Informatica Grafica

Corso di Laurea in Ingegneria Edile – Architettura

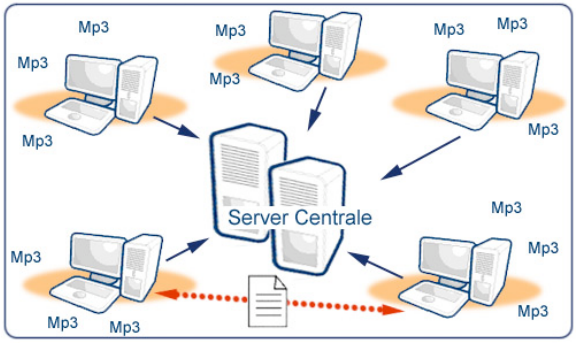
Internet

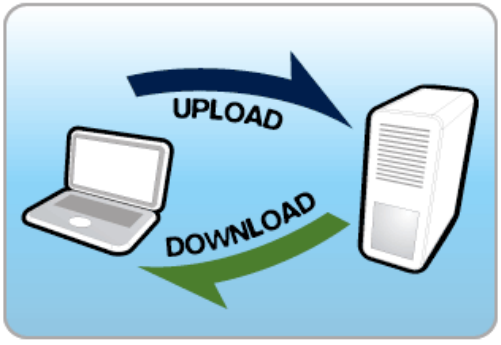
Paolo Torroni

Dipartimento di Elettronica, Informatica e Sistemistica (DEIS)
Università degli Studi di Bologna

Anno Accademico 2011/2012

Internet





- **Internet**
  1. Struttura di Internet
  2. Storia di Internet
  3. Principali protocolli e servizi di Internet
  4. Il Web
  5. I motori di ricerca
  6. Profili giuridici

Parte I

Struttura di Internet

Reti di calcolatori

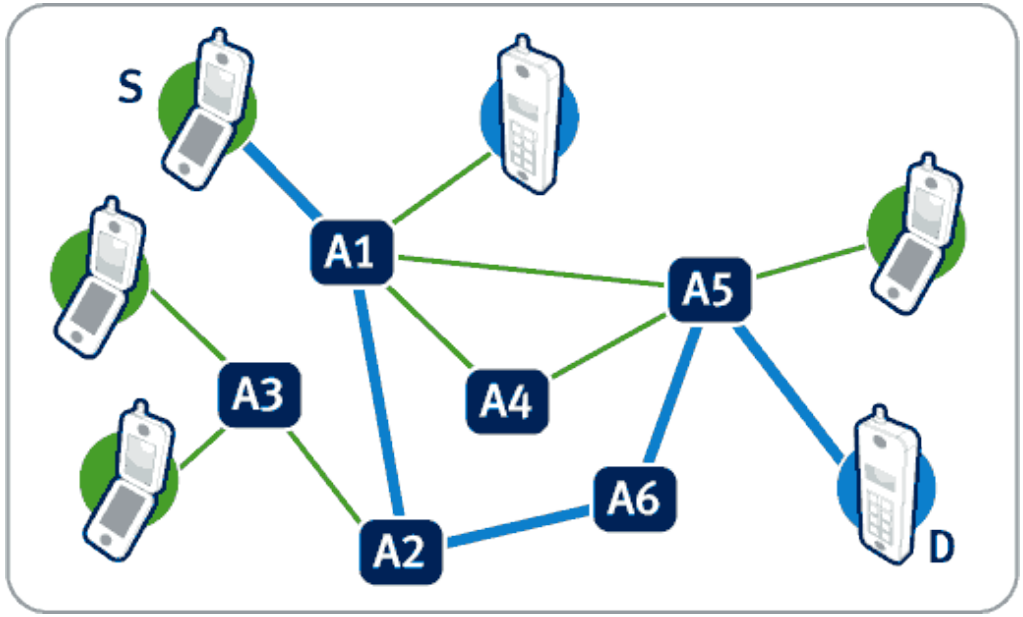
- Una delle più grandi innovazioni dell'informatica è stata la possibilità di collegare tra loro più calcolatori
  - **calcolatori singoli** collegati a risorse (es: stampanti)
  - **rete di calcolatori**: più calcolatori collegati tra loro mediante apposite infrastrutture di comunicazione (cavi, radio, ecc.)
- Esistono moltissime tipologie di reti:
  - **dimensioni**: LAN, MAN, WAN
  - **supporto di telecomunicazione**: Ethernet, coassiale, doppino, fibra, wireless, etc
  - **topologia** della connessione: stella, anello, bus, punto a punto
  - **stabilità** della connessione: connessioni dedicate, commutate, mobili
- Hardware e sistemi operativi possono essere eterogenei
  - per dialogare, necessario avere uno *schema comune*: **protocollo**

Internet e il World Wide Web

- Reti di reti: *internet*.
- **Internet** è la rete planetaria di tutte le reti collegate tra loro e che comunicano attraverso la coppia di procolli TCP/IP.
- Internet rende possibile la trasmissione e condivisione di informazione.
  - La principale architettura informativa basata su Internet: **World Wide Web**
  - Un insieme di **ipertesti collegati tra loro** e che risiedono su nodi fisicamente diversi e distanti tra loro.

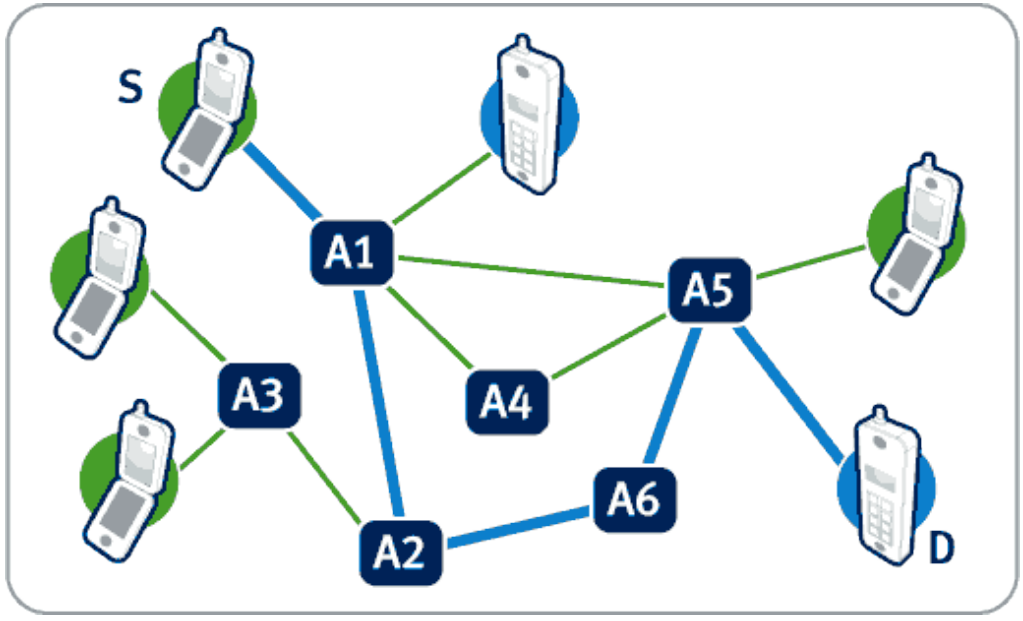
Commutazione di circuito

- Come possono dialogare due nodi di una rete?



- Modalità di connessione della normale telefonia vocale:
  - le risorse (canali di comunicazione, interruttori, ripetitori, ecc.) che si trovano sul percorso S, A1, A2, .., An, D sono **assegnate alla connessione** tra S e D
  - non sono disponibili per altri, fino a quando S e D non le rilasciano, al termine della telefonata

Problemi della commutazione di circuito

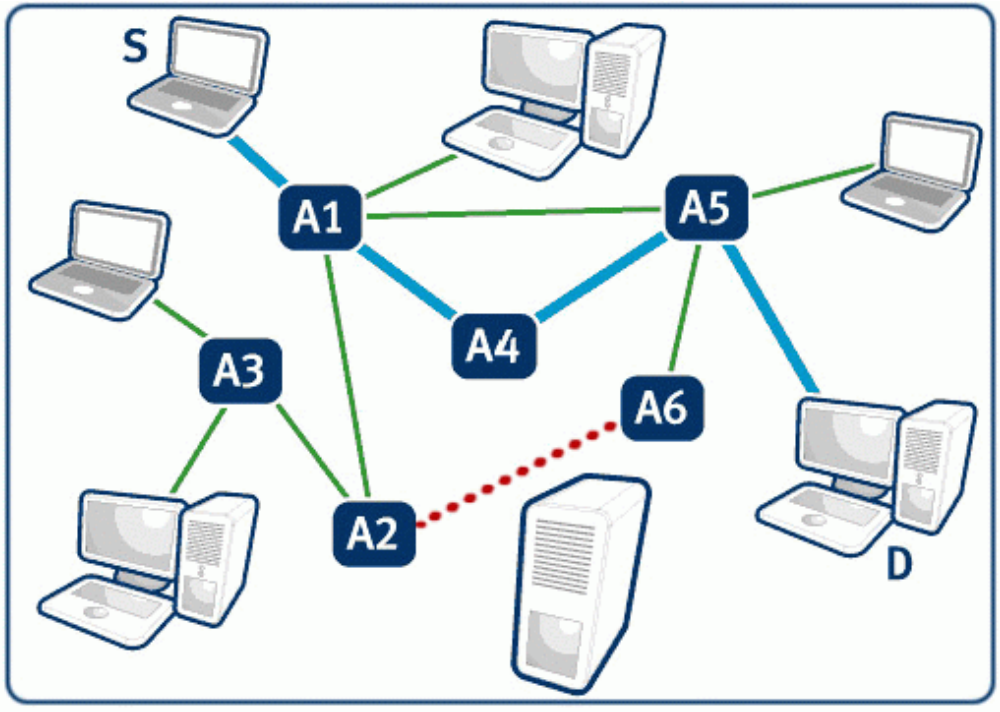


- Inadatta alla comunicazione tra due calcolatori. Motivi:
  1. tempo per realizzare la connessione tra S e D
  2. utilizzo delle risorse
  3. impatto di eventuali guasti al circuito

Una soluzione alternativa (commutazione di pacchetto)

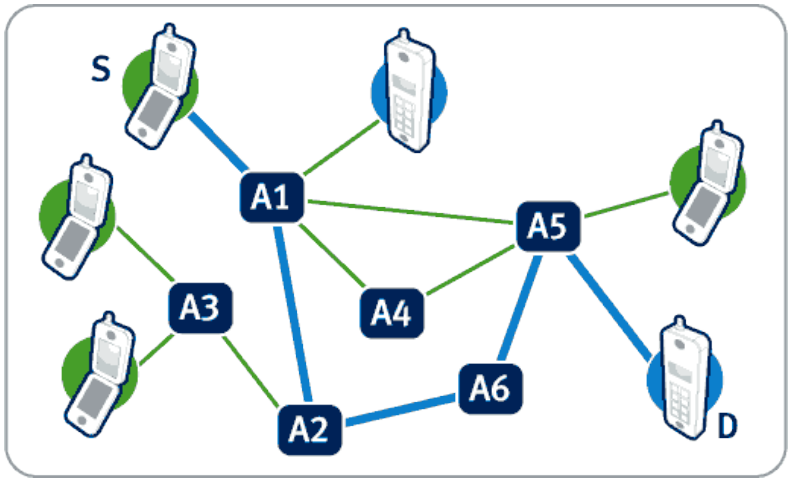
1. Il nodo sorgente (S) suddivide il messaggio in **pacchetti**, ciascuno composto di un numero *fissato (e piccolo)* di caratteri
2. S contraddistingue ogni pacchetto con:
   - la propria **firma**,
   - il **numero d'ordine** del pacchetto all'interno del messaggio,
   - l'**indirizzo** del destinatario (D) sulla rete.
3. S invia ciascuno dei pacchetti a uno dei calcolatori a cui è direttamente collegato.
4. Se questi non è il destinatario finale, quando riceve il pacchetto lo **inoltra** ad uno dei suoi vicini;
   - il comportamento si ripete finché i vari pacchetti non raggiungono D.
5. Via via che D riceve i pacchetti,
   - li **rimette in ordine**,
   - scarta eventuali **duplicati**
   - se necessario, richiede a S la **ritrasmissione** di qualche pacchetto perso.

