MODULE 14: Selected Topics 2

Lecture 14.2
Performance Measurement

Prepared By:

- Scott F. Midkiff, PhD
- Luiz A. DaSilva, PhD
- Kendall E. Giles, PhD

Electrical and Computer Engineering
Virginia Tech

Lecture 14.2 Objectives

- Describe the motivation for performance measurements and analysis
- List and give examples of the three types of performance study methods – analytical, simulation, and measurements – and describe their characteristics
- List the basic steps for a performance study
- Define common metrics used in computer and network systems performance studies, including response time, latency, throughput, blocking probability, and availability

Motivation for Performance Analysis

- Why worry about performance? Answers should be obvious!
  - Budgets are never unlimited – there is always a cost/performance trade-off
  - Excessive periods of poor performance lead to disgruntled users
  - Lack of performance analysis may lead to solutions that do not fix a problem
  - Without performance analysis, one may be responding to perception, not reality
  - Performance is a key parameter in the “interface” to vendors (including service providers) and customers

A Simple Example

- You are buying a PC with a fixed budget
- You are considering two systems – both exactly match your budget and are otherwise the same
    - A PC with a faster processor, but less main memory
    - A PC with a slower processor, but more main memory
- Which one is better? It depends!
    - What operating system is used?
    - What applications are used?
    - Which applications are important?
- Performance analysis can help you decide which PC to buy

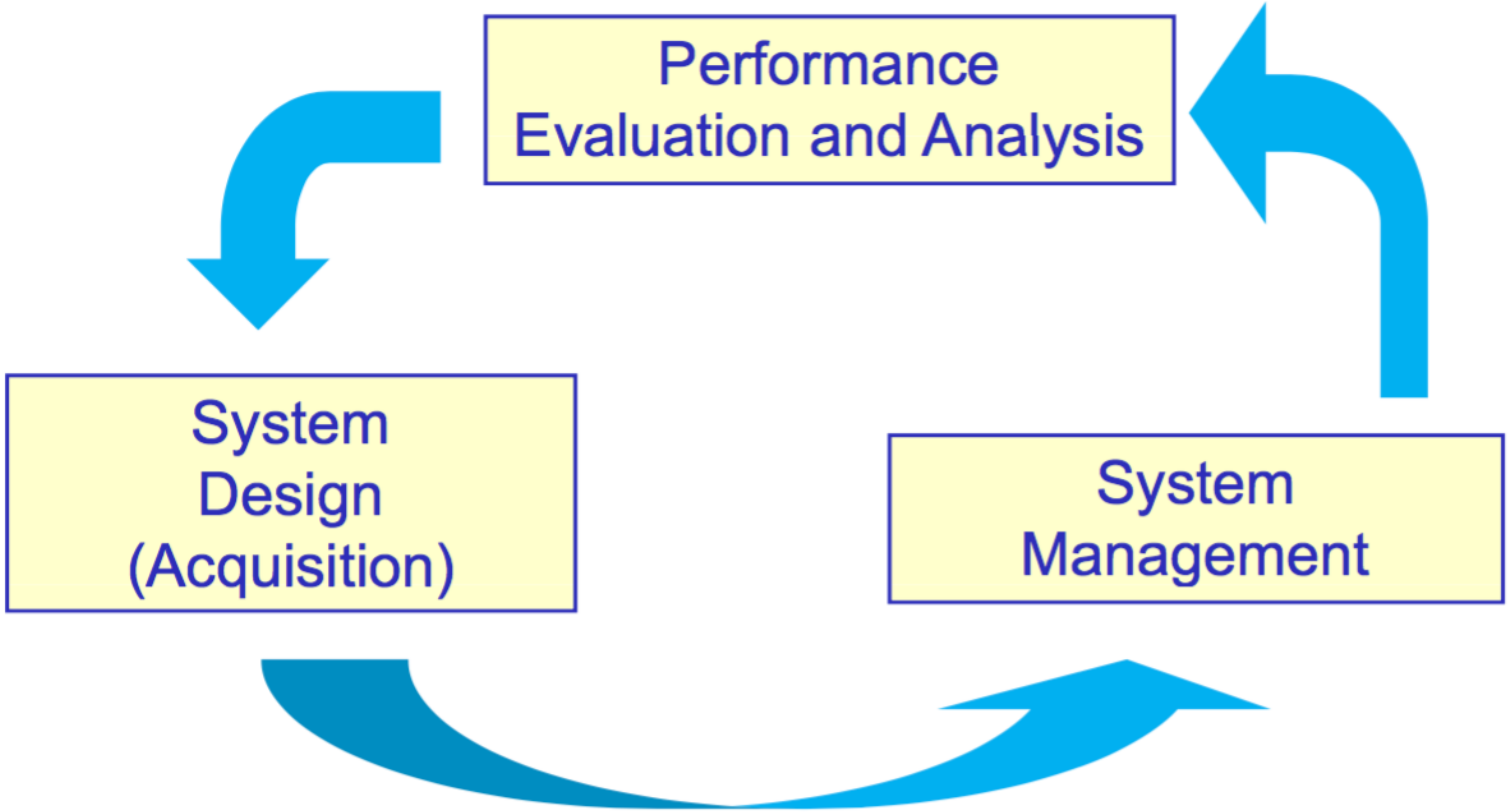
Performance Analysis

- Why?
  - Bottleneck or capacity analysis
  - Sensitivity of performance to parameters
  - Detecting problem areas to guide system tuning
  - Configuration planning and trade-offs
  - Benchmarking
- When?
  - Architecture/system design (before realization or acquisition)
  - Detailed design, implementation, and configuring
  - Operation and management

