ECE560: Computer Systems Performance Evaluation

Lecture 14 -

Embedded Markov-Chain Queueing Systems & Priority Queueing Systems

Instructor: Dr. Liudong Xing

Administration Issues (3/27/2024)

- Homework #5 assigned
  - Due **April 1, Monday**
- Project final report
  - Due: **April 19, Friday**
  - Refer to Report Guidelines

Review (L#11, 13)

- Birth-and-death queueing systems
  - Fit the birth-and-death process
    - A customer arrival is a birth
    - A customer departure after completing service is a death
    - Both arrival process and service times are memoryless!
  - M/M/ Queues
    - M/M/1 Queues
    - M/M/1/N Queues
    - M/M/c Queues
    - M/M/ ∞ Queues
    - M/M/1/k/k Queues
  - Solution
    - Balance equations
    - Little's Law

3

Agenda

- Embedded Markov Chain Queueing Systems
  - M/G/1,
  - M/D/1,
  - GI/M/1
- Priority Queueing Systems

4

Embedded Markov Chain Queueing Systems

- More general arrival process or service times are allowed
  - **M/G/1,**
  - M/D/1,
  - GI/M/1
- Solution (**see extra notes posted for detailed derivations**):
  - Constructing an embedded Markov chain
  - And applying *z*-transform and Laplace-Stieltjes transform methods

5

M/G/1 Queueing Systems

- Assume
  - Poisson arrival process with rate λ
  - General service time distribution with
    - different customers have independent service times
    - E[s] and E[s^2] exist (in order to calculate L,W)

6

Steady-State State Probability

- Steady-state probability distribution:

$$\begin{cases} \pi_0 = 1 - \rho \\ \pi_i = \pi_0 k_i + \sum_{j=1}^{i+1} \pi_j k_{i-j+1} \end{cases}$$

where k_n = Pr[n customers arrive during one service interval]

$$k_n = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} dW_s(t), \quad n = 0,1,2,...$$

$W_s(t)$: cdf of service time s

7

Other state probabilities

- π_i : steady state probability that a departing customer leaves i customers behind – π_i = $Pr[X=i]$
- p_i : probability that there are i customers in the system at arbitrary times – $p_i = Pr[N=i]$
- r_i : probability that an arriving customer finds i customers already in the system

- It can be shown (by Klienrock) that for M/G/1 systems:

$$\pi_i = p_i = r_i$$

8

L for M/G/1 Systems

$$L = \rho + \frac{\rho^2(1+C_s^2)}{2(1-\rho)},$$

where $C_s = \frac{\sqrt{Var[s]}}{E[s]}$ is C.O.V.

C.O.V: Coefficient of Variation
-- the Pollaczek-Khintchine formula

Note: for the exponential service time distribution: C_s=1, then

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

which is the same as in M/M/1!

9

Find W, L_q, W_q

- W (average response time):
  By Little's Law:

$$W = L/\lambda = \frac{\rho + \frac{\rho^2(1+C_s^2)}{2(1-\rho)}}{\lambda}$$

- L_q (average queue length):

$$\begin{aligned} L_q &= L - (1 * P[\text{Server is not empty}] \\ &= L - (1 - P[0 \text{ customer in the system}]) \\ &= L - (1 - \pi_0) = L - (1 - (1 - \rho)) \\ &= L - \rho = \frac{\rho^2(1+C_s^2)}{2(1-\rho)} \end{aligned}$$

- W_q (average waiting time):

$$W_q = L_q/\lambda = \frac{\rho^2(1+C_s^2)}{2(1-\rho)\lambda} = \frac{\rho W_s(1+C_s^2)}{2(1-\rho)}$$

10

Agenda

- Embedded Markov Chain Queueing Systems
  - M/G/1,
  - **M/D/1,**
  - GI/M/1
- Priority Queueing Systems

11

M/D/1 Queueing Systems

- Assume
  - Poisson arrival process with rate λ
  - Deterministic service rate μ
    - Constant service time s = W_s = 1/ μ
- A special case of M/G/1 with Var[s] = 0, E[s] = W_s → C.O.V.

$$C_s = \frac{\sqrt{Var[s]}}{E[s]} = 0$$

12

M/D/1 Queueing Systems (cont'd)

- Performance measures

$$L = \rho + \frac{\rho^2(1+C_s^2)}{2(1-\rho)} = \rho + \frac{\rho^2}{2(1-\rho)} = \frac{\rho(2-\rho)}{2(1-\rho)}$$

$$W = L/\lambda = \frac{\rho(2-\rho)}{2(1-\rho)\lambda} = \frac{W_s(2-\rho)}{2(1-\rho)}$$

