# Section 8.0: Queueing Theory Introduction

"Queues" is the British word for waiting lines, and we have all experienced them. It is estimated that Americans spend over a billion hours per year waiting in queues. Queueing theory is the mathematical study of queues that operations researchers and industrial engineers use to increase the efficiency of queueing systems. The term **queueing system** refers to the people currently waiting in the queue and the person(s) being served. Inefficient queues not only hurt customer satisfaction, but they also have an impact on a nation's economy. If the time we wasted waiting in queues could instead be spent productively, it would amount to the equivalent of half a million additional workers.

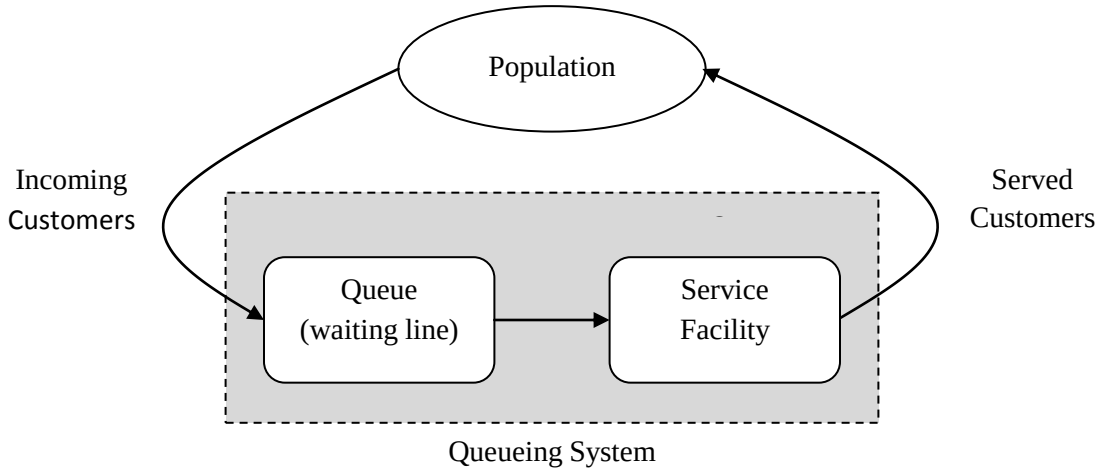


**Figure 8.0.1:** Basic queuing model

Figure 8.0.1 represents the key components of queueing systems. The system begins with customers seeking service and entering the queuing system. They are characterized as an arrival process with an average rate of $\lambda$. The second critical element is the service facility. This can involve one individual, a single server or multiple servers. The basic building block is the rate, $\mu$, at which service is completed by a continuously busy server. The number of servers is simply, $s$. The final element is the queue of customers waiting to be served. When there is more than one server, a decision has to be made as to how to organize the queues. The customers can be organized to form a single queue as most banks do. Alternatively, there may be a separate queue for each server as occurs with cars waiting to pay tolls. Once a customer is served, the customer leaves the queueing system.

The sequence in which customers are served is called the queue discipline. The simplest procedure is FIFO, first-in-first-out. The next customer served is the one at head of the line. However, the customers seeking service may fall into distinct categories that are treated differently. For example, in an emergency room, patients experiencing a heart attack will be given priority over someone with a broken finger. Most airlines give priority boarding to frequent flyers. Some amusement parks allow patrons to pay a premium to reduce their waiting time at rides.

Before discussing the mathematics, we would like the reader to explore his or her own queueing experiences.

Q1. Describe two different queues in which you have spent time waiting in an actual line.

Q2. Describe one situation in which you were waiting for service but the people waiting were not in a physical line that you could see.

Q3. Estimate how much time you spend waiting in queues during a typical week.

Q4. Reflecting on the time you spend in queues, propose two strategies for reducing the amount of time customers spend waiting.

Q5. From a manager's perspective, what are the downsides of your suggestions?

Not all queues are created equal. Some queues are designed to include distractions to make the time you spend waiting more pleasant. Other queues are quite boring and may lead you to frustration as you think about ways you could be using your time more productively. The experience of waiting in line is influenced by the waiting area environment and our expectations as to the length of the wait. Imagine having to wait standing up in a dentist's office for twenty minutes, while a patient is screaming in an adjacent examination room. Now imagine an alternative wait in comfortable chairs with access to the "latest" magazines for a variety of customer tastes. For your ten-year old child there is a video game console, and the area is sound proof.

