



SAPIENZA  
UNIVERSITÀ DI ROMA

UNIVERSITÀ "SAPIENZA" DI ROMA  
FACOLTÀ DI INFORMATICA

---

## Reti di Elaboratori

---

Appunti integrati con il libro "Computer Networking: A Top-Down Approach", J. F. Kurose, K. W. Ross

*Author*  
Simone Bianco

4 maggio 2023

# Indice

<b>0 Introduzione</b>	<b>1</b>
<b>1 Introduzione alle reti</b>	<b>2</b>
1.1 Rete, Host e Collegamenti . . . . .	2
1.2 Struttura di Internet . . . . .	4
1.3 Pacchetti, Forwarding e Routing . . . . .	7
1.4 Misura delle prestazioni . . . . .	9
1.5 Stack protocollare TCP/IP . . . . .	14
<b>2 Livello di Applicazione</b>	<b>18</b>
2.1 Principi delle applicazioni di rete . . . . .	18
2.2 Web e Protocollo HTTP . . . . .	21
2.2.1 Messaggi di richiesta e risposta . . . . .	23
2.2.2 Versioni di HTTP . . . . .	26
2.2.3 Cookies e Web Caching . . . . .	27
2.3 Posta elettronica . . . . .	28
2.3.1 Protocolli SMTP e MIME . . . . .	29
2.3.2 Protocolli POP3 e IMAP . . . . .	31
2.4 Domain Name System (DNS) . . . . .	33
2.4.1 Gerarchia server DNS . . . . .	33
2.4.2 Protocollo DNS . . . . .	36
2.5 Trasferimento di file . . . . .	37
2.5.1 Protocollo FTP . . . . .	37
2.5.2 Protocollo BitTorrent . . . . .	39
<b>3 Livello di Trasporto</b>	<b>40</b>
3.1 Multiplexing e Demultiplexing . . . . .	40
3.2 Protocollo UDP . . . . .	42
3.3 Trasferimento affidabile dei dati . . . . .	44
3.3.1 Protocollo RDT 1.0 e 2.0 . . . . .	46
3.3.2 Protocollo RDT 2.1 e 2.2 . . . . .	47
3.3.3 Protocollo RDT 3.0 . . . . .	49
3.3.4 Go-back-N e Selective repeat . . . . .	51
3.4 Protocollo TCP . . . . .	56
3.4.1 Gestione del timeout e stima del RTT . . . . .	58
3.4.2 Controllo del flusso . . . . .	60
3.4.3 Gestione della connessione . . . . .	61

3.5	Controllo della congestione . . . . .	64
3.5.1	Cause e costi della congestione . . . . .	64
3.5.2	Controllo della congestione nel TCP . . . . .	67
3.6	Equità nei protocolli di trasporto . . . . .	73
<b>4</b>	<b>Livello di Rete</b>	<b>74</b>
4.1	Panoramica del livello di rete . . . . .	74
4.2	Architettura e funzionalità dei router . . . . .	76
4.2.1	Accodamento nelle porte . . . . .	79
4.2.2	Scheduling dei pacchetti . . . . .	81
4.2.3	Frammentazione dei datagrammi . . . . .	82
4.3	Protocollo IP . . . . .	83
4.3.1	Protocollo DHCP e indirizzamento gerarchico . . . . .	85
4.3.2	Servizio NAT e Protocollo IPv6 . . . . .	87
4.4	Protocollo ICMP e Traceroute . . . . .	90
4.5	API OpenFlow e forwarding generalizzato . . . . .	92
4.6	Principi architetturali di Internet . . . . .	94
4.7	Algoritmi di instradamento . . . . .	95
4.7.1	Algoritmo link-state di Dijkstra . . . . .	96
4.7.2	Algoritmo Distance-vector . . . . .	98
4.8	Instradamento intra-AS e inter-AS . . . . .	103
4.8.1	Protocolli RIP e OSPF . . . . .	104
4.8.2	Protocollo BGP . . . . .	107
4.9	Tipologie di instradamento . . . . .	111
4.9.1	Unicast e Broadcast . . . . .	111
4.9.2	Multicast . . . . .	113
4.10	Software Defined Networking (SDN) . . . . .	116
4.11	Gestione della rete . . . . .	121

# Capitolo 0

## Introduzione

# Capitolo 1

## Introduzione alle reti

### 1.1 Rete, Host e Collegamenti

#### Definition 1. Rete e Link

Una **rete** è un'infrastruttura composta da dispositivi detti **nodi della rete** in grado di scambiarsi informazioni tramite dei mezzi di comunicazione, wireless o cablati, detti **link (o collegamenti)**

#### Definition 2. Nodi di una rete

I **nodi** costituenti una rete vengono differenziati in **due macro-categorie**:

- **Sistemi terminali**, differenziati a loro volta in
  - **Host**, ossia un dispositivo di proprietà dell'utente dedicato ad eseguire applicazioni utente
  - **Server**, ossia un dispositivo di elevate prestazioni destinato ad eseguire programmi che forniscono un servizio a diverse applicazioni utente
- **Dispositivi di interconnessione**, ossia dei dispositivi atti a modificare o prolungare il segnale ricevuto, differenziati a loro volta in:
  - **Router**, ossia dispositivi che collegano una rete ad una o più reti
  - **Switch**, ossia dispositivi che collegano più sistemi terminali all'interno di una rete
  - **Modem**, ossia dispositivi in grado di trasformare la codifica dei dati in segnale e viceversa

In particolare, classifichiamo le varie tipologie di rete in:

- **Personal Area Network (PAN)**, avente scala ridotta, solitamente equivalente a pochi metri (es: una rete Bluetooth)

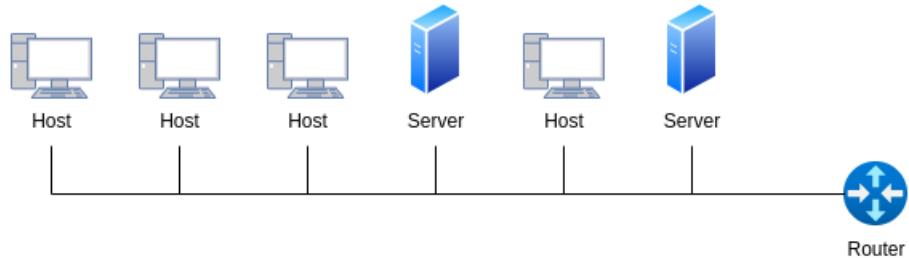
- **Local Area Network (LAN)**, solitamente corrispondente ad una rete privata che collega i sistemi terminali di un appartamento (es: una rete Wi-Fi o Ethernet). Ogni sistema terminale possiede un indirizzo che lo identifica univocamente all'interno della LAN.

Si differenziano in **LAN con cavo condiviso**, ossia dove tutti i dispositivi sono connessi al router tramite un cavo comune, e **LAN con switch**, ossia dove tutti i dispositivi sono connessi ad uno o più switch, i quali a loro volta sono connessi al router

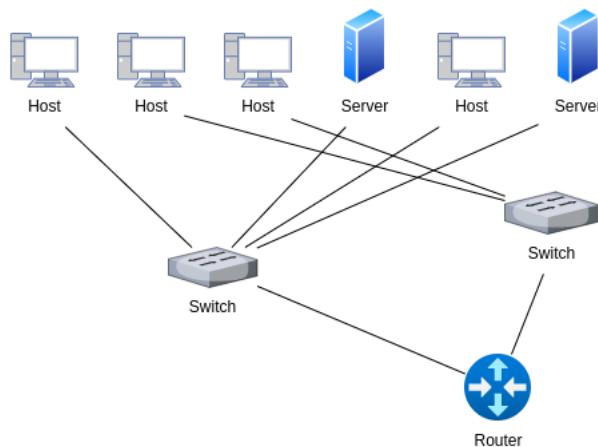
- **Metropolitan Area Network (MAN)**, avente scala pari ad una città
  - **Wide Area Network (WAN)**, avente scala pari ad un paese o una nazione, solitamente gestita da un **Internet Service Provider (ISP)**.
- Si differenziano in **WAN punto-punto**, ossia collegante due reti tramite un singolo mezzo mezzo di trasmissione, e **WAN a commutazione**, ossia collegante più reti tramite più mezzi e dispositivi di collegamento
- **L'Internet**, avente scala globale

Esempi:

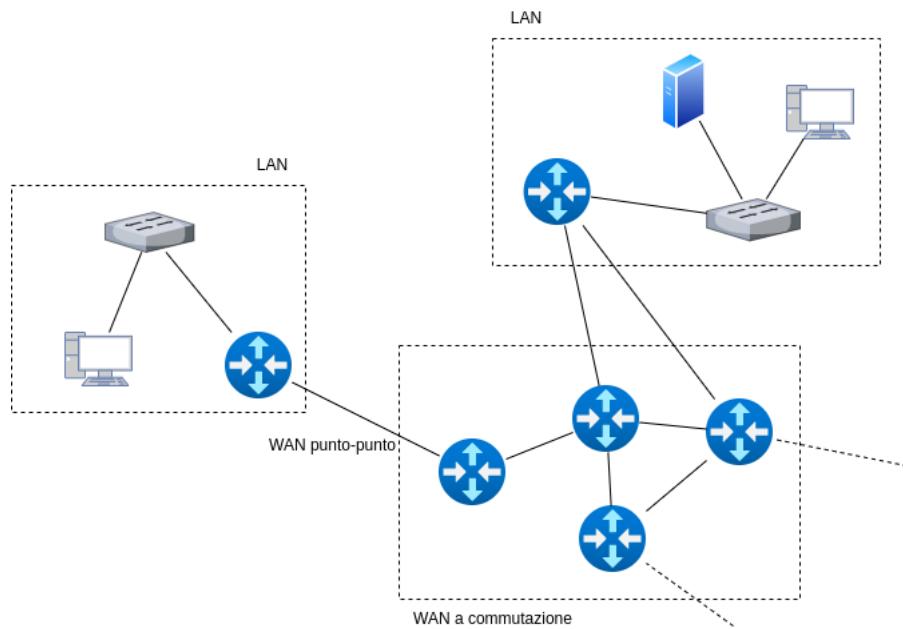
- **LAN a cavo condiviso**



- **LAN con switch**



- **Rete composta**



I supporti fisici utilizzabili per una trasmissione si differenziano in:

- **Doppino intrecciato** (ad esempio un cavo Ethernet), composto da due fili di rame isolati, uno utilizzato per inviare i dati ed uno per riceverli
- **Cavo coassiale**, composto da due conduttori di rame concentrici, entrambi bidirezionali, avente una larghezza di banda maggiore
- **Cavo in fibra ottica**, composto da una fibra di vetro che trasporta impulsi luminosi (dunque alla velocità della luce) al suo interno, ognuno rappresentante un singolo bit
- **Trasmissione wireless**, realizzata tramite l'invio di un segnale radio propagato nell'aria (es: rete cellulare, satellitare o Wi-Fi)

## 1.2 Struttura di Internet

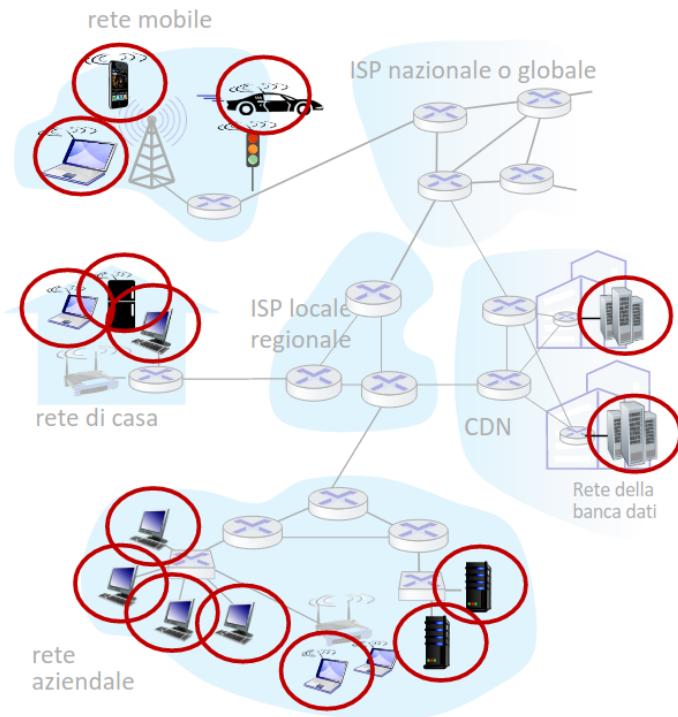
### Definition 3. Rete internet

Definiamo come **internet** (abbreviativo di internetwork) una **rete di reti**, ossia una rete che mette in comunicazione due o più reti tra di loro.

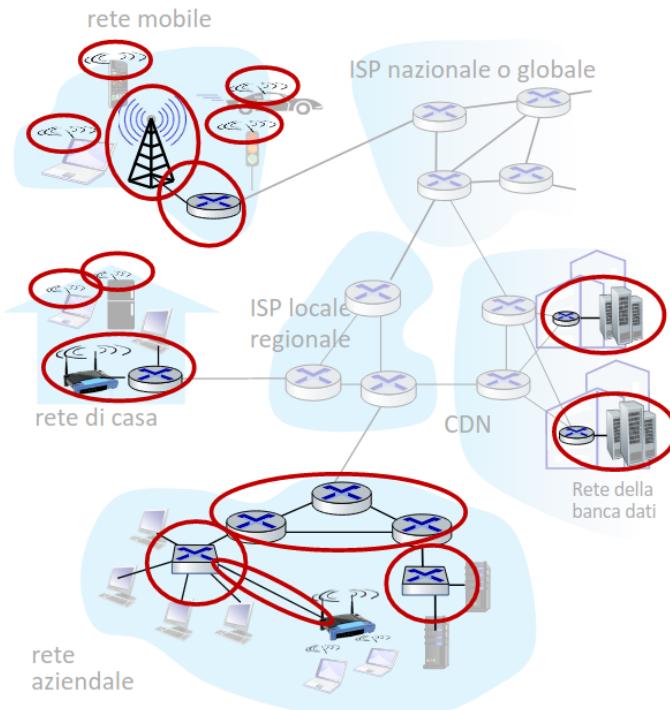
**Attenzione:** nonostante ciò che viene comunemente chiamato l'**Internet** sia una internet, è necessario puntualizzare che con tale termine comune viene indicata la **rete di tutte le reti**.

Al suo interno, la struttura di Internet risulta essere composta da:

- **Periferia della rete (network edge)**, corrispondente all'insieme di tutti i sistemi terminali connessi.

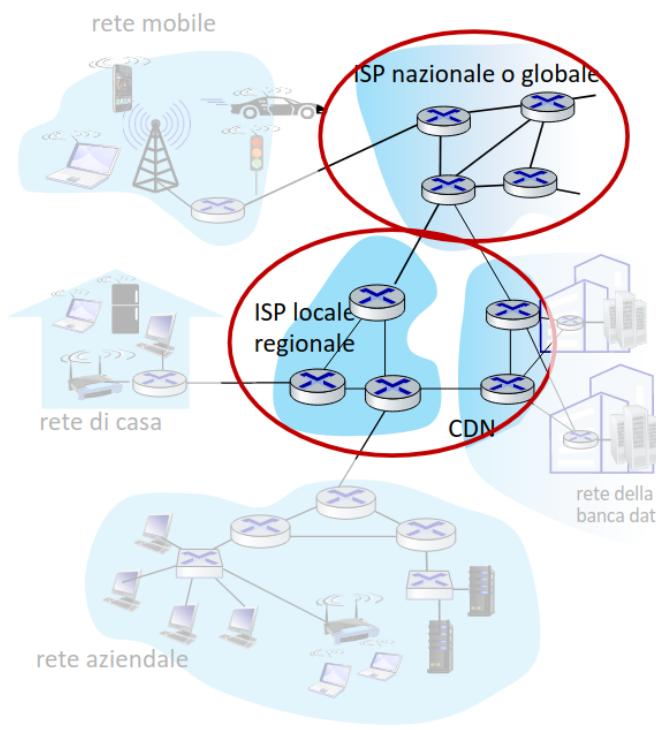


- **Reti di accesso (access network)**, corrispondente ai collegamenti fisici che connettono un sistema terminale al primo **edge router**, ossia il primo router presente nel percorso dal sistema terminale di origine ad un qualsiasi altro sistema terminale di destinazione.



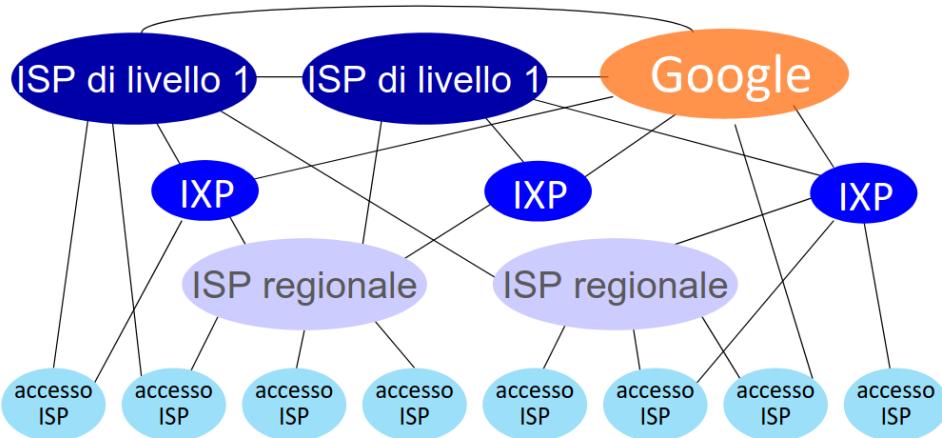
In particolare, l'accesso all'Internet può essere effettuato in più modi:

- **Accesso via cavo**, tramite supporti fisici connessi direttamente ad una rete di distribuzione, detta **cable headend** (es: il centralino di un ISP).
- **Accesso via Digital Subscriber Line (DSL)**, dove viene utilizzata la linea telefonica esistente per collegarsi alla rete dell'ISP
- **Accesso via Wireless LAN (WLAN)**, tramite un collegamento wireless ad una stazione base detta **access point** connessa con il router, a sua volta connesso con un cable headend
- **Accesso via rete cellulare**, dove viene utilizzata la rete cellulare esistente per collegarsi alla rete dell'ISP
- **Accesso via rete aziendale**, tramite una rete aziendale (o universitaria, privata, ...) direttamente connessa ad Internet
- **Nucleo di rete (core o backbone)**, ossia un sistema di router interconnessi tra di loro, corrispondente all'insieme di nodi cui viene realizzata la vera interconnessione tra tutte le reti.



In particolare, all'interno del backbone di Internet sono presenti **più livelli di reti ISP** (es: regionali, nazionali, aziendali, ...), le quali devono essere interconnesse tra di loro tramite degli **Internet Exchange Point (IXP)**.

Inoltre, nel recente periodo, nel backbone di Internet sono state integrate anche delle grandi reti private aziendali, ossia le **reti dei content provider** (es: Google, Netflix, ...), le quali, ormai, funzionano come vere e proprie ISP.



## 1.3 Pacchetti, Forwarding e Routing

### Definition 4. Pacchetto e Velocità di trasmissione

Dato un messaggio  $m$  da trasferire tra due terminali, definiamo come **pacchetti** l'insieme di blocchi di  $L$  bit tali che  $m = \{p_1, \dots, p_k\}$ .

Ogni pacchetto viene trasmesso nella rete ad una **velocità di trasmissione  $R$**  (anche detta larghezza di banda o capacità del collegamento).

### Definition 5. Forwarding e Routing

Le funzioni fondamentali di una rete si dividono in:

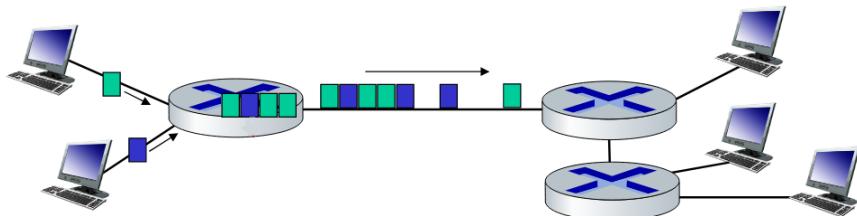
- **Forwarding o Switching (commutazione)**, ossia un'azione locale tramite cui vengono spostati i pacchetti in arrivo dal collegamento di ingresso del router al collegamento appropriato di uscita. Viene effettuato attraverso una **local forwarding table**, contenente gli indirizzi dei nodi locali
- **Routing (instradamento)**, ossia un'azione globale tramite cui vengono determinati i percorsi origine-destinazione seguiti dai pacchetti. Viene effettuato tramite **algoritmi di instradamento**

In particolare, la commutazione può avvenire in due modi:

- **Commutazione di pacchetto:**

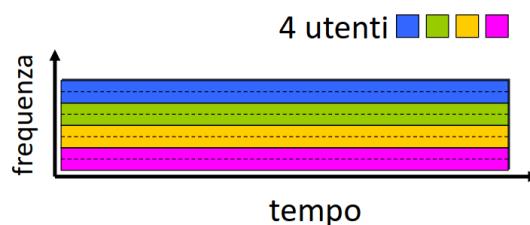
- La rete inoltra i pacchetti da un router all'altro attraverso i collegamenti presenti nell'instradamento dall'origine alla destinazione.
- Una volta inviato, un pacchetto deve completamente raggiungere il nodo a cui sta attualmente venendo inviato prima di poter essere trasmesso al collegamento successivo (**store & forward**)

- Se la velocità di trasmissione sul link di entrata supera la velocità di trasmissione di quello in uscita, i pacchetti verranno messi all'interno di una coda, in attesa di essere trasmessi sul link di uscita
- Se il buffer della coda raggiunge capienza massima, i pacchetti verranno scartati (**perdita di pacchetti**), per poi, se necessario, essere rinviati

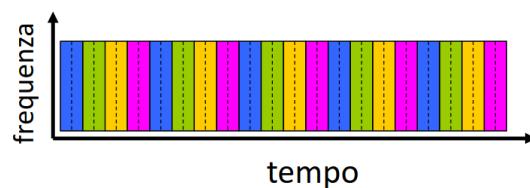


- **Commutazione di circuito:**

- La banda dei mezzi di trasmissione viene **suddivisa** in parti, riservando ognuna di essere ad una comunicazione tra un'origine ed una destinazione.
- Per via di tale suddivisione, il numero di utenti massimo della rete risulta essere **limitato dal numero di suddivisioni**
- La suddivisione può essere effettuata in due modalità:
  - \* **Frequency Division Multiplexing (FDM)**, dove le frequenze del mezzo di trasmissione vengono suddivise in bande di frequenza, ognuna di esse riservata ad una singola comunicazione, la quale può utilizzare al massimo la banda ad essa riservata



- \* **Time Division Multiplexing (TDM)**, dove il tempo viene suddiviso in slot, ognuno di essi riservato ad una singola comunicazione, la quale può utilizzare l'intera banda del mezzo per il breve lasso di tempo dedicato.



Nonostante la **commutazione di pacchetto** permetta l'accesso di un numero maggiore di utenti e non necessiti di stabilire una configurazione del collegamento, la presenza di una possibile perdita di pacchetti rende tale tipo di commutazione prettamente ottimo per **trasmissioni "bursty"**, ossia intermittenti e con lunghi periodi di inattività.

## 1.4 Misura delle prestazioni

### Definition 6. Larghezza di banda e Transmission rate

Con il termine **larghezza di banda (bandwidth)** indichiamo due concetti strettamente legati tra loro:

- La quantità (espressa in  $Hz$ ) rappresentante la **larghezza dell'intervallo di frequenze** utilizzato dal sistema trasmittivo, ossia l'intervallo di frequenze utilizzato dal sistema trasmittivo. Maggiore è tale quantità, maggiore è la quantità di informazioni veicolabili tramite il mezzo di trasmissione.
- La quantità (espressa in  $b/s$ ) detta anche **transmission rate (o bit rate)** rappresentante la **quantità di bit al secondo** che un link **garantisce di trasmettere**. Tale quantità è proporzionale alla larghezza di banda (in  $Hz$ )

### Definition 7. Throughput

Con il termine **throughput** indichiamo la **quantità di bit** al secondo che **passano attraverso un nodo** della rete.

### Observation 1

A differenza del **transmission rate**, il quale fornisce una misura della **potenziale velocità di un link**, il **throughput** fornisce una misura dell'**effettiva velocità di un link**.

In generale, dunque, si ha che

$$T < R$$

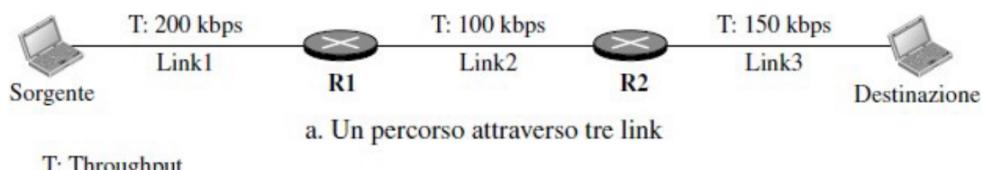
dove  $T$  è il throughput e  $R$  è il transmission rate

### Definition 8. Collo di bottiglia

Dato un percorso end-to-end, ossia tra un dispositivo e un altro, definiamo come **collo di bottiglia** il link limitante il throughput dei link presenti su tale percorso

Esempio:

- Consideriamo il seguente percorso



- Il link  $L_2$  risulta essere il collo di bottiglia di tale percorso, limitando il throughput del percorso a 100 kb/s

**Definition 9. Delay di trasmissione**

Definiamo come **delay (o latenza) di trasmissione** il tempo necessario ad un nodo per immettere un pacchetto su un link, corrispondente a:

$$D_t = \frac{L}{R}$$

dove  $L$  è la dimensione del pacchetto e  $R$  è il transmission rate del link

**Definition 10. Delay di propagazione**

Definiamo come **delay (o latenza) di propagazione** il tempo impiegato dall'**ultimo bit di blocco di dati** posto su un link ad essere propagato fino al nodo di destinazione, corrispondente a:

$$D_p = \frac{k}{v}$$

dove  $k$  è la lunghezza del link e  $v$  è la velocità di propagazione del link

**Definition 11. Delay di un pacchetto**

Definiamo come **delay (o latenza) di un pacchetto** il tempo totale necessario ad un pacchetto per essere inviato completamente da un nodo origine ad un nodo destinatario

$$D_n = D_e + D_q + D_t + D_p$$

dove:

- $D_e$  è il **delay di elaborazione del nodo**, dipendente dalle operazioni di controllo svolte dal nodo
- $D_q$  è il **delay di queueing**, ossia l'attesa del pacchetto all'interno della coda del nodo prima di essere trasmesso, dipendente dalla quantità di pacchetti presenti nella coda
- $D_t$  è il delay di trasmissione del link
- $D_p$  è il delay di propagazione del link

**Proposition 1. Prodotto rate per delay di propagazione**

Dato un link con transmission rate  $R$  e delay di propagazione  $D_p$ , il prodotto

$$B_{max} = R \cdot D_p = \frac{L \cdot k}{D_t \cdot v}$$

rappresenta il **massimo numero di bit distribuiti tutto sul cavo** contemporaneamente

**Esempi:**

1. Si consideri un router A che trasmette pacchetti, ognuno di lunghezza  $L = 4000$  bit, su un canale di trasmissione con rate  $R = 10 \text{ Mb/s}$  verso un router B all'altro estremo del link. Si supponga che il delay di propagazione sia pari a 0.2 ms.

- Quanto impiega il router A a trasmettere un pacchetto al router B?

$$D_t = \frac{L}{R} = \frac{4 \cdot 10^3 \text{ b}}{10^7 \text{ b/s}} = 4 \cdot 10^{-4} \text{ s} = 0.4 \text{ ms}$$

- Quanto impiega il router A a trasmettere un bit al router B?

$$D_{1b} = \frac{1}{R} = \frac{1 \text{ b}}{10^7 \text{ b/s}} = 10^{-7} \text{ s} = 0.1 \mu\text{s}$$

- Qual è il massimo numero di pacchetti al secondo che possono essere trasmessi sul link?

$$\begin{aligned} 1 \text{ P} &= 4000 \text{ b} \implies 1 \text{ b} = \frac{1}{4000} \text{ P} \implies \\ &\implies R = 10^7 \text{ b/s} = \frac{10^7}{4000} \text{ P/s} = \frac{1}{4} \cdot 10^3 \text{ P/s} = 2500 \text{ P/s} \end{aligned}$$

- Supponendo che il router A invii i pacchetti uno dopo l'altro senza introdurre ritardi tra la trasmissione di un pacchetto e il successivo, quanto tempo impiega il router B a ricevere 4 pacchetti?

Poiché i pacchetti vengono inviati senza alcun delay tra di essi, possiamo considerare tali pacchetti come un unico grande pacchetto di dimensione  $4 \cdot L$ , implicando che

$$D_{4t} = \frac{4 \cdot L}{R} = \frac{16 \cdot 10^3 \text{ b}}{10^7 \text{ b/s}} = 16 \cdot 10^{-4} \text{ s} = 1.6 \text{ ms}$$

Inoltre, per lo stesso motivo, il tempo di propagazione rimarrà inalterato, poiché esso non dipende dalla lunghezza del pacchetto, ma solo dalla lunghezza e della velocità di propagazione del link. Di conseguenza, il tempo totale impiegato sarà  $1.6 \text{ ms} + 0.2 \text{ ms} = 1.8 \text{ ms}$

- Qual è il massimo numero di bit e il numero di pacchetti che possono essere presenti sul canale?

$$P_{max} = R \cdot D_p = 10^7 \text{ b/s} \cdot 0.2 \text{ ms} = 2000 \text{ b} = \frac{1}{2} \text{ P}$$

2. Si consideri un host A che vuole inviare un file molto grande, 4 milioni di byte, a un host B. Il percorso tra A e B ha 3 link  $L_1, L_2, L_3$ , ognuno di lunghezza 300 km, ciascuno con rate rispettivo  $R_1 = 500 \text{ kb/s}$ ,  $R_2 = 2 \text{ Mb/s}$  e  $R_3 = 1 \text{ Mb/s}$ .

- Assumendo l'assenza di ulteriore traffico nella rete, qual è il throughput per il file transfer?

Poiché il link  $L_1$  risulta essere il collo di bottiglia del percorso, il throughput risulta essere  $R_1 = 500 \text{ kb/s}$

- Qual è il tempo totale impiegato per trasferire il file all'host B assumendo che i link siano cavi in fibra ottica?

Poiché non vi è specificata la lunghezza di ogni pacchetto, assumiamo che il file venga inviato come un unico grande pacchetto, implicando che  $L = 4 \cdot 10^6 \text{ b} = 32 \cdot 10^6 \text{ b}$ .

Di conseguenza, si ha che:

$$D_t(L_1) = \frac{32 \cdot 10^6 \text{ b}}{5 \cdot 10^5 \text{ b/s}} = 64 \text{ s}$$

$$D_t(L_2) = \frac{32 \cdot 10^6 \text{ b}}{2 \cdot 10^6 \text{ b/s}} = 16 \text{ s}$$

$$D_t(L_3) = \frac{32 \cdot 10^6 \text{ b}}{1 \cdot 10^6 \text{ b/s}} = 32 \text{ s}$$

Poiché  $L_1, L_2, L_3$  sono cavi in fibra ottica, la velocità di propagazione su di essi corrisponde alla velocità della luce, pari a  $\sim 3 \cdot 10^8 \text{ m/s}$ . Dunque, il delay di propagazione di ogni link corrisponderà a:

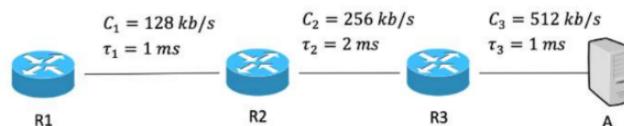
$$D_p = \frac{3 \cdot 10^5 \text{ m}}{3 \cdot 10^8 \text{ m/s}} = 1 \text{ ms}$$

Infine, concludiamo che il tempo totale impiegato corrisponda a:

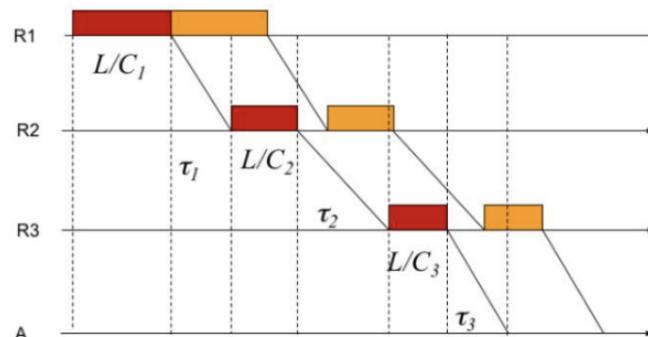
$$D_{tot} = D_t(L_1) + D_t(L_2) + D_t(L_3) + 3 \cdot D_p = 64 \text{ s} + 16 \text{ s} + 32 \text{ s} + 3 \cdot 1 \text{ ms} = 112,003 \text{ s}$$

3. Si consideri la rete nella seguente figura, dove  $C_1, C_2, C_3$  e  $\tau_1, \tau_2, \tau_3$  sono rispettivamente i transmission rate e i delay di propagazione dei tre link.

Al tempo  $t = 0$ , la coda di uscita di  $R_1$  contiene 2 pacchetti diretti ad  $A$ . Assumendo che la lunghezza dei pacchetti sia  $L = 512 \text{ b}$ , si indichi per ciascun pacchetto l'istante in cui esso viene completamente ricevuto da  $A$ .



Per aiutarci durante il calcolo, tracciamo il contenuto di ogni coda al passare del tempo:



Dunque, il tempo totale impiegato dal primo pacchetto corrisponderà a:

$$T_1 = \frac{L}{C_1} + \tau_1 + \frac{L}{C_2} + \tau_2 + \frac{L}{C_3} + \tau_3 = 4 \text{ ms} + 1 \text{ ms} + 2 \text{ ms} + 2 \text{ ms} + 1 \text{ ms} + 1 \text{ ms} = 11 \text{ ms}$$

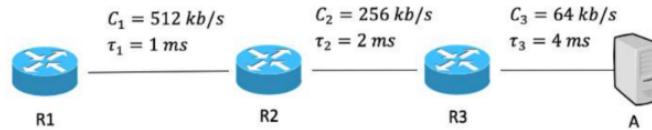
Analogamente, il tempo totale impiegato dal secondo pacchetto corrisponderà a:

$$T_2 = 2 \cdot \frac{L}{C_1} + \tau_1 + \frac{L}{C_2} + \tau_2 + \frac{L}{C_3} + \tau_3 = 8 \text{ ms} + 1 \text{ ms} + 2 \text{ ms} + 2 \text{ ms} + 1 \text{ ms} + 1 \text{ ms} = 15 \text{ ms}$$

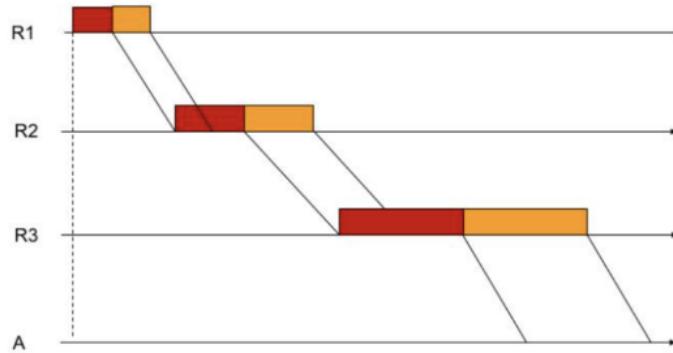
4. Si consideri la rete nella seguente figura, dove  $C_1, C_2, C_3$  e  $\tau_1, \tau_2, \tau_3$  sono rispettivamente i transmission rate e i delay di propagazione dei tre link.

Al tempo  $t = 0$ , la coda di uscita di  $R_1$  contiene 2 pacchetti diretti ad  $A$ . Assumendo che la lunghezza dei pacchetti sia  $L = 512 \text{ b}$ , si indichi per ciascun pacchetto l'istante in cui esso viene completamente ricevuto da  $A$ .

Inoltre, supponendo che vi siano  $n$  pacchetti, si indichi una formula generica descrivente per ciascun pacchetto l'istante in cui esso viene completamente ricevuto da  $A$



Come nel caso precedente, tracciamo il contenuto di ogni coda al passare del tempo:



Il tempo totale impiegato dal primo pacchetto corrisponderà a:

$$T_1 = \frac{L}{C_1} + \tau_1 + \frac{L}{C_2} + \tau_2 + \frac{L}{C_3} + \tau_3 = 1 \text{ ms} + 1 \text{ ms} + 2 \text{ ms} + 2 \text{ ms} + 8 \text{ ms} + 4 \text{ ms} = 18 \text{ ms}$$

Notiamo come, a differenza del caso precedente, il secondo pacchetto giunge nelle code successive mentre il primo pacchetto deve essere ancora completamente spedito, implicando che esso debba essere inserito nella coda di attesa.

Dunque, una volta raggiunta la coda finale, il secondo pacchetto potrà essere inviato solo una volta completato il primo pacchetto.

Di conseguenza, il suo tempo totale di ricezione corrisponde a:

$$T_2 = T_1 + \frac{L}{C_3} = 18 \text{ ms} + 8 \text{ ms} = 26 \text{ ms}$$

Applicando lo stesso ragionamento nel caso di  $n$  pacchetti, la formula generica descrivente l'istante di ricezione dell' $n$ -esimo pacchetto corrisponde a:

$$T_n = T_{n-1} + \frac{L}{C_3} = T_{n-2} + 2 \cdot \frac{L}{C_3} = \dots = T_1 + (n-1) \frac{L}{C_3} = 18 \text{ ms} + 8(n-1) \text{ ms}$$

## 1.5 Stack protocollare TCP/IP

### Definition 12. Protocollo

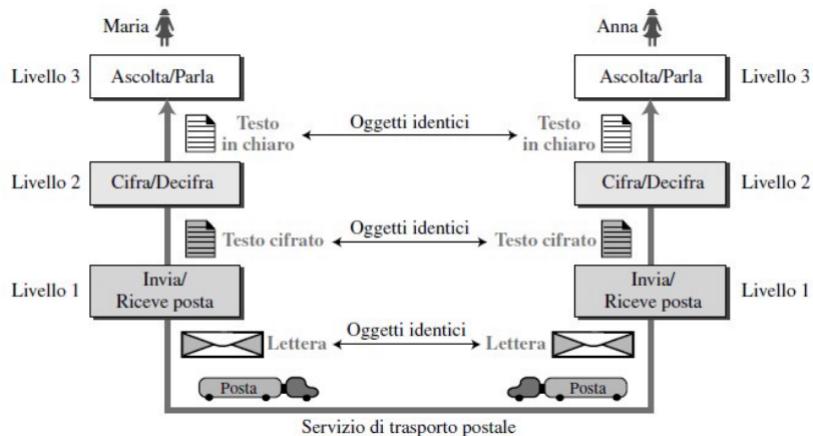
Un **protocollo** definisce l'insieme di **regole** che il dispositivo mittente e il dispositivo destinatario, così come tutti i sistemi intermedi coinvolti, devono rispettare per essere in grado di comunicare.

In situazioni più complesse, potrebbe essere opportuno suddividere i compiti necessari alla comunicazione fra **più livelli (layer)**, nel qual caso è richiesto **un protocollo per ciascun livello**. Tramite un layering dei protocolli, dunque, è possibile suddividere un compito complesso in compiti più semplici, ognuno gestibile da un singolo protocollo.

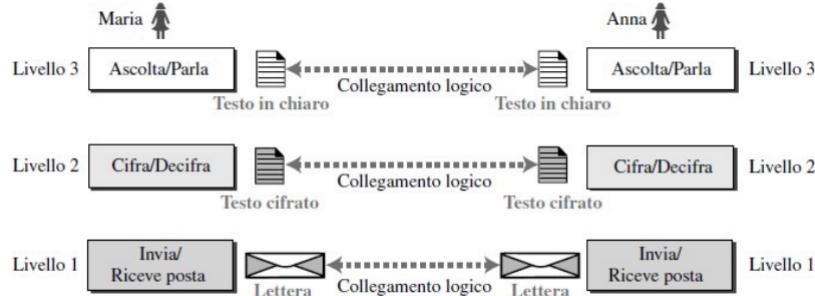
In particolare, ogni layer è **indipendente dagli altri** (modularizzazione), utilizzando i servizi forniti dal layer inferiore e offrendo servizi al layer superiore.

Ogni layer, dunque, può essere considerato come una **black box** con opportuni ingressi ed uscite, senza necessità di essere a conoscenza delle modalità con cui i dati in ingresso vengano trasformati in quelli di uscita.

Quando è richiesta una **comunicazione bidirezionale**, ciascun layer deve essere in grado di effettuare entrambi i compiti richiesti, ossia manipolare i dati in input per inviarli al livello superiore o manipolarli per inviarli al livello inferiore.



In particolare, l'effetto ottenuto tramite una suddivisione in uno stack di layer equivalenti permette l'instaurazione di un **collegamento logico** tra ogni livello dello stack: il protocollo implementato in ciascun livello specifica una comunicazione diretta tra i pari livelli delle due parti: il layer  $N$  di un dispositivo comunica solo ed esclusivamente con il layer  $N$  di tutti i dispositivi.



Inoltre, per via dell'estrema modularizzazione ottenuta, viene facilitata la manutenzione e l'aggiornamento del sistema, poiché il cambiando dell'implementazione del servizio di un layer rimane trasparente al resto del sistema.

### Definition 13. Stack protocollare TCP/IP

La principale forma di stack protocollare utilizzata corrisponde allo **stack protocolare TCP/IP**, la cui struttura a layer corrisponde a:

- **Livello di Applicazione**, il quale fornisce supporto alle applicazioni facente uso della rete (protocolli HTTP, SMTP, FTP, DNS, ...).
- **Livello di Trasporto**, il quale gestisce il trasferimento dei pacchetti dal processo del dispositivo mittente a quello del dispositivo destinatario (protocolli TCP, UDP, ...).
- **Livello di Rete**, il quale gestisce l'instradamento dei pacchetti dall'origine alla destinazione (protocolli IP, ...).
- **Livello di Collegamento (o Link)**, il quale gestisce la trasmissione dei pacchetti da un nodo a quello successivo sul percorso (protocolli Ethernet, Wi-Fi, PPP, ...). Lungo il percorso, un pacchetto può essere gestito da protocolli diversi.
- **Livello Fisico**, dove avviene il vero e proprio trasferimento dei singoli bit

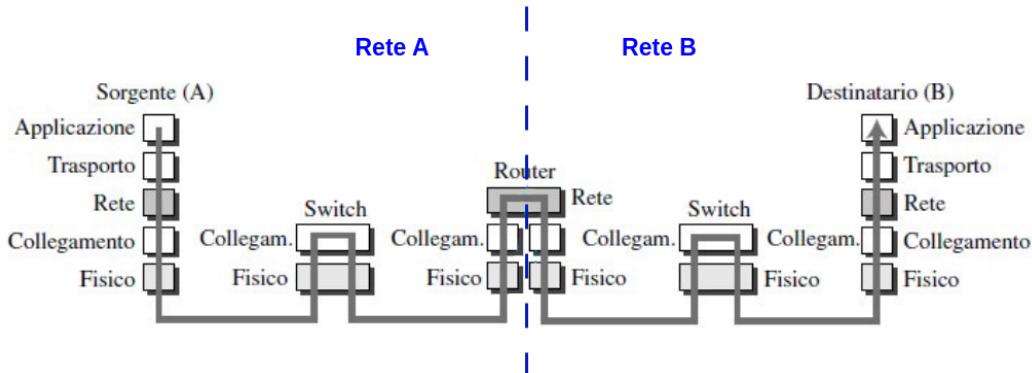


I livelli di Applicazione e di Trasporto sono gestiti tramite **software**, mentre i livelli di Collegamento e Fisico tramite **hardware**.

Durante l'invio di un pacchetto, quest'ultimo, partendo dal livello applicazione del dispositivo sorgente, **percorre tutti i layer dello stack protocollare**, fino a giungere al livello fisico, dove viene effettivamente inviato al nodo successivo.

Tutti i nodi intermedi presenti sul percorso lavoreranno utilizzando solo i livelli necessari. In particolare, ogni dispositivo utilizzerà il livello di collegamento, in modo da poter spedire il pacchetto stesso verso il nodo successivo del percorso.

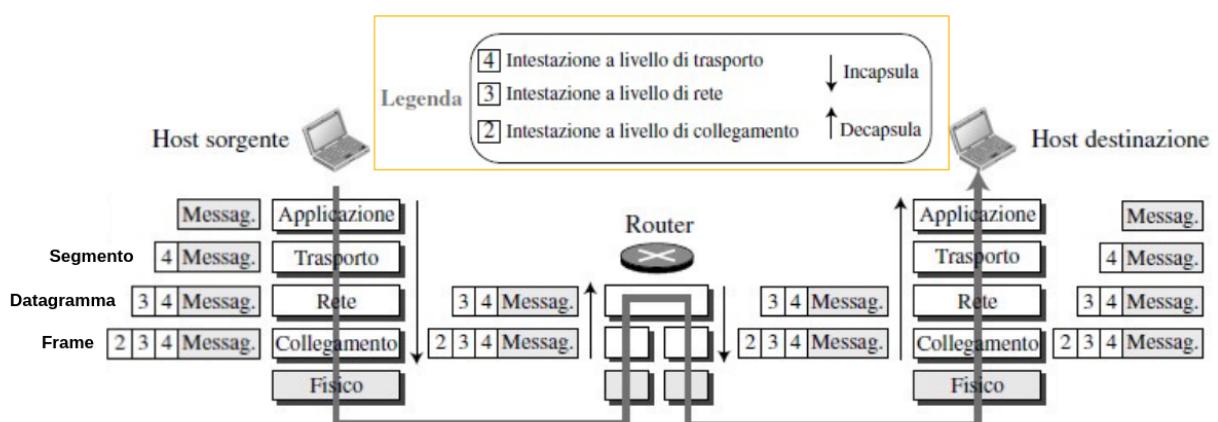
Nel caso in cui si raggiunga il **punto di scambio tra due reti**, solitamente un edge router, verrà utilizzato anche il livello di rete.



Prima di essere spedito al livello inferiore, ogni pacchetto viene **incapsulato**: una volta ricevuto il pacchetto dal layer superiore, il layer attuale applica un proprio **header (o intestazione)**, aggiungendo informazioni necessarie al layer del dispositivo di destinazione corrispondente a quello attuale.

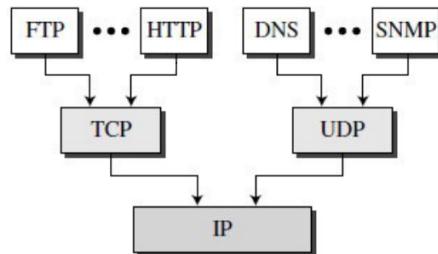
In particolare, ad ogni livello un pacchetto assume il nome di:

- **Messaggio (al livello di applicazione)**, corrispondente al pacchetto originale, senza alcuna intestazione
- **Segmento (al livello di trasporto)**, corrispondente al messaggio ricevuto dal layer superiore a cui viene aggiunto un header di trasporto
- **Datagramma (al livello di rete)**, corrispondente al segmento ricevuto dal layer superiore a cui viene aggiunto un header di rete
- **Frame (al livello di collegamento)**, corrispondente al datagramma ricevuto dal layer superiore a cui viene aggiunto un header di collegamento

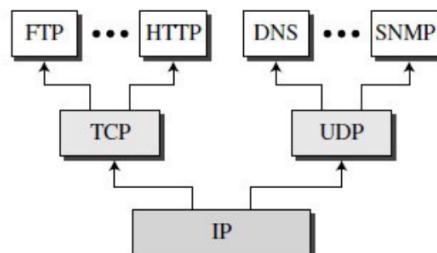


Poiché lo stack protocollare TCP/IP prevede la presenza di **più protocolli nello stesso livello**, ogni livello deve essere in grado di effettuare operazioni di:

- **Multiplexing**, dove ogni protocollo deve essere in grado di encapsulare (uno alla volta) i pacchetti ricevuti da più protocolli presenti al livello superiore



- **Demultiplexing**, dove ogni protocollo deve essere in grado di decapsulare i pacchetti ricevuti ed inviarli a più protocolli presenti nel livello superiore



Per realizzare ciò, nell'header di ogni layer viene inserito un **campo speciale** in grado di identificare quale sia il protocollo di appartenenza di tale pacchetto.

