NYU TANDON SCHOOL OF ENGINEERING

Data Center Networks I

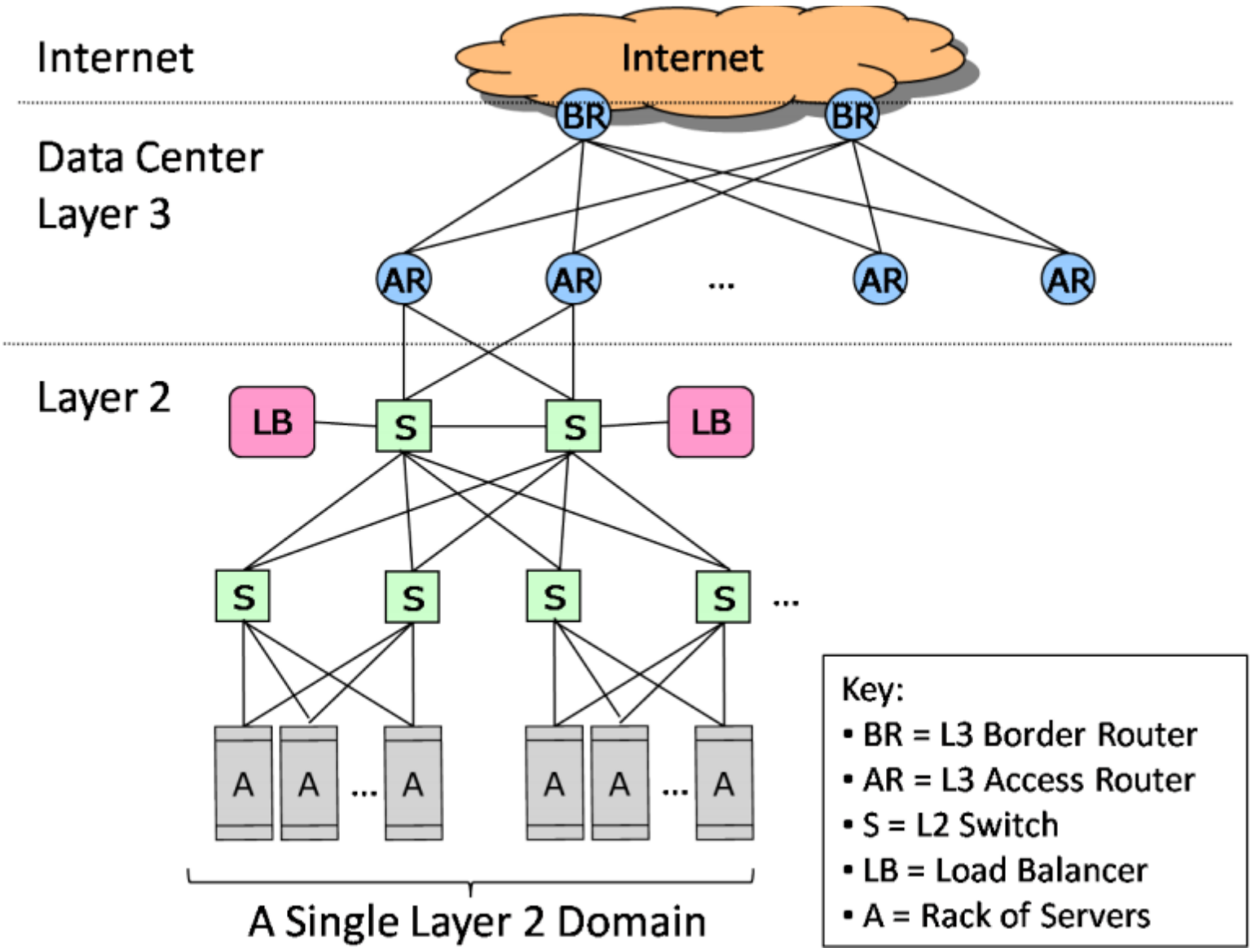
H. Jonathan Chao
ECE Department
chao@nyu.edu

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



Two Important Issues

- Bandwidth bottleneck
  - Oversubscription: ratio of the worst-case achievable aggregate bandwidth among the end hosts to the total bisection bandwidth of a particular communication topology
  - Trade off between cost and the bandwidth provisioning. For instance, lower the total cost of the design with a ratio of 1/8
  - At the aggregation and core layers may create bandwidth bottlenecks among servers
  - Solution 1: optimize VM placement so that communications occur most among the VMs in close proximity to overcome high oversubscription bottleneck
  - Solution 2: redesign a better switch network to
    - Be backwards compatible with existing infrastructure
    - Have low power consumption & heat emission
    - Allow host communication at line speed
- Addressing and routing in a data center
  - Layer 2: flat address space, location-independent
  - Layer 3: structured address space, location-dependent
  - It is ideal to provide a flat address space for mobility, but it is non-trivial for routing

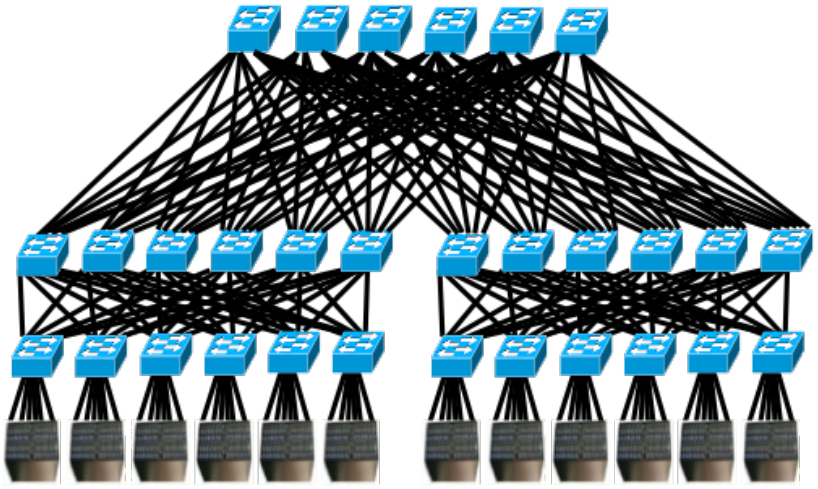
Requirements for a Scalable DC Network

- Any-to-any connectivity with non-blocking fabric
  - Scale to more than 100,000 physical nodes
  - Maximize bi-Sectional bandwidth
- 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

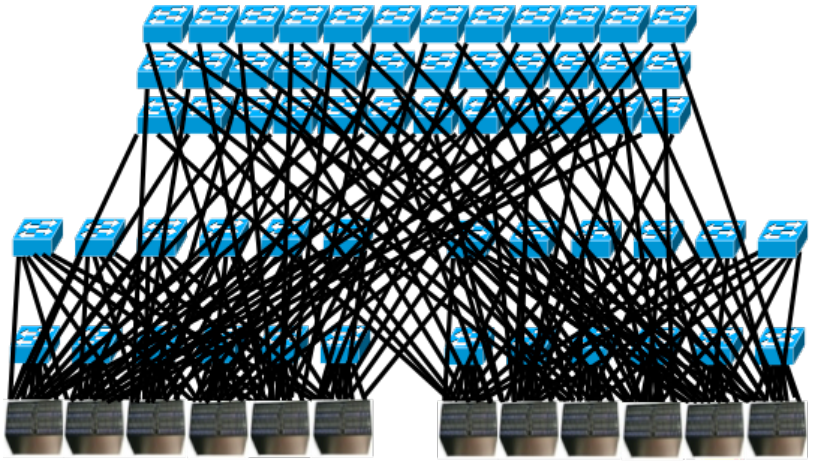
Scale Up or Scale Out?

- 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.

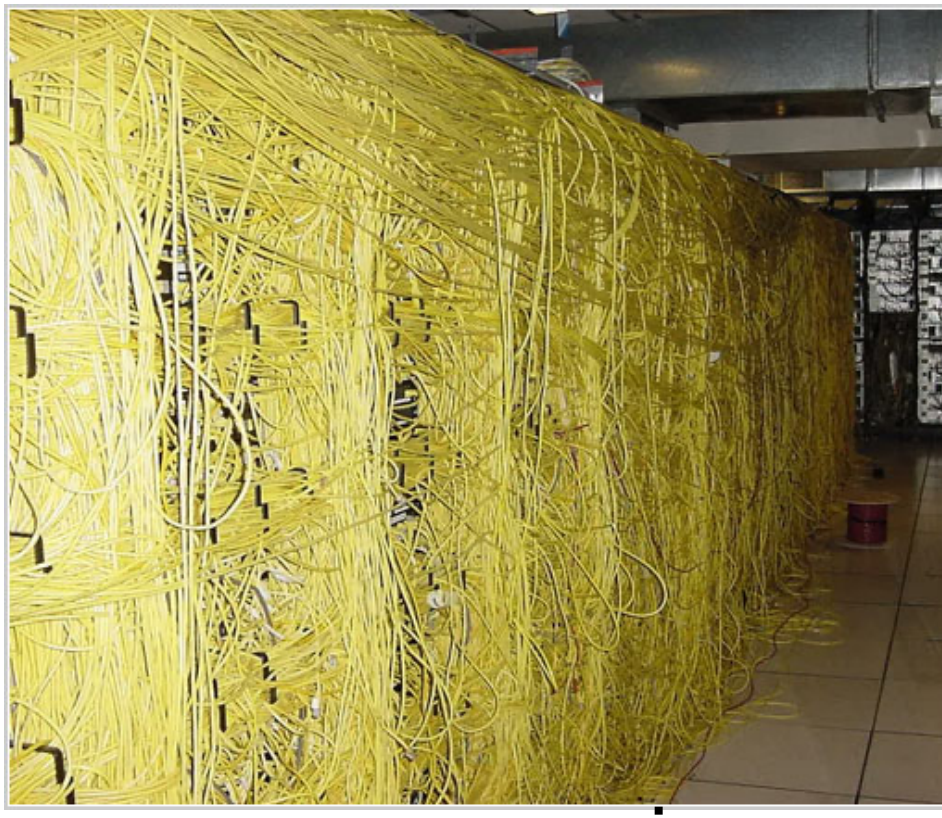
Scale-Out DCN



FatTree



BCube

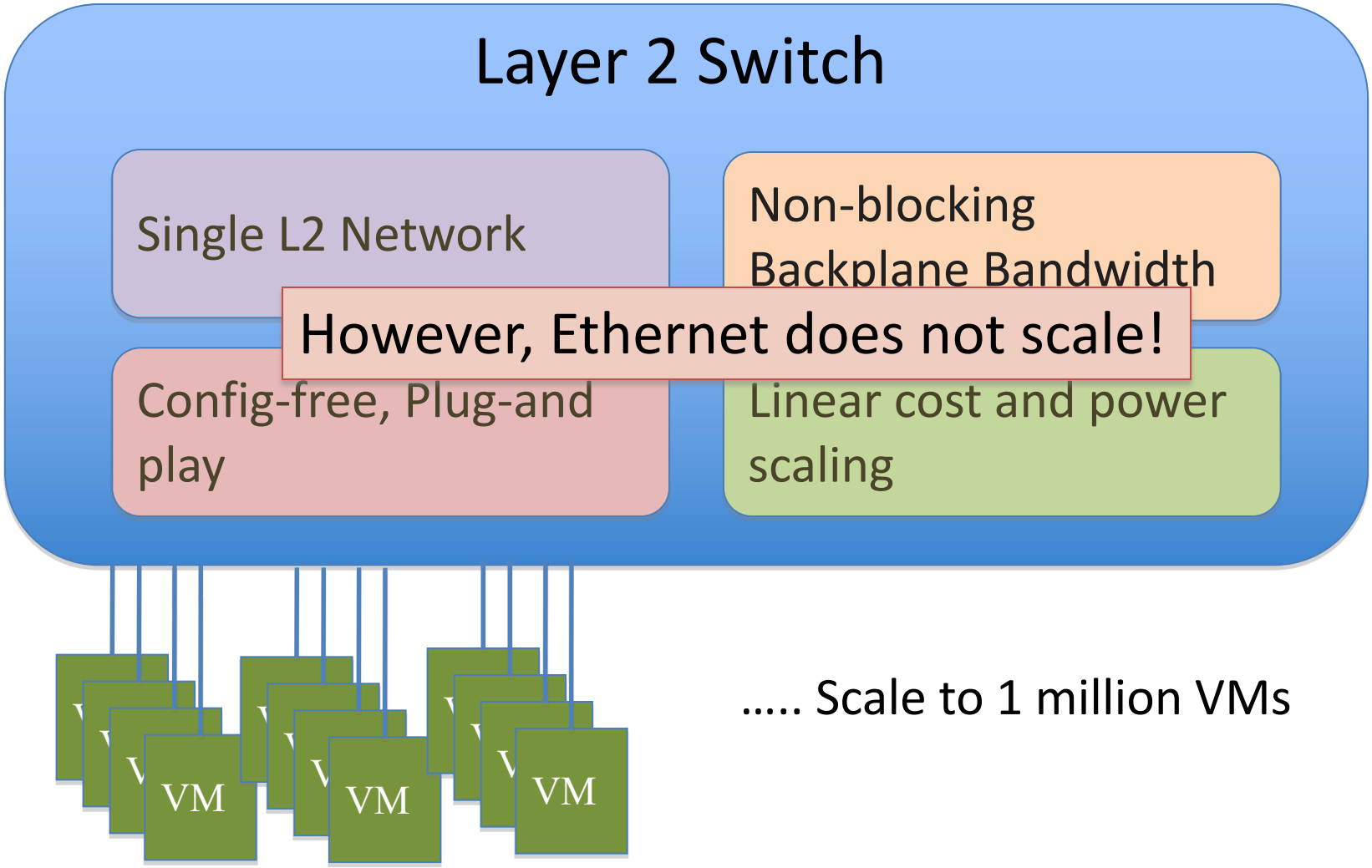


complex wiring

difficult troubleshooting

power consumption

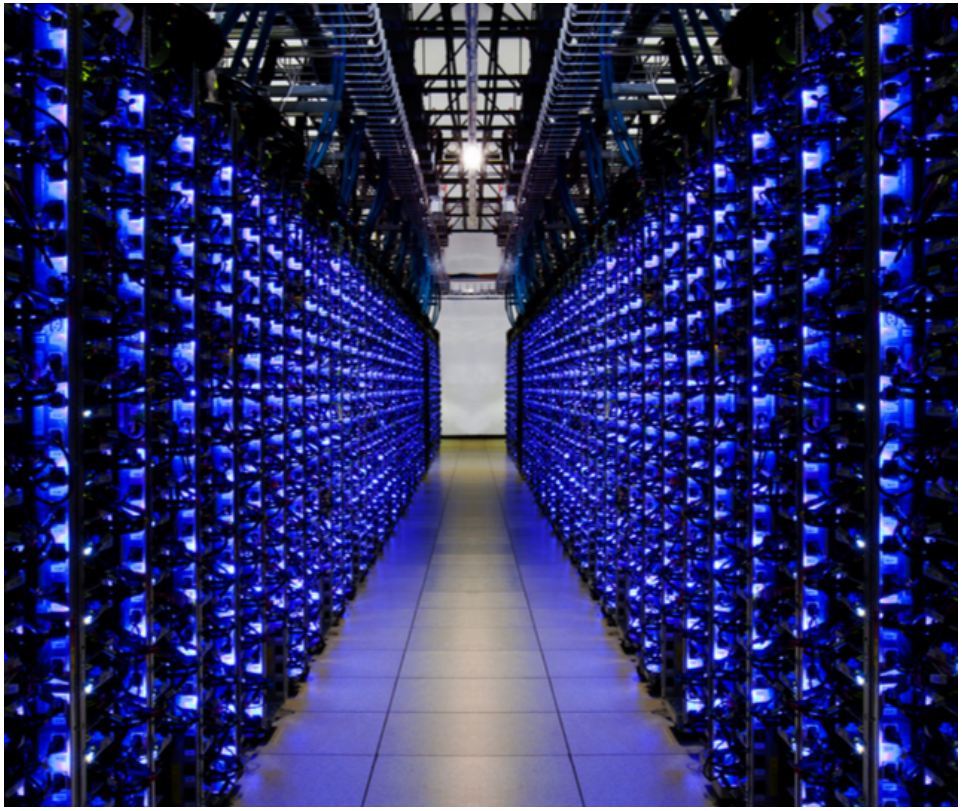
Possible Solution: A Huge L2 Switch!