Disney is one example of a company that has become expert in understanding the psychology of waiting. Waiting in a line that is moving seems less boring than standing still in the same spot. TV monitors with engaging pictures help keep visitors' minds off the clock. In addition, if they can see and hear some of the excitement of those who have completed their wait and enjoyed the attraction, anticipation increases and waiting seems worthwhile. Lastly, expectations are a major factor in determining customer satisfaction. If customers approach a line and are told the wait will be fifteen minutes, at least they have the information to make an informed judgment as to join the line or not. If it turns out to be less than the quoted fifteen minutes, they are pleasantly surprised.

Another dimension to the psychology of waiting relates to fairness. It can be very upsetting to see someone arrive after you in line and end up being served before you. This can happen if there are two separate lines. You might get stuck behind a customer who has a complicated request that takes a long time to service. As a result, people who have joined the other line even after you might ending up waiting less time. Many businesses have addressed this potential inequity by creating one line which all arriving customers enter. Thus, anyone who arrives after you must be further back in line and cannot begin service before you do.

Q6. Propose two strategies for making the time spent in a queue more pleasant.

Q7. Compare your responses to numbers 4 and 6; which approach seems more cost effective?

All of us have experienced the annoyance of having to wait in line. Unfortunately, this phenomenon continues to be common in congested, urbanized and "high-tech" societies. We wait in line in our cars in traffic jams or at toll booths; we wait on hold for a help-line professional waiting for assistance; we wait in line at supermarkets to check out; we wait in line at fast-food restaurants; and we wait in line at banks and post offices. As customers, we do not generally like these waits. The managers of the establishments at which we wait also do not like us to wait, since it may cost them business. Why then is there waiting?

The answer is relatively simple: There is more demand for service than there is facility for service available. Why is this so? There may be many reasons; for example, there may be a shortage of available servers; it may be infeasible economically for a business to provide the level of service necessary to prevent waiting; or there may be a limit to the amount of service that can be provided. Generally, this limitation can be removed with the expenditure of capital.

A key contributor to queue formation is the randomness inherent in the arrival process and the service time. For example, if customers were scheduled every 15 minutes and each customer required exactly 14 minutes of service, no line would form. However, arrival and service of randomness will result in long lines as will be discussed later in this chapter.

The chapter contains a number of formula used to calculate queueing statistics. All of the mathematical models and examples in this chapter have three critical assumptions

- The randomness in the arrival process is described by the **Poisson distribution**.
- The randomness in the service time distribution is re represented by the **exponential distribution**.
- The statistics apply to long-term averages which in the queuing literature is called **steady state**.

To know how much service to make available, a manager would need to answer such questions as, "How long will a customer wait?" and "How many people will form in the line?" Queueing theory attempts to answer these questions through detailed mathematical analysis. The word "queue" is more commonly used in Great Britain and other countries than in the United States, but it is rapidly gaining acceptance in this country. However, it must be admitted that it is just as unpleasant to spend time in a queue as in a waiting line.

A queueing system can be simply described as customers arriving for service, waiting for service if it is not immediate, and if having waited for service, leaving the system after being served. The term customer is used in a general sense and does not imply necessarily a human customer. For example, a customer could be a ball bearing waiting to be polished, an airplane waiting in line to take off, a computer program waiting to be run, or a telephone call waiting to be answered.

Queueing theory can trace its origins back to a Danish mathematician named A. K. Erlang. In 1909, Erlang published *The Theory of Probabilities and Telephone Conversations* based on work he did for the Danish Telephone Company in Copenhagen. In the long gone days, all telephone calls passed through a telephone exchange operator who made the connection and calls were backed-up when all operators were busy. Work continued in the area of telephone applications and was a major factor in managing the first 911 systems in NYC.

There are many valuable applications of the theory, including traffic flow (vehicles, aircraft, people, communications), scheduling (patients in hospitals, jobs on machines, programs on a computer), and facility design (banks, post offices, amusement parks, fast-food restaurants).

## Initiating Activity: Queue at the Pencil Sharpener

The activity will almost always NOT work the first time as designed for many of the following reasons.

- Students forget when their time approaches to go up
- Students forget to bring pencil
- Students forget to bring paper slip
- Students have trouble sharpening pencil – keep checking
- Recorders forget to write down times

As you see the activity break down, stop the activity and discuss an important lesson from the mishap and confusion. Processes rarely work as designed the first time. You can share your own experiences of processes that were implemented and did not work as well as intended especially early in the adoption. You may have to restart the process 2 or 3 times.

