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/
Overview

- 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)
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's 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:

<table>
<thead>
<tr>
<th>File Size</th>
<th>Link A</th>
<th>Link B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>1200</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>1300</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>System</th>
<th>Workload 1</th>
<th>Workload 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>
Possible Solutions

- Compare the average:

<table>
<thead>
<tr>
<th>System</th>
<th>Workload 1</th>
<th>Workload 2</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>20</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>20</td>
<td>15</td>
</tr>
</tbody>
</table>

Conclusion: The two systems are equally good.

- Compare the ratio with system B as the base

<table>
<thead>
<tr>
<th>System</th>
<th>Workload 1</th>
<th>Workload 2</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>0.5</td>
<td>1.25</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Conclusion: System A is better than B.
Solutions (Cont)

- Compare the ratio with system A as the base

<table>
<thead>
<tr>
<th>System</th>
<th>Workload 1</th>
<th>Workload 2</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
<td>2</td>
<td>1.25</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/29/17</td>
<td>Course Introduction</td>
<td></td>
</tr>
<tr>
<td>8/31/17</td>
<td>Common Mistakes</td>
<td>2</td>
</tr>
<tr>
<td>9/5/17</td>
<td>Selection of Techniques and Metrics</td>
<td>3</td>
</tr>
<tr>
<td>9/7/17</td>
<td>Summarizing Measured Data</td>
<td>12</td>
</tr>
<tr>
<td>9/12/17</td>
<td>Comparing Systems Using Random Data</td>
<td>13</td>
</tr>
<tr>
<td>9/14/17</td>
<td>Simple Linear Regression Models</td>
<td>14</td>
</tr>
<tr>
<td>9/19/17</td>
<td>Other Regression Models</td>
<td>15</td>
</tr>
<tr>
<td>9/21/17</td>
<td>Experimental Designs</td>
<td>16</td>
</tr>
<tr>
<td>9/26/17</td>
<td><strong>Mid-Term Exam 1</strong></td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>Topic</td>
<td>Chapter</td>
</tr>
<tr>
<td>------------</td>
<td>------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>9/28/17</td>
<td>$2^k$ Experimental Designs</td>
<td>17</td>
</tr>
<tr>
<td>10/3/17</td>
<td>Factorial Designs with Replication</td>
<td>18</td>
</tr>
<tr>
<td>10/5/17</td>
<td>Fractional Factorial Designs</td>
<td>19</td>
</tr>
<tr>
<td>10/10/17</td>
<td>One Factor Experiments</td>
<td>20</td>
</tr>
<tr>
<td>10/12/17</td>
<td>Two Factor Full Factorial Design w/o Replications</td>
<td>21</td>
</tr>
<tr>
<td>10/17/17</td>
<td>Two Factor Full Factorial Designs with Replications</td>
<td>22</td>
</tr>
<tr>
<td>10/19/17</td>
<td>General Full Factorial Designs</td>
<td>23</td>
</tr>
<tr>
<td>10/24/17</td>
<td>Introduction to Queueing Theory</td>
<td>30</td>
</tr>
<tr>
<td>10/26/17</td>
<td>Analysis of Single Queue</td>
<td>31</td>
</tr>
<tr>
<td>10/31/17</td>
<td>Mid-Term Exam 2</td>
<td></td>
</tr>
</tbody>
</table>
# Tentative Schedule (Cont)

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

- A survey paper on a performance topic
  - 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
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 $\infty$.
☐ 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 $\infty$.
☐ 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 _____ = ______
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=PLjGG94etKypJWOSPMh8Azcg5y5e_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):
  ______________________________________
  ______________________________________