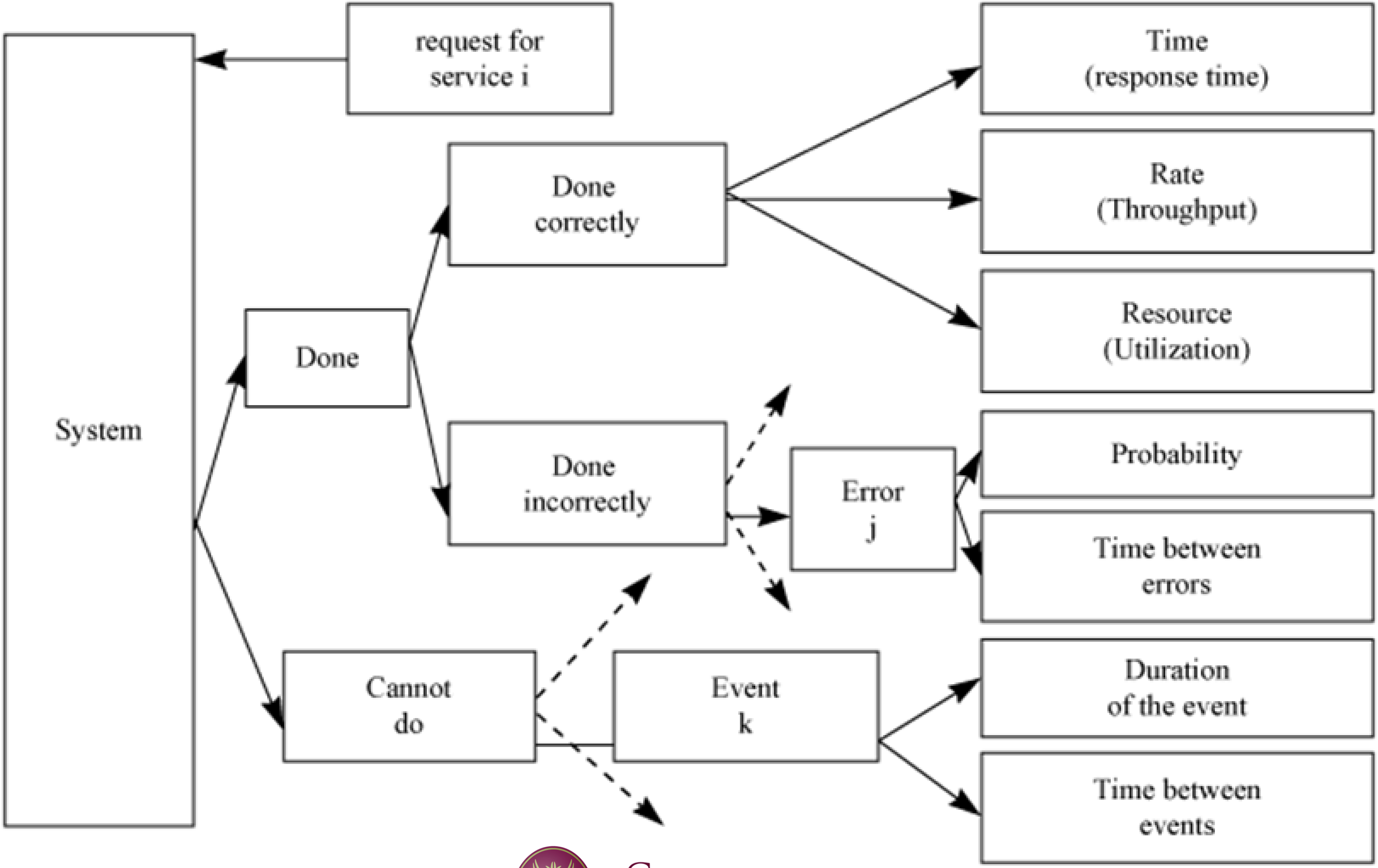
SELECTION OF METRICS (CONT)

