## ECE560: Computer Systems Performance Evaluation



Lecture \# 13 -
Queueing Systems (II)

Instructor: Dr. Liudong Xing

## Administration Issues

- Annotated Bibliography
- Due: March 22, Friday
- Refer to Section 2.2 in the Project Description for the guidelines
- If you are looking for jobs or internship:
- Wednesday, March 27: Job, Internship, and Graduate School Expo, 12-3:00 p.m., Tripp Athletic Center
- Survey results (13 responses)

1. How was the pace of the class?

Too slow $\qquad$ Slow $\qquad$ Just right_13 $\qquad$ Fast _Too fast $\qquad$
2. How was the workload of ECE560 class, when compared to your other classes? Much less_1 $\qquad$ Less__ ${ }^{2}$ $\qquad$ Similar_10 $\qquad$ More $\qquad$ Much more
3. How is the level of detail of material covered in the lectures? Not enough $\qquad$ Just right_ ${ }^{11}$ $\qquad$ Too much _2 $\qquad$
4. How clear are the lectures/examples?

Not clear $\qquad$ Okay _ 0.5 _
5. How was the midterm exam?
Too easy
$\qquad$ Easy _5 $\qquad$ Just right __ -
6. How were the homework assignments?

Too easy ___ Easy _1__ Just right_ ${ }^{11} \ldots$ Hard __ ${ }^{1}$ Too hard _ $\qquad$

## Review of Lecture \#11

- Overview of queueing systems
(Kendall notation, Little's Law)
- Performance measures
- Traffic intensity ( $\alpha$ )
- Server utilization ( $\rho$ )
$-\pi_{\mathrm{n}}$ : steady-state probability that there are $n$ customers in the queueing system
- Throughput ( $\gamma$ ): rate at which jobs successfully depart from the system
- Average number of jobs in the system (L)
- Average time in the system (W)
- Average number of jobs in the queue ( $\mathrm{L}_{\mathrm{q}}$ )
- Average time in the queue $\left(\mathrm{W}_{\mathrm{q}}\right)$
- $\mathrm{D} / \mathrm{D} / 1$ queueing systems
- $M / M / 1$ queueing systems

- Performance measures:

$$
\begin{aligned}
& \alpha=\rho=\lambda / \mu \\
& \pi_{0}=1-\frac{\lambda}{\mu}=1-\rho, \pi_{n}=\left(\frac{\lambda}{\mu}\right)^{n} \pi_{0}=\rho^{n}(1-\rho) \\
& \gamma=\mu P[>0 \text { jobs in the system }] \\
&=\mu(1-P[0 \text { jobs in the system }]) \\
&=\mu\left(1-\pi_{0}\right)=\mu(1-(1-\rho))=\mu \rho=\lambda \\
& L=\sum_{n=0}^{\infty} n \pi_{n}=(1-\rho) \sum_{n=0}^{\infty} n \rho^{n} \\
&=(1-\rho) \rho \sum_{n=1}^{\infty} n \rho^{n-1}=\frac{(1-\rho) \rho}{(1-\rho)^{2}}=\frac{\rho}{1-\rho} \\
& W=L / \lambda=\frac{\rho}{1-\rho} / \lambda=\frac{1}{\mu-\lambda} \\
& L_{q}=L-(1 * P[\text { Server is not empty }] \\
&=L-(1-P[0 \text { jobs in the system }]) \\
&=L-\left(1-\pi_{0}\right)=L-(1-(1-\rho)) \\
&=L-\rho=\frac{\rho}{1-\rho}-\rho=\frac{\rho^{2}}{1-\rho} \\
& W_{q}=L_{q} / \lambda=\frac{\rho^{2}}{1-\rho} \frac{1}{\lambda} \text { or } W_{q}=W-W_{s}=\frac{\rho}{(1-\rho) \lambda}-\frac{1}{\mu}=\frac{\rho^{2}}{1-\rho} \frac{1}{\lambda} \\
& \text { Xing } \odot
\end{aligned}
$$



