Program:	Software Engineering	Class:	VII-A, B
Course Name:	Modeling and Simulations	Topic	Waiting Line Models
Instructor's Name:	M. Iftikhar Mubbashir	Date	May 28, 2024
Course Code:	SE 405	Document	Notes

Introduction:

- Models have been developed to help managers understand and make better decisions concerning waiting line operations. In management science terminology, a waiting line is also known as a **queue,** and the body of knowledge dealing with waiting lines is known as **queueing theory.** In the early 1900s, A. K. Erlang, a Danish telephone engineer, began a study of the congestion and waiting times occurring in the completion of telephone calls. Since then, queueing theory has grown far more sophisticated, with applications in various waiting line situations.
- Waiting line models consist of mathematical formulas and relationships that can be used to determine a waiting line's operating characteristics (performance measures). Operating characteristics of interest include the following:

1. The probability that no units are in the system
2. The average number of units in the waiting line
3. The average number of units in the system (the number of units in the waiting line plus the number of units being served)
4. The average time a unit spends in the waiting line
5. The average time a unit spends in the system (the waiting time plus the service time)
6. The probability that an arriving unit has to wait for service

- Managers with such information can better make decisions that balance desirable service levels against the cost of providing the service.

Structure of a Waiting Line System

- Let's consider the waiting line at the Burger Dome fast-food restaurant.
- Burger Dome sells hamburgers, cheeseburgers, French fries, soft drinks, milkshakes, and a limited number of specialty items and dessert selections. Although Burger Dome would like to serve each customer immediately, at times, more customers arrive than can be handled by the Burger Dome food service staff.
- Thus, customers wait in line to place and receive their orders. Burger Dome is concerned that the current customer service methods result in excessive waiting times. Management wants to conduct a waiting line study to help determine the best approach to reduce waiting times and improve service.

Single-Channel Waiting Line

- In the current Burger Dome operation, a server takes a customer's order, determines the total cost, takes the money from the customer, and then fills the order. Once the first customer's order is filled, the server takes the next customer's order, waiting for service.
- This operation is an example of a **single-channel waiting line.**
- Each customer entering the Burger Dome restaurant must pass through *one* channel—one order-taking and order-filling station—to place an order, pay the bill, and receive the food.
- When more customers arrive than can be served immediately, they form a waiting line and wait for the order-taking and order-filling station to become available. A diagram of the Burger Dome single-channel waiting line is shown in Figure 1.

Distribution of Arrivals

- Defining the arrival process for a waiting line involves determining the probability distribution for the number of arrivals in a given period. For many waiting line situations, the arrivals occur *randomly and independently* of other arrivals, and we cannot predict when an arrival will occur. In such cases, quantitative analysts have found that the **Poisson probability distribution** describes the arrival pattern well.

The Poisson probability function provides the probability of x arrivals in a specific period. The probability function is as follows:

$$P(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, for\ x = 0, 1, 2, \cdots$$

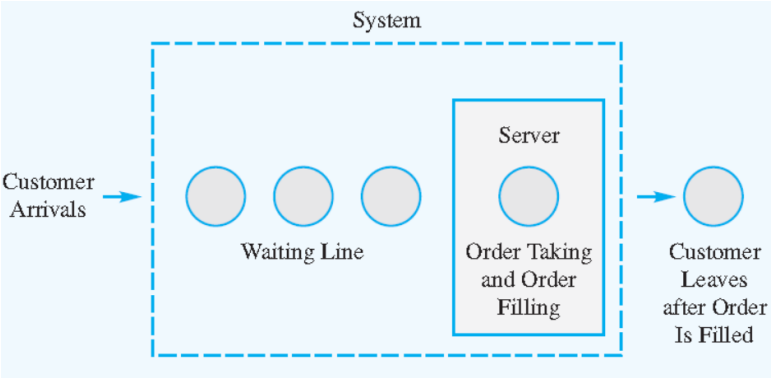

Figure 1: The Burger Dome Single-Channel Waiting Line

- The mean number of arrivals per period, l, is called the **arrival rate.**
- Suppose that Burger Dome analyzed data on customer arrivals and concluded that the arrival rate is 45 customers per hour. For one minute, the arrival rate would be $\lambda = \frac{45\text{customers}}{60\text{minutes}} =$ 0.75 customers / minute. Thus, we can use the following Poisson probability function to compute the probability of *x* customer arrivals during one minute:

The probabilities of 0, 1, and 2 customer arrivals during one minute are		
Customer Number	Formulae	Results
0	$P(0; 0.75) = \frac{0.75^0 e^{-0.75}}{0!}$	0.4724
1	$P(1; 0.75) = \frac{0.75^1 e^{-0.75}}{1!}$	0.3543
2	$P(2; 0.75) = \frac{0.75^2 e^{-0.75}}{2!}$	0.1329

Distribution of Service Times

- The service time is the time a customer spends at the service facility once the service has started.
- At Burger Dome, the service time starts when a customer begins to place an order with the food server and continues until the customer receives the order. Service times are rarely constant.
- At Burger Dome, the number and mix of items ordered vary considerably from one customer to the next.
- Small orders can be handled in seconds, but large orders may require more than two minutes.
- Quantitative analysts have found that if the probability distribution for the service time can be assumed to follow an **exponential probability distribution,** formulas are available to provide helpful information about the operation of the waiting line. Using an exponential probability distribution, the probability that the service time will be less than or equal to a time of length *t*Is:

$$P(x \leq t) = 1 - e^{-\mu t}$$

- Where μ is the mean number of units that can be served per period.
- The mean number of units that can be served per time period, μ, is called the **service rate.**
- Suppose that Burger Dome studied the order-taking and order-filling process and found that the single food server can process an average of 60 customer orders per hour. On a one-minute basis, the service rate would be $\mu = \frac{60\text{customers}}{60\text{minutes}} =$ 1 customers/minute.

Queue Discipline

In a waiting line system, we must define the way the waiting units are arranged for service.

- For the Burger Dome waiting line, and in general for most customer-oriented waiting lines, the units waiting for service are arranged on a **first-come, first-served** basis; this approach is referred to as an **FCFS** queue discipline.
- However, some situations call for different queue disciplines. For example, when people wait for an elevator, the last one on the elevator is often the first one to complete service (i.e., the first to leave the elevator).
- Other types of queue disciplines assign priorities to the waiting units and then serve the unit with the highest priority first.

The probability an order can be processed in $1/_2$ minute or less, 1 minute or less, and 2 minutes or less.		
Customer Number	Formulae	Results
$P(service\ time \leq 0.5)$	$1 - e^{-1(0.5)}$	0.3935
$P(service\ time \leq 1.0)$	$1 - e^{-1(1.0)}$	0.6321
$P(service\ time \leq 2.0)$	$1 - e^{-1(2.0)}$	0.8647

Steady-State Operation

- When the Burger Dome restaurant opens in the morning, no customers are in the restaurant.
- Gradually, activity builds up to a normal or steady state. The beginning or start-up period is referred to as the **transient period.**
- The transient period ends when the system reaches the normal or **steady-state operation.**
- Waiting line models describe the steady-state operating characteristics of a waiting line.

Improving the Waiting Line Operation

- Waiting line models often indicate when improvements in operating characteristics are desirable.
- However, the decision of how to modify the waiting line configuration to improve the operating characteristics must be based on the insights and creativity of the analyst.
- After reviewing the operating characteristics provided by the waiting line model, Burger Dome's management concluded that improvements designed to reduce waiting times are desirable. To make improvements in the waiting line operation, analysts often focus on ways to improve the service rate.
- Generally, service rate improvements are obtained by making either or both of the following changes:

1. Increase the service rate by making a creative design change or by using new technology.
2. Add one or more service channels so that more customers can be served simultaneously.

- Assume that in considering alternative 1, Burger Dome's management decides to employ an order filler who will assist the order taker at the cash register. The customer begins the service process by placing the order with the order taker.
- As the order is placed, the order taker announces the order over an intercom system, and the order filler begins filling the order. When the order is completed, the order taker handles the money, while the order filler continues to fill the order.
- With this design, Burger Dome's management estimates the service rate can be increased from the current 60 customers per hour to 75 customers per hour. Thus, the service rate for the revised system is $\mu = \frac{75\ \text{customers}}{60\text{minutes}} = 1.25\ \text{customers/minute}$.
- . For $\lambda = 0.75$ customers / minute and $\mu = 1.25\frac{\text{customers}}{\text{minute}}$ we look into the statistics again and decide whether intervention worked or not.
- Are any other alternatives available that Burger Dome can use to increase the service rate? If so, and if the mean service rate μ can be identified for each alternative, the revised operating characteristics and any improvements in the waiting line system can be checked.
- The added cost of any proposed change can be compared to the corresponding service improvements to help the manager determine whether the proposed service improvements are worthwhile.
- Another option often available is to add one or more service channels so that more customers can be served simultaneously.
- The extension of the single channel waiting line model to the multiple-channel waiting line model.

