International Journal of Research Radicals in Multidisciplinary Fields (IJRRMF), ISSN: 2960-043X
Volume 2, Issue 1, January-June, 2023, Available online at: www.researchradicals.com

Concepts & Components of Queueing Theory: A Review

Deepak

Research Scholar, Dept. of Mathematics, Baba Mastnath University, Rohtak

ABSTRACT

The study of the behaviour of waiting queues or queues in a variety of different systems is the focus of the mathematical field known as queueing theory. It provides useful insights into the dynamics and performance of queuing systems, which have applications in a wide variety of industries like telecommunications, transportation, healthcare, and customer service, among others. In this abstract, the fundamental ideas and components of queueing theory are discussed. These concepts and components include entities, arrival processes, service facilities, queue disciplines, and performance metrics. If practitioners have a solid grasp of these essential aspects, they will be able to analyse, optimise, and develop effective queueing systems that will increase the quality of service, reduce the amount of time spent waiting, and boost the overall performance of the system.

Keywords: concepts, components, queueing theory.

INTRODUCTION

Queueing theory is a branch of mathematics that deals with the study of queues or waiting lines. It requires conducting an analysis and modelling the behaviour of systems in which entities enter, wait in queue, and are then serviced or processed by one or more servers. The concept of queuing can be useful in a variety of contexts, including but not limited to telecommunications, computer networks, transportation, healthcare, manufacturing, and customer service. According to the notion of queueing, a queue is a queue of customers or things that are waiting to be served. The goal of this theory is to maximise the efficiency with which resources are used by reducing the typical length of wait times as well as the number of customers waiting in queue. The term "queuing theory" refers to a collection of mathematical models of different types of queuing systems. These models accept the parameters of the parts listed above as inputs, and they output quantitative quantities that describe the performance of the system.

Because of the random character of the processes that are being modelled, queueing theory is fairly demanding, and all of the models are dependent on very strong assumptions, which are not always satisfied in practise. Because simulation is the sole method that can be used to study many different systems, particularly queuing networks, which are not solvable in any way.

Nevertheless, queuing systems are practically very significant because of the common trade-off between the different expenses of providing service and the costs associated with waiting for the service (or leaving the system without being serviced). This trade-off is referred to as the "trade-off between the various costs of providing service and the costs associated with waiting for the service." Service of a high quality that is delivered quickly comes at a premium price, but the expenditures incurred as a result of clients waiting in queue are relatively low. On the other side, lengthy lines can result in significant financial losses due to the fact that consumers or machines, for example, do not produce any revenue while they are waiting in line, or customers leave because of lengthy lines. Therefore, a common challenge is to locate the optimal configuration for the system (for example, the optimal number of servers). Either by using the theory of queuing or by simulating the situation, one can find the answer.

The theory of queuing is applicable to many different fields, including the telecommunications industry, the transportation industry, and the manufacturing industry. Nevertheless, one can use the theory to model the behaviour of a single client as well as the behaviour of a group of customers. Additionally, it can be used to mimic the behaviour of a single resource or of multiple resources working together as a group.

The following is a list of important concepts and components that are crucial to queueing theory:

1. The Arrival Process: The arrival process explains how different entities (such as clients, requests, and so on) are brought into the system. Several different kinds of probability distributions, including the Poisson distribution or the exponential distribution, can be used to model it.

2. The service time distribution is a representation of the amount of time necessary to serve or process an entity. It is also possible to simulate it by making use of several probability distributions, such as the exponential distribution or the normal distribution.

3. Queue Discipline: The rules or methods that are adhered to in order to establish the order in which entities are served from the queue are referred to as the queue's queue discipline. Some examples of this are the first-come-first-served (FCFS) system, the last-come-first-served (LCFS) system, and the priority-based scheduling system.

4. Length of the Queue: The length of the queue refers to the number of entities that are currently waiting in the line at any given point in time. Congestion and waiting time in the system can be better understood with the help of an analysis of the queue length.

5. Models of Queuing: There are many different mathematical models that can be used to depict the many kinds of queuing systems. M/M/1 (Markovian arrivals, Markovian service, single server), M/M/c (single queue, multiple servers), M/G/1 (Markovian arrivals, general service time distribution), and a number of other models are examples of common models.

6. Performance Measures: Queueing theory gives a variety of performance measures that may be used to analyse the behaviour of the system. These performance measures include the average waiting time, the length of the queue, the percentage of the system that is being utilised, the throughput, and the probability of an entity having to wait.

7. Little's Law: Little's Law is a fundamental result in queueing theory that asserts the relationship between the average number of entities in a system, the average arrival rate, and the average amount of time spent by an entity in the system. The average number of entities in a system, the average arrival rate, and the average amount of time spent by an entity in the system.

