Laboratorio di «**NET SECURITY E OPEN SOURCE INTELLIGENCE**»

Prof. Franco Sivilli

a.a. 2015-2016

fsivilli@unich.it

Obiettivi

Richiamare i concetti di networking e sicurezza informatica.

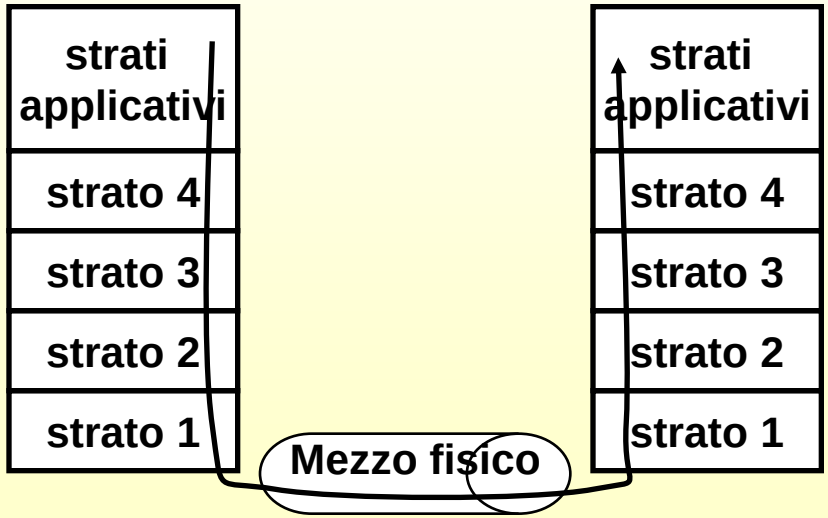
Architetture protocollari e tipologie di reti

Concetti preliminari: protocollo, strato

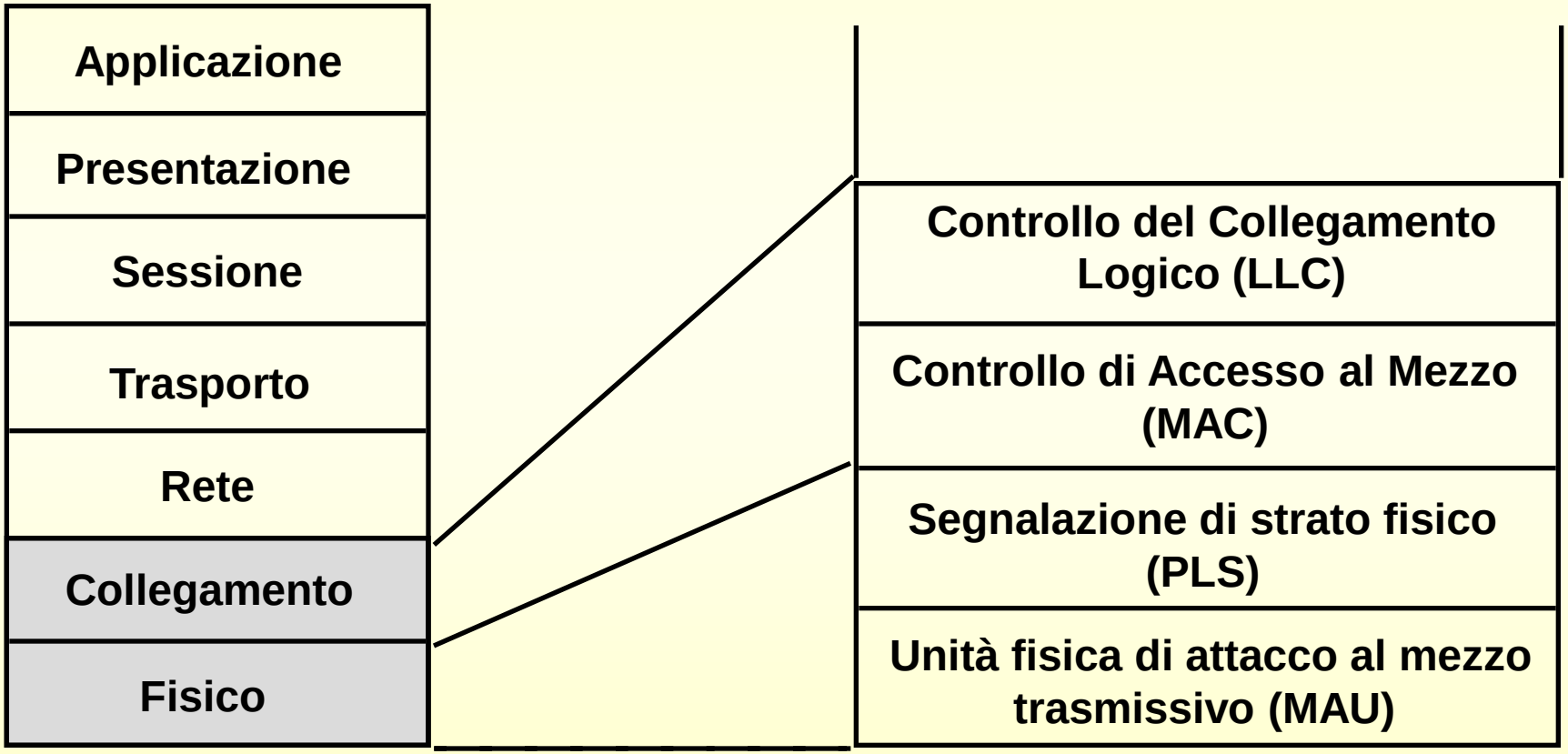
- Protocollo: insieme di regole e modalità di attuazione di una funzione o gruppo di funzioni.
- Le funzioni di un processo di comunicazione possono essere strutturate secondo un *modello a strati*.

Sistema Trasmittente
indirizzo: x1

Sistema Ricevente
indirizzo: x2

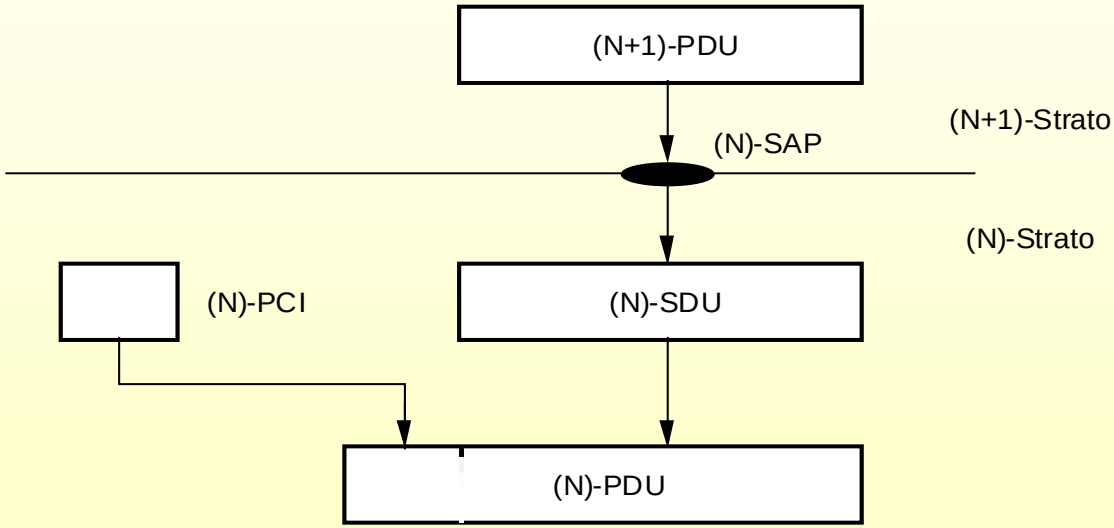


Il modello ISO/OSI



Il modello ISO/OSI

- Ogni strato tratta unità informative dette PDU (Protocol Data Unit) composte dall'informazione utile da trasferire (SDU, Service Data Unit) e da un intestazione (PCI, Protocol Control Information)
- La PDU di uno strato viene incapsulata nella PDU dello strato inferiore



