

Performance Evaluation:

Selection of Techniques and Metrics

Hongwei Zhang

<http://www.cs.wayne.edu/~hzhang>



Acknowledgement: this lecture is partially based on the slides of Dr. Raj Jain.

Outline

- Selecting Evaluation Techniques
- Selecting Performance Metrics
- Commonly Used Performance Metrics
- Setting Performance Requirements

Outline

- **Selecting Evaluation Techniques**
- Selecting Performance Metrics
- Commonly Used Performance Metrics
- Setting Performance Requirements

Criteria for selecting evaluation techniques

Criterion	Analytical		
	Modeling	Simulation	Measurement
1. <u>Stage</u>	Any	Any	Postprototype
2. Time required	Small	Medium	Varies
3. Tools	Analysts	Computer languages	Instrumentation
4. <u>Accuracy</u>	Low	Moderate	Varies
5. <u>Trade-off evaluation</u>	Easy	Moderate	Difficult
6. Cost	Small	Medium	High
7. Saleability	Low	Medium	High

Three rules of validation

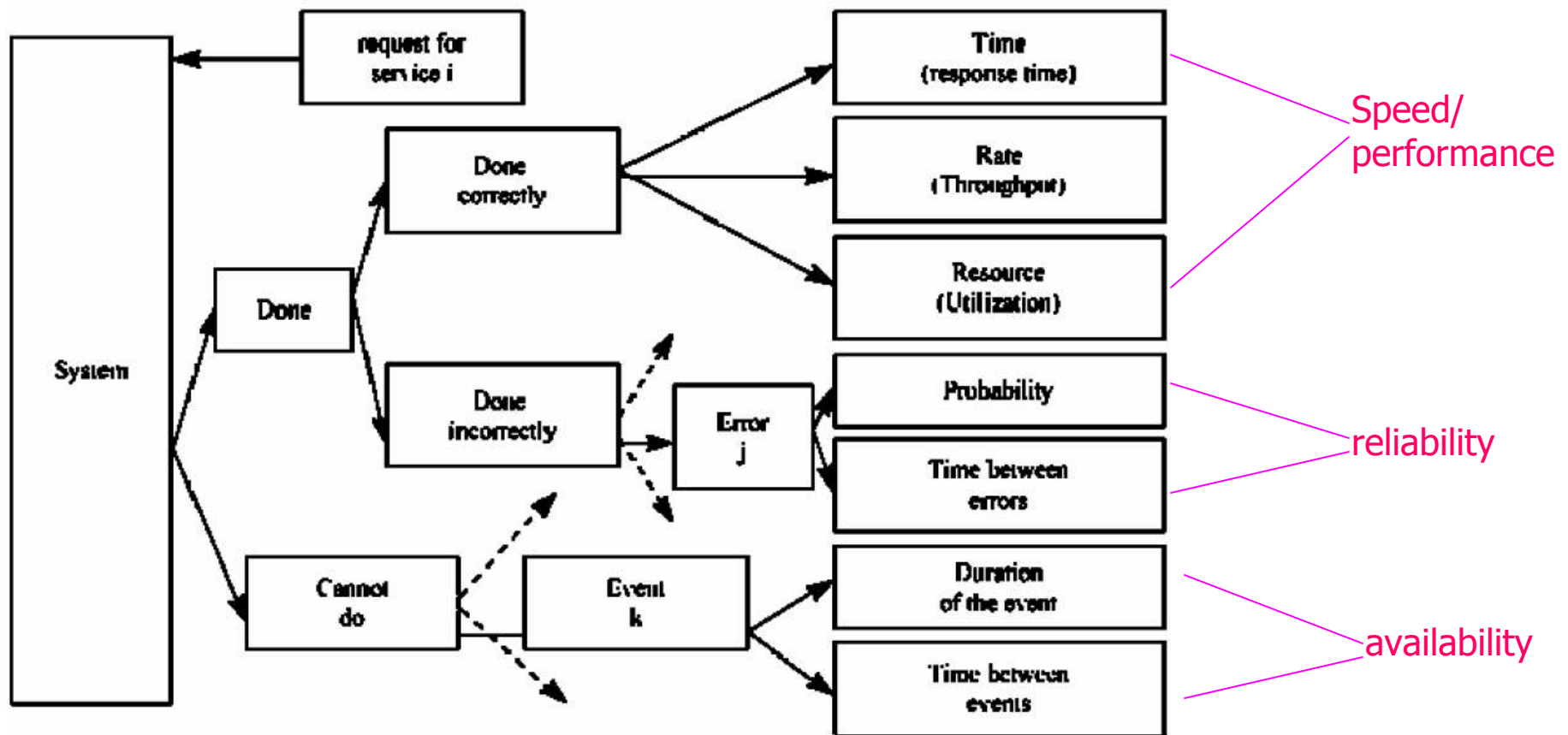
- Until validated, all evaluation results are suspect
 - Do not trust the results of an **analytical model** until they have been validated by a simulation model or measurements
 - Do not trust the results of a **simulation model** until they have been validated by analytical modeling or measurements
 - Do not trust the results of a **measurement** until they have been validated by simulation or analytical modeling

