# Ozcan: Chapter 14.
# Queuing Models and Capacity Planning

ISE 468 ETM 568

Spring 2015 Dr. Burtner

# Outline

- **Queuing System Characteristics**
  - **Population Source**
  - **Servers**
  - **Arrival Patterns**
  - **Service Patterns**
- **Specific Queue Characteristics**
- **Measures of Queuing System Performance**
- **Infinite Source-Models**
- **Model Formulations**
  - **Single Server (M/M/1)**
  - **Multi Server (M/M/s>1)**

# Queuing Models

- A mathematical approach to the analysis of waiting lines
- Weighs the cost of providing a given level of service capacity (i.e., shortening wait times) against the potential costs of having customers

# Costs

- Waiting Costs
  - Salaries paid to employees while they wait for service (e.g., physicians waiting for an x-ray or test result)
  - Cost of waiting space
  - Loss of business due to wait (balking customers)
- Capacity Costs--  the cost of maintaining the ability to provide service

# The Queuing Model

- Main Characteristics
  - The population source
  - The number of servers (channels)
  - Arrival and service patterns
  - Queue discipline (order of service)

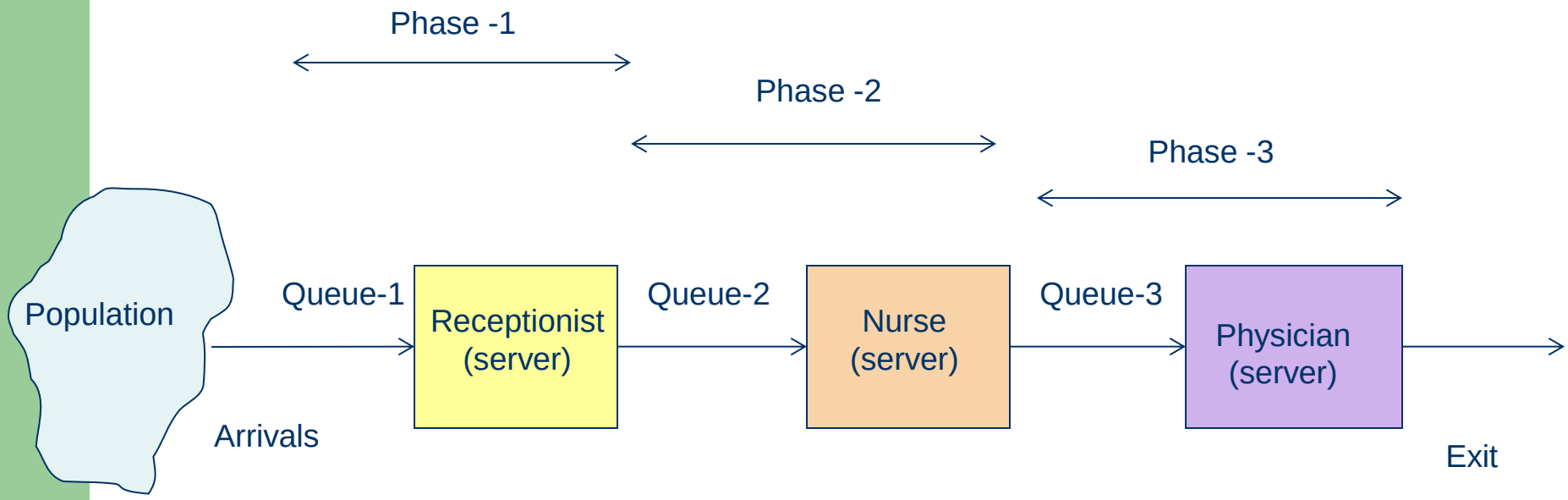
## The Population Source

- Infinite
  - Customer arrivals are unrestricted
  - Customer arrivals exceed system capacity
  - Exists where service is unrestricted
- Finite source
  - Customer population where potential number of customers is limited

## The Number of Servers

- Capacity is a function of the capacity of each server and the number of servers being used
- Servers are also called channels
- Systems can be single or multiple channel, and consist of phases

