Il presente e il futuro delle Reti Locali

Silvano Gai

Consulente

Le spinte per l’innovazione

- Le spinte per l’innovazione delle reti locali vengono da tre aree principali:
  - Datacenter
  - Ethernet in area metropolitana
  - Residenziale
- Questa presentazione si concentra sul Datacenter, dove le reti locali ignorano/convivono con le reti per lo storage

Le reti nel datacenter oggi

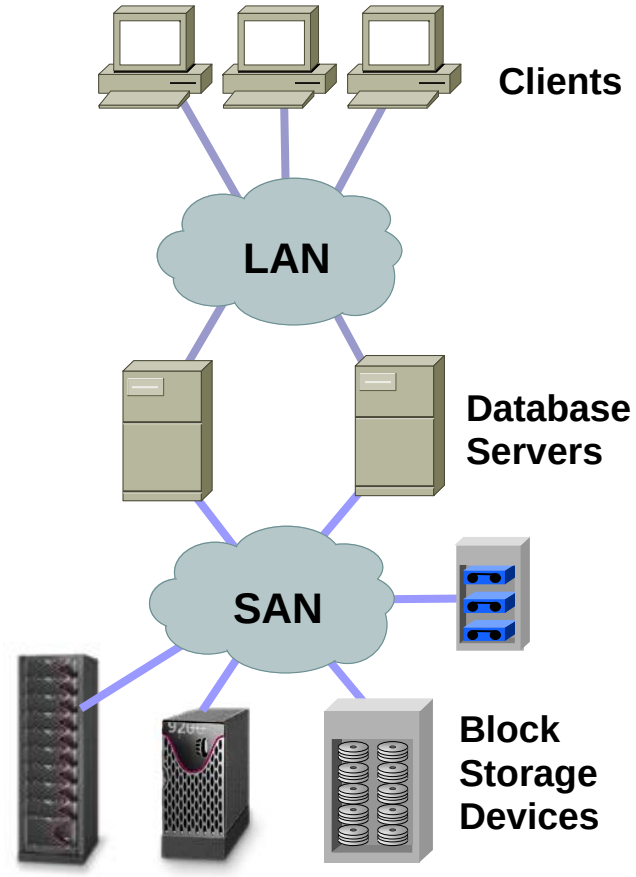
- Ethernet
- Fibre Channel
- Infiniband
- Reti per cluster proprietarie
  - Quadrics
  - Myrinet

Spinte nel Datacenter

- Adottare i modelli 1 RU e Blade server
- Bilanciare il costo del server e delle connessioni
- Utilizzare PCI Express come I/O a 10Gb/s
- Consolidare su un unico I/O esterno il traffico Ethernet, Fibre Channel, Cluster Networks, e FICON
  - costi di acquisizione e gestione
  - spazio occupato e potenza dissipata
- Infrastruttura comune, indipendente dal numero crescente di applicazioni e utenti
  - capacita' di commutazione di molti Terabit/s
  - grande attenzione all'alta affidabilita'

Fibre Channel & SAN

- SAN (Storage area Network): una rete di interconnessione ad alte prestazioni in grado di fornire un alto volume di I/O
- Storage viene visto a livello di blocco su disco tramite il protocollo SCSI
- Riduce i costi di gestione rispetto a storage locale, permettendo di condividere lo storage e di gestirlo centralmente
- Le reti SAN sono oggi quasi totalmente realizzate tramite Fibre Channel
  - L'interoperabilita' multivendor non e' ancora una realta'



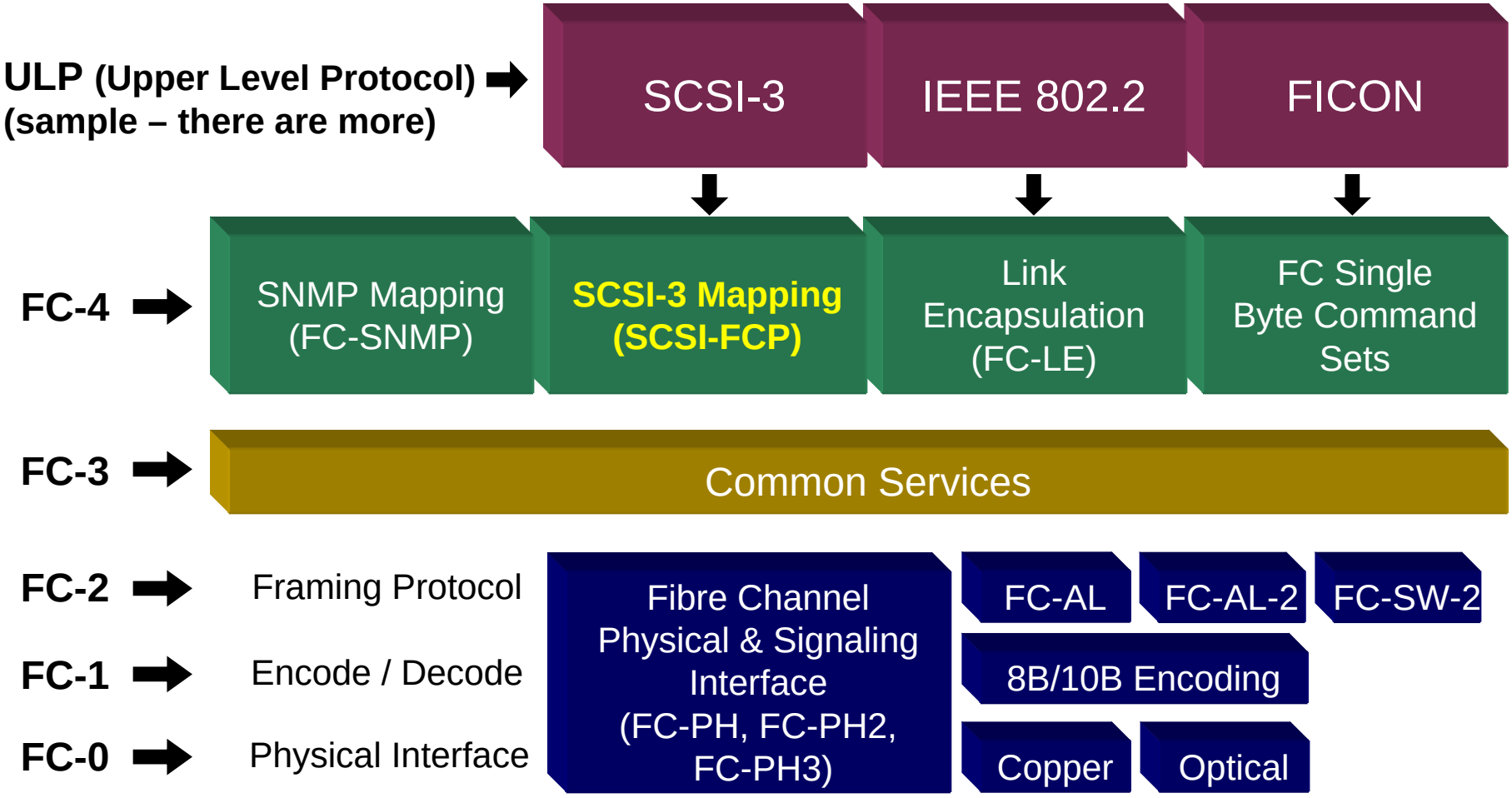
Storage Area Network (SAN)