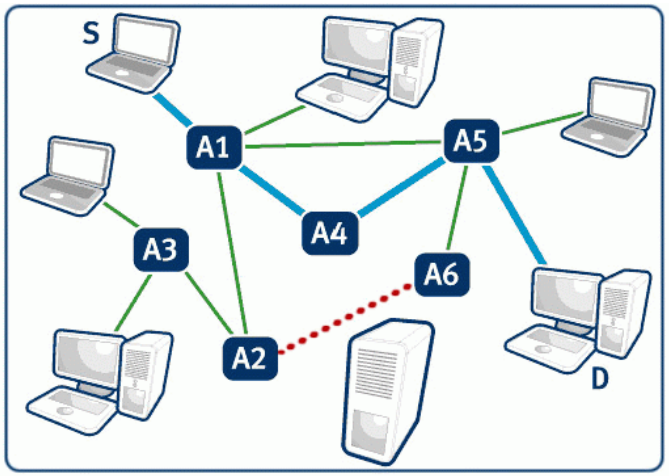
Una soluzione alternativa (commutazione di pacchetto)



- Quali vantaggi?
  1. tempo per realizzare la connessione tra S e D
  2. utilizzo delle risorse
  3. impatto di eventuali guasti al circuito

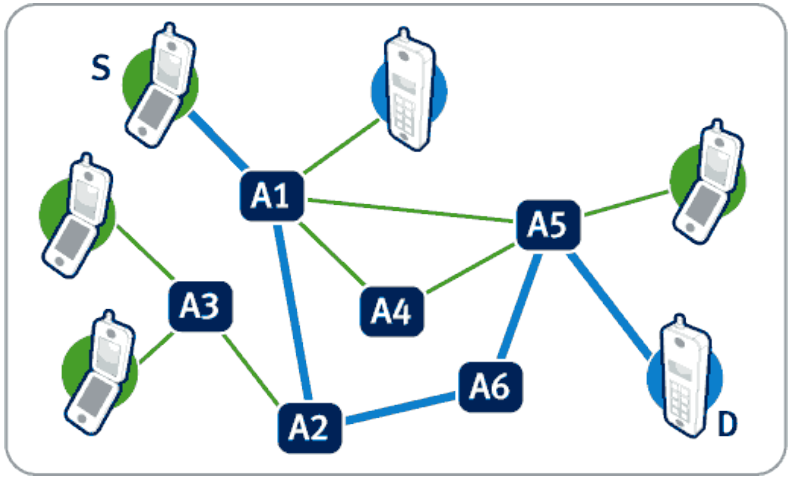
Alcune caratteristiche della commutazione di pacchetto

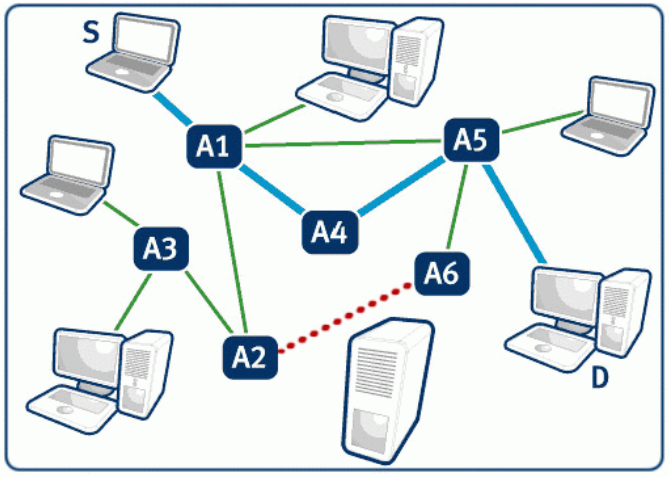




- ogni pacchetto utilizza **tutta la banda di comunicazione disponibile** in quel momento;
- ogni pacchetto che arriva su un nodo viene memorizzato prima di essere ritrasmesso (meccanismo **store-and-forward**);
- molti pacchetti su una specifica connessione possono causare **congestione**: accodamento in attesa di usare la connessione;
- i pacchetti di uno stesso messaggio possono
  - seguire **percorsi diversi**;
  - arrivare a destinazione in **ordine arbitrario**.

Alcune domande sulla commutazione di pacchetto



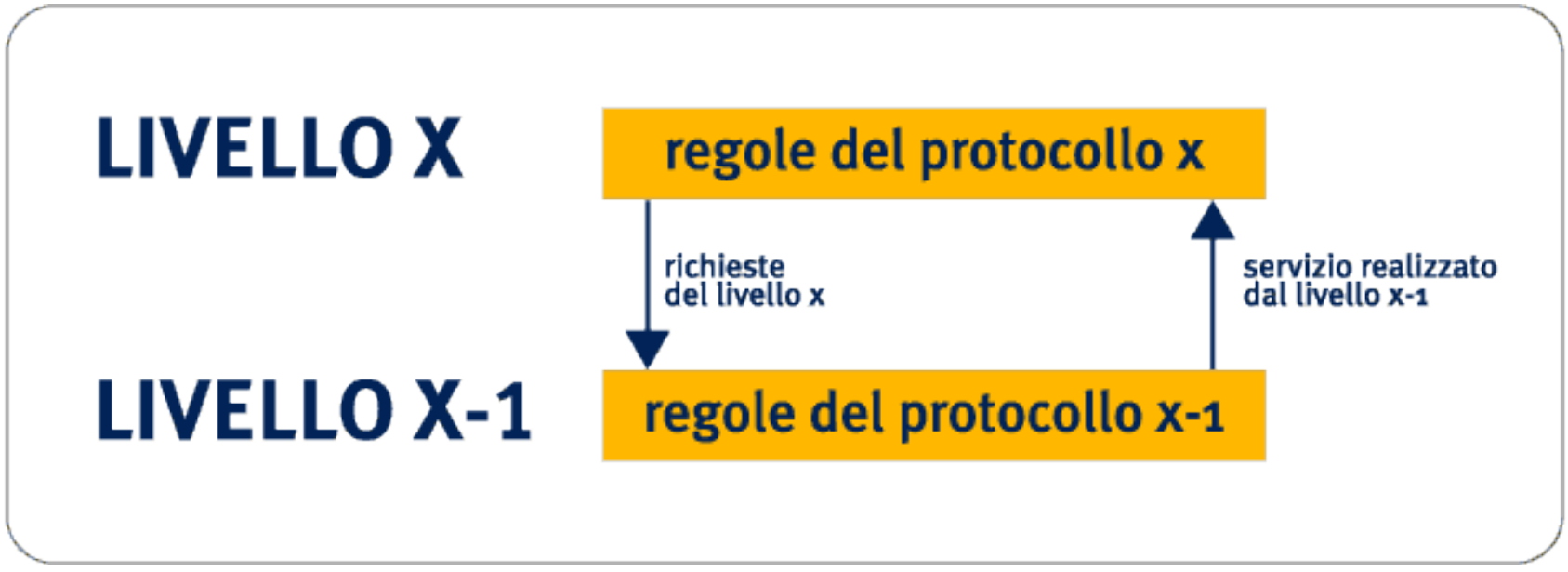


- come decide un nodo a quale vicino trasmettere un pacchetto?
- come viene indicato l'indirizzo che identifica D in modo univoco?
- come conosce S questo indirizzo?
- come vengono gestiti gli errori durante la trasmissione?
- come viene gestita la congestione?
- chi si occupa di tutti questi dettagli?

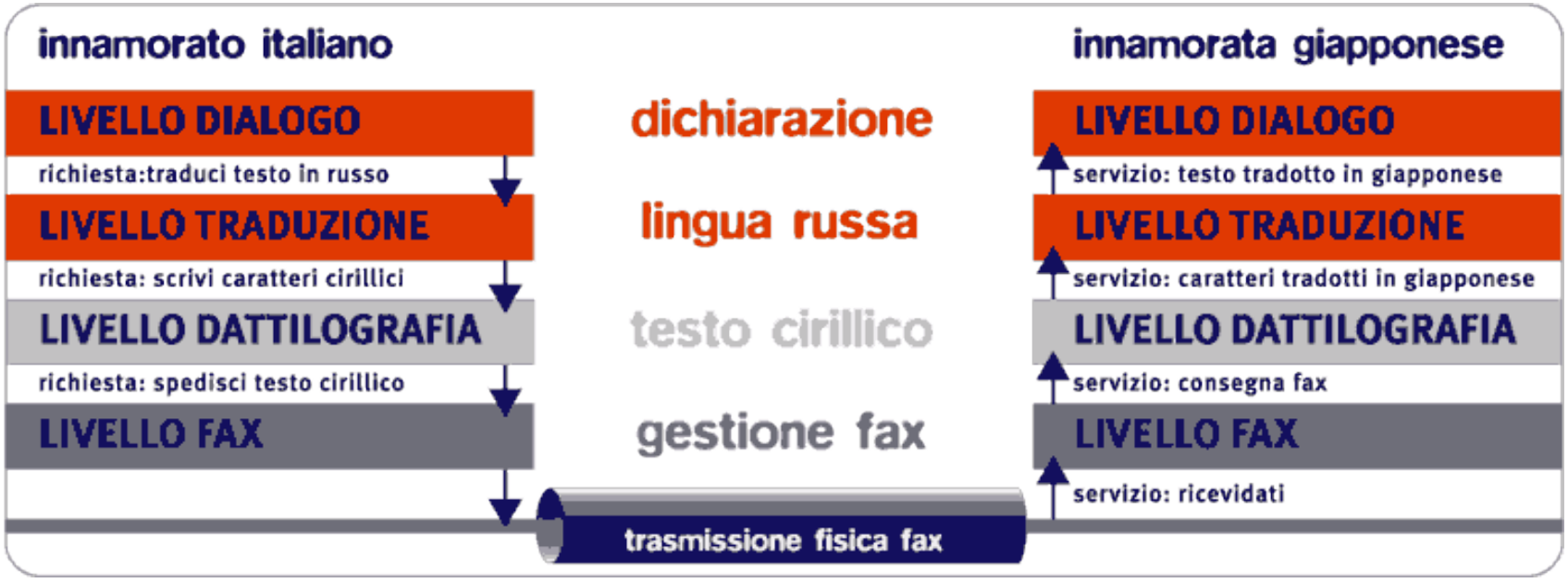
Chi si occupa dei dettagli della comunicazione?

- Un **protocollo** è un accordo tra due parti sulle **modalità** con cui la loro **comunicazione** deve procedere:
  - insieme di **regole** che definiscono
  - il **formato** e l'**ordine** delle **comunicazioni** spedite tra le due parti,
  - le **azioni da compiere** al momento della **trasmissione** e del **ricevimento** della comunicazione
- La rete è organizzata **a livelli**.
- La comunicazione avviene, concettualmente, tra due entità che stanno su due nodi diversi, ma allo stesso livello, secondo il protocollo di quel livello.
- In ogni nodo sono presenti tutti i livelli, e ogni livello nasconde quelli che gli stanno sotto.