Performance Analysis (cont’d)

- How?
  - Analytical models, e.g., queuing models with closed form or numerical solutions
  - Simulation experiments, e.g., discrete-event simulation with statistical results
  - Empirical measurements, e.g., instrumented code and network monitoring

Performance, Design, and Management



Performance Evaluation and Analysis

System Design (Acquisition)

System Management

CHECK POINT

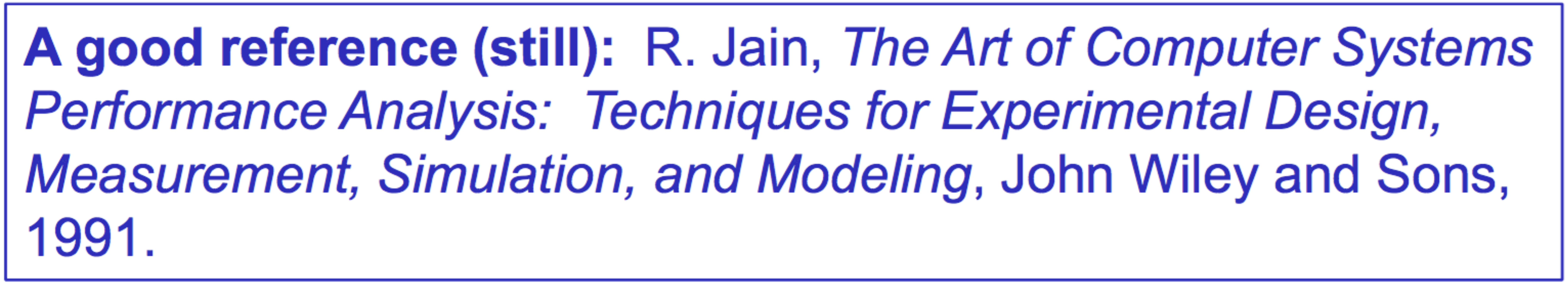
As a checkpoint of your understanding, please pause the video and make sure you can do the following:

- Describe the motivation for performance measurements and analysis
- List and give examples of the three types of performance study methods – analytical, simulation, and measurements – and describe their characteristics

If you have any difficulties, please review the lecture video before continuing.

Systematic Approach

1. State goals of the study and define the system
   - Delineate system boundaries
   - Goals will determine appropriate boundaries
2. List services (functions) and outcomes of services
3. Select metrics
   - Examples: throughput, latency, availability, etc.



A good reference (still): R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*, John Wiley and Sons, 1991.

Systematic Approach (cont'd)

4. List parameters that affect performance
   - System parameters
   - Workload parameters
5. Select factors to study
   - Factors are parameters that will be varied
   - Other parameters will remain fixed
6. Select evaluation technique
   - Analytical
   - Simulation
   - Measurement

Systematic Approach (cont'd 2)

7. Select workload, or list of service requests
    - Probability distributions for analytical study
    - Trace or distribution for simulation study
    - Scripts for measurements
8. Design experiments
    - Phase 1: Large number of factors, few values
    - Phase 2: Small number of factors, many values

Systematic Approach (cont’d 3)

9. Analyze and interpret results
   - Results are random
   - Statistical techniques needed to compare results
10. Present results
    - Graphical techniques
    - Refinement and iteration may be needed

Evaluation Techniques

Criterion	Analytical	Simulation	Measurement
Stage	Any	Any	Post-prototype
Time required	Small	Medium	Varies
Accuracy	Low	Moderate	Varies
Trade-off Evaluation	Easy	Moderate	Difficult
Cost	Low	Medium	High
“*Saleability*”	Low	Medium	High

