Internet Services Performance Part2

prof. Giuseppe Iazeolla

Part2 objectives

- Provide tools for the analysis and design of quality of service (QoS) in internet technology systems, thus
- Internet
- Intranet
- Extranet

Part2 development

- 4 sections
- 1) *Internet model building* or methods for the production of the internet model and the relative fixed and mobile access networks

- 2) *Internet workload characterization* or methods for the representation and analysis of the various types of services and internet traffic

Cont.nd

- 3) *Internet model evaluation* or methods for QoS evaluation (effectiveness, availability, dependability etc.) of the system in the analysis, re-engineering or design stage.

- 4) Experimenting and development of generation tools and evaluation of platform models and internet traffic

When to look at QoS

In the phase of

- Design (engineering)
- Re-engineering
- Governance (management)

Quality vision

- IEEE Definition
- *quality is the degree in which the system possesses a desired combination of* ***attributes***
- Viewpoints:
  - Transcendent
    - innate excellence
  - User
    - fit for use (meet user needs), enjoyable to use, user satisfaction
  - Product
    - desired attributes of software (reliability, correctness, etc.)
    - conformance to requirements
  - Organization
    - costs and profits (increased efficiency, increased effectiveness, added value to organization, marketable product)

Formal definition of quality in systems

- A **vector** of 12 components or quality indices:

$$q = (i_1, i_2, i_3, i_4, i_5, i_6, i_7, i_8, i_9, i_{10}, i_{11}, i_{12})$$

- Its evaluation verifies product ability to meet the desired user goals

quality **indices** (or factors)

i_1 correctness
i_2 reliability
i_3 efficiency
i_4 integrity
i_5 usability

i_6 maintainability
i_7 flexibility
i_8 testability

i_9 portability
i_{10} reusability
i_{11} interoperability
i_{12} evolubility

Importance of quality indices in networks and services

- They are not so visible but they determine the quality of social life and control
    - Other plants
        - nuclear
        - electrical
        - telephone
        - water
        - environmental
        - etc.

Importance of quality indices in networks and services

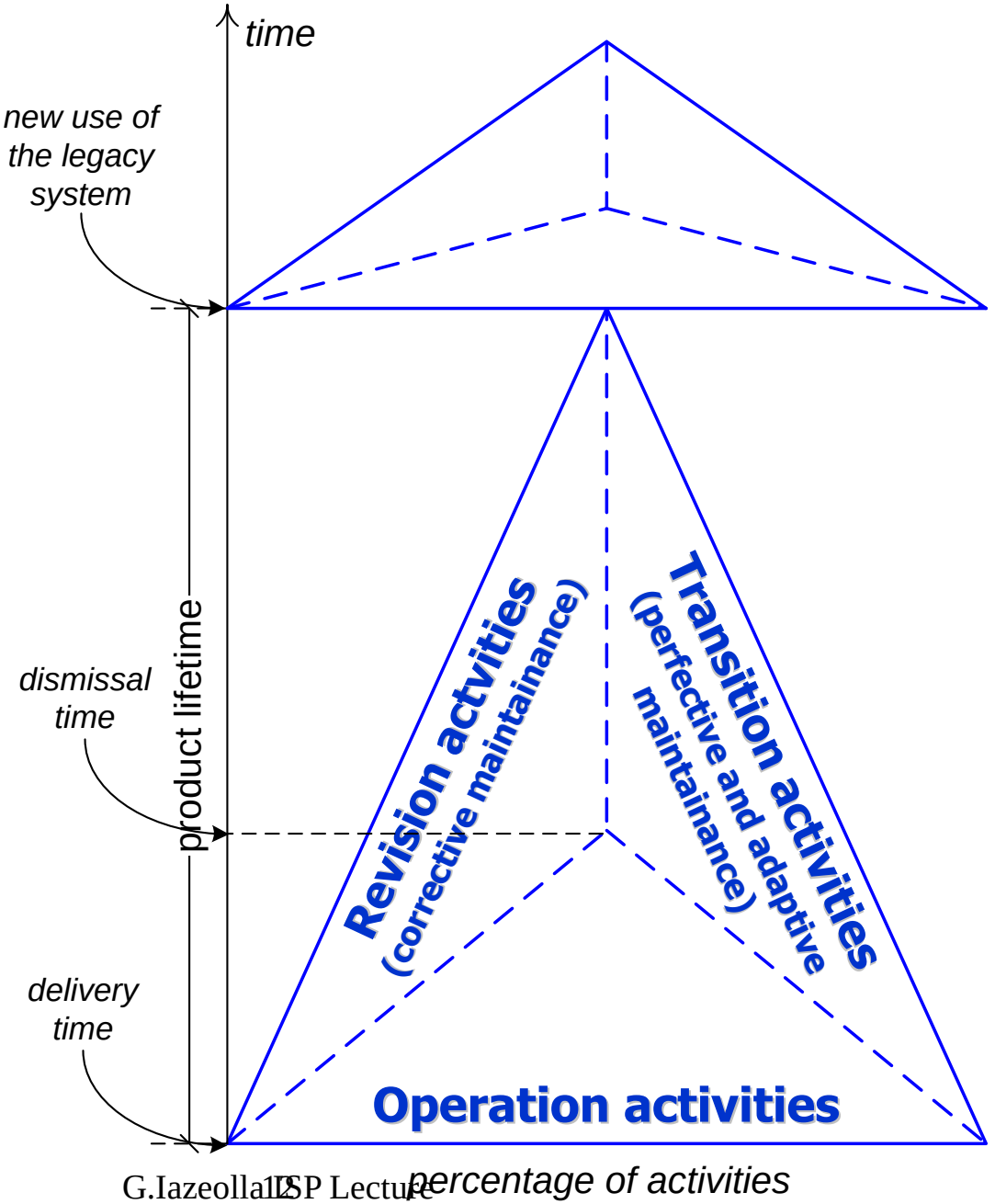
- They are the basis of banking systems
  - production
  - distribution
  - transport
  - traffic (air, land etc.)
  - environmental
  - healthcare
  - medical therapy
  - etc.

