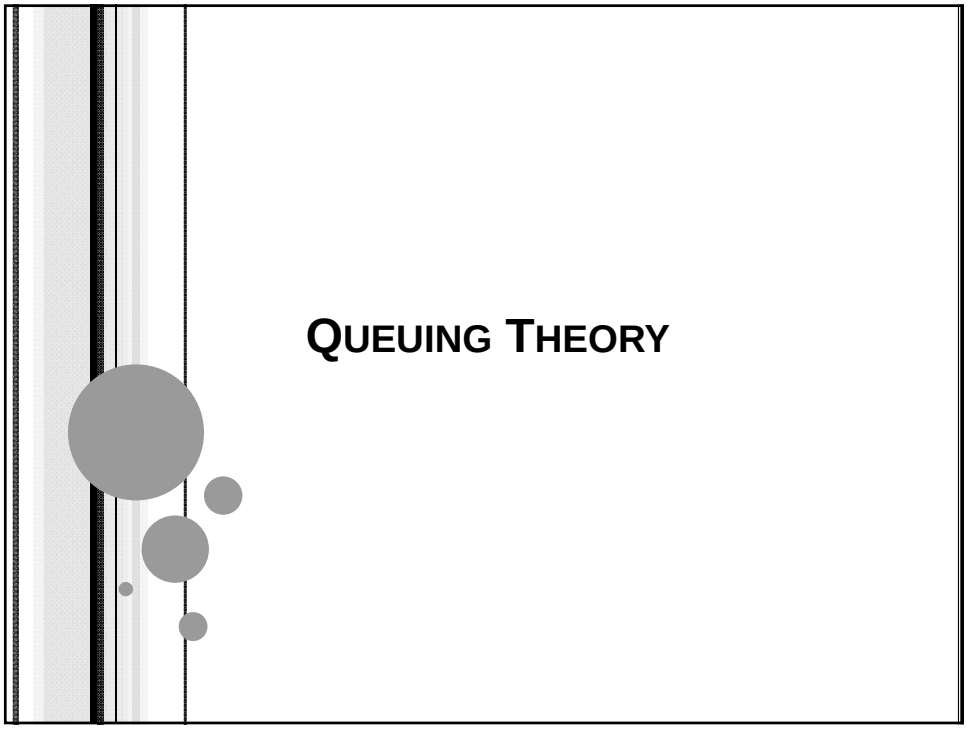


DESCRIPTION

Each of us has spent a great deal of time waiting in lines.

In this chapter, we develop mathematical models for waiting lines, or queues.

LEARNING OBJECTIVES

When you complete this chapter, you should be able to

***Identify or Define*:**

- **The assumptions of the four basic waiting-line models**

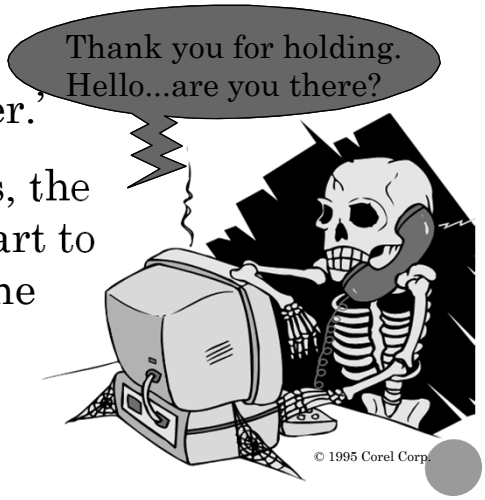
Explain or be able to use:

- **How to apply waiting-line models**
- **How to conduct an economic analysis of queues**

YOU'VE BEEN THERE BEFORE!

'The other line always moves faster.'

'If you change lines, the one you left will start to move faster than the one you're in.'



D-4

SOME QUEUING TERMINOLOGY

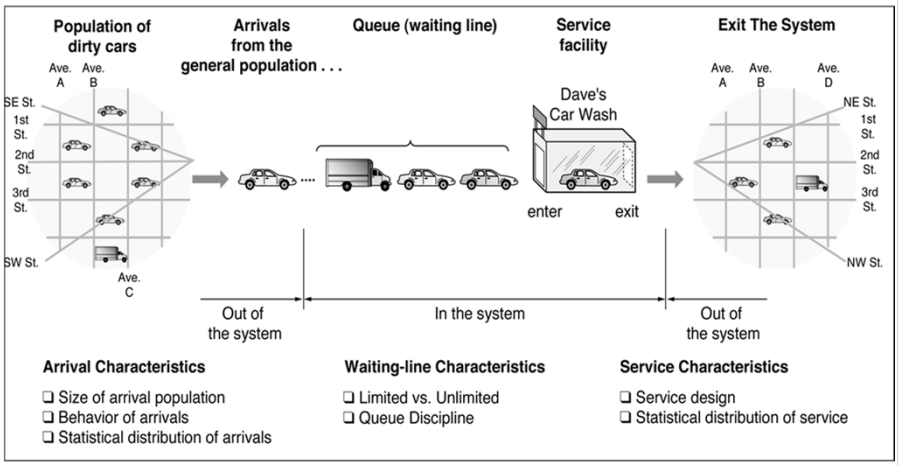
- To describe a queuing system, an input process and an output process must be specified.
- Examples of input and output processes are:

Situation	Input Process	Output Process
Bank	Customers arrive at bank	Tellers serve the customers
Pizza parlor	Request for pizza delivery are received	Pizza parlor send out truck to deliver pizzas

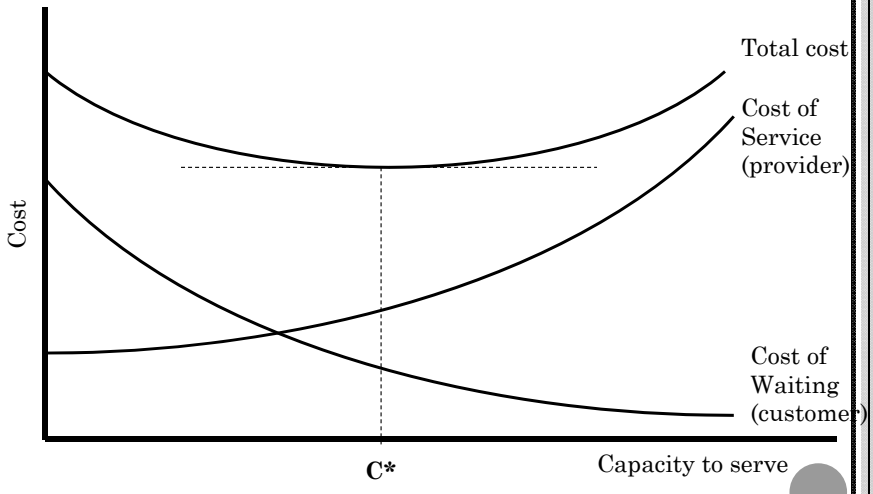
MORE WAITING LINE EXAMPLES

Situation	Arrivals	Servers	Service Process
Bank	Customers	Teller	Deposit etc.
Doctor's office	Patient	Doctor	Treatment
Traffic intersection	Cars	Light	Controlled passage
Assembly line	Parts	Workers	Assembly
Tool crib	Workers	Clerks	Check out/in tools

