Performance Evaluation (of computing and communication systems)

COMPSCI 655PE
Philippe Nain
School of Computer Science
http://www-net.cs.umass.edu/655pe_18/

TA: Gayane Vardoyan
Objectives

- Introduction to *analytical tools* needed to construct/measure/analyze models of resource contention systems
  - computers
  - networks
- Methodology course + applications
Course outline

- Introduction (this lecture)
- Stochastic processes (6 lectures)
  - what are they?
  - Bernoulli/Poisson processes
  - Markov chains/processes
  - applications (Google page rank, 802.11, ...)
- Queueing theory (4 lectures)
  - Little’s law
  - M/M/1 queue and variants
  - transforms and M/G/1 queue
  - applications
Course outline (cont')

- Intermediate queueing theory (3 lectures)
  - priority queues
  - job scheduling
  - bounding techniques
  - applications
- Queueing networks (2 lecture)
- Approximation techniques (2 lectures)
  - aggregation
  - decomposition
  - isolation
  - applications
Course outline (cont')

- Fluid models (1 lecture)
- Modeling TCP (2 lectures)
- Simulations (1 lecture)
- Measurements, traces and analyses (2 lectures)
  - self-similarity
  - estimation
  - measurement design
- Wrapup (1 lecture)
Overview

- Given system
  - Internet, department WLAN, web server, ...

- Want to know performance
  - throughput, average response time, ...

- Utility
  - high-level design phase
    - reliable multicast protocols
  - low-level design
    - bus arbitration algorithms on multiprocessors
  - system configuration
    - how many disks, how much memory, ...
Overview (cont')

Qs:
- Appropriate performance metrics?
- PE methodologies?
- Salient features of PE?
Performance metrics

User point of view

- response time (web server, ssh, scp, Matlab)
  - average
  - variance
  - tail
- quality of result (video on demand, distortion)
  - fraction of frames lost
  - number of halts

System point of view

- throughput
- no. supported sessions
Avg. response time vs. throughput

Note: low delay $\Rightarrow$ low throughput; high throughput $\Rightarrow$ high delay

Once load exceeds threshold, performance often falls apart

E.g., thrashing in paging system
Other metrics

- Device utilization
  - fraction of time device busy
  - useful secondary measure for tracking system problems
  - e.g., 100% CPU util. can explain long response times
- Reliability - prob. of system failure
- Availability, fraction of time system operational
Pb you’ll soon able to solve

- **Computer systems**
  - given choice between single machine with speed $s$, or $n$ machines each with speed $s/n$, which should you choose?
  - if both arrival rate and service rate double, will mean response time stay same?
  - if scheduling policy favors one set of jobs, does it necessarily hurt some other jobs?

- **Communication networks**
  - throughput of wireless protocol where each user accesses channel using same frequency?
  - throughput TCP connection?
  - throughput wifi 802.11 protocol?

- **Call centers**
  - how many operators needed to keep rejection prob. of incoming request low?

... and many others ...
Methods

**Measurement:** measure performance of existing system

**Simulation:** build software emulator of system; execute it; use traces or random numbers to generate workload

**Analysis:** build mathematical model that captures essence of system; use mathematical tools to evaluate performance

**Hybrid:** combinations of above
<table>
<thead>
<tr>
<th></th>
<th>measurement</th>
<th>simulation</th>
<th>analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>when</strong></td>
<td>prototyping, configuration</td>
<td>anytime</td>
<td>anytime</td>
</tr>
<tr>
<td></td>
<td>tuning, monitoring</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>effort</strong></td>
<td>varies</td>
<td>moderate</td>
<td>low</td>
</tr>
<tr>
<td><strong>cost</strong></td>
<td>high</td>
<td>moderate</td>
<td>low</td>
</tr>
<tr>
<td><strong>tools</strong></td>
<td>instrumentation</td>
<td>languages</td>
<td>mathematics</td>
</tr>
<tr>
<td><strong>accuracy</strong></td>
<td>varies</td>
<td>moderate</td>
<td>low</td>
</tr>
<tr>
<td><strong>salability</strong></td>
<td>high</td>
<td>medium</td>
<td>low</td>
</tr>
<tr>
<td><strong>scalability</strong></td>
<td>low</td>
<td>medium</td>
<td>high</td>
</tr>
</tbody>
</table>
Salient system features

