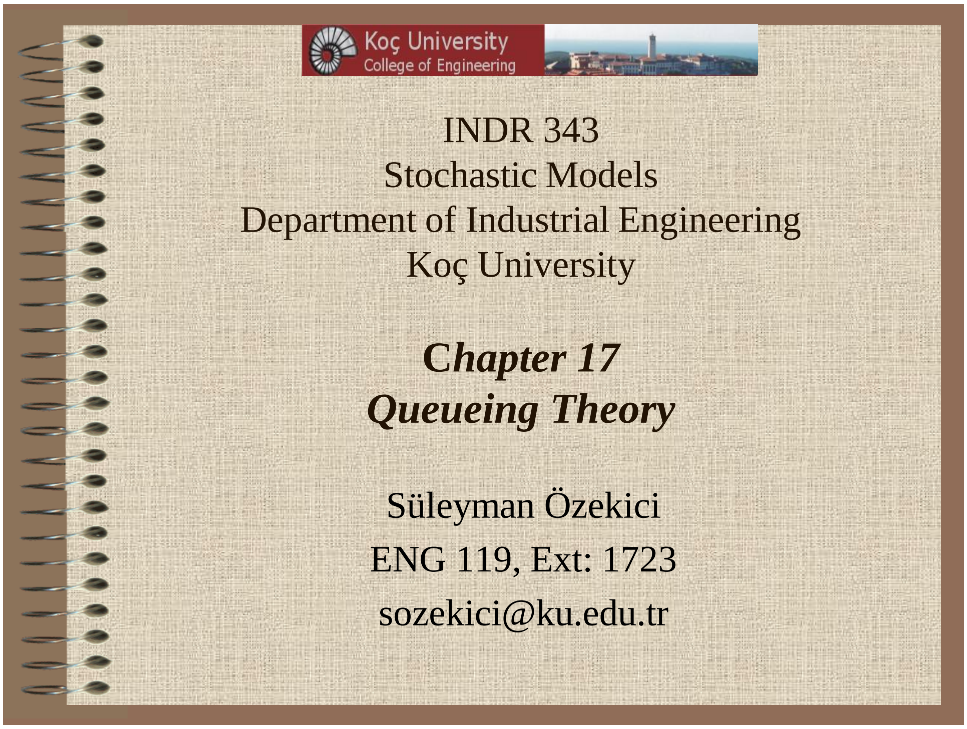
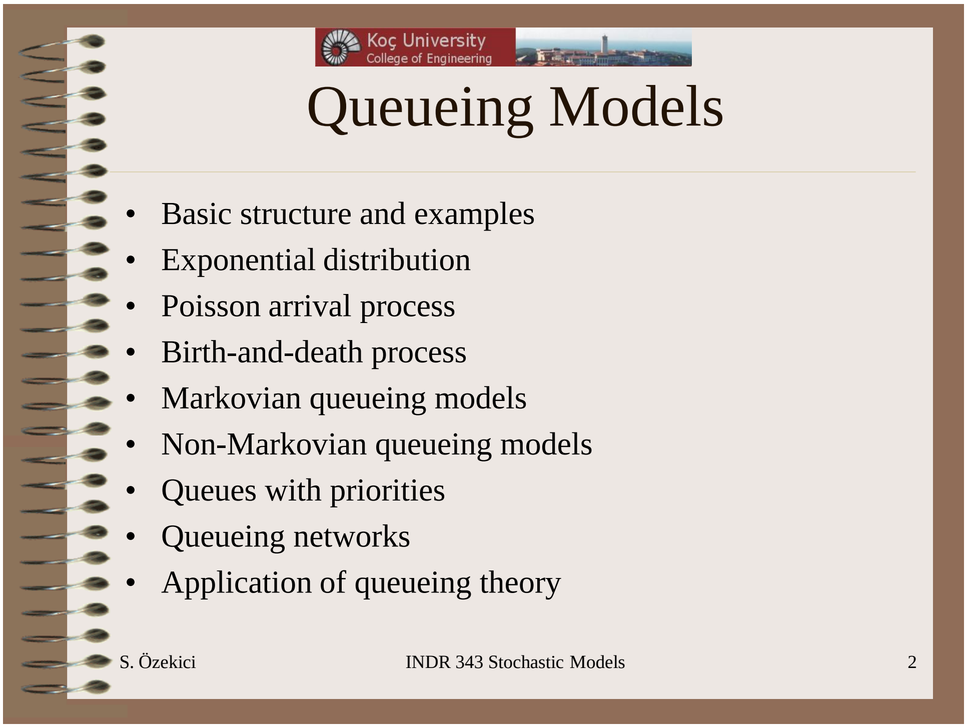


Koç University
College of Engineering

INDR 343
Stochastic Models
Department of Industrial Engineering
Koç University

Chapter 17
Queueing Theory

Süleyman Özekici
ENG 119, Ext: 1723
sozekici@ku.edu.tr



Koç University
College of Engineering

Queueing Models

- Basic structure and examples
- Exponential distribution
- Poisson arrival process
- Birth-and-death process
- Markovian queueing models
- Non-Markovian queueing models
- Queues with priorities
- Queueing networks
- Application of queueing theory

Queueing System

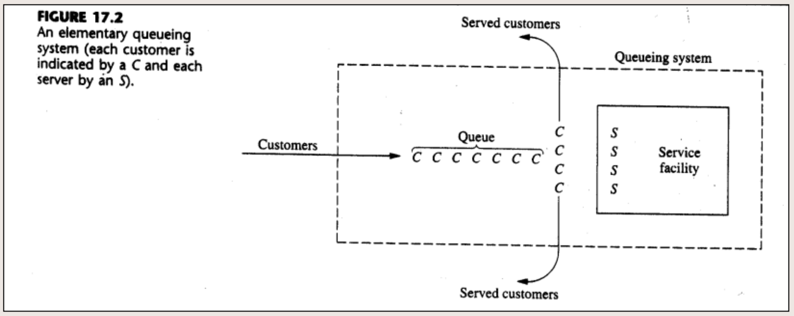
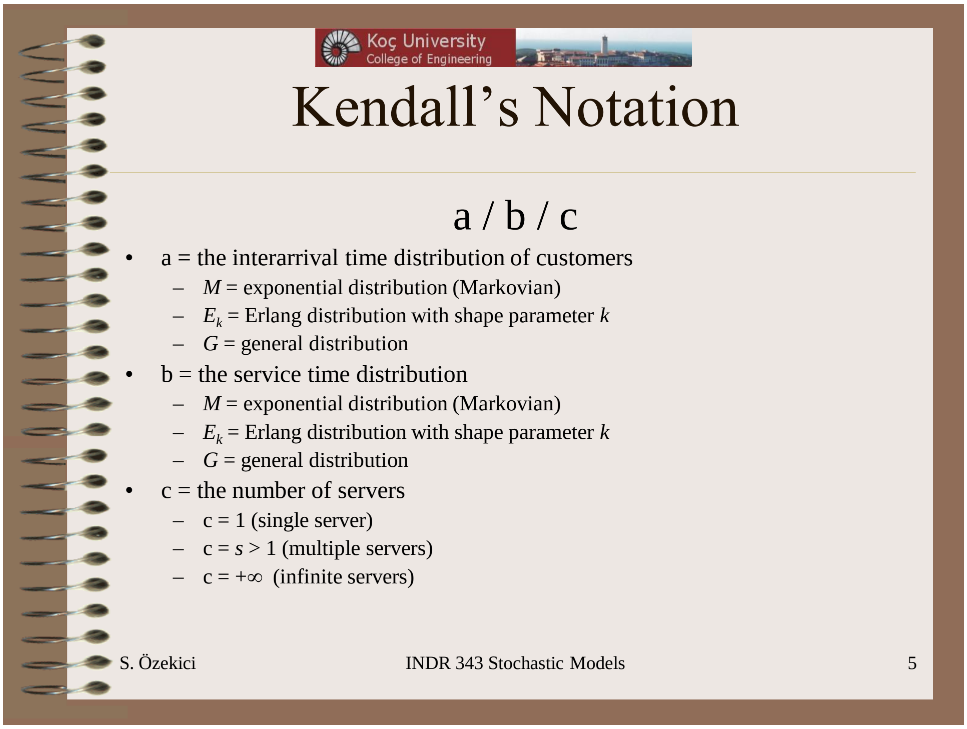


FIGURE 17.2
An elementary queueing system (each customer is indicated by a *C* and each server by an *S*).

Real Life Examples

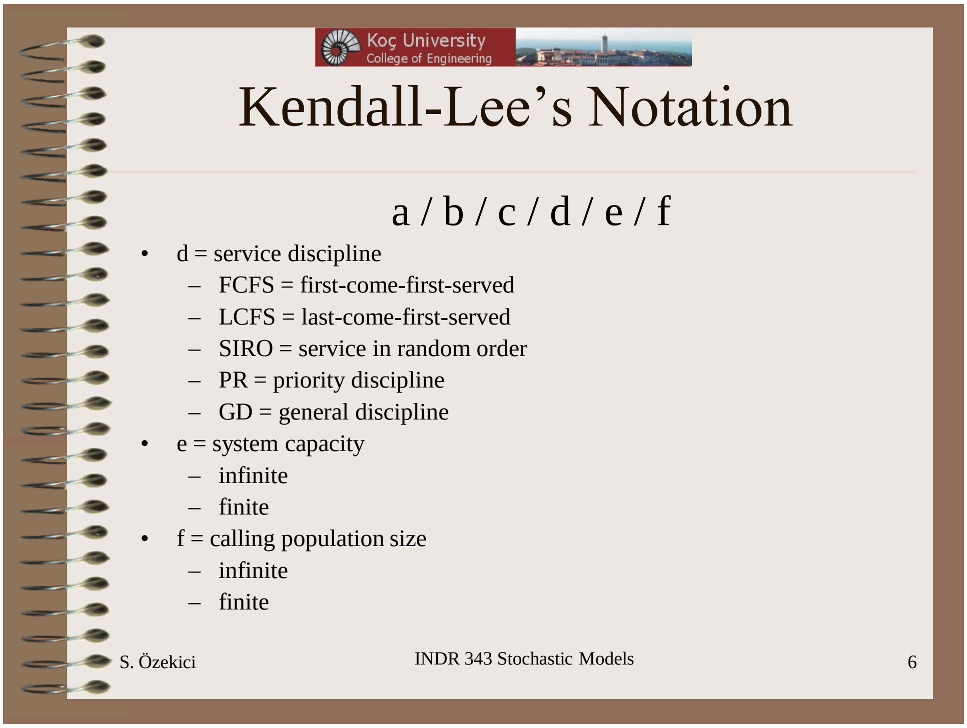
- Commercial service systems
  - Customers arriving at a bank to deposit or withdraw money
  - Customers arriving at a supermarket to buy groceries
- Transportation service systems
  - Cars arriving at the Boğaziçi bridge
  - Airplanes arriving at an airport for landing
- Internal service systems
  - Machines arriving at the maintenance center for repair
  - Products arriving at the quality control station for inspection
- Social service systems
  - Cases arriving at a court of law to be processed by judges
  - Patients arriving at the hospital for health care



Kendall’s Notation

a / b / c

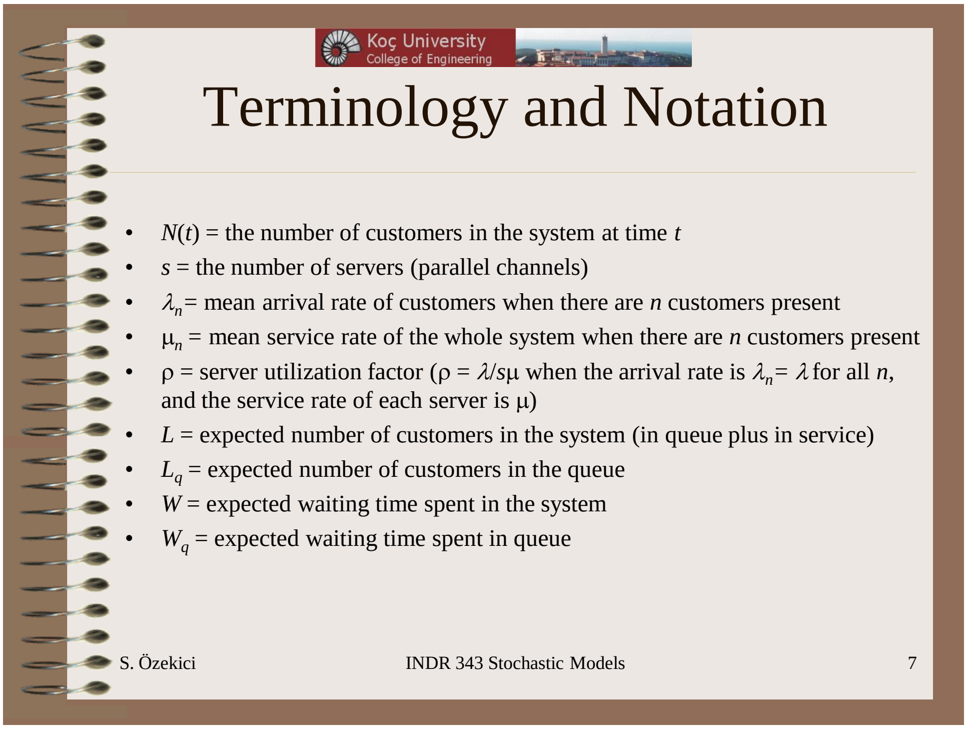
- a = the interarrival time distribution of customers
  - M = exponential distribution (Markovian)
  - E_k = Erlang distribution with shape parameter k
  - G = general distribution
- b = the service time distribution
  - M = exponential distribution (Markovian)
  - E_k = Erlang distribution with shape parameter k
  - G = general distribution
- c = the number of servers
  - c = 1 (single server)
  - c = s > 1 (multiple servers)
  - c = $+\infty$ (infinite servers)