Perche’ Fibre Channel?

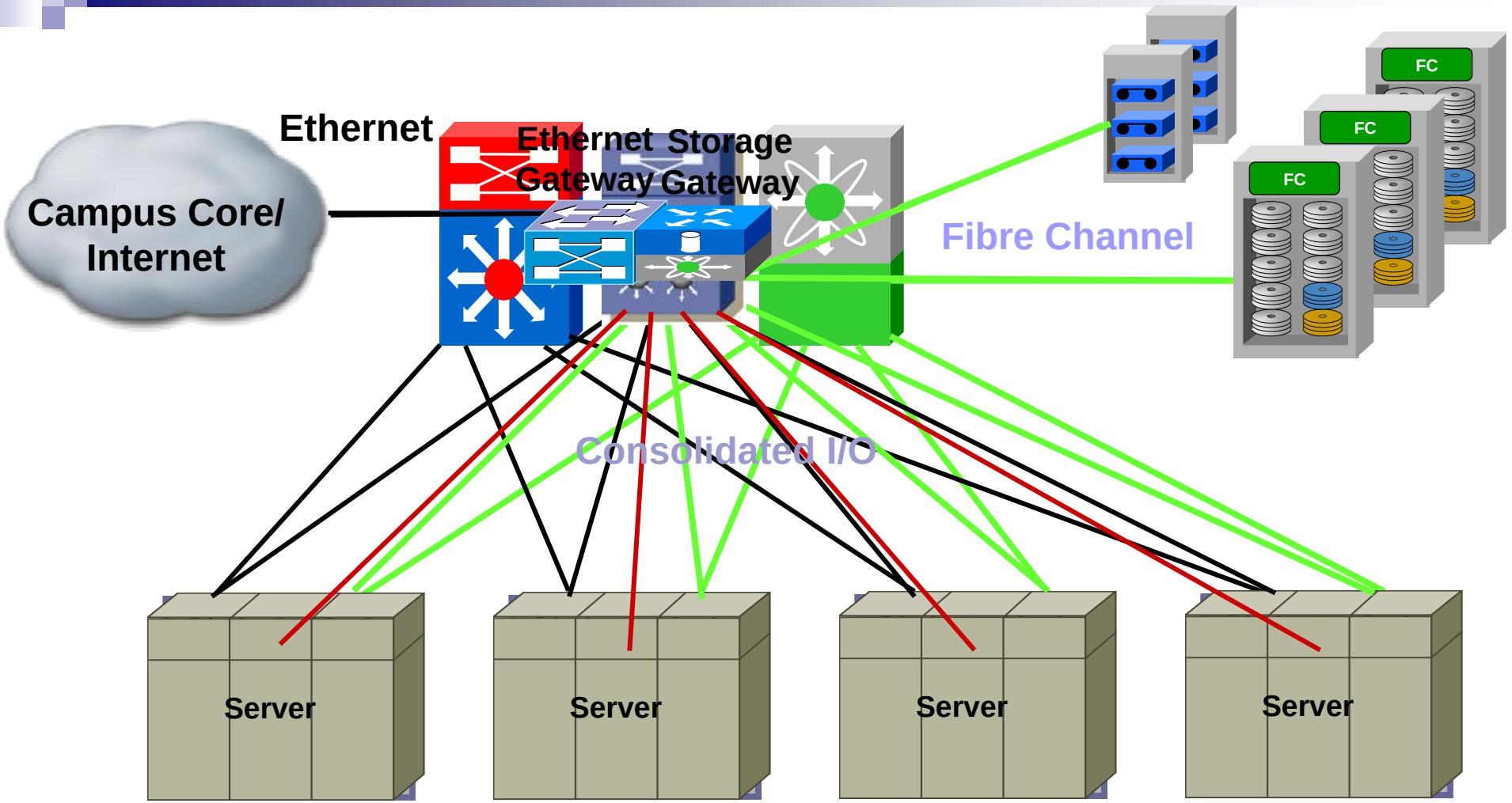
- Negli anni 1990 Fibre Channel era a 1 Gb/s, mentre Ethernet era a 100 Mb/s
- Fibre Channel non perde i pacchetti(*) e quindi SCSI puo’ funzionare su Fibre Channel senza richiedere modifiche
- Le reti che non perdono i pacchetti hanno problemi quali:
  - Deadlock/Livelock
  - Head of Line blocking
- Questi problemi limitano la scalabilita’ delle reti, ma sono stati considerati praticamente accettabili in Fibre Channel