Un'evoluzione dello stack protocollare TCP/IP è il **modello Open Systems Interconnection (OSI)**, dove vengono interposti due livelli tra il livello di applicazione e il livello di trasporto:

- **Livello di Presentazione**, utilizzato per consentire alle applicazioni di interpretare i dati (es: crittografia, compressione, ...)
- **Livello di Sicurezza**, utilizzato per gestire servizi come la sincronizzazione o il ripristino dello scambio di dati



# Capitolo 2

## Livello di Applicazione

### 2.1 Principi delle applicazioni di rete

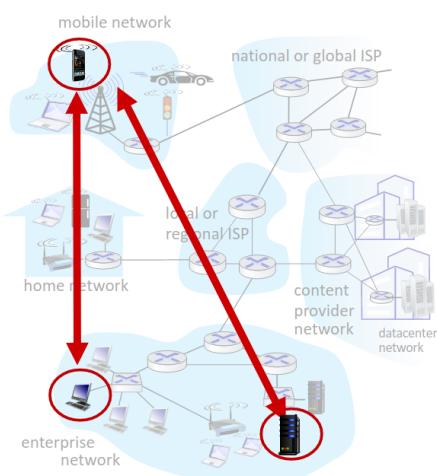
#### Definition 14. Paradigma di comunicazione

Un **paradigma di comunicazione** è una metodologia di scambio informazioni e gestione delle connessioni all'interno di una rete, principalmente all'interno di Internet.

In particolare, i due principali paradigmi utilizzati sono:

- **Paradigma Client-Server**, dove i sistemi terminali vengono divisi in due categorie:
  - **Client**, il quale comunica solo ed esclusivamente con un server, **richiedendo dei servizi** a quest'ultimo, e può rimanere anche inattivo se non necessario, implicando che esso possa avere indirizzi IP dinamici nel tempo. In particolare, per tali caratteristiche, non vi è una comunicazione client-client, ma solo una comunicazione client-server-client
  - **Server**, il quale possiede un indirizzo IP permanente, rimanendo sempre attivo in attesa di **fornire servizi** ai vari client richiedenti

Ad esempio, i protocolli HTTP, FTP e IMAP sono basati su tale paradigma

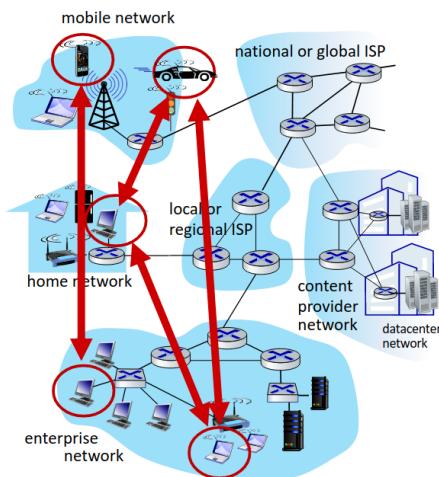


- **Paradigma Peer-to-Peer (P2P)**, dove i sistemi terminali vengono detti **peer** (tradotto: *pari, di equal importanza*) ed ognuno di essi è in grado di comunicare direttamente con ogni altro peer (assumendo quindi sia il compito di client che di server).

In particolare, ogni peer **richiede e fornisce servizi ad altri peer**, rendendo il sistema **estremamente scalabile**: ogni nuovo peer incrementa le capacità di servizio e le richieste di servizio.

Inoltre, come nel caso dei client, ogni peer può connettersi alla rete in modo intermittente utilizzando IP dinamici, diminuendo temporaneamente la quantità di servizi fornibili nella rete. Di conseguenza, la loro gestione risulta estremamente più complessa, ma anche più performante nel caso di un numero elevato di peer.

Un classico esempio di utilizzo del paradigma P2P risultano essere i vari protocolli legati al torrenting e alla condivisione di file di grandi dimensioni.



### Definition 15. Processo

Un **processo** è un programma in esecuzione all'interno di un sistema terminale.

In particolare, un **processo client** è un processo che avvia una comunicazione, mentre un **processo server** è un processo che attende di essere contattato da un processo client.

All'interno dello stesso sistema, due processi comunicano tra di loro utilizzando una comunicazione **inter-process**, definita dal sistema operativo. I processi situati su sistemi diversi, invece, comunicano tra di loro tramite **scambio di messaggi**

### Definition 16. Socket

Un **socket** è un'**astrazione software** tramite cui un processo può inviare e ricevere messaggi tramite il socket di un altro processo. Per poter comunicare, dunque, due processi devono connettersi tramite due socket (uno ciascuno), identificati da una coppia <Indirizzo\_IP, Numero\_Porta>

Ogni protocollo a livello di applicazione definisce:

- Le tipologie di messaggi scambiati (es: richiesta e risposta)
- La sintassi del messaggio
- La semantica del messaggio
- Le regole per come e quando i processi inviano e rispondono ai messaggi

In particolare, i protocolli a tale livello si differenziano in **protocolli aperti**, ossia definiti secondo uno standard pubblico ed adottato comunemente da ogni applicazione (es: HTTP, FTP, ...), e **protocolli proprietari**, ossia non pubblici e fini all'applicazione stessa (es: Skype, ...).

Per poter funzionare correttamente, ogni applicazione di rete necessita di alcuni **servizi di trasporto**. In particolare, esse possono necessitare di:

- **Integrità dei dati**, ossia un trasferimento dei dati affidabile al 100%, senza alcuna perdita di pacchetto o corruzione dei dati
- **Garanzie temporali**, ossia un basso ritardo per la ricezione dei dati
- **Garanzie di throughput**, ossia una quantità minima di throughput dati
- **Sicurezza**, ad esempio crittografia o integrità dei dati a seguito di manomissioni

### Definition 17. Transmission Control Protocol (TCP)

Il **Transmission Control Protocol (TCP)** è un protocollo risiedente sul **layer di trasporto** in grado di fornire **trasporto affidabile**, ossia senza perdita di alcun pacchetto, e controllo del flusso e della congestione, in cambio di un'assenza di garanzie temporali, di throughput e di sicurezza.

Inoltre, il protocollo TCP è **orientato alla connessione**, ossia richiedente una configurazione (**handshaking**) tra il processo client e il processo server

### Definition 18. User Datagram Protocol (UDP)

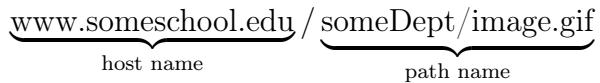
L'**User Datagram Protocol (UDP)** è un protocollo risiedente sul **layer di trasporto** in grado di fornire **trasporto veloce** poiché **non orientato alla connessione** ed **estremamente scarno**, ossia sprovvisto di: trasporto affidabile, controllo del flusso e della congestione e garanzie temporali, di throughput e di sicurezza

Poiché per loro natura i protocolli TCP ed UDP sono privi di garanzie di sicurezza, i messaggi scambiati tra socket TCP e UDP risultano sprovvisti di crittografia, attraversando il percorso instradato completamente in chiaro ed essendo quindi leggibili e manipolabili da qualsiasi dispositivo intermedio.

Per ovviare tale problema, viene implementato a livello di applicazione il protocollo **Transport Layer Security (TLS)** tramite socket realizzati con librerie software specifiche, fornendo connessioni crittografate, integrità dei dati ed autenticazione dell'end-point.

## 2.2 Web e Protocollo HTTP

Una pagina web è composta da **oggetti**, ognuno dei quali può essere archiviato su un diverso web server. In particolare, una pagina web consiste in un **file HTML** il quale include diversi oggetti referenziati tramite vari URL



### Definition 19. Protocollo HTTP

Il **protocollo HTTP (Hypertext Transfer Protocol)** è un protocollo a livello di applicazione utilizzato per la realizzazione di servizi web. La sua porta di riferimento comune all'interno dei socket è la **porta 80**.

Il protocollo HTTP è **stateless**, ossia non conservante alcuna informazione sulle richieste passate, e basato sul **paradigma client-server**, dove il client invia messaggi detti **richieste** e il server invia messaggi detti **risposte**.

Inoltre, il protocollo HTTP fa uso del **protocollo TCP**:

1. Il client avvia una connessione TCP con il server utilizzando la porta 80, rimanendo in attesa che il server accetti la connessione (TCP handshaking)
2. Vengono scambiati messaggi HTTP tra client e server
3. La connessione TCP viene chiusa

Le **connessioni HTTP** si differenziano in due tipologie:

- **Connessione non persistente**, dove viene aperta la connessione TCP e viene inviato massimo un oggetto prima di chiudere la connessione TCP
- **Connessione persistente (HTTP/1.1)**, dove viene aperta la connessione TCP e vengono inviati multipli oggetti in successione prima di chiudere la connessione TCP

**Esempio:**

1. Supponiamo che un utente inserisca l'URL dell'oggetto "www.someSchool.edu/ someDepartment/home.index", contenente del testo e 10 riferimenti ad immagini.
2. Il client HTTP dell'utente (browser, cURL, ...) avvia la connessione TCP con il server HTTP tramite la porta 80
3. Il server HTTP sull'host "www.someSchool.edu" riceve la richiesta di connessione, accettandola e notificando il client
4. Il client HTTP invia un messaggio di richiesta HTTP, contenente il path dell'oggetto desiderato, ossia "/someDept/index.html"
5. Il server HTTP riceve il messaggio di richiesta e invia il messaggio di risposta contenente l'oggetto desiderato, il quale a sua volta contiene i riferimenti alle 10 immagini.

6. A questo punto, si creano due scenari:

- Se la connessione non è persistente, il server chiude immediatamente la connessione TCP, implicando che l'intero processo debba essere ripetuto per tutti e 10 i riferimenti necessari
- Se la connessione è persistente, il client invierà in successione altre 10 richieste al server, richiedendo quindi solo la ripetizione dei passaggi 4 e 5 (per 10 volte), per poi chiudere la connessione TCP

### Definition 20. Round Trip Time (RTT)

Definiamo come **Round Trip Time (RTT)** il tempo impiegato da un pacchetto di piccole dimensioni per compiere il percorso client-server-client

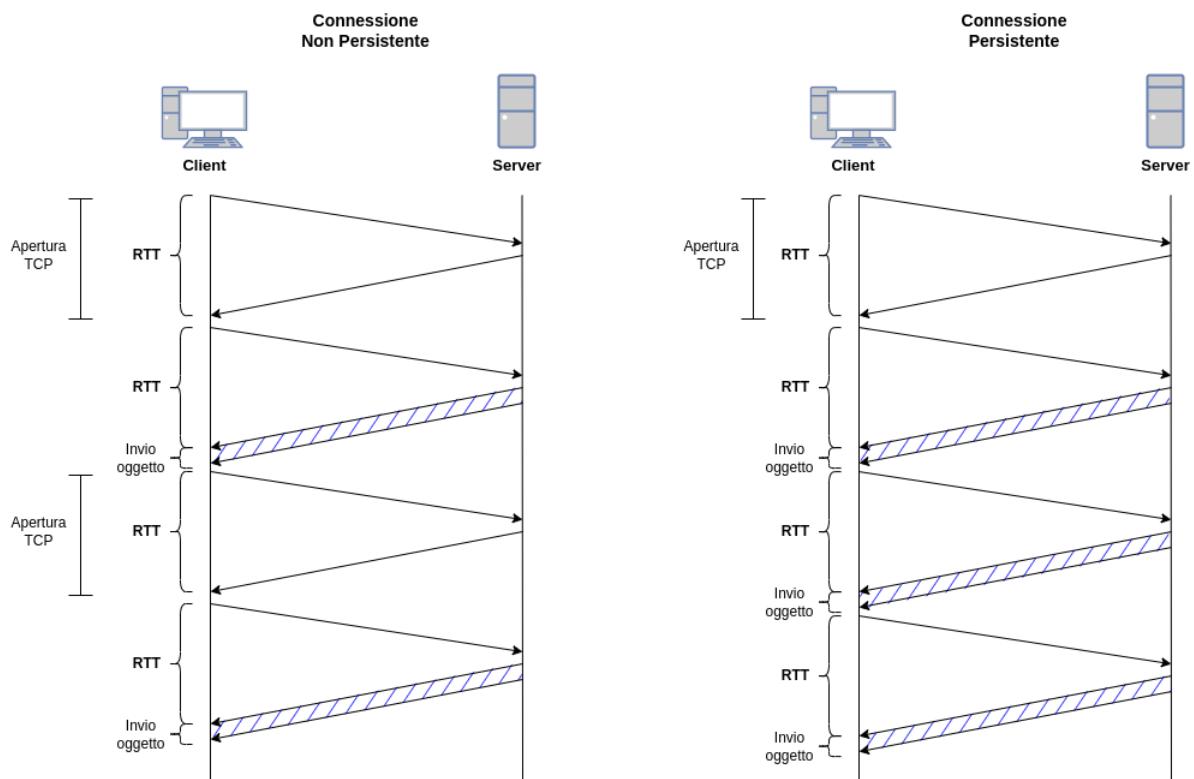
### Observation 2. Tempo di risposta HTTP

Se la **connessione non è persistente**, per ogni oggetto sono necessari due RTT, uno per avviare la connessione TCP ed uno per l'invio di richiesta e risposta, seguiti dal tempo necessario ad inviare l'oggetto

$$T_{tot} = (2 \text{ RTT} + \text{Tempo invio ogg.}) \cdot \text{Num. Oggetti}$$

Se la **connessione è persistente**, invece, saranno necessari un RTT per poter stabilire la connessione TCP, seguiti da un solo RTT per oggetto (con annesso tempo di invio)

$$T_{tot} = 1 \text{ RTT} + (1 \text{ RTT} + \text{Tempo invio ogg.}) \cdot \text{Num. Oggetti}$$



### 2.2.1 Messaggi di richiesta e risposta

I messaggi HTTP di **richiesta** e **risposta** vengono formattati un formato leggibile dall'uomo (in particolare, in codice ASCII).

Ogni messaggio di **richiesta HTTP** viene strutturato nel seguente modo:

- Una **riga di richiesta**, composta dal **metodo** utilizzato, il path richiesto e la versione di HTTP utilizzata, seguiti da un carattere di ritorno a capo, ossia \r, ed un carattere di avanzamento di riga, ossia \n

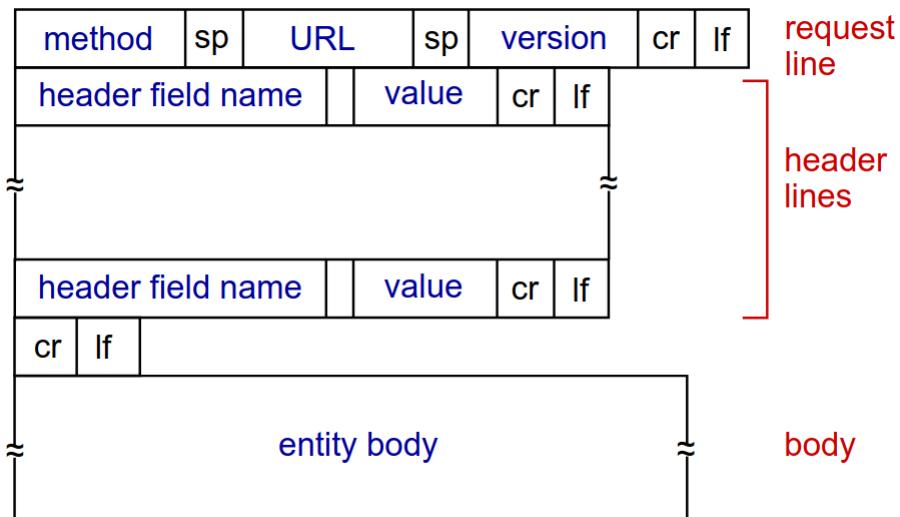
I **metodi** principali inseribili all'interno della riga di richiesta sono:

- **Metodo GET**, utilizzato per l'invio di dati al server, i quali vengono inseriti all'interno dell'URL a seguito di un carattere '?'  
(es: www.mysite.com/search?user=myuser)
- **Metodo POST**, utilizzato per l'invio di dati al server, i quali vengono aggiunti all'interno del body del messaggio (rimanendo quindi parzialmente offuscati all'utente)
- **Metodo HEAD**, utilizzato per richiedere solo l'header di risposta che verrebbe restituito dalla destinazione a seguito di una richiesta GET
- **Metodo PUT**, utilizzato per caricare un nuovo file o sostituirne uno esistente all'interno della destinazione (non più utilizzato poiché estremamente insicuro)
- Un **header (o intestazione)**, composto da varie linee contenenti informazioni utili alla connessione

Alcuni esempi di campi inseribili all'interno di un header di richiesta sono:

Campo Header	Descrizione
User-agent	Indica il programma client utilizzato
Accept	Indica il formato dei contenuti che il client è in grado di accettare
Accept-charset	Famiglia di caratteri che il client è in grado di gestire
Accept-encoding	Schema di codifica supportato dal client
Accept-language	Linguaggio preferito dal client
Authorization	Indica le credenziali possedute dal client
Host	Host e numero di porta del client
Date	Data e ora del messaggio
Upgrade	Specifica il protocollo di comunicazione preferito
Cookie	Comunica un cookie al server
If-Modified-Since	Invia il documento solo se è più recente della data specificata

- Un **body (o contenuto)**, ossia il vero contenuto del messaggio da inviare (solitamente vuoto a meno dell'uso del metodo POST)



Esempio:

```
GET /index.html HTTP/1.1\r\n
Host: www-net.cs.umass.edu\r\n
User-Agent: Firefox/3.6.10\r\n
Accept: text/html,application/xhtml+xml\r\n
Accept-Language: en-us,en;q=0.5\r\n
Accept-Encoding: gzip,deflate\r\n
Accept-Charset: ISO-8859-1,utf-8;q=0.7\r\n
Keep-Alive: 115\r\n
Connection: keep-alive\r\n
\r\n
```

Analogamente, ogni messaggio di **risposta HTTP** viene strutturato in modo simile, ma con alcune differenze:

- Una **riga di stato**, composta dalla versione di HTTP utilizzata, un **codice di status** e una **frase di status** descrivente in breve il codice di status

I **codici di status** si dividono in 5 categorie:

- **Codici 1xx**, indicanti che la risposta ricevuta contiene solamente informazioni  
(es: 100 Continue indica che il server è pronto a ricevere la richiesta del client)
- **Codici 2xx**, indicanti che la richiesta effettuata è andata a buon fine  
(es: 200 OK indica che la richiesta ha avuto successo e l'oggetto richiesto è stato trovato, 204 No Content indica che la richiesta ha avuto successo ma l'oggetto richiesto non contiene nulla al suo interno)
- **Codici 3xx**, indicanti che è stato effettuato un reindirizzamento a seguito della richiesta effettuata  
(es: 301 Moved Permanently indica che l'oggetto richiesto possiede un path diverso da quello richiesto, reindirizzando automaticamente tutte le richieste successive del client)

- **Codici 4xx**, indicanti un errore nella richiesta del client  
(es: **403 Forbidden** indica che il client non possiede i requisiti per accedere all'oggetto richiesto, **404 Not Found** indica che l'oggetto richiesto non esiste )
- **Codici 5xx**, indicanti un errore per cui il server non è riuscito a completare la richiesta  
(es: **500 Internal Server Error** indica un errore sconosciuto all'interno del server, **503 Service Unavailable** indica che il server è attualmente non disponibile)
- Un **header (o intestazione)**, composto da varie linee contenenti informazioni utili alla risposta

Alcuni esempi di campi inseribili all'interno di un header di risposta sono:

Campo Header	Descrizione
Date	Data e ora attuale
Upgrade	Specifica il protocollo di comunicazione preferito
Server	Indica il programma server utilizzato
Set-Cookie	Il server richiede al client di memorizzare un cookie
Content-Encoding	Specifica lo schema di codifica
Content-Language	Specifica la lingua utilizzata nel documento
Content-Length	Indica la lunghezza del documento
Content-Type	Specifica la tipologia del documento
Location	Chiede al client di inviare la richiesta ad un altro sito
Last-modified	Fornisce data e ora dell'ultima modifica del documento

- Un **body (o contenuto)**, ossia il vero contenuto del messaggio da restituire (in particolare, l'oggetto richiesto)

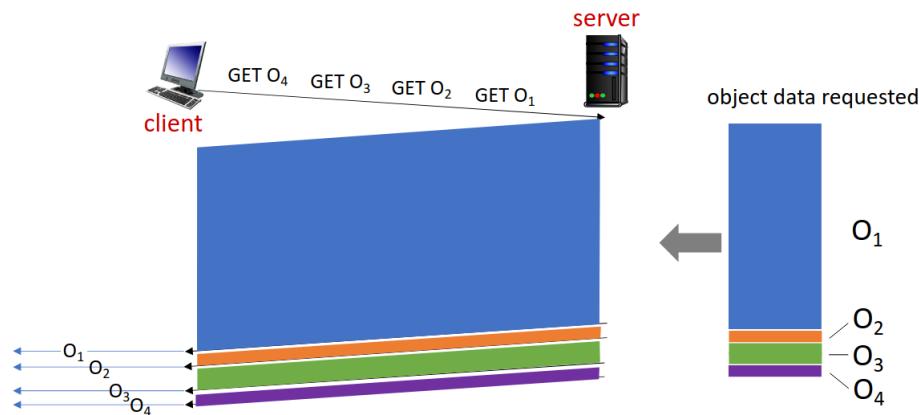
Esempio:

```
HTTP/1.1 200 OK\r\n
Date: Sun, 26 Sep 2010 20:09:20 GMT\r\n
Server: Apache/2.0.52 (CentOS)\r\n
Last-Modified: Tue, 30 Oct 2007 17:00:02 GMT\r\n
Accept-Ranges: bytes\r\n
Content-Length: 2652\r\n
Keep-Alive: timeout=10, max=100\r\n
Connection: Keep-Alive\r\n
Content-Type: text/html; charset=ISO-8859-1\r\n
\r\n
[document content...]
[...]
[document content...]
```

### 2.2.2 Versioni di HTTP

Come già discusso, il **protocollo HTTP/1.1** ha introdotto la possibilità di poter effettuare più richieste GET in successione tramite una singola connessione TCP. Tuttavia, tale modifica ha introdotto ulteriori problematiche:

- Il server risponde alle richieste GET nell'ordine in cui vengono effettuate (**First Come First Served (FCFS)**)
- Un oggetto di piccole dimensioni potrebbe dover attendere la trasmissione di oggetti di grandi dimensioni richiesti prima di esso (**blocco head-of-line (HOL)**)
- La perdita di un segmento TCP causa lo stallo del trasferimento di un oggetto



Per risolvere tali problematiche, il **protocollo HTTP/2** introduce una maggiore flessibilità al server nell'invio di oggetti al client:

- L'ordine di trasmissione degli oggetti richiesti viene stabilito in base alla priorità dell'oggetto specificata dal client
- Gli oggetti vengono **divisi in frame**, schedulati in modo da mitigare il blocco HOL
- Possono essere inviati più oggetti contemporaneamente (**multiplexing**)



Il **protocollo HTTP/3**, invece, risolve le ultime problematiche rimanenti all'interno del protocollo HTTP/2, tramite l'aggiunta di controlli sulla sicurezza, sugli errori e sulla congestione per oggetto, utilizzando il **protocollo QUIC** (basato su UDP) al posto del protocollo TCP.

### 2.2.3 Cookies e Web Caching

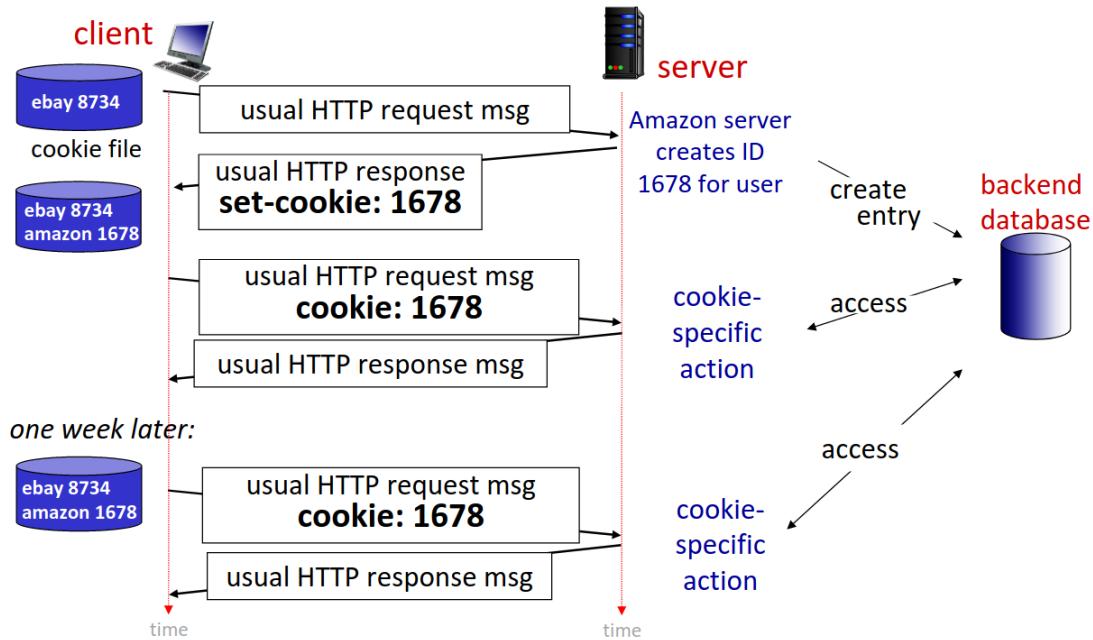
#### Definition 21. Cookie

Un **cookie** è un **piccolo file di testo** contenente brevi informazioni (preferenze sull'utilizzo, parametri preferiti, token di autorizzazione, ...) salvato all'interno di un client da parte di un server web

Poiché il protocollo HTTP è un protocollo **stateless**, i cookie vengono utilizzati all'interno delle applicazioni web per conservare indirettamente alcune informazioni sulle varie comunicazioni client-server effettuate, rendendo ogni richiesta HTTP indipendente dall'altra.

A seguito di un messaggio di risposta da un web server contenente il campo header **Set-Cookie**, il client salva il contenuto del cookie all'interno di un file. Durante le **succesive richieste** effettuate dallo stesso client allo stesso server, tutti i cookie impostati da tale server vengono **allegati ad ogni richiesta HTTP**.

Soltanente, il cookie fornito dal server contiene un ID univoco, in modo da legare una voce nel suo database interno a quel client specifico.



La **durata di un cookie** inviato viene specificata tramite un campo header **Max-Age**, tramite il quale viene specificato il tempo di vita di tale cookie in **secondi**. Allo scadere di tali secondi, il client eliminerà automaticamente tale cookie. Inoltre, non c'è limite alla quantità di secondi specificabili, implicando che sia possibile specificare anche una quantità di secondi pari a mesi o anni

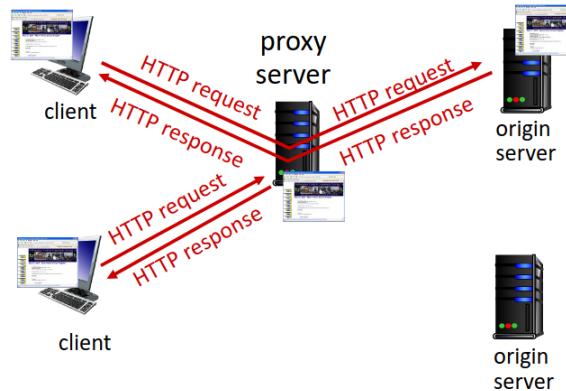
### Definition 22. Proxy Server

Un **proxy server** è un server utilizzato come **intermediario** tra un client e il vero server destinatario.

Solitamente, tale tipologia server viene utilizzato per il **web caching**:

- Se il documento richiesto **è presente** nella cache del proxy server, esso viene restituito al client senza dover raggiungere il server originale
- Se il documento richiesto **non è presente** nella cache, il proxy server inoltra la richiesta del client al server di origine, memorizzando nella sua cache il documento ricevuto nella risposta, restituendolo al client

Tramite il web caching è possibile ridurre notevolmente i tempi di risposta e il traffico nei link di accesso alla rete del server di origine, consentendo ai fornitori di contenuti di essere più efficienti.

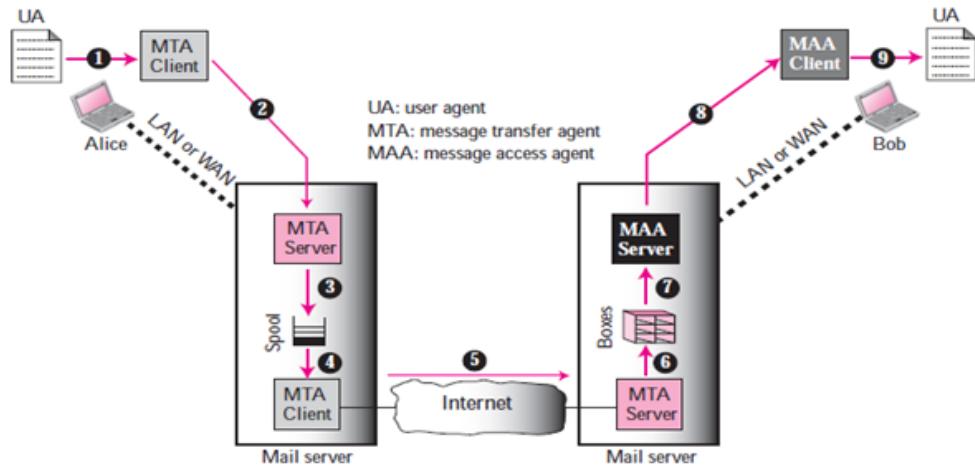


## 2.3 Posta elettronica

Il servizio di **posta elettronica** è costituito da tre entità fondamentali:

- Uno **User agent (UA)**, detto anche *mail reader*, è un processo attivo sul client utente attivato dall'utente stesso o da un timer. Si occupa di informare l'utente nel caso in cui sia disponibile una nuova email da leggere nella sua casella di posta.
- Inoltre, lo user agent permette la composizione, l'editing, l'invio e la lettura di messaggi di posta elettronica. Ogni messaggio di posta inviato da un UA viene passato ad un MTA
- **Mail Transfer Agent (MTA)**, è un processo attivo su un mail server utilizzato per il trasferimento Internet di un messaggio ricevuto da un UA o da un altro MTA
- **Mail Access Agent (MAA)**, è un processo attivo su un mail server utilizzato per leggere i messaggi di posta in arrivo

Ogni **mail server** è dotato di una **casella di posta (mailbox)**, contenente i messaggi in arrivo per l'utente, ed una **coda di messaggi**, contenente i messaggi dell'utente ancora da inviare.



### 2.3.1 Protocolli SMTP e MIME

#### Definition 23. Protocollo SMTP

Il **protocollo SMTP** (Simple Mail Transfer Protocol) è un protocollo a livello di applicazione utilizzato per l'invio di messaggi di posta elettronica in formato ASCII. La sua porta di riferimento comune all'interno dei socket è la **porta 25**.

Il protocollo SMTP effettua un **trasferimento diretto**, ossia dal mail server mittente a quello destinatario (dunque senza mail server intermedi), basato su un'**interazione comando/risposta**: viene inviato un comando in testo ASCII e viene ricevuta una risposta equivalente ad un codice di stato.

Inoltre, il protocollo SMTP fa uso del **protocollo TCP**:

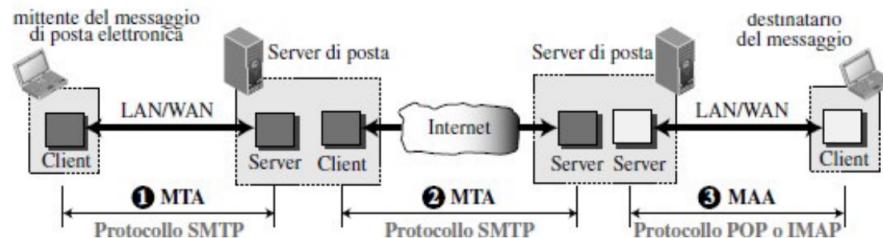
- Il client avvia una connessione TCP con il server utilizzando la porta 25, rimanendo in attesa che il server accetti la connessione (TCP handshaking)
- Vengono scambiati messaggi di posta tra client e server (**connessione persistente**)
- La connessione TCP viene chiusa

#### Esempio:

- Alice usa il suo UA per comporre il messaggio da inviare all'indirizzo di posta elettronica `bob@someschool.edu`
- L'UA di Alice invia il messaggio al mail server di Alice, il quale porrà tale messaggio nella sua coda di messaggi. Successivamente, il client SMTP presente sul mail server di Alice apre una connessione TCP con il mail server di Bob
- Il client SMTP invia il messaggio di Alice sulla connessione TCP tramite il suo MTA

4. Il mail server di Bob riceve il messaggio e lo pone nella casella di posta di Bob
5. Bob invoca il suo UA per leggere il messaggio, il quale preleverà il messaggio tramite l'MAA presente sul suo mail server

(NB: tale operazione *non* è svolta dal protocollo SMTP, bensì dal protocollo POP3 o dal protocollo IMAP che vedremo in seguito)



In particolare, lo scambio di messaggi viene gestito dal protocollo SMTP nel seguente modo:

1. Il client SMTP **tenta di stabilire** una connessione TCP sulla porta 25 con il server STMP. Se il server è attivo, la connessione TCP viene stabilita. Altrimenti, il client riproverà dopo un determinato lasso di tempo.
2. Una volta stabilita la connessione, il client e il server effettuano una **forma aggiuntiva di handshaking**, dove il client indica al server l'indirizzo email del mittente e del destinatario
3. Il client invia il messaggio sulla connessione TCP. Una volta ricevuto il messaggio, se ci sono altri messaggi da inviare viene utilizzata la stessa connessione TCP (**connessione persistente**). Altrimenti, il client invia al server una richiesta di chiusura della connessione.

#### Esempio:

- Di seguito, vediamo un esempio di interazione tra un server SMTP, indicato con S, e un client SMTP, indicato con C.

```

S: 220 hamburger.edu
C: HELO crepes.fr
S: 250 Hello crepes.fr, pleased to meet you
C: MAIL FROM: <alice@crepes.fr>
S: 250 alice@crepes.fr... Sender ok
C: RCPT TO: <bob@hamburger.edu>
S: 250 bob@hamburger.edu ... Recipient ok
C: DATA
S: 354 Enter mail, end with "." on a line by itself
C: Do you like ketchup?
C: How about pickles?
C: .
S: 250 Message accepted for delivery
C: QUIT
S: 221 hamburger.edu closing connection
  
```

Lo **standard RFC 822** definisce la struttura che ogni messaggio di posta elettronica deve assumere:

- Un **header** composto dai seguenti campi:

Campo Header	Descrizione
To	L'indirizzo del destinatario
From	L'indirizzo del mittente
CC	Indirizzi aggiuntivi di mittenti a cui far sapere dell'invio e il contenuto di tale email (abbreviativo di Carbon Copy)
BCC	Analoghi al CC, ma non vengono mostrati al destinatario (abbreviativo di Blind CC)
Subject	L'argomento del messaggio
Sender	Il nome del mittente

- Un **body**, contenente il messaggio da inviare (**solo caratteri ASCII**)

Per poter inviare contenuti diversi dal semplice test ASCII, gli standard RFC 2045 e 2046 definiscono il **protocollo MIME (Multipurpose Internet Mail Extension)**, in grado di estendere i normali messaggi di posta elettronica in messaggi multimediali.

Vengono aggiunte alcune righe all'interno dell'header del messaggio inviato, in particolare una riga **Version**, indicante la **versione del protocollo** MIME utilizzata, e una riga **Type**, indicante il **tipo di dati multimediali** inviati, i quali, prima di essere spediti, vengono convertiti in una **codifica testuale** (solitamente base64), specificata da un campo aggiuntivo **Content-Transfer-Encoding**, in modo da poter essere trasmesso sottoforma di testo ASCII, per poi venir decodificati una volta che il messaggio è giunto al destinatario.

### 2.3.2 Protocolli POP3 e IMAP

#### Definition 24. Protocollo POP3

Il **protocollo POP3 (Post Office Protocol vers. 3)** è un protocollo **stateless** a livello di applicazione utilizzato per il download di messaggi di posta elettronica ricevuti. La sua porta di riferimento comune all'interno dei socket è la **porta 110**.

Per stabilire una connessione, il protocollo POP3 fa uso del **protocollo TCP**, effettuando quindi l'handshake TCP, per poi procedere nelle seguenti tre fasi:

1. **Autorizzazione**, dove lo UA invia nome utente e password per essere identificato dal mail server
2. **Transazione**, dove lo UA recupera i messaggi nella casella di posta dell'utente
3. **Aggiornamento**, dove, successivamente all'invio di un messaggio QUIT da parte dello UA, viene terminata la connessione e vengono rimossi dal mail server i messaggi contrassegnati durante la fase precedente

**Esempio:**

- Se la richiesta effettuata viene eseguita correttamente, il server risponderà con +OK, altrimenti con -ERR. Il comando retr permette di scaricare il messaggio, mentre il comando dele permette di marcare i messaggi da eliminare

```

S: +OK POP3 server ready
C: user rob
S: +OK
C: pass hungry
S: +OK user successfully logged on
C: list
S: 1 498
S: 2 912
S: .
C: retr 1
S: <message 1 contents>
S: .
C: dele 1
C: retr 2
S: <message 1 contents>
S: .
C: dele 2
C: quit
S: +OK POP3 server signing off

```

- Successivamente, viene attivata la fase di aggiornamento, cancellando dal mail server i messaggi marcati tramite dele

Oltre ad essere un protocollo stateless, il protocollo POP3 non fornisce all'utente la possibilità di creare **cartelle remote** tra cui poter suddividere i messaggi, costringendo la creazione di tali cartelle solo a livello locale, implicando che esse non siano condivise tra i vari dispositivi dell'utente.

### Definition 25. Protocollo IMAP

Il **protocollo IMAP (Internet Message Access Protocol)** è un protocollo a livello di applicazione utilizzato per l'accesso ai messaggi di posta elettronica ricevuti. La sua porta di riferimento comune all'interno dei socket è la **porta 143**.

A differenza del protocollo POP3, tutti i messaggi vengono **conservati nel mail server**, permettendo all'utente di avere solo copie locali. Per via di ciò, il protocollo IMAP permette di:

- Associare ogni messaggio ricevuto ad una cartella, detta **inbox**
- Creare cartelle remote e spostare messaggi tra di esse
- Effettuare ricerche nelle cartelle remote
- Conservare lo stato tra le varie sessioni dell'utente (protocollo **stateful**)

## 2.4 Domain Name System (DNS)

### Definition 26. Domain Name System (DNS)

Il **Domain Name System (DNS)** è un sistema utilizzato per mappare singoli nodi di una rete ad un **nome** che li identifichi. Viene realizzato tramite un database distribuito, implementato come una gerarchia di **name server**.

Tra le funzioni fornite dal servizio DNS troviamo:

- **Traduzione** da nome host all'indirizzo IP relativo
- **Distribuzione del carico**, permettendo a più indirizzi IP, ognuno legato ad un server copia di quello originale, di corrispondere ad un unico nome. Quando un client effettua una richiesta, il servizio restituisce l'insieme di indirizzi legati a tale nome in un ordine casuale (rotazione DNS)
- **Host Aliasing**, ossia l'associazione di più sinonimi (alias) allo stesso indirizzo IP, permettendo l'associazione di un nome più semplice rispetto ad uno complesso  
(es: al nome `relay1.west-coast.enterprise.com` associamo l'alias `enterprise.com` e l'alias `www.enterprise.com`)

Per via delle sue funzioni, il servizio DNS risulta essere **fondamentale** per Internet.

In particolare, la **decentralizzazione** del servizio DNS risulta essere critica: se il servizio fosse centralizzato (ossia effettuato da un singolo nodo o rete) sarebbe sufficiente un singolo punto di fallimento affinché il servizio diventi inutilizzabili. Inoltre, se il servizio fosse centralizzato si avrebbe un volume di traffico troppo elevato dovuto alle miliardi di richieste effettuate giornalmente (es: il server DNS Comcast riceve 600 miliardi di richieste al giorno)

### 2.4.1 Gerarchia server DNS

Poiché il mapping DNS è **distribuito** su svariati server, dove in particolare nessuno di essi mantiene il mapping di tutti gli IP possibili (un IP corrisponde a 32 bit, dunque  $2^{32}$  IP possibili), il database tramite cui viene realizzato il servizio DNS è **gerarchico**, seguendo la struttura di un albero:

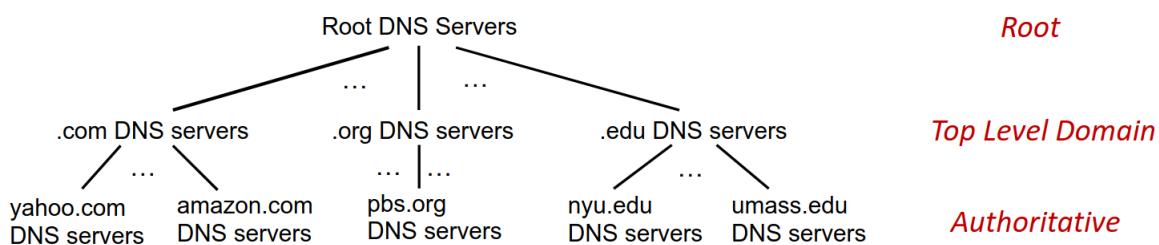
- **Root Server**:
  - Radice dell'albero
  - Viene interrogato da qualsiasi server DNS che non sia in grado di risolvere il nome di un server TLD (ossia restituire l'IP legato ad esso)
- **Server Top-Level Domain (TLD)**:
  - Viene interrogato per risolvere il nome di un server DNS autoritativo
  - Responsabili di domini come `.com`, `.org`, `.net`, ... e tutti i domini nazionali di primo livello, ossia `.it`, `.uk`, `.fr`, ...

- **Server autoritativi (o di competenza):**

- Viene interrogato per risolvere il nome di un host pubblicamente accessibile, solitamente all'interno di un'organizzazione
- Ogni organizzazione con host pubblicamente accessibili deve fornire i record DNS di pubblico dominio che mappano i nomi di tali host ai loro indirizzi IP

- **Server DNS locali (o default name server):**

- Non appartengono alla gerarchia. Ogni ISP ne è dotato
- Possiedono una cache locale delle recenti coppie di mappatura nome-indirizzo (potrebbero non essere aggiornate)
- Funge da proxy iniziale tra il client e il root server: se il nome non è nella cache del server locale, la richiesta viene inoltrata al root server



### Esempio:

1. Il client vuole ottenere l'indirizzo IP dell'host `www.amazon.com`
2. Viene contattato il server DNS locale dell'ISP di riferimento. Se il nome non viene risolto, si procede col passo successivo.
3. Viene contattato il root server per trovare l'indirizzo IP del server TLD `.com`
4. Viene contattato il server TLD `.com` per trovare l'indirizzo IP del server autoritativo `amazon.com`
5. Viene contattato il server autoritativo `amazon.com` per trovare l'indirizzo IP dell'host `www.amazon.com`

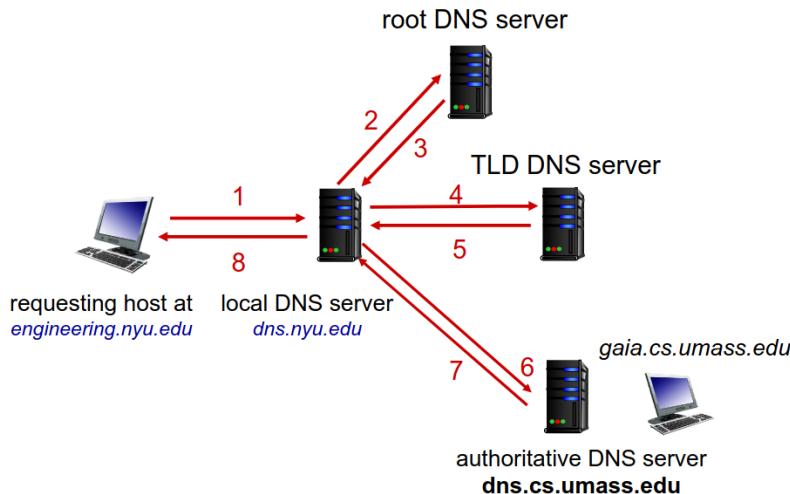
Ogni volta che un server DNS viene a conoscenza di una mappatura, essa viene **memorizzata** all'interno della cache, utilizzando tali record per rispondere a query future. I record presenti nella cache vengono cancellati allo scadere di un **TTL (Time-to-live)** o a seguito di un comando manuale.

Soltanente, all'interno della cache dei server DNS locali sono presenti i server TLD più comuni, implicando che il root server venga interrogato raramente.

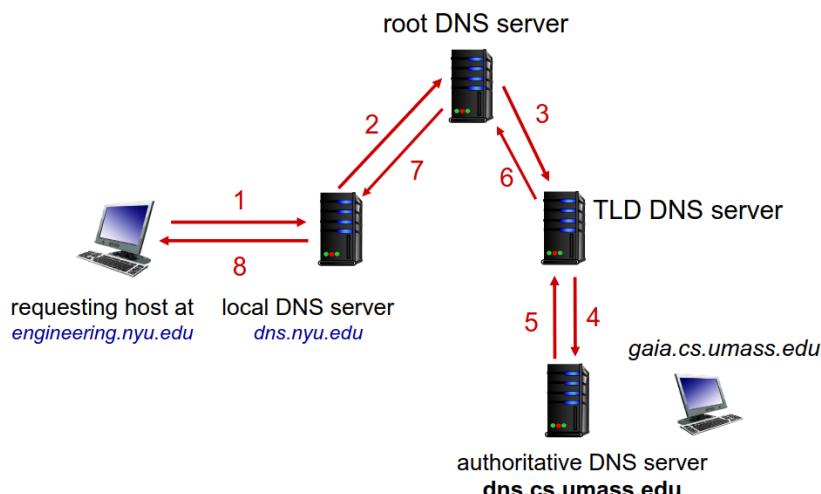
Tuttavia, è necessario notare che i record nella cache **potrebbero non essere aggiornati**: se viene cambiato l'indirizzo IP associato ad un nome presente nella cache, esso potrebbe non essere noto all'interno di Internet fino alla scadenza di tutti i TTL di tutti i server, poiché quest'ultimi risolverebbero la richiesta restituendo l'indirizzo IP precedente.