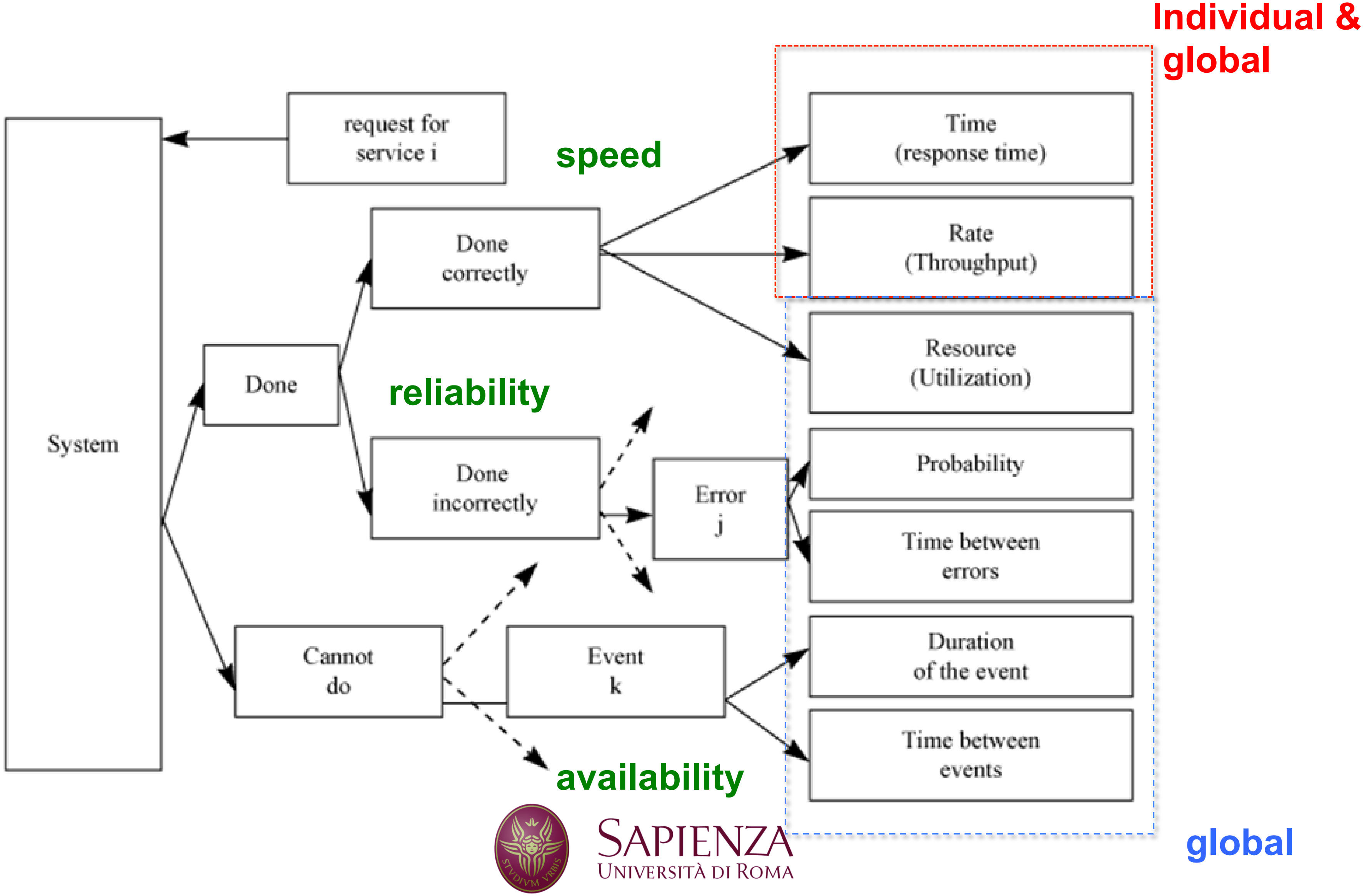
Gaia Maselli
maselli@di.uniroma1.it

SAPIENZA
UNIVERSITÀ DI ROMA

Selecting performance metrics



Selecting performance metrics



SAPIENZA
UNIVERSITÀ DI ROMA

Selecting performance metrics

- Include (by applying the procedure for metric selection):
  - Time, Rate, Resource
  - Error rate, probability
  - Time to failure and duration
- Consider including:
  - Mean and variance
  - Individual and global
- Selection criteria:
  - Low-variability
  - Non-redundancy
  - Completeness

Case study: two congestion control algorithms

- A **network** is composed by a number of *end systems interconnected via intermediate systems*. The end systems *send packets* to other end systems on the network. Intermediate systems forward the packets along the right path. Congestion occurs when the number of packets waiting at an intermediate node exceeds the node's buffering capacity
- **Service**: send packets from specified source to specified destination in order (packet forwarding)
- Possible **outcomes**:
  - some packets are delivered in order to the correct destination
  - Some packet 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 (cont)

