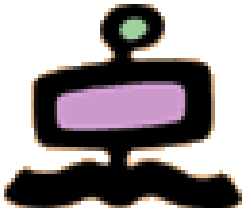
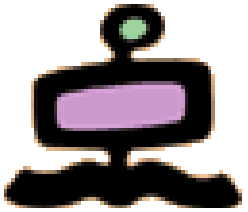
Corso di Biblioteche Digitali



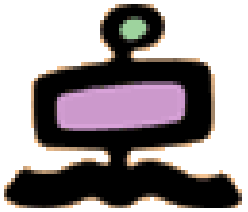
- Vittore Casarosa
  - Mail: casarosa@isti.cnr.it
  - Tel: 050 621 3115 (office) 348 397 2168 (mobile)
  - Skype: vittore1201
- “Ricevimento” at the end of the lessons or by appointment
- Final assessment
  - 70% oral examination
  - 30% project (development of a small digital library))
- Reference material:
  - Ian Witten, David Bainbridge, David Nichols, How to build a Digital Library, Morgan Kaufmann, 2010, ISBN 978-0-12-374857-7 (Second edition)
  - Material provided by the teacher
- **http://cloudone.isti.cnr.it/casarosa/BDG/**

Modules



- Computer Fundamentals and Networking
- A conceptual model for Digital Libraries
- Bibliographic records and metadata
- Information Retrieval and Search Engines
- Knowledge representation
- Digital Libraries and the Web
- Hands-on laboratory: the Greenstone system

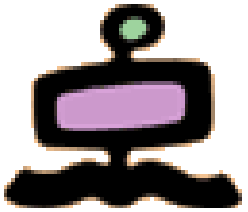
Refresher

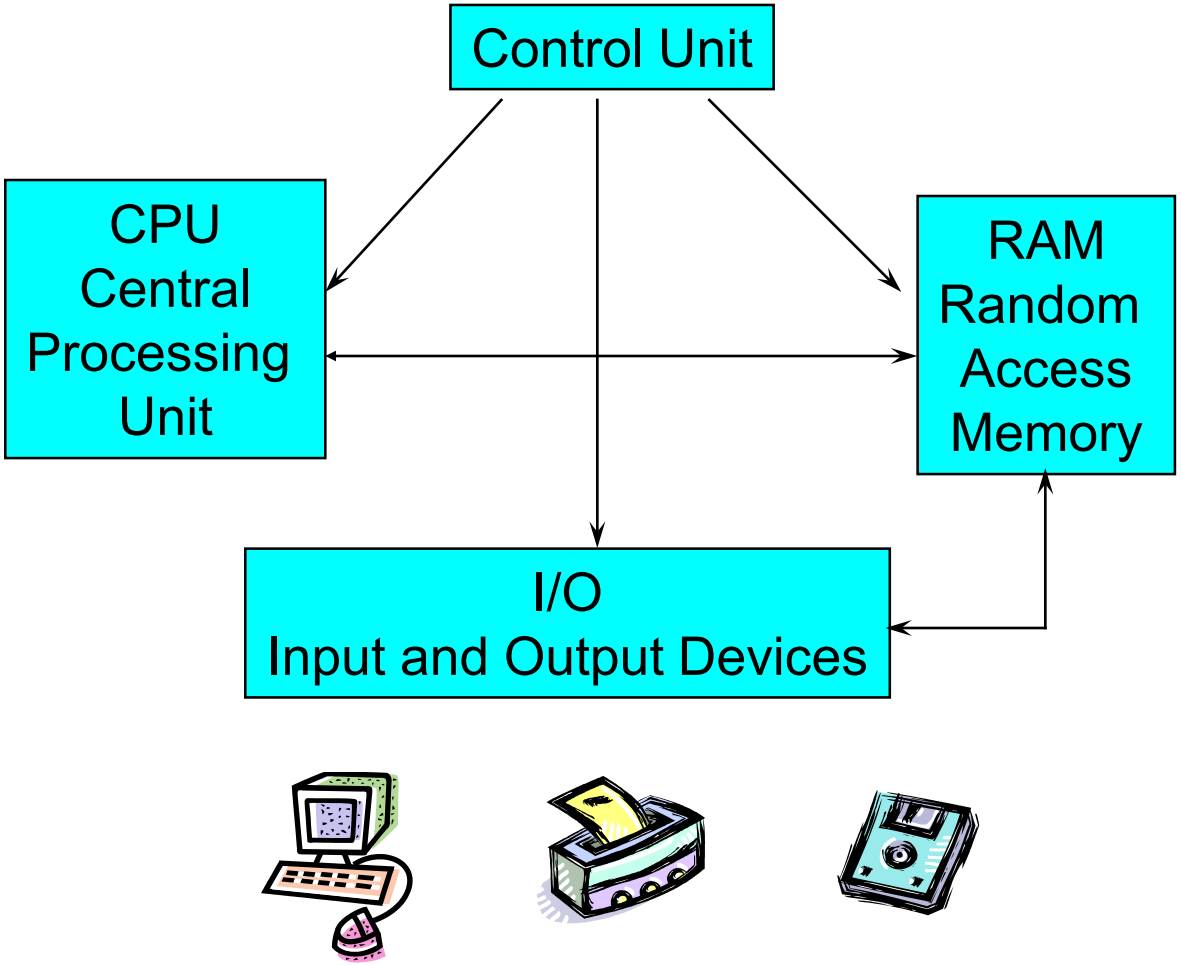


Refresher on Computer Fundamentals and Networking

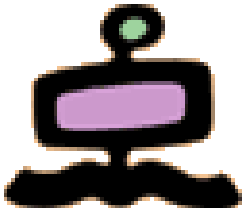
- History of computers
- Architecture of a computer
- Computer networks and the Internet
- Data representation within a computer

Basic components of a computer





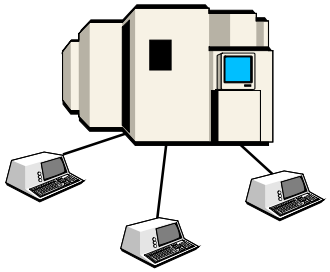
Evolution of technology

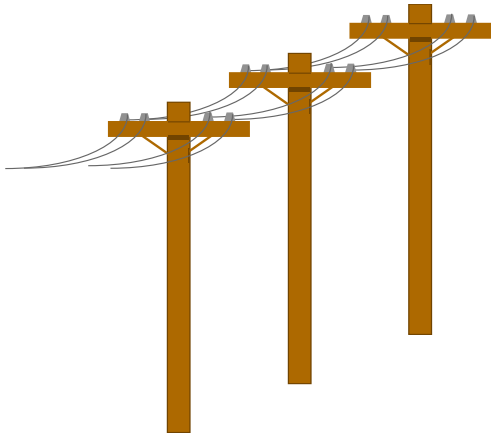


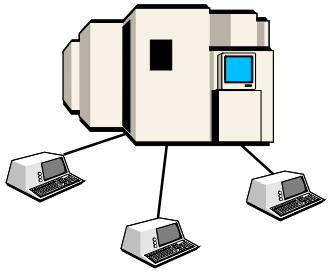
- Computer technology
  - CPU and integrated chips
  - Random Access Memories
    - RAM – from KB to GB
  - External memories
    - Tapes, hard disks, floppy disks
    - Memory sticks
    - CDs
    - DVDs
    - from MB to GB to TB to PB to EB
- Communication technology (networks)
  - (Telephone) line speed
  - Point to point (leased lines)
  - Local Area Networks
  - Inter-networking (TCP/IP)

Early computer communication

From mainframe to mainframe
through telephone lines
(point to point connection)

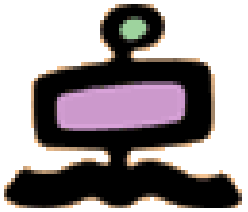






Telephone lines:
slow
expensive
regulated

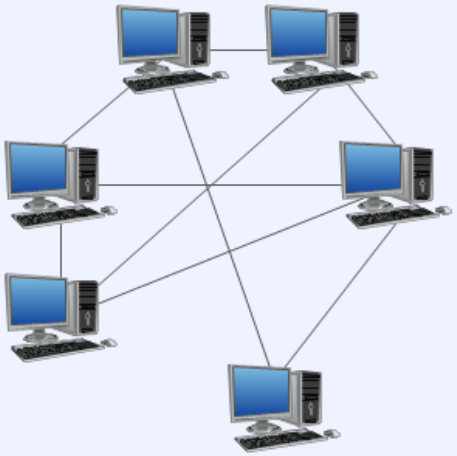
Networking



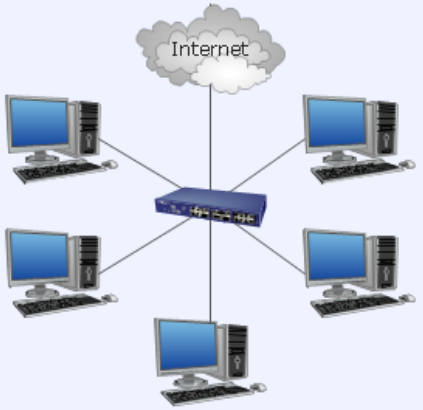
- In the sixties, first studies on “networking”
  - Networking means communication between node A and node B through one or more intermediate nodes
- In the seventies, fragmentation of the market with the arrival of “minicomputers” provided further motivation for research on networking
- At the same time (in the seventies), the arrival of the LANs (Local Area Networks) provided the final impulse for the development of networking

Network topologies

Fully Connected Network Topology

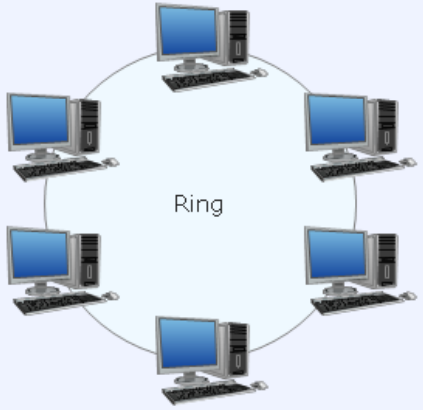


Mesh Network Topology



Star Network Topology

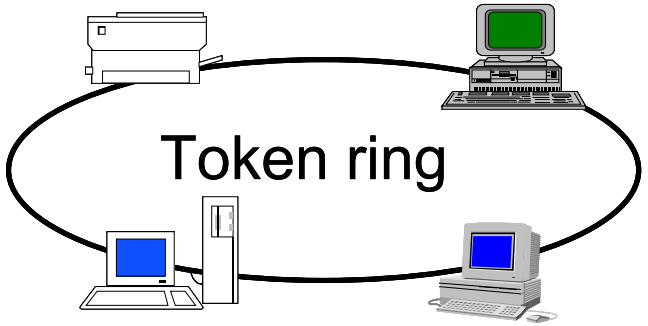
Common Bus Topology



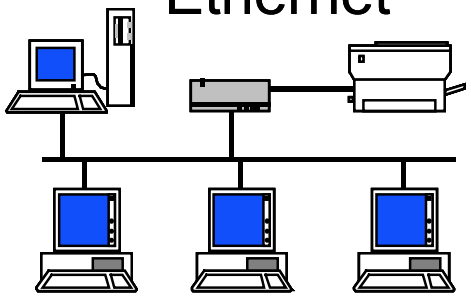
Ring Network Topology

LAN - Local Area Networks

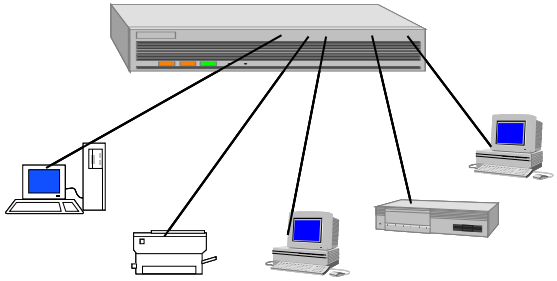
Private networks
Up to several kilometers
Speed up to 100 Mb/sec