Multiple-Channel Waiting Line Model With Poisson Arrivals And Exponential Service Times

- A **multiple-channel waiting line** consists of two or more service channels that are assumed to be identical in terms of service capability. In the multiple-channel system, arriving units wait in a single waiting line and then move to the first available channel to be served.
- The single-channel Burger Dome operation can be expanded to a two-channel system by opening a second service channel. Figure -2 shows a diagram of the Burger Dome two-channel waiting line.
- The formulas are applicable if the following conditions exist:

1. The arrivals follow a Poisson probability distribution.
2. The service time for each channel follows an exponential probability distribution.
3. The service rate _ is the same for each channel.
4. The arrivals wait in a single waiting line and then move to the first open channel for service.

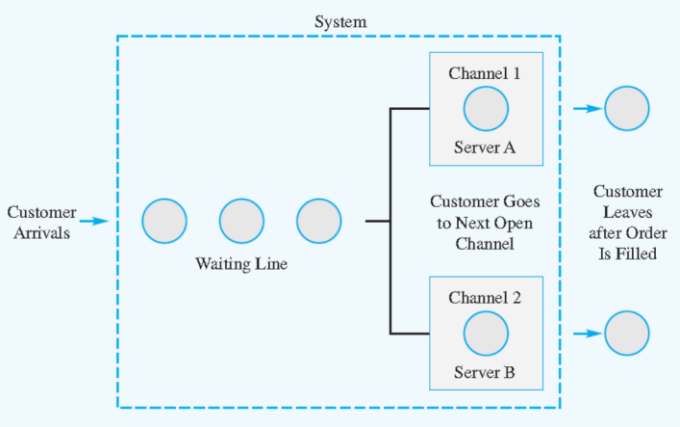

Figure 2:The Burger Dome Two-Channel Waiting Line

Operating Characteristics

- The following formulas can be used to compute the steady-state operating characteristics for multiple-channel waiting lines, where
- λ: the arrival rate for the system
- μ: the service rate for each channel.
- k: the number of channels.
- Because μ is the service rate for each channel, $k\mu$ is the service rate for the multiple channel system.
- The formulas for the operating characteristics of multiple-channel waiting lines can be applied only in situations where the service rate for the system is greater than the arrival rate for the system.
- Some expressions for the operating characteristics of multiple-channel waiting lines are more complex than their single-channel counterparts.

Operating Characteristics for the Burger Dome Problem

- To illustrate the multiple-channel waiting line model, we return to the Burger Dome fast food restaurant waiting line problem.
- Suppose that management wants to evaluate the desirability of opening a second order-processing station so that two customers can be served simultaneously.
- Assume a single waiting line with the first customer in line moving to the first available server. Let us evaluate the operating characteristics for this two channel system.
- We can now compare the steady-state operating characteristics of the two-channel system to the operating characteristics of the original single-channel system.
- After considering these results, what action would you recommend?
- By changing the arrival rate l to reflect arrival rates at different times of the day, and then computing the operating characteristics, Burger Dome's management can establish guidelines and policies that tell the store managers when to schedule service operations with a single channel, two channels, or perhaps even three or more channels.

Some General Relationships For Waiting Line Models

- The operating characteristics of interest included:
- L_q:the average number of units in the waiting line
- L: the average number of units in the system
- W_q:the average time a unit spends in the waiting line
- W: the average time a unit spends in the system
- John D. C. Little showed that several relationships exist among these four characteristics and that these relationships apply to a variety of different waiting line systems.
- Two of the relationships, referred to as *Little's flow equations,* are:

$$L = \lambda W$$

$$L_q = \lambda W_q$$

$$\frac{L_q}{\lambda} = W_q$$

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

Economic Analysis Of Waiting Lines

- Frequently, decisions involving the design of waiting lines will be based on a subjective evaluation of the operating characteristics of the waiting line.
- For example, a manager may decide that an average waiting time of one minute or less and an average of two customers or fewer in the system are reasonable goals.
- The waiting line models presented in the preceding sections can be used to determine the number of channels that will meet the manager's waiting line performance goals.
- On the other hand, a manager may want to identify the cost of operating the waiting line system and then base the decision regarding system design on a minimum hourly or daily operating cost.
- Before an economic analysis of a waiting line can be conducted, a total cost model, which includes the cost of waiting and the cost of service, must be developed.
- To develop a total cost model for a waiting line, we begin by defining the notation to be used:
- C_w: the waiting cost per time period for each unit
- L: the average number of units in the system
- C_s: the service cost per time period for each channel
- k: the number of channels
- TC: the total cost per time period

