

Cloud and Datacenter Networking

Università degli Studi di Napoli Federico II

Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione DIETI

Laurea Magistrale in Ingegneria Informatica

Prof. Roberto Canonico

Datacenter networking infrastructure

Part I





- ▶ Switched Ethernet basic concepts
- ▶ Gigabit Ethernet standard evolution
- ▶ A datacenter's networking infrastructure
- ▶ Organization and topology of a datacenter network
- ▶ Link aggregation
- ▶ VLANs

Connecting N hosts: full mesh

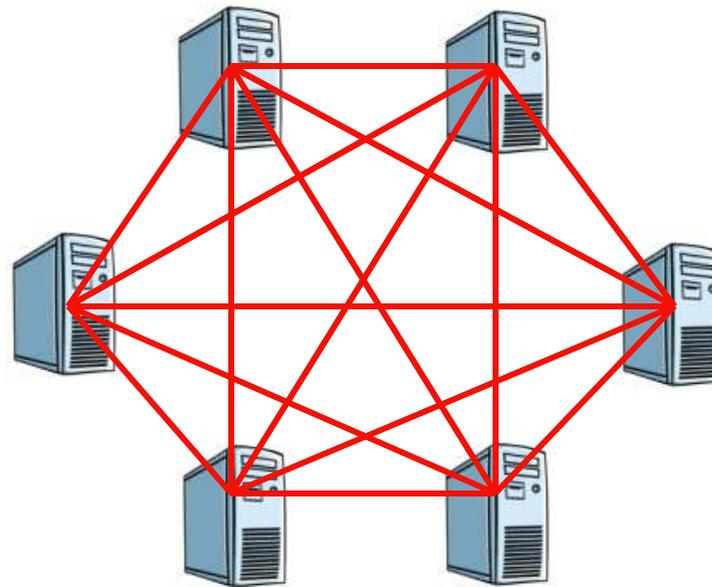


▶ Advantages

- ▶ In case of full-duplex NICs, $N \cdot (N-1)$ simultaneous transmissions are possible

▶ Disadvantages

- ▶ # NICs = $N \cdot (N-1)$ proportional to N^2
- ▶ # bidirectional links = $(N \cdot (N-1) / 2)$ proportional to N^2
- ▶ Cabling is expensive
- ▶ Costly and not scalable

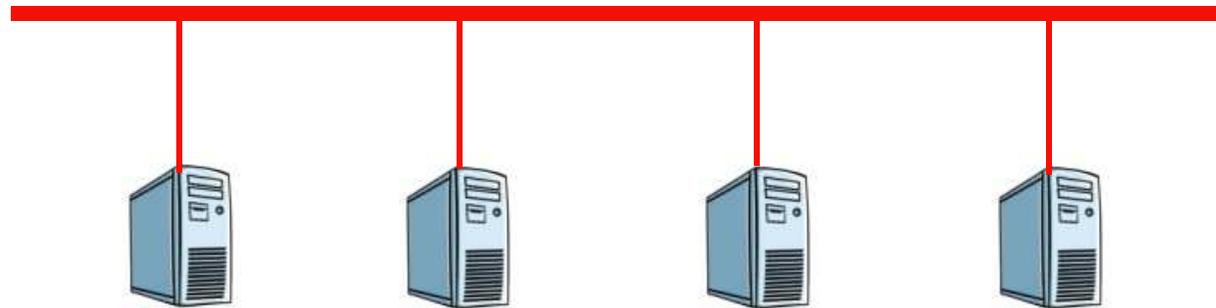


Full mesh

Connecting N hosts: bus



- ▶ **Advantage over full mesh**
 - ▶ Cheaper: 1 NIC per host
 - ▶ Simpler and cheaper cabling
- ▶ **Disadvantages**
 - ▶ Transmission capacity is shared among N hosts
 - ▶ Medium Access Control (CSMA/CD) is needed to regulate access to the shared bus
 - ▶ Cabling a star topology would be simpler in a building



- ▶ **CSMA – *Carrier Sense Multiple Access***
- ▶ **CS: Listen before transmitting**
 - ▶ If a device detects a signal from another device, it waits for a specified amount of time before attempting to transmit
 - ▶ When there is no traffic detected, a device transmits its message
 - ▶ While this transmission is occurring, the device continues to listen for traffic or collisions on the LAN
 - ▶ After the message is sent, the device returns to its default listening mode
- ▶ **CD – *Collision Detection***
 - ▶ When a device is in listening mode, it can detect when a collision occurs on the shared media, because all devices can detect an increase in the amplitude of the signal above the normal level
 - ▶ When a collision occurs, the other devices in listening mode, as well as all the transmitting devices, detect the increase in the signal amplitude



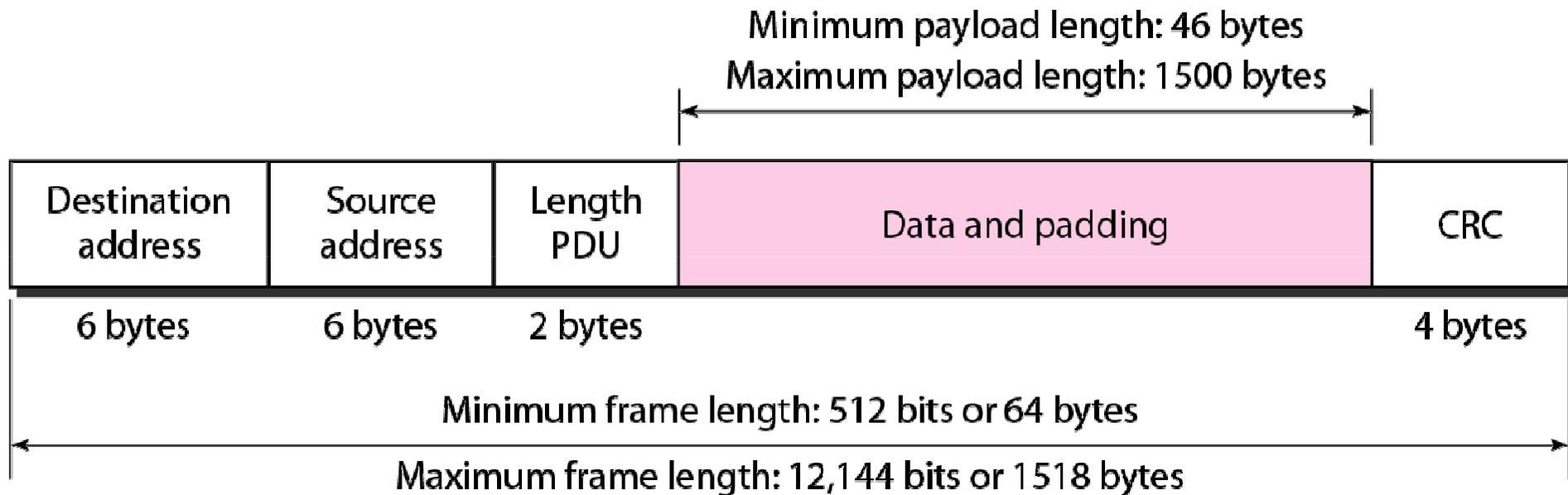
- ▶ **CS: Listen before transmitting**
 - ▶ If a device detects a signal from another device, it waits for a specified amount of time before attempting to transmit
 - ▶ When there is no traffic detected, a device transmits its frame
 - ▶ After the frame is sent, the device returns to its default listening mode
- ▶ **CD – Collision Detection**
 - ▶ During a frame transmission, the device continues to listen for collisions
 - ▶ When a device is in listening mode, it can detect when a collision occurs on the shared media, because all devices can detect an increase in the amplitude of the signal above the normal level
- ▶ **Jam Signal**
 - ▶ When a collision is detected, the transmitting devices send out a jamming signal
 - ▶ The jamming signal notifies the other devices of a collision, so that they invoke an exponential backoff algorithm
 - ▶ This backoff algorithm causes transmitting devices to stop transmitting for a random amount of time, so that the devices that were involved in the collision have a chance that they do not try to send traffic again at the same time

Ethernet frame format



- ▶ **Destination MAC Address (6 bytes)** is the identifier for the intended recipient
 - ▶ The address in the frame is compared to the MAC address in the device
 - ▶ If there is a match, the device accepts the frame
 - ▶ Special destination address FF:FF:FF:FF:FF:FF for broadcast
 - ▶ Special destination addresses for LAN multicast
- ▶ **Source MAC Address Field (6 bytes)** identifies the frame's originating NIC
- ▶ **Length/Type Field (2 bytes)**
 - ▶ If this field's value $\geq 0x0600 = 1536_{10}$, the contents of the Data Field are decoded according to the protocol indicated (works as Type field)
 - ▶ If this field's value $< 0x0600$ then the value represents the length of the data in the frame (works as Length field)

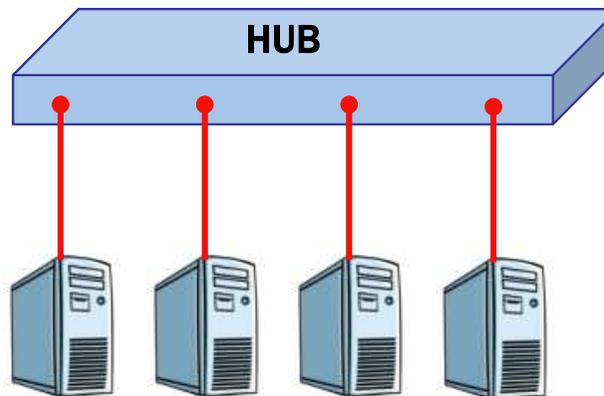
Ethernet frame: min and max length



Connecting N hosts: hub



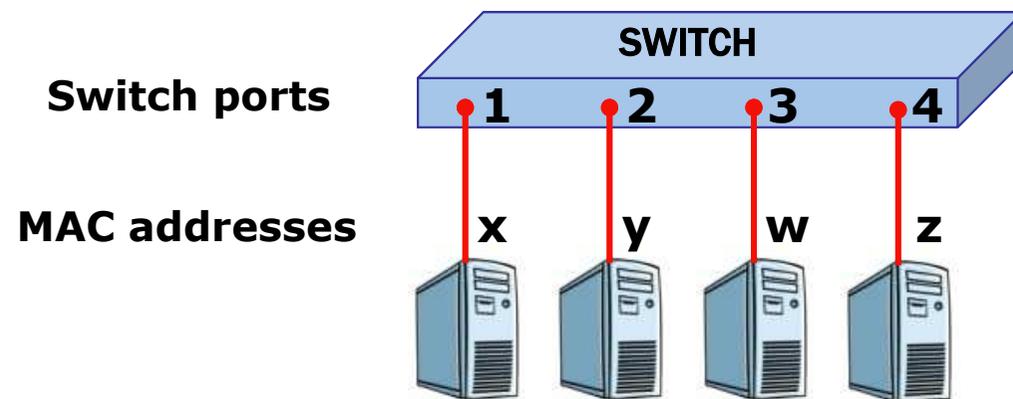
- ▶ An Ethernet hub retransmits a frame to all the ports but the one on which the frame entered the hub
- ▶ Each host compete for the shared capacity with the other $N-1$ hosts attached to the hub, as for the bus topology
- ▶ Advantage over bus
 - ▶ Simpler and cheaper cabling w.r.t. the bus topology (UTP cables)