Importance of quality indices in networks and services

- Improper behavior is the basis of
  - Nuclear accidents
  - System black-outs
    - telephone
    - Energy distribution
  - Management errors
    - Medical equipment
  - System inefficiencies
    - booking
    - Air traffic control etc.

The Quality Triangle

(McCall quality model)



Quality **Indices** in operation activities

- **i_1 Correctness**
  - extent to which the system satisfies spec & fulfills user objectives (does it do what I want?)
- **i_2 Reliability**
  - extent to which the system can be expected to perform its intended function with required precision (does it do it accurately all of the time?)
- **i_3 Efficiency**
  - amount of computing resources and code required by the system to perform a function (will it run on my hardware as well as it can?)
- **i_4 Integrity**
  - extent to which access to system or data by unauthorized persons can be controlled (is it secure?)
- **i_5 Usability**
  - effort required to learn, operate, prepare input and interpret output of a system (can I run it?)

Quality **Indices** in revision activities

i_6 Maintainability

- effort required to locate and fix a defect in the system (can I fix it?)

i_7 Flexibility

- effort required to modify the system (can I change it?)

i_8 Testability

- effort required to test the system to ensure that it performs its intended function (can I test it?)

Quality **Indices** in transition activities

i_9 Portability

- effort required to transfer a software product from one hardware and/or software environment to another (will I be able to use it on another machine?)

i_{10} Reusability

- extent to which a product (or parts thereof) can be reused in other applications (will I be able to reuse some of the software?)

i_{11} Interoperability

- effort required to couple one product with another (will I be able to interface it with another system?)

i_{12} Evolubility

- effort required to update the product in order to fulfill new requirements (is it easy to update when requirements change?)

Summary of Quality **Indices** (or factors)

Operation activities

i_1 correctness
i_2 reliability
i_3 efficiency
i_4 integrity
i_5 usability

Revision activities (corrective maintenance)

i_6 maintainability
i_7 flexibility
i_8 testability

Transition activities (perfective and adaptive maintenance)

i_9 portability
i_{10} reusability
i_{11} interoperability
i_{12} evolubility

Quality **Attributes**

- System quality

$$q = (i_1, i_2, i_3, i_4, i_5, i_6, i_7, i_8, i_9, i_{10}, i_{11}, i_{12})$$

- Each quality index is a **vector** of quality attributes

$$i_j = (a_1, a_2, ..., a_n)$$

- Each attribute is a **vector** of sub-attributes, which are **measured** in order to obtain the attribute value
- Attribute values contribute to quantify the relevant indices
- The evaluation of quality indices allows to verify the system ability to meet the desired levels of quality

Quality attributes

a_1 Complexity

level of understandability and verifiability of elements of the system and their interactions.

a_2 Accuracy

precision of computations and output

a_3 Completeness

full implementation of the required functionalities

a_4 Consistency

use of uniform design and implementation techniques and notations

a_5 Error (fault) tolerance

continuity of operation ensured under adverse conditions

a_6 Traceability

degree to which a relationship can be established between two or more products of the development process

a_7 Expandability

storage or functions can be expanded

a_8 Generality

breadth of potential applications

a_9 Modularity

provisions of highly independent modules

a_{10} Auto-documentation

in-line docs

Impact of attributes on quality indices
(+/- = positive/negative impact)

i_1 Correctness (operations)
- \+ Completeness
- \+ Consistency
- \+ Traceability

i_7 Flexibility (revisions)
- \+ Traceability
- \+ Consistency
- \- Complexity
- \+ Modularity
- \+ Generality
- \+ Auto-documentation

$$i_1 = (a_3, a_4, a_6)$$

$$i_7 = (a_1, a_4, a_6, a_8, a_9, a_{10})$$

i_2 Reliability (operations)
- \+ Error Tolerance
- \+ Consistency
- \+ Accuracy
- \- Complexity

i_{12} Evolubility(transitions)
- \+ Consistency
- \- Complexity
- \+ Modularity
- \+ Expandability
- \+ Generality
- \+ Auto-documentation

$$i_2 = (a_1, a_2, a_4, a_5)$$

$$i_{12} = (a_1, a_4, a_7, a_8, a_9, a_{10})$$

Quality vision in computing networks and services

Vector

$$Q = (f_1, f_2, \ldots, f_n)$$

$\mathbf{f_i}$ = quality ***factor*** (also called quality ***index***)