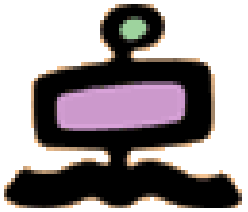
Ethernet



LAN switch

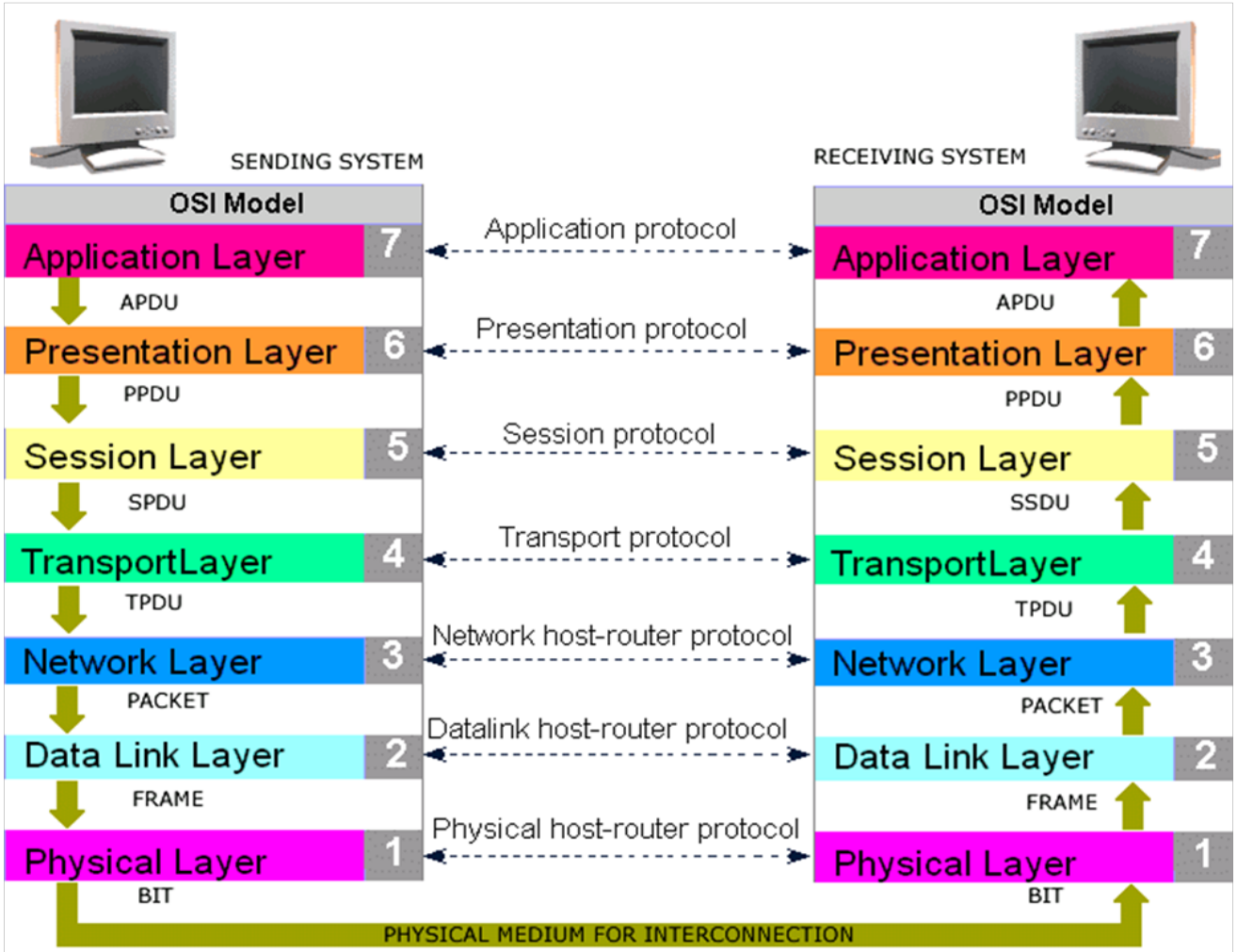


Research on networking



- Starting in the late sixties, many research projects on networking, both from universities and industry
  - Arpanet, Cyclades, SNA (IBM), DECnet
- In the late seventies ISO (International Standard Organization), under pressure of a group of computer manufacturer, started the work for the proposal of a “new” communication standard, called OSI: Open System Interconnection
- The OSI model, though no longer in use today, has established a number of networking concepts and is still used as a “reference model”
- The main concepts introduced by OSI are the “communication layers”

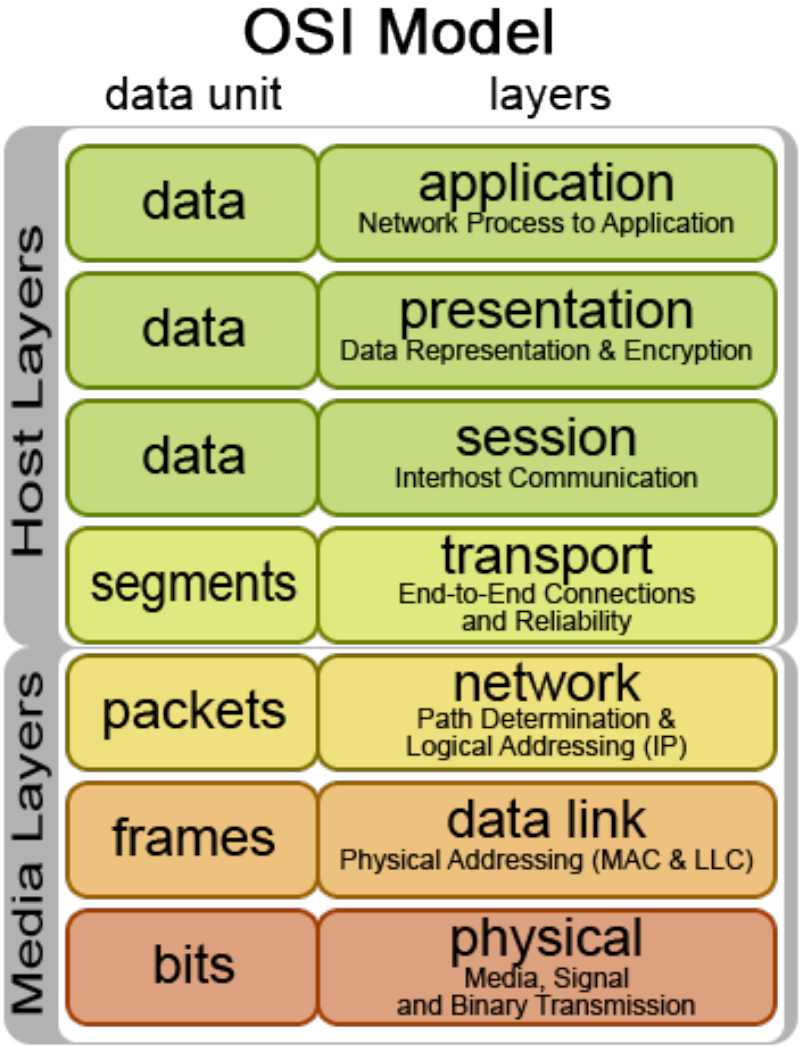
The OSI model 1980-1990



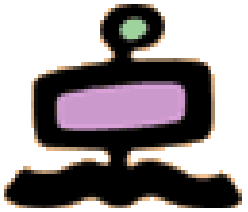
Protocol:
formats and rules for exchanging messages between "partners" (e.g. computers)

Packet switching:
messages are broken down into "packets", and each packet gets to destination independently from the others.

OSI – Open System Interconnection

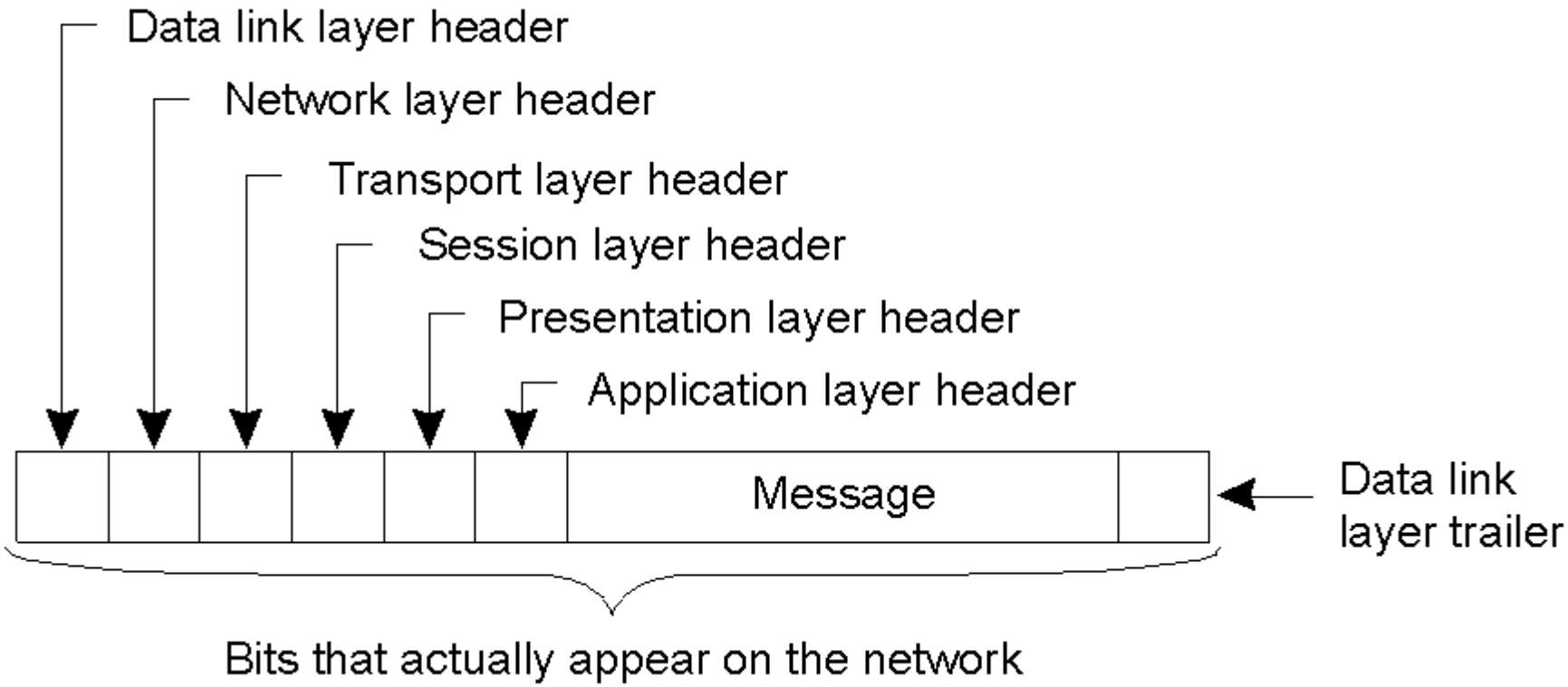


The seven layers

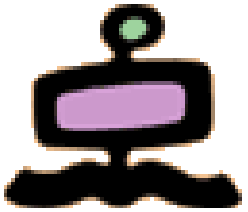


- Layer 7: The application layer...This is the layer at which communication partners are identified, quality of service is identified, user authentication and privacy are considered, and any constraints on data syntax are identified. (This layer is not the application itself, although some applications may perform application layer functions.)
- Layer 6: The presentation layer...This is a layer, usually part of an operating system, that converts incoming and outgoing data from one presentation format to another (for example, from a text stream into a popup window with the newly arrived text). Sometimes called the syntax layer.
- Layer 5: The session layer...This layer sets up, coordinates, and terminates conversations, exchanges, and dialogs between the applications at each end. It deals with session and connection coordination.
- Layer 4: The transport layer...This layer manages the end-to-end control (for example, determining whether all packets have arrived) and error-checking. It ensures complete data transfer.
- Layer 3: The network layer...This layer handles the routing of the data (sending it in the right direction to the right destination on outgoing transmissions and receiving incoming transmissions at the packet level). The network layer does routing and forwarding.
- Layer 2: The data-link layer...This layer provides synchronization for the physical level and does bit-stuffing for strings of 1's in excess of 5. It furnishes transmission protocol knowledge and management.
- Layer 1: The physical layer...This layer conveys the bit stream through the network at the electrical and mechanical level. It provides the hardware means of sending and receiving data on a carrier.