Kendall-Lee’s Notation

a / b / c / d / e / f

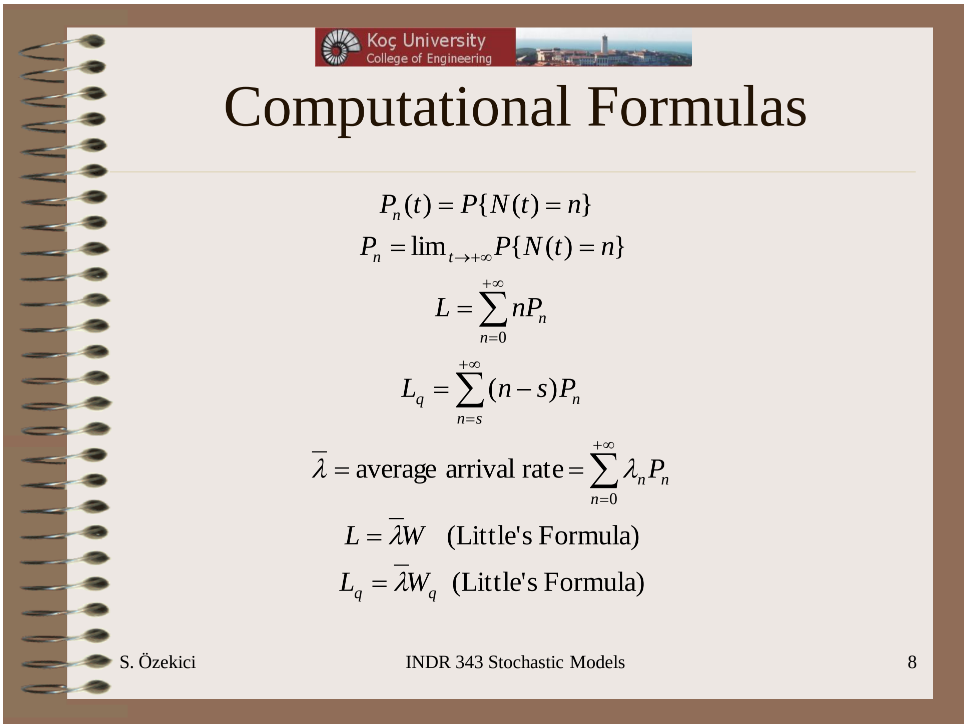
- d = service discipline
  - FCFS = first-come-first-served
  - LCFS = last-come-first-served
  - SIRO = service in random order
  - PR = priority discipline
  - GD = general discipline
- e = system capacity
  - infinite
  - finite
- f = calling population size
  - infinite
  - finite



Koç University
College of Engineering

Terminology and Notation

- $N(t)$ = the number of customers in the system at time t
- s = the number of servers (parallel channels)
- λ_n= mean arrival rate of customers when there are n customers present
- μ_n = mean service rate of the whole system when there are n customers present
- ρ = server utilization factor ($\rho = \lambda/s\mu$ when the arrival rate is $\lambda_n = \lambda$ for all n, and the service rate of each server is μ)
- L = expected number of customers in the system (in queue plus in service)
- L_q = expected number of customers in the queue
- W = expected waiting time spent in the system
- W_q = expected waiting time spent in queue



Koç University
College of Engineering

Computational Formulas

$$P_n(t) = P\{N(t) = n\}$$

$$P_n = \lim_{t\to+\infty} P\{N(t) = n\}$$

$$L = \sum_{n=0}^{+\infty} nP_n$$

$$L_q = \sum_{n=s}^{+\infty} (n-s)P_n$$

$$\bar{\lambda} = \text{average arrival rate} = \sum_{n=0}^{+\infty} \lambda_n P_n$$

$$L = \bar{\lambda}W \quad \text{(Little's Formula)}$$

$$L_q = \bar{\lambda}W_q \quad \text{(Little's Formula)}$$

The Exponential Distribution

- The random variable *T* has the exponential distribution with parameter α if

$$P\{T \le t\} = 1 - e^{-\alpha t}$$

$$P\{T > t\} = e^{-\alpha t}$$

$$f_T(t) = \frac{dP\{T \le t\}}{dt} = \alpha e^{-\alpha t}$$

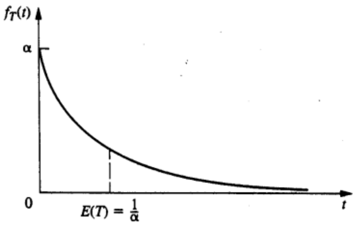
$$E(T) = \frac{1}{\alpha}$$

$$\text{var}(T) = \frac{1}{\alpha^2}$$

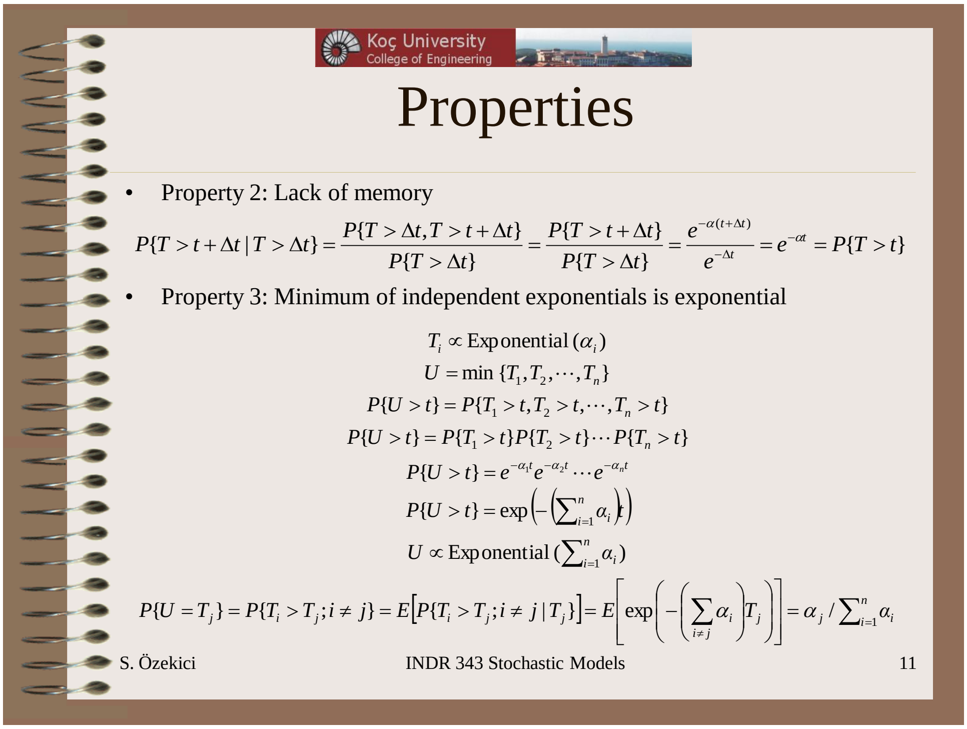
Properties

- Property 1: $f_T(t)$ is strictly decreasing

FIGURE 17.3
Probability density function for the exponential distribution.



$$P\{T > \frac{1}{\alpha}\} = e^{-\alpha\left(\frac{1}{\alpha}\right)} = e^{-1} \cong 0.37$$



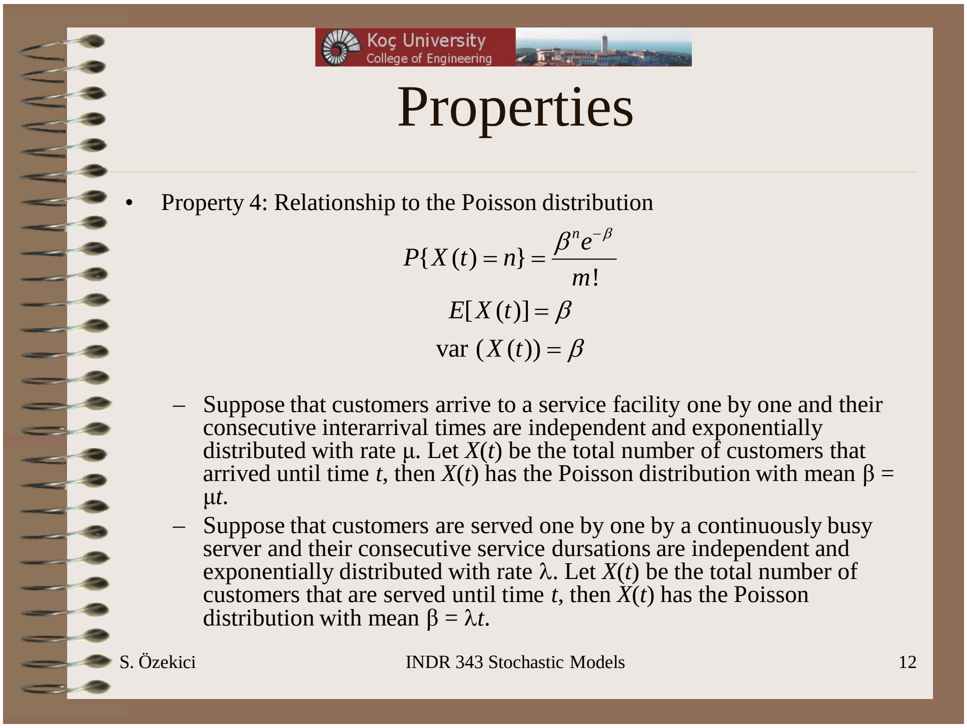
Properties

- Property 2: Lack of memory

$$P\{T > t+\Delta t \mid T > \Delta t\} = \frac{P\{T > \Delta t, T > t+\Delta t\}}{P\{T > \Delta t\}} = \frac{P\{T > t+\Delta t\}}{P\{T > \Delta t\}} = \frac{e^{-\alpha(t+\Delta t)}}{e^{-\Delta t}} = e^{-\alpha t} = P\{T > t\}$$

- Property 3: Minimum of independent exponentials is exponential

$$T_i \propto \text{Exponential}\,(\alpha_i)$$
$$U = \min\,\{T_1, T_2, \cdots, T_n\}$$
$$P\{U > t\} = P\{T_1 > t, T_2 > t, \cdots, T_n > t\}$$
$$P\{U > t\} = P\{T_1 > t\}P\{T_2 > t\}\cdots P\{T_n > t\}$$
$$P\{U > t\} = e^{-\alpha_1 t}e^{-\alpha_2 t}\cdots e^{-\alpha_n t}$$
$$P\{U > t\} = \exp\left(-\left(\sum_{i=1}^{n}\alpha_i\right)t\right)$$
$$U \propto \text{Exponential}\,\left(\sum_{i=1}^{n}\alpha_i\right)$$
$$P\{U = T_j\} = P\{T_i > T_j; i \neq j\} = E\left[P\{T_i > T_j; i \neq j \mid T_j\}\right] = E\left[\exp\left(-\left(\sum_{i \neq j}\alpha_i\right)T_j\right)\right] = \alpha_j \Big/ \sum_{i=1}^{n}\alpha_i$$



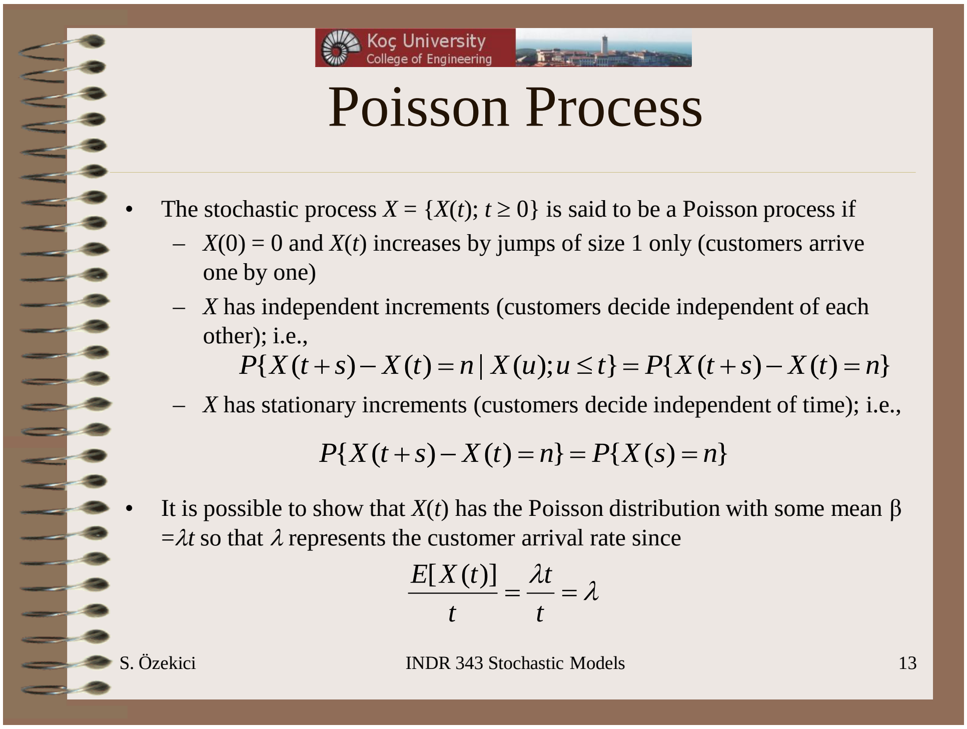
Properties

- Property 4: Relationship to the Poisson distribution

$$P\{X(t) = n\} = \frac{\beta^n e^{-\beta}}{m!}$$
$$E[X(t)] = \beta$$
$$\text{var}\;(X(t)) = \beta$$