….. 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

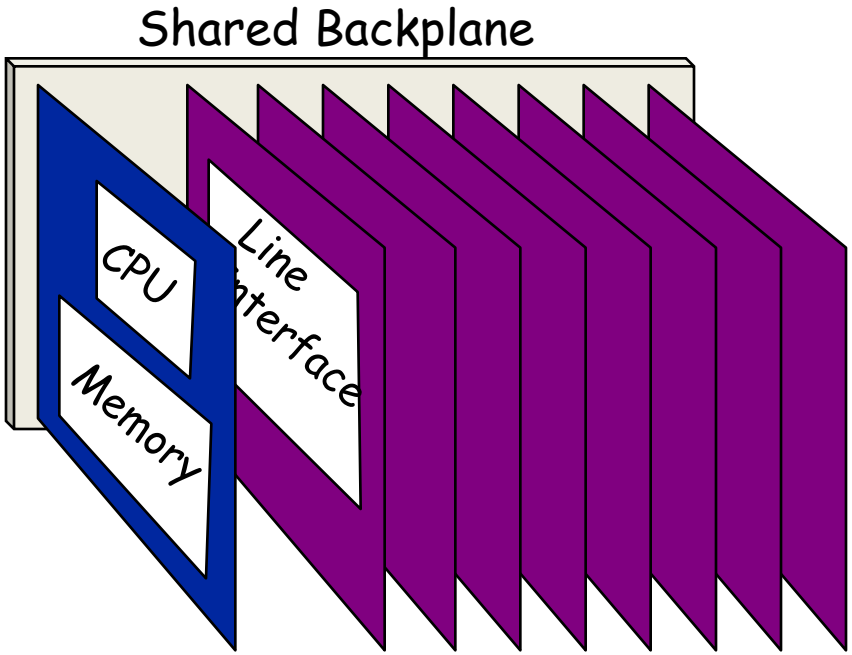
- Scalability:
  - Clos network / Fat-tree to scale out
- Alternative to STP (spanning tree protocol)
  - Link aggregation (Layer 2 trunking), Link Aggregation Control Protocol (LACP), providing a method to control the bundling of several physical ports together to form a single logical channel.
  -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

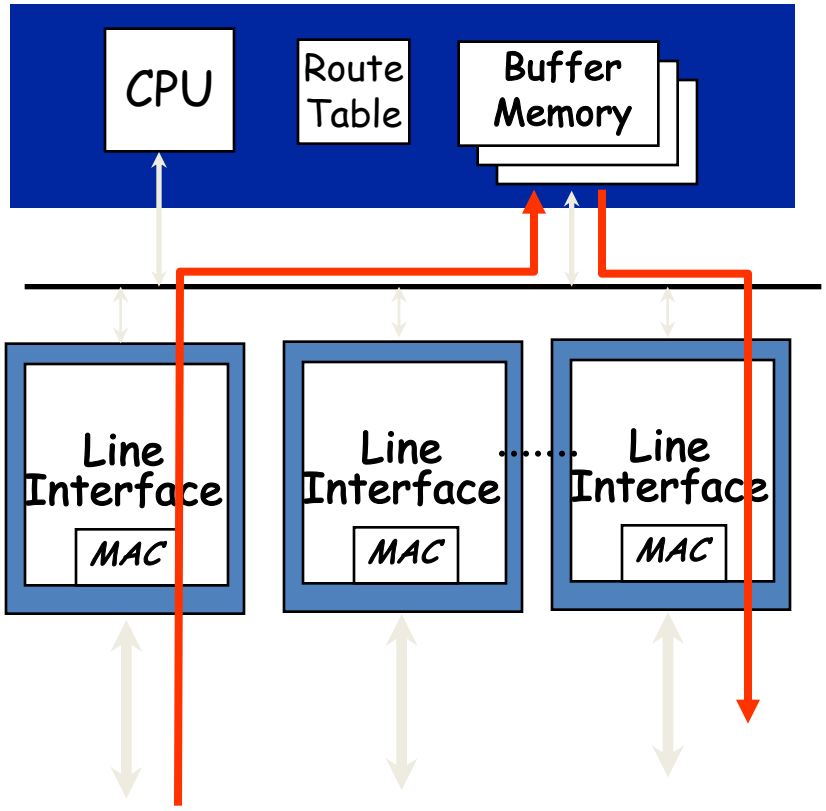
2. Router Structures

Router's Functions

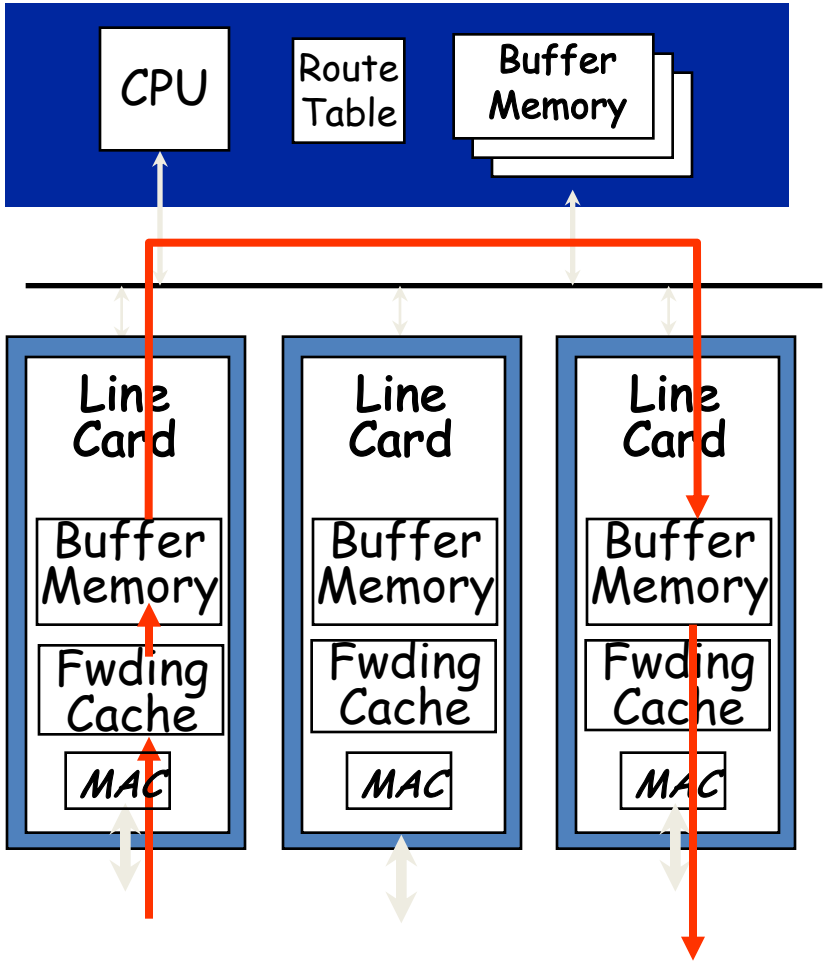
- Look up a forwarding table with many prefixes and masks for the destination IP address of each arriving packet
- Performance objective: Maximize the throughput, average number of bits transferred per second from an input to an output
- Quality of service guarantee to support different applications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

Low-End Router Structure



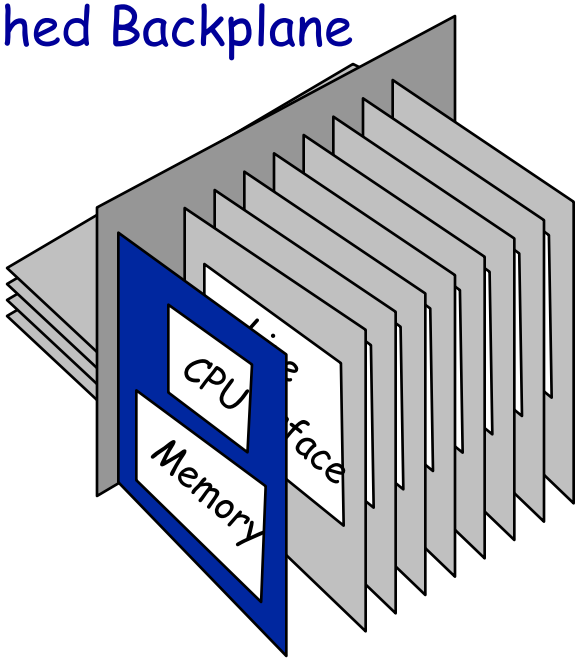


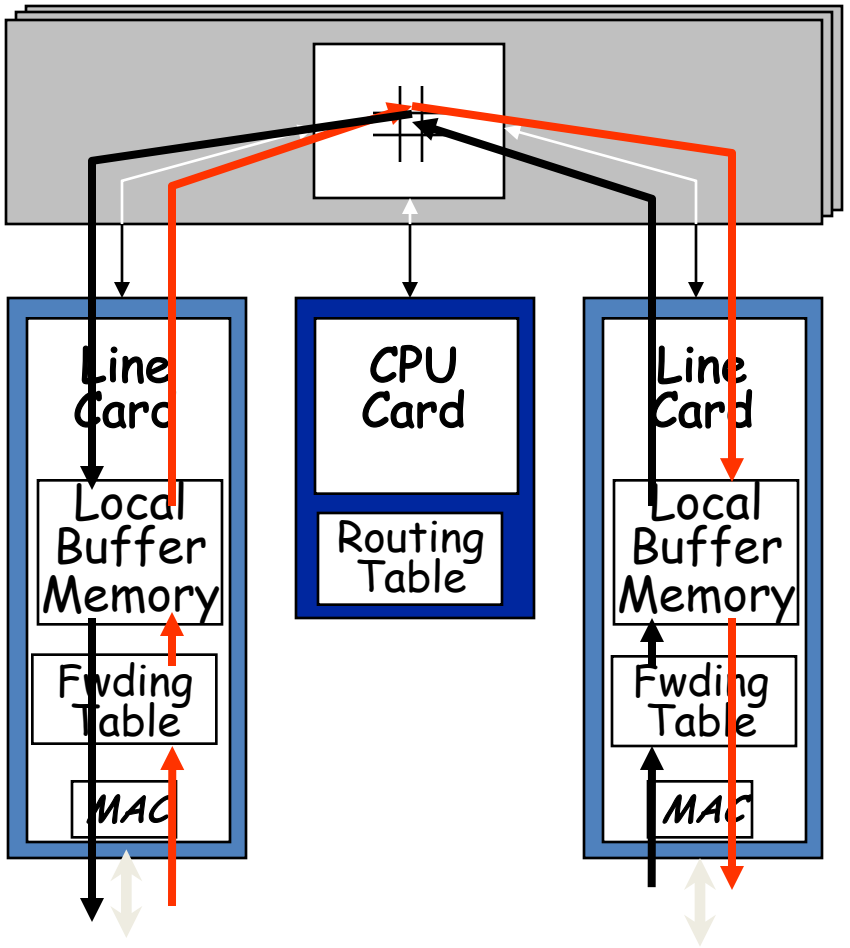
Medium-Size Router Structure



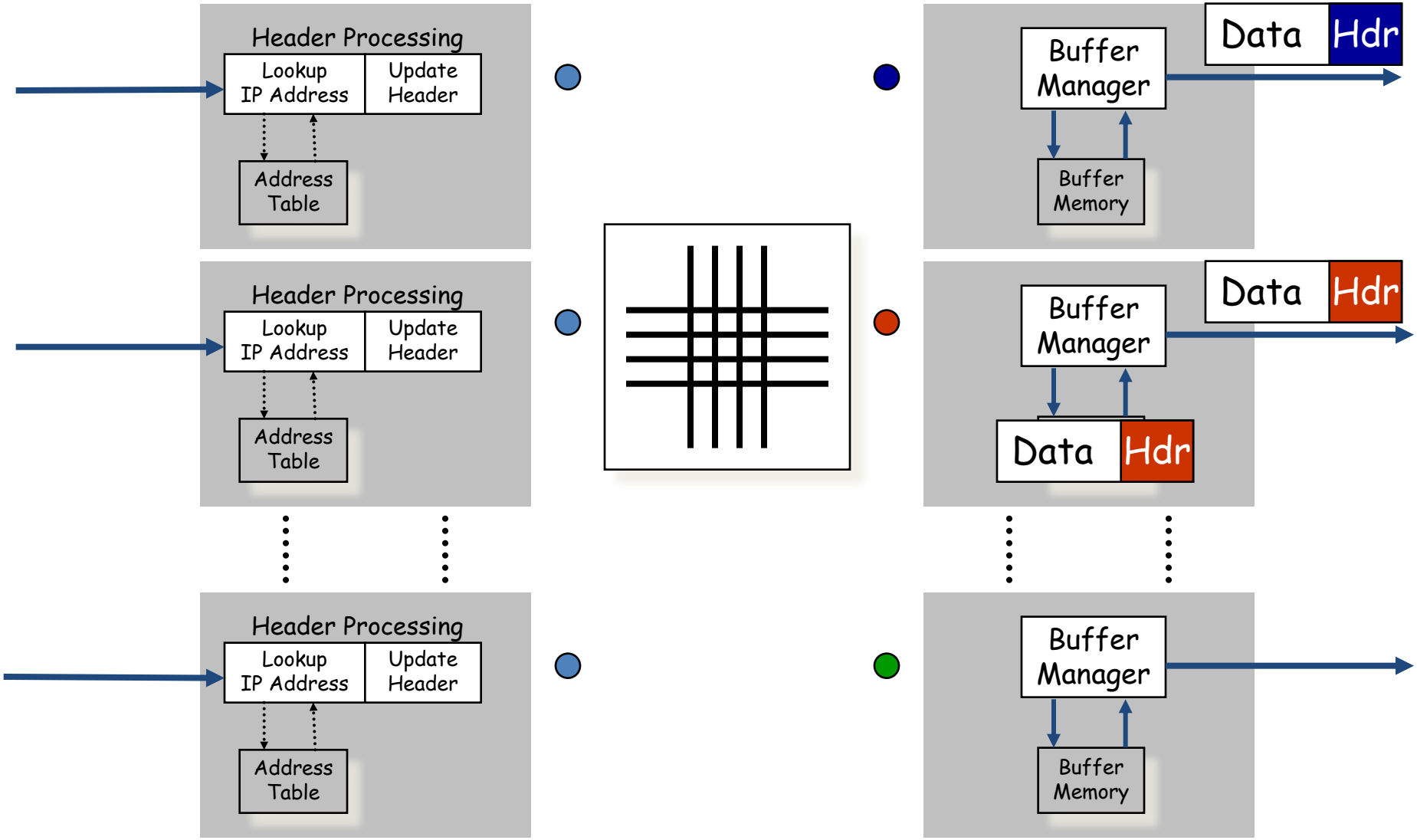
High-End Router Structure

Switched Backplane

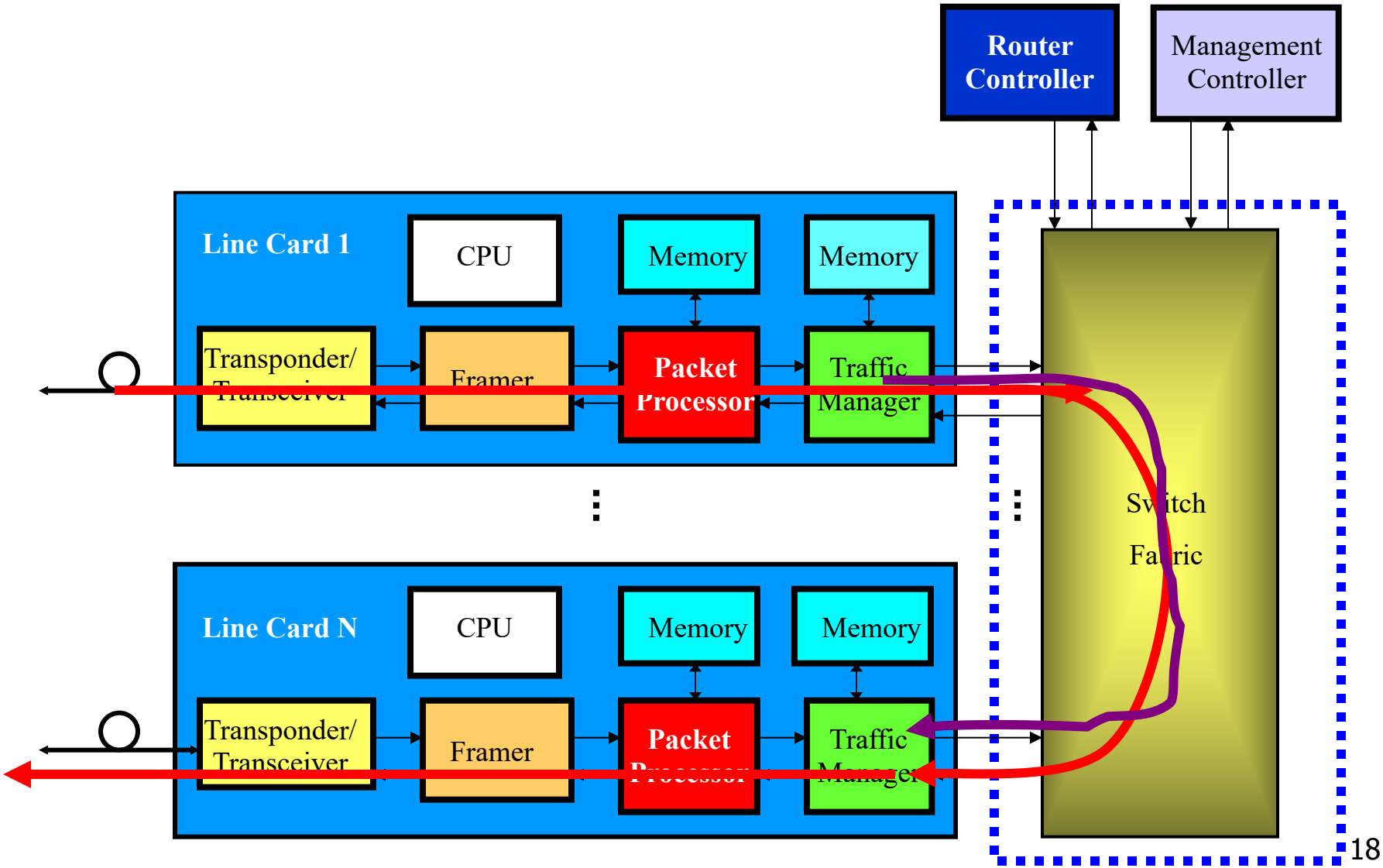




Packet Switched In A Router



A 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 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**
  - 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 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