THREE PARTS OF A QUEUING SYSTEM AT DAVE'S CAR-WASH



CAPACITY DECISIONS: A TRADE-OFF



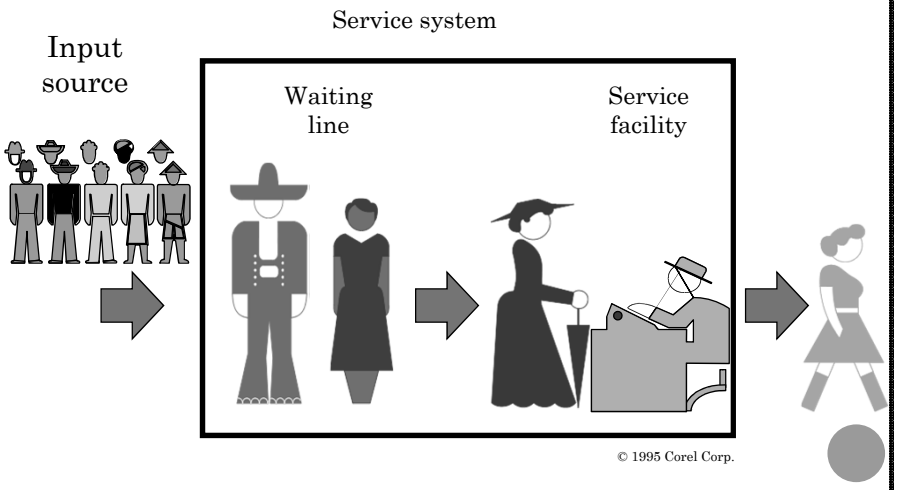
THE INPUT OR ARRIVAL PROCESS

- The input process is usually called the **arrival process**.
- Arrivals are called **customers**.
- We assume that no more than one arrival can occur at a given instant.
- If more than one arrival can occur at a given instant, we say that **bulk arrivals** are allowed.
- Models in which arrivals are drawn from a small population are called **finite source models**.
- If a customer arrives but fails to enter the system, we say that the customer has **balked**.

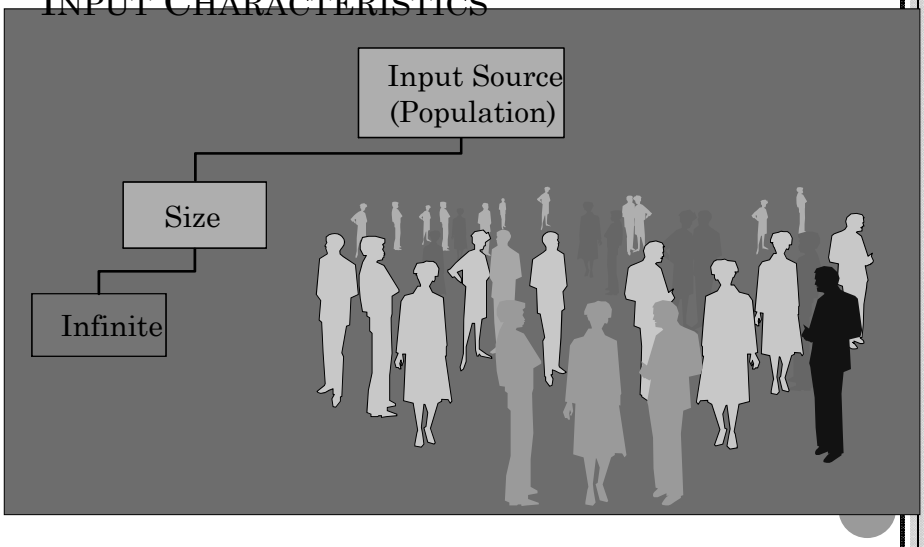
THE OUTPUT OR SERVICE PROCESS

- To describe the output process of a queuing system, we usually specify a probability distribution – the **service time distribution** – which governs a customer's service time.
- We study two arrangements of servers: **servers in parallel** and **servers in series**.
- Servers are in parallel if all servers provide the same type of service and a customer needs only pass through one server to complete service.
- Servers are in series if a customer must pass through several servers before completing service.

