Università degli Studi di Roma "Tor Vergata"
Laurea in Informatica

Sistemi Operativi e Reti
(modulo Reti)
a.a. 2023/2024

Livello di rete: piano dei dati (parte1)

dr. Manuel Fiorelli
manuel.fiorelli@uniroma2.it
https://art.uniroma2.it/fiorelli

Basate sulle slide del libro di testo:
https://gaia.cs.umass.edu/kurose_ross/ppt.php

Livello di rete: i nostri obiettivi

- Capire i principi che stanno dietro i servizi del livello di rete, focalizzandosi sul piano dei dati:
  - modelli di servizio del livello di rete
  - funzioni di inoltro e di instradamento
  - come funziona un router
  - indirizzamento
  - inoltro generalizzato
  - architettura di Internet
- Implementazione in Internet
  - protocollo IP
  - NAT, middlebox

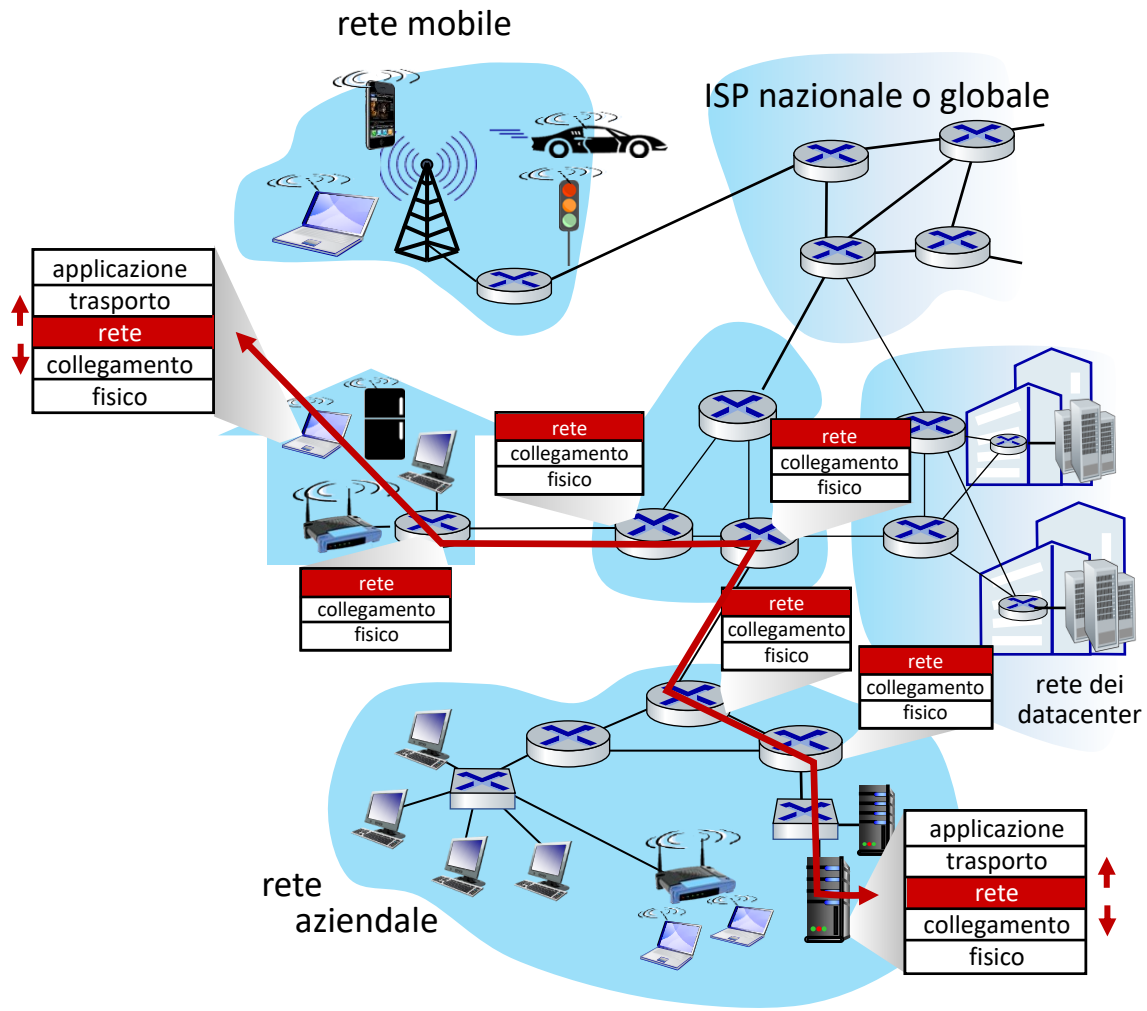
Livello di rete: tabella di marcia sul “piano dei dati”

- Livello di rete: panoramica
  - piano dei dati
  - piano di controllo
- Cosa c'è dentro un router
  - porte di ingresso, struttura di commutazione, porte di uscita
  - buffer management, scheduling
- IP: il Protocollo Internet
  - formato dei datagrammi
  - indirizzamento
  - traduzione degli indirizzi di rete
  - IPv6

- inoltro generalizzato, SDN
  - Match+action
  - OpenFlow: match+action in azione
- middlebox

Servizi e protocolli del livello di rete

- trasporta i segmenti dall'host mittente all'host destinatario
  - mittente: incapsula i segmenti dentro ai datagrammi che passa al livello di collegamento
  - destinatario: consegna i segmenti al protocollo del livello di trasporto
- i protocolli di livello di rete sono implementati da *tutti i dispositivi in Internet*: host, router
- router:
  - esamina i campi dell'intestazione di tutti i datagrammi IP che lo attraversano
  - sposta i datagrammi dalle porte di ingresso alla porta di uscita per trasferire il datagramma lungo il percorso dall'host di origine a quello di destinazione



Due funzioni chiave del livello di rete

funzioni del livello di rete

- *inoltro (forwarding):* trasferisce i pacchetti da un collegamento di ingresso di un router al collegamento di uscita appropriato del router
- *instradamento (routing):* determina il percorso seguito dai pacchetti dall'origine alla destinazione
  - *algoritmi di instradamento*

analogia: fare un viaggio

- *inoltro:* attraversamento di uno svincolo seguendo le indicazioni dei cartelli
- *instradamento:* pianificazione dei percorsi verso tutte le destinazioni scegliendo tra i molteplici possibili e conseguente installazione dei cartelli

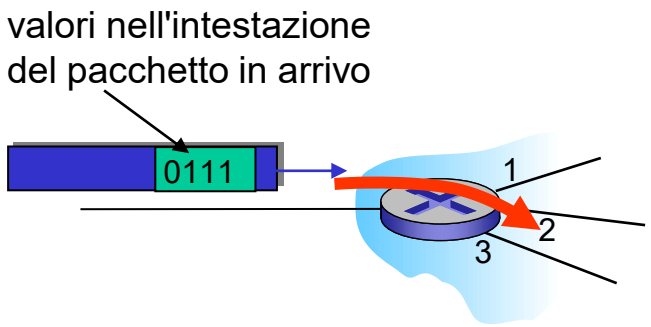
inoltro

instradamento

Livello di rete: piano dei dati e piano di controllo

Piano dei dati:

- funzione *locale*, a livello di singolo router
- determina come i pacchetti in arrivo a una porta di ingresso del router sono inoltrati verso una porta di uscita del router