From this activity students will learn about

- Randomness
- Building blocks of elements of a queueing system
    - Arrival rate
    - Service rate (1/average service time)
    - Number of servers

**Materials**

- Clock showing seconds (preferably a digital counter)
- Two pencil sharpeners
- One pencil per student
- One slip of paper per student
- List of random numbers – one for each student

**Procedure**

First count the number of students who will be given pencils and random numbers. If you have 24 or more students use two pencil sharpeners throughout and plan a three minute (180 second) experiment. Use as many numbers as needed from the table below. Otherwise plan a two minute experiment and only use numbers below 120. All of the numbers in the first three rows are less than 120. The red numbers are the odd values.

| | | | | | | |
|---|---|---|---|---|---|---|
| All values less than 120 | 72 | 103 | 40 | 56 | 117 | 24 |
| | 10 | 49 | 96 | 69 | 26 | 36 |
| | 55 | 14 | 33 | 92 | 106 | 38 |
| Some values greater than 120 | 46 | 141 | 15 | 6 | 132 | 124 |
| | 154 | 65 | 165 | 18 | 171 | 29 |
| | 83 | 145 | 178 | 35 | 136 | 110 |

Data Collection

- One student at each pencil sharpener. He or she is to be handed the slip of paper as the student is ready to start sharpening the pencil.
    - Record time started sharpening
    - Record time finished sharpening
- One student to count and record the number of students in line every 15 seconds. Include the person sharpening a pencil. This student should be able to track even two separate lines. However, if necessary assign separate students to track each line.

| | 15 | 30 | 45 | 60 | 75 | 90 | 105 | 120 | 135 | 150 | 165 | 180 | 195 | 210 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Line 1 | | | | | | | | | | | | | | |
| Line 2 | | | | | | | | | | | | | | |

Student participants

- Give each student a broken point pencil that need sharpening
- Give each student a slip with the time on it

Tell the students to join the line at the time listed on their paper. They are to turn in their paper slips when they are about to pick up the pencil to begin sharpening. The data recorder will then record the time sharpening begins and ends on the same slip of paper.

24 or more students

- Two pencil sharpeners at different locations in the room
- Designate one sharpener for odd numbered slips and one for even numbered.
- In general, distribute the odd and even slips according to the location of their respective sharpeners.
- Carry out the activity with odds and evens separated.
- Repeat the activity with the two sharpeners at one location with a single line formed. Students use which ever sharpener is available. (Remember to dull the pencil points)

Fewer than 24 students

- First, run the activity with just one pencil sharpener.
- Repeat the activity with two pencil sharpeners at one location. (remember to dull the pencils)

Data Analysis

- Calculate the difference between the main time listed on the paper and the time the student began sharpening the pencil. This is the wait time (although it may include other delays depending upon how the student responded to the time on his slip of paper.)
- Calculate the difference between the start and finishing sharpening the pencil. This is service time.
- Count the number of students and divide by the activity duration (120 or 180 seconds) to determine the average arrival rate.
- Calculate the average number of students in the queue.
- Calculate these statistics for both repetitions of the activity.

# Section 8.1: Arm-and-a Leg Tickets: Does This Line Ever Move?

Mr. I. M. Boss is vice president in charge of operations for Arm-and-a-Leg Ticket Sales. He is concerned about complaints regarding long waits at the ticket windows on Friday afternoons at many of the malls. To reduce the number of complaints, Mr. Boss has hired Dr. Hye I. Cue, an expert on queueing theory. Before we begin to analyze Mr. Boss's problem, which is called a single-server model, we will make the following assumptions:

- Individual customers arrive at random to purchase tickets.[1]
- The time to complete a purchase is also random. This might be due to the number of tickets the customer purchases or the customer asking for information about dates and seat locations.[2]

The arrival process is represented as an average rate of $\lambda$. The service time is also represented as a rate, the number of customers served per unit time when the server is continuously busy. For example assume the time to service a customer were 12 minutes. This is converted to a rate of five customers per hour. The Greek letter $\mu$ is the standard symbol for the service rate.

To use queueing theory, Dr. Cue needed to collect data about the customers. She spent several Friday afternoons observing the situation and collecting the data. She found that the average number of customers arriving per hour is 18. She also determined that it takes on average 3 minutes to process a customer. This means the single ticket agent can serve 20 customers per hour when he is continuously busy.

