International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 11 Issue 12, December -2021,
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 7.196| Thomson Reuters ID: L-5236-2015

Queuing Theory and Modeling

Babita
Assistant Professor
Govt. P. G. College for Women Rohtak
e-mail id babita.libra@gmail.com

Abstract:

Queuing theory is a branch of mathematics that is concerned with the study of queues, or waiting lines. It provides mathematical models to analyze and predict various aspects of queues, such as the average waiting time, the average queue length, and the utilization of the system. In order to develop more effective and efficient services and systems, the mathematical field of queuing theory studies congestion and its causes in a process. The arrival process, service process, number of servers, number of system locations, and the number of customers—who may be humans, data packets, cars, or anything else—are all aspects of queueing theory that are examined. Queuing models are frequently used by many organisations, including banks, airlines, telecommunications firms, and police forces, to help manage and allocate resources so that demand may be met quickly and affordably. Despite being employed in hospitals and other healthcare facilities, queuing analysis is not frequently applied in this industry. Queuing analysis is also a useful tool for estimating capacity requirements and managing demand for any system in which the timing of service needs is random. In addition to providing examples of their use, this paper discusses fundamental queuing theory and models as well as some straightforward changes and extensions that are especially helpful in the healthcare industry. Along with the important topic of data requirements, topics like as model selection, model construction, and the interpretation and application of results are also covered.

Keywords:

Population, Queue, Models, Customer, Cost

Introduction:

Queuing theory is widely applied in various fields, including telecommunications, transportation, healthcare, and computer science, where queues are commonly found. For example, in a call center, queuing theory can be used to determine the optimal number of operators required to handle incoming calls, ensuring minimal waiting time for customers. One of the fundamental concepts in queuing theory is the arrival process, which represents the rate at which customers arrive at the system. This process can be modeled using a variety of probability distributions, such as the Poisson distribution. Additionally, queuing theory considers the service process, which represents the rate at which customers are served. This process can also be modeled using different distributions, such as the exponential distribution. The act of standing in queue at bus stops, gas stations, restaurants, ticket booths, doctor's offices, bank counters, traffic lights and other locations is commonplace. Incoming calls wait to mature in the telephone exchange, trucks wait to be unloaded, aeroplanes wait to take off or land, and so on. Queues (waiting queues) are also present in workshops where the machines wait to be repaired; at a tool crib where the mechanics wait to receive tools; in a warehouse where items wait to be used; and so on. A queue typically forms at a production or operation system when either customers (human beings or physical entities) who need service wait because there are more customers than there are service facilities, or service facilities are inefficient or take longer than necessary to serve a customer. The queueing theory can be used in a number of circumstances when it is difficult to forecast the rate (or time) at which consumers will arrive and the rate (or time) at which a facility will provide services. It can be used in particular to estimate the level of service (either the service rate or the quantity of service facilities) that strikes a balance between the two competing expenses listed below:

(i) The price of providing the service

(ii) Costs incurred as a result of service delivery delays.

Agner Krarup Erlang, a Danish engineer, statistician, and mathematician, conducted a study of the Copenhagen telephone exchange in the early 1900s, which is where queuing theory first emerged. His work paved the way for the development of telephone network analysis and the Erlang theory of effective networks. Erlangs are still used today as the basic building block of voice system telecommunications traffic.Contrary to simulation approaches, queueing models just need a small amount of data and produce very straightforward equations for forecasting a variety of performance indicators, such as mean delay or chance of waiting longer than a certain length of time before being

serviced. This indicates that they are simpler to create and more affordable to use. Additionally, rather than only evaluating performance for a specific case, they offer a quick approach to undertake "what-if" evaluations, discover tradeoffs, and uncover appealing alternatives.Because queuing models require little data, are easy to apply, and are quick, queueing theory is a very effective and useful technique. They can be used to quickly assess and contrast a variety of service provision choices because to their simplicity and speed.

Parameters of Queuing Model:

Customers' population can be thought of either restricted or endless. A bank on a busy street or a filling station along the road are two examples of systems with a big potential client base that exploit the abstraction of unlimited population. The number of users in systems with unbounded populations has no impact on the arrival process. A few operations that a computer is to manage (serve), or a few devices that a service technician is to fix, are examples of limited populations. In systems with a small population, the quantity of customers has an impact on arrivals (more customers indicate less frequent arrivals; if everyone is in, there are no arrivals at all). The word "customer" must be interpreted extremely broadly. Customers can include individuals, different kinds of machines, computer programmes, calls, data packets, manufactured goods, etc.

Arrival establishes how users access the system. Customers typically arrive at random times, with random gaps between two consecutive arrivals. Typically, the arrival is represented by an Arrival Pattern, which is a random distribution of intervals. Both single and batch (bulk) arrival models are available.