Ogni livello nasconde quelli che gli stanno sotto



Comunicazione tra innamorati



I livelli dei protocolli nelle reti di calcolatori (caso generale)

I livelli dei protocolli di Internet

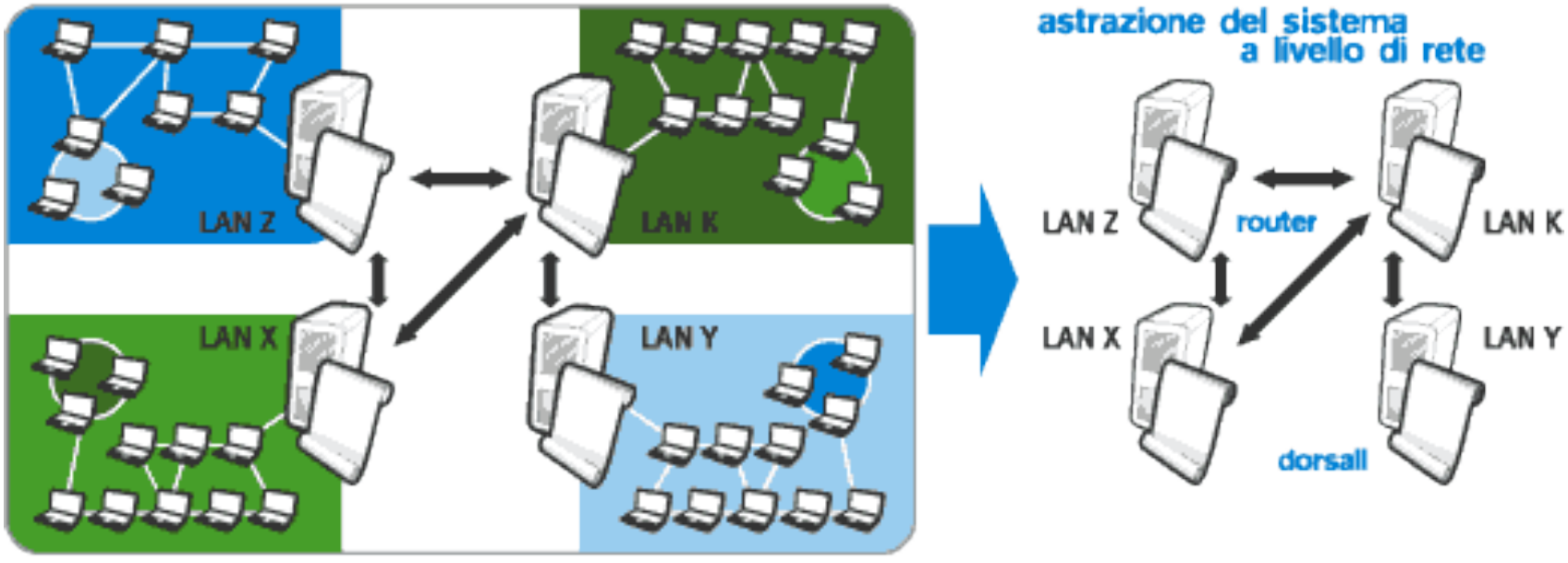
Internet è progettata su cinque livelli

- **Livello applicazione** Richiesta, scambio e ricezione dati.
  - **HTTP** (scambio di servizi sul WWW),
  - **SMTP** (posta elettronica), ecc.
- **Livello trasporto** Impacchettamento dati del livello applicazione e loro invio sulla rete; ricostituzione del messaggio, richiesta di rinvio pacchetti perduti.
  - **TCP (Transmission Control Protocol)**.
- **Livello rete** Instradamento (*routing*) dei messaggi (a quale nodo inoltrare un pacchetto?)
  - **IP (Internet Protocol)**.
- **Livello data-link** (MAC/LLC) Accesso al mezzo di comunicazione, usando un apposita codifica dei dati digitali.
- **Livello fisico** Voltaggio e durata dei segnali, ecc.

I livelli dei protocolli di Internet

- Caratterizzati dalla coppia TCP/IP
- Mancano i livelli di sessione e presentazione
- Per i livelli più bassi e applicativo: non un solo protocollo

Una rete di reti eterogenee



- Grazie ai livelli, è possibile collegare reti eterogenee: basta
  - eleggere in ciascuna rete un nodo rappresentante (**router**),
  - collegare i router con un canale di comunicazione,
  - far passare dai router tutte le comunicazioni tra le due reti.

Una rete di reti eterogenee

- La comunicazione tra router segue un protocollo comune.
- I router gestiscono ai livelli inferiori le differenze tra reti:
  - velocità di trasmissione,
  - dimensione dei pacchetti,
  - condizioni d'errore, ecc.
- Ciascun router vede un'altra rete come costituita soltanto dal suo router.
- Non occorre sincronizzare le reti (commutazione di pacchetto)
- È una soluzione **scalabile**:
  - per aggiungere una nuova rete basta poter indirizzare un pacchetto verso un nodo della nuova rete
  - il destinatario (non la rete) ha la responsabilità di rimettere insieme il messaggio e scartare i pacchetti duplicati

Come identificare un nodo su Internet?

- I nodi delle reti sono individuati da *numeri univoci*, detti **indirizzi**, assegnati come segue:
  - l'universo Internet è suddiviso in **reti fisiche**;
  - ad ogni rete fisica è assegnato **in modo centralizzato** un certo indirizzo (es. 234);
  - se questa è divisa in **sottoreti**, gli indirizzi ai suoi nodi vengono assegnati **in modo gerarchico**, mediante concatenazione
    - Ad es: due sottoreti $\Rightarrow$ indirizzi 234.1 e 234.2;
  - infine, l'indirizzo di ciascun nodo è la concatenazione dell'indirizzo della sua sottorete con un numero che lo individua in modo **univoco** nella sottorete
    - Ad es: 234.2.27.
- I numeri ai nodi delle reti fisiche, o alle sottoreti, sono **assegnati dai gestori** delle reti stesse.
- La gestione degli indirizzi è del **livello di rete**, quindi del **protocollo IP**

Indirizzi IP

- Un indirizzo IP è una sequenza di 4 numeri decimali $\in [0..255]$, separati da un punto
- Es: `130.136.2.37`
  - Indirizzo della rete fisica: `130.136`
  - Identificativo della sottorete: `2`
  - Identificativo del nodo: `37`
- I numeri alle **reti fisiche** vengono assegnati dalla **Internet Assigned Number Authority** (IANA), `www.iana.org`
- IANA delega analoghi **organismi regionali** all'assegnamento dei numeri IP all'interno delle relative zone geografiche
- Per l'Europa: RIPE NCC (Réseaux IP Européens Network Coordination Centre), `www.ripe.net`
- I gestori delle singole reti fisiche (es: CeSIA) assegnano i numeri ai loro nodi.

Indirizzi simbolici di dominio (nomi logici)

- Un indirizzo IP identifica univocamente un dispositivo fisico in Internet, ai fini del protocollo IP.
- Per l'uso umano, non è pratico usare gli indirizzi IP
- Si usano degli **indirizzi simbolici**, o **nomi logici**: caratteri invece di numeri.
  - Es: `www.myspace.com`, `mail.unibo.it`
- L'insieme e la struttura di questi nomi costituiscono il **Domain Name System**, o **DNS**, di Internet.
  - L'ultima parte del nome logico identifica il **Top Level Domain** (dominio di primo livello): es. `it`, `com`
  - La penultima identifica il dominio di secondo livello, etc.
- Non esistono limiti al numero di livelli di un nome logico

Corrispondenza tra nomi logici e indirizzi IP

- Ad un nome logico corrisponde **un unico indirizzo IP**
- Ad un indirizzo IP possono corrispondere **più nomi logici**
- **Attenzione**: la struttura dei nomi logici non ha **nulla a che vedere** con la gerarchia degli indirizzi IP
  - `www.miur.it` $\rightarrow$ `193.206.6.24`
  - `pop.cs.unibo.it` $\rightarrow$ `130.136.1.110`
  - `lia.deis.unibo.it` $\rightarrow$ `137.204.46.194`
- Alcuni nodi della rete (**DNS**, Domain Name *Server*) si occupano di tradurre i nomi logici in indirizzi IP
  - Quindi per ottenere l'indirizzo IP di `www.myspace.com` basta sapere l'indirizzo IP di un DNS che lo conosce
- Se un DNS non conosce l'indirizzo IP di un determinato nodo, inoltra la richiesta a un altro DNS, via via fino a uno dei **DNS root server**

DNS root server

- Ci sono molti DNS root server sparsi un po' in tutto il mondo (130 posizioni in 53 paesi a settembre 2007)
- Si tratta di un servizio fornito da organizzazioni, selezionate dallo IANA, che si assumono l'obbligo di fornirlo

Corrispondenza tra indirizzi IP e nomi logici

- Ad un indirizzo IP possono corrispondere **più nomi logici**
- Non è sempre possibile ottenere informazioni significative da un semplice indirizzo IP
- Esistono però dei servizi di *lookup* (**WHOIS**)
  - RIPE, `http://www.db.ripe.net/whois`
  - APNIC, `www.apnic.net/apnic-bin/whois.pl`
  - ARIN, `www.arin.net/whois/`
  - Network Solutions, `www.networksolutions.com/en_US/whois`
- **Reverse DNS lookup**
  - `http://remote.12dt.com/`
- **Localizzazione geografica**
  - `http://www.geobytes.com/ipLocator.htm`

Chi assegna i nomi logici di dominio?

- Domini di primo livello: esiste un organismo internazionale indipendente che sovrintende alla loro ripartizione e definizione
  - **Internet Corporation for Assigned Names and Numbers** `www.icann.org`
  - Primi domini di primo livello:
    - `edu` (educational: università e scuole);
    - `com` (commerciale);
    - `mil` (militare);
    - `gov` (governativo);
    - `int` (internazionale);
    - `net` (fornitori di connettività).
  - Successivamente:
    - Domini nazionali e regionali, `it`, `fr`, `jp`, `eu`, ...
    - Altri domini: `biz`, `coop`, `museum`, `name`, `org`, ...
- Ogni dominio di primo livello ha un organismo di gestione
  - **Registration Authority** per il dominio `it`: `www.nic.it`
- I domini di terzo livello sono assegnati dal titolare del dominio di secondo livello, etc.

Chi coordina Internet?

- Internet non è proprietà di nessuno, ma si basa su **scelte tecniche** condivise.
  - Solo per quanto riguarda l'**interoperabilità** della rete.
- Organismi che definiscono standard e procedure per Internet:
  - **Internet Society** (`www.isoc.org`)
    - sviluppo di Internet e il supporto di IETF e IRTF
    - associazione di privati e enti pubblici
    - Internet Architecture Board (`www.iab.org`): nomina i membri dei due organismi tecnici;
    - Internet Engineering Task Force (IETF, `www.ietf.org`): per questioni tecniche, come la definizione degli standard;
    - Internet Research Task Force (IRTF, `www.irtf.org`): coordina la ricerca di medio periodo;
  - **Internet Corporation for Assigned Names and Numbers**:
    - organismo internazionale indipendente
    - gestione dello spazio di nomi e indirizzi nel medio periodo

Riassumendo...

- Con "*Internet*" si indica il sistema informativo globale che:
  - (i) è logicamente connesso mediante **un unico spazio globale di indirizzi** basato sul protocollo IP o sulle sue estensioni;
  - (ii) permette di **supportare le comunicazioni** utilizzando la coppia di protocolli **TCP/IP** o le sue estensioni e/o altri protocolli compatibili con IP;
  - (iii) fornisce, utilizza o rende accessibili, in modo pubblico o privato, **servizi ad alto livello** sfruttando i livelli di comunicazione e le infrastrutture che sono stati descritti ai punti precedenti.