(*) il termine piu’ corretto e’ trame, ma pacchetti e’ il termine usato colloquialmente

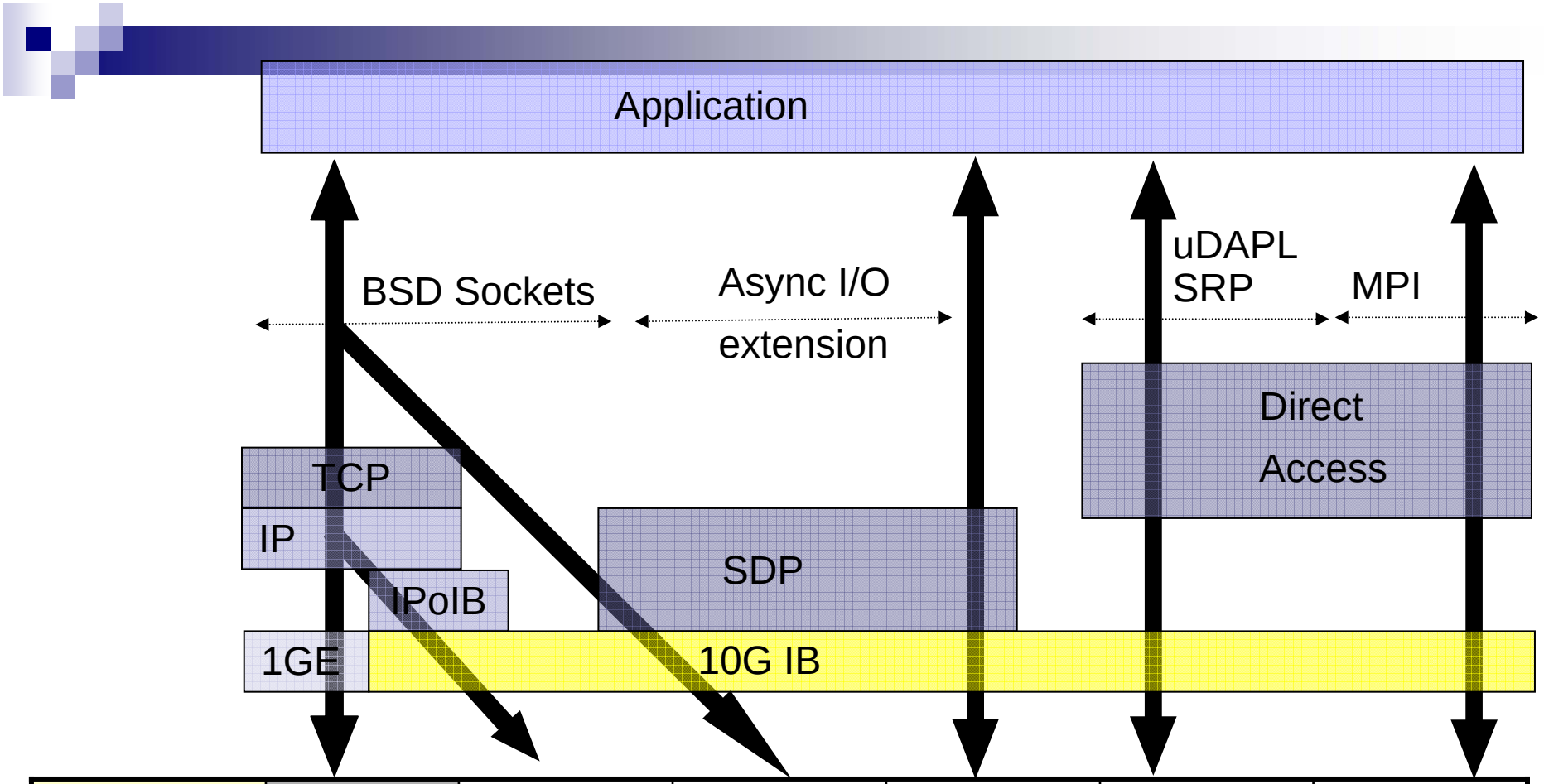
Architettura Fibre Channel



Evoluzione del Data Center



Il tentativo di Infiniband



Throughput	1-3 Gb/s	4.1Gb/s	4.5 Gb/s	7.9 Gb/s	8 Gb/s	8 Gb/s
Latency	40-60 usec	30 usec	18 usec	18 usec	8 usec	5 usec

