

EE/CpE 345

Modeling and Simulation

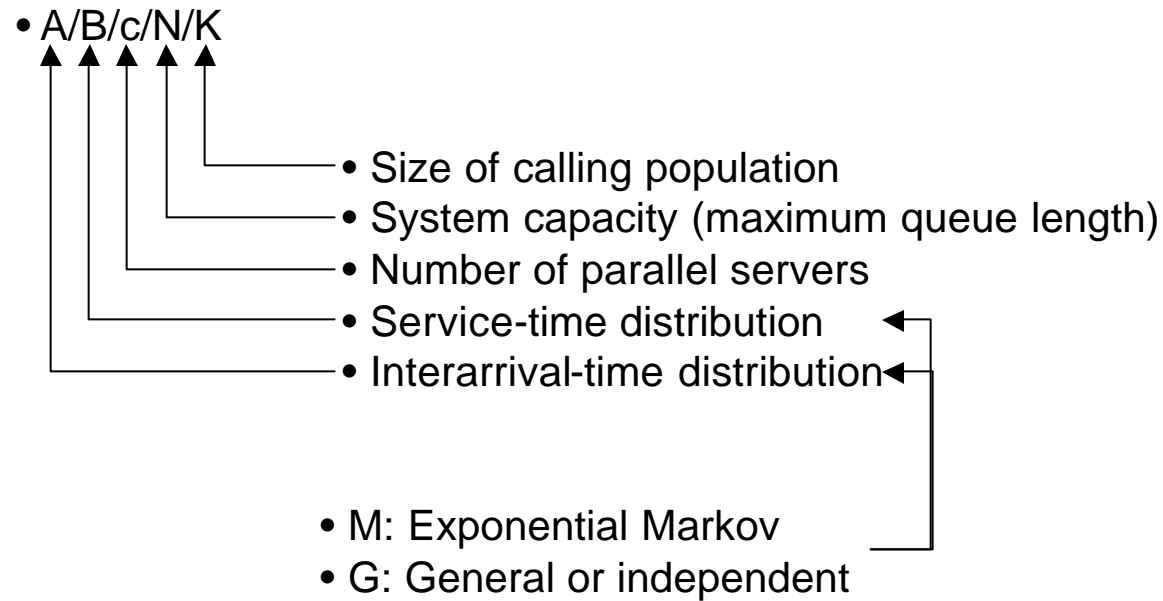
Fall 2003

Class 6

This week

- Steady-state behavior of Infinite-population Markovian Models
- Steady-state behavior of Finite-population Markovian Models
- Networks of queues

Queuing Notation (review)



Steady-State Behavior of Infinite-Population Markovian Models

- $M/M/c/N/8$, $M/G/c/N/8$
- Assumptions:
 - Infinite population
 - Arrivals: Poisson process with rate λ arrivals per time unit
 - Interarrival time: Exponentially distributed with mean $1/\lambda$
 - Service time: Exponentially distributed or arbitrary
 - Queue discipline: FIFO
- Markovian model: Exponentially distributed arrival process
- Static equilibrium (steady-state):
 - $$P(L(t) = n) = P_n(t) = P_n$$
 - System state (number of customers in systems) is independent of time
 - If system is stable, it is generally either
 - approaching static equilibrium
 - staying in static equilibrium

Steady-State Behavior of Infinite-Population Markovian Models

- Average number of customers in system, L :

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

– time average number of customers = average over number of customers

- Given L , apply Little's equation to find other parameters: $L = \mathbf{l} w$

– average customer time in system

$$w = \frac{L}{\mathbf{l}}$$

– average customer time in queue
(time in system-service time)

$$w_Q = w - \frac{1}{\mathbf{m}}$$

– average number of customers in queue

$$L_Q = \mathbf{l} w_Q$$

- For **infinite** calling population, for system to be stable:

$$\mathbf{r} = \frac{\mathbf{l}}{c\mathbf{m}} < 1$$

Single Server Queues with Poisson Arrivals and Unlimited Capacity: $M/G/1$

- Assume server has service times with mean $1/m$ and variance s^2
- If $r = \lambda/m < 1$, system is stable and has steady-state characteristics:

r	$\frac{\lambda}{m}$
L	$\lambda + \frac{\lambda^2 \left(\frac{1}{m^2} + s^2 \right)}{2(1-r)} = \lambda + \frac{\lambda^2 (1 + s^2 m^2)}{2(1-r)}$
W	$\frac{1}{m} + \frac{\lambda \left(\frac{1}{m^2} + s^2 \right)}{2(1-r)}$
W_Q	$\frac{\lambda \left(\frac{1}{m^2} + s^2 \right)}{2(1-r)}$
L_Q	$\frac{\lambda^2 \left(\frac{1}{m^2} + s^2 \right)}{2(1-r)} = \frac{\lambda^2 (1 + s^2 m^2)}{2(1-r)}$
P_0	$1 - r$