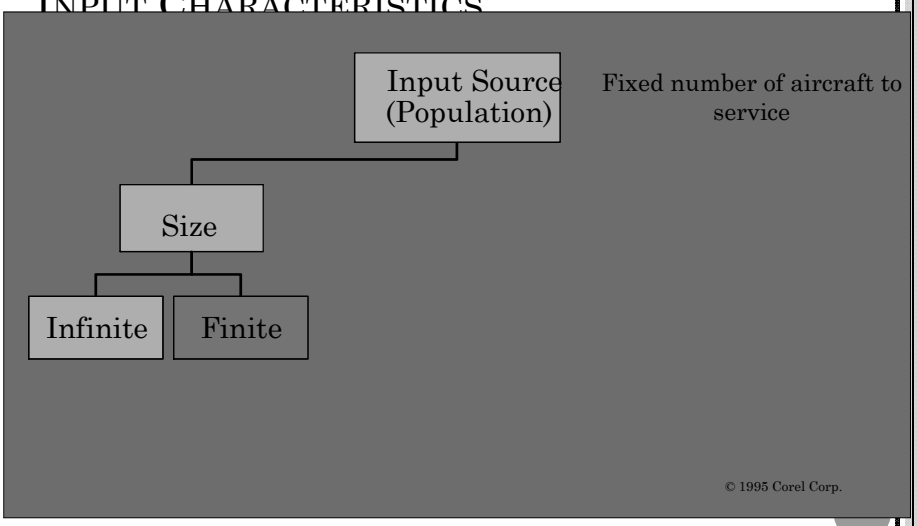
WAITING LINE SYSTEM



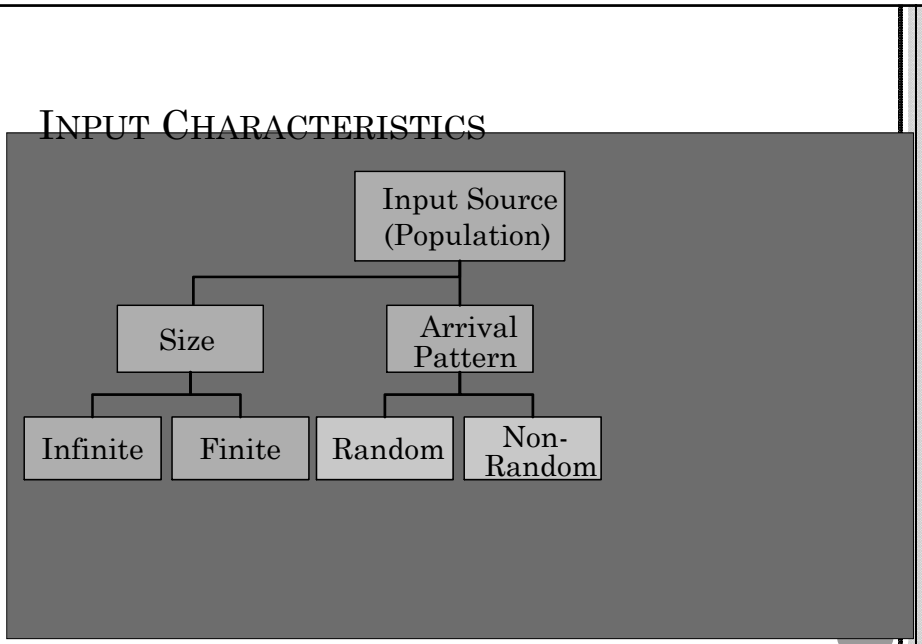
INPUT CHARACTERISTICS



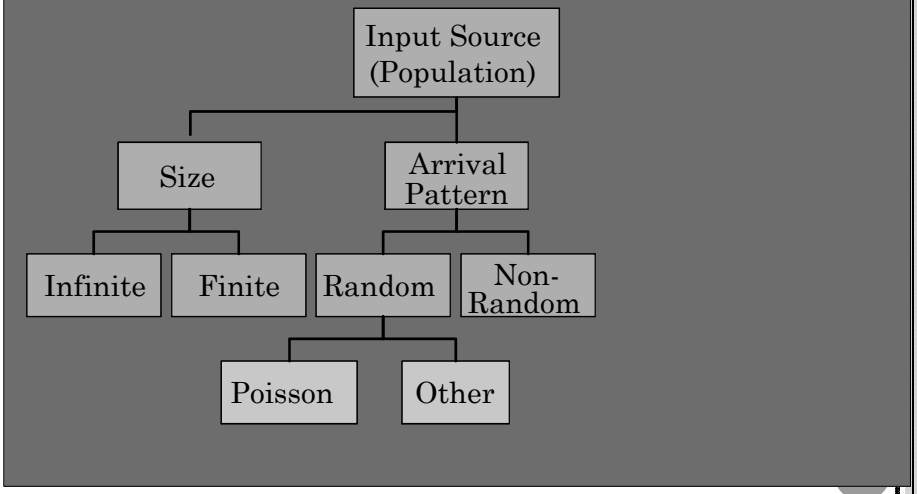
Input Characteristics



Input Characteristics



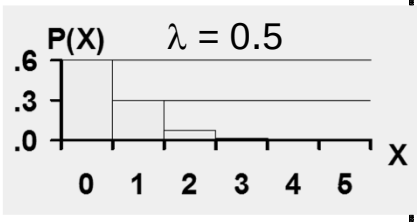
INPUT CHARACTERISTICS

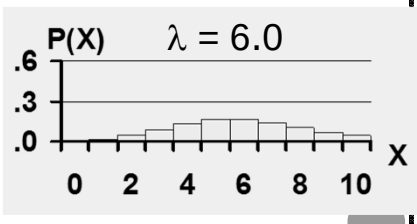


POISSON DISTRIBUTION

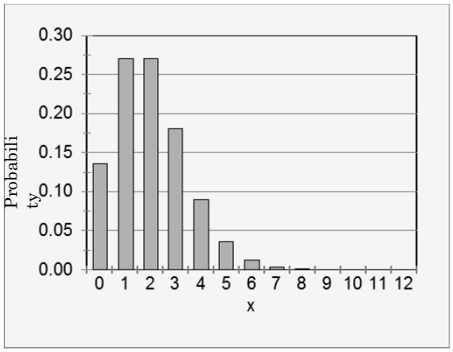
- Number of events that occur in an interval of time
  - Example: Number of customers that arrive in 15 min.
- Mean = λ (e.g., 5/hr.)
- Probability:

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

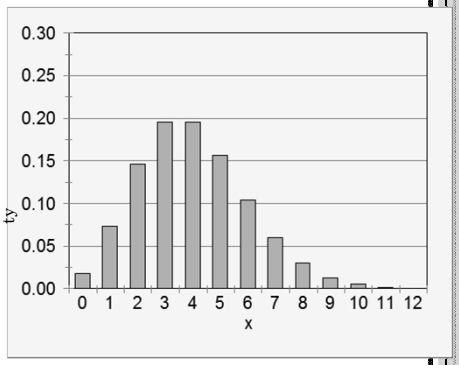




Poisson Distributions for Arrival Times

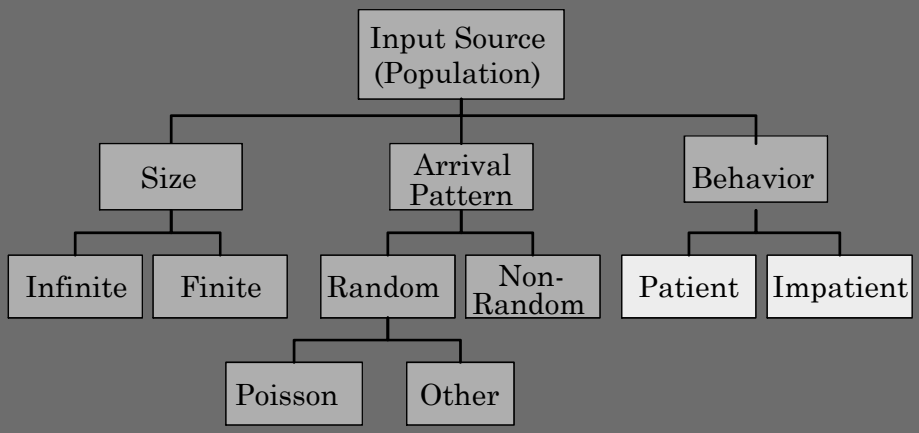


λ=2

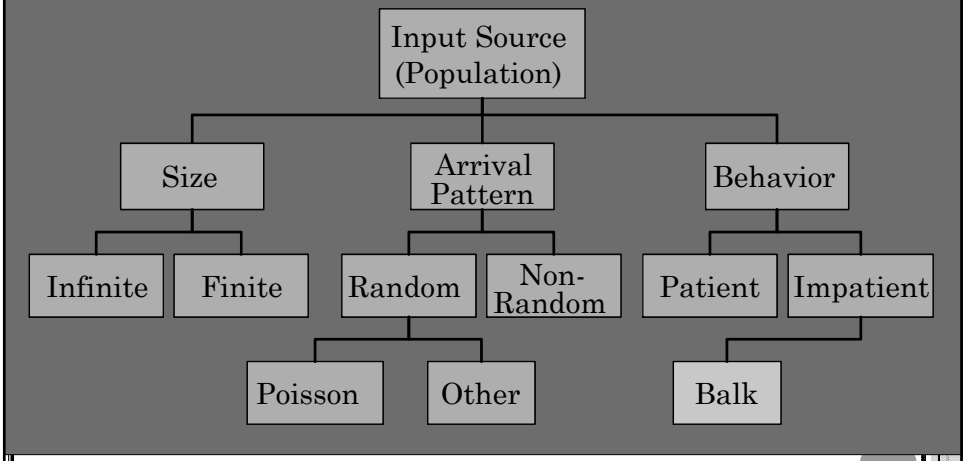


λ=4

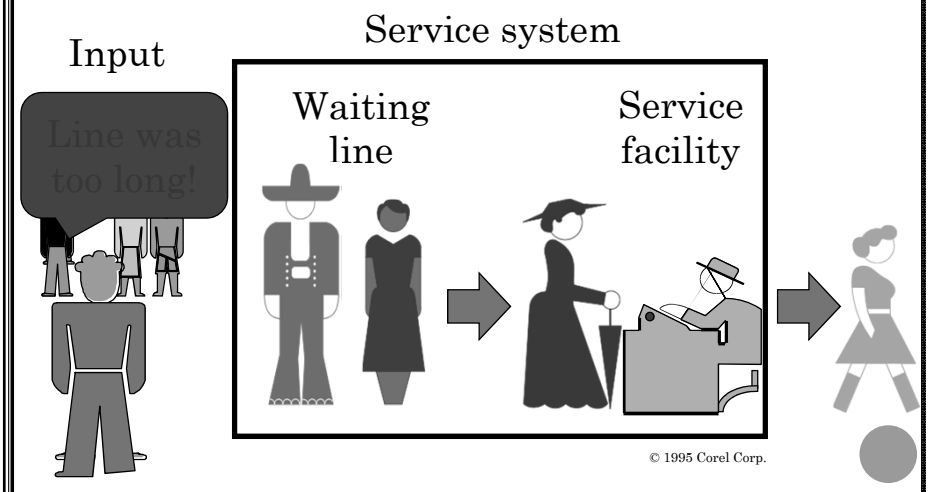
Input Characteristics



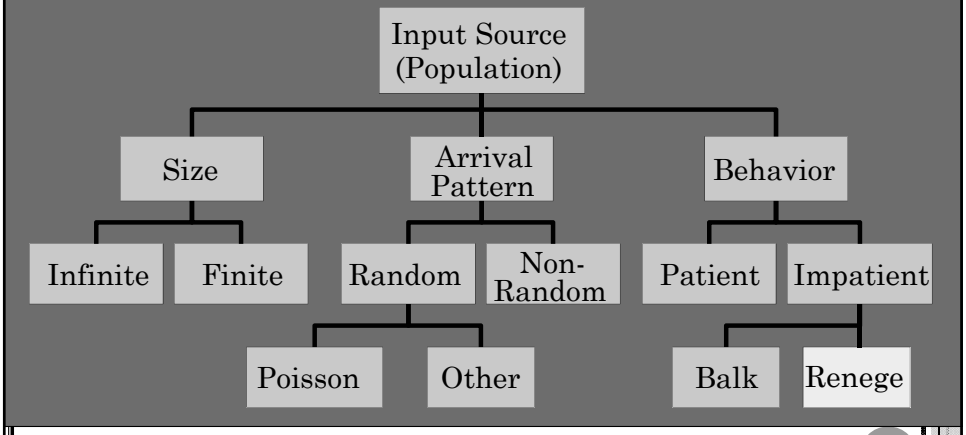
INPUT CHARACTERISTICS



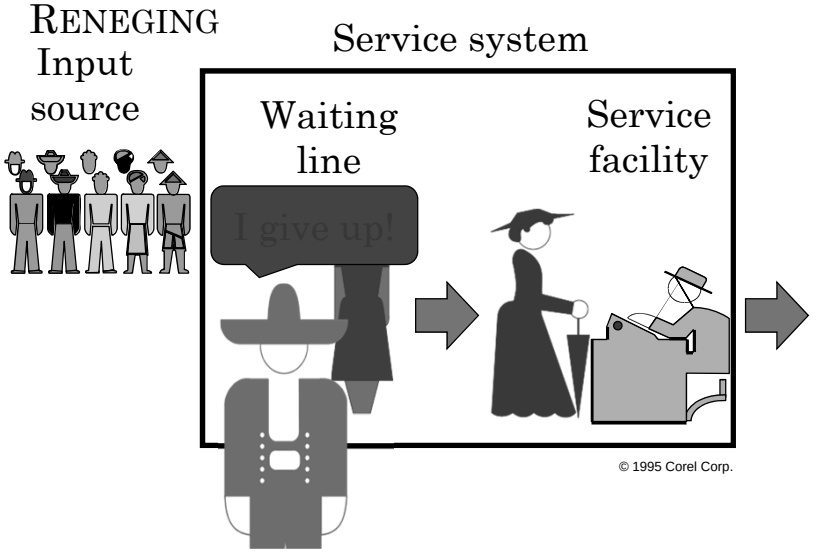
BALKING



INPUT CHARACTERISTICS



RENEGING



QUEUE DISCIPLINE

- The **queue discipline** describes the method used to determine the order in which customers are served.
- The most common queue discipline is the **FCFS discipline** (first come, first served), in which customers are served in the order of their arrival.
- Under the **LCFS discipline** (last come, first served), the most recent arrivals are the first to enter service.
- If the next customer to enter service is randomly chosen from those customers waiting for service it is referred to as the **SIRO discipline** (service in random order).

- Finally we consider **priority queuing disciplines**.
- A priority discipline classifies each arrival into one of several categories.
- Each category is then given a priority level, and within each priority level, customers enter service on a FCFS basis.
- Another factor that has an important effect on the behavior of a queuing system is the method that customers use to determine which line to join.

Modeling Arrival and Service Processes

- We define t_i to be the time at which the ith customer arrives.
- In modeling the arrival process we assume that the T's are independent, continuous random variables described by the random variable **A**.
- The assumption that each interarrival time is governed by the same random variable implies that the distribution of arrivals is independent of the time of day or the day of the week.
- This is the assumption of stationary interarrival times.

- Stationary interarrival times is often unrealistic, but we may often approximate reality by breaking the time of day into segments.
- A negative interarrival time is impossible. This allows us to write

$$P(\mathbf{A} \le c) = \int_0^c a(t)dt \text{ and } P(\mathbf{A} > c) = \int_c^\infty a(t)dt$$

- We define1/λ to be the mean or average interarrival time.

$$\frac{1}{\lambda} = \int_0^\infty ta(t)dt$$

- We define λ to be the **arrival rate**, which will have units of arrivals per hour.
- An important question is how to choose **A** to reflect reality and still be computationally tractable.
- The most common choice for **A** is the **exponential distribution**.
- An exponential distribution with parameter λ has a density $a(t) = \lambda e^{-\lambda t}$.
- We can show that the average or mean interarrival time is given by $E(\mathbf{A}) = \frac{1}{\lambda}$.

- Using the fact that var $\mathbf{A} = E(\mathbf{A}^2) - E(\mathbf{A})^2$, we can show that

$$\text{var}\,\mathbf{A} = \frac{1}{\lambda^2}$$

- Lemma 1: If **A** has an exponential distribution, then for all nonnegative values of t and h,

$$P(\mathbf{A} > t + h \mid \mathbf{A} \geq t) = P(\mathbf{A} > h)$$

- A density function that satisfies the equation is said to have the **no-memory property**.
- The no-memory property of the exponential distribution is important because it implies that if we want to know the probability distribution of the time until the next arrival, then *it does not matter how long it has been since the last arrival.*

Relations between Poisson Distribution and Exponential Distribution

- If interarrival times are exponential, the probability distribution of the number of arrivals occurring in any time interval of length t is given by the following important theorem.
- Theorem 1: Interarrival times are exponential with parameter λ if and only if the number of arrivals to occur in an interval of length t follows the Poisson distribution with parameter λt.

- A discrete random variable **N** has a Poisson distribution with parameter λ if, for n=0,1,2,…,

$$P(\mathbf{N} = n) = \frac{e^{-\lambda}\lambda^n}{n!} \quad (n = 0,1,2,\ldots)$$

- What assumptions are required for interarrival times to be exponential? Consider the following two assumptions:
  - Arrivals defined on nonoverlapping time intervals are independent.
  - For small Δt, the probability of one arrival occurring between times t and $t + \Delta t$ is $\lambda\Delta t + o(\Delta t)$ refers to any quantity satisfying

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$$

- Theorem 2: If assumption 1 and 2 hold, then **N**t follows a Poisson distribution with parameter λt, and interarrival times are exponential with parameter λ; that is, $a(t) = \lambda e^{-\lambda t}$.
- Theorem 2 states that if the arrival rate is stationary, if bulk arrives cannot occur, and if past arrivals do not affect future arrivals, then interarrival times will follow an exponential distribution with parameter λ, and the number of arrivals in any interval of length t is Poisson with parameter λt.