8. Queuing Networks: Queuing networks are comprised of a number of different queues that are all connected to one another. In these networks, entities travel between different queues based on the specific service requirements they have. Complex systems that consist of several phases or processes can be modelled with their help.

The purpose of queueing theory is to analyse and optimise the performance of queuing systems. This is accomplished by developing a knowledge of the trade-offs that exist between elements such as service capacity, arrival rate, and queue length. This mathematical framework assists in the design of efficient systems, the prediction of performance under a variety of situations, and the making of informed decisions regarding the allocation of resources and the development of processes.

KENDALL'S NOTATION

David George His contributions to the fields of probability, statistical shape analysis, ley lines, and queueing theory earned Kendall a reputation as a prominent statistician and mathematician in England. The usual method for describing and categorising a queueing node is called Kandall's Notation, and it was developed by John Kandall.

The following are characteristics of a queueing model:

a) The procedure of customers checking in at the front desk.
b) The actions of the company's clients.
b) The timings of the services.
d) The service discipline.
g) The capacity of the service.
f) The place where people wait.

For the purpose of describing a queuing system, a unique notation that is known as Kendall's notation is utilised. The form of the notation is described as A/B/c/K.

->A represents the interarrival time distribution. ->B defines the service time distribution. ->c provides the number of servers. ->K provides the size of the system capacity (including the servers).

The letters A and B are represented by the traditional symbols of:

->M for the exponential distribution (Markov is an abbreviation for the word) ->D for the deterministic distribution

->G denotes a general or widespread distribution.

When the total capacity of the system is known to be unlimited (K = 1), one need just use the symbol A/B/c. Examples of relatively frequent queuing systems include M/M/1, M/M/c, M/G/1, and G/M/1.

CONCEPTS OF QUEUEING THEORY

Let's go deeper into the fundamental ideas behind queueing theory, which are as follows:

1. The Arrival Process The arrival process explains how different entities enter the queuing system at different points in time. Either the pace at which individual entities arrive or the amount of time that passes between two consecutive arrivals is used to define it. Poisson processes and exponential inter-arrival periods are examples of often encountered arrival processes.

2. The service time distribution is a representation of the amount of time necessary to serve or process an entity. In most cases, it is defined by the pace at which entities are provided or by the distribution of the amount of time spent on each service. There are a number of possible probability distributions that can be followed by the service time distribution. These distributions can be exponential, normal, or any other distribution that is based on empirical data.

3. Queue Discipline: The rules or techniques that are utilised to define the order in which entities are served from the queue are collectively referred to as queue discipline. The efficiency and equity of the system are both susceptible to the influence of a variety of disciplines. Some popular queue disciplines include:

• First-Come-First-Served (FCFS) refers to a system in which the individual or entity that arrives first receives service first.
• Last-Come-First-Served (LCFS) refers to a system in which the person or thing that comes last receives their service first.
• Priority-based Scheduling: Different entities have varying levels of priority, and the higher-priority entities are served before the ones with a lower priority.

4. Length of the Queue: The length of the queue refers to the number of entities that are currently waiting in the line at a specific point in time. It is an important indicator of both the level of congestion and the level of service that the system provides. Understanding the operation of the system and how it behaves can be aided by doing an analysis of the queue length.

5. Queueing Models: Queueing models are mathematical representations of the many distinct kinds of queueing systems. They describe the characteristics of the arrival process, including the number of servers, the distribution of service times, and the discipline of the queue. The M/M/1 model represents a system with a Markovian (Poisson) arrival process, Markovian (exponential) service times, and a single server. This model is one of the most prevalent types of queueing models.

• M/M/c is a notation that represents a system with Markovian arrivals, Markovian service times, and numerous servers.
• A system is said to be M/G/1 if it has Markovian arrivals, a general service time distribution, and a single server.

6. Performance Measures: The theory of queueing provides a variety of performance measures that can be used to evaluate the operation and effectiveness of a queueing system. The following are some examples of common performance measures:

• The Average Waiting Time refers to the typical amount of time an entity spends waiting in the queue.
The average or maximum number of entities that are waiting in the line is referred to as the "Queue Length."
• System Utilisation refers to the percentage of time that the server(s) are actively processing requests from entities.
• Throughput is defined as the typical number of entities that are processed in a given amount of time.

• The possibility that an entity will have to wait in the queue is referred to as the "probability of waiting."

7. Little's Law: Little's Law is a significant contribution to the field of queueing theory. It asserts that in a stable queueing system, the average number of entities in the system is equal to the average arrival rate multiplied by the average time spent in the system. Additionally, it states that the average time spent in the system is equal to the average number of entities in the system. It is possible to express it mathematically using the equation L = W, where L is the typical number of entities, is the typical arrival rate, and W is the typical amount of time spent in the system.