Example 6.9: Able and Baker Again

- Able: $1/\bar{m}=24$ minutes, $s^2=400$ minutes² faster service with high variability
- Baker: $1/\bar{m}=25$ minutes, $s^2=4$ minutes² slower service with low variability
- $I=1/30$ per minute
- Which has the shortest average queue length (L_Q)
- Which has highest $P(\text{no delay})$

	Able	Baker
r	.8	.833
L_Q	2.711	2.097
P_0	.2	.167

An M/M/1 Queue

- An M/G/1 queue with exponential service time is an M/M/1 queue
- Mean service time = $1/m$ variance, $s^2=1/m^2$
- Steady state parameters:

L	$\frac{l}{m-l} = \frac{r}{1-r}$
W	$\frac{1}{m-l} = \frac{1}{m(1-r)}$
W_Q	$\frac{l}{m(m-l)} = \frac{r}{m(1-r)}$
L_Q	$\frac{l^2}{m(m-l)} = \frac{r^2}{1-r}$
P_n	$(1-r)r^n$

An $M/M/1$ Queue - Example 6.11

- An $M/M/1$ queue with service rate $m=10$ per hour
- Examine variation of L and w with increasing arrival rate $I=\{5,6,7.2,8.64,9.99,10\}$

I	5.0	6.0	7.2	8.64	9.99	10.0
r	.500	.600	.720	.864	.999	1.0
L	1.00	1.50	2.57	6.35	999	8
w	.20	.25	.36	.73	100	8

Utilization vs. Service Variability

- For any $M/G/1$ queue - to decrease L_Q , queue length:
 - reduce server utilization, r
 - increase service rate, m
 - decrease arrival rate, I
 - increase number of servers, c
 - reduce service time variability, s^2
- For a random variable X , define coefficient of variation, (cv) :

$$(cv)^2 = \frac{V(X)}{[E(X)]^2}$$

The ratio of variance to expectation squared

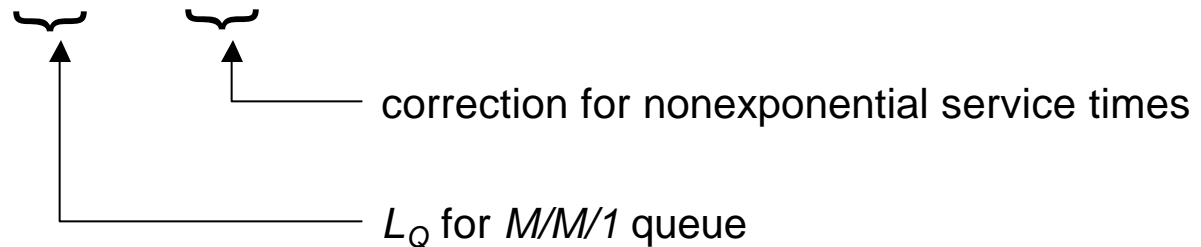
Utilization vs. Service Variability

- For an $M/G/1$ queue, define L_Q in terms of cv :

$$(cv)^2 = \frac{V(X)}{[E(X)]^2} = \frac{s^2}{1/m^2} = s^2 m^2$$

$$L_Q = \frac{r^2 (1 + s^2 m^2)}{2(1 - r)}$$

$$L_Q = \left(\frac{r^2}{(1 - r)} \right) \left(\frac{1 + (cv)^2}{2} \right)$$



Utilization vs. Service Variability

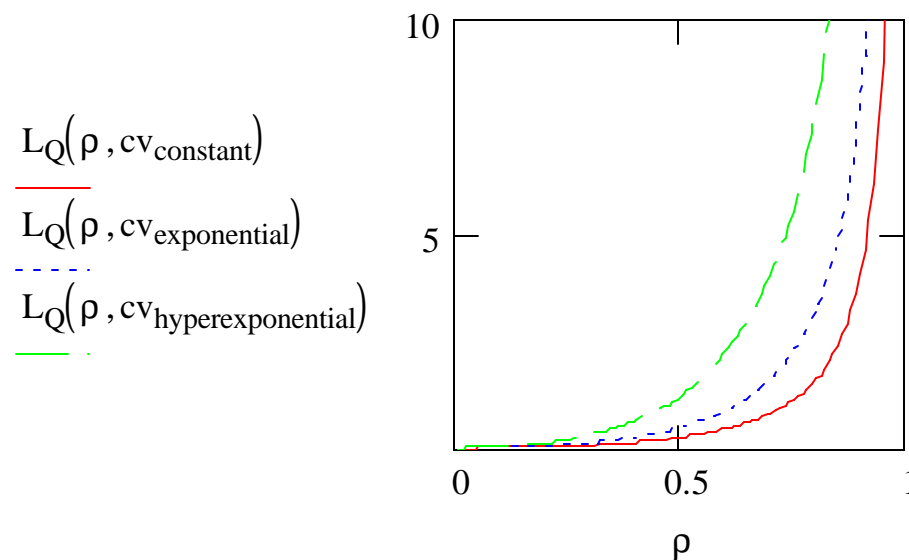
$$L_Q(\rho, cv) := \left(\frac{\rho^2}{1-\rho} \right) \left[\frac{1+(cv)^2}{2} \right]$$

