NYU TANDON SCHOOL OF ENGINEERING

Data Center Networks I

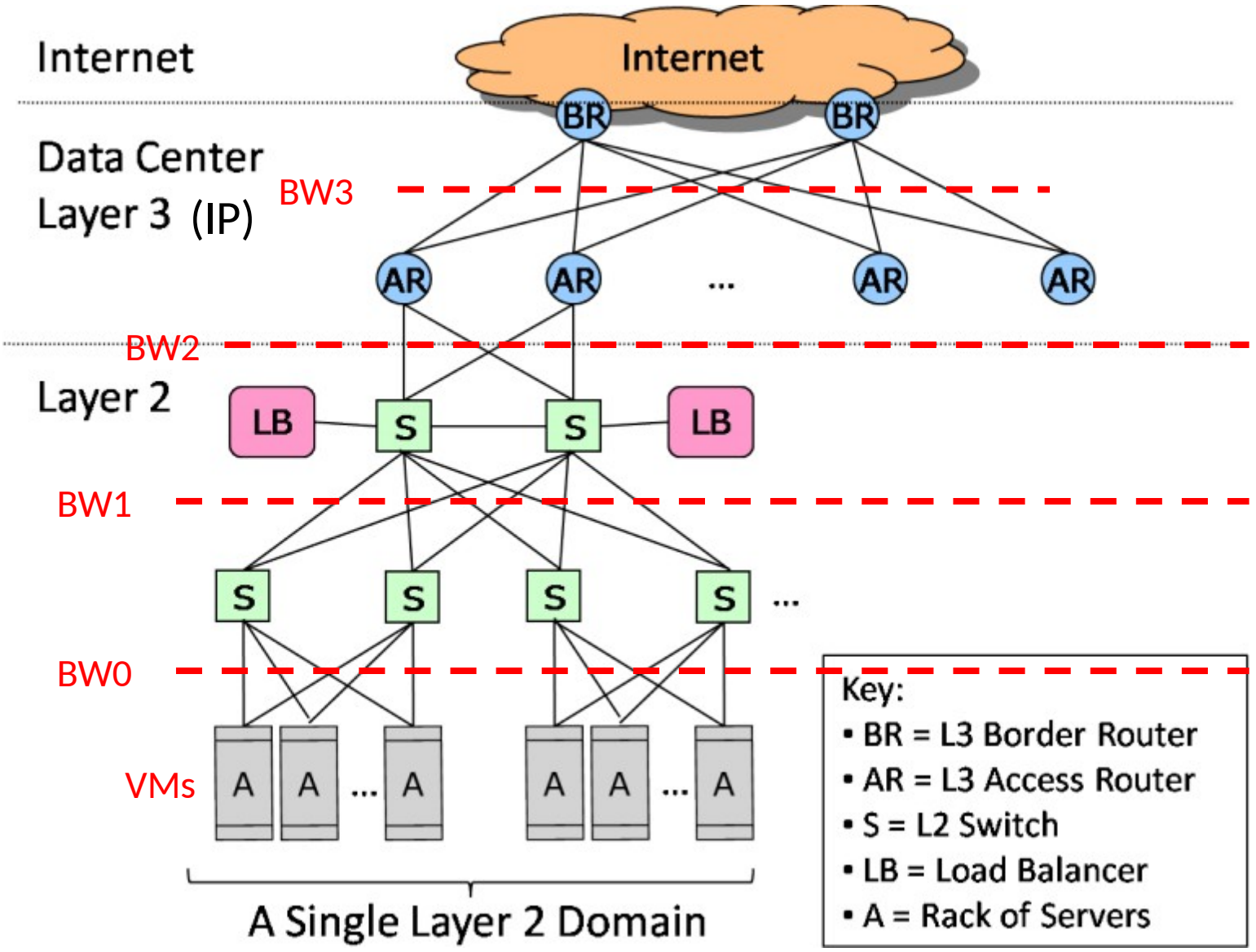
H. Jonathan Chao
ECE Department
chao@nyu.edu

Outline

1. Today's Data Center Networks
2. Router Structures
3. Crossbar Switch
4. Clos Network
5. Fat-Tree Network
6. Commercial Switches used in Data Centers
7. High-Speed Switch Chips