Topics

- More Birth-and-Death Queues (Cont'd)
- M/M/1/N Queues
- M/M/c Queues
- M/M/ $\infty$ Queues
- M/M/1/k/k Queues

Dr. Xing ©
Q-Systems

## M/M/1/N Queues

- The infinite buffer/queue assumption is unrealistic in practice!
- $N$ : the system capacity/the maximum number of customers allowed in the system: 1 in service and ( $\mathrm{N}-1$ ) in the queue
- State transition diagram:


M/M/1/N: Steady-State Probabilities


- The same Balance Equations as $\mathrm{M} / \mathrm{M} / 1$, thus,

$$
\left.\begin{array}{c}
\begin{array}{l}
\lambda \pi_{0}=\mu \pi_{1} \\
\lambda \pi_{1}=\mu \pi_{2} \\
\ldots \ldots \ldots \ldots \\
\lambda \pi_{N-1}=\mu \pi_{N}
\end{array} \\
\stackrel{\begin{array}{l}
\text { In general } \\
\Rightarrow \\
\pi_{n}
\end{array}}{ }=\frac{\lambda}{\mu} \pi_{n-1}=\left(\frac{\lambda}{\mu}\right)^{n} \pi_{0}=\alpha^{n} \pi_{0} \\
\text { for } n=0,1,2, \ldots, N, \quad \alpha=\lambda / \mu
\end{array}\right] \begin{aligned}
& \sum_{n=0}^{N} \pi_{n}=\pi_{0} \sum_{n=0}^{N} \alpha^{n}=1 \Rightarrow \pi_{0}=\frac{1}{\sum_{n=0}^{N} \alpha^{n}}=\frac{1}{\frac{1-\alpha^{N+1}}{1-\alpha}}=\frac{1-\alpha}{1-\alpha^{N+1}} \\
& \text { Dr. Xing © }
\end{aligned}
$$

## M/M/1/N: Performance Measures (I)

## M/M/1/N: Performance Measures (II)

$$
\begin{aligned}
\alpha & =\lambda / \mu \\
\pi_{0} & =\frac{1-\alpha}{1-\alpha^{N+1}} \\
\pi_{n} & =\alpha^{n} \pi_{0}=\frac{\alpha^{n}(1-\alpha)}{1-\alpha^{N+1}} \text { for } n=0,1,2, \ldots, N
\end{aligned}
$$

Blocking probability $\left(\mathrm{P}_{\mathrm{B}}\right)$ : probability an arriving job
The effective / average arrival rate of customers into the system:

$$
\begin{gathered}
\lambda_{e f f}=\lambda\left(1-P_{B}\right) \\
P_{\mathrm{B}}=\pi_{N}=\alpha^{N} \pi_{0}=\frac{\alpha^{N}(1-\alpha)}{1-\alpha^{N+1}}
\end{gathered}
$$

The true server utilization rate $\rho$ - probability that is turned away due to a full buffer

$$
P_{\mathrm{B}}=\pi_{N}=\alpha^{N} \pi_{0}=\frac{\alpha^{N}(1-\alpha)}{1-\alpha^{N+1}}
$$

For example: for $\alpha=0.5$, if want a blocking probability of $<10^{-6}$, what is the minimum system capacity?

M/M/1/N: Performance Measures (III)
Throughput ( $\gamma$ ): rate at which jobs successfully depart
M/M/1/N: Performance Measures (IV)
from the system

$$
\begin{aligned}
\gamma & =\mu P[>0 \text { jobs in the system }]+0 \mathrm{P}[0 \text { jobs }] \\
& =\mu(1-P[0 \text { jobs in the system }]) \\
& =\mu\left(1-\pi_{0}\right)=\mu\left(1-\left(\frac{1-\alpha}{1-\alpha^{N+1}}\right)\right) \\
& =\mu \frac{\alpha\left(1-\alpha^{N}\right)}{1-\alpha^{N+1}}=\frac{\lambda\left(1-\alpha^{N}\right)}{1-\alpha^{N+1}}
\end{aligned}
$$