Connecting N hosts: switch



- ▶ A switch determines how to handle incoming frames by using its *MAC address table*
- ▶ A switch builds its MAC address table by recording the source MAC addresses of the nodes connected to each of its ports (*learning*)
- ▶ Once a specific node's MAC address is associated to a specific switch port in the MAC address table, the switch knows where (i.e. on which port) to send subsequent frames destined for that specific MAC address
- ▶ Before a switch learns the port on which a given MAC address is reachable, the switch transmits a frame destined for that unknown MAC address to all the ports but the one on which the frame entered the switch

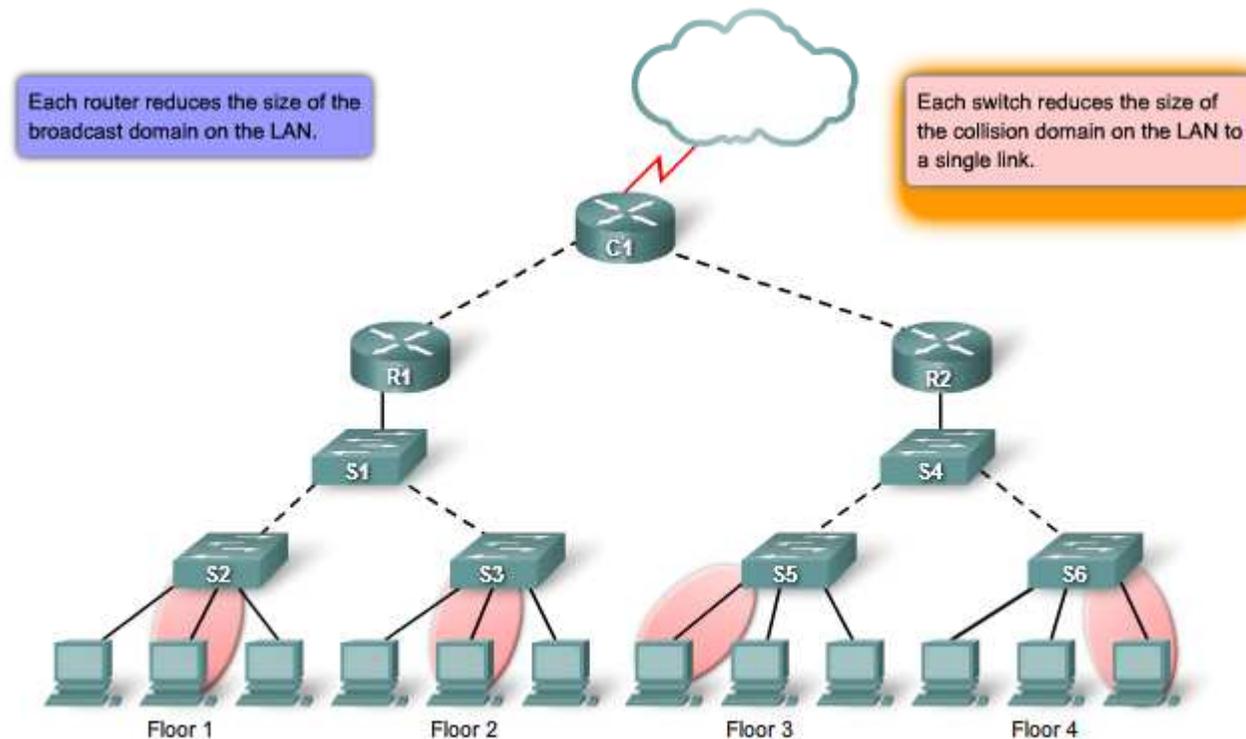


MAC address	port
x	1
y	2
w	3
z	4

Switches and collision domains



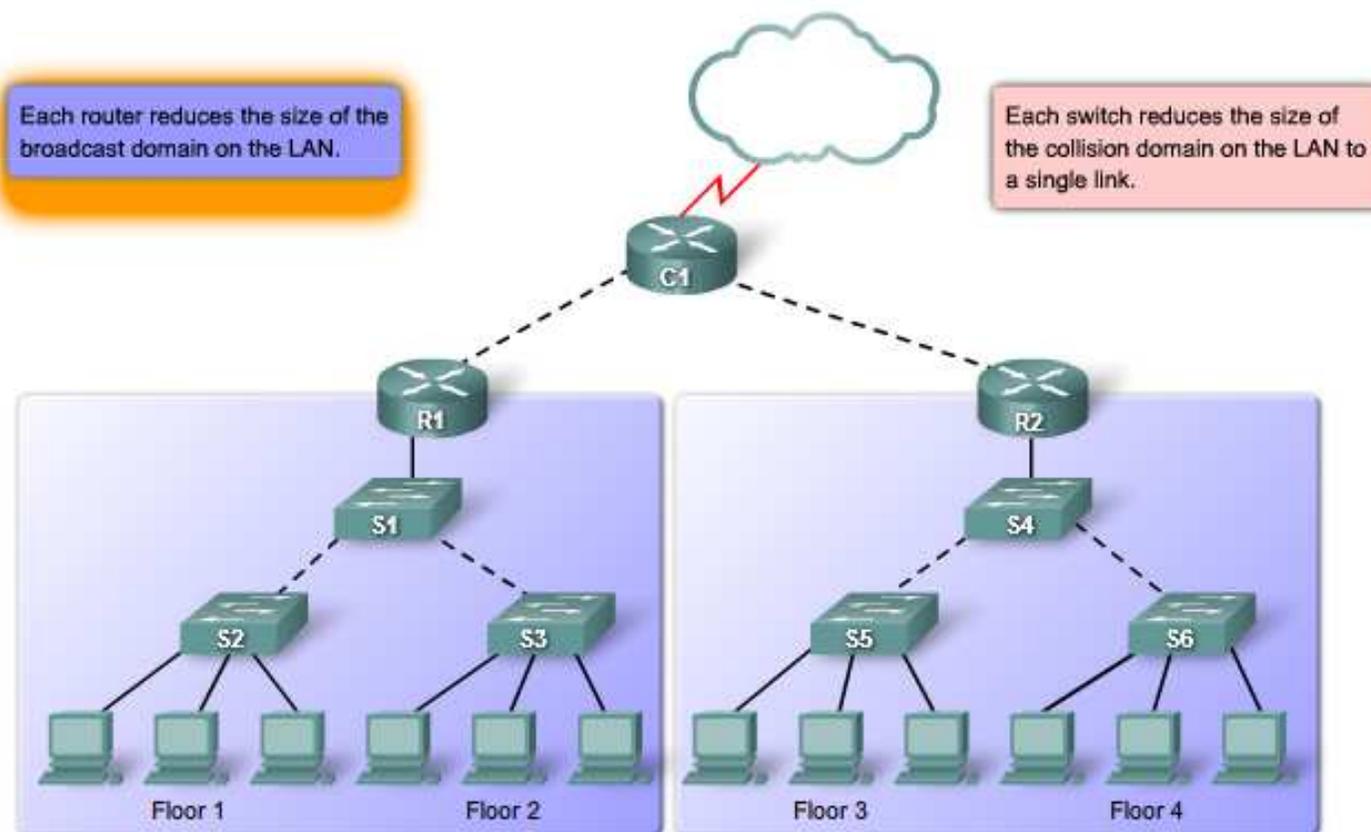
- ▶ In store-and-forward switching, when the switch receives the frame, it stores the data in buffers until the complete frame has been received
- ▶ In a switched network, collision domains shrink to single links
- ▶ If the links between switches and hosts are full-duplex, no collisions may occur
- ▶ During the storage process, the switch also performs an error check using the Cyclic Redundancy Check (CRC) trailer portion of the frame



Switches and broadcast domains



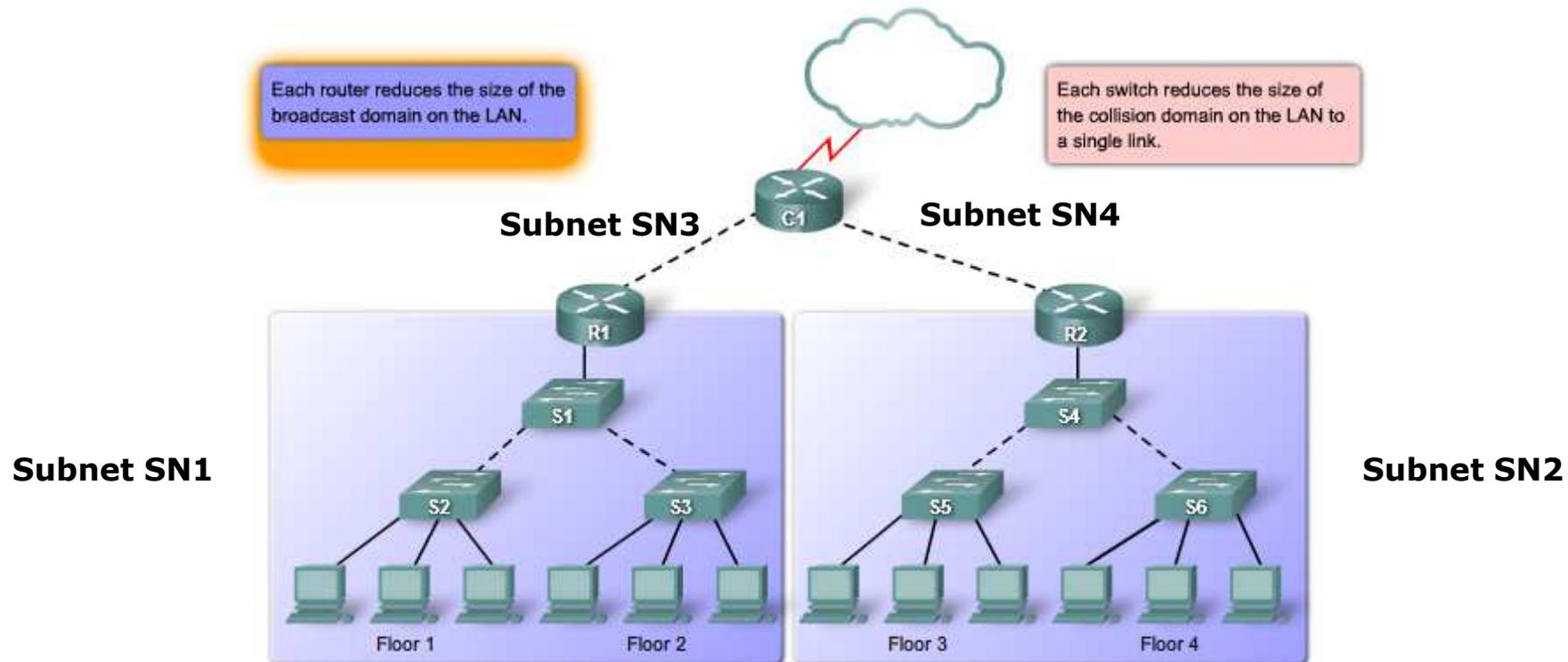
- ▶ Although switches filter most frames based on MAC addresses, they do not filter broadcast frames
- ▶ A collection of interconnected switches forms a single broadcast domain