1. Today’s Data Center Networks

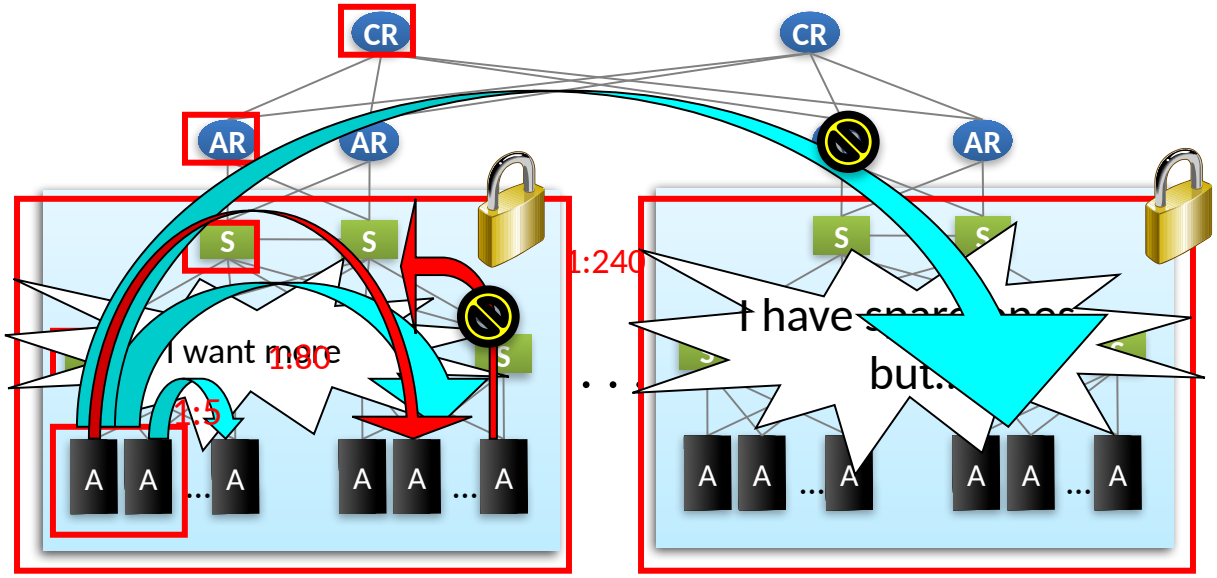
Architecture of Data Center Networks (DCNs)



Two Important Issues in DCN

- Bandwidth bottleneck
  - Oversubscription: the ratio of the aggregated bandwidth of all end hosts to a smallest bisection bandwidth of a particular network topology (e.g., BW0/BW3)
  - Oversubscription is a trade off between the cost and the bandwidth provisioning. For instance, a topology with an oversubscription of 8 has a lower cost than that of a topology with an oversubscription of 1
  - With a large oversubscription, the aggregation and core layers may cause a bandwidth bottleneck for the communications among the servers
    - Solution 1: place VMs in a proximity so that communications among them can avoid a high oversubscription bottleneck ␣ lacking flexibility of using computing resources
    - Solution 2: redesign a better data center network that
      - is backwards compatible with existing equipment and infrastructure
      - has a low power consumption & heat emission
      - allows hosts to communicate at the line speed
- Addressing and routing in a data center
  - Layer 2: flat address space, location-independent, suitable for mobility
  - Layer 3: structured address space, location-dependent, better routing scalability
  - It is desired to have a flat address space for mobility and scalability in routing

Conventional DCN Problems



- Static network assignment
- Fragmentation of resources

- Poor server to server connectivity
- Traffics affects each other
- Poor reliability and utilization

Hakim Weatherspoon, High Performance Systems and Networking Lecture Slides
https://www.cs.cornell.edu/courses/cs5413/2014fa/lectures/09-vl2.pptx

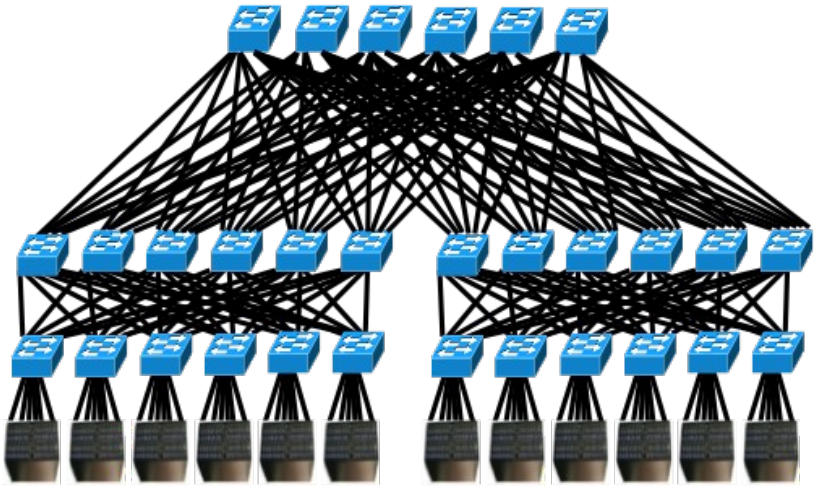
Requirements for a Scalable DCN

- Any-to-any connectivity with non-blocking fabric
  - Scale to more than 100,000 physical nodes
  - Maximize bi-Sectional bandwidth, especially when the 80-20 rule, where 80% traffic remains in the cluster while 20% across the clusters, does not hold any longer
- High availability
  - Resilient control-plane
  - Fast convergence upon failure (quick failure detection and recovery)
  - Fault-domain isolation
- Load balancing routing
  - Efficiently use all available links
  - Multi-path/multi-topology
- Facilitate application deployment
  - Support for multi-tenancy
  - Share resources between different customers
  - Workload mobility, clustering, etc.
- Virtual machine mobility
  - Scalable layer 2 domain

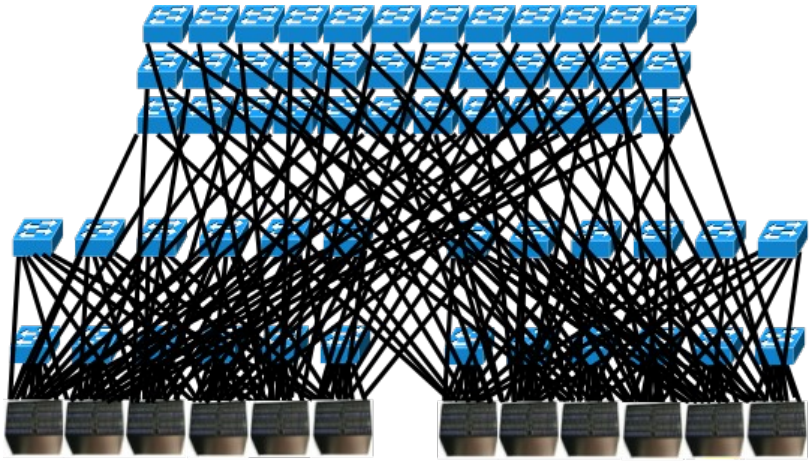
Scale Up or Scale Out?