Alternative way to compute $\gamma$ (by looking at the input side)
Everything that arrives and is not blocked must
eventually depart

$$
\gamma=\lambda\left(1-P_{B}\right)=\lambda\left(1-\alpha^{N} \frac{1-\alpha}{1-\alpha^{N+1}}\right)=\frac{\lambda\left(1-\alpha^{N}\right)}{1-\alpha^{N+1}}
$$

Average number of customers in the system $L$ :

$$
\begin{aligned}
& L=\sum_{n=0}^{N} n \pi_{n}=\frac{1-\alpha}{1-\alpha^{N+1}} \sum_{n=0}^{N} n \alpha^{n} \\
& =(1-\alpha) \alpha / 1-\alpha^{N+1} \sum_{n=1}^{N} n \alpha^{n-1} \\
& =\frac{(1-\alpha) \alpha}{1-\alpha^{N+1}}\left(\sum_{n=0}^{N} \alpha^{n}\right)^{\prime}=\frac{(1-\alpha) \alpha}{1-\alpha^{N+1}}\left(\frac{1-\alpha^{N+1}}{1-\alpha}\right)^{\prime} \\
& =\frac{\alpha}{1-\alpha}-(N+1) \frac{\alpha^{N+1}}{1-\alpha^{N+1}}
\end{aligned}
$$

Average response time $W$ :

$$
W=L / \lambda_{e f f}=\left[\frac{\alpha}{1-\alpha}-\frac{(N+1) \alpha^{N+1}}{1-\alpha^{N+1}}\right] / \lambda_{e f f}
$$

Note: effective arrival rate $\lambda_{e f f}$ used

$$
\text { in Little's Law! } \lambda_{\text {eff }}=\lambda\left(1-P_{B}\right)
$$

Dr. Xing ©
Q-Systems

M/M/1/N: Performance Measures (V)

$$
\begin{aligned}
& \mathrm{M} / \mathrm{M} / 1 / \mathrm{N}: \text { Performance Measures } \\
& \alpha=\lambda / \mu \quad \text { (Summary) } \\
& \pi_{0}=\frac{1-\alpha}{1-\alpha^{N+1}}, \\
& \pi_{n}=\alpha^{n} \pi_{0}=\frac{\alpha^{n}(1-\alpha)}{1-\alpha^{N+1}} \text { for } n=0,1,2, \ldots, N \\
& P_{\mathrm{B}}=\pi_{N}=\alpha^{N} \pi_{0}=\frac{\alpha^{N}(1-\alpha)}{1-\alpha^{N+1}} \\
& \lambda_{e f f}=\lambda\left(1-P_{B}\right) \\
& \rho=1-\pi_{0}=1-\frac{1-\alpha}{1-\alpha^{N+1}}, \\
& \gamma=\lambda\left(1-P_{B}\right)=\lambda\left(1-\alpha^{N} \frac{1-\alpha}{1-\alpha^{N+1}}\right)=\frac{\lambda\left(1-\alpha^{N}\right)}{1-\alpha^{N+1}} \\
& L=\frac{\alpha}{1-\alpha}-(N+1) \frac{\alpha^{N+1}}{1-\alpha^{N+1}} \\
& W=L / \lambda_{e f f}=\left[\frac{\alpha}{1-\alpha}-\frac{(N+1) \alpha^{N+1}}{1-\alpha^{N+1}}\right] / \lambda_{e f f} \\
& L_{q}=L-\left(1-\pi_{0}\right) \\
& W_{q}=L_{q} / \lambda_{e f f} \text { or } W_{q}=W-W_{s}=W-\frac{1}{\mu}
\end{aligned}
$$

