

CPE 345: Modeling and Simulation

Lecture 6

Today's topics

- More definitions for the queueing system
- Steady-state behavior of infinite population Markovian models
- Steady-state behavior of finite population Markovian models
- Network of queues

- Midterm
 - March 14 – open books and notes (printed, not laptop)
 - Material covered, including today's lecture

The conservation equation (Little equation)

- Little's equation: $L = \lambda w$ as $T \rightarrow \infty, N \rightarrow \infty$

The average number of customers in the system at an arbitrary point in time is equal to the average number of arrivals per unit time, times the average time spent in the system

- Holds true for almost all queueing systems or subsystems, regardless of the number of servers, arrival/departure process, queueing discipline, etc.
- Very useful for analyzing any type of queueing system or for a quick check of your results from a simulation

Server Utilization

- Def: percentage of time that server is busy (serving customers)
- Notation: ρ
- For a general queue: G/G/1 (λ = arrival rate, μ = service rate)

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

- Recall stability condition: arrival rate < service rate

$$\lambda < \mu \Rightarrow \rho < 1$$

Server Utilization: G/G/c

- Server utilization

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

- Stability condition: arrival rate < service rate

$$\lambda < c\mu \Rightarrow \rho < 1$$

Steady-state behavior of infinite-population Markovian models

- Infinite population – the arrival rate not influenced by the users already in the system
- Arrivals: Poisson with arrival rate λ (arrivals per unit time)
 - Exponential distribution for the inter-arrival times (mean = $1/\lambda$)
→ Markovian models
- Service time: exponentially distributed or following a general distribution
- Queueing discipline: FIFO
- Steady-state: statistical equilibrium → system state (number of customers in the system) is independent of time

$$P(L(t) = n) = P_n(t) = P_n$$

- If a system is stable
 - Is approaching stat. equilibrium given any starting state
 - Remains in stat. equilibrium, once the equilibrium is reached

Fundamental metrics

- Average number of customers in the system

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

- Average customer time in the system

$$w = \frac{L}{\lambda} \quad (\text{we have used Little's equation})$$

- Average customer time in queue

$$w_Q = w - \frac{1}{\mu} \quad (\text{time in service} - \text{service time})$$

- Average number of customers in queue

$$L_Q = \lambda w_Q \quad (\text{again Little's equation})$$

General remarks

- All previous results valid for any M/G/c/N/∞ queue
- Specific formulas are derived for different service distributions, by explicitly determining the expression for P_n (all the previous results depend on P_n)
- Also, for infinite calling population, we have to make sure that the system is stable:

$$\rho = \frac{\lambda}{c\mu} < 1$$

Steady-state formulas for M/G/1

Mean service time $1/\mu$, service variance σ^2

ρ	λ / μ
L	$\rho + \frac{\lambda^2(1/\mu^2 + \sigma^2)}{2(1-\rho)} = \rho + \frac{\rho^2(1 + \sigma^2\mu^2)}{2(1-\rho)}$
W	$\frac{1}{\mu} + \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1-\rho)}$
W_Q	$\frac{\lambda(1/\mu^2 + \sigma^2)}{2(1-\rho)}$
L_Q	$\frac{\lambda^2(1/\mu^2 + \sigma^2)}{2(1-\rho)} = \frac{\rho^2(1 + \sigma^2\mu^2)}{2(1-\rho)}$
P_0	$(1-\rho)$

Able-Baker example

- Able and Baker competing for a job
 - Able – faster on average but with a larger standard deviation for service (σ)

$$1 / \mu_A = 24 \text{ min.}; \sigma = 20 \text{ min}$$

- Baker – slower on average, but more consistent

$$1 / \mu_B = 25 \text{ min.}; \sigma = 2 \text{ min}$$

- Arrivals: Poisson with $\lambda=2$ per hour = $1/30$ per minute
- Hiring criteria: queue length
- Able or Baker?

$$L_Q^A = \frac{\lambda^2 (1/\mu^2 + \sigma^2)}{2(1-\rho)} = \frac{1/30^2 (24^2 + 20^2)}{2(1-24/30)} = 2.71$$

$$L_Q^B = \frac{\lambda^2 (1/\mu^2 + \sigma^2)}{2(1-\rho)} = \frac{1/30^2 (25^2 + 2^2)}{2(1-25/30)} = 2.097$$