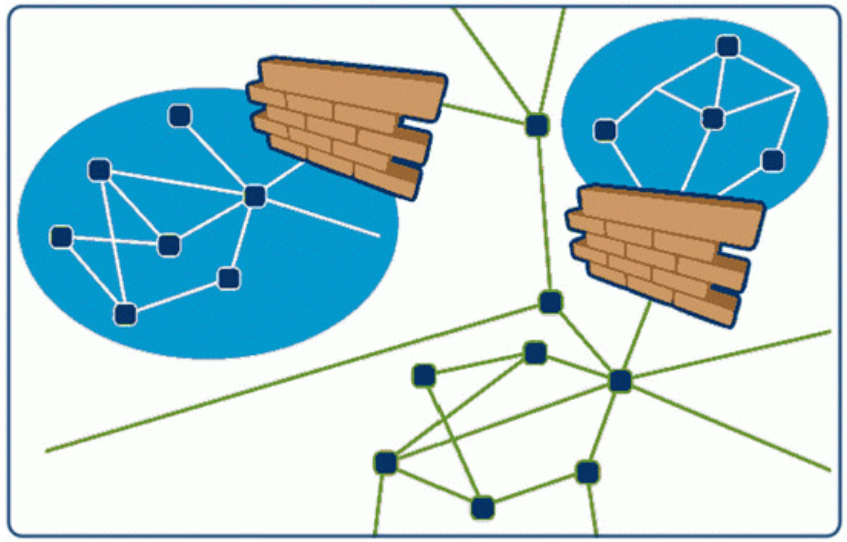
Federal Networking Council, 24 ottobre 1995

Per collegarsi ad Internet

- Occorre un computer che supporti i protocolli TCP/IP
- Bisogna ottenere un indirizzo IP
- Bisogna collegare il computer a un router collegato a Internet
  - Collegandolo a una LAN, ad esempio con un cavo Ethernet, o con una scheda Wireless
  - Collegandosi a un ISP attraverso un modem, tradizionale o ADSL, o attraverso un cavo per la banda larga

Intranet

- TCP/IP può essere usato anche per implementare una rete aziendale (**intranet**)
- Se la intranet è collegata a una rete pubblica, l'accesso è di solito protetto da un **firewall**
- Due sottoreti di una rete aziendale possono comunicare attraverso Internet
  - Esempio: Università di Bologna e sedi della Romagna.



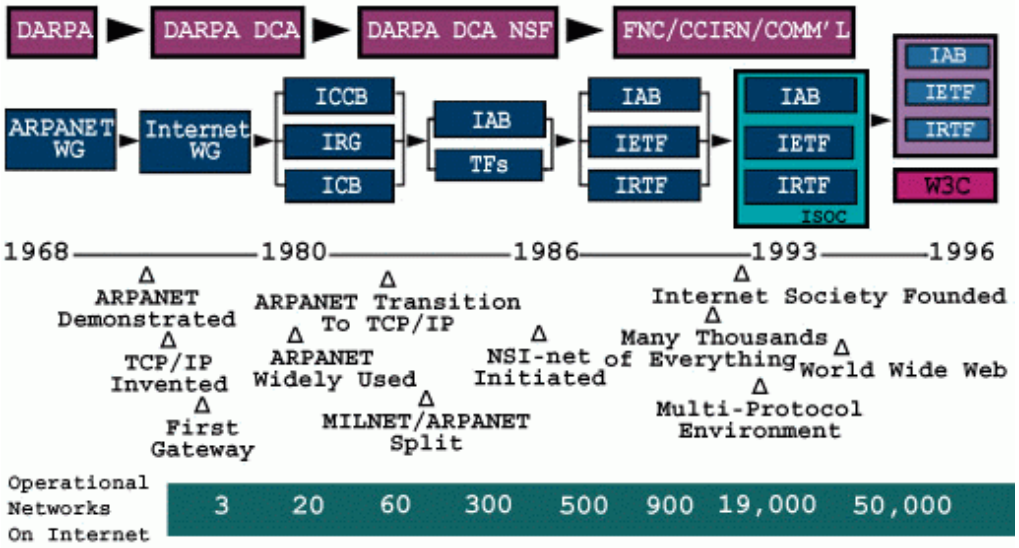
Costi di Internet

- Costi dei servizi e delle **informazioni** messe a disposizione sulla rete stessa.
- Costi della **struttura di interconnessione** e dell'effettivo trasferimento dei dati.
  - Ripartiti tra tutti gli utenti, secondo il tipo di connessione di cui dispongono
  - Ciascun utente paga la connessione al proprio ISP
  - L'ISP compra connettività da altri ISP o dai proprietari della connessione fisica
- I prezzi dipendono in gran parte dal regime di mercato (monopolio/concorrenza)

Parte II

Un po' di storia

Breve storia di Internet: i primi anni



1968 DARPA (Defense Advanced Research Projects Agency) incarica BBN (Bolt, Beranek & Newman) di creare ARPAnet

1970 Primi 5 nodi: UCLA, Stanford, UC Santa Barbara, Utah University, e BBN

1974 Specifiche del protocollo TCP (Vint Cerf)

1984 Internet (1000 nodi) adotta TCP/IP in massa

Le scelte cruciali

- Idee chiave nel progetto di TCP/IP (Robert Kahn):
  1. **nessuna modifica interna** deve essere richiesta per collegarsi a Internet;
  2. **best effort**: se un pacchetto non raggiunge la destinazione sarebbe, sta al mittente originario (non ad altri nodi intermedi) ritrasmetterlo;
  3. **semplicità dei router**, che non devono mantenere alcuna informazione sul flusso dei singoli pacchetti attraverso di essi
  4. **nessun controllo globale della rete** al livello della sua operatività.

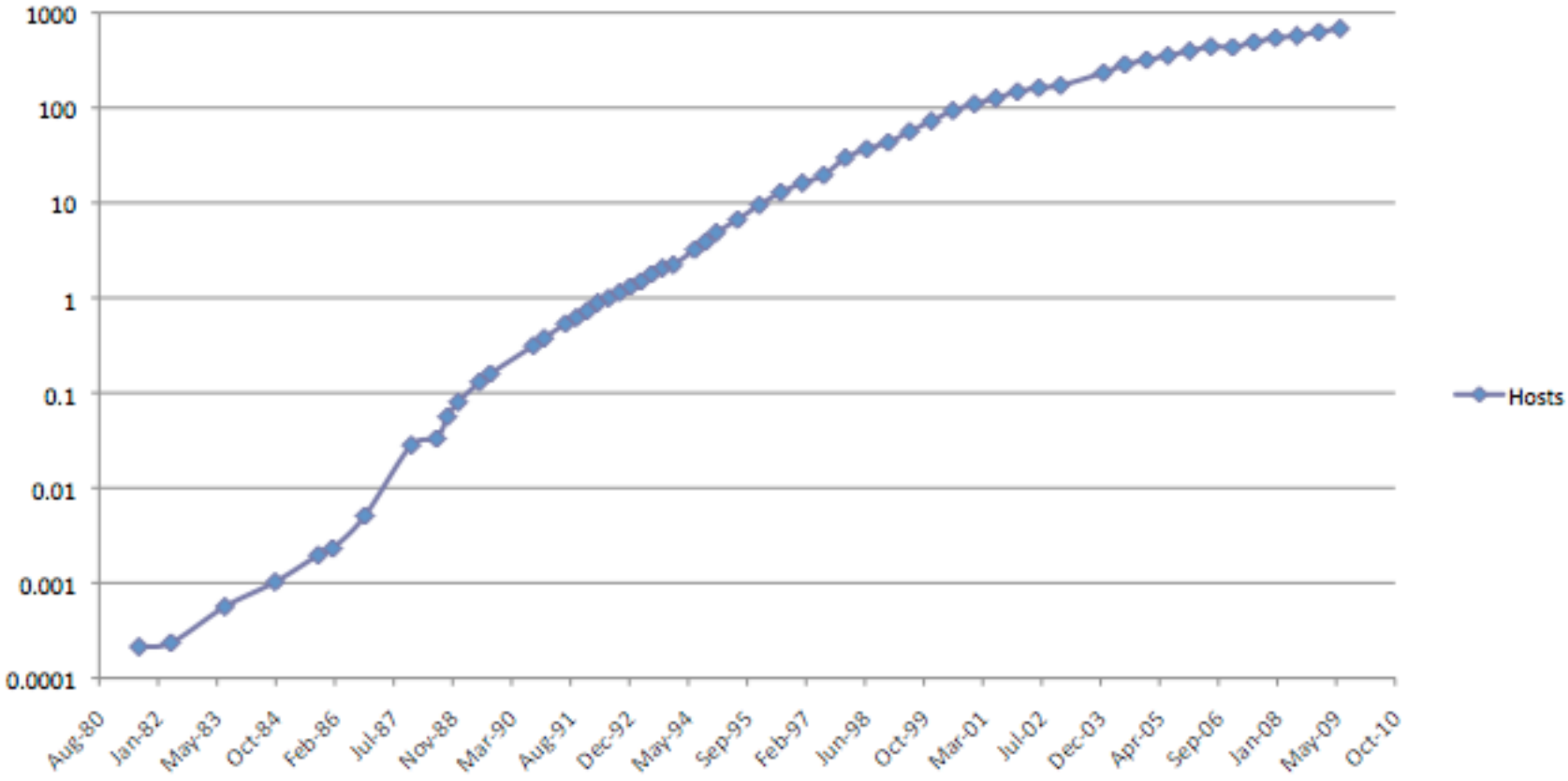
La crescita di Internet

- Agli inizi degli anni '80 esistono molte reti, pubbliche e private (DEC, IBN, Xerox)
- Alla metà degli anni '80 le reti pubbliche, CSNET e NSFNET operano alcune scelte di grande importanza:
  - **scelgono TCP/IP** come protocollo;
  - forniscono dorsali di connessione (**backbones**) tra i nodi principali delle reti;
  - si accordano per l'uso di ARPANET, con un costo a forfait e non a consumo, e vietano il traffico per usi commerciali.
- Fattori di maggiore impatto per la crescita di Internet:
  - oculate scelte di **gestione** incentivano la nascita di reti private;
  - **quantità di finanziamento** veicolato in NSFNET (200 milioni di dollari dal 1986 al 1995);
  - **qualità tecnica** dei protocolli TCP/IP.
- Per saperne di più: `http://www.isoc.org/internet/history/`

Un bilancio storico

- Internet ha sempre coinvolto insieme ricerca pubblica e iniziativa privata.
- Allo stesso tempo, è sempre stata una **rete non proprietaria**, le cui **specifiche** e **caratteristiche** erano e sono **pienamente disponibili e utilizzabili da chiunque** intenda investirvi.
- Aspetti chiave nella storia di Internet:
  1. **tecnologia**: commutazione di pacchetto mediante protocolli in continuo miglioramento;
  2. **gestione operativa e strategica** di una infrastruttura globale complessa;
  3. **aspetti sociali**: vasta comunità di persone che sviluppano la tecnologia anche grazie alla rete stessa;
  4. **aspetto economico**: transizione da struttura di ricerca ad infrastruttura informativa di vasta scala.

Quanto è grande Internet



- Grafico in **scala logaritmica** (milioni di nodi)
- Fonte: Internet Systems Consortium, Internet Host Count History `https://www.isc.org/solutions/survey/history`

Ipertesti: la storia

1945 Vannevar Bush, MIT: idea di organizzare la conoscenza su base reticolare e associativa, invece che sequenziale

1965 Ted H. Nelson usa per la prima volta la parola "ipertesto"

> *un corpus di materiali scritti o grafici interconnessi in un modo così complesso da non poter essere ragionevolmente presentato o rappresentato su carta. Può contenere sommari, o schemi dei suoi stessi contenuti e delle loro relazioni reciproche; può contenere annotazioni, aggiunte e note [...] Un tale sistema potrebbe crescere senza limiti, inglobando gradualmente una parte sempre più ampia della conoscenza scritta esistente al mondo.*

1962 Doug Engelbart, SRI, inventa il mouse.

- Ancora nessuno aveva pensato di sfruttare Internet come supporto.

Dagli ipertesti al Web

1989 Tim Berners-Lee, CERN, specifica un sistema di documenti ipertestuali distribuiti, per la collaborazione di gruppi di ricerca in tutto il mondo

- Idea nuova: un documento possa essere indicato in modo univoco attraverso un *Universal Document Identifier* (oggi, URL).
- Progetto di HTML per specificare documenti
- Progetto del protocollo HTTP per recuperare ipertesti in Internet.

1993 Marc Andreessen, U. Illinois, scrive Mosaic, il primo browser multimediale

1994 Mosaic ha un bacino di milioni utenti nel mondo. Andreessen fonda Netscape.

oggi 1 miliardo di persone usano il Web attraverso browser multimediali.