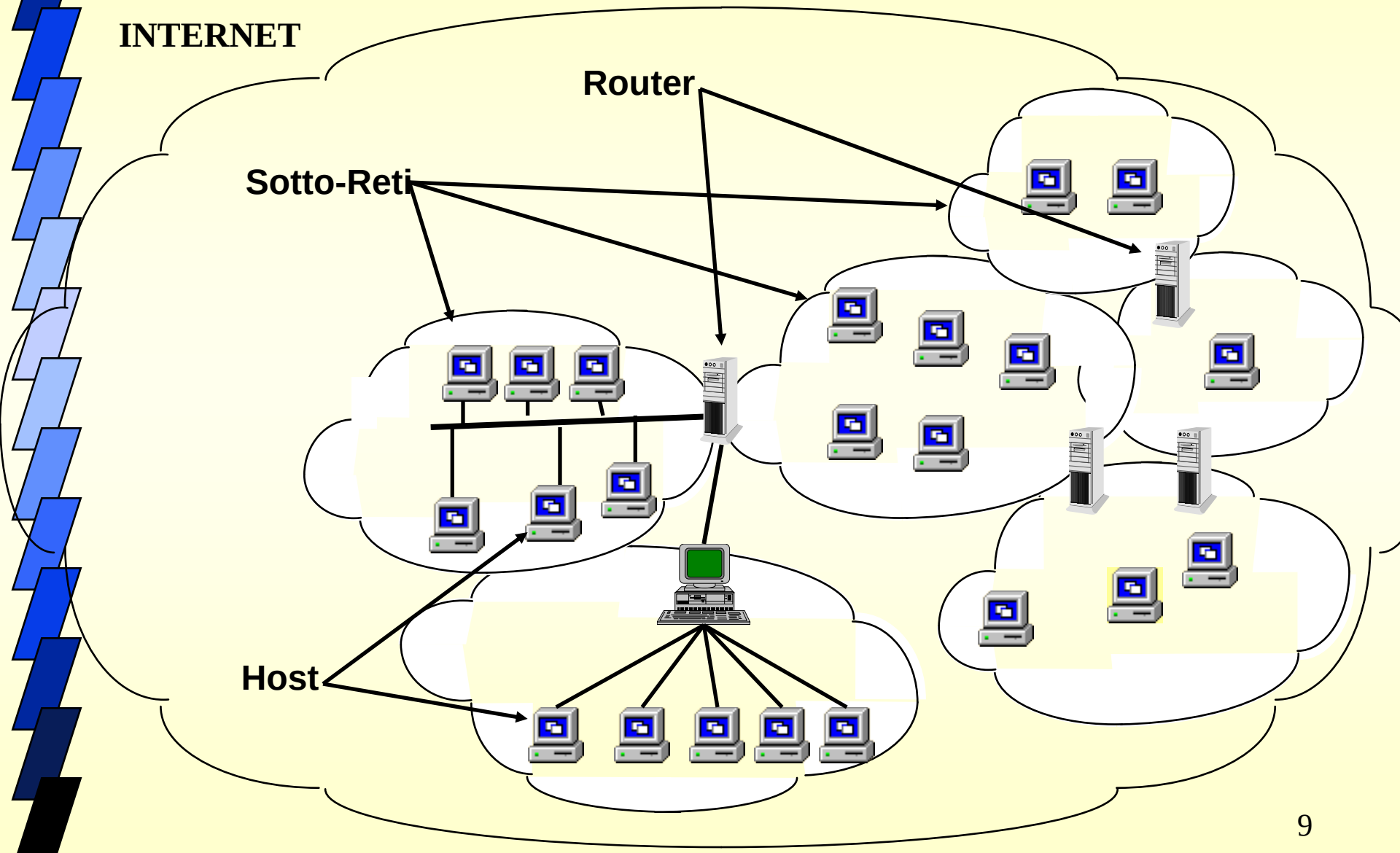
LAN, MAN E WAN

- Una LAN è l'interconnessione di apparecchiature di calcolo in area geografica limitata
- Quando l’area geografica abbraccia una città si parla di MAN, quando riguarda un’area geografica più estesa (nazione o più nazioni) si parla di WAN

Internet

- Cos'è Internet?
  - È una rete a commutazione di pacchetto costituita dall'interconnessione di reti eterogenee ed indipendenti, ognuna delle quali gestita, finanziata ed amministrata autonomamente
  - Offre un servizio di trasferimento gratuito e non affidabile (per ora…)
    - *best effort* = senza alcuna garanzia di integrità informativa e/o trasparenza temporale
  - È attualmente in crescita per estensione ($\rightarrow$ mondiale), capillarità ($\rightarrow$ home), capacità ($\rightarrow$ Gigabit/sec)

Struttura di rete



Struttura di rete

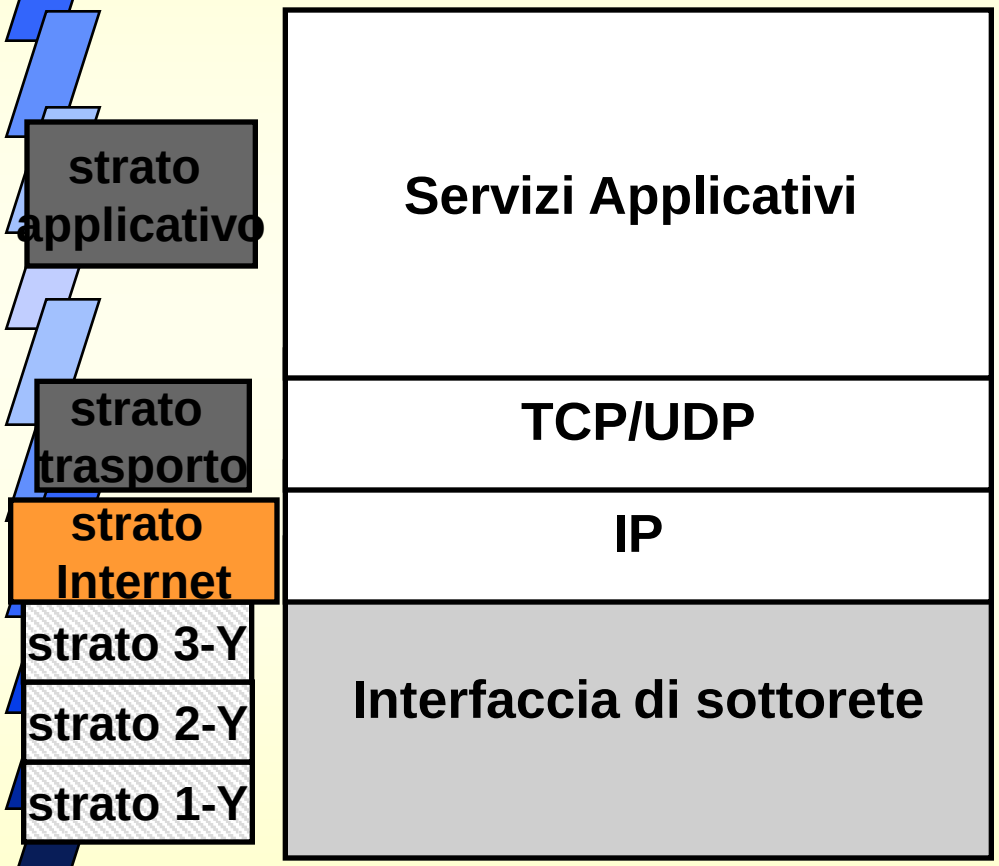
- Internet è la particolare inter-rete basata sui protocolli TCP/IP ad estensione mondiale gestita da appositi organi regolatori
- Internet non è quindi una nuova rete ma un insieme di risorse e di convenzioni per interconnettere delle reti (che sono quindi viste da Internet come sotto-reti)
- Scopo di Internet è quindi consentire a hosts (postazioni) appartenenti a sotto-reti disomogenee (per topologia, struttura fisica, modi di trasferimento e prestazioni) di comunicare tra loro

Architettura protocollare TCP/IP

TCP/IP è una suite di protocolli standard industriali progettati per le reti WAN

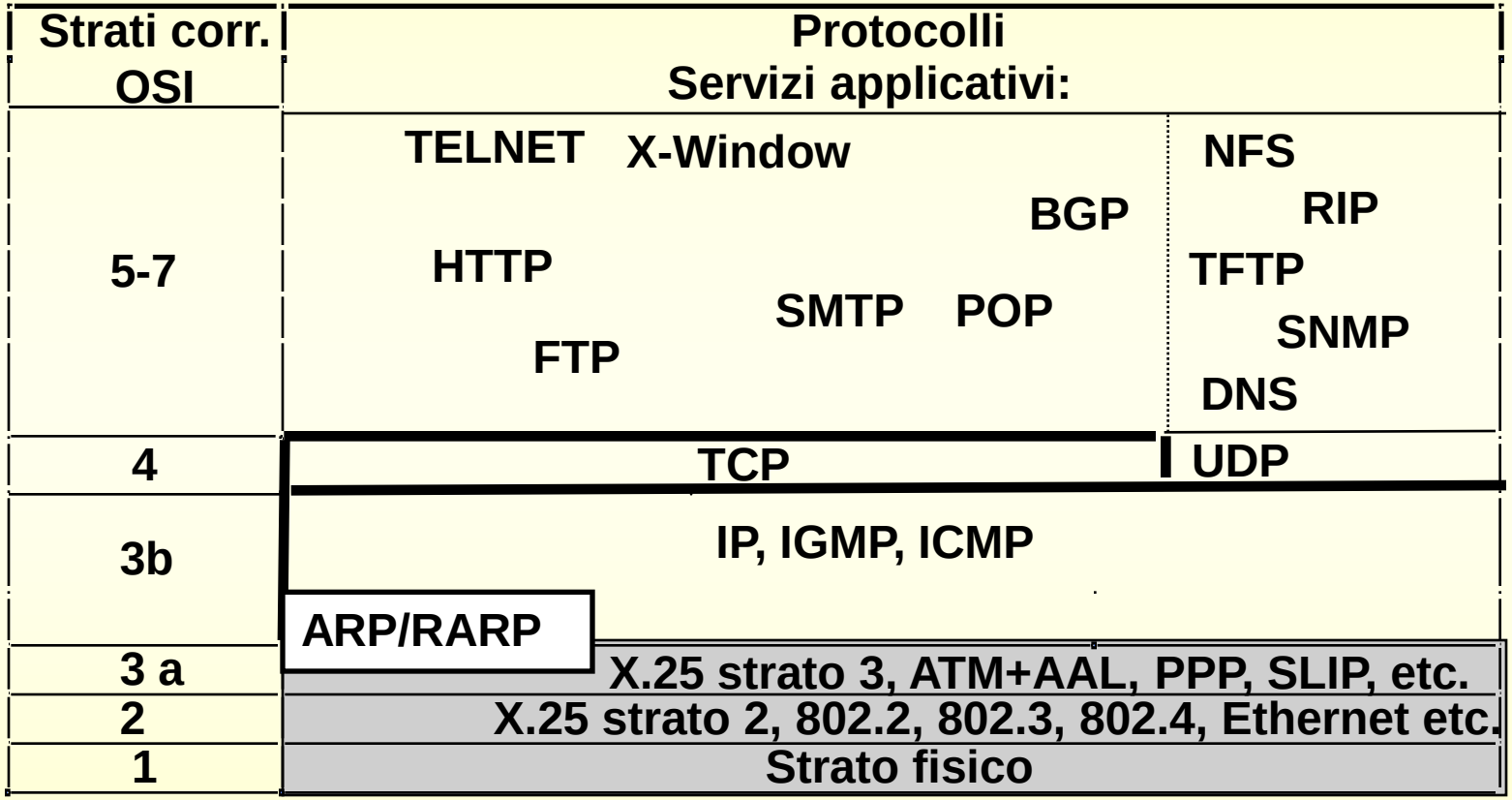
Deriva dagli esperimenti con le reti a commutazione di pacchetto condotti negli Stati Uniti dall'agenzia DARPA tra gli anni '60 e '70

Architettura protocollare TCP/IP (4 livelli)



- **I protocolli applicativi definiscono modalità e formati per lo scambio di messaggi di ogni dimensione e scopo**
- **TCP manipola flussi di dati in modo orientato alla connessione, recuperando i datagrammi persi e risequenziandoli. TCP e UDP indirizzano i singoli processi**
- **IP è responsabile dell'invio di datagrammi attraverso la Rete, ciascun datagramma è trattato indipendentemente**
- **Interfaccia di sottorete: indica genericamente i protocolli propri della sotto-rete (Ethernet, PPP,..)**

Architettura protocollare TCP/IP suite



Caratteristiche

- La struttura di rete è non gerarchica
- I protocolli TCP/IP trattano tutte le sotto-reti in modo uguale; ad esempio, ognuno dei seguenti sistemi di comunicazione è visto da Internet come una singola sotto-rete
  - una rete in area locale (es. Ethernet)
  - una rete geografica (es. la rete telefonica, rete ATM,..)
  - una connessione punto-punto dedicata (es. PPP su circuito telefonico via modem verso l'ISP)

Internet Protocol (IP)

E' il principale responsabile dell'indirizzamento e del routing di pacchetti tra host o reti. IP è connectionless.