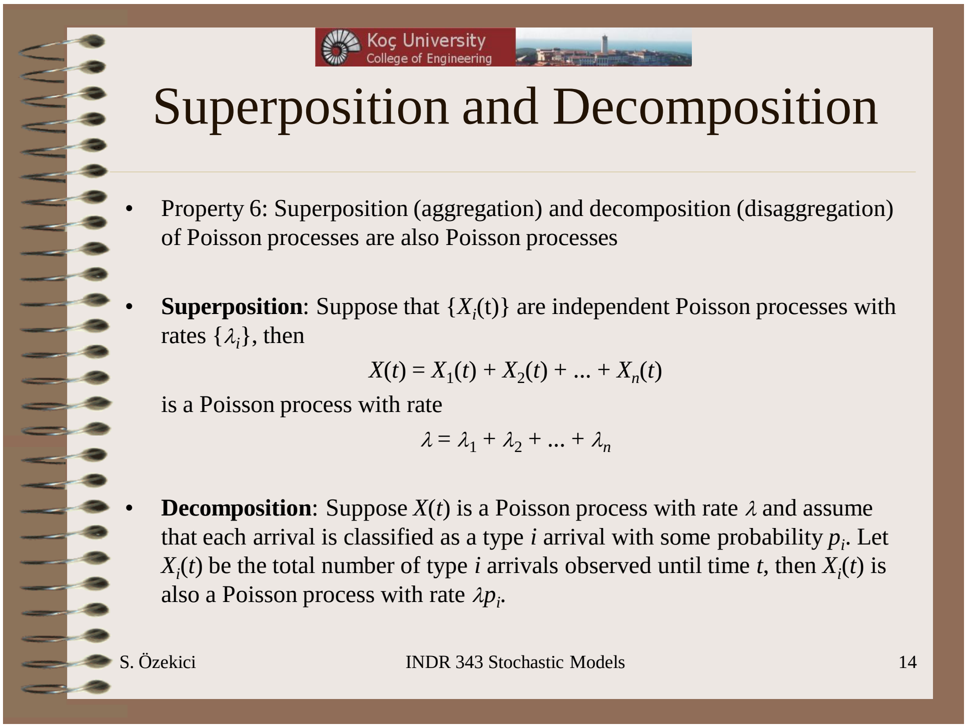
  – Suppose that customers arrive to a service facility one by one and their consecutive interarrival times are independent and exponentially distributed with rate μ. Let *X*(*t*) be the total number of customers that arrived until time *t*, then *X*(*t*) has the Poisson distribution with mean β = μ*t*.

  – Suppose that customers are served one by one by a continuously busy server and their consecutive service dursations are independent and exponentially distributed with rate λ. Let *X*(*t*) be the total number of customers that are served until time *t*, then *X*(*t*) has the Poisson distribution with mean β = λ*t*.



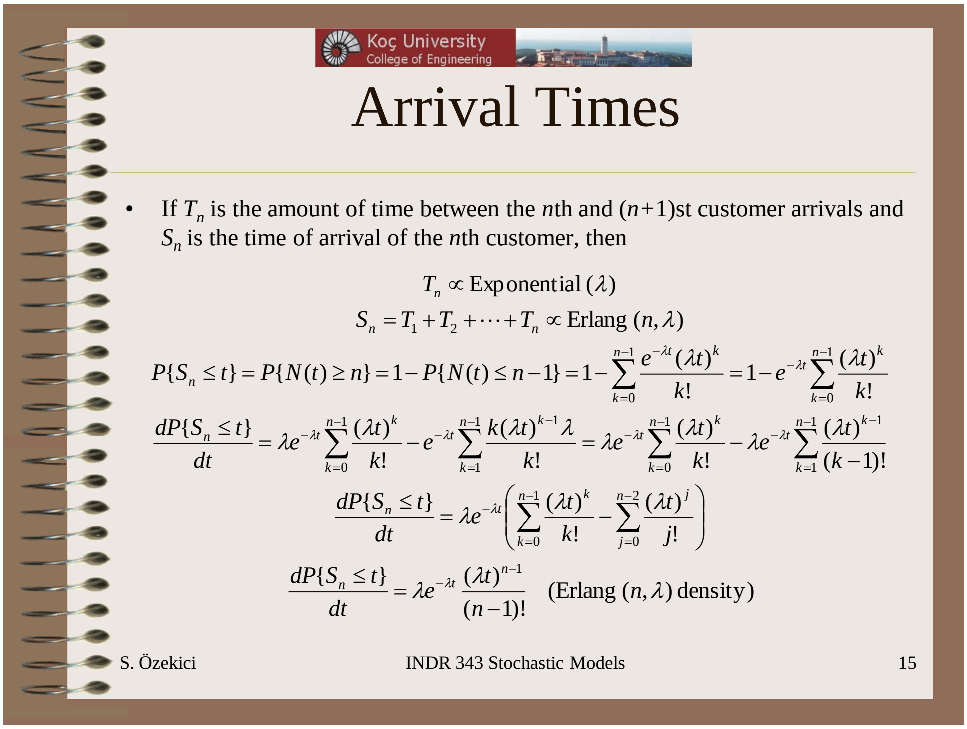
Poisson Process

- The stochastic process $X = \{X(t); t \geq 0\}$ is said to be a Poisson process if
  - $X(0) = 0$ and $X(t)$ increases by jumps of size 1 only (customers arrive one by one)
  - X has independent increments (customers decide independent of each other); i.e.,
    $$P\{X(t+s) - X(t) = n \mid X(u); u \leq t\} = P\{X(t+s) - X(t) = n\}$$
  - X has stationary increments (customers decide independent of time); i.e.,
    $$P\{X(t+s) - X(t) = n\} = P\{X(s) = n\}$$
- It is possible to show that $X(t)$ has the Poisson distribution with some mean $\beta = \lambda t$ so that λ represents the customer arrival rate since
  $$\frac{E[X(t)]}{t} = \frac{\lambda t}{t} = \lambda$$



Superposition and Decomposition

- Property 6: Superposition (aggregation) and decomposition (disaggregation) of Poisson processes are also Poisson processes
- **Superposition**: Suppose that $\{X_i(\mathrm{t})\}$ are independent Poisson processes with rates $\{\lambda_i\}$, then
  $$X(t) = X_1(t) + X_2(t) + \ldots + X_n(t)$$
  is a Poisson process with rate
  $$\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n$$
- **Decomposition**: Suppose $X(t)$ is a Poisson process with rate λ and assume that each arrival is classified as a type i arrival with some probability p_i. Let $X_i(t)$ be the total number of type i arrivals observed until time t, then $X_i(t)$ is also a Poisson process with rate λp_i.



Arrival Times

- If T_n is the amount of time between the nth and $(n+1)$st customer arrivals and S_n is the time of arrival of the nth customer, then

$$T_n \propto \text{Exponential}\,(\lambda)$$

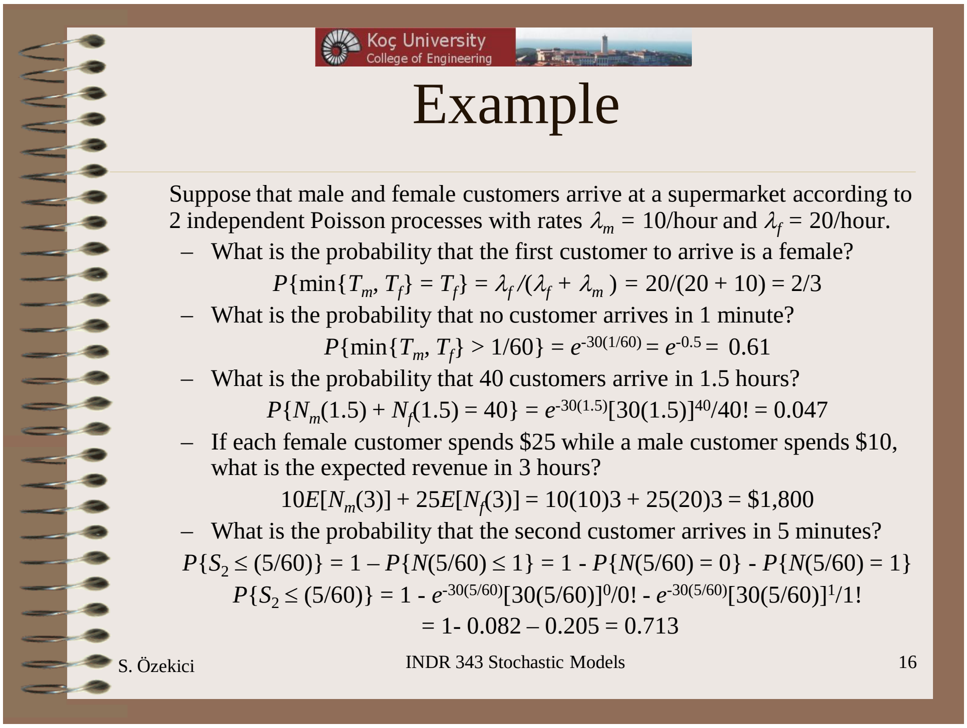
$$S_n = T_1 + T_2 + \cdots + T_n \propto \text{Erlang}\,(n, \lambda)$$

$$P\{S_n \le t\} = P\{N(t) \ge n\} = 1 - P\{N(t) \le n-1\} = 1 - \sum_{k=0}^{n-1} \frac{e^{-\lambda t}(\lambda t)^k}{k!} = 1 - e^{-\lambda t} \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!}$$

$$\frac{dP\{S_n \le t\}}{dt} = \lambda e^{-\lambda t} \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} - e^{-\lambda t} \sum_{k=1}^{n-1} \frac{k(\lambda t)^{k-1}\lambda}{k!} = \lambda e^{-\lambda t} \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} - \lambda e^{-\lambda t} \sum_{k=1}^{n-1} \frac{(\lambda t)^{k-1}}{(k-1)!}$$

$$\frac{dP\{S_n \le t\}}{dt} = \lambda e^{-\lambda t} \left(\sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} - \sum_{j=0}^{n-2} \frac{(\lambda t)^j}{j!} \right)$$

$$\frac{dP\{S_n \le t\}}{dt} = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \quad (\text{Erlang}\,(n, \lambda)\,\text{density})$$



Example

Suppose that male and female customers arrive at a supermarket according to 2 independent Poisson processes with rates λ_m = 10/hour and λ_f = 20/hour.

- What is the probability that the first customer to arrive is a female?

$P\{\min\{T_m, T_f\} = T_f\} = \lambda_f/(\lambda_f + \lambda_m) = 20/(20 + 10) = 2/3$

- What is the probability that no customer arrives in 1 minute?

$P\{\min\{T_m, T_f\} > 1/60\} = e^{-30(1/60)} = e^{-0.5} = 0.61$

- What is the probability that 40 customers arrive in 1.5 hours?

$P\{N_m(1.5) + N_f(1.5) = 40\} = e^{-30(1.5)}[30(1.5)]^{40}/40! = 0.047$

- If each female customer spends \$25 while a male customer spends \$10, what is the expected revenue in 3 hours?

$10E[N_m(3)] + 25E[N_f(3)] = 10(10)3 + 25(20)3 = \$1,800$

- What is the probability that the second customer arrives in 5 minutes?

$P\{S_2 \le (5/60)\} = 1 – P\{N(5/60) \le 1\} = 1 - P\{N(5/60) = 0\} - P\{N(5/60) = 1\}$

$P\{S_2 \le (5/60)\} = 1 - e^{-30(5/60)}[30(5/60)]^0/0! - e^{-30(5/60)}[30(5/60)]^1/1!$

$= 1- 0.082 – 0.205 = 0.713$

Birth-and-Death Process

- The process $N(t)$ representing the number of customers in the system is a birth-and-death process if:
  - Given $N(t) = n$, the remaining time to the next birth (customer arrival) is exponential with parameter λ_n,
  - Given $N(t) = n$, the remaining time to the next death (service completion) is exponential with parameter μ_n,
  - The remaining times to the next birth and death are independent.
- The birth-and-death process $N(t)$ is a Markov process with transition rate diagram

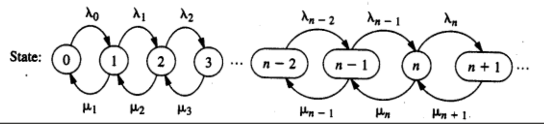


FIGURE 17.4 Rate diagram for the birth-and-death process.