Routers and IP subnets

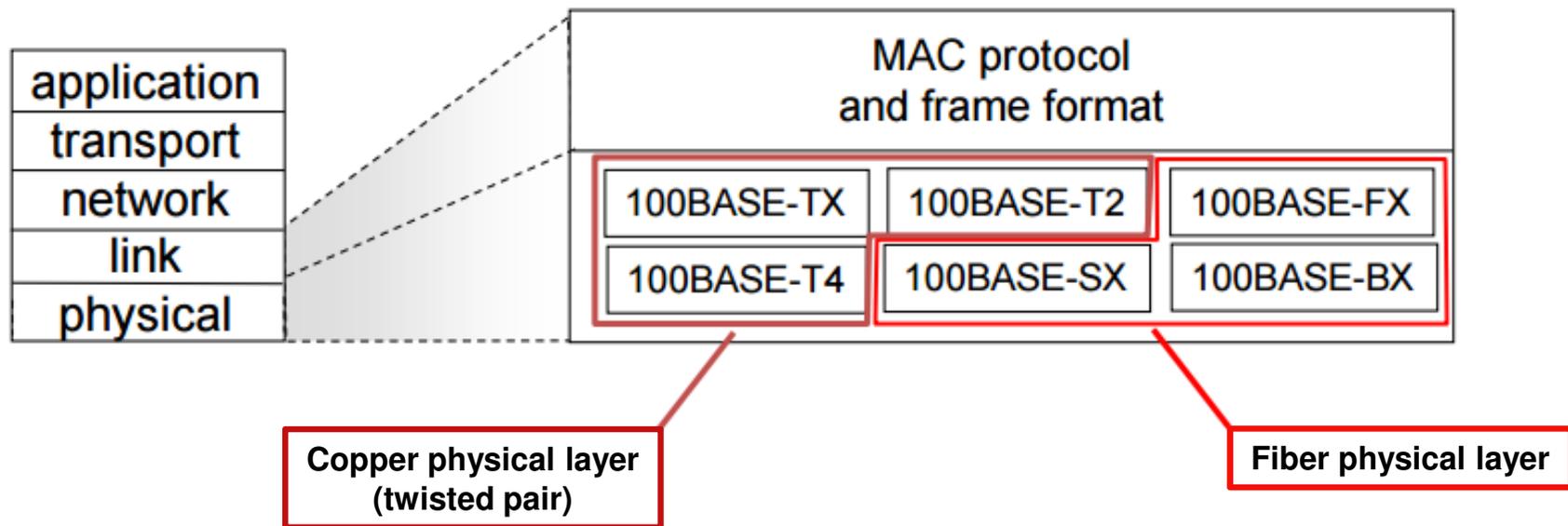


- ▶ To partition a large network in multiple isolated broadcast domains, routers are needed
 - ▶ Routers split a network in multiple IP subnets
 - ▶ A broadcast transmission does not cross the IP subnet boundary
 - ▶ Approach possible only if IP subnets are physically separated as in the picture below
 - ▶ Subnet SN1 on the left, SN2 on the right

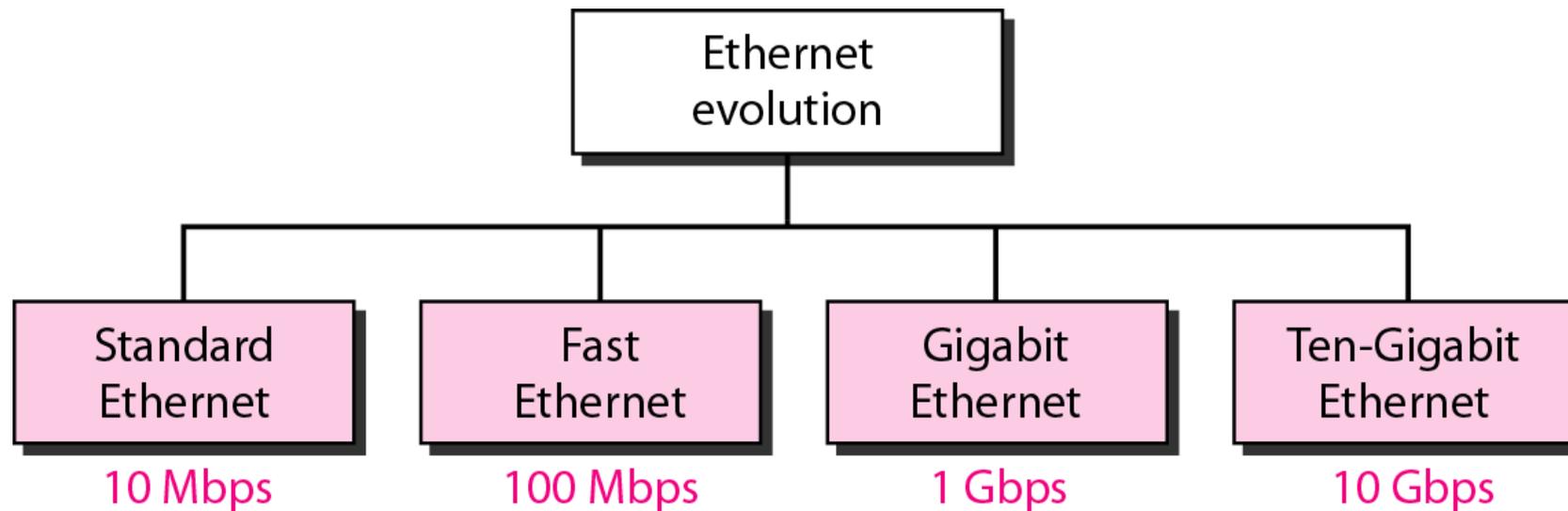




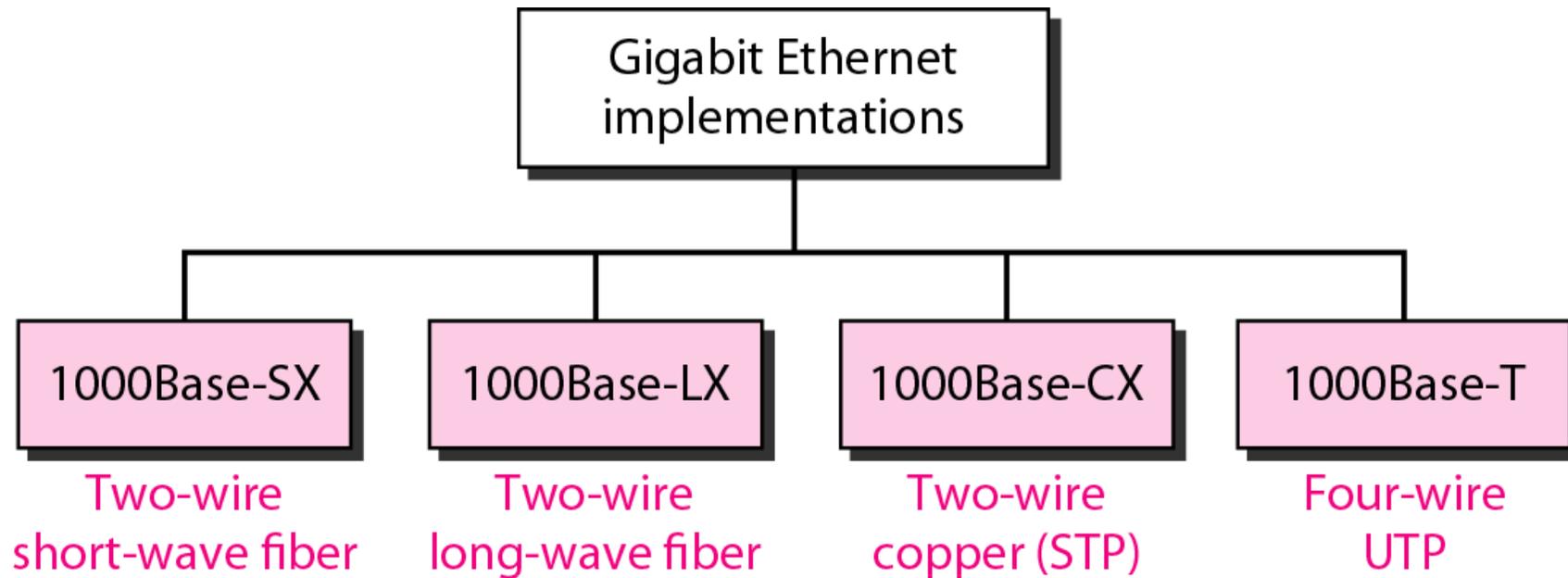
- ▶ IEEE 802.3 is actually a collection of many different standards
 - ▶ common MAC protocol and frame format
 - ▶ different speeds: 2 Mbps, 10 Mbps, 100 Mbps, 1Gbps, 10Gbps, 40Gbps, 100Gbps
 - ▶ different physical layer media: fiber, cable



Ethernet technology evolution



Gigabit Ethernet implementations and media



<i>Characteristics</i>	<i>1000Base-SX</i>	<i>1000Base-LX</i>	<i>1000Base-CX</i>	<i>1000Base-T</i>
Media	Fiber short-wave	Fiber long-wave	STP	Cat 5 UTP
Number of wires	2	2	2	4
Maximum length	550 m	5000 m	25 m	100 m
Block encoding	8B/10B	8B/10B	8B/10B	
Line encoding	NRZ	NRZ	NRZ	4D-PAM5

10-Gigabit Ethernet implementations and media



- ▶ IEEE 802.3ae standard for fiber optic cables
- ▶ IEEE 802.3ak for twinaxial copper cables
- ▶ IEEE 802.3an for UTP cat 6A and cat7 cables

- ▶ How does 10GbE compare to other varieties of Ethernet?
 - ▶ Frame format is the same, allowing interoperability between all varieties of legacy, fast, gigabit, and 10 Gigabit, with no reframing or protocol conversions
 - ▶ Bit time is now 0.1 ns - all other time variables scale accordingly
 - ▶ Only full-duplex fiber connections are used, CSMA/CD is not necessary