MODELING THE SERVICE PROCESS

- We assume that the service times of different customers are independent random variables and that each customer's service time is governed by a random variable **S** having a density function *s(t)*.
- We let 1/μ be the mean service time for a customer.
- The variable 1/μ will have units of hours per customer, so μ has units of customers per hour. For this reason, we call μ the service rate.
- Unfortunately, actual service times may not be consistent with the no-memory property.

- For this reason, we often assume that $s(t)$ is an Erlang distribution with shape parameters k and rate parameter $k\mu$.
- In certain situations, interarrival or service times may be modeled as having zero variance; in this case, interarrival or service times are considered to be **deterministic**.
- For example, if interarrival times are deterministic, then each interarrival time will be exactly 1/λ, and if service times are deterministic, each customer's service time is exactly 1/μ.

THE KENDALL-LEE NOTATION FOR QUEUING SYSTEMS

- Standard notation used to describe many queuing systems.
- The notation is used to describe a queuing system in which all arrivals wait in a single line until one of *s* identical parallel servers is free. Then the first customer in line enters service, and so on.
- To describe such a queuing system, Kendall devised the following notation.
- Each queuing system is described by six characters:

1/2/3/4/5/6

- The first characteristic specifies the nature of the arrival process. The following standard abbreviations are used:

M = Interarrival times are independent, identically distributed (iid) and exponentially distributed

D = Interarrival times are iid and deterministic

E_k = Interarrival times are iid Erlangs with shape parameter *k*.

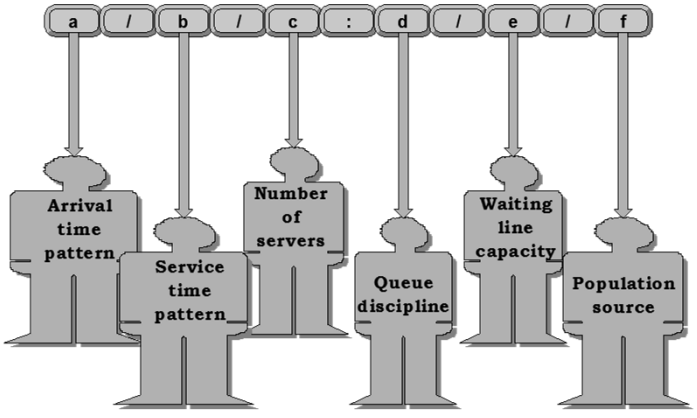
GI = Interarrival times are iid and governed by some general distribution

- The second characteristic specifies the nature of the service times:
  - M = Service times are iid and exponentially distributed
  - D = Service times are iid and deterministic
  - E_k = Service times are iid Erlangs with shape parameter k.
  - G = Service times are iid and governed by some general distribution

- The third characteristic is the number of parallel servers.
- The fourth characteristic describes the queue discipline:
  - FCFS = First come, first served
  - LCFS = Last come, first served
  - SIRO = Service in random order
  - GD = General queue discipline
- The fifth characteristic specifies the maximum allowable number of customers in the system.
- The sixth characteristic gives the size of the population from which customers are drawn.

- In many important models 4/5/6 is $GD/\infty/\infty$. If this is the case, then 4/5/6 is often omitted.
- $M/E_2/8$/FCFS/10/∞ might represent a health clinic with 8 doctors, exponential interarrival times, two-phase Erlang service times, a FCFS queue discipline, and a total capacity of 10 patients.

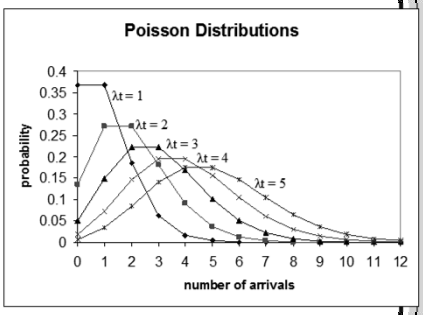
Summary



40

Model Nomenclature & Terms

- 1/2/3 = arrival dist/ service dist/ # of servers
  - Poisson arrivals (M)
    - Distribution of independent arrivals over time
  - Exponential service (M) (time between poisson arrivals)
  - General service (G) (other distributions)



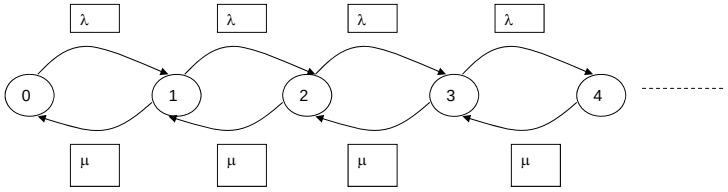
Birth-Death Processes

- We subsequently use birth-death processes to answer questions about several different types of queuing systems.
- We define the number of people present in any queuing system at time t to be the **state** of the queuing systems at time t.
- We call π_j the **steady state**, or equilibrium probability, of state j.
- The behavior of $P_{ij}(t)$ before the steady state is reached is called the **transient behavior** of the queuing system.

- A **birth-death process** is a continuous-time stochastic process for which the system's state at any time is a nonnegative integer.

THE "FLOW-BALANCING APPROACH" (ENTRY-EXIT RATE BALANCING APPROACH)

- In the "rate diagram" given below, think of the following:



- Each circle representing a state (i.e., number of customer in the system) has an unknown probability p_j, j= 0, 1, 2, … associated with it

Derivation of Steady-State Probabilities for Birth-Death Processes

- We now show how the π_j's may be determined for an arbitrary birth-death process.

$$\pi_{j-1}\lambda_{j-1} + \pi_{j+1}\mu_{j+1} = \pi_j(\lambda_j + \mu_j) \quad (j = 1,2,...)$$

$$\pi_1\mu_1 = \pi_0\lambda_0$$

- The above equations are often called the **flow balance equations,** or **conservation of flow equations**, for a birth-death process.

- We obtain the flow balance equations for a birth-death process:

$$\begin{aligned}
&(j = 0) && \pi_0\lambda_0 = \pi_1\mu_1 \\
&(j = 1) && (\lambda_1 + \mu_1)\pi_1 = \lambda_0\pi_0 + \mu_2\pi_2 \\
&(j = 2) && (\lambda_2 + \mu_2)\pi_2 = \lambda_1\pi_1 + \mu_3\pi_3 \\
&\quad \vdots \\
&(j\text{th equation}) && (\lambda_j + \mu_j)\pi_j = \lambda_{j-1}\pi_{j-1} + \mu_{j+1}\pi_{j+1}
\end{aligned}$$

$C_j = (\lambda_0 \lambda_1 \lambda_2 \ldots\ldots \lambda_{j-1})/(\mu_1 \mu_2 \mu_3 \ldots\ldots \mu_j)$

$$\pi_j = \pi_0 c_j \quad (j = 1,2,\ldots)$$

Solution of Birth-Death Flow Balance Equations

- If $\sum_{j=1}^{j=\infty} c_j$ is finite, we can solve for π_0:

$$\pi_0 = \frac{1}{1 + \sum_{j=1}^{j=\infty} c_j}$$

- It can be shown that if $\sum_{j=1}^{j=\infty} c_j$ is infinite, then no steady-state distribution exists.
- The most common reason for a steady-state failing to exist is that the arrival rate is at least as large as the maximum rate at which customers can be served.