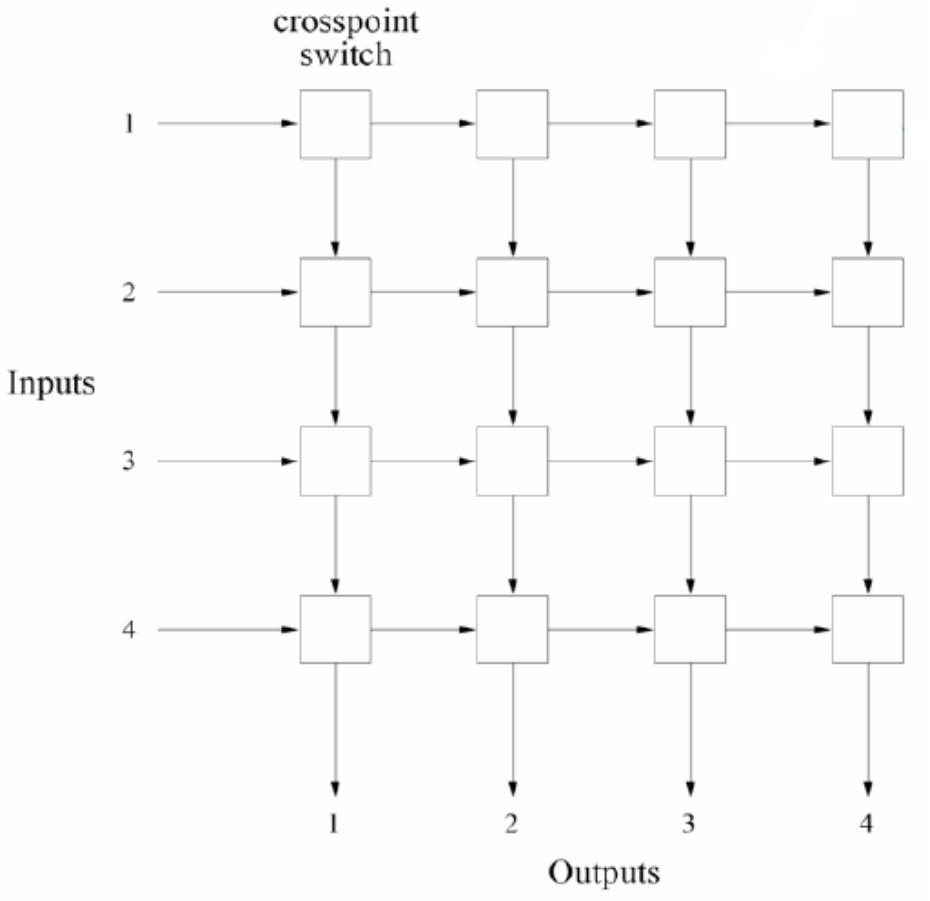
System Capacity	24 Tbit/s
Slot Capacity	3 Tbit/s
Chassis per rack	1
Dimensions (W x H x D)	17.5 x 62.5 x 33.1 in (44.5 x 158.8 x 84.1 cm)
Maximum Weight	1294 lbs (587.0 kg)
Mounting	Front or center rack mount
Power System Rating (Max)	217 A @ -48 VDC
Operating Temperature	32° to 104° F (0° to 40° C)
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

3. Crossbar Switch

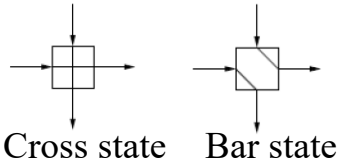
Switch fabric

- Different switch fabric architectures
  - Crossbar (non-blocking, the ideal switch fabric)
  - Clos
  - Banyan, Benes …
- Blocking and output contention
  - Non-blocking
    - when there is a request generated between an idle input port and an idle output port, a connection can always be set up.
    - an idle port refers to a port not connected nor requested.
  - Blocking
    - it is possible that no connection can be set up between an idle input/output pair.
  - Output contention

Crossbar Switch Fabric

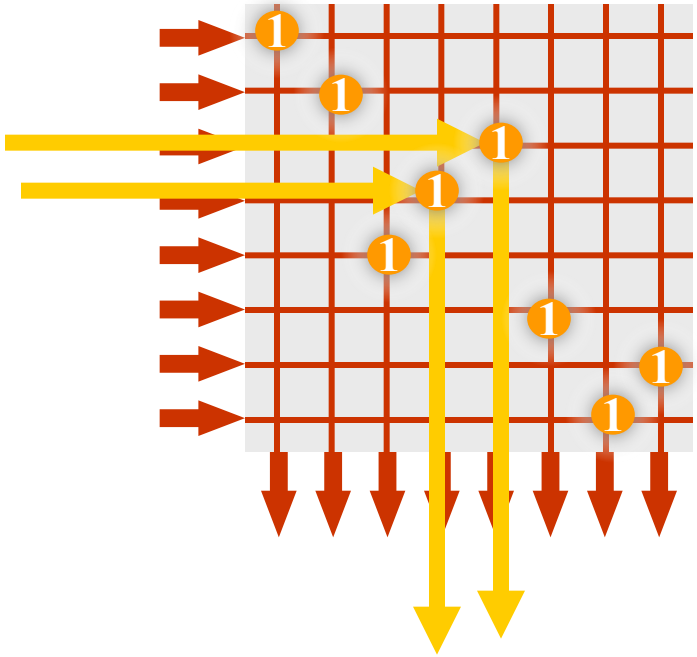


- An *N* x *N* crossbar switch consists of a 2-dimensional array of N^2 cross-points, one corresponding to each input-output pair.



- Each cross-point has two possible states: cross (default) and bar. A connection between input port i and output port j is established by setting the (i, j)th cross-point switch to the bar state while letting other cross-points along the connection remain in the cross state

Crossbar Switch Fabric



Permutation

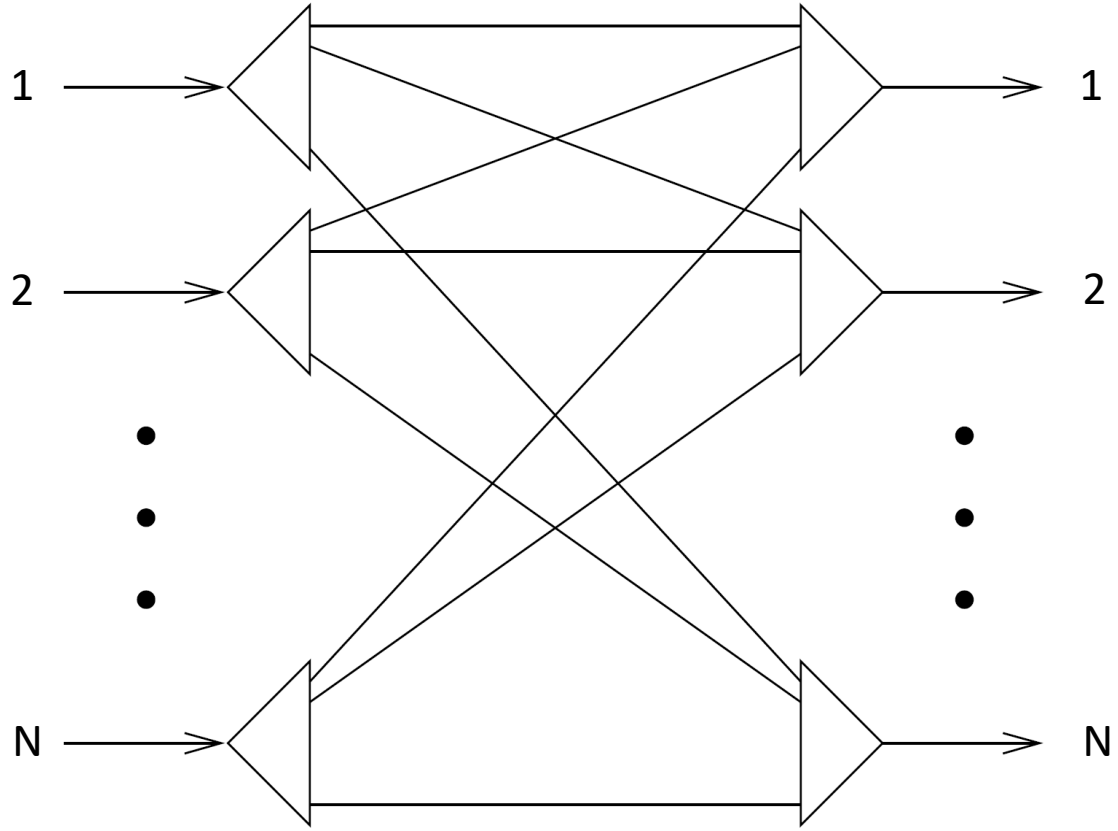
Outputs

Inputs

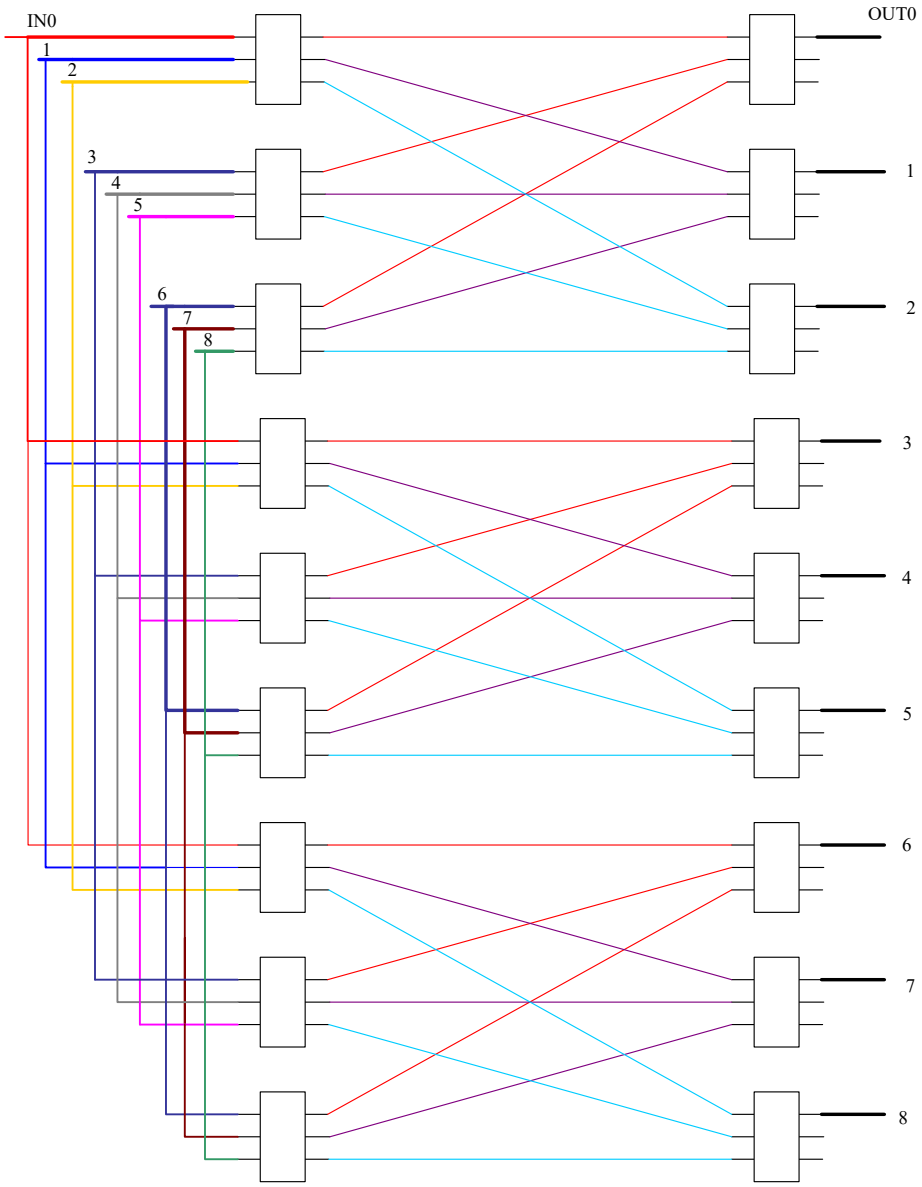
1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0
0	0	0	1	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1
0	0	0	0	0	0	1	0

- A crosspoint switch supports all permutations
- It is non-blocking
- It needs N^2 crosspoints
- 2N I/O pins or connectors cause the scalability issue