La **risoluzione dei nomi**, ossia la restituzione dell'indirizzo IP ad esso legato, può avvenire in due modalità:

- **Risoluzione a query iterativa**, dove il server contattato dal client risponde con il nome del prossimo server da contattare, il quale (probabilmente) sarà in grado di risolvere il nome



- **Risoluzione a query ricorsiva**, dove l'onere della risoluzione del nome viene affidato al server contattato, ricorsivamente



Ogni mappatura nome-indirizzo viene inserita all'interno di un **resource record (RR)**, il quale assume la struttura (name, value, type, ttl), dove a seconda del valore del campo **type** si ha che:

- **type = A**, indica che il campo **name** contiene il nome di un host interno ad un dominio (hostname) e il campo **value** contiene l'indirizzo IP di tale host  
(es: name = relay1.bar.foo.com, value = 45.37.93.126)
- **type = NS**, indica che il campo **name** contiene il nome di un dominio e il campo **value** contiene l'hostname del server autoritativo associato a tale dominio  
(es: name = foo.com, value = dns.foo.com)

- **type = CNAME**, indica che il campo **name** contiene un alias del nome canonico e il campo **value** contiene il nome canonico stesso

(es: name = www.ibm.com, value = servereast.backup2.ibm.com)

- **type = MX**, indica che il campo **name** contiene il nome di un mail server interno ad un dominio e il campo **value** contiene l'hostname di tale mail server

#### Esempio:

- Un server autoritativo per un hostname contiene un record di tipo A per l'hostname stesso, ad esempio

(corsi.di.uniroma1.it, 131.111.45.68, A)

- Un server non autoritativo per un dato hostname contiene un record di tipo NS per il dominio che include l'hostname e un record di tipo A che fornisce l'indirizzo IP del server DNS nel campo **value** del record NS.

Ad esempio, un server TLD .it che non è autoritativo per l'host corsi.di.uniroma1.it, contiene i due record

(uniroma1.it, dns.uniroma1.it, NS)

(dns.uniroma1.it, 128.119.40.111, A)

## 2.4.2 Protocollo DNS

### Definition 27. Protocollo DNS

Il **protocollo DNS** è un protocollo a livello di applicazione utilizzato per la risoluzione di hostname e nomi di dominio. La sua porta di riferimento comune all'interno dei socket è la **porta 53**.

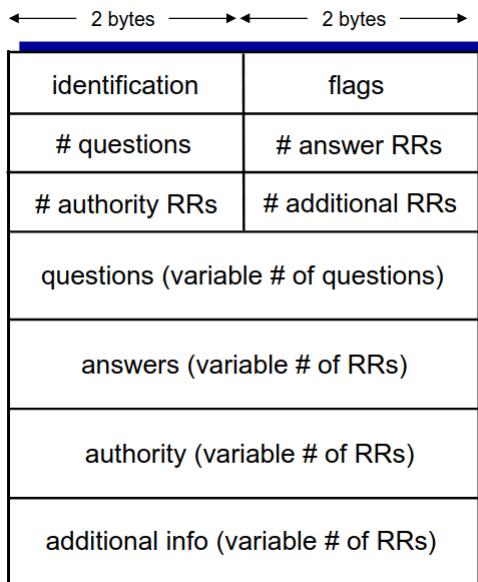
Al fine di rendere il trasferimento il più rapido possibile, il protocollo DNS utilizza il **protocollo UDP**, richiedendo l'invio di un singolo messaggio, evitando necessità della creazione del collegamento. Se un messaggio non giunge a destinazione dopo un determinato timeout, esso viene semplicemente rinvia.

Inoltre, il protocollo DNS è un protocollo **stateless**, (semplicemente poiché non è necessario salvare alcuno stato)

La **richieste** e le **risposte** DNS assumono la stessa struttura:

- Un **header** lungo 32 bit, composto da due campi da 16 bit:
  - **Identification**, contenente informazioni del richiedente
  - **Flags**, contenente flag di stato indicanti se il messaggio sia di richiesta o risposta, se la risoluzione ricorsiva sia preferita o disponibile e se la risposta sia di un server autoritativo

- Un campo **questions** di dimensione variabile contenente le informazioni per una richiesta
- Un campo **answers** di dimensione variabile contenente i RR da inviare come risposta
- Un campo **authority** di dimensione variabile contenente i RR autoritativi da inviare come risposta
- Un campo per le **informazioni aggiuntive**



## 2.5 Trasferimento di file

### 2.5.1 Protocollo FTP

#### Definition 28. Protocollo FTP

Il **protocollo FTP (File Transfer Protocol)** è un protocollo a livello di applicazione utilizzato per il trasferimento di file basato sul **paradigma client-server**.

Per gestire il trasferimento dei file, il protocollo FTP utilizza **due connessioni TCP**:

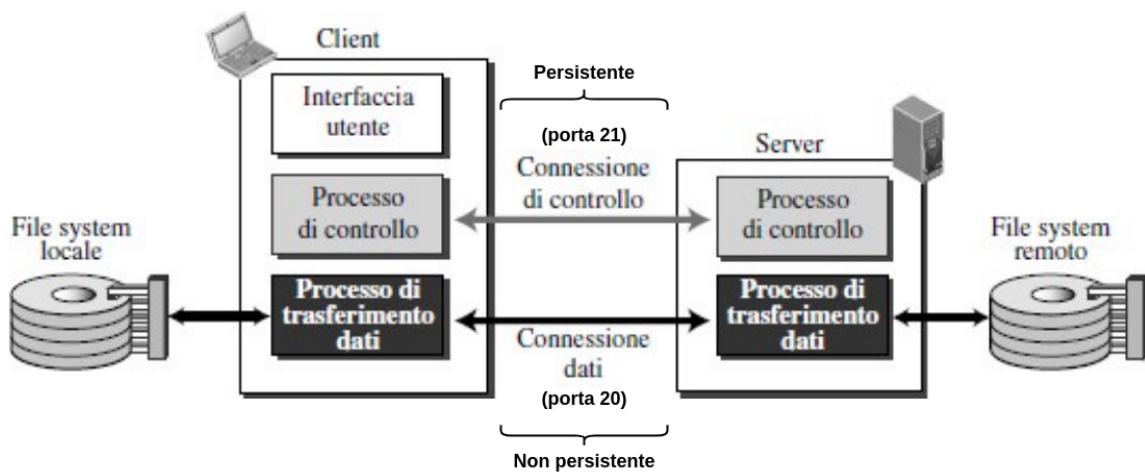
- Una **connessione di controllo** (porta 21), utilizzata per trasferire le informazioni per il controllo del trasferimento (es: nome utente, password, comandi per cambiare directory e per il trasferimento)
- Una **connessione dati** (porta 20), la quale viene aperta ogni qualvolta sia necessario trasferire un file, per poi chiuderla successivamente

Inoltre, il protocollo FTP è un protocollo **stateful**, conservando la directory corrente e l'autenticazione della sessione precedente

Nel protocollo FTP, il **client** corrisponde al dispositivo avviante il trasferimento verso un dispositivo remoto, mentre il **server** corrisponde al dispositivo remoto stesso.

Quando l'utente fornisce al proprio client il nome del server a cui connettersi tramite il comando `ftp <nome host>`, il processo client FTP stabilisce la **connessione di controllo** sulla porta 21.

Successivamente, il client trasferisce nome utente e password sulla porta 21, autenticandosi. Una volta ottenuta l'autorizzazione dal server, il client può **trasferire uno o più file** memorizzati nel file system locale verso quello remoto (o viceversa), aprendo e chiudendo la connessione dati sulla porta 20 ad ogni trasferimento.



I principali comandi del protocollo FTP sono:

Comando e Argomenti	Descrizione
ABOR	Interrompe il comando precedente
CDUP	Torna alla directory del livello precedente
CWD «nome directory»	Cambia directory corrente
DELE «nome file»	Elimina il file
LIST «nome directory»	Elenca i file nella directory
MDK «nome directory»	Crea una directory
PASS «password»	Invia la password dell'utente
PASV	Il server sceglie la porta della connessione
PORT «porta»	Il client sceglie la porta della connessione
PWD	Mostra nome directory corrente
QUIT	Termina la comunicazione
RETR «nomi dei file»	Trasferisce uno o più file dal server al client
RMD «nome directory»	Elimina la directory
RNTO «vecchio nome» «nuovo nome»	Rinomina il file specificato dal vecchio nome
STOR «nomi dei file»	Trasferisce uno o più file dal client al server
USER «nome utente»	Invia il nome dell'utente

## 2.5.2 Protocollo BitTorrent

### Definition 29. Protocollo BitTorrent

Il **protocollo BitTorrent** è un protocollo a livello di applicazione utilizzato per il trasferimento di file basato sul **paradigma peer-to-peer (P2P)**. Nonostante non abbia una porta standard, solitamente vengono utilizzate le **porte nel range 6881-6889** assieme al **protocollo TCP**.

Ogni peer entra a far parte di un **torrent**, ossia un gruppo di peer scambianti frammenti di file tra loro, registrandosi su un **tracker**, ossia un dispositivo che tiene traccia dei peer partecipanti al torrent, per poi connettersi ad un sottoinsieme di peer "vicini".

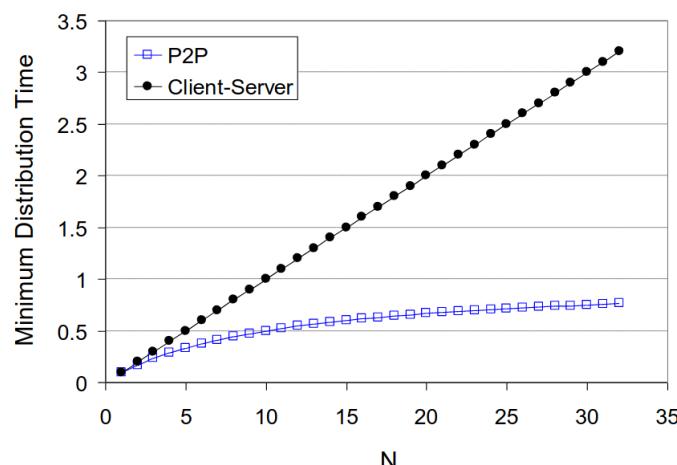
Durante il download di file, il peer svolge anche la funzione di uploader (**seeder**) di blocchi verso altri peer. Una volta ricevuto il file, il peer può scegliere se uscire dal torrent o rimanerne all'interno, continuando a svolgere la funzione di seeder.

In un dato momento, peer diversi possiedono diversi sottoinsiemi di blocchi componenti un file. Per richiedere tali blocchi, un peer chiede periodicamente agli altri l'**elenco dei blocchi** attualmente posseduti. Successivamente, il peer richiede i blocchi mancanti, dando precedenza ai più rari.

Per favorire l'altruismo tra i peer e sfavorire la presenza di **leecher**, ossia dispositivi che egoisticamente escono dal torrent una volta scaricato un file, il protocollo BitTorrent usa un approccio **tit-for-tat** (traduzione più vicina: *do ut des*, "io ti do e tu mi dai" ):

- Ogni peer seeder invia blocchi agli ulteriori quattro peer seeder che attualmente stanno uploadando i blocchi richiesti alla velocità maggiore
- Gli altri peer non appartenenti alla top 4 vengono "strozzati" (**choked**) dal peer seeder, bloccando l'invio dei blocchi ad essi.
- Ogni 10 secondi, tale top 4 viene rivalutata. Inoltre, ogni 30 secondi viene sbloccato casualmente un peer strozzato (**optimistic un-choking**), il quale può entrare o meno a far parte della top 4

Per via di tale approccio, il trasferimento di un file ad  $N$  dispositivi risulta più ottimale nel caso dell'applicazione del paradigma P2P



# Capitolo 3

## Livello di Trasporto

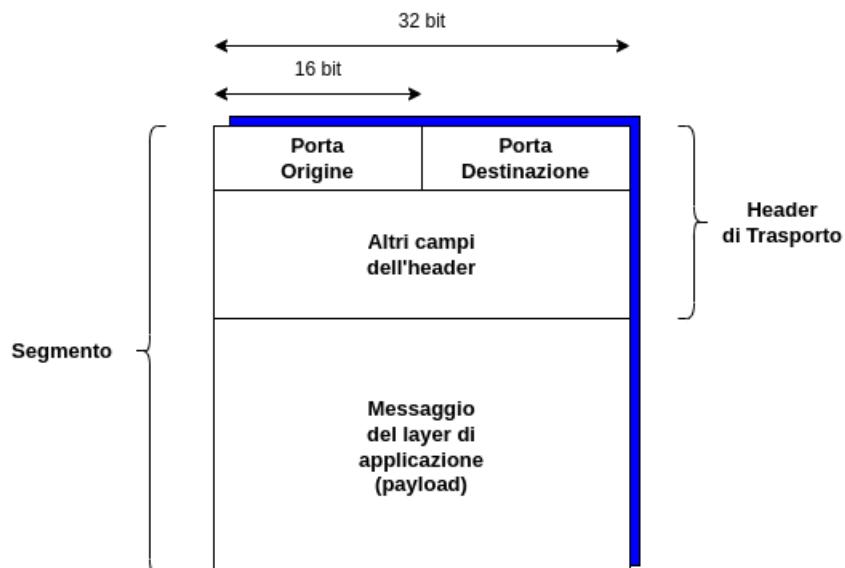
I servizi e protocolli situati nel **livello di trasporto** forniscono **comunicazione logica** tra processi applicativi in esecuzione su dispositivi diversi, a differenza del **livello di rete**, il quale si occupa della comunicazione logica direttamente tra i dispositivi stessi.

In particolare, il dispositivo mittente suddivide i messaggi dell'applicazione in segmenti, passandoli al livello di rete, mentre il dispositivo destinatario riassembra i segmenti in messaggi, passandoli al livello di applicazione.

### 3.1 Multiplexing e Demultiplexing

Per implementare le funzionalità di **multiplexing** e **demultiplexing** al livello di trasporto, ogni host utilizza **indirizzi IP** e **numeri di porta** per indirizzare correttamente un segmento al socket appropriato del destinatario:

- L'header di ogni segmento possiede un numero di porta per l'origine e la destinazione
- Ogni datagramma del livello di rete trasporta un segmento del livello di trasporto
- L'header di ogni datagramma possiede l'indirizzo IP dell'origine e l'indirizzo IP della destinazione

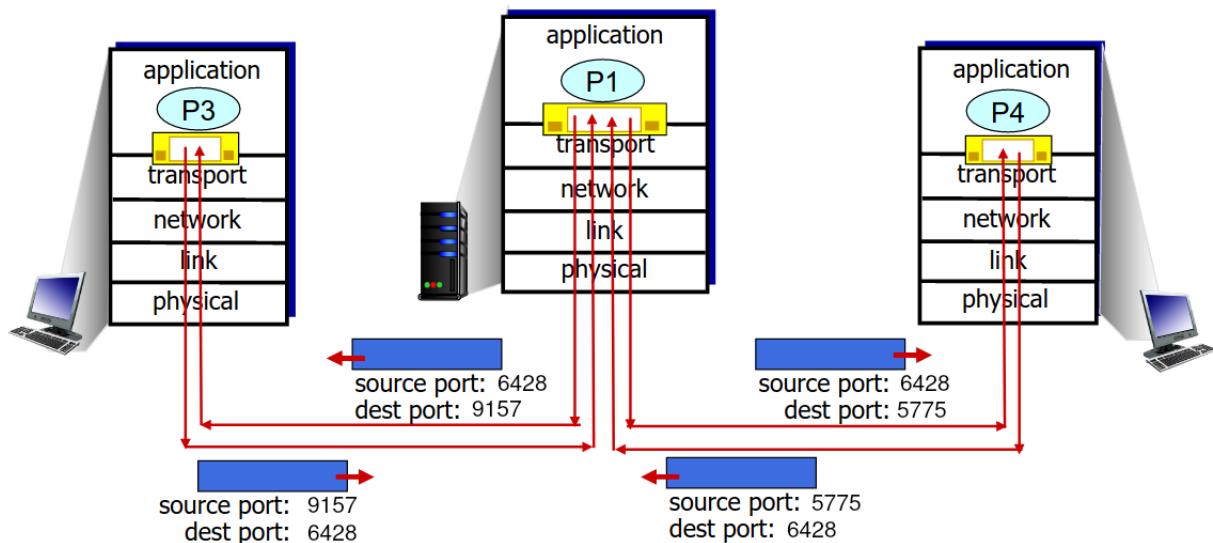


Nel caso di un **demultiplexing senza connessione** (es: protocollo UDP), durante la creazione di un socket all'interno di un processo è necessario specificare la porta locale dell'**host** con cui identificare tale socket.

(es: DatagramSocket d\_soc = new DatagramSocket(12534);)

Successivamente, qualsiasi dispositivo che voglia comunicare con tale host invierà un datagramma al cui interno sia specificata la coppia **indirizzo IP di destinazione** e la **porta di destinazione**. Una volta giunto alla destinazione, verrà letto il numero di porta di destinazione presente nell'header del segmento contenuto all'interno del datagramma ricevuto, indirizzando il segmento al **socket con tale numero di porta**.

In particolare, è necessario sottolineare che, in tal modo, il socket su tale host per la comunicazione con più mittenti sia **unico**. Di conseguenza, qualsiasi datagramma inviato su tale porta apparterrà allo **stesso stream dati**. Tuttavia, essi saranno comunque distinti univocamente dal numero di porta e l'indirizzo IP del mittente presenti nel datagramma.

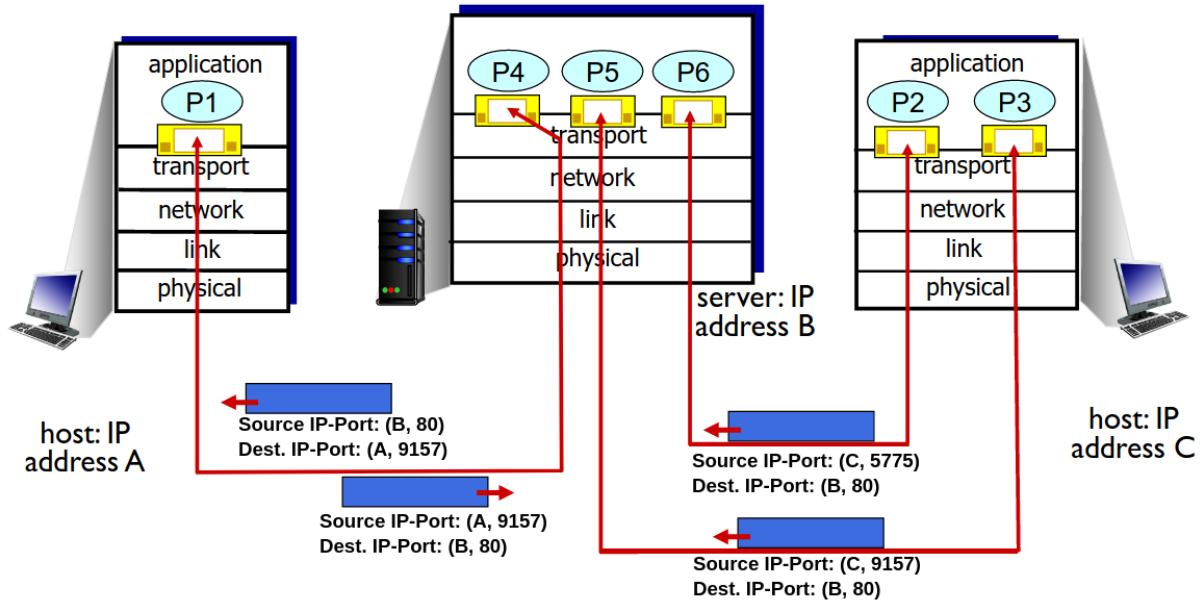


Nel caso del **demultiplexing con connessione** (es: protocollo TCP), invece, **ogni socket viene identificato univocamente come una quadrupla**

(IP\_Orig., Porta\_Orig., IP\_Dest., Porta\_Dest.)

Una volta ricevuto il datagramma, vengono utilizzati tutti e quattro i valori per indirizzare il segmento al socket appropriato. In tal modo, ogni **connessione** è identificata in modo univoco da una **coppia di socket** (una sul primo host ed una sul secondo host), permettendo di implementare le garanzie previste dai protocolli.

Inoltre, per via dell'identificazione univoca dei socket, un host può avere **più socket legati alla stessa porta** con una comunicazione diversa: se due host A e C avviano una connessione avente come destinazione la porta 80 dell'host B, su quest'ultimo verranno creati **due socket diversi**.



## 3.2 Protocollo UDP

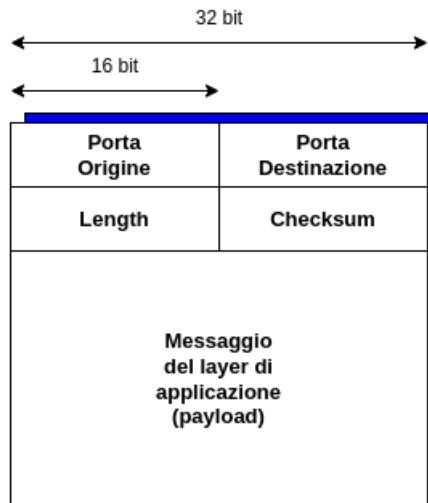
Come già accennato, il **protocollo UDP** è un protocollo di trasporto "senza fronzoli" (**bare bone**) e **senza connessione**. Per tanto, non avviene alcun handshake tra mittente e destinatario, implicando che ogni segmento UDP venga gestito indipendentemente dagli altri. Inoltre, il protocollo UDP svolge un servizio **best-effort**, dunque i segmenti UDP possono essere persi o consegnati in modo non ordinato.

Tuttavia, tali caratteristiche rendono UDP **vantaggioso** in alcune casistiche:

- Poiché non vi è alcuna connessione, il protocollo risulta semplice, oltre all'**assenza del ritardo RTT** necessario per l'handshake richiesto
- La dimensione dell'header è minima, rendendo il **pacchetto più leggero**
- L'**assenza di controllo della congestione** permette al protocollo UDP di tentare la trasmissione senza alcun limite di velocità e il funzionamento anche in casi di congestione dovuti ad un carico elevato sui nodi della rete
- Se si vuole rendere il trasferimento affidabile anche utilizzando UDP, basta implementare l'affidabilità necessaria al livello di applicazione (es: HTTP/3 tramite il protocollo QUIC), piuttosto che al livello di trasporto

In particolare, l'header utilizzato dal protocollo UDP, oltre a contenere le porte di origine e di destinazione, contiene solamente due campi aggiuntivi:

- Un campo **length** (16 bit), indicante la lunghezza del contenuto
- Un campo **checksum** (16 bit), utilizzato per rilevare errori nel segmento trasmesso



Il valore di **checksum** viene calcolato tramite una **somma in complemento ad uno con wrap-around**:

1. Il mittente considera il contenuto del segmento (compresi gli altri campi dell'header e gli indirizzi IP) come una sequenza di numeri interi a 16 bit
2. Il mittente calcola il checksum sommando in **complemento ad 1** (ossia sommando e poi invertendo tutti i bit) i numeri interi della sequenza.

In particolare, se è presente un riporto finale generato dalla somma del bit più significativo, viene sommato anche tale riporto (**wrap-around**)

3. Il valore del checksum viene inserito nel campo dell'header e il pacchetto continua il processo di trasmissione
4. Una volta giunto a destinazione, l'host ricevente **calcola nuovamente il checksum**, verificando che sia uguale a quello inserito nell'header del segmento.

Se il checksum è differente, viene rilevato un **errore** nella trasmissione

Tuttavia, è necessario notare che se il **checksum è uguale non è detto** che la trasmissione non abbia generato alcun errore:

- Consideriamo il calcolo del checksum tra i seguenti numeri interi a 16 bit:

$$\begin{array}{r}
 1 \ 1 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ + \\
 1 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ = \\
 \hline
 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1
 \end{array}$$

- Poiché è presente un riporto, viene effettuato il **wrap-around** sommando tale riporto ai restanti 16 bit:

$$\begin{array}{r}
 1 \ 1 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ + \\
 1 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1 \ = \\
 \hline
 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1
 \end{array}$$

↓      ↓      ↓ \\
 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0

- Infine, vengono **invertiti i bit** del risultato per ottenere la somma in complemento ad 1

$$\begin{array}{cccccccccccccccccc}
 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
 & & & & & & \downarrow & \downarrow & \downarrow & & & & & & & & \\
 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1
 \end{array}$$

- Tuttavia, nel caso in cui i due bit meno significativi di entrambi i numeri fossero stati invertiti (per qualche motivo sconosciuto) durante la trasmissione, il **checksum calcolato sarebbe identico**

$$\begin{array}{cccccccccccccccccc}
 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & + \\
 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & = \\
 \hline
 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1
 \end{array}$$

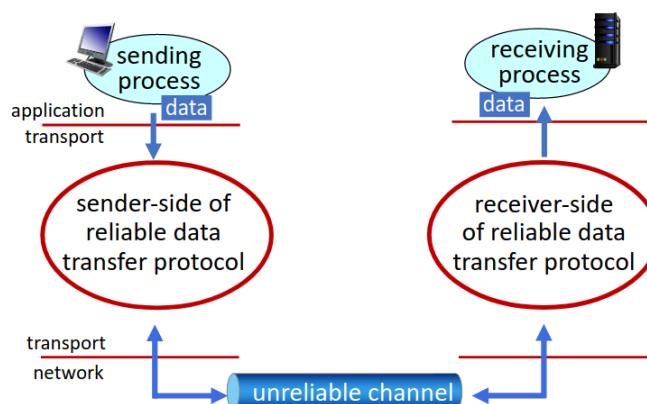
$$\begin{array}{cccccccccccccccccc}
 & & & & & & \downarrow & \downarrow & \downarrow & & & & & & & & \\
 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
 & & & & & & \downarrow & \downarrow & \downarrow & & & & & & & & \\
 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1
 \end{array}$$

### 3.3 Trasferimento affidabile dei dati

Per realizzare un trasferimento affidabile dei dati, è necessario implementare un **canale sicuro** al cui interno non vengano perse o corrotte informazioni.



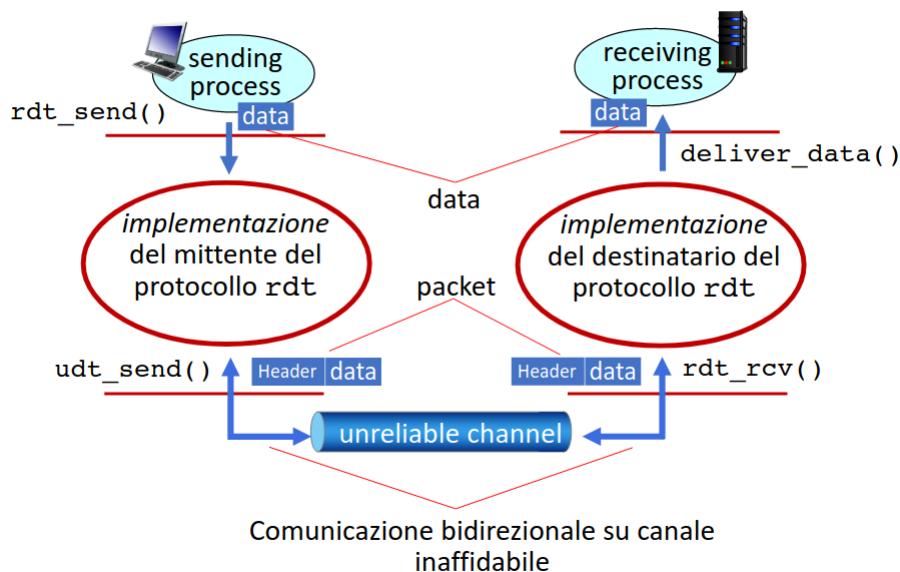
Tuttavia, un canale fisico che possa svolgere tale funzione risulta essere irrealizzabile. Per tale motivo, la complessità del **protocollo di trasferimento dati affidabile (RDT - Reliable Data Transfer)** dipende fortemente dalle caratteristiche del canale inaffidabile utilizzato.



Inoltre, è necessario puntualizzare che il mittente e il destinatario **non conoscono lo stato l'uno dell'altro** (es: se la ricezione sia andata a buon fine), a meno che non gli venga comunicato tramite un ulteriore messaggio.

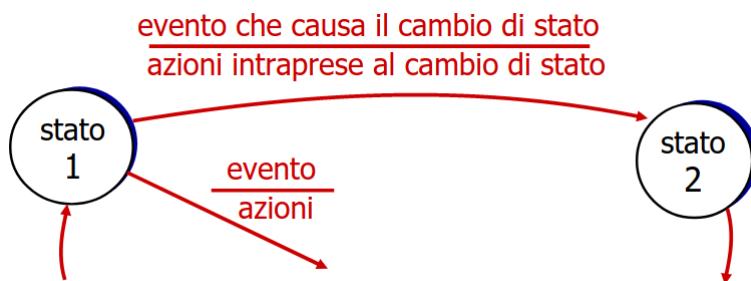
Il protocollo RDT presenta delle **interfacce** per il suo utilizzo:

- **rdt\_sent(data)**, il quale viene chiamato dal livello di applicazione e il cui argomento corrisponde ai dati da inoltrare al destinatario
- **udt\_send()**, dove UDT è acronimo di Unreliable Data Transfer, il quale viene chiamato dal protocollo RDT sul mittente per trasferire il pacchetto sul canale inaffidabile
- **rdt\_rcv()**, il quale viene chiamato alla ricezione del pacchetto dal destinatario
- **deliver\_data()**, il quale viene chiamato dal protocollo RDT sul destinatario per inoltrare al livello di applicazione i dati ricevuti



Poiché i dispositivi comunicanti non sono a conoscenza dello stato altrui, il protocollo RDT si basa su un **trasferimento dei dati unidirezionale**, dunque come se uno solo dei due sia il mittente ed uno solo sia il destinatario (sebbene in realtà sia bidirezionale).

Per rappresentare le operazioni e le decisioni effettuate dal protocollo, utilizzeremo le **macchine a stati finiti** (FSM - Finite State Machine) e in particolare la seguente notazione:



### 3.3.1 Protocollo RDT 1.0 e 2.0

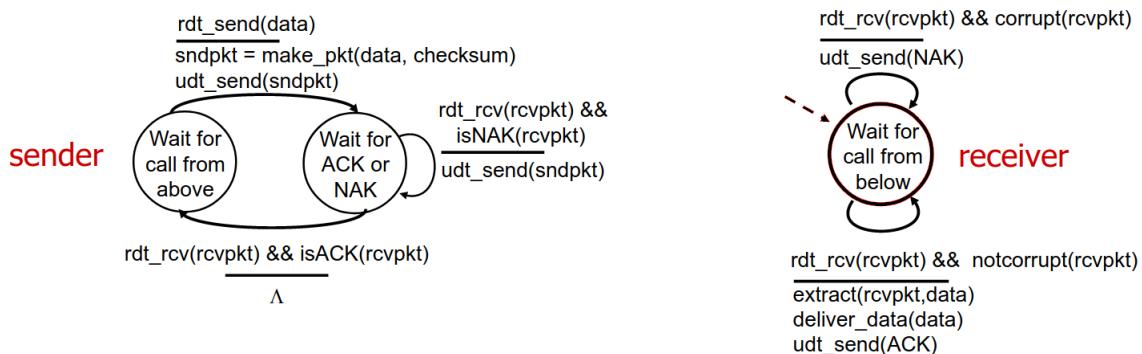
All'interno del **protocollo RDT 1.0**, viene assunto che il canale sottostante utilizzato per il trasferimento sia **perfettamente affidabile**, implicando che il mittente invii i dati nel canale e il ricevitore li legga direttamente, senza alcuna operazione aggiuntiva



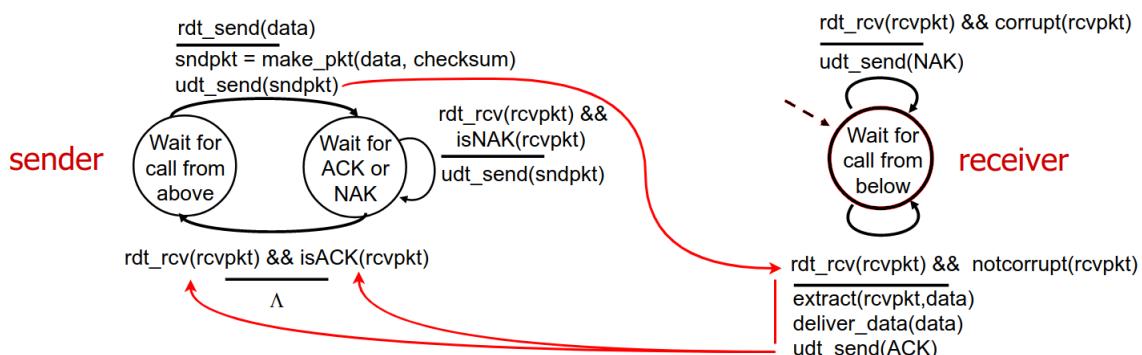
Nel **protocollo RDT 2.0**, invece, viene assunto che il canale sottostante possa invertire alcuni bit nel pacchetto inviato. Analogamente al protocollo UDP, viene utilizzato un **checksum** per rilevare la presenza di errori. Nel caso in cui venga rilevato uno di quest'ultimi, il destinatario comunicherà al mittente l'esito dell'operazione:

- **Acknowledgements (ACK)**, dove il destinatario dice esplicitamente al mittente che il pacchetto è stato ricevuto senza problemi
- **Negative acknowledgements (NAK)**, dove il destinatario dice esplicitamente al mittente che il pacchetto ricevuto presenta degli errori

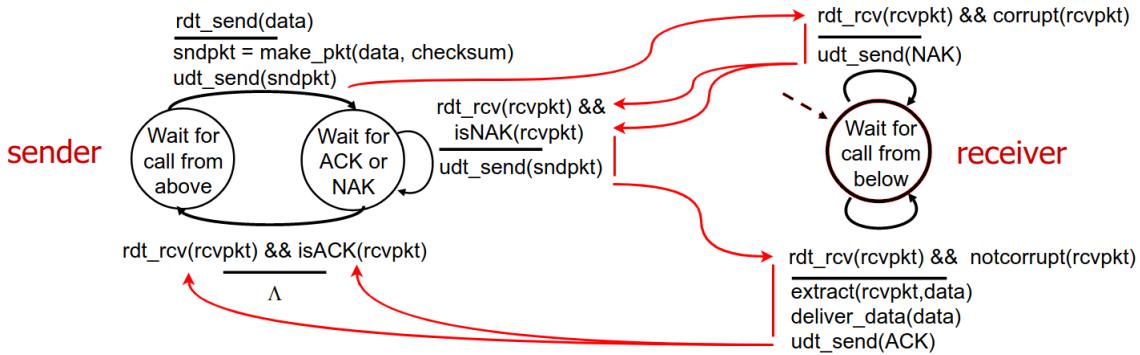
Successivamente all'invio di un pacchetto, il mittente rimane in attesa della risposta del destinatario (meccanismo **stop and wait**)



Se la risposta ricevuta è un **ACK**, il mittente torna il stato di attesa del prossimo pacchetto da parte del livello applicativo.



Se invece la risposta è un **NAK**, il mittente rinvia il pacchetto generante l'errore e rimane in attesa della risposta del destinatario, ripetendo nuovamente tale processo nel caso in cui si riceva nuovamente un NAK.

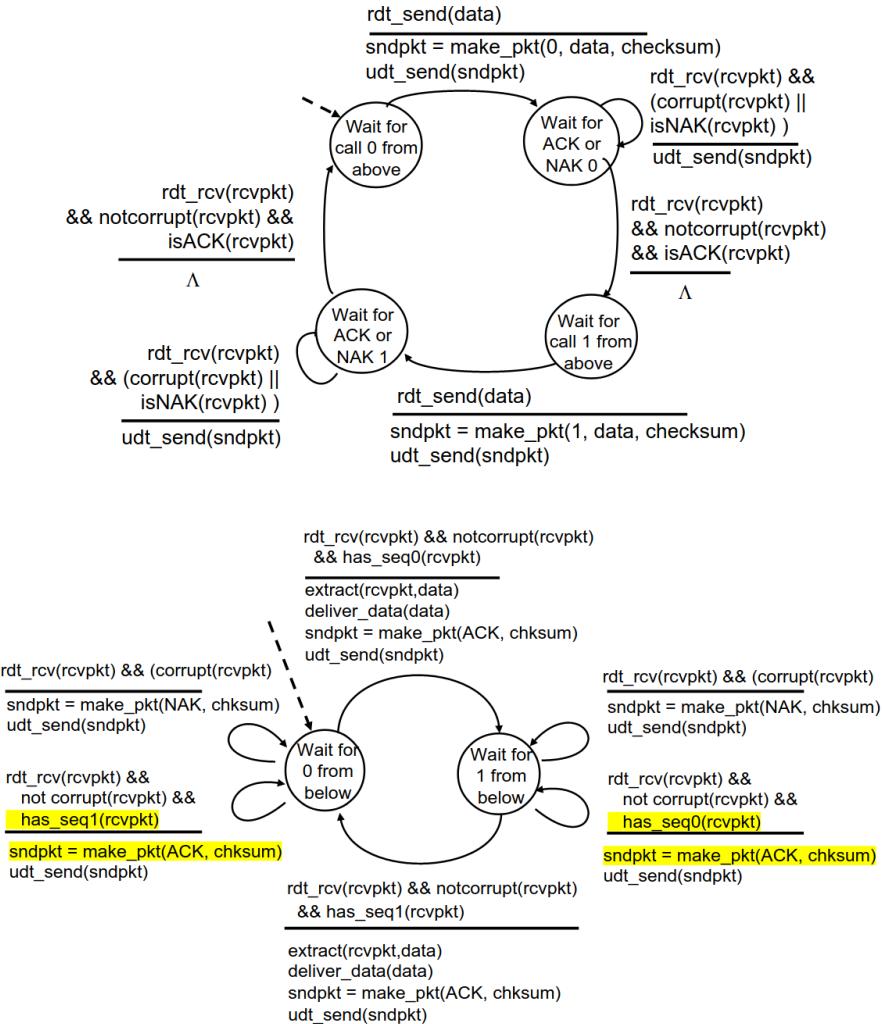


Tuttavia, la versione 2.0 del protocollo RDT presenta un **difetto fatale**: se la risposta ACK/NAK è corrotta, il mittente non è più a conoscenza di cosa sia accaduto al destinatario. Inoltre, non è sufficiente ritrasmettere il pacchetto per risolvere tale difetto, poiché il destinatario potrebbe ricevere due pacchetti duplicati ed inoltrarli al livello di applicazione.

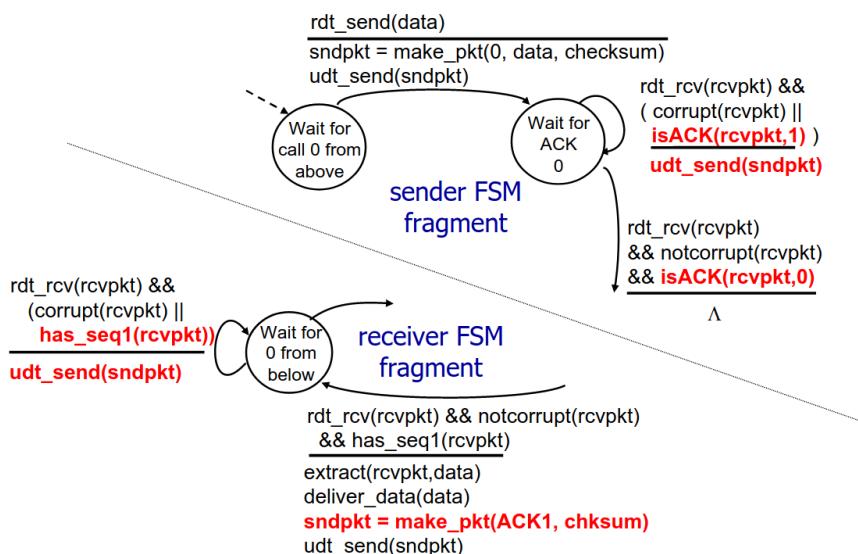
### 3.3.2 Protocollo RDT 2.1 e 2.2

Per risolvere il difetto fatale della versione 2.0, il **protocollo RDT 2.1**:

- Viene controllato se la risposta ACK/NAK sia **corrotta**. Nel caso in cui lo sia, il pacchetto viene rinvia.
- Il destinatario non è a conoscenza della possibile corruzione del pacchetto ACK/-NAK
- Viene aggiunto un **numero di sequenza** al pacchetto inviato. In particolare, sono necessari i numeri di sequenza 0 ed 1 affinché il protocollo stop and wait possa funzionare correttamente:
  - Assieme alla risposta di ACK, il destinatario invia un **numero di riscontro**, il quale, per convenzione, indica sempre il numero di sequenza del prossimo pacchetto atteso dal destinatario
  - Se il destinatario ha ricevuto correttamente il pacchetto 0, invia un riscontro con valore 1 (dunque il prossimo pacchetto atteso è il pacchetto 1)
  - Analogamente, se il destinatario ha ricevuto correttamente il pacchetto 1, invia un riscontro con valore 0 (dunque il prossimo pacchetto atteso è il pacchetto 0)
- Se il pacchetto ricevuto dal destinatario è un duplicato, esso viene automaticamente scartato senza essere inviato al livello di applicazione



In aggiunta alle modifiche della versione 2.1, il protocollo RDT 2.2 **elimina** la necessità di una risposta **NAK**: il ricevitore invia come numero di riscontro il numero di sequenza dell'**ultimo pacchetto correttamente**. In tal modo, un ACK duplicato al mittente comporta la stessa azione di un NAK, ossia la ritrasmissione del pacchetto corrente.

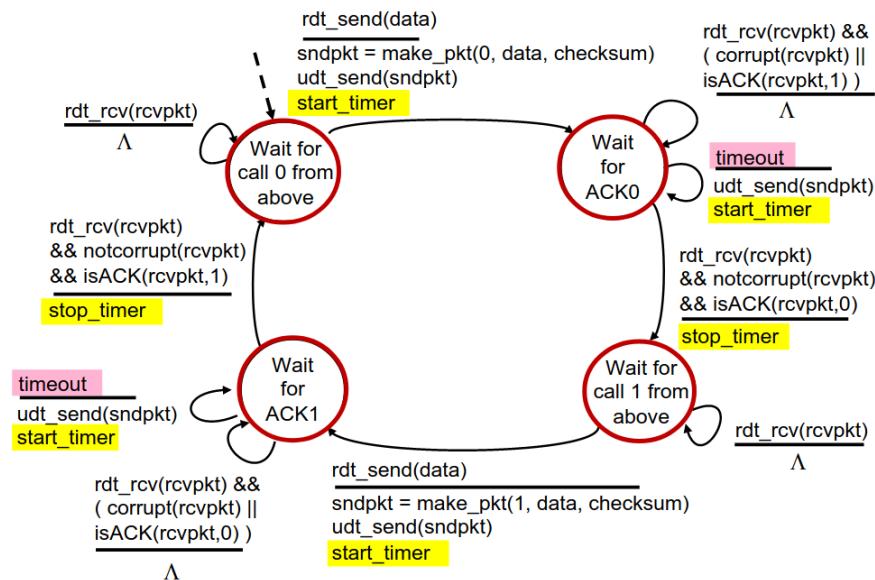


### 3.3.3 Protocollo RDT 3.0

Oltre all'assunzione di possibili bit invertiti, il **protocollo RDT 3.0** assume la possibilità di una **perdita di pacchetti**, sia dati che ACK. Per risolvere tale problematica, il mittente **attende un lasso di tempo**:

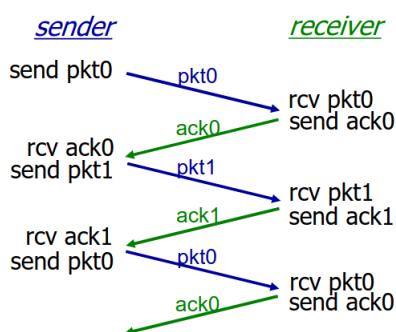
- Il destinatario deve specificare il **numero di sequenza del pacchetto** per il quale sta inviando un ACK
- Se non viene ricevuto alcun ACK allo scadere del lasso di tempo, il pacchetto dati (che indicheremo con pkt) viene ritrasmesso
- Se pkt o ACK arrivano successivamente allo scadere del tempo, il pacchetto verrà ritrasmesso, implicando che la trasmissione verrà duplicata (problema già gestito dai numeri di sequenza)

La FSM associata al mittente corrisponde a:

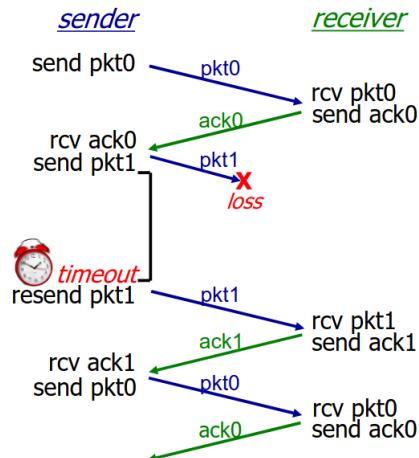


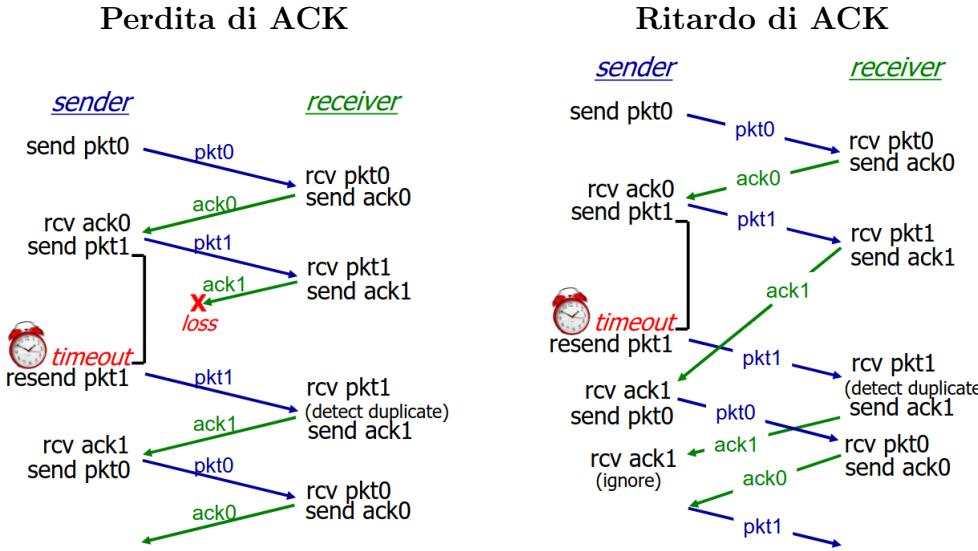
Di seguito, vengono mostrati alcuni esempi di gestione tramite protocollo RDT 3.0:

Nessuna perdita



Perdita di pkt



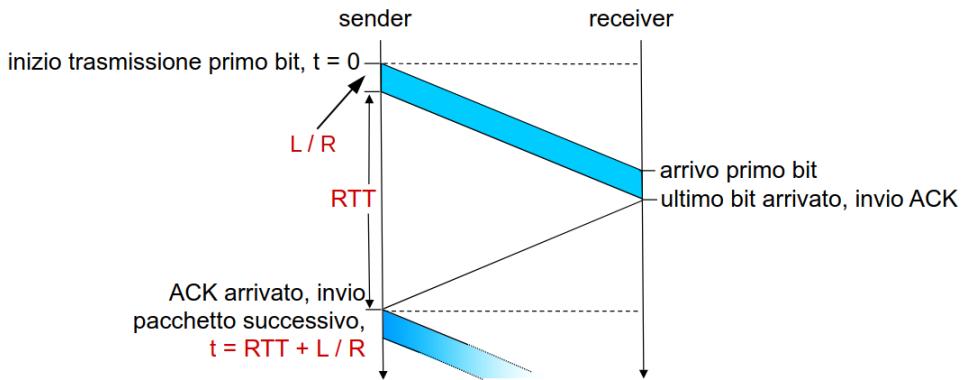


Tuttavia, in cambio dei notevoli benefici del meccanismo stop and wait, le **prestazioni** del protocollo RDT 3.0 risultano essere **infime**, limitando le prestazioni dell'infrastruttura sottostante, ossia il canale.

