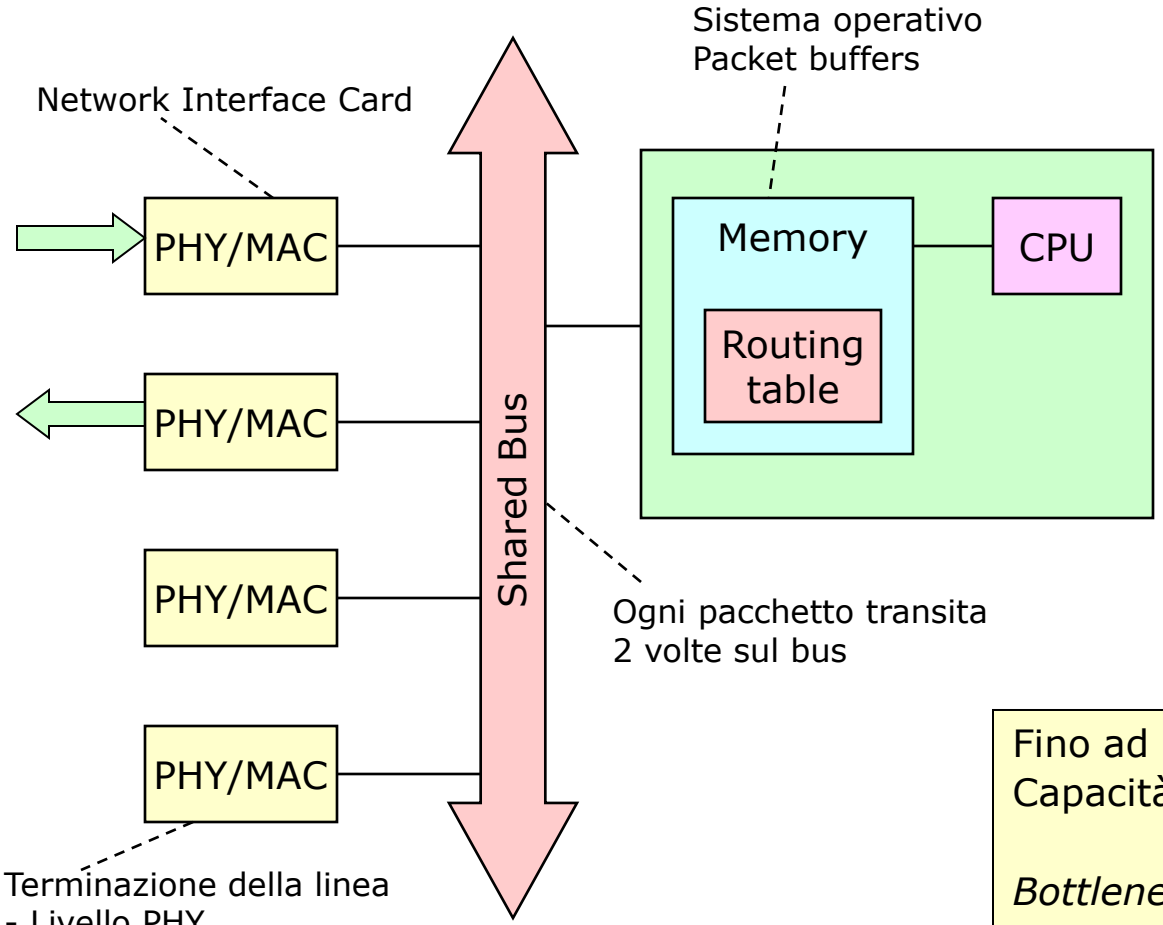
Cenni sull’architettura degli apparati di rete

Panoramica sulle principali architetture e problematiche degli apparati di rete

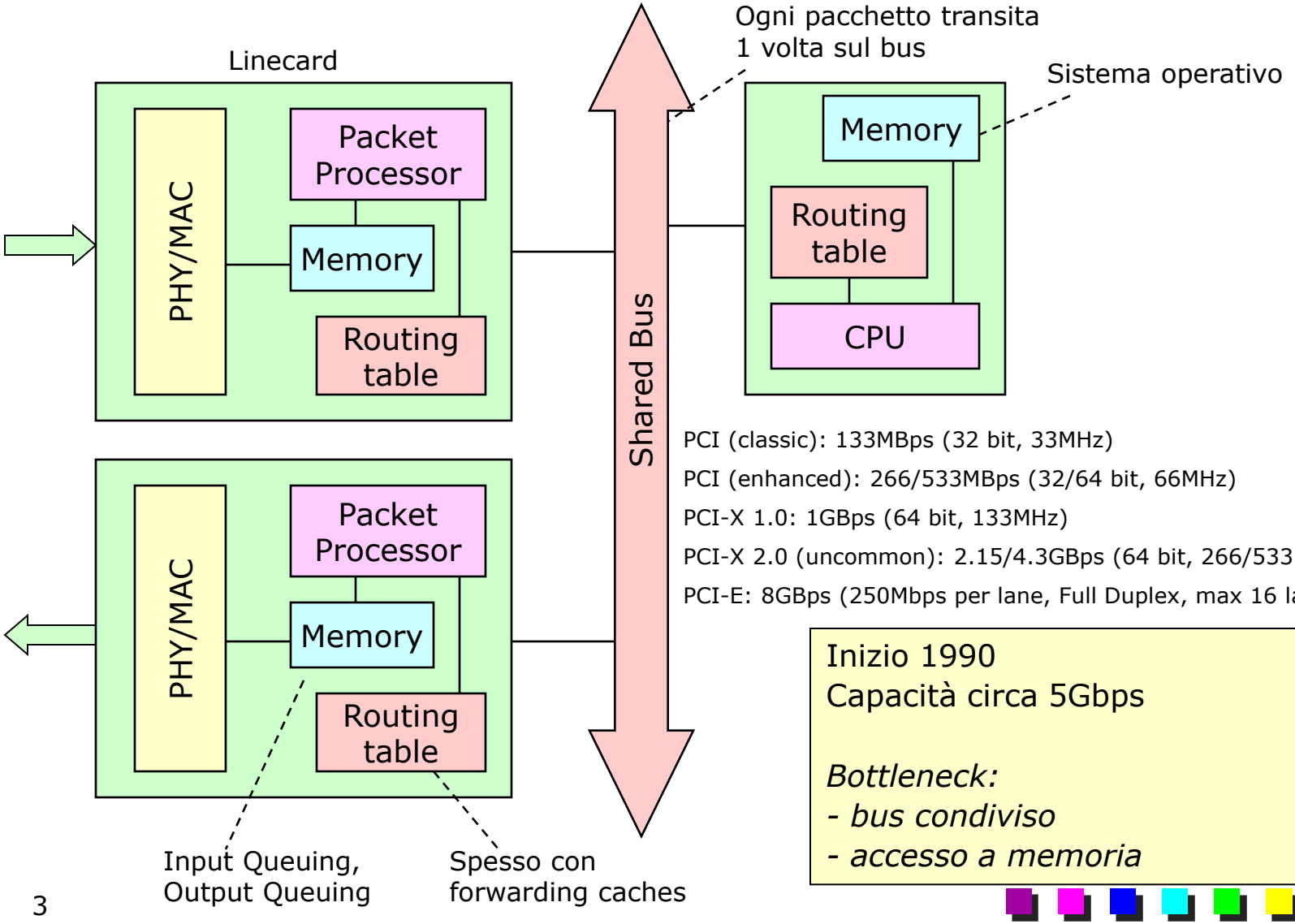
Router di 1ª generazione: PC modificati



Fino ad inizio 1990
Capacità inferiore a 500Mbps

Bottleneck:
- bus condiviso
- accesso a memoria
- capacità di processing

Router di 2ª generazione



PCI (classic): 133MBps (32 bit, 33MHz)