Layered Protocols



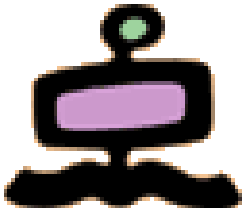
Mnemonics for OSI layers



All	Application
People	Presentation
Seem	Session
To	Transport
Need	Network
Data	Data Link
Processing	Physical

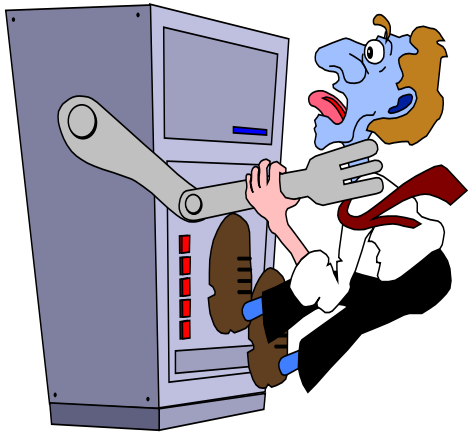
Please	Physical
Do	Data Link
Not	Network
Throw	Transport
Sausage	Session
Pizza	Presentation
Away	Application

OSI and Internet

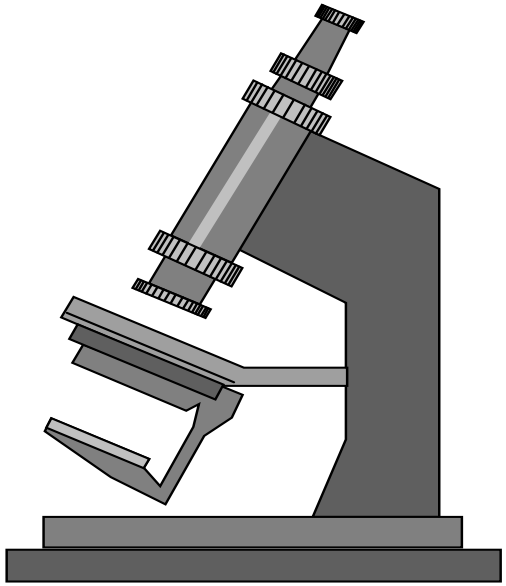


- The OSI effort provided a sound and durable foundation for networking, but never became a “market leader”
  - Slow development (late ’70s to early ’90s)
    - Initial opposition from IBM
    - “Designed by a Committee”
  - Expensive development
  - Heavy and slow in operation
- In the same period the Internet was defining a number of “light weight” protocols
- Most of the market preferred them to OSI

Internet evolution
1960-1990



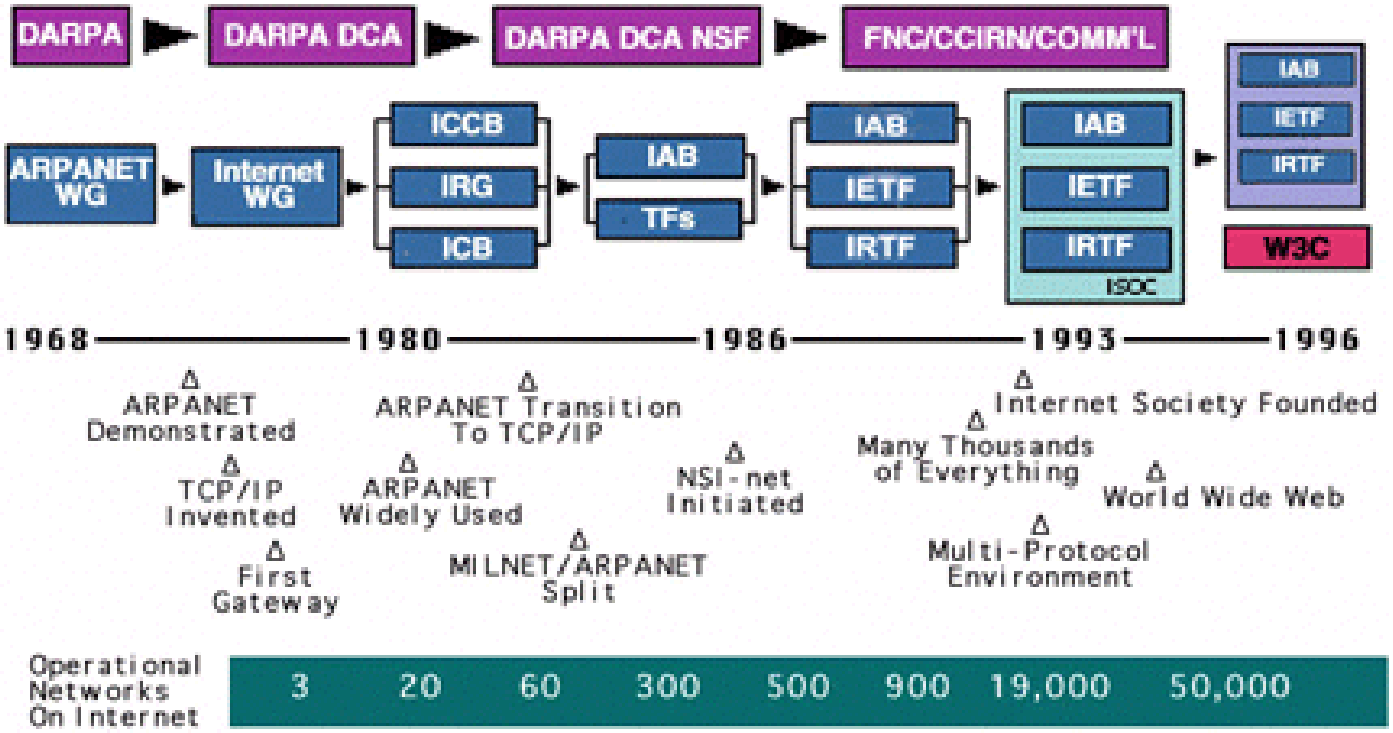
Reserach
Network
NSF
Internet



Experimental
Network
DARPA
Arpanet

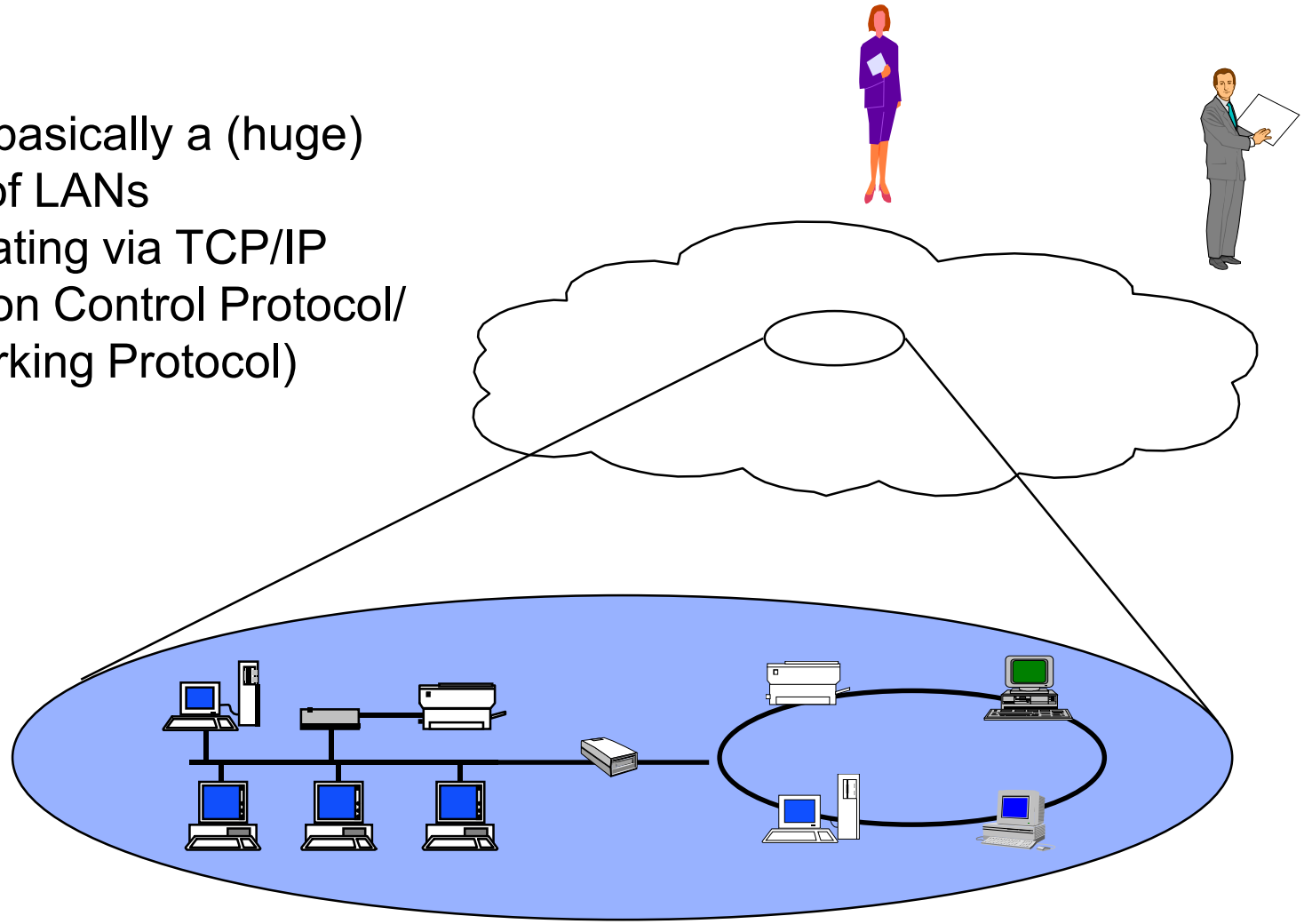
Communication
Infrastructure
Private and public
sectors
The Web