In particolare, la **percentuale di utilizzo**  $U_{mit}$  della comunicazione da parte del mittente, ossia la frazione di tempo in cui il mittente è impegnato nell'invio corrisponde a:

$$U_{mit} = \frac{D_t}{RTT + D_t}$$

dove  $D_t$  è il delay di trasmissione (dunque  $D_t = \frac{L}{R}$  con  $L$  la lunghezza del pacchetto e  $R$  il transmission rate del link)



### Esempio:

- Considerando un link avente un rate pari a  $R = 1 \text{ Gb/s}$ , una lunghezza di pacchetto pari a  $L = 8000 \text{ b}$  e un ritardo di propagazione sia pari a 15 ms, la percentuale di utilizzo del mittente corrisponde a:

$$D_t = \frac{8 \cdot 10^3 \text{ b}}{10^9 \text{ b/s}} = 8 \mu\text{s}$$

$$U_{mit} = \frac{8 \mu\text{s}}{30 \text{ ms} + 8 \mu\text{s}} = 27 \cdot 10^{-5} = 0.027\%$$

### 3.3.4 Go-back-N e Selective repeat

Per migliorare le prestazioni del protocollo RDT 3.0, viene utilizzato il **pipelining**, dove il mittente consente la presenza di molteplici trasferiti senza aver ricevuto un ACK precedente.

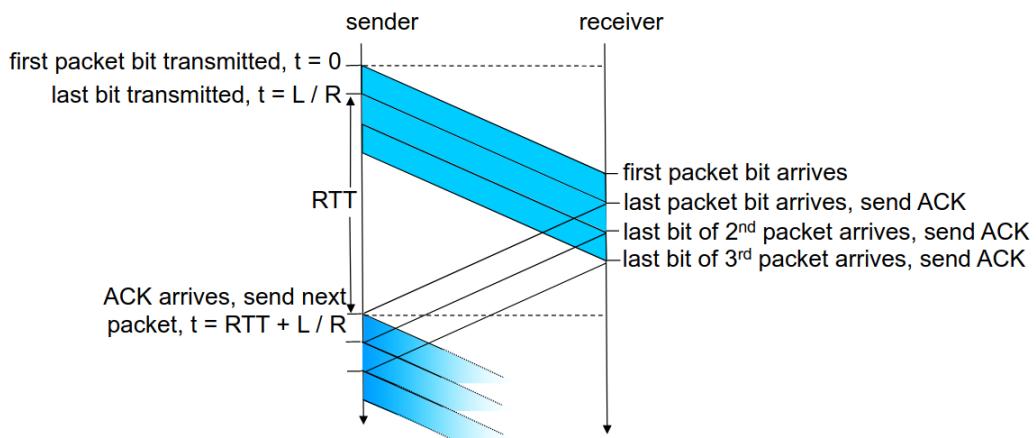
Per realizzare il pipelining, l'**intervallo di numeri di sequenza** deve essere **aumentato**, poiché è necessario tener traccia di più pacchetti simultaneamente, richiedendo inoltre la presenza di un **buffer** interno al mittente e al destinatario.

Poiche i pacchetti successivi al primo vengono inviati durante contemporaneamente al RTT del primo pacchetto, è sufficiente considerare un solo RTT, incrementando notevolmente la percentuale di utilizzo del mittente:

**Esempio:**

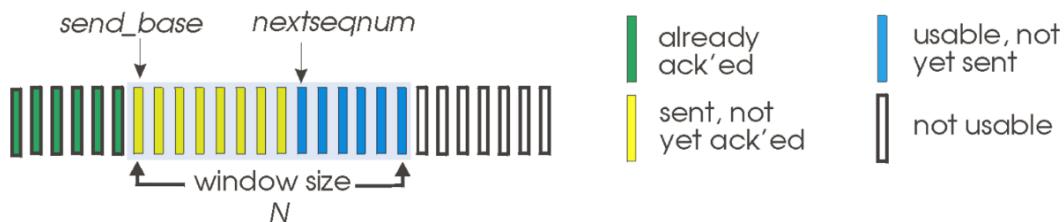
- Riprendendo i dati dell'esempio precedente, effettuando il pipelining con 3 pacchetti si ha che

$$U_{mit} = \frac{3 \cdot D_t}{RTT + D_t} = 3 \cdot \frac{8 \mu s}{30 ms + 8 \mu s} = 81 \cdot 10^{-5} = 0.081\%$$



Una delle metodologie con cui viene implementato il pipelining è il **Go-back-N**:

- Il mittente ha una "finestra" di  $N$  pacchetti consecutivi trasmessi senza ACK (**ACK cumulativo**). La ricezione del pacchetto **ACK(n)** viene interpretato dal mittente come un ACK per ognuno dei singoli  $N$  pacchetti, implicando che alla sua ricezione la finestra venga spostata in avanti in modo che essa abbia il pacchetto  $N + 1$  come primo pacchetto



- Viene mantenuto attivo un **timer** per il pacchetto della finestra inviato e senza ACK **più vecchio**. Una volta scaduto tale timeout, viene ritrasmesso il pacchetto e tutti i pacchetti con numero di sequenza maggiore presenti all'interno della finestra

- Il destinatario invia sempre l'**ACK con numero di sequenza maggiore** (in ordine) per i pacchetti attualmente ricevuti **correttamente**, implicando che non vi siano pacchetti con numero di sequenza minore mancanti.

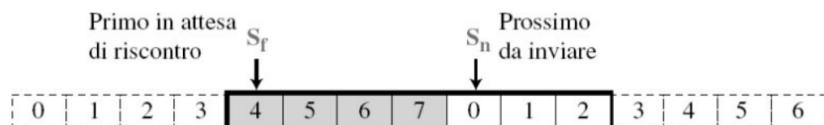
Tale procedura potrebbe generare ACK duplicati e richiede di ricordare solamente un **valore `rcv_base`** (a differenza della finestra del mittente), corrispondente al numero di sequenza del pacchetto di cui si è in attesa

- Se il destinatario riceve un pacchetto fuori ordine, può, a seconda dell'implementazione, scartare tale pacchetto (**politica don't buffer**) o conservarlo (**politica buffer**), inviando in entrambi i casi un ACK con il più alto numero di sequenza che si trovi nell'ordine corretto, richiedendo quindi la trasmissione di tutti i pacchetti con numero di sequenza maggiore.

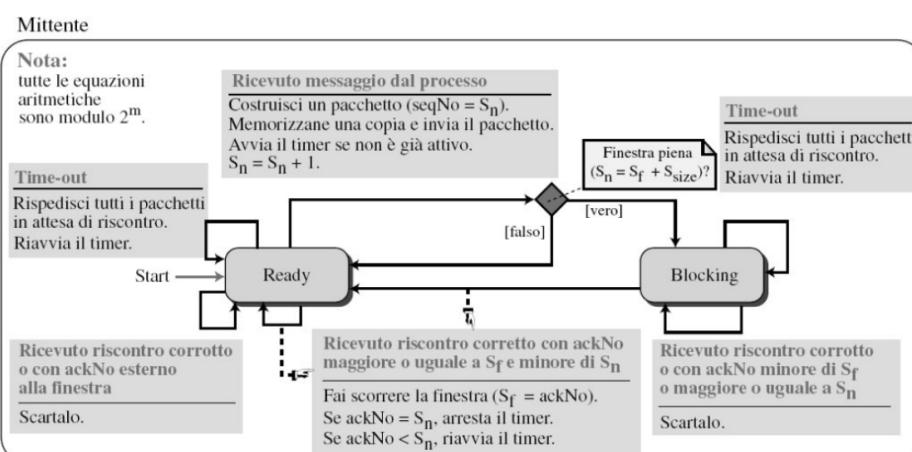
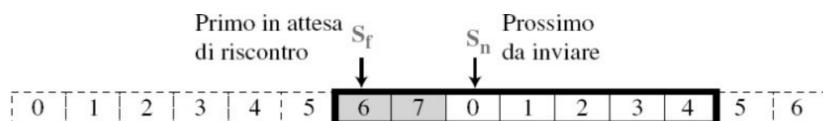


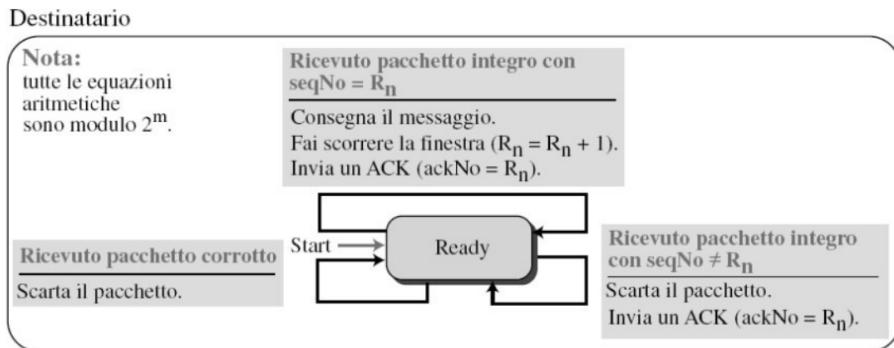
### Esempio:

- Consideriamo la seguente finestra di 7 pacchetti

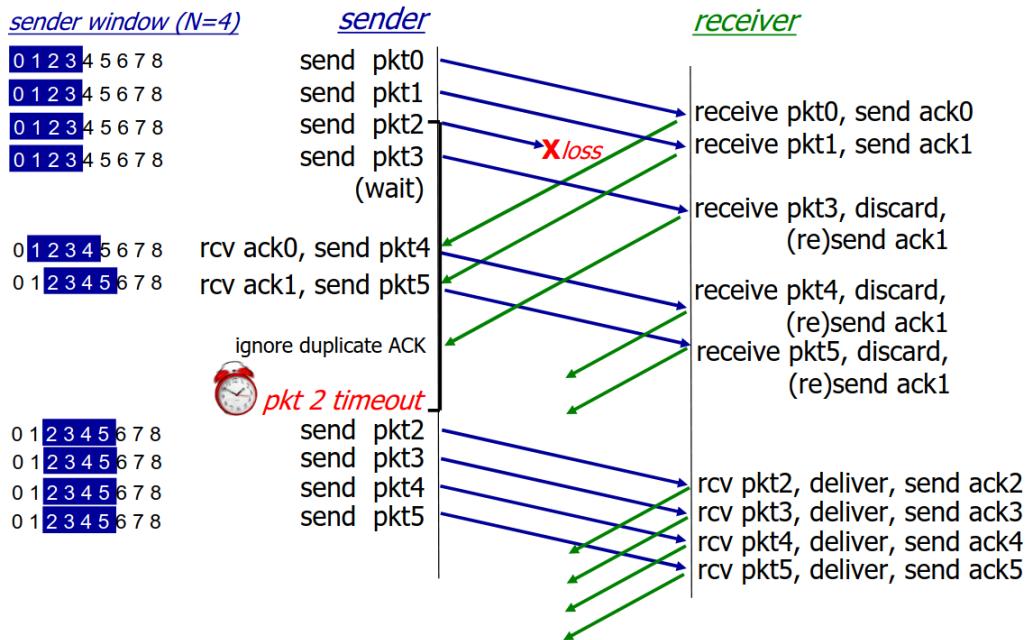


- Una volta ricevuto ACK(5), i pacchetti 4 e 5 vengono considerati come arrivati a destinazione, scorrendo la finestra in avanti





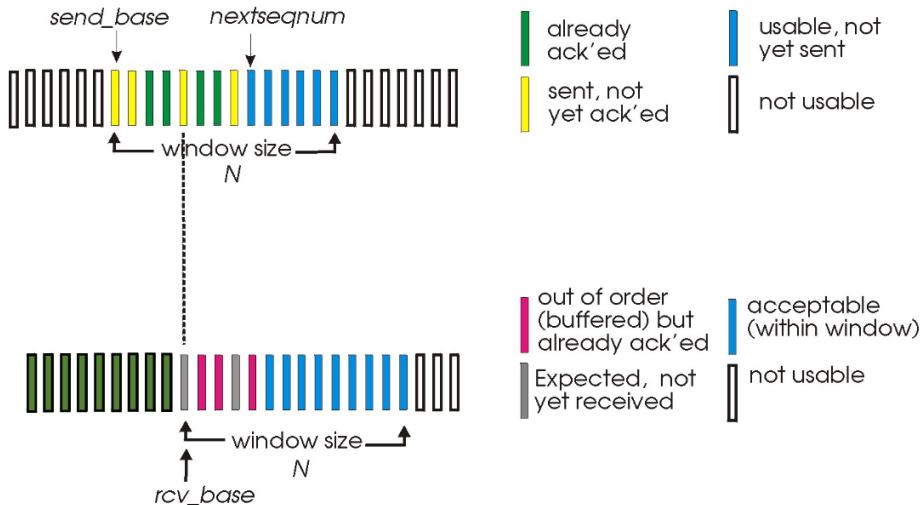
Esempio (protocollo Go-back-N con politica don't buffer):



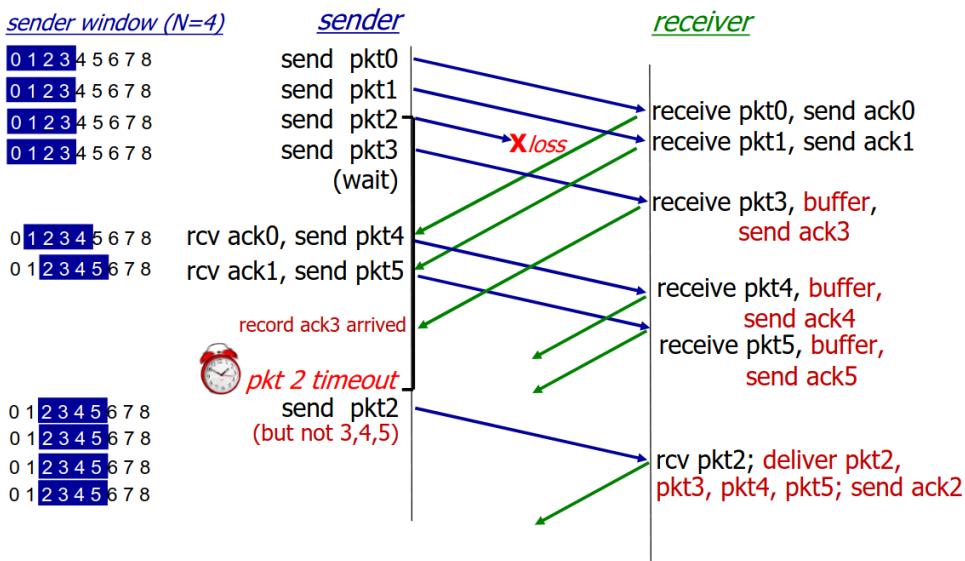
Poiché nel caso in cui un singolo pacchetto venga perso o corrotto è necessario rinviare tutti i pacchetti successivi già inviati nella pipeline, il protocollo Go-back-N può **peggiорare la congestione della rete**.

Contrariamente, il **protocollo Selective Repeat** è in grado di gestire tale problematica:

- Oltre al mittente, anche il destinatario è dotato di una **finestra di  $N$  pacchetti**
- Il destinatario conferma **individualmente** tutti i pacchetti ricevuti correttamente, anche nel caso in cui essi siano fuori sequenza, **bufferizzandoli** per l'eventuale consegna in ordine al livello superiore
- Il mittente mantiene un **timer per ogni pacchetto** inviato senza ACK, rinvia ogni pacchetto individualmente alla scadenza del suo timeout
- La finestra del mittente scorre a partire dal **pacchetto più alto confermato in ordine** (senza pacchetti non confermati prima di esso). Alla ricezione dell'ACK di un pacchetto, dunque, se tale pacchetto era il più piccolo pacchetto non ancora confermato, la finestra avanza fino al prossimo pacchetto non confermato

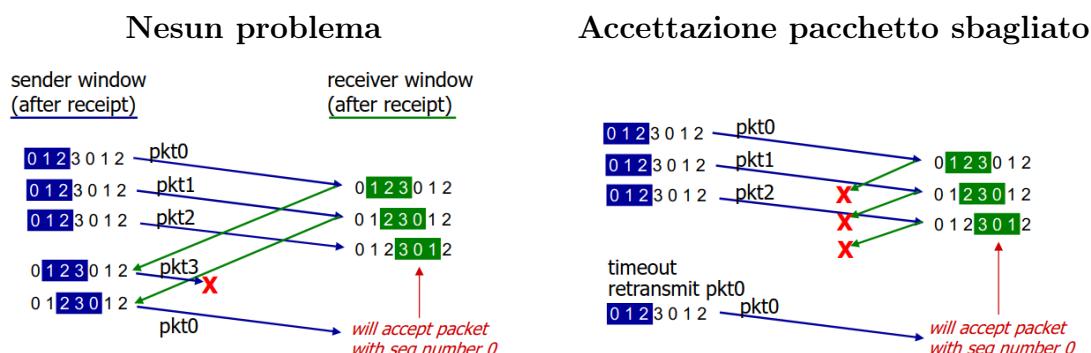


Esempio:



Tuttavia, anche il protocollo Selective Repeat non è privo di problematiche. In particolare, se la dimensione della finestra è troppo piccola, si può andare in contro a casi sfavorevoli (**dilemma della finestra**).

Ad esempio, con un range di numeri di sequenza pari a 0, 1, 2, 3 e una finestra di dimensione 3, si ha che:



**Esempio:**

- In una rete con un valore fisso  $m > 1$  (numero di bit della sequenza), è possibile utilizzare entrambi i meccanismi Go-Back-N e Selective Repeat, si indichino i vantaggi e gli svantaggi dell'impiego di ciascuno di essi. Quali altre considerazioni si devono fare per decidere quale meccanismo utilizzare?

- **Go-back-N**

- Ritrasmette tutti i frame inviati dopo il frame che si sospetta essere danneggiato o perso
- Se il tasso di errore è alto, spreca molta larghezza di banda
- Meno complicato
- Window size  $N - 1 = 2^m - 1$
- Riordinamento non è richiesto né lato mittente né lato destinatario
- Il destinatario non memorizza i frame ricevuti dopo il frame corrotto finché esso non viene ritrasmesso (dipende dall'implementazione)
- Non è richiesta alcuna ricerca di frame né lato mittente né destinatario

- **Selective Repeat**

- Ritrasmette solo i frame sospettati di essere persi o danneggiati
- Comparativamente meno larghezza di banda viene sprecata nella ritrasmissione
- Più complesso in quanto richiede l'applicazione di logica aggiuntiva, ordinamento e archiviazione, lato mittente e destinatario
- Window size  $\frac{N+1}{2} = 2^{m-1}$
- Il destinatario deve essere in grado di ordinare in quanto deve mantenere la sequenza dei frame
- Il destinatario memorizza i frame ricevuti dopo il frame danneggiato nel buffer finché il frame danneggiato non viene sostituito
- Il mittente deve essere in grado di cercare e selezionare il frame richiesto

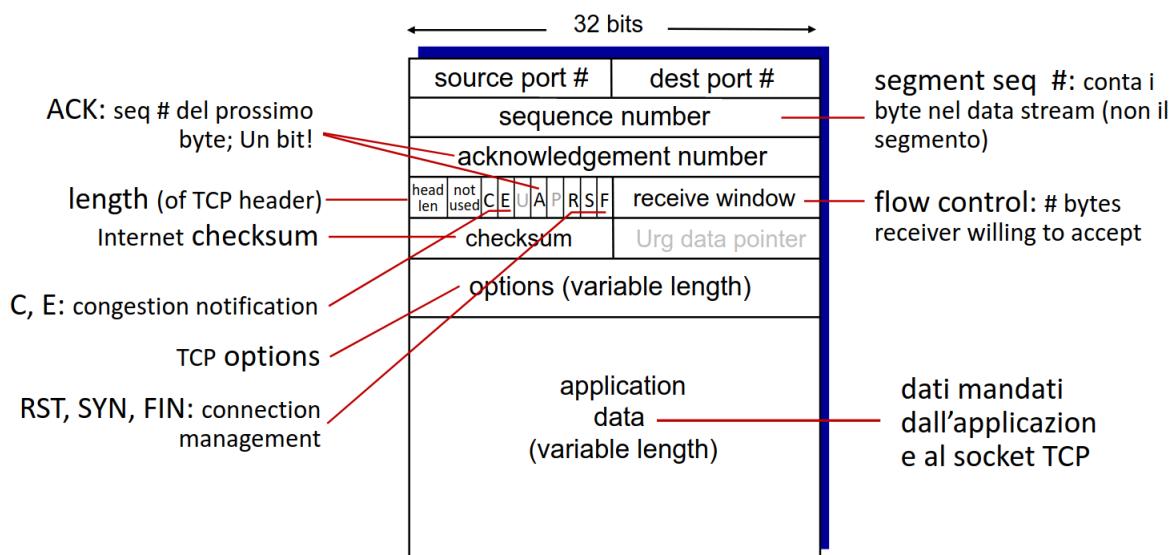
Un'idea aggiuntiva implementata nei **trasferimenti bidirezionali** (dunque dove entrambi i dispositivi sono sia mittente sia destinatario, coincidenti con la vita reale) è il **piggybacking**, dove nel momento in cui un pacchetto stia trasportando dati dal dispositivo A al dispositivo B, vengono trasportati anche i riscontri ricevuti da A inerenti ai pacchetti ricevuti da B, in modo che entrambi i dispositivi ne siano a conoscenza, gestendo efficientemente il rinvio dei pacchetti.

## 3.4 Protocollo TCP

Il **protocollo TCP** è un protocollo **end-to-end**, ossia con un solo mittente ed un solo destinatario, offrendo un **byte stream affidabile e in ordine**, dove i messaggi del livello di applicazione vengono concatenati in un unico stream, a differenza dell'UDP, dove ogni messaggio è un segmento diverso.

Come già discusso, il protocollo TCP è **orientato alla connessione**, dove l'**handshaking** inizializza lo stato del mittente e del destinatario prima dello scambio dei dati.

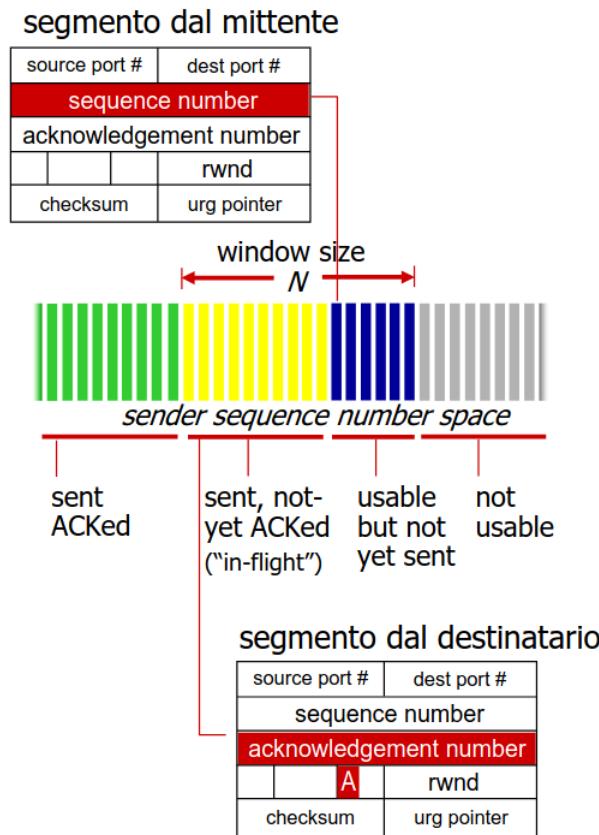
Il flusso dati, inoltre è **full duplex**, ossia bidirezionale all'interno della stessa connessione (dati full duplex), limitati tuttavia da un **Maximum Segment Size (MSS)**. Tuttavia, è necessario sottolineare che si tratta di un paradigma diverso dalla commutazione di circuito, poiché la rete non è a conoscenza dello stabilimento della connessione.



Per gestire il trasporto affidabile, il protocollo TCP utilizza:

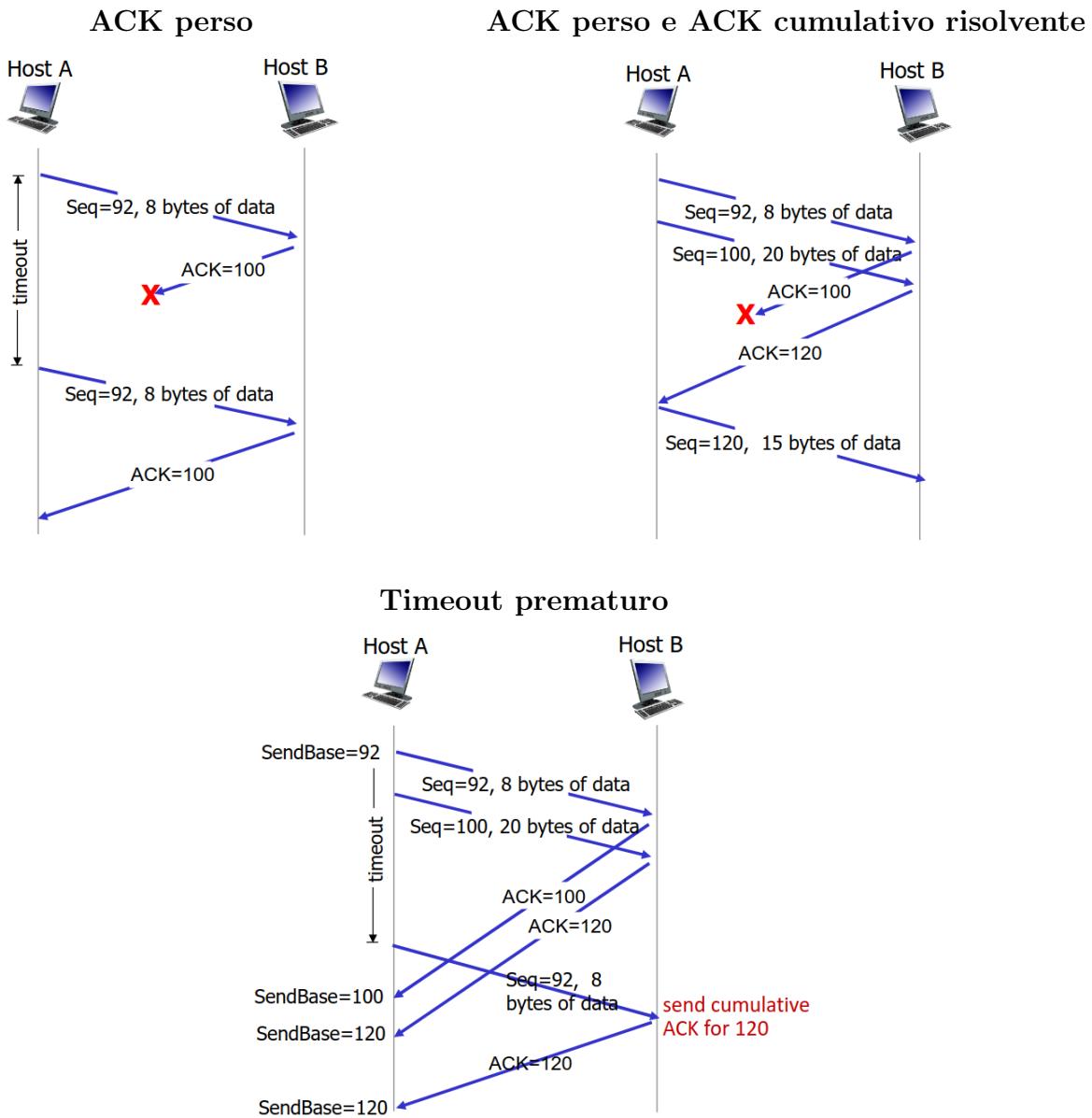
- **ACK cumulativi** e **pipelining**, dove il window size dipende dal **flow control**, ossia una garanzia sul non sovraccarico del destinatario da parte del mittente, e dal **congestion control**, ossia una garanzia sol non sovraccarico della rete da parte del mittente
- Il **numero di sequenza** dei segmenti del protocollo TCP corrisponde al numero di sequenza del **primo byte del settore data** del segmento stesso. Per quanto riguarda l'**ACK**, viene utilizzato il numero di sequenza del **byte successivo aspettato**.
- Nel momento in cui il mittente riceve dati dal livello di applicazione, viene creato il segmento e il suo numero di sequenza, avviando il timer a meno che esso non sia già in esecuzione. Inoltre, il **timer** utilizzato è **singolo** e collegato al **segmento non confermato più vecchio**. Allo scadere del **TimeoutInterval**, viene ritrasmesso il segmento che ha causato il timeout, riavviando il timer.

- Alla **ricezione di un ACK**, invece, se quest'ultimo copre segmenti precedentemente non confermati, vengono aggiornate le informazioni di tali pacchetti, avviando il timer se vi sono ancora segmenti non confermati, il quale sarà collegato al nuovo segmento più vecchio (ibrido tra Go-back-N e Selective Repeat)



Per quanto riguarda il destinatario, invece, si hanno quattro scenari:

- All'arrivo di un **segmento in ordine**, con **numero di sequenza atteso** e tutti i segmenti **precedenti già confermati**, viene inviato un **delayed ACK**: dopo aver atteso 500 ms per il prossimo segmento, se quest'ultimo non è stato ricevuto viene inviato l'ACK
- All'arrivo di un **segmento in ordine**, con **numero di sequenza atteso** e ma con un segmento precedente **non ancora confermato**, viene immediatamente inviato un ACK cumulativo confermando entrambi i segmenti
- All'arrivo di un **segmento fuori ordine** e con numero di sequenza maggiore di quello atteso (dunque vi è un **gap**), viene inviato immediatamente un ACK duplicato
- All'arrivo di un segmento che in parte o totalmente **riempie in ordine un gap**, viene immediatamente inviato l'ACK



### 3.4.1 Gestione del timeout e stima del RTT

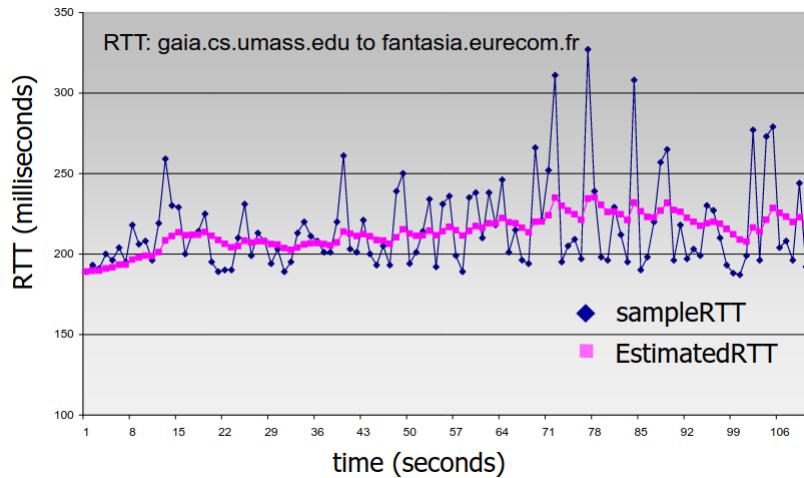
Il valore di timeout impostato deve essere **più lungo di un RTT**. Tuttavia, poiché il RTT è variabile, è necessario **stimarlo**.

Se il timeout scelto è **troppo corto**, si verificheranno troppi timeout prematuri, creando una serie di ritrasmissioni non necessarie. Se invece è **troppo lungo**, vi è una reazione troppo lenta a seguito della perdita di un pacchetto.

Per stimare il RTT, viene campionato un valore **SampleRTT**, ossia il tempo misurato dalla trasmissione del segmento fino alla ricezione dell'ACK (ignorando le ritrasmissioni). Poiché SampleRTT varia, viene utilizzata una **media delle misurazioni recenti** e non solo dell'ultimo SampleRTT (EWMA - Exponential Weighted Moving Average):

$$\text{EstimatedRTT} = (1 - \alpha) \cdot \text{PreviousEstimatedRTT} + \alpha \cdot \text{SampleRTT}$$

dove tipicamente si ha  $\alpha = 0.125$  e dove l'influenza del campione passato diminuisce in modo esponenziale



Il valore del **TimeoutInterval**, dunque, corrisponderà al valore attuale dell'EstimatedRTT sommato ad un **margine di sicurezza**

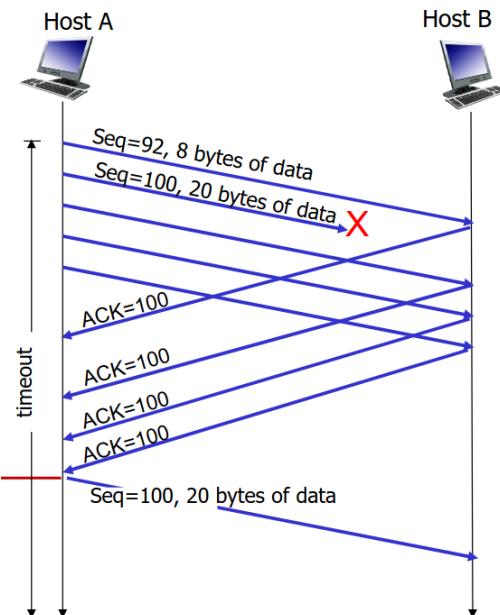
$$\text{TimeoutInterval} = \text{EstimatedRTT} + 4 \cdot \text{DevRTT}$$

dove DevRTT è l'EWMA della deviazione di SampleRTT da EstimatedRTT

$$\text{DevRTT} = (1 - \beta) \cdot \text{PreviousDevRTT} + \beta |\text{SampleRTT} - \text{EstimatedRTT}|$$

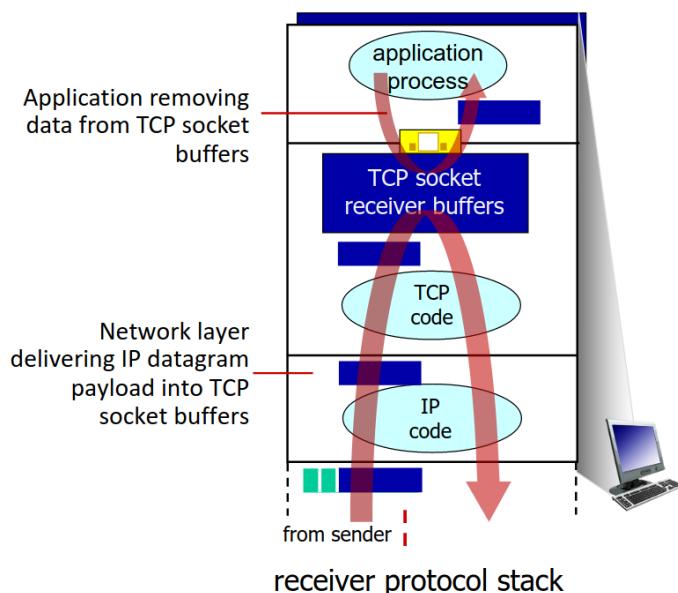
con un valore tipico  $\beta = 0.25$

Un'ottimizzazione ulteriore del protocollo TCP prevede l'implementazione del **fast retransmit**: se il mittente riceve 3 ACK aggiuntivi per gli stessi dati (dunque **tre ACK duplicati**), viene nuovamente inviato il segmento non confermato con numero di sequenza più piccolo poiché probabilmente tale segmento è andato perso, dunque non è necessario aspettare il timeout



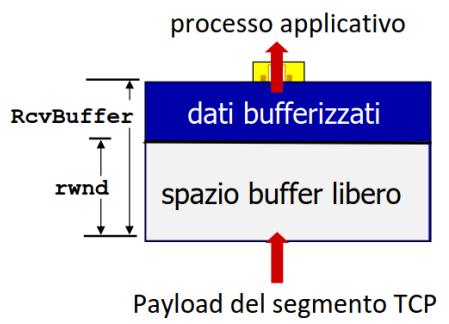
### 3.4.2 Controllo del flusso

Per poter funzionare correttamente, il protocollo TCP necessita di un **controllo del flusso**. Ad esempio, se la velocità con cui il livello di rete del destinatario fornisce i dati è maggiore rispetto a quella con cui il suo livello di applicazione rimuove i dati dal buffer del socket, il buffer andrà in **overflow**, implicando che i dati in eccesso vengano necessariamente **scartati**, risultando tuttavia come ricevuti correttamente dal destinatario.



Di conseguenza, è necessario che il **destinatario controlli il mittente**, impedendo che quest'ultimo possa riempire il buffer del destinatario trasmettendo troppi dati velocemente. Per gestire il flusso, quindi, viene utilizzata un campo **rwnd** (**receive window**) all'interno del segmento TCP:

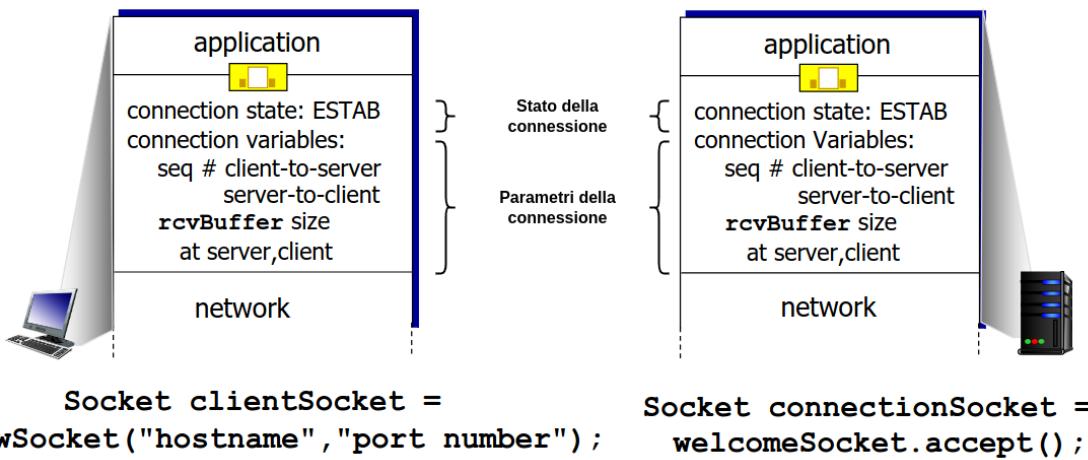
- Il destinatario inserisce in **rwnd** il numero di byte che è disposto ad accettare (dunque lo spazio rimanente nel buffer del socket)
- La dimensione del **RcvBuffer** impostata tramite le opzioni socket (predefinito a 4096 byte) o gestita automaticamente dal sistema operativo
- Il mittente limita la quantità di dati inviati senza ACK al valore di **rwnd**, garantendo che il buffer di ricezione non vada in overflow



Buffering lato destinatario TCP

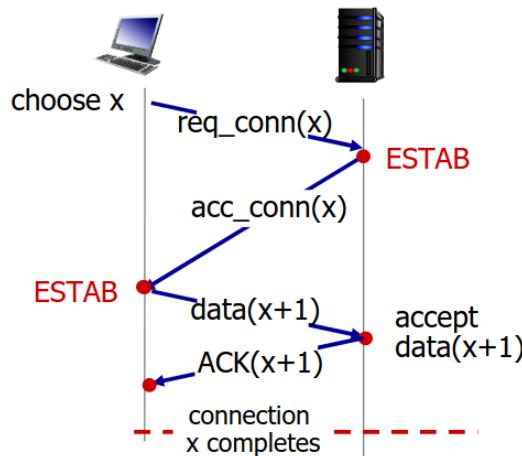
### 3.4.3 Gestione della connessione

Prima di effettuare lo scambio di dati, il mittente e il destinatario effettuano un **handshake**, dove viene determinata la disponibilità dell'un e dell'altro ad accettare di stabilire una connessione, concordando i parametri di quest'ultima (es: l'inizio del numero di sequenza)



Una prima implementazione dell'handshake è l'**handshake a 2 vie**, dove il mittente invia la richiesta di connessione al destinatario, il quale invia successivamente l'accettazione di tale richiesta, assumendo lo stato di connessione **ESTAB** (established).

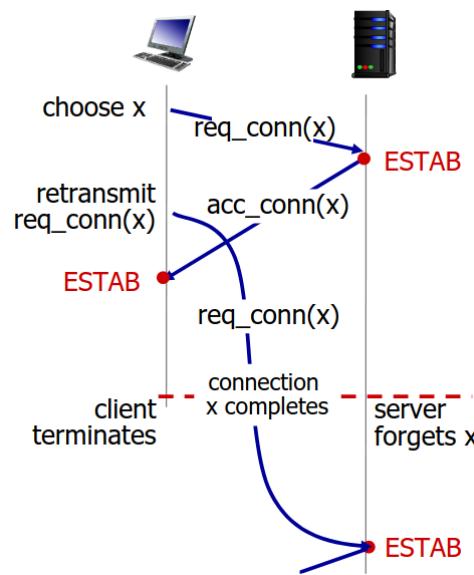
Una volta ricevuta l'accettazione, anche il mittente assumerà lo stato ESTAB, per poi procedere con l'invio effettivo dei dati.



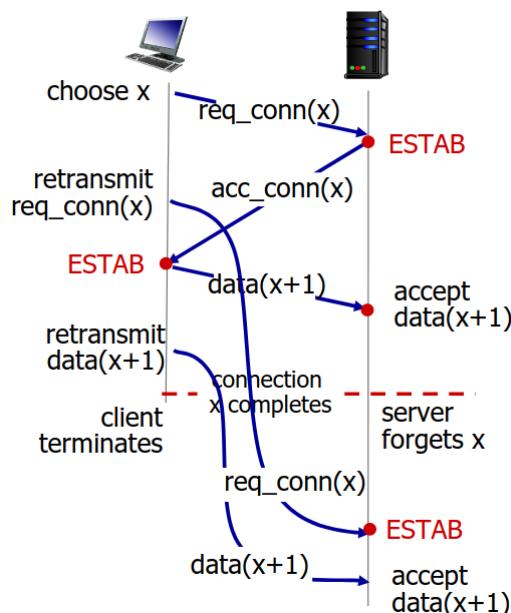
Tuttavia, in tale implementazione il destinatario **non è a conoscenza** della **ricezione** da parte del mittente del pacchetto di **accettazione** della connessione, presentando quindi **due problematiche fondamentali**:

- Nel caso in cui il mittente **rinvii la richiesta** di connessione allo scadere del timer TCP e l'accettazione del destinatario inerente alla prima richiesta giunge comunque dopo lo scadere del timer, il mittente suppone che la connessione sia andata a buon fine (nonostante il RTT sia estremamente basso), stabilendo quindi una **prima connessione**.

- Se tale connessione viene **terminata** prima che la seconda richiesta del mittente sia giunta al destinatario, quest'ultimo interpreterà la richiesta come una richiesta appartenente ad un'**seconda connessione**.
- Tuttavia, poiché tale richiesta era solo un rinvio della prima richiesta connessione, il client ignorerà la seconda richiesta di accettazione del server, creando quindi una **connessione fantasma senza client**



- Inoltre, nel caso in cui venga stabilita una connessione fantasma, si potrebbe andare incontro ad accettazioni di pacchetti dati duplicati

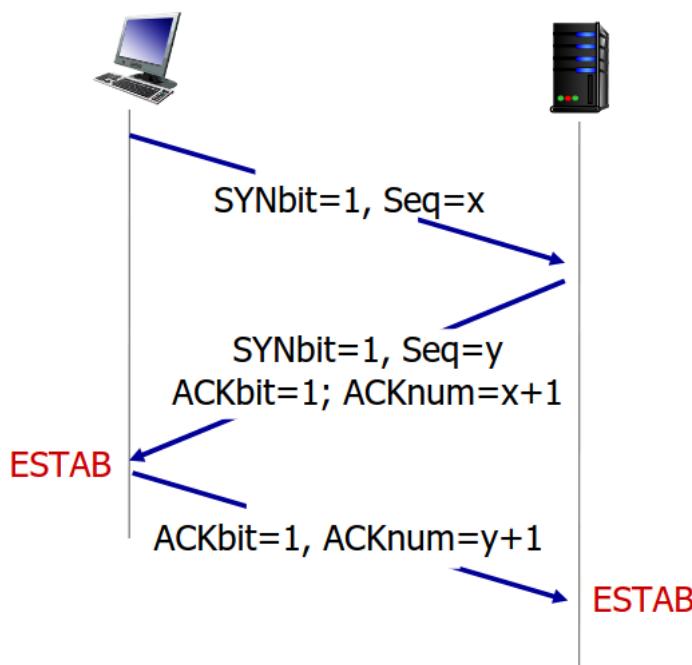


Di conseguenza, l'handshake TCP viene implementato attraverso uno scambio di 3 messaggi (**handshake a 3 vie**):

- Il mittente sceglie un numero di sequenza iniziale  $x$  e invia un pacchetto di tipo **SYN (synchronize)** al destinatario, richiedendo di stabilire una connessione.

Per inviare un pacchetto di tipo SYN, è sufficiente impostare il campo **SYN = 1** all'interno dell'header

- Una volta ricevuto il pacchetto SYN, il destinatario sceglie un numero di sequenza iniziale  $y$  e invia un pacchetto di tipo **SYN/ACK (synchronize and ACK)** al mittente, impostando i campi **SYN = 1** e **ACK = 1** nell'header
- Una volta ricevuto il pacchetto SYN/ACK, il mittente invia un pacchetto di tipo ACK (dunque con solo **ACK = 1**), passando in stato **ESTAB**
- Infine, una volta ricevuto il pacchetto ACK, anche il destinatario passerà in stato **ESTAB**



In questo modo, il destinatario sarà a conoscenza dello stato finale del mittente, risolvendo le due problematiche.

Per effettuare la **chiusura di una connessione**, il primo dispositivo (mittente o destinatario che sia) invia al secondo dispositivo un pacchetto di tipo **FIN (finished)**. Una volta ricevuto il pacchetto FIN, il secondo dispositivo risponderà con un pacchetto FIN/ACK, per poi inviare, dopo un breve lasso di tempo, un secondo pacchetto FIN. Analogamente, anche il primo dispositivo una volta ricevuto il pacchetto FIN invierà un pacchetto FIN/ACK.

Utilizzando tale **handshake a 4 vie**, entrambi i dispositivi riescono accertarsi il corretto termine della connessione. Inoltre, in tal modo entrambi i dispositivi possono terminare la connessione simultaneamente (creando una sorta di doppio handshake a 2 vie)

## 3.5 Controllo della congestione

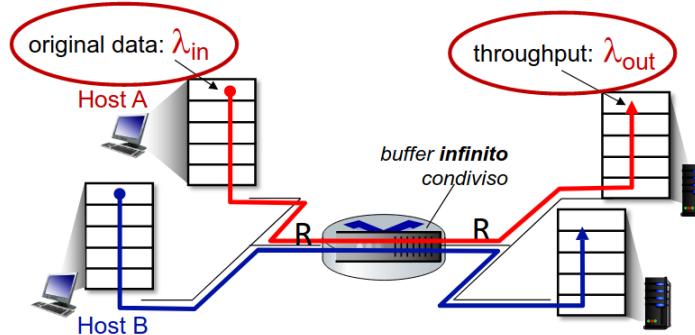
A differenza del **controllo del flusso**, il quale si occupa di gestire un mittente troppo veloce per un destinatario, il **controllo della congestione** si occupa di gestire situazioni in cui vi sono troppe fonti che inviano una grande quantità di dati troppo velocemente per poter essere gestiti correttamente dalla rete.

In presenza di congestione della rete, si manifestano **lunghi ritardi**, dovuti all'accodamento di troppi pacchetti nel buffer dei router, e **perdita di pacchetti**, dovuti agli overflow dei buffer dei vari router.