Balance Equations

$$\lambda_0 P_0 = \mu_1 P_1 \Rightarrow P_1 = \frac{\lambda_0}{\mu_1} P_0$$

$$\lambda_1 P_1 = \mu_2 P_2 \Rightarrow P_2 = \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0$$

$$\vdots$$

$$\lambda_{n-1} P_{n-1} = \mu_n P_n \Rightarrow P_n = \frac{\lambda_{n-1}}{\mu_n} P_{n-1} = \frac{\lambda_{n-1} \cdots \lambda_1 \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1} P_0$$

$$P_n = C_n P_0 \qquad \text{where } C_n = \frac{\lambda_{n-1} \cdots \lambda_1 \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1}$$

Solution

$$\sum_{n=0}^{+\infty} P_n = 1 \Rightarrow \left(\sum_{n=0}^{+\infty} C_n \right) P_0 = 1$$

$$P_0 = \frac{1}{\left(\sum_{n=0}^{+\infty} C_n \right)}$$

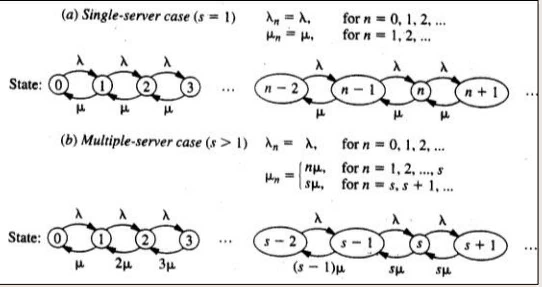
$$P_n = \frac{C_n}{\left(\sum_{n=0}^{+\infty} C_n \right)} = C_n P_0$$

- There is a limiting (steady-state) distribution if

$$\sum_{n=0}^{+\infty} C_n < +\infty \quad \text{(stability condition)}$$

M/M/s Model

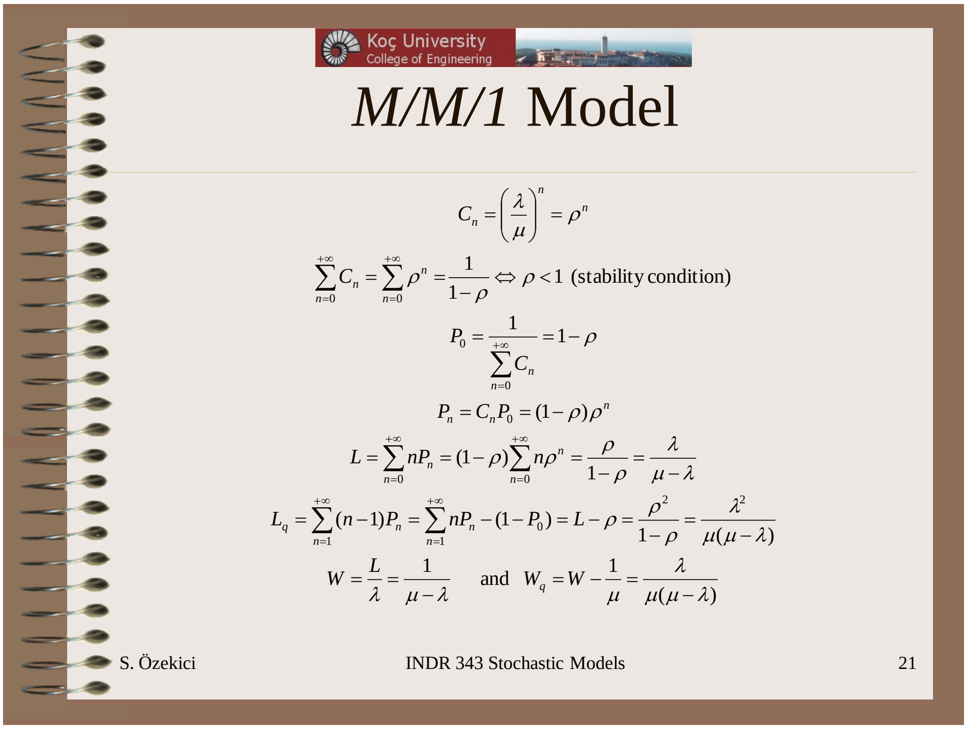
FIGURE 17.5
Rate diagrams for the *M/M/s* model.



(*a*) *Single-server case* ($s = 1$) $\lambda_n = \lambda$, for $n = 0, 1, 2, \ldots$ $\mu_n = \mu$, for $n = 1, 2, \ldots$

(*b*) *Multiple-server case* ($s > 1$) $\lambda_n = \lambda$, for $n = 0, 1, 2, \ldots$

$$\mu_n = \begin{cases} n\mu, & \text{for } n = 1, 2, \ldots, s \\ s\mu, & \text{for } n = s, s+1, \ldots \end{cases}$$



M/M/1 Model

$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n$$

$$\sum_{n=0}^{+\infty} C_n = \sum_{n=0}^{+\infty} \rho^n = \frac{1}{1-\rho} \Leftrightarrow \rho < 1 \text{ (stability condition)}$$

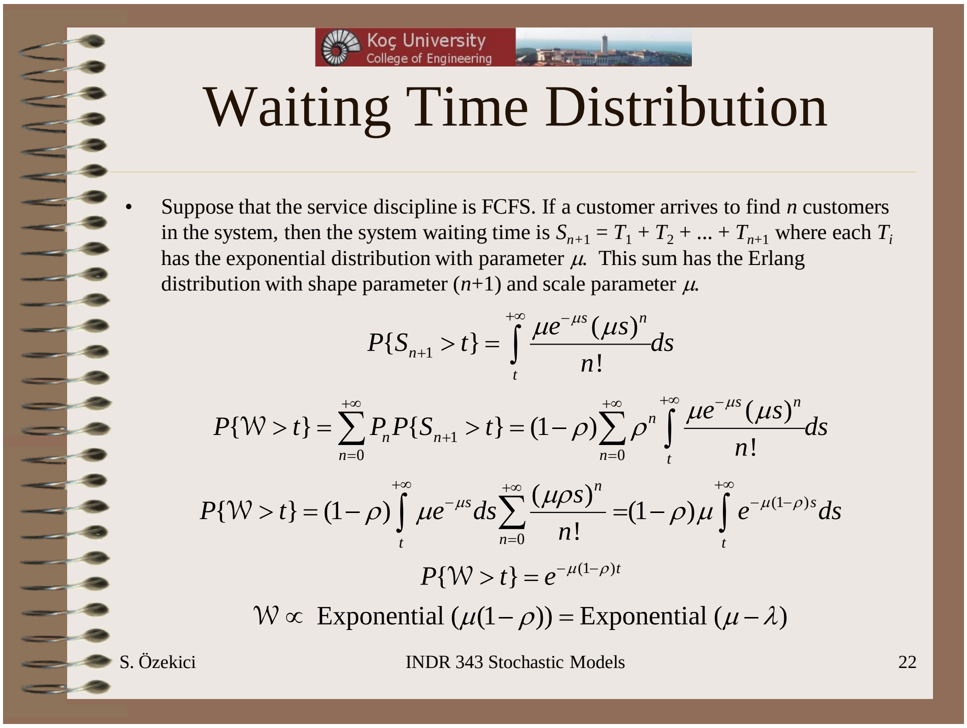
$$P_0 = \frac{1}{\sum_{n=0}^{+\infty} C_n} = 1 - \rho$$

$$P_n = C_n P_0 = (1-\rho)\rho^n$$

$$L = \sum_{n=0}^{+\infty} nP_n = (1-\rho)\sum_{n=0}^{+\infty} n\rho^n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

$$L_q = \sum_{n=1}^{+\infty}(n-1)P_n = \sum_{n=1}^{+\infty} nP_n - (1-P_0) = L - \rho = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

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



Waiting Time Distribution

- Suppose that the service discipline is FCFS. If a customer arrives to find *n* customers in the system, then the system waiting time is $S_{n+1} = T_1 + T_2 + ... + T_{n+1}$ where each T_i has the exponential distribution with parameter μ. This sum has the Erlang distribution with shape parameter (n+1) and scale parameter μ.

$$P\{S_{n+1} > t\} = \int_t^{+\infty} \frac{\mu e^{-\mu s}(\mu s)^n}{n!} ds$$

$$P\{\mathcal{W} > t\} = \sum_{n=0}^{+\infty} P_n P\{S_{n+1} > t\} = (1-\rho)\sum_{n=0}^{+\infty} \rho^n \int_t^{+\infty} \frac{\mu e^{-\mu s}(\mu s)^n}{n!} ds$$

$$P\{\mathcal{W} > t\} = (1-\rho)\int_t^{+\infty} \mu e^{-\mu s} ds \sum_{n=0}^{+\infty} \frac{(\mu\rho s)^n}{n!} = (1-\rho)\mu \int_t^{+\infty} e^{-\mu(1-\rho)s} ds$$

$$P\{\mathcal{W} > t\} = e^{-\mu(1-\rho)t}$$

$$\mathcal{W} \propto \text{ Exponential } (\mu(1-\rho)) = \text{Exponential } (\mu - \lambda)$$

Waiting Time Distribution

- If a customer arrives to find n customers in the system, then the waiting time in the queue is $S_n = T_1 + T_2 + \ldots + T_n$ where each T_i has the exponential distribution with parameter μ. This sum has the Erlang distribution with shape parameter n and scale parameter μ.

$$P\{S_n > t\} = \int_t^{+\infty} \frac{\mu e^{-\mu s}(\mu s)^{n-1}}{(n-1)!} ds$$

$$P\{\mathcal{W}_q = 0\} = P_0 = 1 - \rho$$

$$P\{\mathcal{W}_q > t\} = \sum_{n=1}^{+\infty} P_n P\{S_n > t\} = (1-\rho)\sum_{n=1}^{+\infty} \rho^n \int_t^{+\infty} \frac{\mu e^{-\mu s}(\mu s)^{n-1}}{(n-1)!} ds$$

$$P\{\mathcal{W}_q > t\} = (1-\rho)\rho \int_t^{+\infty} \mu e^{-\mu s} ds \sum_{n=1}^{+\infty} \frac{(\mu \rho s)^{n-1}}{(n-1)!} = \rho(1-\rho)\mu \int_t^{+\infty} e^{-\mu(1-\rho)s} ds$$

$$P\{\mathcal{W}_q > t\} = \rho e^{-\mu(1-\rho)t}$$

M/M/s Model Results

The queue is stable if $\rho = \lambda / s\mu < 1$

$$P_0 = \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1-\lambda/(s\mu)} \right]^{-1}$$

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 & \text{if } 0 \le n \le s \\ \dfrac{(\lambda/\mu)^n}{s!\, s^{n-s}} P_0 & \text{if } n \ge s \end{cases}$$

$$L_q = \frac{P_0 (\lambda/\mu)^s \rho}{s!(1-\rho)^2} \Rightarrow W_q = \frac{L_q}{\lambda}$$

$$W = W_q + \frac{1}{\mu} \Rightarrow L = \lambda W = \lambda\left(W_q + \frac{1}{\mu}\right) = L_q + \frac{\lambda}{\mu}$$

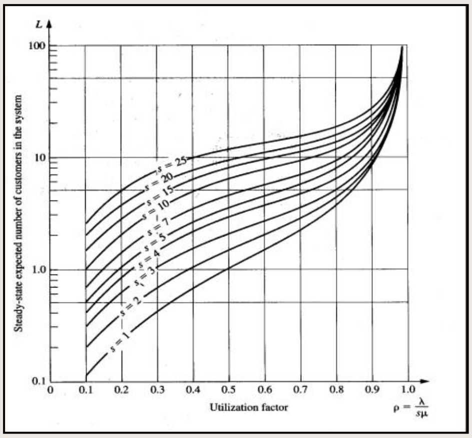
M/M/s Model Results

$$P\{\mathcal{W} > t\} = e^{-\mu t}\left[\frac{1+P_0\left(\lambda/\mu\right)^s}{s!(1-\rho)}\left(\frac{1-e^{-\mu t(s-1-(\lambda/\mu)}}{s-1-(\lambda/\mu)}\right)\right]$$

$$P\{\mathcal{W}_q = 0\} = \sum_{n=0}^{s-1} P_n$$

$$P\{\mathcal{W}_q > t\} = (1-P\{W_q = 0\})e^{-s\mu(1-\rho)t}$$

Queueing Chart for *L*



Koç University
College of Engineering

County Hospital Example

- The emergency room of the County Hospital provides quick medical care for emergency cases. At any hour there is always one doctor on duty. However, because of a growing tendency for emergency cases, the hospital is experiencing a continuing increase in the number of emergency patients. As a result, it has become quite common for patients to have to wait to be treated. The management is considering hiring an additional doctor.
- Analysis of past data reveals that patients arrive according to a Poisson process at a rate of 1 every ½ hour. This implies that $\lambda = 2$ customers per hour.
- A doctor requires an average of 20 minutes per patient and the service duration is exponential. So, $\mu = 3$ customers per hour.
- We need to compare the two *M/M/s* models with $s = 1$ ($\rho = 2/3$) and $s = 2$ ($\rho = 2/2(3)=1/3$).

Koç University
College of Engineering

Comparison of Alternatives

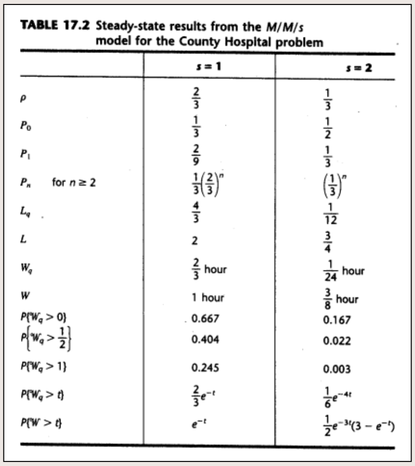


TABLE 17.2 Steady-state results from the *M/M/s* model for the County Hospital problem

	$s = 1$	$s = 2$
ρ	$\frac{2}{3}$	$\frac{1}{3}$
P_0	$\frac{1}{3}$	$\frac{1}{2}$
P_1	$\frac{2}{9}$	$\frac{1}{3}$
P_n for $n \geq 2$	$\frac{1}{3}\left(\frac{2}{3}\right)^n$	$\left(\frac{1}{3}\right)^n$
L_q	$\frac{4}{3}$	$\frac{1}{12}$
L	2	$\frac{3}{4}$
W_q	$\frac{2}{3}$ hour	$\frac{1}{24}$ hour
W	1 hour	$\frac{3}{8}$ hour
$P\{\mathcal{W}_q > 0\}$	0.667	0.167
$P\left\{\mathcal{W}_q > \frac{1}{2}\right\}$	0.404	0.022
$P\{\mathcal{W}_q > 1\}$	0.245	0.003
$P\{\mathcal{W}_q > t\}$	$\frac{2}{3}e^{-t}$	$\frac{1}{6}e^{-4t}$
$P\{\mathcal{W} > t\}$	e^{-t}	$\frac{1}{2}e^{-3t}(3 - e^{-t})$

M/M/1/GD/K Model

- This is the *M/M/1* model when the system capacity is finite (*K*). In the birth-and-death model, we need to take

$$\lambda_n = \begin{cases} \lambda & \text{for } n = 0,1,2,\cdots,K-1 \\ 0 & \text{for } n \geq K \end{cases}$$

$$C_n = \begin{cases} \left(\dfrac{\lambda}{\mu}\right)^n = \rho^n & \text{for } n = 0,1,2,\cdots,K \\ 0 & \text{for } n > K \end{cases}$$

$$P_0 = \frac{1-\rho}{1-\rho^{K+1}} \quad \text{(The queue is always stable)} \quad P_n = \frac{1-\rho}{1-\rho^{K+1}}\rho^n$$

$$L = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} \Rightarrow L_q = L - (1-P_0)$$

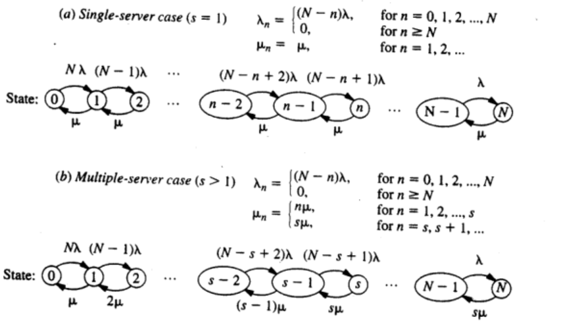
$$\overline{\lambda} = \sum_{n=0}^{+\infty} \lambda_n P_n = \sum_{n=0}^{K-1} \lambda P_n = \lambda(1-P_K)$$

$$W = \frac{L}{\overline{\lambda}}, W_q = \frac{L_q}{\overline{\lambda}}$$

M/M/s/GD/+∞/N Model

- This is the *M/M/s* model with finite calling population (*N*). It is also known an the machine-repair model.