Crossbar Switch = Full Interconnected Switch



Build a 9x9 switch using many 3x3 switch chips

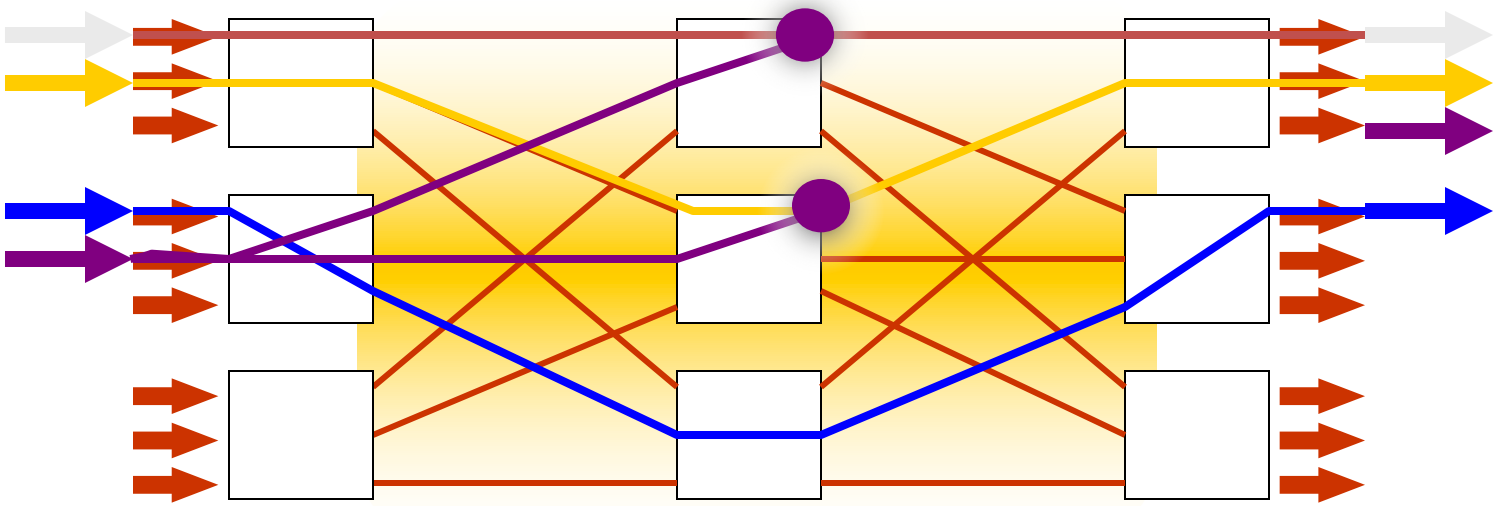


4. Clos Network

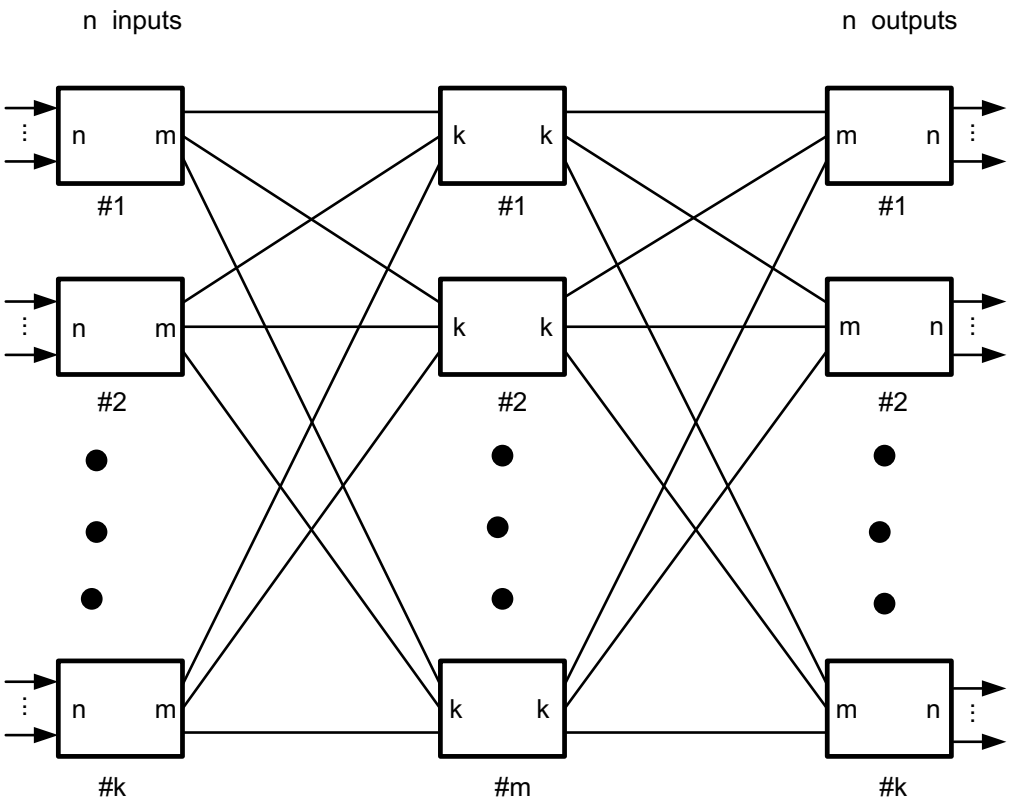
A Clos Network

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

- If given the permutation, you can route it.
- If given each connection one at a time, you may not.



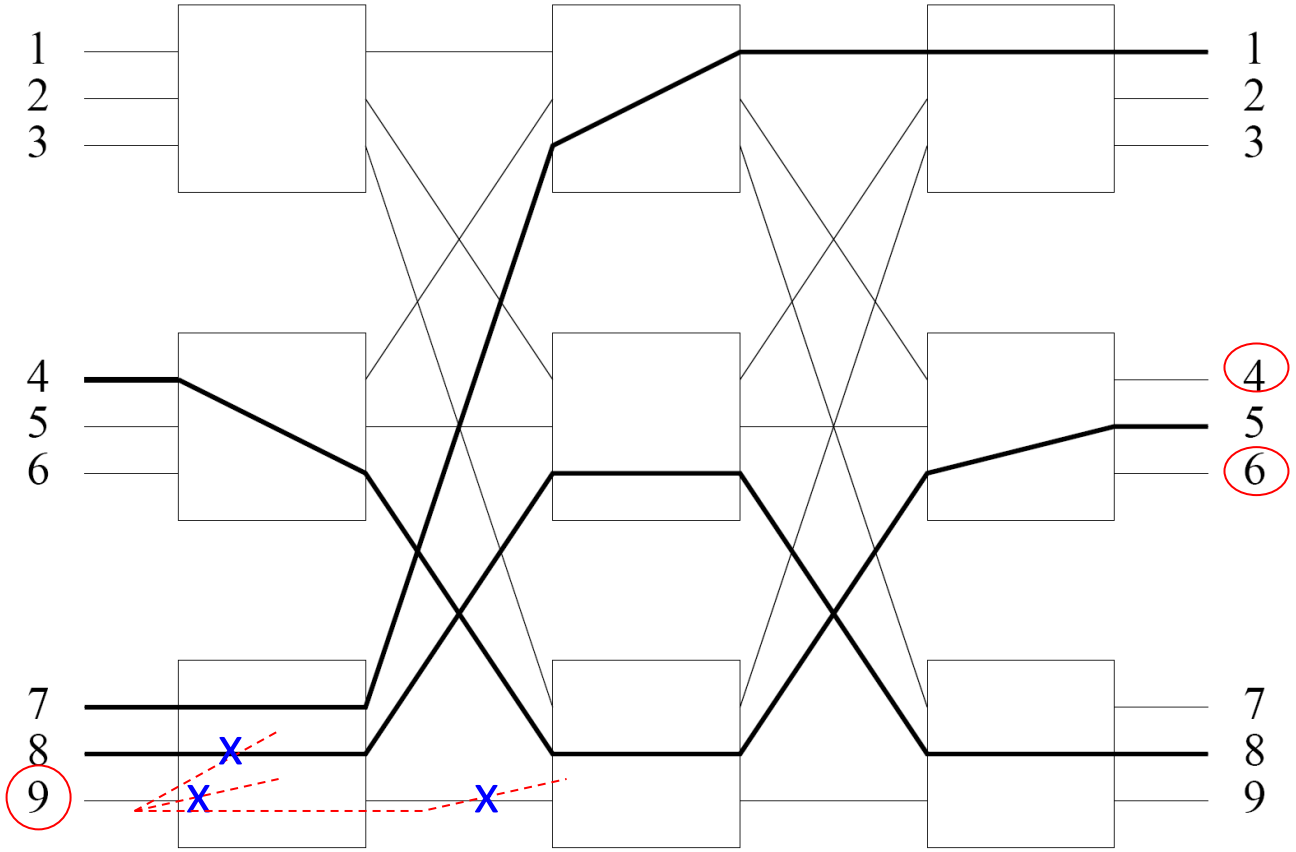
A 3-stage Clos Network



- Each switch is a crossbar switch.
- Each input has up to m possible paths to each output.
- Clos networks can be generalized to any odd number of stages. By replacing each center stage crossbar switch with a 3-stage Clos network, a 5-stage Clos networks can be constructed. By applying the same process repeatedly, 7, 9, 11,... stages are possible.

Clos, Charles (Mar 1953). "A study of non-blocking switching networks". *Bell System Technical Journal*. **32** (2): 406–424.

Internal blocking in the 3-stage Clos network



Input 9 cannot be connected to output 4 or 6, even though both of them are idle.

Nonblocking of a Clos network

- **Strictly Non-Blocking Network**

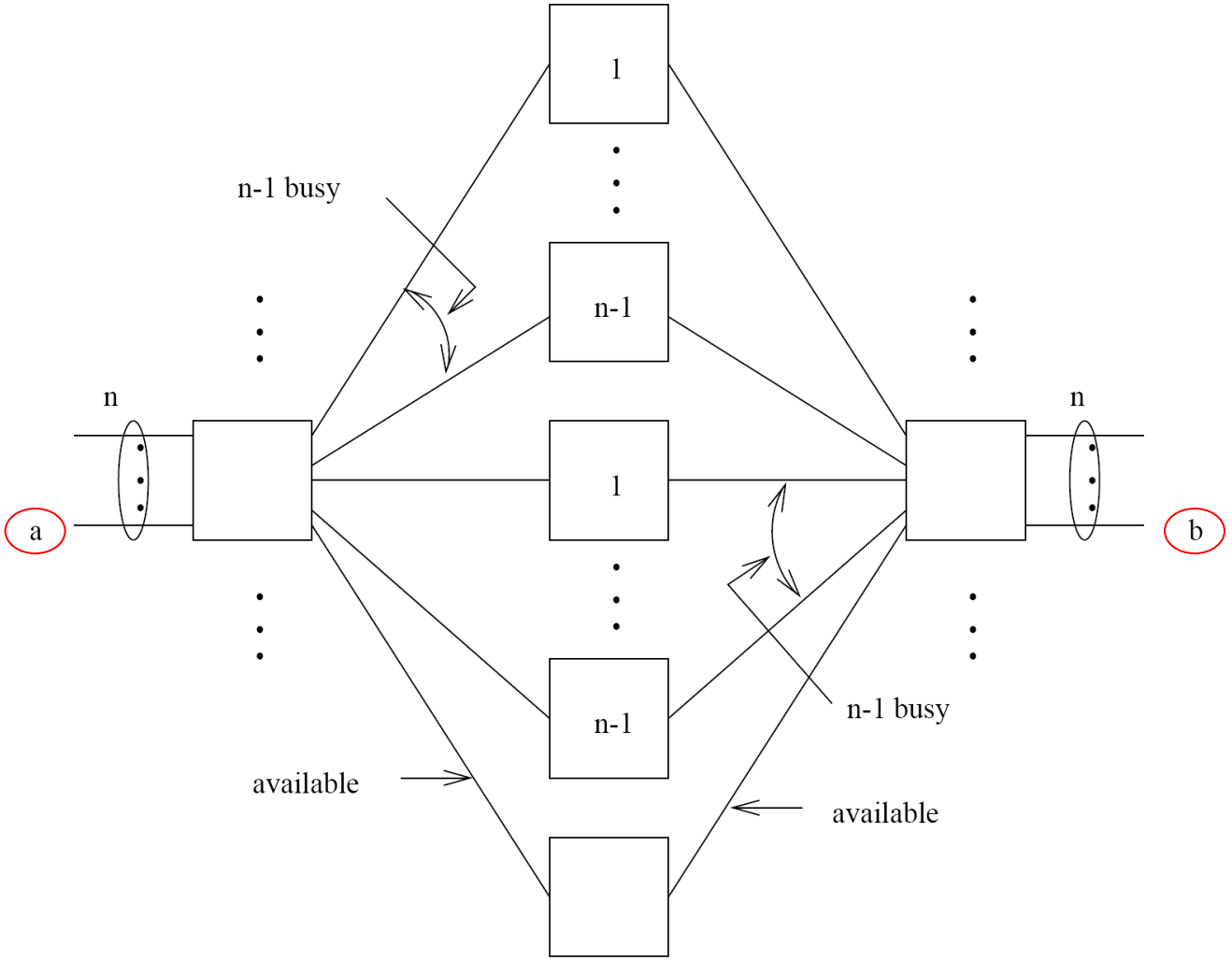
  A network is strictly Non-Blocking if it can route a request from the available paths without shuffling any existing connections. It simply means that the number of communication paths should be greater than the number of inputs.

- **Rearrangeable Non-Blocking Network**

  A network is rearrangeable Non-Blocking network when it can route a request by rearranging the existing connections. Thus, the Clos network is rearrangably nonblocking.

- By increasing the value of *m* (the number of middle-stage switch modules), the probability of blocking is reduced.
- To find the value of *m* needed for a strictly nonblocking three-stage switch, let us refer to next slide.

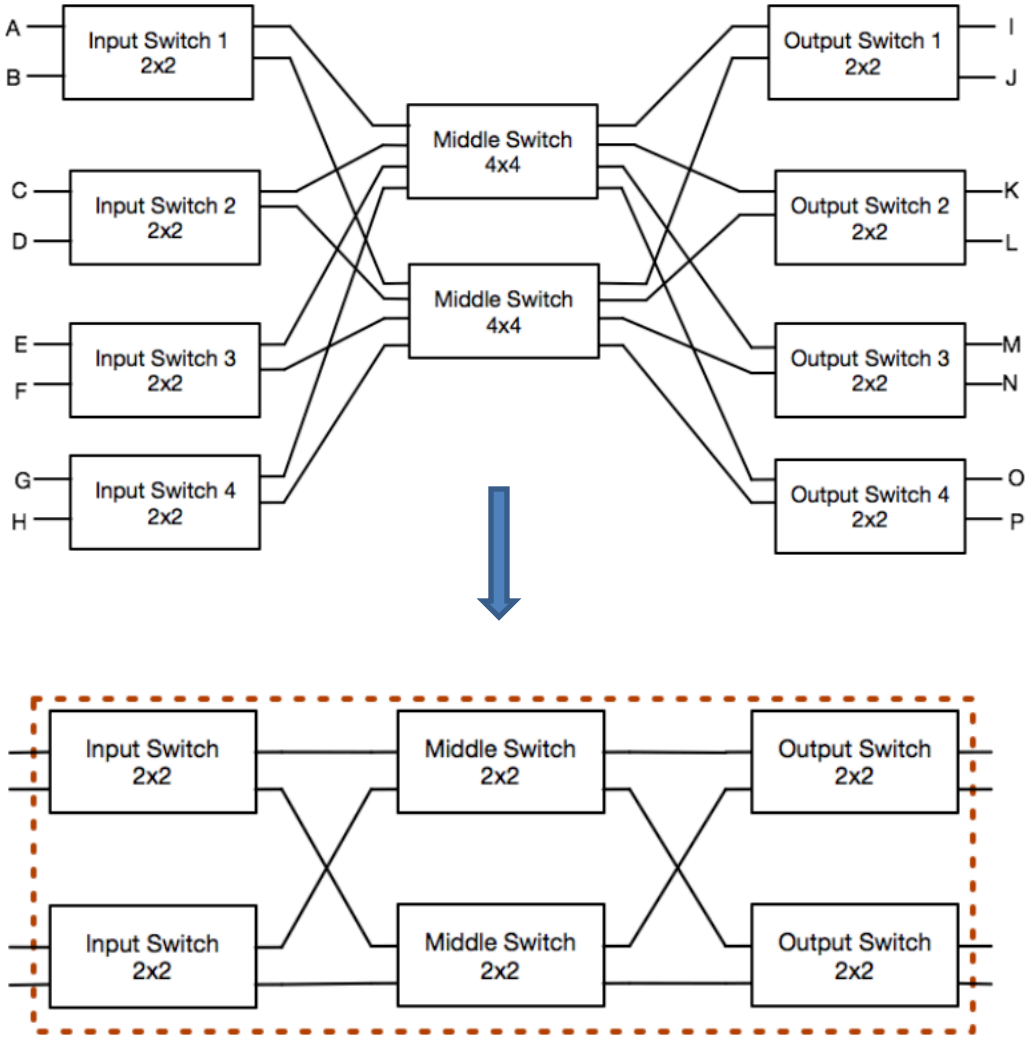
Nonblocking condition for the Clos switch



Nonblocking condition for the Clos network

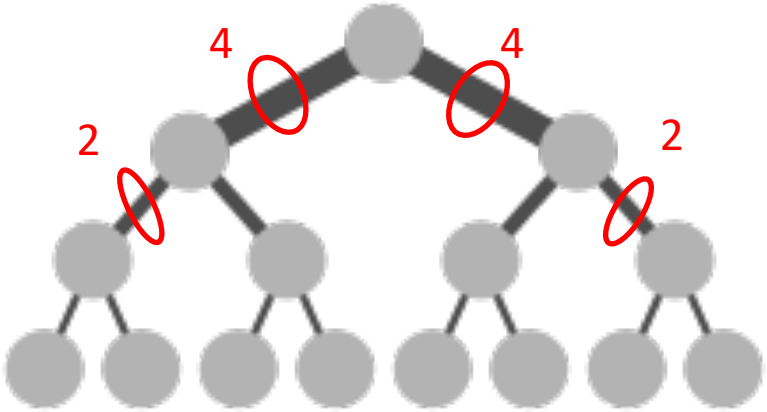
- Worst case situation for blocking occurs if all of the remaining n - 1 input lines and n - 1 output lines are busy and are connected to different middle-stage switch modules.
- Thus a total of (n - 1) + (n - 1) = $2n$ – 2 middle-stage switch modules are unavailable for creating a path from *a* to *b*.
- However, if one more middle-stage switch module exists, an appropriate link must be available for the connection.
- Thus, a three-stage Clos switch will be nonblocking if $m \geq (2n-2)+1 = 2n-1$.