- Performance: for packets delivered in order
  - Time-rate-resource ➔
    1. Response time to deliver packets (the time inside the network for individual packets)
    2. Throughput: the number of packets per unit of time
    3. Processor time per packets on the source end system
    4. Processor time per packets on the destination end system
    5. Processor time per packets on intermediate systems
  - Variability of response time ➔ retransmissions
    - Response time: the delay inside the network
    6. Variance of response time

Case study (cont)

- Out-of-order packets consume buffers

7. Probability of out-of-order arrivals

- Duplicate packets consume the network resources

8. Probability of duplicate packets

- Lost packets require retransmission

9. Probability of lost packets

- Too much loss cause excessive retransmissions disconnection

10. Probability of disconnect

- The network is a multiuser system: all users have to be treated fairly

10. Fairness: for any given set of user throughputs

Case study (cont)

- Shared resource ➔fairness

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

- Fairness index properties
  - Always lies between 0 and 1
  - Equal throughput ➔ fairness =1
  - If *k* of *n* receive *x* and *n*-*k* users receive zero throughput: the fairness index is *k/n*

Example on fairness

Suppose one is asked to distribute 20 dollars among 100 persons

Case 1: Give 20 cents to each of the 100 persons

$$x_i = 0.2 \quad i = 1, 2, \cdots, 100$$

$$Fairness\ Index = 1.0$$

➔Scheme is totally fair

Case 2: discrimination criteria

$$x_i = \begin{cases} 2 & i = 1, 2, \cdots 10 \\ 0 & i = 11, 12, \cdots 100 \end{cases}$$

$$Fairness\ index = 0.10$$

➔ Scheme is only 10% fair

Case study (cont)

Changes on the metric set

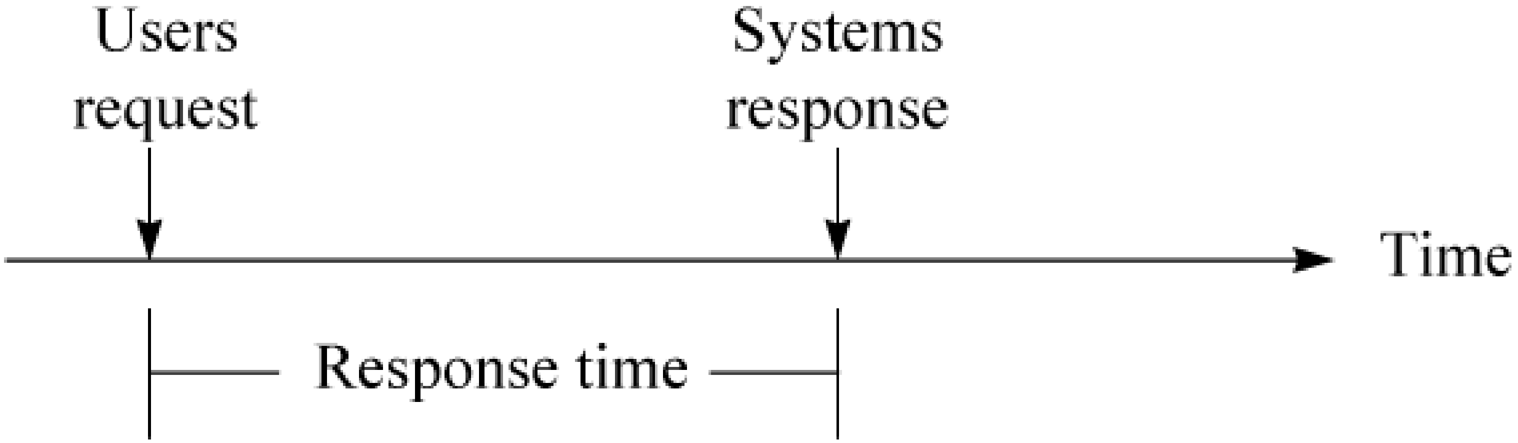
- Often it happens that *throug*hput and *delay* are redundant metrics
  - The schemes that result in higher throughput also result in higher delay
  - The two metrics may be removed and combined in a single metric called **power**, which is defined as the ratio of throughput to response time

$$\text{Power} = \frac{\text{Throughput}}{\text{Response Time}}$$

- The variance in response time can also be dropped since it is redundant with the probability of duplication and the probability of disconnection
  - A higher variance results in a higher probability of duplications and a higher probability of premature disconnection