8. Queuing Networks: The concept of queuing networks entails the existence of many, interconnected queues. Entities will travel between different lines depending on the type of service they want. Modelling complicated systems that consist of several phases or processes typically requires the utilisation of queuing networks. They make it possible to conduct an analysis of the performance of the entire system and to comprehend the interactions that take place between the various queues.

These fundamental ideas help in analysing, optimising, and building effective queueing systems for a variety of real-world applications. Additionally, they serve as the basis for the queueing theory that underpins it. By having a grasp of these principles, a person is able to obtain insights into the behaviour of waiting lines, anticipate the performance of the system, and make decisions that are based on accurate information to improve the situation.

COMPONENTS OF QUEUEING THEORY

Let's explore the basic components of queueing systems in more detail:

1. Entities: The terms "entities" and "entities" refer to the items or entities that enter and move through the queueing system, respectively. They can stand in for clients, service requests, jobs, or any other unit that has a need for assistance. Entities make their way to the system, where they may be required to wait in a line before being eventually served or processed.

2. The Arrival Process: The arrival process defines the manner in which entities enter the queuing system at various points in time. It outlines the sequence as well as the speed at which the entities arrive. It is possible to model the arrival process with probability distributions such as the Poisson distribution or the exponential distribution. It provides assistance in comprehending the inter-arrival times as well as the arrival patterns of entities.

3. Service Facility The resources or servers that are made available to process entities are represented by the service facility. Depending on the way the system is configured, it may include one or more servers as its component parts. It is possible for the service facility to consist of a single server, numerous servers that run in parallel, or multiple servers with varying capabilities or capacities.

4. Service Time Distribution: The service time distribution is the representation of the amount of time needed by the service facility to either serve or process an entity. It gives a description of how long each entity has been providing service. The time distribution of the service can be modelled with probability distributions such as exponential and normal distributions, as well as other empirical distributions that are determined by the nature of the service.

5. Queue: A queue is a waiting area where individuals or organisations wait for service when the facility providing that service is busy. It is one of the primary elements that make up a queuing system. When they arrive, entities are added to the queue, and they are served in a particular order determined by the queue discipline. The length of the queue is an indication of the number of entities that are currently waiting in the line at any one time.

6. Queue Discipline Refers to the Rules or tactics Used to Determine the Order in Which Entities Are Served from the Queue Queue discipline refers to the rules or tactics used to determine the order in which entities are served from the queue. Different queue disciplines can be applied, such as:

   • First-Come-First-Served (FCFS) refers to a system in which the individual or entity that arrives first receives service first.
   • Last-Come-First-Served (LCFS) refers to a system in which the person or thing that comes last receives their service first.

• Priority-based Scheduling: Different entities have varying levels of priority, and the higher-priority entities are served before the ones with a lower priority.

7. Departure is the event that occurs when an entity finishes its service and exits the system. Departure is also known as "exit." Following service, an entity leaves the facility where the service was performed and either continues through the system's processing steps or leaves the system entirely.

8. Performance Measures Performance measures are utilised in order to assess the functioning and productivity of a queuing system. These metrics offer useful insights into the performance of the system and contribute to the process of decision-making. The following are some examples of common performance measures:

• The Average Waiting Time refers to the typical amount of time an entity spends waiting in the queue.
The average or maximum number of entities that are waiting in the line is referred to as the "Queue Length."
• System Utilisation refers to the percentage of time that the server(s) are actively processing requests from entities.
• Throughput is defined as the typical number of entities that are processed in a given amount of time.
• Service Level: Measures of Customer Satisfaction, Such as the Proportion of Customers Who Experience a Predetermined Waiting Time Threshold.

9. For the purposes of analysing, creating, and optimising queueing systems in a variety of areas, having a fundamental understanding of these components is essential. One can obtain a full grasp of the workings of queueing systems by focusing on the entities, arrival process, service facility, queue, queue discipline, departure, and performance metrics. This will allow one to make the most of the time spent waiting in queue.

CONCLUSIONS

The theory of queueing is a useful resource for gaining an understanding of and effectively managing waiting lines in real-world settings. This study gives a complete review of the fundamental components of queuing systems by examining entities, arrival processes, service facilities, queue disciplines, and performance metrics. Specifically, this research focuses on how these aspects interact with one another. The application of the ideas of queueing theory enables practitioners to make educated decisions regarding the allocation of resources, the design of systems, and the improvement of processes. Increasing the satisfaction of customers, boosting operational efficiency, and making the most of available resources are all possible outcomes for businesses that successfully optimise their queueing systems. The principles of queueing theory will continue to play an important role in optimising the performance of various systems and providing better service experiences as technology continues to improve and new challenges emerge.

