Introduction

John C.S. Lui

Department of Computer Science & Engineering
The Chinese University of Hong Kong
www.cse.cuhk.edu.hk/∼cslui

Outline

Outline

Introduction

- Introduction to *analytical tools* which are needed to construct/analyze models of resource contention systems.
  - computers
  - networks
- It is a course on methodologies **PLUS** applications
- Not covered: *measurement*

Course Outline

- **Introduction**
- **Probability and Random Variable (review)**
- **Stochastic Processes**
  - What are they?
  - Bernoulli/Poisson processes
  - Markov chains
  - applications
- **Elementary queueing theory**
  - Little's Law
  - M/M/1 queue and variants
  - transforms and M/G/1 queue
  - stochastic differential equations and M/G/1 queue
  - applications

Course Outline (continue)

- **Intermediate queueing theory**
  - priority queues, queues with vacation
  - bounding techniques
  - matrix solution methods
  - applications
- **Bounds, inequalities and approximation techniques**
  - aggregation and decomposition
  - isolation
  - applications
- **Priority Queueing Systems**
- **Large deviation theory**
- **Fluid Analysis**
- **Introduction to Stochastic Dynamic Programming**
- **Queueing networks**

If time allows

- **Stochastic comparison**
- **Markov-modulated Processes and their analysis**
- **Online Stochastic Combinatorial Optimization**
- **Economic Models of Communication Networks**
- **Non-product form networks, dynamic control of queueing systems, multi-armed bandit problems and statistical issues**

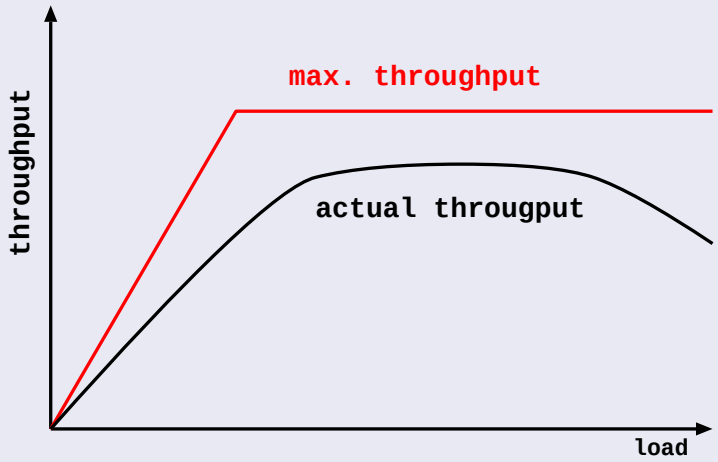
Overview

- **given a system**
  - Internet, department LAN, web server,....
- **want to know its performance**
  - throughput, avg. response time,...
- **utility**
  - high-level design phase (e.g., reliable multicast protocols,...)
  - low-level design phase (e.g., bus arbitration algorithms on multiprocessors,...)
  - system configuration (e.g., how many disks, how much memory,...)
- **questions:**
  - appropriate performance metrics?
  - performance evaluation methodologies?
  - salient features of performance evaluations?

Performance Metrics

- **User's point of view**
  - **response time (web server, telnet, ftp)**
    - average
    - variance
    - tail
  - **quality of result** (video on demand, imprecise computation)
    - fraction of frames lost
- **System's point of view**
  - throughput
  - number of supported sessions

Average response time vs. throughput



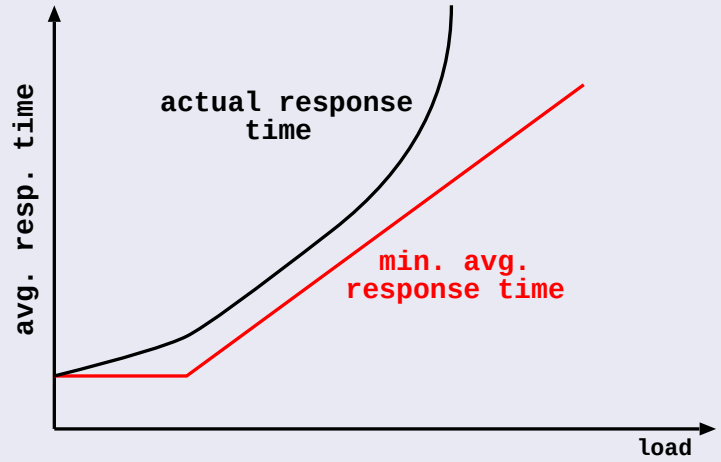


Figure: An illustration of throughput and response time

- low delay, low throughput; high throughput, 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 utilization. can explain long response time
- **reliability**: prob. of system failure
- **availability**: fraction of time system operational

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

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
- **suggest to use**
  - probabilistic methods
  - queueing theory
  - stochastic analysis

Grading

- Homework 30%
- Project 30%
- Final Exam 40%