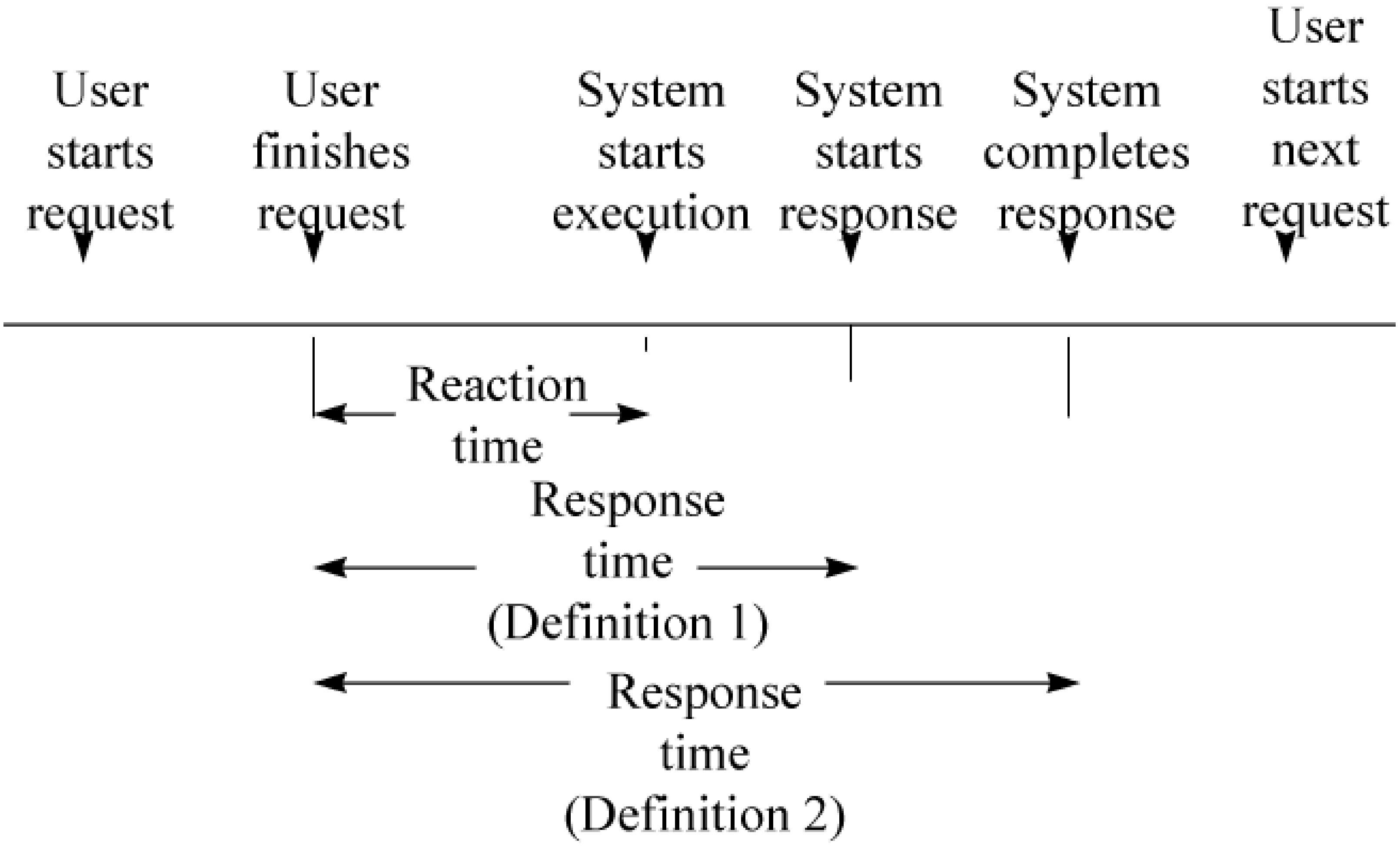
Commonly used performance metrics

Response time and **Reaction Time**

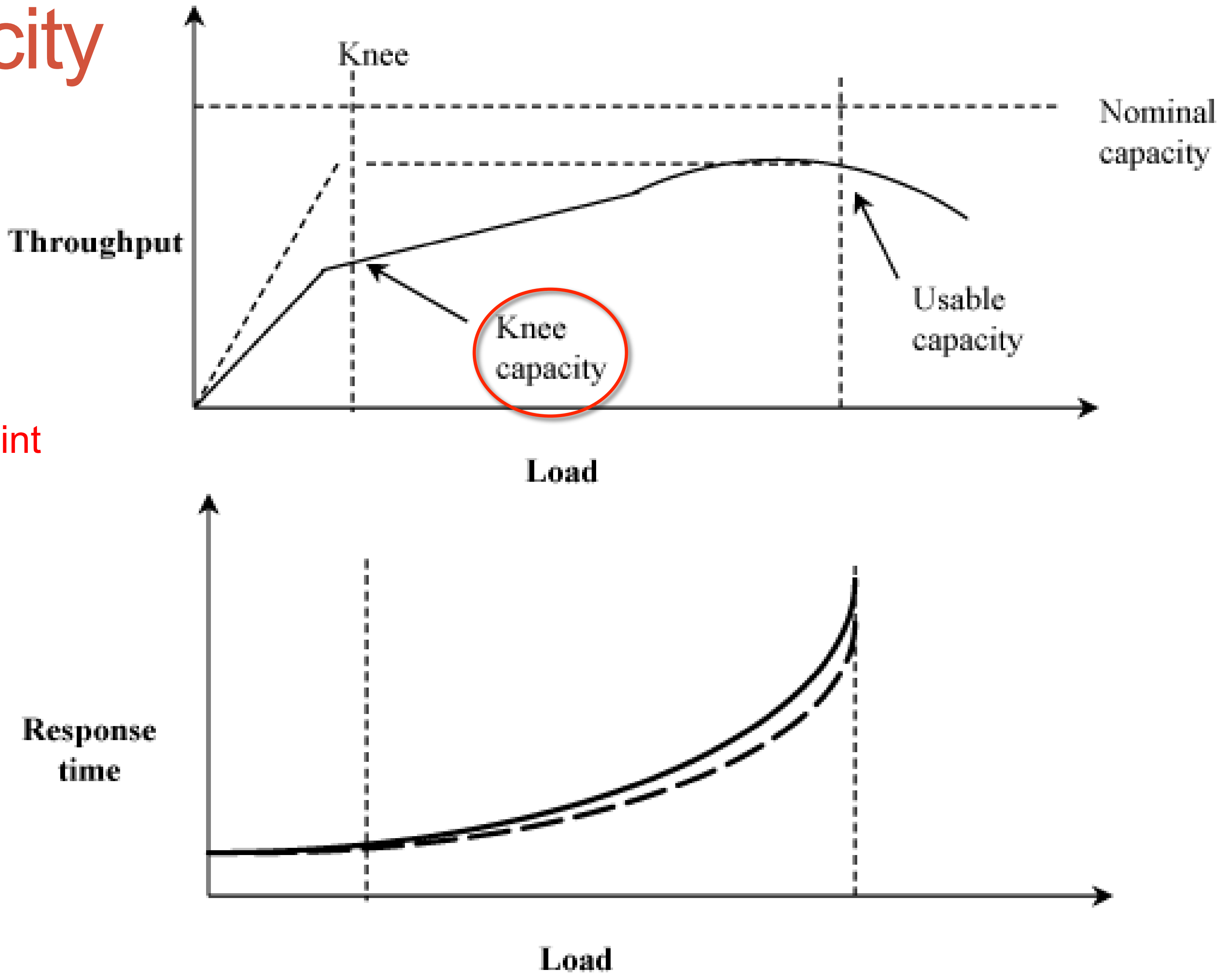


SAPIENZA
UNIVERSITÀ DI ROMA

Response Time (cont)



Capacity



Optimal
operating point

Common Performance Metrics (cont)