PCI (enhanced): 266/533MBps (32/64 bit, 66MHz)

PCI-X 1.0: 1GBps (64 bit, 133MHz)

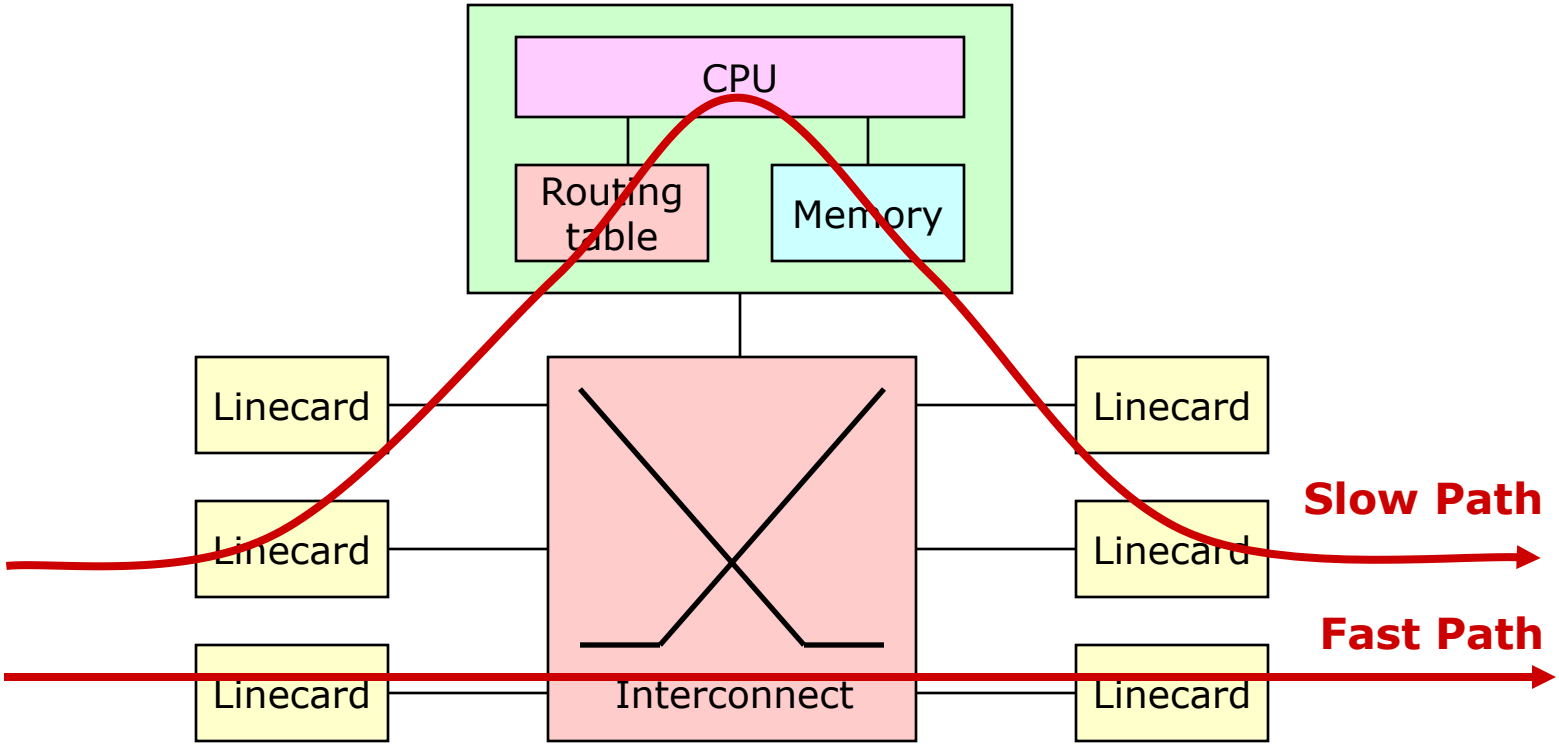
PCI-X 2.0 (uncommon): 2.15/4.3GBps (64 bit, 266/533MHz)

PCI-E: 8GBps (250Mbps per lane, Full Duplex, max 16 lanes)