$$TC = C_w L + C_s k$$

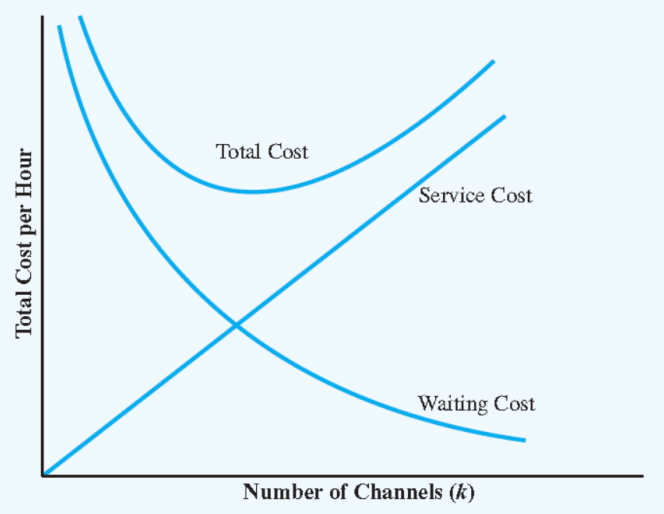

Figure 3: The General Shape Of Waiting Cost, Service Cost, and Total Cost Curves in Waiting Line Models

Other Waiting Line Models

- D. G. Kendall suggested a notation that is helpful in classifying the wide variety of different waiting line models that have been developed.
- The three-symbol Kendall notation is as follows:

$$A/B/k$$

where

- Depending on the letter appearing in the A or B position, a variety of waiting line systems can be described. The letters that are commonly used are as follows:
- M designates a Poisson probability distribution for the arrivals or an exponential probability distribution for service time.
- D designates that the arrivals or the service time is deterministic or constant
- G designates that the arrivals or the service time has a general probability distribution with a known mean and variance
- Using the Kendall notation, the single-channel waiting line model with Poisson arrivals and exponential service times is classified as an $M/M/1$ model.
- The two-channel waiting line model with Poisson arrivals and exponential service times presented would be classified as an $M/M/2$ model.
- A: denotes the probability distribution for the arrivals
- B: denotes the probability distribution for the service time
- k: denotes the number of channels

Operating Characteristics for the M/G/1 Model

- The notation used to describe the operating characteristics for the $M/G/1$ model is
- λ: the arrival rate
- μ: the service rate
- s: the standard deviation of the service time

An Example: Retail sales at Hartlage's Seafood Supply are handled by one clerk. Customer arrivals are random, and the arrival rate is 21 customers per hour, or $\lambda = \frac{21}{60} = 0.35$ customers per minute. A study of the service process shows that the service time is 2 minutes per customer, with a standard deviation of $\sigma = 1.2$ minutes. The mean time of 2 minutes per customer shows that the clerk has a service rate of $\mu = \frac{1}{2} = 0.50$ customers per minute. Compute operating characteristics of this $M/G/1$ waiting line system?

Constant Service Times

- The single-channel waiting line model that assumes random arrivals but constant service times line can occur in production and manufacturing environments where machine-controlled service times are constant.
- This waiting line is described by the $M/D/1$ model, with the D referring to the deterministic service times.

Multiple-Channel Model With Poisson Arrivals, Arbitrary Service Times, And No Waiting Line

- An interesting variation of the waiting line models involves a system in which no waiting is allowed. Arriving units or customers seek service from one of several service channels.
- If all channels are busy, arriving units are denied access to the system.
- In waiting line terminology, arrivals occurring when the system is full are **blocked** and are cleared from the system.
- Such customers may be lost or may attempt a return to the system later.
- The specific model considered is based on the following assumptions:

1. The system has k channels.
2. The arrivals follow a Poisson probability distribution, with arrival rate λ.

3. The service times for each channel may have any probability distribution.
4. The service rate μ is the same for each channel.

- An arrival enters the system only if at least one channel is available. An arrival occurring when all channels are busy is blocked—that is, denied service and not allowed to enter the system.
- With G denoting a general or unspecified probability distribution for service times, the appropriate model for this situation is referred to as an $M/G/k$ model with "blocked customers cleared."
- The question addressed in this type of situation is, How many channels or servers should be used?
- A primary application of this model involves the design of telephone and other communication systems where the arrivals are the calls and the channels are the number of telephone or communication lines available. In such a system, the calls are made to one telephone number, with each call automatically switched to an open channel if possible.
- When all channels are busy, additional calls receive a busy signal and are denied access to the system.