### 3.5.1 Cause e costi della congestione

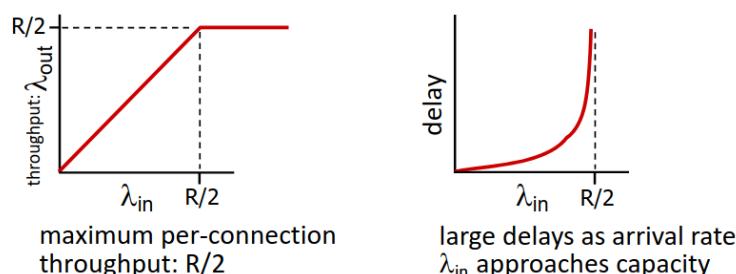
Consideriamo il seguente scenario:

- Vi sono due connessioni aperte passanti per un router con **buffer di dimensione infinita** e il transmission rate dei link è  $R$
- $\lambda_{in}$  è l'**arrival rate del router**, la quantità di dati inviati da un host della prima rete al router
- $\lambda_{out}$  è il **throughput del router**, la quantità di dati inviati dal router ad un host della seconda rete



In tal caso, poiché il buffer è infinito ci troviamo in una situazione in cui non sono necessarie ritrasmissioni dovute alla perdita del pacchetto. Di conseguenza, l'**arrival rate** riesce ad essere equivalente al **throughput**, corrispondente alla quantità di dati in uscita dal router.

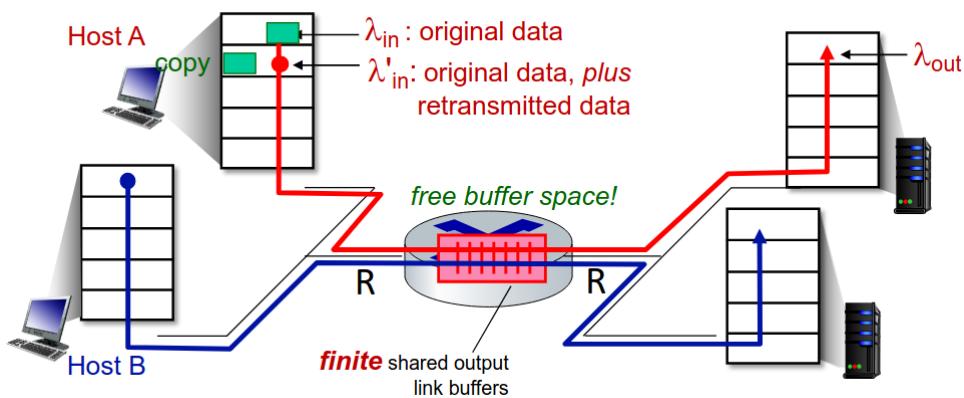
Tuttavia, poiché sono aperte due connessioni passanti per il router, il throughput massimo di ognuna di esse corrisponde a  $\frac{R}{2}$ . Inoltre, anche in tale scenario perfetto, man mano che  $\lambda_{in}$  si avvicina a  $\frac{R}{2}$ , il **delay cresce notevolmente**, per via carico eccessivo sui link stessi della rete.



Nella vita reale, ovviamente, la dimensione dei buffer è **finita**, implicando che alcuni pacchetti possano andar persi, venendo ritrasmessi dal mittente a seguito dello scadere del timeout.

Dato l'arrival rate  $\lambda'_{in}$  dei **dati originali sommati ai dati ritrasmessi**, è necessario sottolineare che:

- Al livello di applicazione, la quantità di dati inviati è equivalente a quella dei dati ricevuti, dunque si ha che  $\lambda_{in} = \lambda_{out}$
- Al livello di trasporto, tuttavia, l'input contiene anche i dati ritrasmessi, implicando che  $\lambda'_{in} \geq \lambda_{in}$



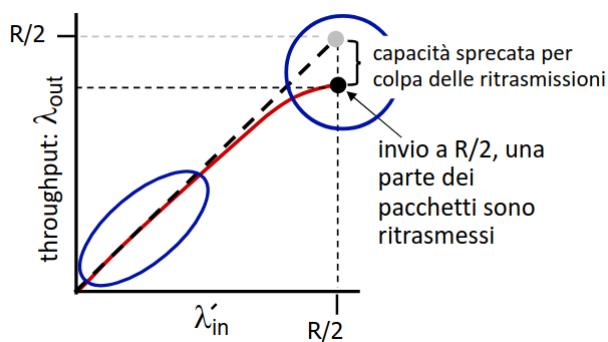
A questo punto, procediamo per **assunzioni** per studiare come la congestione influenzi l'infrastruttura:

- Idealmente, possiamo assumere che il mittente vada ad inviare i dati **soltamente** nel caso in cui esso sappia che i buffer dei router abbiano abbastanza spazio per ricevere il pacchetto (assunzione **perfect knowledge**)

In tal caso, ci troveremmo in una situazione identica allo scenario perfetto, poiché la rete sarebbe in grado di gestire il carico senza problemi, inviando i dati da un router all'altro al loro arrivo.

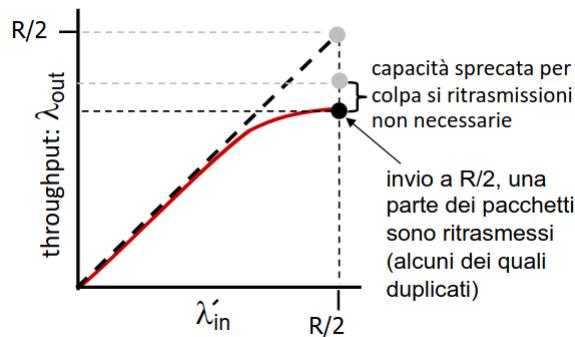
- In uno scenario più realistico, assumiamo che i pacchetti possano essere scartati a seguito di un **overflow** di un buffer e che il mittente sia a conoscenza perfetta di quali pacchetti siano andati persi, ritrasmettendoli (**perfect knowledge parziale**).

In tal caso, parte della capacità dei link verrebbe sprecata per via delle ritrasmissioni, **diminuendo il throughput**

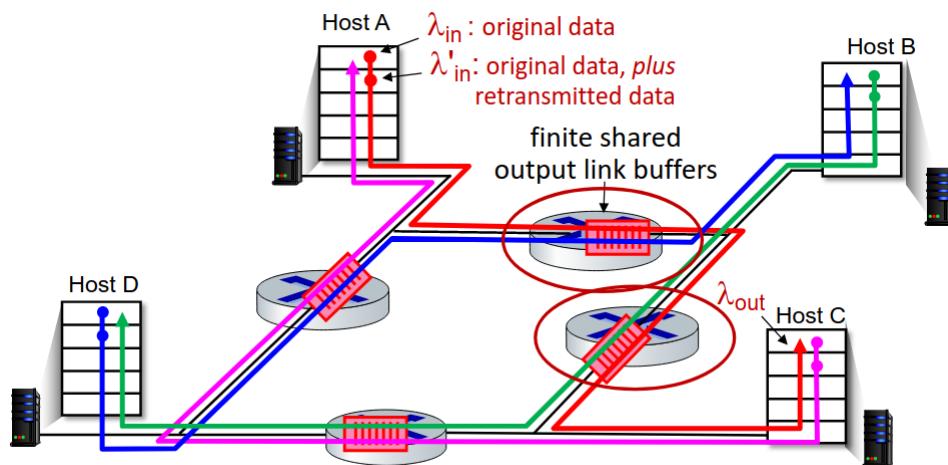


- In uno scenario reale, oltre alla perdita di pacchetti dovuta ad un overflow dei buffer, il **timer** del mittente può scadere prematuramente, inviando due copie dello stesso pacchetto ritrasmesso (**duplicati non necessari**)

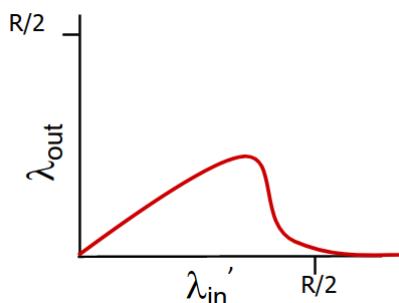
In tal caso, ulteriore parte della capacità dei link verrebbe sprecata per via delle ritrasmissioni non necessarie, **diminuendo il throughput ulteriormente**



Consideriamo invece ora una rete più realistica in cui tutti e quattro i dispositivi sono mittenti e vi siano più router colleganti le loro reti (rete multi-hop).



All'aumentare dei valori  $\lambda_{in}^{red}$  e  $\lambda_{in}^{red'}$  del collegamento rosso mostrato in figura, **tutti i pacchetti del collegamento blu vengono scartati**, poiché il buffer del router viene riempito dai pacchetti del collegamento rosso rinviati, portando il valore  $\lambda_{out}^{blue}$  a tendere a 0, diminuendo il throughput generale della rete



Possiamo quindi riassumere il comportamento della rete nei seguenti punti:

- Il throughput non può mai superare la capacità
- Il ritardo aumenta con l'avvicinarsi alla capacità
- La perdita e ritrasmissione riduce il throughput effettivo
- I duplicati non necessari riducono ulteriormente il throughput effettivo
- Viene sprecata capacità di trasmissione e buffering upstream per i pacchetti persi downstream

Infine, utilizziamo tali punti per definire i **costi della congestione**:

- È necessario un **lavoro** (numero di ritrasmissioni) **maggiori** per un dato throughput durante la congestione
- Il collegamento trasporta **più copie** dello stesso pacchetto **non necessarie**, riducendo il throughput massimo ottenibile
- Quando un pacchetto viene scartato, tutta la capacità di trasmissione upstream e la porzione di buffer utilizzata per esso viene **sprecata**

### 3.5.2 Controllo della congestione nel TCP

Per tentare di gestire la congestione, vengono principalmente utilizzati due approcci:

- **Controllo della congestione end-to-end**, dove non viene ricevuto alcun feedback esplicito dalla rete e la congestione viene **dedotta** dalle perdite e ritardi osservati dal mittente e il destinatario.

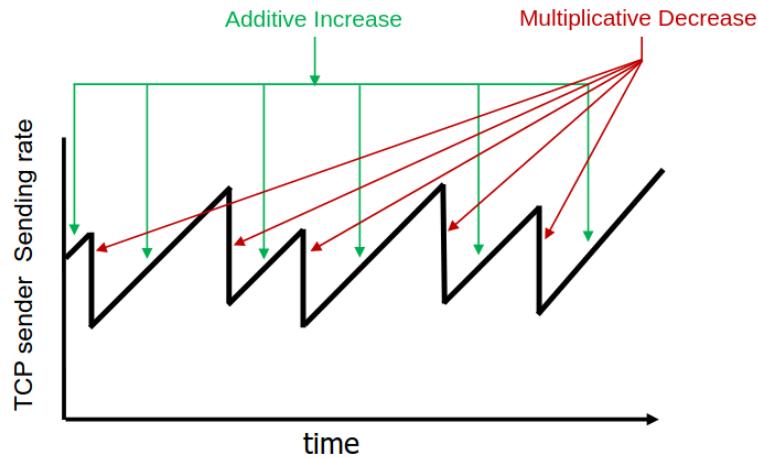
Tale approccio è adottato dal protocollo TCP

- **Controllo della congestione assistito dalla rete**, dove i router forniscono un feedback **diretto** agli host di invio/ricezione con flussi che passano attraverso router congestionati, indicando, in alcuni casi, direttamente il livello di congestione o la velocità di invio impostata

#### Definition 30. Algoritmo AIMD

L'algoritmo **AIMD** (**Additive Increase, Multiplicative Decrease**) è un algoritmo utilizzato da alcune versioni del protocollo TCP per **prevenire la congestione**, dove i mittenti possono **aumentare la velocità di invio** fino a quando si verifica una **perdita di pacchetti**, per poi diminuirla:

- **Additive Increase**: il rate viene aumentato di 1 MSS (Maximum Segment Size) ad ogni RTT fino a quando una perdita non viene osservata
- **Multiplicative Decrease**: ad ogni perdita osservata, il rate di invio viene dimezzato



### Definition 31. Congestion avoidance e Congestion window

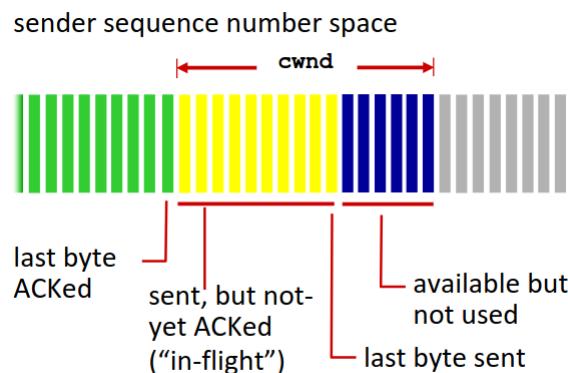
Approssimativamente, il protocollo TCP utilizza la seguente procedura di **prevenzione della congestione (congestion avoidance)**:

- Viene utilizzato un valore **cwnd** (**congestion window**), corrispondente alla quantità di byte inviata ad ogni RTT, da cui ne segue che

$$\text{Rate di invio} \approx \frac{\text{cwnd}}{\text{RTT}} \text{ B/s}$$

- Il mittente limita la trasmissione a  $\text{LastByteSent} - \text{LastByteAcked} \leq \text{cwnd}$
- Il valore cwnd varia dinamicamente reagendo alla congestione osservata. In particolare, utilizzando l'**Additive Increase** ad ogni ACK ricevuto si ha che:

$$\text{cwnd} \approx \text{prev\_cwnd} + \left( \text{MSS} \cdot \frac{\text{MSS}}{\text{prev\_cwnd}} \right)$$



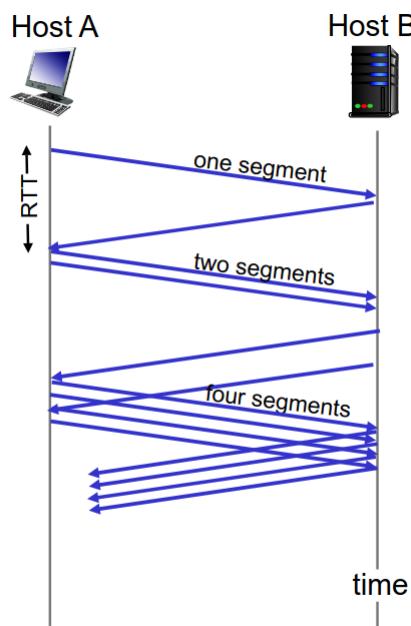
**Definition 32. Fast recovery**

Il **fast recovery** è una procedura utilizzata da alcune versioni del protocollo TCP per **prevenire la congestione**, dove ad ogni perdita rilevata a seguito di un **triplo ACK** duplicato viene **dimezzato il rate di invio** (ottenuto circa dimezzando il valore cwnd)

**Definition 33. Slow start**

Lo **slow start** è una procedura utilizzata da alcune versioni del protocollo TCP **prevenire la congestione**, dove:

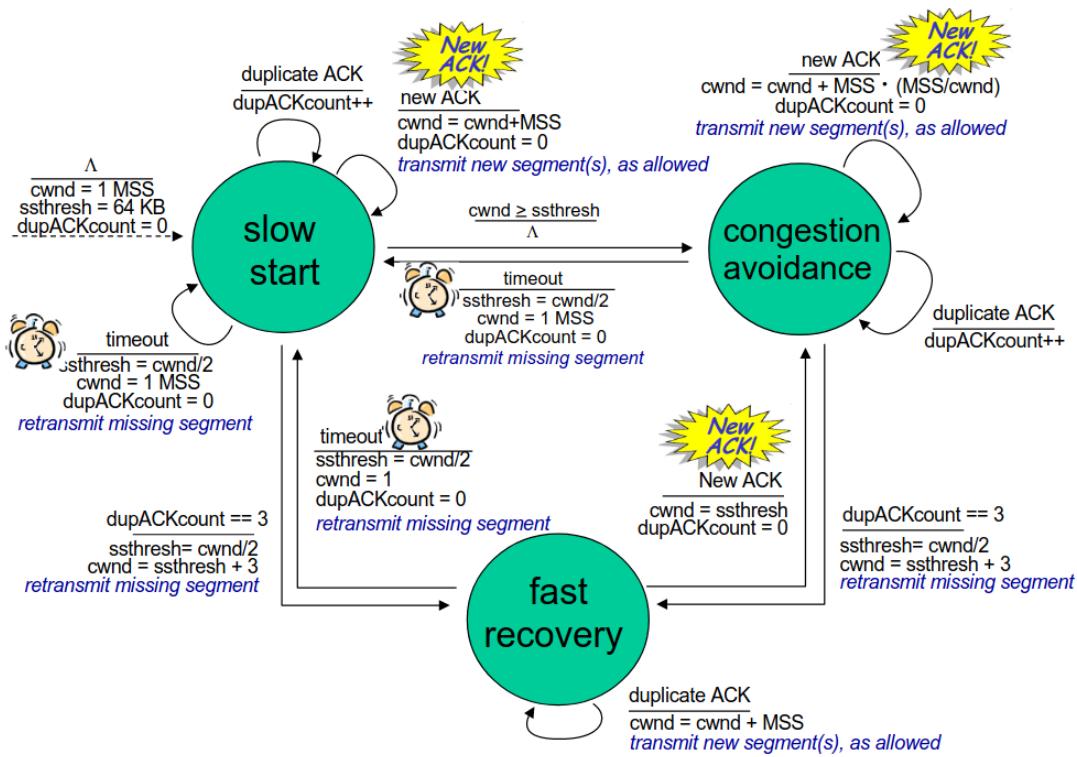
- Il valore cwnd viene **impostato ad 1 MSS** all'inizio della connessione o a seguito di un **timeout**
- Successivamente, il valore di cwnd viene **raddoppiato ad ogni RTT**, fino al rilevamento della prima perdita di pacchetto (**incremento esponenziale**)



*Nota: l'immagine superiore è un'approssimazione del vero comportamento, poiché raddoppiare cwnd non implica che venga raddoppiata la quantità di segmenti inviati, bensì che raddoppi la quantità di segmenti di dimensione massima inviati*

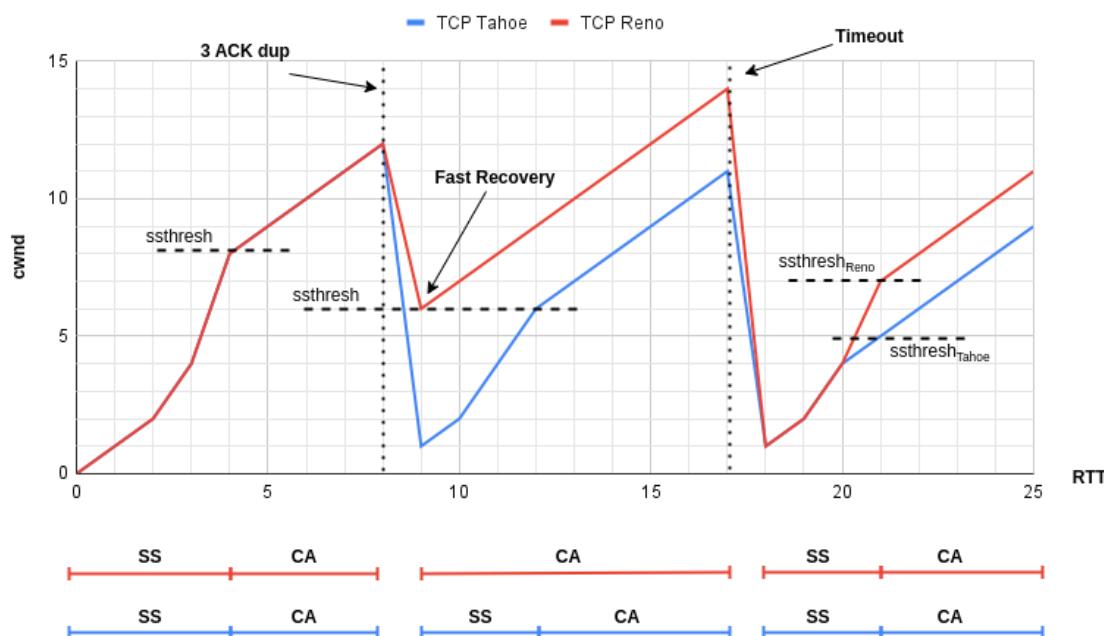
Tali meccanismi di prevenzione della congestione di rete vengono utilizzati principalmente da due versioni del protocollo TCP:

- **TCP Tahoe**: composto da un'algoritmo di **congestion avoidance** (tramite l'**Additive Increase**), l'algoritmo **slow start** e il **fast retransmit**
- **TCP Reno**: analogo al TCP Tahoe, ma con l'utilizzo aggiuntivo del **fast recovery** (FSM riportata in seguito)



In particolare, per effettuare il passaggio tra **slow start** e **congestion avoidance**, viene utilizzato un valore **ssthresh** (slow start threshold).

- A seguito del rilevamento di una perdita di pacchetto, il valore **ssthresh** viene impostato a  $\frac{cwnd}{2}$  (con il valore di **cwnd** precedente alla perdita). Se al prossimo RTT si verificasse che  $cwnd \geq ssthresh$ , viene posto  $cwnd = ssthresh$  e si passa dallo slow start al congestion avoidance
- Nel caso particolare del TCP Reno, se si verifica un triplo ACK duplicato, il valore **cwnd** viene dimezzato dal fast recovery, dunque si ha direttamente  $cwnd = ssthresh$



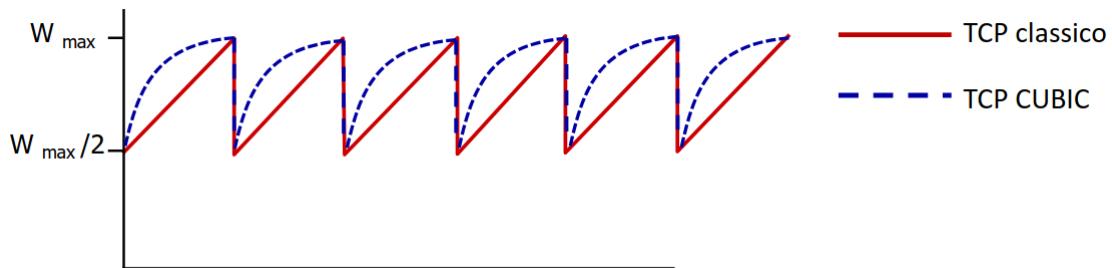
### Definition 34. TCP CUBIC

Il **TCP CUBIC** è una versione di TCP utilizzante l'algoritmo **CUBIC** per **prevenire la congestione**, il quale è definito dalla seguente logica:

- Viene utilizzato un valore  $W$ , corrispondente alla **velocità di invio** (o alla quantità di dati inviati dal mittente, ossia  $cwnd$ ). Ad ogni perdita rilevata, viene **dimezzato** tale valore.
- Il limite superiore  $W_{max}$  corrisponde alla **velocità di invio** nel momento in cui è stata rilevata una **perdita di pacchetto** (viene assunto che a seguito della perdita lo stato di congestione della rete non sia variato di molto)
- Viene utilizzato un valore  $K$  corrispondente al **momento di tempo stimato** in cui il valore  $W$  raggiungerà il limite superiore  $W_{max}$  (la modalità di stima viene definita dallo standard RFC 8312)
- Il valore  $W$  viene **incrementato in funzione del cubo della distanza tra il tempo corrente e  $K$** .

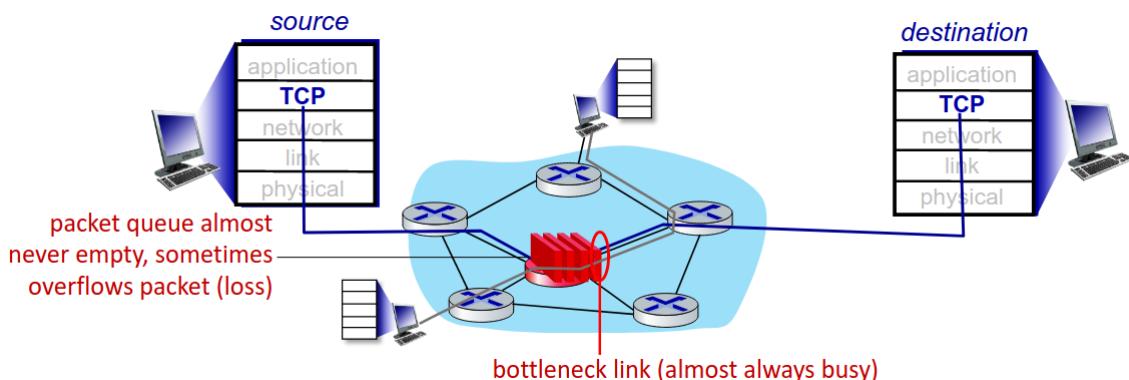
Di conseguenza, il valore  $W$  **crescerà molto rapidamente** quando il tempo corrente è lontano dal valore  $K$ , mentre **crescerà lentamente** al suo avvicinarsi.

Supponendo, ad esempio, che il valore di  $W_{max}$  rimanga sempre lo stesso nel tempo, il throughput del TCP CUBIC rispetto a quello standard risulta più elevato:

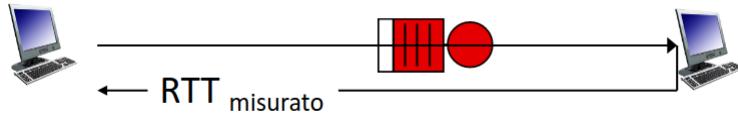


In particolare, essendo predefinito sul sistema operativo Linux, il TCP CUBIC è la versione del protocollo TCP più diffusa nei web server.

Una volta aver visto algoritmi e tecniche per prevenire la congestione, possiamo concentrarci sul link di uscita del stesso in cui si verifica la perdita (**bottleneck link**)



In presenza di un bottleneck link, **aumentando il rate di invio** non aumenterà il throughput per via del bottleneck link ma **aumenterà il RTT misurato**.



Una soluzione ottimale, dunque, è quella di mantenere il percorso end-to-end **quasi pieno**, ma senza superare la soglia stabilità, utilizzando un approccio **delay-based**:

- Il valore  $RTT_{min}$  è il RTT minimo osservato dal mittente, corrispondente quindi al RTT osservato quando il percorso **non è congestionato**
- Approssimativamente, il **throughput misurato** ad ogni RTT corrisponde a:

$$\text{Throughput}_{\text{misurato}} = \frac{\text{Byte}_{\text{RTT}}}{\text{RTT}_{\text{misurato}}}$$

dove  $\text{Byte}_{\text{RTT}}$  è il numero di byte inviati nell'ultimo RTT, mentre il **throughput non congestionato** è pari a:

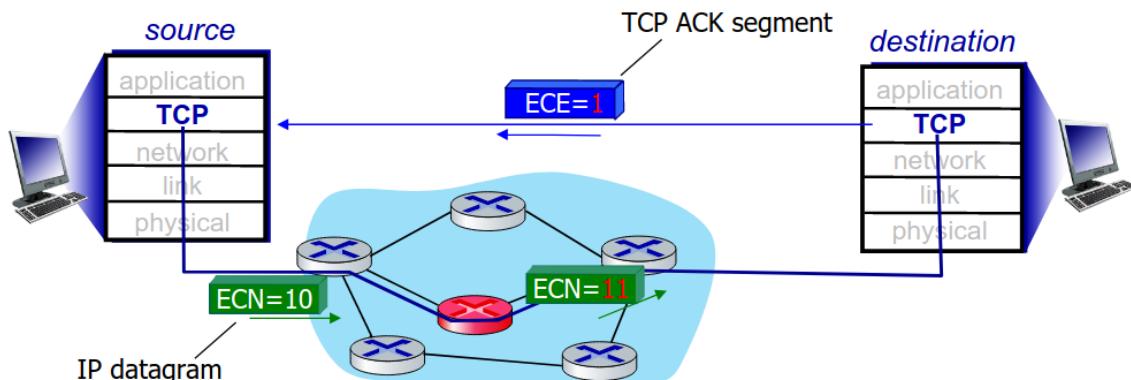
$$\text{Throughput}_{\text{non cong}} = \frac{\text{cwnd}}{\text{RTT}_{\text{min}}}$$

- Se il valore del throughput misurato è **molto vicino** a quello non congestionato, il valore di **cwnd** viene **incrementato** in modo lineare (dunque +1 MSS ad ogni RTT), mentre se è **molto inferiore** viene **decrementato** in modo lineare

Tramite tale approccio, alcune versioni di TCP (es: protocollo BBR) riescono ad indurre un controllo della congestione senza forzare delle perdite di pacchetto, massimizzando il throughput ma mantenendo basso il ritardo.

Altre versioni di TCP (es: TCP ECN), invece, implementano anche la seconda modalità di controllo della congestione, ossia il **controllo della congestione assistito dalla rete**, dove:

- Due bit dell'**header del livello di rete** vengono contrassegnati dal **router** per indicare lo stato della congestione
- Una volta raggiunto il destinatario, quest'ultimo imposterà il bit ECE sul segmento ACK per notificare al mittente lo **stato della congestione**



## 3.6 Equità nei protocolli di trasporto

### Proposition 2. Equità nelle connessioni

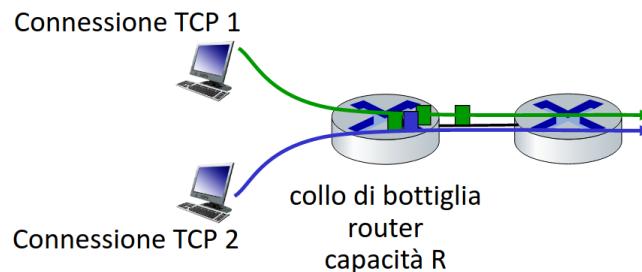
Affinché un protocollo di trasporto sia definibile **equo**, se  $K$  sessioni di tale protocollo condividono lo stesso **bottleneck link** con larghezza di banda  $R$ , ciascuna delle  $K$  sessioni deve avere una velocità media pari a  $\frac{R}{K}$

Notiamo con facilità che il **protocollo UDP** sia un protocollo **non equo** per via dell'assenza di un controllo della congestione e di limiti sulla banda utilizzabile.

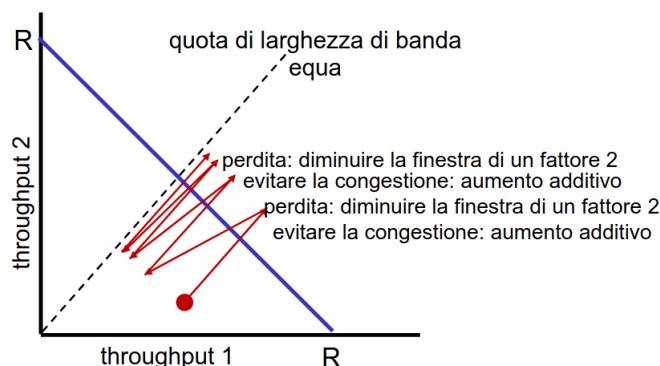
Per tale motivo, spesso applicazioni multimediali utilizzanti il protocollo UDP (es: i servizi streaming) "rubano" velocità di connessione ad altre applicazioni.

Per quanto riguarda il protocollo TCP, invece, è necessario effettuare uno studio:

- Consideriamo il seguente scenario con due sessioni TCP con algoritmo AIMD concorrenti sullo stesso bottleneck link



- Tramite l'**additive increase** viene generata una pendenza pari ad 1
- Tramite il **multiplicative decrease** viene ridotto proporzionalmente il throughput



Sotto **ipotesi idealizzate**, dunque, il **protocollo TCP** risulta essere **equo** (es: stesso RTT, numero fisso di sessioni, ...).

Tuttavia, molte applicazioni moderne utilizzano **più di una connessione TCP parallela** tra due host (es: un web browser). Di conseguenza, anche se la larghezza di banda fosse equamente distribuita tra tutte le connessioni possibili tra i due host, tale applicazione otterrebbe comunque una quantità di banda superiore alle altre applicazioni.

# Capitolo 4

## Livello di Rete

### 4.1 Panoramica del livello di rete

Come già accennato, a differenza del livello di trasporto, il **livello di rete** si occupa della comunicazione logica tra i dispositivi stessi tramite il trasporto di segmenti dall'host di invio a quello di ricezione.

In particolare, ogni **router** della rete si occupa di esaminare i campi header di tutti i **datagrammi** IP che lo attraversano, spostandoli dalle porte di ingresso alle porte di uscita per trasferirli lungo il percorso end-to-end

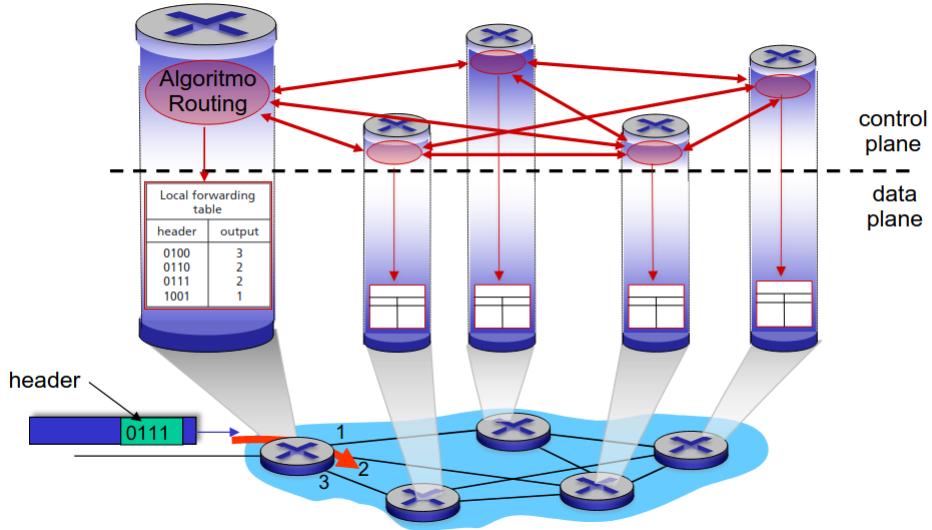
#### Definition 35. Forwarding e Routing

All'interno del livello di rete distinguiamo **due funzionalità fondamentali**:

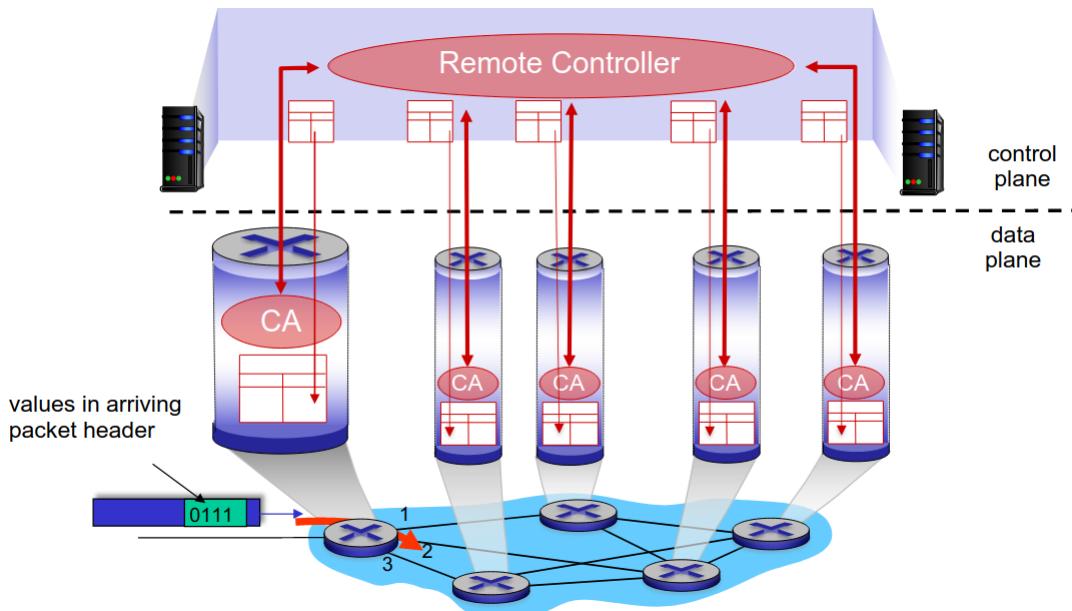
- **Forwarding (inoltro)**, ossia il trasferimento dei pacchetti dal link di ingresso di un router al link appropriato di uscita tramite la **gestione delle porte**
- **Routing (instradamento)**, ossia la determinazione (solitamente tramite **algoritmi** di routing) del percorso seguito dai pacchetti dalla sorgente alla destinazione

Per poter realizzare correttamente il servizio di trasferimento dei datagrammi, il livello di rete deve essere in grado di realizzare entrambe di tali funzioni fondamentali. Distinguiamo quindi il servizio di rete in **due strati**:

- **Data plane**, dove viene determinato come il datagramma in arrivo sulla porta di ingresso di un router venga inoltrato alla porta di uscita del router (**lavoro in locale**)
- **Control plane**, dove viene determinato come il datagramma venga instradato tra i router lungo il percorso end-to-end dall'host di origine all'host di destinazione (**logica a livello di rete**)



Oltre agli **algoritmi di routing** implementati all'interno dei singoli router, per il control plane può essere utilizzato anche il **Software-Defined Networking (SDN)**, dove un server remoto, detto **controller remoto**, calcola preventivamente tutte le **forwarding table** dei router, ossia le tabelle contenenti le regole di inoltro, i quali poi si connetteranno con il controller stesso per ottenere ed installare la propria tabella



Nella gestione dei "canali" di trasporto dei datagrammi dal mittente al destinatario viene utilizzato un modello **best effort**:

- Non vi è garanzia sull'effettiva **consegna** del datagramma a destinazione
- Non vi è garanzia sulle **tempistiche** o sull'**ordine** di consegna dei datagrammi
- Non vi è garanzia sulla **larghezza di banda** disponibile per il flusso end-to-end

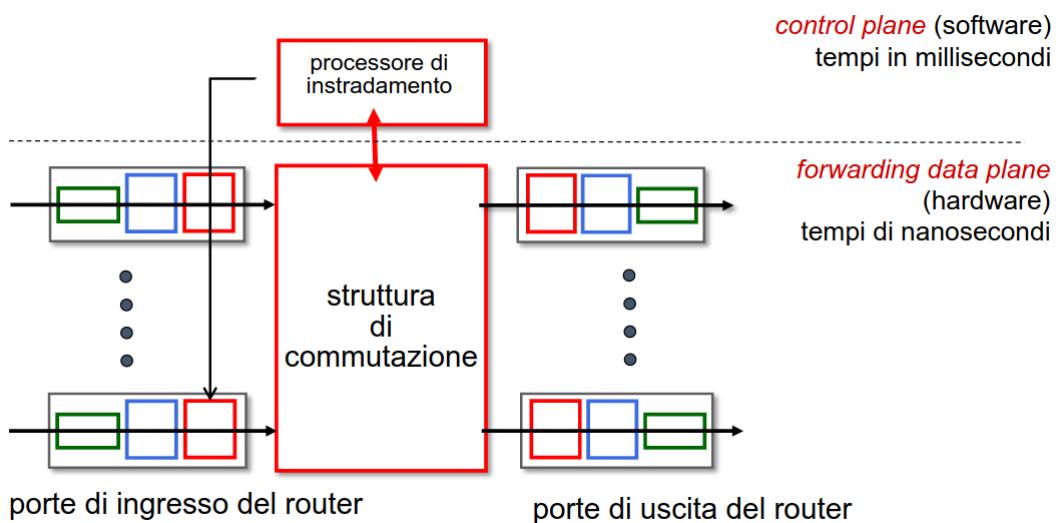
Nonostante i difetti, il modello best effort ha raggiunto un ottimo **successo**:

- La sua **semplicità** ha permesso ad Internet di essere ampiamente adottato ed implementato
- Una **larghezza di banda sufficiente** consente alle prestazioni delle applicazioni in tempo reale di essere per lo più sufficientemente ottime
- I **servizi distribuiti replicati a livello di applicazione** che si connettono vicino alle reti dei clienti (data center, reti di distribuzione di contenuti, ...) consentono di fornire servizi da più posizioni

## 4.2 Architettura e funzionalità dei router

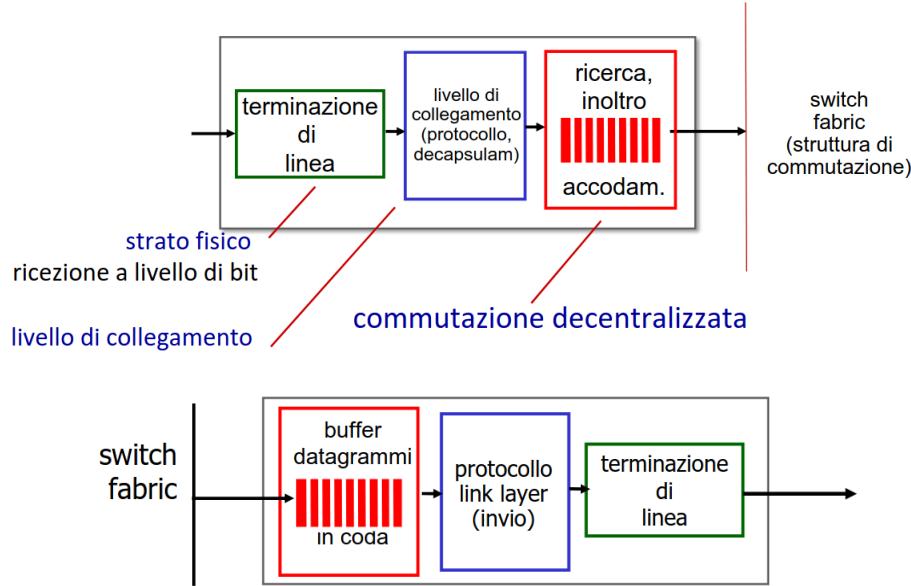
L'architettura interna di un router può essere riassunta in:

- Una serie di **porte di ingresso e di uscita** connesse a dei link
- Una **struttura di commutazione (switching fabric)** interposta tra le porte di entrata ed uscita
- Un **processore di instradamento**, il quale si occupa di effettuare i calcoli necessari per il routing



Le **porte di ingresso e di uscita** sono dotate di:

- Una propria **memoria** contenente le **forwarding table**
- Una **terminazione di linea** (ossia il termine/inizio del link ad esse associate)
- Un'interfaccia con il **livello di collegamento** tramite cui viene gestito il protocollo utilizzato (es: Ethernet)
- Una **coda di ingresso/uscita** in cui vengono inseriti temporaneamente i pacchetti appena vengono ricevuti o prima di essere spediti.



In particolare, all'interno della coda della porta di ingresso viene effettuata una **commutazione decentralizzata**, dove per ogni datagramma al suo interno viene cercata la porta di uscita utilizzando i valori del campo di intestazione e della forwarding table nella memoria della porta di input stessa (**match plus action**).

Per realizzare tale commutazione, vengono utilizzati il **destination-based forwarding**, dove l'inoltro è basato solo sull'indirizzo IP di destinazione presente negli header, e il **generalized forwarding**, dove l'inoltro è basato su un insieme di valori dell'header dei datagrammi.

#### Definition 36. Longest prefix matching

Il **longest prefix matching** è un algoritmo di forwarding basato sul destination-based forwarding: durante la ricerca della voce della forwarding table per un indirizzo di destinazione, viene selezionata l'entrata il cui indirizzo ha il **prefisso più lungo corrispondente** a quello dell'indirizzo di destinazione

**Esempio:**

- Consideriamo la seguente forwarding table, dove gli asterischi, detti **wildcard**, indicano che un qualsiasi valore tra 0 o 1 possa occupare tale posizione (**intervalli di indirizzi IP**)

Intervallo di indirizzi di destinazione	Porta di uscita
11001000 00010111 00010*** *****	0
11001000 00010111 00011000 *****	1
11001000 00010111 00011*** *****	2
altrimenti	3

- Utilizzando il longest prefix matching, il datagramma contenente l'indirizzo di destinazione 11001000 00010111 00010110 10100001 verrà inoltrato alla porta di uscita 0, poiché l'intervallo 11001000 00010111 00010\*\*\* \*\*\*\*\* possiede il prefisso corrispondente più lungo tra le tutte le entrate

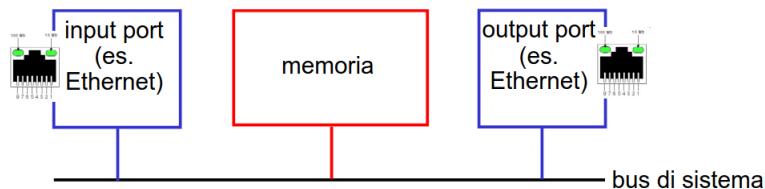
- Analogamente, il datagramma contenente l'indirizzo di destinazione `11001000 00010111 00011000 10101010` verrà inoltrato alla porta di uscita 1, poiché l'intervallo `11001000 00010111 00011000 *****` possiede il prefisso corrispondente più lungo tra le tutte le entrate (24 cifre rispetto alle 21 della porta 2)

Per quanto riguarda gli **switching fabric**, essi si occupano di effettuare il vero e proprio inoltro, trasferendo ogni pacchetto dai link di input al link di output appropriato.

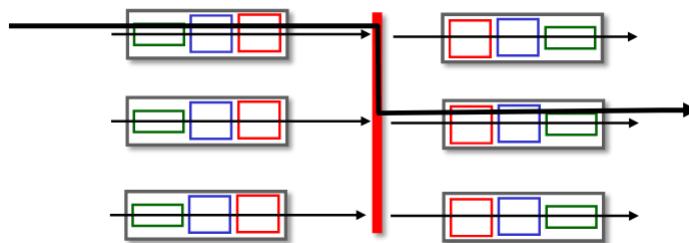
In particolare, lo **switching rate** dello switching fabric corrisponde alla velocità con cui i pacchetti possono essere trasferiti dagli ingressi alle porte (idealmente pari a  $N \cdot R$ , dove  $N$  è il numero di porte di ingresso/uscita ed  $R$  è il transmission rate dei link connessi alla porta, supponendo che esso sia uguale per tutti i link)

Le principali tre modalità di implementazione degli switching fabric prevedono:

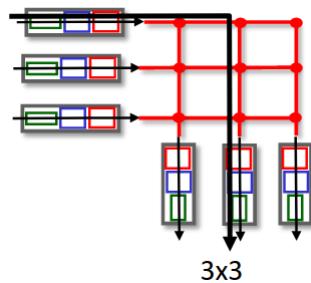
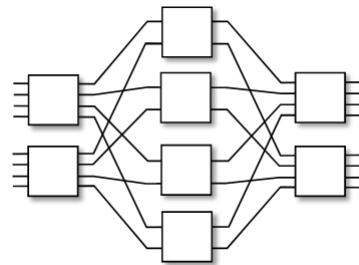
- Commutazione tramite memoria**, dove la commutazione è sotto diretto controllo della CPU, copiando i pacchetti in arrivo nella **memoria del sistema**, per poi inviarli sulle porte di uscita, implicando che lo switching rate sia limitato dalla larghezza di banda della memoria.