- Scale up: using high-end switches and routers to construct DCNs
- Scale out: using commodity switches and routers to construct DCNs
- Edge switch cost: $7,000 for each 48-port GigE switch
- Aggregation and core switch cost: $700,000 for 128-port 10GigE switches
- As of today, many people favor the scale out approach.

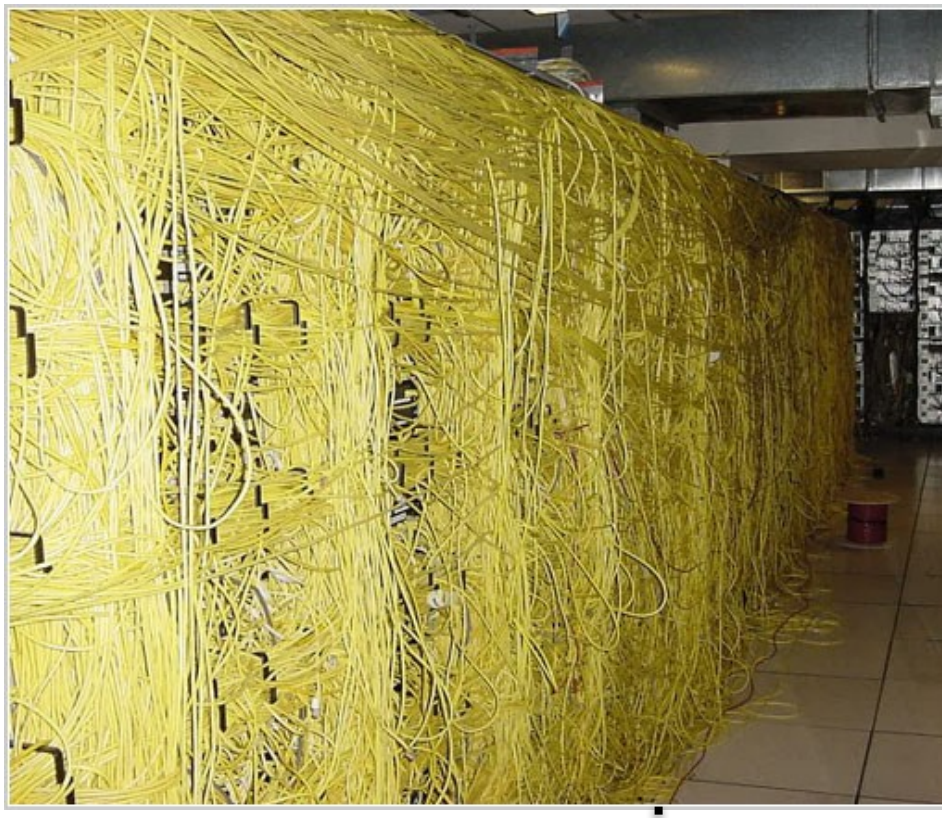
Scale-Out DCN



FatTree