Q1. Dr. Cue observed that $\lambda$ = ____ and $\mu$ = _____. (Be sure to include units.)

Q2. If 18 customers arrive per hour, on average, how much time is there between successive arrivals?

Q3. If $\lambda$ customers arrive per hour, on average, how much time is there between arrivals?

Q4. If $\mu$ customers can be served per hour, on average, how long does it take to serve one customer?

To develop a mathematical model of our queue, we must have an idea of the utilization, $\rho$. This is the ratio of the average rate of customer arrivals per unit of time, $\lambda$, to the average rate of customers who can be served per unit of time, $\mu$. In order for this ratio to make sense, the time units of $\lambda$ must be the same as those of $\mu$.

Let $\rho = \frac{\lambda}{\mu}$.

[1] The pattern of random arrivals is assumed to follow the Poisson distribution.
[2] The pattern of random duration of service is assumed to follow the exponential distribution.

Q5. What is the utilization on Friday afternoons at the Arm-and-a Leg Ticket counter?

The average number of customers in the system is represented by $L$. This includes those in line and the customer at the ticket window. Experts in queueing theory have determined the following relationship between $L$ and $\rho$.[3]

$$L = \frac{\rho}{1-\rho}$$

For example, there are times during the week when the number of arriving customers is much less than 18. For example, on Tuesday evenings the arrival rate is only 12 per hour. In that case the utilization, $\rho$, is equal to 0.6. Then

$$L = \frac{0.6}{1-0.6} = 1.5 \text{ customers}$$

There would be on average 1.5 customers in the line including a customer being served. (This is an average. Therefore, there is nothing wrong with this average not being an integer value.)

Q6. Use the value of $\rho$ for Friday afternoons, to calculate $L$ for Friday afternoons.

Customer satisfaction actually is more dependent upon the length of time it takes to get a ticket than the length of the line.

Let $W$ = the average time a customer waits in a system including the time to be served by the ticket agent.

There is a well-known equation known as Little's formula, that relates $W$ to $L$.

$$L = \lambda W$$

This equation is easily manipulated to determine $W$ in terms of $L$.

$$W = \frac{L}{\lambda}$$

When the arrival rate was only 12 per hour, $L$ was equal to 1.5.

$$W = \frac{1.5}{12} = 0.125 \text{ hours}$$

[3] The calculation of $L$ assumes the system is approaching steady state and the statistics are long-term averages. This calculation involves the following steps. Create of an infinite series of state rate balance equations. These are solved by recursion. Then $L$ is calculated with a formula for an infinite geometric series. F.S. Hillier and G. J. Lieberman (2012) Introduction to Operations Research, McGraw Hill.

Thus, the average time in the system is 0.125 hours. The unit of measurement is hours because the arrival rate was given in arrivals per hour. This number is multiplied by 60 to translate the waiting time into an average of 7.5 minutes. This total waiting time to complete service can be split into its two components. The first component is the time waiting to begin service. This is referred to as the wait in queue, $W_q$. The second component is the service time.

$$\begin{aligned} W &= \text{total waiting time} \\ &= \text{average time waiting to begin service + average service time} \\ &= W_q + \frac{1}{\mu} \end{aligned}$$

Thus to determine $W_q$ we simply subtract the average service time from $W$.

$$W_q = W - \frac{1}{\mu}$$

Earlier it was calculated that W was 7.5 minutes and the average service time was three minutes. Thus, the average wait in queue on Tuesday evenings is 4.5 minutes.

$$\begin{aligned} W_q &= W - \frac{1}{\mu} \\ &= \left(0.125 - \frac{1}{20}\right) \text{ hours} \\ &= (7.5 - 3) \text{ minutes} \\ &= 4.5 \text{ minutes} \end{aligned}$$

Q7. What is the average waiting time, $W$, on Friday afternoons? (The first calculation will be in units of hours. Convert this to a more readily understandable unit of minutes.)

Q8. What is the average waiting time in queue, $W_q$, on Friday afternoons?

Dr. Cue was interested in understanding waiting time on other days. She asked the person in charge of the computer system to show her what data was recorded with each ticket sale. She found that the system kept track of the actual time the ticket was sold. She decided that she could use those times to estimate the number of customers arriving per hour.

Q9. Why is the time stamp on the ticket not a perfect indication as to when the customer came to purchase a ticket?

Dr. Cue found that the arrival rate varied by day of week and time of day. The average arrival rate ranged from a low of 4 per hour on Monday afternoons to a high of 18 per hour on Friday afternoons. She also noticed that around the holiday season, the rate could even exceed 19 customers per hour.