Converting a 3 stage to a 5 stage Clos network



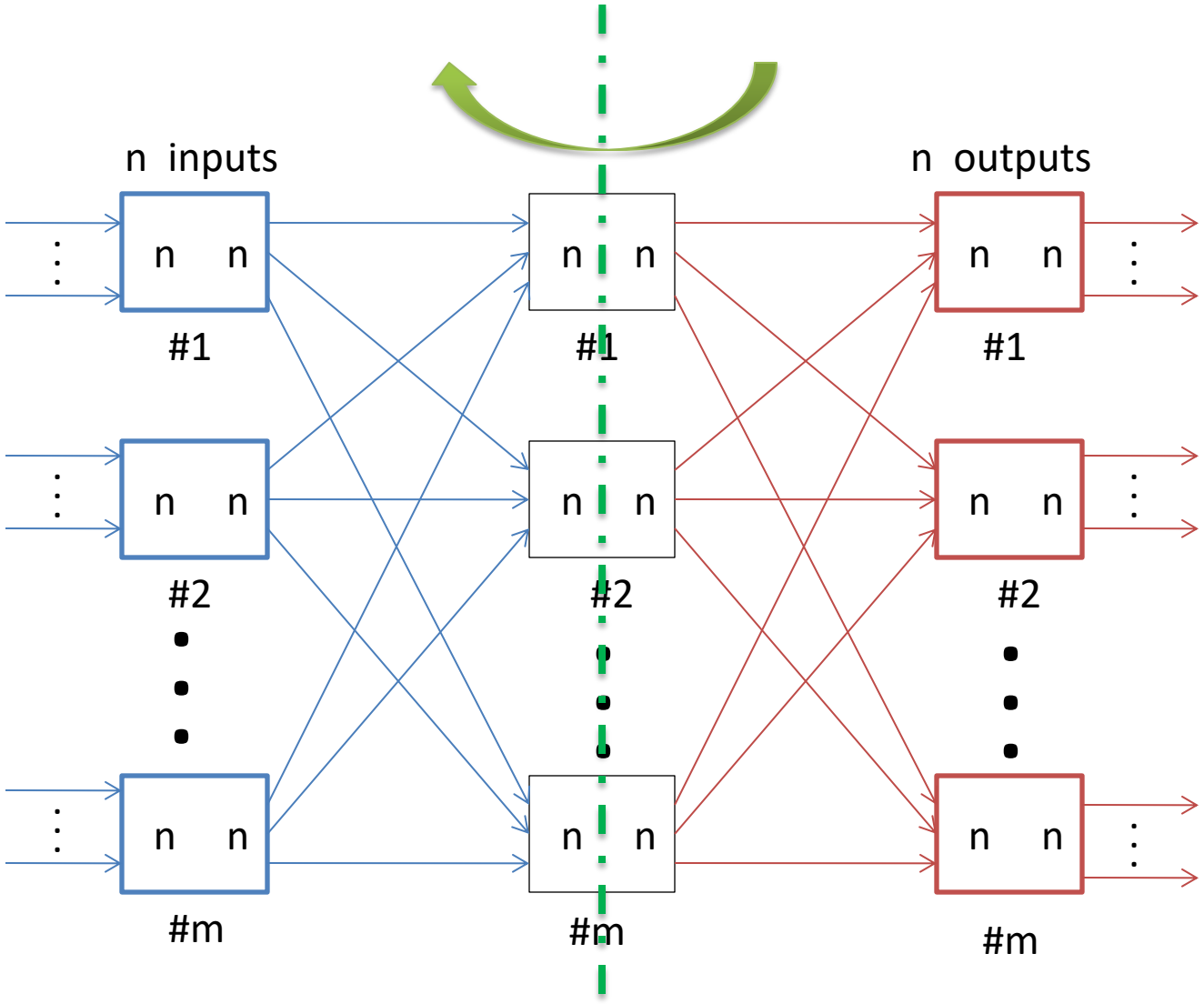
5. Fat-Tree Network

Fat-Tree Network

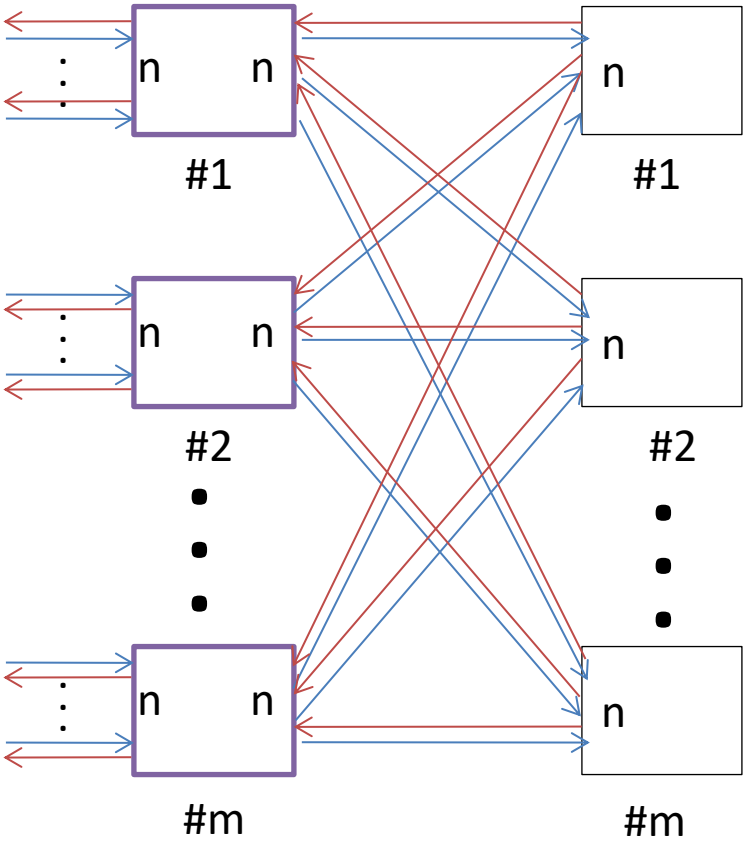


- The **fat tree network** is a universal network for provably efficient communication. It was invented by Charles E. Leiserson of MIT in 1985.
- In a tree data structure, every branch has the same thickness, regardless of their place in the hierarchy.
- In a fat tree, branches nearer the top of the hierarchy are "fatter" (thicker) than branches further down the hierarchy.
- Fat tree has identical bandwidth at any bisections. In other words, it has the same aggregated bandwidth at each level.
- Each port supports same speed as end host
- All devices can transmit at line speed if packets are distributed uniform along available paths

A folded Clos network



Fat Tree as a special case of Clos network



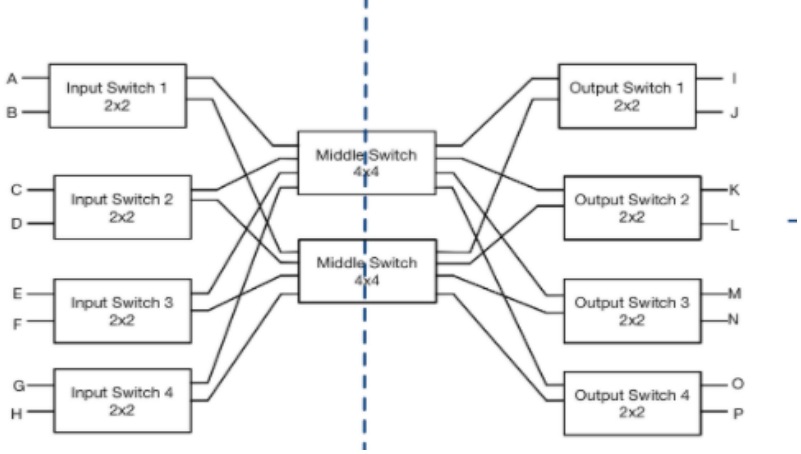
Three-stage Clos Network:
3k switches, each 2n un-directional (UD) ports

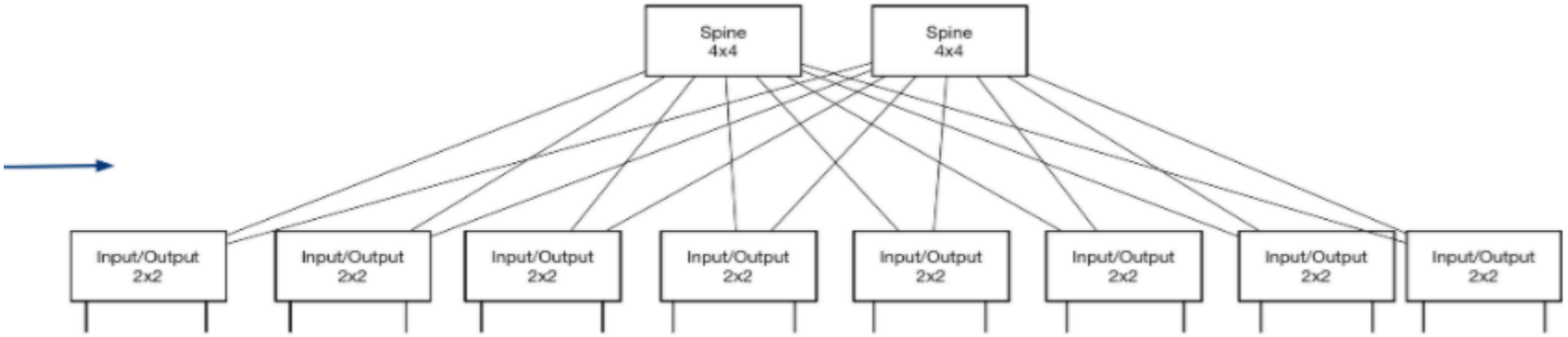
Fat –tree Network:
Layer one: m switches, each 2n BD ports
Layer two: m switches, each n BD ports

A Fat-Tree is generally represented by FT(k, L) where k is the radix of the switch and L is the levels of the Fat-Tree.

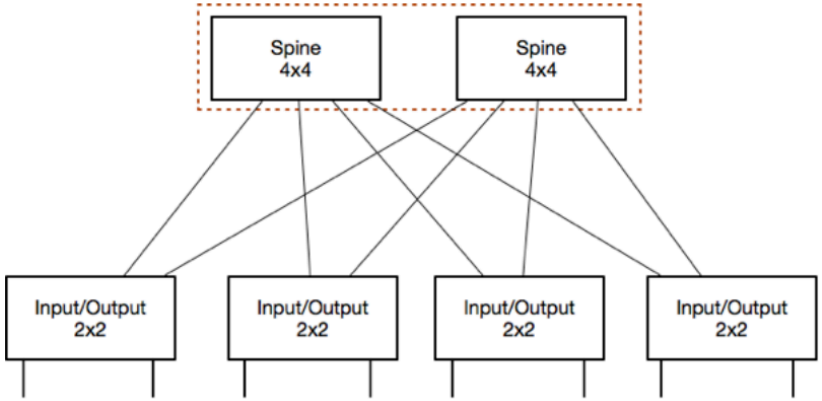
For the same switch size of k BD ports:
Layer one: m switches, each k BD ports
Layer two: m switches, each k BD ports

