Queuing Theory Models
or
Waiting Lines

Learning Objectives

After completing this chapter, we will be able to:

1. **Describe the trade-off curves for cost-of-waiting time and cost of service.**
2. **Understand the three parts of a queuing system: the calling population, the queue itself, and the service facility.**
3. **Describe the basic queuing system configurations.**
4. **Understand the assumptions of the common models dealt with in this chapter.**
5. **Analyze a variety of operating characteristics of waiting lines.**

Chapter Outline

1. Introduction
2. Waiting Line Costs
3. Characteristics of a Queuing System
4. Single-Channel Queuing Model with Poisson Arrivals and Exponential Service Times (*M/M/1*)

Introduction

- ***Queuing theory*** **is the study of** ***waiting lines.***
- **It is one of the oldest and most widely used quantitative analysis techniques.**
- **The three basic components of a queuing process are arrivals, service facilities, and the actual waiting line.**
- **Analytical models of waiting lines can help managers evaluate the cost and effectiveness of service systems.**

Waiting Line Costs

- Most waiting line problems are focused on finding the ideal level of service a firm should provide.
- In most cases, this service level is something management can control.
- When an organization ***does*** have control, they often try to find the balance between two extremes.

Waiting Line Costs

- There is generally a trade-off between cost of providing service and cost of waiting time.
  - A *large staff* and *many* service facilities generally results in high levels of service but have high costs.
  - Having the *minimum* number of service facilities keeps *service cost* down but may result in dissatisfied customers.
- Service facilities are evaluated on their *total expected cost* which is the sum of *service costs* and *waiting costs.*
- Organizations typically want to find the service level that minimizes the total expected cost.