10 Gigabit Ethernet over Fiber: IEEE802.3ae



- ▶ Ratified in June 2002, the IEEE802.3ae LAN standard was developed to update the preexisting IEEE802.3 standard for 10GbE fiber transmission
- ▶ With the new standard, new media types were defined for LAN, metropolitan area network (MAN) and wide area network (WAN) connectivity
 - ▶ **10GBASE-SR (*Short Reach*)** – uses the lowest cost optics (850nm) to support 10GbE transmission over standard multimode fiber for distances up to 300 meters
10GBASE-SR is often the standard of choice to use inside the datacenters where fiber is already deployed and widely used
 - ▶ **10GBASE-LR (*Long Reach*)** – uses higher cost optics (1310nm) and requires more complex alignment of the optics to support single-mode fiber up to 10 km
 - ▶ **10GBASE-LRM (*Long Reach Multimode*)** – operating at 1310 nm, can span up to 220 meters with a multimode fiber using a technology called EDC (*Electronic Dispersion Compensation*)
10GBase-LRM is targeted for those customers who have older fiber already in place but need extra reach for their network
 - ▶ **10GBASE-ER (*Extended reach*)** – uses the most expensive optics (1550nm) and single-mode fiber for a link length up to 40 km
 - ▶ **10GBASE-SW, 10GBASE-LW, 10GBASE-EW** – defined for use with a WAN PHY, these standards were defined to operate at the same baud rate as OC-192/STM-64 SONET/SDH equipment
 - ▶ **10GBASE-LX4** – supports traditional FDDI grade multimode fiber for distances up to 300 meters using Coarse Wavelength Division Multiplexing (CWDM), which lowers the transmission rate of each wavelength to 3.125Gbaud; the LX4 standard also supports single-mode fiber for up to 10 Km

10 Gigabit Ethernet over Copper: IEEE 802.3ak & 802.3an



- ▶ IEEE802.3ak is a low-cost 10GbE solution intended for copper cabling with short distance connectivity that makes it ideal for wiring closet and data center connectivity
 - ▶ Approved in 2004
 - ▶ Also known as 10GBASE-CX4
 - ▶ The CX4 standard transmits 10GbE over four channels using twin-axial cables derived from Infiniband connectors and cable
- ▶ IEEE802.3an is the latest proposed 10GbE standard for use with unshielded twisted-pair (UTP) style cabling
 - ▶ Approved in 2006
 - ▶ Also known as 10GBASE-T
 - ▶ At least Category 6A (Cat 6A) or Category 7 (Cat 7) UTP cables are required

Transceivers



- ▶ Transceivers are hot-swappable devices used to connect a variety of physical media to Ethernet switches and NICs
- ▶ Transceivers are also referred to as *Medium Attachment Units* (MAUs)
- ▶ Gigabit Ethernet has two types of transceivers:
 - ▶ Gigabit Interface Connector (GBIC)
 - ▶ Small Form Factor Pluggable (SFP) or “mini-GBIC”
- ▶ 10Gb Ethernet (10 GbE) has several defined transceiver types:
 - ▶ XENPAK – mainly used in LAN switches; the first 10GbE pluggable transceivers on the market to support the 802.3ae standard transmission optics; these transceivers also support the 802.3ak copper standard to connect CX4 cables
 - ▶ XPAK – used primarily in Network Interface Cards (NIC) and Host Bus Adapter (HBA)
 - ▶ X2 – smaller form factor (about 2/3 the size of the XENPAK)
 - ▶ XFP – the closest in size to SFP
 - ▶ SFP+ - an enhanced version of SFP that supports data rates up to 16 Gbit/s and can be used for both 8 Gbit/s Fibre Channel and 10Gb Ethernet for both copper and optical cables
- ▶ 40Gb Ethernet (40 GbE) uses the following transceiver types:
 - ▶ QSFP/QSFP+ – allows data rates of 4x10 Gbit/s for Ethernet, Fibre Channel, InfiniBand and SONET/SDH links providing four channels of data in one pluggable interface



A switch with 4 SFP ports



XFP transceiver

Direct-Attach Active Optical Cable with SFP+ Connectors



QSFP to 4 SFP+ Breakout Cable

Twinaxial cabling or "Twinax"



- ▶ A type of cable similar to coaxial cable, but with two inner conductors instead of one
- ▶ Applied in SFP+ Direct-Attach Copper (10GSFP+Cu), a popular choice for 10G Ethernet
- ▶ On SFP+ it is possible to transmit at 10 Gigabits/second full duplex over 5 m distances
- ▶ Twinax with SFP+ offers 15 to 25 times lower transceiver latency than current 10GBASE-T Cat 6/Cat 6a/Cat 7 cabling systems: 0.1 μ s versus 1.5 to 2.5 μ s
- ▶ The power draw of Twinax with SFP+ is around 0.1 watts, which is also much less than 4–8 watts for 10GBASE-T
- ▶ 40GBASE-CR4 and 100GBASE-CR10 physical layers using 7 m twin-axial cable are being developed as part of 100 Gbit Ethernet specifications by IEEE 802.3bj workgroup



**Direct Attach Twinax Copper (DAC)
with SFP+ Connectors**

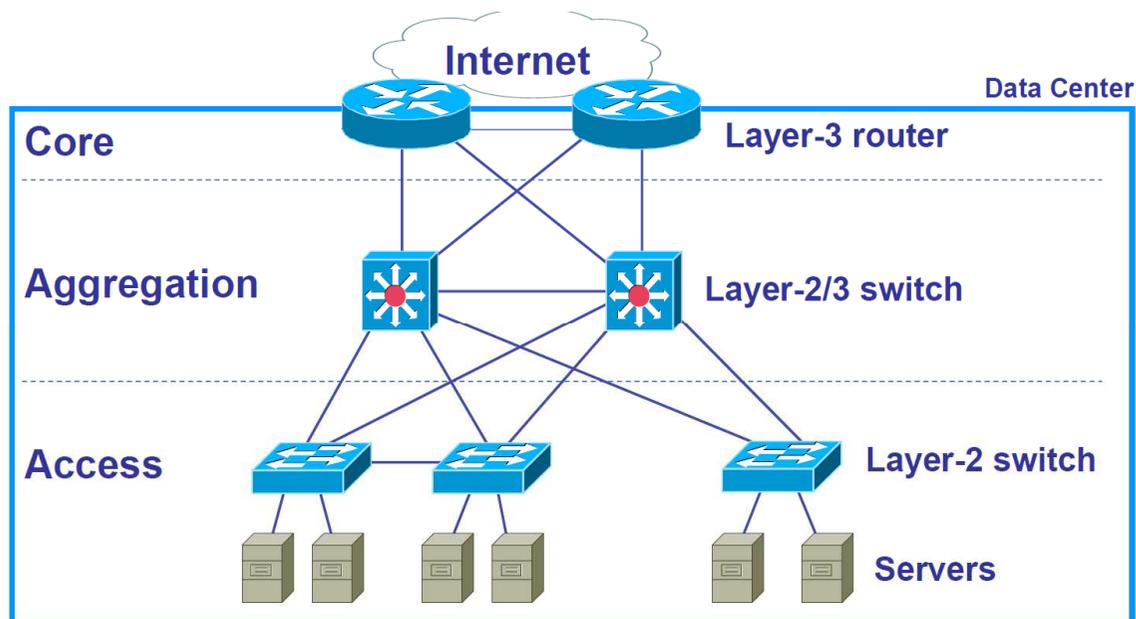


- ▶ *A datacenter's networking infrastructure allows servers to communicate among them, access shared storage resources as well as to communicate to the rest of the world through the Internet*
- ▶ **As for any networking infrastructure, it comprises:**
 - ▶ Active equipments
 - ▶ A cabling infrastructure (cables, pathways, patch panels, etc.)
- ▶ **Active equipments include:**
 - ▶ L2 switches
 - ▶ Routers and L3 switches
 - ▶ Firewalls and other special-purpose devices (load balancers, IDS, etc.)
- ▶ **Moreover, a DC usually comprises storage devices and a specialized networking infrastructure (SAN) used to connect servers with storage equipments**
 - ▶ Fibre Channel
- ▶ *High Performance Computing* HPC datacenters usually include a low-latency communications infrastructure to better support parallel computing
 - ▶ InfiniBand

DC networking architecture: 3-tier model



- ▶ In a DC, servers are physically organized in racks for a more efficient space utilization and for ease of management
- ▶ The datacenter networking infrastructure is designed according to a hierarchical architecture
- ▶ Servers' NICs (*network interface cards*) (2/4 NICs per server) are connected to a first layer infrastructure called *access layer*
- ▶ *Access layer's switches*, in turn, are connected to a second layer infrastructure, called *aggregation layer*
- ▶ The whole DC is connected to the Internet through a third layer infrastructure, called *core layer*, typically operating at layer 3 (*IP routing*)

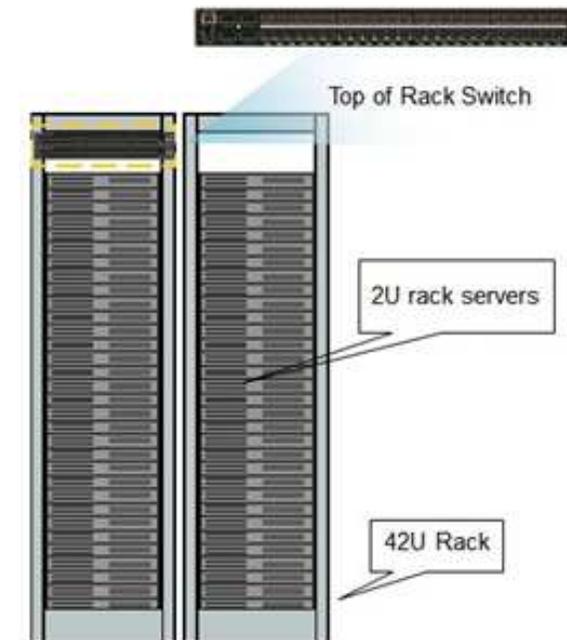
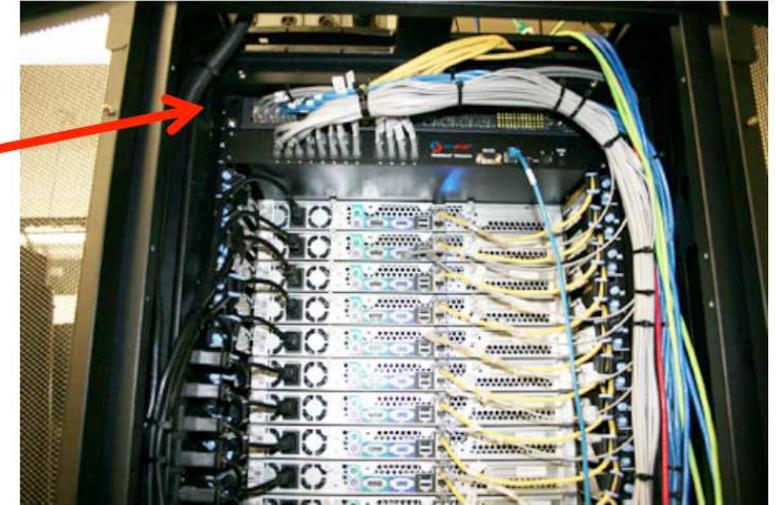