## Approximation of a Finite-Buffer System by the Infinite Buffer Model

## Hands-On Problem

- Consider a computer system with one processor and a queue with 2 buffers. The job requests arrive to the processor at the rate of 16 requests per second with Poisson pattern. The time to service a job request at the processor is exponentially distributed with a mean of 50 milliseconds. Answer the following questions:
- What is the probability of the processor being busy?
- What is the effective arrival rate of job requests into the system?
- Assume a job request in the queue and not being serviced can depart without service; this behavior is called "defect". Assume the defect process is also exponential with the constant rate of $\delta=2$ requests/second.
- Draw the complete state-transition diagram for the queueing computer system with the above described defect behavior.
- Pick any one state and write down the balance equation for that state.
very good approximation of a finite buffer system $\mathrm{M} / \mathrm{M} / 1 / \mathrm{N}$, even for moderate buffer sizes.

Dr. Xing ©
Q-Systems
Agenda

- More Birth-and-Death Queues
- M/M/1/N Queues
$-\underline{M / M / c}$ Queues
$-\mathrm{M} / \mathrm{M} / \infty$ Queues
$-\mathrm{M} / \mathrm{M} / 1 / \mathrm{k} / \mathrm{k}$ Queues
Q-Systems
Dr. Xing ©



## M/M/c Queues (I)

- Have c identical servers

- A birth-and-death process with coefficients:

$$
\begin{aligned}
& \lambda_{n}=\lambda, n=0,1,2,3, \ldots \ldots . \\
& \mu_{n}=\left\{\begin{array}{l}
n \mu, n=1,2, \ldots, c \\
c \mu, n \geq c
\end{array}\right.
\end{aligned}
$$

- Traffic intensity and server utilization:

$$
\begin{aligned}
& \alpha=\lambda / \mu \\
& \rho=\alpha / c=\lambda / c \mu
\end{aligned}
$$

$$
\begin{aligned}
& \text { M/M/c Queues (II) } \\
& \begin{array}{l}
(\lambda+\mu) \pi_{1}=\lambda \pi_{0}+2 \mu \pi_{2} \\
(\lambda+2 \mu) \pi_{2}=\lambda \pi_{1}+3 \mu \pi_{3} \\
1 \\
1 \\
1 \\
\cdots \\
\mathrm{c}+1
\end{array} \\
& \begin{array}{l}
\lambda \pi_{2}=\mu \pi_{1} \\
(\lambda+c \mu) \pi_{c}=\lambda \pi_{c-1}+c \mu \pi_{c+1} \\
(\lambda+c \mu) \pi_{c+1}=\lambda \pi_{c}+c \mu \pi_{c+2} \\
\cdots \cdots \cdots \cdots
\end{array}
\end{aligned}
$$

By adding each two consecutive equations:

$$
\begin{array}{ll}
\lambda \pi_{0}=\mu \pi_{1} & \pi_{1}=\frac{\lambda}{\mu} \pi_{0} \\
\lambda \pi_{1}=2 \mu \pi_{2} & \pi_{2}=\frac{\lambda}{2 \mu} \pi_{1} \\
\ldots \ldots \ldots . . & \ldots \ldots \ldots . . . \\
\lambda \pi_{c-1}=c \mu \pi_{c} & \text { i.e., } \\
\lambda \pi_{c}=c \mu \pi_{c+1} & \pi_{c}=\frac{\lambda}{c \mu} \pi_{c-1} \\
\lambda \pi_{c+1}=c \mu \pi_{c+2} & \pi_{c+1}=\frac{\lambda}{c \mu} \pi_{c} \\
\ldots \ldots \ldots \ldots & \pi_{c+2}=\frac{\lambda}{c \mu} \pi_{c+1}
\end{array}
$$

