Introduction to Queueing Theory for Computer Scientists

Raj Jain
Washington University in Saint Louis
Jain@eecs.berkeley.edu or Jain@wustl.edu

A Mini-Course offered at UC Berkeley, Sept-Oct 2012

These slides and audio/video recordings are available on-line at:

http://amplab.cs.berkeley.edu/courses/queue

and http://www.cse.wustl.edu/~jain/queue

- Goals of this Course
- Contents of the course
- Tentative Schedule
- Pre-requisites

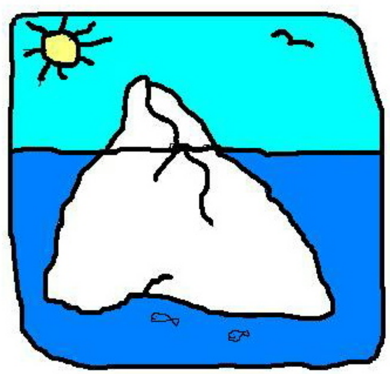
Queueing vs. Queuing

- Queueing is one character longer than Queuing
- Oxford English dictionary (England) is much thicker than Webster English dictionary (American) because English add extra letters to words: Colour, Flavour, Humour, Neighbour
- It is not American vs. English.
  There are no queues in England. They form a line.
- Queueing is unique - the only word with 5 vowels together
- Queueing is original until 1950's.
- MS word dictionary has only queuing. Corrects queueing to queuing.
  $\Rightarrow$ Now both are equally used.
- Amazon has 1176 books on queueing and 1260 books on queuing
- Google Scholar has 184000 papers on queueing and 212000 on queuing.
- Queueing is used by most respected computer scientists including Kleinrock, e.g., Queueing Systems Journal.

Goals of This Course

- Introductory course on **Applications** of Queueing Theory for Computer Scientists

1. Introduction to Queueing Theory
2. Analysis of A Single Queue
3. Queueing Networks
4. Operational Laws
5. Mean Value Analysis and Related Techniques



Queueing Models: What You will learn?

- What are various types of queues.
- What is meant by an *M/M/m/B/K* queue?
- How to obtain response time, queue lengths, and server utilizations?
- How to represent a system using a network of several queues?
- How to analyze simple queueing networks?
- How to obtain bounds on the system performance using queueing models?

Example

- **Exercise 31.3**: The average response time of a server is three seconds. During a one-minute observation interval, the idle time on the system was ten seconds.

Using a queueing model for the system, determine the following:

- System utilization
- Average service time per query
- Number of queries completed during the observation interval
- Average number of jobs in the system
- Probability of number of jobs in the system being greater than 10
- 90-percentile response time
- 90-percentile waiting time

Examples of Recent Applications

- Server virtualized system with live **VM migration**
- Service delivery improvements for **cloud service** providers
- Trading **power consumption** against performance by reserving blocks of servers
- Optimal partitioning of a **multi-core** server processor
- Modeling and optimizing the **delay-energy tradeoff** in TDM systems with sleep mode
- Optimal **inter-cell coordination** for multiple user classes with elastic traffic

Prerequisite

- Basic Probability and Statistics:
  - Mean, variance, standard deviation
  - Density function, Distribution function
  - Coefficient of variation
    Correlation coefficient
  - Median, mode, quantile
  - Normal distribution, Exponential distribution

Tentative Schedule

1	09/26/12	Introduction, Notation
2	10/03/12	Single Queue
3	10/10/12	Queueing Networks
4	10/17/12	Operational Laws
5	10/24/12	Operational Laws
6	10/31/12	Mean Value Analysis

Homeworks

- Application of the concepts to a system of your choice.
- Due by Monday noon time by email.

Text Book

- R. Jain, “Art of Computer Systems Performance Analysis,” Wiley, 1991, ISBN:0471503363
(Winner of the “1992 Best Computer Systems Book” Award from Computer Press Association”)

Other Related Topics

- Measurement techniques:
  - Workload selection
  - Workload characterization
- Probability and Statistics:
  - Use of mean, median, modes, confidence Intervals
  - Regression
- Experimental Design
  - Maximum information from minimum number of experiments
- Simulation

Quiz 0: Prerequisites

True or False?

T F

❑ ❑ The mean of a uniform(0,1) variate is 1.

❑ ❑ The sum of two normal variates with means 4 and 3 has a mean of 7.

❑ ❑ The probability of a fair coin coming up head once and tail once in two throws is 1.

❑ ❑ The density function f(x) approaches 1 as x approaches ∞.

❑ ❑ Given two variables, the variable with higher median also has a higher mean.

❑ ❑ The probability of a fair coin coming up heads twice in a row is 1/4.

❑ ❑ The difference of two normal variates with means 4 and 3 has a mean of 4/3.

❑ ❑ The cumulative distribution function F(x) approaches 1 as x approaches ∞.

❑ ❑ High coefficient of variation implies a low variance and vice versa.

Marks = Correct Answers _____ - Incorrect Answers _____ = ______

http://amplab.cs.berkeley.edu/courses/queue/quiz0.html

Quiz 1: Post Quiz

True or False?

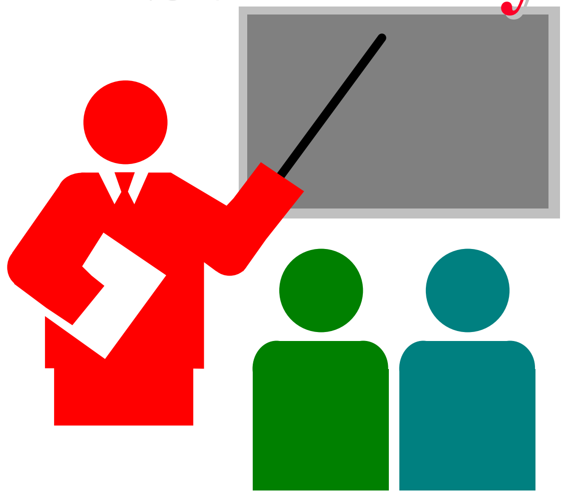
T F

- ❑ ❑ M/M/1/3/100 queue has 3 servers
- ❑ ❑ A single server queue with arrival rate of 1 jobs/sec and a service time of 0.5 seconds has server utilization of 0.5
- ❑ ❑ The delay in an G/G/∞ system is equal to the job service time.
- ❑ ❑ In a product form queueing network, the probability of a state can be obtained by multiplying state probabilities of individual queues.
- ❑ ❑ During a 10 second observation period, 400 jobs were serviced by a processor which can process 200 jobs per second. The processor utilization is 50%.
- ❑ ❑ MVA can be used to compute response times for non-product form networks.

Marks = Correct Answers _____ - Incorrect Answers _____ = ______

http://amplab.cs.berkeley.edu/courses/queue/quiz1.html

Summary



- Queueing in computer systems is quite common
- Understanding queueing theory will help you make design decisions
- Simple models are often more useful than sophisticated complex expressions with invalid assumptions.