Le prospettive: Internet2

- Internet progettato per collegare grandi calcolatori e fornire semplici servizi a migliaia di utenti.
- Oggi: cambiati i numeri e la tipologia di utenti (wireless) e il tipo e qualità dei servizi offerti.
- Problemi:
  1. infrastruttura di comunicazione lenta e troppo sfruttata;
  2. non gestisce problematiche di sicurezza, identità, autorizzazione e autenticazione degli utenti;
  3. protocolli hanno raggiunto i loro limiti fisici (indirizzi IP disponibili);
  4. applicazioni limitate dalla tecnologia di rete invece che ruolo guida per lo sviluppo di tecnologie innovative.
- Progetto Internet2 `http://www.internet2.edu`
- IPv6

Parte III

Servizi di Internet

Client/server e peer-to-peer

- Esistono due modelli di interazione/comunicazione tra calcolatori su Internet che permettono di fornire un servizio.
  - "architetture"
- **Client/server**: quando un computer (client) ha bisogno di un servizio di Internet (es: scaricare un file), chiede ad un altro computer (server) di fornire il servizio in questione
  - asimmetria
  - Esempio: navigazione Web.
- **Peer-to-Peer**: i nodi della rete che comunicano/collaborano giocano ruoli interscambiabili e svolgono le stesse funzioni
  - simmetria
  - comunicazione diretta tra nodi, senza intermediari
  - Esempio: Napster

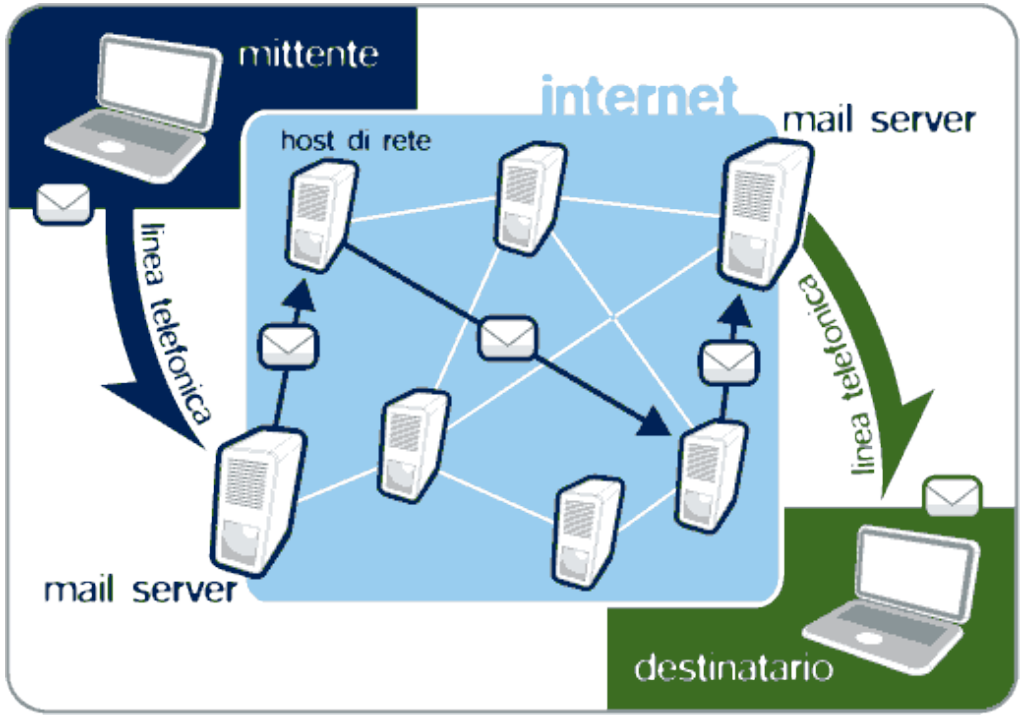
Glossario di Internet

- **Protocollo** insieme di regole e di convenzioni da rispettare perché due calcolatori possano comunicare.
- **Servizio**: prestazione fornita da una macchina servente (server) a una macchina cliente (client) che ne fa richiesta.
- **Applicazione**: programma.
- Attenzione: a volte uno stesso termine può essere usato per indicare un protocollo, un servizio, un'applicazione
  - Esempio: FTP

Posta Elettronica

- È un **servizio**.
- Consente di scambiarsi sia messaggi di testo sia, utilizzando opportuni strumenti, ogni altro tipo di file.
- La comunicazione è **asincrona**
  - Ovvero, non è necessario che il destinatario sia collegato nel momento in cui il messaggio viene inviato.
- Ogni messaggio ha un mittente e uno o più destinatari.
- È necessario avere un **indirizzo**, dato da un fornitore di servizio, per **ricevere** posta

Posta Elettronica



- Ad ogni indirizzo e-mail corrisponde una mailbox conservata in una macchina di proprietà del fornitore di servizio (mail server)
- Nella mailbox vengono depositati automaticamente i messaggi di posta indirizzati al proprietario della casella (**mail daemon**)
  - Quindi, non è necessario essere connessi a Internet affinché i nostri messaggi raggiungano la nostra mailbox.

Posta Elettronica: inoltro

`paolo.torroni@unibo.it`

- `unibo.it` identifica in maniera **univoca in Internet** un particolare **mail server**, che ospita l'utente `paolo.torroni`.
- `paolo.torroni` identifica in maniera **univoca all'interno del mail server** un determinato **utente**.
- Il **protocollo SMTP** (Simple Mail Transfer Protocol) è il protocollo di livello applicazione che gestisce il trasferimento della posta elettronica.
- La composizione e invio di un messaggio avviene attraverso un **client**, cioè un'applicazione che prepara il messaggio e lo invia nella rete usando SMTP.
  - Thunderbird, Mail, Eudora, Outlook, ...
- Alcuni fornitori di servizio email offrono anche un'**interfaccia Web**, per cui non è richiesto un client sulla propria macchina
  - GMail, Hotmail, Yahoo, Libero, ...

Posta Elettronica: ricezione

`paolo.torroni@unibo.it`

- Lo stesso cliente di posta o interfaccia Web permette anche di leggere la posta elettronica che giace nella nostra mailbox.
- Due protocolli:
  - **POP** (Post Office Protocol): posta viene "scaricata" sul computer dove è installato il client.
  - **IMAP** (Internet Message Access Protocol): si può scegliere di lasciare la posta sul server
- POP, IMAP, Web?
  - accesso alla posta quando non si è collegati a Internet (*off-line*)
  - accesso da postazioni diverse
  - accesso da computer in cui è installato solo il browser

Come viene confezionato un messaggio per SMTP?

- Headers:

```
From: Barak Obama <admin@internet.com>
Date: Mer gen 15, 2003  17:25:47 Europe/Rome
To: Paolo Rossi <p.rossi@CS.UniBO.IT>
Subject: Prova
Return-Path: <admin@internet.com>
Received: by le (mbox p.rossi) (with Cubic Circle's
  cucipop (v1.31 1998/05/13) Wed Jan 15 17:25:55 2003)
Received: from source ([69.9.251.177]) by CS.UniBO.IT
  (8.9.3/8.9.3/Debian 8.9.3-6) with ESMTP id RAA29182
  for <p.rossi@cs.unibo.it>; Wed, 15 Jan 2003 17:25:45 +0100
X-From_: admin@internet.com Wed Jan 15 17:25:46 2003
User-Agent: Microsoft-Entourage/10.0.0.1309
Message-Id: <BA4B4A1B.D4BE%g.verdi@cs.unibo.it>
Mime-Version: 1.0
Content-Type: text/plain; charset="US-ASCII"
Content-Transfer-Encoding: 7bit

Questo e' il contenuto o corpo del messaggio di posta elettronica.
```

- Non necessariamente l'informazione contenuta è veritiera.

Alcune considerazioni sull'uso della email

- Uso corretto degli strumenti: **Netiquette**
  - **Firmare** i propri messaggi con **nome e cognome**
  - Non modificare il testo dei messaggi che si inoltrano
    - Inoltro di messaggi privati: opportuno chiedere permesso all'autore
  - Non inviare messaggi aggressivi (*flames*).
  - Rileggere prima di spedire.
  - Essere tolleranti in quel che si riceve.
    - Se provocati, meglio non rispondere
    - o, per lo meno, dormirci sopra
  - Usare maiuscole e minuscole.
    - SE SI USANO SOLO LE MAIUSCOLE È COME SE SI STESSE URLANDO.

Alcune considerazioni sull'uso della email

- Uso degli strumenti giusti: quando è meglio non usare l'email
  - Non usare la inbox come **backup** o storage di **dati sensibili** (es: password)
    - Servizi di remote/online/secure backup
  - Evitare di spedire per email **file voluminosi** (es: foto, video)
    - **Strumenti per la condivisione**
    - Flickr, YouTube, P2P
  - Non usare la inbox come **to-do-list**
    - **Strumenti di produttività**
    - Thinking Rock `http://www.trgtd.com.au/`

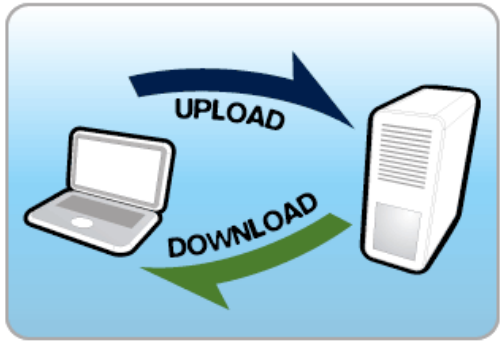
Alcune considerazioni sull'uso della email

- Aspetti tecnici
  - Non tutti i client visualizzano i messaggi allo stesso modo
  - Cautela nell'uso della formattazione
- Sicurezza, autenticità e privacy
  - SMTP trasmette i messaggi **in chiaro**
  - Possibile usare strumenti per proteggere e "certificare" il contenuto dei messaggi.
    - PGP, `www.pgp.com`
  - Alcuni server consentono/richiedono l'**autenticazione** (SSL).
  - In Italia: Posta elettronica certificata (PEC)

Mailing list

- **Servizio di comunicazione uno-a-molti** che si basa sulla posta elettronica.
- A volte: comunicazioni periodiche (*newsletter*)
  - A.Word.A.Day, `www.wordsmith.org/awad`
  - Crypto-Gram Newsletter, `http://www.schneier.com/crypto-gram.html`
- Gestione delle mailing list:
  - manuale: **moderatore** che mantiene una lista di indirizzi di posta elettronica
  - automatica: servizio che accetta iscrizioni via posta elettronica.

FTP



- **File Transfer Protocol** è il protocollo più efficace e veloce per **trasferire file** da un computer ad un altro.
  - Due macchine: **locale** (lato **client**) e **remota** (lato **server**)
  - Due operazioni principali: **upload** e **download**
- Necessario avere un'applicazione: **cliente FTP**, e conoscere l'indirizzo IP o nome logico del server FTP.
- Molti **browser** implementano il protocollo FTP
- Per distribuire file: occorre installare un server FTP
  - FileZilla `http://filezilla-project.org/`
- Possibile accesso anonimo o con password
- Per trasferimento sicuro: Secure FTP (SFTP)

Sistemi di chat e instant messaging

- ICQ, `http://www.icq.com/`
- 120.000.000 iscritti a dicembre 2001.
- due funzioni principali:
  1. visualizza l'elenco degli utenti ICQ (nella nostra *contact list*)
  2. permette ai propri utenti collegati alla rete di comunicare.
- Così come ICQ, tanti altri
  - Sistemi integrati per chat, VOIP, SMS, scambio file, link, etc.
  - Soprattutto in modalità **sincrona**.
  - Skype (eBay) `skype.com`

Newsgroup

- **Bacheche elettroniche**. Raccolte di messaggi in cui si discute qualche argomento
- Servizio **asincrono**
- Archiviazione dei messaggi
- Organizzazione in **thread** di discussione
- Modalità di accesso:
  - tramite news client abilitato a ricevere messaggi da un News server
    - spesso integrato nell'email client
  - iscrivendosi ad apposite mailing list;
  - via Web: usando siti specializzati nell'archiviare i newsgroup.