Outline

- Selecting Evaluation Techniques
- **Selecting Performance Metrics**
- Commonly Used Performance Metrics
- Setting Performance Requirements

Selecting performance metrics

- Depend on services offered by the system; for each service request ...



Selecting metrics (contd.)

- Include:
 - Speed/performance: Time, Rate, & Resource
 - Reliability: Error rate & probability
 - Availability: Time to failure & duration
- Consider including:
 - Mean and variance
 - Individual and Global metrics
- Selection Criteria:
 - Low-variability: so that few # of experiments are required
 - Non-redundancy
 - Completeness: reflect all possible outcomes

Case study: compare two congestion control algorithms

- Service: Send packets from specified source to specified destination in order and without duplicate
- Possible outcomes:
 - Some packets are delivered *in order* to the correct destination
 - Some packets are delivered *out-of-order* to the destination
 - Some packets are delivered more than once (duplicates)
 - Some packets are dropped on the way (lost packets)

Case study (contd.)

- Performance: For packets delivered in order,
 - Time-rate-resource
 - Response time to deliver the packets (delay inside the network)
 - Throughput: the number of packets per unit of time
 - Resource: processor time per packet on the source, the destination, and the intermediate systems
 - Variability of the response time
 - Highly variant response time can result in unnecessary retransmissions

Case study (contd.)

- Probability of out-of-order arrivals
 - Out-of-order packets consume buffers
- Probability of duplicate packets
 - Duplicate packets consume the network resources
- Probability of lost packets
 - Lost packets require retransmission
- Probability of disconnection
 - Disconnection interrupt service

Case study (contd.)

- Fairness: because resources are shared
 - Fairness index

$$f(x_1, x_2, \dots, x_n) = \frac{(\sum_{i=1}^n x_i)^2}{n \sum_{i=1}^n x_i^2}$$

- Fairness Index Properties
 - Always lies between 0 and 1
 - Equal throughput: fairness index is 1
 - If k of n receive x and $n-k$ users receive zero throughput: fairness index is k/n

Case study (contd.)

- Throughput and delay were found redundant =>
Use Power = throughput/response-time
- Variance in response time redundant with the probability of *duplication* and the probability of *disconnection* => drop "variance in response time"

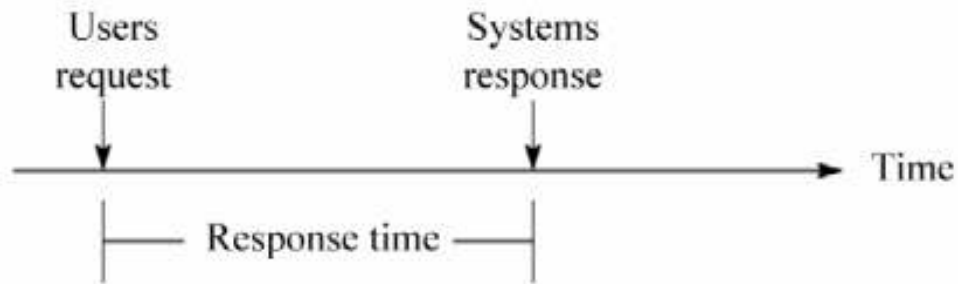
Outline

- Selecting Evaluation Techniques
- Selecting Performance Metrics
- **Commonly Used Performance Metrics**
- Setting Performance Requirements

