

EE/PEP 345

Modeling and Simulation

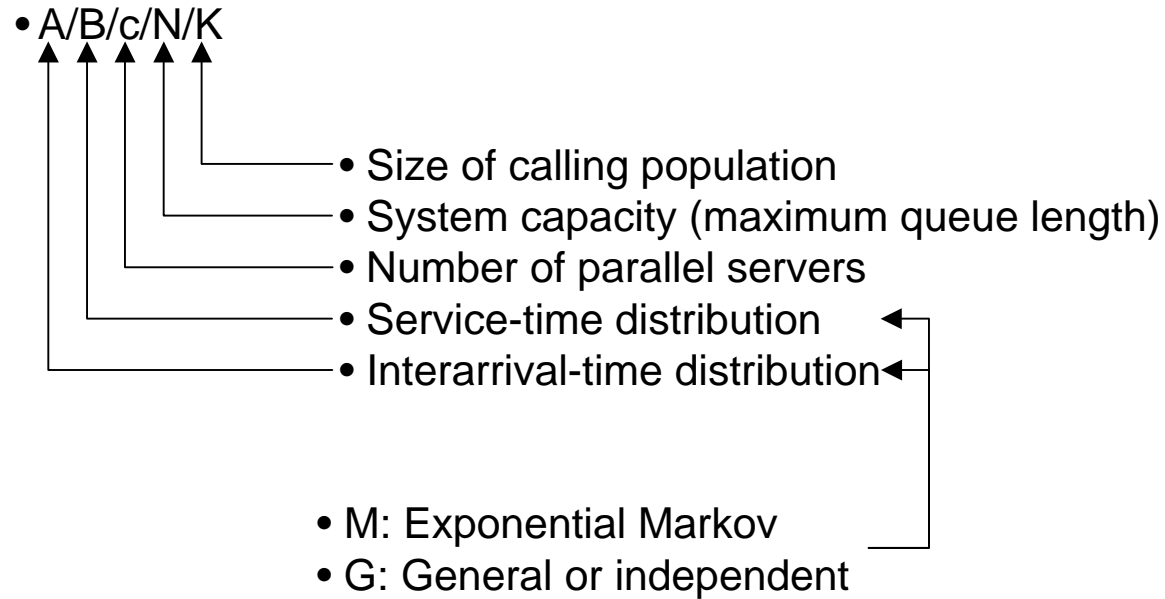
Spring 2004

Class 6

This week

- Service variability
- Multiserver queues
- Finite capacity queues
- Steady-state behavior of Finite-population Markovian Models
- Networks of queues
- Review for Midterm – next week.

Queuing Notation (review)



Utilization vs. Service Variability

- For any $M/G/1$ queue - to decrease L_Q , queue length:
 - reduce server utilization, ρ
 - increase service rate, μ
 - decrease arrival rate, λ
 - increase number of servers, c
 - reduce service time variability, σ^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 :

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

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

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



correction for nonexponential service times

L_Q for $M/M/1$ queue

Utilization vs. Service Variability

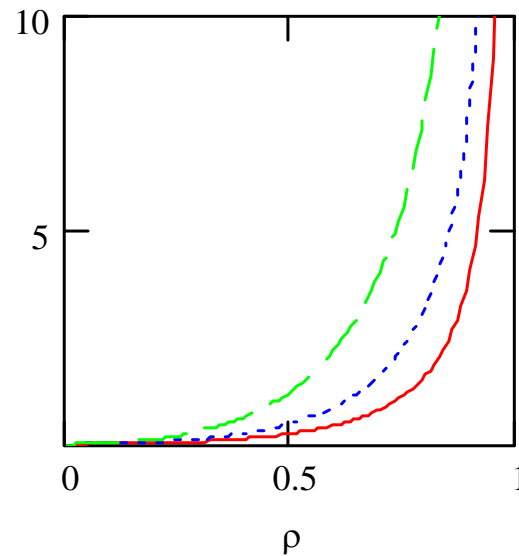
$$L_Q(\rho, cv) := \left(\frac{\rho^2}{1-\rho} \right) \left[\frac{1+(cv)^2}{2} \right] \quad cv_{\text{constant}} := 0$$