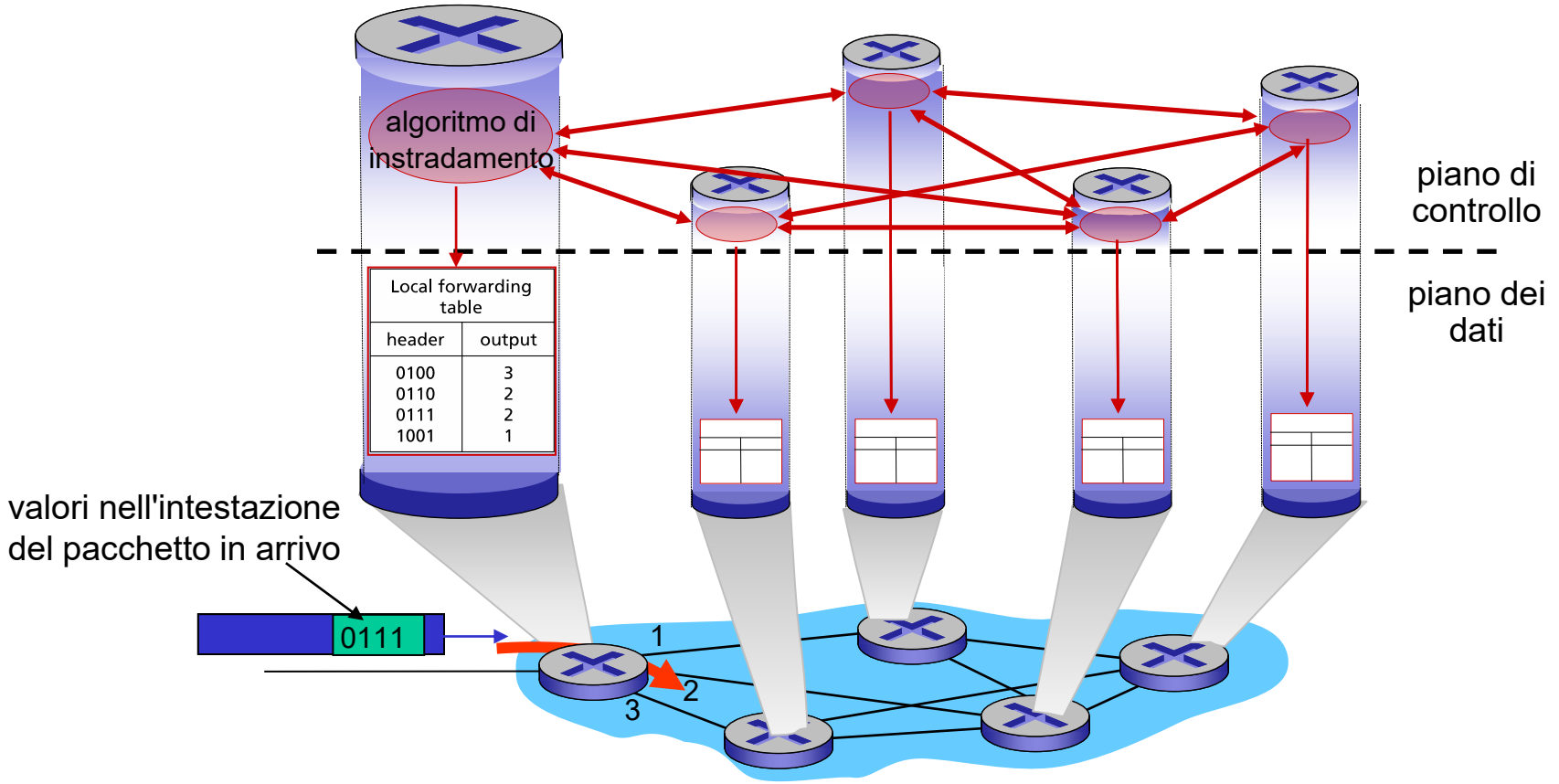


Piano di controllo

- *logica di rete*
- determina come i pacchetti sono instradati tra i router lungo un percorso dall'host di origine all'host di destinazione
- due approcci per il piano di controllo:
  - *algoritmi di instradamento tradizionali:* implementati nei router
  - *software-defined networking (SDN)*: implementato nei server (remoti)

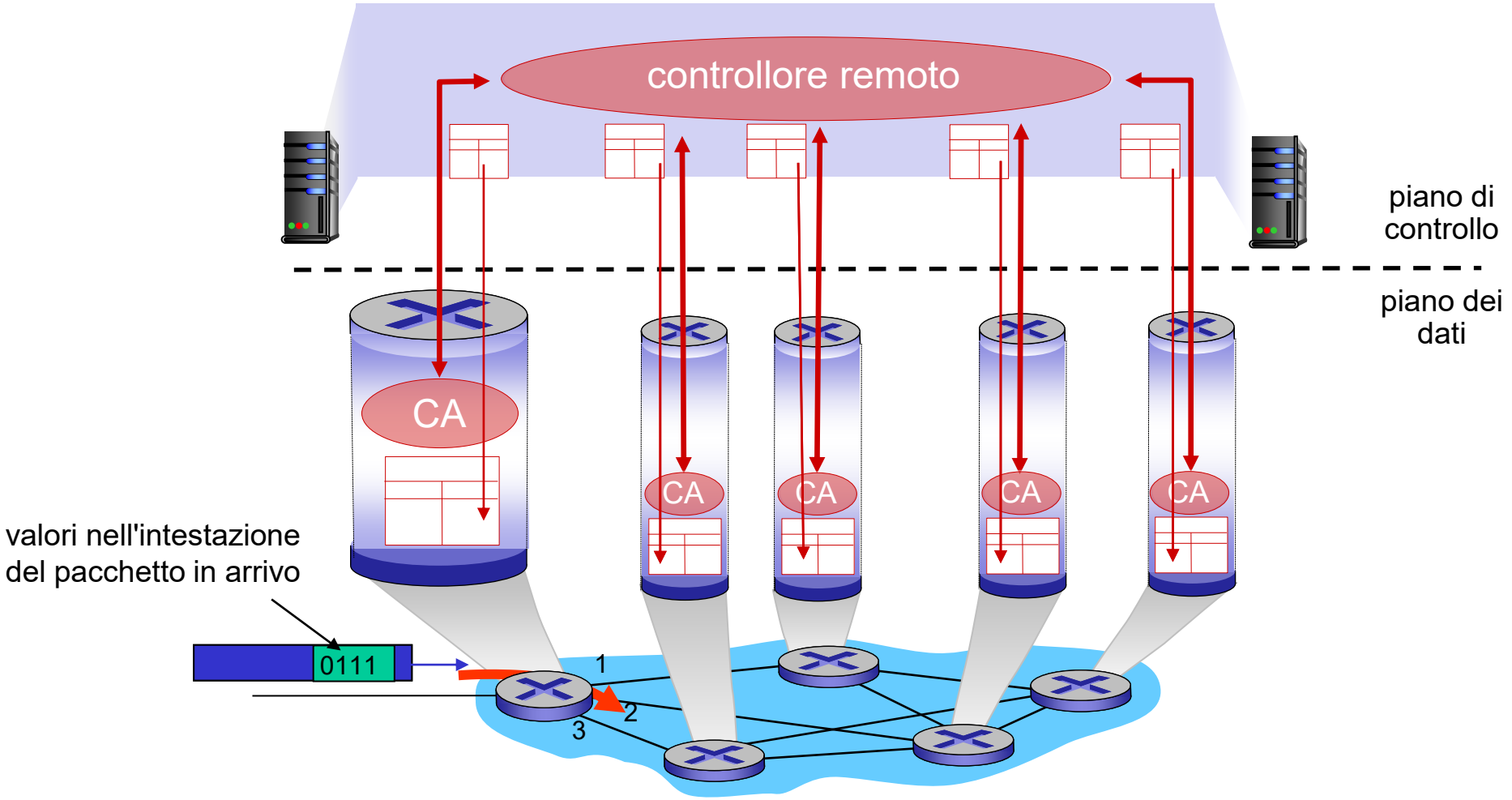
Piano di controllo per router

I singoli componenti dell'algoritmo di routing *in ogni singolo router*.



Software-Defined Networking (SDN)

Un *controllore remoto* calcola e installa le tabelle di inoltro nei router



Modello di servizio del livello di rete

D: Qual è il *modello di servizio* per il "canale" che trasporta i datagrammi dal mittente al destinatario (ovvero le sue caratteristiche)?

Esempi di servizi per un singolo datagramma

- consegna garantita
- consegna garantita con un ritardo inferiore a 40 ms

Esempi di servizi per un *flusso* di datagrammi:

- consegna in ordine
- minima ampiezza di banda garantita
- restrizioni sulle modifiche della spaziatura tra i pacchetti

Modelli di servizi del livello di rete

Architettura di rete	Modello di servizio	Garanzie di qualità del servizio, *quality of service* (QoS)?			
		Banda	Consegna	Ordine	Temporizzazione
Internet	best effort	nessuna	no	no	no
ATM	Constant Bit Rate	tasso costante	sì	sì	sì
ATM	Available Bit Rate	min. garantita	no	sì	no
Internet	Intserv Guaranteed (RFC 1633)	sì	sì	sì	sì
Internet	Diffserv (RFC 2475)	possibile	possibilmente	possibilmente	no

Modelli di servizi del livello di rete

Architettura di rete	Modello di servizio	Garanzie di qualità del servizio, *quality of service* (QoS)?			
		Banda	Consegna	Ordine	Temporizzazione
Internet	best effort	nessuna	no	no	no

Modello di servizio "best effort" di Internet

Nessuna garanzia circa*:*

i. consegna del datagramma alla destinazione con successo
ii. tempi o ordine di consegna
iii. larghezza di banda disponibile per il flusso da un capo all'altro

Riflessioni sul servizio best effort

- la semplicità del meccanismo ha consentito l'ampia diffusione di internet
- una dotazione sufficiente di larghezza di banda e protocolli in grado di adattarsi alla banda disponibile consentono alle prestazioni delle applicazioni in tempo reale (ad esempio, voce interattiva, video) di essere "sufficientemente buone" per la "maggior parte del tempo"
- servizi replicati e distribuiti a livello applicativo (datacenter, reti di distribuzione dei contenuti) che si collegano alle reti dei clienti e consentono di fornire servizi da più luoghi
- il controllo della congestione dei servizi "elastici" aiuta