Vantaggi/Svantaggi di Infiniband

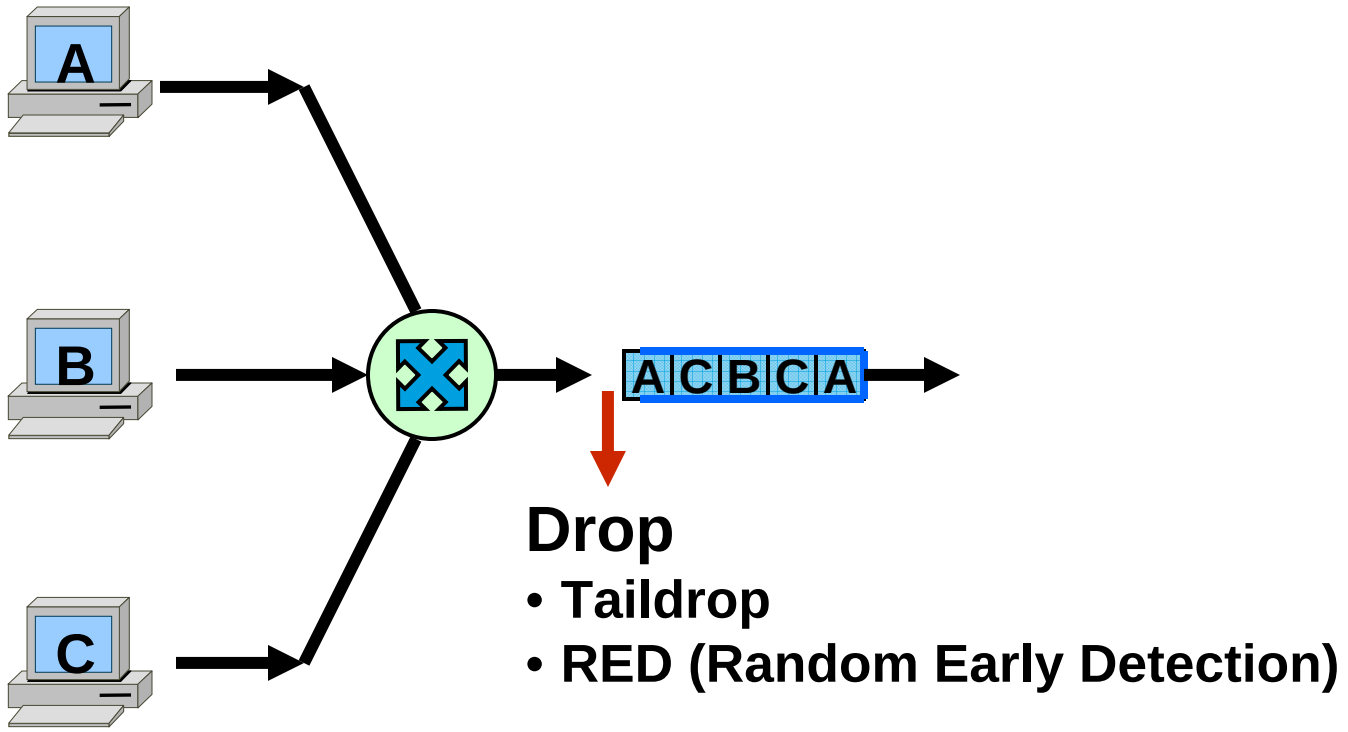
- Vantaggi
  - IO consolidato
  - Alta velocita’
  - Bassa latenza
  - Basso Costo
- Svantaggi
  - Una nuova rete
  - Difficile compatibilita’ con Ethernet e Storage
    - Gateway “pesanti”
  - Mezzo fisico non standard
  - Driver non certificati

Ethernet

- Ieri
  - Mezzo trasmissivo condiviso
  - Collisioni e ritrasmissioni
  - Ritardo di propagazione teoricamente non limitato
  - Pacchetti limitati a 1500 bytes
- Oggi
  - Punto-punto full-duplex
  - Pacchetti Jumbo (9 Kbytes)

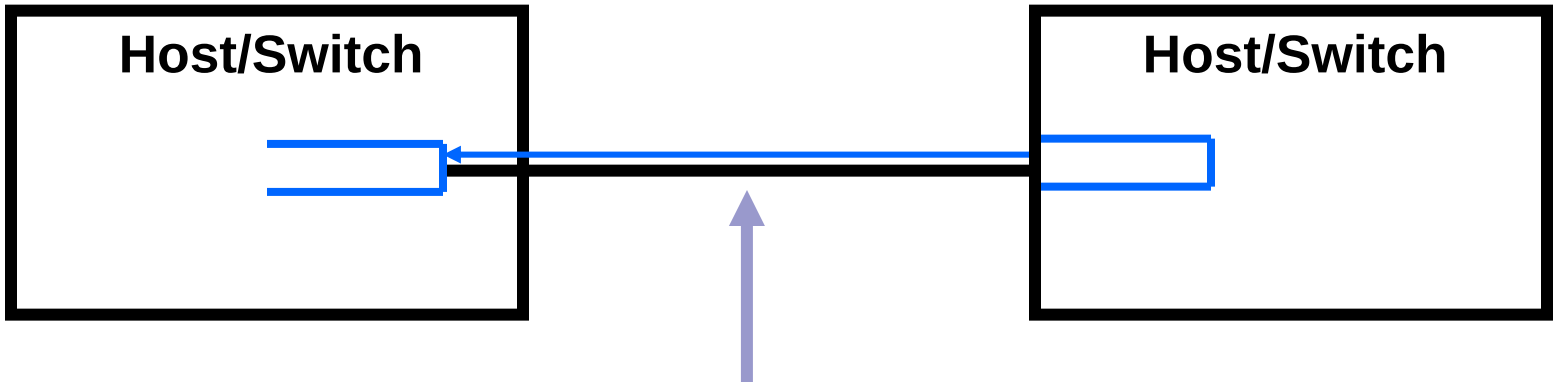
Perche' le reti locali perdono i pacchetti

- Per l'overflow delle code in presenza di congestione
  - Utilizzato da TCP/IP per gestire la congestione



Come evitare l'overflow delle code

- Fibre Channel e Infiniband
  - Il controllo di flusso a livello link e' obbligatorio:
    - Prende la forma di crediti tra i buffer
  - Permette di utilizzare buffer piccoli (8-32 pacchetti) e quindi poca memoria integrabile negli ASICs.
- Ethernet
  - E' definito come Pause Frame, ma non utilizzato



Inviare dei feedback alla sorgente di un link indicando lo stato del buffer a valle