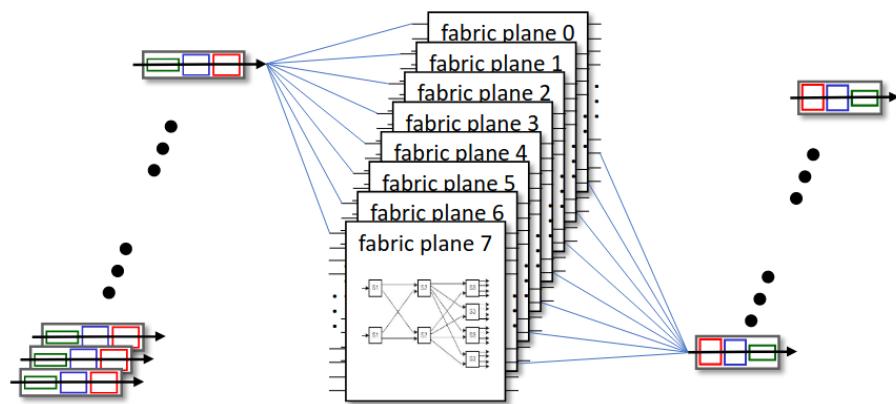
- Commutazione tramite bus**, dove i datagrammi vengono trasferiti dalla memoria della porta di ingresso alla memoria della porta di uscita tramite un **singolo bus** condiviso, implicando che lo switching rate sia limitato dalla larghezza di banda del bus e che vi sia una contesa tra le porte per l'utilizzo del bus



- Commutazione tramite reti di interconnessione**, dove vengono utilizzate reti di interconnessione (es: Crossbar, reti Clos) inizialmente sviluppate per connettere processori tra di loro. In particolare, vengono utilizzati **interruttori multistadio**, ossia interruttori  $N \times N$  formati da più stadi di interruttori più piccoli, e il **parallelismo**, frammentando i diagrammi in celle di lunghezza fissa all'ingresso per poi commutarle attraverso la rete di interconnessione e riassemblarle una volta raggiunta la porta di uscita

**Interruttore  $3 \times 3$** **Interruttore multistadio  $8 \times 8$   
formato interruttori più piccoli**

Inoltre, l'uso di reti di interconnessione permette un maggiore **scaling** utilizzando più "piani" di commutazione in parallelo

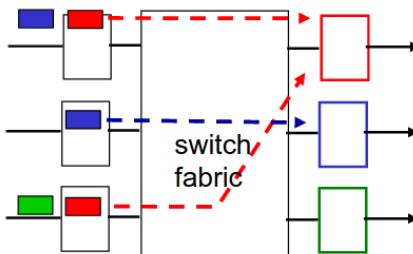


#### 4.2.1 Accodamento nelle porte

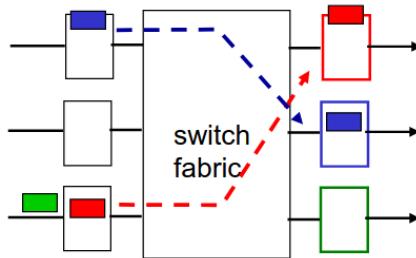
Nel caso in cui lo switch fabric sia più lento delle porte di input combinate, dunque se lo switching rate è minore del transmission rate dei link di entrata, potrebbe verificarsi dell'**accodamento** nei buffer di coda delle **porte di input**, generando un **queueing delay** e una possibile **perdita** dovuta all'overflow del buffer.

In particolare, possono verificarsi due situazioni sfavorevoli che possano generare accodamento nella porta di input:

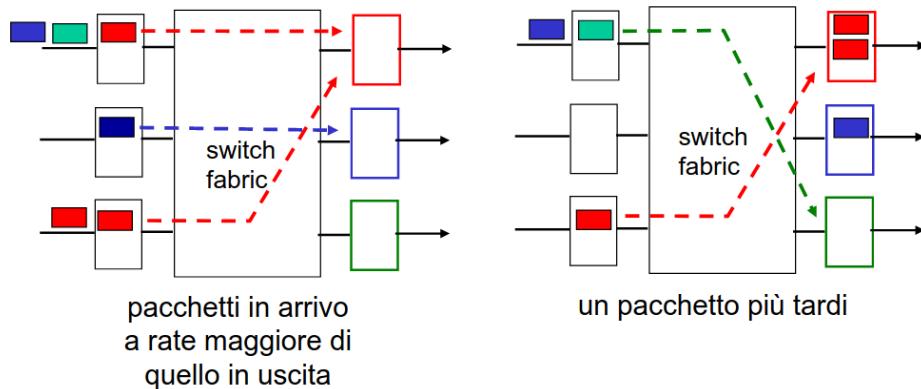
- **Contesa della porta di uscita:** in ogni istante può essere commutato un solo pacchetto verso una determinata porta di uscita, impedendo agli altri pacchetti di proseguire, bloccando di conseguenza la loro coda



- **Blocco HOL (Head-of-Line):** all'interno di ogni coda il datagramma nella parte anteriore della coda impedisce agli altri datagrammi in coda di poter proseguire



Analogamente, può verificarsi dell'**accodamento** all'interno nei buffer di coda delle **porte di output** nel caso in cui lo switch rate superi il transmission rate del link di uscita, generando ritardo e perdite di pacchetti.



Per gestire tali accodamenti, dunque, è necessario gestire minuziosamente i **buffer**. In particolare, la raccomandazione standard più recente prevede un buffering pari a

$$\text{Buffer} = \frac{\text{RTT} \cdot C}{\sqrt{N}}$$

dove  $N$  è il numero di flussi e  $C$  è la capacità dei collegamenti.

Tuttavia, un buffering eccessivo può **aumentare i ritardi** (in particolare all'interno dei router domestici):

- Con RTT lunghi si ottengono scarse prestazioni per le app in tempo reale ed una risposta TCP troppo lenta
- I buffer dovrebbero solo assorbire le fluttuazioni statistiche di occupazione in mancanza di congestione, senza creare un collo di bottiglia troppo pieno

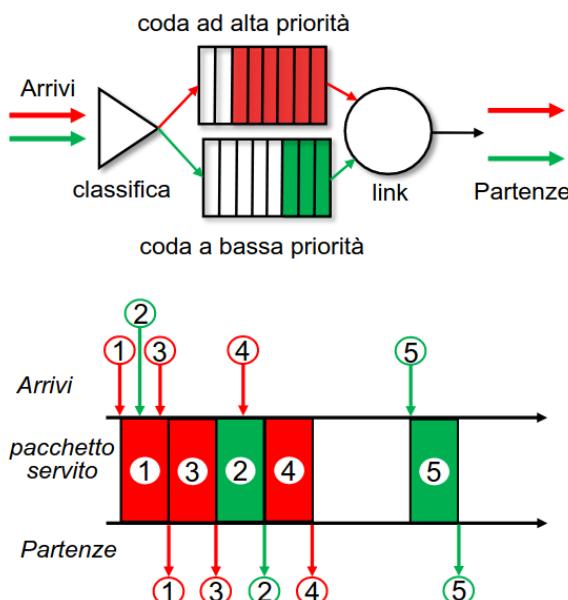
Di conseguenza, è necessario utilizzare un **protocollo di scarto** per scegliere quale pacchetti inserire nella coda e quali scartare quando il buffer è pieno. In particolare, vengono utilizzati principalmente il **tail drop**, dove viene scartato l'ultimo pacchetto in arrivo, e il **priority drop**, dove i pacchetti vengono scartati selettivamente in base alla priorità.

### 4.2.2 Scheduling dei pacchetti

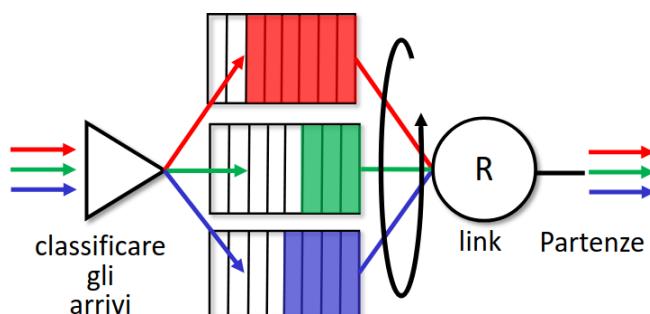
Per decidere quale sia il prossimo pacchetto da inviare, le porte di uscita vengono gestite tramite **politiche di scheduling**, cercando di ottenere le **migliori prestazioni** possibili mantenendo una **neutralità della rete**, ossia la modalità con cui un ISP dovrebbe allocare le proprie risorse.

In particolare, vengono principalemtnie utilizzate quattro politiche:

- **First come, First served (FCFS)**, dove i pacchetti vengono trasmessi in ordine di arrivo alla porta di uscita
- **Priority scheduling**, dove il traffico in arrivo viene classificato ed inserito in una classe di coda in base alla sua priorità, inviando il pacchetto della coda con la priorità più alta contenente pacchetti nel buffer (FCFS all'interno delle classi di priorità). La priorità viene determinata utilizzando un insieme di campi presenti nell'intestazione del pacchetto.



- **Round Robin (RR)**, dove il traffico in arrivo viene sempre classificato in code di classe utilizzando più code e l'invio dei pacchetti viene effettuato ciclicamente: viene ciclicamente eseguita una scansione delle code di classe, inviando a turno un pacchetto completo da ciascuna classe (se disponibile)

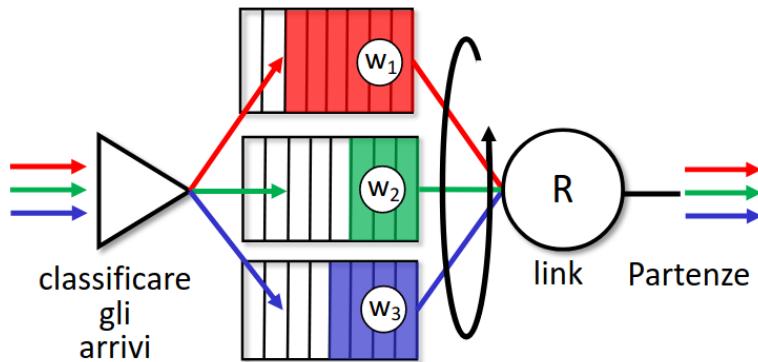


- **Weighted fair queueing (WFQ)**, dove il traffico in arrivo viene sempre classificato in code di classe utilizzando più code e l'invio dei pacchetti viene effettuato in base al peso delle classi: ogni classe  $i$  ha un peso  $w_i$  e riceve una quantità ponderata di servizio ad ogni ciclo, equivalente a

$$\frac{w_i}{\sum_{j=1}^k w_j}$$

dove  $k$  è il numero di classi.

In tal modo, si ottiene una gestione pari ad un Round Robin generalizzato e viene garantita una larghezza di banda minima per ogni classe di traffico.



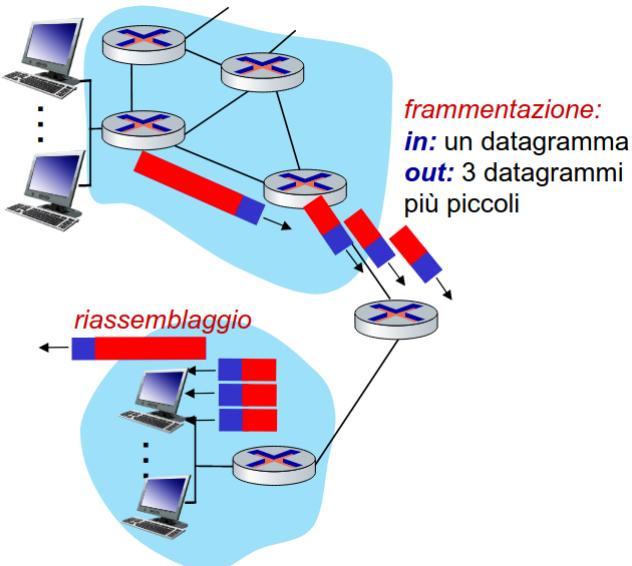
Inoltre, con la terminologia "neutralità della rete" vengono anche contrassegnati i **principi sociali/economici** e i **provvedimenti legali** atti a stabilire regole e politiche di gestione. In particolare, tale neutralità viene basata su tre principi:

- **No blocking**: non devono essere bloccati contenuti, applicazioni, servizi o dispositivi leciti e non dannosi soggetti ad una ragionevole gestione della rete
- **No throttling**: non si deve compromettere o degradare il traffico Internet legittimo sulla base di contenuti, applicazioni o servizi Internet o l'uso di un dispositivo non dannoso, soggetto a una ragionevole gestione della rete
- **No payed priority**: non deve essere prevista la prioritizzazione retribuita

### 4.2.3 Frammentazione dei datagrammi

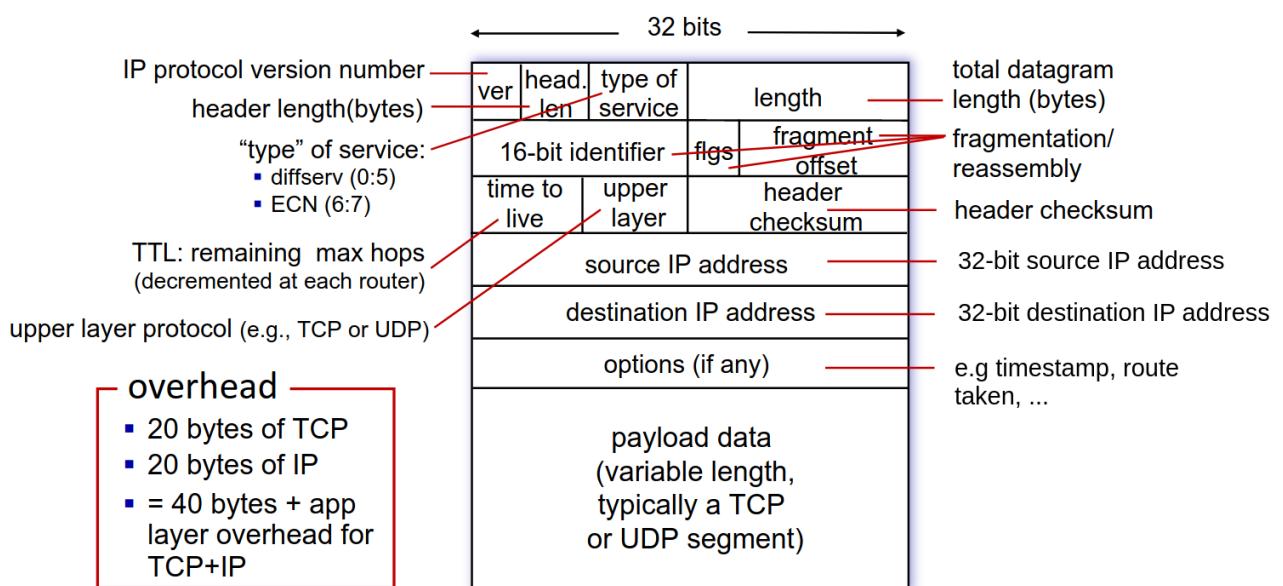
Per gestire meglio la **congestione**, i collegamenti di rete possiedono **Maximum Transmission Unit (MTU)**, ossia la dimensione massima comunicabile in una singola trasmissione del livello di rete all'interno del link stesso (dunque variabile a seconda del tipo di collegamento).

Di conseguenza, ogni datagramma di grandi dimensioni viene **frammentato** lungo la sua trasmissione a seconda dei link attraverso cui avviene il forwarding, venendo **ri-assemblato** solamente una volta raggiunta la destinazione, richiedendo quindi un campo all'interno dell'header per mantenere traccia dell'**ordine**.



### 4.3 Protocollo IP

Principalmente, all'interno del livello di rete viene utilizzato un solo protocollo standard, ossia il **protocollo IP (Internet Protocol)**.



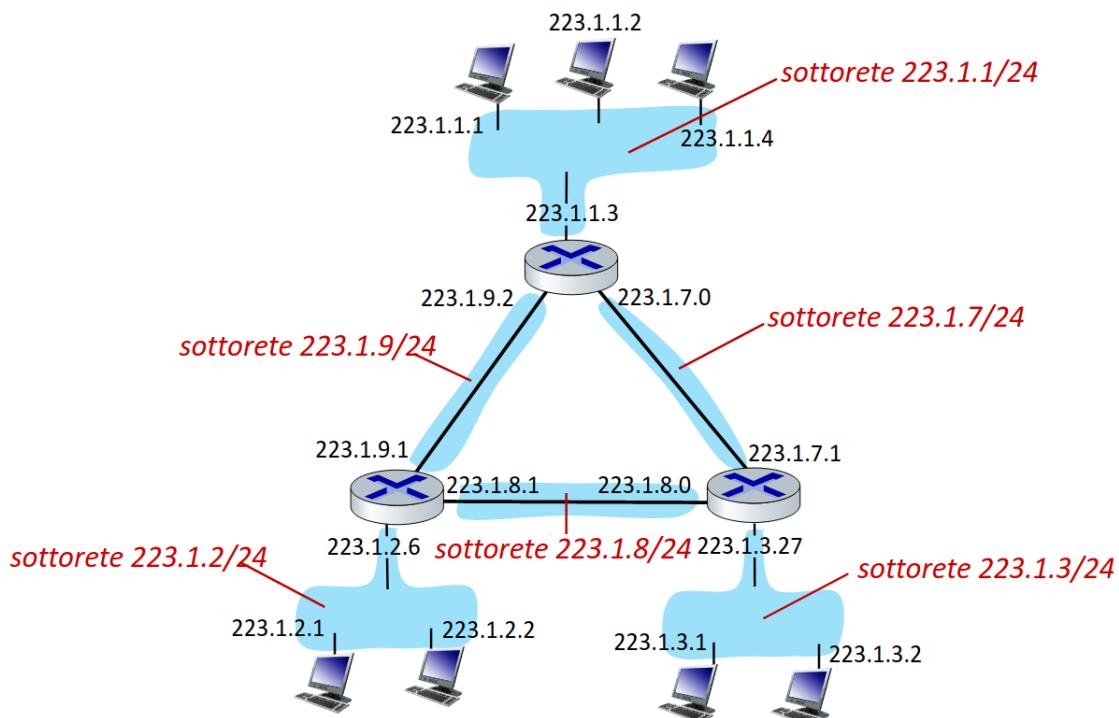
Il protocollo IP è basato su tre costrutti fondamentali:

- **Interfaccia**, ossia una **connessione** tra host e router associata ad un **collegamento fisico**. Solitamente, i router possiedono più interfacce, mentre un host possiede una o due interfacce (es: interfaccia Ethernet cablata ed interfaccia wireless Wi-Fi). Vengono gestite e determinate dal livello di collegamento.

- **Sottorete**, ossia un insieme di interfacce di dispositivi che possono raggiungersi fisicamente l'un l'altro senza passare attraverso un router intermedio (dunque tramite uno switch o altri mezzi).
- **Indirizzo IP**, ossia un identificatore a 32 bit associato ad un'interfaccia. Per facilitare la gestione agli umani, viene interpretato con una **notazione decimale puntata**, dove l'indirizzo viene suddiviso in quattro ottetti ed ogni ottetto viene interpretato come un valore decimale  
(es: l'indirizzo 11011111 00000001 00000001 00000001 viene interpretato come 223.1.1.1)

La struttura degli indirizzi IP viene definita dal **Classless Inter Domain Routing (CIDR, letto "cider")**:

- Ogni indirizzo possiede una **parte di sottorete**, condivisa tra i dispositivi della stessa sottorete e costituita da un determinato numero di bit più significativi (dunque più a sinistra). Gli  $n$  bit utilizzati per la parte di sottorete vengono definiti all'interno della **subnet mask**, i cui primi  $n$  bit più significativi sono posti ad 1 e i restanti posti a 0 (formata da 32 bit totali)
- Ogni indirizzo possiede una **parte di host**, identificante l'host stesso all'interno della sottorete e costituita dai bit meno significativi rimanenti
- Il formato degli indirizzi "ciderized" segue la struttura  $a.b.c.d/x$ , dove  $a.b.c.d$  è l'indirizzo IP e  $x$  è la quantità di bit posti ad 1 della subset mask  
(es: l'indirizzo 200.23.16.0/23 viene interpretato come 11001000 00010111 00010000 00000000, dove i primi 23 bit rappresentano la sottorete)



L'utilizzo della **subnet mask** risulta essere estremamente comodo per ottenere in modo efficiente informazioni sul **blocco di indirizzi** definito dalla sottorete:

- Il **numero di indirizzi del blocco** corrisponde al bitwise NOT della maschera sommato ad 1

$$\text{Num. indirizzi} = \text{NOT(mask)} + 1$$

- Il **primo indirizzo del blocco** corrisponde al bitwise AND tra un indirizzo qualsiasi del blocco e la maschera

$$\text{Primo indirizzo} = (\text{qualsiasi indirizzo del blocco}) \text{ AND } (\text{mask})$$

(solitamente viene utilizzato per indicare direttamente la sottorete)

- L'**ultimo indirizzo del blocco** corrisponde al bitwise OR tra un indirizzo qualsiasi del blocco e il bitwise NOT della maschera

$$\text{Ultimo indirizzo} = (\text{qualsiasi indirizzo del blocco}) \text{ OR } (\text{NOT(mask)})$$

### Proposition 3. Indirizzi IP speciali

Alcuni **indirizzi IP speciali** vengono utilizzati come scorciatoie per ottenere determinati comportamenti:

- L'indirizzo **0.0.0.0** viene utilizzato per indicare **qualsiasi indirizzo possibile**
- Gli indirizzi IP la cui parte di rete è impostata completamente a 0 si riferiscono alla **sottorete corrente**
- L'indirizzo **255.255.255.255** permette la **trasmissione broadcast**, ossia ad ogni dispositivo, sulla rete **locale**
- Gli indirizzi con parte di rete opportuna e la parte di host impostata completamente ad 1 permettono l'invio di pacchetti **broadcast a reti distanti**
- Gli indirizzi in cui il primo ottetto è impostato a 127, dunque gli indirizzi **127.x.y.z**, vengono riservati al **loopback**. Tali pacchetto non vengono trasmessi sul mezzo di trasmissione ma vengono elaborati localmente dall'host e trattati come se fossero pacchetti in arrivo.

#### 4.3.1 Protocollo DHCP e indirizzamento gerarchico

Ogni **host** può ottenere il suo indirizzo IP all'interno della sua rete, dunque la sua **parte host** da associare alla parte di sottorete, in modalità **statica** tramite una sua configurazione interna (es: il file `/etc/rc.config` sul sistema operativo Unix) o in modalità **dinamica** ottenendo tale indirizzo IP da un server addetto, non richiedendo alcuna configurazione.

### Definition 37. Protocollo DHCP

Il **protocollo Dynamic Host Configuration Protocol (DHCP)** è un protocollo a livello di applicazione in grado di assegnare dinamicamente gli indirizzi IP agli host interni ad una rete:

1. Nel momento in cui si unisce alla rete, l'host effettua una richiesta broadcast interna alla rete, cercando un server DHCP al suo interno (**DHCP discover**)
2. Il server DHCP (solitamente collocato all'interno del router stesso) risponde alla richiesta offrendo all'host un possibile indirizzo IP (**DHCP offer**)
3. L'host risponde al server accettando tale indirizzo IP, prendendolo "in prestito" fino a quando esso non si disconnetterà dalla rete (**DHCP request**)
4. Il server DHCP risponde all'host confermando la presa in prestito di tale indirizzo (**DHCP ack**)

### Observation 3. Riutilizzo di indirizzi precedenti

Se l'host appena unitosi alla rete **ricorda** e desidera **riutilizzare** il precedente indirizzo IP, verrà inviato direttamente il messaggio di DHCP request in modalità broadcast, permettendo al server DHCP di rispondere immediatamente con un DHCP ack

Oltre all'indirizzo IP, solitamente il protocollo DHCP restituisce anche altre informazioni sulla sottorete, come l'indirizzo del router di **gateway**, ossia il primo router raggiungibile dal client per comunicare in rete, il nome e l'indirizzo IP del server DNS (se ve ne è uno) e la subnet mask della sottorete.

Per quanto riguarda le **reti** invece, ognuna di esse ottiene il suo indirizzo IP, dunque la sua **parte di sottorete**, effettuando una richiesta al proprio ISP, il quale allocherà una **porzione del proprio spazio di indirizzi**.

**Esempio:**

- Un ISP possiede il blocco di indirizzi **200.23.16.0/20**
- L'ISP decide di suddividere il suo blocco in 8 blocchi, associando ciascuno di essi ad un'organizzazione richiedente un blocco di indirizzi.

Proprietario	Indirizzo IP
ISP	<u>11001000 00010111 00010000 00000000</u> 200.23.16.0/20
Organiz. 1	<u>11001000 00010111 00010000 00000000</u> 200.23.16.0/23
Organiz. 2	<u>11001000 00010111 00010010 00000000</u> 200.23.18.0/23
...	...
Organiz. 7	<u>11001000 00010111 00011110 00000000</u> 200.23.30.0/23

In tal modo, è possibile creare una struttura di **indirizzamento gerarchico**, permettendo un'instradamento più efficiente tramite la pubblicizzazione di percorsi più specifici per poter raggiungere le sottoreti.

### Esempio:

- Nell'esempio precedente, l'ISP pubblicizza un percorso più specifico per poter raggiungere le varie organizzazioni tra cui ha diviso il suo blocco di indirizzi, richiedendo all'esterno che gli venga inviato qualsiasi pacchetto avente un'indirizzo ricadente in tali range

In particolare, l'indirizzamento gerarchico è anche uno dei motivi per cui il **longest prefix matching** risulti essere così efficiente, cercando di inviare il pacchetto seguendo l'orientamento gerarchico.

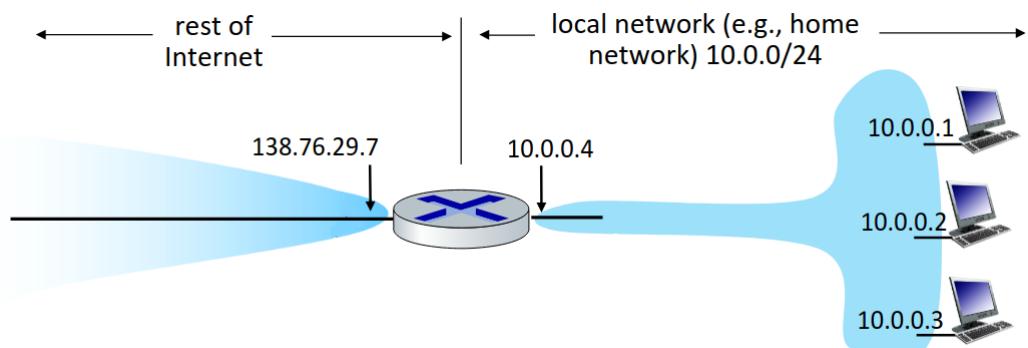
Per ottenere un blocco di indirizzi, ogni ISP deve effettuare una richiesta all'**Internet Corporation for Assigned Names and Numbers (ICANN)**, la quale alloca gli indirizzi IP attraverso **5 registri regionali** e gestisce la zona radice del DNS, inclusa la delega della gestione dei singoli TLD.

Tuttavia, nel 2011 l'ICANN ha assegnato l'ultima parte disponibile di **indirizzi IPv4** (ossia la versione trattata fino ad adesso), **esaurendo** a tutti gli effetti **gli indirizzi assegnabili**.

### 4.3.2 Servizio NAT e Protocollo IPv6

Per aggirare il problema dell'esaurimento degli indirizzi IPv4, è stato idealizzato un **escamotage** tramite l'implementazione del **Network Address Translation (NAT)**:

- Tutti i dispositivi interni ad una sottorete condividono **un solo IPv4 pubblico**, il quale viene utilizzato per identificare l'intera sottorete al mondo esterno
- Tutti i dispositivi interni ad una sottorete possiedono un proprio **indirizzo IPv4 privato**, il quale può essere utilizzato solo all'interno della sottorete stessa
- Tutti i datagrammi che escono dalla rete locale hanno lo **stesso indirizzo IPv4 pubblico di origine** ma **diversi numeri di porta di origine**, utilizzando quindi la porta del livello di trasporto come un identificatore univoco per un indirizzo IPv4 privato

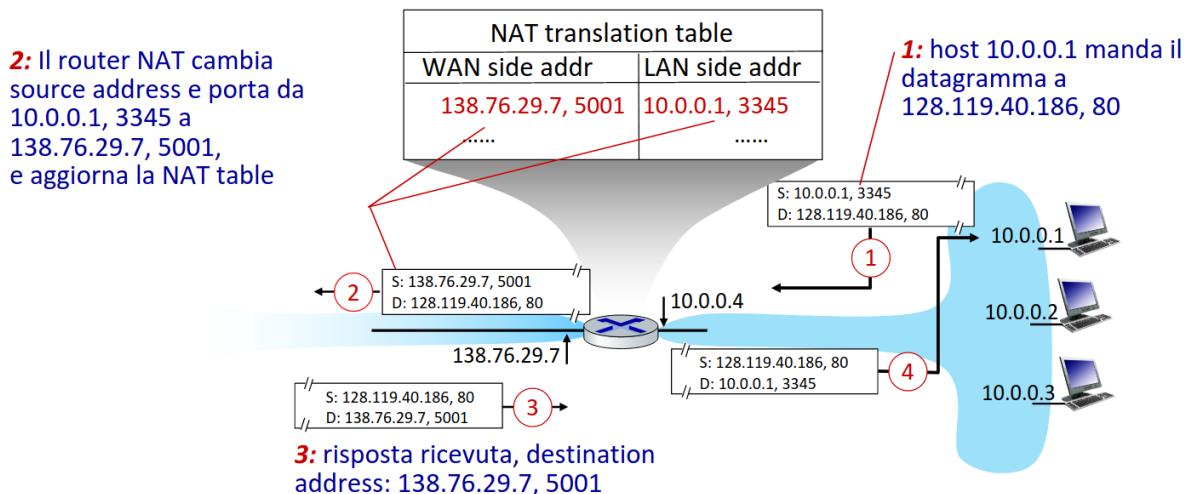


Tramite il NAT, dunque, l'ISP può utilizzare un **singolo indirizzo IPv4** per identificare ogni dispositivo interno ad una sottorete. Inoltre, gli indirizzi della rete locale possono essere gestiti separatamente, permettendo di cambiarli senza dover avvisare il mondo

esterno e fornendo una **maggior sicurezza**, poiché i dispositivi della rete locale non sono direttamente indirizzabili e visibili dall'esterno.

Per implementare il NAT, dunque, è necessario:

1. **Sostituire** la coppia <IP origine, Porta Origine> di ogni datagramma in uscita con la coppia IP NAT, Nuova Porta>, implicando che i client/server remoti risponderanno utilizzando la nuova coppia come indirizzo di destinazione
2. **Memorizzare** all'interno di una **tabella di traduzione NAT** ogni coppia di conversione
3. **Sostituire** nuovamente la coppia di tutti i datagrammi in arrivo con la coppia originale, per poi spedire il datagramma al destinatario interno alla rete

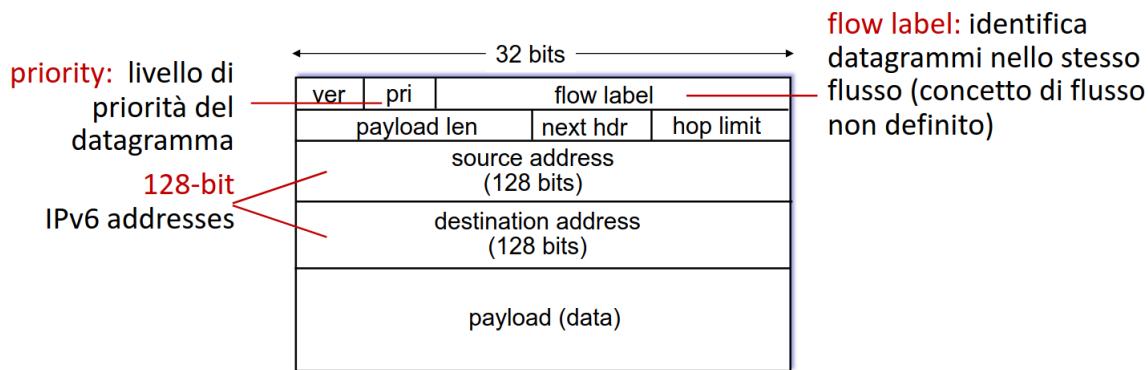


Per via della sua implementazione, l'utilizzo del NAT ha generato molte controversie:

- Per loro definizione stessa, i router dovrebbero processare i pacchetti solo **fino al livello di rete**, mentre per attuare il NAT è necessario che essi adoperino anche il livello di trasporto per modificare le porte dei datagrammi
- La **carenza di indirizzi** può essere risolta anche tramite gli **indirizzi IPv6** (che vedremo in seguito)
- Viene **violata** la modalità di trasmissione **end-to-end**, poiché è necessario manomettere il pacchetto per effettuare le traduzioni

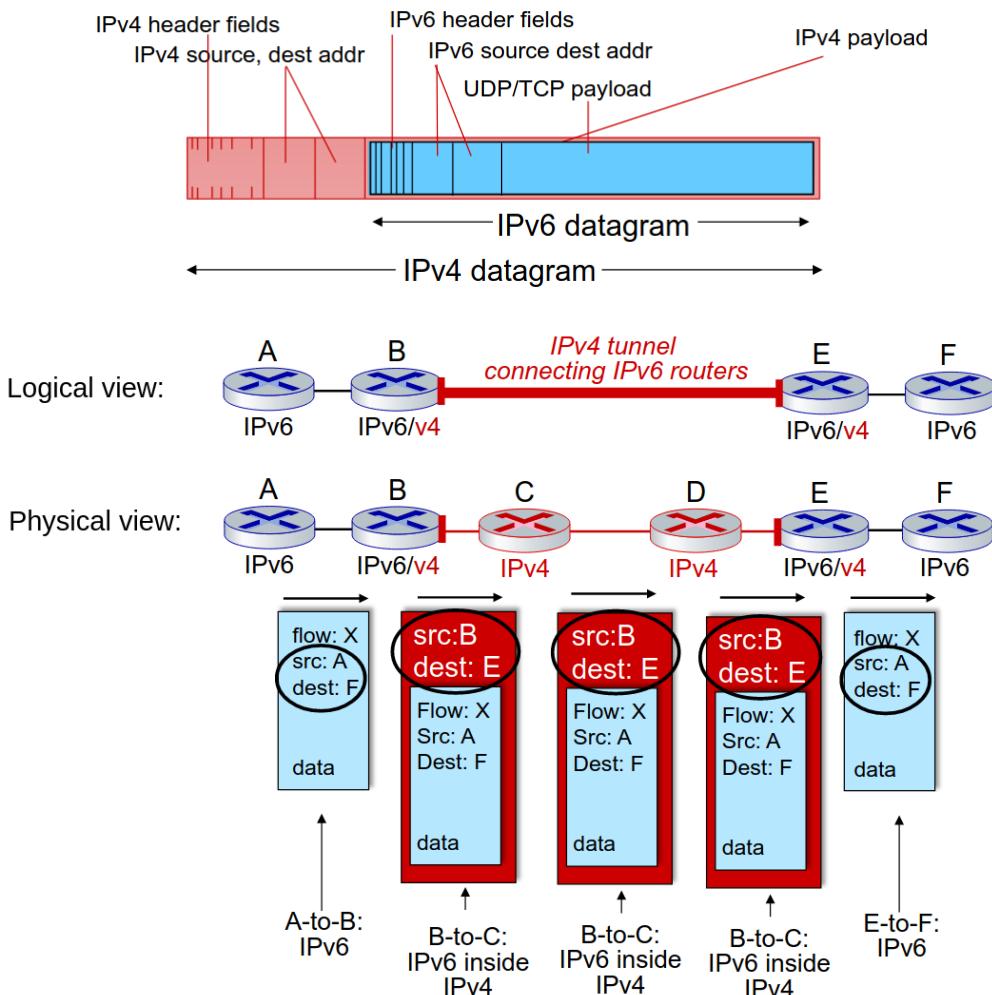
Tuttavia, il NAT risulta ormai essere attualmente ampiamente utilizzato in reti domestiche, istituzionali e cellulari, rendendo **lenta** la sua sostituzione con strumenti più moderni.

In particolare, il principale strumento che (man mano) sostituirà l'uso dell'IPv4 e il NAT è il **protocollo IPv6**, dove vengono utilizzati **128 bit** per gli indirizzi invece di 32 bit e vengono **rimossi** il **checksum** per i datagrammi (non per gli altri livelli superiori) per velocizzare l'elaborazione nei router, la **frammentazione** dei datagrammi e i **campi opzione** (implementabili tramite protocolli superiori).



Poiché non tutti i router possono essere aggiornati contemporaneamente, attualmente vengono utilizzati un **indirizzamento misto**, utilizzando sia l'IPv4 che l'IPv6:

- Per le comunicazioni tra due router IPv4 viene utilizzato direttamente il protocollo IPv4. Analogamente, per le comunicazioni tra due router IPv6 viene utilizzato direttamente il protocollo IPv6
- Per le comunicazioni tra un router IPv6 e un router IPv4 viene utilizzato il **tunneling**, dove un datagramma IPv6 viene trasportato come payload all'interno di un datagramma IPv4 ("datagramma dentro un datagramma").



## 4.4 Protocollo ICMP e Traceroute

### Definition 38. Protocollo ICMP

Il **protocollo Internet Control Message Protocol (ICMP)** è un protocollo a livello di rete utilizzato da host e router per scambiarsi informazioni a livello di rete (es: report degli errori come un host irraggiungibile).

I **messaggi ICMP** hanno un campo **tipo** e un campo **codice**, contenendo l'header e i primi 8 byte del datagramma IP che ha provocato la generazione del messaggio.

Il protocollo ICMP viene considerato "parte" del protocollo IP, nonostante quest'ultimo venga utilizzato da ICMP per inviare i suoi messaggi. Per tale motivo, esso viene considerato come "superiore" a IP all'interno dello stack TCP/IP.

<b>Tipo</b>	<b>Codice</b>	<b>Descrizione</b>
0	0	Risposta echo (a ping)
3	0	Rete destin. irraggiungibile
3	1	Host destin. irraggiungibile
3	2	Protocollo dest. irraggiungibile
3	3	Porta destin. irraggiungibile
3	6	Rete destin. sconosciuta
3	7	Host destin. sconosciuto
4	0	Riduzione (controllo di congestione)
8	0	Richiesta echo
9	0	Annuncio del router
10	0	Scoperta del router
11	0	TTL scaduto
12	0	Errata intestazione IP

Uno dei programmi utilizzante lo scambio di messaggi echo di richiesta e risposta del protocollo ICMP è il **programma ping**, presente su (quasi) ogni dispositivo, utilizzato per calcolare rapidamente il RTT.

**Esempio:**

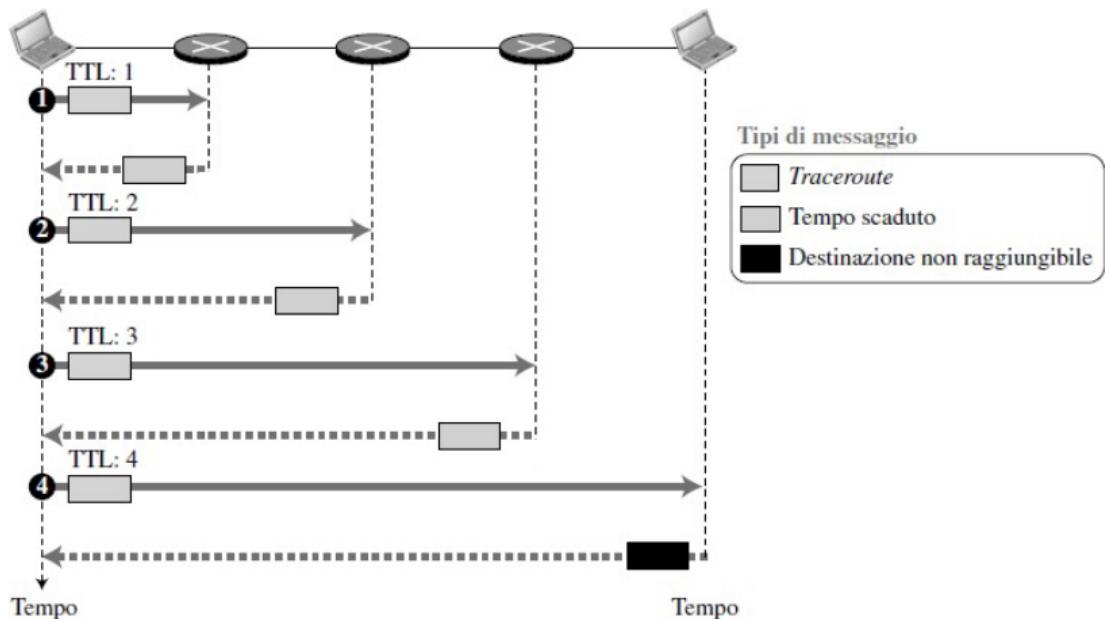
```
$ping google.it

PING google.it (142.250.180.163) 56(84) bytes of data.
64 bytes from mil04s44-in-f3.1e100.net (142.250.180.163):
icmp_seq=1 ttl=114 time=17.0 ms
64 bytes from mil04s44-in-f3.1e100.net (142.250.180.163):
icmp_seq=2 ttl=114 time=16.5 ms
64 bytes from mil04s44-in-f3.1e100.net (142.250.180.163):
icmp_seq=3 ttl=114 time=16.9 ms
--- google.it ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2003ms
rtt min/avg/max/mdev = 16.452/16.800/17.026/0.250 ms
```

Un ulteriore programma utilizzato per vedere il percorso effettuato dal traffico per raggiungere un determinato host è il **programma traceroute**.

Il programma invia una **serie di datagrammi IP** alla destinazione, ciascuno contenente un segmento UDP con un numero di porta inutilizzato. All'interno di ogni datagramma viene inserito un **valore incrementale** (partendo da 1) nel campo header **Time-to-live (TTL)**, corrispondente al numero di router attraversabili prima che il datagramma venga considerato come **scaduto**:

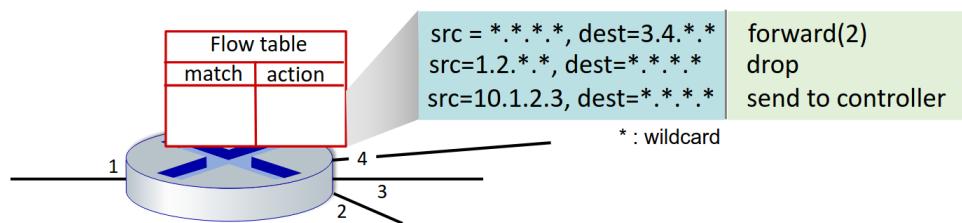
- Per ogni datagramma inviato, il mittente avvia un **timer**
- Se l'*n*-esimo **datagramma** arriva all'*n*-esimo **router**, esso scarterà il datagramma, inviando al mittente un messaggio di allerta ICMP (tipo 11, codice 0), contenente inoltre il nome del router e il suo indirizzo IP. Quando il messaggio ICMP arriverà al mittente, esso calcolerà anche il RTT. Tale processo viene ripetuto per tre volte.
- Se invece il segmento UDP arriva all'**host di destinazione**, esso restituirà un messaggio ICMP segnalando che la porta sia irraggiungibile (tipo 3, codice 3), utilizzato solo come "valore simbolico". Quando l'origine riceverà tale messaggio, verrà arrestato l'invio di dei datagrammi.
- Una volta arrestato l'invio, verranno utilizzati tutti i messaggi di risposta ricevuti per ricostruire il **percorso effettuato**



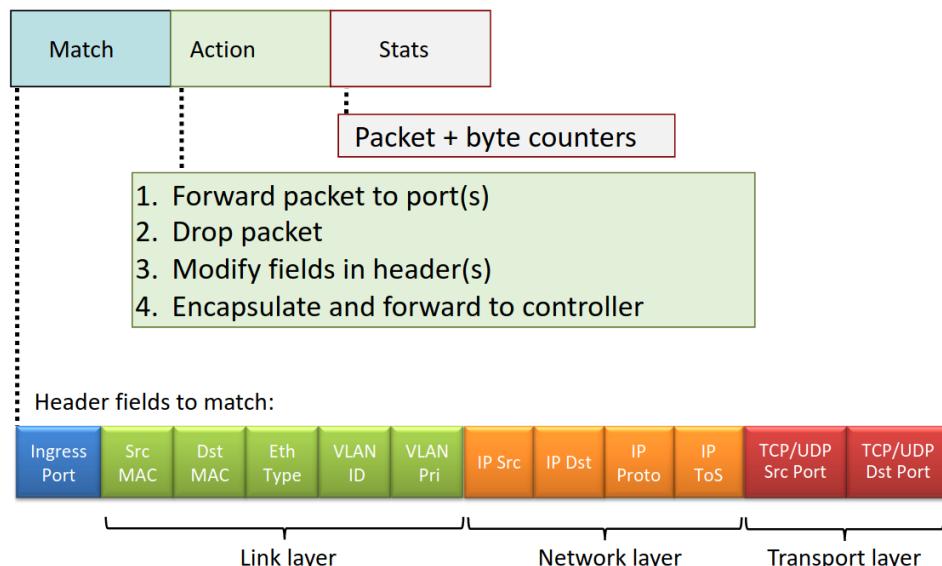
## 4.5 API OpenFlow e forwarding generalizzato

Come già discusso in precedenza, ogni router è dotato di una propria **forwarding table** (anche detta **flow table**). L'uso delle forwarding table può essere astratto tramite il concetto di **match plus action**:

- Ad ogni **match** (ad esempio usando il longest prefix matching), viene eseguita un'**azione**
- Le **azioni** eseguibili consistono in **forward**, **drop**, **modify** e **send to controller**
- Per disambiguare pattern sovrapposti (ossia match multipli), vengono utilizzate **regole di priorità**
- Vengono utilizzati anche dei **contatori** per il numero di byte e il numero di pacchetti



Le **API OpenFlow** consentono l'**accesso al data plane** di un host o router attraverso la rete, venendo utilizzato per dettare le **regole** implementate all'interno delle forwarding table. Ogni regola dettata è composta da un campo **match**, un campo **action** ed un campo **statistics**



Esempi:

1. **Destination-based forwarding:** per inoltrare sulla porta di output 6 tutti i datagrammi IP destinati all'indirizzo IP 51.6.0.8 verrà utilizzata la seguente regola:

Switch Port	MAC src	MAC dst	Eth type	VLAN ID	VLAN Pri	IP Src	IP Dst	IP Prot	IP ToS	TCP s-port	TCP d-port	Action
*	*	*	*	*	*	*	*	51.6.0.8	*	*	*	port6

2. **Firewall:** per bloccare tutti i datagrammi IP la cui porta di destinazione è la porta TCP/22 (corrispondente ad un protocollo non visto, ossia il protocollo SSH) verrà utilizzata la seguente regola:

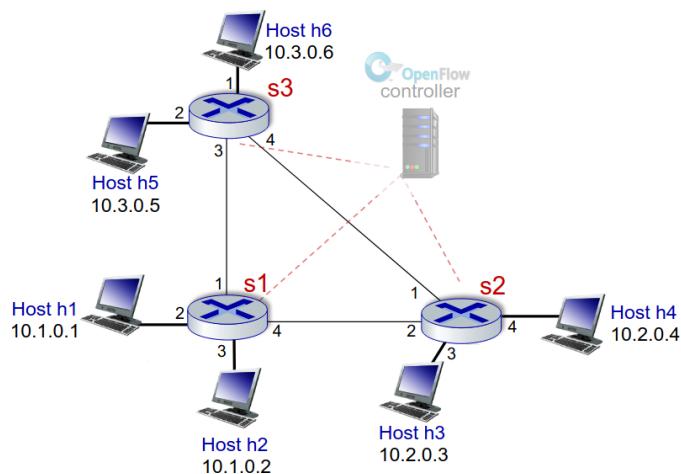
Switch Port	MAC src	MAC dst	Eth type	VLAN ID	VLAN Pri	IP Src	IP Dst	IP Prot	IP ToS	TCP s-port	TCP d-port	Action
*	*	22:A7:23: 11:E1:02	*	*	*	*	*	*	*	*	*	port3

3. **Forwarding a livello di collegamento:** per inoltrare sulla porta di output 3 tutti i datagrammi IP destinati all'indirizzo MAC 22:A7:23:11:E1:02 (vedremo in seguito il protocollo MAC) verrà utilizzata la seguente regola:

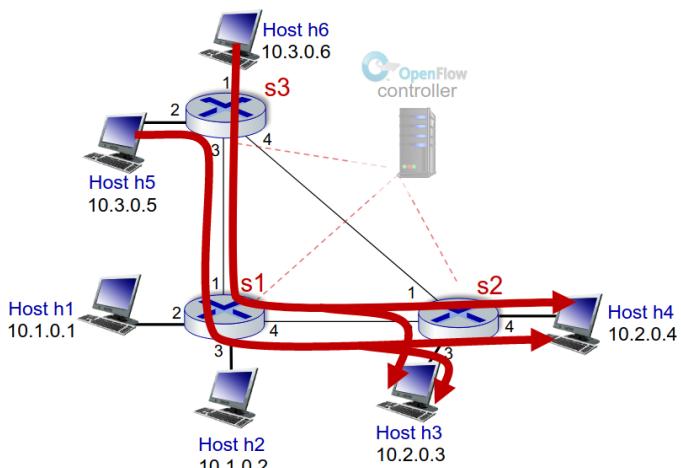
Switch Port	MAC src	MAC dst	Eth type	VLAN ID	VLAN Pri	IP Src	IP Dst	IP Prot	IP ToS	TCP s-port	TCP d-port	Action
*	*	22:A7:23: 11:E1:02	*	*	*	*	*	*	*	*	*	port3

#### 4. Gestione del flusso

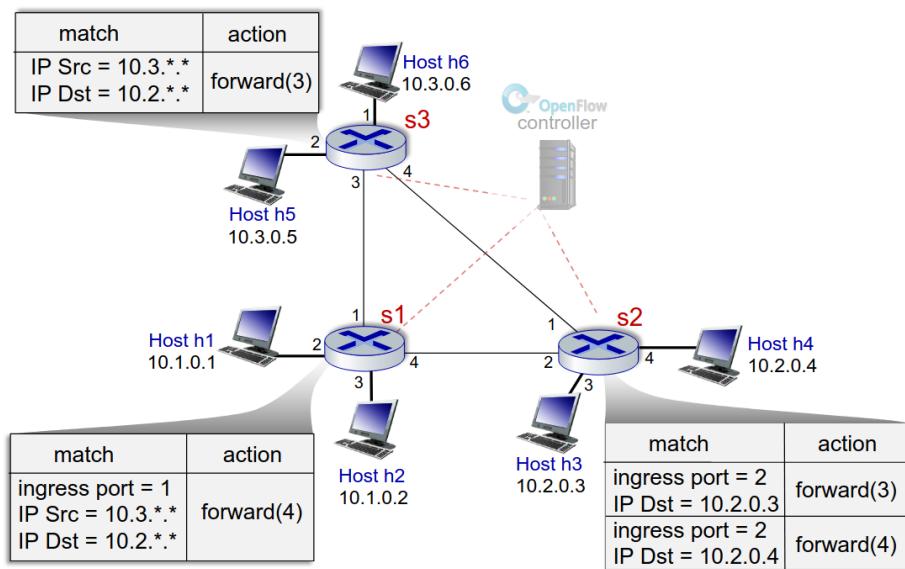
- Consideriamo la seguente rete



- Vogliamo far sì che i datagrammi dagli host h5 e h6 inviati verso gli host h3 e h4 passino prima per il router s1 e poi per il router s2



- Per ottenere tale flusso, impostiamo le seguenti flow table all'interno dei router



## 4.6 Principi architetturali di Internet

### Definition 39. Middleboxes

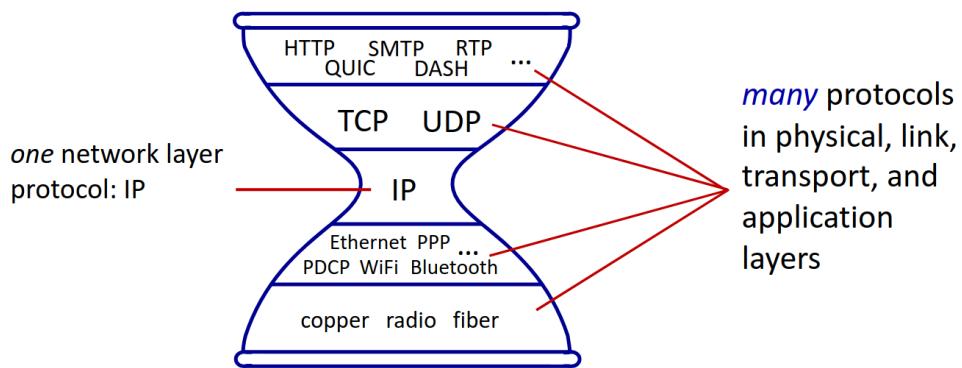
Un **middlebox** è un qualsiasi dispositivo **intermediario** tra mittente e destinatario che esegue **funzioni diverse** dalle normali funzioni standard di un router (es: NAT, Firewall, Cache servers, Load balancers, ...)