Queuing Costs and Service Levels

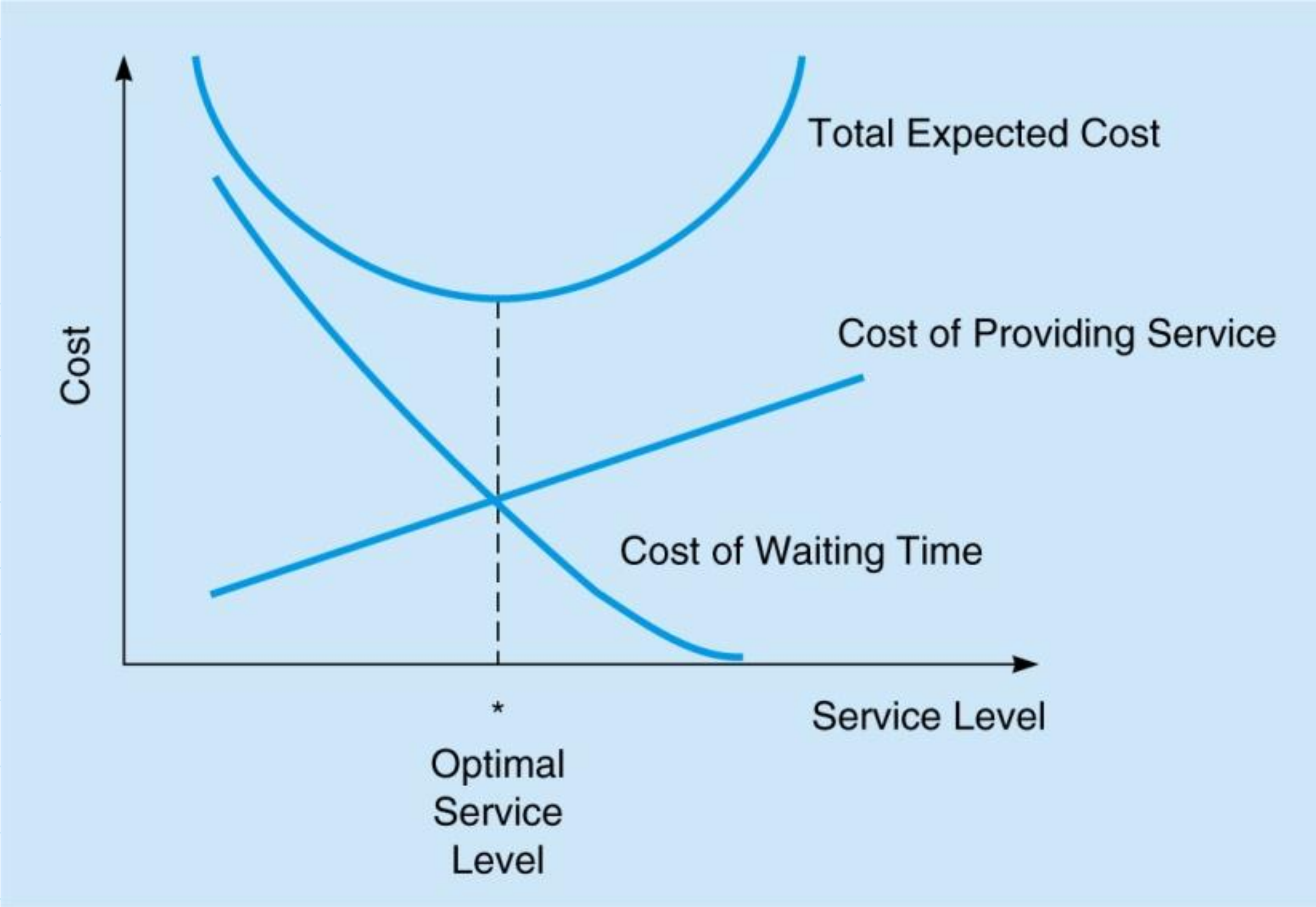

Figure 1

Characteristics of a Queuing System

- **There are three parts to a queuing system:**
  1. **The arrivals or inputs to the system (sometimes referred to as the *calling population*).**
  2. **The queue or waiting line itself.**
  3. **The service facility.**
- **These components have their own characteristics that must be examined before mathematical models can be developed.**

Characteristics of a Queuing System

Arrival Characteristics have three major characteristics: *size*, *pattern*, and *behavior.*

- **The size of the calling population can be either unlimited (essentially *infinite*) or limited (*finite*).**
- **The pattern of arrivals can arrive according to a known pattern or can arrive *randomly.***
  - **Random arrivals generally follow a *Poisson distribution.***

Characteristics of a Queuing System

Behavior of arrivals

- Most queuing models assume customers are patient and will wait in the queue until they are served and do not switch lines.
- ***Balking*** refers to customers who refuse to join the queue.
- ***Reneging*** customers enter the queue but become impatient and leave without receiving their service.
- That these behaviors exist is a strong argument for the use of queuing theory to managing waiting lines.

Characteristics of a Queuing System

Waiting Line Characteristics

- Waiting lines can be either ***limited*** or ***unlimited.***
- Queue discipline refers to the rule by which customers in the line receive service.
  - The most common rule is ***first-in, first-out*** (***FIFO***).
  - Other rules are possible and may be based on other important characteristics.
- Other rules can be applied to select which customers enter which queue, but may apply FIFO once they are in the queue.

Characteristics of a Queuing System

Service Facility Characteristics

- Basic queuing system configurations:
  - Service systems are classified in terms of the number of channels, or servers, and the number of phases, or service stops.
  - A ***single-channel system*** with one server is quite common.
  - ***Multichannel systems*** exist when multiple servers are fed by one common waiting line.
  - In a ***single-phase system,*** the customer receives service form just one server.
  - In a ***multiphase system,*** the customer has to go through more than one server.