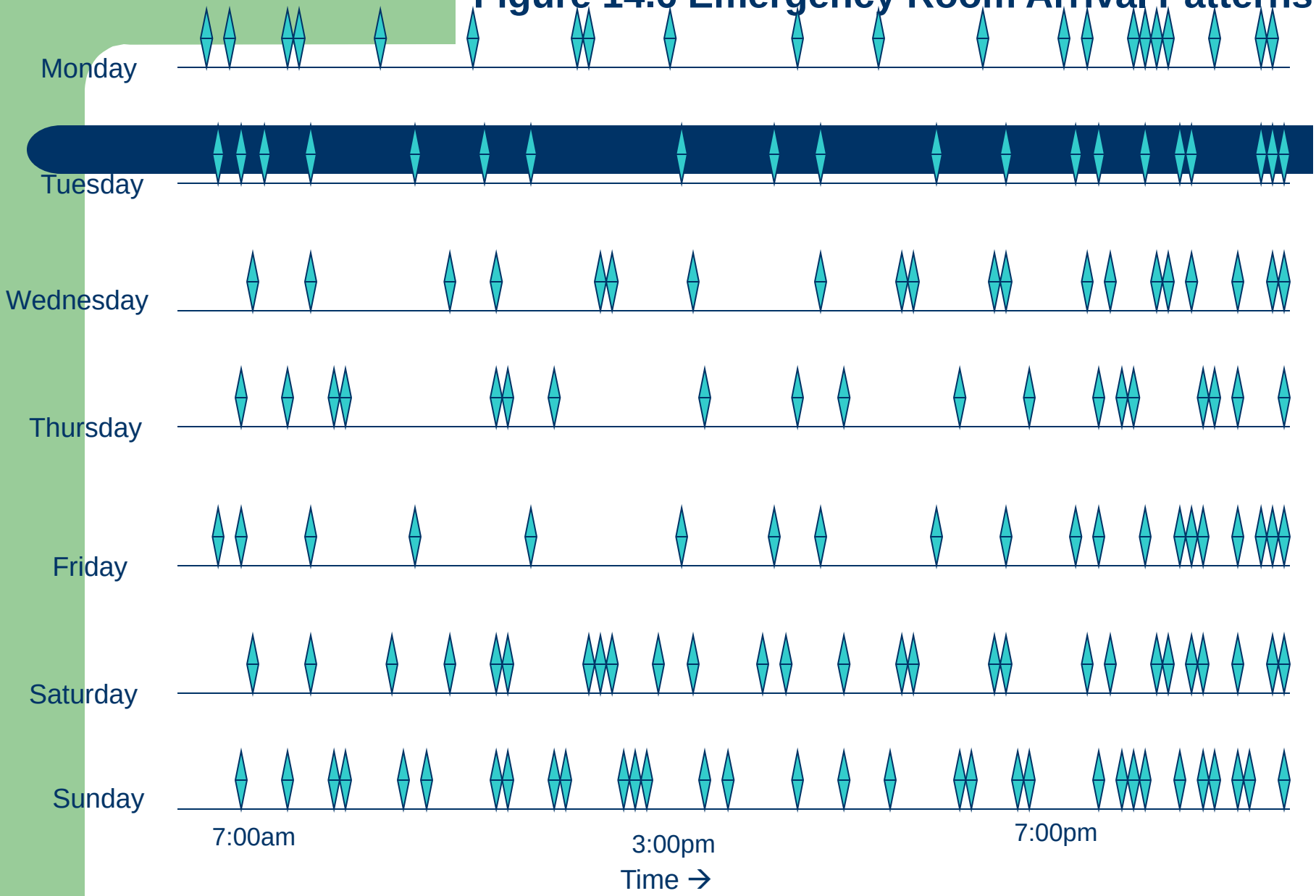
# Ozcan: Figure 14.4 Conceptualization of a Single-line, Multi-phase System