Dr. Xing ©
Q-Systems
$\qquad$

## M/M/c (III): Solution to Steady-

State Probabilities

## M/M/c (IV): Performance Measures (I)

Average number of customers in the system:

$$
L=\sum_{n=0}^{\infty} n \pi_{n}
$$

Average response time:

$$
W=L / \lambda
$$

Average waiting time:

$$
W_{q}=W-W_{s}
$$

Average queue length:

$$
L_{q}=\lambda W_{q}
$$

- Therefore:
$\pi_{0}\left[\sum_{n=0}^{c-1} \frac{\alpha^{n}}{n!}+\sum_{n=c}^{\infty} \frac{\alpha^{n}}{c!c^{n-c}}\right]=1 \Rightarrow \pi_{0}=\left[\sum_{n=0}^{c-1} \frac{\alpha^{n}}{n!}+\sum_{n=c}^{\infty} \frac{\alpha^{n}}{c!c^{n-c}}\right]^{-1}$

$$
\therefore \pi_{0}=\left[\sum_{n=0}^{c-1} \frac{\alpha^{n}}{n!}+\frac{\alpha^{\alpha}}{c!(1-\alpha / c)}\right]^{-1}
$$

P[wait] ?
Dr. Xing ©
Q-Systems

## M/M/c(V): Performance Measures (II)

- A customer must queue for service iff there are $c$ or more customers already in the system:

$$
\begin{aligned}
& P[\text { wait }]=P[n \geq c]=\sum_{n=c}^{\infty} \pi_{n}=\pi_{0}\left[\sum_{n=c}^{\infty} \frac{\alpha^{n}}{c!c^{n-c}}\right] \\
& =\pi_{0} \frac{\alpha^{c}}{c!(1-\alpha / c)}=\frac{\pi_{c}}{1-\alpha / c}=\frac{\pi_{c}}{1-\rho}
\end{aligned}
$$

$$
\therefore P[\text { wait }] \frac{\alpha^{c} / c!}{(1-\rho)\left(\sum_{n=0}^{c-1} \frac{\alpha^{n}}{n!}+\frac{\alpha^{c}}{c!(1-\alpha / c)}\right)}=\frac{\alpha^{c} / c!}{(1-\rho) \sum_{n=0}^{c-1} \frac{\alpha^{n}}{n!}+\frac{\alpha^{c}}{c!}}=C[c, \alpha]
$$

-- Erlang's C formula or Erlang's delay formula

- It can be shown that

$$
\begin{aligned}
& W_{q}(t)=P[\text { time in queue } \mathrm{q} \leq \mathrm{t}]=1-C[c, \alpha] e^{-\mu t(c-\alpha)} \\
& \text { and } \\
& W_{q}=E[q]=C[c, \alpha] W_{s} / c(1-\rho), L_{q}=\lambda W_{q}=C[c, \alpha] \rho /(1-\rho) \\
& \text { Dr. Xing © } \quad \text { Q-Systems }
\end{aligned}
$$

## Hands-On Problem

- A storage system consists of two disk drives sharing a common queue with infinite capacity. The I/O requests arrive to the storage system at the rate of 40 requests per second with Poisson pattern. The time to service an I/O request at each disk drive is exponentially distributed with a mean of 45 milliseconds.
- Draw the state transition diagram (show at least the first 4 states).
- What is the probability that each disk drive is busy (i.e., average disk drive utilization)?
- What is the probability that the entire storage system is idle?
- What is the probability that both disk drives are busy and exactly two I/O requests are waiting in the queue.

Dr. Xing ©
Q-Systems

## Agenda

- More Birth-and-Death Queues
- M/M/1/N Queues
- M/M/c Queues
$-\mathrm{M} / \mathrm{M} / \infty$ Queues
- M/M/1/k/k Queues


## M/M/ $\infty$ Queues

- Have an infinite number of servers
- A server is immediately provided for each arriving customer
- State transition diagram:

- State probabilities:

$$
\begin{gathered}
\pi_{n}=\frac{(\lambda / \mu)^{n}}{n!} \pi_{0}=\frac{\alpha^{n}}{n!} \pi_{0}, n=0,1,2,3, \ldots \ldots \\
\sum_{n=0}^{\infty} \pi_{n}=1 \Rightarrow \pi_{0} \sum_{n=0}^{\infty} \frac{\alpha^{n}}{n!}=\pi_{0} e^{\alpha}=1 \\
\therefore \pi_{0}=e^{-\alpha} \\
\pi_{n}=e^{-\alpha} \frac{\alpha^{n}}{n!} \Rightarrow L=\alpha
\end{gathered}
$$

Dr. Xing ©
Q-Systems

## M/M/ $\infty$ Queues: Example

- Calls in a telephone arrive randomly (Poisson) at an exchange at the rate of 140 per hour. If there is a very large number of lines available to handle the calls that last an average of 3 minutes, what is the average number of lines in use?


## Agenda

- More Birth-and-Death Queues
- M/M/1/N Queues
- M/M/c Queues
- M/M/ $\infty$ Queues
- M/M/1/k/k Queues

Dr. Xing ©
Q-Systems

## Kendall Notation (review)

Standard notation for queueing systems:

$$
\mathrm{A} / \mathrm{B} / \mathrm{c} / \mathrm{K} / \mathrm{m} / \mathrm{Z}
$$

- A: arrival process or inter-arrival time distribution
- 'M' = Poisson arrival process
- ' $\mathrm{D}^{\prime}=$ Deterministic (constant) arrival rate
- 'G' = General arrival process
- B: service process or service time dist.
- ' $\mathrm{M}^{\prime}=$ Exponential service time dist.
- 'D' = Deterministic (constant) service time
- 'G' = General service time
- c: number of servers
- K: the capacity of the system (queue+server(s)) (default: $\infty$ )
- m: total job/customer population (default: $\infty$ )
- Z: scheduling discipline (default: FIFO)


## M/M/1/k/k Queues (I)

- Machine repair model

- There are only k customers (devices / machines)
- The machines break at a rate of $\beta$
- The operating time between breakdowns: $\tau$

$$
\operatorname{Pr}[\tau \leq t]=1-e^{-\beta t}
$$

- Mean operating time between breakdowns is $1 / \beta$
- One server (repairman)
- State transition diagram?


## M/M/1/k/k Queues (II)

- State transition diagram


$$
\left\lvert\, \begin{array}{lc}
\text { State } & \text { Rate out=Rate in } \\
0 & k \beta \pi_{0}=\mu \pi_{1} \\
1 & {[(k-1) \beta+\mu] \pi_{1}=k \beta \pi_{0}+\mu \pi_{2}} \\
\cdots & \ldots \ldots \ldots \cdots \\
\mathrm{k}-1 & (\beta+\mu) \pi_{k-1}=2 \beta \pi_{k-2}+\mu \pi_{k} \\
\mathrm{k} & \mu \pi_{k}=\beta \pi_{k-1}
\end{array}\right.
$$

$$
\begin{aligned}
& \pi_{1}=\frac{k \beta}{\mu} \pi_{0} \\
& \pi_{2}=\frac{(k-1) \beta}{\mu} \pi_{1} \\
& \ldots \ldots \ldots . \omega_{k-1}=\frac{2 \beta}{\mu} \pi_{k-2} \\
& \pi_{k-1}=\frac{\beta}{\mu} \pi_{k-1}
\end{aligned}
$$

## M/M/1/k/k Queues (III)

- State probabilities

$$
\begin{aligned}
& \pi_{n}=\frac{(k-n+1) \beta}{\mu} \pi_{n-1}=\frac{(k-n+1)(k-n+2) \ldots k \beta^{n}}{\mu^{n}} \pi_{0} \\
& \quad=\frac{k!}{(k-n)!}\left(\frac{\beta}{\mu}\right)^{n} \pi_{0} \quad \mathrm{n}=0,1,2, \ldots \mathrm{k} \\
& \sum_{n=0}^{k} \pi_{n}=1 \Rightarrow \pi_{0} \sum_{n=0}^{k} \frac{k!}{(k-n)!}\left(\frac{\beta}{\mu}\right)^{n}=1 \\
& \therefore \pi_{0}=\left[\sum_{n=0}^{k} \frac{k!}{(k-n)!}\left(\frac{\beta}{\mu}\right)^{n}\right]^{-1}=B[k, z]
\end{aligned}
$$

called Erlang's B formula, where $z=\mu / \beta$.