Inizio 1990
Capacità circa 5Gbps

Bottleneck:
- bus condiviso
- accesso a memoria

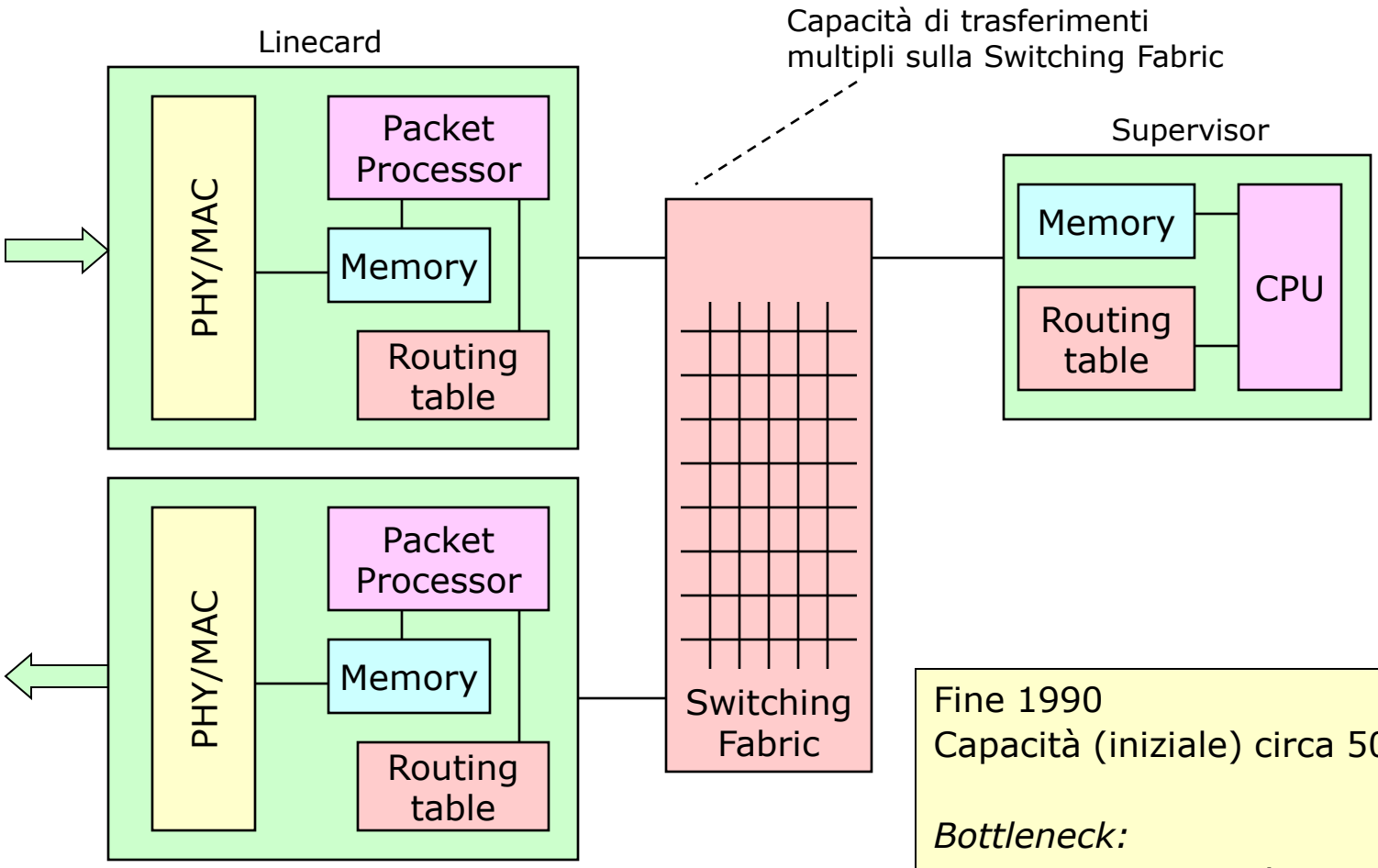
Router di 2ª generazione: fast/slow path



La capacità della card di controllo diventa secondaria

Gli sforzi vengono concentrati nell'ottimizzazione del *Fast Path*

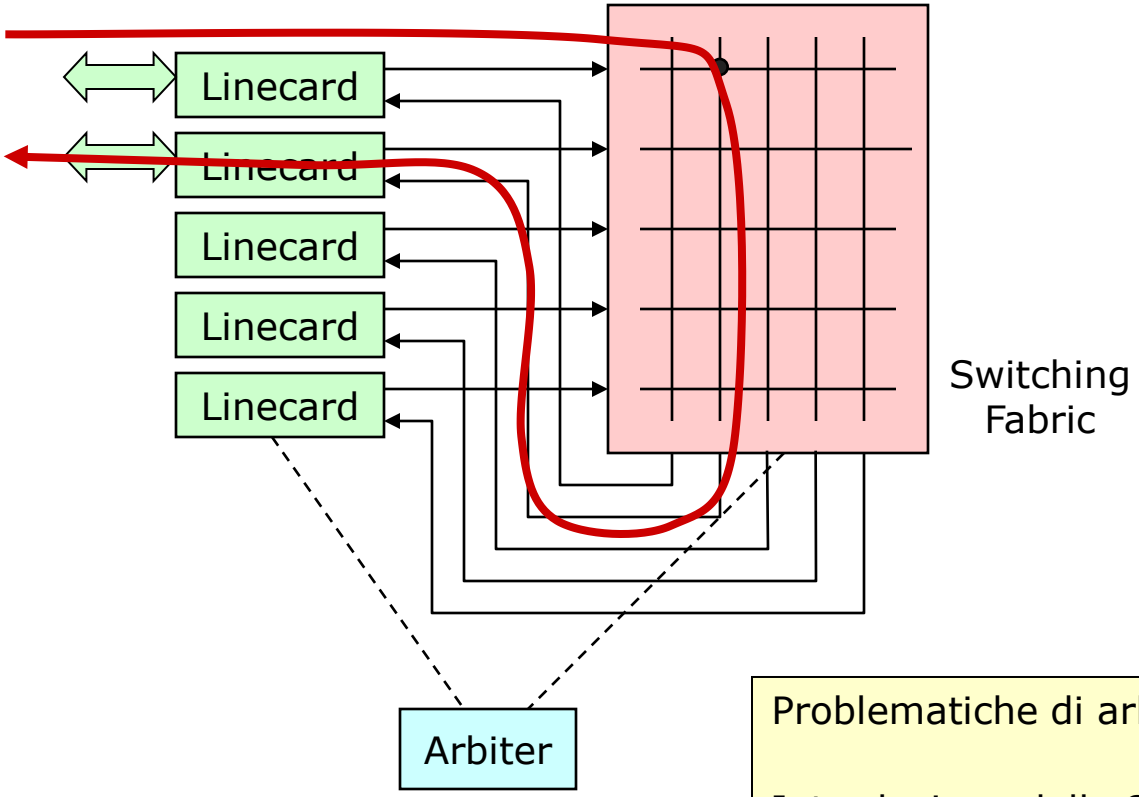
Router di 3ª generazione



Fine 1990
Capacità (iniziale) circa 50Gbps

Bottleneck:
- accesso a memoria
- processing (forwarding + …)

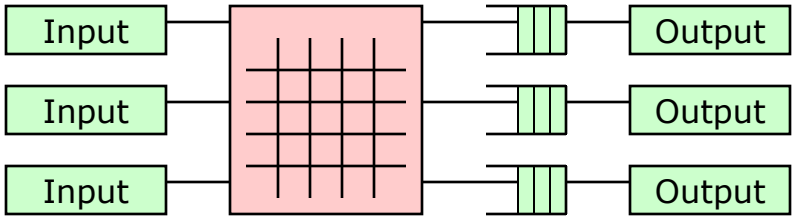
Switching Fabric (or Crossbar)



Problematiche di arbitraggio

Introduzione dello Switching Arbiter
- decisioni molto veloci (velocità pari al throughput aggregato)
- necessità di input/output buffer
- integrazione con meccanismi di QoS