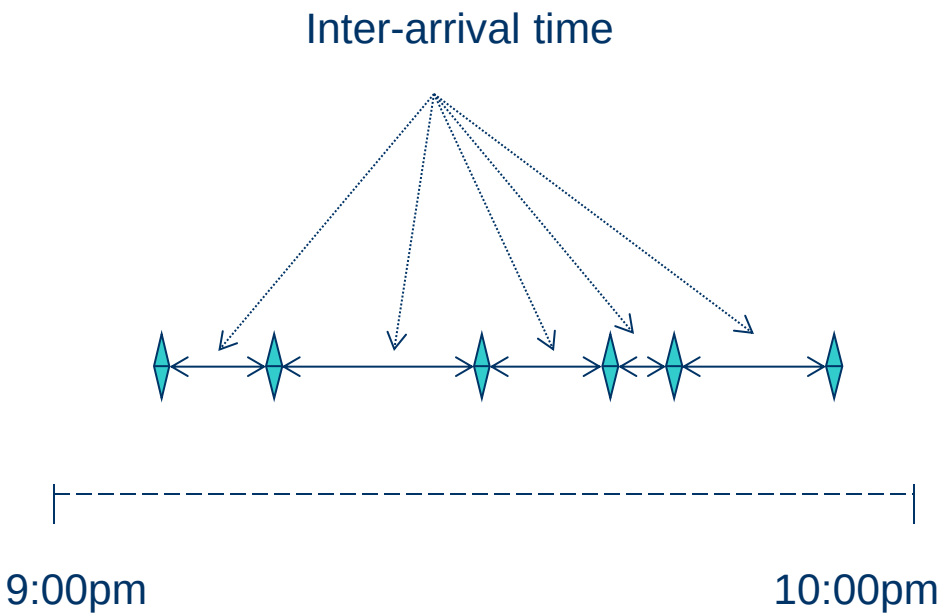
## Arrival and Service Patterns

- Both arrival and service patterns are random, thus causing waiting lines
- Models assume that:
  - Customer arrival rate can be described Poisson distribution
  - Service time can be described by a negative exponential distribution

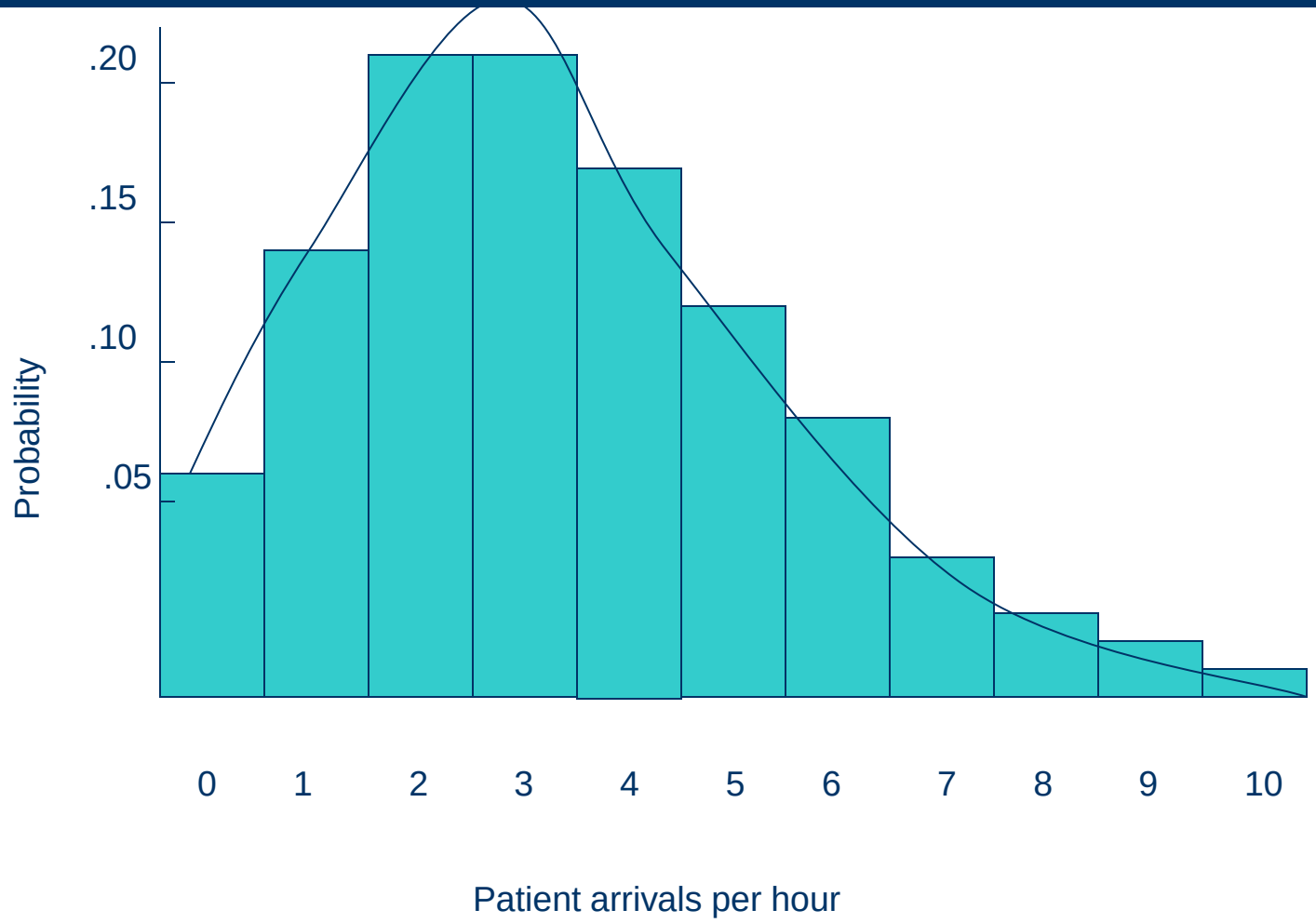
# Figure 14.6 Emergency Room Arrival Patterns