Newsgroup

- **USENET**: Users Network
- Decine di migliaia di gruppi, organizzati in gerarchie di notiziari:

Comp	Computer, ricerca e industria informatica
Sci	Scienze fisiche ed ingegneristiche
Humanities	Letteratura e studi umanistici
Rec	Attività ricreative, compresi sport e musica
...	etc.

- Ogni gerarchia è divisa in sottoargomenti
  - `rec.sport` si occupa di sport,
    - `rec.sport.basketball` di pallacanestro
    - etc.
- Documentarsi su Netiquette e comportarsi correttamente.
- Funzionalità newsgroup con accesso Web: **Forum**
  - UniversiBO, `www.universibo.unibo.it`
  - CAD Forums, `www.cadforums.net`

Protocolli e Sicurezza

- Esistono informazioni che richiedono di essere **protette**
  - Password, coordinate bancarie, numeri di carte di credito, etc.
  - Documenti, es. visure catastali, denuncia dei redditi, etc.
- Le informazioni che viaggiano su Internet **non sono in generale protette**.
- Problemi principali:
  - **Disponibilità.** Ciò che inviamo viene ricevuto dal destinatario?
  - **Confidenzialità.** Ciò che inviamo viene letto solo dal destinatario?
  - **Autenticità.** Sappiamo con certezza chi è il mittente?
  - **Integrità.** Sappiamo che il documento non è stato modificato nel tragitto?

Secure Socket Layer (Transport Layer Security)

- SSL (TLS) è un protocollo che fornisce un servizio di **protezione delle informazioni** che viaggiano in rete.
- Strato tra **livello TCP e livello applicativo**
- Fa uso di **algoritmi crittografici**
- Utilizzato in combinazione con protocolli visti su TCP
  - HTTP**S**, IMAP**S**, **S**SMTP, ...
- SSL fornisce un servizio di autenticazione degli *endpoint* della comunicazione
  - autenticazione del server
  - opzionalmente: autenticazione del client
  - negoziazione di materiale crittografico tramite un canale che garantisce riservatezza e integrità.

Meccanismi di cifratura a chiave pubblica

- SSL usa la **cifratura a chiave pubblica** per generare delle chiavi segrete di comunicazione
  - Esempio di uso della cifratura a chiave pubblica:
    1. S conosce la chiave pubblica di D
    2. S cifra il messaggio M usando la chiave pubblica di D:
    3. D riceve decifra il messaggio usando la sua chiave privata

$$M \xrightarrow{public\_key_D} M' \xrightarrow{private\_key_D} M$$

- Le chiavi segrete sono poi usate da client e server per scambiare messaggi riservati.
- Garanzie fornite da SSL:
  - Confidenzialità
  - Autenticità
  - Integrità
- SSL non offre la sicurezza che un messaggio venga ricevuto

Problemi di identità...

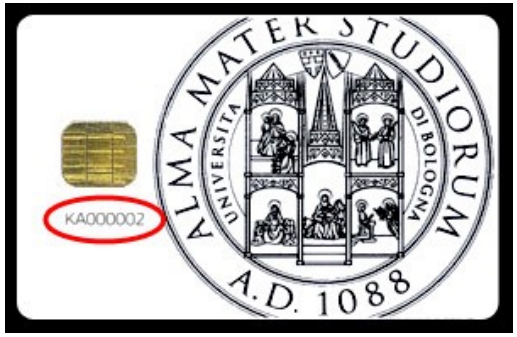
- Una chiave pubblica non è di per sé associata a una "persona", ma esclusivamente ad una chiave privata.
- Come associare una chiave pubblica a una persona?
- Nei documenti cartacei, l'associazione con un'indentità avviene mediante firma...
  - la firma dovrebbe essere difficile da imitare;
  - il firmatario non può ripudiare la propria firma;
  - il destinatario può accertare l'identità del firmatario;
  - il destinatario non può alterare un documento firmato;
  - il tutto è verificabile da una terza parte (giudice).
- Che possibilità abbiamo nel mondo digitale?
  - Conformità di una **caratteristica fisiologica o comportamentale** con un dato "biometrico" di riferimento.
  - Conoscenza di un **dato segreto concordato** in precedenza (password o personal identification number)
  - Possesso di un **oggetto riconoscibile** da parte della macchina (una scheda a banda magnetica o una smart card)

Certificati

- I **certificati** sono documenti elettronici che usano una firma digitale per **associare un'identità a una chiave pubblica**.
  - Emessi da **autorità di certificazione** (**certification authority**)
  - Grandi organizzazioni delle quali le persone "si fidano"
  - Di solito: rilascio di certificato e **chiave privata** su *smart card*
- Contenuto di un certificato digitale:
  - Un **numero di serie** che identifica il certificato in modo univoco;
  - Il **proprietario del certificato** (**subject**), cioè la persona o ente identificato;
  - L'**algoritmo** usato per creare la firma;
  - L'**ente** che ha verificato l'informazione e prodotto il certificato;
  - Le **date** di validità (da..a)
  - La **chiave pubblica** per cifrare un messaggio per il subject
  - **Firma della Certification Authority** e altre informazioni per verificare che il certificato non sia stato manomesso.

Algoritmi a chiave pubblica

- È possibile eseguire la firma digitale di documenti elettronici mediante una semplice attrezzatura:
  - Smart card con certificato valido
  - Lettore di smart card + software installato
- Es: AlmaEsami



- Il **software** che esegue la firma digitale di un documento/file:
  - riceve in input il certificato digitale, la password associata al certificato, e il file da firmare,
  - restituisce come output un file modificato (documento firmato).

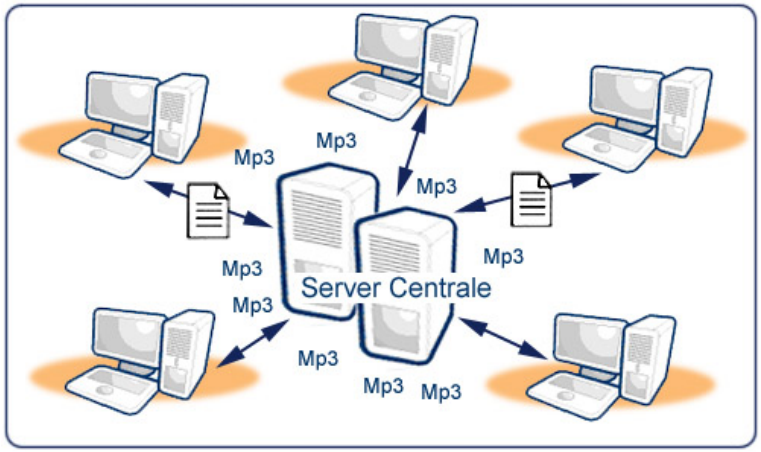
Valore legale dei certificati

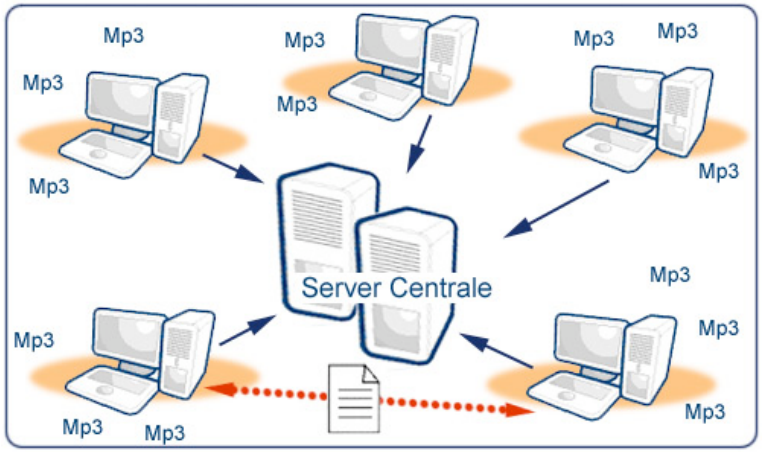
- Attenzione alla legislazione!
- In Italia questo tipo di firma ha **pieno valore legale**, tanto quanto (e per certi aspetti "di più") di una firma autografa tradizionale!!
- Alternativa molto più debole (es. transizioni bancarie): **firma elettronica**
  - invio di username e password.

Posta elettronica certificata (PEC)

- Regole tecniche custodite, gestite e pubblicate da DigitPA (ex CNIPA)
- Gestori: lista su `http://www.digitpa.gov.it/`
  - Actalis, Fastweb, Notariato, Poste Italiane, IWBank, Regione Marche, Innova Puglia, etc.
- Aspetti tecnici e normativi
  - Certezza dell'invio e della consegna (o meno) dei messaggi al destinatario
  - Garanzia dell'identità del mittente
  - Integrità e autenticità dei messaggi mediante firma elettronica apposta dai gestori
  - Possibilità di inviare messaggi a qualunque indirizzo
  - Se destinata a indirizzo PEC: **stesso valore legale della raccomandata A/R**
  - Gestori conservano traccia delle operazioni per 30 mesi
  - Gestori tenuti a verificare presenza di virus
- `www.postacertificata.gov.it`

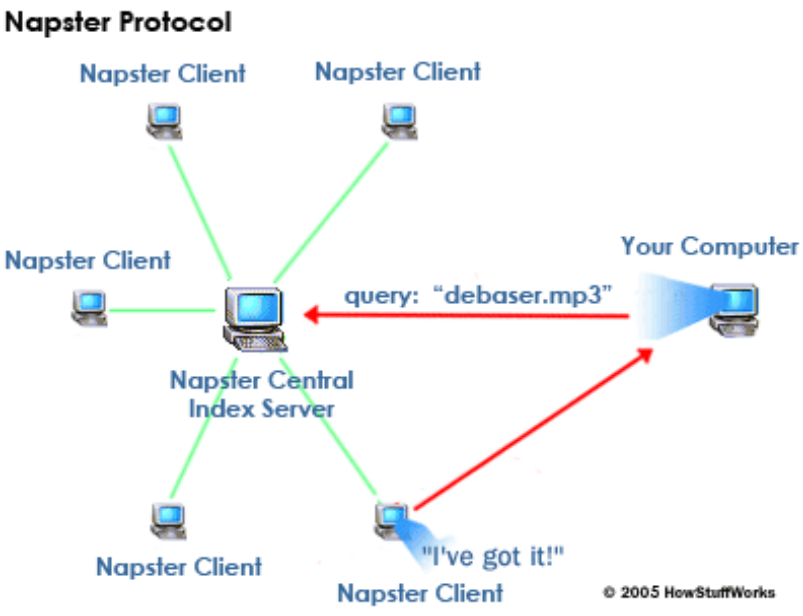
Peer-to-Peer





- Client/server: versione informatica di una biblioteca.
- P2P: comunicazione diretta tra due applicazioni che risiedono su due nodi della rete, senza server intermedi.
- Differenze principali:
  - Scalabilità e costi di manutenzione
  - Distribuzione del carico e dei costi
  - Presenza di colli di bottiglia nelle connessioni di rete
  - Robustezza e disponibilità dell'informazione
  - Facilità nella ricerca dell'informazione
  - Controllo su cosa/come viene distribuito

Applicazioni P2P: File sharing

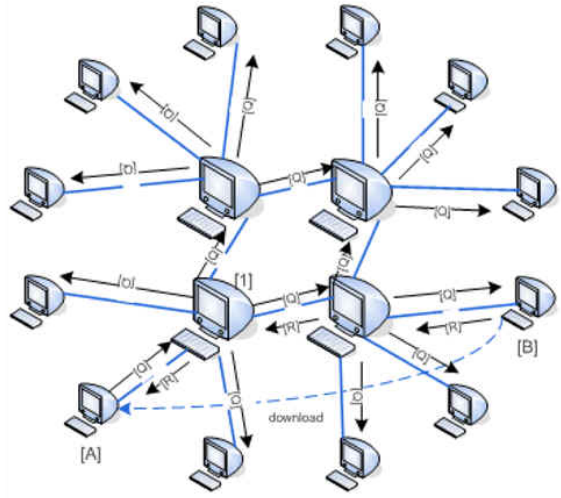


- **Napster** (Shawn Fanning 1999)
  - Solo per **mp3**.
  - Utilizzo di un server centrale che **non mantiene i file veri e propri, ma solo il nome e dove sono localizzati**
  - 26.4ML utenti registrati a febbraio 2001
  - ! Chiuso a giugno 2001 per problemi legali con artisti e case discografiche (Metallica, Madonna, A&M Records) e la Recording Industry Association of America (RIAA)

Applicazioni P2P: File sharing

- **Gnutella** (Justin Frankel & Tom Pepper, Nullsoft, 2000)
  - **Nessun server**.
  - Lista dei file disponibili creata a ogni connessione.
  - Client comunica le proprie informazioni ai propri "vicini", che le propagano ai loro vicini, etc.
  - Per iniziare: **bisogna conscere almeno un vicino**.
  - 1.8ML computer a giugno 2005.
  - Fine 2007: la rete di condivisione file più usata in Internet (40% share)
  - Client: LimeWire, BearShare, BearFix, Gnucleus, Shareaza, Acquisition, FrostWire, Morpheus, Phex, etc.