Q10. Complete the table below to determine the average utilization ($\rho$), queue length ($L$), waiting time ($W$), and waiting time in queue ($W_q$) for different arrival rates ($\lambda$). (The initial calculation of W will be measured in hours since both the arrival and service rates are defined in units per hour. Convert this number into minutes.)

| $\lambda$ | $\rho = \frac{\lambda}{\mu}.$ | $L = \frac{\rho}{1-\rho}$ | $W = \frac{L}{\lambda}$ | $W_q = W - \frac{1}{\mu}$ |
|---|---|---|---|---|
| 4 | | | | |
| 7 | | | | |
| 10 | | | | |
| 12 | 0.6 | 1.5 | 0.125 hrs. = 7.5 min. | 4.5 min. |
| 14 | | | | |
| 16 | | | | |
| 18 | | | | |
| 19 | | | | |
| 19.5 | | | | |

**Table 8.1.1:** Queues for different arrival rates

Dr. Cue asked her assistant to graph $L$ as a function of $\rho$ using the data in Table 8.1.1. After reviewing the graph, she noticed there appeared to be two asymptotes.

Q11. Describe the vertical line that appears to be an asymptote.

Q12. Describe the horizontal line that appears to be an asymptote.

Now let's consider other values of $\rho$ not mentioned in Table 8.1.1.

Q13. Suppose $\rho$ = - 0.5. What would be the value of $L$?

Q14. Why do values of $\rho$ that are negative make no sense in this problem context? Why do the corresponding values of $L$ also have no meaning?

In the Table 8.1.1 the highest value of $\lambda$ was 19.5 customers per hour. It is possible that during the busiest periods the number of arrivals might be a as high 25 customers per hour. If that occurs

$$\rho = \frac{25}{20} = 1.25$$

When the utilization is greater than 1, the arrival rate exceeds the service capacity. In this instance, there are five more customers per hour than can be serviced. This is 25% more than the service capacity. If you insert 1.25 into the equation for $L$, the equation yields the value -5.0.

$$L = \frac{1.25}{1-1.25} = -5 \text{ customers}$$

There can be no such thing as negative five customers waiting in line. In other words this equation is not meaningful if $\rho$ is greater than one.

What happens to the equation when the arrival rate equals 20, the same as the service rate? If $\rho$ were equal to 1, the denominator would be 0 and the expression for $L$ would be undefined. In fact $\rho$ equal to 1 is an asymptote. As $\rho$ approaches 1 from below, $L$ grows infinitely.

Thus, the equation for $L$ provides usable and meaningful values over the domain $0 \leq \rho < 1$. To represent this fact, we provide the domain as well as the equation when specifying $L$ as a function of $\rho$.

$$L = \frac{\rho}{1-\rho} \qquad \text{for} \qquad 0 \leq \rho < 1$$

Why is it that the equation does not work for values of $\rho$ greater than or equal to 1? The mathematics used to develop the equation of $L$ applies when the system fluctuates around a long-term average value, $L$. However, when the arrival rate exceeds the service rate, the lines simply continue to increase hour after hour. For example if $\lambda$ is 25, then each hour there will be five more people than the system can handle. As a result, the queue would continually increase.

## 8.1.1 Reducing Waiting Time at Peak Hours

Dr. Cue and Mr. Boss began discussing ways to reduce the waiting time during peak hours. One short term solution is to provide a minimum wage assistant to help the ticket agent speed up processing during peak hours. With the assistant, the average processing is estimated to decrease by 25% to two minutes and 15 seconds.

Q15. On average, what is the maximum number of customers the agent can process in an hour with the assistant's help?

Q16. With this solution, what is the total waiting time, *W*, when customers are arriving at rate of 18 per hour on Friday afternoons?

During the holiday season, the arrival rate can increase to as much as 19.5 customers per hour.

Q17. With this solution would the waiting times be reasonable even during the holiday season?

A longer term solution is to provide a self-service terminal alongside the ticket booth. By placing it near the booth, people who are having difficulty can always buy from the ticket agent. It is estimated that 20 percent of the customers will be able to use the terminal to make a simple purchase. The other customers will buy from the ticket agent.

Q18. If this solution is used, what will be the total waiting time, *W*, for customers on Friday afternoons?

Q19. With the addition of a terminal, would the waiting times be reasonable even during the holiday season?