THE *M*/*M*/1/*GD*/∞/∞ QUEUING SYSTEM AND THE QUEUING FORMULA L=ΛW

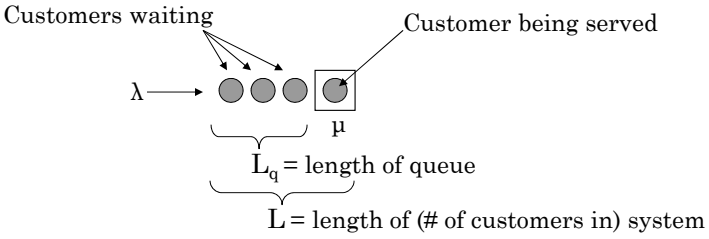
- We define $p = \frac{\lambda}{\mu}$. We call p the **traffic intensity *(utilization)*** of the queuing system.
- We now assume that $0 \le p < 1$ thus

$$\pi_0 = 1 - p \quad (0 \le p < 1)$$

$$\pi_j = p^j (1 - p) \quad (0 \le p < 1)$$

If $p \ge 1$, however, the infinite sum "blows up". Thus, if $p \ge 1$, no steady-state distribution exists.

M/M/1



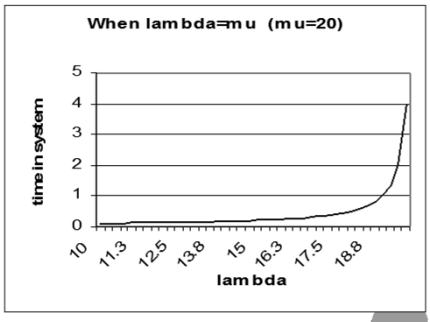
- λ = average customer arrivals per hour
- μ = average service rate per hour

M/M/1

$$\text{Length in System (L)} = \frac{\lambda}{\mu-\lambda}$$

$$\text{Time in System (W)} = \frac{1}{\mu-\lambda}$$

In an effort to maximize efficiency, the manager of the airport tried to get the average time per take off to equal the time between arrivals. What will happen?



DERIVATION OF *L*

- Throughout the rest of this section, we assume that $p<1$, ensuring that a steady-state probability distribution does exist.
- The steady state has been reached, the average number of customers in the queuing system (call it *L*) is given by

$$L = \sum_{j=0}^{j=\infty} j\pi_j = \sum_{j=0}^{j=\infty} jp^j(1-p)$$

$$= (1-p)\sum_{j=0}^{j=\infty} jp^j$$

and

$$L = (1-p)\frac{p}{(1-p)^2} = \frac{p}{1-p} = \frac{\lambda}{\mu-\lambda}$$

DERIVATION OF L_Q

- In some circumstances, we are interested in the expected number of people waiting in line (or in the queue).
- We denote this number by L_q.

$$L_q = \frac{p}{1-p} - p = \frac{p^2}{1-p} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

$$L_q = \sum_{j=1}^{j=\infty} (j-1)\pi_j$$

DERIVATION OF L_S

- Also of interest is *Ls*, the expected number of customers in service.

$$L_s = 0\pi_0 + 1(\pi_1 + \pi_2 + \cdots) = 1 - \pi_0 = 1 - (1-p) = p$$

$$L_q = L - L_s = \frac{p}{1-p} - p = \frac{p^2}{1-p}$$

$$\pi_j = \frac{\lambda+\mu}{\mu}\left(\frac{\lambda}{\mu}\right)^{j-1}\pi_0$$

THE QUEUING FORMULA $L=\lambda W$

- We define W as the expected time a customer spends in the queuing system, including time in line plus time in service, and W_q as the expected time a customer spends waiting in line.
- By using a powerful result known as **Little's queuing formula**, W and W_q may be easily computed from L and L_q.
- We first define the following quantities L
  - λ = average number of arrivals *entering* the system per unit time

- L = average number of customers present in the queuing system
- L_q = average number of customers waiting in line
- L_s = average number of customers in service
- W = average time a customer spends in the system
- W_q = average time a customer spends in line
- W_s = average time a customer spends in service

- **Theorem 3** – For any queuing system in which a steady-state distribution exists, the following relations hold:

$$L = \lambda W$$
$$L_q = \lambda W_q$$
$$L_s = \lambda W_s$$

EXAMPLE 4

- Suppose that all car owners fill up when their tanks are exactly half full.
- At the present time, an average of 7.5 customers per hour arrive at a single-pump gas station.
- It takes an average of 4 minutes to service a car.
- Assume that interarrival and service times are both exponential.
  1. For the present situation, compute L and W.

2. Suppose that a gas shortage occurs and panic buying takes place.
   - To model the phenomenon, suppose that all car owners now purchase gas when their tank are exactly three-fourths full.
   - Since each car owner is now putting less gas into the tank during each visit to the station, we assume that the average service time has been reduced to 3 1/3 minutes.
   - How has panic buying affected L and W?

SOLUTIONS

1. We have an $M/M/1/GD/\infty/\infty$ system with $\lambda = 7.5$ cars per hour and $\mu = 15$ cars per hour. Thus $p = 7.5/15 = .50$. $L = .50/1-.50 = 1$, and $W = L/\lambda = 1/7.5 = 0.13$ hour. Hence, in this situation, everything is under control, and long lines appear to be unlikely.
2. We now have an $M/M/1/GD/\infty/\infty$ system with $\lambda = 2(7.5) = 15$ cars per hour. Now $\mu = 60/3.333 = 18$ cars per hour, and $p = 15/18 = 5/6$. Then

   Thus, panic buying has cause long lines.

$$L = \frac{\frac{5}{6}}{1-\frac{5}{6}} = 5 \text{ cars and } W = \frac{L}{\lambda} = \frac{5}{15} = \frac{1}{3} \text{ hours} = 20 \text{ minutes}$$

- Problems in which a decision maker must choose between alternative queuing systems are called **queuing optimization problems**.

MORE ON $L = \lambda W$

- The queuing formula $L = \lambda W$ is very general and can be applied to many situations that do not seem to be queuing problems.
  - L = average amount of quantity present.
  - λ = Rate at which quantity arrives at system.
  - W = average time a unit of quantity spends in system.
- Then $L = \lambda W$ *or* $W = L/\lambda$

A SIMPLE EXAMPLE

- Example
  - Our local MacDonalds' receives an average of 10,000 pounds of potatoes per week.
  - The average number of pounds of potatoes on hand is 5000 pounds.
  - On the average, how long do potatoes stay in the restaurant before being used?
- Solution
  - We are given that L=5000 pounds and λ = 10,000 pounds/week. Therefore W = 5000 pounds/(10,000 pounds/week)=.5 weeks.

A Queueing Model Optimization

- Problems in which a decision maker must choose between alternative queueing systems
- **Example**: An average of 10 machinists per hour arrive seeking tools. At present, the tool center is staffed by a clerk who is paid $6 per hour and who takes an average of 5 minutes to handle each request for tools. Since each machinist produces $10 worth of goods per hour, each hour that a machinists spends at the tool center costs the company $10. The company is deciding whether or not it is worthwhile to hire (at $4 per hour) a helper for the clerk. If the helper is hired the clerk will take an average of only 4 minutes to process requirements for tools. Assume that service and arrival times are exponential. Should the helper be hired?