Switching Fabric: Output Queuing

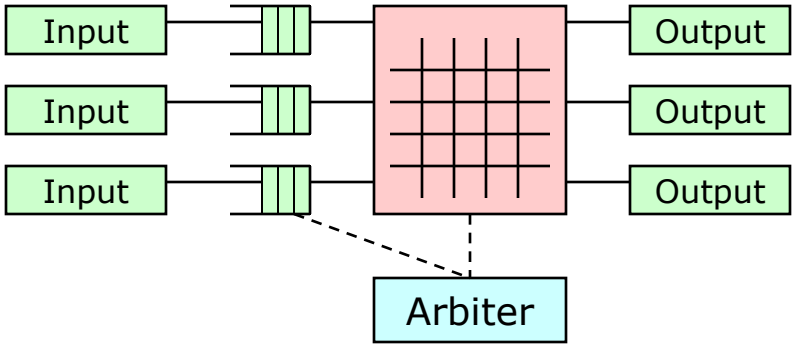


Switching Fabric Speedup: N
Velocità di accesso ai buffer: N*R

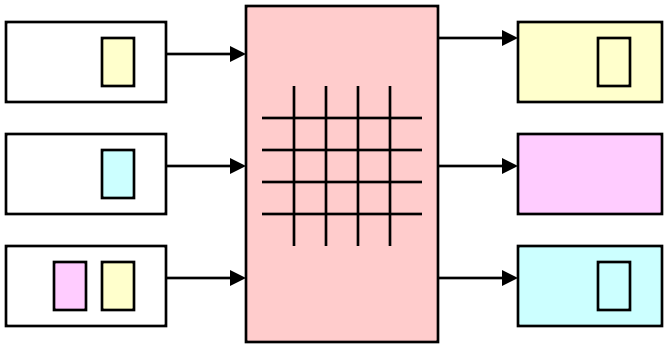
La velocità di accesso ai buffer può essere ridotta arbitrando l'accesso alla Switching Fabric
Questo porta a creare architetture con memorizzazione distribuita (output + qualche altro posto)
Altra soluzione: limitare la velocità di accesso ai buffer, e, in caso di contesa grave, scartare il pacchetto

Numero ingressi: N
Velocità di ogni ingresso: R

Switching Fabric: Input Queuing



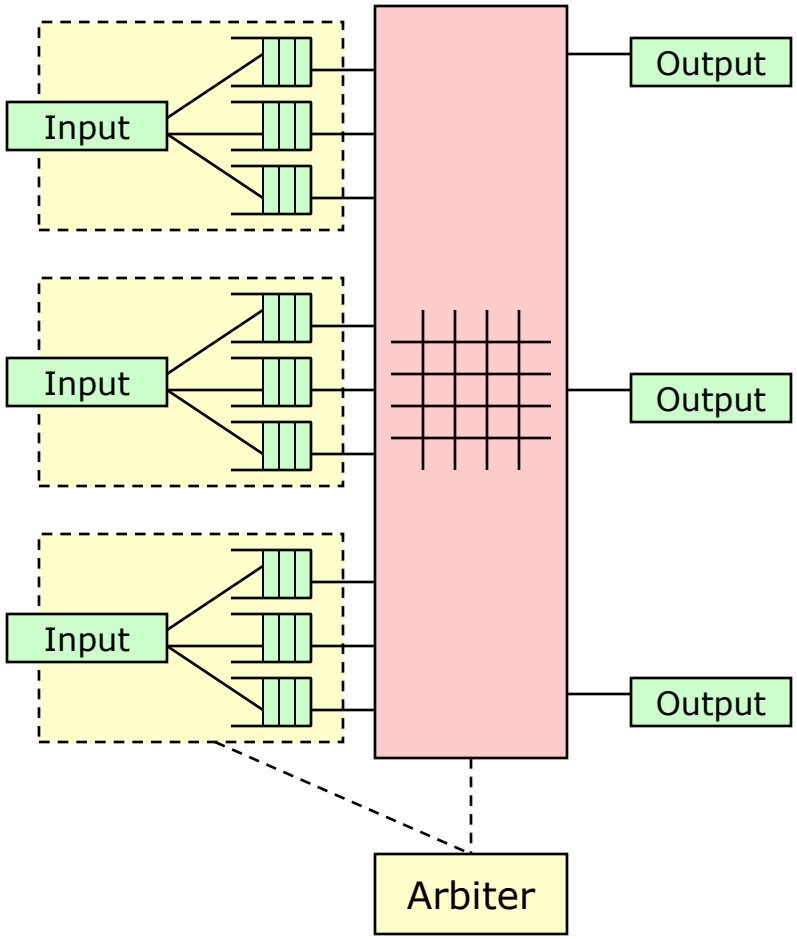
Switching Fabric Speedup: 1
(non dipende dal numero di porte)
Velocità di accesso ai buffer: R
Head-of-Line Blocking
Si dimostra che con pacchetti distribuiti uniformemente, l'utilizzazione max è 58.6%



HOL Blocking: il pacchetto fucsia non può essere trasmesso anche se la porta fucsia è libera, perchè il pacchetto giallo che lo precede è bloccato

Necessità di un arbitro per decidere quale dei pacchetti gialli ha diritto ad essere trasmesso
- random (necessario un random engine)
- round-robin (necessario un puntatore all'ultima uscita servita)
- longest queue (necessario indicatori di occupazione)
- least served (necessario un service time per input)

Switching Fabric: Virtual Output Queuing



Switching Fabric Speedup: 1
Velocità di accesso ai buffer: R

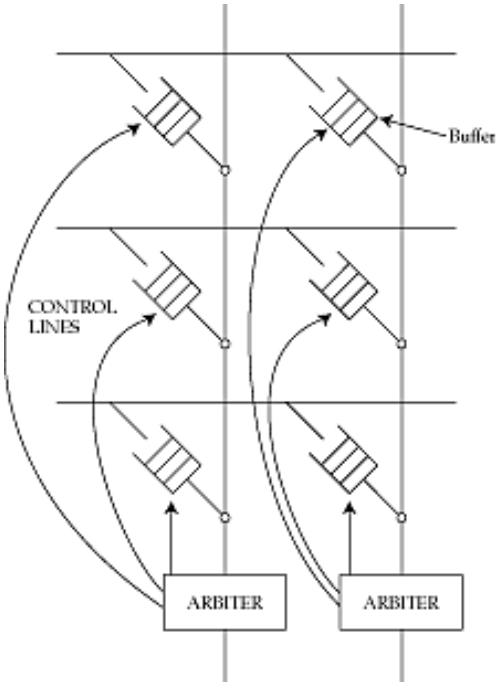
Risolve l'Head-of-Line Blocking

Ad ogni ingresso viene mantenuta una coda per uscita
– Ad ogni ciclo il controllore decide quali VOQ possono inoltrare un pacchetto e configura la crossbar
  – In IQ la scelta ad ogni ciclo è tra N pacchetti HoL
  – In VOQ è tra N^2 pacchetti HoL

– Algoritmo efficiente per configurare la CrossBar
  – Dato un grafo di richieste (da ogni ingresso fino a N archi) estrarre un sottografo non output blocking
  – Attenzione: anche in VOQ da ogni ingresso (con N HoL) può partire un solo pacchetto