Come far evolvere Ethernet

- Una evoluzione, non una rivoluzione
  - Alta Velocita’ (10Gb/s),
  - Bassa Latenza (pochi microsecondi),
  - Costo piu’ basso e competitivo
  - Aggiungere il concetto di Virtual Lane
  - Aggiungere il concetto di Layer 2 Multipath
- Gruppi attivi
  - IEEE 802.3 e 802.1
  - IETF TRILL

Virtual Lane

- Estensione del concetto di priorita’
  - 8 VL ciascuna con risorse dedicate
    - e.g. input buffer e output queue
  - L’indicatore di appartenenza e’ nel pacchetto
- Il comportamento della rete e’ negoziabile per VL

Alta Velocita’

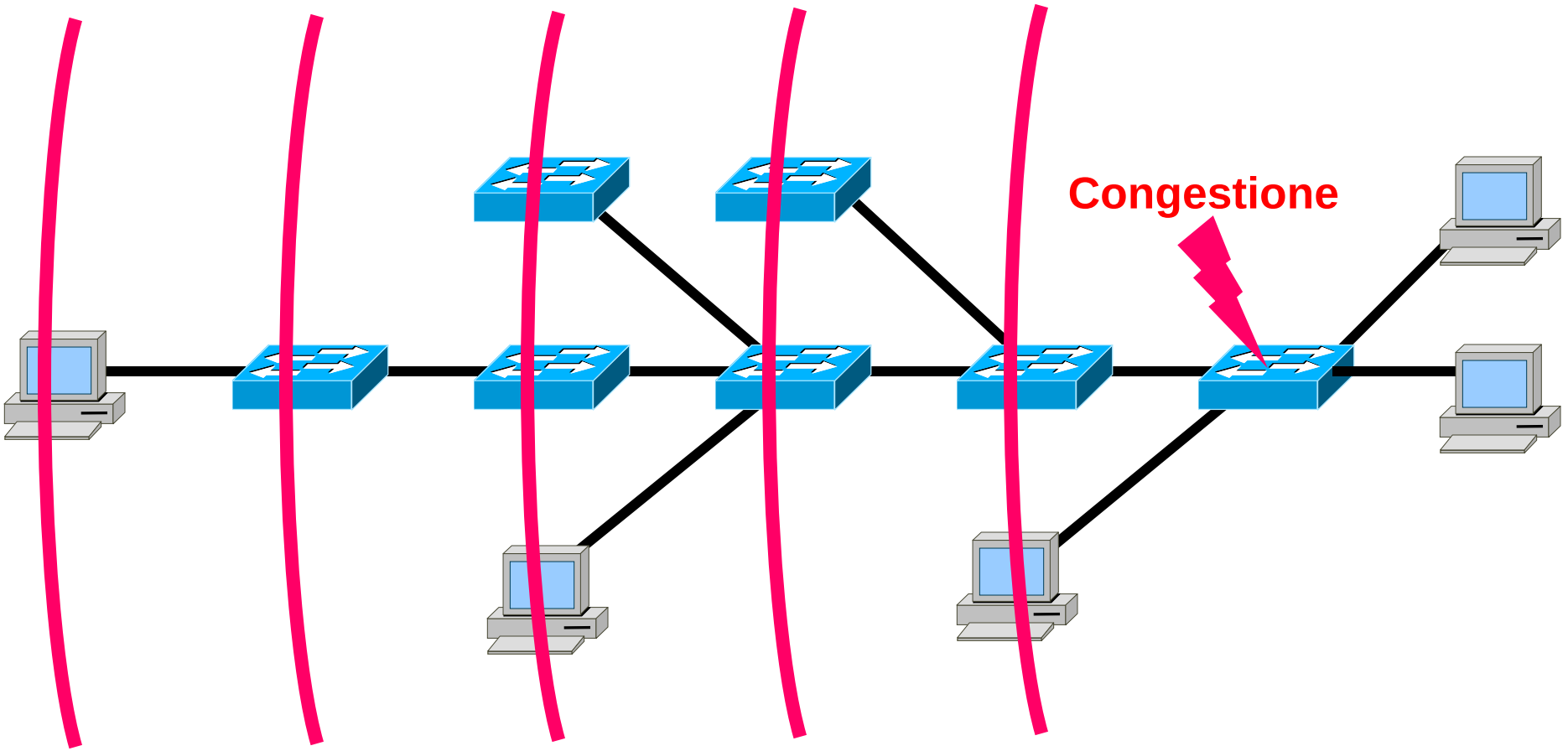
- Ethernet evolve storicamente per multipli di 10
  - 10 GE e’ 10 Gb/s a livello data link layer
- I costi alti hanno tre ragioni:
  - Buffer
  - Mezzo trasmissivo
  - Struttura di forwarding

Buffer

- Non si puo' certo pensare di dimensionare il buffer in accordo alla classica formula pubblicata sui libri
  - B = RTT * C
  - Per esempio, RTT = 100 ms, C = 10 Gb/s
    - B = 1 Gb, circa 100MB di buffer esterno per porta
  - Studi piu' recenti quali quelli di Appenzeller et al. propongono
    - B = RTT * C/$\sqrt{N}$, dove N e' il numero di flussi
  - Aiutano molto in ambiente service provider, ma non in ambiente Datacenter
    - La sfida e' condividere efficientemente pochi MB d memoria on chip tra 16-32 porte!
- Il controllo della congestione aiuta a ridurre il buffer!