Example: Q = (f1, f2, ... , f6)

f1 = performance

f2 = usability

f3 = flexibility

f4 = scalability

f5 = adequacy

f6 = evolability

Quality factor (or quality index) f_i

$$f_i = (\mathbf{a}_{i1}, \mathbf{a}_{i2}, \ldots\ldots, \mathbf{a}_{in})$$

$\mathbf{a}_{ix}$= i-th ***attribute*** of quality

Quality index of ISP interest:

Performance index =
$f_1 = (a_{11}, a_{12})$ = (effectiveness, predictability)

Quality attribute a_{xx}

- $a_{xx} = (\mathbf{s}_{xx1}, \ldots\ldots, \mathbf{s}_{xxn})$
- $\mathbf{s_{xxx}}$ = ***sub-attribute*** of quality

Example

- a_{11} = **effectiveness** = $(s_{111}, \ldots, s_{11n})$ = (waiting time, ..., throughput)
- a_{12} = **predictability** = $(s_{121}, \ldots, s_{12k})$ = =(availability,….., safety)

Quality metric

metric =

= measurements of the last level sub-attribute

Examples

- Waiting time = time unit
- throughput = no. elements/time
- availability = % functioning time
- safety = prob. of non catastrophic events

attributes of index f_1 (performance)

Quality factor	First level attribute	Second level attribute	metrics
performance	**effectiveness**	Waiting time	Time unit
		Response time	Time unit
		Queue length	No. of elements
		space	Space unit
		throughput (capacity)	No. elements/time
		utilization	% (timeper_occ)
	dependability	availability	% (timeper_perf)
		dependability	prob(time_perf)
		safety	prob(no_accid)
		security	prob(no_intrus)

sub-attributes of the efficiency attribute

waiting time

response time

queue length

occupied space

throughput

utilization

metrics of sub-attributes efficiency

of *waiting time, response time* → time unit

(sec, hours, etc.)

of *queue length* → number of elements

(programs, commands, requests in queue)

of *throughput* → no. elements / time

(number of elements served per time unit)

of *utilization* → % (timeper_occ)

percentage of time in a given period (timeper), during which an installation or its component is occupied (occ)

Sub-attributes of the attribute predictability

availability

reliability

safety

security

Sub-attributes of the attribute predictability

of *availability* → % (timeper_perf)

% of time in a given period (timeper) during which an installation maintains its performance (perf)

of *reliability* → prob(time_perf)

probability that an installation maintains its performance (perf) at a given time instance (time)

of *safety* → prob(no_accid)

probability that no catastrophic events occur (accid) while the installation is functioning as requested

of *security* → prob(no_intrus) =

probability that no unwanted access (intrus) occurs while the installation is functioning as requested

Internet network

- = *computer network*
- A *computer network* is: a *distributed computer installation* (also called distributed *configuration* of the installation).
- A distributed computer installation is:

An interconnection network of centralized installations (called nodes or host computers) which satisfies 3 specific *requirements*.

Requirements for the distributed installation configurations

Configuration is distributed if it satisfies three requirements:

- ***requirement 1)*** presence of many central processing units (CPU)

-requirement 2) presence of a "computer network" (**LAN, MAN, WAN**) or combination of these

-requirement 3) presence of a **system software** which manages the installation in order to give it:

1- resource sharing
2- openness
3- concurrency
4- scalability
5- fault tolerance
6- transparency

System software of a distributed installation

A distributed software system installation:

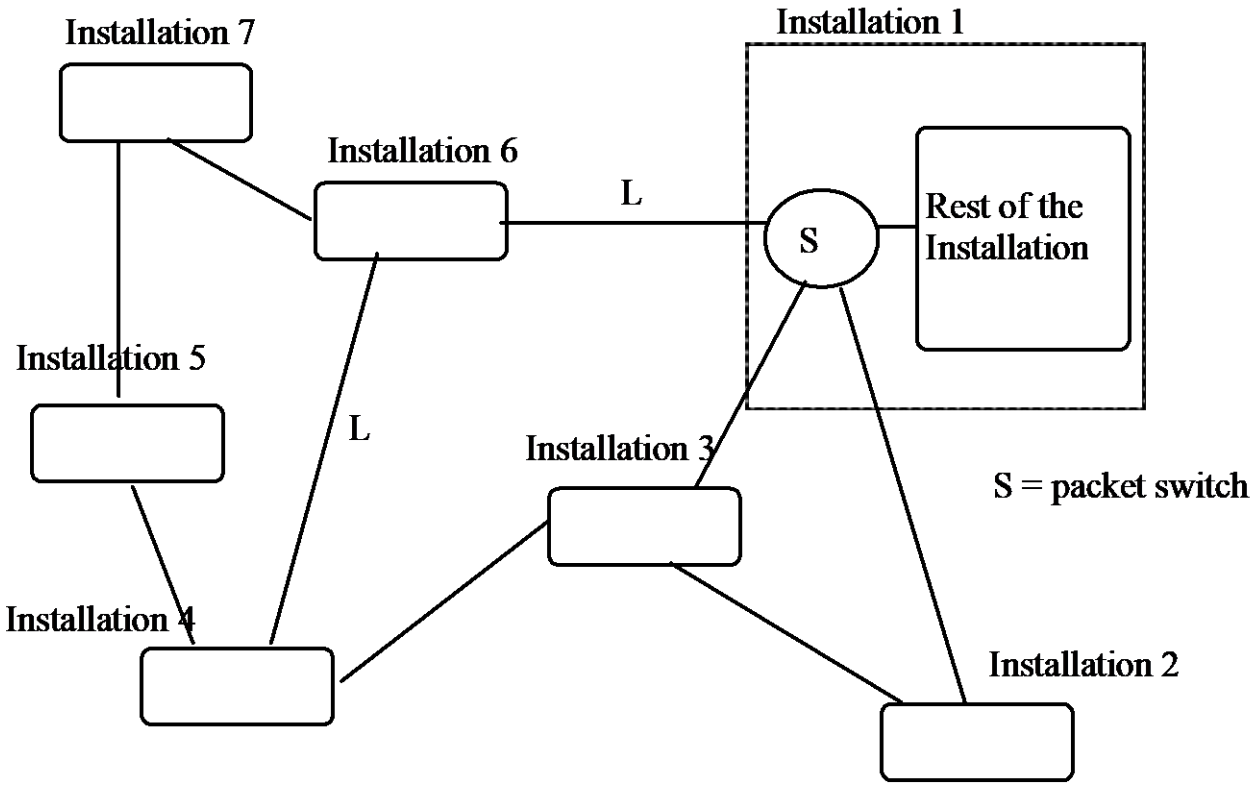
- host nodes ability to co-ordinate activities and share common resources (hardware, software and data).

- Show users a single installation even if made up of several hosts

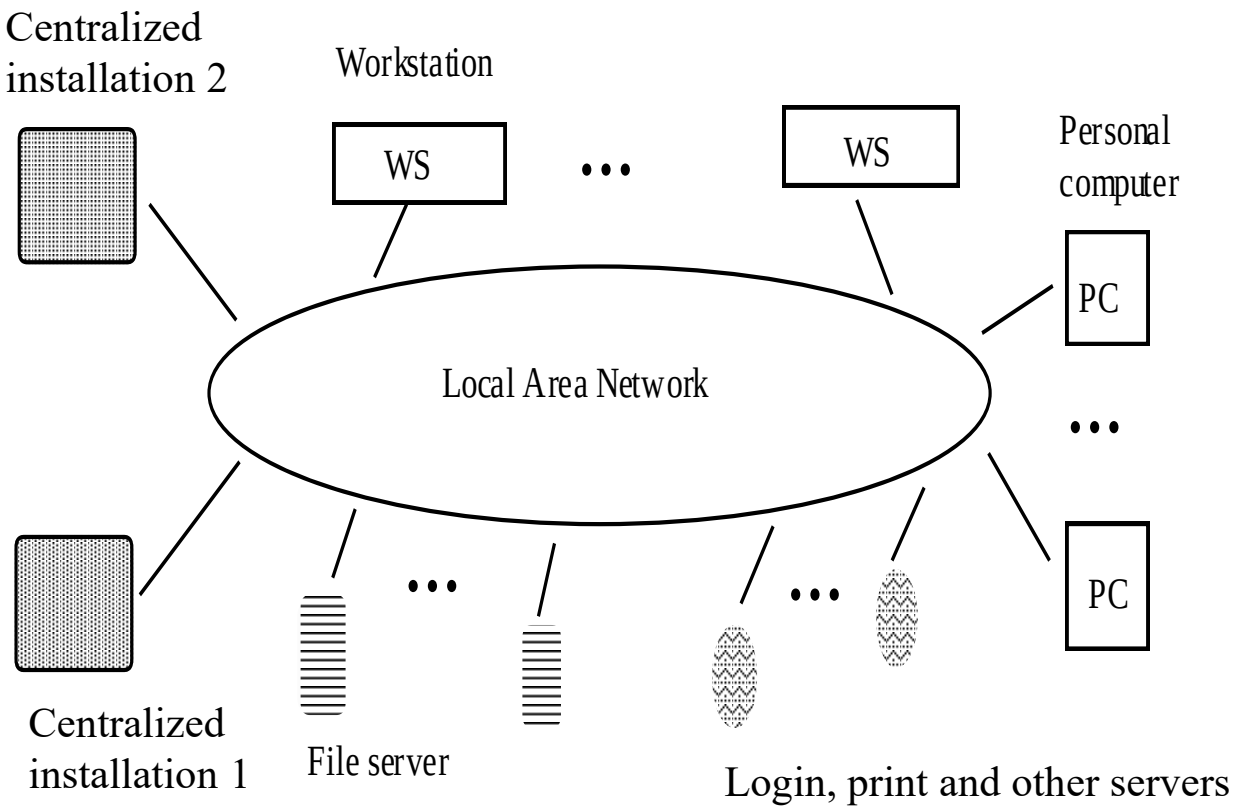
The most important component of the *system software of a distributed installation is the* ***distributed operating system*** (level between VM3 and VM4)