Switching Fabric: Buffered Fabric

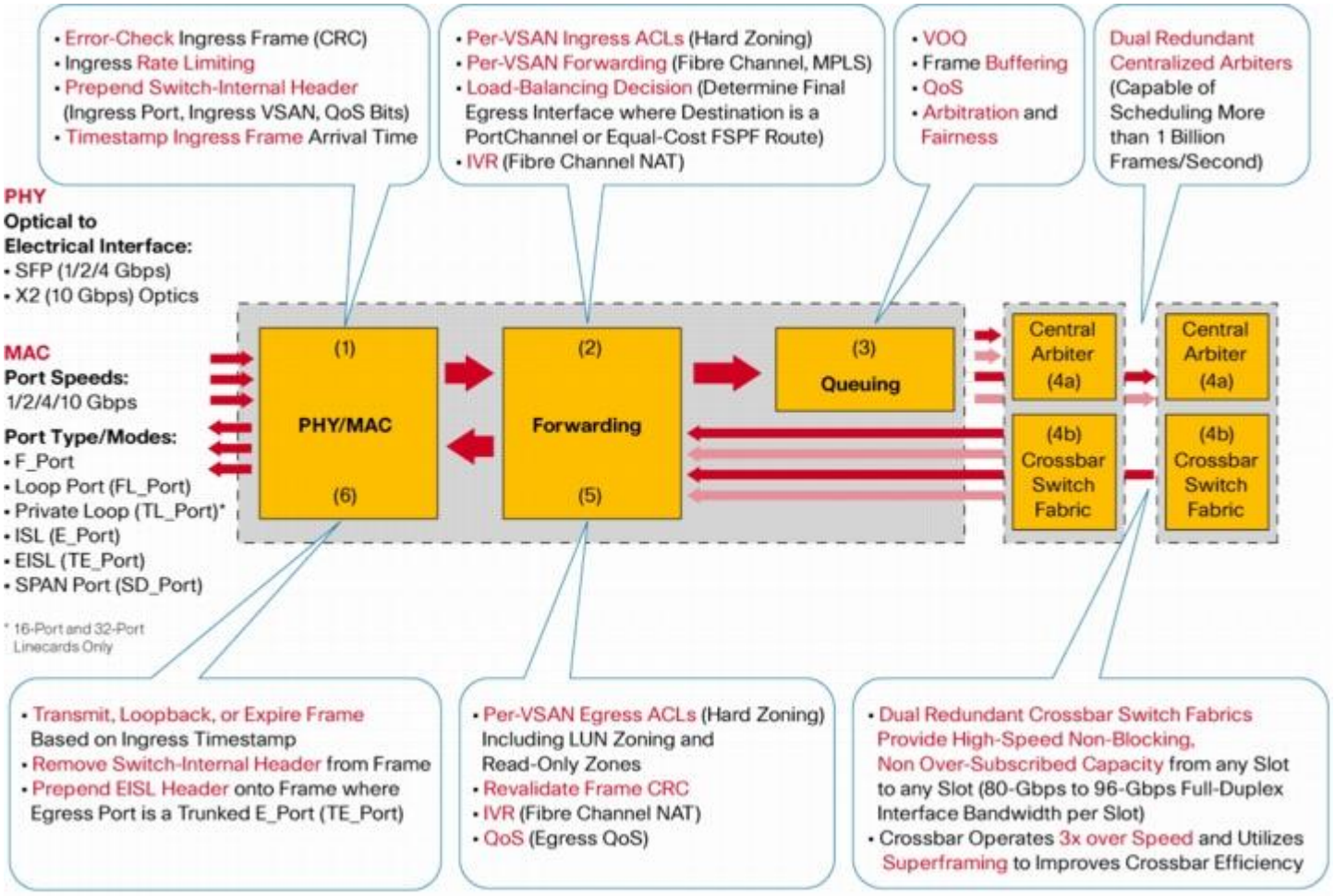
- Inserisce il buffering all'interno della crossbar
  - Se due input port vogliono accedere allo stesso output, uno dei due pacchetti verrà memorizzato nei buffer interni alla Fabric
  - L'arbitro dovrà essere in grado di pilotare la scelta
  - Criticità nella gestione di QoS (la scelta diventa complessa)
- Speed-up: può essere 1
- Costoso
  - a meno di memorie on-chip, che sono necessariamente piccole
- Soluzioni ibride (IQ, OQ, …)
  - Difficili da analizzare, ma comuni in pratica



More recent trends

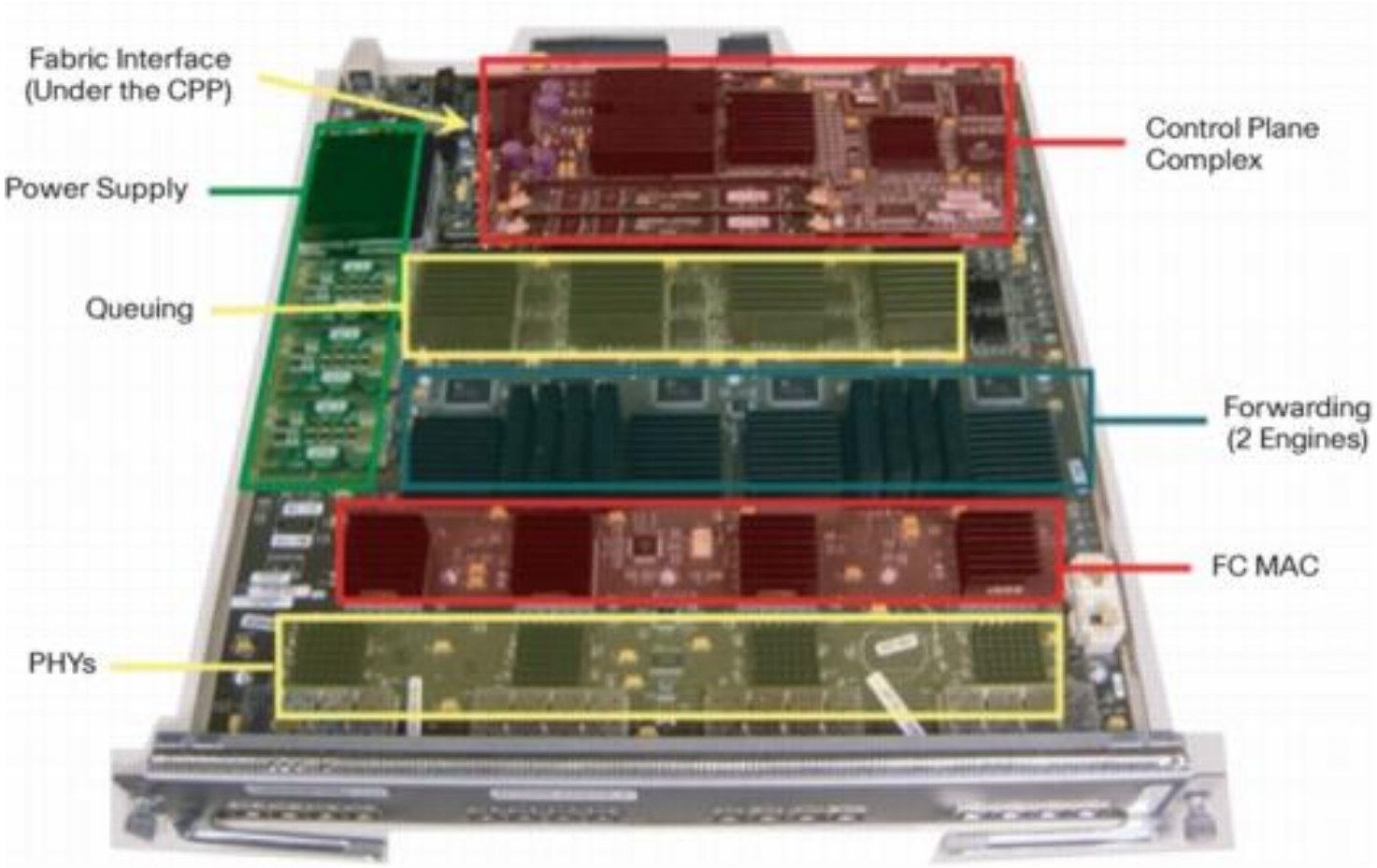
- Scalability
  - Multi-chassis routers
- Flexibility
  - By far the most important issue
  - Routers are no longer “pure” routers, as many other functions are being added that do not belong to L3
    - DHCP server, NAT, ACL (at layer 4 and beyond), firewalls, network monitors (e.g., NetFlow), etc.
  - Service cards are becoming increasingly common
    - Mostly for security purposes
    - Network processors or even general purpose processors
    - Reduce the total cost of ownership compared to discrete (dedicated) appliances

