1 Introduction

The idea of a queue is one that has been around for as long as anyone can remember. It is a concept that has driven the establishments throughout our history in an orderly fashion. Especially today, we encounter this concept without even realizing it because it has become such a natural core requisite for the processing of our everyday lives. Imagine a world without the idea of a queue. There would be absolute chaos, and things would get done in a terribly inefficient way. In a world with a rapidly increasing population, queues are becoming more and more important. When it comes to something like a simple list of chores, we think to ourselves and utilize all our possible resources to accomplish our task as soon as humanly possible. When it comes to more complicated subjects, ranging from people being served inside a store to the processing of data inside a computer, the Queueing Theory becomes necessary so that we may construct a model to improve the system that is handling the queues. In the real life situation, an investment of some form must be made in order to reduce the waiting of time, which in almost all cases is what businesses strive to do. Queueing models and techniques are required so that whoever is in charge can make the best judgement call possible on whether or not the investment is a good idea or a bad idea. Some specific systems that can be worked on and improved by these queueing models include and are not limited to production, supply, transportation, communication, and information
processing. There are countless examples out in the real world where the Queueing Theory and models are extremely important for success. In a supermarket, the managers must be alert about several elements that directly affect how the business is run. A manager must be concerned about when the store is busiest so that he has enough people working to service the people trying to buy groceries. This also involves knowing how long each customer must wait in order to be served and the difference in efficiency with each employee that works at the cash register. For an insurance agency, the agents must be ready to handle every single phone call that comes in regarding questions about different policies, quotes, and any other question. If they don’t have enough trained people stationed and a customer doesn’t get the chance to get his or her question answered, that customer becomes angered and the insurance agents risk the loss of business. In some businesses, computers are connected to a large mainframe computer that processes all financial transactions. If the mainframe computer couldn’t store enough data, then the business would be in serious trouble and lose information. If the amount of transactions were to increase, the business would have to be sure that mainframe computer could handle the increase in productivity. Another classic example would be a traffic light. According to the flow of traffic, priority would have to come into play so that the design would be fair. Making the street that is less busy have more time with the green light would cause delay for the busier street and ultimately cause some serious inefficiency.

2 Simple Elements of a Queueing Model

With all these examples, you can pull out some elements that build the characteristics of a very basic queueing model:
2.1 The arrival process of the input:

It is extremely important to understand and analyse the arrival times of customers or data input. Customer arrival time can always vary, as they are not required to be dependent of one another. Independent arrivals may follow common distribution such as the Poisson distribution, or the exponential interarrival times. The customers may arrive solo or come in batches. An example of a batch would be several friends entering an ice cream store for dessert. They arrived as a group and must all be serviced. By understanding and analysing the arrival times of customers, a business can manage their time and efforts appropriately to service and satisfy the customers successfully.

2.2 The behaviour of the input:

Many businesses are concerned with customer satisfaction. No one can blame them, as that is a key factor in how well the business runs. The behaviour of the customers must be studied so that the business knows how to handle them. In many different businesses, the behaviour varies quite a bit. An example of a huge difference would be an office of the Department of Motor Vehicles and any typical restaurant. At the DMV office, the customers tend to be there for official business that must be done, and so although the waiting times may be uncomfortably long, they must stay there and deal with it. However, at a restaurant, if the waiting time begins to get too long, the customers may simply choose a different and better restaurant to go to and leave. The idea that customers may leave any time they want is a huge deal for businesses in those situations. The leaving of a customer because of the business’s disability to service the customer on time can cause a company to lose valuable reputation and business. To prevent any negatively affecting events from happening, the company should study their customers and understand their behaviours.
2.3 The service times:

This element refers to none other than the amount of time it takes to service an object in a queue. The importance of this subject should be trivial. If a business is purposely taking a very long time to service its customers, then the satisfaction of the customers drastically fall. The service time for any system can and should be adjusted depending on the length of the queue or the rate in which the queue increases. Having ten cashiers at a supermarket on a very slow day with a couple customers every hour would be very inefficient and irresponsible. Again it is important to study the behaviour of the input in order to set the appropriate service time. From past and future elements discussed, it is clear to see that other elements influence each other in specific ways that should be carefully examined.

2.4 The order of service:

In order to maintain the peace and prevent chaos, the order of service should be taken into consideration. A normal queue would have the first person to arrive into the queue serviced first, like a typical ice cream store. Some queues have priorities, such as mail. If you pay the extra cost, the post office would deliver your mail in a shorter time than normal. More rare queues have a completely random selection of who gets priority or not. Another type of priority is last come first served.