# Figure 14.7 Measures of Arrival Patterns



# Ozcan: Figure 14.8 Poisson Distribution



# Exponential and Poisson Distributions

If service time is exponential, then the service rate is Poisson. Further, if the customer arrival rate is Poisson, then the inter-arrival time (i.e., the time between arrivals) is exponential.

For instance: If a lab processes 10 customers per hour (rate), the average service time is 6 minutes. If the arrival rate is 12 per hour, then the average time between arrivals is 5 minutes.

Thus, service and arrival rates are described by the Poisson distribution and inter-arrival times and service times are described by a negative exponential distribution.

## Queue Characteristics

- **Balking**
  - **Patients who arrive and see big lines (the flu shot example) may change their minds and not join the queue, but go elsewhere to obtain service; this is called balking.**
- **Reneging**
  - **If they do join the queue and are dissatisfied with the waiting time, they may leave the queue; this is called reneging.**

# Queue Discipline

Refers to the order in which patients are processed, for instance:

- ✓ First Come First Served
- ✓ Most Serious First Served
- ✓ Highest Costs (waiting) First Served

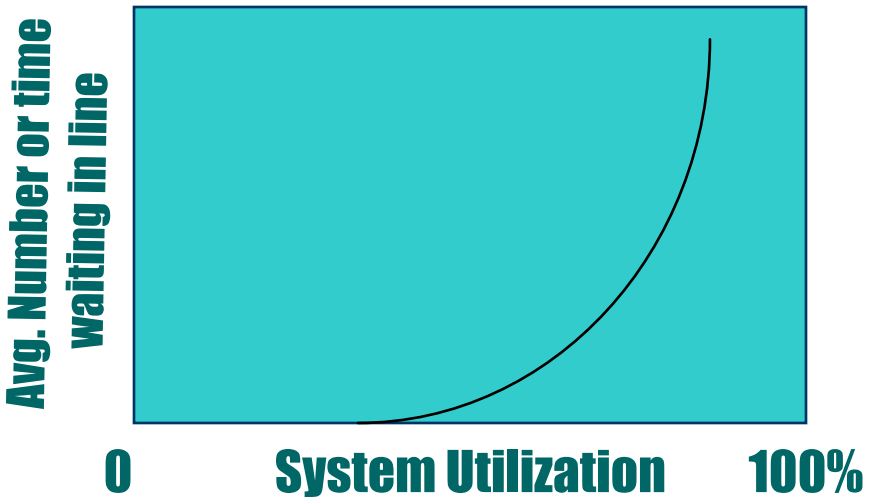
# System Metrics

- Average number of customers waiting (in line or in system)
- Average time customers wait (in line or in system)
- System utilization (% of capacity utilized)
- Implied costs of a given level of capacity and its related waiting line
- Probability that an arrival will have to wait for service

# Tradeoffs

Increases in system utilization are achieved at the expense of increases in both the length of the waiting line and the average waiting time.

Most models assume that the system is in a steady state where average arrival and service rates are stable.