Linecards: example (logical view)



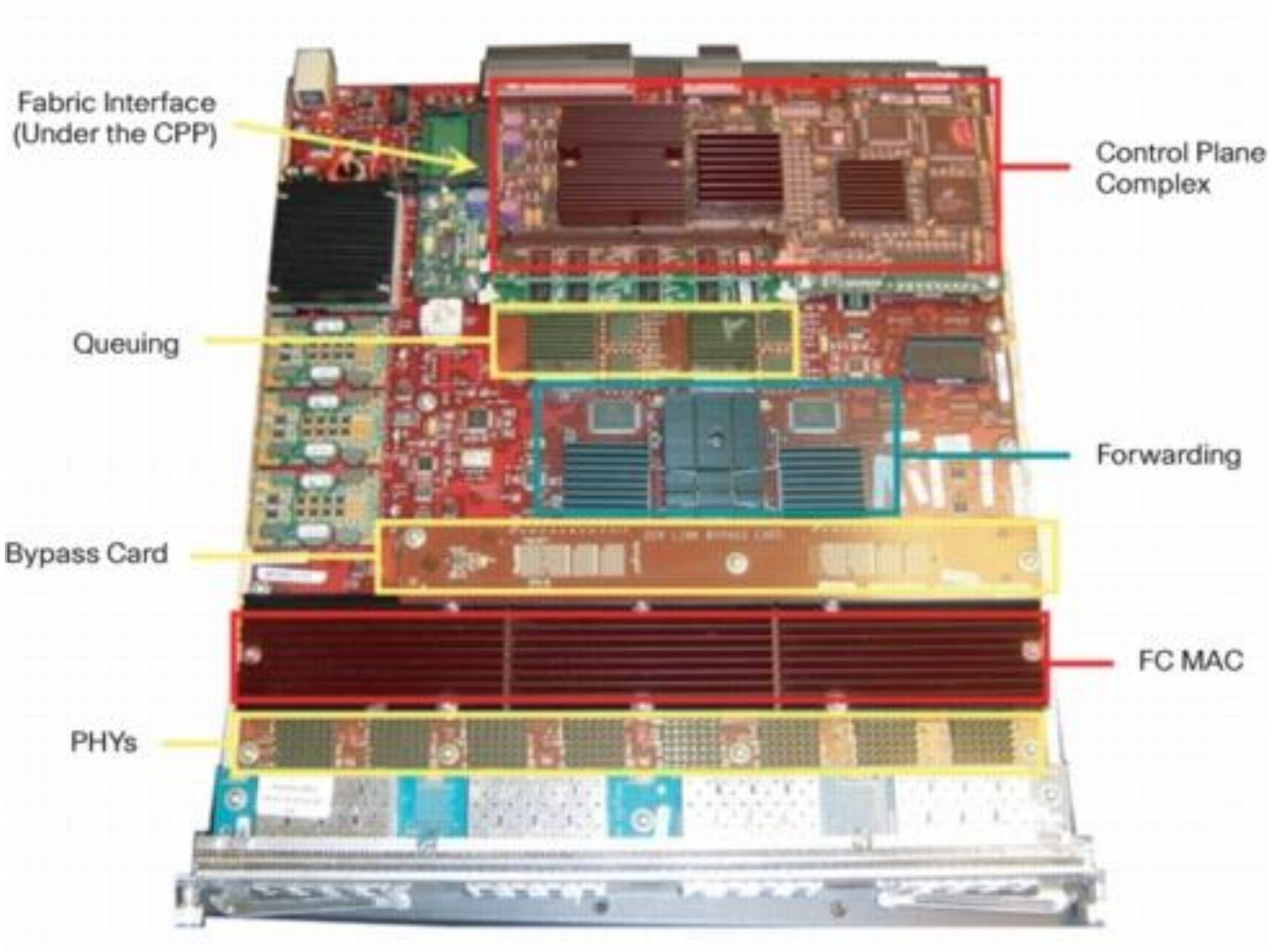
Courtesy of Cisco Systems, "Cisco MDS 9000 Family Switch Architecture". Available online at: http://www.cisco.com/en/US/prod/collateral/modules/ps5991/prod_white_paper0900aecd8044c7e3_ps5987_Products_White_Paper.html

Linecards: example (physical view)



16-Port 1/2-Gbps Non Over-Subscribed, Non-Blocking Linecard

Linecards: example (physical view)