2.5 The capacity of service

With a direct relationship to the service times, the capacity of service refers to the number of workers servicing the customers. Generally speaking, the larger the capacity of service, the faster the service times. A business would typically use the service times to determine if the capacity of service should be increased or decreased to satisfy the customers.
2.6 The waiting room for the input not being serviced:

In many systems, there can only be a finite number of units that can be serviced. For example, at a dance club there can only be a certain amount of people (the capacity of the club). The people waiting to get in after the club has reached capacity must enter the system’s waiting room. Another example is a data communication network. In these networks, only a limited amount of cells can be buffered at a time, and a good design of the buffer size is important for the success of the network.

3 Kendall’s Notation:

David George Kendall, a British statistician, developed a system to characterize a queueing model. The notation is of the form $A/B/C$. $A$ represents the arrival process of the units that need to be serviced. There are quite a few different types of symbols that can be written for this spot, and some are M for a Poisson arrival process, D for an unchanging and constant inter-arrival time, and G for a general distribution. $B$ represents the service time distribution, which actually shares almost the same symbol representations: M for an exponential distribution, G for general distribution, and D for deterministic distribution. Also note that these are just a few examples for the B factor of Kendall’s notation. The last factor, $C$, represents the number of servers that are servicing the units for the queue (Adan,24).

4 Little’s Law:

Little’s Law is an equation that states that the average number of units in the system is equal to the average sojourn time multiplied by the average number of units entering the system per unit of time. The sojourn time is the time spent waiting in the waiting room plus the time spent being serviced. If we let,
\( L \) = Average number of units in the queueing system.
\( \lambda \) = Average number of arrivals per unit of time.
\( W \) = Average sojourn time.

Then the following equation represents Little’s Law:

\[
L = \lambda W
\]  \hspace{1cm} (1)

Although the equation is extremely simple and general, it still proves to be very useful. In fact, it is because of its simplicity that it is so useful. With only three unknowns, we do not have to consider small details such as the number of servers in the queueing system, or what the distributions for the service time and the arrival process. All we need are two measurements to be able to calculate the third. When it comes to queueing systems, measuring all three unknowns can be difficult, which allows for Little’s Law to come into play.

Application of Little’s Law seem almost endless because many businesses take into account the variables that are a part of the equation. An example of application could be the calculation of how long it takes a business to respond to an email. Suppose a company receives on average one-hundred emails a day. we can immediately point out that our \( \lambda = 100 \) messages/day. Assuming the business deletes an email once it has been read, then the remaining emails in the inbox are those that are waiting to be read. If over the past year the average number of emails in her inbox was approximately five hundred, then we can let \( L = 200 \) messages. Using Little’s Law, we may now calculate the time it takes on average for the company to respond to be \( W = 2 \) days.

Over the years, Little’s Law has faced some serious development and evolution to be applied to operations management. However, some of the terms are changed or replaced so that the equation is more applicable to the business. The original equation also assumed that the system was relatively stable and the arrival process was consistent, which is something we
definitely cannot assume for any business in life. Conditions must be met by altering the operations of the business so that Little’s Law can be applied and prove to be as useful as it truly is (Chhajed, 92).

5 Server Utilization

A concept that is important to be studied is the busyness of a queueing system, otherwise known as server utilization. The server utilization, often symbolized as \( \rho \), is a fraction that represents how active a system is. When the server utilization reaches one, the system is constantly busy. However, when \( \rho \geq 1 \), the mean queue length explodes, meaning the queue grows to infinity. This result would lead to a system that simply cannot handle its job. This would prove to be a huge problem for real life businesses, and adjustments would have to be made for the success of the business.

6 Priorities:

When priorities come into play for queueing, we must start taking into account different types of customers or input, each which receives a different amount of attention. There are two different priorities to be considered: non-preemptive priority rule and preemptive-resume priority rule. Non-preemptive priority rule states that even if a customer with higher priority has just arrived, that customer cannot interrupt the servicing of another customer with less priority. Preemptive-resume priority, on the other hand, allows for the interruption of ongoing services to attend to the highest priority customers first (Adan,87).

7 The Simple M/M/1 Queue:

In this queueing model, we let,

\[ \lambda = \text{the average arrival rate}, \]
\[ \mu = \text{the service rate}, \]
\[ \frac{1}{\lambda} = \text{the mean inter-arrival rate}, \]
\[ \frac{1}{\mu} = \text{the mean service rate}, \]

There is a single server,

There is an infinite amount of space in the waiting room,

The server utilization \( \rho = \frac{\lambda}{\mu} \) is always less than one.