Folded Clos, Spine and Leaf for 3 stage Clos network



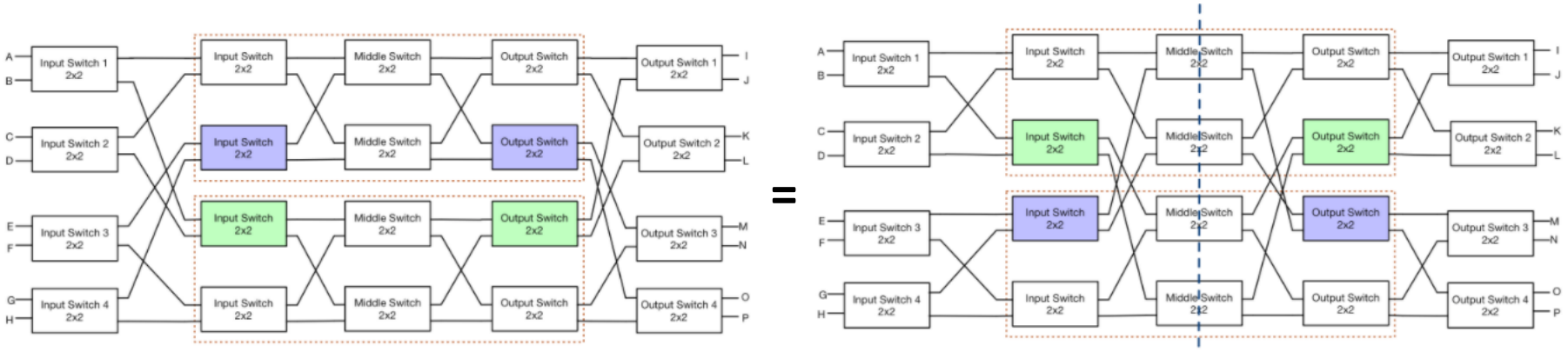


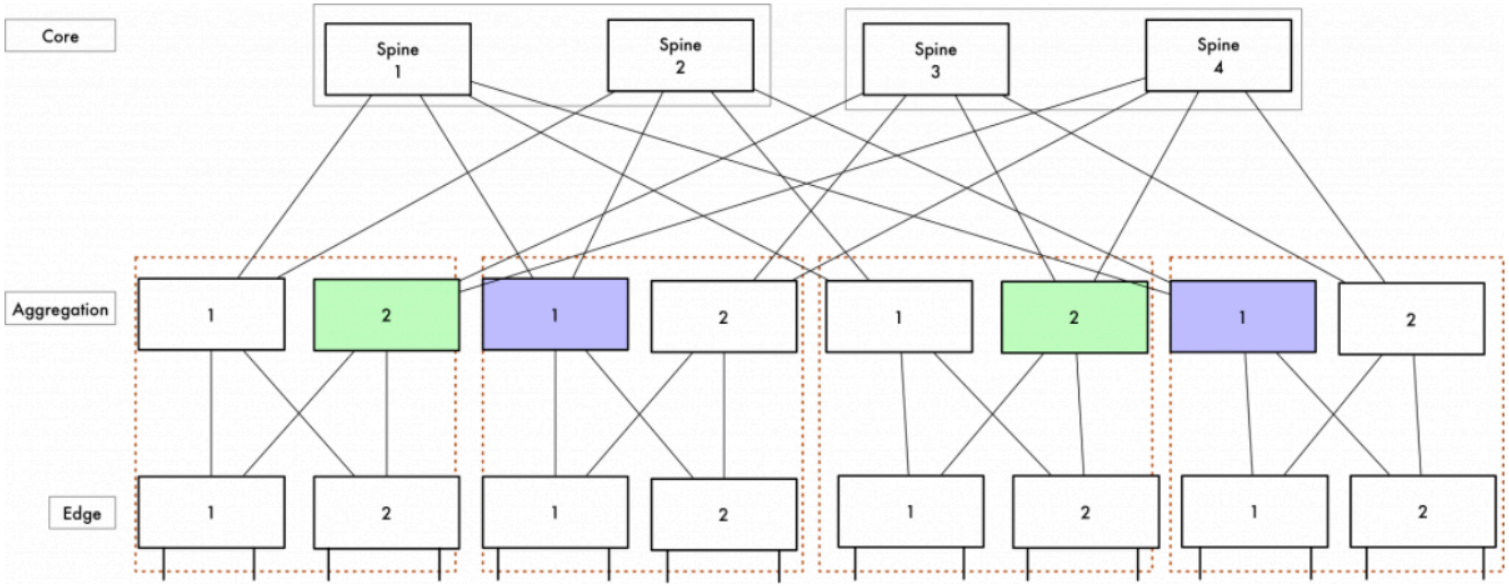
Fat-Tree with k=4, L=2



https://packetpushers.net/demystifying-dcn-topologies-clos-fat-trees-part2/

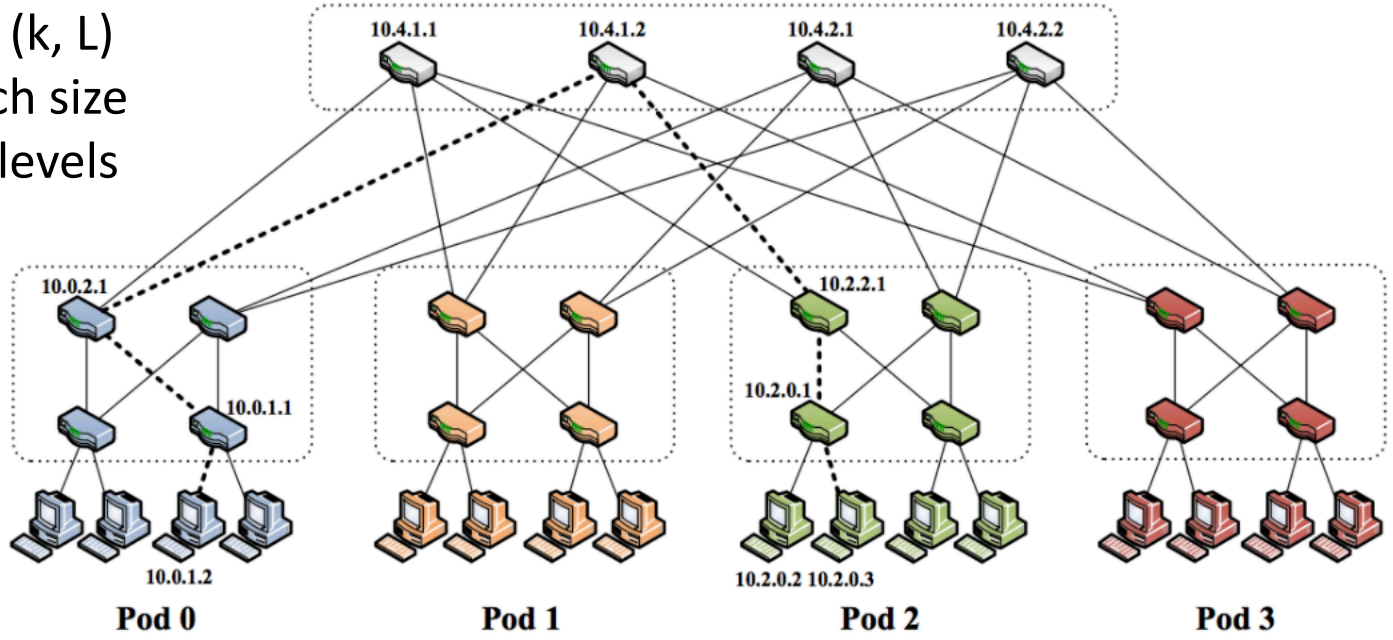
Fat-Tree with k=4, L=3





https://packetpushers.net/demystifying-dcn-topologies-clos-fat-trees-part2/

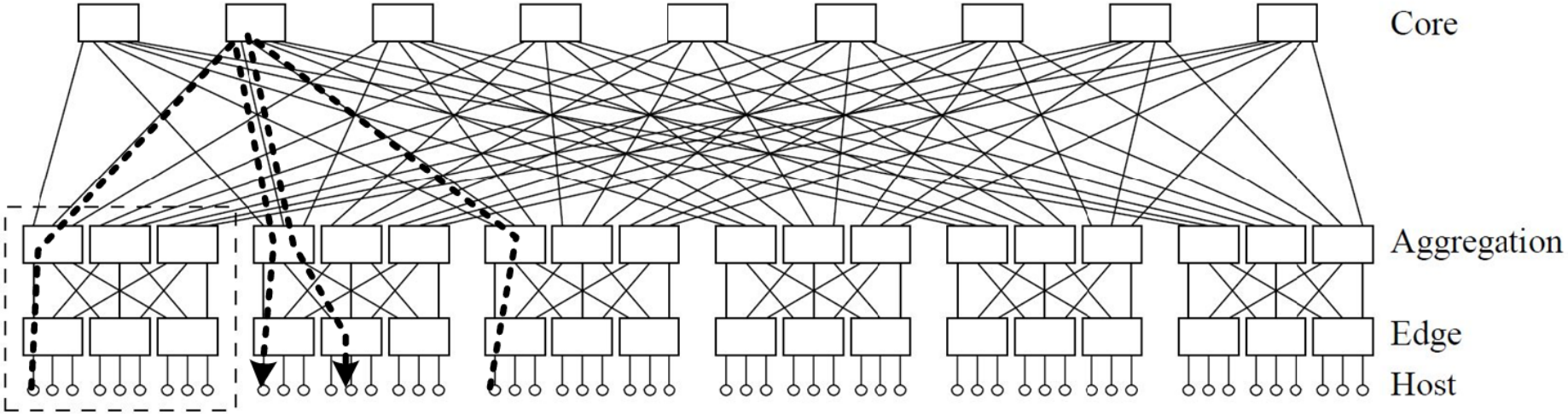
Fat-Tree (k, L)
k = switch size
L = tree levels



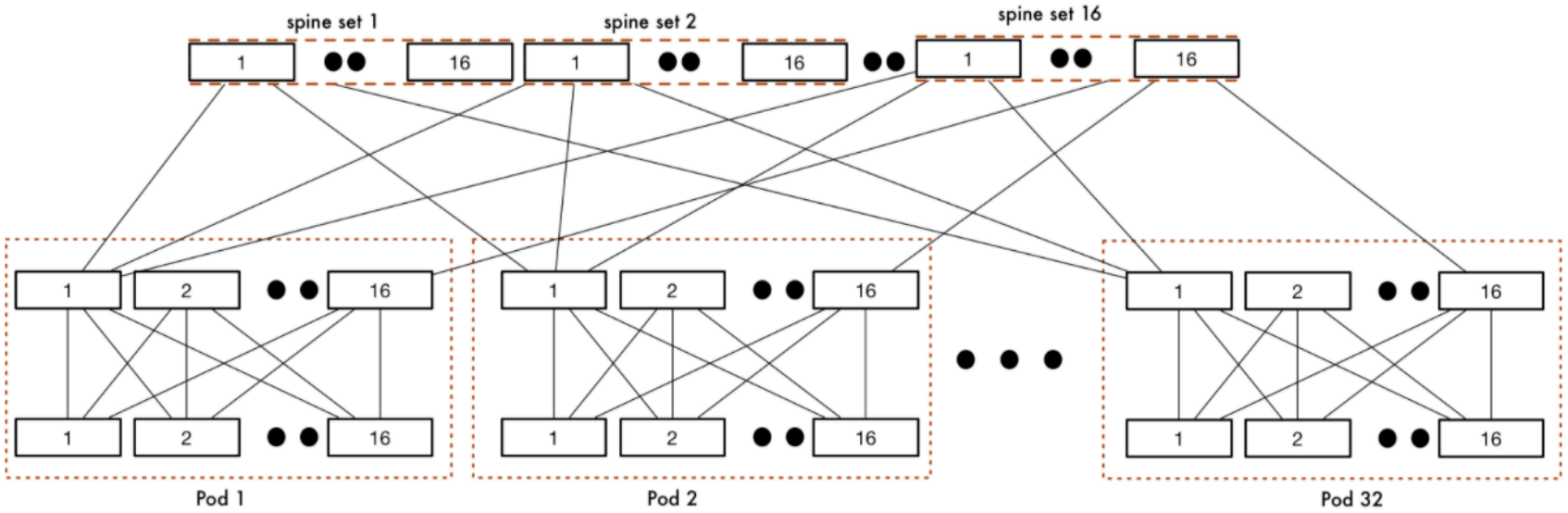
	Fat-tree with L levels	Two level Fat Tree L=2	Three Level Fat Tree L = 3	Four Level Fat Tree L=4
Number of Core switches	$\left(\frac{k}{2}\right)^{L-1}$	$\frac{k}{2}$	$\left(\frac{k}{2}\right)^{2}$	$\left(\frac{k}{2}\right)^{3}$
Number of Hosts supported	$2\left(\frac{k}{2}\right)^{L}$	$2\left(\frac{k}{2}\right)^{2}$	$2\left(\frac{k}{2}\right)^{3}$	$2\left(\frac{k}{2}\right)^{4}$
Total Switches	$(2L-1)\left(\frac{k}{2}\right)^{L-1}$	$3(\frac{k}{2})$	$5\left(\frac{k}{2}\right)^{2}$	$7\left(\frac{k}{2}\right)^{2}$
Number of Edge Switches	$2\left(\frac{k}{2}\right)^{L-1}$	k	$2\left(\frac{k}{2}\right)^{2}$	$2\left(\frac{k}{2}\right)^{3}$
Number of Pods	$2\left(\frac{k}{2}\right)^{L-2}$	NA	k	$k^2/2$

https://packetpushers.net/demystifying-dcn-topologies-clos-fat-trees-part2/

A Fat Tree using 6-port Switches



Fat-Tree FT(32, 3)

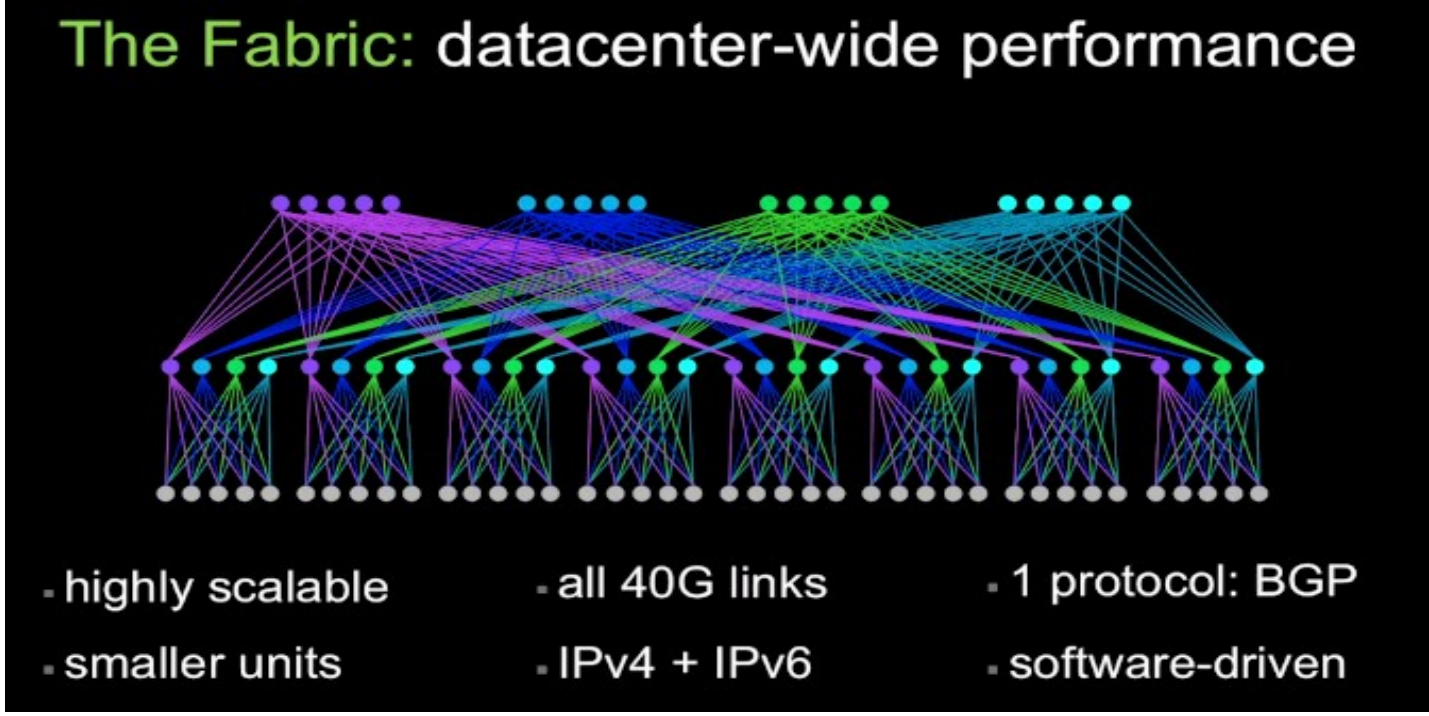


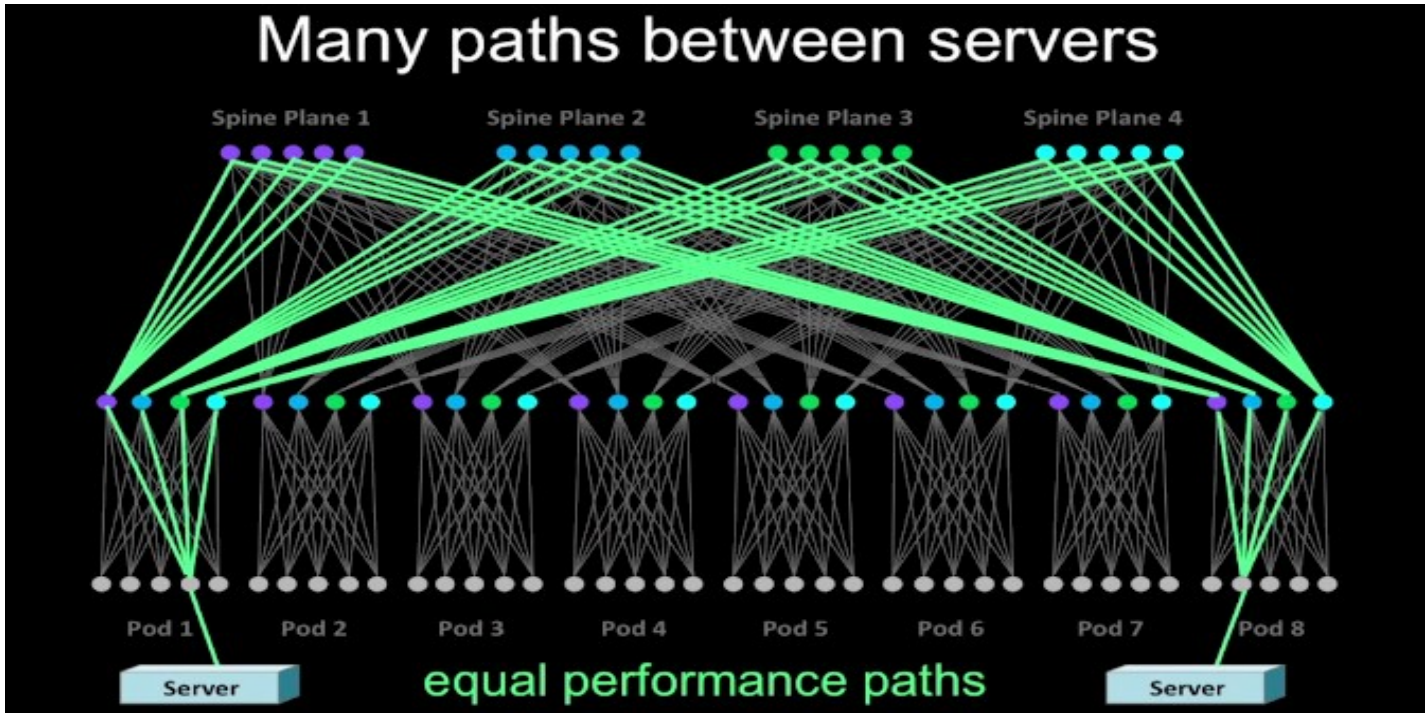
- We often want to grow the fabric incrementally as the traffic ramps up rather than deploying the full scale fabric at once.
- We could divide all the 256 spines in sets of 16 (k/2) spines which gives us a total of 16 spine sets.
- We can then just start by deploying few spine sets, e.g., 4, based on the oversubscription we are comfortable with.