FIGURE 17.7
Rate diagrams for the finite calling population variation of the *M/M/s* model.



(*a*) *Single-server case* ($s = 1$)

$$\lambda_n = \begin{cases} (N-n)\lambda, & \text{for } n = 0, 1, 2, \ldots, N \\ 0, & \text{for } n \geq N \end{cases}$$

$$\mu_n = \mu, \quad \text{for } n = 1, 2, \ldots$$

(*b*) *Multiple-server case* ($s > 1$)

$$\lambda_n = \begin{cases} (N-n)\lambda, & \text{for } n = 0, 1, 2, \ldots, N \\ 0, & \text{for } n \geq N \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & \text{for } n = 1, 2, \ldots, s \\ s\mu, & \text{for } n = s, s+1, \ldots \end{cases}$$

M/M/s/GD/+∞/N Results

$$P_n = \begin{cases} \dfrac{N!}{(N-n)!n!}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & 0 \le n \le s \\ \dfrac{N!}{(N-n)!s!s^{n-s}}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & s \le n \le N \end{cases}$$

$$P_0 = \left[\sum_{n=0}^{s-1} \frac{N!}{(N-n)!n!}\left(\frac{\lambda}{\mu}\right)^n + \sum_{n=s}^{N} \frac{N!}{(N-n)!s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}$$

$$L_q = \sum_{n=s}^{N}(n-s)P_n, L = \sum_{n=0}^{s-1} nP_n + L_q + s(1 - \sum_{n=0}^{s-1} P_n)$$

$$\overline{\lambda} = \sum_{n=0}^{+\infty} \lambda_n P_n = \sum_{n=0}^{N}(N-n)\lambda P_n = \lambda(N-L)$$

$$W = \frac{L}{\overline{\lambda}}, W_q = \frac{L_q}{\overline{\lambda}}$$

M/G/1 Model

- The service durations are independent and identically distributed with some mean $1/\mu$ and variance σ^2

$$\rho = \lambda / \mu < 1 \quad \text{(Stability condition)}$$

$$P_0 = 1 - \rho$$

$$L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)} \quad \text{(Pollaczek - Khintchine formula)}$$

$$L = \rho + L_q$$

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

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

Koç University
College of Engineering

Special Cases of *M/G/1*

- If $G = D$ is deterministic so that $\sigma^2 = 0$, then

$$L_q = \frac{\rho^2}{2(1-\rho)}$$

- If $G = M$ is exponential so that $\sigma^2 = 1/\mu^2$, then

$$L_q = \frac{\lambda^2\sigma^2+\rho^2}{2(1-\rho)} = \frac{(\lambda/\mu)^2+\rho^2}{2(1-\rho)} = \frac{\rho^2}{(1-\rho)} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

- If $G = E_k$ is Erlang $(k, k\mu)$ so that $\sigma^2 = 1/k\mu^2$, then

$$L_q = \frac{\lambda^2/(k\mu^2)+\rho^2}{2(1-\rho)} = \frac{1+k}{2k}\frac{\rho^2}{1-\rho} = \frac{1+k}{2k}\frac{\lambda^2}{\mu(\mu-\lambda)}$$

$$W_q = \frac{L_q}{\lambda} = \frac{1+k}{2k}\frac{\lambda}{\mu(\mu-\lambda)}$$

$$W = W_q + \frac{1}{\mu}, L = \lambda W$$

Koç University
College of Engineering

GI/M/1 Model

- The customer interarrival times are independent and identically distributed with mean $1/\lambda$

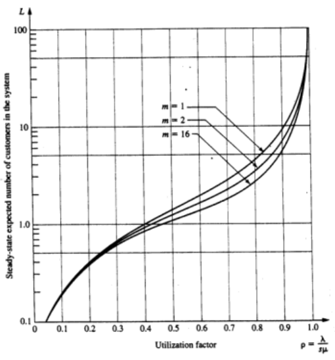
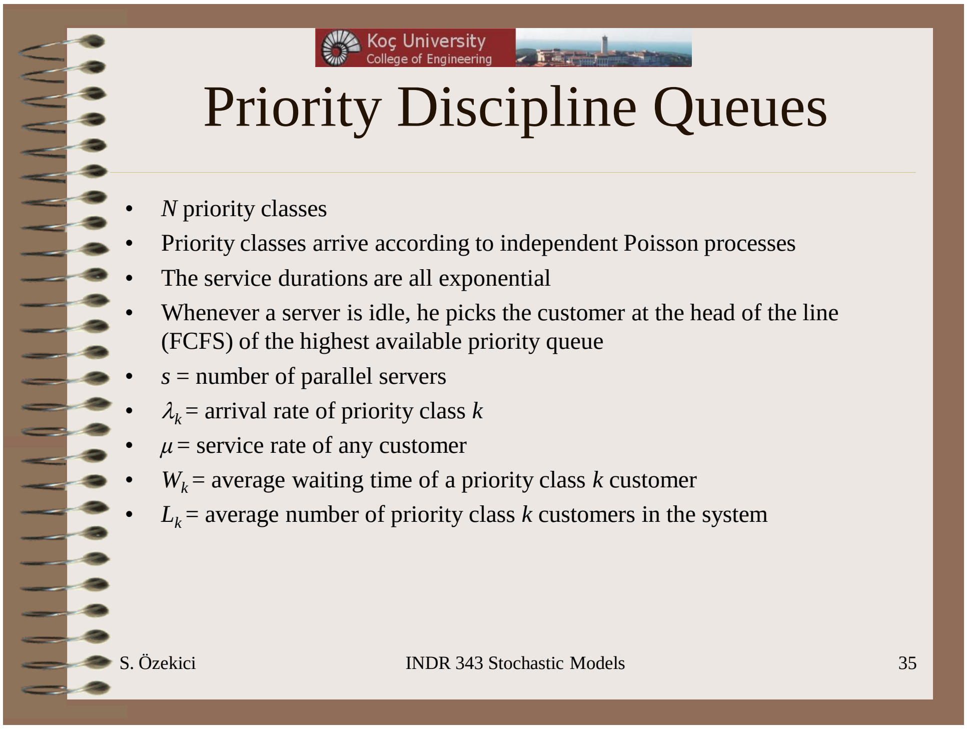
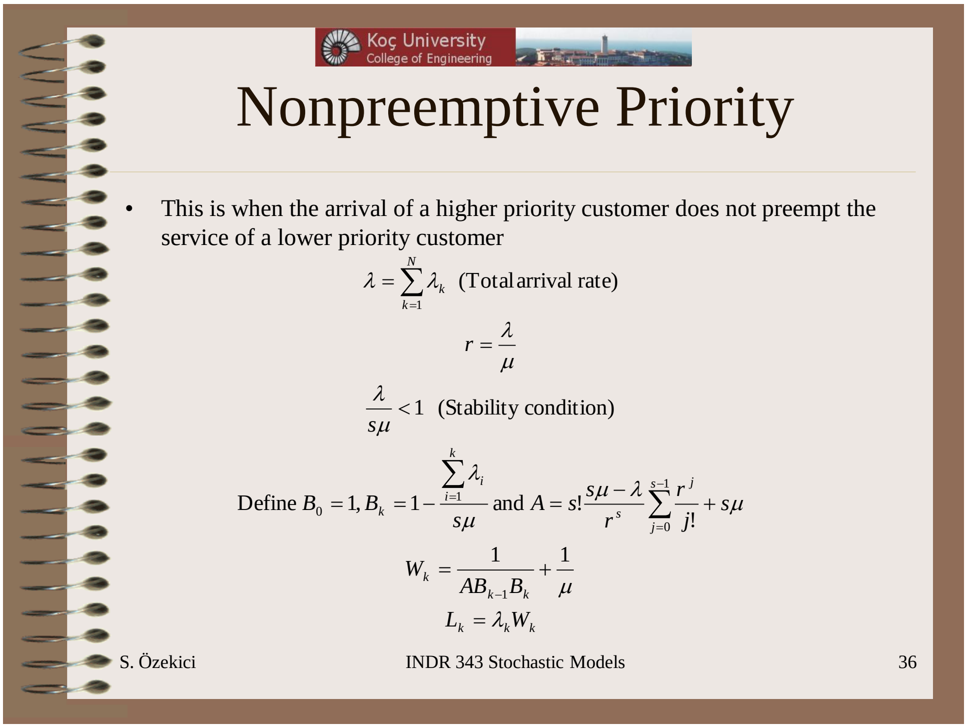


FIGURE 17.12
Values of L for the $E_k/M/2$ model (Sec. 17.7).



Priority Discipline Queues

- N priority classes
- Priority classes arrive according to independent Poisson processes
- The service durations are all exponential
- Whenever a server is idle, he picks the customer at the head of the line (FCFS) of the highest available priority queue
- s = number of parallel servers
- λ_k = arrival rate of priority class k
- μ = service rate of any customer
- W_k = average waiting time of a priority class k customer
- L_k = average number of priority class k customers in the system



Nonpreemptive Priority

- This is when the arrival of a higher priority customer does not preempt the service of a lower priority customer

$$\lambda = \sum_{k=1}^{N} \lambda_k \quad \text{(Total arrival rate)}$$

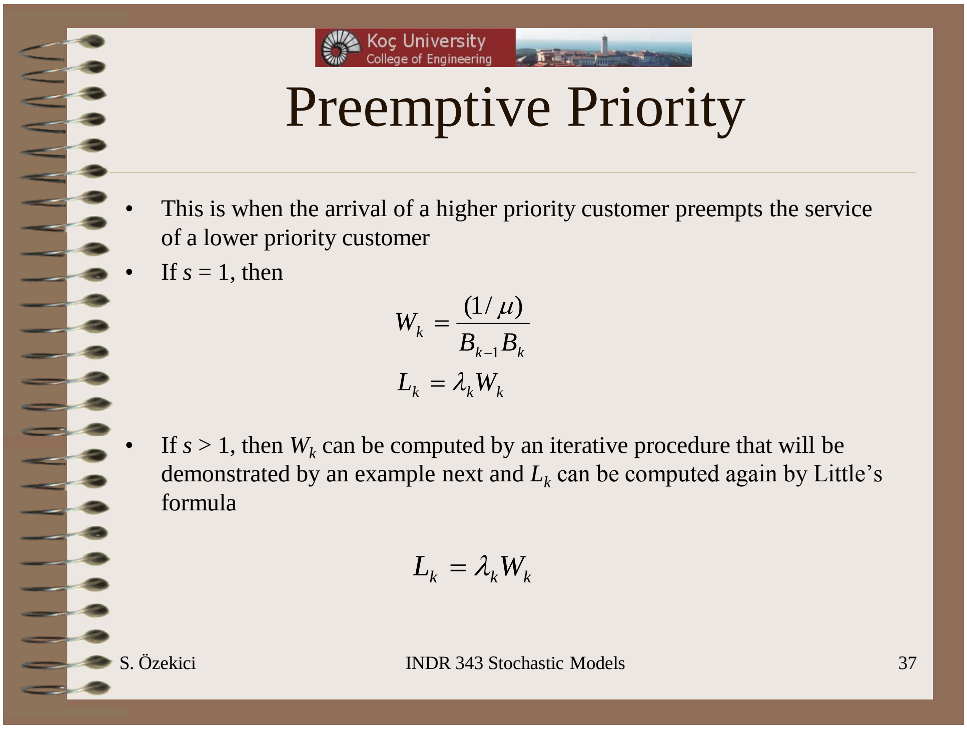
$$r = \frac{\lambda}{\mu}$$

$$\frac{\lambda}{s\mu} < 1 \quad \text{(Stability condition)}$$

$$\text{Define } B_0 = 1, B_k = 1 - \frac{\sum_{i=1}^{k} \lambda_i}{s\mu} \text{ and } A = s! \frac{s\mu - \lambda}{r^s} \sum_{j=0}^{s-1} \frac{r^j}{j!} + s\mu$$