Le funzioni principali del protocollo IP sono:

- in trasmissione
  - incapsula in datagrammi i dati provenienti dallo strato di trasporto
  - predispone l'opportuna intestazione (indirizzi src e dst,…)
  - applica algoritmo di routing
  - invia i dati verso l'opportuna interfaccia di rete
  - non richiede conferma dell'avvenuta ricezione
- in ricezione
  - verifica la validità dei datagrammi in arrivo
  - esamina l'intestazione
  - verifica se sono dati da rilanciare
  - se sono dati locali, consegna il contenuto del datagramma all'opportuno protocollo.

Schema di indirizzamento

- Internet è stata definita sistema di comunicazione universale perché consente ad ogni calcolatore di comunicare con ogni altro calcolatore
- Al tal fine è necessario stabilire un metodo globalmente accettato per identificare ed indirizzare in modo univoco tutti gli host
- Ciò ha richiesto definire un nuovo schema di indirizzamento dato che ognuna delle sotto-reti ha un suo proprio, diverso e quindi non univoco (a livello globale), schema di indirizzamento (indirizzi Ethernet, indirizzi X.25, numeri telefonici etc.)

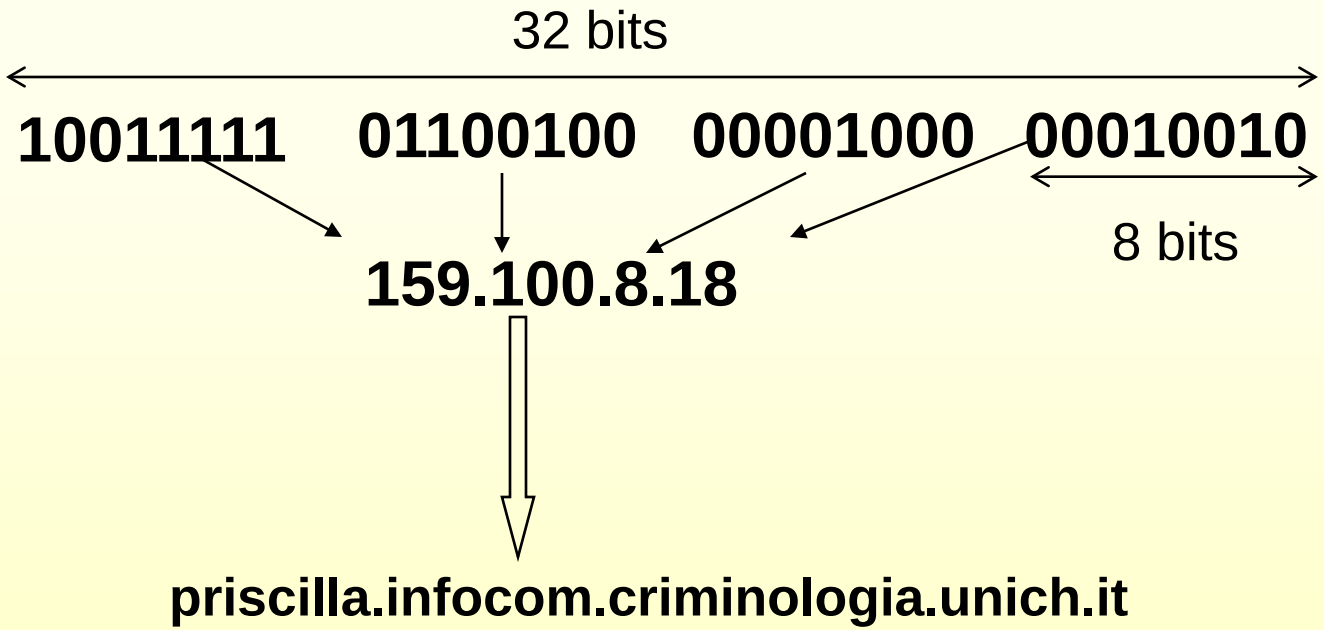
Schema di indirizzamento

- Gli indirizzi devono essere unici in tutta la rete (è possibili attribuire indirizzi arbitrari ad una sub-rete TCP/IP solo se questa non è connessa con altre reti)
- Un indirizzo IP identifica un host e non uno specifico utente. L'identificazione di un utente (in senso OSI) all'interno di un host è affidata ai protocolli di strato superiore (TCP o UDP)
- Lo schema di indirizzamento IP è stato progettato per consentire un efficiente instradamento, per una rete con dimensioni decisamente inferiori alle attuali
- Un indirizzo IP identifica prima la rete a cui un host è connesso e poi l'host all'interno di quella rete

IP_Address=Net_Id.Host_Id

Schema di indirizzamento

- **L'indirizzo IP utilizzato dal protocollo è espresso in stringhe di 32 bits (no 4 byte; ogni bit è significativo)..**
- **..che possono essere espresse in notazione puntata (dotted):**
- **a ogni indirizzo IP può essere associato un nome (DNS)**



Indirizzi IP

- Ogni host IP ha un indirizzo diviso in due parti:

  IP_Address=Net_ID.Host_ID

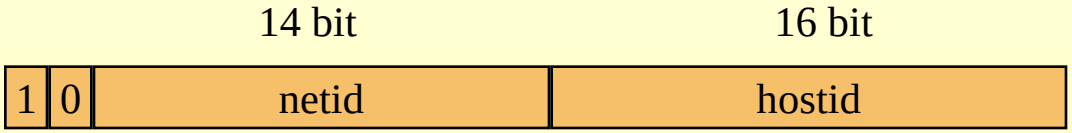
  – Host_ID identifica l'host all'interno della sottorete
  – Net_ID identifica la sottorete su Internet

Sono state inizialmente definite 5 classi di indirizzi:

Classe A **(0.x.x.x→127.x.x.x)**

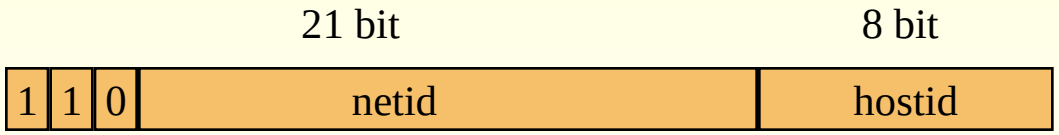
127.0.0.0 riservato

Classe B **(128.x.x.x→191.x.x.x)**

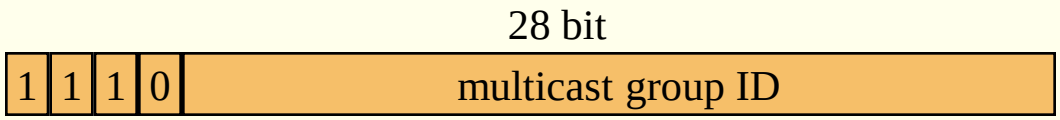


Indirizzi IP

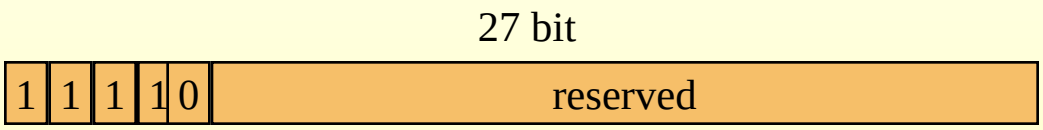
Classe C **(192.x.x.x - 223.x.x.x)**



Classe D (per multicast) **(224.x.x.x - 239.x.x.x)**



Classe E (per sperimentazione) **(240.x.x.x - 255.x.x.x)**



Le classi si distinguono dai primi bit del primo ottetto

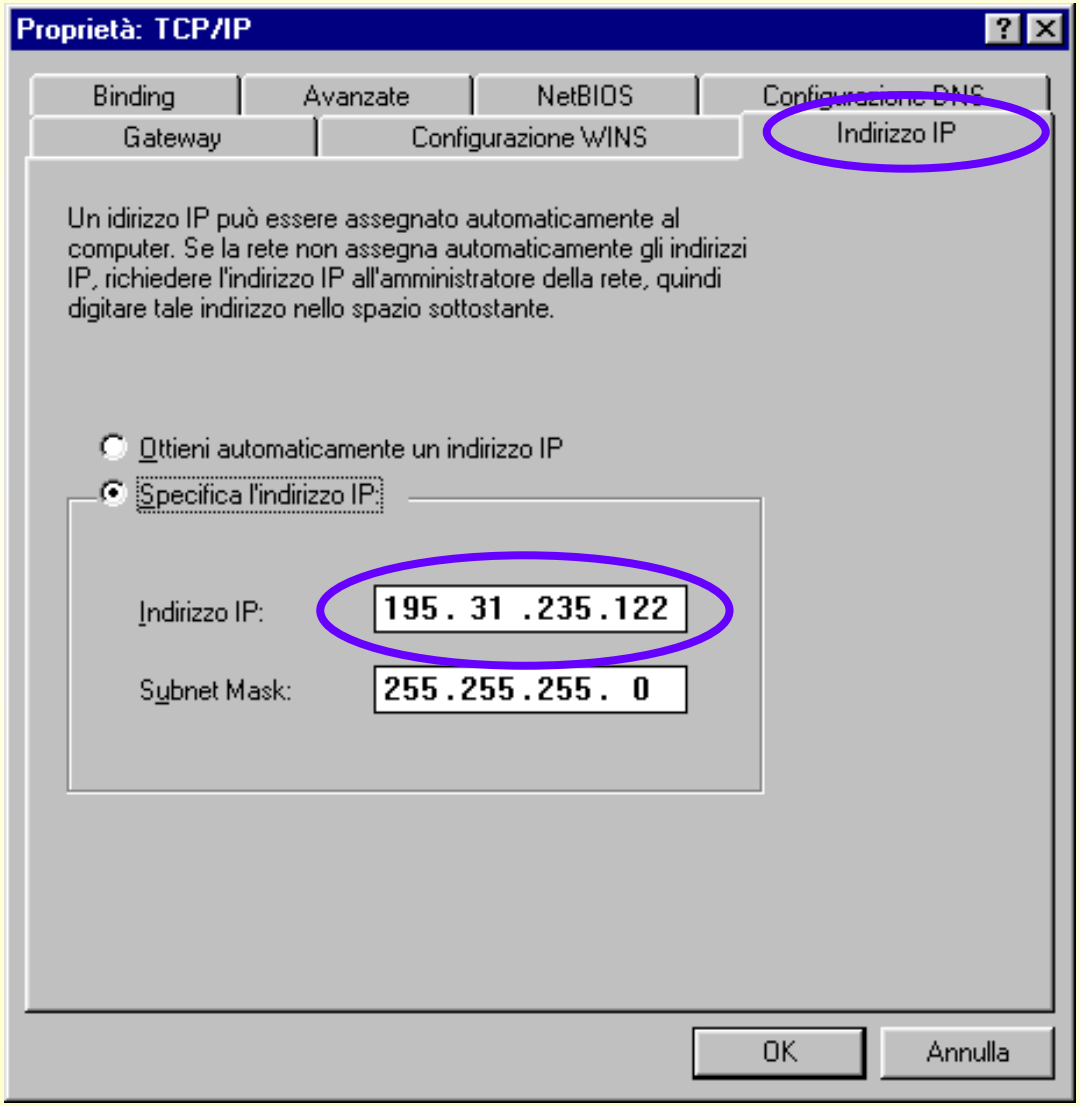
Parametri di configurazione

Indirizzo IP → identifica l'host

Maschera di sottorete→ se l'host appartiene alla rete locale o ad una remota (l'AND tra IP address e subnet mask fornisce l'indirizzo di rete)