REFERENCES

[1]. Asmussen, S. R.; Boxma, O. J. (2009). "Editorial introduction". Queueing Systems. 63 (1–4): 1–2. doi:10.1007/s11134-009-9151-8. S2CID 45664707.
[2]. Agner Krarup Erlang (1878-1929) | plus.maths.org". Pass.maths.org.uk. 1997-04-30. Archived from the original on 2008-10-07. Retrieved 2013-04-22.Erlang, Agner Krarup (1909). "The theory of probabilities and telephone conversations" (PDF). Nyt Tidsskrift for Matematik B. 20: 33–39. Archived from the original (PDF) on 2011-10-01.
[3]. Kingman, J. F. C. (2009). "The first Erlang century—and the next". Queueing Systems. 63 (1–4): 3–4. doi:10.1007/s11134-009-9147-4. S2CID 38588726.
[4]. Whittle, P. (2002). "Applied Probability in Great Britain". Operations Research. 50 (1): 227–239. doi:10.1287/opre.50.1.227.17792. JSTOR 3088474.
[5]. Kendall, D.G.:Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain, Ann. Math. Stat. 1953
[6]. Kingman, J. F. C.; Atiyah (October 1961). "The single server queue in heavy traffic". Mathematical Proceedings of the Cambridge Philosophical Society.
[7]. Ramaswami, V. (1988). "A stable recursion for the steady state vector in markov chains of m/g/1 type". Communications in Statistics. Stochastic Models. 4: 183–188. doi:10.1080/15326348808807077.
[8]. Morozov, E. (2017). "Stability analysis of a multiclass retrial system with coupled orbit queues". Proceedings of 14th European Workshop. Lecture Notes in Computer Science. Vol. 17. pp. 85–98. doi:10.1007/978-3-319-66583-2_6. ISBN 978-3-319-66582-5.
[9]. "Simulation and queueing network modeling of single-product production campaigns". Science Direct.

[10]. Manuel, Laguna (2011). Business Process Modeling, Simulation and Design. Pearson Education India. p. 178. ISBN 9788131761359. Retrieved 6 October 2017.
[11]. Harchol-Balter, M. (2012). "Scheduling: Non-Preemptive, Size-Based Policies". Performance Modeling and Design of Computer Systems. pp. 499–507. doi:10.1017/CBO9781139226424.039. ISBN 9781139226424.
[12]. Andrew S. Tanenbaum; Herbert Bos (2015). Modern Operating Systems. Pearson. ISBN 978-0-13-359162-0.
[13]. Harchol-Balter, M. (2012). "Scheduling: Preemptive, Size-Based Policies". Performance Modeling and Design of Computer Systems. pp. 508–517. doi:10.1017/CBO9781139226424.040. ISBN 9781139226424.
[14]. Harchol-Balter, M. (2012). "Scheduling: SRPT and Fairness". Performance Modeling and Design of Computer Systems. pp. 518–530. doi:10.1017/CBO9781139226424.041. ISBN 9781139226424.
[15]. Dimitriou, I. (2019). "A Multiclass Retrial System With Coupled Orbits And Service Interruptions: Verification of Stability Conditions". Proceedings of FRUCT 24. 7: 75–82.
[16]. Sundarapandian, V. (2009). "7. Queueing Theory". Probability, Statistics and Queueing Theory. PHI Learning. ISBN 978-8120338449.
[17]. Lawrence W. Dowdy, Virgilio A.F. Almeida, Daniel A. Menasce. "Performance by Design: Computer Capacity Planning by Example". Archived from the original on 2016-05-06. Retrieved 2009-07-08.
[18]. Schlechter, Kira (March 2, 2009). "Hershey Medical Center to open redesigned emergency room". The Patriot-News. Archived from the original on June 29, 2016. Retrieved March 12, 2009.
[19]. Mayhew, Les; Smith, David (December 2006). Using queuing theory to analyse completion times in accident and emergency departments in the light of the Government 4-hour target. Cass Business School. ISBN 978-1-905752-06-5. Archived from the original on September 7, 2021. Retrieved 2008-05-20.
[20]. Tijms, H.C, Algorithmic Analysis of Queues, Chapter 9 in A First Course in Stochastic Models, Wiley, Chichester, 2003
[21]. Kendall, D. G. (1953). "Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain". The Annals of Mathematical Statistics. 24 (3): 338–354. doi:10.1214/aoms/1177728975. JSTOR 2236285.
[22]. Hernández-Suarez, Carlos (2010). "An application of queuing theory to SIS and SEIS epidemic models". Math. Biosci. 7 (4): 809–823. doi:10.3934/mbe.2010.7.809. PMID 21077709.