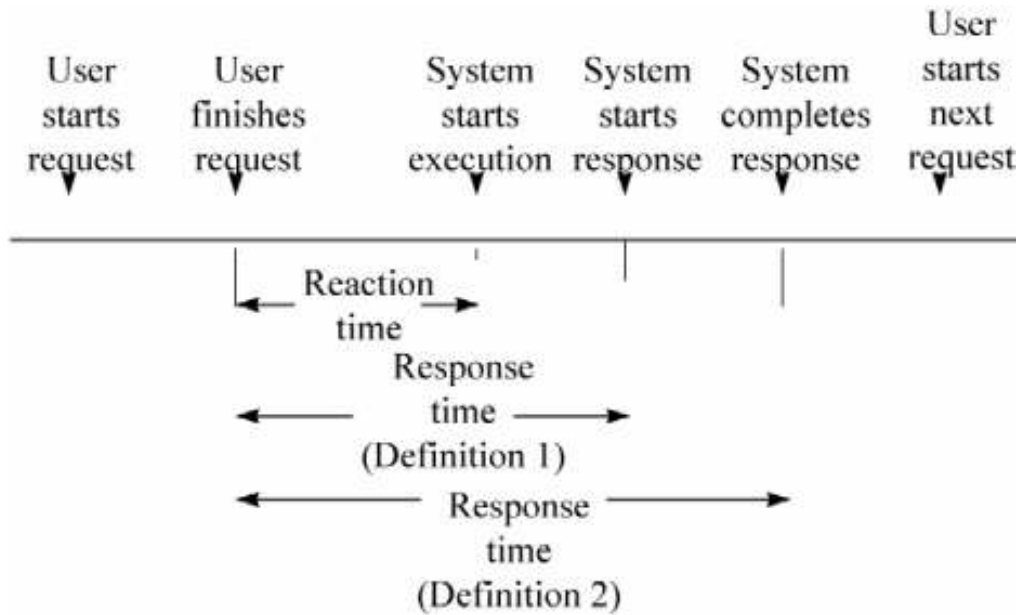
Commonly used performance metrics

- Response time
- Throughput
- Utilization
- Reliability
- Availability

Response time



Simple definition



Precise definition

Response time (contd.)

- Turnaround time
 - the time between the submission of a batch job and the completion of its output.
- Stretch Factor
 - the ratio of the response time with multiprogramming to that without multiprogramming

Throughput (rate)

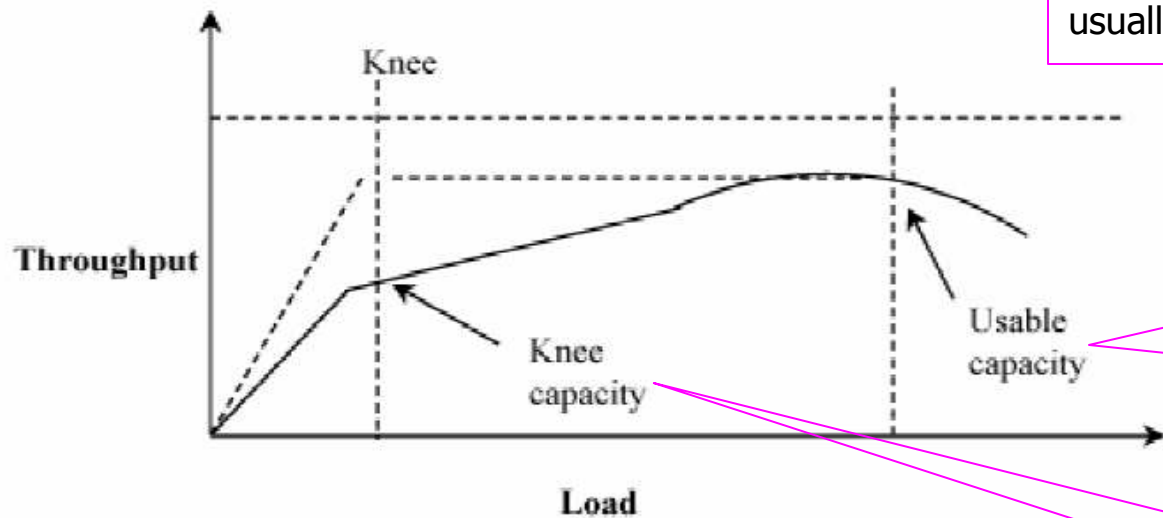
- Requests served per unit of time
- Examples:
 - Packets Per Second (PPS)
 - Bits per second (bps) Jobs per second
 - Millions of Instructions Per Second (MIPS)
 - Millions of Floating Point Operations Per Second (MFLOPS)
 - Requests, jobs, or transactions per second

Throughput (contd.)