Default Gateway→per comunicare con un host di altra rete (router IP)

Configurazione Windows indirizzo IP



Router IP

Il router è un apparato attivo di rete di livello 3 (rete) che ha il compito di instradare il traffico IP.

Per instradare un router deve:

- Conoscere l'indirizzo di destinazione
- Identificare le sorgenti da cui può imparare il percorso di destinazione
- Scoprire i possibili percorsi
- Selezionare il miglior percorso
- Aggiornare i percorsi conosciuti (tabella di routing)

Router IP

Le tabelle di routing possono essere

- Statiche (impostate manualmente)
- Dinamiche (il router “impara” i percorsi migliori scambiando con altri Router le sue tabelle di routing)

Router IP

- I router IP ricevono datagrammi IP da un'interfaccia e li inoltrano su un'altra
- Si distinguono dagli Host perchè:
  - hanno in genere più di un'interfaccia
  - utilizzano "protocolli di routing" più sofisticati (RIP, IGRP, OSPF)
- I router IP hanno (normalmente) un indirizzo IP per ogni interfaccia.

Esaurimento degli indirizzi IP

Il progressivo esaurimento degli indirizzi IP unitamente alla rapida crescita delle dimensioni delle tabelle di routing ha spinto l'IETF (*Internet Engineering Task Force*) ad intraprendere delle azioni preventive

Tali misure preventive possono essere raggruppate nelle seguenti categorie:

- Assegnazione razionale degli indirizzi IP (InterNIC)
- Indirizzi privati e *Network Address Translation* (proxy NAT ovvero ogni indirizzo privato esce con un unico indirizzo pubblico, se non ho proxy ogni privato viene tradotto in un pubblico diverso, questìultimo è un meccanismo di sicurezza)
- PAT (Port Address Translation): molti IP privati sono tradotti in unico IP pubblico associando a ciascuno dei molti IP anche una porta diversa da IP a IP. Le porte devono essere superiori a 1024 in quanto quelle <1024 sono le porte Server. Con il PAT sono protetto in quanto i servizi sono su porte <1024, mentre il destIP è >1024.
- IP versione 6 (IPv6). Utilizza 16 ottetti (128 bit), ciascuno in esadecimale, divisi in 8 coppie. Es.4A3F:AE67:F240:56C4:3409:AE52:440F:1403

Ip v.6 :innovazioni

Stateless autoconfiguration o “plug & ping”: il terminale si autoconfigura anche senza la presenza di un server DHCP

Spazio infinito di indirizzi: possibilità di utilizzare l’Ip come un numero telefonico assegnato alla persona (256 indirizzi per ogni persona del pianeta!!)

IPV6 formalizza 3 tipi di indirizzi:

- unicast: singola interfaccia (Es. NIC del PC)
- Multicast: insieme di interfaccie a cui inviare un insieme di pacchetti;
- Anycast (nuovo): una qualunque interfaccia (una sola) di un insieme di interfacce. Chi invia lascia al sistema di instradamento il compito di scegliere l’interfaccia più vicina.

Ip v.6: problematiche

Compatibilità con IPv4: Tunneling (incapsulare IPv6 in IPv4) o Traduzione

Aggiornamento delle tabelle di routing dei Router

L’Ipv6 è la premessa per l’Internet of things

Una unica rete a cui collegare il cellulare e gli elettrodomestici di casa (frigorifero, lavatrice, forno).

Convergenza tra le ICT e domotica.

TCP e UDP

TCP è un servizio di trasmissione **affidabile** (l'host ricevente manda un ACK entro un tempo prestabilito per ogni segmento ricevuto) **orientato alla connessione** (viene stabilita una sessione tra host).

UDP è un servizio di trasmissione senza connessioni che non garantisce la trasmissione dei pacchetti. Viene utilizzato da applicazioni che non richiedono il riconoscimento della ricezione dei dati (Voice over IP).

I protocolli applicativi di Internet

- Il modello Client-Server
- DNS
- FTP, Telnet
- La posta elettronica
  - SMTP, POP3
- World Wide Web e HTTP

Il paradigma client/server

CLIENT **SERVER**

programma **client**:

- quando necessario, si connette al server (calcolatore) sulla porta specifica associata al server (programma)
- invia dei messaggi composti secondo il protocollo di comunicazione
- aspetta i risultati

programma **server**:

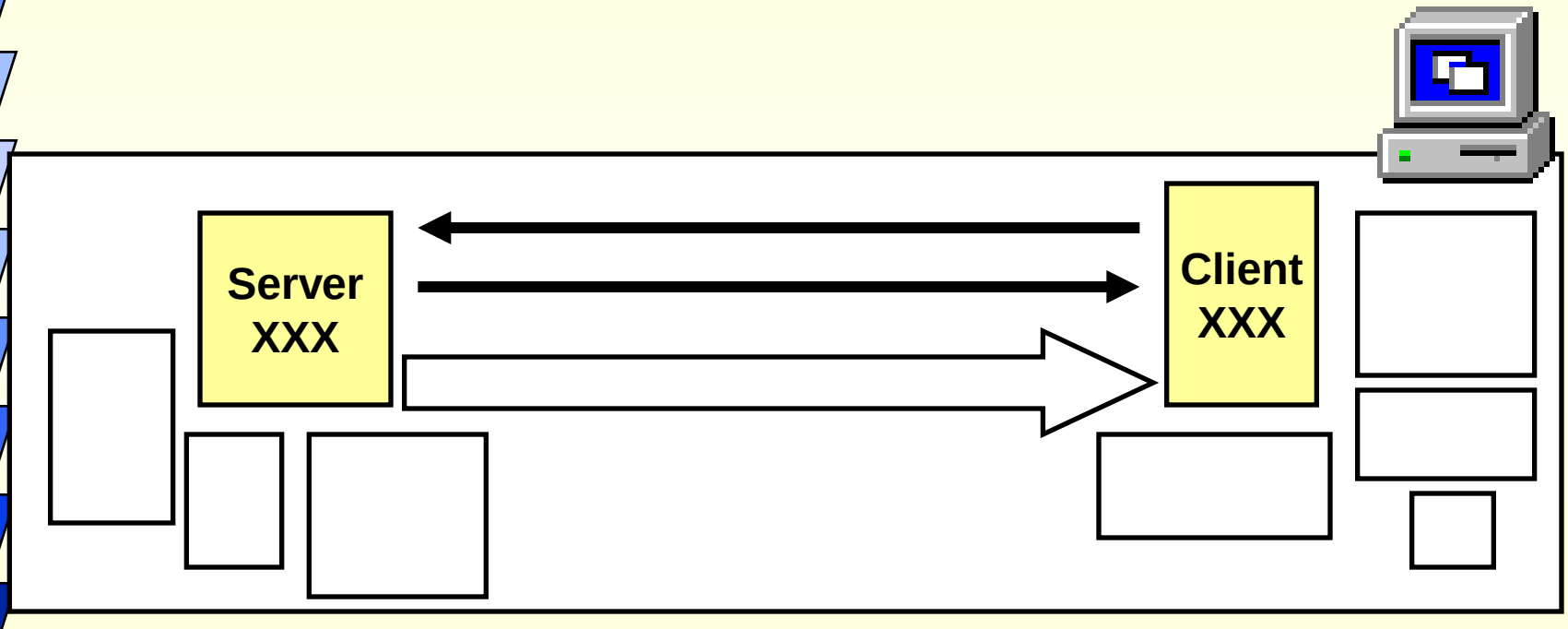
- è sempre attivo (*daemon*)
- "ascolta" i messaggi in arrivo su una porta
- li interpreta (grazie al protocollo) ed effettua il servizio richiesto
- rispedisce indietro i risultati

I protocolli di Internet

- come accennato, sono “stratificati”
- i protocolli applicativi (ma non solo) rispettano tutti lo stesso principio “client/server”
- definiscono:
  - il formato dei messaggi scambiati
  - il significato di alcune parti del messaggio (es. richiesta servizi)
- sono descritti negli "RFC" (Request for Comments), liberamente disponibili su Internet

