COMP 190-088: Systems Performance Analysis

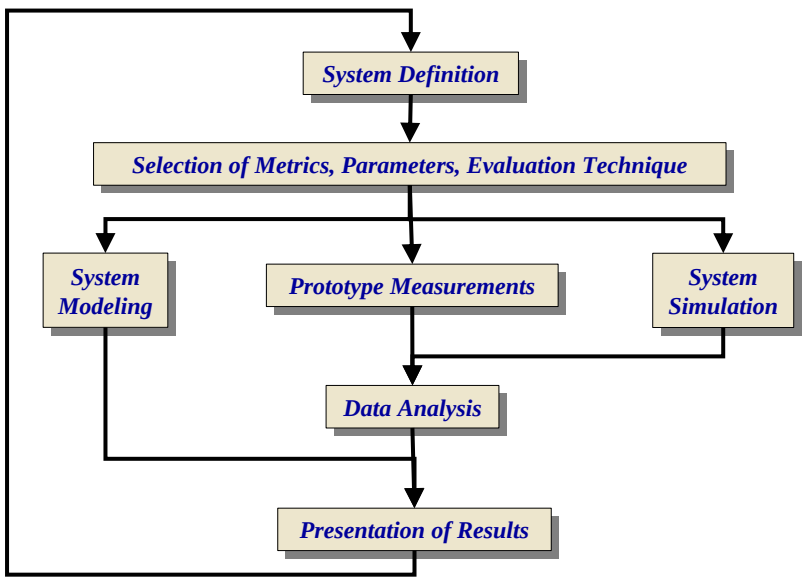
System Modeling - I

Jasleen Kaur

Department of Computer Science
The University of North Carolina at Chapel Hill

Spring 2005

Performance Analysis Methodology

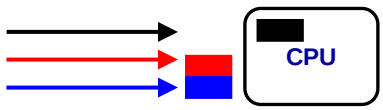


System Modeling Overview

- **Queuing Basics**
- Single-server Analysis
- Multiple-server Analysis
- Operational Laws
- Case Studies
  - Processor Scheduling
  - Disk Scheduling
  - Memory Management
- Network of Queues

Wait Queues in Systems

- Systems have several resources from which processes receive service
  - eg, CPU, Disk I/O, memory sub-system, etc
- Most resources can be used by only one job at a time
  - Queues form when current demand for a resource exceeds its capacity



- Large queues result in large response times
  - Systems that reduce waiting time are preferable

System performance modeling involves analysis of queues

What System Characteristics to Analyze?

- Example of a service-provider that needs to select a CPU for its server
  - What would it need to consider?
- Mean service time of each CPU
  - A: 20 s
  - B: 30 s
- Cost of procurement and operation of CPUs
  - A: $12/hour
  - B: $6/hour
- Revenue for each job completed: $3
- Objective: Maximize profit

Which CPU to select?

Maximizing Profit

CPU	Mean Service Time	Cost / hour	Max. Revenue / hour	Max Profit / hour	Revenue / hour	Profit / hour
A	20 s	$12	$540	**$528**	$216	$204
B	30 s	$6	$360	$324	$216	**$210**

- A yields higher maximum profit
- Analysis assumes jobs are always in line
  - Large inter-arrival times (eg, 50 s)
    - CPU is not the bottleneck!
  - Some jobs may be lost due to overflow of finite queue buffers
    - Limited memory resources available in systems
    - Limited buffers provisioned to ensure small response times

  CPUs may see a different **effective** job arrival rate!

Analysis should consider job inter-arrival times and buffer sizes

Incorporating Randomness in Service/Arrival

- Commonly used distributions for service and inter-arrival times:
  - Normal distribution:
    - Average value represents the typical job
    - ☹: assumes symmetric distribution around the average
    - Mean service time may be skewed by some house-keeping jobs
  - Exponential distribution:
    - Median is smaller than the mean
    - Memoryless distribution, which simplifies analysis
    - Often, a reasonable approximation

Recap: Characteristics Relevant for Analysis

- Arrival process:
  - Generally assumed to be independent and identically distributed (IID) random variable
  - τ : Inter-arrival times (IAT)
  - $1/\lambda$: Mean IAT
  - Commonly used distribution: Poisson arrivals
    - IATs are exponentially distributed
- Service time (S):
  - Generally assumed to be IID random variables
  - Commonly used distribution: exponential
  - $1/\mu$: Mean service time
- System queue capacity (B):
  - Includes the jobs in service
  - If buffers are large, convenient for analysis to assume infinite

Characteristics Relevant for Analysis

- Number of servers (m):
  - eg, number of CPUs in a multi-processor environment
  - When non-identical servers:
    - Divide into groups of identical servers
    - Separate queues for each group
- Population size (K):
  - Total number of jobs that can ever arrive at the system
  - Finite for most real systems
  - If large, assumed infinite to facilitate easy analysis

Characteristics Relevant for Analysis

- Service discipline:
  - Order in which queued requests are served
  - Common choices: First-come-first served (FCFS/FIFO)
  - Others:
    - Last-come-first-served (LCFS),
    - LCFS with Preemt and Resume (LCFS-PR)
    - Round-robin (RR)
      - Fixed-sized quantum
    - Processor Sharing (PS)
      - Infinitesimally small quantum size
    - Delay Center
      - Infinite servers with fixed service time
      - eg, satellite links
  - Some based on service time
    - Shortest Processing Time First (SPT)