WAN distribution

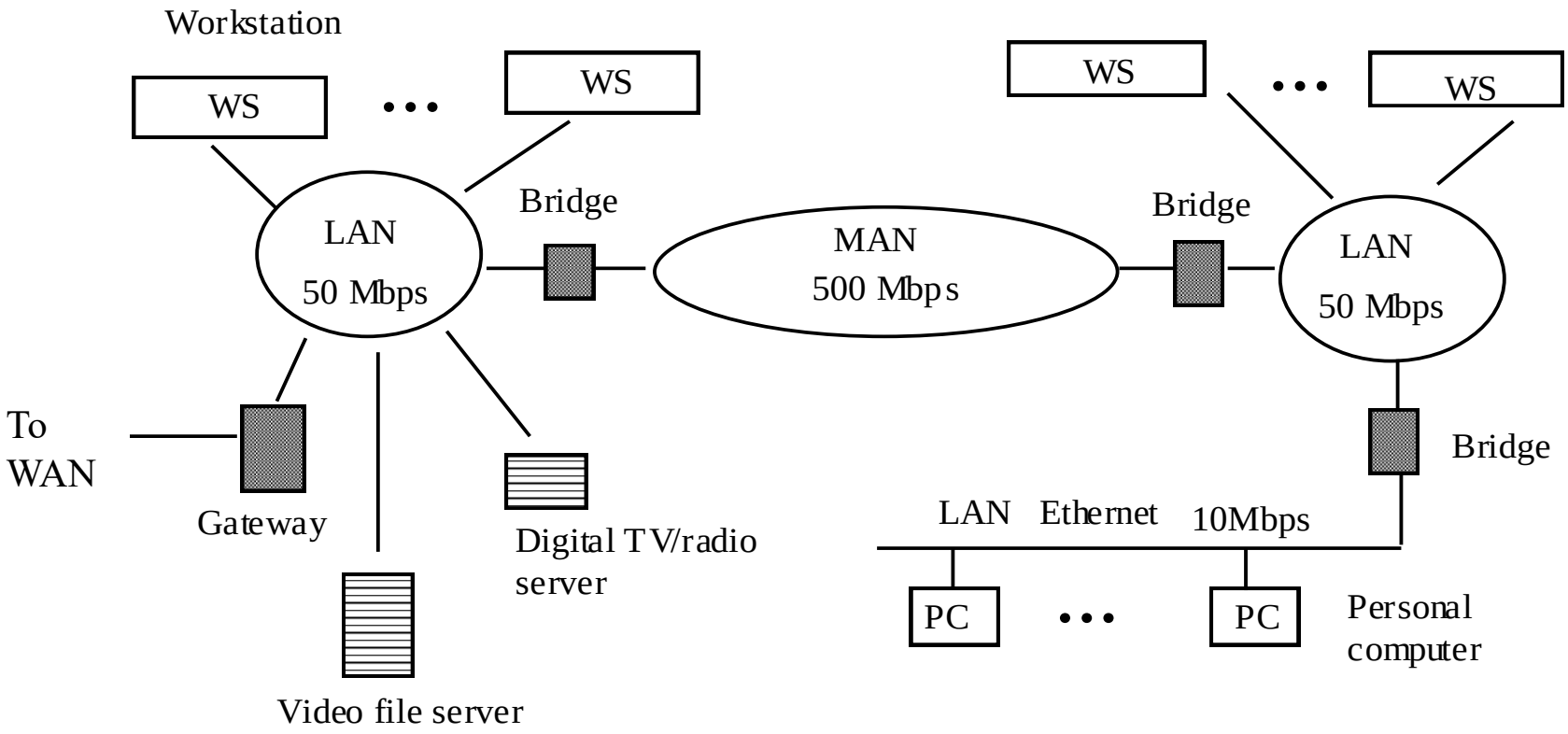
- Geographical networks or WAN



LAN distribution



LAN, MAN and WAN distribution



LAN characterization

- local networks or LAN

- *transmission medium* = coaxial cable or fiber optic
- inside each single building or a series of departmental buildings
- installations of medium-low complexity
- relatively high transmission speed

Network technology

- *Ethernet* (from 10 to 100 megabit/sec)
- *FDDI* (100 megabit/sec)
- *TokenRing IBM* (from 4 to 10 megabit/sec)
- *Apple LocalTalk* (0,23 megabit/sec)

WAN characterization

- Geographical networks or WAN

- *mixed transmission medium*: telephone, coaxial cable, fiber optic, satellite
- connect installations of medium-high complexity localized in different cities, countries or continents
- on average, low transmission speeds

Connected installations called *host computers*

Network technology

ISDN (64K/128K - 2M)

FRN (1.5M - 45M)

ATM (25M/155M - 2.4G)

MAN characterization

- metropolitan networks or MAN

- *transmission medium:* fiber optic

- video, audio and data transfer over distances of up to 50 km

Network technology

ISDN -ATM

Characterization summary
Transmission medium and network technologies for distributed installations