Modello Client / Server

- Le componenti Client / Server possono risiedere nello stesso sistema



- Domain Name System

Domain Name System

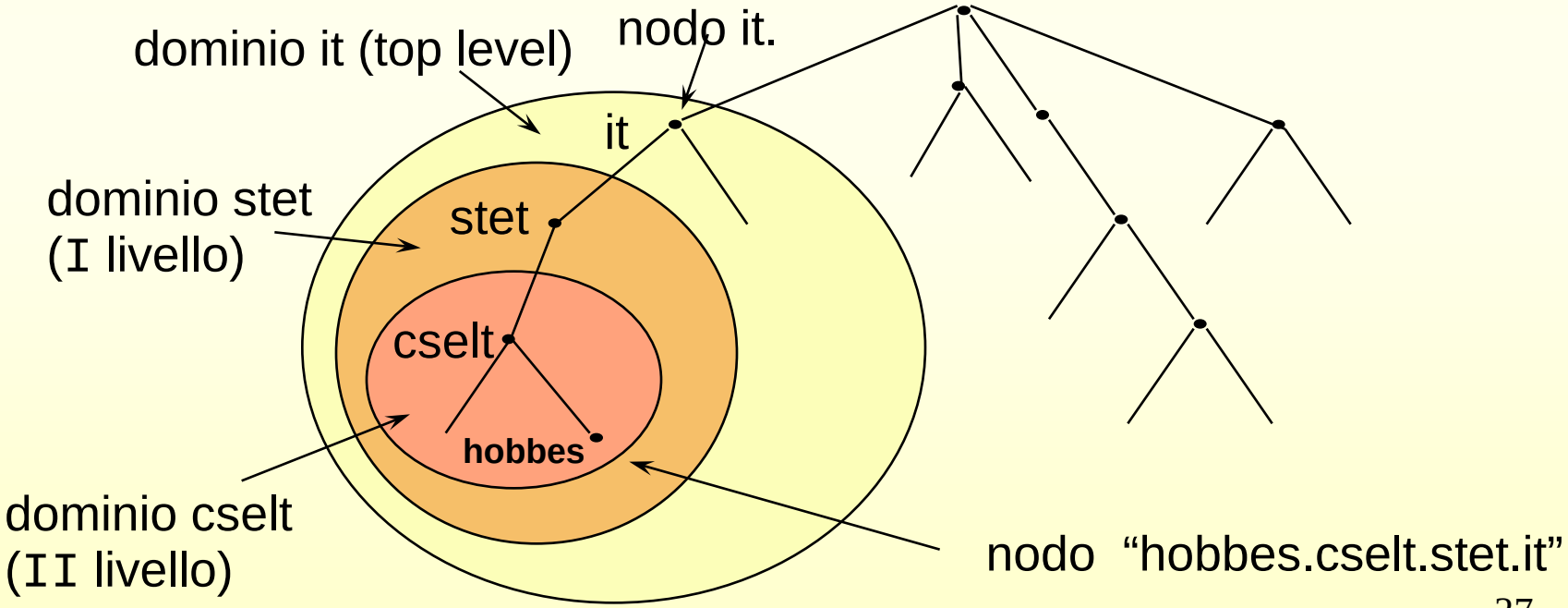
- Il software implementato in Internet consente di utilizzare oltre alla notazione dotted anche un altro tipo di notazione (mnemonica):

  “151.100.8.18”=“aulainfo.economia.unich.it”

- E’ necessario che un opportuno software di rete traduca nomi in indirizzi e viceversa (il passaggio da notazione dotted a indirizzo di 32 bits è banale in quanto implica una semplice conversione decimale-binario)
- Questa traduzione è attuata da un protocollo di alto livello implementato in un meccanismo noto come Domain Name System (DNS)

Domain Name System

- In Internet i nomi sono organizzati gerarchicamente in Domini
  - I nomi sono costituiti da stringhe separate da “.”
  - La parte più significativa è a destra



Attribuzione dei nomi

- L’insieme dei nomi è prima partizionato in un certo numero di sotto-insiemi dal Network Information Center (NIC); il compito di assegnare i nomi all’interno di un sotto-insieme è delegato ad un’autorità di livello inferiore e così via
- Un nome è composto da una serie di sotto-nomi separati da un punto. Ogni punto separa un’autorità da quella che gli è gerarchicamente inferiore

aulainfo.sociologia.unich.it

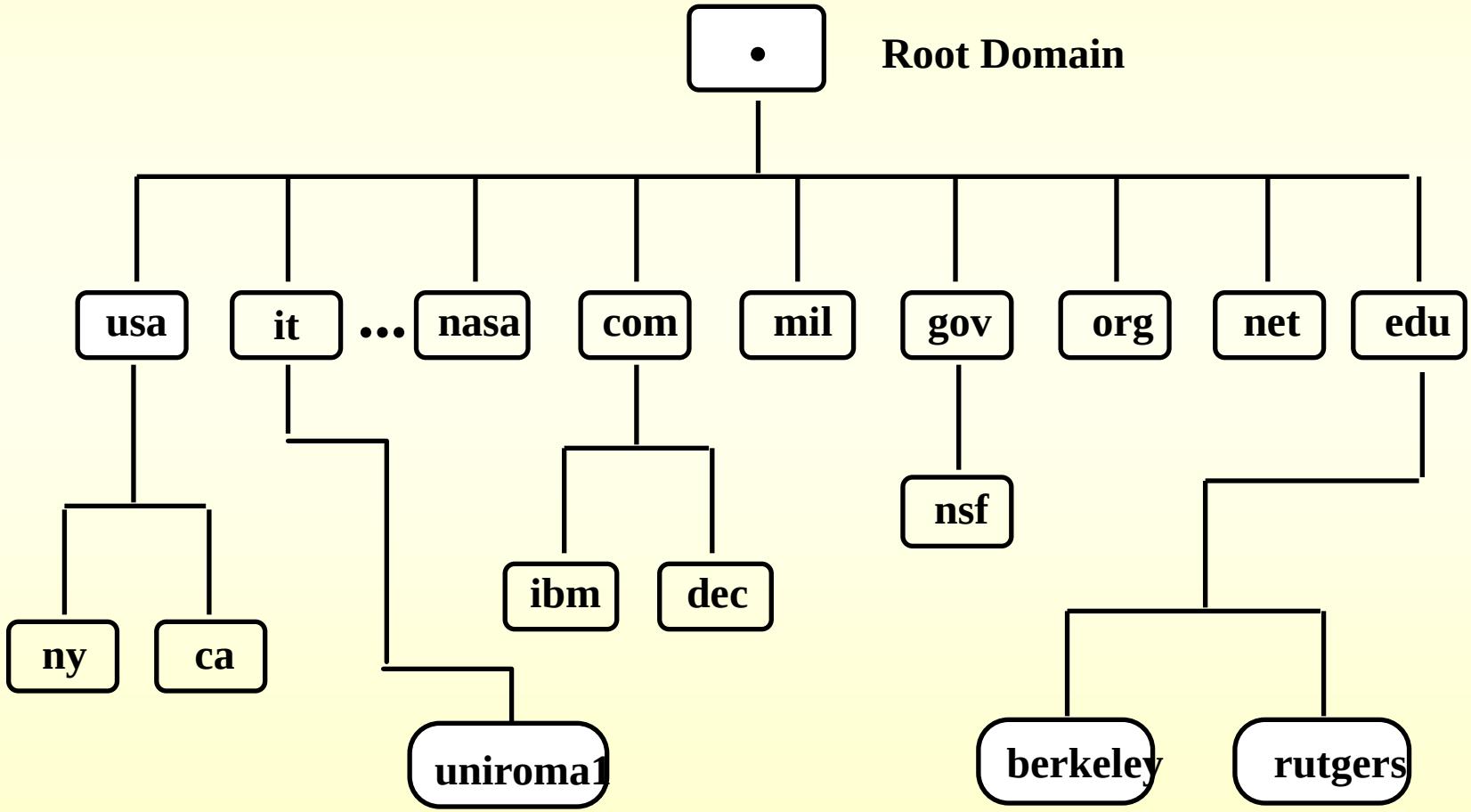
Attribuzione dei nomi

classificazione per tipologia	
Nome del dominio	**Tipo di organizzazione**
COM	**Commerciali**
EDU	**Accademiche e didattiche**
GOV	**Statali**
MIL	**Militari**
NET	**Centri di Gestione di Internet**
ARPA	**ARPANET (obsoleto)**
INT	**Organizzazioni internazionali**
ORG	**Altre organizzazioni**
FIRM	**Aziende, affari**
STORE	**Merce in vendita**
WEB	**enfatizzante WWW**
ARTS	**enfatizzante arte e cultura**
REC	**enfatizzante intrattenimento e divertimenti**
INFO	**enfatizzante fornitori di informazione**
NOM	**enfatizzante nomenclature personali**

Attribuzione dei nomi

classificazione geografica							
Nome del dominio	**USA**	**IT**	**DE**	**FR**	**UK**	**JP**	**etc.**
nazione	**USA**	**Italia**	**Germ.**	**Franc.**	**G.Br.**	**Giapp.**	**etc.**