Internet timeline



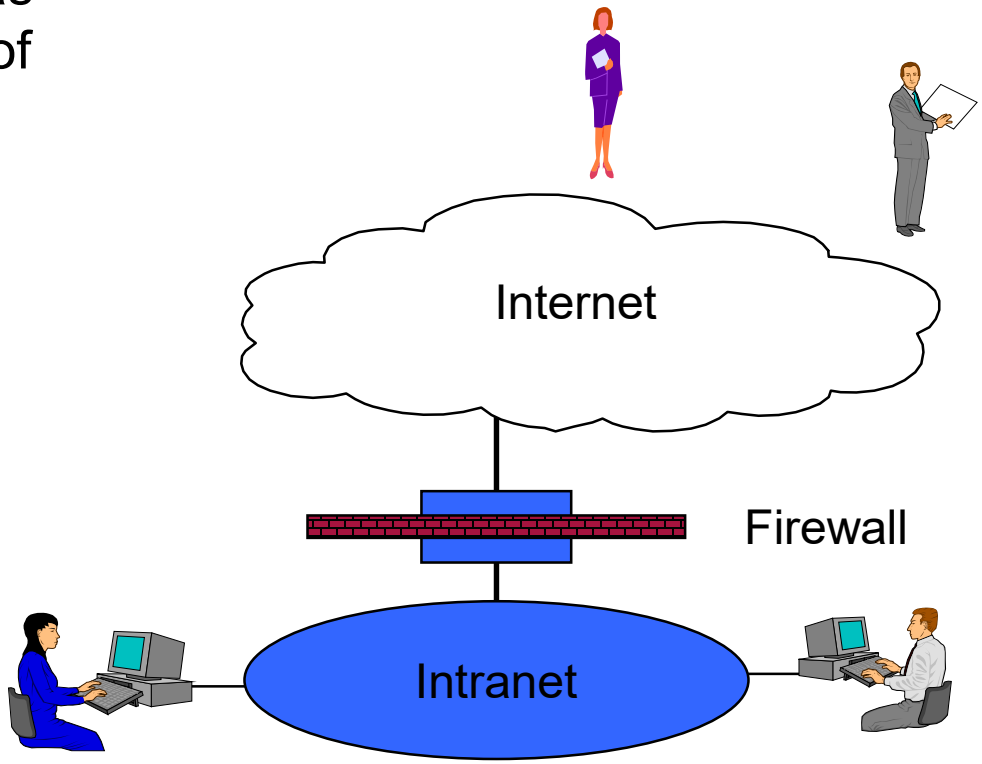
Inter-networking

Internet is basically a (huge) collection of LANs communicating via TCP/IP (Trasmission Control Protocol/ Internetworking Protocol)

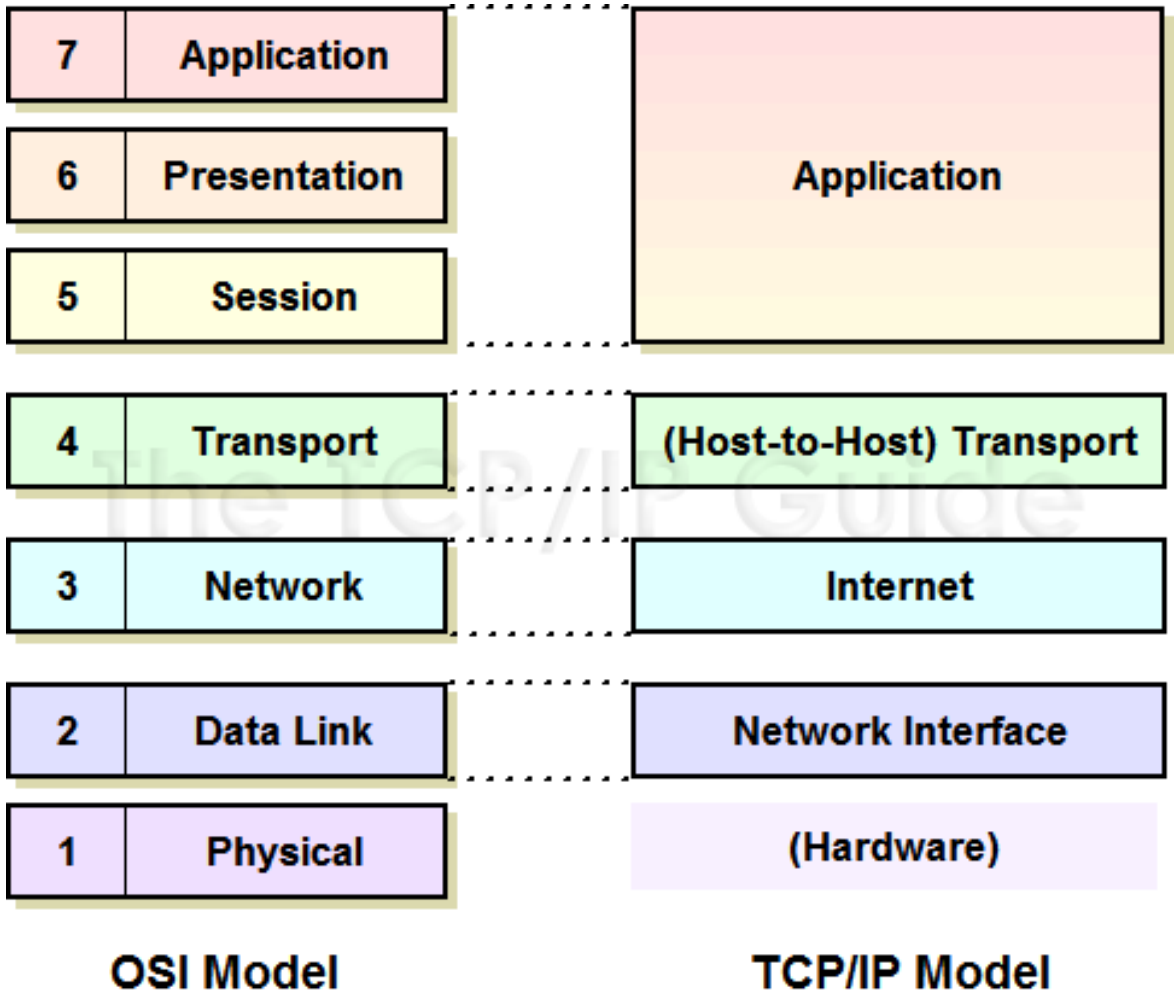


Internet and Intranets

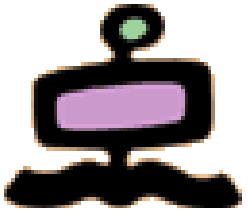
The growth of Internet was also due to the adoption of the Internet protocols by private companies



