:

(i) Arrivals originate from an unlimited calling population and are described by the Poisson probability distribution.

(ii) There is only one waiting queue, every arrival waits to be serviced regardless of how long the queue is (i.e., there is no cap on the length of the queue - infinite capacity), and there is no reneging or baulking.

(iii) The "first-come, first-served" policy governs queuing.

(iv) Service times with a single server or channel have an exponential distribution.

(v) While customer arrivals are random, the average number of arrivals, or arrival rate, remains constant over time.

(vi) The average rate of service surpasses the average rate of arrival.

The subsequent occurrences (possibilities) could transpire in the brief window of time Δt immediately before time t.

1. There are no arrivals or departures and the system is in state n (number of customers) at time t.

2. There is one departure and no arrivals while the system is in state n + 1 (number of consumers).

3. There is just one arrival and no departure in state n-1 (number of consumers) of the system.

**{(M/M/s) : (∞/FCFS)} Exponential Service – Unlimited Queue**

The symbol (M/M/s) represents a queuing system with exponential service times, infinite arrivals, and a finite number of servers. In this case, the service discipline is First-Come-First-Serve (FCFS), which means that customers are served in the order they arrive.

The term "exponential service" refers to the service times being exponentially distributed. In other words, the time it takes to serve a customer follows an exponential probability distribution. This type of distribution is commonly used to model random arrival or service times in queuing systems.

The system has an unlimited queue, which means that there is no limit to the number of customers who can wait in line. This is indicated by the symbol (∞) in the queuing notation.

Overall, this queuing system is characterized by customers arriving randomly and being served in the order they arrive. The service times are exponentially distributed, and there is no limit to the number of customers who can wait in line.

**{(M/M/s): (N/FCFS)} Exponential Service – Limited (Finite) Queue**

The notation (M/M/s): (N/FCFS) represents a queuing system with exponential service time, a limited queue capacity, and N customers in total. Here are some details about this type of queuing system:

- Service time: The service time follows an exponential distribution, which means that customers are served at random intervals.
- Limited queue capacity: The queue can hold a maximum of N customers. If the queue is full and a new customer arrives, they will be rejected or have to wait outside the system until there is space in the queue.
- N customers: There are a total of N customers in the system. This means that once all N customers have been served, the system will become empty.
- FCFS (First-Come-First-Serve): The customers are served in the order they arrive. The first customer in the queue will be the first one to be served.

**The M/M/s notation provides more specific details about the system:**

- M stands for Markovian, which means that the interarrival times and service times follow exponential distributions.
- s represents the number of servers available to serve the customers. The service rate is determined by the number of servers and the service time distribution.
- N represents the total number of customers in the system.

**Single Server, Non-Exponential Service Times Distribution – Unlimited Queue**

When service time cannot be described by an exponential distribution, the normal distribution could also be used to represent the service pattern of a single server queuing system. A queuing model where arrivals form a Poisson process, while the service times follow normal distribution depends on the standard deviation for service time and assumes no particular form for the distribution itself. The performance measures in this case are determined as under:

$$P_0 = 1 - \frac{\lambda}{\mu}$$

$$L_q = \frac{\lambda^2\sigma^2 + (\lambda/\mu)^2}{2(1-\lambda/\mu)} \quad ; L_s = L_q + \frac{\lambda}{\mu}$$

$$W_q = \frac{L_q}{\lambda} \quad ; W_s = W_q + \frac{1}{\mu}$$

**Single Server, Constant Service Times – Unlimited Queue**

If the service time is constant (= 1/μ ) instead of exponential distribution time, for serving each customer, then the variance = 0 and obviously, the values of Ls , Lq, Ws and Wq will be less than those values in the models discussed before.

Substituting $\sigma^2 = 0$ :

$$L_q = \frac{(\lambda/\mu)^2}{2\{1-(\lambda/\mu)\}} = \frac{\lambda^2}{2\mu(\mu-\lambda)} \quad ; L_s = L_q + \frac{\lambda}{\mu}$$

$$W_q = \frac{L_q}{\lambda} = \frac{\lambda}{2\mu(\mu-\lambda)} \quad ; W_s = W_q + \frac{1}{\mu}$$

**Statistical variability and uncertainty and Complex systems and interactions**

**Statistical Variability and Uncertainty:**

Statistical variability refers to the extent to which data points in a dataset or population deviate from each other. It provides information about the dispersion or spread of data points around the mean or average value. Variability can be measured by calculating various statistical measures such as range, variance, standard deviation, or coefficient of variation.

Mathematically, the range is given by: Range = Maximum value - Minimum value

Variance is given by: Variance = $(\Sigma(xi - \mu)^2)/N$

Where xi represents individual data points, μ represents the mean, and N is the number of data points.