- Throughput generally increases as load initially increases, but will stop increasing after certain threshold load
- As load increases, response time usually increases

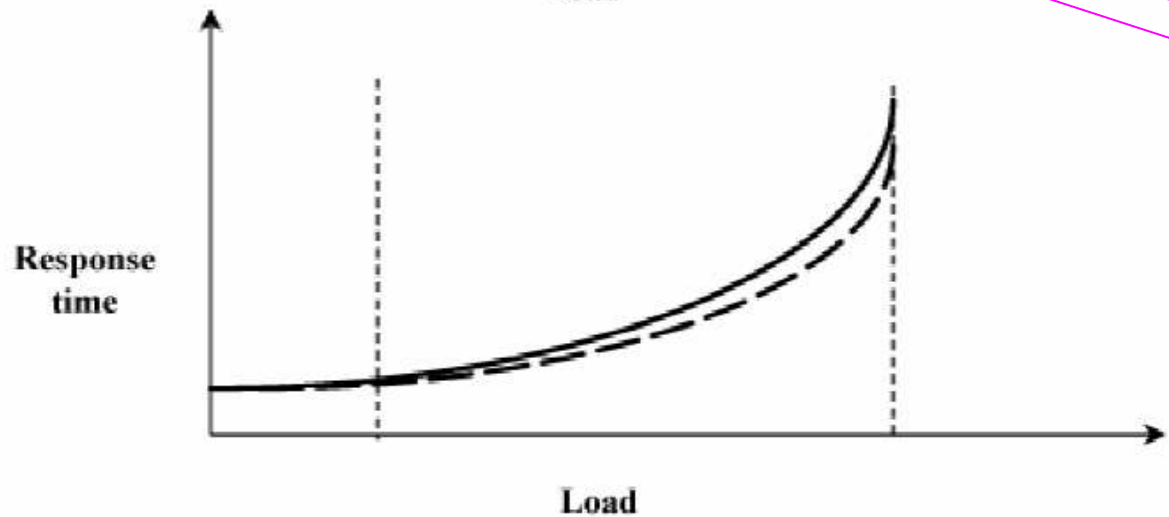
Throughput (contd.)

- Maximum achievable throughput under ideal workload conditions. E.g., bandwidth in bits per second
- The response time at maximum throughput is usually too high



Maximum throughput achievable without exceeding a pre-specified response-time limit

Knee = Low response time and High throughput



Utilization, reliability, availability

- Utilization

- the fraction of time the resource is busy servicing requests; or for memory, average fraction that is used

- Reliability

- Probability of errors
- Mean time between errors (error-free seconds)

- Availability

- Mean Time to Failure (MTTF)
- Mean Time to Repair (MTTR)
- $= \text{MTTF} / (\text{MTTF} + \text{MTTR})$

Utility classification of metrics

- Higher is better (HB)
 - E.g., throughput
- Lower is better (LB)
 - E.g., response time
- Nominal is better (NB)
 - E.g., Utilization (too high utilization leads to too high response time)

Outline

- Selecting Evaluation Techniques
- Selecting Performance Metrics
- Commonly Used Performance Metrics
- **Setting Performance Requirements**

Setting performance requirements

- Examples

- “The system should be both processing and memory efficient. It should not create excessive overhead”
- “There should be an extremely low probability that the network will duplicate a packet, deliver a packet to the wrong destination, or change the data in a packet.”

- Problems

- Non-Specific
- Non-Measurable
- Non-Acceptable
- Non-Realizable
- Non-Thorough

=> SMART: specific, measurable, acceptable & realizable, thorough

Case study: LAN

- Service
 - Send frame to D
- Outcomes
 - Frame is correctly delivered to D
 - Incorrectly delivered
 - Not delivered at all
- Requirements on **Speed**
 - The access delay at any station should be less than one second.
 - Sustained throughput must be at least 80 Mbits/sec

Case study (contd.)

- Requirements on **Reliability**

- Five different *error modes*

- Different amount of damage
 - Different level of acceptability

- Corresponding requirements

- The probability of any *bit being in error* must be less than 1E-7
 - The probability of any *frame being in error* (with error indication set) must be less than 1%
 - The probability of a *frame in error being delivered without error indication* must be less than 1E-15
 - The probability of a *frame being misdelivered* due to an undetected error in the destination address must be less than 1E-18
 - The probability of a frame being delivered more than once (*duplicate*) must be less than 1E-5
 - The probability of *losing a frame* on the LAN (due to all sorts of errors) must be less than 1%

Case study (contd.)

- Requirements on **Availability**
 - Two fault modes: Network reinitializations, permanent failures
 - Corresponding requirements
 - The mean time to initialize the LAN must be less than 15 milliseconds
 - The mean time between LAN initializations must be at least one minute
 - The mean time to repair a LAN must be less than one hour. (LAN partitions may be operational during this period.)
 - The mean time between LAN partitioning must be at least one-half a week

Summary

- Selecting Evaluation Techniques
- Selecting Performance Metrics
- Commonly Used Performance Metrics
- Setting Performance Requirements

Exercise

- Make a complete list of metrics to compare
 - Two network topologies
 - Two MAC protocols
 - Two routing protocols
 - Two transport protocols