Type of transmission medium	***Max speed (bit/sec)***
Telephone loop (unshielded/shielded)	2M / 100M
Coaxial cable (standard/broadband)	264M / 550M
Satellite/land microwaves	200M
Wireless LAN radio	3.3M
Infrared LAN	4M
Fiber optic cable	320G
Network technology	***Typical max speed (bit/sec)***
Apple local talk	0.23M
Token ring standard	4M-6M
Ethernet standard or fast Token ring	10M-16M
Fast Ethernet	100M-1G
FDDI (fiber distributed data interface)	100M
DDN (digital data network)	2.4K-2M
PSN (packet switching network)	2.4K-64K
FRN (frame relay network)	1.5M-45M
ISDN (integrated services digital network)	64K/128K - 2M
ATM (asynchronous transfer mode)	25M/155M - 2.4G
OC (optical carrier)	52M-10G

Centralized installation

- Historically centralized precede distributed, but defined:

  all of what isn't distributed is centralized

- Installation that *doesn't* respond to *requirement 3*, remains centralized even if *requirement 1* is present (many Pc type units), and *requirement 2* (network presence)

Detail
6 properties (requirement 3) of the distributed installation
(given by the distributed system sw)

1.resource sharing

- Allows use of *resources* residing in any host different from the one it works on

- *Resources* = hardware component (disc, printer etc.)
  or software (application program, data, archives, databases)

-*Resource manager* : offers the user access interface to resources capable of operating from one plant to any other one.

- ***2. openness***

Interfaces made *public*, in a way that new hardware resources are developed or software functions have necessary development documents and makes functions integrated into the systems

- ***3. concurrency***

Capacity of controlling parallels of job steps or tasks

- ***4. scalability***

In operating capacity terms as well as physical

- 5. fault tolerance

Capacity of the installation to behave correctly also in the presence of faults or limit the damage

They are attributed to the installation through two ways,

- use of redundancy hardware components (*hardware redundancy*):

- Design of programs for the recovery of errors (*software recovery*): the intermediate results are saved at periodical time instance (checkpoint times).

- 6. Transparency

Capacity of hiding the distribution and independence of services of the installation, making them perceive them as coming from a set of concentrated components.

- 6.1 ***network transparency***

6.1.1. ***access transparency:*** ensures the uniformity of access rules, or if they don't exist, it simulates by hiding the diversities to the user.

6.1.2. ***location transparency:*** allows access to information objects without knowing when they are connected.

- ***6.2 concurrency transparency:*** invisibility of competition of sharing common resources.

- *transparency (cont.)*

- ***6.3 replication transparency:*** hides redundant components

- ***6.4 failure transparency:*** transparency of faults to the hw/sw unit and their replacement

- ***6.5 migration transparency:*** transparency of the movement of informative objects from one site to another of the installation

- ***6.6 performance transparency:*** transparency of reconfiguration of the installation to improve performance

- ***6.7 scaling transparency:*** transparency of broadening operations

Characterization of the interconnection networks

Main performance attributes of a network:

bandwidth (physical and logic)
latency time

Depend on:

- physical circuit characteristics (physical and topological bandwidth)
- connecting mode = transfer characteristics among nodes

  - Operating mode
  - *Control strategy*
  - *switching method*
  - *Communication protocol*

Bandwidth

Bandwidth: Transmission speed or band width (bit/sec), which can be reached in the data transmission of data among units

Performance index of throughput, or **capacity** (from a few thousand bits/sec up to tens or hundreds of Mbit/sec)

physical bandwidth = speed allowed between two nodes from the physical circuit

Logic bandwidth = effective transmission speed obtained in the practice between two nodes (< of physical one, keeping in mind the competition delays)

Latency Time

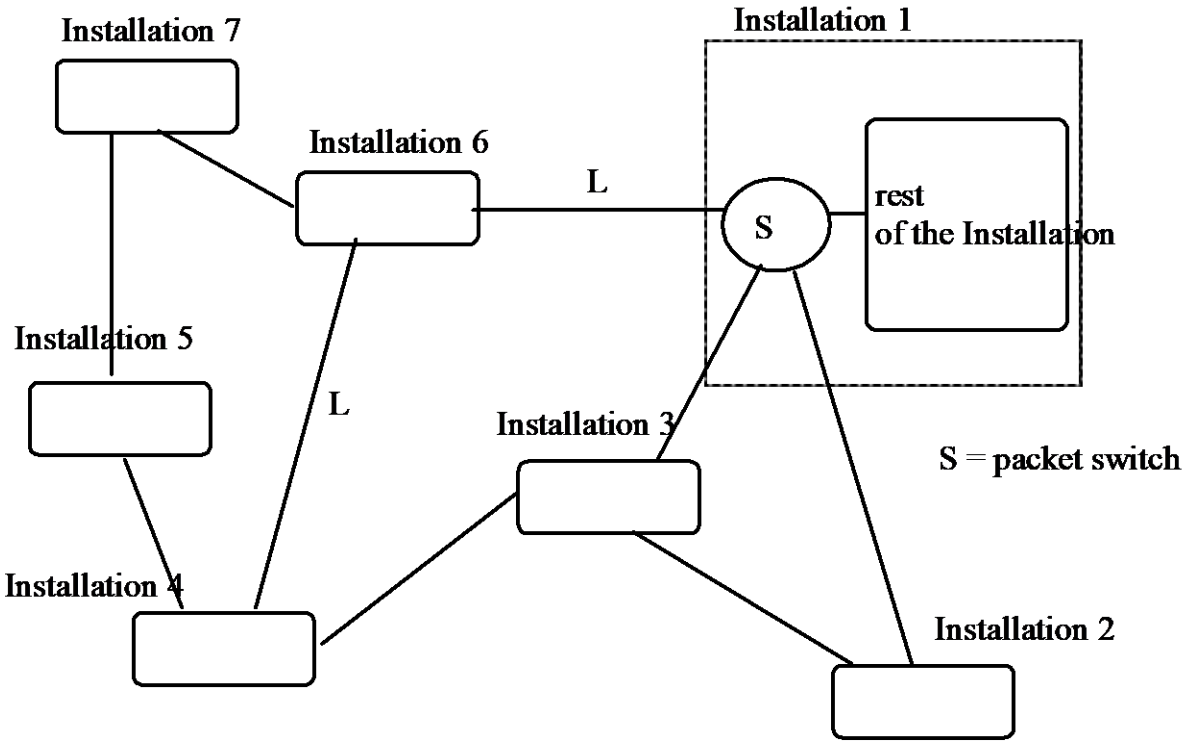
***latency time*: average access time to a common base of information blocks (i.b.).**