https://packetpushers.net/demystifying-dcn-topologies-clos-fat-trees-part2/

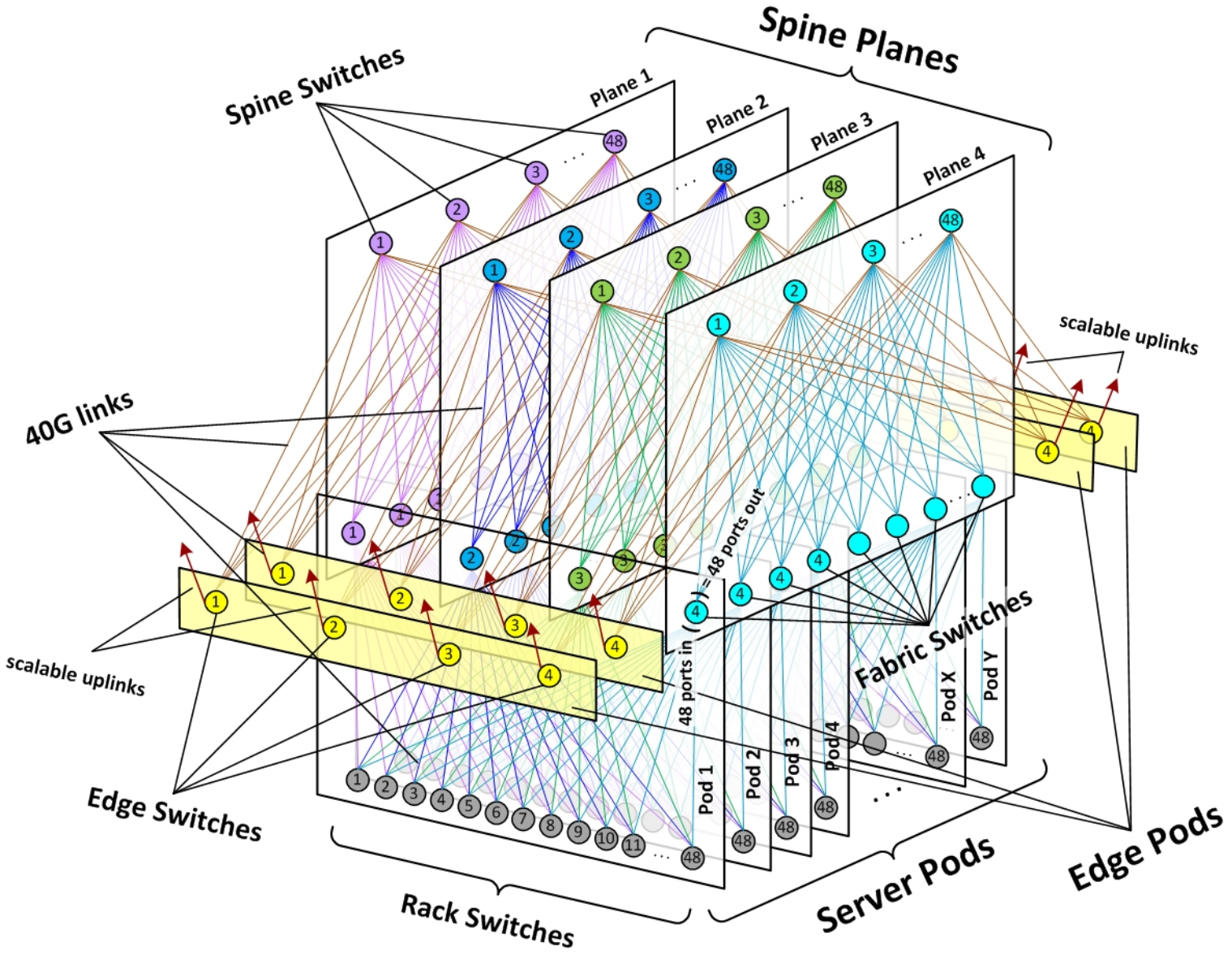
Facebook's Datacenter

- A folded 5-stage Clos
- Switch sizes:
  10x10, 8x8
- Total switch capacity:
  40G x 4 x 40 = 6.5T





FaceBook Altoona Data Center Fabric

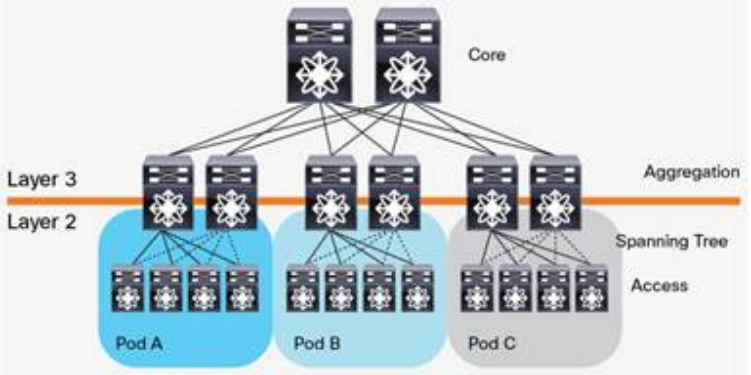


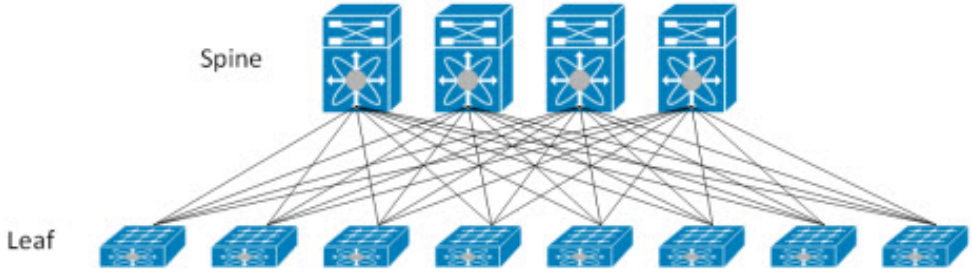
5. Commercial Switches Used in Data Centers

ARISTA

- Arista Networks is a computer networking company headquartered in Santa Clara, California, USA.
- Arista was founded on October 2004 by three former Cisco employees: Andy Bechtolsheim, David Cheriton and Kenneth Duda.
- Arista Networks is a leader in building scalable, high-performance and ultra-low latency cloud networks for modern datacenters.
- The company designs and sells multilayer network switches to deliver software-defined networking (SDN) solutions for large datacenter, cloud computing, high-performance computing and high-frequency trading environments.
- Arista's advantage: **Highest Densities**, **Lowest Power Consumption**, and **Superior buffering**
- Revenue in 2017: $1,646.2 million

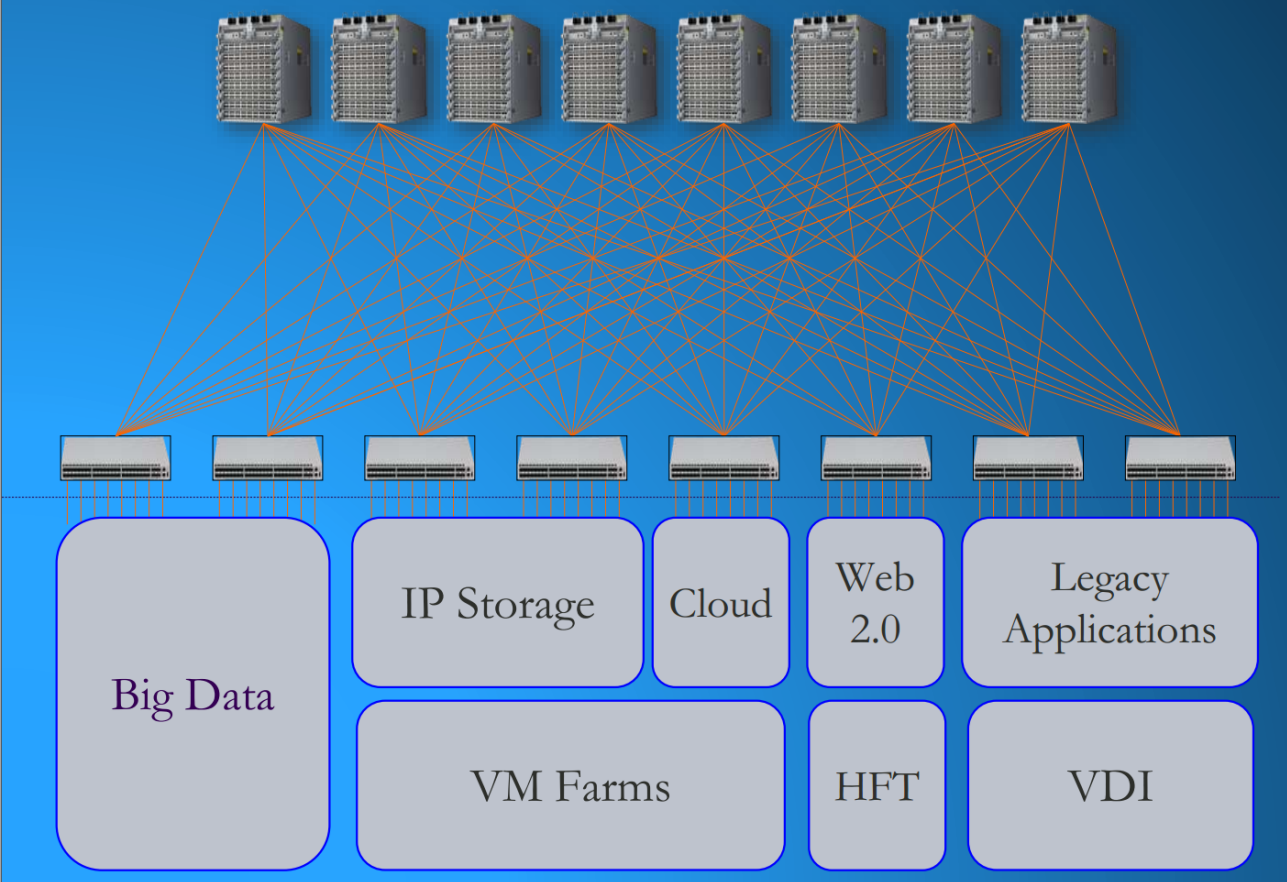
Leaf-Spine Network Topology





- The **Leaf layer** consists of access switches that connect to devices like servers.
- The **Spine layer** (made up of switches that perform routing) is the backbone of the network, where every Leaf switch is interconnected with each and every Spine switch.
- With Leaf-Spine configurations, all devices are exactly the same number of segments away and contain a predictable and consistent amount of latency for traveling information.
- If oversubscription of a link occurs, an additional spine switch can be added. If device port capacity becomes a concern, a new leaf switch can be added.
- The ease of expansion optimizes the IT department's process of scaling the network.

Built for ANY Application – Universal Cloud Architecture



Architectural Benefits

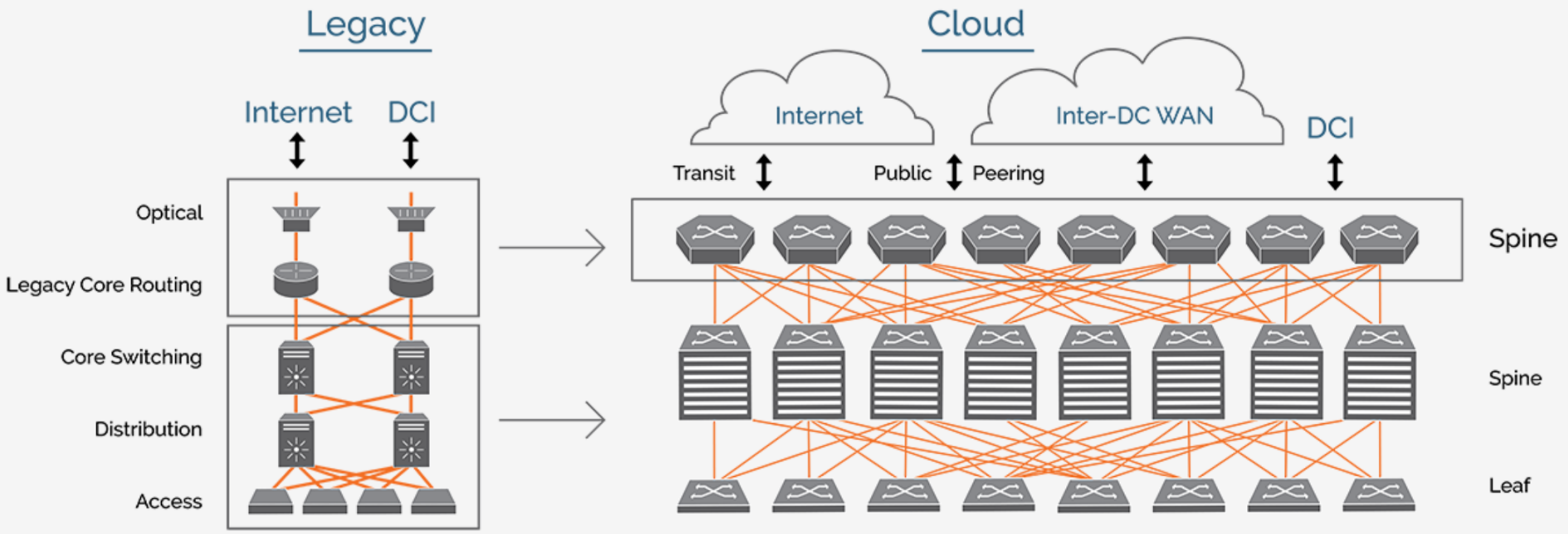
- East-West Traffic Flow
- Zero Oversubscription
- 5-10 micro second Inter-rack latency
- 100s of Racks Maximum Scaling

Network Applications

The DC Core Needs A Spine

- The first phase of cloud-scale transformation within the DC, focused on eliminating tiers and adopting a resilient Leaf Spine Layer 3 ECMP (equal-cost multi-path) Universal Cloud network design.
- However, in most DCs, there still exists a tier above the Spine called the **DC Core**, which comprises of a pair of purpose built and oversubscribed routers facing the Internet or Data Center Interconnect (DCI)
- These platforms tend to be limited in system capacity and difficult to upgrade, unable to easily adapt to a rapid growth in application traffic demands or changing traffic patterns.
- Besides, the **inter-DC traffic** traversing the DC core is growing several-fold per year, due to increased levels of bulk data transfers between application clusters spanning multiple DCs.
- These shifts in traffic patterns drive the need for a **flexible, cost-effective, scale-out** design at the DC Core that is non-blocking, supports large scale ECMP for multi-pathing, enables IP routing with Internet route scale and delivers best in class routing convergence.

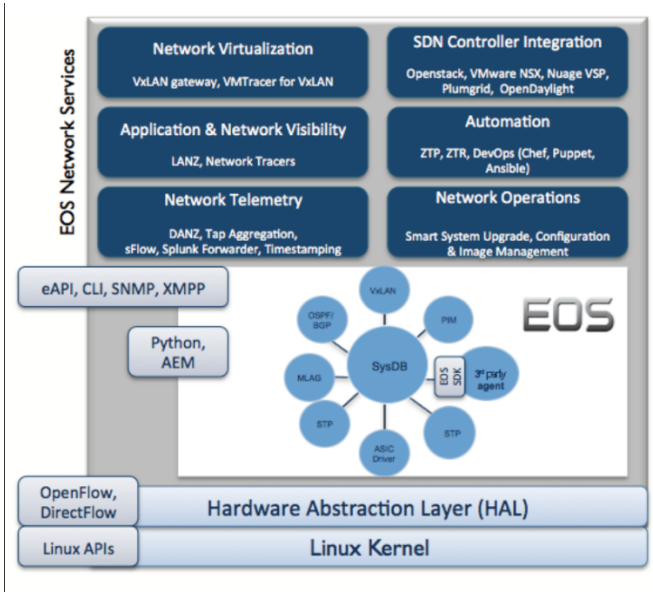
The Universal Spine: DC Spine and Core Convergence

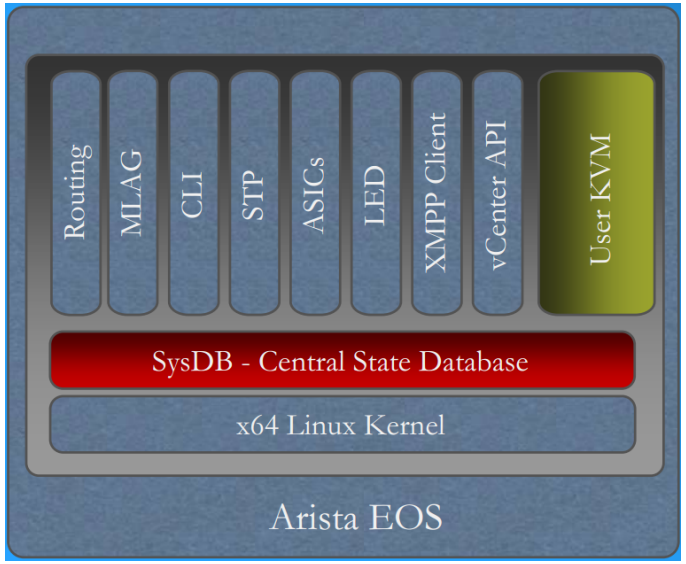


Software Architecture

Arista's own Linux-based network operating system, EOS (Extensible Operating System), is fully programmable and highly modular.