Controllo della Congestione

Necessita' di definire un meccanismo di controllo della congestione che limiti il traffico generato dalle sorgenti



Mezzo trasmissivo

- I transceivers ottici 10 GE sono costosi
- Realizzazioni in cavo di rame abbattono i costi
  - Costi contro distanze
  - IEEE 802.3ak aka 10Base-CX4
  - Alcune startup lavorano a 10GE su Cat.6/7
  - Altre startup lavorano a soluzioni in fibra poco costose

Bassa Latenza

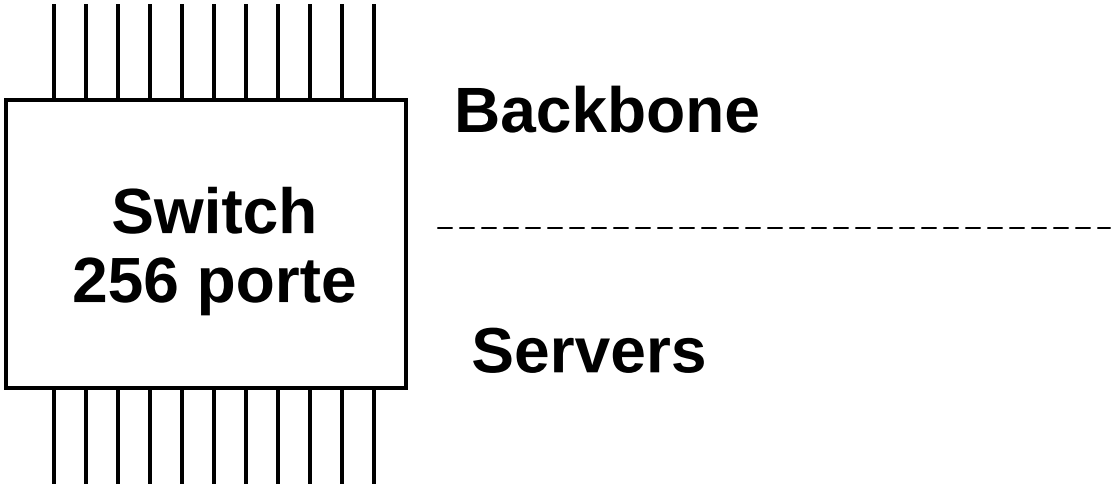
- Bisogna lavore a livello di architettura dello switch, ma soprattutto a livello di architettura dell'host:
  - Non si puo' fare nulla per ridurre il ritardo di serializzazione e di propagazione
- In presenza di congestione, la latenza nello switch e' dominata dal buffering:
  - In assenza di congestione, si possono usare tecniche di cut-through per ridurre la latenza dello switch
- Il programmatore e' interessato alla latenza da una memoria utente all'altra. Occorre un'ottima NIC con:
  - Terminazione in hardware dei protocolli
  - Efficace integrazione con il chip-set del processor
  - Meccanismi di Kernel bypass

Instradamento solo a livello 2

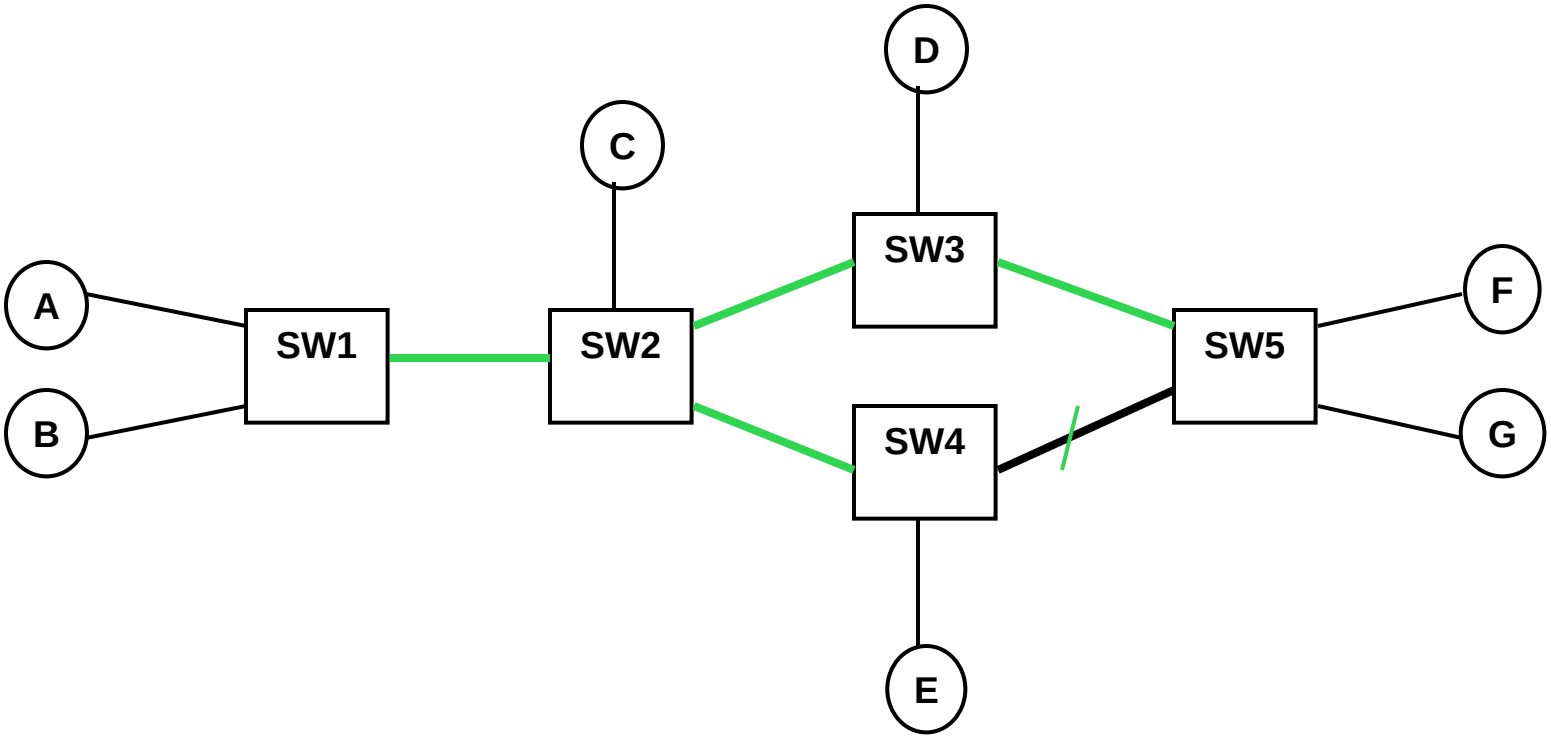
- Gli switch Ethernet disponibili sul mercato hanno centinaia di prestazioni non essenziali per Ethernet, ad esempio forwarding di livello 3 (IPv4 and IPv6), ACLs, layer 3 multicast, MPLS.
  - Queste prestazioni richiedono RAMs, CAMs, TCAMs esterne
- Gli switch per Metro-Ethernet hanno necessita' di un gigantesco database (>= 512 K entries)
- Switch per Datacenter possono evitare la maggior parte di queste prestazioni e quindi avere costi piu' bassi.