### 7.1 Behaviour with respect to time:

Thanks to a nice exponential distribution, we can simply describe the state of the queue, which means the total amount of customers inside a system. We let \( p_n(t) \) stand for the probability of \( n \) customers at time \( t \). Using the characteristics of an exponential distribution, we get for \( \Delta t \to 0 \),

\[
p_0(t + \Delta t) = (1 - \lambda \Delta t)p_0(t) + \mu \Delta t p_1(t) + o(\Delta t) \tag{2}
\]

\[
p_0(t + \Delta t) = \lambda \Delta t p_{n-1}(t) + (1 - (\lambda + \mu) \Delta t)p_n(t) + \mu \Delta t p_{n+1}(t) + o(\Delta t), \quad n = 1, 2, ... \tag{3}
\]

These equations are subject to the normalisation condition that the sum of all the probabilities equals 1, we can take the limits as \( \Delta t \to 0 \), giving us the rather complicated differential equations:

\[
p'_0(t) = -\lambda p_0(t) + \mu p_1(t), \tag{4}
\]

\[
p'_n(t) = \lambda p_{n-1}(t) - (\lambda + \mu)p_n(t) + \mu p_{n+1}(t), \quad n = 1, 2, ... \tag{5}
\]

It is surprising how complicated one of the most simple queueing models can be when referring to its behaviour against time. When it comes to more complicated models, we can only expect much more complicated expressions of their behaviour (Adan, 29).
7.2 The equilibrium solutions:

In order to get the equilibrium solutions, we must have very specific initial conditions. Those conditions are:

\[ p'_i(t) = 0 \]  \hspace{1cm} (6)
\[ p_i(t) = p_i \]  \hspace{1cm} (7)

By defining \( \rho = \frac{\lambda}{\mu} \), with \( \rho < 1 \), we get:

\[ p_1 = \rho p_0 \]  \hspace{1cm} (8)

\[ p_{N+1} = (1 + \rho)p_N - \rho p_{N-1} = \rho p_N = \rho^{N+1}p_0, \quad N \geq 1 \]  \hspace{1cm} (9)

After applying the normalisation condition:

\[ p_i = \rho^i(1 - \rho), \quad i = 0, 1, 2, ... \]  \hspace{1cm} (10)

This is the equilibrium solution for the exponential distribution with the previously stated initial conditions. However, more importantly, the average arrival time must be more than the average service time (Bose, 18).

7.3 The mean waiting time:

After finding the equilibrium probabilities, we can obtain expressions for the average number of customers in the system, \( E(L) \), and the average sojourn time, \( E(S) \). At first, we have:

\[ E(L) = \sum_{n=0}^{\infty} np_n = \frac{\rho}{1 - \rho}, \]  \hspace{1cm} (11)

Apply Little’s Law to get:

\[ E(S) = \frac{1/\mu}{1 - \rho}. \]  \hspace{1cm} (12)
We can get $E(S)$ and $E(L)$ without knowing probabilities by using both Little’s Law and the PASTA property, which basically states that "the fraction of customers finding on arrival the system in some state A is exactly the same as the fraction of time the system is in state A" (Adan, 27). Using these tools, we get:

$$E(S) = E(L) \frac{1}{\mu} + \frac{1}{\mu}. \quad (13)$$

We also have:

$$E(L) = \lambda E(S). \quad (14)$$

And in conclusion, we can obtain the mean number of customers in the queue, $E(L^q)$, by subtracting the mean number of customers in service from $E(L)$:

$$E(L^q) = E(L) - \rho = \frac{\rho^2}{1 - \rho}. \quad (15)$$

Similarly, we get the mean waiting time, $E(W)$, by subtracting the mean service time from $E(S)$.

$$E(W) = E(S) - \frac{1}{\mu} = \frac{\rho}{\mu} \frac{\mu}{1 - \rho}. \quad (16)$$

8 Conclusion

In this world, one can never run away from eventually facing a queue of some form. As the population of our world increases, the demand for services increases as well. Without putting a large amount of effort into creating stability for such a demand, inefficiency will skyrocket and eventually destroy any sense of order that existed before. The understanding of the queueing theory is extremely useful for the organization of any system that requires servicing. With this knowledge, you can analyze the weaknesses of a system, and make
logical decisions to improve it. This progression into a perfect system is what everyone should strive for to create a more perfect world.
Bibliography


Wishart, David M. G. *Queueing Systems in Which the Discipline is 'Last-Come, First-Served*. Hanover, MD: INFORMS, 1960.