Access layer: Top of Rack switch



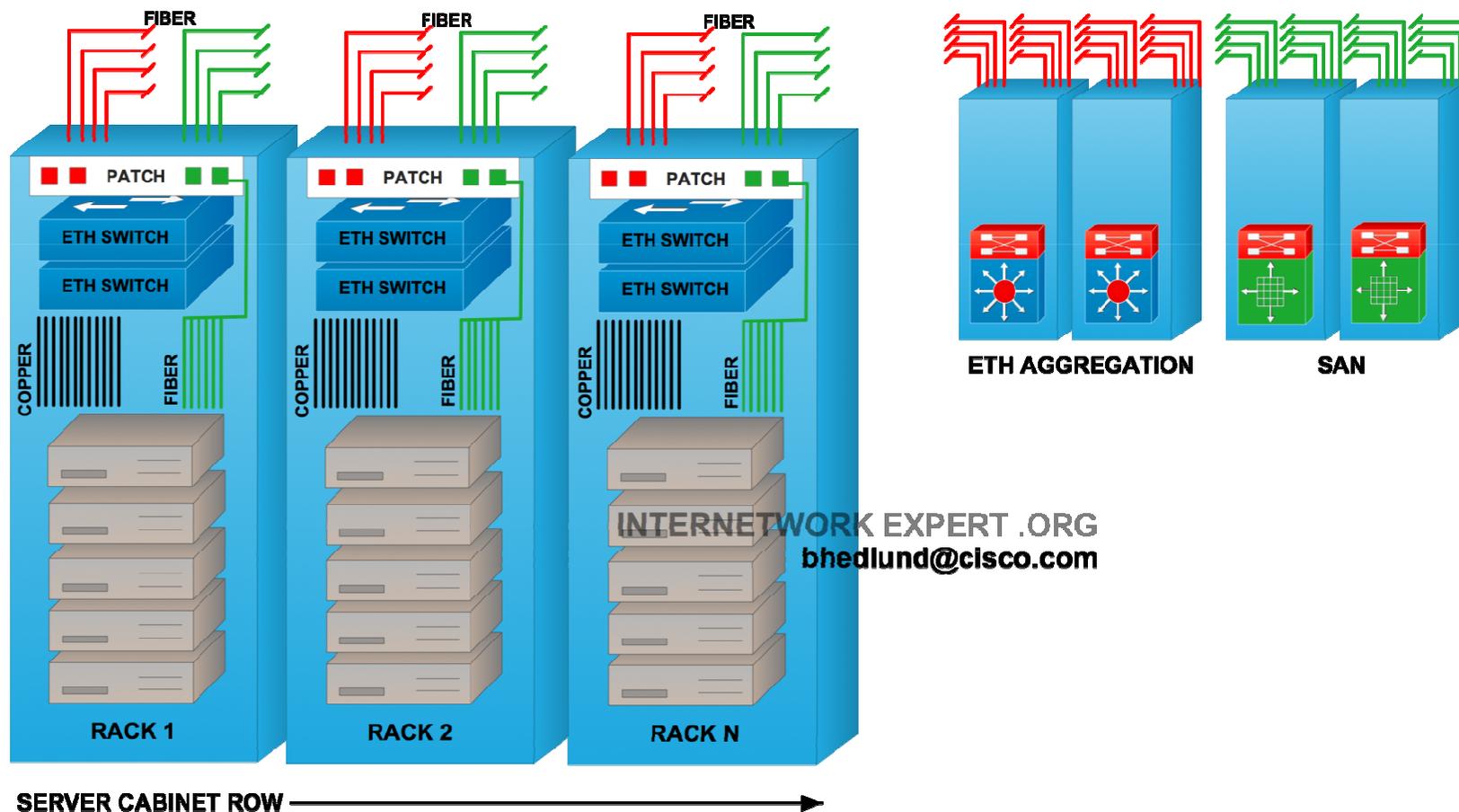
- ▶ The DC's *access layer* is typically formed by one or two Top-of-Rack (ToR) Ethernet switches located in each rack
 - ▶ N server / rack
 - ▶ $N \approx 20$ (≈ 40) for 2U servers (or 1U) and 42U racks
 - ▶ 2 NIC Ethernet / server
 - ▶ 1 or 2 48-ports switch + uplink in each rack
 - ▶ Until 2015:
 - ▶ 1 GbE server connectivity, 10 GbE uplink
 - ▶ Recent tren:
 - ▶ 10 GbE server connectivity, 40 GbE uplink
- ▶ Remote management NICs are also to be taken into account
 - ▶ HP iLO, Dell DRAC, etc.
 - ▶ 1 remote management NIC per server
 - ▶ 1 switch dedicated to management connections

Top-of-Rack switch



Top of Rack (or In-Rack) design

- ▶ Switches do not need to be necessarily in top part of the rack
 - ▶ Sometimes a convenient location is in the middle of the rack
- ▶ Copper (e.g. UTP) cabling for in-rack connections
- ▶ Fibers to connect racks to aggregation layer switches and to SAN



Brad Hedlund. *Top of Rack vs End of Row Data Center Designs*. <http://bradhedlund.com>

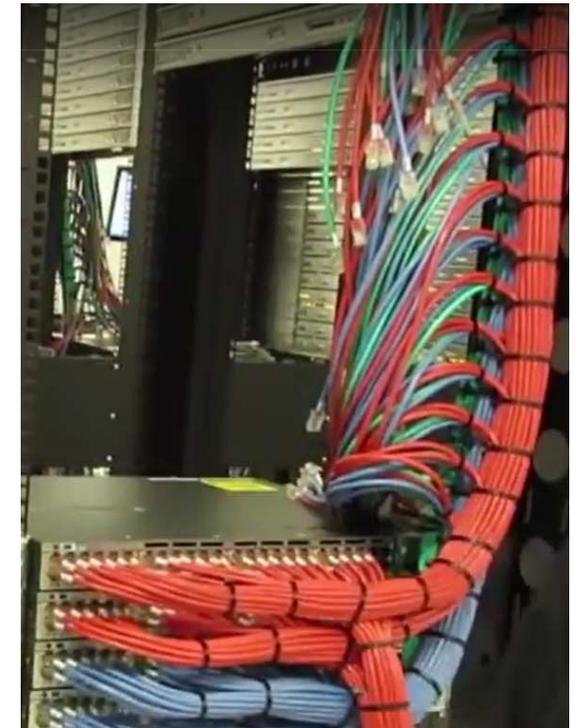
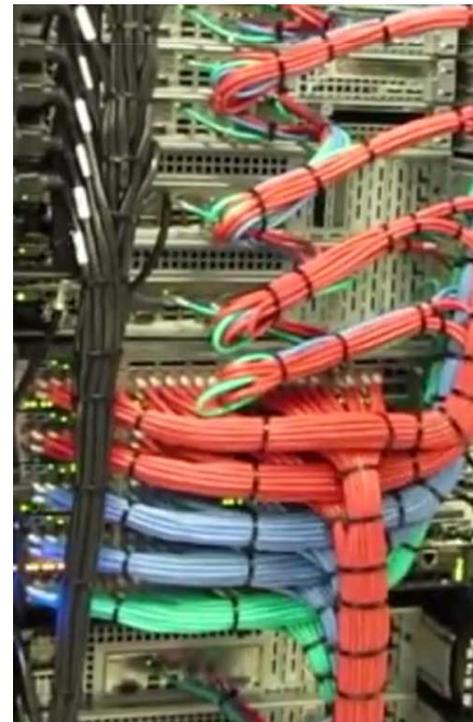
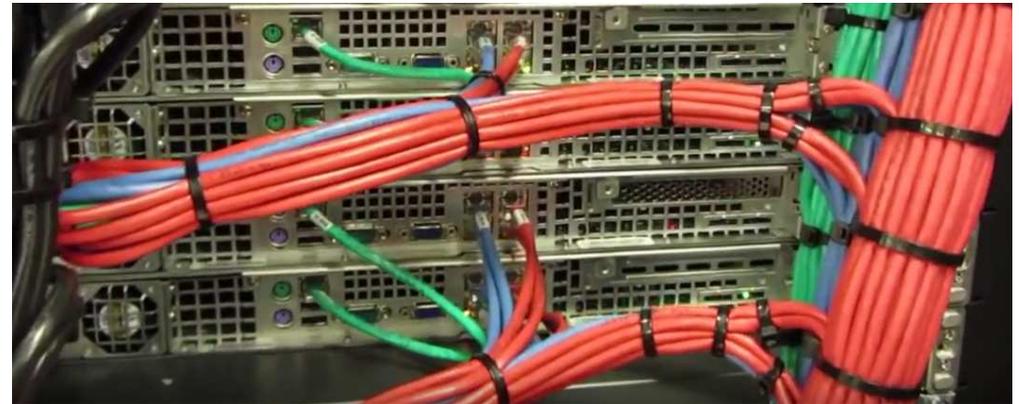
In-Rack switches: cabling



A few pictures showing racks with in-rack switches and servers connections
(unstructured cabling)

Each server has several connections:

- Dual data connections
 - Dual connections to storage SAN
 - Remote management
-
- Switches are mounted in the middle of the rack
 - Cables are bundled and tied together

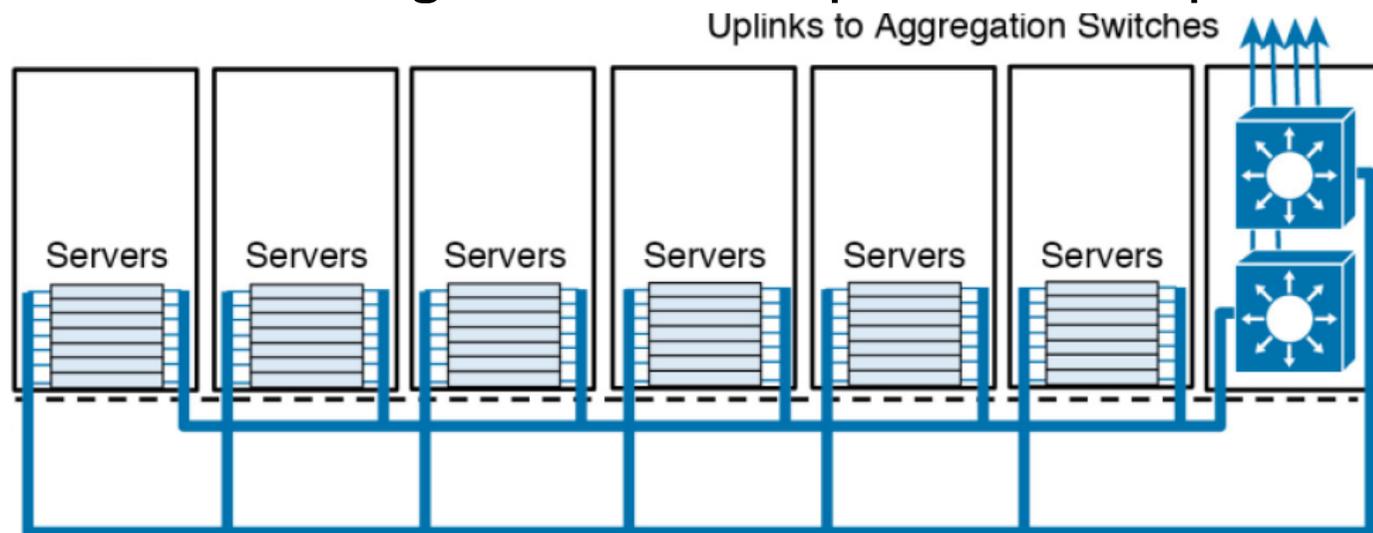


Source: [Softlayer Amsterdam DC video tour](#)