### Proposition 4. Principi architetturali di Internet

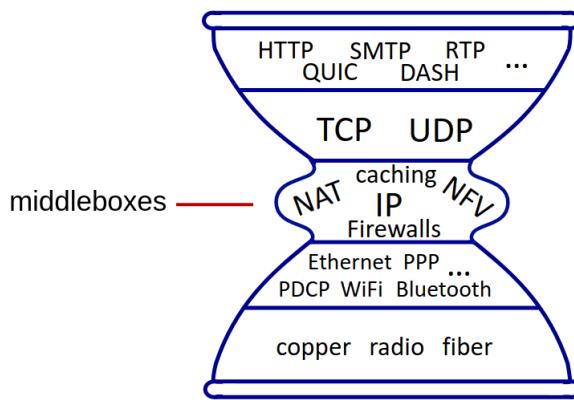
Come dettato all'interno del documento RFC 1958, non vi è una vera e propria architettura standard per Internet, bensì solamente delle "**tradizioni**". In termini generici, il servizio Internet è basato su tre principi fondamentali:

- **Connettività semplice**, rendendo il servizio facile da implementare nella maggior numero di dispositivi possibili
- Mantenere la **clessidra TCP/IP** con il **minor girovita possibile**, cercando di ridurre al minimo il numero di servizi svolti dal livello di rete, aumentando la quantità verso i livelli superiori e inferiori
- Complessità ed intelligenza (ossia lo svolgimento delle operazioni, il mantenimento dei dati, ...) deve essere implementata sulla **periferia della rete** utilizzando il **principio end-to-end**.

### Clessidra TCP/IP ottimale



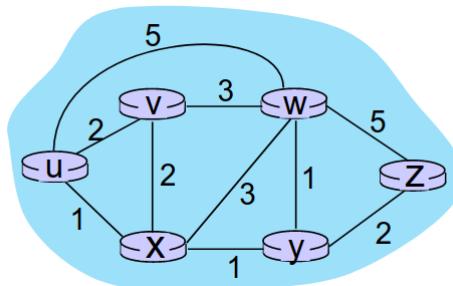
### Clessidra TCP/IP dopo 40 anni



## 4.7 Algoritmi di instradamento

Gli **algoritmi di instradamento** vengono utilizzati per determinare il **percorso migliore**, ossia una sequenza di router che i pacchetti devono attraversare, da una sorgente ad una destinazione.

Per determinare tali percorsi, la rete viene modellata come un **grafo**  $G = (N, E)$  i cui vertici  $V(G) = \{v_1, \dots, v_n\}$  corrispondono ai singoli router e/o host e gli archi  $E(G)$  corrispondono ai collegamenti tra tali dispositivi. Ad ogni arco  $(u, v) \in E(G)$  viene attribuito un **costo** (o **peso**) il quale può essere dettato da **più valori** (maggior velocità, minore congestione, ...).



Gli algoritmi di routing vengono classificati in:

- **Statici**, ossia determinanti percorsi poco soggetti al cambiamento, oppure **dinamici**, ossia determinanti percorsi soggetti ad un aggiornamento periodico in risposta a variazioni dei costi
- **Globali**, dove al termine dell'algoritmo tutti i router conoscono completamente la topologia e il costo dei link della rete, oppure **decentralizzati**, dove lo scambio di informazioni avviene tra router vicini che non conoscono l'intero stato della rete

Per le successive sezioni, utilizzeremo la seguente **notazione**:

- $c_{x,y}$  è il **costo del link diretto** tra i nodi  $x$  e  $y$ . Viene posto uguale a  $\infty$  se tale link diretto non esiste.
- $D(v)$  è la **stima corrente** del costo del percorso a minor costo dalla nodo sorgente al nodo destinazione  $v$
- $p(v)$  è il **nodo predecessore** lungo il percorso dal nodo sorgente al nodo  $v$

### 4.7.1 Algoritmo link-state di Dijkstra

L'algoritmo **link-state** è un algoritmo **dinamico globale** basato sull'**algoritmo di Dijkstra**. Dato un **nodo sorgente**  $u \in V(G)$ , per ogni nodo  $u \neq v \in V(G)$  viene calcolato il percorso a **distanza minore** (ossia di minor costo) da un **nodo sorgente** a tutti gli altri nodi della rete, per poi fornire una forwarding table alla sorgente in base ai percorsi calcolati.

---

**Algorithm 1:** Algoritmo Link-State (basato su Dijkstra)

---

**Function** linkStateDijkstra( $G, u$ ):

```

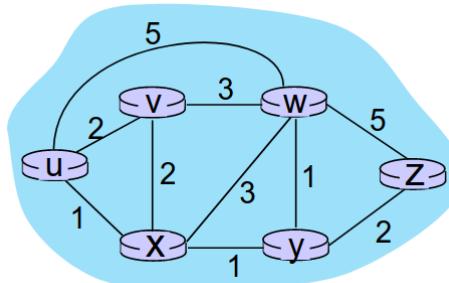
 $R = \{u\};$ 
for  $v \in V(G)$  do
    if  $\exists(u, v) \in E(G)$  then
         $| D(v) = c_{u,v};$ 
    else
         $| D(v) = \infty;$ 
    end
end
while  $R \neq V(G)$  do
     $w := \arg \min_{w \in V(G) - R} [D(w)];$ 
     $R.\text{add}(w);$ 
    for  $x \in V(G) - R$  do
        if  $\exists(w, x) \in E(G)$  then
             $| D(x) = \min(D(x), D(w) + c_{w,x});$ 
             $| p(x) = w;$ 
        end
    end
end

```

---

**Esempio:**

- Consideriamo la seguente rete



- Vogliamo calcolare la forwarding table del nodo sorgente  $u$ . Inizializziamo quindi la tabella delle distanze dalla sorgente  $u$  verso ogni nodo, ponendo la distanza di ogni nodo adiacente a  $u$  pari al costo del link diretto e pari a  $\infty$  per ogni altro nodo

Step	$R$	$D(v), p(v)$	$D(w), p(w)$	$D(x), p(x)$	$D(y), p(y)$	$D(z), p(z)$
0	{u}	2, u	5, u	1, u	$\infty$	$\infty$

- A questo punto, consideriamo il nodo avente distanza minore dal nodo attualmente analizzato (ossia  $u$ ), corrispondente al nodo  $x$ . Per ogni nodo  $a$  adiacente a  $x$  non ancora analizzato (ossia non in  $R$ ), poniamo la nuova distanza calcolata pari al minimo tra la distanza dalla sorgente ad  $x$ , ossia  $D(x)$ , e la somma tra la distanza tra la sorgente ed  $a$  e il costo del link tra  $a$  ed  $x$ , ossia  $D(a) + c_{a,x}$

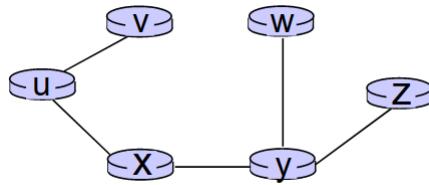
$R$	$D(v), p(v)$	$D(w), p(w)$	$D(x), p(x)$	$D(y), p(y)$	$D(z), p(z)$
{u}	2, u	5, u	1, u	$\infty$	$\infty$
{u, x}	2, u	4, x		2, x	$\infty$

- Proseguendo analogamente, le distanze finali calcolate saranno pari a:

$R$	$D(v), p(v)$	$D(w), p(w)$	$D(x), p(x)$	$D(y), p(y)$	$D(z), p(z)$
{u}	2, u	5, u	1, u	$\infty$	$\infty$
{u, x}	2, u	4, x		2, x	$\infty$
{u, x, y}	2, u	3, y			4, y
{u, x, y, v}		3, y			4, y
{u, x, y, v, w}					4, y
{u, x, y, v, w, z}					

- Una volta ottenute le distanze finali, verrà costruita la forwarding table di  $u$ :
  - I nodi  $v$  e  $x$  sono direttamente raggiungibili da  $u$  con distanza minima
  - Il nodo  $y$  è raggiungibile tramite  $x$  con distanza minima

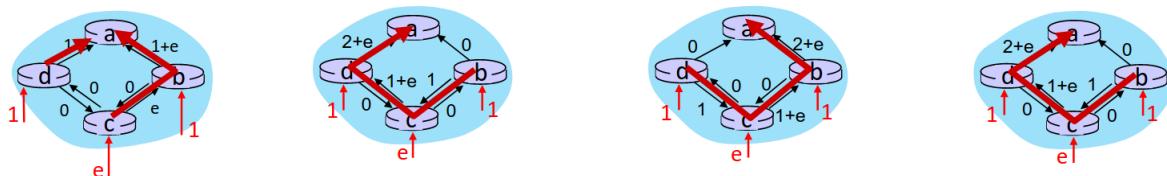
- I nodi  $w$  e  $z$  sono raggiungibili tramite  $y$  con distanza minima, necessitando dunque di passare anche per  $x$



Forwarding table di $u$	
Destinazione	Link di uscita
$v$	$(u, v)$
$x$	$(u, x)$
$y$	$(u, x)$
$w$	$(u, x)$
$z$	$(u, x)$

L'algoritmo link-state ha una **complessità computazionale** pari a  $O(n^2)$  (anche se è possibile implementarlo ottimamente in  $O(n \log n)$ ) ed una **complessità di comunicazione** pari a  $O(n^2)$ , poiché ogni router deve trasmettere in broadcast il suo stato dei costi a tutti gli altri router (richiedendo  $O(n)$  tramite algoritmi efficienti).

Inoltre, poiché i costi dei link dipendono dal volume di traffico, può verificarsi un **caso patologico** per via delle **oscillazioni del percorso**, richiedendo di essere costantemente ricalcolati.



## 4.7.2 Algoritmo Distance-vector

L'algoritmo **distance-vector** è un algoritmo **dinamico decentralizzato** basato sulla **equazione di Bellman-Ford**.

### Theorem 5. Equazione di Bellman-Ford

Dati  $x, y \in V(G)$ , sia  $D_x(y)$  la **distanza minima da  $y$  ad  $x$** .

In tal caso, si ha che:

$$D_x(y) = \min_{v \in V(G)} [c_{x,v} + D_v(y)]$$

Al verificarsi di un determinato **evento** (es: lo scadere di un timer), ogni nodo invia ai propri vicini la propria **stima del distance-vector**, ossia un vettore contenente le distanze verso tutti i nodi della rete. Quando un nodo riceve una stima da parte di un vicino, utilizza tale stima per aggiornare il proprio distance-vector tramite l'**equazione di Bellman-Ford**. Sotto determinate condizioni ottimali, la distanza stimata **converge** dopo un determinato numero di interazioni alla **distanza minima**.

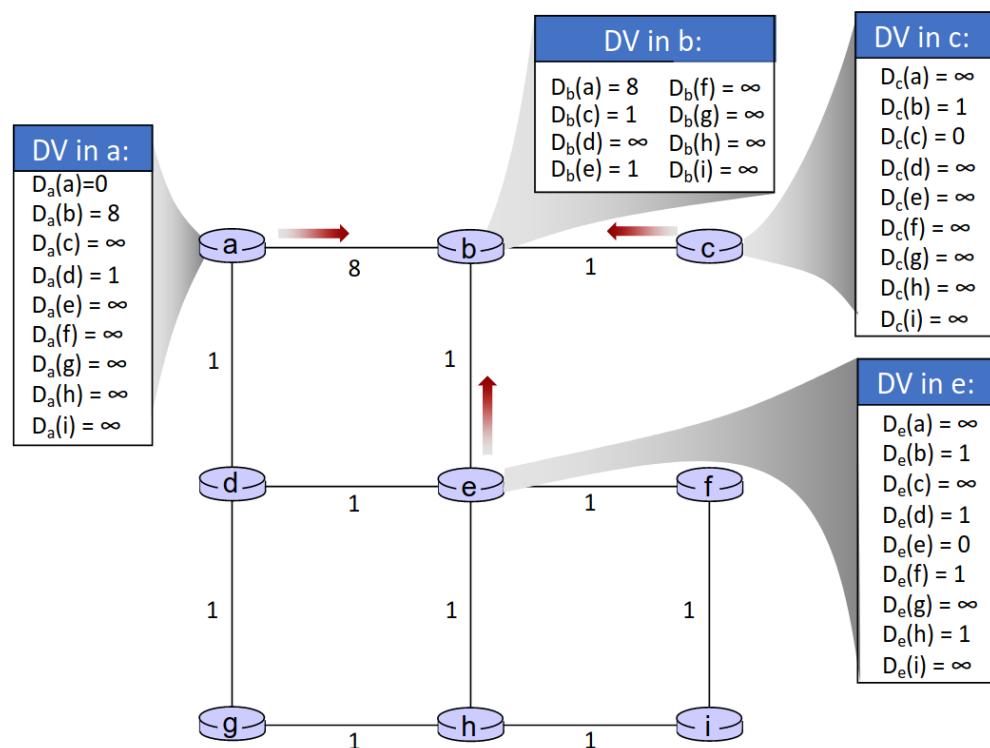
Solitamente, l'aggiornamento locale del vettore di un nodo viene effettuato solo a seguito dell'**aggiornamento** del costo di un **link diretto** da tale nodo verso un suo vicino o a seguito della **ricezione di un vettore aggiornato** inviato da un vicino (**Iterativo ed asincrono**). Inoltre, ogni nodo invia il proprio distance-vector ai vicini solo quando esso viene aggiornato (**Distribuito, self-stopping e responsive**).

Dunque, l'algoritmo è riassumibile nei seguenti passi:

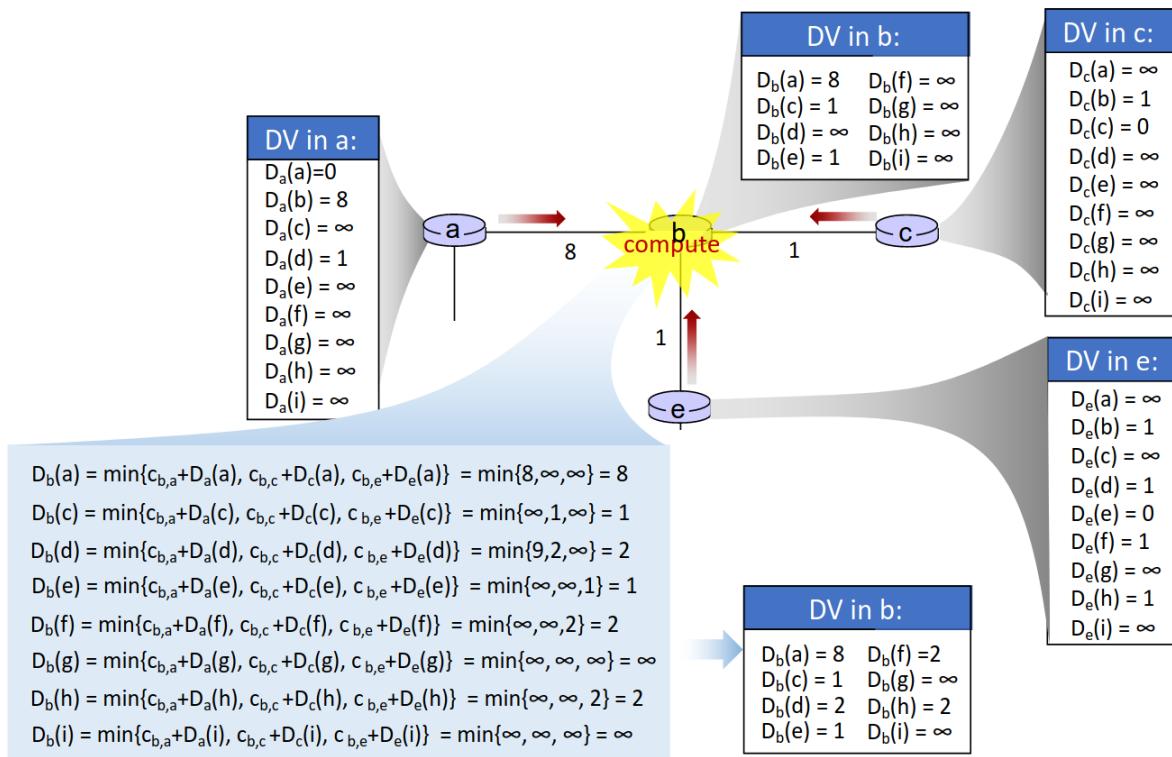
1. **Inizializzazione:** il DV di ogni nodo contiene il costo diretto verso tutti i suoi vicini e  $\infty$  per ogni altro nodo
2. **Attesa dell'evento:** ogni nodo attende il cambio di un costo diretto locale o la ricezione del vettore di un vicino
3. **Ricalcolo del DV:** se l'evento viene attivato per un nodo, esso ricalcola il proprio DV utilizzando i valori precedenti e quelli ricevuti
4. **Invio solo se modificato:** se al termine del calcolo il DV del nodo è stato aggiornato, esso viene inviato ai vicini del nodo
5. Viene ripetuto il tutto in loop tornando al passo 2

**Esempio:**

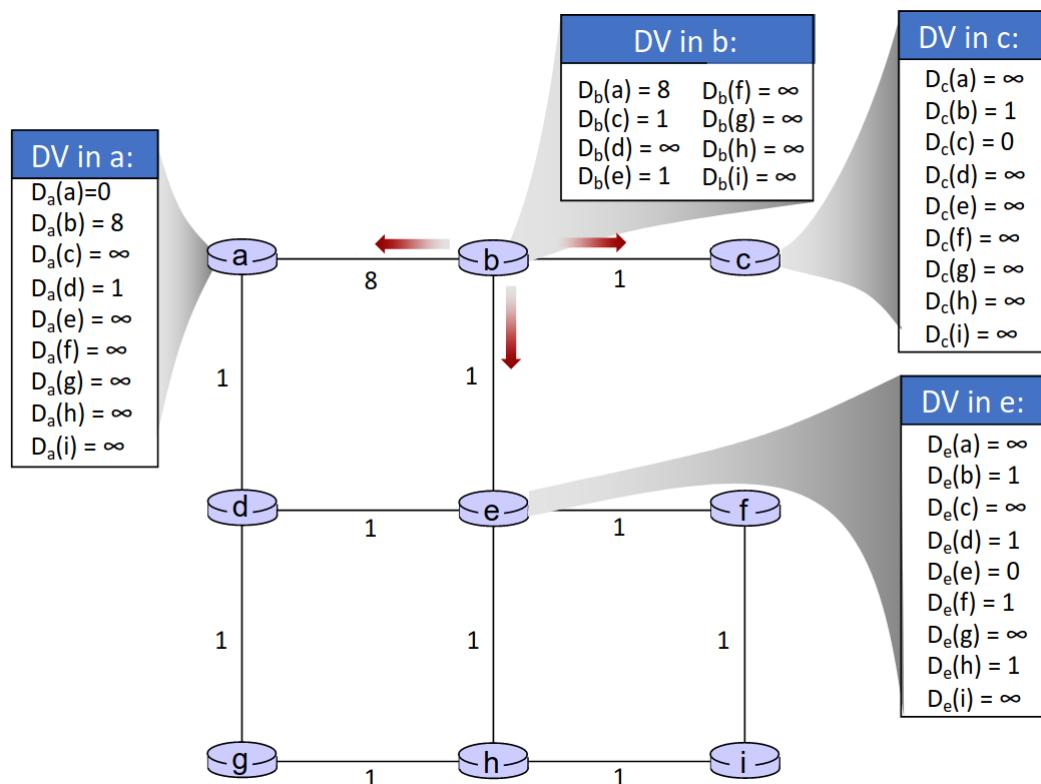
- Supponiamo che i router *a*, *c* ed *e* inviano il proprio DV al router *b*



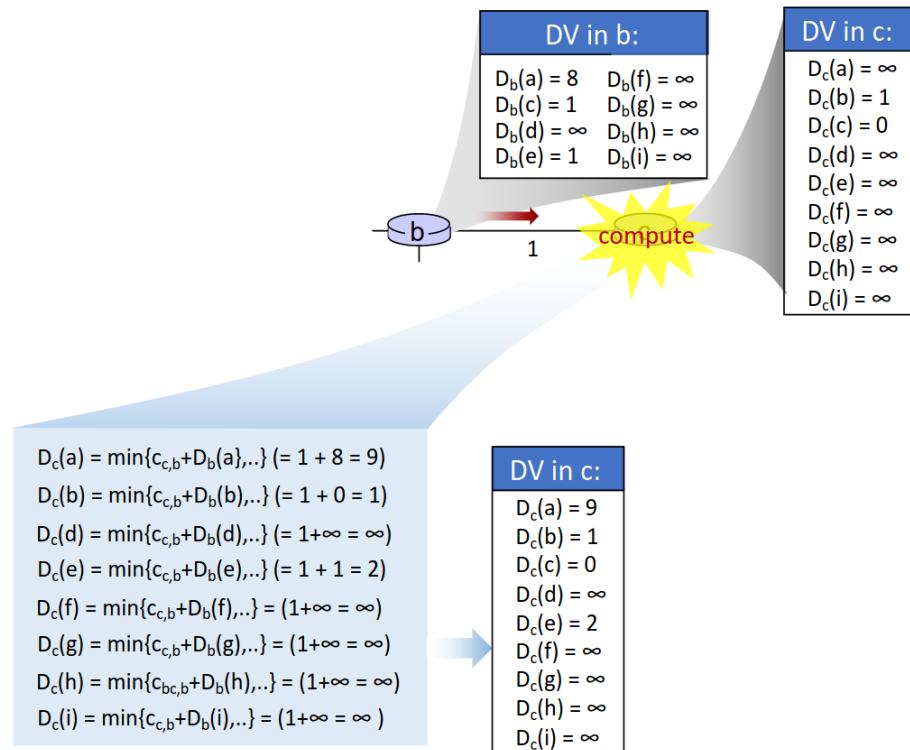
- Una volta ricevuti i vettori, il router *b* ricalcola il proprio DV utilizzando l'equazione di Bellman-Ford



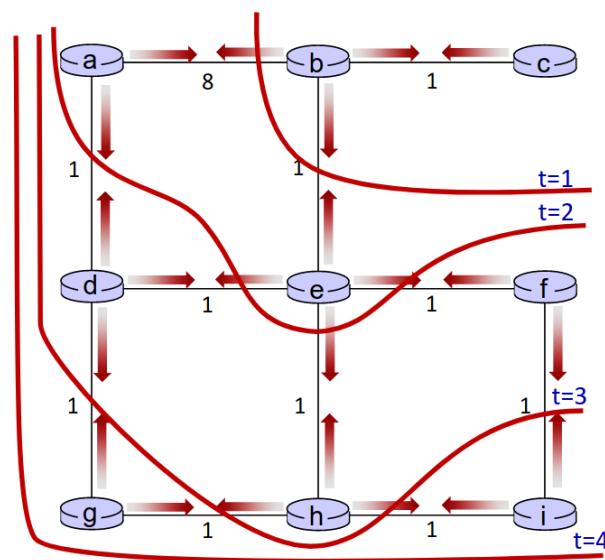
- Successivamente, il nuovo DV viene inviato ai vicini del router *b*



- Una volta ricevuto il DV di  $b$ , il router  $c$  (e anche i router  $a$  ed  $e$ ) procederà a ricalcolare il proprio DV utilizzando le nuove distanze ricevute

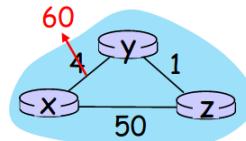


- Al passare degli istanti di tempo, i cambiamenti effettuati all'istante  $t = 0$  verranno propagati su tutti gli altri router della rete



Come l'algoritmo link-state, anche l'algoritmo distance-vector è soggetto a **comportamenti patologici**, in particolare il **conteggio all'infinito**:

1. Data la seguente rete, supponiamo che il costo del link  $(x, y)$  venga modificato da 4 a 60



2. Di conseguenza, il router  $y$  nota il nuovo costo del collegamento diretto verso  $x$  sia 60. Tuttavia, il nodo  $y$  ha precedentemente ricevuto il DV del router  $z$ , venendo a sapere che tramite  $z$  sia possibile raggiungere  $x$  con un costo pari a 6, aggiornando quindi il proprio DV ed inviandolo ai suoi vicini

$$D_y(x) = 4 \quad \xrightarrow{\text{diventa}} \quad D_y(x) = c_{y,z} + D_z(x) = 1 + 5 = 6$$

**(Attenzione:** è necessario ricordare che l'algoritmo distance-vector è decentralizzato dunque il vertice  $y$  non sa che il percorso da  $z$  a  $x$  passi per  $y$  stesso)

3. Successivamente, il vertice  $z$  riceverà il DV di  $y$ , notando che la distanza del percorso da  $y$  a  $x$  tramite cui  $z$  possa raggiungere  $x$  è stato modificato, aggiornando quindi il proprio DV ed inviandolo ai vicini

$$D_z(x) = 5 \quad \xrightarrow{\text{diventa}} \quad D_z(x) = c_{z,y} + D_z(y) = 1 + 6 = 7$$

4. Analogamente,  $y$  riceverà il DV di  $z$ , ricadendo nella stessa casistica

$$D_y(x) = 6 \quad \xrightarrow{\text{diventa}} \quad D_y(x) = c_{y,z} + D_z(x) = 1 + 7 = 8$$

5. ...

### Proposition 6. Soluzioni al conteggio all'infinito

Per risolvere il comportamento patologico del **conteggio all'infinito**, l'algoritmo DV adotta due politiche aggiuntive:

- **Split horizon**, dove, invece che inviare l'intera tabella attraverso ogni interfaccia, ogni nodo invia solo una porzione della propria tabella a seconda dell'interfaccia (es: se il nodo  $x$  riceve il DV del nodo  $y$ , nel DV di  $x$  aggiornato inviato verso  $y$  verranno omesse le informazioni ricevute da  $y$ )
- **Poisoned reverse**, dove, durante l'invio del proprio DV, il nodo mittente pone a  $\infty$  la distanza dei percorsi passanti attraverso il vicino a cui sta inviando il nuovo DV (es: se il nodo  $x$  deve inviare il suo DV al nodo  $y$  e un percorso per di  $x$  verso un nodo  $z$  passa per  $y$ , nel DV inviato viene posto  $D_x(z) = \infty$ )

A differenza dell'algoritmo link-state, l'algoritmo distance-vector possiede una **complessità di comunicazione** pari a  $O(n)$ . Per quanto riguarda la **velocità di convergenza**, invece, l'algoritmo LS necessita di  $O(n^2)$  computazioni (a meno di oscillazioni) mentre l'algoritmo DV richiede condizioni troppo ottimali (l'instradamento potrebbe divenire ciclico).

Inoltre, per propria natura stessa, l'algoritmo DV risulta essere **meno robusto**:

- Supponiamo che un router subisca un malfunzionamento o un attacco esterno, peggiorando notevolmente il costo dei propri link diretti
- Nell'algoritmo LS, tale router pubblicherà un **costo errato dei link diretti**. Tuttavia, poiché ogni router calcola solamente la propria tabella, gli altri router non verranno influenzati.
- Nell'algoritmo DV, invece, tale router pubblicherà un **costo errato dei percorsi (black-holing)**. Di conseguenza, poiché ogni altro router userà il DV di tale router per i calcoli, l'errore verrà **propagato sull'intera rete**

## 4.8 Instradamento intra-AS e inter-AS

Gli algoritmi di routing precedentemente visti si basano su una *concezione irrealistica* della rete, poiché viene assunto che tutti i router siano identici e che non vi sia alcuna gerarchia al suo interno, oltre all'evidente problema di scala dovuto alle miliardi di destinazioni che porterebbe ad intasare la rete con messaggi di scambio di forwarding table.

### Proposition 7. Instradamento intra-AS e inter-AS

Per risolvere tali problematiche, i router vengono **aggregati** in regioni note come **autonomous systems (AS)** o **domini**. Ogni AS costituisce una rete composta da soli router.

Distinguiamo quindi due tipologie di instradamento:

- **Instradamento intra-AS** ossia instradamento all'interno di un AS, dove tutti i router all'interno dell'AS devono eseguire lo stesso protocollo di instradamento intra-AS, implicando che router di diversi AS possano scegliere il proprio protocollo.
- **Instradamento inter-AS**, ossia instradamento tra diversi AS, dove ogni AS possiede un gateway router posto sul bordo e connesso ai gateway router degli altri AS (i gateway partecipano comunque all'instradamento intra-AS)

Le **forwarding table**, dunque, verranno configurate sia da algoritmi di instradamento intra-dominio sia da algoritmi di instradamento inter-dominio.

Di conseguenza, ogni router deve essere in grado di apprendere quali destinazioni siano raggiungibili tramite gli AS esterni e propagare tali informazioni all'interno del proprio AS.

### 4.8.1 Protocolli RIP e OSPF

#### Definition 40. Protocollo RIP

Il protocollo Routing Information Protocol (RIP) è un protocollo di **instradamento intra-AS** basato sull'**algoritmo distance-vector**.

La metrica di costo utilizzata è la **distanza misurata in hop**, ossia il numero di router necessari da attraversare per raggiungere la destinazione. Il **valore massimo** per tale metrica è **15 hop** (il valore 16 corrisponde a  $\infty$ )

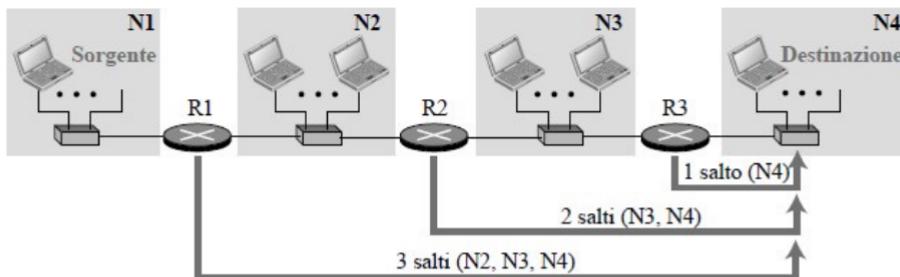


Tabella d'inoltro per R1

Rete di destinazione	Prossimo router	Costo (in hop)
N1	—	1
N2	—	1
N3	R2	2
N4	R2	3

Tabella d'inoltro per R2

Rete di destinazione	Prossimo router	Costo (in hop)
N1	R1	2
N2	—	1
N3	—	1
N4	R3	2

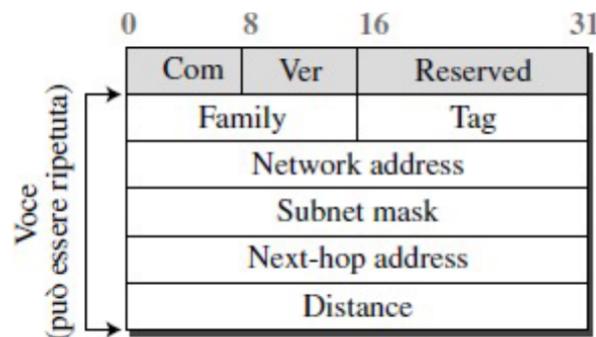
Tabella d'inoltro per R3

Rete di destinazione	Prossimo router	Costo (in hop)
N1	R2	3
N2	R2	2
N3	—	1
N4	—	1

Periodicamente, dunque dopo un **prefissato lasso di tempo**, ogni router utilizzante l'algoritmo RIP invierà il proprio distance-vector assieme ad alcune informazioni aggiuntive, fornendo agli altri router informazioni sugli host e le altre reti (ossia gli altri AS) raggiungibili.

Se all'interno del DV di un router  $x$  vi è un'entrata indicante la possibilità di raggiungere la rete  $A$  con un costo pari a  $N$  hop, ogni altro router all'**interno della rete** di  $x$  saprà di poter raggiungere la rete  $A$  con un costo pari a  $N + 1$  passando tramite  $x$ .

Ogni messaggio può contenere **più voci**, ognuna di esse corrispondenti ad un'entrata del distance-vector del router mittente.



I messaggi di **RIP request** vengono inviati dai router al momento della loro immissione all'interno di un AS oppure a fini diagnostici (es: per richiedere una voce specifica).

Per quanto riguarda i messaggi di **RIP response**, invece, essi vengono inviati in risposta ad un messaggio di richiesta (**solicited response**) o a seguito dello scadere di un timer di 25-35 secondi (**unsolicited response**). Oltre a tale timer periodico, vengono utilizzati due ulteriori timer:

- Un **timer di scadenza** (150-210 secondi), dove se allo scadere del tempo non è stato ricevuto alcun aggiornamento per un percorso, esso viene considerato come scaduto, venendo posto a 16 (dunque a  $\infty$ ).

Se un router non riceve messaggi da un suo vicino per circa 180 secondi (media tra 150 e 210), tale vicino viene considerato **spento o guasto**, impostando il costo di tale percorso a 16 e propagando l'informazione sugli altri nodi della rete.

- Un **timer per il garbage collection** (120 secondi), dove se allo scadere del tempo il router continua ad annunciare un percorso con costo pari a 16, tale percorso viene completamente rimosso

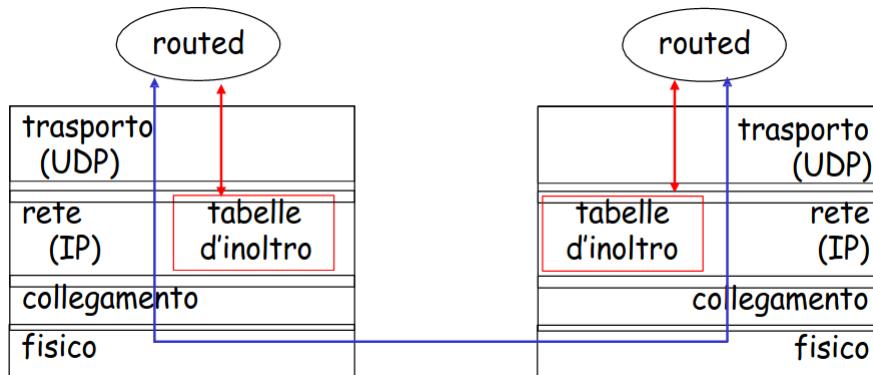
Essendo basato sull'algoritmo distance-vector, il protocollo RIP presenta gli stessi comportamenti patologici, i quali vengono mediati utilizzando sia lo **split horizon** sia il **poisoned reverse**. Inoltre, quando viene ricevuta un'informazione da una rotta non più valida (dunque posta a 16), viene avviato un timer e tutti i messaggi arrivati prima del timeout e riguardanti tale rotta vengono ignorati (**hold-down**)

#### Observation 4

Il **protocollo RIP** viene implementato tramite un processo a livello di applicazione chiamato **routed (route daemon)**, il quale utilizza **protocollo UDP** sulla **porta 520** per l'invio dei messaggi.

Per tale motivo, seppur considerato un protocollo al livello di rete, sarebbe più corretto considerare il protocollo RIP come un protocollo a livello di applicazione.

Tuttavia, è necessario sottolineare che l'utilizzo del protocollo UDP non sia necessario, bensì solo una comodità a livello di implementazione.



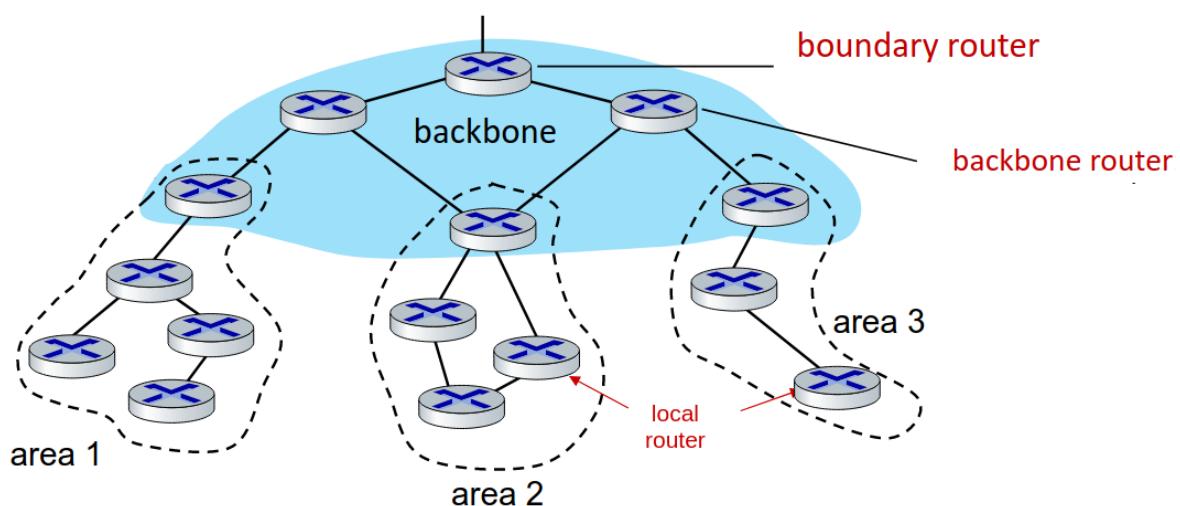
### Definition 41. Protocollo OSPF

Il protocollo Open Shortest Path First (OSPF) è un protocollo open-source di instradamento intra-AS basato sull'**algoritmo link-state**.

Per gestire il costo dei singoli link, vengono utilizzate **più metriche possibili** (es: larghezza di banda, ritardo, ...). Inoltre, tutti i messaggi OSPF sono **autenticati** per prevenire intrusioni.

Il protocollo OSPF utilizza una **gerarchia a due livelli**, composta da una **backbone** e varie **local area**:

- Gli annunci link-state vengono propagati solo nella backbone o all'interno di un'area locale, riducendo la quantità di messaggi in base alla gerarchia
- Ogni nodo conosce, a seconda di dove si trova, la **topologia dettagliata** della propria area o del backbone, mentre conosce solo la **direzione** necessaria per raggiungere le altre destinazioni
- I router vengono distinti in quattro tipologie:
  - **Backbone router**, situato all'interno del backbone, esegue la propagazione solo all'interno del backbone
  - **Local router**, situato all'interno di un'area locale, esegue la propagazione solo all'interno dell'area stessa
  - **Boundary router**, ossia il backbone router tramite cui l'intero AS si connette ad altri AS (gateway router)
  - **Area border router**, situato sia nel backbone sia all'interno di un'area locale (punto di scambio), "riepiloga" le distanze verso le altre destinazioni nella propria area e le pubblica nel backbone



#### 4.8.2 Protocollo BGP

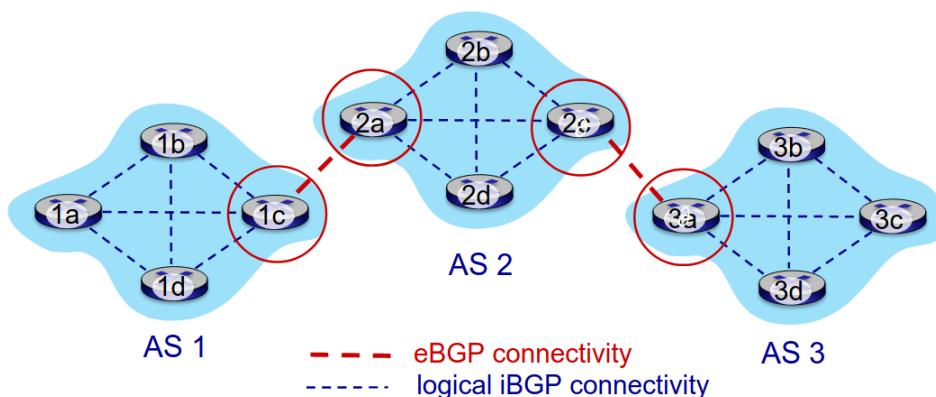
#### Definition 42. Protocollo BGP

Il protocollo **Border Gateway Protocol (BGP)** è un protocollo di **istradamento inter-AS** basato sull'algoritmo path-vector (non visto precedentemente).

Consente ad un'AS di pubblicizzare alle altre AS la propria esistenza e le destinazioni che essa può raggiungere.

Il protocollo BGP fornisce due tipologie di connettività ad ogni AS:

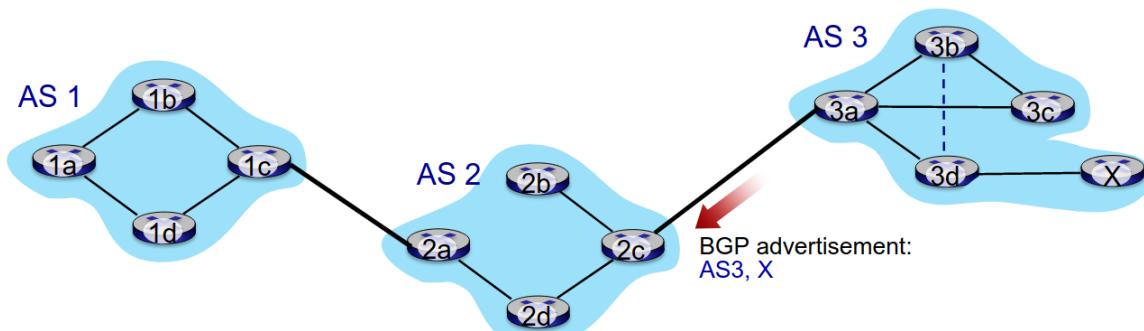
- **eBGP (external BGP)**, permettendo di ottenere informazioni sulla raggiungibilità di una sottorete tramite gli AS vicini
  - **iBGP (internal BGP)**, permettendo di propagare tali informazioni sulla raggiungibilità a tutti i router interni all'AS



All'interno di una **sessione BGP**, due router BGP (detti *peer*) si scambiano messaggi BGP su una connessione TCP semipermanente, pubblicizzando i percorsi verso diversi prefissi di rete di destinazione.

Esempio:

- Quando il gateway 3a di AS3 annuncia il percorso AS3, X al gateway 2c di AS2, l'AS3 promette ad AS2 di inoltrare tutti i datagrammi diretti verso X



I messaggi BGP possono essere di quattro tipologie:

- **OPEN**, dove viene aperta una sessione BGP tra due peer BGP, autenticando prima il peer che apre la connessione
- **UPDATE**, dove viene pubblicizzato un nuovo percorso o ritirato uno precedente
- **KEEPALIVE**, dove viene richiesto di mantenere viva la connessione in assenza di messaggi UPDATE (viene utilizzato anche come ACK per il messaggio OPEN)
- **NOTIFICATION**, dove vengono segnalati errori nei messaggi precedenti (viene utilizzato anche per chiudere la sessione)

Per quanto riguarda i **percorsi BGP pubblicizzati**, essi sono composti da un prefisso ed una serie di attributi:

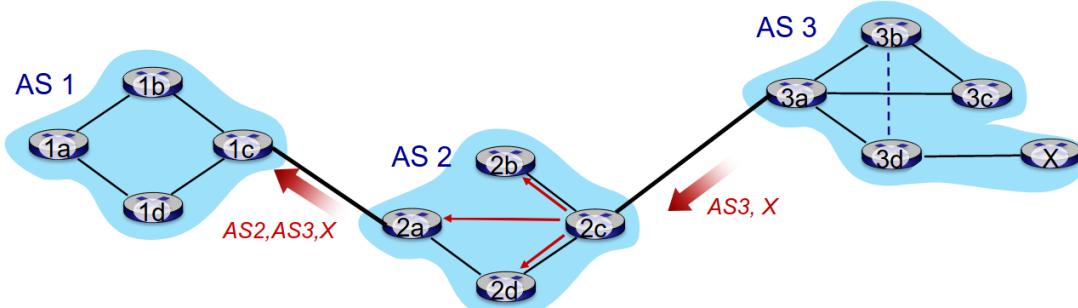
- Il **path prefix**, ossia la destinatario che viene pubblicizzata dall'AS mittente (in formato *ciderized*)
- L'attributo **AS-PATH**, contenente l'elenco di AS attraverso cui è passato l'annuncio, utilizzato dall'AS destinatario come percorso per raggiungere l'AS mittente
- L'attributo **NEXT-HOP**, contenente l'indirizzo dell'hop successivo dell'AS destinatario per poter raggiungere il gateway, ossia l'**egress router**, (es: corrispondente al router 2c nell'esempio precedente)

Alla ricezione di un percorso pubblicizzato, il router destinatario sceglie se **accettare** o **rifiutare** il percorso in base ad una propria politica (**policy-based routing**). Ad esempio, una politica di routing potrebbe essere basata sul rifiutare qualsiasi percorso passante attraverso un determinato AS o un determinato paese (ulteriore utilizzo dell'attributo AS-PATH).