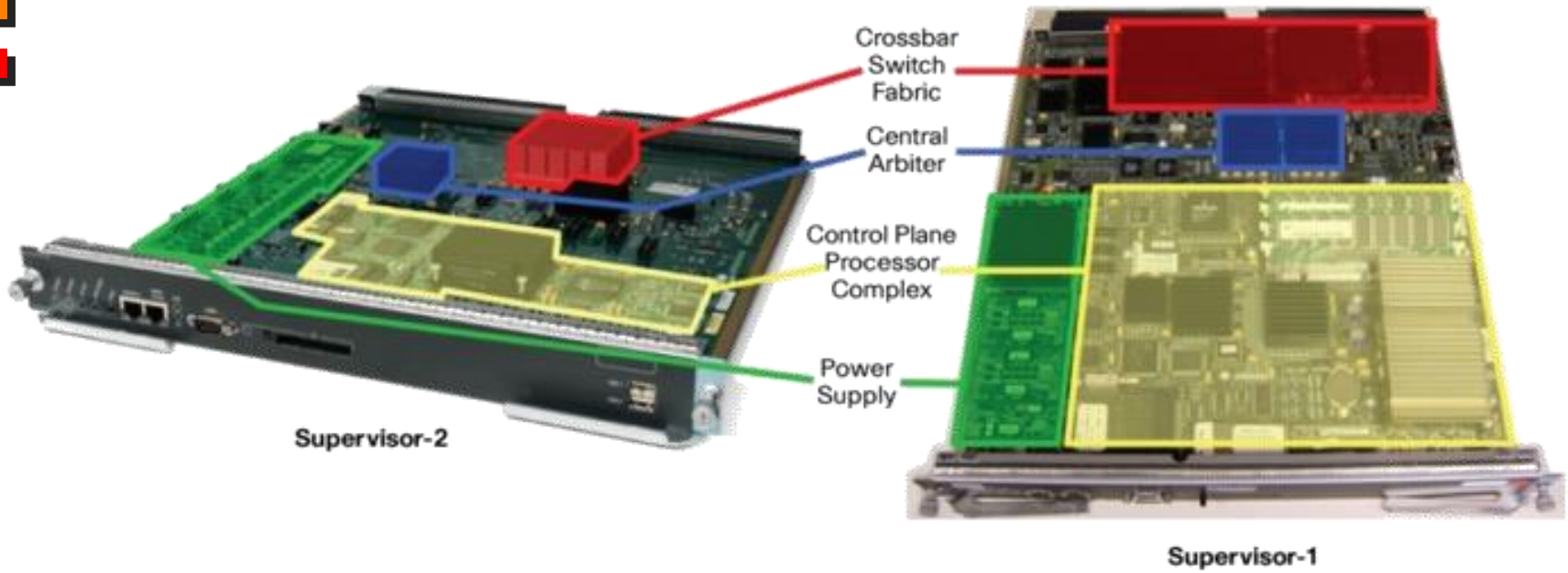


32-Port 1/2-Gbps Host-Optimized Over-Subscribed, Non-Blocking Linecard

Crossbar: example

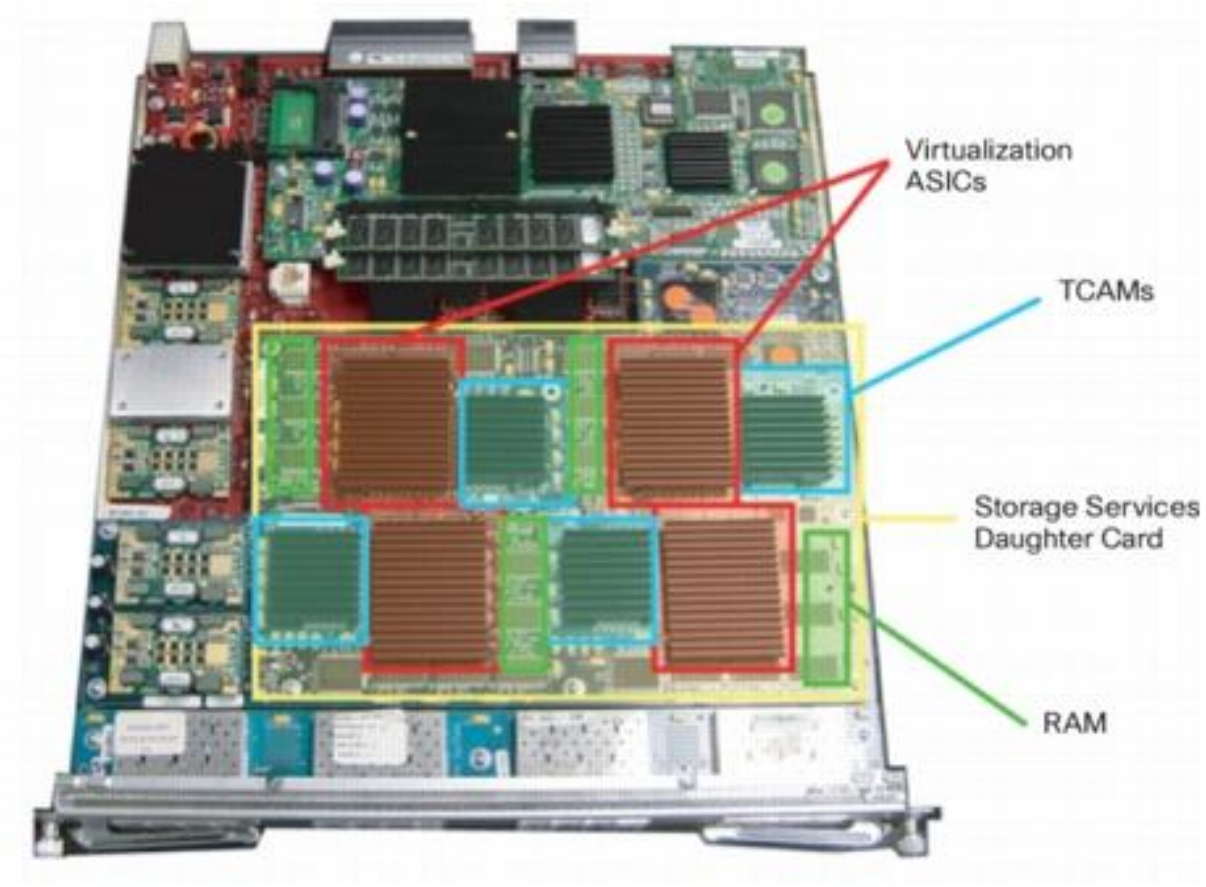
Cisco MDS 9513 Crossbar Switch Fabric Module

Supervisor: example



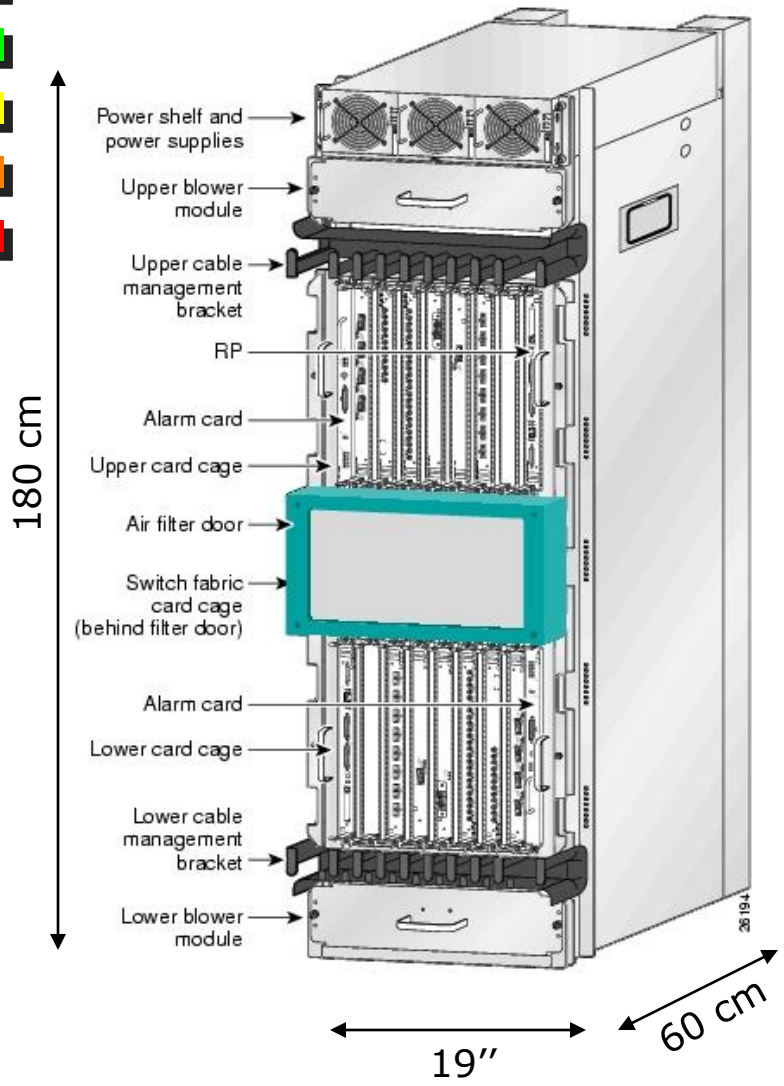
Cisco MDS 9000 Supervisor-1 and Supervisor-2 modules