OSI and TCP/IP

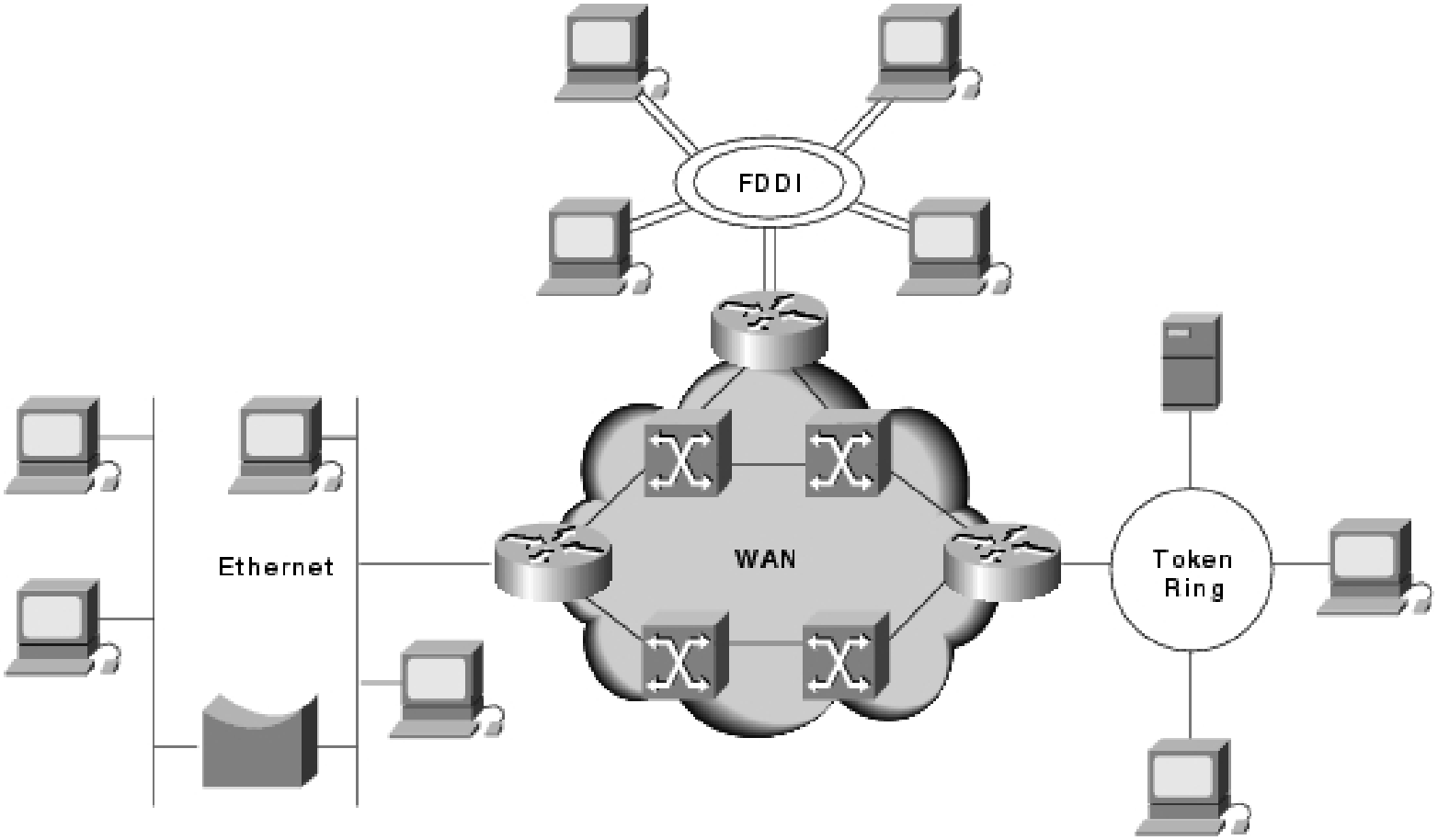


The Internet layers

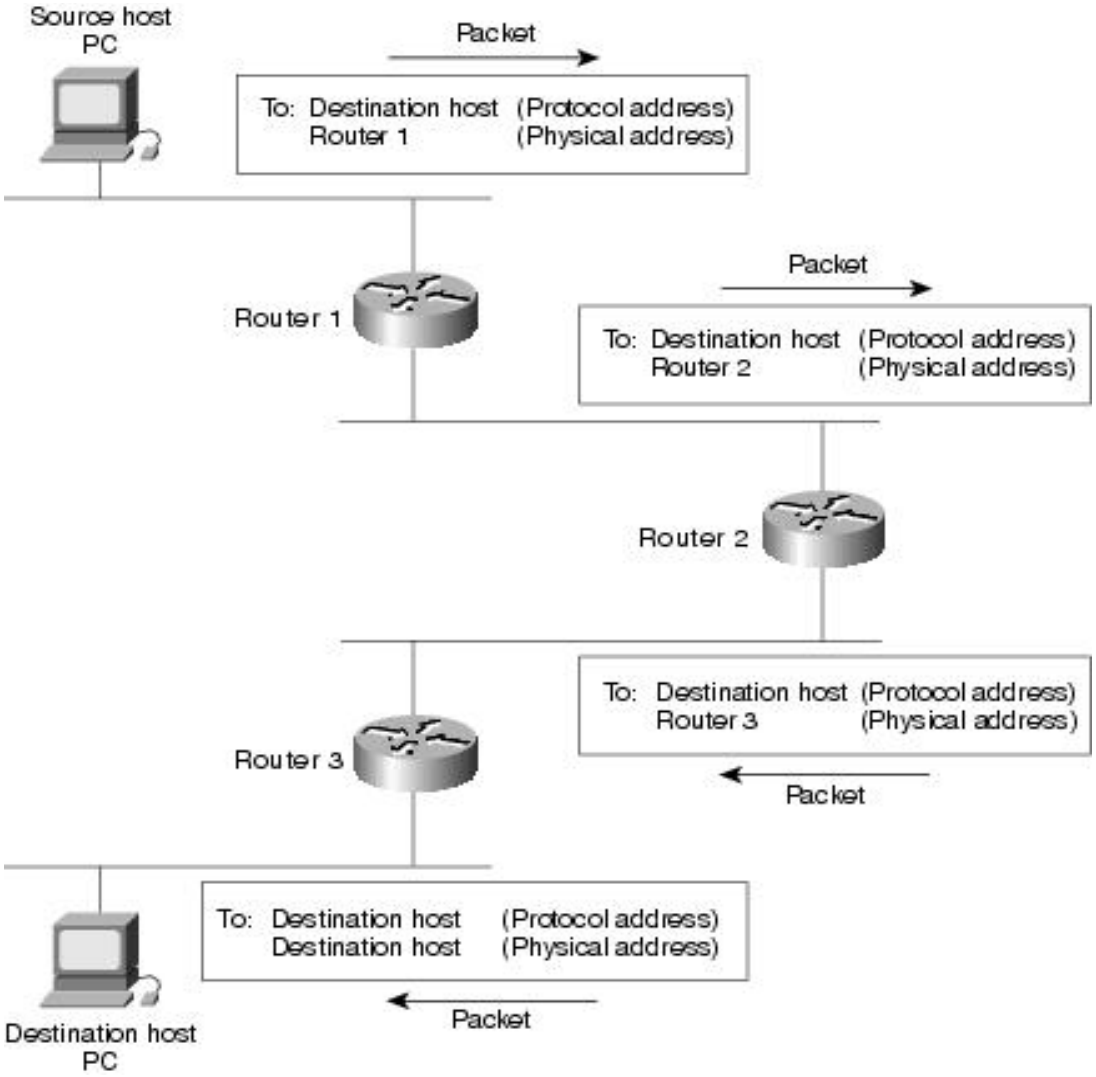


Layer	Function
Application Services	Email (SMTP), network management (SNMP), WWW (HTTP), and other services
Transport	End-to-end message integrity (UDP, TCP) and error recovery (TCP)
Internet	Connectionless, best-effort packet delivery (IP)
Network Interface	Almost any network technology

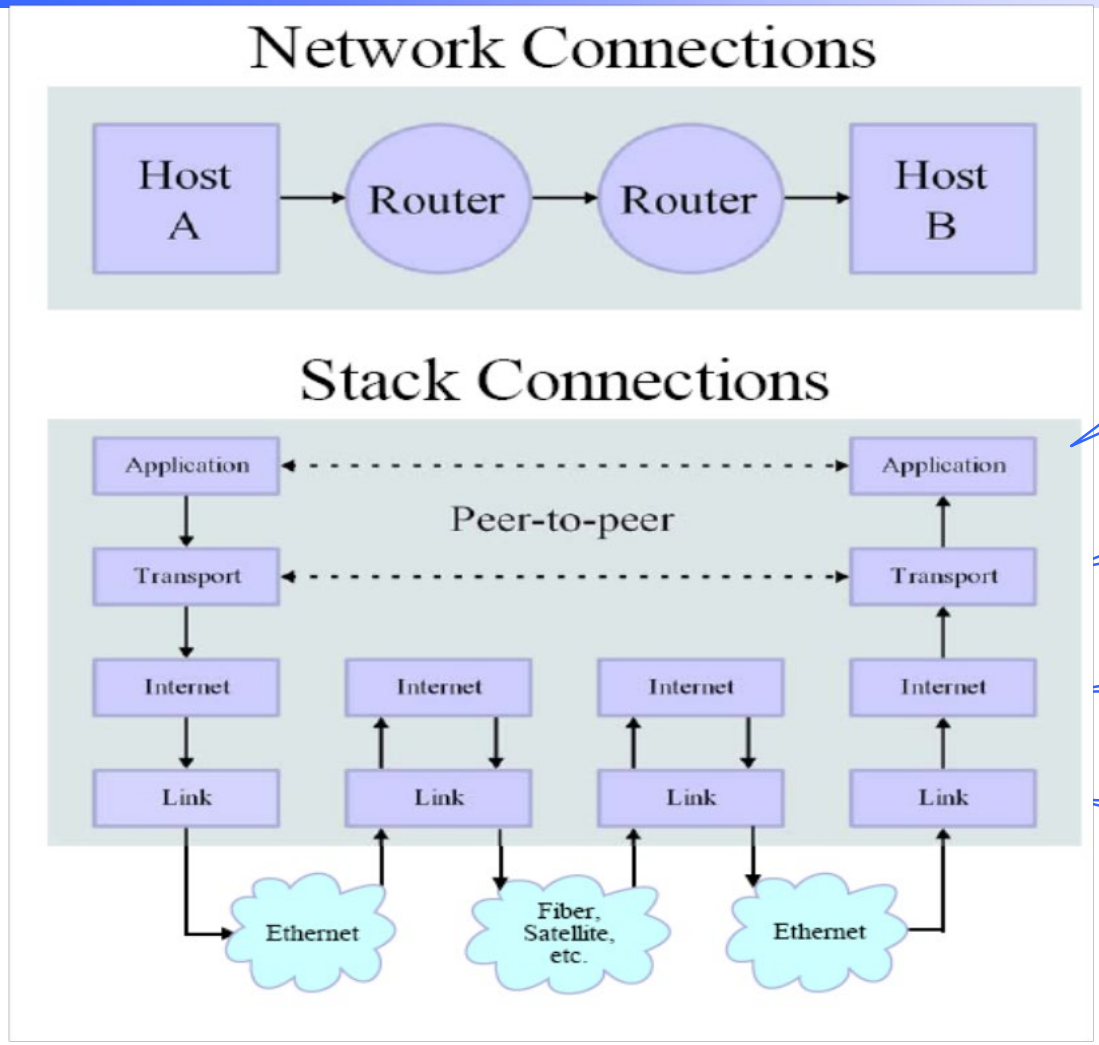
Internetworking



Routing in Internet



Internet protocols



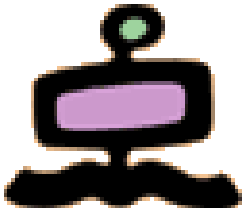
Application protocols

TCP/UDP

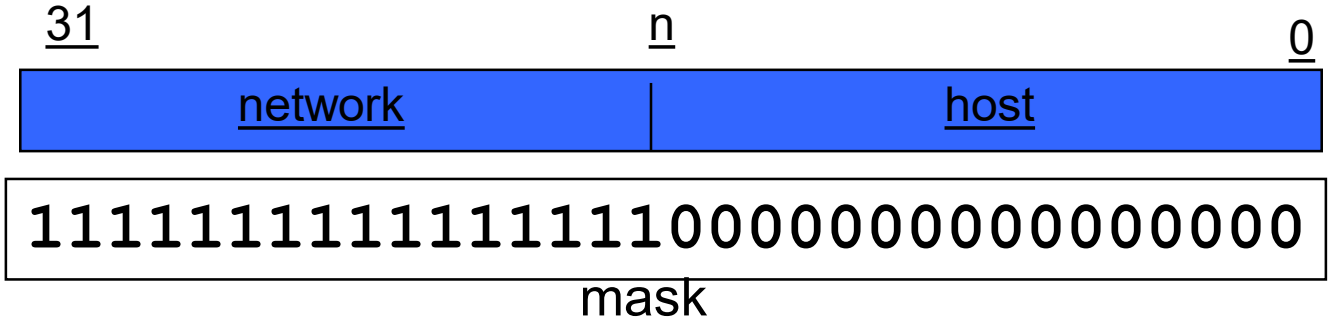
IP

ETHERNET

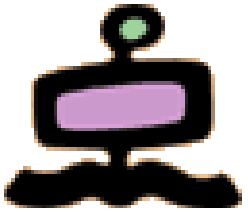
IPv4 addressing



- Each node in the Internet is identified by (one or more) IP address, and each IPv4 address has 32 bits (4 bytes)
- An IP address is (was) made of two parts:
  the network address and the node address within the network
- The boundary between the parts is variable, and is identified by the “network mask”
- The 1s in the mask identify the net portion and the 0s the host portion



IPv4 addresses

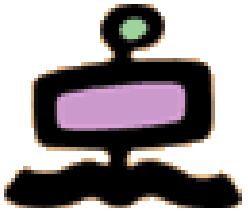


- An IP address is usually indicated with four numbers (from 0 to 255) corresponding to the 4 bytes of the address

  IP address: 131.114.1.30
  mask: 255.255.255.0
  network address 131.114.1
  host address 30

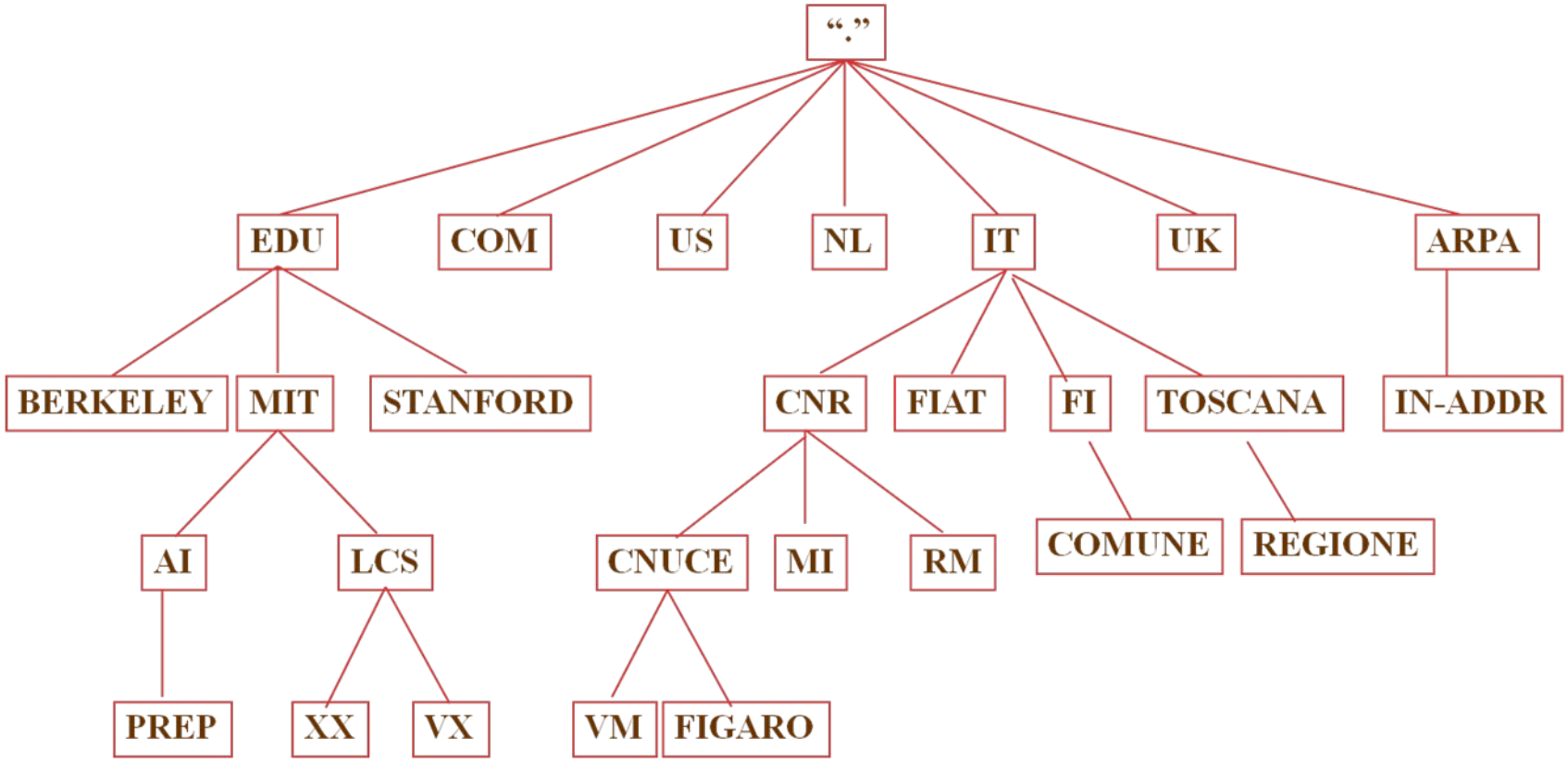
- Three classes of network addresses (255.0.0.0, 255.255.0.0, 255.255.255.0)
- No more IPv4 addresses available today
  - Network Address Translation (NAT) commonly used
- IPv6 (128 bits) replacing IPv4