BCube

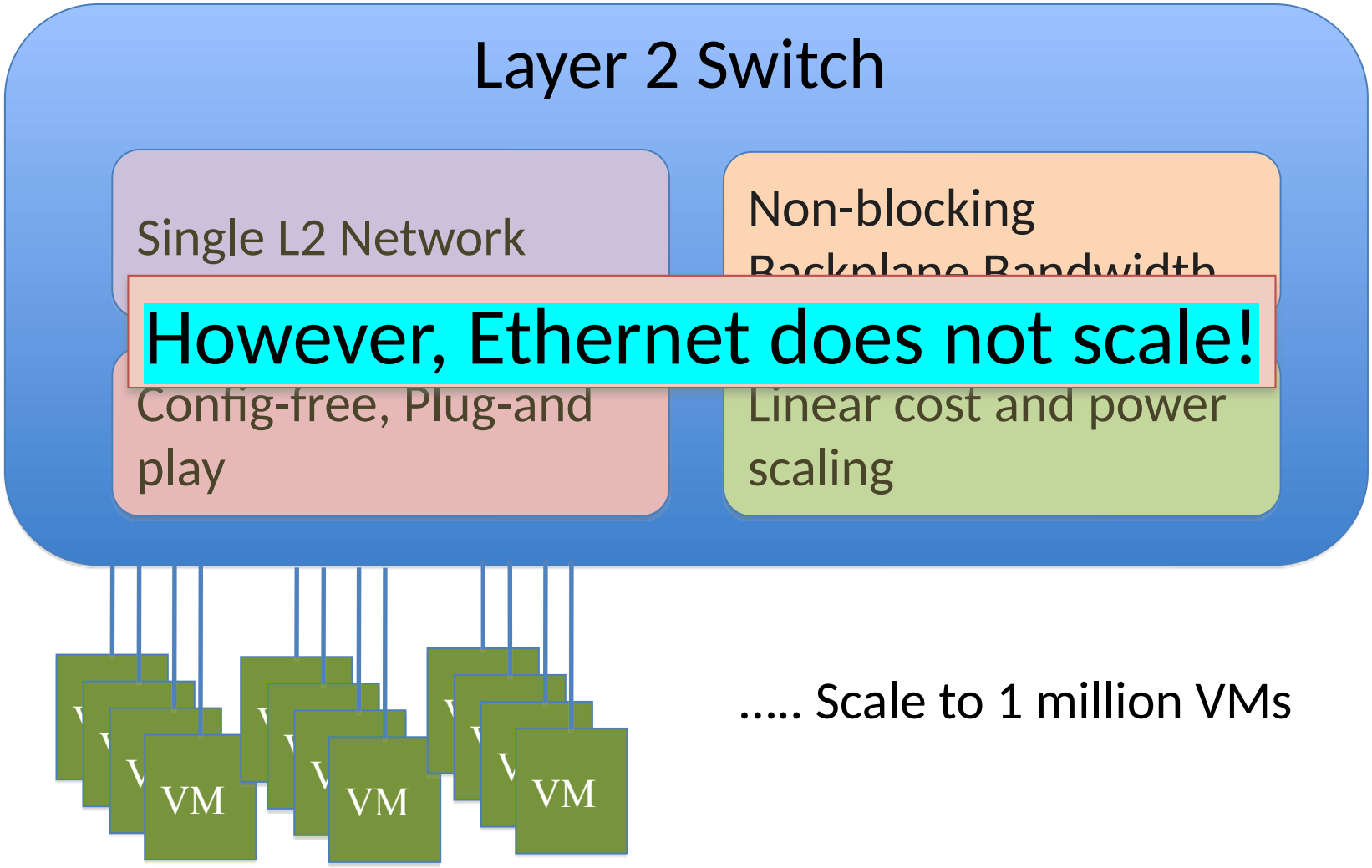


complex wiring

difficult troubleshooting

power consumption

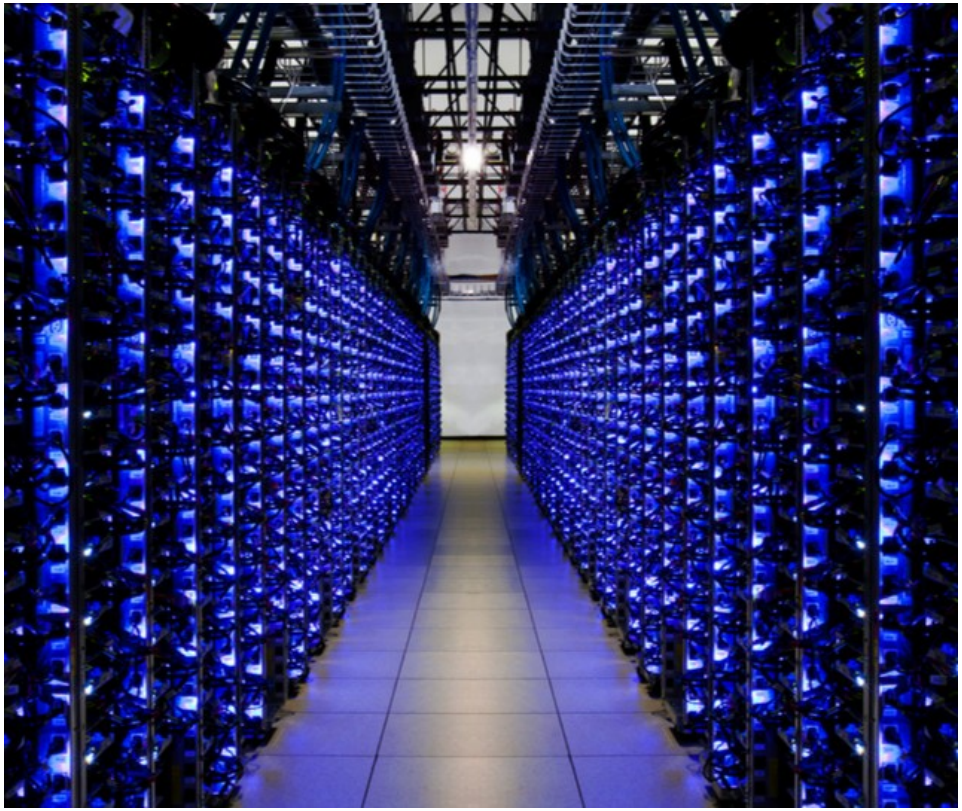
Possible Solution: A Huge L2 Switch!



….. Scale to 1 million VMs

Design Objectives for Mega Data Center Networks (DCNs)

- Huge bisection capacity
- Flat layer 2 address space
- High resilience
- Low latency
- Good manageability
- …….