- The server (repairman) utilization:

$$
\rho=1-\pi_{0}=1-B[k, z]
$$

Dr. Xing ©
Q-Systems

## M/M/1/k/k Queues (IV):

 Performance Measures- The average rate of customers to the queueing system $\lambda$

$$
\begin{aligned}
& \rho=\alpha=\lambda / \mu=\lambda W_{s} \Rightarrow \\
& \lambda=\rho / W_{s}=(1-B[k, z]) / W_{s}
\end{aligned}
$$

## M/M/1/k/k Queues (V):

Performance Measures

- Average time in the queue $W_{q}$ :
- Each customer (machine) has a complete cycle consisting of
.
Operating time $(\mathrm{o})+$ Queueing time $(\mathrm{q})+$ Service/repair time $(\mathrm{s})$
- Average time to complete the cycle:

$$
E[o]+E[q]+E[s]=1 / \beta+W_{q}+W_{s}
$$

- Overall "arrival" rate from k independent identical customer (average rate at which k machines break down and enter the queueing system) $\lambda$ is:

$$
\begin{aligned}
& \lambda=k /(E[o]+E[q]+E[s])=k /\left(1 / \beta+W_{q}+W_{s}\right) \\
& \therefore W_{q}=k / \lambda-1 / \beta-W_{s}=k / \lambda-1 / \beta-1 / \mu
\end{aligned}
$$

## M/M/1/k/k Queues (VI):

 Performance Measures- Average queue length $L_{q}$ :

$$
\begin{aligned}
L_{q} & =\lambda W_{q} \\
& =\lambda(k / \lambda-1 / \beta-1 / \mu) \\
& =k-\lambda / \beta-\lambda / \mu
\end{aligned}
$$

- Average time in the system $W$ :

$$
\begin{aligned}
W & =W_{q}+W_{s} \\
& =k / \lambda-1 / \beta-W_{s}+W_{s} \\
& =k / \lambda-1 / \beta
\end{aligned}
$$

- Average number of customers in the system $L$ :

$$
\begin{aligned}
L & =\lambda W \\
& =\lambda(k / \lambda-1 / \beta) \\
& =k-\lambda / \beta
\end{aligned}
$$

M/M/1/k/k Queues (VI): Example

- Users of a time-sharing system have exponentially distributed think times with an average value of 20 seconds.
- Service time are exponentially distributed with a mean of 1.5 seconds.
- Q: how many terminals can the system support if we want the average stretch factor is no more than 5 ?

Example Solution (Hint)

- Use M/M/1/k/k model

$\beta=$ arrival rate per user $=1 / 20$ per second
$\mu=$ service rate $=1 / 1.5=2 / 3$ per second
Find maximum $k$ satisfying $\mathrm{W} / \mathrm{Ws}<=5$ ?

More in the textbook...

- M/M/c/c Queues
- Also called the $\mathrm{M} / \mathrm{M} / \mathrm{c}$ loss systems because customers who arrive when all the c servers are busy are not allowed to wait for service and thus are lost to the system
- Ch 5.2.4
- $\mathrm{M} / \mathrm{M} / \mathrm{c} / \mathrm{k} / \mathrm{k}$ Queues
- Machine repair, multiple repairmen
- Ch 5.2.7

Dr. Xing ©
Q-Systems

## Next Topics

- Embedded Markov Chain

Queueing Systems

## Things to Do

- Homework
- Keep working on the project