DNS: Domain Name System



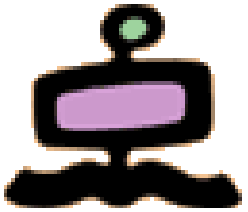
- In order to be connected to the Internet, a host (computer) must have one (or more) IP address (static or dynamic)
- At the beginning (and more human friendly), hosts on the Internet were identified by names through a simple text file (hosts.txt)
- Very simple system, but not scalable to thousand of hosts (impossible to keep updated)
- Definition of a hierarchical name system based on domains
  - Distributed data base (zones)
  - Quick access to data more important than real time updating (cache)
- In each zone there is a table with the IP address of the (last part of the) domain name

Hierarchical structure



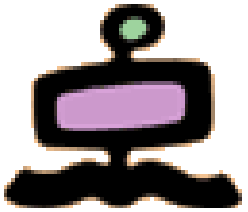
The complete name of each domain is the sequence of the node names (going towards the root) separated by “dot”
For example: VM●CNUCE●CNR●IT

Name domains



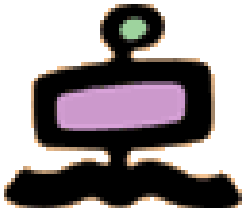
- Each node is responsible to ensure the uniqueness of the first level domain names under it
- Top Level domains (tld)
- Generic TLD (gtld), defined initially
  - com, edu, org, net, int, gov, mil (1984)
  - biz, info, name, pro, aero, coop, museum (2000)
  - eu, xxx, firenze (?), roma (?) (2013 ???)
- Country code TLD (cctld), defined by ISO 3166 (about 300)
  - it, de, nl, etc
  - gp (Guadalupe), pa (Palestine)
- Cybersquatting and trade marks

Evolution of technology



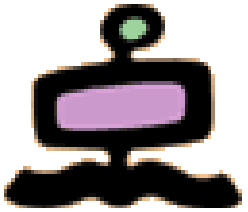
- Computer technology
  - CPU and integrated chips
  - Random Access Memories
    - RAM – from KB to GB
  - External memories
    - Tapes, hard disks, floppy disks
    - Memory sticks
    - CDs
    - DVDs
    - from MB to GB to TB to PB to EB
- Communication technology (networks)
  - Point to point through (possibly leased) telephone lines (problems were speed and cost)
  - Local Area Networks
  - Inter-networking (TCP/IP)

Evolution of computer market



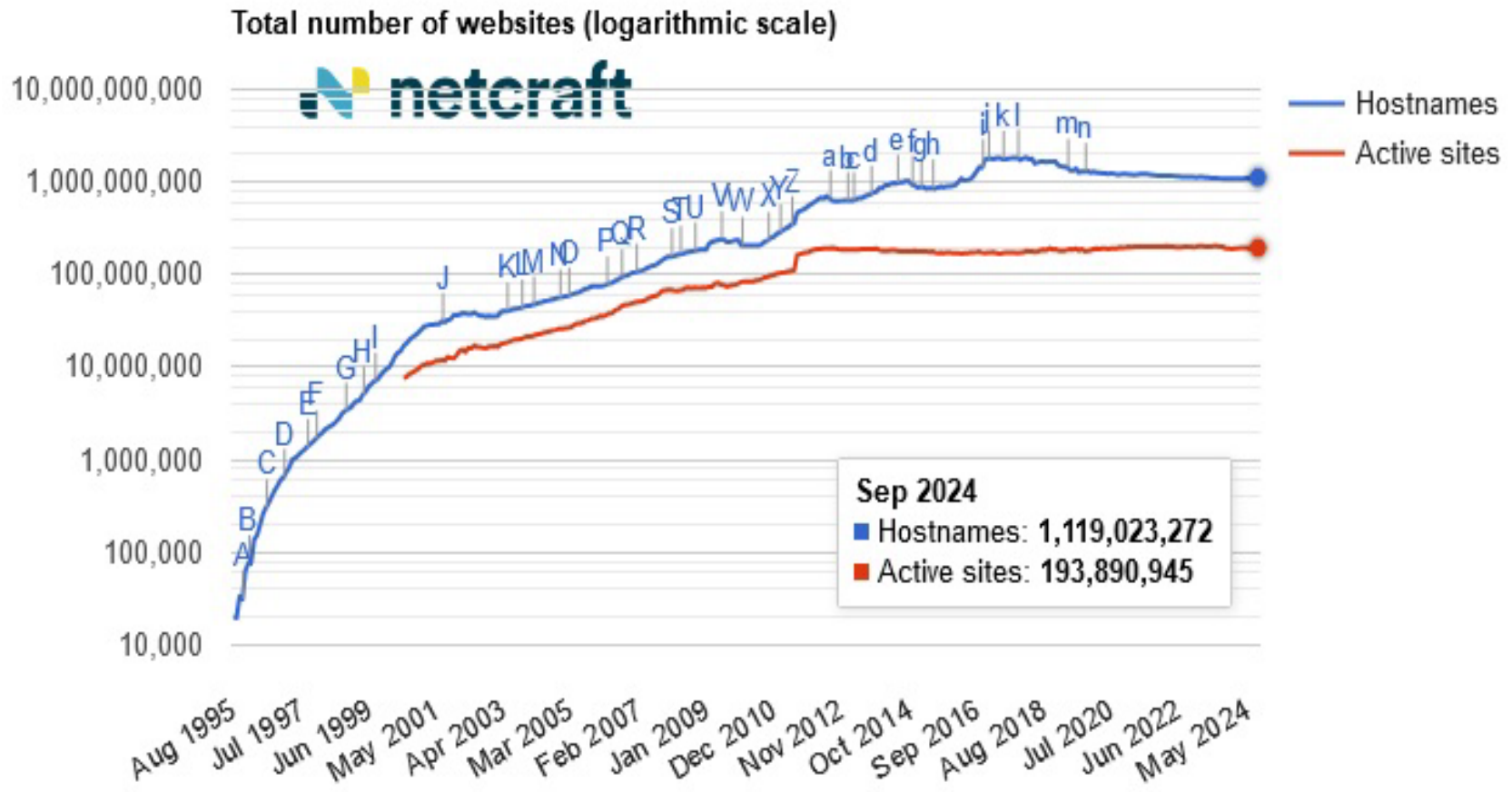
- Military applications in early 40s
- Scientific/research applications in late 40s
- Commercial applications appear in early 50s
- Monopoly of IBM starts with 650, 701, 702
- Monopoly of IBM continues with 7070, 7090 and the 360 series, starting the “mainframe era” (in the 60s)
- Arrival of the “minicomputers” in the 70s
- Arrival of the PC in the 80s
- Arrival of the (public) Internet in the 90s
- Arrival of the Web in the 90s

The Internet and the Web



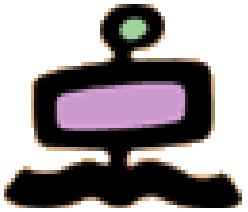
- Internet is a worldwide network of computers
  - It started in 1969 as a university research network (funded by DARPA) with 4 computers
  - By the end of the 80's, when it was opened to "the world", it already had more than 20000 hosts in universities and research centers worldwide
  - As of May 2023, the total number of web servers is estimated to be in the order of 1100 millions hosts
  - As of June 2022, the total number of Internet users is estimated to be 5,473 billions users
- The Web is the information space accessible through the Internet
  - As of September 2022, the number of "visible" Web pages (indexed by Google) is estimated to be between 50 and 60 billions pages
  - There is also a Deep Web, whose content is not indexed by any search engine, and whose size is completely unknown
- The Web has been made possible by a combination of computer technology and communication technology