Related Solution Strategies

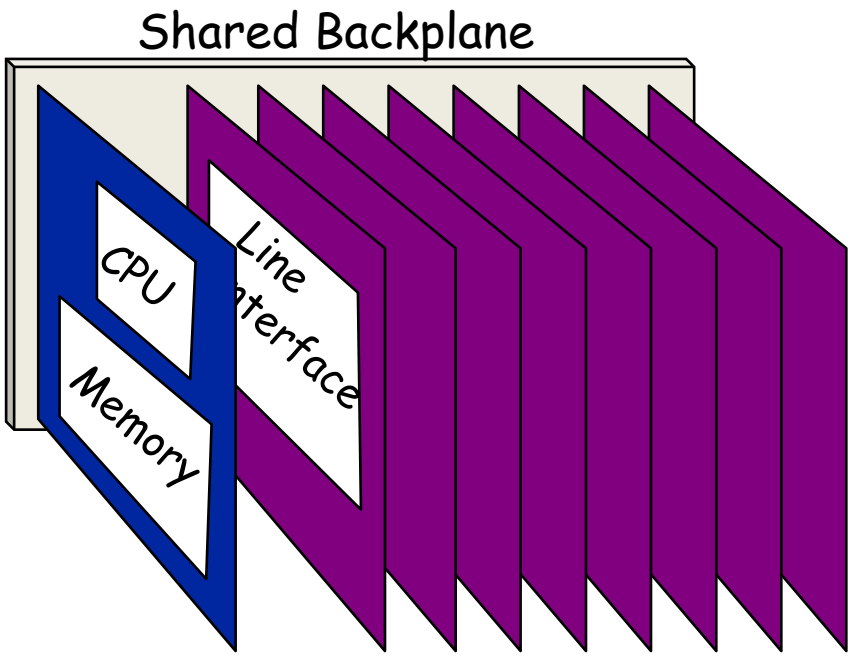
- Scalability:
  - Clos network / Fat-tree to scale out
- Alternative to STP (spanning tree protocol)
  - Link aggregation (Layer 2 trunking), Link Aggregation Control Protocol (LACP), providing a method to control the bundling of several physical ports together to form a single logical channel.
  - Routing protocols to layer 2 network
- Load balancing
  - Randomness or traffic engineering approach

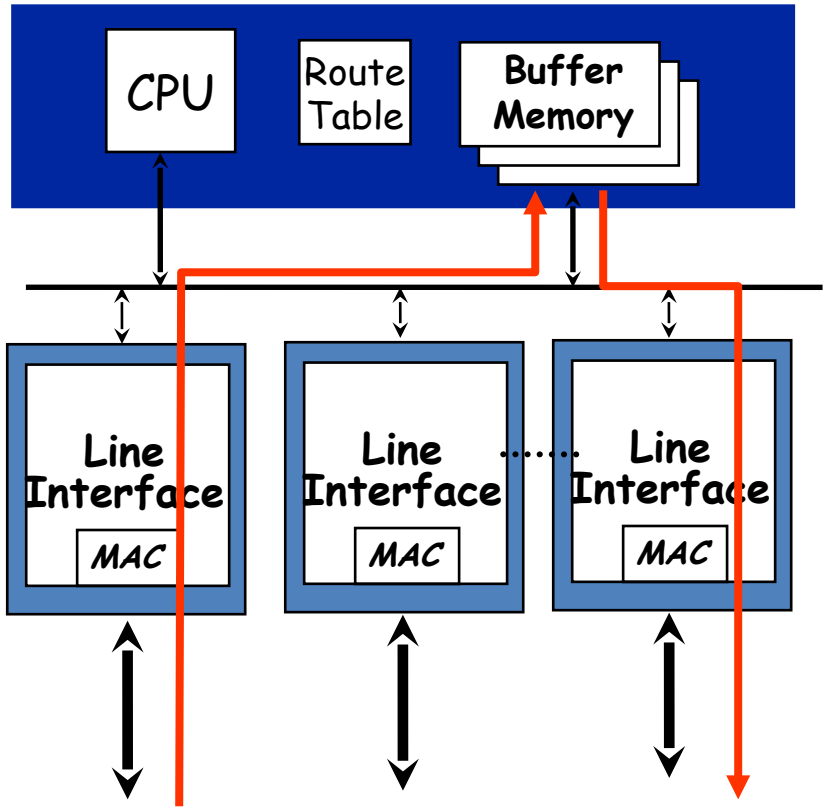
2. Router Structures

Router's Functions

- Look up a forwarding table with many prefixes and masks for the destination IP address of each arriving packet
- Performance objective: Maximize the throughput, average number of bits transferred per second from an input to an output
- Quality of service guarantee to support different applications with packet loss and latency requirements

