CSE 567M: Computer Systems Analysis also known as Experimental Data Analysis

Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

These slides are available on-line at:
http://www.cse.wustl.edu/~jain/cse567-17/

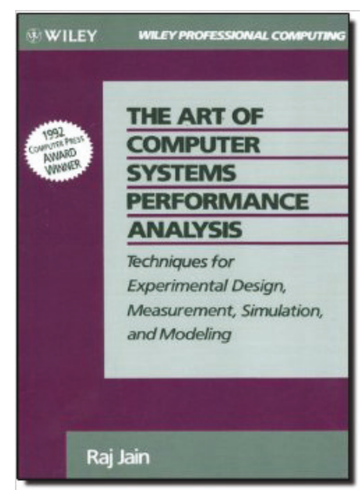
- Goal of this Course
- Contents of the course
- Tentative Schedule
- Project
- Grading

Goal of This Course

- Comprehensive course on analysis any system, algorithm, or component
- Includes measurement, statistical modeling, experimental design, simulation, and queuing theory
- How to avoid common mistakes in performance analysis
- Graduate course: (Advanced Topics)
  ⇒ Lot of independent reading and writing
  ⇒ Project/Survey paper (Research techniques)

Text Book

- R. Jain, "Art of Computer Systems Performance Analysis," Wiley, 1991, ISBN:0471503363



Objectives: What You Will Learn

- Specifying performance requirements
- Evaluating design alternatives
- Comparing two or more systems
- Determining the optimal value of a parameter (system tuning)
- Finding the performance bottleneck (bottleneck identification)
- Characterizing the load on the system (workload characterization)
- Determining the number and sizes of components (capacity planning)
- Predicting the performance at future loads (forecasting).

Basic Terms

- **System:** Any collection of hardware, software, and firmware
- **Metrics:** Criteria used to evaluate the performance of the system. components.
- **Workloads:** The requests made by the users of the system.

Main Parts of the Course

- Part I: An Overview of Performance Evaluation
- Part II: Measurement Techniques and Tools
- Part III: Probability Theory and Statistics
- Part IV: Experimental Design and Analysis
- Part V: Simulation
- Part VI: Queueing Theory
- Part VII: Stochastic Processes

Part I: An Overview of Performance Evaluation

- Introduction
- Common Mistakes and How To Avoid Them
- Selection of Techniques and Metrics

Example I

- What performance metrics should be used to compare the performance of the following systems:
  - Two disk drives?
  - Two transaction-processing systems?
  - Two packet-retransmission algorithms?

Part II: Measurement Techniques and Tools

- Types of Workloads
- Popular Benchmarks
- The Art of Workload Selection
- Workload Characterization Techniques
- Monitors
- Accounting Logs
- Monitoring Distributed Systems
- Load Drivers
- Capacity Planning
- The Art of Data Presentation
- Ratio Games

Example II

- Which type of monitor (software or hardware) would be more suitable for measuring each of the following quantities:
  - Number of Instructions executed by a processor?
  - Degree of multiprogramming on a timesharing system?
  - Response time of packets on a network?

Part III: Probability Theory and Statistics

- Probability and Statistics Concepts
- Four Important Distributions
- Summarizing Measured Data By a Single Number
- Summarizing The Variability Of Measured Data
- Graphical Methods to Determine Distributions of Measured Data
- Sample Statistics
- Confidence Interval
- Comparing Two Alternatives
- Measures of Relationship
- Simple Linear Regression Models
- Multiple Linear Regression Models
- Other Regression Models

Example III

- The number of packets lost on two links was measured for four file sizes as shown below:

File Size	Link A	Link B
1000	5	10
1200	7	3
1300	3	0
50	0	1

Which link is better?

Part IV: Experimental Design and Analysis

- Introduction to Experimental Design
- 2^k Factorial Designs
- $2^k r$ Factorial Designs with Replications
- 2^{k-p} Fractional Factorial Designs
- One Factor Experiments
- Two Factors Full Factorial Design without Replications
- Two Factors Full Factorial Design with Replications
- General Full Factorial Designs With *k* Factors

Example IV

- The performance of a system depends on the following three factors:
  - Garbage collection technique used: G1, G2, or none.
  - Type of workload: editing, computing, or AI.
  - Type of CPU: C1, C2, or C3.

How many experiments are needed? How does one estimate the performance impact of each factor?

Part V: Simulation

- Introduction to Simulation
- Types of Simulations
- Model Verification and Validation
- Analysis of Simulation Results
- Random-Number Generation
- Testing Random-Number Generators
- Random-Variate Generation
- Commonly Used Distributions

Example V

- In order to compare the performance of two cache replacement algorithms:
  - What type of simulation model should be used?
  - How long should the simulation be run?
  - What can be done to get the same accuracy with a shorter run?
  - How can one decide if the random-number generator in the simulation is a good generator?