Considerazioni Topologiche

- 10 GE e’ eccezionale per i server, ma paradossalmente e’ un collo di bottiglia per il backbone!
- Le reti per il Datacenter non hanno la tipica struttura ad imbuto delle reti per il wiring closet:
  - Hanno necessita’ di una “high bisectional bandwidth”
- Tecniche di Link Aggregation (e.g Etherchannel) sono di aiuto, ma non risolutive

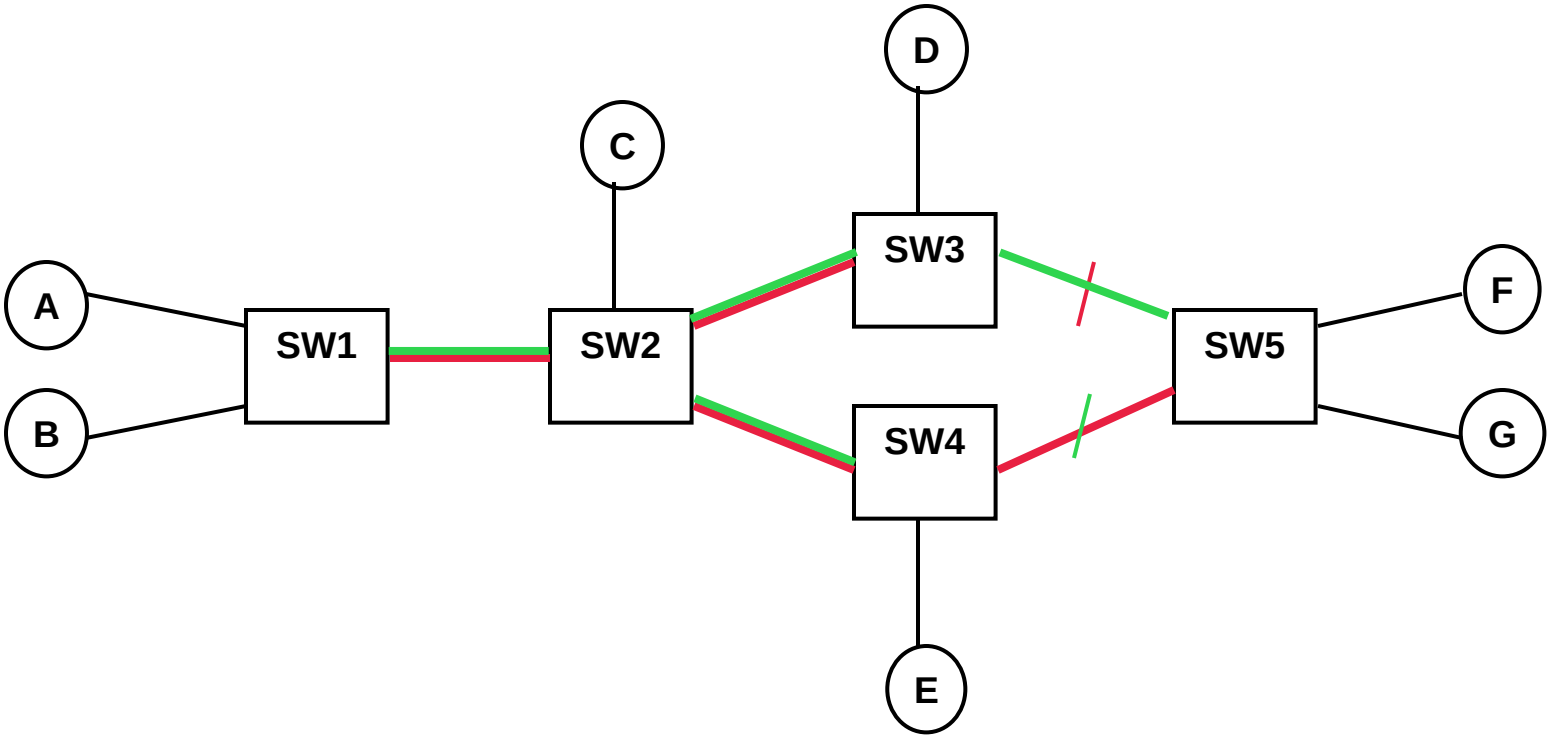


Classical Ethernet – Spanning Tree



- Il link tra SW4 & SW5 non e’ usato

Classical Ethernet – ST e VLAN

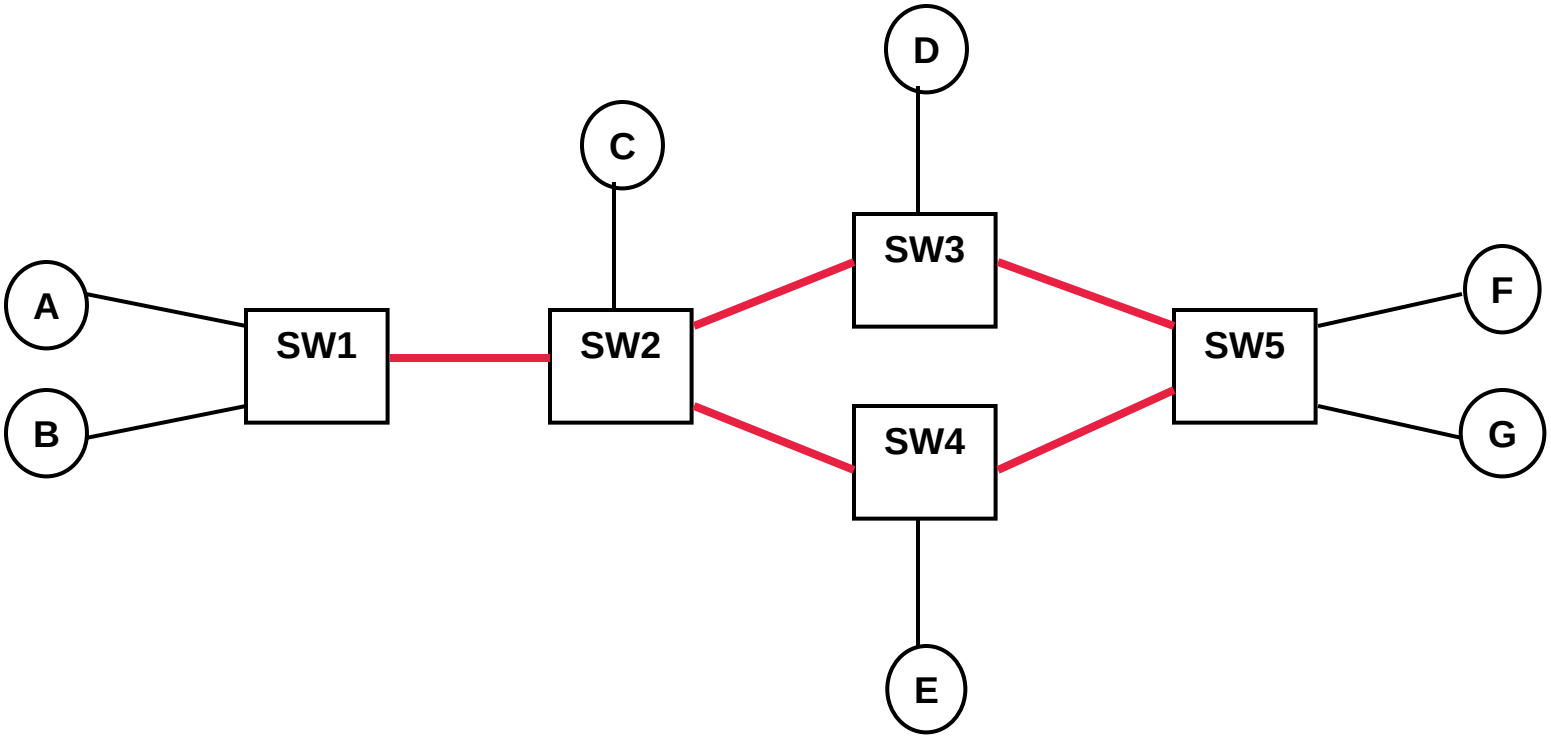


- Il link tra SW4 & SW5 e’ usato solo dalla VLAN rossa

Multi-path a livello 2

- Per superare i limiti di Spanning Tree, l'IETF ha creato il gruppo di lavoro TRILL
  - IS-IS viene utilizzato per l'instradamento a livello 2 del traffico unicast con destinazione nota
- Vantaggi
  - IS_IS supporta cammini paralleli di costo identico
- Limitazioni
  - Il traffico Unicast per destinazione ignota e il traffico multicast broadcast continueranno ad andare su alberi
  - Sfortunatamente gli indirizzi MAC non sono aggregabili

Multipathing per Unicast con destinazione nota