- Built on unmodified Linux Kernel: Standardized, Simple and Open.
- Single binary SW image across entire platform: Simplicity, lower OpEx.
- Unique “SysDB” database manages process state and process communication





Arista Networks: The Best Data Center Portfolio

EOS: Extensible Operating System

vEOS Available for Free

7010

48-port Data Center Class Gigabit Ethernet Switch

7048T

48-port Data Center Class Gigabit Ethernet Switch with Deep Buffering

7050 S/T/Q

1/10G & 10/40G Data Center Switches
10G SFP+ / 10G-T
Dense Virtualization
10GbE / 40GbE DC

7150S

Ultra Low Latency
24,52,64-port SFP+
1G-40GbE Switches

Intelligent Application Switch

7050X & 7280X

Dense Low Latency
32 & 64-port QSFP+
96xSFP+/8xQSFP+
48-port 10Gb w/ 100Gb Uplinks

Advanced Virtualization
Scale-out
Visibility

7300X

High Density, Modular System supporting up to 512 40GbE

Cloud Scale
Leaf and Spine
10GbE-40GbE

7500E

Lossless, High Density, Modular Switching System supporting up to 1152 Wire speed 10GbE Ports

Spine
10-40-100GbE

ARISTA

Universal Spine -- 7500R Series Modular Switches

- An extra layer intended for hyperscale deployments
- The universal spine is a higher-end system compared to the regular spine switch, connecting the data center to the outside world, dealing with large routing tables and supporting optics capable of long-haul reaches.
- The 7500R series are built for massive amounts of traffic, with the largest of them, the 7516R, capable of 576 100-Gb/s ports. The port-to-port latency is 3.5-13 micro-sec, with low power consumption as low as 25W per 100GbE port. It is ideal for high performance data center networks.
- Routing tables in the 7500R can handle 2 million routes with the latest enhancements to Arista's EOS software.

System Performance				
	7516R	7512R	7508R	7504R
				
Switching Capacity	150Tbps	115Tbps	75Tbps	38Tbps
Linecard Capacity	9.6 Tbps	9.6 Tbps	9.6 Tbps	9.6 Tbps
10GbE Interfaces	2,304	1,728	1,152	576
25GbE Interfaces	2,304	1,728	1,152	576
40GbE Interfaces	576	432	288	144
50GbE Interfaces	1,152	864	576	288
100GbE Interfaces	576	432	288	144
Forwarding Rate	69 Bpps	51 Bpps	34.5 Bpps	17.3 Bpps
Total Buffer	384GB	288GB	192GB	96GB
Rack Units	29	18	13	7
Airflow	Front to Rear	Front to Rear	Front to Rear	Front to Rear

Universal Leaf – 7280R Series Fixed Configuration switches

- A leaf needs to offer high availability, telemetry, dynamic and deep buffering, and different routing options, etc. But, most of the leaf switches are purpose built to be good at one or two of these things, meaning customers must buy a wide variety of leaf switches.
- The concept behind the Universal Leaf is fairly simple – build a switch that can be a common leaf for any use case, such as routing, general computing, IP storage and digital media.
- The Arista 7280R Series provides a combination of dynamic and deep buffers, high performance and full internet scale routing in a high density and flexible fixed configuration switch combined with extensive features such as VXLAN and LANZ.
- The 7280R leverage the Broadcom Jericho chipset and Arista's CloudVision management software to handle demanding workloads at scale and maintain performance.

	7280CR-48	7280QR-C72	7280QR-C36	7280SR-48C6	7280TR-48C6	7280QRA-C36S	7280SRA-48C6	7280TRA-48C6	7280SRAM-48C6
									
Description	48-Port QSFP100 and 8-Port QSFP+	72-Port QSFP (56x40G and 16x100G)	36-Port QSFP (24x40G and 12-Port 100G)	48-Port SFP+ and 6-Port QSFP100	48-Port RJ45 and 6-Port QSFP100	36-Port QSFP (18x40G and 12-Port 100G)	48-Port SFP+ and 6-Port QSFP100	48-Port RJ45 and 6-Port QSFP100	48-Port SFP+ and 6-Port QSFP100
L2/3 Throughput	10.24 Tbps	6.4 Tbps	4.32 Tbps	2.16 Tbps	2.16 Tbps	3.84 Tbps	2.16 Tbps	2.16 Tbps	2.16 Tbps
L2/3 PPS	5.76 Bpps	2.88 Bpps	1.44 Bpps	720 Mpps	720 Mpps	1.44 Bpps	720 Mpps	720 Mpps	720 Mpps
Latency	From 3.8us	From 3.8us	3.8us	3.8us	3.8us	From 3.8us	From 3.8us	From 3.8us	From 3.8us
Typical Power Draw	1363W	725W	324W	263W	290W	419W	313W	290W	382W
Total System Buffer	32GB	16GB	8GB	4GB	4GB	8GB	4GB	4GB	4GB
Rack Units	2	2	1	1	1	1	1	1	1

Other Series

- 7060X and 7260X Series: Cloud Optimized Switches

  - The Arista 7060X and 7260X Series are a range of 1RU and 2RU high performance 25GbE, 40GbE, 100GbE and 400GbE high density, fixed configuration, data center switches with wire speed layer 2 and layer 3 features, and advanced features for software driven cloud networking. The Arista 7060X and 7260X deliver a rich choice of interface speed and density allowing networks to seamlessly evolve from 10GbE/40GbE to 25GbE/100GbE and 200GbE/400GbE.

- 7100 Series: Ultra Low Latency Switches (350-380ns)

  - The Arista 7100 Series features the industry's highest density and lowest latency 10 Gigabit Ethernet switching solution and the first with an extensible modular network operating system. With breakthrough price-performance, the Arista 7100 Series enables 10 Gigabit Ethernet to be deployed everywhere in the data center, which can significantly improve server utilization and consequently data center power efficiency.

	7060PX4-32	7060DX4-32
Description	32-Port 400G OSFP and 2 SFP+	32-Port 400G QSFP-DD and 2 SFP+
Maximum 400G Ports	32	32
L2/L3 Throughput	12.8 Tbps	12.8 Tbps
L2/3 PPS	Up to 8 Bpps	Up to 8 Bpps
Latency	700ns	700ns
Typical Power Draw	640W	690W
AC and DC	Yes	Yes
Total System Buffer	64MB	64MB
RU	1	1

	7150 Switches		
	7150S-24	7150S-52	7150S-64
Description	7150 Switch 24-Port SFP+	7150 Switch 52-Port SFP+	7150 Switch 48-Port SFP+ 4 QSFP+
Total Ports	24	52	64
SFP+ Ports	24	52	48
L2/3 Throughput	480 Gbps	1.04 Tbps	1.28 Tbps
L2/3 PPS	360 Mpps	780 Mpps	960 Mpps
Latency	350ns	380ns	380ns
Typical Power Draw	191 W	191 W	224 W

6. High-Speed Switch Chips

Broadcom was an American fableless semiconductor company that made products for the wireless and broadband communication industry.
Broadcom Corporation was founded from UCLA in 1991.
It was acquired by Avago Technologies in 2016 and currently operates as a wholly owned subsidiary of the merged entity Broadcom Inc.
-- Wikipedia

Broadcom's product portfolio serves multiple applications within seven primary target markets: data center, networking, software, broadband, wireless, storage and industrial.
-- https://www.broadcom.com/products/

Ethernet Connectivity, Switching, and PHYs

Designed for today's enterprise, cloud, and data center environments, Broadcom is a leading supplier of Ethernet network adapters, controllers, switches, PHYs and PHY transceivers.

SELECT PRODUCTS

Select

Stingray™ SmartNIC adapters and IC

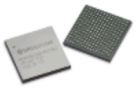
Stingray™ Ethernet SmartNIC adapters and SoCs provide high performance, programmable networking in compact PCIe form factors for virtualized and bare metal services

SEE PRODUCTS

Automotive Ethernet PHYs

Broadcom offers the industry's most comprehensive portfolio of copper PHY transceivers for the automotive market. These PHYs support 100M and 1000M transmission rates and are conforming to the BroadR-Reach®, IEEE 100BASE-T1, 100BASE-Tx, 1000BASE-T1 standards.

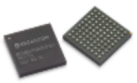
SEE PRODUCTS



Automotive Switches

Offered in both low and high port configurations supporting both 100BASE-T1 and 1000BASE-T1 Ethernet, Broadcom's Automotive Ethernet Switching solutions are highly integrated, cost effective, and based on a field-proven, industry-leading architecture.

SEE PRODUCTS



Automotive Microcontrollers

Building on Broadcom's extensive BroadR-Reach 100BASE-T1 Ethernet portfolio, Broadcom's Automotive Microcontroller solutions are highly integrated, cost effective, and designed for automotive applications. These Automotive Ethernet microcontrollers provide a wide range of interfaces to enable customers to design Ethernet based endpoint applications such as cameras, audio amplifiers, graphic displays etc.

SEE PRODUCTS

Industrial BroadR-Reach

Innovative BroadR-Reach technology extends the range of twisted-pair connections so that Ethernet and IP services can be deployed for longer reach with more cost-effective cabling.

SEE PRODUCTS

Ethernet Switches and Switch Fabric Devices

Broadcom switching technology, along with our leading-edge software solutions, can be found in a wide array of products for home and small business, data center, enterprise, and service provider networks.

SELECT PRODUCTS

Select

Four switch chip families

RoboSwitch
Highly integrated, cost effective and smart-managed. Combine all the functions of a high-speed switch system including packet buffer, PHY transceiver, address management, etc. into a single CMOS device.

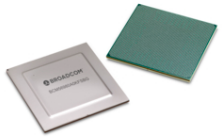
StrataConnect
Supply switches support 1Gigabit, 2.5 Gigabit, 5 Gigabit, 10 Gigabit and 25 Gigabit I/O speeds covering switch bandwidths from tens of gigabits to hundreds of gigabits.

StrataXGS
- Tomahawk®
- Trident ®

Covers multiple markets. The StrataXGS family of products scales from multi-gigabit to multi-terabit switches.

StrataDNX
Offers the greatest extensibility and scalability, with the ability to scale both tables and buffering with external lookup engines or memory.

Roadmap of Switch Chips



12.8 T/s BCM56980
in 2018

Since Trident in 2010….
a 10x increase in bandwidth

BCM56970 TOMAHAWK2 — 64 port 100GE

BCM56960 TOMAHAWK — 32 port 100GE

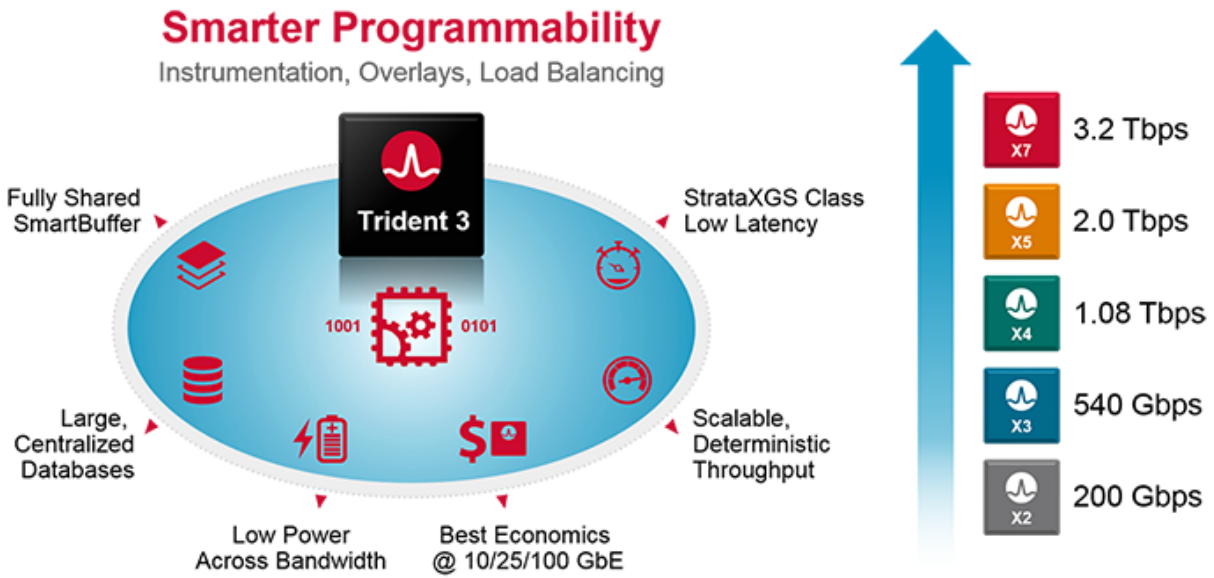
BCM56850 TRIDENT2 — 32 port 40GE

BCM56840 TRIDENT — 64 port 10GE

2010	2012	2014	2016
640G/s	1280G/s	3200G/s	6400G/s

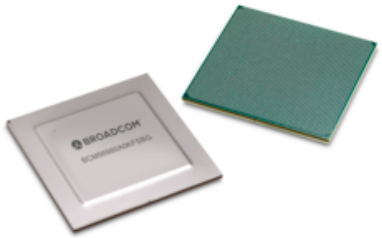
Trident 3 (BCM56870 series)

- With support for up to 32 ports of 100GbE, the StrataXGS® Trident 3 switch series can be used to build highly scalable, low-power, feature-rich Top-of-Rack (ToR), Aggregation and Spine switches.
- The programmable pipeline allows new features to be added after deployment by means of in-field upgrade, providing capex investment protection for operators.
- Trident 3 programmability enables the latest, cloud-scale instrumentation and telemetry features, for example in-band telemetry.
- Trident 3 programmability can also be used to add new software-defined forwarding and database reconfiguration features.



BCM56980 SERIES

12.8 Tb/s StrataXGS® Tomahawk® 3 Ethernet Switch Series



- Provides 256 integrated SerDes with 56G-PAM4 (4-level pulse amplitude modulation) , supporting 200GbE and 400GbE
- **Up to 32x400GbE, 64x200GbE or 128x100GbE ports**
- Over 40 percent reduction in power per port compared to previous Tomahawk devices
- New instrumentation features such as inband telemetry, latency histograms, switch utilization monitors and more
- New elephant flow feature can detect high-bandwidth, long-lived flows and take QoS actions to protect Mice Flows
- More than doubles the IP route forwarding scale compared to previousTomahawk.
- Dynamic Load Balancing and Dynamic Group Multipathing enhance ECMP
- New shared-buffer architecture offers 4X higher burst absorption
- PCIe Gen3 x4 host CPU interface with on-chip accelerators improves control-plane performance by up to 5X
- Compatibility with previous generation Tomahawk and Tomahawk 2 devices
- Single-chip solution for data center Top-of-Rack, Aggregation and Spine Switches
- Data center Fixed and Chassis/Modular switches

Who Needs 400 Gbit/s?

1. **Cloud service provider**
   - Vendors like Amazon, Microsoft and IBM are building massive cloud-scale data centers and they need the efficiency and density that 400GE provides, in addition to the massive performance upgrade.

2. **Telecommunications provider**
   - The telecom operators are dealing with a flood of traffic. With the average person having 3.4 connected mobile devices by 2020, telecoms have to deal more traffic through their already massive data center.

3. **Distributed businesses and campuses**
   - Traffic from mobile devices is actually moving off cellular networks to Wi-Fi, putting additional strain in offices and campus networks.

4. **Bandwidth-hungry applications**
   - Bandwidth-hungry applications, such as streaming video, gaming and digital marketing that rely on rich media require higher speeds.