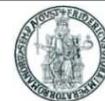
Access layer: End-of-Row or Middle-of-Row



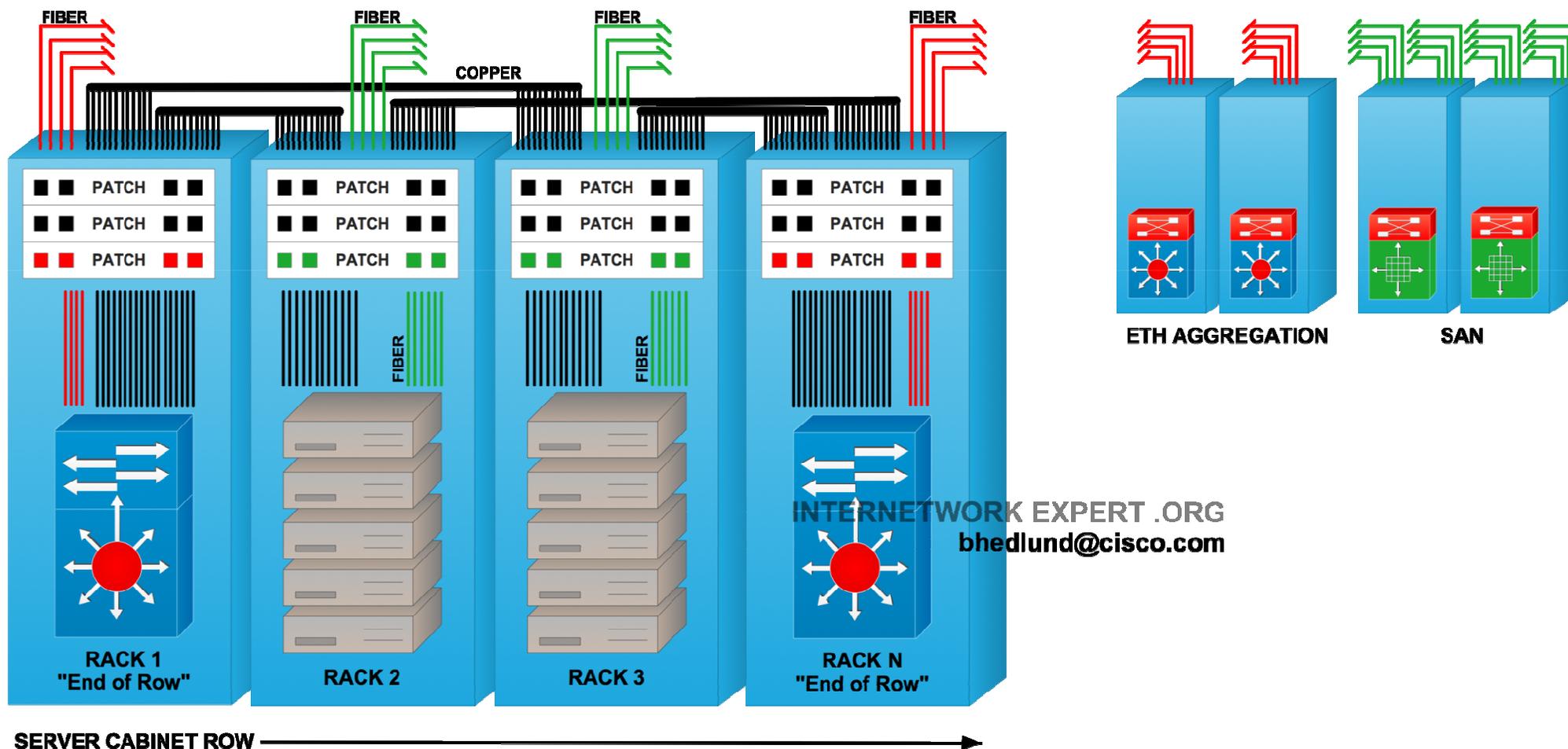
- ▶ The *access layer*, alternatively to a ToR arrangement, may be organized by sharing two bigger switches between all the servers of a row of racks
- ▶ These two shared access switches are usually mounted in a rack of their own, physically located either at one end of a row of racks (*End-of-Row, EoR*) or at the center of a row of racks (*Middle-of-Row, MoR*)
- ▶ Advantages:
 - ▶ Network devices located in a separate rack → easier management and maintenance
 - ▶ Power and control subsystems are shared → greater energy efficiency
- ▶ Disadvantages:
 - ▶ Longer links
 - ▶ Access switches with a greater number of port → more expensive



End-of-Row design



- ▶ When an End-of-Row design is used, structured cabling is preferred
- ▶ Both copper and fibers used for inter-rack cabling

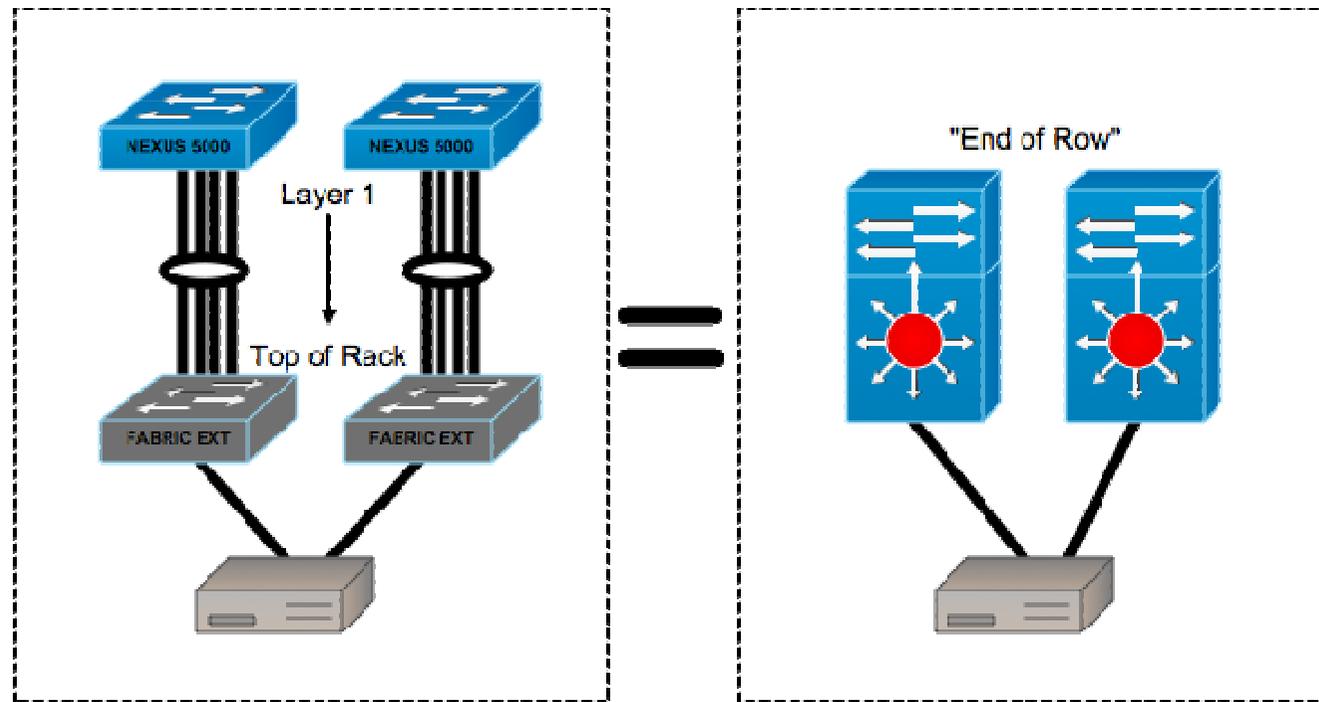


Brad Hedlund. *Top of Rack vs End of Row Data Center Designs*. <http://bradhedlund.com>

Access layer: mixed solutions with Fabric Extender (1)



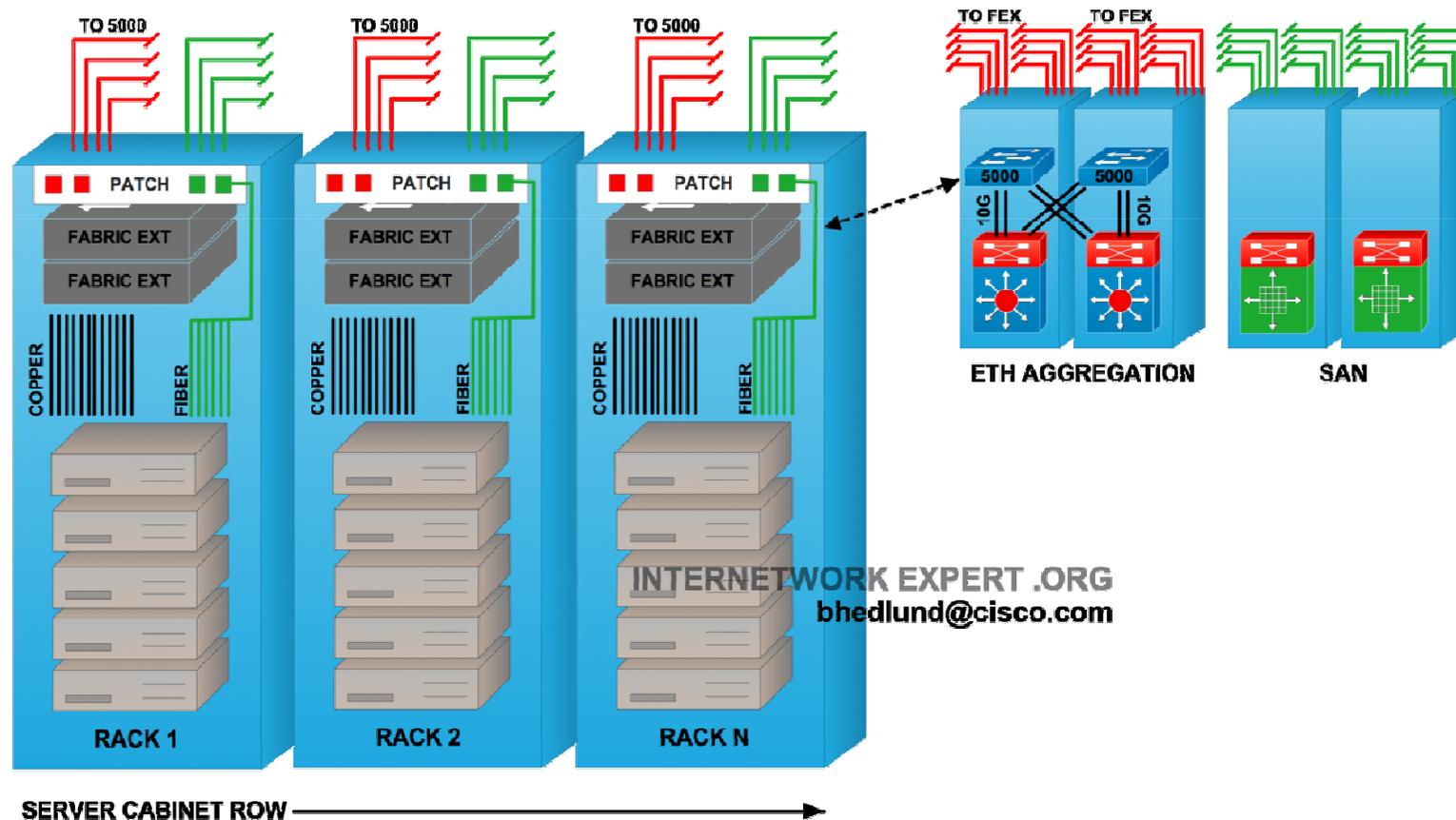
- ▶ Some network device vendors recommend solutions in which the access layer is built by combining in-rack switches (*Fabric Extenders*) with End-of-Row switches