$cv_{\text{constant}} := 0$

$cv_{\text{exponential}} := 1$

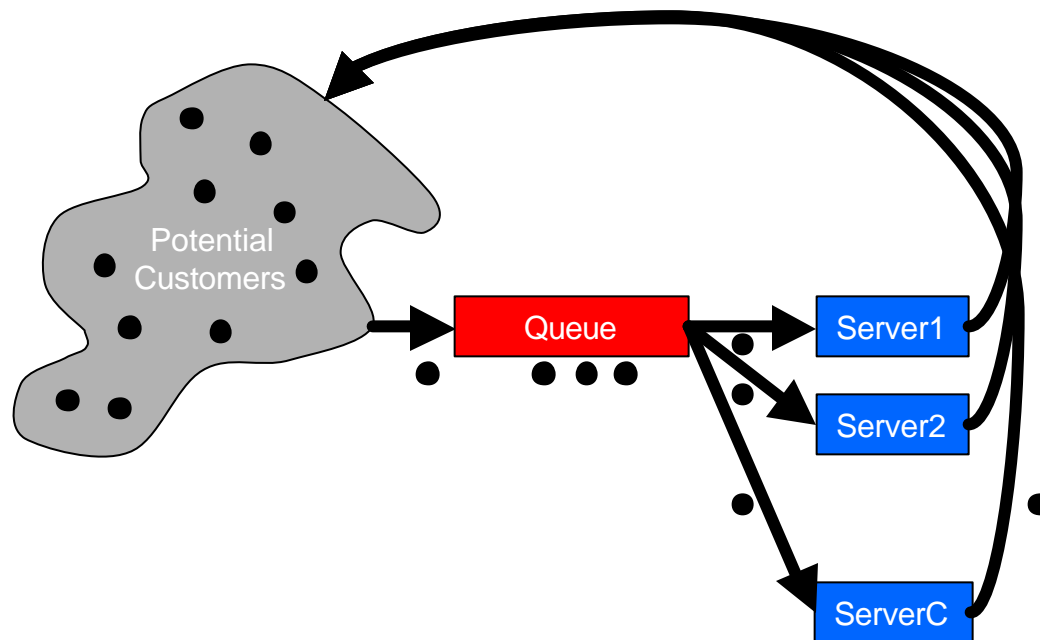
$\rho := 0, .01 \dots .999$

$cv_{\text{hyperexponential}} := 2$



Multiserver Queue: $M/M/c$

- Assume c servers operating in parallel
 - Each server has independent, identically exponential service distribution
- Poisson arrival process
- Customers form a single queue and wait for first available server
- Static equilibrium requires $\lambda < mc$

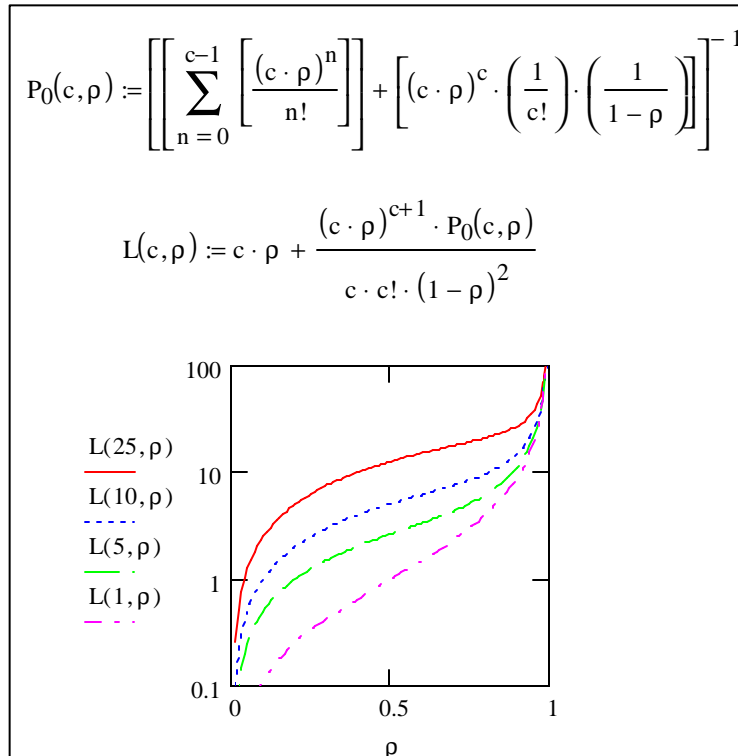


Multiserver Queue: $M/M/c$

- Steady-state parameters:

r	$\frac{l}{cm}$
P_0	$\left\{ \left[\sum_{n=0}^{c-1} \frac{(l/m)^n}{n!} \right] + \left[\left(\frac{l}{m} \right)^c \left(\frac{1}{c!} \right) \left(\frac{cm}{cm-1} \right) \right] \right\}^{-1} = \left\{ \left[\sum_{n=0}^{c-1} \frac{(cr)^n}{n!} \right] + \left[\left(\frac{l}{m} \right)^c \left(\frac{1}{c!} \right) \frac{1}{1-r} \right] \right\}^{-1}$
$P(L(\infty) \geq c)$	$\frac{(l/m)^c P_0}{c!(1-l/cm)} = \frac{(cr)^c P_0}{c!(1-r)}$
L	$cr + \frac{(cr)^{c+1} P_0}{c(c!)(1-r)^2} = cr + \frac{rP(L(\infty) \geq c)}{(1-r)}$
W	$\frac{L}{l}$
W_Q	$w - \frac{1}{m}$
L_Q	$L_{w_Q} = \frac{(cr)^{c+1} P_0}{c(c!)(1-r)^2} = \frac{rP(L(\infty) \geq c)}{(1-r)}$
$L - L_Q$	$\frac{l}{m} = cr$

Multiserver Queue: $M/M/c$



An Approximation for $M/G/c$ Queue

- L_Q and w_Q can be found for $M/G/1$ by applying a correction factor to the $M/M/1$ queue parameters
- There is no exact formula to do the same for $M/G/c$ queues with finite c , but the same correction factor is approximately correct, especially for (cv) near 1

Multiserver Queue with Infinite Number of Servers: $M/G/\infty$

- Number of servers can be considered infinite if:
 - Self-service (as many servers as needed per customer)
 - When there is ample system capacity, so servers are rarely all busy
 - When you are trying to determine the number of servers needed to insure customers are rarely delayed.

Multiserver Queue with Infinite Number of Servers: $M/G/\infty/\infty$

- Steady-state parameters

P_0	$e^{-1/m}$
w	$1/m$
w_Q	0
L	$1/m$
L_Q	0
P_0	$(e^{-1/m}(1/m)^n)/n!$

Example 6.16: M/G/8 queue

- An ISP is planning number of ports needed
 - Customer arrival rate, $\lambda=500/\text{hour}$
 - Hold time (service time), $1/\mu=3$ hours
 - How many ports are needed to serve customers 95% of the time?
- Assume an infinite user population
- $L=\lambda/\mu=1500$
- find a minimum c (>1500) such that:

$$P(L(\infty) \leq c) = \sum_{n=0}^c P_n = \sum_{n=0}^c \frac{e^{-1500} (1500)^n}{n!} \geq .95$$

- $c=1564$

Multiserver Queue with Poisson Arrivals and Limited Capacity: $M/G/c/N/8$

- If an arrival occurs and the system is full, customer is turned away
- The effective arrival rate $I_e = I(1 - P_N)$ - actual arrival rate reduced by probability that system capacity is reached.
- Practical consideration: server utilization is reduced if the system restricts the number of customers waiting. (see example 6.17)

Steady-State Behavior of Finite-Population Models: $M/M/c/K/K$

- System state influences arrival rate:
 - For infinite customer population, arrival rate is independent of system state (infinite supply of customers remains, even if a large number are in system)
 - For finite customer population, one or more customers in system means that there are fewer customers who can be generating service requests.

Network of Queues

- Fundamental principles (infinite customer population, infinite capacity):
 - Conservation of customers:
 - single queues: average departure rate = average arrival rate
 - tandem queues: arrival rate at $i+1^{\text{st}}$ queue is arrival rate at i^{th} queue times fraction routed to $i+1^{\text{st}}$ queue
 - “Kierkov’s current law”: arrival rate into queue is sum of arrival rates from all sources
 - Load splitting across servers:
 - parallel server utilization is inversely proportional to server capacity
 - Poisson arrivals into tandem queues with exponential service times:
 - each queue behaves like a $M/M/c$ queue

Homework

Prepare for midterm: Next week 10/20 during class