A Queueing Model Optimization

- Goal: Minimize the sum of the hourly service cost and expected hourly cost due to the idle times of machinists
- Delay cost is the component of cost due to customers waiting in line
- Goal: Minimize Expected cost/hour = service cost/hour + expected delay cost/hour
- Expected delay cost/hour = (expected delay cost/customer) (expected customers/hour)
- Expected delay cost/customer = ($10/machinist-hour)(average hours machinist spends in the system) = 10W
- Expected delay cost/hour = $10W\lambda$
- Now compute expected cost/hour if the helper is not hired and also compute the same if the helper is hired

A QUEUEING MODEL OPTIMIZATION

- If the helper is not hired $\lambda = 10$ machinists per hour and $\mu = 12$ machinists per hour
- $W = 1/(\mu-\lambda)$ for M/M/1/GD/∞/∞. Therefore, $W = 1/(12-10) = ½ = 0.5$ hour
- Service cost /hour = \$6/hour and expected delay cost/hour = 10(0.5)(10) = \$50
- Without the helper, the expected hourly cost is \$6 + \$50 = \$56
- With the helper, $\mu = 15$ customers/hour. Then $W = 1/(\mu-\lambda)= 1/(15-10) = 0.2$ hour and the expected delay cost/hour = 10(0.2)(10) = \$20
- Service cost/hour = \$6 + \$4 = \$10/hour
- With the helper, the expected hourly cost is \$10 + \$20 = \$30

MSOffic

EXAMPLE (BASIC MODEL)

A tugboat serves ships arriving in a harbor. The average time between ship arrivals is 2.5 hours. The average time required to provide service (i.e. tow a ship to its berth) is 2.0 hours. Studies have shown that arrival rates are approximately Poisson distributed and service time is exponential distributed

66

EXAMPLE – CONTINUED..

- System parameters:
  - Single-channel, single-phase system
  - Arrival rate (*lambda*) = 0.4 ships per hour
  - Service rate (*mu*) = 0.5 ships per hour
  - Number of servers (*s*) = 1

67

EXAMPLE – CALCULATE THE FOLLOWING

1. Utilization rate of the towing service
2. Percentage of the time the service is idle
3. Probability of having one ship in the system
4. Probability of having 2 ships in the system
5. The average number of ships waiting to be towed
6. The average number of ships waiting to be towed and being towed
7. The average amount of time a ships spends waiting to be towed
8. The average amount of time a ships spends in the towing service (waiting to be towed and being towed).

68

SOLUTIONS:

1. Utilization rate of the towing service

$$p = \frac{\lambda}{\mu} = \frac{0.4}{0.5} = 0.80$$

2. Percentage of the time the service is idle

$$1 - p = 0.2$$

4. Probability of having 2 ships in the system

$$\pi_2 = 0.2\left(\frac{0.4}{0.5}\right)^2 = 0.128$$

69

SOLUTIONS: CONTINUED....

5. The average number of ships waiting to be towed

$$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{0.4^2}{0.5(0.5-0.4)} = 3.20$$

6. The average number of ships waiting to be towed and being towed

$$L = \frac{\lambda}{(\mu-\lambda)} = \frac{0.4}{(0.5-0.4)} = 4.0$$

7. The average amount of time a ship spends waiting to be towed

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{0.4}{0.5(0.5-0.4)} = 8.0$$

70

SOLUTIONS: CONTINUED.....

8. The average amount of time a ships spends in the towing service (waiting to be towed and being towed).

$$W_{\square} = \frac{1}{(\mu - \lambda)} = \frac{1}{(0.5 - 0.4)} = 10$$

71

EXAMPLE: SJJT, INC. (A)

- *M*/*M*/1 Queuing System

  Joe Ferris is a stock trader on the floor of the New York Stock Exchange for the firm of Smith, Jones, Johnson, and Thomas, Inc. Stock transactions arrive at a mean rate of 20 per hour. Each order received by Joe requires an average of two minutes to process.

  Orders arrive at a mean rate of 20 per hour or one order every 3 minutes. Therefore, in a 15 minute interval the average number of orders arriving will be $\lambda = 15/3 = 5$.

EXAMPLE: SJJT, INC. (A)

- Arrival Rate Distribution

Question

What is the probability that no orders are received within a 15-minute period?

Answer

$$P(x = 0) = (5^0 e^{-5})/0! = e^{-5} = .0067$$

EXAMPLE: SJJT, INC. (A)

- Arrival Rate Distribution

Question

What is the probability that exactly 3 orders are received within a 15-minute period?

Answer

$$P(x = 3) = (5^3 e^{-5})/3! = 125(.0067)/6 = .1396$$

EXAMPLE: SJJT, INC. (A)

- Arrival Rate Distribution

Question

What is the probability that more than 6 orders arrive within a 15-minute period?

Answer

$$P(x > 6) = 1 - P(x = 0) - P(x = 1) - P(x = 2) - P(x = 3) - P(x = 4) - P(x = 5) - P(x = 6)$$
$$= 1 - .762 = .238$$

EXAMPLE: SJJT, INC. (A)

- Service Rate Distribution

Question

What is the mean service rate per hour?

Answer

Since Joe Ferris can process an order in an average time of 2 minutes (= 2/60 hr.), then the mean service rate, μ, is μ = 1/(mean service time), or 60/2.

$$\mu = 30/\text{hr.}$$

EXAMPLE: SJJT, INC. (A)

- Service Time Distribution

Question

What percentage of the orders will take less than one minute to process?

Answer

Since the units are expressed in hours,

$P(T \leq 1 \text{ minute}) = P(T \leq 1/60 \text{ hour})$.

Using the exponential distribution, $P(T \leq t) = 1 - e^{-\mu t}$.

Hence, $P(T \leq 1/60) = 1 - e^{-30(1/60)}$

$= 1 - .6065 = .3935 =$ 39.35%

EXAMPLE: SJJT, INC. (A)

- Service Time Distribution

Question

What percentage of the orders will be processed in exactly 3 minutes?

Answer

Since the exponential distribution is a continuous distribution, the probability a service time exactly equals any specific value is 0 .

EXAMPLE: SJJT, INC. (A)

- Service Time Distribution

Question

What percentage of the orders will require more than 3 minutes to process?

Answer

The percentage of orders requiring more than 3 minutes to process is:

$P(T > 3/60) = e^{-30(3/60)} = e^{-1.5} = .2231 =$ 22.31%

EXAMPLE: SJJT, INC. (A)

- Average Time in the System

Question

What is the average time an order must wait from the time Joe receives the order until it is finished being processed (i.e. its turnaround time)?

Answer

This is an $M/M/1$ queue with $\lambda = 20$ per hour and $\mu = 30$ per hour. The average time an order waits in the system is: $W = 1/(\mu - \lambda)$

$= 1/(30 - 20)$

$= 1/10$ hour or 6 minutes

EXAMPLE: SJJT, INC. (A)

- Average Length of Queue

Question

What is the average number of orders Joe has waiting to be processed?

Answer

Average number of orders waiting in the queue is:

$$L_q = \lambda^2/[\mu(\mu - \lambda)]$$
$$= (20)^2/[(30)(30-20)]$$
$$= 400/300$$
$$= 4/3$$

EXAMPLE: SJJT, INC. (A)

- Utilization Factor

Question

What percentage of the time is Joe processing orders?