Access layer: mixed solutions with Fabric Extender (2)



- ▶ In such arrangement, in-rack switches are managed as “extensions” (*line-cards*) of the EoR switch
 - ▶ Configuration only needed in EoR switches → faster and easier to manage



Brad Hedlund. *Top of Rack vs End of Row Data Center Designs*. <http://bradhedlund.com>

Access layer organization and cabling

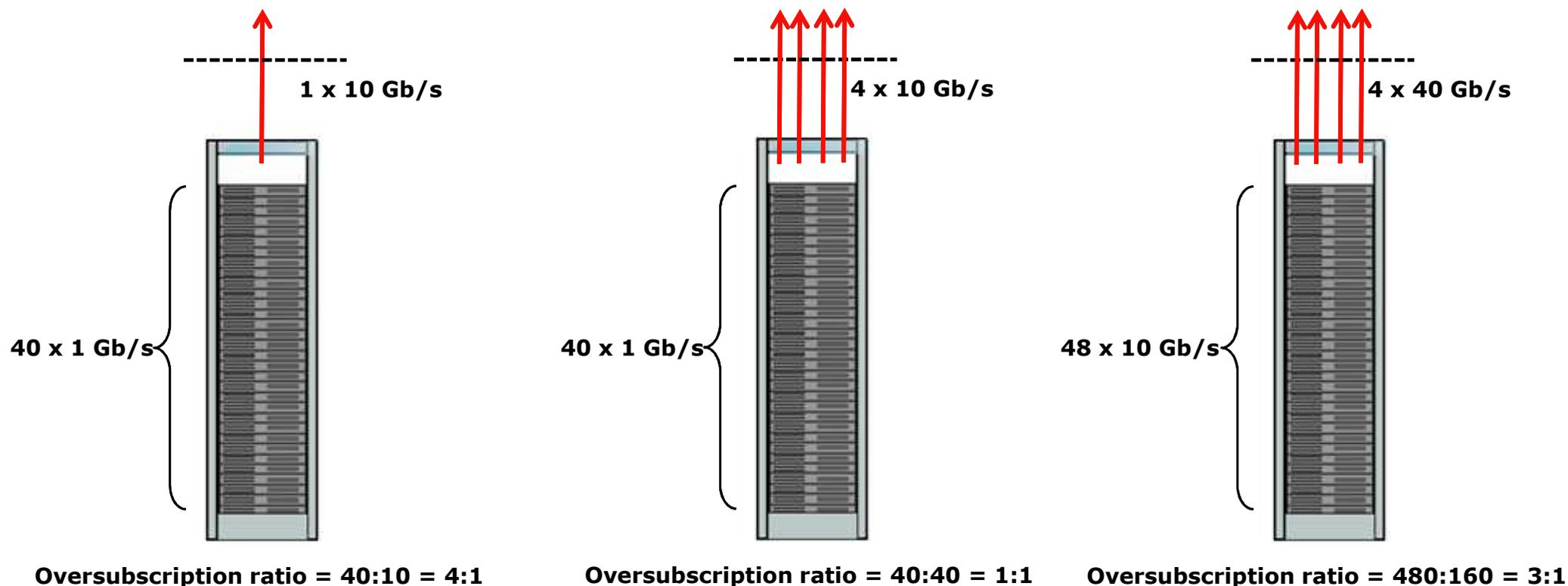


- ▶ The way access layer is organized has an impact on rack and DC cabling
- ▶ In a Top-of-Rack arrangement, servers are usually directly connected to access switches, without intermediate patch panels (*unstructured cabling*)
- ▶ In End-of-Row and Middle-of-Row arrangement, *structured cabling* solutions are typically preferred, with patch panels decoupling servers from inter-rack connections to access layer switches

Access-aggregation uplink: oversubscription



- ▶ Access layer switches connected to the rest of DC (*aggregation layer*) through a number of uplink connections (typically based on optical fibers)
- ▶ The ratio between the aggregated capacity of server links and the capacity of uplink links is usually referred to as *oversubscription ratio*
- ▶ Some examples for a ToR-based access layer:

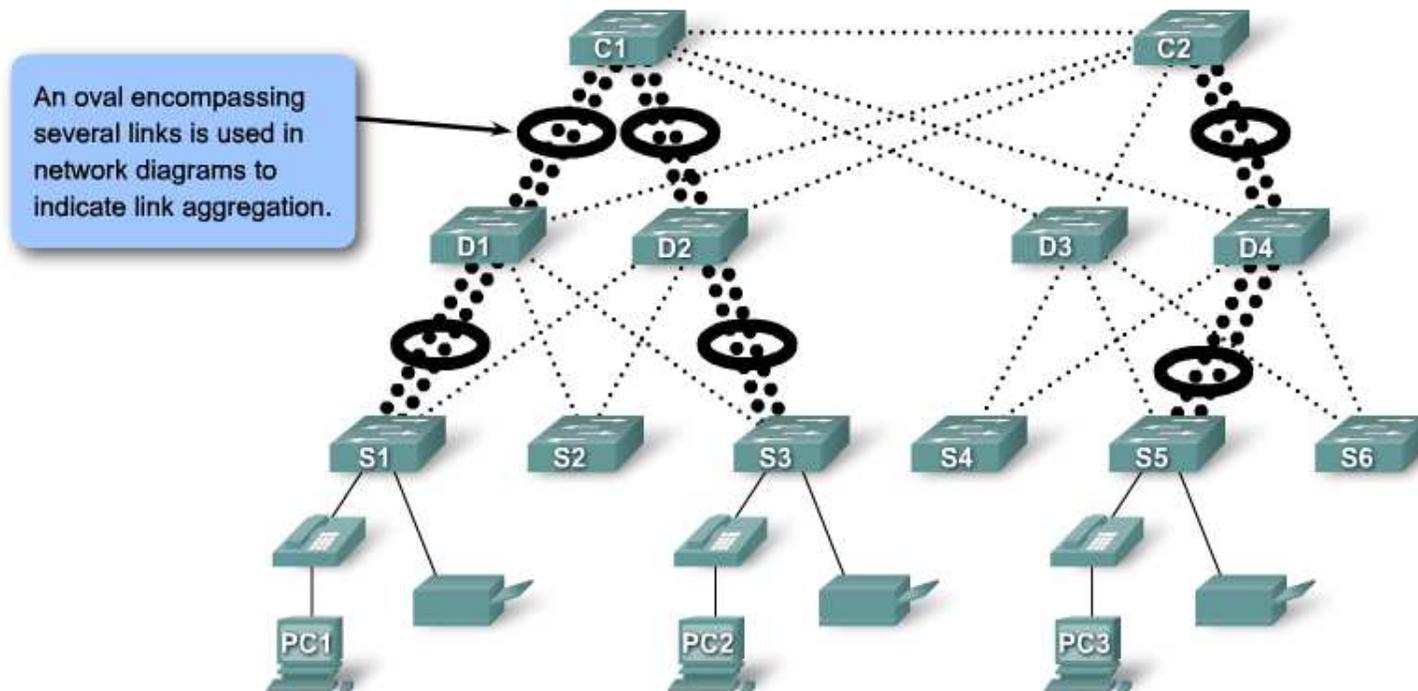


Bandwidth aggregation



- ▶ In order to reduce the oversubscription ratio, it is common to connect two switches with bunches of parallel links
- ▶ Beware: multiple parallel links form loops !
- ▶ Loop-avoidance techniques, such as STP, disable all links but one in a bundle
- ▶ To effectively use the aggregated bandwidth, special techniques are needed
 - ▶ E.g. Cisco's EtherChannel or the IEEE 802.3ad standard

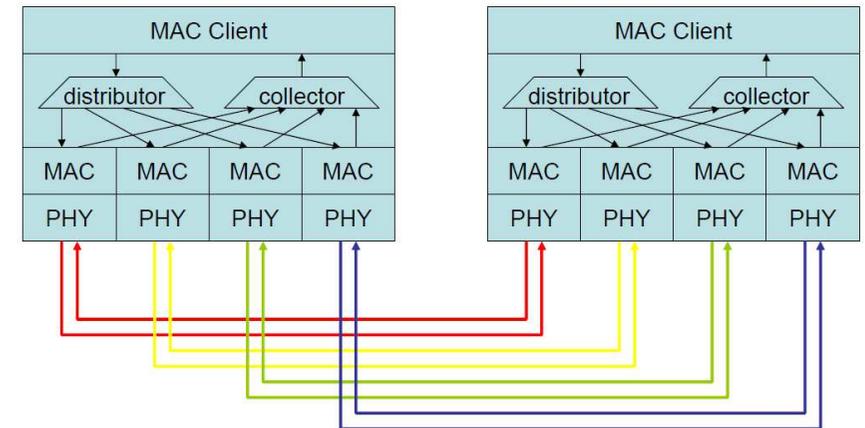
Bandwidth aggregation is normally implemented by combining several parallel links between two switches into one logical link.