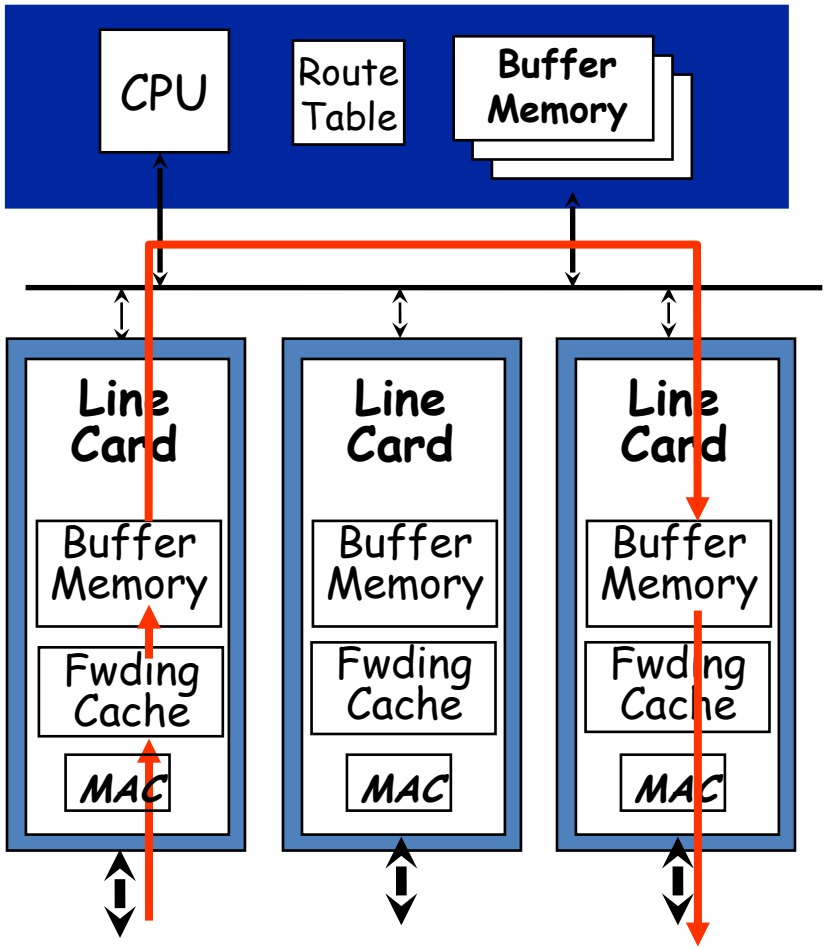
Low-End Router Structure





A CPU/Memory card with multiple Line Interfaces, or called Line Cards

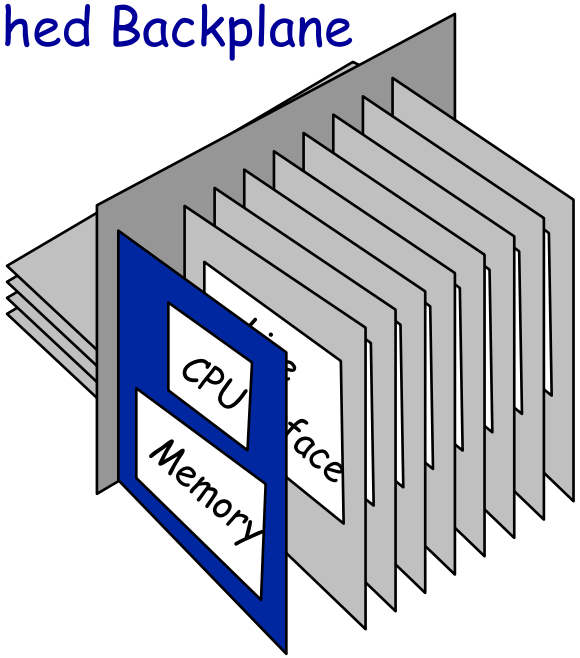
Medium-Size Router Structure



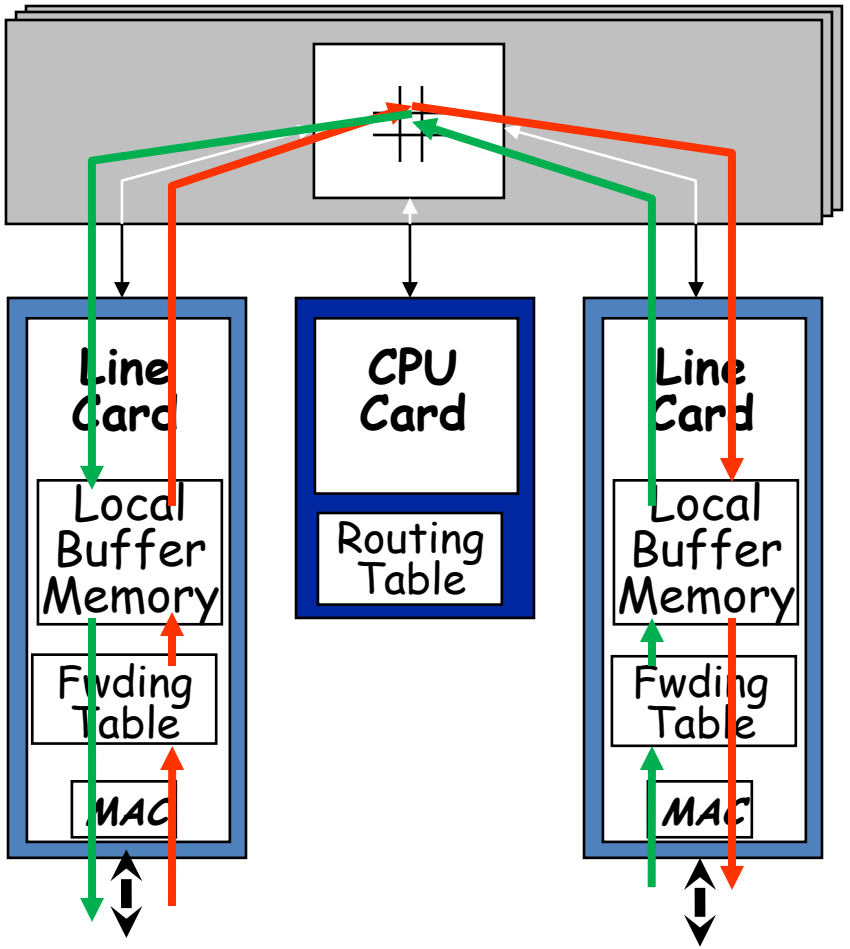
- Individual line buffer vs the shared buffer in the CPU card