Il successo del modello di servizio "best-effort" è difficilmente contestabile

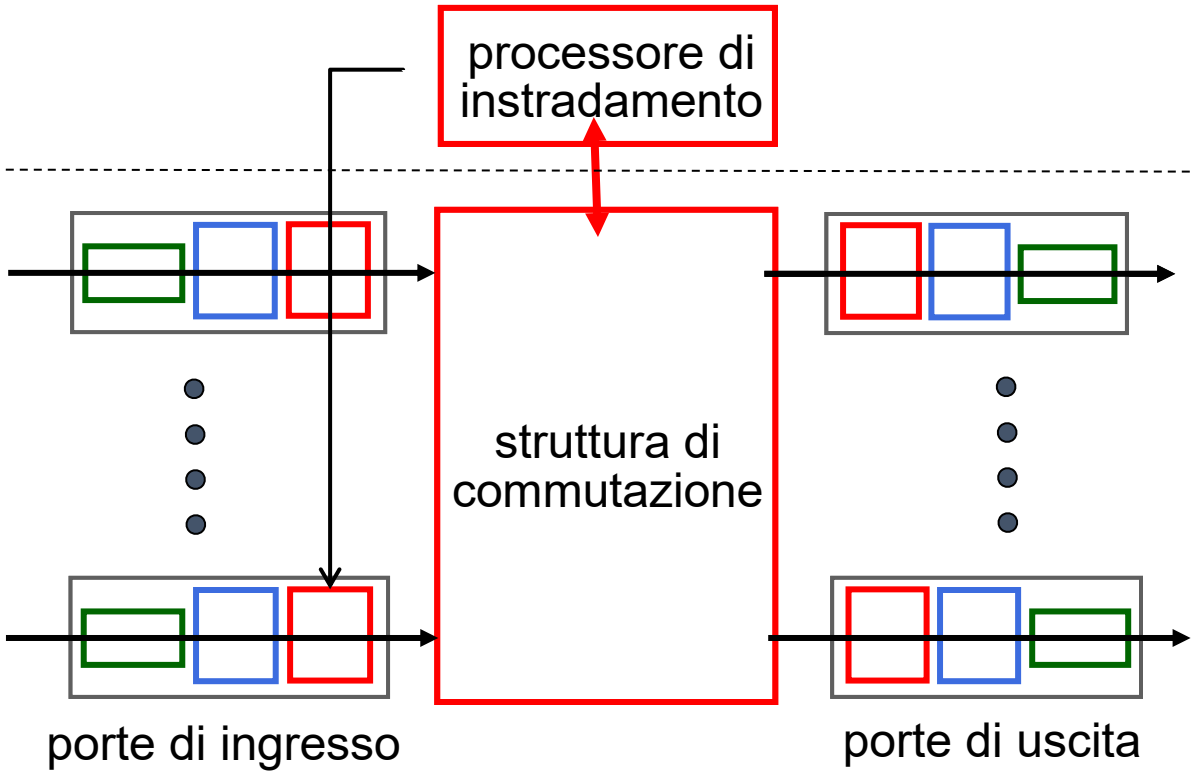
Livello di rete: tabella di marcia sul “piano dei dati”

- Livello di rete: panoramica
  - piano dei dati
  - piano di controllo
- Cosa c'è dentro un router
  - Porte di ingresso, struttura di commutazione, porte di uscita
  - buffer management, scheduling
- IP: il Protocollo Internet
  - Formato dei datagrammi
  - indirizzamento
  - Traduzione degli indirizzi di rete
  - IPv6

- Inoltro generalizzato, SDN
  - Match+action
  - OpenFlow: match+action in azione
- Middlebox

Architettura del router

visione ad alto livello di una generica architettura di router:



piano di controllo (instradamento, risposta a malfunzionamenti e gestione)
(software) opera sulla scala temporale dei millisecondi o dei secondi

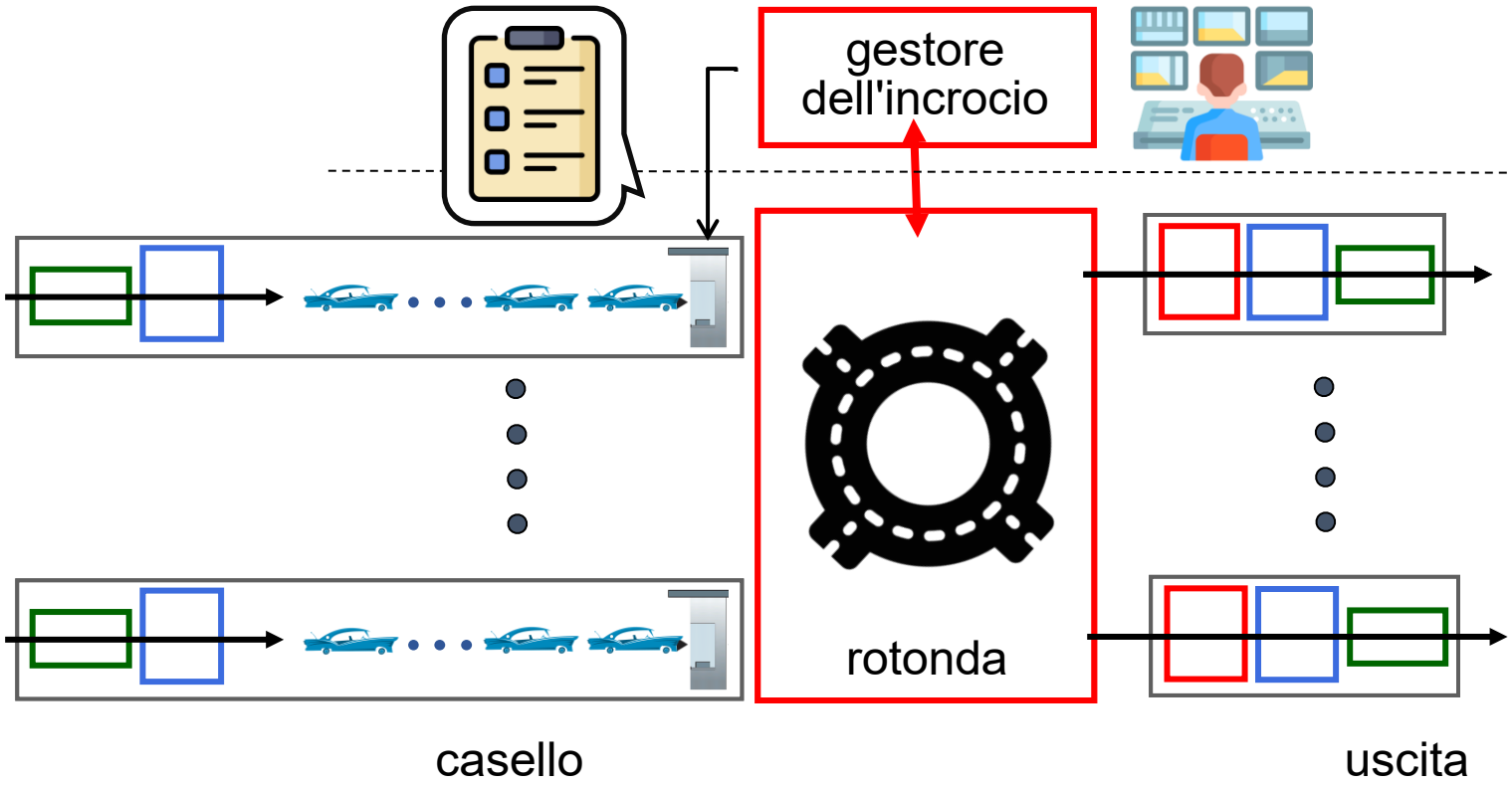
piano dei dati (inoltro)
(hardware) opera sulla scala temporale dei nanosecondi

Si consideri un collegamento a 100 Gbps e un datagramma da 64 byte. Il prossimo arriverà tra:

$$\frac{64 \cdot 8\ bit}{100\ Gb/s} = \frac{512\ bit}{100 \cdot 10^9 bit/s} = 5.12\ ns$$

Architettura del router

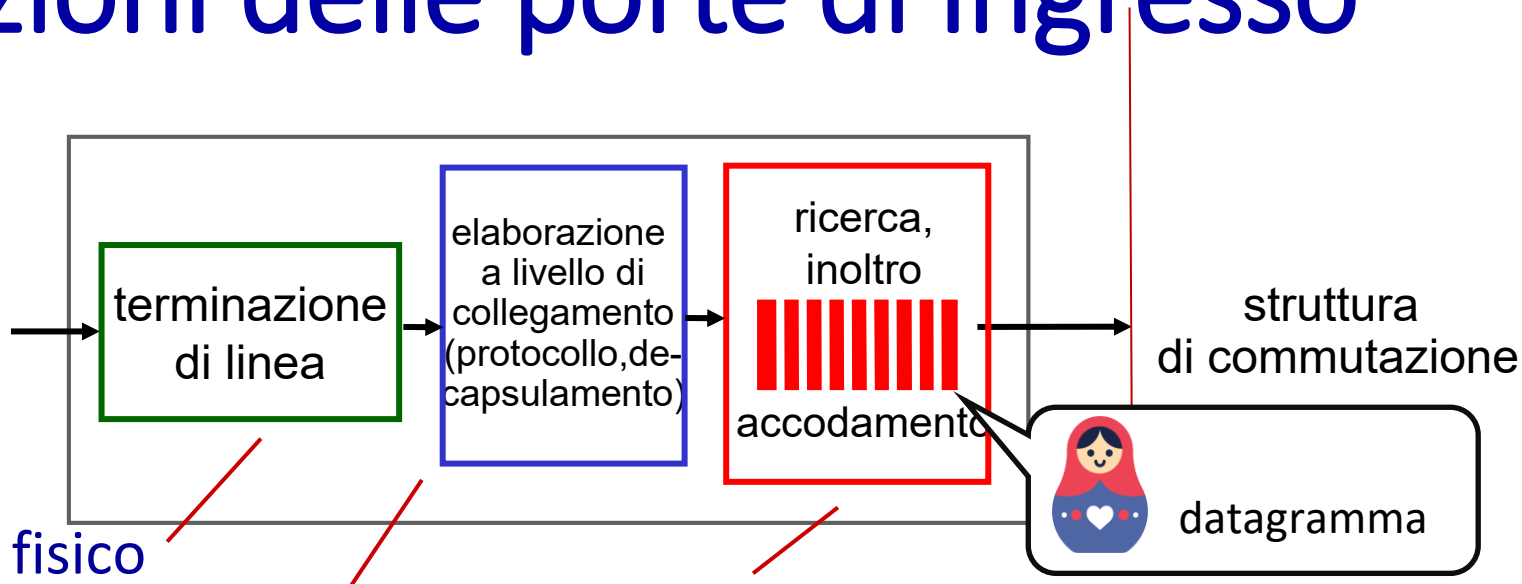
analogia per la architettura generica di router



piano di controllo (instradamento, risposta a malfunzionamenti e gestione)
(software) opera sulla scala temporale dei millisecondi o dei secondi