Internet Web servers



https://news.netcraft.com/archives/category/web-server-survey/

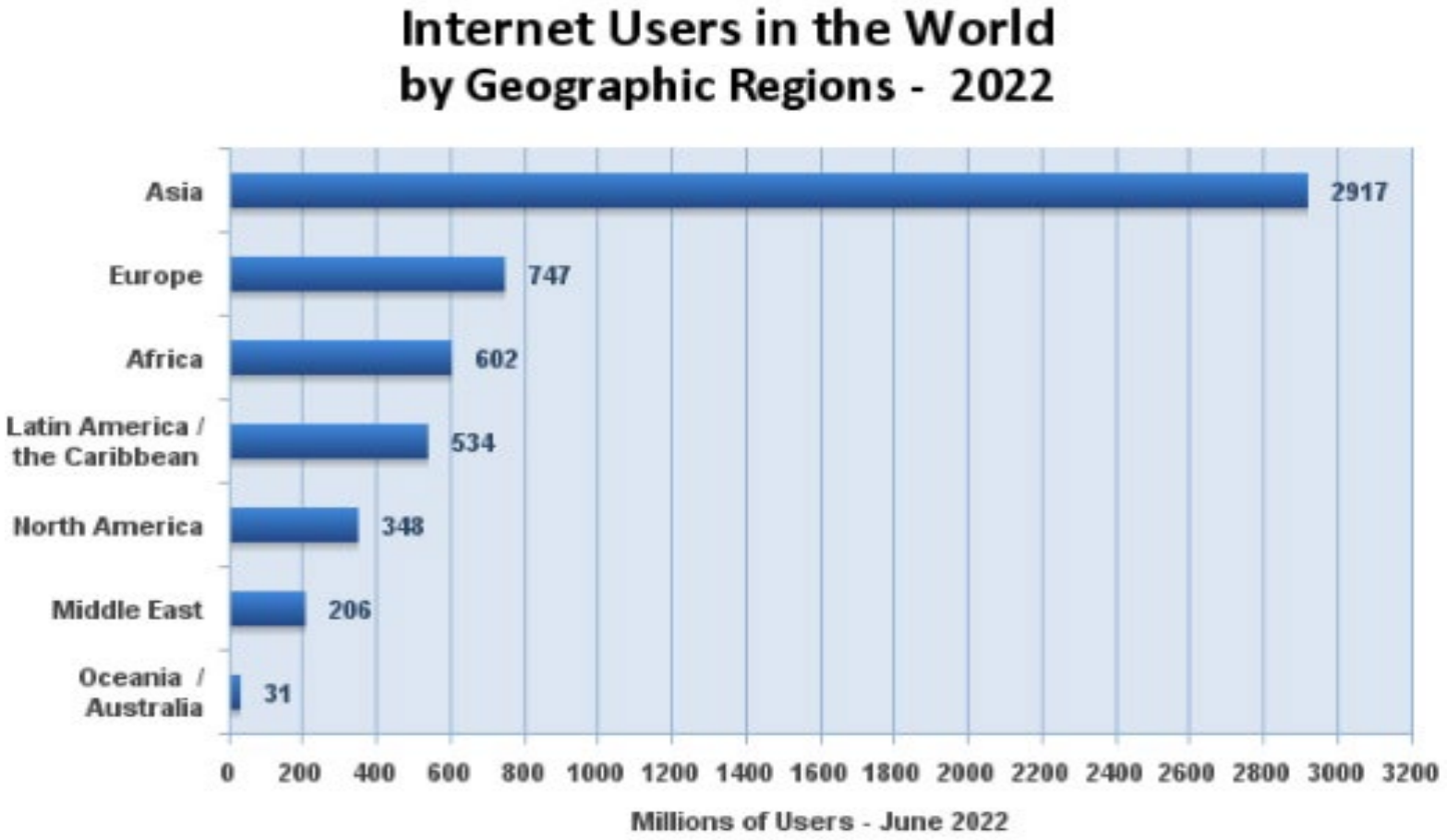
Internet users in the World



WORLD INTERNET USAGE AND POPULATION STATISTICS 2023 Year Estimates						
World Regions	Population (2022 Est.)	Population % of World	Internet Users 31 Dec 2021	Penetration Rate (% Pop.)	Growth 2000-2023	Internet World %
Africa	1,394,588,547	17.6 %	601,940,784	43.2 %	13,233 %	11.2 %
Asia	4,352,169,960	54.9 %	2,916,890,209	67.0 %	2,452 %	54.2 %
Europe	837,472,045	10.6 %	747,214,734	89.2 %	611 %	13.9 %
Latin America / Carib.	664,099,841	8.4 %	534,526,057	80.5 %	2,858 %	9.9 %
North America	372,555,585	4.7 %	347,916,694	93.4 %	222 %	6.5 %
Middle East	268,302,801	3.4 %	206,760,743	77.1 %	6,194 %	3.8 %
Oceania / Australia	43,602,955	0.5 %	30,549,185	70.1 %	301 %	0.6 %
WORLD TOTAL	7,932,791,734	100.0 %	5,385,798,406	67.9 %	1,392 %	100.0 %

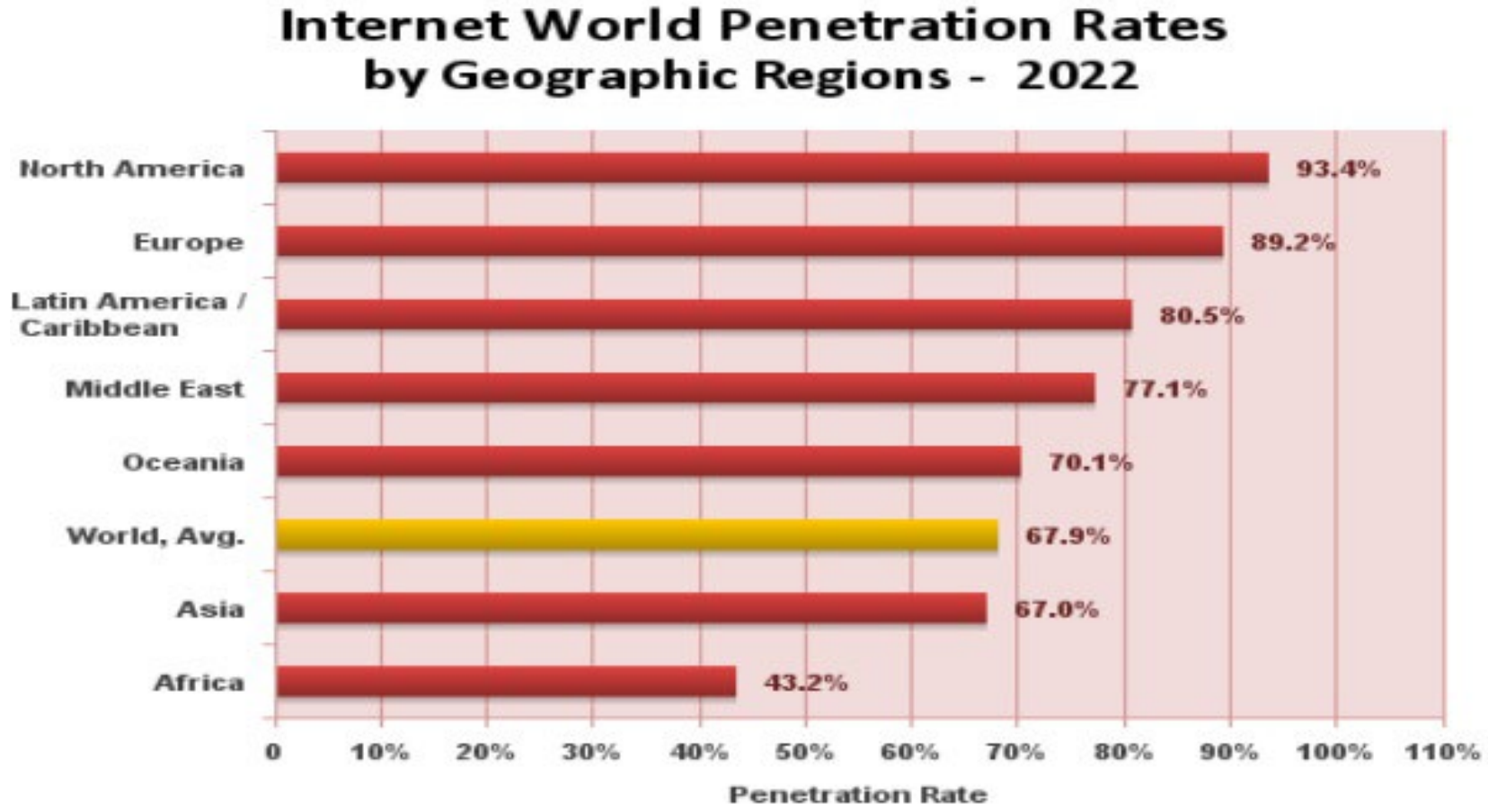
NOTES: (1) Internet Usage and World Population Statistics estimates are for June 30, 2022. (2) CLICK on each world region name for detailed regional usage information. (3) Demographic (Population) numbers are based on data from the United Nations Population Division. (4) Internet usage information comes from data published by Nielsen Online, by the International Telecommunications Union, by GfK, by local ICT Regulators and other reliable sources. (5) For definitions, navigation help and disclaimers, please refer to the Website Surfing Guide. (6) The information from this website may be cited, giving the due credit to www.internetworldstats.com.

Internet users in the World



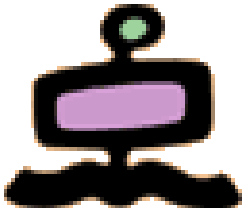
Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis:5,385,798,406 Internet users estimated in June 30, 2022

Internet World penetration rates



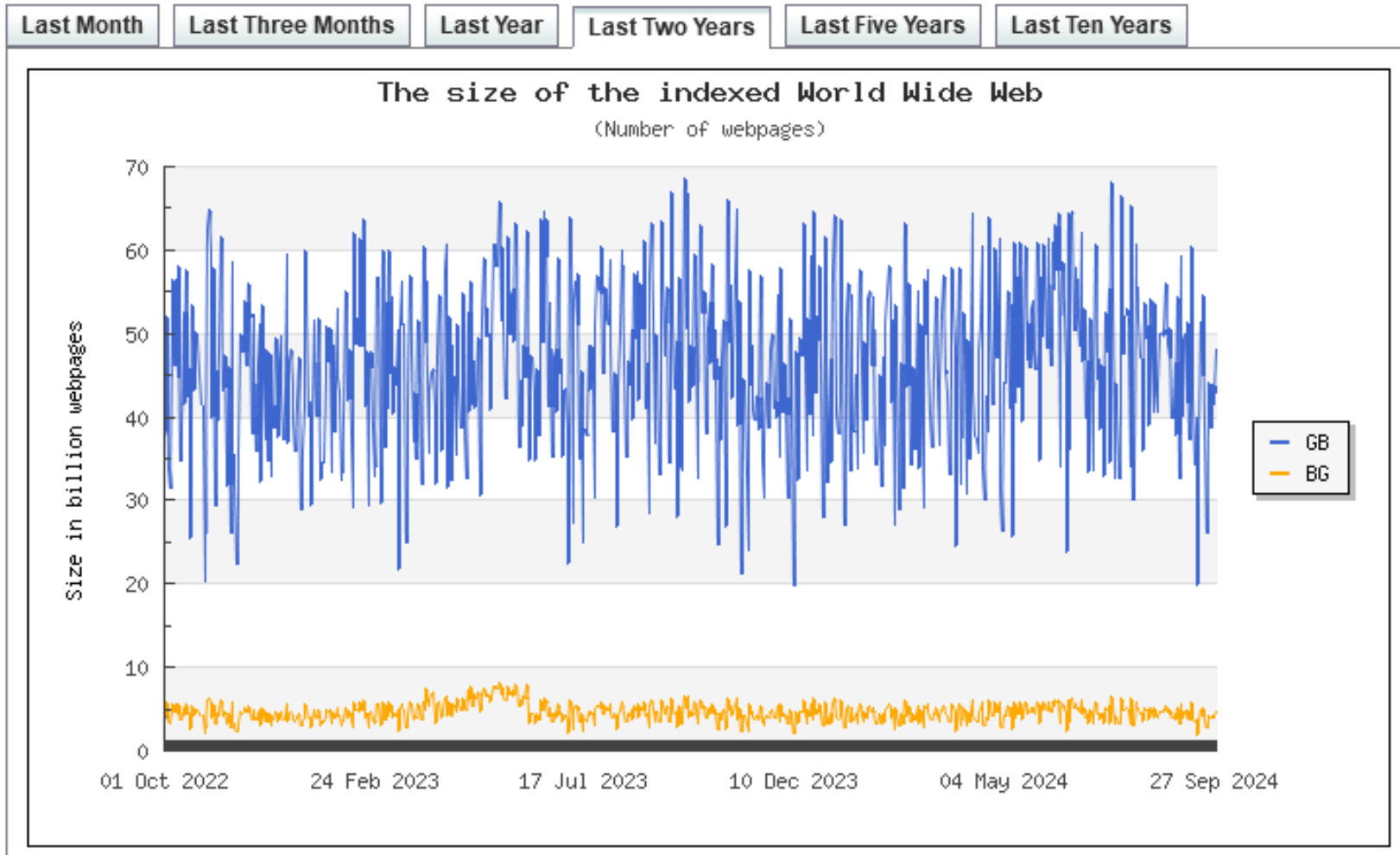
Source: Internet World Stats - www.internetworldstats.com/stats.htm
Penetration Rates are based on a world population of 7,932,791,734 and 5,385,798,406 estimated Internet users in June 30, 2022.