Part VI: Queueing Theory

- Introduction to Queueing Theory
- Analysis of A Single Queue
- Queueing Networks
- Operational Laws
- Mean Value Analysis and Related Techniques
- Convolution Algorithm
- Advanced Techniques

Example VI

- The average response time of a database system 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:

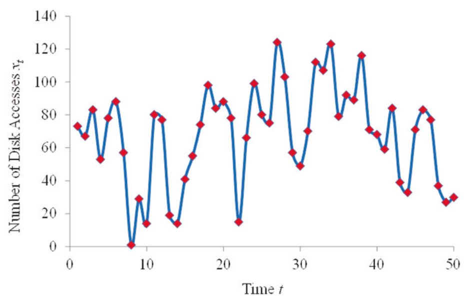
  - System utilization
  - Average service time per query
  - Number of queries completed during the observation interval
  - Average number of jobs in the system
  - Probability of number of jobs in the system being greater than 10
  - 90-percentile response time
  - 90-percentile waiting time

Part VII: Stochastic Processes

- What are different types of time series models?
- How do you fit a model to a series?
- How do you model a series that has a periodic or seasonal behavior as is common in video streaming?
- What are heavy-tailed distributions and why they are important?
- How to check if a sample of observations has a heavy tail?
- What are self-similar processes?
- What are short-range and long-range dependent processes?
- Why does long-range dependence invalidate many conclusions based on previous statistical methods?
- How do you check if a sample has a long-range dependence?

Example VII

- What is the right model for the following measurements on # of disk accesses



The Art of Performance Evaluation

- Given the same data, two analysts may interpret them differently.

Example:

- The throughputs of two systems A and B in transactions per second is as follows:

System	Workload 1	Workload 2
A	20	10
B	10	20

Possible Solutions

- Compare the average:

System	Workload 1	Workload 2	Average
A	20	10	15
B	10	20	15

Conclusion: The two systems are equally good.

- Compare the ratio with system B as the base

System	Workload 1	Workload 2	Average
A	2	0.5	1.25
B	1	1	1

Conclusion: System A is better than B.

Solutions (Cont)

- Compare the ratio with system A as the base

System	Workload 1	Workload 2	Average
A	1	1	1
B	0.5	2	1.25

Conclusion: System B is better than A.

- Similar games in: Selection of workload, Measuring the systems, Presenting the results.
- Common mistakes will also be discussed.

Grading

- Exams (Best of 2 mid terms + Final) 60%
- Class participation 5%
- Homeworks 15%
- Project 20%

Prerequisites

- CSE 131: Computer Science I
- CSE 126: Introduction To Computer Programming
- CSE 260M: ~~Introduction To Digital Logic And Computer Design~~ (Not required)
- Basic Probability and Statistics
- Matrix multiplication and inversion

Prerequisite

- Statistics:
  - Mean, variance
  - Normal distribution
  - Density function, Distribution function
  - Coefficient of variation Correlation coefficient
  - Median, mode, Quantile
- Programming

Tentative Schedule

Date	Topic	Chapter
8/29/17	Course Introduction	
8/31/17	Common Mistakes	2
9/5/17	Selection of Techniques and Metrics	3
9/7/17	Summarizing Measured Data	12
9/12/17	Comparing Systems Using Random Data	13
9/14/17	Simple Linear Regression Models	14
9/19/17	Other Regression Models	15
9/21/17	Experimental Designs	16
9/26/17	**Mid-Term Exam 1**	

Tentative Schedule (Cont)

Date	Topic	Chapter
9/28/17	2**k Experimental Designs	17
10/3/17	Factorial Designs with Replication	18
10/5/17	Fractional Factorial Designs	19
10/10/17	One Factor Experiments	20
10/12/17	Two Factor Full Factorial Design w/o Replications	21
10/17/17	Two Factor Full Factorial Designs with Replications	22
10/19/17	General Full Factorial Designs	23
10/24/17	Introduction to Queueing Theory	30
10/26/17	Analysis of Single Queue	31
10/31/17	**Mid-Term Exam 2**	

Tentative Schedule (Cont)

Date	Topic	Chapter
11/2/17	Queueing Networks	32
11/7/17	Operational Laws	33
11/9/17	Mean-Value Analysis	34
11/14/17	Time Series Analysis	37
11/16/17	Heavy Tailed Distributions,Self-Similar Processes, and Long-Range Dependence	38
11/21/17	Random Number Generation	26
11/23/17	***Thanks Giving Break***	
11/28/17	Analysis of Simulation Results	34
11/30/17	Art of Data Presentation	10
12/5/17	Clustering Techniques	
12/7/17	**Final Exam**	

Projects

- A survey paper on a performance topic
  - Workloads/Metrics/Analysis: Databases, Networks, Computer Systems, Web Servers, Graphics, Sensors, Distributed Systems
  - Comparison of Measurement, Modeling, Simulation, Analysis Tools: NS2
  - Comprehensive Survey:
    Technical Papers, Industry Standards, Products
- A real case study on performance of a system you are already working on
- Average 6 Hrs/week/person on project + 9 Hrs/week/person on class
- Recent Developments: Last 2 to 4 years ⇒ Not in books
- Better ones may be submitted to magazines or journals

Projects (Cont)

- **Goal:** Provide an insight (or information) not obvious before the project.
- **Real Problems:** Thesis work, or job
- **Homeworks:** Apply techniques learnt to your system.