piano dei dati (inoltro)
(hardware) opera sulla scala temporale dei nanosecondi

Funzioni delle porte di ingresso



livello fisico
ricezione di bit

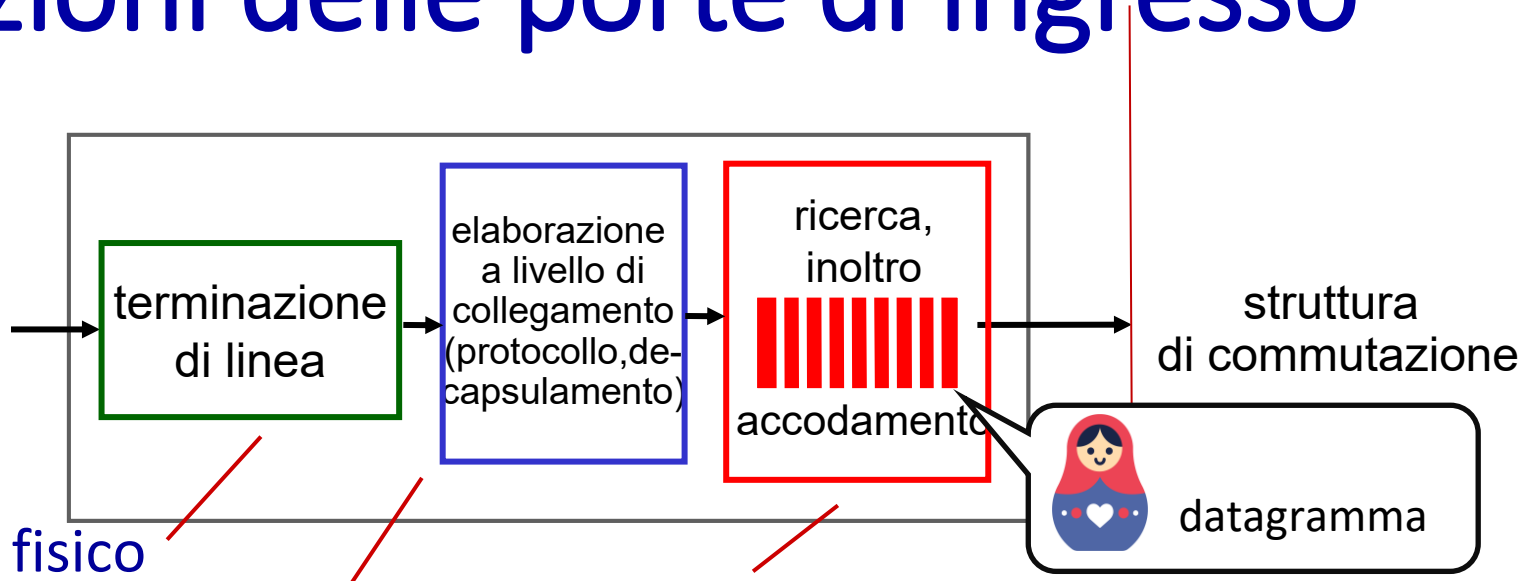
Livello di collegamento:
Es., Ethernet

Frame

commutazione decentralizzata:

- usando i valori dei campi di intestazione, trova la porta di uscita usando la tabella di inoltro nella memoria della porta di ingresso *("match plus action")*
- obietto: completare l'elaborazione nella porta di ingresso alla "velocità della linea"
- accodamento presso la porta di ingresso: se i datagrammi arrivano più velocemente di quanto la struttura di commutazione possa trasferirli

Funzioni delle porte di ingresso



livello fisico
ricezione di bit

Livello di collegamento:
Es., Ethernet

Frame

commutazione decentralizzata:

- usando i valori dei campi di intestazione, trova la porta di uscita usando la tabella di inoltro nella memoria della porta di ingresso *("match plus action")*
- inoltro basato sulla destinazione: inoltro basato esclusivamente sull'indirizzo IP di destinazione (tradizionale)
- inoltro generalizzato: inoltro basato su più campi di intestazione

Destinazione basata sull'indirizzo di destinazione

forwarding table

Destination Address Range	Link Interface
11001000 00010111 00010000 00000000 through 11001000 00010111 00010111 11111111	0
11001000 00010111 00011000 00000000 through 11001000 00010111 00011000 11111111	1
11001000 00010111 00011001 00000000 through 11001000 00010111 00011111 11111111	2
otherwise	3

D: ma cosa succede se gli intervalli non si dividono così bene?

Destinazione basata sull'indirizzo di destinazione

forwarding table

Destination Address Range	Link Interface
11001000 00010111 00010000 00000000 through 11001000 00010111 00010111 11111111	0
11001000 00010111 00010000 00000100 through 11001000 00010111 00010000 00000111	3
11001000 00010111 00011000 00000000 through 11001000 00010111 00011000 11111111	1
11001000 00010111 00011001 00000000 through 11001000 00010111 00011111 11111111	2
otherwise	3

Corrispondenza a prefisso più lungo

Corrispondenza a prefisso più lungo

quando si cerca una voce della tabella di inoltro per un dato indirizzo di destinazione, si usa il prefisso di indirizzo *più lungo* che corrisponde all'indirizzo di destinazione.

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

esempi:

11001000	00010111	00010110	10100001	quale interfaccia?
11001000	00010111	00011000	10101010	quale interfaccia?

Corrispondenza a prefisso più lungo

Corrispondenza a prefisso più lungo

quando si cerca una voce della tabella di inoltro per un dato indirizzo di destinazione, si usa il prefisso di indirizzo *più lungo* che corrisponde all'indirizzo di destinazione.

Intervallo di indirizzi di destinazione	Interfaccia
11001000 00010111 00010*** ********	0
11001000 00010111 00011000 ********	1
11001 0011*** ********	2
altrimenti	3

corrispondenza!

esempi:

11001000	00010111	00010110	10100001	quale interfaccia?
11001000	00010111	00011000	10101010	quale interfaccia?

Corrispondenza a prefisso più lungo

Corrispondenza a prefisso più lungo

quando si cerca una voce della tabella di inoltro per un dato indirizzo di destinazione, si usa il prefisso di indirizzo *più lungo* che corrisponde all'indirizzo di destinazione.

Intervallo di indirizzi di destinazione	Interfaccia
11001000 00010111 00010*** ********	0
11001000 00010111 00011000 ********	1
11001000 00010111 00011*** ********	2
otherwise	3

corrispondenza!

esempi:

11001000 00010111 00010110 10100001 quale interfaccia?

11001000 00010111 00011000 10101010 quale interfaccia?

Corrispondenza a prefisso più lungo

Corrispondenza a prefisso più lungo

quando si cerca una voce della tabella di inoltro per un dato indirizzo di destinazione, si usa il prefisso di indirizzo *più lungo* che corrisponde all'indirizzo di destinazione.

Intervallo di indirizzi di destinazione	Interfaccia
11001000 00010111 00010*** ********	0
11001000 00010111 00011000 ********	1
11001000 00010111 00011*** ********	2
altrimen	3

corrispondenza!

esempi:

11001000 00010111 00010110 10100001	quale interfaccia?
11001000 00010111 00011000 10101010	quale interfaccia?

Corrispondenza a prefisso più lungo

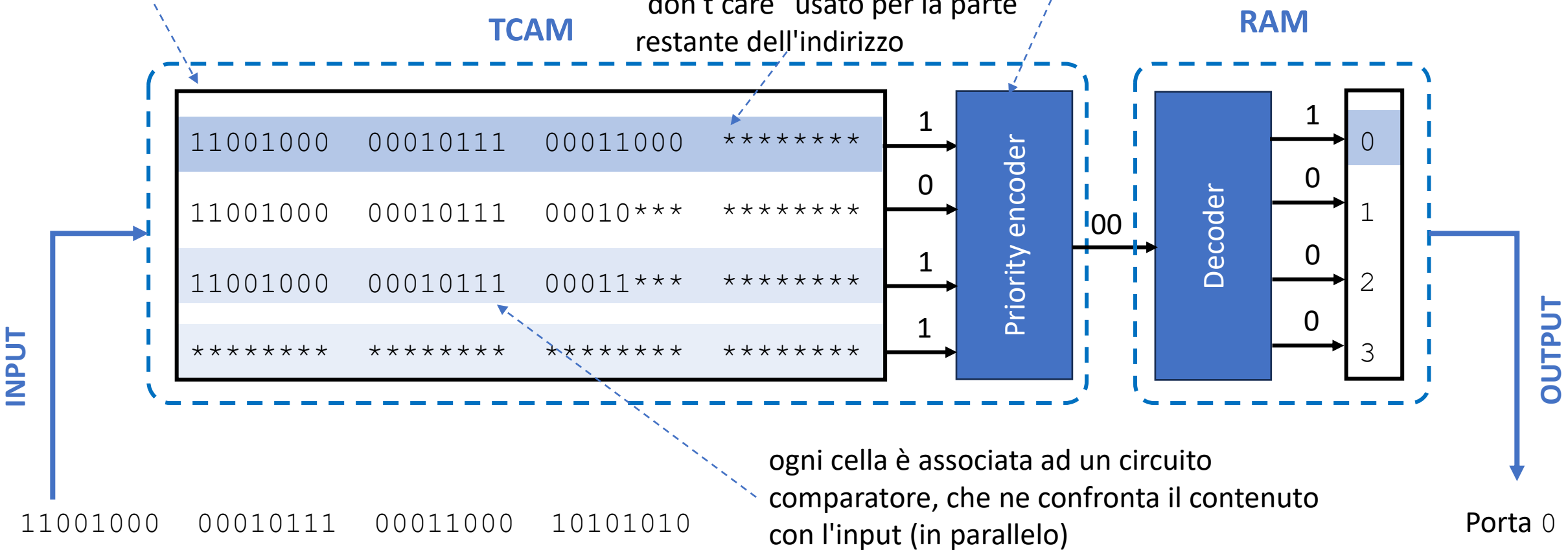
- Vedremo a breve *perché* viene usata la corrispondenza a prefisso più lungo, quando studieremo l'indirizzamento
- corrispondenza a prefisso più lungo: spesso eseguito con le ternary content addressable memories (TCAMs)
  - *content addressable:* un indirizzo IP a 32 bit è passato alla memoria che restituisce il contenuto della tupla nella tabella di inoltro corrispondente a quell'indirizzo in un tempo essenzialmente costante
  - Cisco Catalyst: ~1M voci nella tabella di inoltro in TCAM