Conventionally defined as the

worst-case minimum access time

Physical and **logic**, where the 1st is the superior limit reachable by the 2nd

worst-case minimum access time



Inst1 → Inst1 = 10 msec
Inst3 → Inst1 = from min20 to max30 (according to the path)
Inst5 → Inst1 = from min100 to max200
wcase.min.acc.time = 100 msec

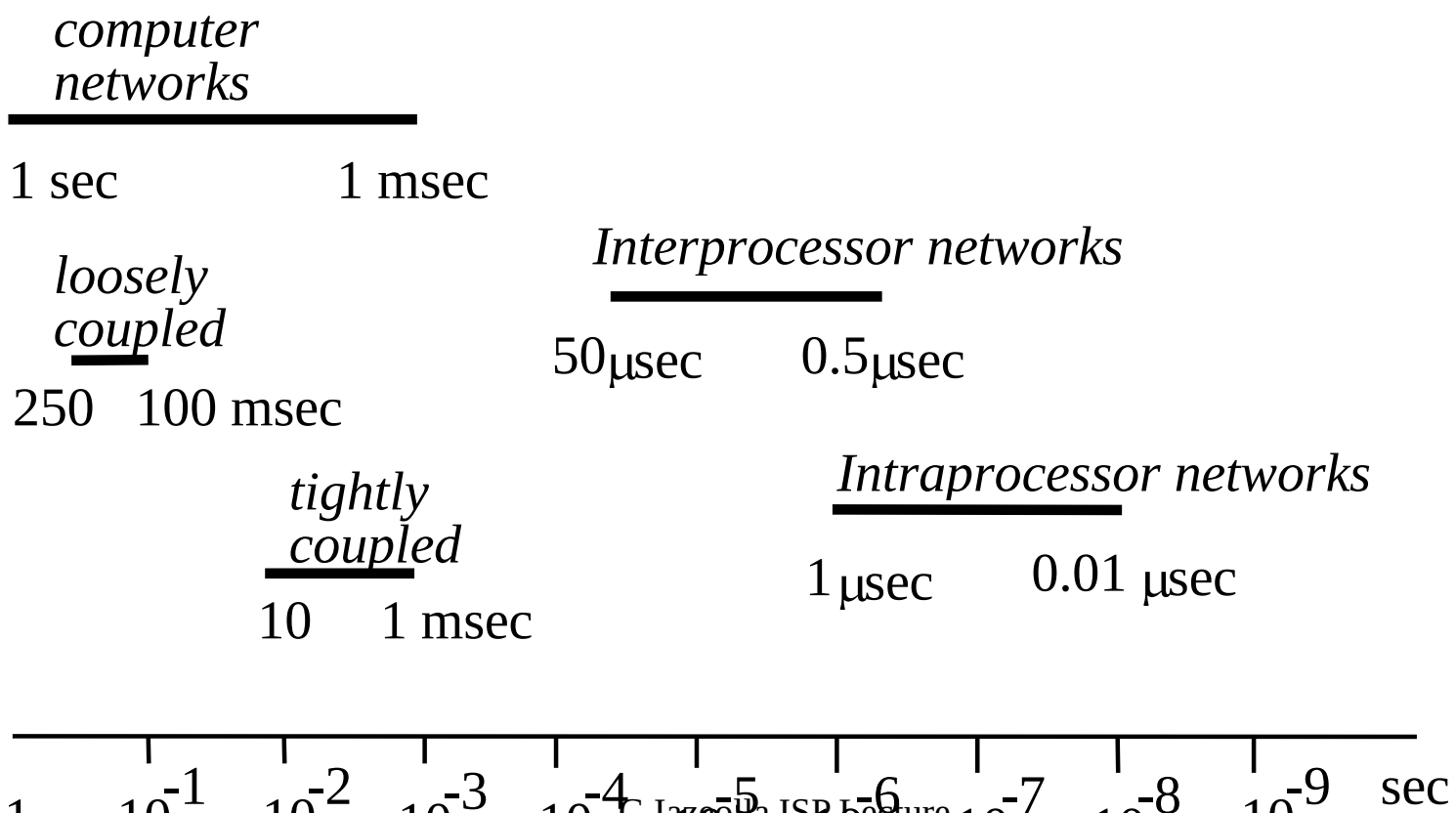
worst-case minimum access time

We presume that common information resides in the main memory of installation 1.