Se un percorso viene accettato, esso viene **propagato** all'interno dell'AS, affinché i router interni ne siano a conoscenza, e verso le **altre AS raggiungibili**.

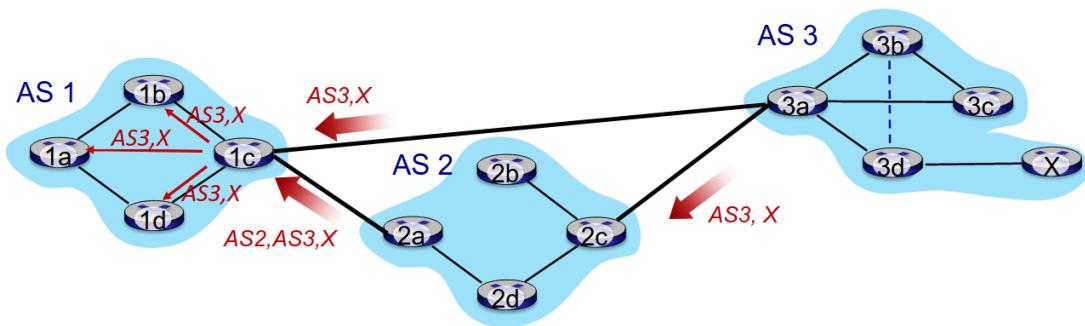
**Esempi:**

1. Consideriamo la seguente situazione:



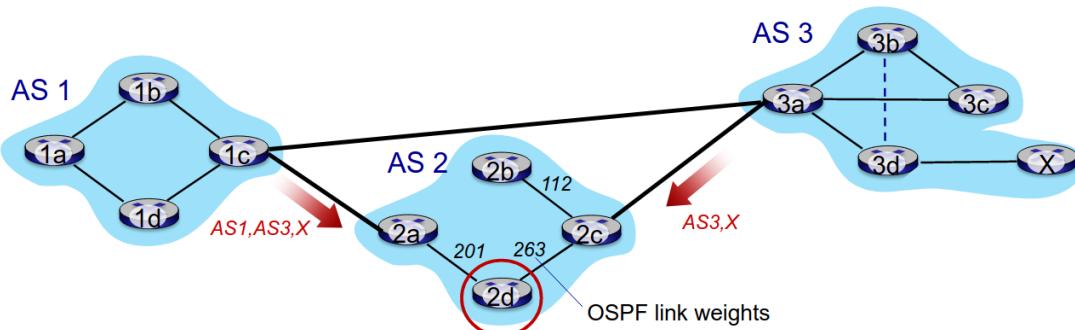
- Il router 2c di AS2 riceve (tramite eBGP) l'annuncio del percorso AS3, X dal router 3a di AS3.
- Basandosi sulla propria politica, il router 2c accetta il percorso e lo propaga (tramite iBGP) all'interno del proprio AS
- Successivamente, basandosi sempre sulla propria politica, il router 2c annuncia (tramite eBGP) il percorso AS2, AS3, X al router 1c di AS1

2. Consideriamo la seguente situazione (separata dalla precedente):



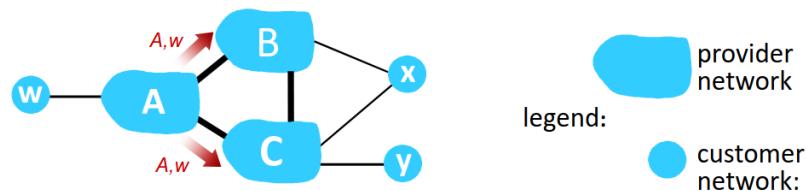
- Il router 1c di AS1 apprende il percorso AS2, AS3, X da parte del router 2a di AS2 e il percorso AS3, X da parte del router 3a di AS3
- Basandosi sulla propria politica, il router 1c decide di accettare il percorso AS3, X e rifiutare il percorso AS2, AS3, X (es: poiché necessita meno hop), propagando il percorso all'interno del suo AS

3. Consideriamo la seguente situazione (separata dalla precedente):



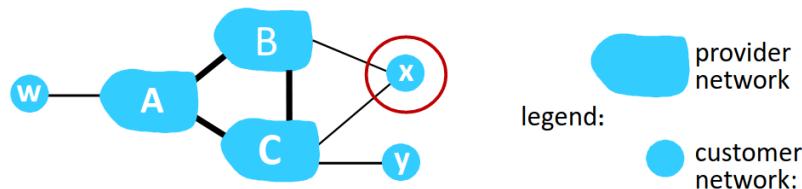
- Il router 2d di AS2 apprende (tramite iBGP) la possibilità di raggiungere il router X di AS3 sia tramite il router 2a sia tramite il router 2c (utilizzo dell'attributo NEXT-HOP)
- In tal caso, il router 2d sceglierà il gateway locale avente il minor costo intra-AS (**hot potato routing**) nonostante il percorso passante tramite 2c abbia un numero di hop inferiore, favorendo l'utilizzo di minor traffico all'interno di AS1 al prezzo di avere un minor controllo sul traffico esterno

4. Consideriamo la seguente situazione (separata dalla precedente):



- L'AS dell'ISP A pubblicizza il percorso A, w agli AS degli ISP B e C
- Poiché l'ISP B non trae alcun vantaggio per l'instradamento C, B, A, w (poiché nè C, nè A e nè w sono clienti di B), decide di non pubblicizzare tale percorso verso C (policy basata sull'*egoismo*)
- Di conseguenza, l'ISP C apprenderà solo il percorso C, A, w ricevuto da A stesso

5. Consideriamo la seguente situazione (separata dalla precedente):



- L'AS X è **dual-homed**, ossia connesso a due ISP
- Poiché B e C sono già direttamente connessi, per evitare il transito di traffico tra di essi passando attraverso l'AS x, quest'ultimo non pubblicherà il percorso C, x, B e il percorso B, x, C (policy basata sulla rimozione di intermediari tra ISP)

### Proposition 8. Scelta dei percorsi BGP

Se un router apprende più di un percorso verso la stessa destinazione, verrà selezionato il percorso in base a:

- Attributo del valore di preferenza locale
- AS-PATH più breve
- Percorso verso il NEXT-HOP con peso minore (**hot-potato routing**)
- Criteri aggiuntivi dettati dalla policy

## 4.9 Tipologie di instradamento

### 4.9.1 Unicast e Broadcast

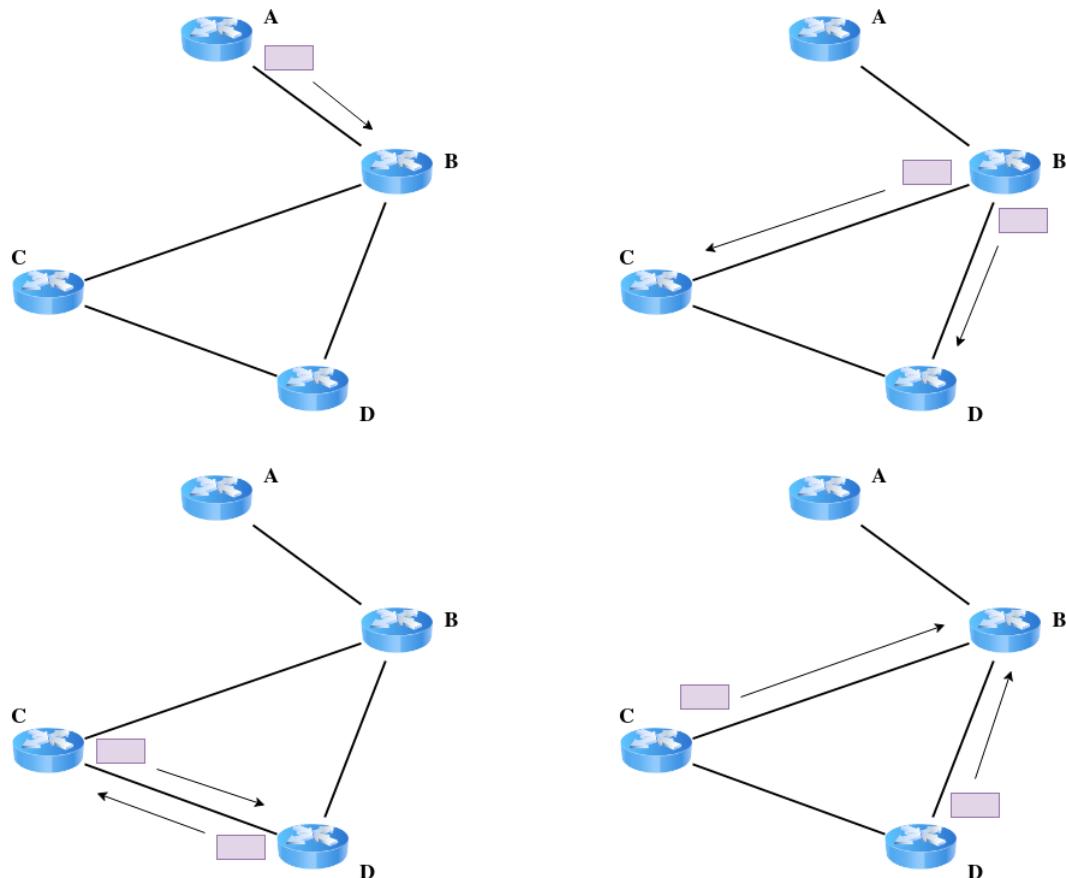
#### Definition 43. Unicast e Broadcast

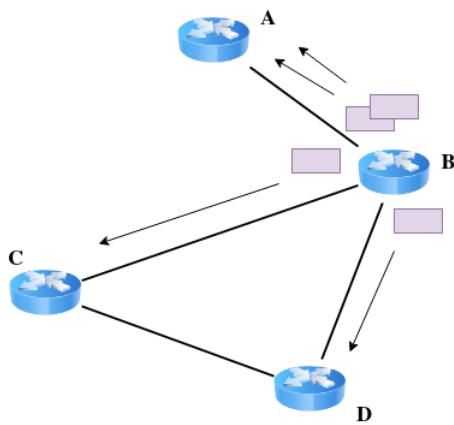
Nell'ambito delle comunicazioni in rete, definiamo come:

- **Unicast** la comunicazione tra **una sorgente ed una destinazione**, solitamente effettuata tramite la coppia <IP sorgente, IP destinazione>
- **Broadcast** la comunicazione tra **una sorgente e tutti i nodi di una rete**, solitamente effettuata tramite la coppia <IP sorgente, IP broadcast destinazione> (es: come già discusso, l'IP speciale 255.255.255.255 effettua il broadcast sulla rete locale)

Alla ricezione di un pacchetto broadcast da parte di un nodo, esso verrà **duplicato** ed inviato su tutti i nodi adiacenti al ricevente, fatta eccezione del nodo tramite cui è stato ricevuto il pacchetto. Tale tipologia di invio di messaggi viene detto **uncontrolled flooding** (tradotto *inondazione incontrollata*).

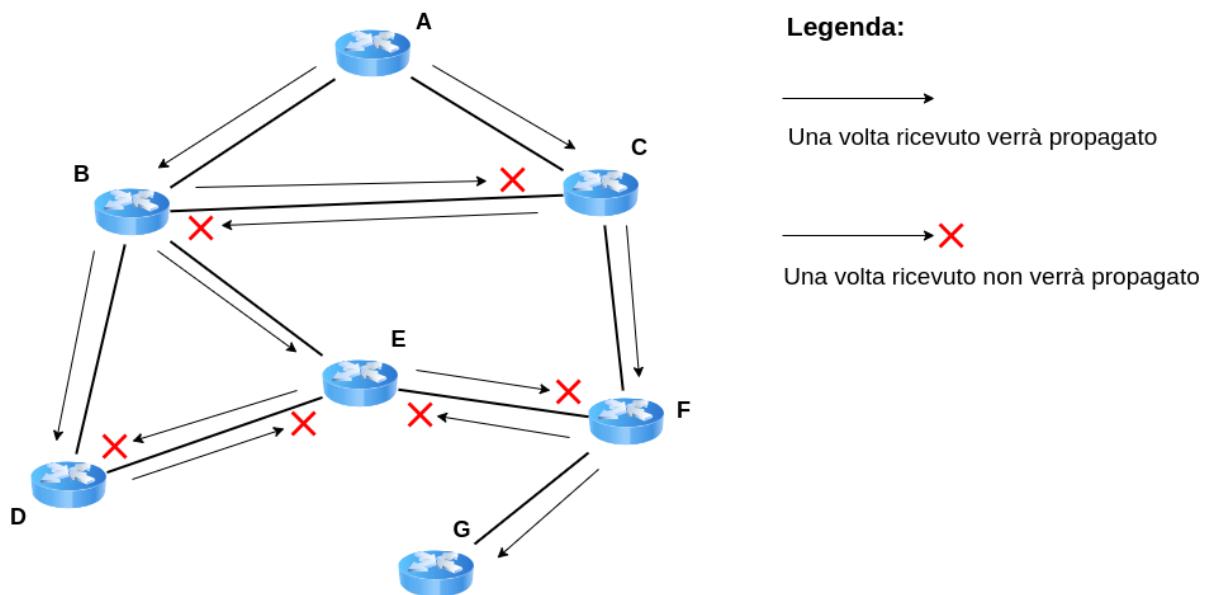
L'utilizzo dell'uncontrolled flooding può portare ad un **grave peggioramento della rete**, soprattutto nel caso in cui vi siano **cicli nel grafo**, portando il pacchetto broadcast ad essere duplicato ed inviato all'infinito.





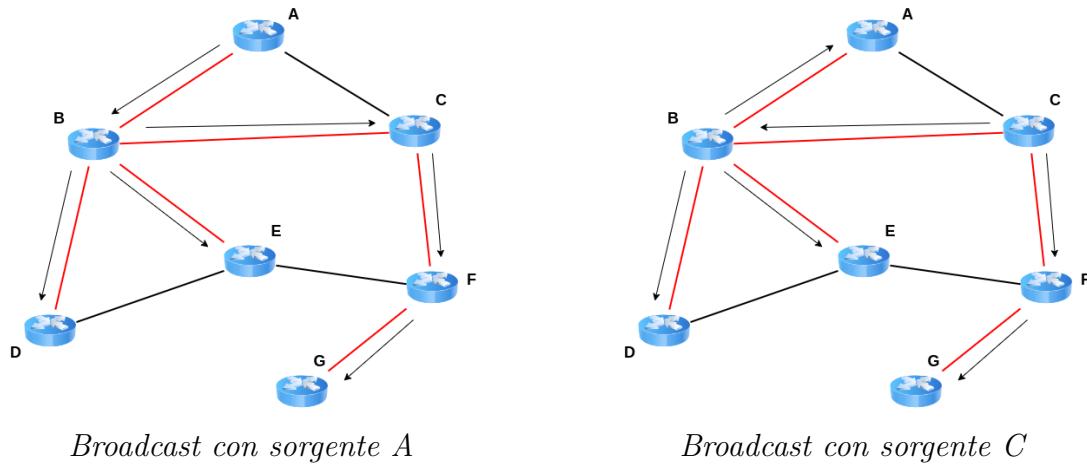
Di conseguenza, è necessario utilizzare una strategia di **controlled flooding** affinché si possa ridurre il traffico sulla rete:

- **Sequence number controlled flooding**, dove ogni nodo contiene una lista dei pacchetti già ricevuti, duplicati ed inviati, inoltrando il pacchetto broadcast ricevuto solo se non è già stato inviato
- **Reverse path forwarding (RPF)**, dove il pacchetto broadcast ricevuto viene inoltrato solo se è stato inoltrato dal link appartenente al **percorso più breve verso la sorgente** dell'invio



Nonostante le due strategie eliminino il problema di inondare la rete di pacchetti all'infinito, esse non eliminano completamente la trasmissione di pacchetti ridondanti, poiché essi verranno comunque inviati/ricevuti più volte prima di essere scartati.

Per risolvere definitivamente la ridondanza dei pacchetti, viene utilizzato uno **spanning tree**, ossia un albero in cui ogni nodo può essere raggiunto da un solo link, propagando i pacchetti broadcast **soltamente all'interno dell'albero stesso**.



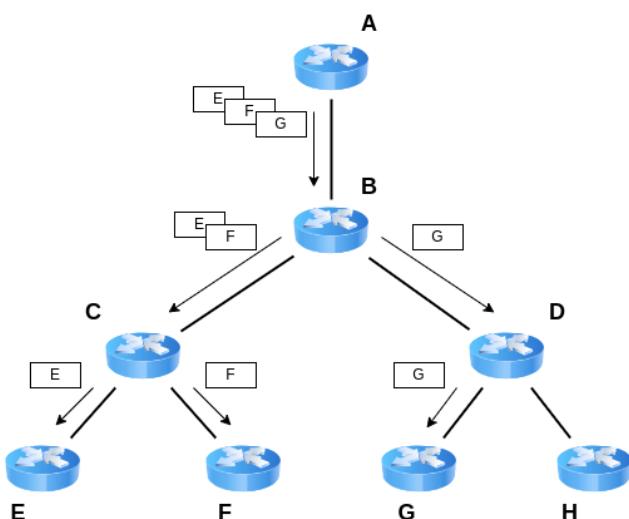
Per creare lo spanning tree, viene scelto un **nodo radice** come centro della rete. Successivamente, ogni altro nodo invierà un messaggio di **join** in modalità unicast verso la radice. Tale messaggio viene propagato finché esso non arriva ad un nodo che già appartiene all'albero o finché non arriva alla radice. Una volta "toccato" l'albero, il percorso mancante verrà aggiunto allo spanning tree.

#### 4.9.2 Multicast

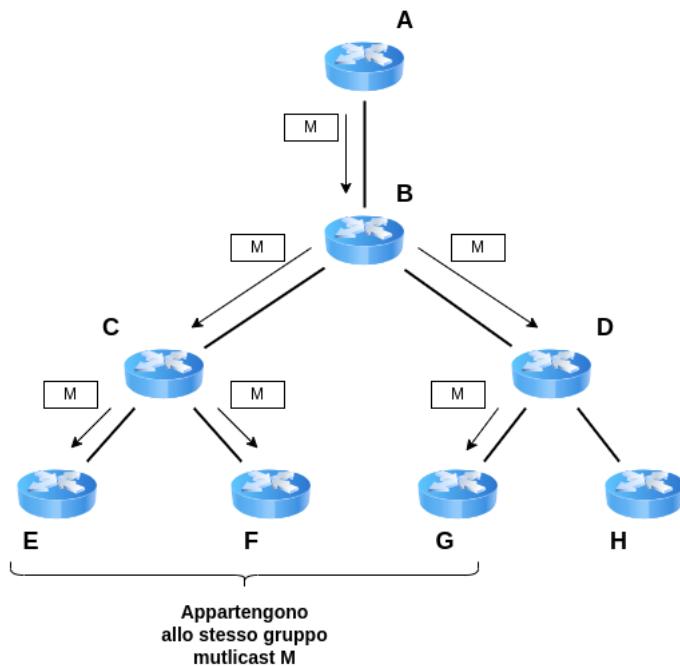
##### Definition 44. Multicast

Nell'ambito delle comunicazioni in rete, definiamo come **Multicast** la comunicazione tra **una sorgente e un gruppo di nodi della rete**

Molte applicazioni richiedono il trasferimento di pacchetti da uno o più mittenti verso un gruppo di destinatario (es: il trasferimento di un aggiornamento verso un gruppo di macchine, streaming ad un gruppo di utenti, ...). Effettuare tali trasferimenti utilizzando dei **pacchetti unicast multipli** risulta essere **estremamente inefficiente** per via dell'aggiunta di ritardi nella rete.

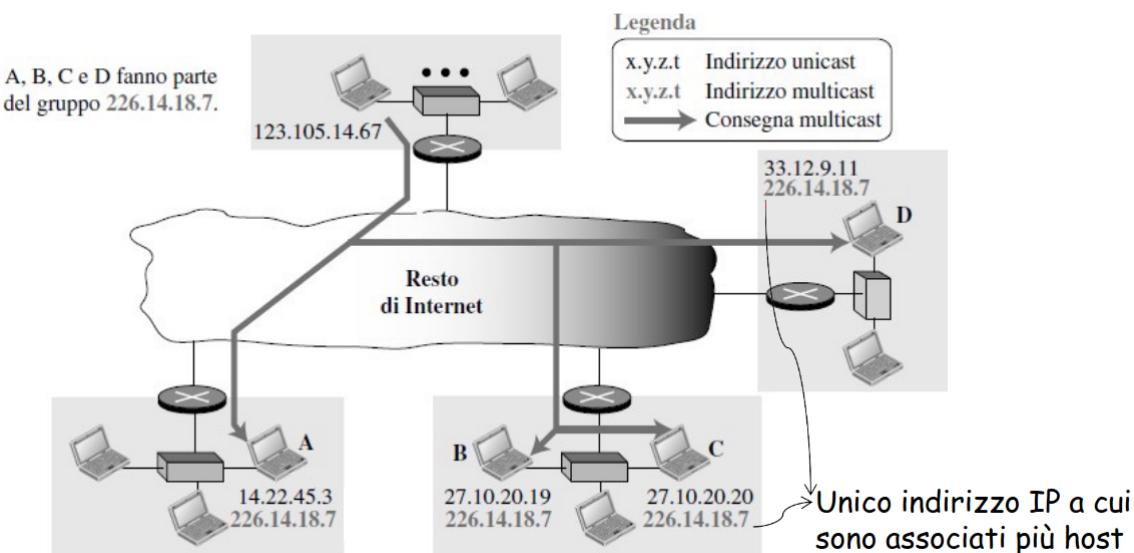


La soluzione più ottimale, dunque, risulta essere quella di trattare l'insieme di destinatari come un **gruppo multicast**, necessitando di un singolo pacchetto che verrà man mano sdoppiato per poter raggiungere tutte le destinazioni.



Tuttavia, poiché il protocollo IP è in grado di gestire un singolo indirizzo IP di destinazione, è necessario identificare tutti i membri del gruppo attraverso un **indirizzo multicast** (aggiuntivo rispetto al normale indirizzo IP).

In particolare, viene utilizzato un **blocco di indirizzi riservati** per il multicast. Per l'IPv4, ad esempio, il blocco di indirizzi da 224.0.0.0 a 239.255.255.255 ( $2^{28}$  gruppi possibili). Dunque, qualsiasi **indirizzo IP secondario** "appartenente alla rete" 224.0.0.0/4 viene considerato come un indirizzo multicast valido (dunque qualsiasi indirizzo nel formato 1110-**identificatore gruppo-**)



Dunque, l'appartenenza ad un gruppo multicast non ha alcuna relazione con il prefisso associato alla rete. Inoltre, l'appartenenza ad un gruppo è **variabile** (ad esempio il gruppo potrebbe avere un periodo di appartenenza limitato).

Per tanto, un router deve essere in grado di venire a conoscenza di quali gruppi siano raggiungibili su ciascuna delle sue interfacce per poter propagare l'informazione.

#### Definition 45. Protocollo IGMP

Il **protocollo Internet Group Management Protocol (IGMP)** è un protocollo di comunicazione utilizzato per offrire agli host la possibilità di comunicare al proprio router direttamente connesso la volontà di **aderire** ad uno specifico **gruppo multicast**.

I messaggi del protocollo IGMP (inviai con TTL pari a 1) si suddividono in:

- **Membership query**, inviato dal router agli host per determinare a quali gruppi multicast hanno aderito gli host (inviai periodicamente)
- **Membership request**, inviato da un host al router per informarlo dell'adesione ad un gruppo multicast
- **Leave group**, inviato da un host al router per informarlo dell'abbandono di un gruppo multicast

Ogni router multicast mantiene una **lista** per ciascuna sottorete di gruppi multicast (a patto che almeno un elemento del gruppo faccia parte della sottorete) impostando un **timer** per ogni **membership**. Se la membership non viene aggiornata (da request o leave) prima dello scadere del tempo, essa viene rimossa dalla lista.

Fra la popolazione complessiva di router, solo alcuni di essi, in particolare quelli collegati agli host del gruppo multicast, si occuperanno del traffico multicast (**multicast router**). Di conseguenza, è necessario un protocollo che coordini i vari router multicast per instradare il traffico multicast all'interno di Internet.

Per realizzare ciò, viene mantenuto un **albero** che colleghi i vari router multicast, instradando il traffico multicast solamente all'interno dell'albero stesso. Un albero può essere unico per tutto il gruppo o diverso a seconda della sorgente.

I principali protocolli per l'instradamento multicast sono:

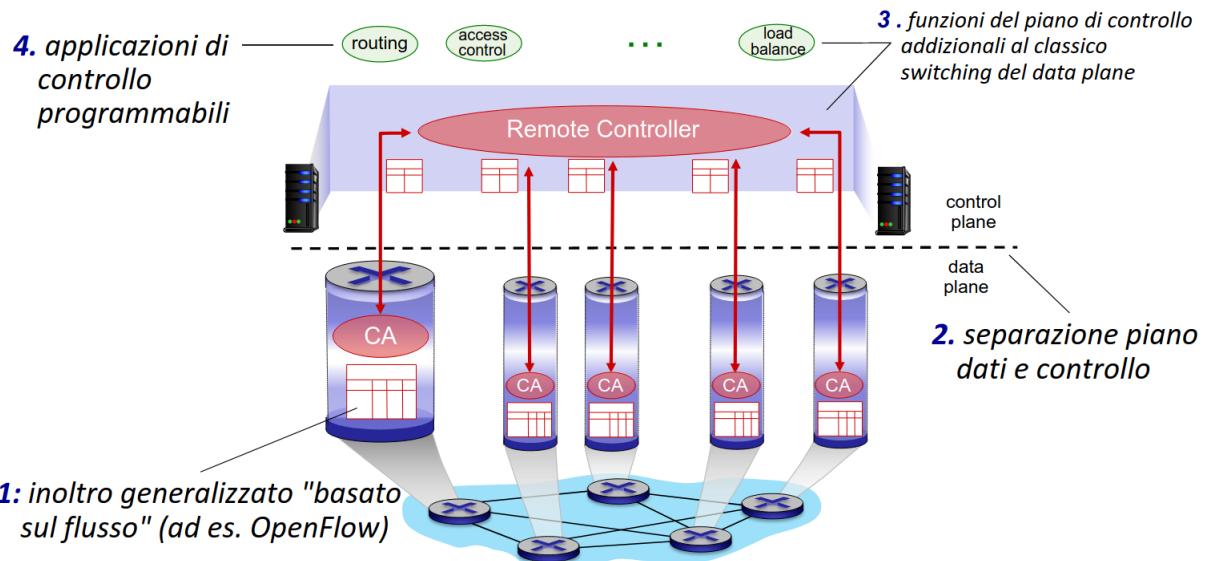
- **Instradamento multicast intra-AS:**
  - Distance-vector multicast routing protocol (DVMRP)
  - Multicast open shortest path first (MOSPF)
  - Protocol independent multicast (PIM)
- **Instradamento multicast inter-AS:**
  - Multicast border gateway protocol (MBGP)

## 4.10 Software Defined Networking (SDN)

Come precedentemente trattato, il livello di rete è stato storicamente implementato tramite un approccio di controllo distribuito sui router:

- Un **router monolitico** contiene hardware di commutazione, esegue implementazioni proprietarie dei protocolli standard Internet (es: IP, RIP, OSPF, BGP, ...) su sistemi operativi proprietari (es: Cisco IOS, ...)
- Diversi **middleboxes** per diverse funzionalità aggiuntive del livello di rete: firewall, load balancing, NAT, ...

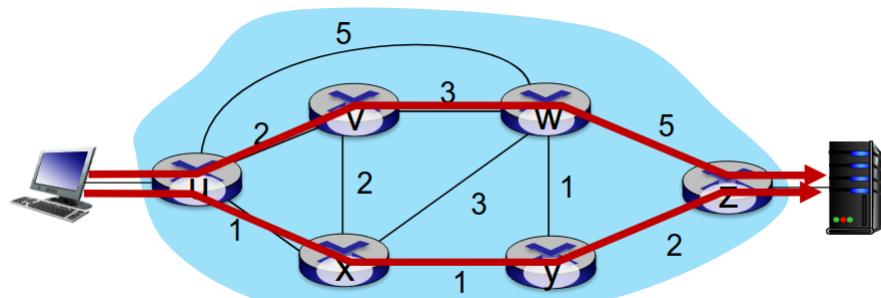
A differenza di un piano di controllo basato sull'approccio distribuito tra i vari router, dunque, il **Software Defined Networking (SDN)** permette l'implementazione di un **singolo piano di controllo** tramite un controller remoto che calcola e poi installa le tabelle di inoltro tramite le **API OpenFlow**, permettendo una gestione della rete più semplice, evitando errori di configurazione dei router e permettendo una **maggior flessibilità dei flussi di traffico**.



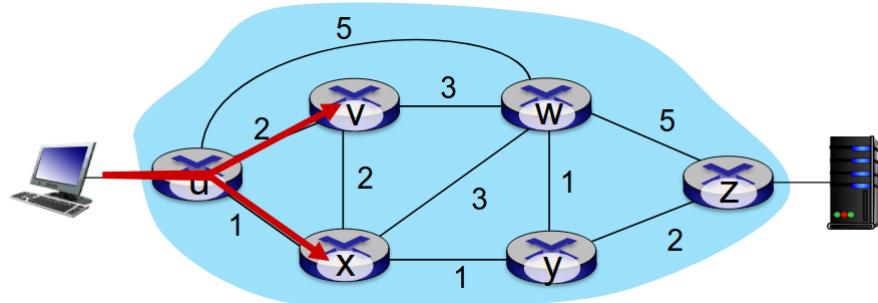
36

Esempi:

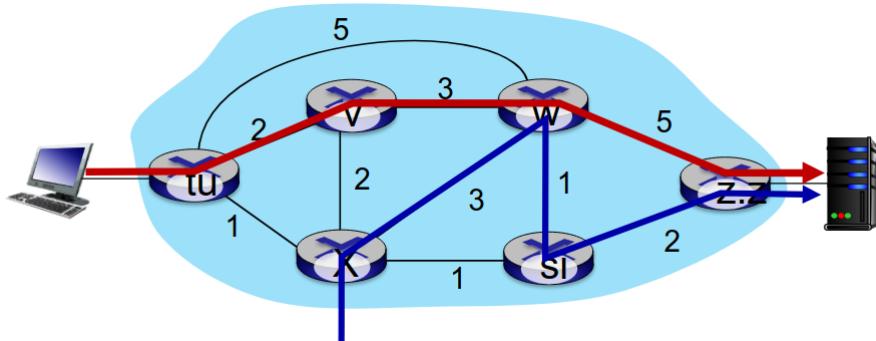
1. • L'ISP vuole far sì che il traffico dal router  $u$  verso il router  $z$  scorra sul percorso  $u, v, w, z$  anziché sul percorso  $u, x, y, z$ .



- Per ottenere ciò, utilizzando il normale approccio distribuito, sarebbe necessario ridefinire i pesi dei collegamenti in modo che l'algoritmo di instradamento del traffico calcoli il percorso desiderato.
  - Alternativamente, sarebbe necessario realizzare un nuovo algoritmo di routing.
2. • L'ISP vuole far sì che il traffico dal router  $u$  verso il router  $z$  venga bilanciato (**load balancing**) sui percorsi  $u, v, w, z$  e  $u, x, y, z$ .



- Utilizzando il normale approccio distribuito, ciò sarebbe impossibile se non tramite un nuovo algoritmo di routing.
- 3. • Il router  $w$  vuole instradare il traffico blu verso il router  $z$  e il traffico rosso verso il router  $z$  tramite due percorsi diversi



- Utilizzando il normale approccio distribuito, ciò sarebbe impossibile se non tramite una tipologia di forwarding diversa dal destination-based forwarding e un nuovo algoritmo di routing.

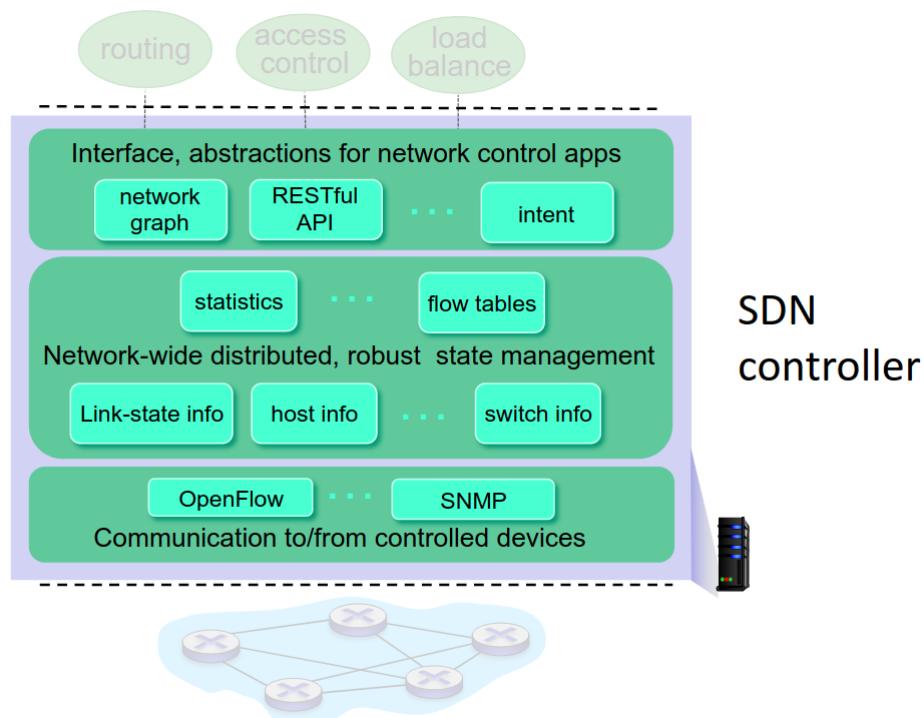
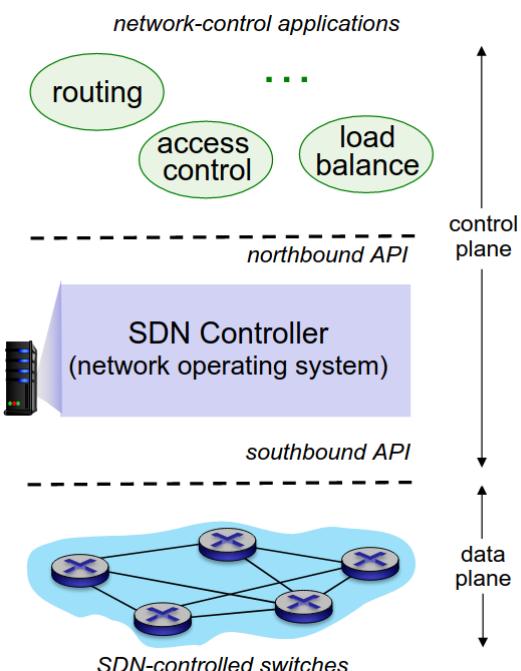
Tramite l'uso di un SDN, inoltre, il forwarding può essere realizzato tramite **switch di rete** veloci e semplici (al posto di normali router più complessi), i quali implementano il **generalized forwarding** all'interno del data plane, dove la **forwarding table** degli switch viene calcolata e installata sotto la supervisione del controller SDN tramite le API OpenFlow e un protocollo per la comunicazione con il controller.

Per quanto riguarda le **applicazioni di controllo della rete** presenti all'interno del control plane, invece, esse implementano funzioni di controllo utilizzando i servizi di livello inferiore (ossia le API fornite dal controller SDN). Possono essere fornite da un fornitore distinto rispetto a quello degli switch e del controller SDN.

Il **controller SDN** viene gestito da un sistema operativo di rete, mantenendo informazioni sullo stato della rete e interagendo con gli altri livelli attraverso delle API:

- **API Northbound**, utilizzate per interagire con le applicazioni di controllo della rete presenti nel control plane
- **API Southbound**, utilizzate per interagire con gli switch di rete all'interno del data plane tramite le **API Southbound**

Per evitare la presenza di un **single point of failure**, ossia un dispositivo il cui malfunzionamento porterebbe al malfunzionamento dell'intera rete, i controller SDN vengono implementati come **sistema distribuito**



#### Definition 46. Protocollo OpenFlow

Il **protocollo OpenFlow** è un protocollo di comunicazione utilizzato per la comunicazione tra controller SDN e switch di rete attraverso il **protocollo TCP** (con crittografia opzionale).

**Attenzione:** il protocollo OpenFlow è **diverso** dalle API OpenFlow, nonostante quest'ultime vengano utilizzate dal protocollo stesso.

I messaggi del protocollo OpenFlow si suddividono in tre categorie:

- **Controller-to-switch:**

- Messaggi di **feature**, ossia una query del controller per conoscere le funzionalità supportate dallo switch
- Messaggi di **configure**, dove il controller modifica parametri di configurazione dello switch
- Messaggi di **modify-state**, dove vengono utilizzate le API OpenFlow per aggiungere, eliminare o modificare campi della forwarding table dello switch
- Messaggi di **packet-out**, dove il controller invia un pacchetto da una specifica porta dello switch

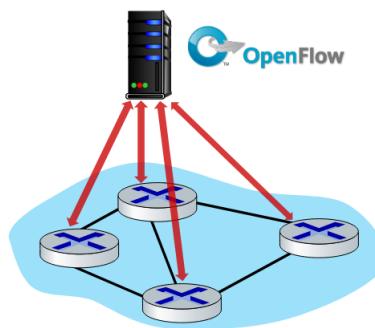
- **Asincroni (Switch-to-controller):**

- Messaggi di **packet-in**, dove viene trasferito un pacchetto al controller
- Messaggi di **flow-removed**, dove lo switch elimina una riga della forwarding table e notifica il controller
- Messaggi di **port-status**, dove lo switch informa il controller della modifica o problematica su una porta

- **Simmetrici (C-to-S & S-to-C)**

- Messaggi di hello, messaggi echo, messaggi di errore, ...

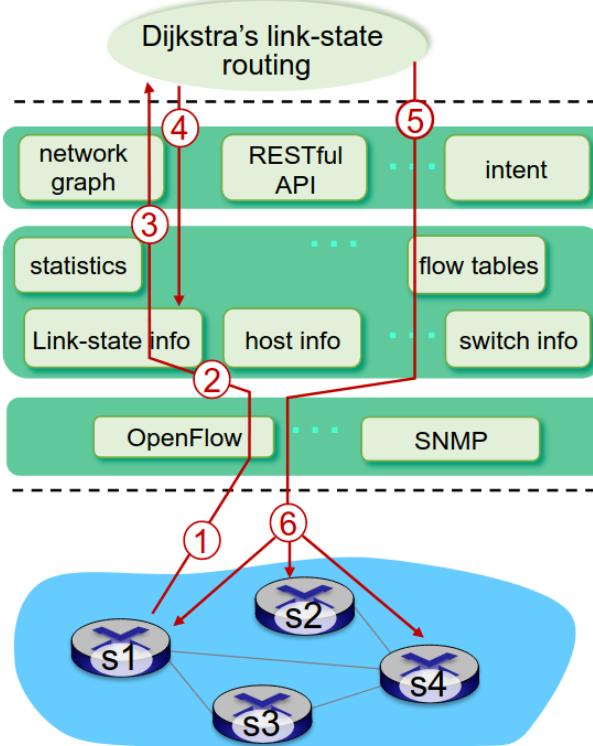
### Controller OpenFlow



### Esempio di interazione piano dati/controllo:

1. All'interno dello switch di rete S1 si verifica un errore sul collegamento verso lo switch di rete S2. Di conseguenza, lo switch S1 invia un messaggio OpenFlow di port-status per informare il controller
2. Il controller SDN riceve il messaggio OpenFlow, aggiornando le informazioni sullo stato del collegamento
3. L'applicazione di controllo della rete inerente al calcolo dei percorsi tramite l'algoritmo link-state di Dijkstra viene richiamata

4. L'applicazione accede alle informazioni sul grafo di rete e alle informazioni sullo stato dei collegamenti, calcolando i nuovi percorsi
5. L'applicazione di routing interagisce con il componente di calcolo delle forwarding table presente all'interno del controller SDN, calcolando le nuove forwarding table degli switch di rete necessarie
6. Il controller utilizza il protocollo OpenFlow e le API OpenFlow per installare le nuove tabelle negli switch che devono essere aggiornati



Per via della sua complessità di gestione, l'utilizzo dei controller SDN è attualmente confinato ai singoli AS, venendo quindi utilizzati come **"sostituto"** del normale **routing intra-AS tradizionale**. Gli obiettivi futuri, dunque, prevedono una maggiore scalabilità tramite la sostituzione anche del routing inter-AS, nonché una maggior robustezza ai guasti ed una maggior affidabilità/sicurezza (versioni aggiornate di OpenFlow usano l'autenticazione).

## 4.11 Gestione della rete

La **gestione della rete** prevede la gestione dei vari AS distribuiti all'interno di Internet attraverso quattro componenti fondamentali:

- **Managing server**, ossia un server tipicamente gestito da amministratori della rete
- **Managed device**, ossia un qualsiasi dispositivo della rete con componenti hardware o software configurabili
- **Dati** (es: dati di configurazione dello stato dei dispositivi, dati operativi, statistiche dei dispositivi, ...)
- **Protocollo di network management**, utilizzato dal managing server per interrogare, configurare e gestire i managed device e utilizzato da quest'ultimi per inviare dati o eventi rilevati al server

### Definition 47. Management Information Base (MIB)

Un **Management Information Base (MIB)** è un database presente all'interno di un managed device memorizzante dati sullo stato e la configurazione del dispositivo stesso attraverso il linguaggio **Structure of Management Information (SMI)**

Object ID	Name	Type	Comments
1.3.6.1.2.1.7.1	UDPIInDatagrams	32-bit counter	total # datagrams delivered
1.3.6.1.2.1.7.2	UDPNoPorts	32-bit counter	# undeliverable datagrams (no application at port)
1.3.6.1.2.1.7.3	UDInErrors	32-bit counter	# undeliverable datagrams (all other reasons)
1.3.6.1.2.1.7.4	UDPOutDatagrams	32-bit counter	total # datagrams sent
1.3.6.1.2.1.7.5	udpTable	SEQUENCE	one entry for each port currently in use

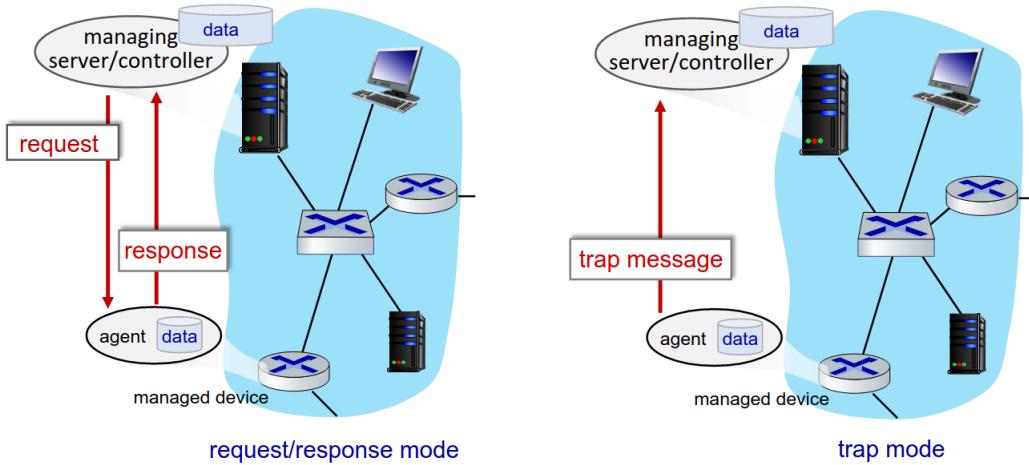
*Esempio di variabili MIB per il protocollo UDP*

### Definition 48. Protocollo SNMP

Il **protocollo Simple Network Management Protocol (SNMP)** è un protocollo di gestione della rete utilizzato per interrogare/impostare i dati presenti nei MIB dei dispositivi. Viene implementato tramite il **protocollo UDP** sulla porta nota 161.

Message type	Function
GetRequest GetNextRequest GetBulkRequest	manager-to-agent: "get me data" (data instance, next data in list, block of data).
SetRequest	manager-to-agent: set MIB value
Response	Agent-to-manager: value, response to Request
Trap	Agent-to-manager: inform manager of exceptional event

I messaggi del protocollo SNMP si differenziano in **messaggi request/response**, dove il server invia una richiesta al dispositivo e quest'ultimo risponde, e **messaggi trap**, dove il dispositivo informa il server a seguito di un'eccezione.



#### Definition 49. Protocollo NETCONF

Il **protocollo Network Configuration Protocol (NETCONF)** è un protocollo di gestione della rete utilizzato per gestire/configurare **attivamente** i dispositivi a livello di rete.

Il protocollo NETCONF sul **paradigma di remote procedure call (RPC)**, inviando messaggi NETCONF codificati in linguaggio XML attraverso un protocollo di trasporto sicuro (es: tramite TLS).

Inoltre, il protocollo NETCONF è in grado di recuperare, modificare, interrogare e attivare configurazioni sui managed devices attraverso dei **commit atomici** su più dispositivi (ossia un singolo commit in grado di modificare simultaneamente tutti i dispositivi)

NETCONF	Operation Description
<get-config>	Retrieve all or part of a given configuration. A device may have multiple configurations.
<get>	Retrieve all or part of both configuration state and operational state data.
<edit-config>	Change specified (possibly running) configuration at managed device. Managed device <rpc-reply> contains <ok> or <rpccerror> with rollback.
<lock>, <unlock>	Lock (unlock) configuration datastore at managed device (to lock out NETCONF, SNMP, or CLIs commands from other sources).
<create-subscription>, <notification>	Enable event notification subscription from managed device

**Esempio:**

```

01 <?xml version="1.0" encoding="UTF-8"?>
02 <rpc message-id="101" note message id
03   xmlns="urn:ietf:params:xml:ns:netconf:base:1.0">
04     <edit-config> change a configuration
05       <target>
06         <running/> change the running configuration
07       </target>
08     <config>
09       <top xmlns="http://example.com/schema/
10         1.2/config">
11           <interface>
12             <name>Ethernet0/0</name> change MTU of Ethernet 0/0 interface to 1500
13             <mtu>1500</mtu>
14           </interface>
15         </top>
16       </config>
17     </edit-config>
18   </rpc>

```

Per facilitare la scrittura di messaggi NETCONF RPC, viene utilizzato il **linguaggio di modellazione YANG**. In particolare, ogni documento XML descrivente il dispositivo e le sue funzionalità può essere generato a partire da una descrizione YANG, esprimendo anche vincoli tra dati che devono essere soddisfatti da una configurazione NETCONF valida.