Attribuzione dei nomi

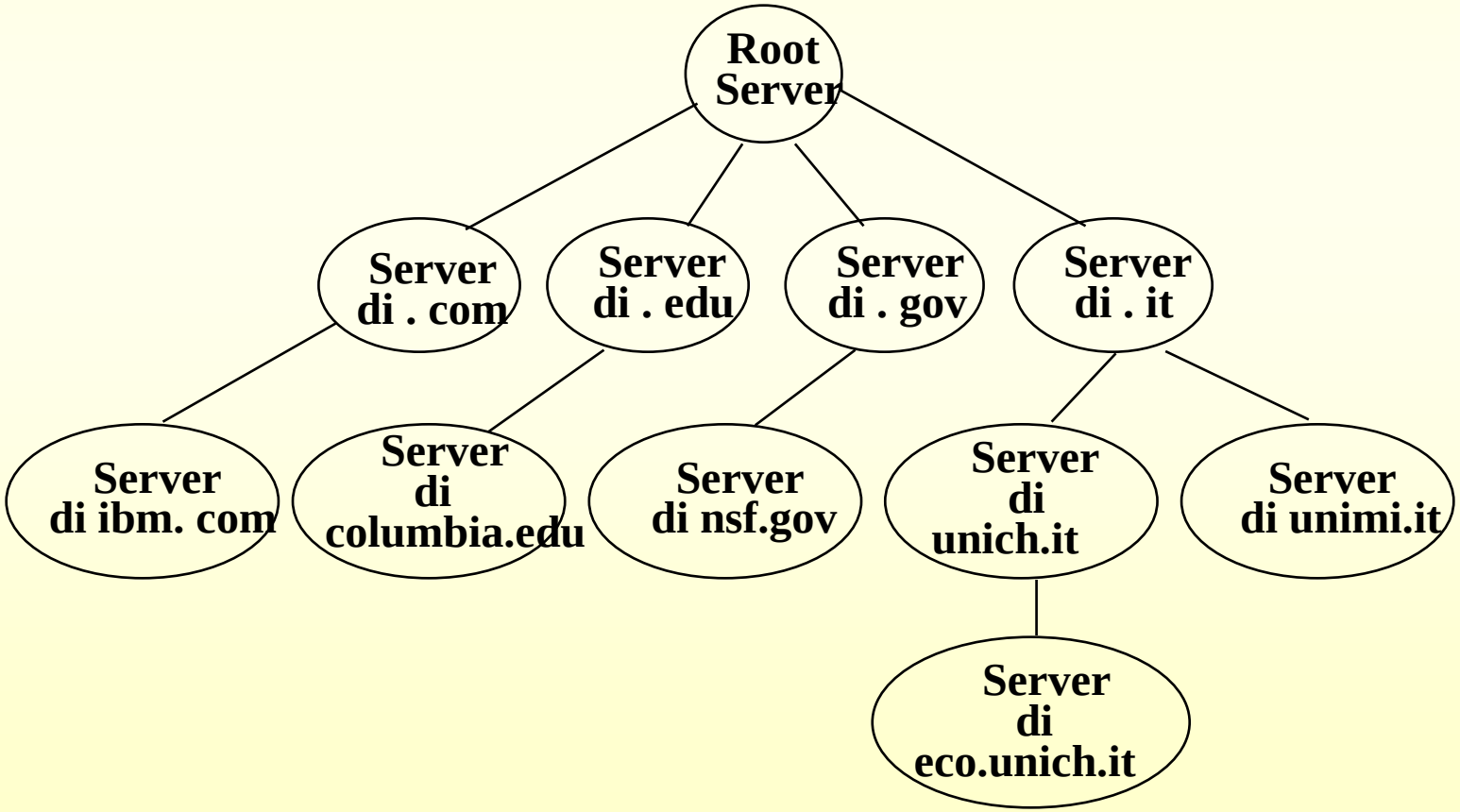


Traduzione dei nomi in indirizzi

- DNS include un efficiente ed affidabile algoritmo distribuito per tradurre nomi in indirizzi
- E’ distribuito in quanto è costituito da una molteplicità di servers che co-operano tra loro
- E’ efficiente in quanto molti nomi possono essere tradotti localmente senza generare traffico in Internet
- DNS è costituito da un certo numero di sistemi indipendenti e co-operanti chiamati name servers

Traduzione dei nomi in indirizzi

La risoluzione di un indirizzo avviene in modo top-down, iniziando dalla radice dell'albero e procedendo lungo i servers di livello gerarchico inferiore



Traduzione dei nomi in indirizzi

- L'algoritmo appena descritto ha tre svantaggi:
  - la gran parte delle richieste fa riferimento a nomi locali, risalire ogni volta fino al root server è inefficiente
  - il root server è sottoposto ad un carico di elaborazione molto rilevante (anche se più calcolatori lavorano in parallelo per svolgere tale compito)
  - un guasto del root server o di server di alto livello pregiudicherebbe il funzionamento dell'intero DNS

Traduzione dei nomi in indirizzi

- Per ovviare a questi problemi l'algoritmo è stato integrato con delle funzionalità dette di "cache"
- Ogni server memorizza i nomi che è riuscito a risolvere insieme all'indirizzo del name server che ha operato la traduzione
- Se gli viene richiesta di nuovo la stessa traduzione non ha bisogno di rivolgersi nuovamente al root server. Tale meccanismo funziona a tutti i livelli gerarchici
- ogni traduzione comprende l'indicazione di un TimeToLive (es. 3 giorni) allo scadere del quale l'associazione NOME-INDIRIZZO viene automaticamente cancellata.

Traduzione dei nomi in indirizzi

- Se, ad esempio, un host locale vuole risolvere l'indirizzo dell'host “sivilli.aulainfo.psicologia.unich.it”
  - per prima cosa controlla nella sua cache
  - se non lo trova interroga il name server della sua zona, il quale cerca nella sua cache.
  - Se non lo trova cerca, sempre nella cache, un server di uno dei sub-domini dell'host cercato. Nell’ordine: “aulainfo.psicologia.unich.it”, “psicologia.unich.it”, “unich.it”,...”
  - e da questo avvierà una ricerca iterativa
  - alla fine ritornerà la traduzione all’host, il quale aggiornerà la sua cache
- Il name server memorizzerà nella sua cache l’indirizzo ottenuto e tutti quelli incontrati in questo iter per future eventualità.

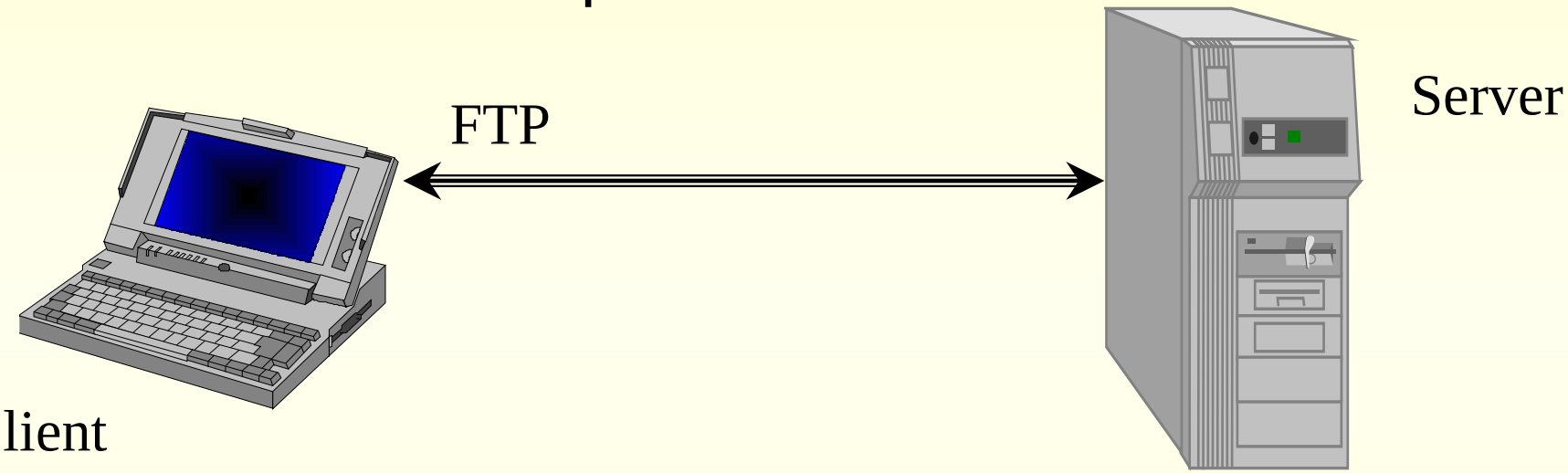
Questioni sull'indirizzamento

...

- Tre livelli di indirizzamento**:**
  - **Indirizzi MAC (6 bytes per reti IEEE 802)**
  - **Indirizzi IP (4 bytes)**
  - **Indirizzi DNS (nome degli host)**
- URL **(*Uniform Resource Locator*), il metodo di indirizzamento usato nel WWW, generalizza i nomi degli host DNS**
- Overlay network model**: nessun mapping diretto o algoritmico tra indirizzi a differenti livelli (a parte nomi di host ed URL); da qui la necessità di un protocollo di risoluzione degli indirizzi:**

Il servizio FTP

protocollo/software per il trasferimento di file



Due modalità di accesso:

- full service (username + password)
- anonymous FTP

FTP

- protocollo per il trasferimento di files tra sistemi remoti
- è uno dei protocolli applicativi più vecchi
- permette la gestione locale del file system del sistema remoto tramite una serie di comandi per la navigazione nel file system, ed il trasferimento dei files in entrambe le direzioni
- si basa su convenzioni tipiche di Unix (file system, comandi e loro output)
- rispetta i privilegi di accesso ai files del sistema remoto
- è un protocollo a connessione persistente (a differenza di HTTP...)

FTP

- Utilizzato per trasferire file quando è già nota la loro posizione
- permette:
  - trasferimento di file in blocco
  - autenticazione
  - accesso interattivo
- architettura client-server: molti clients possono connettersi ad un server

FTP per la condivisione di risorse

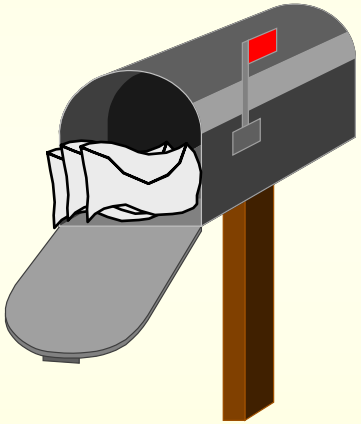
- ftp è stato il primo metodo per distribuire pubblicamente risorse informatiche, e tutt'ora rappresenta uno dei modi più diffusi per la realizzazione di grossi depositi di file
- Come?
  - Sui sistemi ove si vogliano rendere disponibili dei files si definisce un'area pubblica (cioé una porzione di file system, di solito `/pub`)
  - l'utente esterno può accedere con un nome particolare per vedere quell'area:
    - nessun nome (non tutti i sistemi accettano)
    - utente `anonymous` (di solito per controllo chiedono il nome utente reale nel campo password)

TELNET

- **Fornisce un “terminale virtuale” su una rete TCP/IP**
- ***E’ un PROTOCOLLO* che fornisce una “comunicazione ad 8 bit bidirezionale”**
- ***E’ anche una APPLICAZIONE che supporta il protocollo TELNET***
- **connessione TCP client-server (server ascolta sulla porta 23)**
- **I flussi di controllo e dati viaggiano sulla stessa connessione**
- **alla connessione vengono negoziate alcune opzioni**

La posta elettronica

- Caratteristiche:
  - velocità
  - versatilità
  - economicità
  - Indipendenza dal tempo e dallo spazio

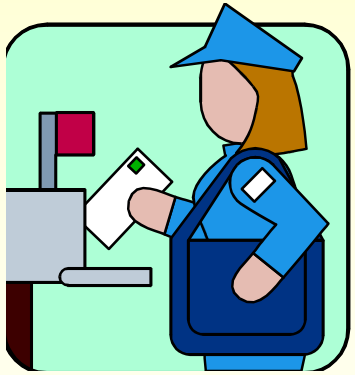


gli strumenti necessari

- Mailbox (casella postale)
- indirizzo posta elettronica

- PC connesso ad Internet
- programma “client” sul PC

La mailbox



E’ il contenitore elettronico dei messaggi ricevuti

- Normalmente risiede su un calcolatore potente e sempre connesso alla rete
- Ha associato un indirizzo di posta elettronica

Come leggere ed inviare la posta?

