1

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) How was the pace of the class? 		
Too slow Slow Just right_13 Fast Too fast		
2. How was the workload of ECE560 class, when compared to your other classes?		
Much less_1Less_2Similar_10MoreMuch more		
3. How is the level of detail of material covered in the lectures?		
Not enough Just right _11 Too much _2		
4. How clear are the lectures/examples?		
Not clear Okay0.5 Clear1.5 Very clear11		
5. How was the midterm exam?		
Too easy Easy _5 Just right _8 Hard Too hard		
6. How were the homework assignments?		
Too easy Easy _1 Just right _11 Hard _1 Too hard		



- Overview of queueing systems (Kendall notation, Little's Law)
- Performance measures
 - Traffic intensity (α)
 - Server utilization (ρ)
 - $-\pi_n$: steady-state probability that there are *n* customers in the queueing system
 - Throughput (γ): 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 (L_g)
 - Average time in the queue (W_q)
- D/D/1 queueing systems
- M/M/1 queueing systems

Dr. Xing ©

Periode (M/M/1)

$$\begin{array}{c}
\overbrace{\mu}\\ \overbrace{\mu}\\$$











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

The effective / average arrival rate of customers into

the system: $\lambda_{eff} = \lambda (1 - P_B)$

$$P_{\rm B} = \pi_{\rm N} = \alpha^{\rm N} \pi_0 = \frac{\alpha^{\rm N} (1-\alpha)}{1-\alpha^{\rm N+1}}$$

The true server utilization rate ρ – probability that the server is busy:

$$\rho = 1 - \pi_0 = 1 - \frac{1 - \alpha}{1 - \alpha^{N+1}}, \quad \text{or}$$

$$\rho = \lambda_{eff} W_s = \frac{\lambda_{eff}}{\mu} = \lambda (1 - P_B) / \mu = \alpha (1 - P_B)$$

$$\int_{P} \int_{P} \frac{\alpha (1 - \alpha^N)}{1 - \alpha^{N+1}}$$
Dr. Xing © Q-Systems 10

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

Throughput (γ): rate at which jobs successfully depart from the system

 $\gamma = \mu P[>0 \text{ jobs in the system}] + 0P[0 \text{ jobs}]$ $= \mu (1 - P[0 \text{ jobs in the system}])$ $= \mu (1 - \pi_0) = \mu (1 - (\frac{1 - \alpha}{1 - \alpha^{N+1}}))$ $= \mu \frac{\alpha (1 - \alpha^N)}{1 - \alpha^{N+1}} = \frac{\lambda (1 - \alpha^N)}{1 - \alpha^{N+1}}$

Alternative way to compute γ (by looking at the input side)

Everything that arrives and is not blocked must eventually depart

$$\gamma = \lambda(1 - P_B) = \lambda(1 - \alpha^N \frac{1 - \alpha}{1 - \alpha^{N+1}}) = \frac{\lambda(1 - \alpha^N)}{1 - \alpha^{N+1}}$$

Dr. Xing ©

Q-Systems

11

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

Average number of customers in the system *L*:

$$L = \sum_{n=0}^{N} n \pi_{n} = \frac{1-\alpha}{1-\alpha^{N+1}} \sum_{n=0}^{N} n \alpha^{n}$$

= $\frac{(1-\alpha)\alpha}{1-\alpha^{N+1}} \sum_{n=1}^{N} n \alpha^{n-1}$
= $\frac{(1-\alpha)\alpha}{1-\alpha^{N+1}} \left(\sum_{n=0}^{N} \alpha^{n}\right)^{'} = \frac{(1-\alpha)\alpha}{1-\alpha^{N+1}} \left(\frac{1-\alpha^{N+1}}{1-\alpha}\right)^{'}$
= $\frac{\alpha}{1-\alpha} - (N+1) \frac{\alpha^{N+1}}{1-\alpha^{N+1}}$

Average response time *W*:

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

Note: effective arrival rate λ_{eff} used

in Little's Law! $\lambda_{eff} = \lambda(1 - P_B)$

Dr. Xing ©

 $\lambda_{eff} = \lambda (1 - P_B)$ Q-Systems





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

• For M/M/1 with an infinite buffer:

$$\pi_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho = 1 - \alpha, \text{ where } \rho = \alpha = \frac{\lambda}{\mu}$$
$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0 = \rho^n (1 - \rho) = \alpha^n (1 - \alpha)$$

• For M/M/1/N with a finite buffer:

$$\pi_0 = \frac{1-\alpha}{1-\alpha^{N+1}}, \text{ where } \alpha = \frac{\lambda}{\mu}$$
$$\pi_n = \alpha^n \pi_0 = \frac{\alpha^n (1-\alpha)}{1-\alpha^{N+1}} \text{ for } n = 0, 1, 2, ..., N$$

For $\alpha = 0.8$ and N=16, these probabilities differ by less than 2.3%

For $\alpha = 0.8$ and N=32, the difference is only 0.06%

Therefore, the infinite buffer model M/M/1 is a very good approximation of a finite buffer system M/M/1/N, even for moderate buffer sizes.

Dr. Xing ©

Q-Systems

15

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.

Q-Systems

Dr. Xing ©	
------------	--











Average number of customers in the system:

$$L=\sum_{n=0}^{\infty}n\,\pi$$

Average response time:

$$W = L / \lambda$$

Average waiting time:

$$W_q = W - W_s$$

Average queue length:

 $L_q = \lambda W_q$

 $1 - \pi_0$

Probability that at least one of the *c* servers is busy:

The probability that an arriving customer has to wait

P[wait] ?

Dr. Xing ©

for service

Q-Systems



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







• 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?

Dr. Xing ©

Q-Systems





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



Q-Systems