- Resource contention
  - need to determine waiting times
  - need to evaluate different scheduling policies

- Unknown service requirements
  - use statistical description, e.g., average, variance

Suggests use of
- probabilistic methods
- queueing theory
- statistical methods
Modeling cycle

- Construct model
  - Abstraction of essential features
  - Ignore non-essential features
- Model evaluation
- Model validation
  - Simulation: construct program
  - Analysis: solve model
  - Measurement + statistics
  - Intuition
  - Measurements
  - Simulation
  - Analysis
- Done
Course materials

- No required text
- Recommended reading material:
  - Sheldon Ross, "Stochastic Processes"
  - Mor Harchol Balter, "Performance Modeling and Design of Computer Systems"
  - course notes (see course website)
- website
Course requirements

Prerequisite: B+ or better in a solid course on probability theory
Try practice problem set on web. If in doubt drop course

Requirements:
5 homeworks: 4 best enter 50% of course grade
Take-home midterm exam: 25% of course grade
In-class final exam: 25% of course grade
Will use Moodle for assignments and exams
Miscellaneous

- Prof. office hours: Tu/Th 11:30am-12:30pm, Room TBA + by appointment
  email address: pnain@umass.edu

- TA office hours: Tu/Th 9am-10am, Room CS207
  email address: gvardoyan@cs.umass.edu

- Class email list: TBA
Probability: what you need to know

Sample space $S$ - set of all possible events

- set of individuals in room
- set of UMass routers
- $\{1,2,...,6\}$ - sides of a die
- real line
- $\{0,1\}$

Event $E \subseteq S$

$E=[0,10], \ S = [0,\infty)$

e.g. service times less than or equal to 10
Probability $P$ defined on events

- $0 \leq P[E] \leq 1$
- if $E = \emptyset$, $P[E] = 0$; if $E = S$, $P[E] = 1$
- $P[\bigcup_n E_n] = \sum_n P[E_n]$ for any countable, mutually exclusive set of sets $\{E_n\}$
random variable, a function $X : S \rightarrow \mathbb{R}$

- $X = x$
- $X \leq t$

$P(X = x) =$

$P(\{\omega : X(\omega) = x\})$

$P(X \leq t)$

$= P(\{\omega : X(\omega) \leq t\})$
random variable, a function $X: S \rightarrow \mathbb{R}$

- $X = x$
- $X \leq t$

- $P(X = x) = P(\{\omega : X(\omega) = x\})$

- $P(X \leq t)$
  $= P(\{\omega : X(\omega) \leq t\})$

or $X^{-1}((-\infty,t])$

$\{\omega : X(\omega) \leq t\}$

$X(\omega) \leq t$

or $X^{-1}(x)$

$\{\omega : X(\omega) = x\}$
Discrete random variables (rvs)

If $X$ takes discrete values, i.e., $0, 1, ...$

$$P_X(x) = P(X = x), \quad x = 0, 1, ...$$

Note: if $X$ is continuous valued and $F_X$ is continuous,

$$P_X(x) = 0, \quad -\infty < x < \infty$$
Continuous-valued rvs

- $f_X(x) = dF_X(x)/dx$ provided that $F_X$ is continuous

- $P(x \leq X \leq x + \Delta x) \approx f_X(x) \Delta x$

- example, $X$ exponential random variable (rv) with rate $\lambda$

\[
f_X(x) = \begin{cases} 
\lambda e^{-\lambda x}, & x \geq 0 \\
0, & x < 0
\end{cases}
\]

\[
F_X(x) = \begin{cases} 
1 - e^{-\lambda x}, & x \geq 0 \\
0, & x < 0
\end{cases}
\]
Q: How can one generate instances of $X$ with given pdf $F_X(x)$?