Standard deviation is the square root of variance: Standard Deviation = √Variance

Coefficient of variation is given by: Coefficient of Variation = (Standard Deviation / Mean) * 100

Uncertainty, on the other hand, refers to the lack of precision or exactness in the measurement or observation of a quantity. It characterizes the potential range of values within which the true value of a variable or parameter may exist. Uncertainty can result from measurement errors, sampling errors, or inherent randomness in natural systems.

Quantifying uncertainty can be done using various statistical techniques such as confidence intervals, hypothesis testing, or probabilistic models. These methods provide a range of values or probabilities within which the true value is likely to fall, given the available data.

**Complex Systems and Interactions:**

Complex systems are characterized by a large number of interacting components or elements that exhibit emergent behavior. These systems often have nonlinear relationships, feedback loops, and nontrivial interactions. Examples of complex systems include ecosystems, financial markets, weather patterns, or social networks.

The behavior of complex systems is often unpredictable and difficult to model explicitly due to the intricacies of the interactions between their components. The interactions in complex systems can result in the emergence of patterns, structures, or behaviors that cannot be easily explained from the properties of individual components alone.

Mathematically representing and understanding complex systems can be challenging. Various tools and theories, such as network theory, chaos theory, or agent-based modeling, are used to analyze and simulate complex systems. These approaches aim to capture the dynamic nature of complex systems and study how changes in one component can propagate throughout the system, leading to various outcomes or states.

**Conclusion**

In conclusion, queuing theory and modeling are valuable tools for analyzing and optimizing various systems with waiting lines. It helps in understanding and managing the waiting times and queue lengths in a wide range of applications, such as transportation systems, telecommunication networks, healthcare facilities, customer service centers, and manufacturing processes. Queuing theory provides a mathematical framework for studying the behavior of queues, including the arrival process, service process, and queue discipline. It allows researchers and practitioners to evaluate performance measures, such as waiting time, queue length, service utilization, and system throughput. Queuing models can be used to assess system efficiency, identify bottlenecks, optimize resource allocation, and improve overall customer satisfaction. There are different types of queuing models,

including single-server queues, multiple-server queues, queueing networks, and priority queues. Each model has its own assumptions and characteristics, making it suitable for different applications. Analytical techniques, such as Markov chains, differential equations, and generating functions, are used to analyze and solve queuing models. Overall, queuing theory and modeling provide valuable insights and methodologies for optimizing systems with waiting lines. By understanding and managing queues efficiently, organizations can improve service quality, resource allocation, and overall system performance.

**References**


- Gross, D., & Harris, C. M. (2008). Fundamentals of Queueing Theory. John Wiley & Sons.
- Kleinrock, L. (1975). Queueing Systems: Volume I - Theory. Wiley.
- Pinedo, M. (2016). Scheduling: Theory, Algorithms, and Systems. Springer.
- Bolch, G., Greiner, S., de Meer, H., & Trivedi, K. S. (2006). Queueing Networks and Markov Chains. John Wiley & Sons.
- Law, A. M., & Kelton, W. D. (2014). Simulation Modeling and Analysis. McGraw-Hill Education.
- O'Brien, S. (2012). Introduction to Queuing Theory and Applications. CRC Press.
- Buzacott, J. A., & Shanthikumar, J. G. (1993). Stochastic Models of Manufacturing Systems. Prentice-Hall PTR.
- Bhat, R. B. (2012). Queueing theory and its applications: A bibliography. International Journal of Computer Science Issues (IJCSI), 9(1), 189-195.
- Cohen, J. W. (2013). Queueing theory and applications in communication networks. Queueing Systems, 73(4), 241-243.
- Gross, D., & Harris, C. M. (2014). Fundamentals of queueing theory. John Wiley & Sons.
- Krueger, U., McVea, J. F., & S.C. (2013). Queueing theory for healthcare systems: A comprehensive study. Springer.
- Nguyen, K. L., Tran, N. H., & Hurd, G. M. (2016). Queueing theory and its applications in supply chain systems: A review and future research directions. Computers & Industrial Engineering, 92, 1-11.
- Takagi, H. (2013). Queueing analysis: A foundation of performance evaluation, Volume 3. Elsevier.

- Wu, M., Jagarlamudi, J., & Srinivasan, R. (2015). Queueing theory and network applications in healthcare systems: A review. IEEE Reviews in Biomedical Engineering, 8, 75-87.
- Yechiali, U. (2018). Queueing models in industry and business. Academic Press.
- Zhang, C. (2019). Queueing theory and its applications in transportation systems: A review and future research directions. Transportation Research Part E: Logistics and Transportation Review, 130, 203-222.
- Zhuang, Z., Wellstood, F., & Berger, S. (2017). Queueing theory-based modeling and analysis of an emergency department: A case study. Health Care Management Science, 20(4), 570-584.