 Hire Baker

Able and Baker - cont

- Other criteria?
 - Percentage of customers that find the server idle and have zero waiting time

$$P_0^A = (1 - \rho_A) = 1 - \frac{24}{30} = 0.2$$

$$P_0^B = (1 - \rho_B) = 1 - \frac{25}{30} = 0.17$$

M/M/1 queue

- Service times – also exponential, with mean $1/\mu$
 - From exponential distribution: $\sigma^2 = 1/\mu^2$

L	$\frac{\lambda}{\mu - \lambda} = \frac{\rho}{1 - \rho}$
W	$\frac{1}{\mu - \lambda} = \frac{1}{\mu(1 - \rho)}$
W_Q	$\frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu(1 - \rho)}$
L_Q	$\frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{\mu(1 - \rho)}$
P_n	$\left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n = (1 - \rho)\rho^n$

M/M/1 Example

- Effect of ρ on system performance (delay, average number of users in the system)
- M/M/1 with $\mu = 10$ customers/hour, λ varies

λ	5.0	6.0	7.2	8.64	10.00
ρ	0.5	0.6	0.72	0.864	1.0
L	1.00	1.50	2.57	6.35	∞
w	0.20	0.25	0.36	0.73	∞

Utilization and service variability

- For M/G/1 queues, queue lengths can be reduced by
 - Decreasing server utilization ρ
 - Reduce arrival rate
 - Increase service rate
 - Increasing the number of servers
 - Decreasing service time variability σ^2
 - We can define the coefficient of variation for a r.v. X as

$$L_{M/G/1} = \rho + \frac{\rho^2(1 + \sigma^2\mu^2)}{2(1 - \rho)}$$

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

Length of queue and coefficient of variation

$$(cv)^2 = \frac{\sigma^2}{1/\mu^2} = \sigma^2 \mu^2$$

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

correction term

L_Q for M/M/1

- As $\rho \uparrow$, $L_Q \uparrow$
- For fixed ρ , as $(cv)^2 \uparrow$, $L_Q \uparrow$
- $(cv)^2$ is characteristic for a given service distribution

Multiserver Queue (M/M/c)

- c servers operating in parallel
 - Each has an iid service distribution (identical and independent distributed service), mean service for one server $1/\mu$
- Poisson arrivals (mean λ)
- Define offered load $\rightarrow \lambda/\mu$.
- When number of customers in system $n < c$, customer goes directly in service at the next available server, $n > c$, customer is queued
 - *Convention: Service selected in the order of index for servers*
- Statistical equilibrium: $\lambda/\mu < c$

M/M/c parameters

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

Example M/M/c

- Poisson arrivals: 2 mechanics/min
- Exponentially distributed service with mean $1/\mu = 40$ sec.

$$\mu = 60 / 40 = 3 / 2 = 1.5 \text{ service per minute}$$

- Offered load $\lambda / \mu = 2 / 1.5 > 1 \rightarrow$ more than one server needed
 - Steady state requirement $\lambda / \mu < c \rightarrow c > 1.33 \rightarrow$ at least $c=2$
- For $c=2$
 - Probability that a new arrival finds both servers idle

$$P_0 = \left\{ \sum_{n=0}^1 \frac{(4/3)^n}{n!} + \left(\frac{4}{3}\right)^2 \left(\frac{1}{2!}\right) \left[\frac{2(3/2)}{2(3/2) - 2} \right] \right\}^{-1} = \frac{1}{5} = 0.2$$

- Probability that a new arrival finds both servers busy

$$P(L(\infty) \geq 2) = \frac{(4/3)^2}{2!(1 - 2/3)} \left(\frac{1}{5}\right) = \frac{8}{15} = 0.533$$

Example M/M/c – cont.