Answer

The percentage of time Joe is processing orders is equivalent to the utilization factor, λ/μ. Thus, the percentage of time he is processing orders is:

$$\lambda/\mu = 20/30$$
$$= 2/3 \text{ or } 66.67\%$$

THE $M/M/s/GD/\infty/\infty$ QUEUING SYSTEM

- We now consider the $M/M/s/GD/\infty/\infty$ system.
- We assume that interarrival times are exponential (with rate λ), service times are exponential (with rate μ), and there is a single line of customers waiting to be served at one of the s parallel servers.
- If $j \leq s$ customers are present, then all j customers are in service; if $j > s$ customers are present, then all s servers are occupied, and $j - s$ customers are waiting in line.

j	λ_j	μ_j
0	λ	0
1	λ	μ
2 .	λ .	2μ .
s	λ	$s\mu$
s+1 . ∞	λ . λ	$s\mu$. $s\mu$

- Summarizing, we find that the $M/M/s/GD/\infty/\infty$ system can be modeled as a birth-death process with parameters

$$\lambda_j = \lambda \quad (j = 0,1,\ldots)$$
$$\mu_j = j\mu \quad (j = 0,1,\ldots,s)$$
$$\mu_j = s\mu \quad (j = s+1, s+2,\ldots)$$

we define $p=\lambda / s\mu$. For $p<1$, the following steady-state probabilities

AN M/M/S QUEUEING OPTIMIZATION EXAMPLE

$$\pi_0 = \frac{1}{\sum_{i=0}^{(s-1)} \frac{(p)^i}{i!} + \frac{(p)^s}{s!(1-p/s)}}$$

$$\pi_j = \frac{(p)^j \pi_0}{j!} \quad (\mathrm{j} = 1,2,\ldots., \mathrm{s})$$

$$\pi_{\mathrm{j}} = \frac{(p)^j \pi_0}{s! s^{j-s}} \quad (\mathrm{j} = \mathrm{s}, \mathrm{s}+1, \mathrm{s}+2\ldots..)$$

$$\mathrm{P}(\mathrm{j} \geq \mathrm{s}) = \frac{(p)^s \pi_0}{s!(1-p/s)}$$

$$L_q = \frac{p^{s+1}}{(s-1)!(s-p)^2}\ \pi_0, \quad W_q = L_q/\lambda$$

$$L = L_q + \lambda/\mu$$

$$W = L/\lambda = L_q/\lambda + 1/\mu = W_q + 1/\mu$$

- A crew of mechanics at the Highway Department garage repair vehicles that break down at an average of λ = 8 vehicles per day (approximately Poisson in nature). The mechanic crew can service an average of μ= 11 vehicles per day with a repair time distribution that approximates an exponential distribution. The crew cost is approximately $300 per day. The cost associated with lost productivity from the breakdown is estimated at $150 per vehicle per day (or any fraction thereof). Which is cheaper, the existing system with one service crew, or a revised system with two service crews?
- **Answer:**
- L for the single server is 8 / (11-8) = 8/3 = 2.667. The single-server system server cost is $300 per day; wait cost is $150 x 2.667 = $400, for a total of $700.
- For the two-server system, L = 0.8381. The two-server system will double the server cost to $600, but reduce the wait cost to $150 x .8381 = $125.72, for a total of $725.72.
- The single-server system is cheaper.

THE *M/M/1/GD/c/∞* QUEUING SYSTEM

- The *M/M/1/GD/c/∞* queuing system is identical to the *M/M/1/GD/∞/∞* system except for the fact that when *c* customers are present, all arrivals are turned away and are forever lost to the system.

$$W = \frac{L}{\lambda(1-\pi_c)} \quad \text{and} \quad W_q = \frac{L_q}{\lambda(1-\pi_c)}$$

$$\bar{\lambda} = \lambda(1-\pi_c) = \sum_{j=0}^{\infty} \lambda_j \pi_j$$

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

$\overline{\lambda} =$ Effective Arrival Rate

j	λ_j	μ_j
0	λ	0
1	λ	μ
.	.	μ
.	.	μ
c	0	μ

$$\pi_j = \left(\frac{\lambda}{\mu}\right)^j \pi_0$$

$$\pi_0 = \frac{1-p}{1-p^{c+1}}, ifp \neq 1$$

$$\pi_0 = \frac{1}{c+1}, ifp = 1$$

- For the $M/M/1/GD/c/\infty$ system, a steady state will exist even if $\lambda \geq \mu$.
- This is because, even if $\lambda \geq \mu$, the finite capacity of the system prevents the number of people in the system from "blowing up".

$$
\begin{aligned}
L &= \sum_{j=0}^{c} j\,\pi_j \\
&= \sum_{j=0}^{c} j\,p^j\,\pi_0 \\
&= \pi_0\,p\sum_{j=1}^{c} j\,p^{j-1} \\
&= \pi_0\,p\sum_{j=1}^{c} \frac{d\,p^j}{d\,p} \\
&= \pi_0\,p\,\frac{d\sum_{j=1}^{c} p^j}{d\,p} \\
&= \frac{1-p}{1-p^{c+1}}\,p\,\frac{1-(c+1)\,p^c+c\,p^{c+1}}{(1-p)^2} \\
&= \frac{p\,(1-(c+1)\,p^c+c\,p^{c+1})}{(1-p)(1-p^{c+1})} \quad \text{if} \quad p \neq 1
\end{aligned}
$$

Example 15.6-2

Automata car wash facility operates with only one bay. Cars arrive according to a Poisson distribution with a mean of 4 cars per hour, and may wait in the facility's parking lot if the bay is busy. The time for washing and cleaning a car is exponential, with a mean of 10 minutes. Cars that cannot park in the lot can wait in the street bordering the wash facility. This means that, for all practical purposes, there is no limit on the size of the system. The manager of the facility wants to determine the size of the parking lot.

For this situation, we have $\lambda = 4$ cars per hour, and $\mu = \frac{60}{10} = 6$ cars per hour. Because $\rho = \frac{\lambda}{\mu} < 1$, the system can operate under steady-state conditions.

The TORA or excelPoissonQ.xls input for this model is

Lambda	Mu	c	System limit	Source limit
4	6	1	infinity	infinity

Example 15.6-4

Consider the car wash facility of Example 15.6-2. Suppose that the facility has a total of four parking spaces. If the parking lot is full, newly arriving cars balk to other facilities. The owner wishes to determine the impact of the limited parking space on losing customers to the competition.

In terms of the notation of the model, the limit on the system is $N = 4 + 1 = 5$. The following input data provides the output in Figure 15.6.

Lambda	Mu	c	System limit	Source limit
4	6	1	5	infinity

Because the limit on the system is $N = 5$, the proportion of lost customers is $p_5 = .04812$, which, based on a 24-hour day, is equivalent to losing $(\lambda p_5) \times 24 = 4 \times .04812 \times 24 = 4.62$ cars a day. A decision regarding increasing the size of the parking lot should be based on the value of lost business.

Looking at the problem from a different angle, the expected total time in the system, W_s, is .3736 hour, or approximately 22 minutes, down from 30 minutes in Example 15.6-3 when all arriving cars are allowed to join the facility. This reduction of about 25% is secured at the expense of losing about 4.8% of all potential customers because of the limited parking space.

Scenario 1: (M/M/1) : (GD/5/infinity)

Lambda =	4.00000	Mu =	6.00000
Lambda eff =	3.80752	Rho/c =	0.66667
Ls =	1.42256	Lq =	0.78797
Ws =	0.37362	Wq =	0.20695

n	Probability pn	Cumulative Pn	n	Probability pn	Cumulative Pn
0	0.36541	0.36541	3	0.10827	0.87970
1	0.24361	0.60902	4	0.07218	0.95188
2	0.16241	0.77143	5	0.04812	1.00000

FIGURE 15.6

TORA output of Example 15.6-4 (file toraEx15.6-4.txt)