$$W_k = \frac{1}{AB_{k-1}B_k} + \frac{1}{\mu}$$

$$L_k = \lambda_k W_k$$



Preemptive Priority

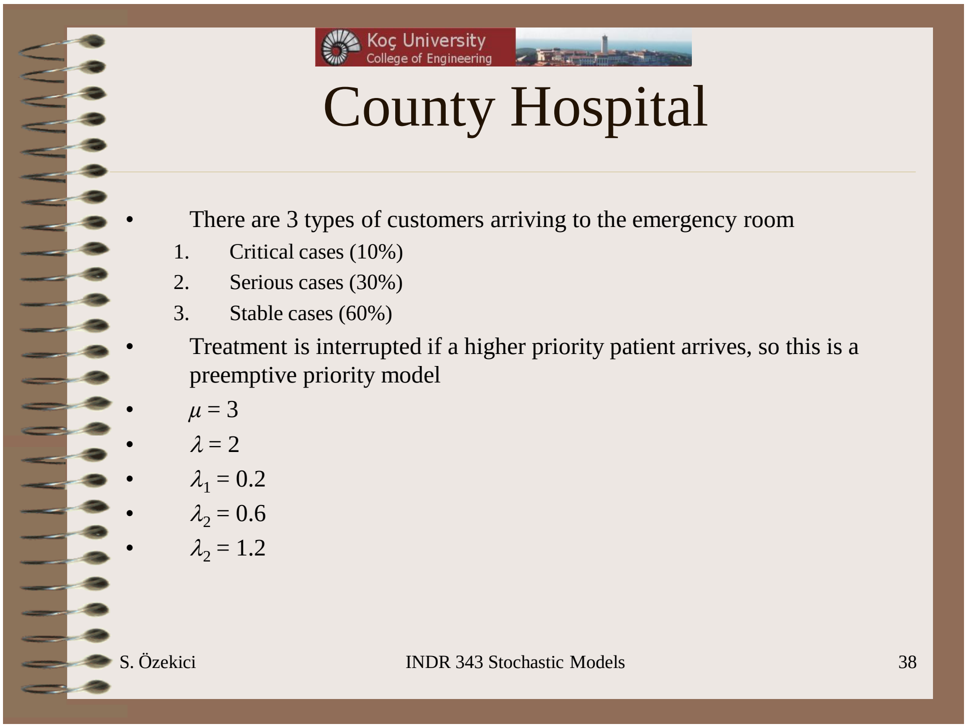
- This is when the arrival of a higher priority customer preempts the service of a lower priority customer
- If s = 1, then

$$W_k = \frac{(1/\mu)}{B_{k-1}B_k}$$

$$L_k = \lambda_k W_k$$

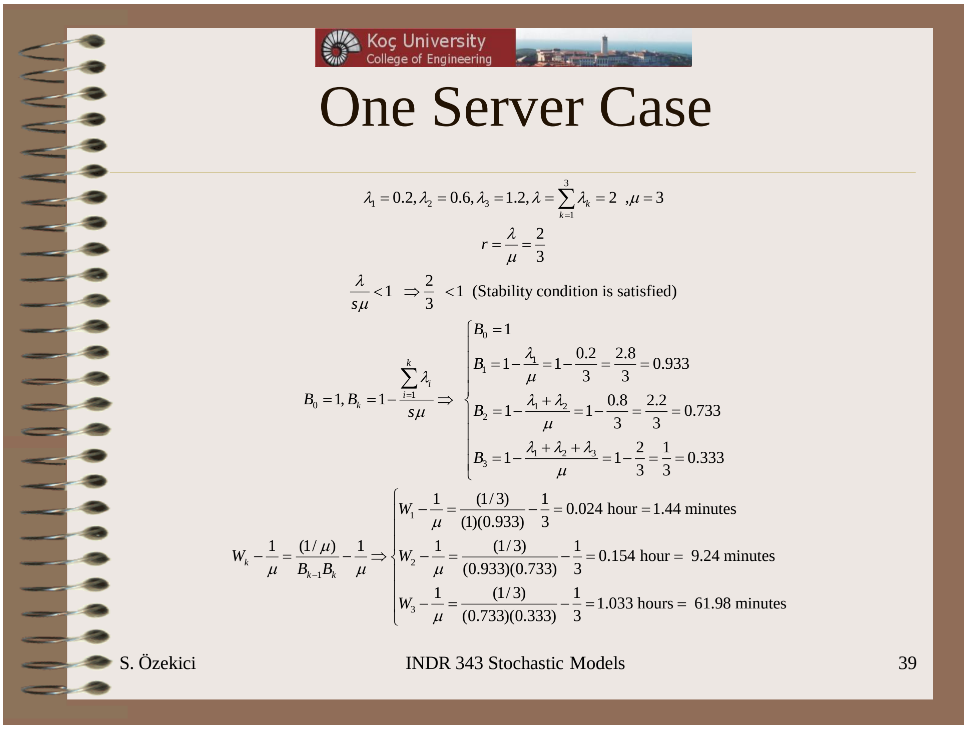
- If s > 1, then W_k can be computed by an iterative procedure that will be demonstrated by an example next and L_k can be computed again by Little's formula

$$L_k = \lambda_k W_k$$



County Hospital

- There are 3 types of customers arriving to the emergency room
  1. Critical cases (10%)
  2. Serious cases (30%)
  3. Stable cases (60%)
- Treatment is interrupted if a higher priority patient arrives, so this is a preemptive priority model
- $\mu = 3$
- $\lambda = 2$
- $\lambda_1 = 0.2$
- $\lambda_2 = 0.6$
- $\lambda_2 = 1.2$



Koç University
College of Engineering

One Server Case

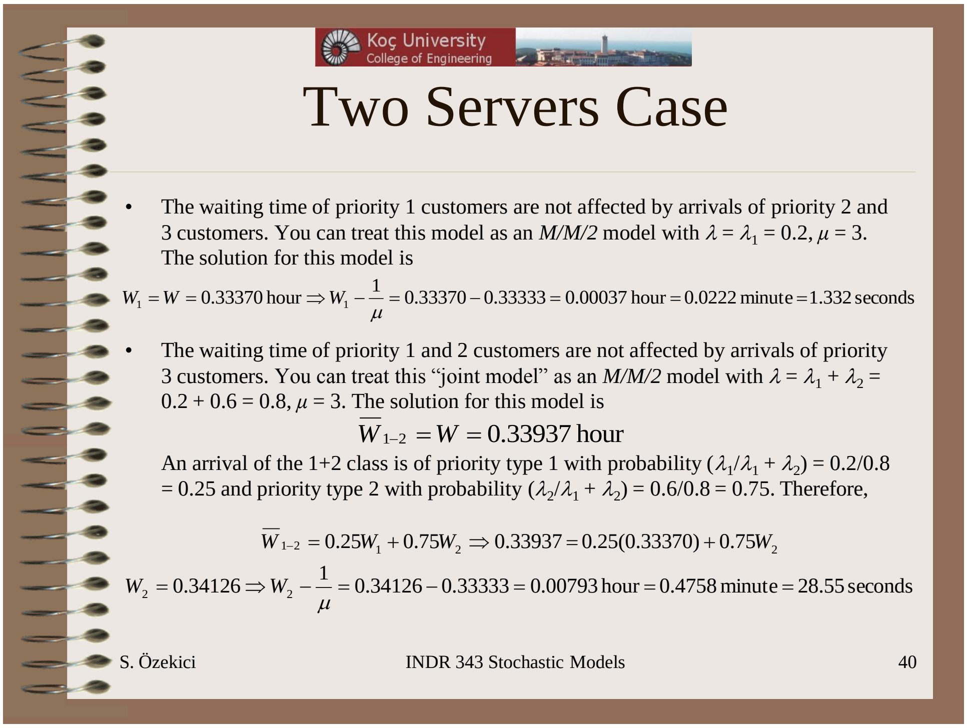
$$\lambda_1 = 0.2, \lambda_2 = 0.6, \lambda_3 = 1.2, \lambda = \sum_{k=1}^{3} \lambda_k = 2 \ \ , \mu = 3$$

$$r = \frac{\lambda}{\mu} = \frac{2}{3}$$

$$\frac{\lambda}{s\mu} < 1 \ \Rightarrow \frac{2}{3} \ < 1 \ \text{ (Stability condition is satisfied)}$$

$$B_0 = 1, B_k = 1 - \frac{\sum_{i=1}^{k} \lambda_i}{s\mu} \Rightarrow \begin{cases} B_0 = 1 \\ B_1 = 1 - \frac{\lambda_1}{\mu} = 1 - \frac{0.2}{3} = \frac{2.8}{3} = 0.933 \\ B_2 = 1 - \frac{\lambda_1 + \lambda_2}{\mu} = 1 - \frac{0.8}{3} = \frac{2.2}{3} = 0.733 \\ B_3 = 1 - \frac{\lambda_1 + \lambda_2 + \lambda_3}{\mu} = 1 - \frac{2}{3} = \frac{1}{3} = 0.333 \end{cases}$$

$$W_k - \frac{1}{\mu} = \frac{(1/\mu)}{B_{k-1} B_k} - \frac{1}{\mu} \Rightarrow \begin{cases} W_1 - \frac{1}{\mu} = \frac{(1/3)}{(1)(0.933)} - \frac{1}{3} = 0.024 \text{ hour} = 1.44 \text{ minutes} \\ W_2 - \frac{1}{\mu} = \frac{(1/3)}{(0.933)(0.733)} - \frac{1}{3} = 0.154 \text{ hour} = \ 9.24 \text{ minutes} \\ W_3 - \frac{1}{\mu} = \frac{(1/3)}{(0.733)(0.333)} - \frac{1}{3} = 1.033 \text{ hours} = \ 61.98 \text{ minutes} \end{cases}$$



Koç University
College of Engineering

Two Servers Case

- The waiting time of priority 1 customers are not affected by arrivals of priority 2 and 3 customers. You can treat this model as an *M/M/2* model with $\lambda = \lambda_1 = 0.2$, $\mu = 3$. The solution for this model is

$$W_1 = W = 0.33370 \text{ hour} \Rightarrow W_1 - \frac{1}{\mu} = 0.33370 - 0.33333 = 0.00037 \text{ hour} = 0.0222 \text{ minute} = 1.332 \text{ seconds}$$

- The waiting time of priority 1 and 2 customers are not affected by arrivals of priority 3 customers. You can treat this "joint model" as an *M/M/2* model with $\lambda = \lambda_1 + \lambda_2 = 0.2 + 0.6 = 0.8$, $\mu = 3$. The solution for this model is

$$\overline{W}_{1-2} = W = 0.33937 \text{ hour}$$

An arrival of the 1+2 class is of priority type 1 with probability $(\lambda_1/\lambda_1 + \lambda_2) = 0.2/0.8 = 0.25$ and priority type 2 with probability $(\lambda_2/\lambda_1 + \lambda_2) = 0.6/0.8 = 0.75$. Therefore,

$$\overline{W}_{1-2} = 0.25 W_1 + 0.75 W_2 \Rightarrow 0.33937 = 0.25(0.33370) + 0.75 W_2$$

$$W_2 = 0.34126 \Rightarrow W_2 - \frac{1}{\mu} = 0.34126 - 0.33333 = 0.00793 \text{ hour} = 0.4758 \text{ minute} = 28.55 \text{ seconds}$$

Two Servers Case

- The waiting time of priority 1, 2 and 3 customers can be found by treating this “joint model” as an *M/M/2* model with $\lambda = \lambda_1 + \lambda_2 + \lambda_3 = 0.2 + 0.6 + 1.2 = 2$, $\mu = 3$. The solution for this model is

$$\overline{W}_{1-3} = W = 0.375 \text{ hour}$$

We know that an arrival of the 1+2 +3 class is of priority types 1, 2 or 3 with probabilities 0.1, 0.3 and 0.6 respectively. Therefore,

$$\overline{W}_{1-3} = 0.1W_1 + 0.2W_2 + 0.6W_3 \Rightarrow 0.375 = 0.1(0.33370) + 0.3(0.34126) + 0.6W_3$$

$$W_3 = 0.39875 \Rightarrow W_3 - \frac{1}{\mu} = 0.39875 - 0.33333 = 0.06542 \text{ hour} = 3.93 \text{ minutes}$$

- There is a substantial improvement in the waiting times if 2 doctors are employed.