Applicazioni P2P: File sharing

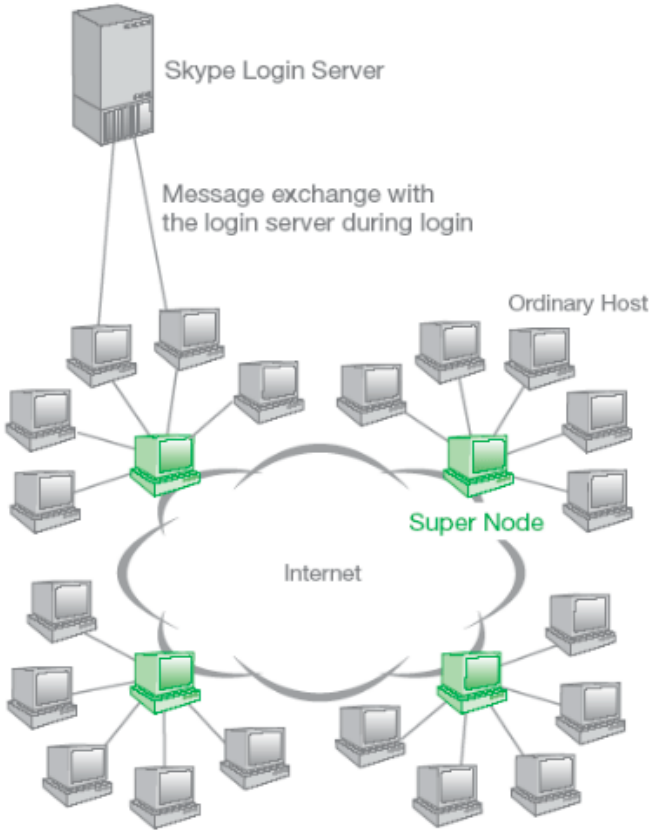


- **Kazaa** (Niklas Zennström, Janus Friis & Priit Kasesalu, 2001)
  - Architettura simile a Gnutella.
  - Organizzazione gerarchica a **supernodi**
    - i calcolatori più potenti e con le connessioni più veloci
    - contengono lista di alcuni file disponibili e dove sono localizzati
    - **Ricerca mediante comunicazione tra supernodi**
  - Possibilità di download parallelo
  - Stessi autori di **Skype** e Joost
  - Client: Grokster, Morpheus, LimeWire etc.
  - ! Sicurezza (prudenza: dove tenere i file da condividere)
  - ! Privacy (spyware? adware?)
  - ! Questioni legali (RIAA, MPAA, azioni contro singoli utenti)

Applicazioni P2P: Content distribution

- **BitTorrent** (Bram Cohen 2001).
  - Condivisione di file di grandi dimensione.
  - Idea: replicare un file rapidamente in un gran numero di nodi
  - Meccanismo basato su suddivisione dei file e seeding.
  - Architettura basata su trackers e indexers:
    - S crea un piccolo file chiamato "torrent" (`MyFile.torrent`) che contiene **metadati** sul file da condividere e sul **tracker** T che coordina la distribuzione
    - D deve ottenere un torrent per il file da scaricare, e connettersi a T,
    - T comunica a D in quali altri nodi si trovano i pezzi del file
  - Nessun peer è sottoposto a carico eccessivo, tutti possono contribuire.
  - A febbraio 2009: **27-55% di tutto il traffico su Internet**
  - ! Azioni legali contro siti che ospitano tracker BitTorrent: `ThePirateBay`, `Suprnova.org`, `Torrentspy`, `LokiTorrent`, `Demonoid`, `OiNK.cd`, etc.
- Procolli simili: Ares, Emule/Edonkey

Applicazioni P2P: VOIP



- **Skype** (Niklas Zennström, Janus Friis & Priit Kasesalu, 2001)
  - Architettura simile a Kazaa (**supernodi**).
    - I client Skype mantengono una tabella locale che contiene gli indirizzi IP dei supernodi
    - Promozione a supernodo: trasparente
  - Organizzazione gerarchica con Login Server.
  - Connessione alla rete Skype:
    1. connettersi a un supernodo
    2. autenticarsi presso il server di login di Skype.
  - 276ML utenti a gennaio 2008

Applicazioni P2P: P2PTV

- **Joost** (Niklas Zennström & Janus Friis, 2007-2008).
  - Contributo della comunità (1ML beta tester!)
  - I server inviano lo stream video a un gruppo ristretto di client, che lo propagano a loro volta ad altri clienti, etc.
- Non solo Joost
  - Contributo di grandi società (FOX, Warner Music, Paramount Pictures, Yahoo, Google, YouTube, CBS)
  - Modello di business: annunci pubblicitari (Sony Pictures, BMW, Sprite, etc.)
  - Broadcaster: Babelgum, BBC iPlayer, LiveStation, Miro, ReelTime, Zattoo, etc.
  - Client: TVUPlayer, Abroadcasting (US), Zattoo (CH/US), Octoshape (DK), LiveStation (UK), etc.

Applicazioni P2P: Calcolo scientifico

- **Seti@HOME** (Sullivan et al., 1997)
  - GRID Computing
  - *Una forma di calcolo distribuito in cui un super-computer virtuale è composto di un cluster di computer in rete, debolmente connessi, che agiscono in modo coordinato per eseguire calcoli computazionalmente molto dispendiosi*
  - Esperimento scientico, Berkeley `http://setiathome.ssl.berkeley.edu/`: Search for Extraterrestrial Intelligence (SETI)

Parte IV

World Wide Web

Ipertesti

- Ipertesto: *documento che contiene al suo interno collegamenti ad altri documenti o a sezioni dello stesso documento*.
- Informazione organizzata in modo reticolare
- **World Wide Web** (**WWW**, o **Web**): un gigantesco ipertesto multimediale distribuito dotato di un'interfaccia di facile uso
  - tecnologia informatica per realizzare link facili da usare
  - documenti composti di testo, immagini, video, audio, ecc.
  - diverse parti di un ipertesto possono risiedere su calcolatori diversi e distanti tra loro (in modo trasparente all'utente)
  - **browser**, visualizza in modo uniforme i dati multimediali e risolve il problema del raggiungimento dei dati remoti.

Organizzazione del Web

- Basata su servizi client/server
  - **server HTTP**: mette a disposizione **le informazioni** (documenti ipertestuali), residenti sul server stesso **client**: accede ai documenti attraverso un **indirizzo**
- Quando un client effettua una richiesta di un ipertesto a un server HTTP, il server invia al client le varie componenti che costituiscono il documento
- Principali client: Firefox, Safari, Opera, Chrome, Internet Explorer, etc.

Come è descritto un ipertesto?

- 2 tipi di informazione:
  - Informazione (testo)
  - la struttura (titolo, corpo del testo, link, etc.)
- **Hyper Text Markup Language** (**HTML**): Linguaggio per la marcatura di ipertesti.
- Informazioni su come il documento deve essere visualizzato
  - `http://www.w3schools.com/html/tryit.asp?filename=tryhtml_intro`
- Diversi browser possono visualizzare in modo diverso la struttura descritta da HTML

Hyper Text Markup Language (HTML)

```
<HTML>
  <HEAD>
    <meta name="description" content="Esemplificazione di HTML">
    <title>Documento di prova</title>
 </HEAD>
<BODY>
  <H1> Questo &egrave; il titolo</H1>
  Questo documento descrive <I>con un esempio</I> l'uso di HTML.
  <H2> Essenza di HTML </H2>
  Un documento HTML consiste di testo "immerso" in <b>tag</b> HTML.
  I comandi descrivono la struttura del documento.
  HTML permette:
  <UL>
<LI> di descrivere la struttura del documento;</LI>
<LI> inserire elementi non testuali, come questo stemma:
      <IMG src="./logoUnibo.gif"
    ALT="logo dell'universita di Bologna">;</LI>
<LI> inserire riferimenti ipertestuali: <A HREF="http://www.cs.unibo.it">
Dipartimento di Scienze dell'Informazione</A></LI>
  </UL>
 </BODY>
</HTML>
```

Uniform Resource Locator (URL)

`http://user:pass@lia.deis.unibo.it:80/~pt`

- Ogni **risorsa** sul Web è univocamente individuata da una sequenza di caratteri che ne costituisce l'indirizzo
  - **Uniform Resource Identifier** (**URI**)
  - **Uniform Resource Locator** (**URL**)
- Contenuto di un URL:
  - il **tipo della risorsa**, che definisce il **protocollo** da utilizzare per recuperare il documento
    - `http`, `ftp`, `https`, `news`, ...
  - l'**indirizzo** della risorsa sul Web:
    - opzionalmente, *username* e *password* (es per `ftp`)
    - il **dominio** (nome simbolico o IP), che identifica un **server**;
    - opzionalmente, un numero intero (*porta*) $\in$ [0..65535]
    - un **percorso locale** che individua uno specifico documento all'interno del server
  - Il dominio *è case insensitive*
  - Il percorso locale *può essere case sensitive* (dipende dal server)

! Cautela (phishing)

Contenuti non testuali

- Informazioni per il browser

```
<meta name="description"
      content="Esemplificazione di HTML">
<meta http-equiv="Content-Type"
      content="text/html; charset=iso-8859-1">
```

- Informazioni sul contenuto del documento

```
<meta name="keywords" content="university
      teaching computers logics" >
```

- Link a risorse multimediali e varie
  - **immagini** (gif, jpg, tiff, bmp),
  - **audio** (mov, wav, mp3),
  - **video** (wmv, mpg, mov, rm, mpeg, avi),
  - **documenti in formato proprietario** (doc, pdf, swf, etc.)
- **Codice**: **javascript**, **applet**.

Hyper Text Transmission Protocol (HTTP)

- **Lato client**
  - Il client (tipicamente un *browser*) effettua una richiesta a un server HTTP di una **risorsa** corrispondente a una **URL**
    - ad esempio: l'utente segue un link durante la "navigazione"
  - A questo punto, il client rimane in attesa della risorsa
  - Quando (se) la riceve, la analizza per capire se ha bisogno di ulteriori risorse per riprodurla in modo completo
    - ad esempio: immagini, codice javascript, file audio, etc.
  - Nel caso, richiede separatamente ogni singola risorsa di cui si compone il documento
  - A mano a mano che riceve i vari elementi, riproduce la risorsa
- **Lato server**
  - Il server HTTP sta in attesa di una comunicazione su una determinata *porta* di comunicazione
  - ogni volta riceve una richiesta da parte di un client (**URL**), la soddisfa inviando la risorsa
  - abbandona la connessione e si rimette in ascolto

Helper e plug-in

- Per riprodurre documenti di altro formato (es. pdf), è possibile
  - ricorrere a programmi esterni al browser (**helper**)
  - estendere le funzionalità del browser con dei componenti esterni ma incorporati nel browser (**plug-in**)
- È possibile specificare che azioni intraprendere (quale helper o plugin) per ciascun tipo di risorsa
- I tipi delle risorsa sono specificati in un formato chiamato **MIME type**
- Mentre un helper resta aperto dopo la chiusura del browser, un plug-in vive in funzione delle azioni di navigazione del browser

Il web è sicuro?