High-End Router Structure

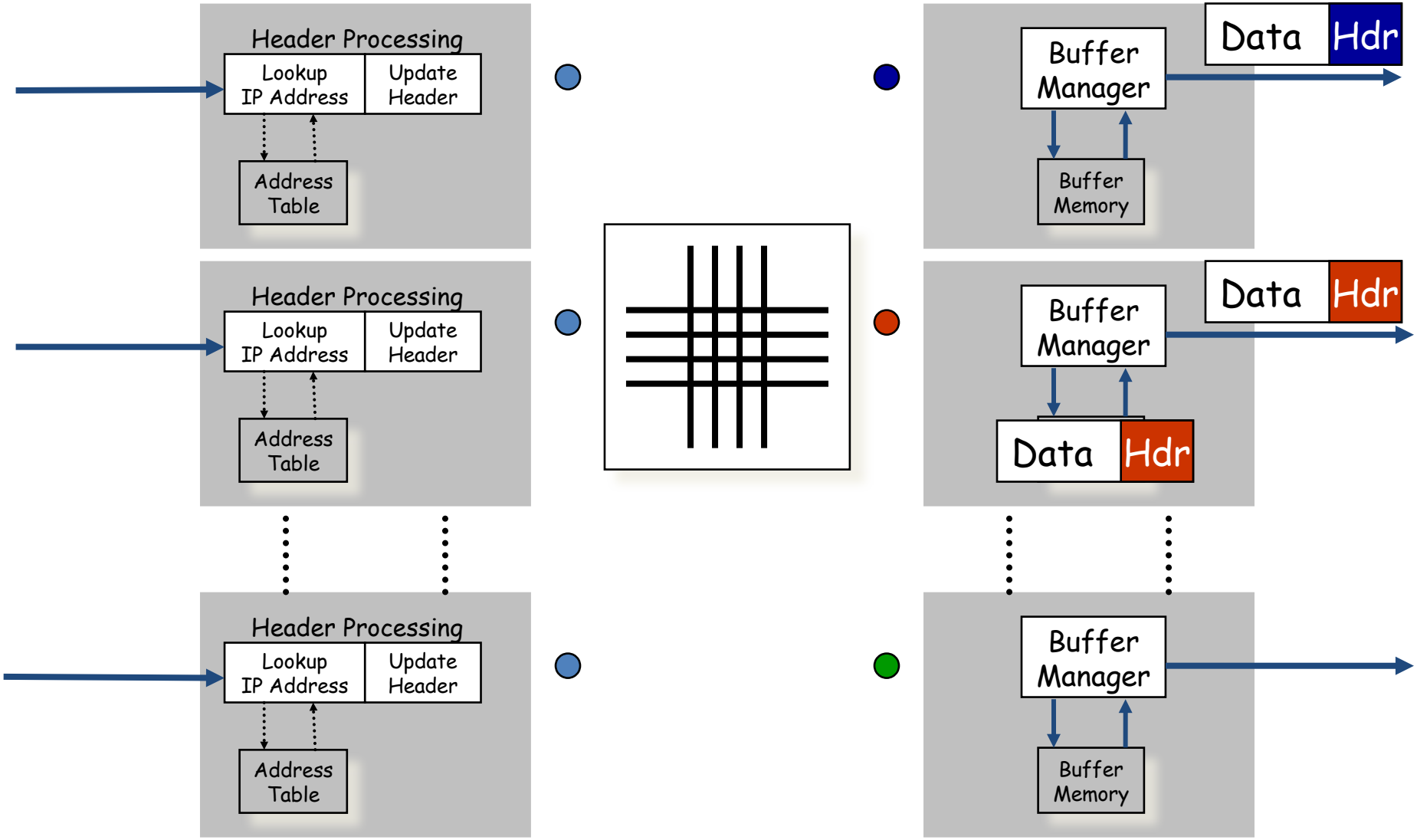
Switched Backplane



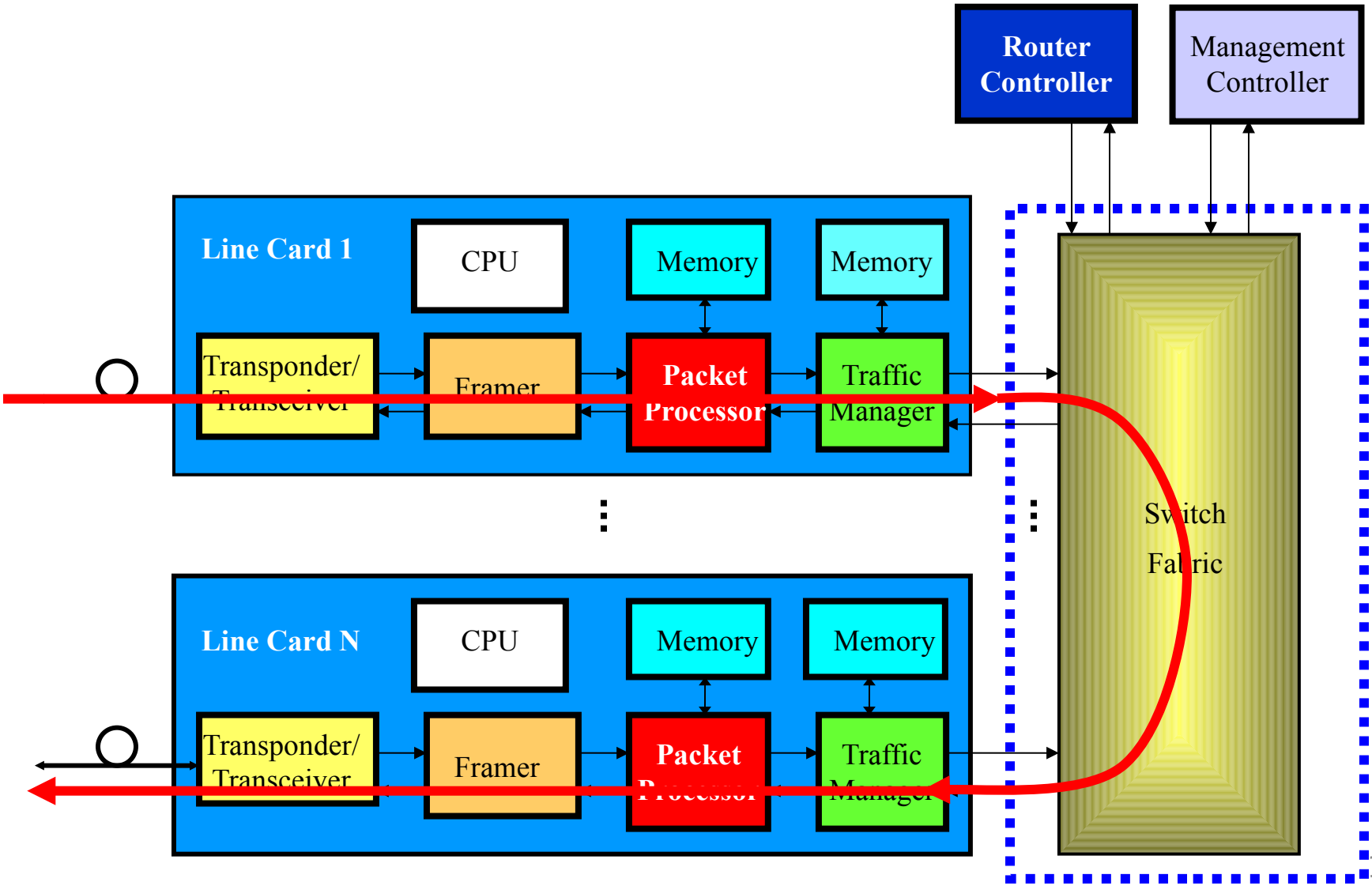
Fwding Table vs Fwding Cache



Packets Switched In A High-End Router



A High-End Router Structure



Components of a Line Card

- Transponder/transceiver
  - optical-to-electrical and electrical-to-optical conversions
  - Serial-to-parallel and parallel-to-serial conversions
- Framer
  - Receiver side - Synchronization, frame overhead processing, and cell or packet delineation
  - Transmit side - frame pattern insertion and/or scrambling
- Packet processor
  - Packet header processing
  - **IP route lookup**
  - **Packet classification**
- Traffic manager
  - Traffic access control, buffer management, and scheduling
- Central processing unit
  - Perform control plane functions
  - Routing table updates, buffer management, and exception handling

Cisco NRS 6008 Single-Chassis System

- **Software compatibility**
  - Cisco IOS XR Software Release 5.0 or later
- **System capacity**
  - Total switching capacity, in+out line rate x no. of line cards
  - 4 Tbit/s per line card capability (2 Tbps per slot ingress and 2 Tbit/s per slot egress) for a total switching capacity of 32 Tbit/s in a Single Chassis configuration
  - Up to 1 Peta bit/sec total switching capacity in a multi-chassis configuration
- **Fabric Cards**