Comparison

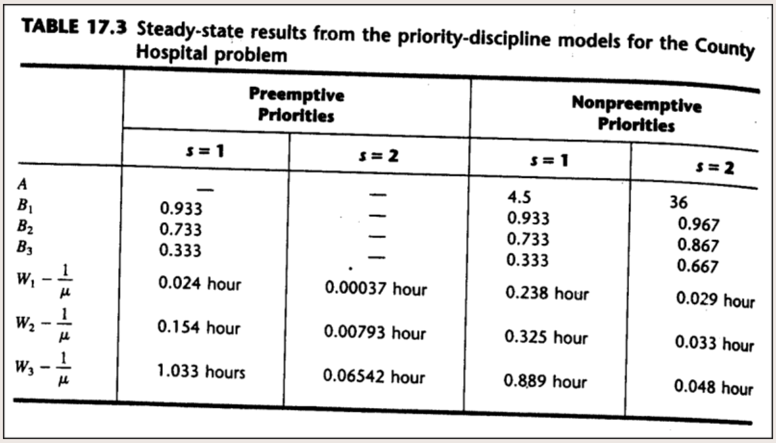


TABLE 17.3 Steady-state results from the priority-discipline models for the County Hospital problem

	Preemptive Priorities		Nonpreemptive Priorities	
	$s = 1$	$s = 2$	$s = 1$	$s = 2$
A	—	—	4.5	36
B_1	0.933	—	0.933	0.967
B_2	0.733	—	0.733	0.867
B_3	0.333	—	0.333	0.667
$W_1 - \frac{1}{\mu}$	0.024 hour	0.00037 hour	0.238 hour	0.029 hour
$W_2 - \frac{1}{\mu}$	0.154 hour	0.00793 hour	0.325 hour	0.033 hour
$W_3 - \frac{1}{\mu}$	1.033 hours	0.06542 hour	0.889 hour	0.048 hour

Queueing Networks

- **Equivalence Property**: For the *M/M/s* queueing model with infinite capacity and customer arrival rate λ, the steady-state output process is also a Poisson process with rate λ.
- **Infinite Queues in Series**: Suppose that there are *m* service facilities connected in series where customers arrive at the first one according to a Poisson process with rate λ and pass through all of the facilites in the same order. There are s_i servers in the *i*th facility who work exponentially at rate μ_i. Then the joint distribution of the number of customers in the *m* service facilities of the network has the following **product form solution**

$$P\{(N_1, N_2, \cdots, N_m) = (n_1, n_2, \cdots, n_m)\} = P\{N_1 = n_1\}P\{N_2 = n_2\}\cdots P\{N_m = n_m\} = P_{n_1}P_{n_2}\cdots P_{n_m}$$

- In other words, under steady-state conditions, the *i*th facility can be treated as an independent *M/M/s* queueing system with arrival rate λ and service rate μ_i

Jackson Networks

- A Jackson network is a system of *m* service facilities where facility *i* has
  - infinite queue capacity
  - customers arrive from the outside according to a Poisson process with rate a_i
  - s_i servers with an exponential service time distribution with rate μ_i
  - a customer who leaves facility *i* is routed to facility *j* with probability P_{ij} or departs the system with probability

$$q_i = 1 - \sum_{j=1}^{m} P_{ij}$$

- Under steady-state conditions, each facility *j* behaves as an independent *M/M/s* queueing system with arrival rate

$$\lambda_j = a_j + \sum_{i=1}^{m} \lambda_i P_{ij} \quad (s_j \mu_j > \lambda_j)$$

Example

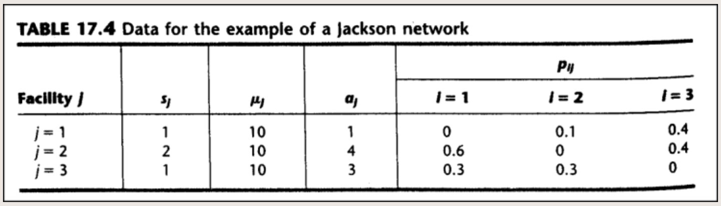


TABLE 17.4 Data for the example of a Jackson network

Facility j	s_j	μ_j	a_j	p_{ij}		
				$i = 1$	$i = 2$	$i = 3$
$j = 1$	1	10	1	0	0.1	0.4
$j = 2$	2	10	4	0.6	0	0.4
$j = 3$	1	10	3	0.3	0.3	0

$$\begin{array}{lllllllll}
\lambda_1 & = & 1 & + & & + & 0.1\lambda_2 & + & 0.4\lambda_3 \\
\lambda_2 & = & 4 & + & 0.6\lambda_1 & + & & + & 0.4\lambda_3 \\
\lambda_3 & = & 3 & + & 0.3\lambda_1 & + & 0.3\lambda_2 & + &
\end{array}$$

$$\lambda_1 = 5, \lambda_2 = 10, \lambda_3 = 7.5$$

Example

$$\rho_i = \frac{\lambda_i}{s_i \mu_i} = \begin{cases} \frac{1}{2} & \text{for } i = 1 \\ \frac{1}{2} & \text{for } i = 2 \\ \frac{3}{4} & \text{for } i = 3 \end{cases}$$

$$P_{n_1} = \frac{1}{2}\left(\frac{1}{2}\right)^{n_1} \quad \text{(Facility 1)}$$

$$P_{n_2} = \begin{cases} \frac{1}{3} & \text{for } n_2 = 0 \text{ or } 1 \\ \frac{1}{3}\left(\frac{1}{2}\right)^{n_2 - 1} & \text{for } n_2 \geq 2 \end{cases} \quad \text{(Facility 2)}$$

$$P_{n_3} = \frac{1}{4}\left(\frac{3}{4}\right)^{n_3} \quad \text{(Facility 3)}$$

Example

$$L_1 = 1, L_2 = \frac{4}{3}, L_3 = 3$$

$$L = L_1 + L_2 + L_3 = 5\frac{1}{3}$$

$$\lambda = a_1 + a_2 + a_3 = 1 + 4 + 3 = 8$$

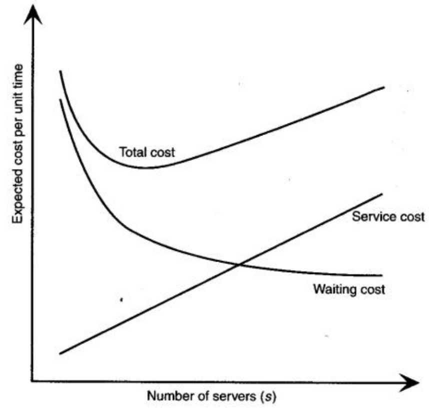
$$W = \frac{L}{\lambda} = \frac{16/3}{8} = \frac{2}{3} \quad \text{(Little's formula)}$$

Application of Queueing Theory

- There are many parameters associated with a queueing system that are in fact decision variable, like
  - The number of servers at a service facility
  - The service rate or efficiency of the servers
  - The number of service facilities
  - The system capacity
  - The size of the calling population
  - The service discipline

- There are tradeoffs between the service cost (SC) of providing the service and the waiting cost (WC) of customers

The Costs



■ FIGURE 17.13
The shape of the expected cost curves for determining the number of servers to provide.

Waiting Cost

- If $g(n)$ is the waiting cost incurred per unit time when there are n customers in the system, then

$$E[WC]=E[g(N)]=\sum_{n=0}^{+\infty} g(n)P_n$$

- In the linear case, if $g(n) = C_w n$ then

$$E[WC]=E[g(N)]=E[C_w N]=C_w E[N]=C_w L$$

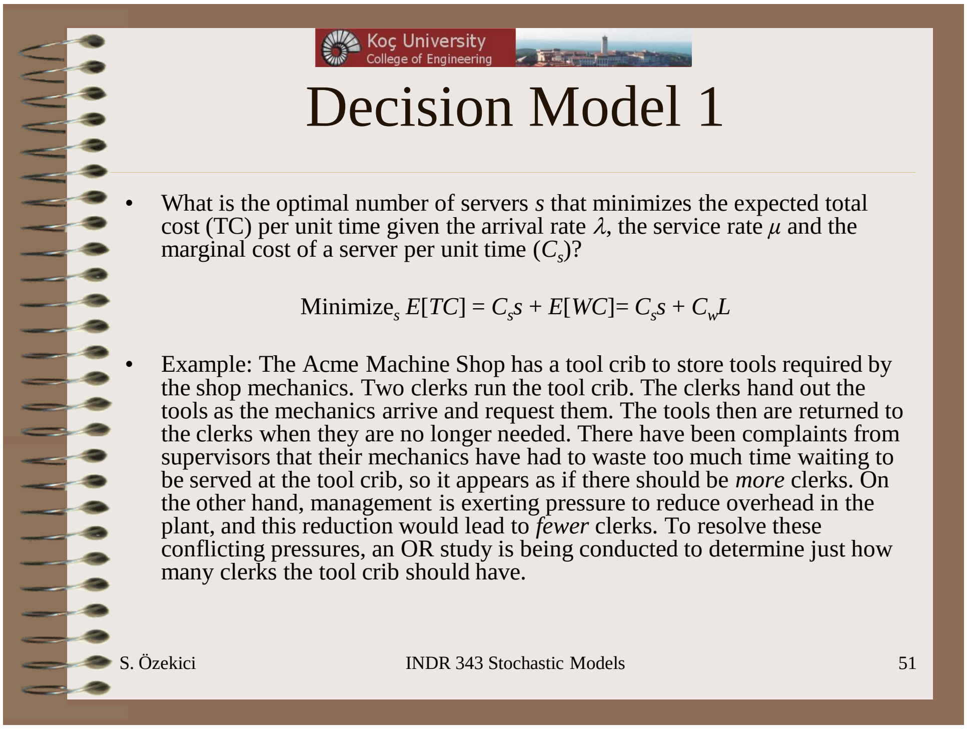
- If $h(w)$ is the waiting cost incurred if a customer waits for w units of time in the system, then

$$E[h(\mathcal{W})]=\int_0^{+\infty} h(w)f_{\mathcal{W}}(w)dw$$

$$E[WC]=\lambda E[h(\mathcal{W})]=\lambda\int_0^{+\infty} h(w)f_{\mathcal{W}}(w)dw$$

- In the linear case, if $h(w) = C_w w$ then

$$E[WC]=\lambda E[h(\mathcal{W})]=\lambda E[C_w\mathcal{W}]=\lambda C_w E[\mathcal{W}]=C_w\lambda W=C_w L$$

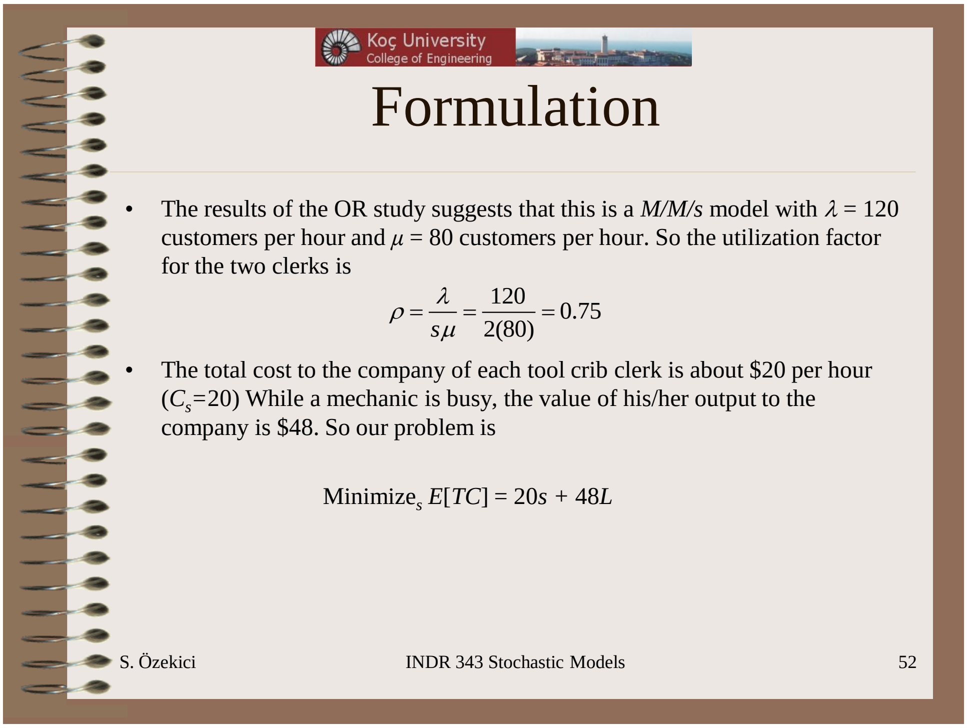


Decision Model 1

- What is the optimal number of servers s that minimizes the expected total cost (TC) per unit time given the arrival rate λ, the service rate μ and the marginal cost of a server per unit time (C_s)?

$$\text{Minimize}_s\ E[TC] = C_s s + E[WC] = C_s s + C_w L$$

- Example: The Acme Machine Shop has a tool crib to store tools required by the shop mechanics. Two clerks run the tool crib. The clerks hand out the tools as the mechanics arrive and request them. The tools then are returned to the clerks when they are no longer needed. There have been complaints from supervisors that their mechanics have had to waste too much time waiting to be served at the tool crib, so it appears as if there should be *more* clerks. On the other hand, management is exerting pressure to reduce overhead in the plant, and this reduction would lead to *fewer* clerks. To resolve these conflicting pressures, an OR study is being conducted to determine just how many clerks the tool crib should have.