- Time-average length of the waiting line

$$L_Q = \frac{(2/3)(8/15)}{1 - 2/3} = 1.07$$

- Time-average number in the system:

$$L = L_Q + \frac{\lambda}{\mu} = 2.4$$

- Average time a mechanic spends in service

$$w = \frac{L}{\lambda} = 1.2 \text{ minutes}$$

- Average time a mechanic spends waiting in queue

$$w_Q = w - \frac{1}{\mu} = 0.533 \text{ minutes}$$

An approximation for M/G/c queue

- For M/G/1 – formulas for L_Q and w_Q can be obtained from the corresponding M/M/1 formulas, by multiplying with the correction factor
$$\frac{(1 + (cv)^2)}{2}$$
- For M/G/c – no exact formulas can be determined
 - Approximate formulas for L_Q and w_Q can be obtained by applying the same correction factor for corresponding the formulas for the M/M/c queue

When the number of servers is infinite:

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

- Each customer is its own server (self-service)
- When service capacity far exceeds server demand (ample-server system)
- When you want to determine the number of servers needed, such that customers are rarely delayed

Steady-state parameters for M/G/∞ queue

P_0	$e^{-\lambda/\mu}$
w	$1/\mu$
w_Q	0
L	λ/μ
L_Q	0
P_n	$\frac{e^{-\lambda/\mu}(\lambda/\mu)^n}{n!}, n = 0,1,2,\dots$

Example M/G/ ∞ / ∞

- Example 6.16

- New online computer inf. service -> plan capacity in terms of how many users can be logged in simultaneously
- Users arrive with Poisson distribution, with $\lambda = 500$ per hour, and stay logged for $1/\mu = 180$ min (3 h)
- For planning purposes, assume that the number of simultaneously connected users is ∞ :

$$L = \lambda / \mu = 500 \cdot 3 = 1500 \Rightarrow \text{more than 1500 servers are req.}$$

- Insure adequate capacity 95% of the time:

$$P(L(\infty) \leq c) = \sum_{n=0}^c P_n = \frac{e^{-1500} (1500)^n}{n!} \geq 0.95 \Rightarrow c = 1564$$

Multi-server queue with Poisson arrivals and limited capacity: M/G/c/N/ ∞

- If an arrival occurs when system is full, that arrival is turned away and does not enter the system
- Effective arrival rate
$$\lambda_e = \lambda(1 - P_N)$$
- P_N = probability that system is full
- If Little equations are used to compute waiting times, effective arrival rate should be considered

Steady state parameters for the M/M/c/N queue:

$$a = \lambda/\mu, \rho = \lambda/c\mu$$

P_0	$\left[1 + \sum_{n=1}^c \frac{a^n}{n!} + \frac{a^c}{c!} \sum_{n=c+1}^N \rho^{n-c} \right]^{-1}$
P_N	$\frac{a^N}{c!c^{N-c}} P_0$
L_Q	$\frac{P_0 a^c \rho}{c!(1-\rho)^2} \left[1 - \rho^{N-c} - (N-c)\rho^{N-c}(1-\rho) \right]$
λ_e	$\lambda(1 - P_N)$
w_Q	L_Q / λ_e
w	$w_Q + 1/\mu$
L	$\lambda_e w$

Steady-state behavior of finite population models

- The system state influences arrival rate
 - Example: small group of machines that break down from time to time and require repair
 - **If all machines are broken, no new arrival for service is possible**

Networks of queues

- Some fundamental principles for infinite calling population and no limit on system capacity
 - Departure rate out of a queue is the same as the arrival rate into the queue, over the long run
 - If customers arrive to queue i with λ_i , and a fraction $0 \leq p_{ij} \leq 1$ is routed to queue j upon departure, then the arrival rate from queue i to queue j is $\lambda_i p_{ij}$, over the long run.
 - The overall arrival rate into a queue is the sum of the arrival rate from all the sources.
 - If queue j has c_j parallel servers, each working at rate μ_j , then, the long run utilization of each server is
$$\rho_j = \frac{\lambda_j}{c_j \mu_j}$$
 - Each queue behaves like an M/M/ c_j queue.

Homework

- Problem 11, chapter 5
- Problem 1, page 247 (chapter 6)
- Problem 5, page 248 (chapter 6)