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

$$W_q = L_q/\lambda = \frac{\rho^2}{2(1-\rho)\lambda} = \frac{\rho W_s}{2(1-\rho)}$$

13

Agenda

- Embedded Markov Chain Queueing Systems
  - M/G/1,
  - M/D/1,
  - **GI/M/1**
- Priority Queueing Systems

14

GI/M/1 Queueing Systems

- Assume
  - Renewal arrival process
    - The inter-arrival times are i.i.d. *r.v.*s
  - Exponential service time with a mean of $1/\mu$
    - μ is the average service rate

Renewal Processes (Chapter 4.5)

- A Poisson process can be characterized as a counting process for which the inter-arrival times (times between successive events) are *i.i.d*, exponential *r.v.*s
- A renewal process is a generalization of the Poisson process
  - A counting process for which the inter-arrival times (times between successive events) are *i.i.d r.v.*s

Steady-State Probabilities

- π_n: the steady-state probability that **an arriving customer** finds *n* customers in the system, for n = 0, 1, 2, ...
- Laplace-Stieltjes Transform (LST) of a random variable *X*:

$$A^*(\theta) = L_X[\theta] = E[e^{-\theta X}] =$$

$$\begin{cases} \sum_{x_i} e^{-\theta x_i} p(x_i) & \text{if } X \text{ is discrete} \\ \int_0^\infty e^{-\theta x} f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

π_n for GI/M/1

- Wolff showed that
  - π_0 is the unique solution of equation

$$1 - \pi_0 = A^*[\mu\pi_0]$$

such that $0 < \pi_n < 1$, where A*[θ] is the LST of the inter-arrival time τ

  - General expression of π_n in terms of π_0:

$$\pi_n = \pi_0(1-\pi_0)^n, n = 0,1,2,...$$

A geometric distribution (L#6) with q=1- π_0, p= π_0, thus,

$$E[X] = \frac{q}{p} = \frac{1-\pi_0}{\pi_0}, Var[X] = \frac{q}{p^2} = \frac{1-\pi_0}{\pi_0^2}$$

Example

Consider the **M/M/1** queueing system

- First, find **A*[θ]:**
  - the inter-arrival time τ is exponentially distributed with rate λ, that is,

$$\Pr[\tau \le t] = 1 - e^{-\lambda t}$$

$$pdf: f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Second, find π_0

$$1 - \pi_0 = A^*[\mu\pi_0]$$

- Third, find π_n

$$\pi_n = \pi_0 (1 - \pi_0)^n$$

p_n for GI/M/1

- Distinction between
  - π_n: the steady-state probability that an arriving customer finds *n* customers in the system
    - From "an arriving customer" point of view
  - p_n: the steady-state probability that there are *n* customers in the system
    - From "a random observer" point of view

An Illustrating Example (π_n *vs* P_n)

- A D/D/1 with E[τ]=10 min, W_s=5 min → ρ=1/2
  - E[τ] > W_s → an arriving customer never sees another customer, hence

$$\pi_0 = 1, \pi_n = 0 \text{ for } n \geq 1$$

  - ρ=1/2 → the server is busy half of the time, i.e., the system contains one customer half the time and is empty half the time as observed by an outside observer, hence

$$p_0 = 0.5, p_1 = 0.5, p_n = 0 \text{ for } n \geq 2$$

$\pi_n = p_n$ *iff* the arrival process is Poisson (shown by Wolff)

GI/M/1 (Cont'd)

- **p_n ?**
- Kleinrock showed that for GI/M/1:

$$p_0 = 1 - \rho$$
$$p_n = \rho\pi_0(1 - \pi_0)^{n-1}, n = 1,2,...$$

GI/M/1 (Cont'd)

- Find L, W, L_q, W_q?

$$L = \sum_{n=0}^{\infty} np_n = \rho\pi_0 \sum_{n=0}^{\infty} n(1-\pi_0)^{n-1}$$
$$= \rho\pi_0 \left(\sum_{n=0}^{\infty} (1-\pi_0)^n \right)' = \rho\pi_0 \frac{1}{\pi_0^2} = \frac{\rho}{\pi_0}$$

$$L_q = L - (1 * P[\text{Server is not empty}]$$
$$= L - (1 - P[0\text{customer in the system}])$$
$$= L - (1-p_0) = L - (1-(1-\rho))$$
$$= L - \rho = \frac{\rho}{\pi_0} - \rho = \frac{\rho(1-\pi_0)}{\pi_0}$$

$$W = L/\lambda = \frac{\rho}{\pi_0}/\lambda = \frac{W_s}{\pi_0}$$
$$W_q = L_q/\lambda = \frac{(1-\pi_0)W_s}{\pi_0}$$

Hands-On Problem

Considering a computer subsystem that can be modeled as the GI/M/1 queueing system. Specifically, the service time is exponentially distributed with a constant rate of 60 jobs per second. The Laplace-Stieltjes transform of the job inter-arrival time τ is assumed to be $A^*[\theta] = \frac{\theta+\lambda}{3\lambda}$ with λ = 30.