Formulation

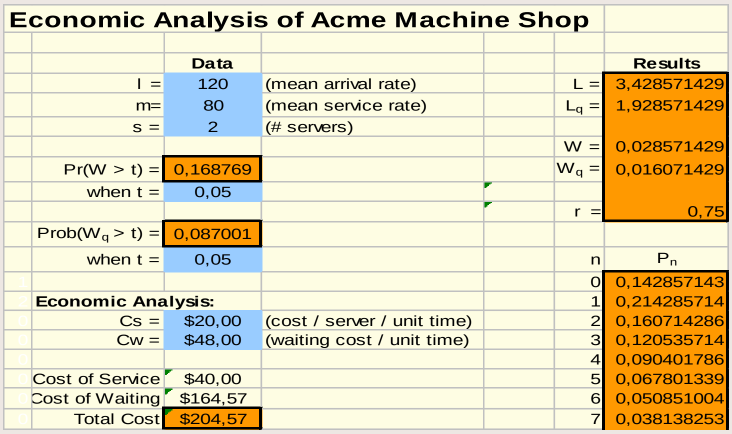
- The results of the OR study suggests that this is a *M/M/s* model with $\lambda = 120$ customers per hour and $\mu = 80$ customers per hour. So the utilization factor for the two clerks is

$$\rho = \frac{\lambda}{s\mu} = \frac{120}{2(80)} = 0.75$$

- The total cost to the company of each tool crib clerk is about $20 per hour ($C_s$=20) While a mechanic is busy, the value of his/her output to the company is $48. So our problem is

$$\text{Minimize}_s\ E[TC] = 20s + 48L$$

Economic Analysis



Economic Analysis of Acme Machine Shop					
	Data				Results
l =	120	(mean arrival rate)		L =	3,428571429
m=	80	(mean service rate)		L_q =	1,928571429
s =	2	(# servers)			
				W =	0,028571429
Pr(W > t) =	0,168769			W_q =	0,016071429
when t =	0,05				
				r =	0,75
Prob(W_q > t) =	0,087001				
when t =	0,05			n	P_n
				0	0,142857143
Economic Analysis:				1	0,214285714
Cs =	$20,00	(cost / server / unit time)		2	0,160714286
Cw =	$48,00	(waiting cost / unit time)		3	0,120535714
				4	0,090401786
Cost of Service	$40,00			5	0,067801339
Cost of Waiting	$164,57			6	0,050851004
Total Cost	$204,57			7	0,038138253

Optimal Solution

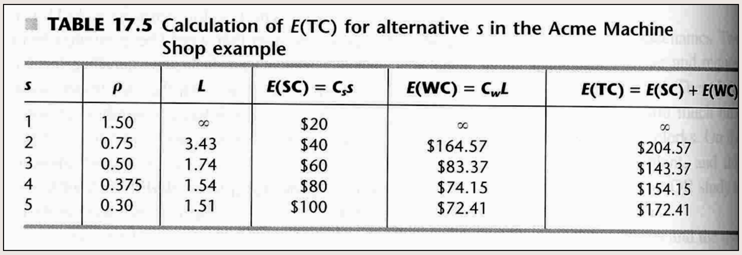
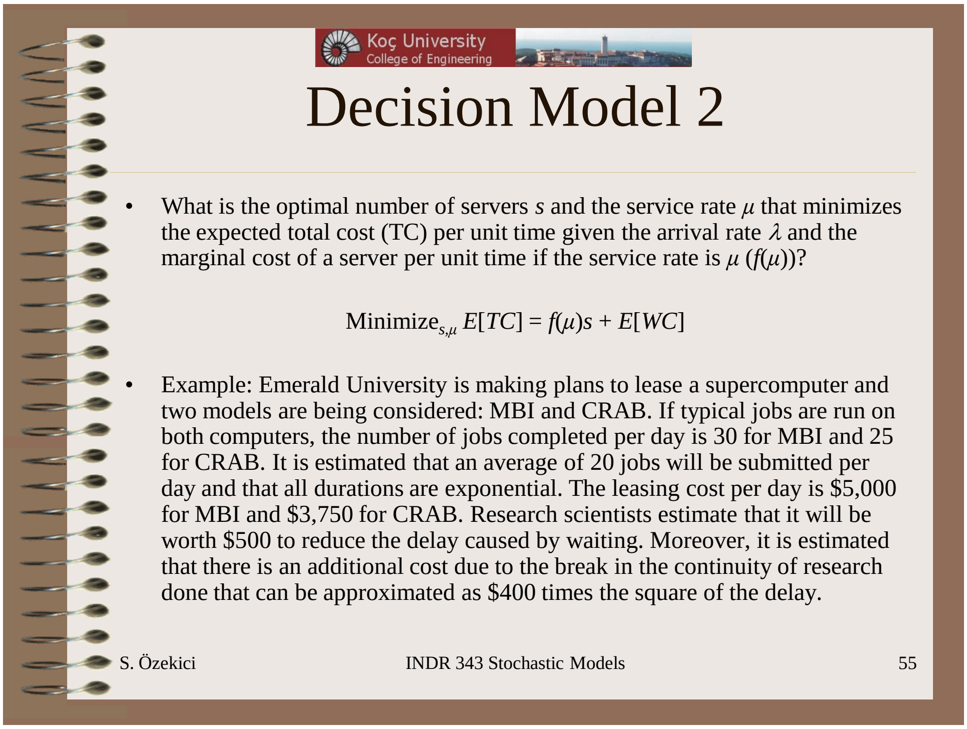


TABLE 17.5 Calculation of $E(TC)$ for alternative s in the Acme Machine Shop example

s	ρ	L	$E(SC) = C_s s$	$E(WC) = C_w L$	$E(TC) = E(SC) + E(WC)$
1	1.50	∞	$20	∞	∞
2	0.75	3.43	$40	$164.57	$204.57
3	0.50	1.74	$60	$83.37	$143.37
4	0.375	1.54	$80	$74.15	$154.15
5	0.30	1.51	$100	$72.41	$172.41

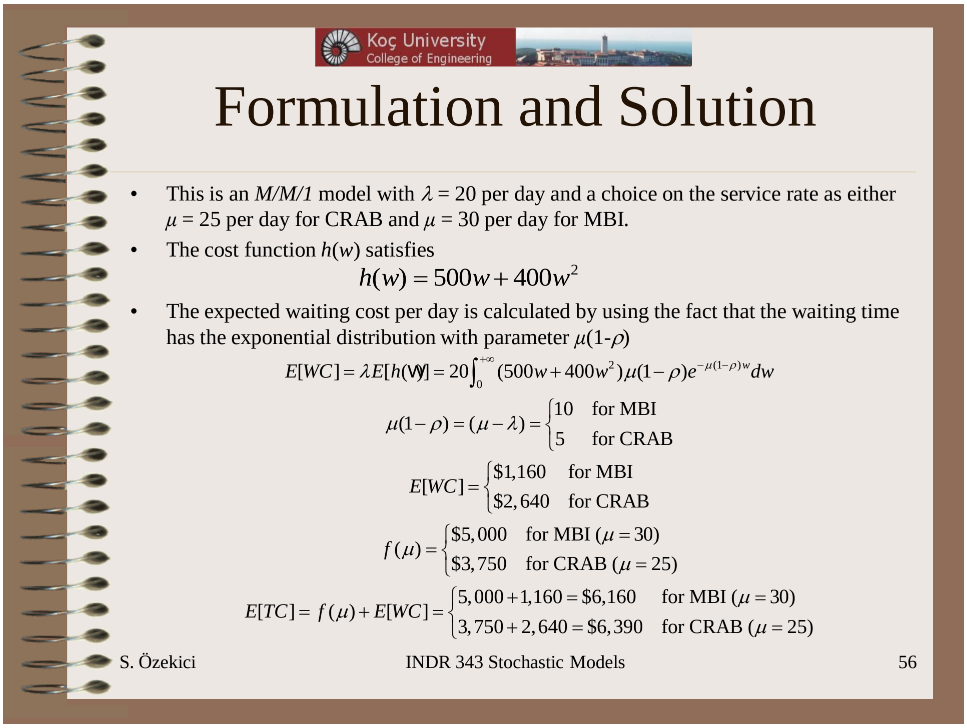


Decision Model 2

- What is the optimal number of servers s and the service rate μ that minimizes the expected total cost (TC) per unit time given the arrival rate λ and the marginal cost of a server per unit time if the service rate is μ ($f(\mu)$)?

$$\text{Minimize}_{s,\mu}\ E[TC] = f(\mu)s + E[WC]$$

- Example: Emerald University is making plans to lease a supercomputer and two models are being considered: MBI and CRAB. If typical jobs are run on both computers, the number of jobs completed per day is 30 for MBI and 25 for CRAB. It is estimated that an average of 20 jobs will be submitted per day and that all durations are exponential. The leasing cost per day is $5,000 for MBI and $3,750 for CRAB. Research scientists estimate that it will be worth $500 to reduce the delay caused by waiting. Moreover, it is estimated that there is an additional cost due to the break in the continuity of research done that can be approximated as $400 times the square of the delay.



Formulation and Solution

- This is an *M/M/1* model with λ = 20 per day and a choice on the service rate as either μ = 25 per day for CRAB and μ = 30 per day for MBI.
- The cost function $h(w)$ satisfies

$$h(w) = 500w + 400w^2$$

- The expected waiting cost per day is calculated by using the fact that the waiting time has the exponential distribution with parameter $\mu(1-\rho)$

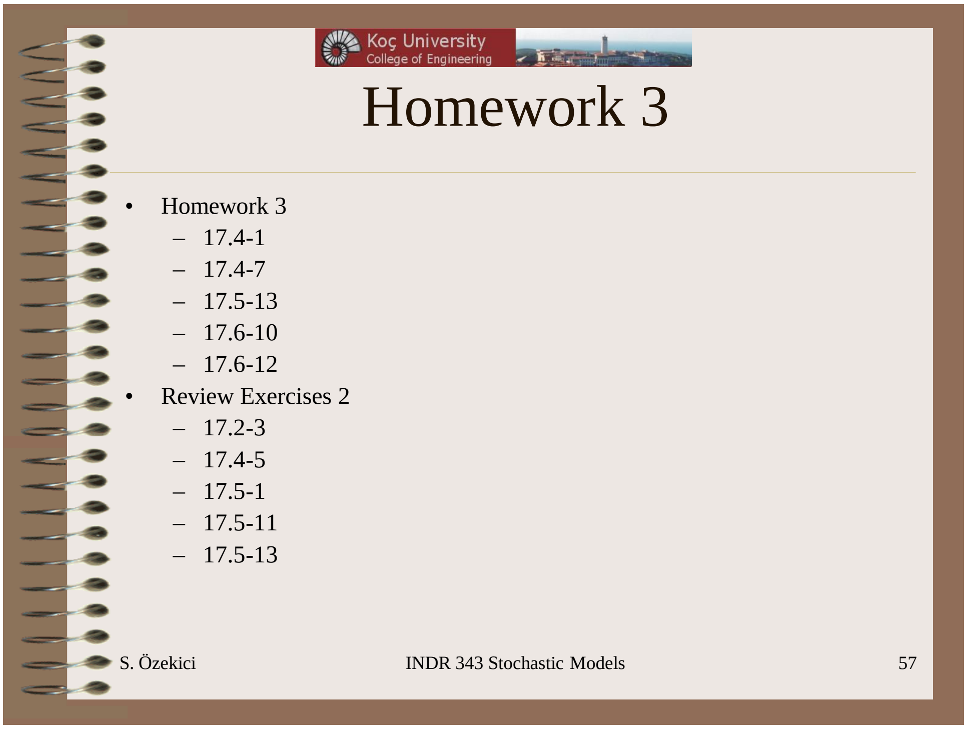
$$E[WC] = \lambda E[h(W)] = 20\int_0^{+\infty}(500w + 400w^2)\mu(1-\rho)e^{-\mu(1-\rho)w}dw$$

$$\mu(1-\rho) = (\mu-\lambda) = \begin{cases} 10 & \text{for MBI} \\ 5 & \text{for CRAB} \end{cases}$$

$$E[WC] = \begin{cases} \$1{,}160 & \text{for MBI} \\ \$2{,}640 & \text{for CRAB} \end{cases}$$

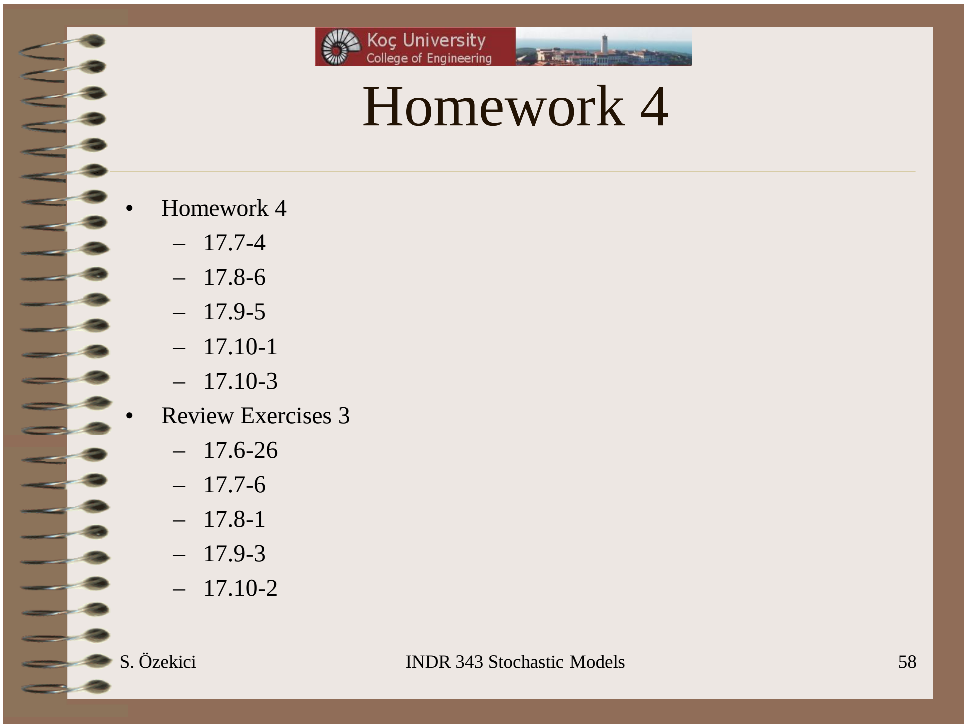
$$f(\mu) = \begin{cases} \$5{,}000 & \text{for MBI } (\mu = 30) \\ \$3{,}750 & \text{for CRAB } (\mu = 25) \end{cases}$$

$$E[TC] = f(\mu) + E[WC] = \begin{cases} 5{,}000 + 1{,}160 = \$6{,}160 & \text{for MBI } (\mu = 30) \\ 3{,}750 + 2{,}640 = \$6{,}390 & \text{for CRAB } (\mu = 25) \end{cases}$$



Homework 3

- Homework 3
  - 17.4-1
  - 17.4-7
  - 17.5-13
  - 17.6-10
  - 17.6-12
- Review Exercises 2
  - 17.2-3
  - 17.4-5
  - 17.5-1
  - 17.5-11
  - 17.5-13



Homework 4

- Homework 4
  - 17.7-4
  - 17.8-6
  - 17.9-5
  - 17.10-1
  - 17.10-3
- Review Exercises 3
  - 17.6-26
  - 17.7-6
  - 17.8-1
  - 17.9-3
  - 17.10-2