Corrispondenza a prefisso più lungo

Occorre memorizzare i prefissi dal più lungo al più corto, in modo che la prima corrispondenza selezionata dal *priority encoder* sia quella per il prefisso più lungo

restituisce l'indirizzo della prima linea a 1

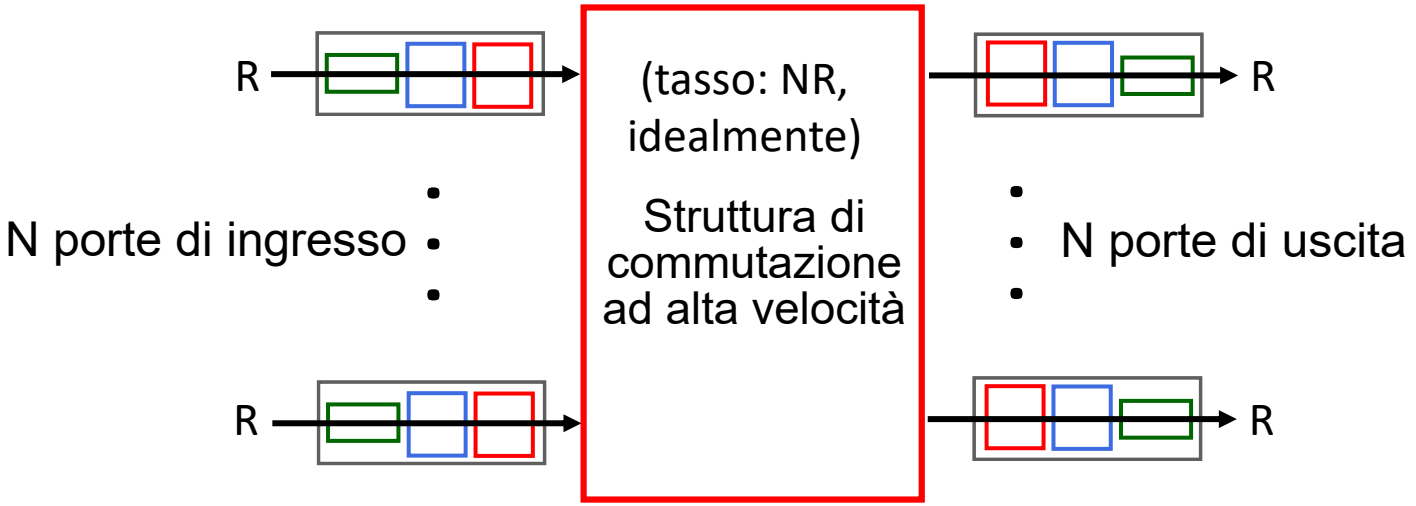
"don't care" usato per la parte restante dell'indirizzo



Vagamente basato su:
Irfan, Muhammad & Ullah, Dr. Zahid & Cheung, Ray C.C.. (2019). A High-performance Distributed RAM based TCAM Architecture on FPGAs. IEEE Access. PP. 10.1109/ACCESS.2019.2927108.

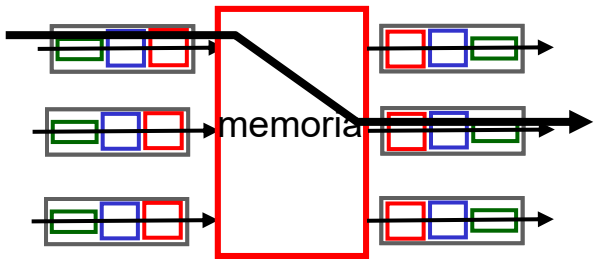
Struttura di commutazione (*switching fabric*)

- trasferisce i pacchetti dal collegamento di ingresso al collegamento di uscita appropriato
- tasso di trasferimento: tasso al quale i pacchetti vengono trasferiti dalla porta di input a quella di output
  - Spesso misurato come multiplo del tasso di trasmissione delle linee di input/output
  - N input: si desidera avere un tasso di trasferimento della struttura di commutazione N volte il tasso delle linee di input/output

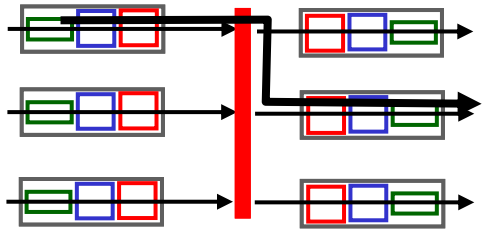


Struttura di commutazione (*switching fabric*)

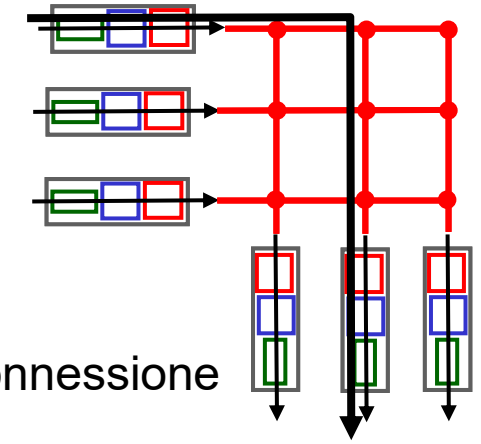
- trasferisce i pacchetti dal collegamento di ingresso al collegamento di uscita appropriato
- tasso di trasferimento: tasso al quale i pacchetti vengono trasferiti dalla porta di input a quella di output
  - Spesso misurato come multiplo del tasso di trasmissione delle linee di input/output
  - N input: si desidera avere un tasso di trasferimento della struttura di commutazione N volte il tasso delle linee di input/output



memoria



bus

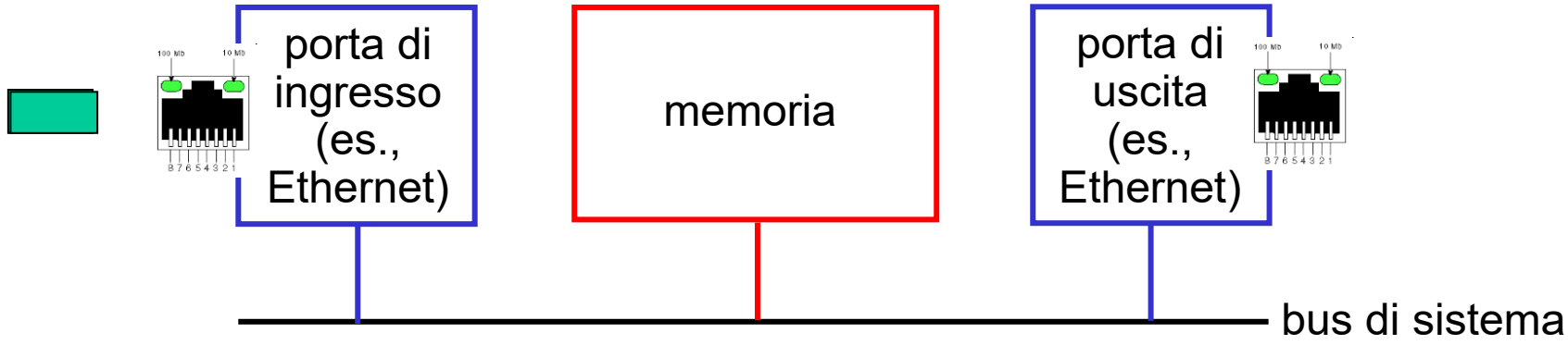


rete di interconnessione

Commutazione in memoria

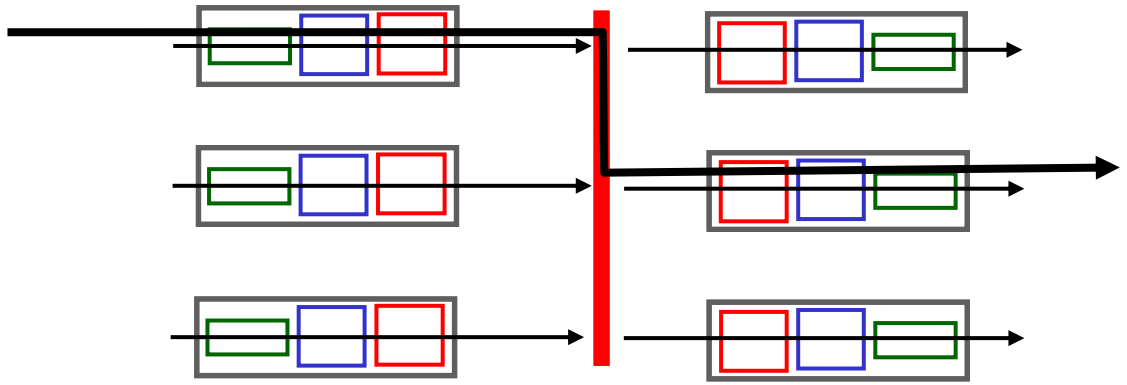
router di prima generazione:

- computer tradizionali con commutazione sotto il diretto controllo della CPU
- pacchetti copiati nella memoria del sistema
- velocità limitata dall'ampiezza di banda della memoria (2 attraversamenti del bus per datagramma)



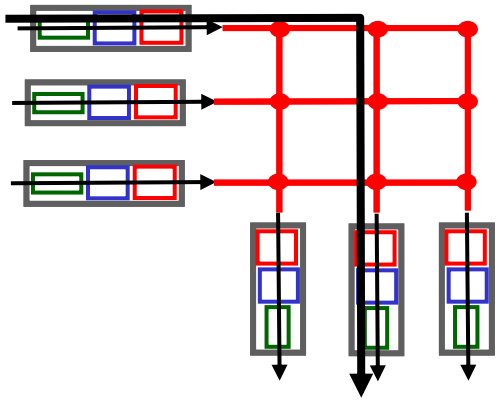
Commutazione tramite bus

- le porte di ingresso trasferiscono un pacchetto direttamente alle porte di uscita tramite un bus condiviso
- *bus contention:* velocità di commutazione limitata dalla velocità del bus
- bus a 32 Gbps, Cisco 5600: velocità sufficiente per router di accesso

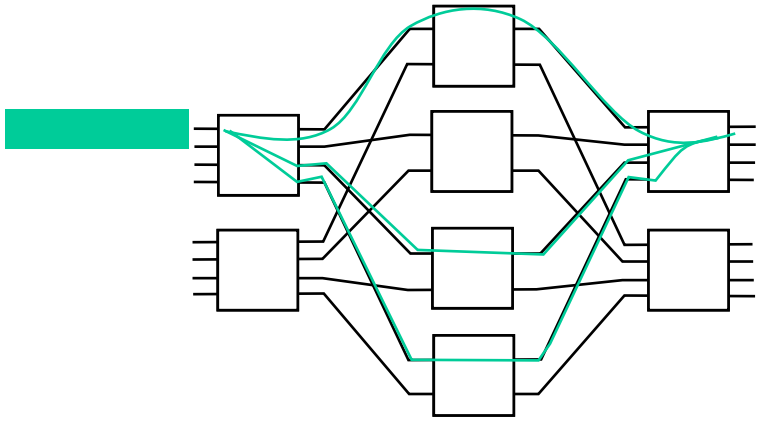


Commutazione attraverso rete di interconnessione

- Crossbar (matrice di commutazione), reti Clos, altre reti di interconnessione sviluppate originariamente per architetture multiprocessore
- multistage switch: switch *nxn* da più stadi di switch più piccoli
- sfruttare il parallelismo:
  - frammenta il datagramma in celle di lunghezza fissa all'ingresso
  - commutare le celle attraverso la rete di commutazione, riassemblare il datagramma in uscita



3x3 crossbar switch

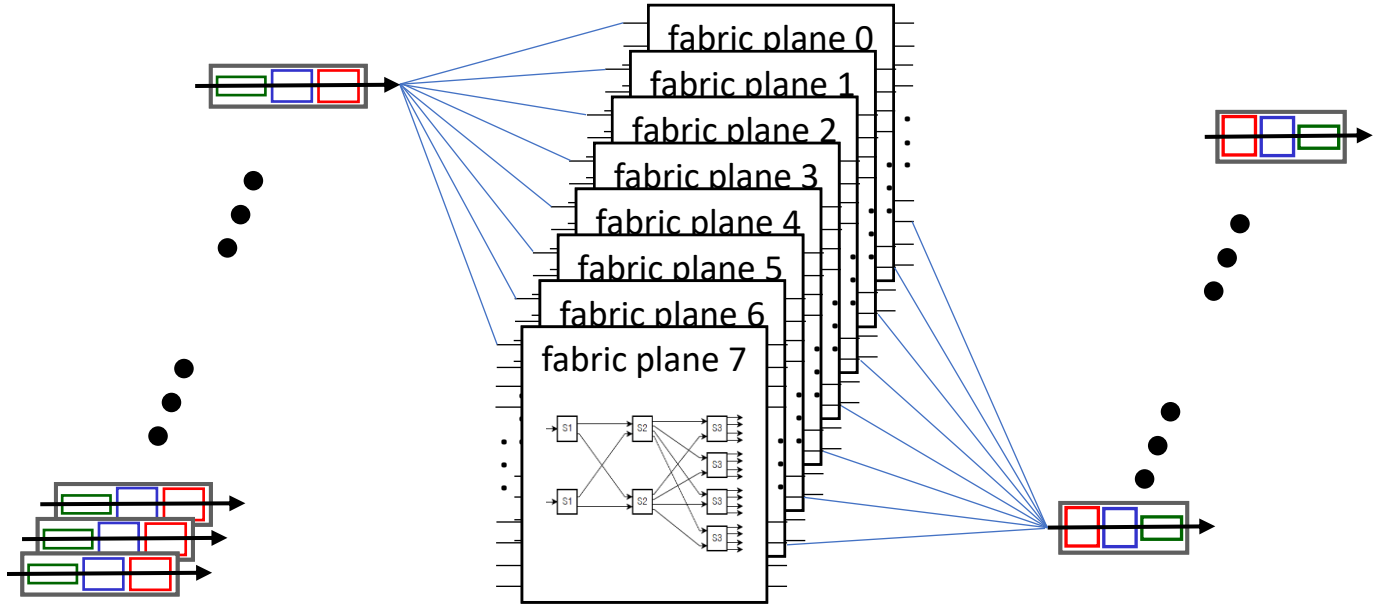


8x8 multistage switch
costruito da switch più piccoli

Commutazione attraverso rete di interconnessione

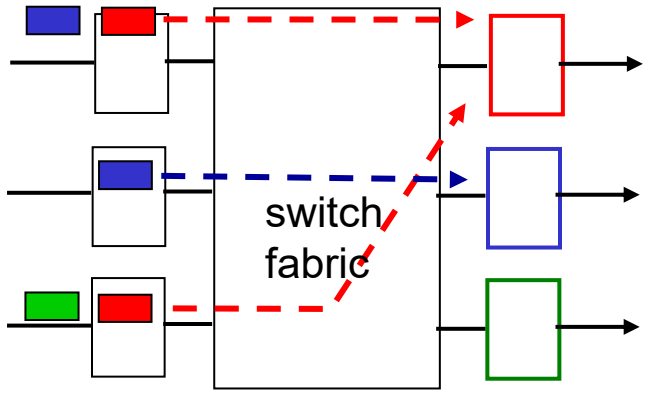
- scalare usando molteplici piani di commutazione in parallelo:
  - speedup, scaleup attraverso il parallelismo

- Cisco CRS router:
  - unità di base: 8 switching plane
  - ogni plane: rete di interconnessione a 3 stadi
  - Capacità di commutazione fino a centinaia di Tbps

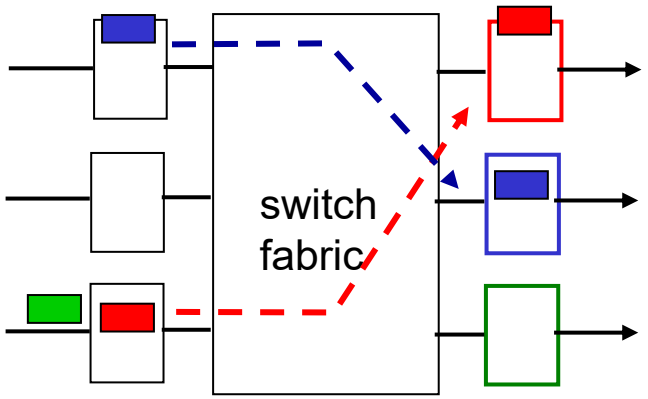


Accodamento sulle porte di ingresso

- Se la struttura di commutazione è più lenta della porte di ingresso combinate -> può verificarsi accodamento sulle porte di ingresso
  - ritardo di accodamento e perdite dovute all'overflow dei buffer di input!
- Blocco in testa alla coda [*Head-of-the-Line (HOL) blocking*]: il datagramma accodato all'inizio della coda impedisce agli altri in coda di avanzare

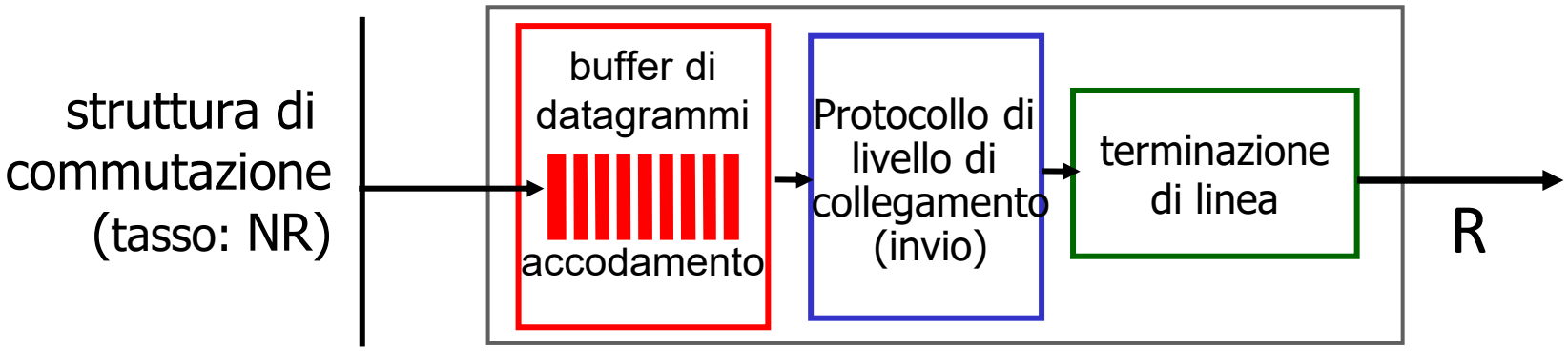


contesa della porta di uscita: soltanto un datagramma rosso può essere trasferito. Il pacchetto rosso in basso è *bloccato*



dopo il trasferimento di un pacchetto: il pacchetto verde sta sperimentando il *blocco in testa alla coda*