Four basic queuing system configurations

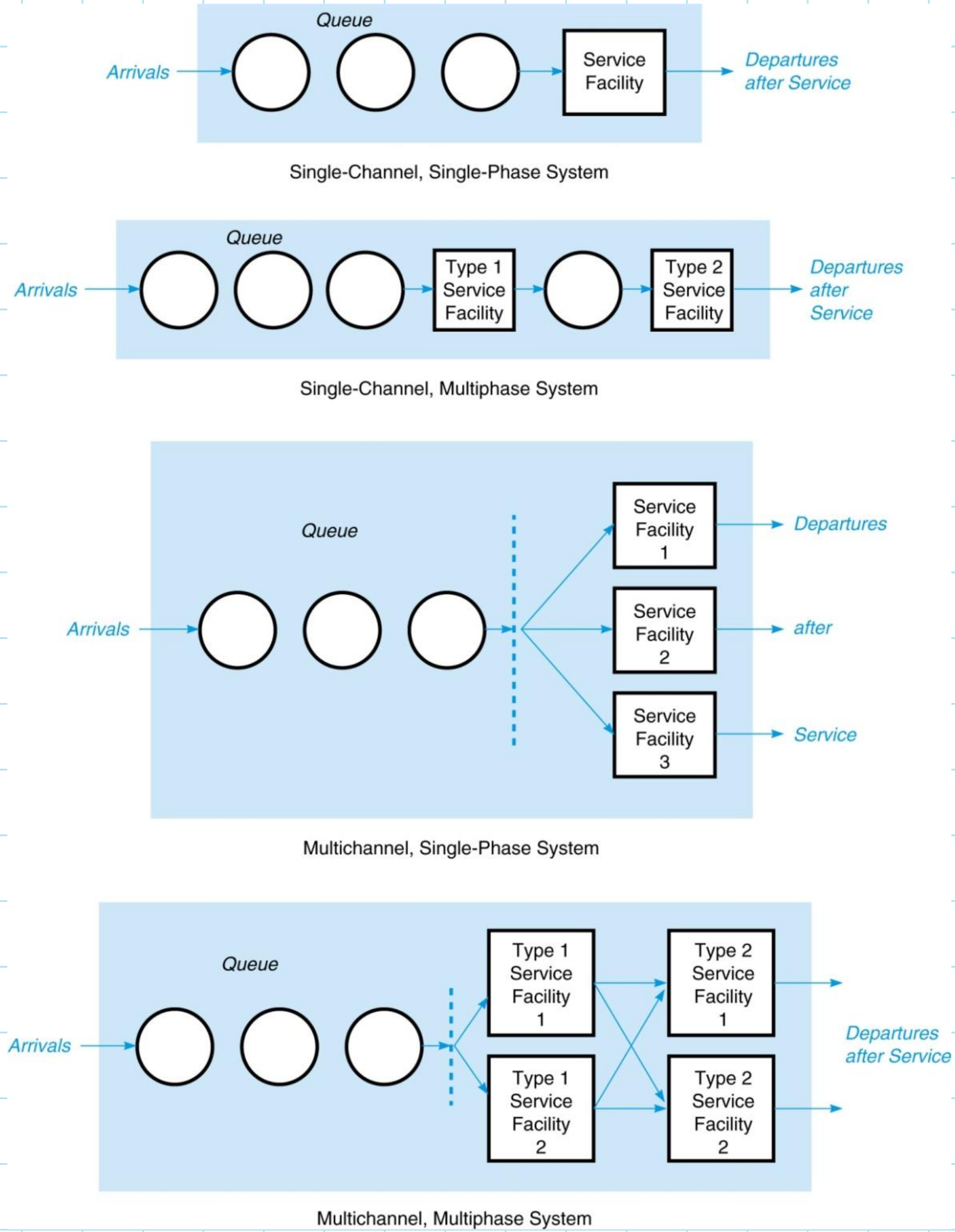

Figure 2

Characteristics of a Queuing System

Service time distribution

- **Service patterns can be either constant or random.**
- **Constant service times are often machine controlled.**
- **More often, service times are randomly distributed according to a *negative exponential probability distribution.***
- **Analysts should observe, collect, and plot service time data to ensure that the observations fit the assumed distributions when applying these models.**

Identifying Models Using Kendall Notation

- D. G. Kendall developed a notation for queuing models that specifies the pattern of arrival, the service time distribution, and the number of channels.
- Notation takes the form:

Arrival distribution	/	Service time distribution	/	Number of service channels open

- Specific letters are used to represent probability distributions.

  M = Poisson distribution for number of occurrences

  D = constant (deterministic) rate

  G = general distribution with known mean and variance

Identifying Models Using Kendall Notation

- A single-channel model with Poisson arrivals and exponential service times would be represented by:

$$M/M/1$$

- If a second channel is added the notation would read:

$$M/M/2$$

- A three-channel system with Poisson arrivals and constant service time would be

$$M/D/3$$

- A four-channel system with Poisson arrivals and normally distributed service times would be

$$M/G/4$$

Single-Channel Model, Poisson Arrivals, Exponential Service Times (M/M/1)

Assumptions of the model:

- Arrivals are served on a FIFO basis.
- There is no balking or reneging.
- Arrivals are independent of each other but the arrival rate is constant over time.
- Arrivals follow a Poisson distribution.
- Service times are variable and independent but the average is known.
- Service times follow a negative exponential distribution.
- Average service rate is greater than the average arrival rate.

Single-Channel Model, Poisson Arrivals, Exponential Service Times (M/M/1)

- **When these assumptions are met, we can develop a series of equations that define the queue's *operating characteristics.***
- **Queuing Equations:**

Let

λ = **mean number of arrivals per time period**

μ = **mean number of customers or units served per time period**

The arrival rate and the service rate must be defined for the same time period.