Example of Previous Case Studies

- Performance of Google App Engine and Amazon Web Service
- Availability and Sensitivity of Smart Grid Components
- Modeling and Analysis Issues in x86-based Hypervisors
- Image Sensor Performance
- Performance of Solving Laplace's Equation using Auto-Pipe
- Performance Modeling of Multi-core Processors
- Performance of Named Data Networking
- A Measurement Study of Packet Reception using Linux
- Performance Analysis of Robotics Systems
- Performance and Measurement Issues of Smart Phones Design
- Analysis of Online Social Networks
- Measurement Study on the BitTorrent File Distribution System
- A Survey of Wireless Sensor Network Simulation Tools

Project Schedule

Tue 10/03	Topic Selection
Tue 10/10	References Due
Tue 10/17	Outline Due
Tue 11/07	First Draft Due → Peer reviewed
Tue 11/14	Reviews Returned
Tue 11/21	Final Report Due

Office Hours

- Tuesday/Thursday:11 AM to 12 noon
- Office: Jolley 208
- Teaching Assistant:
  - Maede Zolanvari, maede.zolanvari@wustl.edu
  - Office Hours: Monday/Friday 1-2PM
  - Jolley 323

Exams

- Exams consist of numerical, fill-in-the-blank and multiple-choice (true-false) questions.
- There is negative grading on incorrect multiple-choice questions. Grade: +1 for correct. -1/(n-1) for incorrect.
  ⇒ For True-False: +1 for Correct, -1 for Incorrect
  This ensures that random marking will produce an average of 0.
- Everyone including the graduating students are graded the same way.
- Highest score achieved becomes 100% for that exam.

Exams (Cont)

- All exams are closed book.
  One 8.5"X11" cheat sheet with your notes on both sides is allowed.
- No smart phones allowed.
  Only simple TI-30 or equivalent calculator allowed for calculations.
- Exam dates are fixed and there are no substitute exams
  ⇒ Plan your travel accordingly.
- Best of the two mid-terms is used.

Homework Submission

- All homeworks are due on the following Tuesday at the beginning of the class unless specified otherwise.
- Any late submissions, if allowed, will *always* have a penalty.
- All homeworks should be submitted in hardcopy
- All homeworks are identified by the class handout number.
- All homeworks should be on a separate sheet.
  Your name should be on every page.
- Please write CSE567 in the subject field of all emails related to this course.
- Use word "Homework" in the subject field on emails related homework. Also indicate the homework number.
- The first page of all homeworks submitted should be blank with only your name on the top-right corner

Homework Grading

- Grading basis: Method + Correct answer
- Show how you got your answer
  - Show intermediate calculations.
  - Show equations or formulas used.
  - If you use a spreadsheet, a statistical package, or write a program, print it out and turn it in with the homework.
  - For Excel, set the print area and scale the page accordingly to fit to a page. (See Page Setup)

Quizzes

- There may be a short 5-minute quiz at the beginning of each class to check if you have read the topics covered in the last class.

Academic Integrity

- Academic integrity is expected in homeworks
- All solutions submitted are expected to be yours and not copied from others or from solution manuals or from Internet
- All integrity violations will be reported to the department and action taken

Class Discussions

- We will use Piazza for class discussion.
- Find our class page at: https://piazza.com/wustl/fall2017/cse567m/home
- You can sign up at: https://piazza.com/wustl/fall2017/cse567m

Summary

- Goal: To prepare you for correct analysis and modeling of any system
- There will be a self-reading and writing
- Get ready to work hard

Quiz 0: Prerequisites

True or False?

T F

□ □ The mean of uniform(0,1) variates 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 high variance and vice versa.
□ □ If x is 0, then after x++, x will be 1.

Marks = Correct Answers _____ - Incorrect Answers _____ = ______

Scan This to Download These Slides

Raj Jain
http://rajjain.com

Related Modules

CSE567M: Computer Systems Analysis (Spring 2013),
https://www.youtube.com/playlist?list=PLjGG94etKypJEKjNAa1n_1X0bWWNyZcof

CSE473S: Introduction to Computer Networks (Fall 2011),
https://www.youtube.com/playlist?list=PLjGG94etKypJWOSPMh8Azcgy5e_10TiDw

Wireless and Mobile Networking (Spring 2016),
https://www.youtube.com/playlist?list=PLjGG94etKypKeb0nzyN9tSs_HCd5c4wXF

CSE571S: Network Security (Fall 2011),
https://www.youtube.com/playlist?list=PLjGG94etKypKvzfVtutHcPFJXumyyg93u

Video Podcasts of Prof. Raj Jain's Lectures,
https://www.youtube.com/channel/UCN4-5wzNP9-ruOzQMs-8NUw

Student Questionnaire

- Name: ____________________
- Email: ____________________
- Phone: ____________________
- Degree: __________ Expected Date: __________
- Technical Interest Area(s):
  ____
  ____
- Prior probability/statistics related courses/activities:
  ____
  ____
- Prior computer systems related courses (Max 5):
  ____
  ____