The queue represents the lineup of clients awaiting service, yet it could also be empty. Customers who are being served are typically not thought of as being in queue. Customers will occasionally literally form a queue to wait for a bank teller. Sometimes a line represents an abstraction, such as when a plane is waiting for a runway to land or a machine is in need of maintenance.

A queue has two crucial characteristics: Maximum Size and Queuing Discipline.

The maximum number of people that can wait in queue is known as the maximum queue size.

When the number of service channels is added to the maximum queue size, the term "system capacity" is used. Although queues are usually capped, some theoretical models presumptively have an infinitely long queue. Some consumers are compelled to give up without being serviced if the queue length is restricted.

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 11 Issue 12, December -2021,
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 7.196| Thomson Reuters ID: L-5236-2015

The queueing discipline (rules for adding and removing consumers to/from the queue) describes how the queue is set up. There are several common methods:

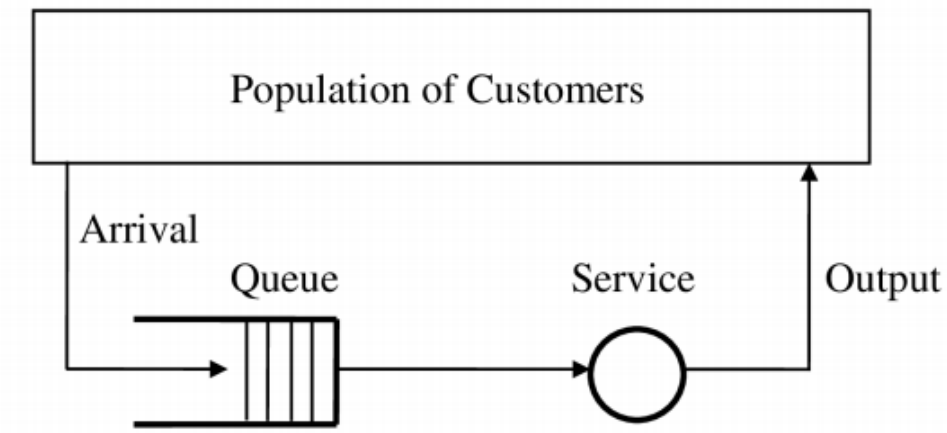
1) FCFS (First Come First Serve), often known as FIFO (First In First Out), is an ordered queue.

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

3) Serve In Random Order (SIRO).

4) A priority queue, which can be thought of as a group of FIFO queues having different priorities.

There are several additional more sophisticated queuing techniques that often adjust the position of the consumer in the queue based on the length of time they have been waiting, the anticipated length of the service, and/or their priority. The majority of computer multi-access systems use these techniques. The discipline of queuing is not a factor in many quantitative indicators (such as average queue length and average time spent in the system). Because of this, the majority of models simply assume the standard FIFO queue or do not consider the queuing discipline at all. The variation of the waiting time is actually the most significant parameter that depends on the queuing discipline if there are no priority.



Queuing Process

The number, kind, and length of queues—single, multiple, or priority queues—are all considered in the queuing process. The design of the service mechanism determines the type of queue, and the size (or length) of the queue is determined by operational factors including available space, governmental regulations, and client attitudes. A service system may occasionally be unable to handle more clients at once than is necessary. Until there is enough room to accommodate additional

consumers, no more customers are allowed to enter. Finite (or limited) source queue describes these kinds of scenarios. Movie theatres, restaurants, and other establishments are examples of finite source lineups. Contrarily, a service system is referred to as having an endless (or unlimited) source queue if it can handle any number of clients at once. There is no limit on how many orders can be placed, for instance, in the sales department where customers' orders are taken. Customers frequently choose not to enter a service system if there are lengthy lines in front of the facility even though there is additional waiting room available if they arrive to a service system and discover huge lines. In these scenarios, client attitude determines how long the queue will be. For instance, most of the time a driver will not stop at a petrol station where there are numerous vehicles waiting and will instead look for service elsewhere [1] when he notices that station has a long queue of cars.

No queue is permitted to build in some finite source queuing systems. For instance, drivers are redirected to another parking place (service facility) if the existing location cannot accommodate more arriving automobiles (customers). A service facility may have a finite number of lines or an endless one. But this has certain benefits as well, such the ability to divide the workforce, the ability for customers to join any queue, and the ability to manage customers' stalling behaviour.

Division of Service Time

The server's service time is the amount of time spent providing service to a client from the moment it begins to the time it is finished. The two possible definitions of a random service time are:

a) The Standard Service Rate With regard to consumers served per unit of time, the service rate gauges the facility's service capability. The anticipated number of clients serviced between time intervals 0 and t will be t if is the average service rate. In the case of an exponential distribution for the service time, a Poisson distribution will best reflect the service rate. When a service is initiated at time zero, the likelihood that it will not be finished by time t is given by

(b) Average Length of Service Time: The changing service time is symbolised by 1/, which is the negative exponential probability distribution that describes it.