Operating Characteristics for the M/G/k Model with Blocked Customers Cleared

- We approach the problem of selecting the best number of channels by computing the steady-state probabilities that j of the k channels will be busy. These probabilities are:
- λ: the arrival rate
- μ: the service rate for each channel
- k: the number of channels
- P_j:the probability that j of the k channels are busy for j = 0, 1, 2, . . . , k
- The most important probability value is P_k, which is the probability that all k channels are busy.
- On a percentage basis, P_k indicates the percentage of arrivals that are blocked and denied access to the system.
- Another operating characteristic of interest is the average number of units in the system; note that this number is equivalent to the average number of channels in use.

An Example: Microdata Software, Inc., uses a telephone ordering system for its computer software products. Callers place orders with Microdata by using the company's 800 telephone number. Assume that calls to this telephone number arrive at a rate of $\lambda = 12$ calls per hour. The time required to process a telephone order varies considerably from order to order. However, each Microdata sales representative can be expected to handle $\mu = 6$ calls per hour. Currently, the Microdata 800 telephone number has three internal lines, or channels, each operated by a separate sales representative. Calls received on the 800 number are automatically transferred to an open line, or channel, if available.

Whenever all three lines are busy, callers receive a busy signal. In the past, Microdata's management assumed that callers receiving a busy signal would call back later. However, recent research on telephone ordering showed that a substantial number of callers who are denied access do not call back later. These lost calls represent lost revenues for the firm, so Microdata's management requested an analysis of the telephone ordering system. Specifically, management wanted to know the percentage of callers who get busy signals and are blocked from the system. If management's goal is to provide sufficient capacity to handle 90% of the callers, how many telephone lines and sales representatives should Microdata use?

Waiting Line Models With Finite Calling Populations

- For the waiting line models introduced so far, the population of units or customers arriving for service has been considered to be unlimited.
- In technical terms, when no limit is placed on how many units may seek service, the model is said to have an **infinite calling population.**
- Under this assumption, the arrival rate λ remains constant regardless of how many units are in the waiting line system.
- This assumption of an infinite calling population is made in most waiting line models.
- In other cases, the maximum number of units or customers that may seek service is assumed to be finite.

- In this situation, the arrival rate for the system changes, depending on the number of units in the waiting line, and the waiting line model is said to have a **finite calling population.**
- The formulas for the operating characteristics of the previous waiting line models must be modified to account for the effect of the finite calling population.
- The finite calling population model discussed in this section is based on the following assumptions:

1. The arrivals for *each unit* follow a Poisson probability distribution, with arrival rate λ.
2. The service times follow an exponential probability distribution, with service rate μ.
3. The population of units that may seek service is finite.

- With a single channel, the waiting line model is referred to as an $M/M/1$ model with a finite calling population.
- The arrival rate for the $M/M/1$ model with a finite calling population is defined in terms of how often *each unit* arrives or seeks service.
- This situation differs from that for previous waiting line models in which λ denoted the arrival rate for the system.
- With a finite calling population, the arrival rate for the system varies, depending on the number of units in the system. Instead of adjusting for the changing system arrival rate, in the finite calling population model λ indicates the arrival rate for each unit.
- One of the primary applications of the $M/M/1$ model with a finite calling population is referred to as the *machine repair problem.* In this problem, a group of machines is considered to be the finite population of "customers" that may request repair service.
- Whenever a machine breaks down, an arrival occurs in the sense that a new repair request is initiated. If another machine breaks down before the repair work has been completed on the first machine, the second machine begins to form a "waiting line" for repair service.
- Additional breakdowns by other machines will add to the length of the waiting line.
- The assumption of first-come, first-served indicates that machines are repaired in the order they break down. The $M/M/1$ model shows that one person or one channel is available to perform the repair service. To return the machine to operation, each machine with a breakdown must be repaired by the single-channel operation.

Example The Kolkmeyer Manufacturing Company uses a group of six identical machines; each machine operates an average of 20 hours between breakdowns. Thus, the arrival rate or request for repair service for each machine is $\lambda = \frac{1}{20} = 0.05$ per hour. With randomly occurring breakdowns, the Poisson probability distribution is used to describe the machine breakdown arrival process. One person from the maintenance department provides the single-channel repair service for the six machines. The exponentially distributed service times have a mean of two hours per machine or a service rate of $\mu = \frac{1}{2} = 0.5$ machines per hour. to compute the operating characteristics for this system.