A: given $X$, $F_X(x)$

• generate $x_1$, $x_2$, $x_3$, ...
• assume can generate numbers, uniformly distributed between 0 and 1 (rand)
• inverse function method
  
  $x_1 = F_X^{-1}(\text{rand})$

  
  $F_X^{-1}(y) = \inf\{x: F_X(x) \geq y\}$
Generating rvs (cont')

Example: \( F_X(x) = 1 - \exp(-ax), \ x > 0 \)

\( y = 1 - \exp(-ax) \) gives \( x = (-1/a)\log(1-y) \)

Therefore \( F^{-1}_X(y) = (-1/a)\log(1-y) \)

If \( u_1, u_2, u_3 \) uniform (independent) rvs in \([0,1]\) then

\( x_1 = (-1/a)\log(1-u_1), \ x_2 = (-1/a)\log(1-u_2), \ x_3 = (-1/a)\log(1-u_3) \)

Exponential (independent) rvs with rate \( a \)
Aside

If \( F_X(u) < F_Y(u) \), for all \( u \), then

\[
X \prec_d Y
\]

\( X \) is less than \( Y \) in distribution.
Expectation

- if $X$ is discrete

$$E[X] = \sum_{\{x_i\}} P[(X = x_i)]x_i$$

- if $X$ is continuous

$$E[X] = \int_{-\infty}^{\infty} xf_X(x)dx$$


**Expectation (cont')**

- more generally,
  
  \[
  E[h(X)] = \sum h(x_i)P[X = x_i]
  \]
  
  \[
  E[h(X)] = \int h(x)f_X(x)dx
  \]

- often interested in
  
  \[E[X^k], \quad k = 2, 3, \ldots\]

  \[\sigma_X^2 = E[(X - E[X])^2] = E[X^2] - E[X]^2, \text{ (variance of } X)\]
Multiple rvs

- \( F_{X,Y}(x, y) = P[X \leq x, Y \leq y], \quad f_{X,Y}(x, y) \)

- \( E[h(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f_{X,Y}(x, y) \, dx \, dy \)

- Note \( E[X + Y] = E[X] + E[Y] \)

Independence: \( X \) and \( Y \) are independent iff

\[ F_{X,Y}(x, y) = F_X(x)F_Y(y) \]
Multiple rvs (cont')

\[ \text{var} (X+Y) = ? \]

\[
= E[Z^2 - (E[Z])^2] \quad \text{with } Z = X+Y \\
= E[X^2] + E[Y^2] - (E[X])^2 - (E[Y])^2 \\
\quad + 2E[XY] - 2E[X]E[Y] \\
= \text{var}(X) + \text{var}(Y) + 2(E[XY] - E[X]E[Y])
\]

\[ \text{var} (X+Y) = \text{var}(X)+\text{var}(Y) \text{ iff } X \text{ and } Y \text{ independent rvs as in this case } E[XY]=E[X]E[Y] \]
Bayes formula

Given two events $A$ and $B$

$$P(A \text{ and } B) = P(A|B) \ P(B) = P(B|A) \ P(A)$$

therefore

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \text{ provided } P(B) > 0$$
Conditioning

- if $X, Y$ discrete valued, say $X, Y \in \mathbb{N} = \{1, 2, \ldots\}$
  \[
P_{X,Y}(x,y) = P(X = x, Y = y), \quad x, y \in \mathbb{N}
  \]

  \[
P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}, \quad x, y \in \mathbb{N}
  \]

  where

  \[
P_Y(y) = \sum_{x \in \mathbb{N}} P_{X,Y}(x,y), \quad y \in \mathbb{N}
  \]
Conditioning (cont')

May want density of $X$ conditioned on $Y = y$

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Note:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)dy$$

or

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$
Interesting and weird feature of exponential rv

Q: \( P(X > s+t \mid X > s) = ? \quad s>0, \ t>0 \)
   with \( P(X>x) = \exp(-ax) \) (X is an exp rv)

A: Bayes formula with \( A=\{X>s+t\} \), \( B=\{X>s\} \) gives

\[
P(X > s+t \mid X > s) = \frac{P(X>s+t \text{ and } X>s)}{P(X>s)}
= \frac{P(X>s+t)}{P(X>s)}
= \frac{\exp(-a(s+t))}{\exp(-as)}
= \exp(-at) \quad \text{independent of } s!
\]

Exponential rvs are memoryless