- ogni utente ha una sua **mailbox** (casella di posta elettronica) che corrisponderà ad uno spazio nel file system; in quello spazio il server SMTP locale, salverà i messaggi ricevuti per lo specifico utente che lo stesso utente per il tramite del suo client di posta (Outlook, Eudora etc) andrà a leggere. In questo caso il client di posta dell'utente si connetterà al server POP3 o IMAP per leggere tali messaggi. Il Client di posta si preoccuperà anche di inviare la posta in uscita al Server SMTP.

  NOTE

  - POP3: permette di scaricare i messaggi come tali, senza funzionalità di gestione
  - IMAP: accesso alla mailbox, i messaggi rimangono sul server e sono quindi accessibili da più sistemi

Messaggio e-mail

- È indipendente da come l'utente compone il messaggio!
- formato base: **RFC 822**
- testo ASCII organizzato genericamente in righe logicamente suddiviso in:
  - **Intestazione** (header)
    - mittente (From:)
    - destinatario (To:)
    - oggetto (Subject:)
    - ...
  - **Corpo** del messaggio
    - testo qualsiasi, separato dall'intestazione con una linea vuota, e terminato da un punto singolo

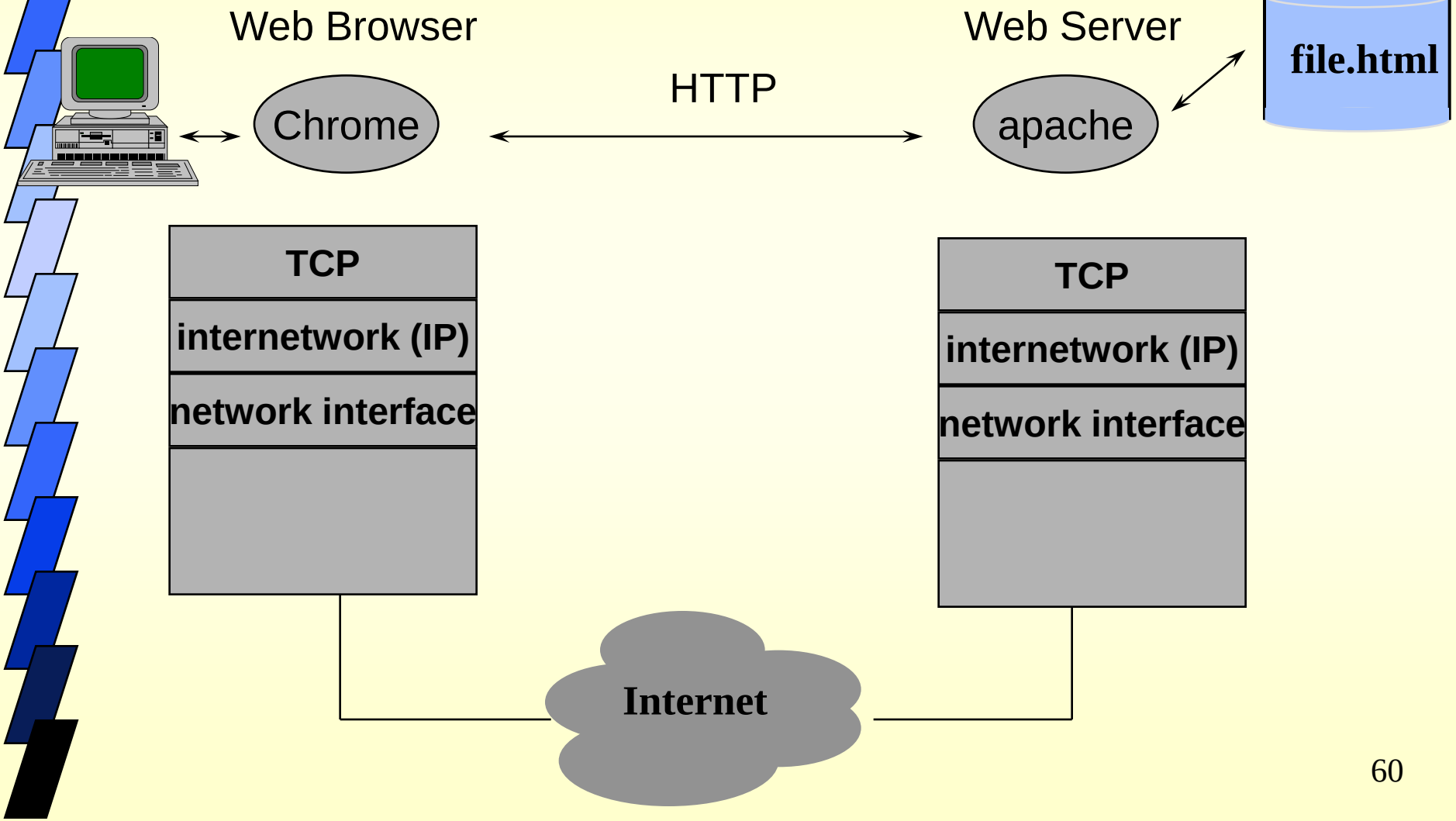
La modalità Webmail

- Molti Provider offrono l'accesso alla mailbox in modalità Webmail. In questo caso l'utente necessita di un browser per collegarsi, via http al sito del provider (Es. www.virgilio.it), lì troverà un apposito spazio della pagina dove inserire nome utente e pwd. L'accesso alla propria mailbox, dal pdv dell'utente, avviene mediante accesso ad un web server e non ad un server di posta (SMTP o POP). In realtà sarà il Web Server ad intercedere con l'SMTP Server per l'invio della posta ed con il POP3 Server per poter leggere la posta. A quest'ultimo verranno passate le credenziali (user e pwd) inserite dall'utente attraverso il suo browser.

Un altro protocollo applicativo di Internet: HTTP

- principale protocollo utilizzato nel World Wide Web
- utilizzato per la realizzazione di sistemi ipermediali distribuiti
- è essenzialmente un protocollo per il trasferimento di file
- porta standard: 80

WWW: HTTP


Web Browser
Web Server
file.html
HTTP
Chrome
apache
TCP
internetwork (IP)
network interface
TCP
internetwork (IP)
network interface
Internet

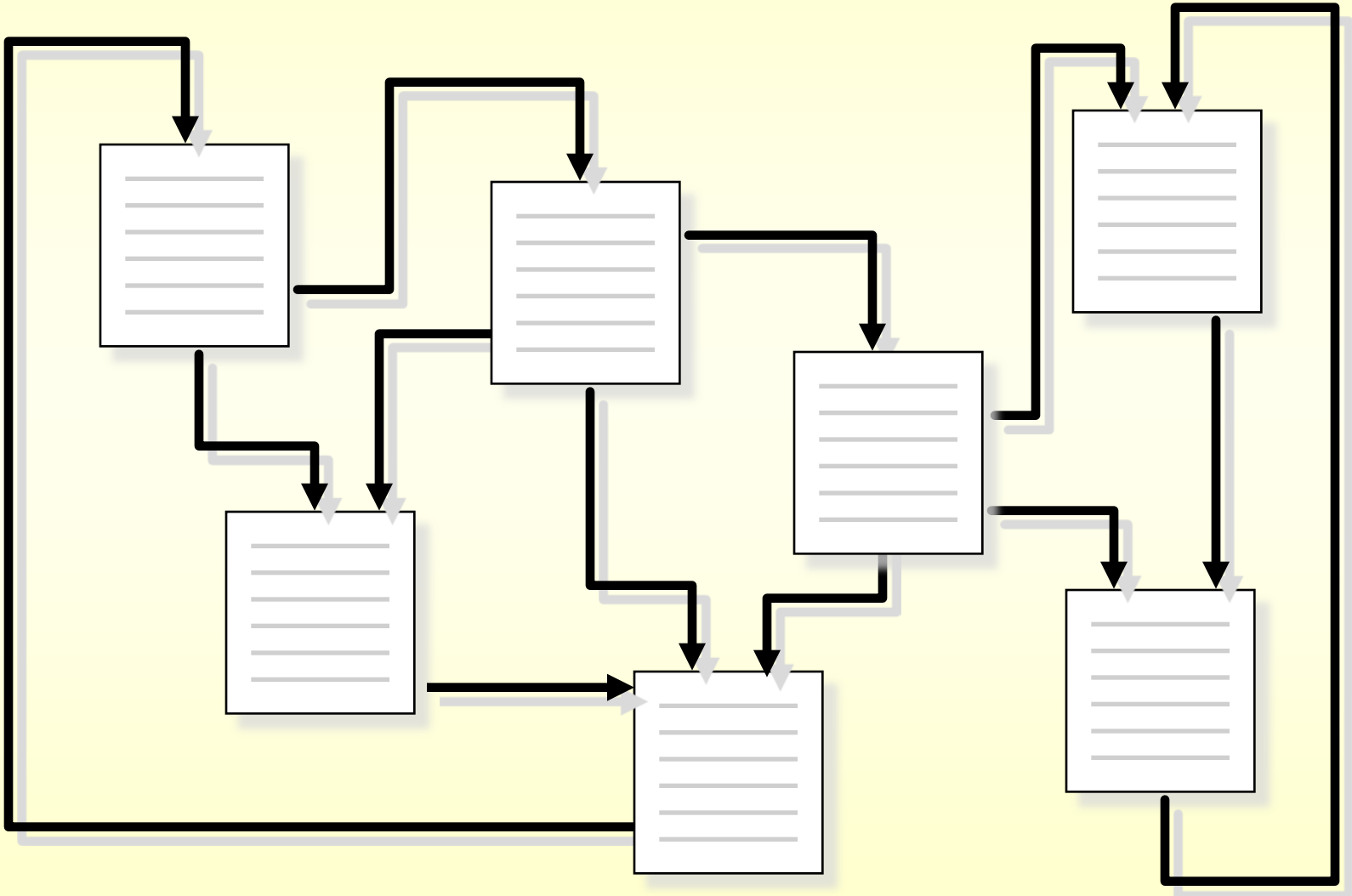
World Wide Web (Web, WWW, w^3)

- è un unico ipertesto distribuito sulla rete Internet
- si basa su diversi formati di memorizzazione delle informazioni, e qualsiasi tipo di informazione
  - quello che dà le caratteristiche ipertestuali è HTML
- utilizza diversi protocolli applicativi di Internet (potenzialmente tutti)
  - quello principalmente usato è HTTP
- Storicamente:
  - nasce nel 1989 al CERN per la collaborazione su progetti di ricerca internazionali di fisica, con condivisione di documenti (di testo)
  - 1991: prima demo pubblica, solo testo
  - 1993: prima interfaccia grafica (NCSA Mosaic) -> successo
  - 1994: nasce Netscape, CERN e MIT fondano il W3 Consortium
  - 1995: Netscape, sebbene in rosso, raccoglie investimenti

WWW: cos'è? (per l'utente...)