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

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

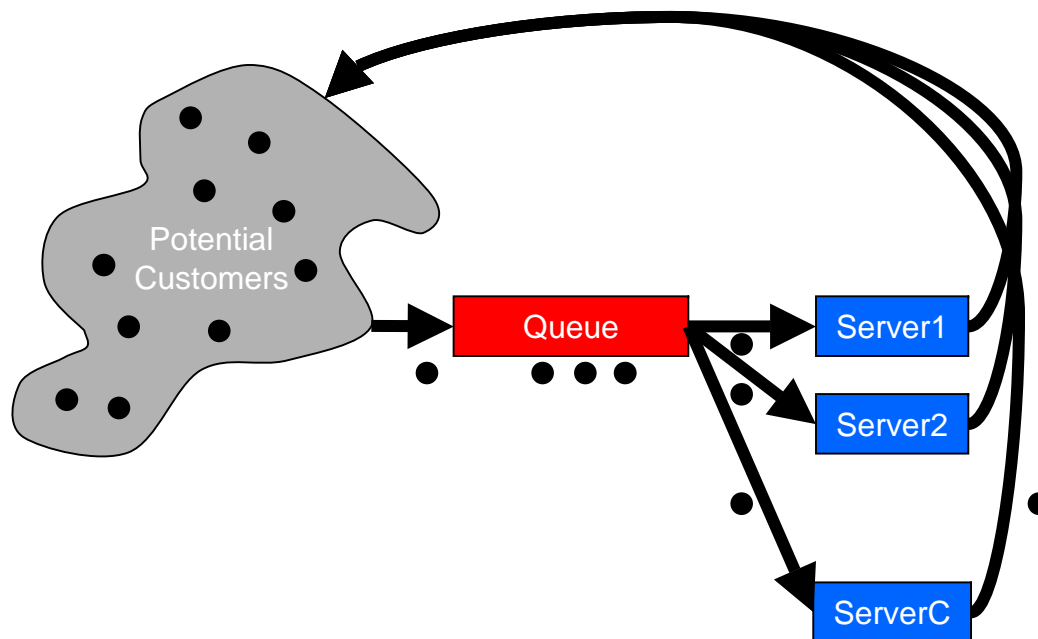
$$cv_{\text{hyperexponential}} := 2$$

$L_Q(\rho, cv_{\text{constant}})$
—
 $L_Q(\rho, cv_{\text{exponential}})$
- - -
 $L_Q(\rho, cv_{\text{hyperexponential}})$
- . -



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 $\lambda/\mu < c$

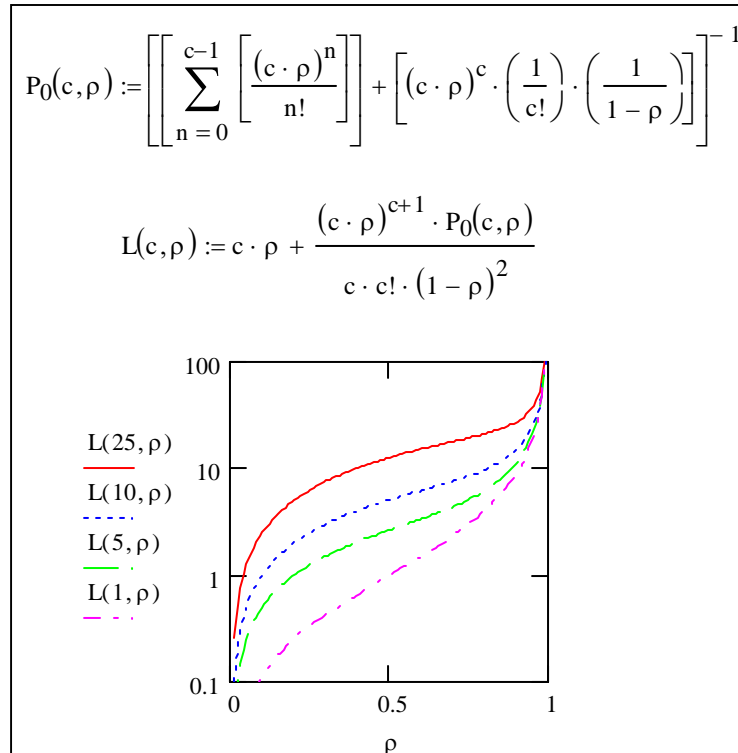


Multiserver Queue: $M/M/c$

- Steady-state parameters:

ρ	$\frac{\lambda}{c\mu}$
P_0	$\left\{ \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} \right] + \left[\left(\frac{\lambda}{\mu} \right)^c \left(\frac{1}{c!} \right) \left(\frac{c\mu}{c\mu - \lambda} \right) \right] \right\}^{-1} = \left\{ \left[\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} \right] + \left[\left(\frac{\lambda}{\mu} \right)^c \left(\frac{1}{c!} \right) \frac{1}{1-\rho} \right] \right\}^{-1}$
$P(L(\infty) \geq c)$	$\frac{(\lambda/\mu)^c P_0}{c!(1-\lambda/c\mu)} = \frac{(c\rho)^c P_0}{c!(1-\rho)}$
L	$c\rho + \frac{(c\rho)^{c+1} P_0}{c(c!)(1-\rho)^2} = c\rho + \frac{\rho P(L(\infty) \geq c)}{(1-\rho)}$
W	$\frac{L}{\lambda}$
W_Q	$w - \frac{1}{\mu}$
L_Q	$\lambda w_Q = \frac{(c\rho)^{c+1} P_0}{c(c!)(1-\rho)^2} = \frac{\rho P(L(\infty) \geq c)}{(1-\rho)}$
$L - L_Q$	$\frac{\lambda}{\mu} = c\rho$

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: $M/G/\infty/\infty$

- 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	$e^{-\lambda/\mu}$
w	$1/\mu$
w_Q	0
L	λ/μ
L_Q	0
P_n	$(e^{-\lambda/\mu}(\lambda/\mu)^n)/n!$

Example 6.16: M/G/∞ 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: $M/G/c/N/\infty$

- If an arrival occurs and the system is full, customer is turned away
- The effective arrival rate $\lambda_e = \lambda(1 - P_N)$ - 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 during class