  - 6 Fabric Cards support 5+1 redundancy
  - System can sustain more than one fabric card failure
  - Universal Fabric card (2T & 1T) OR Single-chassis Fabric Card (1T)

Cisco NRS 6008 Single-Chassis System

- **Line cards**
  - 100-G line cards
    - 20 x 100 Gigabit Ethernet multiservice cards with combo optics (CPAK & QSFP)
    - 20 x 100 Gigabit Ethernet LSR (Label Switch Router) cards with combo optics (CPAK & QSFP)
    - 10 x 100 Gigabit Ethernet -multiservice cards with CPAK optics
    - 10 x 100 Gigabit Ethernet LSR cards with CPAK optics
  - 10-G line cards
    - 60 x 10 - Gigabit Ethernet multiservice cards with Enhanced Small Form-Factor Pluggable (SFP+) optics
    - 60 x 10 - Gigabit Ethernet LSR cards with Enhanced Small Form-Factor Pluggable (SFP+) optics
- **Connectivity**
  - 100 and 10 Gigabit Ethernet on 100-Gbps line cards using breakout or patch panel solutions

Juniper’s PTX 5000 Spec

System Capacity	24 Tbit/s
Slot Capacity	3 Tbit/s
Chassis per rack	1
Dimensions (W x H x D)	17.5 x 62.5 x 33.1 in (44.5 x 158.8 x 84.1 cm)
Maximum Weight	1294 lbs (587.0 kg)
Mounting	Front or center rack mount
Power System Rating (Max)	217 A @ -48 VDC
Operating Temperature	32° to 104° F (0° to 40° C)
Humidity	Relative humidity operating: 5 to 90% (noncondensing)
Altitude	Up to 10,000 ft. (3,048 m)
Port density 10G/40G/100G	1536/384/240