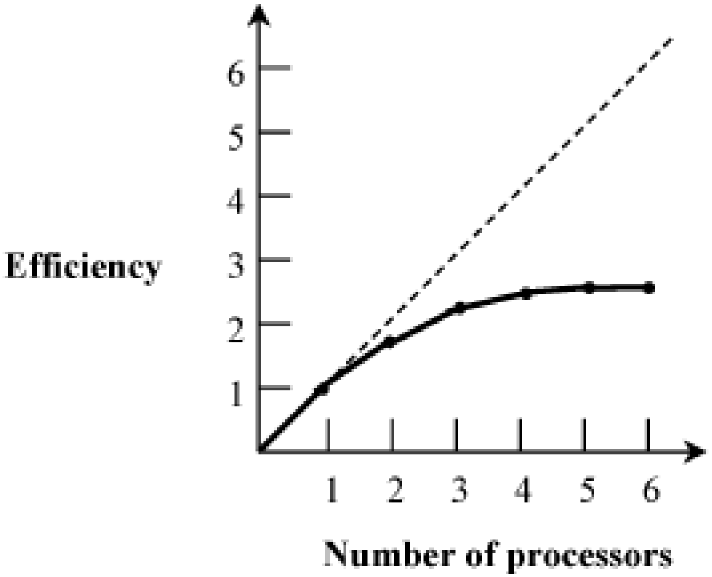
- **Nominal Capacity**: Maximum achievable throughput under ideal workload conditions. E.g., bandwidth in bits per seconds. The response time at maximum throughput is too high

- **Usable capacity**: Maximum throughput achievable without exciding a pre-specified response-time limit

- **Knee capacity**: Knee= low response time and high throughput

Common performance metrics

- **Efficiency**: Ratio of usable capacity to nominal capacity
- **Utilization**: The fraction of time the resource is busy servicing requests. Average fraction used for memory



Commonly used performance networks

Let us identify *some* common metrics in computer networks

Commonly used performance metrics

Response time: interval between a user's request and the system response

In computer networks:

Packet delay: the delay experienced by a packet as it passes through the network, given by the sum of

1. *Routing delay (processing + queuing)*: time a packet spends inside a router (time between the arrival of the trailing bit at the router and the moment the first bit of the packet is placed on the output link)
2. *Transmission delay*: time required to place a packet onto a link (packetSize/rate)
3. *Propagation delay*: time required for a packet to pass from one end of a link to the other end (distance/propSpeed)

Commonly used performance metrics

Throughput: rate at which the requests are serviced by the system

In computer networks:

Throughput: the rate at which traffic flow through the network.

- In a given interval T, the throughput is calculated as the number of packets the pass at a point *without loss* over time T, and is measured in packets per seconds (**pps**) or bits per second (**bps**)
- Throughput over a sequence of hops is determined by the element with minimum available capacity
- The maximum achievable throughput under ideal workload conditions is called **nominal capacity** of the system and corresponds to the **bandwidth**
- The ratio of maximum achievable throughput (usable capacity) to nominal capacity is called **efficiency**. Example: if the maximum throughput from a 100Mbps LAN is 85 Mbps, its efficiency is 85%

Commonly used performance metrics

Reliability: probability of errors or the mean time between errors

In computer networks:

Packet Loss: fraction of packets lost in a time period

Relative Packet loss rate: if C_n is the number of packets entering a network element in time period n, and L_n is the number of packets lost during that time period, the relative packet loss rate can be estimated as L_n/C_n

Commonly used performance metrics

A metric related to throughput and packet loss

Goodput: the rate at which the application endpoint successfully receives data

- The rate at which TCP sends packets is the load it places on the network per unit of time
- Packets being sent may be retransmissions
- Bytes are not necessarily delivered to the application at the rate that the connection sends them

Utility classification of performance metrics

- Three classes depending upon the utility function of a performance metric

1. **Higher is better** (**HB**): systems users and systems managers prefer higher values of such metrics
   - throughput is an HB metric
2. **Lower is better (LB**): systems users and systems managers prefer lower values of such metrics
   - packet delay is an HB metric
3. **Nominal is best** (**NB**): Both high and low values are undesirable. A particular value in the middle is considered best
   - Utilization

Setting performance requirements

- Specifying performance requirements for a system is often a problem
- Typical requirements
  - The system should be both processing and memory efficient
  - There should be an extremely low probability that the network will duplicate a packet

➔unacceptable !!!

Setting Performance Requirements

Problems:

- *Non-specific*: no clear numbers are specified
- *Non-measur*able: there is no way to measure a system and verify that it meets the requirements
- Non-acceptable: non realistic
- Non-realizable: too high to be realizable
- *Non-thorough*: no attempt is made to specify all the possible requirements

- Requirements must be SMART
  - Specific
  - Measurable
  - Acceptable
  - Realizable
  - Thorough