☑ Use of multiple techniques allows validation

CHECK POINT

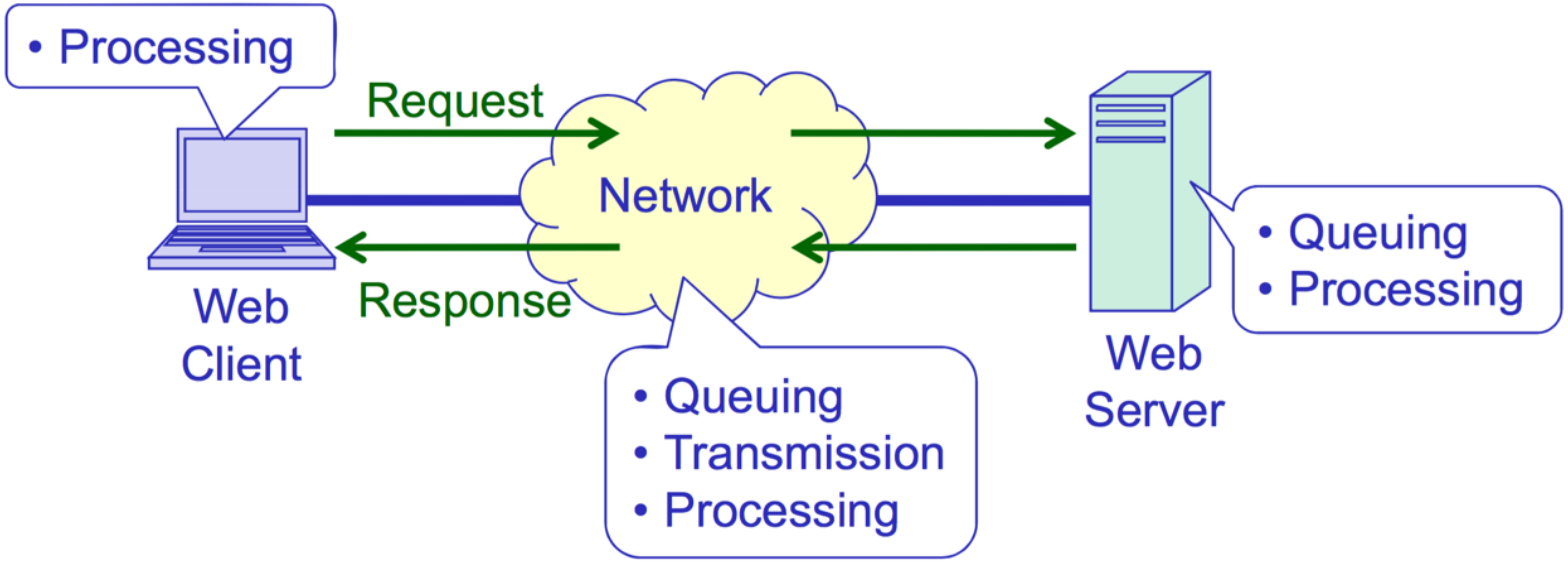
As a checkpoint of your understanding, please pause the video and make sure you can do the following:

- List the basic steps for a performance study

If you have any difficulties, please review the lecture video before continuing.

Metrics: Response Time

- Given a request or other event, the response time is the wait until the system responds or completes the action
- Often depends on a number of performance characteristics for a number of system components
- Example: time for a web page to load after clicking a link



Metrics: Latency

- Latency is just the delay from some event until another event (caused by the first)
- Response time is a special case of latency
- Often depends on a number of performance characteristics and it can contribute to other metrics
- Example: time from when one host (a client) until the message (a request) is received by the destination host (a server)

Metrics: Throughput

- Throughput is a measure of productivity – how many units of work (computations, packet transmissions, queries processed, etc.) are performed in a unit of time
- Higher throughput may lead to:
  - Low response time if higher throughput is applied to the tasks of a fixed number of users, or
  - Similar response time if throughput is used to support an increased number of users
- Examples:
  - Instructions per second
  - Packets per second
  - Database queries per second

Metrics: Blocking Probability

- “Blocking systems” limit the number of users or sessions
  - Fixed resources (e.g., fixed number of software license connections or channels)
  - Assured service levels require limits on users (e.g., throughput for a batch computing system)
- Performance measure is the probability that a request for service is accepted (the call goes through) or is blocked (a system busy signal is received)
- Depends on:
  - Arrival rate (number of requests)
  - Time in service (time to complete accepted requests)

Metrics: Availability

- Availability is used to express the probability that a system is available for use
  - Maybe unavailable due to failure (including due to intentional or unintentional attack)
  - Maybe unavailable for planned maintenance
- Availability depends on:
  - Mean time to failure (MTTF)
  - Mean time to repair (MTTR)

CHECK POINT

As a checkpoint of your understanding, please pause the video and make sure you can do the following:

- Define common metrics used in computer and network systems performance studies, including response time, latency, throughput, blocking probability, and availability

If you have any difficulties, please review the lecture video before continuing.

Summary

- Performance measurements are important to many aspects of designing, acquiring, operating, and/or using computer and network systems
- Performance studies should have well defined goals and scope and should be based on an understanding of metrics, parameters, and factors
- Three type of analysis are commonly used:
    - Analytical – for basic understanding and early analysis
    - Simulation – measure a system that is not available
    - Measurements – requires a system for the study
- Common metrics used in computer and network systems performance studies include response time, latency, throughput, blocking probability, and availability

MODULE 14: Selected Topics 2

Lecture 14.2
Performance Measurement

Prepared By:

- Scott F. Midkiff, PhD
- Luiz A. DaSilva, PhD
- Kendall E. Giles, PhD

Electrical and Computer Engineering
Virginia Tech