- Access to these by the same installation will cost as much as a memory fetch. Let's presume at least 10msec.
- Access by installation 3 will also require the sending of a message to installation 1, with an overall cost of min 20msec.
- Access by installation 5 requires a triple skip through installation 4 and 3 to send to installation 1 what could cost a min of 100msec.
- By applying the definition of *worst-case minimum access time* we have that the network in consideration has a latency time of 100 msec, a value which falls within the band indicated above from 1msec to 1 sec.

Latency times for the various types of network

The nodes have rare interactions in **loosely coupled** computer networks (e.g. they host 2 independent o.s.)
The nodes have frequent interactions in tightly coupled computer networks (they host two parts of 1 similar o.s.)



Connecting ways: operating

Way of communicating among nodes

Synchronous: the nodes communicate in a way that the recipient node makes sure it does it contemporarily to the one which transmits.

Asynchronous: the node which transmits also does it while the receiver is occupied in other activities (information is registered in a memory area of the receiver and detected as soon as possible)

Mixed: both ways.

Connecting ways: Control strategy

Way with which the components of S and L type are managed for the communication among nodes

Control can be

centralized (governed by a central control),

distributed (governed by the same single components S and L)

Connecting ways: Switching method

Way with which data is routed to transfer it from one node to another

Circuit switching

The connection remains active during the transmission between 2 nodes which operate in a synchronous way. The intermediate nodes aren't involved.

Packet switching

Data is divided into packets, each of which can follow a different route. The nodes operate in a asynchronous way. The intermediate nodes are involved.

Integrated switching

Combines the first two

Connecting ways: Communication protocol

Way in which the elements S and L of the network accomplish the control strategy and switching method

It carries out 2 functions:

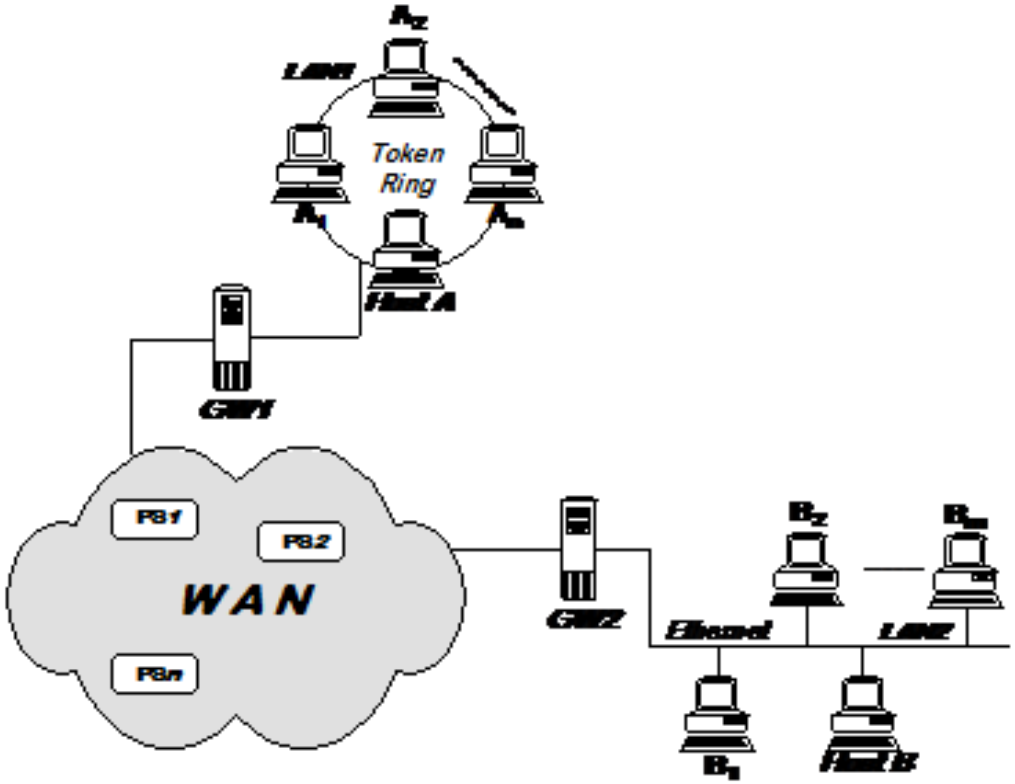
routing, concerns the control of elements S to obtain the routing of data from the source to the destination.

handshaking, concerns the control of elements of elements L, to obtain the cooperation among points S connected by the L themselves.

Internet platform =

= Computer network (or distributed computer network) but ”heterogeneous”

(**in** transmission medium, **in** network technologies and **in** connecting ways)



(PS= packet node switch of a WAN)