- collezione di documenti (testo, immagini, etc) che risiede su calcolatori sparsi in giro, e connessi tramite Internet
- I documenti vengono chiamati **pagine**
- ogni pagina può contenere dei collegamenti (**link**) ad altre pagine correlate, situate ovunque
- l'utente può seguire questi collegamenti, cliccando sui link e muovendosi di pagina in pagina
- Le pagine sono viste grazie ad un programma detto **browser**, che recupera e mostra le pagine, interpreta le richieste dell'utente, etc.

Ipertesto



Multimedia/Ipermedia

- i documenti multimediali comprendono informazioni provenienti da sorgenti diverse: testo, immagini, audio, video, ...
- i documenti ipermediali sono documenti multimediali con capacità ipertestuali, ovvero con possibilità di accesso non lineare
- un nodo di un documento ipermediale sarà quindi costituito da testo, immagini, etc., ed un link potrà puntare a frammenti costituiti da questi particolari tipi di documenti

WWW: ipermedia distribuiti

- WWW è un documento ipermediale distribuito:
  - distribuito perché lo spazio di memorizzazione fisica è costituito dai file system di computer siti ovunque e connessi ad Internet
  - ipermediale perché i nodi possono essere costituiti da qualsiasi tipo di documento, il cui "tessuto" ipertestuale è dato da HTML
  - i nodi ed il modo di accesso sono identificati da nomi appositi (**URL**)
- si basa sul principio client/server
  - il client si occupa della visualizzazione dei nodi, e dell'interpretazione delle richieste di navigazione dell'utente
  - i dati sono invece mantenuti e distribuiti da appositi server in modalità diverse

Alcune caratteristiche di WWW

- è un grande ipertesto
  - costituito da decine di milioni di pagine
  - non molto ordinato
  - eterogeneo per contenuti e stile
- (Internet) non ha un proprietario centrale
  - è possibile pubblicare/trasmettere tutto (**netiquette**)
  - non viene garantita la qualità del servizio
  - gli utenti si collegano da ovunque ed accedono ovunque
  - pubblicazione/trasmissione sono a basso costo
  - facile ed ampia diffusione di qualsiasi prodotto elettronico
- problemi
  - difficoltà di reperimento delle informazioni utili
  - non tutti gli “utenti” sono in buona fede
    - pagine a contenuto impreciso/falso (-> truffa)
    - email non sollecitata
    - virus

URL - Uniform Resource Locator

- in generale, i nodi di un ipertesto devono essere identificati in qualche modo
- se l'ipertesto è anche distribuito, allora i problemi sono:
  - dare un nome al nodo
  - identificare dove il nodo è memorizzato
  - indicare come accedere al nodo
- l'URL serve a questo scopo, ed è fatto così:
- schema://indirizzo.su.Internet/identificatore/locale/della/pagina
  - schema è il protocollo per l'accesso
  - l'indirizzo identifica il calcolatore su Internet ove risiede il nodo (pagina)
  - l'identificatore locale è usualmente la posizione del nodo nel file system locale, o sua abbreviazione

URL e indirizzamento univoco

URL: Uniform Resource Locator

Sintassi: *servizio**://host.domain/nomeoggetto***

***servizio**:nomeoggetto*

Esempi: **http://www.unich.it/www/welcome.html**

gopher://gopher.unich.it/

ftp://ftp.switch.ch/mirror/msdos/

news:it.comp.aiuto

mailto:brunetta@cc.unich.it

La componente ipertestuale: HTML

- **HTML** (HyperText Markup Language) è un linguaggio di markup, cioè i cui comandi sono inseriti esplicitamente all'interno del testo
- è un'applicazione di SGML Standard Generalized Markup Language, ISO 8879)
- i comandi permettono la formattazione del testo similmente a quel che accade in un word processor, e però anche la creazione di link
- in aggiunta a ciò, alcuni comandi hanno anche valore semantico, in quanto indicano porzioni di documento di significato specifico (es. titolo, intestazioni, ecc.).
- è comunque un formato testuale

Il browser

- funge da interfaccia uniforme ad ogni servizio di rete
- gestisce i protocolli indicati negli schemi degli URL,
- visualizza dinamicamente il testo ricevuto dai server secondo le istruzioni di formattazione specificate dai tag di HTML
- visualizza almeno una serie di formati multimediali standard (jpeg, gif, wav,...)
- interpreta le richieste dell'utente di selezione dei link, nascondendo ogni passo necessario alla selezione del protocollo, alla ricerca del server ed alla richiesta della risorsa
- Esempi: Safari, Internet Explorer, Chrome

I server

- ogni calcolatore che funge da server ha in genere in funzione un programma per ogni protocollo servito (http, ftp, ...)
- le risorse distribuite da HTTP di solito consistono in
  - files (HTML, JPEG, GIF, MPEG, ...) presenti sul file system locale in appositi sottoalberi
  - a volte programmi di cui viene distribuito l'output (es. database, ed in generale CGI)
  - a volte programmi che vanno ad essere eseguiti sul calcolatore client (classi JAVA)

Information retrieval: i motori di ricerca

I **motori di ricerca** sono strumenti per mezzo dei quali è possibile ricercare alcuni termini (parole) all'interno di una grande quantità di siti web. In seguito ad una ricerca, i motori di ricerca riportano una lista di siti che contengono i termini cercati.

Information retrieval: i motori di ricerca

Una **directory** contiene una raccolta di indirizzi di siti web, catalogati per tipologia dei contenuti, che sono stati espressamente selezionati da personale umano.

I **motori di ricerca**, invece, scandagliano continuamente l'intero WWW (World Wide Web) e includono nel proprio archivio di indirizzi **tutti** i siti web che riescono ad individuare

Information retrieval: i motori di ricerca nel web 1.0

I migliori motori di ricerca hanno adottato delle tecniche per mezzo delle quali ad ogni sito archiviato viene attribuito un valore che rappresenta una sorta di "indice di qualità" del sito web.

In questo modo è possibile, in seguito ad una ricerca, offrire all'utente una lista di siti ordinata in base all'indice di qualità dei siti elencati, partendo dal sito che presenta il valore più alto.

Information retrieval: i motori di ricerca nel web 1.0 (visibilità)

I motori di ricerca sono uno degli strumenti migliori per riuscire ad acquisire un'**alta visibilità** su Internet, in quanto consentono di veicolare gli utenti proprio verso quei siti web che gli utenti stessi sono interessati a trovare.

Il miglior passo da compiere per risultare visibili sui motori di ricerca è quello di realizzare un sito web di qualità.

Information retrieval: i motori di ricerca (come sono analizzati i siti web)

Ogni motore di ricerca utilizza alcuni programmi chiamati "spiders" (ragni) il cui unico compito è quello di visitare continuamente una grande quantità di siti web, leggere il testo contenuto nelle pagine ed estrarre quelle parole/termini che rappresentano al meglio i contenuti del sito.

Per ogni pagina letta, lo spider cerca al suo interno e memorizza ogni link (collegamento) ad altri siti, aggiungendoli ad una lista di siti da visitare. In questo modo, attraverso un processo a catena, lo spider è in grado di ottenere una quantità enorme di indirizzi di siti e pagine web, riuscendo ad incrementare il numero di siti conosciuti molto più di quanto possa essere fatto dalle *directory*, che si basano su un lento meccanismo di iscrizione e valutazione dei siti, operato da esseri umani.

Information retrieval: i motori di ricerca (spider)

Quanto appena detto sugli spider ci porta a fare una prima importante considerazione: **i siti web acquistano "corposità" agli occhi dei motori di ricerca solo se contengono buone quantità di testo**

Information retrieval: i motori di ricerca nel (posizionamento)

Per **posizionamento di un sito** si intende quell'insieme di tecniche che possono portare un sito a raggiungere posizioni prominenti nei risultati delle ricerche sui motori.

Parametri in base ai quali i siti vengono "promossi“:

•**Contenuti testuali**

Senza ombra di dubbio il fattore che più di ogni altro incide sul posizionamento di un sito web nei motori di ricerca e lo aiuta a scalare le loro "classifiche" è rappresentato dai contenuti testuali del sito stesso.

Più l'argomento ricercato dall'utente viene trattato sul sito, e più il motore di ricerca spingerà il sito verso i primi posti della lista.

È dunque consigliato produrre considerevoli quantità di testo e trattare qualunque argomento in maniera estesa e approfondita.

•**Keyword (parole chiavi)**

Ogni volta che un utente effettua una ricerca su un motore di ricerca, inserisce alcuni termini che, a suo giudizio, ritiene attinenti all'argomento di suo interesse. Nel realizzare un sito web bisogna porre attenzione al fatto che nei testi delle pagine siano presenti anche quelle parole-chiave che presumibilmente gli utenti useranno come termini di ricerca sui motori.

Se le keyword rappresentano una buona percentuale del testo complessivo di una pagina, il motore di ricerca tenderà a far salire il sito nelle liste.

Creare una pagina contenente solo keyword e nessuna frase di senso compiuto rappresenta un trucco di posizionamento controproducente in quanto può spingere il motore di ricerca a bandire definitivamente il sito dai suoi archivi.

•**Popolarità (numero di link)**

La tecnica attraverso la quale i motori di ricerca calcolano l'indice di popolarità di un sito si basa sul numero di link sparsi per il web che puntano ad esso. Più sono i link che puntano al sito (in un certo senso "consigliandolo") e più il sito è considerato popolare. Va inoltre notato che i link non possiedono tutti ugual peso; un link presente sul sito di una importante e conosciuta società ha peso maggiore rispetto a un link presente su una semplice home-page personale.

Information retrieval: i motori di ricerca (popolarità)

Più un sito riceve link e più è considerato popolare dai motori di ricerca.

Si può essere sicuri della bontà di un link se il medesimo proviene da una normale pagina web, ad esempio un articolo, se fa uso di normale tecnologia HTML e se la sua nascita non è conseguenza di automatismi o schemi di scambio organizzato o affiliazione.

Il link che vale di più è in pratica quello che idealmente testimonia, nei confronti di un sito web, una genuina e disinteressata attenzione verso i suoi contenuti.

FINE

fsivilli@unich.it