Applications:

Take into account a communication line that is shared by numerous stations and has slotted time. One data packet's transmission time is equal to the length of the slot. A collision occurs when two or more stations broadcast packets at the same time, meaning that all packets are lost and must be

retransmitted. A collision will undoubtedly occur if the stations at odds try to retransmit damaged packets in the closest available slot. To prevent this, each station broadcasts the packet independently of the other stations and delays actions until the next slot with probability 1p, or, alternatively, each station introduces a random delay before the next attempt to send the packet.

Banking

Most banks employed common queuing techniques. It is advantageous to avoid spending a lot of time in a queue. A bank is an example of an infinitely long wait, customers that arrive at random times, and three different services that each take a different amount of time to complete: opening an account, processing a transaction, and checking your balance.

Systems for Computers

The use of queues is widespread in computer systems. To respond to various inquiries relevant to the inquiries in the queue, computer systems communicate with one another. Queuing systems aid in computing the service facilities with one and more servers in computer systems. Additionally, it aids in buffers or waiting areas with unlimited and limited capacities. By using the queueing system, various individuals from various communities attempt to obtain some sort of service. The term "customer" can also refer to a computer system programme, a packet in a communication network, a job in a computer system, or any type of query or request in the system. An individual consumer quits the queueing system after receiving service. When service centres are busy, customers enter the computerised waiting system right away. In this notation, A represents the inter arrival time distribution. In this idea, B stands for the distribution of service times, c for the number of servers, and K for the amount of the system's overall capacity, which includes the servers.[2]

M frequently takes the place of the letter A since it stands for the exponential distribution described by Markov. D is for deterministic distribution, while G or GI stands for generic distribution. Many computer system applications give various client classes preferential treatment, which means that service is offered in a line, and the most important customers are served first. Additionally, they operate in accordance with the preemptive priority policy and the non-preemptive priority policy, which are the two fundamental priority policies. Multiple resource systems are included in even the smallest computer systems. As a result, each of these numerous computer systems is connected to various queues.

Delays, Utilization and System Size

Utilisation, which is calculated as the average number of busy servers divided by the total number of servers multiplied by 100, is a crucial metric in queuing theory. Utilisation is frequently viewed from a managerial standpoint as a gauge of production; as a result, a high utilisation rate is regarded as desirable. For instance, in the planning of hospital beds, utilisation is referred to as occupancy level, and historically, an average hospital occupancy level of 85% has been regarded as the minimum threshold for the states to decide if new beds may be required under Certificate of Need (CON) laws. There is a generally held belief in the medical profession that there are too many hospital beds because the real average occupancy rate for nonprofit hospitals has recently been below 70%. The number of hospital beds has declined by about 25% over the last 20 years, mostly as a result of this impression. However, calculating bed capacity based on occupancy levels can lead to extremely long bed waiting times. In all queuing systems, wait times increase as average [3] utilisation level rises.

It's crucial to remember that this relationship is not linear.Two crucial aspects of the system, unpredictability and size, determine exactly where the elbow will be in the curve. Both the length of service periods and the time between arrivals are typically subject to variation, which is typically quantified by the coefficient of variation (CV), or ratio of standard deviation to mean. The elbow will be more to the left the more variable the system is, making delays worse for the same level of utilisation. The ratio of average demand to average service time, known as system size, determines how many servers are required. For a given level of utilisation, delays will be shorter the larger the system, as the elbow will be closer to 100%. Planning or assessing capacity in a service system is significantly impacted by these fundamental queuing concepts in a number of significant ways. First, there must be a strict difference between the average demand and the average total capacity, which is determined by multiplying the number of servers by the speed at which each server can serve consumers. In other words, the system will be "unstable" and the queue will keep expanding unless average utilisation is strictly less than 100%. Although on the surface this fact can seem counter-intuitive, operations professionals have understood it for years. Therefore, a minimum of 6 healthcare professionals are required if an emergency room sees an average of 10 patients per hour and each doctor or physician assistant can treat an average of 2 patients per hour. (Of course, in many situations, arriving guests may decide not to join the queue or they may back out after waiting a while. If so, stability might be feasible even if average demand exceeds average capacity.) Second,

for a given level of utilisation, the delays will be longer the smaller the system. For example, larger hospitals can function at higher utilisation levels than smaller ones while maintaining equal levels of congestion and delays thanks to queuing systems' economies of scale. Finally, the delays will be longer at any given utilisation level the more variable the service duration is (such as length of stay). Therefore, a clinic or doctor's office that focuses on certain services, like mammography or eye tests, will have lower patient wait times than a university-based clinic of similar size and provider utilisation that handles a wide range of illnesses and injuries. When we explore queuing model applications, these characteristics will be more fully highlighted.[4]

Model M/G

We assume a Poisson arrival with rate *X* and sample service where all customers are served immediately. The service takes on average ***V/u*** and we assume that service times and arrival are all independent. Note that the only assumption on the service time is a finite mean ***V/u.*** There is no queue, so the system state represents the number of busy channels.