Accodamento in uscita



questa è una slide importante

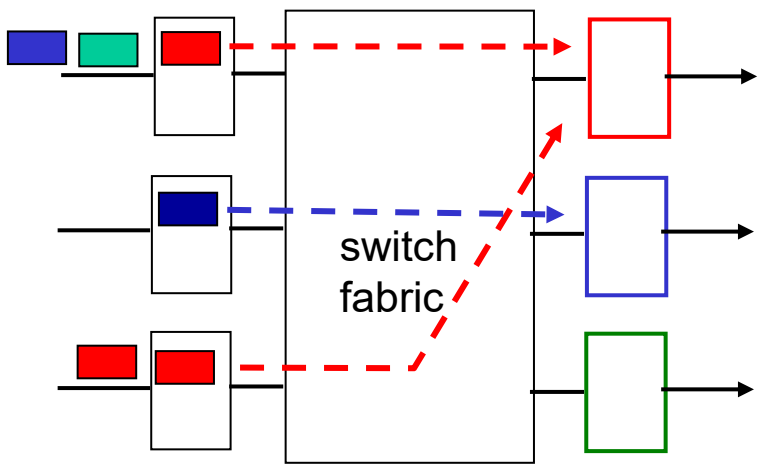
- *Buffering* richiesto quando i datagrammi arrivano dalla struttura di commutazione più velocemente del tasso di trasmissione del collegamento. *Drop policy:* quale datagramma scartare se il buffer non è sufficiente?

→ I datagrammi possono essere persi a causa di congestione, mancanza di buffer

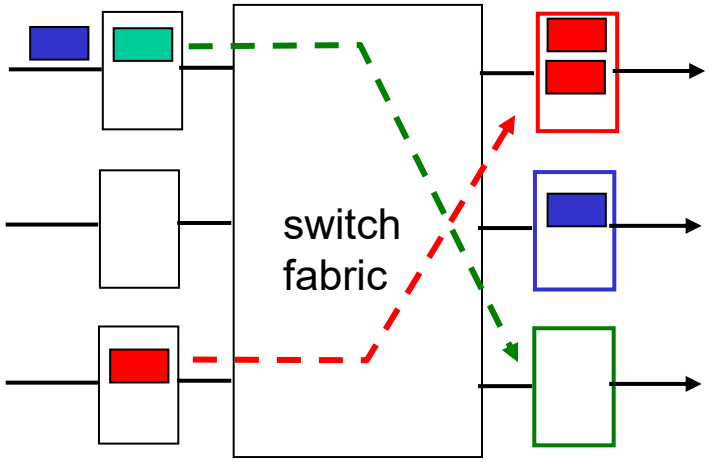
- *Disciplina di scheduling* sceglie tra i datagrammi in coda quale trasmettere

→ Schedulazione con priorità - chi ottiene le migliori prestazioni, neutralità della rete

Accodamento in uscita



contenuto della porta di uscita all'istante t



una unità di tempo più tardi

- buffering quando il tasso di arrivo attraverso la struttura di commutazione supera la velocità delle linea di uscita
- *accodamento (ritardo) e perdite causata dall'overflow del buffer della porta di uscita!*

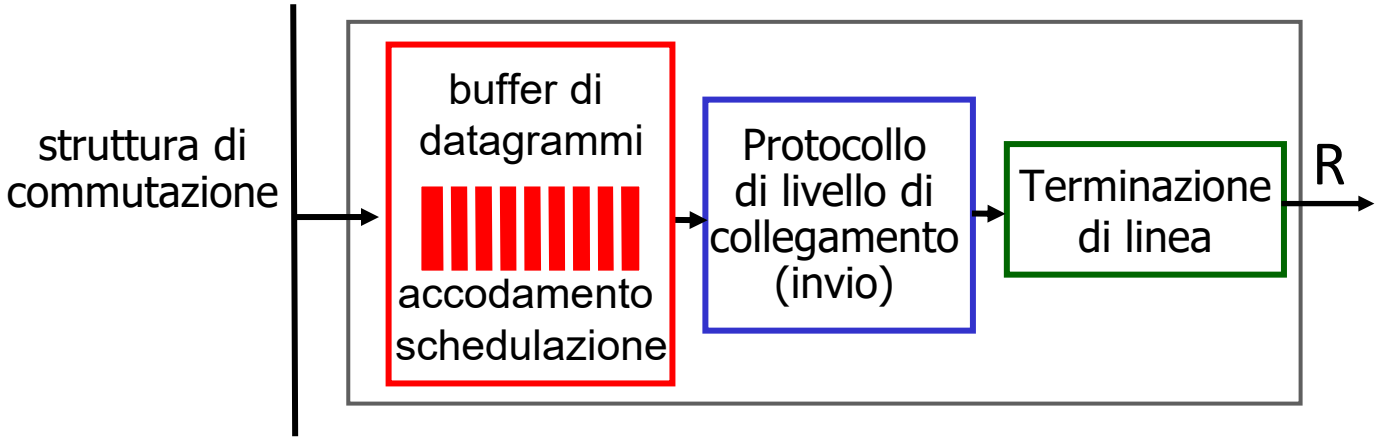
Quanta memoria di buffer è necessaria?

- RFC 3439 rule of thumb: buffering medio uguale al prodotto del RTT "tipico" (diciamo 250 ms) per la capacità del collegamento C
  - es., capacità del collegamento C = 10 Gbps: buffer di 2.5 Gbit
- raccomandazione più recente: con *N* flussi, dimensione del buffer

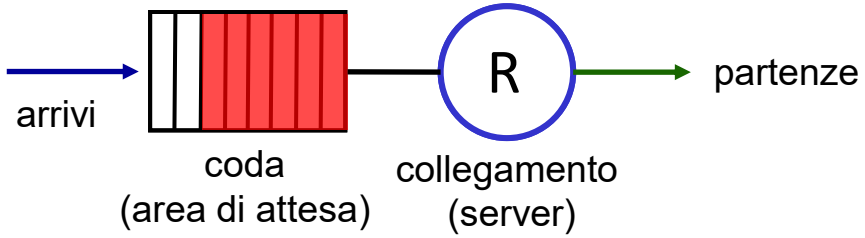
$$\frac{RTT \cdot C}{\sqrt{N}}$$

- ma *troppo* buffering può aumentare i ritardi (soprattutto nei router domestici)
  - RTT elevato: prestazioni scarse delle applicazioni real-time, mittenti TCP meno reattivi alla congestione e alla perdita dei pacchetti
  - ricordiamoci del controllo di congestione basato sul ritardo: "mantenere il collegamento collo di bottiglia sufficientemente pieno (occupato) ma non più pieno"

Gestione del buffer



Astrazione: coda



gestione del buffer:

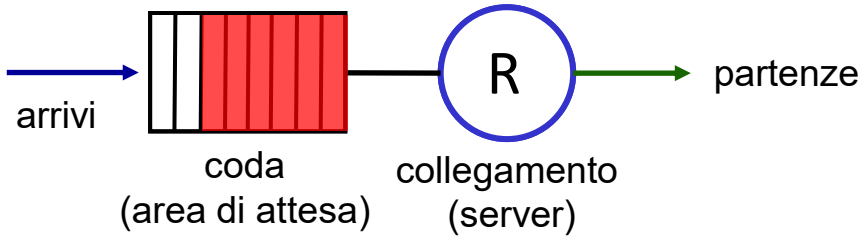
- politica di scarto (drop): quale pacchetto eliminare quando la coda è piena
  - tail drop: scarta il pacchetto in arrivo
  - priorità: scarta/rimuovi in base alla priorità
- marcatura: quali pacchetti marcare per segnalare la congestione (ECN, RED)

Schedulazione dei pacchetti: FCFS

Schedulazione dei pacchetti: decidere quale pacchetto inviare successivamente sul collegamento

- first come, first served
- priority
- round robin
- weighted fair queueing

Astrazione: coda



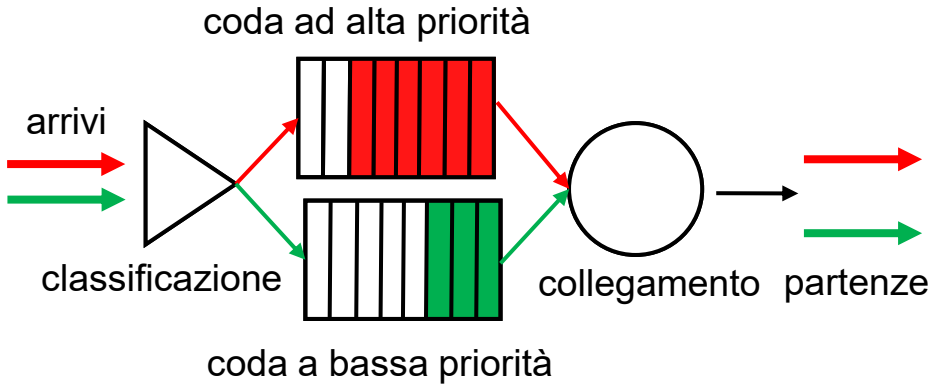
FCFS: pacchetti trasmessi in ordine di arrivo alla porta di uscita