Single-Channel Model, Poisson Arrivals, Exponential Service Times (M/M/1)

1. The average number of customers or units in the system, L_s :

$$L_s = \frac{\lambda}{\mu - \lambda}$$

2. The average time a customer spends in the system, W_s :

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

3. The average number of customers in the queue, L_q:

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

Single-Channel Model, Poisson Arrivals, Exponential Service Times (M/M/1)

4. **The average time a customer spends waiting in the queue, W_q:**

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

5. **The *utilization factor* for the system, ρ, the probability the service facility is being used:**

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

Single-Channel Model, Poisson Arrivals, Exponential Service Times (M/M/1)

6. The percent idle time, P_0, or the probability no one is in the system:

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

7. The probability that there are *n* number of customers in the system :

$$P_n = \rho^n (1 - \rho)$$

8. The probability that the number of customers in the system is greater than *k*, $P_{n>k}$:

$$P_{n>k} = \left(\frac{\lambda}{\mu}\right)^{k+1}$$

Arnold's Muffler Shop

- Arnold's mechanic can install mufflers at a rate of 3 per hour.
- Customers arrive at a rate of 2 per hour.
- So:

λ = 2 cars arriving per hour

μ = 3 cars serviced per hour

$$L = \frac{\lambda}{\mu - \lambda} = \frac{2}{3-2} = \frac{2}{1} = 2 \text{ cars in the system on average}$$

$$W = \frac{1}{\mu - \lambda} = \frac{1}{3-2} = 1 \text{ hour that an average car spends in the system}$$

Arnold's Muffler Shop

$$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{2^2}{3(3-2)} = \frac{4}{3(1)}$$ = 1.33 cars waiting in line on average

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{2}{3}\text{ hour}$$ = 40 minutes average waiting time per car

$$\rho = \frac{\lambda}{\mu} = \frac{2}{3} = 0.67$$ = percentage of time mechanic is busy

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{2}{3} = 0.33$$ = probability that there are 0 cars in the system

Arnold's Muffler Shop

Probability of more than *k* cars in the system

k	$P_{n>k} = (^2/_3)^{k+1}$	
0	0.667	Note that this is equal to $1 - P_0 = 1 - 0.33 = 0.667$
1	0.444	
2	0.296	
3	0.198	Implies that there is a 19.8% chance that more than 3 cars are in the system
4	0.132	
5	0.088	
6	0.058	
7	0.039	

Single-Channel Model, Poisson Arrivals, Exponential Service Times (M/M/1:N)

Assumptions of the model:

- **Arrivals are served on a FIFO basis.**
- **There is no balking or reneging.**
- **Arrivals are independent of each other but the arrival rate is constant over time.**
- **Arrivals follow a Poisson distribution.**
- **Service times are variable and independent but the average is known.**
- **Service times follow a negative exponential distribution.**
- **Average service rate is greater than the average arrival rate.**
- **Capacity of system is limited to N customers**

Single-Channel Model, Poisson Arrivals, Exponential Service Times (M/M/1:N)

1. **The average number of customers or units in the system, L_S :**

$$L_S = \sum_{n=1}^{N} nP_n = \frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}$$

2. **The average time a customer spends in the system, W_S :**

$$W_S = \frac{L_S}{\lambda(1-P_N)}$$

3. **The average number of customers in the queue, L_q:**

$$L_q = L_S - \rho$$

Single-Channel Model, Poisson Arrivals, Exponential Service Times (M/M/1:N)

4. **The average time a customer spends waiting in the queue, W_q:**

$$W_q = \frac{L_q}{\lambda(1 - P_N)}$$

5. **The *utilization factor* for the system, ρ, the probability the service facility is being used:**

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

Single-Channel Model, Poisson Arrivals, Exponential Service Times (M/M/1:N)

6. **The percent idle time, P_0, or the probability no one is in the system:**

$$P_0 = \frac{1-\rho}{1-\rho^{N+1}}$$

7. **The probability that there are *n* number of customers in the system :**

$$P_n = \frac{(1-\rho)\rho^n}{1-\rho^{N+1}} = P_0\rho^n$$

7. **The probability that the number of customers in the system is greater than *k*, $P_{n>k}$:**

$$P_{n>k} = \left(\frac{\lambda}{\mu}\right)^{k+1}$$