- Lo scopo e’ di utilizzare tutti i cammini per il traffico unicast verso destinazioni note.

Transporto affidabile su Ethernet

- RDMA richiede un trasporto affidabile:
  - TCP e' la risposta di IETF
    - Pesante e costoso da realizare in SW
      - I TOE a 10 Gb/s sono solo agli inizi
      - Intel IOAT
  - Un'alternativa e' un trasporto affidabile simile a quello realizzato da Infiniband
    - Finestra fissa, simile ad LLC2
    - Piu' facile da realizzare in HW
- Le nuove NIC devono realizzare in HW:
  - Un trasporto affidabile
  - Supporto per memoria virtuale (true zero copy)
  - Interfaccia MPI 2.0
  - Supporto per Storage

Conclusioni

- Ethernet e’ una rete locale che e’ riuscita ad evolversi nel corso degli anni
  - Ha incluso al suo interno buone idee derivate da altre reti
- Chiunque guardi ad Ethernet come ad una realta’ immobile non si rende conto del grande fervore di attivita’ che esiste dietro questo standard
  - Datacenter, Metro e Residential sono le tre aree principali di innovazione
- Sino a quando Ethernet continuera’ ad evolvere, i clienti continueranno a favorirla rispetto a nuovi standard

Grazie

Silvano Gai