The Internet and the Web



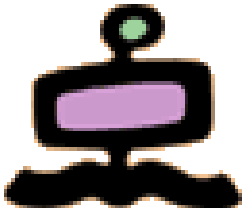
- Internet is a worldwide network of computers
  - It started in 1969 as a university research network (funded by DARPA) with 4 computers
  - By the end of the 80's, when it was opened to "the world", it had more than 20000 hosts in universities and research centers worldwide
  - As of September 2022, the total number of web servers is estimated to be in the order of 1,130 billions hosts
  - As of June 2022, the total number of Internet users is estimated to be 5,473 billions users
- The Web is the information space accessible through the Internet
  - As of September 2022, the number of "visible" Web pages (indexed by Google) is estimated to be between 40 and 50 billions pages
  - There is also a Deep Web, whose content is not indexed by any search engine, and whose size is completely unknown
- The Web has been made possible by a combination of computer technology and communication technology

The size of the indexed Web



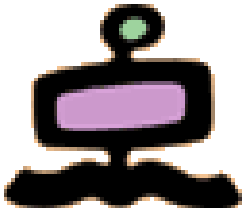
https://www.worldwidewebsize.com/

The Internet and the Web



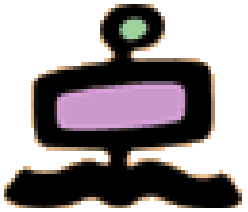
- Internet is a worldwide network of computers
  - It started in 1969 as a university research network (funded by DARPA) with 4 computers
  - By the end of the 80's, when it was opened to "the world", it had more than 20000 hosts in universities and research centers worldwide
  - As of September 2022, the total number of web servers is estimated to be in the order of 1,130 billions hosts
  - As of June 2022, the total number of Internet users is estimated to be 5,473 billions users
- The Web is the information space accessible through the Internet
  - As of September 2022, the number of "visible" Web pages (indexed by Google) is estimated to be between 50 and 60 billions pages
  - There is also a Deep Web, whose content is not indexed by any search engine, and whose size is completely unknown
- The Web has been made possible by a combination of computer technology and communication technology

The World Wide Web



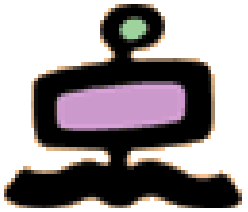
- Combination of computer technology and communication technology
- It all started with the “hyperlink” (late eighties)
- Then came the “browser” (Mosaic) (early nineties)
- Then came the “information explosion” (mid-nineties)
- Then came the “dot come, dot gone” (late nineties)
- Then came the second wave (early 2000)
- Then came the Web 2.0 (around 2004)
- Then came the Web 3.0 (around 2010)
- Today we have:
  - An estimate of about 1200 million hosts
  - An estimate of 40 to 50 billion pages on line

The editors



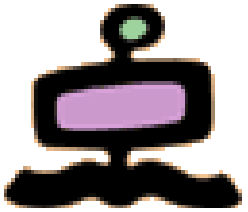
- Text processing applications started already in the early days of the computers (sixties)
- A “text processor” (or editor) has two main functions:
  - processing the text (delete, replace, insert, etc.)
  - specifying the format (bold, center, new line, etc.)
- The first editors were using a “mark up” language (i.e. commands intermixed with the text) to provide formatting instructions (only limited interactivity available through typewriter-like terminals)
- The “second generation” editors were using the WYSIWYG paradigm: What You See Is What You Get (much better interactivity available with display and mouse)

The hyperlink



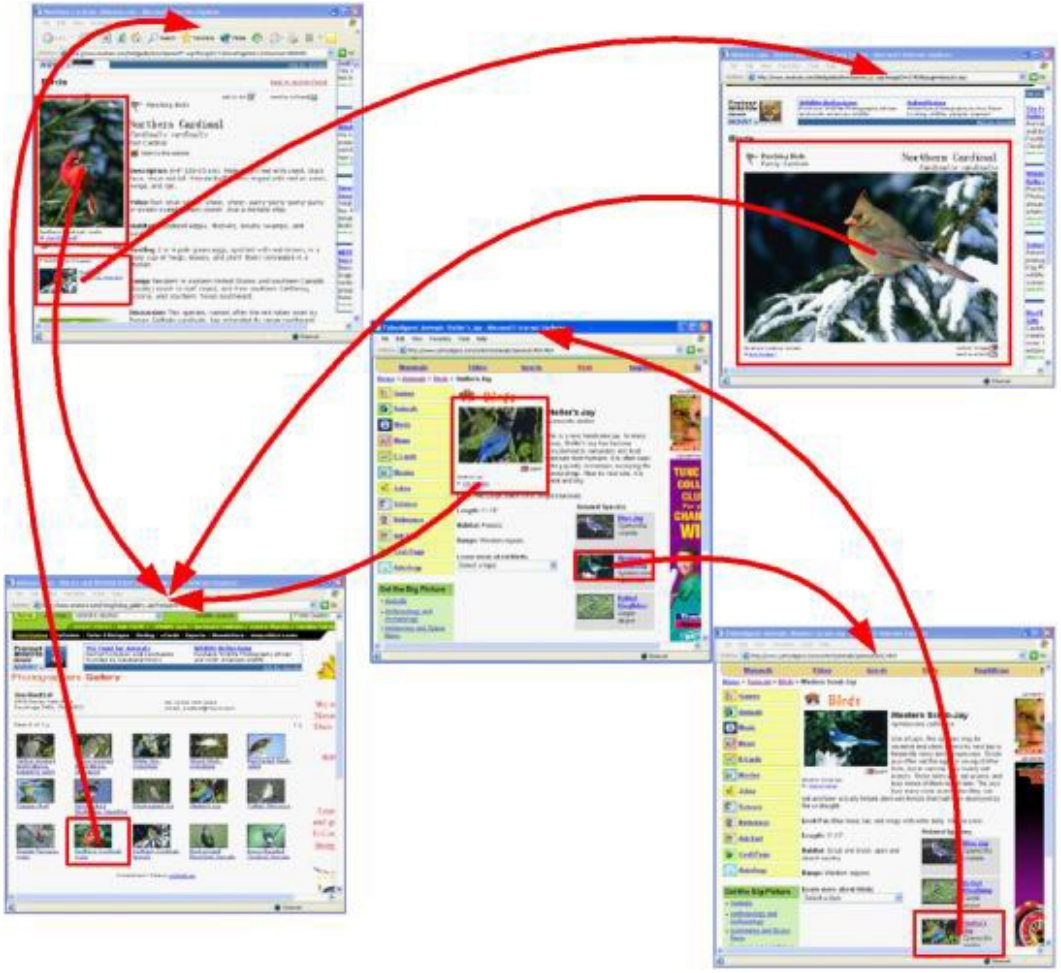
- The idea of the “hyperlink” was (experimentally) proposed in the sixties, as a feature of a “smart editor”
  - selecting a portion of the text, it was possible to open a second document, in addition to the one being edited (very awkward to use on a typewriter-like terminal)
- With the arrival of display screens and the mouse (eighties) the hyperlink came back in “3D documents”
  - clicking on a portion of the text it was possible to open a second document, which was maintained as a second (virtual) screen behind the first one
- With the arrival of the (fast) internet, it became the “web hyperlink”
  - clicking on a portion of the text it was possible to open a second document, coming from a different computer

The browser



- With the arrival of the (web) hyperlink, the problem was then how to properly display a (web) page that had been generated on a different computer, possibly with a different (wysiwyg) editor
- The solution was the definition of HTML (Hyper Text Markup Language), i.e. a standard mark up language for formatting a page, and the implementation of smart editors (called browsers – the most popular was Mosaic, released in 1993) capable of correctly displaying pages formatted with HTML, regardless of where they were coming from
- At the same time it was defined the HTTP protocol (Hyper Text Transfer Protocol) for the exchange of information between the browser and the Web server

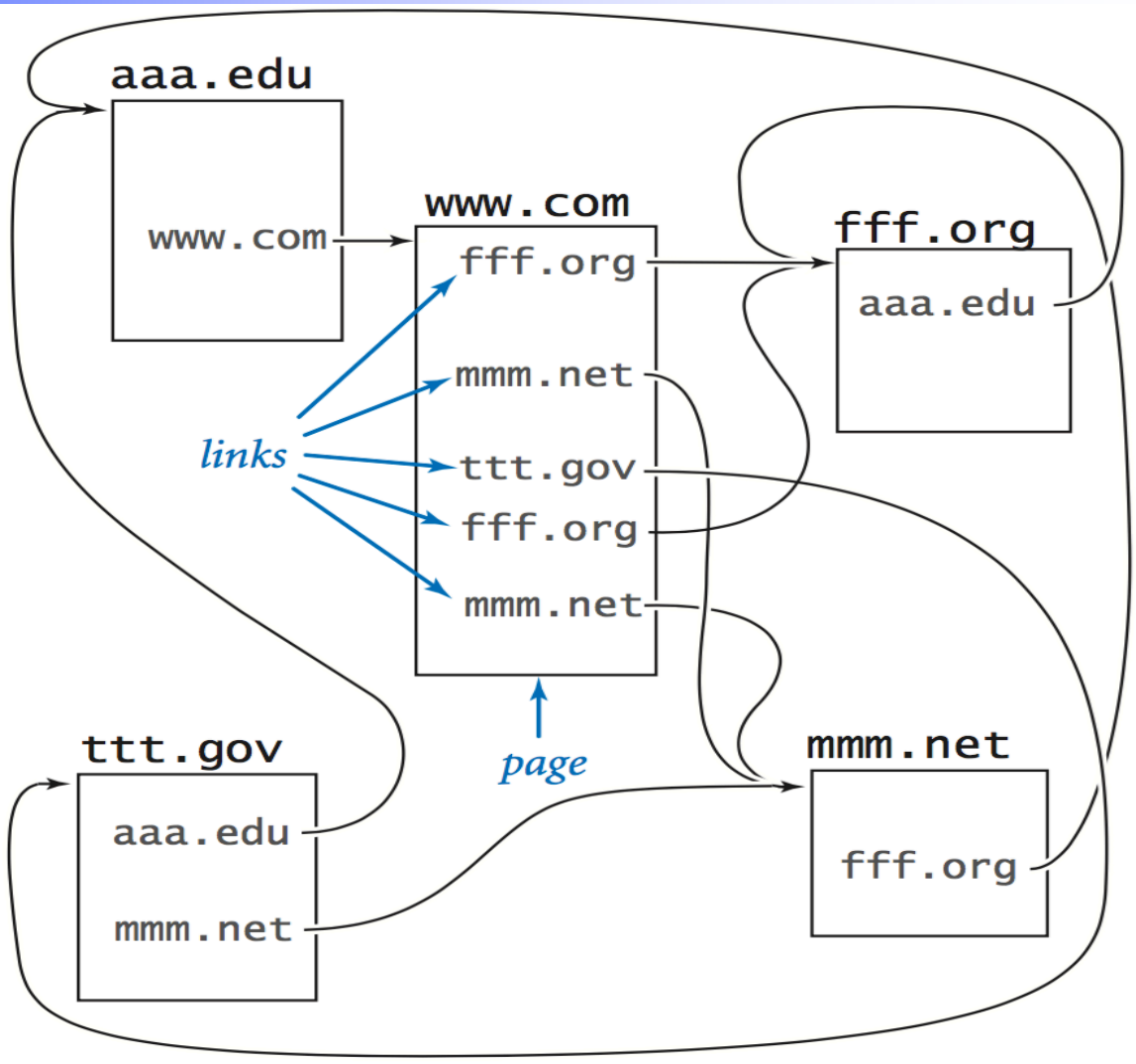
The World Wide Web



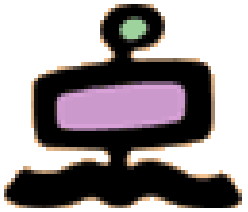
The hyperlinks

A link is made of two parts:

- the visible part (text or image)
- the link to the resource (typically a web page) to be looked for when clicking on the visible part



Evolution of the Web



- Web 1.0 (1993-2003/4)
  - Web is a “publishing medium”
  - Users (humans) can only read
- Web 2.0 (2003/4-today)
  - Web is a “social medium”
  - Users (humans) can publish and interact (e.g. Youtube, Wiki, Flickr, Facebook, etc.)
- Web 3.0 (2010/1-today, more often called IoT - Internet of Things)
  - In addition to humans, users of the Web are “programs” that can interact
  - Users of the Web are also “things”, whose programs interact with other things