- What is the probability that an arriving job finds the system is busy?
- What is the probability that an arriving job finds 3 customers in the system?
- What is the average number of jobs in the system queue?
- What is the average time a job spends in the system?

Agenda

- Embedded Markov Chain Queueing Systems
  - M/G/1,
  - M/D/1,
  - GI/M/1
- **Priority Queueing Systems**

Kendall Notation (review)

Standard notation for queueing systems:

A/B/c/K/m/Z

- **A**: arrival process or inter-arrival time distribution
  - 'M' = Poisson arrival process
  - 'D' = Deterministic (constant) arrival rate
  - 'G' = General arrival process
- **B**: service process or service time dist.
  - 'M' = 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: ∞)
- **m**: total job/customer population (default: ∞)
- **Z**: scheduling discipline (default: FIFO)

Priority Queueing Systems

- Concepts
- Two control policies
- Results on M/G/1, M/M/c priority queueing systems

Priority Queueing Systems

- Before we focused on non-priority queueing systems: FCFS/FIFO
- Priority queueing systems: queueing systems in which some customers get preferential treatment
  - Customers divided into priority classes, numbered from *1* to *n*
  - The lower the priority class number, the higher the priority
  - Customers within a class served by FCFS

Assumption

- All queue schedule disciplines are *work conserving*
  - Customers do not leave without completing service
  - The server is never idle if there are customers present requiring service

Control Policies

- Control policies to resolve the situation wherein a customer of class i arrives to find a customer of class j in service ($i<j$)
- Non-preemptive
  - The newly arrived customer always waits until the customer in service completes service before gaining access to the service facility → a head-of-the-line (HOL) system
- Preemptive

Control Policies (Cont'd)

- Preemptive
  - service of the customer of class *j* is interrupted, and the newly arrived customer of higher priority begins service
  - The interrupted customer returns to the head of the *j*th class
  - The interrupted customer resumes the service at the point of interruption – *preemptive-resume,* or
  - The interrupted customer repeats the entire service from the beginning -- *preemptive-repeat*

M/G/1 Priority Q Systems

- Consider an M/G/1 queueing system with any queue discipline that chooses customers by an algorithm that **does not consider customer service times or any measure of them**. Then the performance measures L, W, Lq, Wq will be the same as for the FCFS queue discipline (shown by Kleinrock)

$$L = \rho + \frac{\rho^2(1+C_s^2)}{2(1-\rho)}, \text{ where } C_s = \frac{\sqrt{Var[s]}}{E[s]}$$

$$W = L/\lambda = W_s + \frac{\rho W_s(1+C_s^2)}{2(1-\rho)}$$

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

$$W_q = L_q/\lambda = \frac{\rho^2(1+C_s^2)}{2(1-\rho)\lambda} = \frac{\rho W_s(1+C_s^2)}{2(1-\rho)}$$

M/M/c Priority Q Systems

- Also called multi-server priority systems
- Consider M/M/c non-preemptive systems with *n* priority classes
  - The customers arrival to class *i* is Poisson with rate λ_i. Then the overall arrival pattern is Poisson with mean (superposition property)

$$\lambda = \lambda_1 + \lambda_2 + \ldots \lambda_n$$

  - Each customer has the same exponential service time requirement with a mean of 1/μ
  - Cobham showed that

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

$$E[q_1] = \frac{C[c,\alpha]W_s}{c(1-\lambda_1 W_s / c)}$$

$$\text{Ref.(M/M/c with FCFS in L\#13)}: W_q = E[q] = C[c,\alpha]W_s \Big/ c(1-\rho)$$

$$E[q_i] = \frac{C[c,\alpha]W_s}{c\left[1-\left(W_s\sum_{j=1}^{i-1}\lambda_j\right)\Big/ c\right]\left[1-\left(W_s\sum_{j=1}^{i}\lambda_j\right)\Big/ c\right]}, \quad j = 2,\ldots,n$$

Next Topic

- Queueing Networks

Things to Do

- Homework
- Class Project