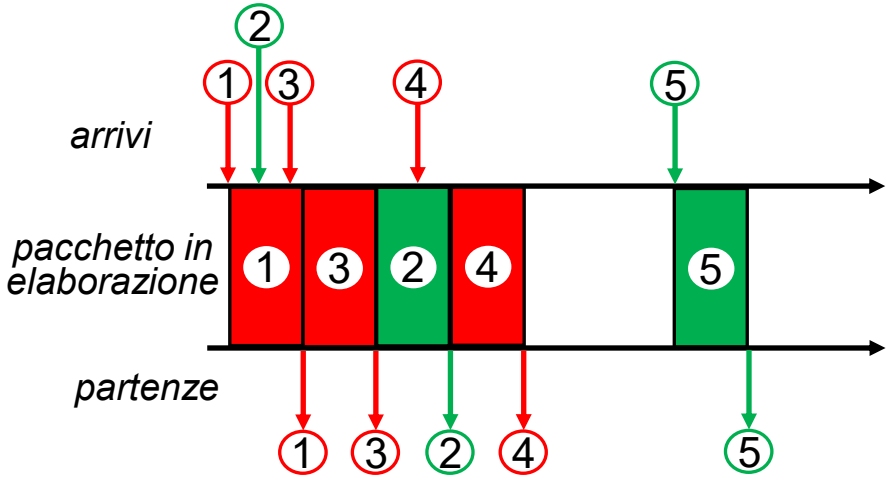
- conosciuta anche come: First-in-first-out (FIFO)
- esempi del mondo reale?

Schedulazione dei pacchetti : priority

Schedulazione con priorità:

- traffico in arrivo classificato, accodato per classi
  - qualsiasi campo di intestazione può essere usato per la classificazione
- invia il pacchetto dalla coda non vuota con priorità piùà alta
  - FCFS all'interno di ciascuna classe



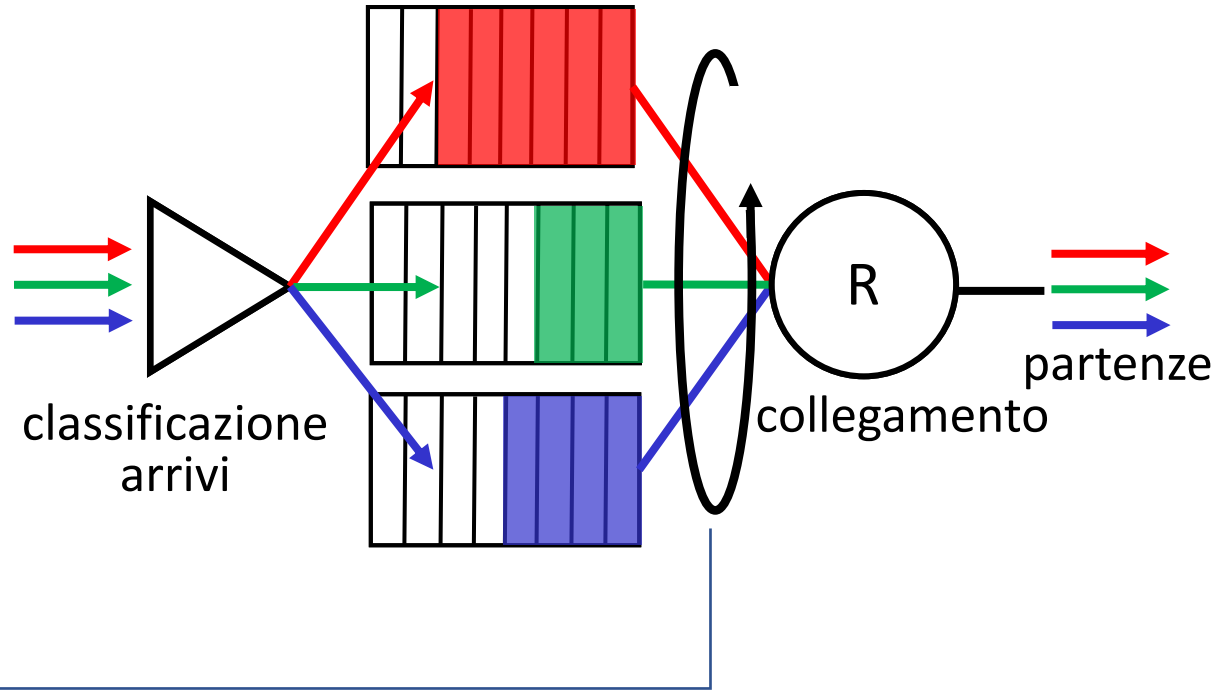


accodamento a priorità senza prelazione
(*non-preemptive priority queuing*)

Schedulazione dei pacchetti: round robin

Round Robin (RR) scheduling:

- Traffico in arrivo classificato, accodato per classi
  - qualsiasi campo di intestazione può essere usato per la classificazione
- Il server esegue ciclicamente e ripetutamente la scansione delle code di classe, inviando a turno un pacchetto completo di ogni classe (se disponibile).



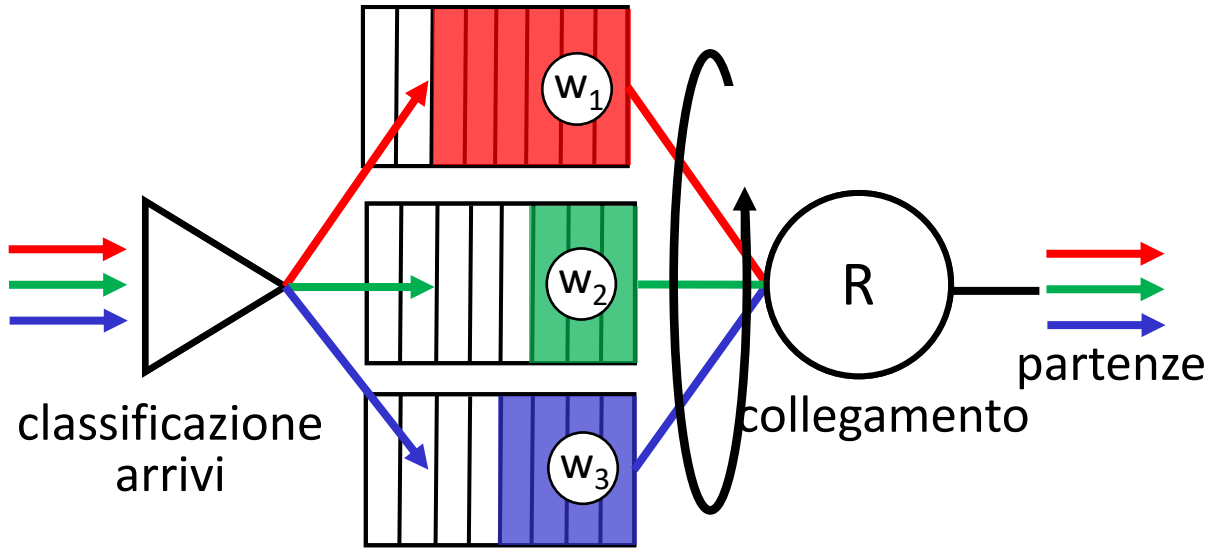
Schedulazione dei pacchetti: weighted fair queueing

Weighted Fair Queuing (WFQ):

- generalizza Round Robin
- Ciascuna classe, *i,* ha un peso, w_i, e riceve una quantità ponderata di servizio in ogni ciclo :

$$\frac{w_i}{\sum_j w_j}$$

- garanzia di larghezza di banda minima (per classe di traffico)



Barra laterale: Neutralità della rete

Cos'è la neutralità della rete (*net neutrality*)?

- *tecnica:* come un ISP dovrebbe condividere/allocare le proprie risorse
  - la schedulazione dei pacchetti e la gestione dei buffer sono i *meccanismi*
- Principi *sociali e economici*
  - proteggere la libertà di espressione
  - Incoraggiare l'innovazione, la competizione
- Far rispettare *politiche* e *leggi*

Ogni paese ha il proprio approccio alla neutralità della rete

Barra laterale: Neutralità della rete

2015 US FCC *Order on Protecting and Promoting an Open Internet:* tre regole definite "clear, bright line":

- no blocking … "non bloccherà i contenuti, le applicazioni, i servizi o i dispositivi non dannosi leciti, fatta salva una ragionevole gestione della rete."
- no throttling … "non devono pregiudicare o degradare il traffico Internet lecito sulla base del contenuto, dell'applicazione o del servizio Internet o dell'uso di un dispositivo non dannoso, fatta salva una ragionevole gestione della rete."
- no paid prioritization. … "non deve impegnarsi nella prioritizzazione a pagamento"

Nel 2017, la *Restoring Internet Freedom Order* ha annullato questi divieti, concentrandosi invece sulla trasparenza degli ISP.

ISP: telecommunications or information service?

Un ISP è un "servizio di telecomunicazioni" o un fornitore di "servizi di informazione"?

- la risposta *è* importante dal punto di vista normativo!

US Telecommunication Act del 1934 e 1996:

- *Titolo II:* impone “common carrier duties” ai *servizi di telecomunicazione*: tariffe ragionevoli, non discriminazione e richiede una regolamentazione
- *Titolo I:* si applica ai *servizi di informazione:*
  - no common carrier duties (*non regolamentato*)
  - ma concede alla FCC l'autorità "... necessaria per l'esecuzione delle sue funzioni "