Service card: example



Cisco MDS 9000 32-Port 1/2-Gbps Storage Services Module

Cisco 12816



16-slots, 40 Gbps/slot

➔ *640Gbps*
➔ *1,28Tbps (commercial)*

187 Kg

Chassis fully configured, using all card slots, AC-input power shelf, and 3 AC-input power supplies

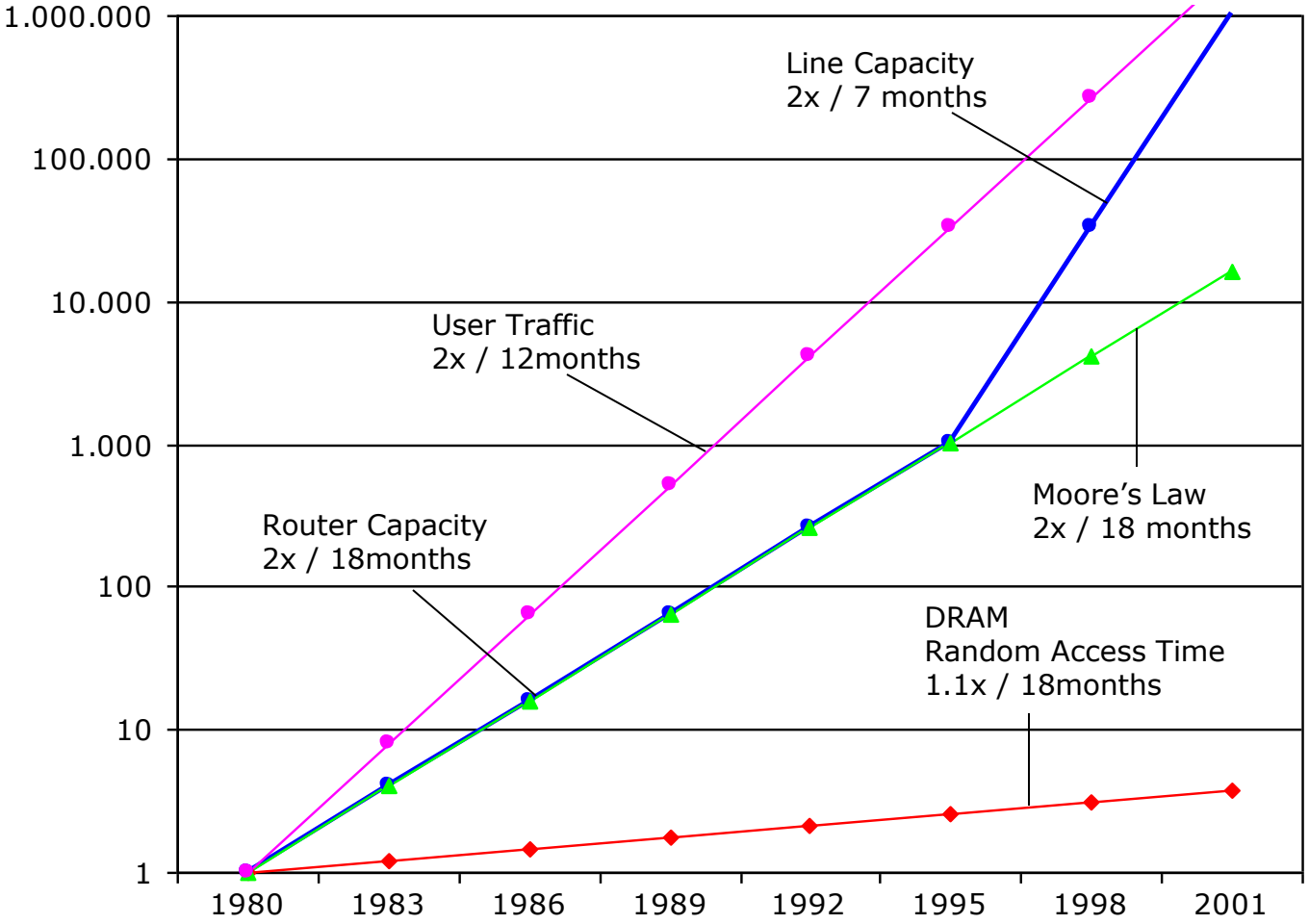
4800W maximum

3 AC-input power supplies—N+1 redundancy

Cisco CSR-1

- Chassis from 320-Gbps, 640-Gbps, and 1.2-Tbps
- Slots with 40 Gbps
- Multi-chassis, from 1,2 to 92 Tbps
  - Max 72 chassis per linecard
  - Max 8 chassis per fabric switching

Trends in Technology, Routers & Traffic



Source: Nick McKeown, "Network Processors and their memory", Network Processor Workshop, Madrid, Feb 2004.

Maggiori problematiche attuali

- Processori
  - Nuovi servizi che richiedono per-packet processing
    - QoS, VPN, Header translation (NAT), L7 Classification, L7 inspection (es. sicurezza), Mobilità (?)
- Memoria
  - Maggiori velocità di linea ➔ aumento delle capacità di buffering
    - Non è un problema sostanziale
  - Tempi di accesso
    - Negli ultimi anni il tempo di accesso alla memoria è rimasto sostanzialmente costante
    - "*Cache in SRAM, Store in DRAM*" non è utilizzabile
      - Il "cache miss" non è tollerato
      - Questo assunto inizia a subire alcune critiche
- Nuove architetture distribuite
  - Possibilità di suddividere il processamento su più blocchi elementari distribuiti
- Consumo e dissipazione termica
  - Alcuni apparati arrivano a consumare 1200W per ogni linecard