# Ozcan: Exhibit 14.1 Queuing Model Classification

A: specification of arrival process, measured by inter-arrival time or arrival rate.
- M: negative exponential or Poisson distribution
- D: constant value
- K: Erlang distribution
- G: a general distribution with known mean and variance

B: specification of service process, measured by inter-service time or service rate
- M: negative exponential or Poisson distribution
- D: constant value
- K: Erlang distribution
- G: a general distribution with known mean and variance

C: specification of number of servers -- "s"

D: specification of queue or the maximum numbers allowed in a queuing system

E: specification of customer population

# Typical Infinite-Source Models

- Single channel, M/M/s

- Multiple channel, M/M/s>1,
where "s" designates the number of channels (servers).

- These models assume steady state conditions and a Poisson arrival rate.

# Infinite Source Models

- Single channel, exponential service time
- Single channel, constant service time
- Multiple channel, exponential service time
- Multiple priority service

# Exhibit 14.2 Queuing Model Notation

| | |
|---|---|
| $\lambda$ | arrival rate |
| $\mu$ | service rate |
| $L_q$ | average number of customers waiting for service |
| $L$ | average number of customers in the system (waiting or being served) |
| $W_q$ | average time customers wait in line |
| $W$ | average time customers spend in the system |
| $\rho$ | system utilization |
| $1/\mu$ | service time |
| $P_o$ | probability of zero units in system |
| $P_n$ | probability of n units in system |

# Infinite Source Models: Model Formulations

Five key relationships that provide basis for queuing formulations and are common for all infinite-source models:

1.The average number of patients being served is the ratio of arrival to service rate.

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

2. The average number of patients in the system is the average number in line plus the average number being served.

$$L = L_q + r$$

3. The average time in line is the average number in line divided by the arrival rate.

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

4. The average time in the system is the sum of the time in line plus the service time.

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

5. System utilization is the ratio of arrival rate to service capacity.

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

# Single Channel, Poisson Arrival and Exponential Service Time (M/M/1).

The average number waiting

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

The probability of n units in system

$$P_n = P_0 \left( \frac{\lambda}{\mu} \right)^n$$

The probability of 0 units in system

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

# Single Channel, Poisson Arrival and Exponential Service Time (M/M/1).

## Example 14.1

A hospital is exploring the level of staffing needed for a booth in the local mall where they would test and provide information on diabetes. Previous experience has shown that on average, every 15 minutes a new person approaches the booth.

A nurse can complete testing and answering questions, on average, in 12 minutes. If there is a single nurse at the booth, calculate system performance measures including the probability of idle time and of 1 or 2 persons waiting in the queue. What happens to the utilization rate if another workstation and nurse are added to the unit?

**Solution:**

Arrival rate: λ = 1(hour) ÷ 15 = 60(minutes) ÷ 15 = 4 persons per hour.
Service rate: μ = 1(hour) ÷ 12 = 60(minutes) ÷ 12 = 5 persons per hour.

Average persons served at any given time

$$r = \frac{\lambda}{\mu} = \frac{4}{5} = .8$$

Persons waiting in the queue

$$L_q = \frac{4^2}{5(5-4)} = 3.2$$

Number of persons in the system

$$L = L_q + \frac{\lambda}{\mu} = 3.2 + .8 = 4 \ persons.$$

Minutes of waiting time in the queue

$$W_q = \frac{L_q}{\lambda} = \frac{3.2}{4} = 0.8 = 48$$

Minutes in the system (waiting and service)

$$W = W_q + \frac{1}{\mu} = 48 + \frac{60}{5} = 48 + 12 = 60$$

Calculate queue lengths of 0, 1, and 2 persons

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{4}{5} = 1 - .8 = .2$$

$$P_1 = P_0\left(\frac{\lambda}{\mu}\right)^1 = (.2)\left(\frac{4}{5}\right)^1 = (.2)(.8)^1 = (.2)(.8) = .16$$

$$P_2 = P_0\left(\frac{\lambda}{\mu}\right)^2 = (.2)\left(\frac{4}{5}\right)^2 = (.2)(.8)^2 = (.2)(.64) = .128$$

# Utilization of Servers

Current system utilization (s=1)

$$\rho = \frac{\lambda}{s\mu} = \frac{4}{1*5} = 80\%.$$

System utilization with an additional nurse (s=2)

$$\rho = \frac{\lambda}{s\mu} = \frac{4}{2*5} = 40\%.$$

System utilization decreases as we add more resources to it.

## Assumptions and Limitations

- Steady state of arrival and service process
- Independence of arrival and service is assumed
- The number of servers is assumed fixed
- Little ability to incorporate important exceptions to the general flow
- Difficult to model all but simple processes