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

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

This model is based on certain assumptions about the queuing system:

(i) Arrivals are described by Poisson probability distribution and come from an infinite calling population.

(ii) Single waiting line and each arrival waits to be served regardless of the length of the queue (i.e. no limit on queue length – infinite capacity) and that there is no balking or reneging. [5]

(iii) Queue discipline is 'first-come, first-served'.

(iv) Single server or channel and service times follow exponential distribution.

(v) Customers arrival is independent but the arrival rate (average number of arrivals) does not change over time.

(vi) The average service rate is more than the average arrival rate.

The following events (possibilities) may occur during a small interval of time, Δt just before time t.

1. The system is in state n (number of customers) at time t and no arrival and no departure.
2. The system is in state n + 1 (number of customers) and no arrival and one departure.
3. The system is in state n – 1 (number of customers) and one arrival and no departure. [6]

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

In this case instead of a single server, there are multiple but identical servers in parallel to provide service to customers. It is assumed that only one queue is formed and customers are served on a first-come, first-served basis by any of the servers. The service times are distributed exponentially with an average of μ customers per unit of time. If there are n customers in the queuing system at any point in time, then the following two cases may arise:

(i) If $n < s$, (number of customers in the system is less than the number of servers), then there will be no queue. However, $(s – n)$ number of servers will not be busy. The combined service rate will then be $\mu n = n\mu$.

(ii) If $n \geq s$, (number of customers in the system is more than or equal to the number of servers) then all servers will be busy and the maximum number of customers in the queue will be $(n – s)$. The combined service rate will be $\mu n = s\mu$.

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

This model is an extension of Model IV. However, the assumption of an unlimited waiting area for customers is not valid in certain cases:

(i) The parking area once full to its capacity, turns away arriving vehicles.

(ii) In a production facility, parts arriving from a previous production stage to a machine for further processing wait on a conveyer belt, with limited capacity. If the waiting parts fill the belt to its capacity, the production at the previous stage must come to a halt. [7]

In such a situation, the arriving customers turned away may or may not come back. Hence, the cost associated with losing a customer should be taken into consideration, along with the cost per server and the cost of waiting.

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:

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 11 Issue 12, December -2021,
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 7.196| Thomson Reuters ID: L-5236-2015

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

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

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

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 $\sigma^2 = 0$:

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

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

Conclusion:

In conclusion, queuing theory and modeling provide powerful tools for analyzing and optimizing queues in various real-world scenarios. By understanding the dynamics of queues and utilizing mathematical models, organizations can enhance their operations, improve customer satisfaction, and increase efficiency. Because it serves a number of crucial purposes as a process, standing in queue is a common occurrence. When there are little resources available, queues are a fair and necessary method of controlling the flow of customers. If a queuing method is not built to deal with overcapacity, negative results result. Because it offers the tools for queue optimisation and helps to define queue characteristics like average wait time, queuing theory is crucial. Queuing theory in operation research provides guidance for the creation of effective workflow systems from a commercial perspective. Since simulation techniques have both benefits and drawbacks, queueing theory should be taken into account as an option. All models of queuing theory are founded on very rigid presumptions that are infrequently met by actual systems. On the other hand, employing straightforward single queue system formulas yields results quickly. Consider utilising the proper queuing theory model for rough estimates. A simulation research is required for a thorough analysis based on specific empirically gathered data.

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 11 Issue 12, December -2021,
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 7.196| Thomson Reuters ID: L-5236-2015

References:

[1] AT&T's call processing simulator (CAPS) operational concept for inbound call centres was published in Interfaces 24: 6-28 in 1994 by Brigandi, A.J., Dargon,

[2]D.R., Sheehan, and T. Spencer.

Priority assignment in waiting queue problems was addressed by A. Cobbam in 1954 in Operations Research, 2: pp. 70–76.

[3] A queueing linear programming solution to scheduling police cars was described by P.J. Kolesar, K. Rider, T. Crabill, and W. Walker in 1975.

[4] Devalakshmi M. Kumar M.R., "APPLICATIONS OF QUEUEING THEORY," IJCRT, Volume 6(2), pp. 1–7, 2018.

[5] Operation Research Theory and Applications by J.K. Sharma. Publications by Laxmi Foundations of Queueing Theory, International Series in OR & Management Sciences, Prabhu, N.U. (1997).

[6] Allen, A.O., 1978, Probability, statistics, and queueing theory with applications to computer science. The Academic Press in New York.