- **Sniffing**: intercettazione (senza interferenza) dei pacchetti in transito
- Possibile in una rete a commutazione di pacchetto (all'interno della sottorete fisica)
- Così come SMTP, anche HTML è un protocollo **in chiaro**
  - È possibile ricostruire una comunicazione tra browser e server
- Evidente problema di **riservatezza** (dati sensibili e altre informazioni private)
- Quali dati sono esposti?
- Quali soluzioni?

Informazioni trasmesse dal browser al server

- Il nostro browser conosce molte cose di noi...
  - siti preferiti, *history*
  - alcuni username password usati per la navigazione
  - i documenti nel nostro file system!
- Nelle comunicazioni HTTP, trasmette:...
  - data e ora
  - tipo di browser (Firefox, Safari, Chrome, ...);
  - il sistema operativo (Mac OSX, Linux, Solaris, ...);
  - l'**indirizzo IP** da cui proviene la richiesta;
  - la **pagina precedentemente visitata**.

I server possono essere "pericolosi"?

- Certamente bisogna "fidarsi" del browser
  - Spesso viene considerato **root-of-trust** per i **certificati**
  - Importanza delle soluzioni verificabili (**open source**)
  - Considerare l'utilizzo di un **firewall**
- Punto critico: **esecuzione di codice**
  - applet, javascript
  - ambiente protetto

  ! attenzione ai possibili exploit

  $\rightarrow$ tenere il proprio browser aggiornato (**security updates**)
- Possibile trasmettere al server informazioni specifiche sull'utente tramite la gestione delle **cookie**

Informazioni sulla navigazione: cookie

- Una cookie è un **piccolo file** che il server chiede al browser di memorizzare sul disco dell'utente
- Spesso contiene **un numero, diverso per ogni utente che si collega col server**
- Due tipi di cookie:
  - **temporanei**: creati al momento in cui inizia una **sessione** con un server e sono **cancellati** al termine della sessione stessa
  - **permanenti**: creati al primo collegamento con un certo server e **rimangono sul disco** anche dopo la chiusura della sessione
- Molto comuni sono le **third party cookie**: ricevute da un sito che non è quello che stiamo vedendo.
  - Situazione molto comune con la pubblicità (*banners*)
  - Problemi di riservatezza (**user profiling**)

Chi coordina il WWW?

- Il Web nasce aperto, senza formati proprietari (copyright)
- Importante garantire l'esistenza unitaria del Web e il suo carattere aperto e libero
- 1994: **World Wide Web Consortium**, (**W3C**), `http://www.w3.org/` presso il MIT
  - Sede europea: ERCIM, `http://www.ercim.org/`
- Scopo principale: sviluppare protocolli, specifiche, software e strumenti che garantiscano l'interoperabilità del Web.

Parte V

Ricerca delle informazioni sul WWW

Ricerca delle informazioni sul WWW

- L'accesso a risorse su Web avviene tramite una URL
- Se non si conosce la URL, bisgona avere uno strumento per **cercare** informazioni (URL) sul Web
- Caratteristiche del Web da tenere in considerazione:
  - dimensioni della rete e numero delle risorse
  - dinamica della rete e delle risorse
  - eterogeneità nelle strutture dei documenti
  - varietà dei contenuti
- Non basta un semplice "elenco telefonico"!

Motori di ricerca e directory

Motori veri e propri	**Directory**
Automatici	Risultato di lavoro umano
Indicizzano URL	Classificano interi siti
Mirano a indicizzare tutto il Web	Mirano a diventare guide alle migliori risorse Web
Si basano su algoritmi matematici	Si basano su giudizi di valore qualitativi ed umani
Interrogabili attraverso combinazioni di parole chiave	Percorribili per categorie e sotto-categorie

- Search engine: `http://www.google.com/`
- Directory: `http://dir.yahoo.com/`

Com'è fatto un motore di ricerca?

- Un motore di ricerca si basa su diversi componenti:
- **Back end:**
  - Un programma che interroga periodicamente il Web per scaricare le pagine e catalogarle.
    - Usa un pool di **agenti autonomi** chiamati **spider**
  - Un **database** delle pagine catalogate.
- **Front end:**
  - Un'**interfaccia** di interrogazione (**query**).
    - Ricerche per **parole chiave**
    - Linguaggio di interrogazione: `Or`, `And`, `But Not`, `+`, `-`, `Near`, ''...'', *

Google PageRank

- Il cuore della tecnologia di ricerca di Google è un algoritmo: **PageRank**
- PageRank assegna un voto (**ranking**) alle pagine Web
- Si basa sulla struttura ipertestuale del Web (link) come indicatore del valore di una data pagina
  - Link da A a B $\Rightarrow$ A "vota" per B
- Considera il **numero dei voti** e il **contenuto delle pagine che esprimono questi voti**.
  - Voti che provengono da pagine più importanti hanno un peso maggiore
  - Siti Web di alta qualità ricevono un voto più alto di altri
- Sofisticate tecniche di **text-matching** per trovare pagine importanti e **rilevanti per una determinata ricerca**

Meta-Motori

- Rappresentazione grafica dei risultati della ricerca sotto forma di una o più **mappe**
- Es: **Kartoo**, `http://www.kartoo.com` (fino al 2010)
- **Motori dedicati**: es. viaggi
  - Orbitz `http://www.orbitz.com/`
  - eDreams `http://www.edreams.it/`
  - etc.
- **Motori embedded**
  - Ricerca in siti web (es: UniBo `www.unibo.it`, Repubblica `www.repubblica.it`, etc.)

Motori semantici

- **Web Semantico**
  - Tim Berners-Lee, James Hendler and Ora Lassila: **The Semantic Web**. Scientific American, May 2001, `http://www.scientificamerican.com/`
- Idea: passaggio da ricerca sintattica a ricerca semantica
  - consentire alle macchine di "capire" il significato degli ipertesti
  - esprimere i dati e il loro significato in formati standard direttamente elaborabili dalle macchine.
- Uso di **ontologie** (Ingegneria della Conoscenza).

"The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The first steps in weaving the Semantic Web into the structure of the existing Web are already under way. In the near future, these developments will usher in significant new functionality as machines become much better able to process and understand the data that they merely display at present."

Computazione scientifica: WolframAlpha

- Steven Wolfram, `http://www.wolframalpha.com/`
- Componenti in WorlframAlpha:
  - **Raccolta dati** e loro inserimento nel sistema in modo che siano computabili.
    - Correlazione delle fonti, identificazione di dati sospetti, cross-check, etc.
    - Importante apporto umano (**data curation**, intervento di **esperti di dominio**)
  - **Calcolo**: codifica metodi e algoritmi scientifici e ingegneristici
    - calcolo di maree, velocità media, piani di ammortamento, soluzione di equazioni differenziali, etc.
  - **Traduzione dell'input**
    - da linguaggio naturale svincolato a una precisa rappresentazione simbolica (ontologie, scoperte linguistiche, corpora)
  - **Presentazione dei risultati**
    - generazione automatica dei report
    - rappresentazione cognitivamente ottimale, promozione delle informazioni più utili

Come far conoscere il proprio sito a più utenti possibili?

- Alcuni concetti di base:
  - Fondamentale inserirlo in motori di ricerca
  - Proporre il proprio sito per la catalogazione nelle directory
    - può costare denaro, soprattutto per directory note
  - Informazioni più informazioni possibili sul sito che vogliamo registrare (parole chiave, categorie)
  - Usare gli strumenti di HTML (tag `<meta>`).
  - Essere linkati da siti importanti (PageRank)

Parte VI

Profili giuridici

Privacy, sicurezza e diritto d'autore

- Possibilità tecnica $\neq$ liceità
- I rapporti giuridici e i conflitti che hanno luogo su Internet sono disciplinati dal diritto.
- **Privacy**: protezione dei dati personali
- **Sicurezza**: obblighi di chi tratta dati personali e legislazione in materia di accesso abusivo
- **Diritto d'autore**: tutela delle opere dell'ingegno

Protezione dei dati personali ("privacy")

1996 **Legge sulla privacy** disciplina la raccolta dei dati personali

- Obbligo di chi li raccoglie di informarne il soggetto e riconoscergli il potere di esercitare un controllo sui dati raccolti

2004 Entra in vigore il **Codice in materia di protezione dei dati personali**, `http://www.garanteprivacy.it/`

- **Dato personale**: qualunque informazione riferibile a qualunque soggetto
- **Dato anonimo**: non riconducibile a un interessato identificato o identificabile, né in origine né a seguito di trattamento
- **Dato sensibile**: riguardate la personalità etico-sociale dell'individuo e le sue caratteristiche psico-sanitarie.

- Il soggetto che tratta dati personali deve garantire
  - che i dati siano esatti,
  - che siano utilizzati solo per le finalità del trattamento,
  - che siano conservati solo per il tempo strettamente necessario

Protezione dei dati personali ("privacy")

- *La libertà e la segretezza della corrispondenza e di ogni altra forma di comunicazione sono inviolabili* (Costituzione, Art.15)
  - Email di lavoro $\longrightarrow$ non è strettamente "personale"
- Esempi di illeciti nel trattamento dei dati personali (in assenza di consenso del soggetto):
  - **Spamming**: invio di comunicazioni commerciali
  - **Tracking**: registrazione di siti visitati o acquisti effettuati
  - Lettura o trasmissione a terzi, da parte dei fornitori di connettività, del contenuto di messaggi di posta elettronica

Sicurezza Informatica

2004 **Codice in materia di protezione dei dati personali**

- Obblighi per i soggetti che trattano dati personali.
- La mancata adozione delle **misure minime** di sicurezza dà luogo a **responsabilità penale**.
  - gestione di autenticazione e autorizzazione
  - aggiornamento periodico delle procedure di sicurezza
  - protezione degli strumenti elettronici e dei dati
  - garanzie sulla disponibilità dei dati e dei sistemi
  - tutela dei dati sensibili volte a garantire l'anonimato
- **Hacking**: accesso **abusivo** ad un sistema informatico o telematico **protetto da misure di sicurezza**
- Punito indipendentemente dal fatto che all'accesso segua un danneggiamento, un furto o un altro reato.

Firma digitale e firme elettroniche

- Molte disposizioni recenti volte a individuare un equivalente digitale della sottoscrizione autografa.
  - Obiettivi: promuovere commercio elettronico ed e-government
- **Acquisti su Internet**: di solito non richiedono la forma scritta, quindi non richiedono firma digitale o firma elettronica

1997 **Bassanini**: il documento informatico, recante alcuni tipi di **firme elettroniche**, può assumere il **valore giuridico** dei documenti firmati con la **sottoscrizione autografa**

- **Rapporti fra PA e cittadino**: è consentito inviare istanze alle PA semplicemente per posta elettronica.
- Livelli di sicurezza
  - **Firma elettronica** elettronica/qualificata/avanzata/digitale
  - **Certificati** elettronici/qualificati
  - **Certificatori** elettronici/qualificati/accreditati

Disposizioni sul diritto d'autore

- Da distinguere:
  - **invenzioni industriali**: tutelate dal **brevetto** (1939)
  - **opere dell'ingegno**: tutelate dal **diritto d'autore** (1941)
    - Tutela fino a 70 anni dopo la morte dell'autore

  → Software? Opere multimediali?
- Programmi: due modelli
  - **Software proprietario**: licenza d'uso a pagamento
  - **Software libero**: open source, licenze **Creative Commons** `http://creativecommons.org/`
    - Attribuzione della paternità dell'opera
    - Uso commerciale? [Sì/no]
    - Modifiche consentite? [Sì/no/*share alike*]