IEEE 802.3ad Link Aggregation



- ▶ LAG is performed above the MAC
- ▶ LAG assumes all links are:
 - ▶ full duplex
 - ▶ point to point
 - ▶ same data rate
- ▶ Traffic is distributed packet by packet
- ▶ All packets associated with a given “conversation” are transmitted on the same link to prevent mis-ordering
- ▶ Does not change packet format
- ▶ Does not add significant latency
- ▶ Does not increase the bandwidth for a single conversation
- ▶ Achieves high utilization only when carrying multiple simultaneous conversations



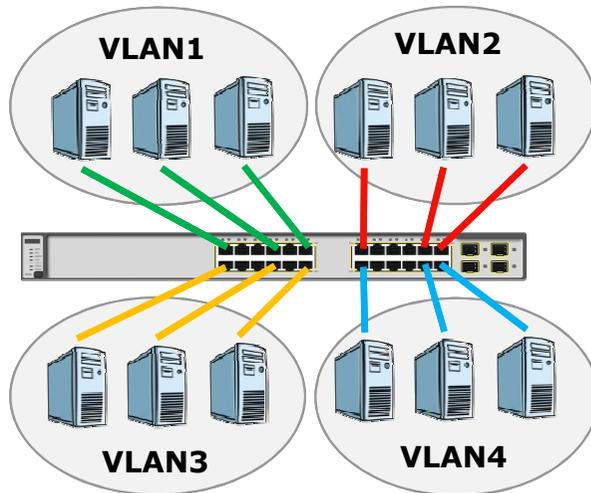


- ▶ LACP provides a method to control the bundling of several physical ports together to form a single logical channel
- ▶ LACP allows a network device to negotiate an automatic bundling of links by sending LACP packets to the peer (directly connected device that also implements LACP)
- ▶ Maximum number of bundled ports allowed in the port channel: 1 to 8
- ▶ LACP packets are sent with multicast group MAC address 01:80:c2:00:00:02
- ▶ During LACP detection period LACP packets are transmitted every second
- ▶ Keep alive mechanism for link member: (default: slow = 30s, fast=1s)
- ▶ Advantages deriving from LACP over static configuration
 - ▶ Failover occurs automatically
 - ▶ Dynamic configuration: the device can confirm that the configuration at the other end can handle link aggregation
- ▶ CISCO's switches support both LACP and the proprietary *Port Aggregation Protocol* (PAgP)

VLANs



- ▶ VLANs create separate broadcast domains within the same switch
 - ▶ Needed if multiple IP subnets need to coexist in the same switch
 - ▶ A router is needed to route traffic between VLANs
- ▶ In a single switch network, VLANs are typically assigned to ports by the admin

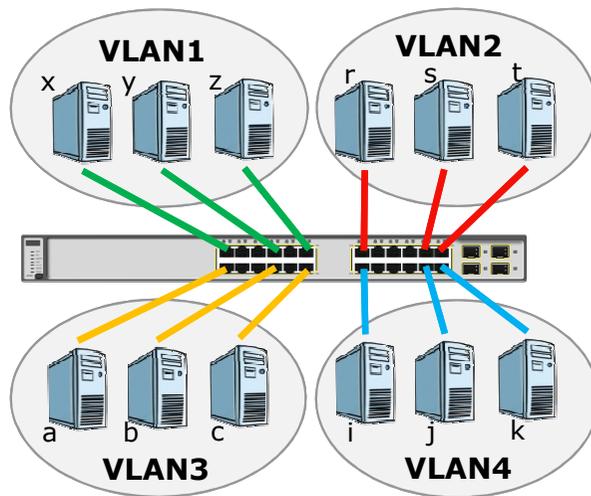


- ▶ Each switch port could be assigned to a different VLAN
- ▶ Ports assigned to the same VLAN share broadcasts
- ▶ Ports that do not belong to the same VLAN do not share broadcasts
- ▶ The default VLAN for every port in the switch is the “native VLAN”
 - ▶ The native VLAN is always VLAN 1 and may not be deleted
- ▶ All other ports on the switch may be reassigned to alternate VLANs

VLAN bridging tables



- ▶ Implementing VLANs on a switch causes the following to occur
 - ▶ The switch maintains a separate *bridging table* for each VLAN
 - ▶ If a frame comes in on a port in VLAN x, the switch searches the bridging table for VLAN x
 - ▶ When a frame is received, the switch adds the source address to the bridging table if it is currently unknown
 - ▶ The destination is checked so a forwarding decision can be made
 - ▶ For learning and forwarding the search is made against the address table for that VLAN only



VLAN1 bridging table

MAC address	port
x	1
y	7
z	11

VLAN2 bridging table

MAC address	port
r	13
s	21
t	23

VLAN3 bridging table

MAC address	port
a	2
b	8
c	12

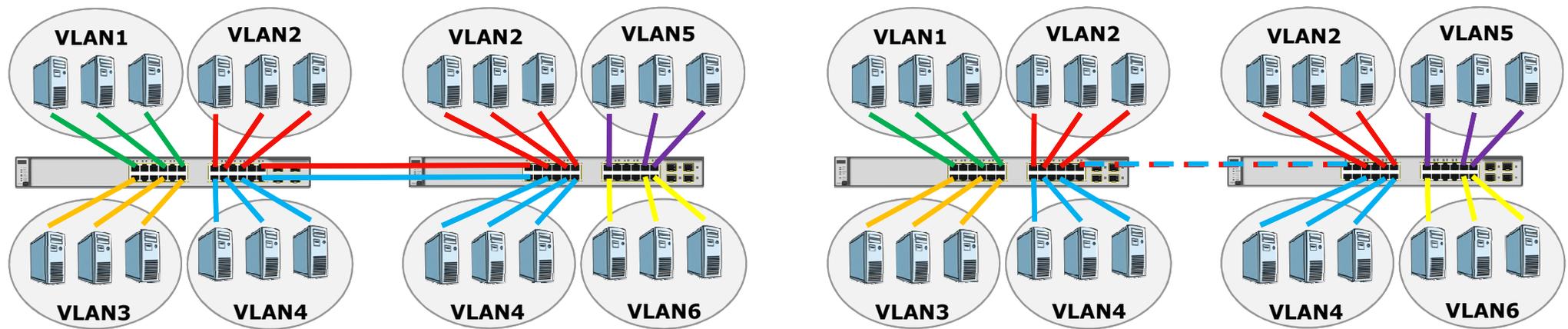
VLAN4 bridging table

MAC address	port
i	14
j	22
k	24

VLANs spanning multiple switches



- ▶ Problem: how to extend multiple VLANs over two distinct switches ?
- ▶ Solution #1
 - ▶ one link connecting the two switches for each VLAN that needs to be extended
 - ▶ costly and inefficient
- ▶ Solution #2 – *port trunking*
 - ▶ a single link (*trunk*) connects the two switches and carries traffic for all the VLANs that live in both switches
 - ▶ To associate each frame to the corresponding VLAN, a special tag is required in the frame header (*VLAN tagging*)
- ▶ In general, a *trunk* is a link carrying traffic for several VLANs and a switch may have several trunking ports



Two pairs of ports dedicated to extend VLANs,
one for VLAN2 and another for VLAN4

VLANs extended by means of port trunking

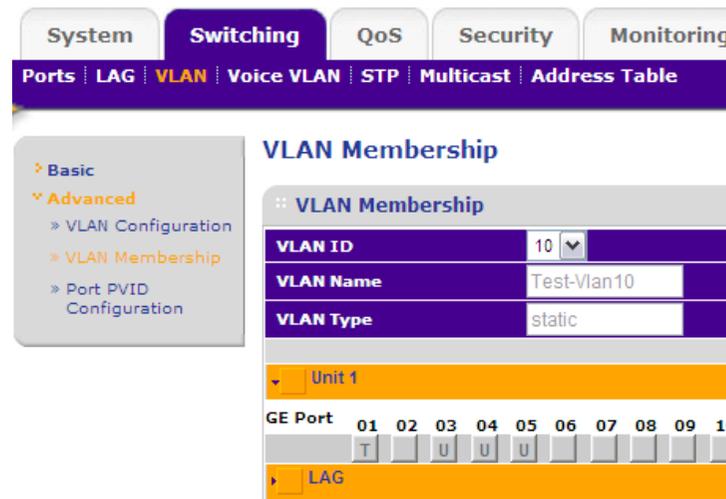
VLAN tagging



- ▶ VLAN Tagging is used when a link connecting two different switches needs to carry traffic for more than one VLAN
- ▶ A unique packet identifier is added within each header to designate the VLAN membership of each packet
- ▶ When a packet enters a trunk port with a given VLAN ID:
 - ▶ VLAN ID is removed from the packet
 - ▶ Packet is forwarded to the appropriate port based on the VLAN ID and destination MAC address
 - ▶ If the destination MAC address is FF:FF:FF:FF:FF:FF, the packet is forwarded to all the VLAN ports
- ▶ 2 major methods of VLAN tagging: Cisco proprietary Inter-Switch Link (ISL) and IEEE 802.1Q
- ▶ IEEE 802.1Q inserts VLAN ID (12 bits) in a new header field

Port 01 is configured as a trunk port for VLAN 10
(T stands for Tagged)

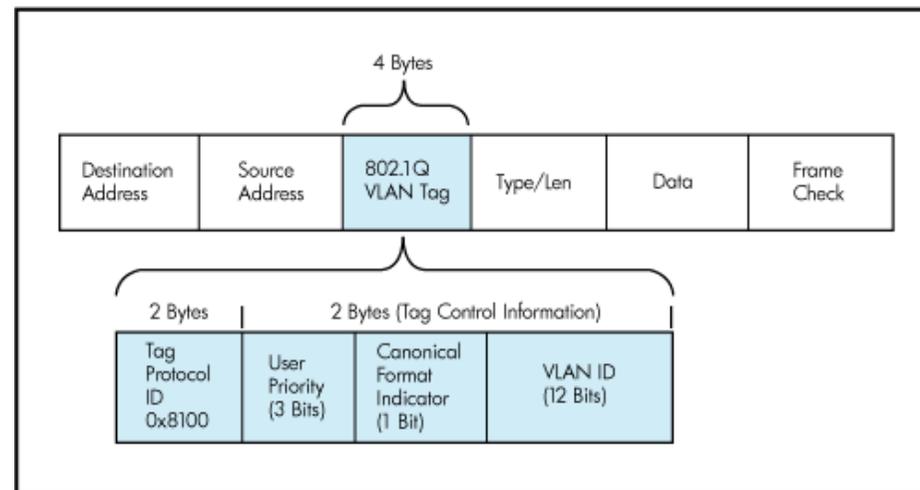
Ports 03, 04 and 05 are statically associated to VLAN 10
without any tagging (U stands for Untagged)



IEEE 802.1Q header



- ▶ IEEE 802.1Q adds a 4-byte header field:
- ▶ 2-byte tag protocol identifier (TPID) with a fixed value of 0x8100
- ▶ 2-byte tag control information (TCI) containing the following elements:
 - ▶ Three-bit user priority (8 priority levels, 0 thru 7)
 - ▶ One-bit canonical format (CFI indicator), 0 = canonical, 1 = noncanonical, to signal bit order in the encapsulated frame (see IETF RFC2469)
 - ▶ Twelve-bit VLAN identifier (VID) - Uniquely identifies the VLAN to which the frame belongs
 - ▶ defining 4,096 VLANs, with 0 and 4095 reserved values

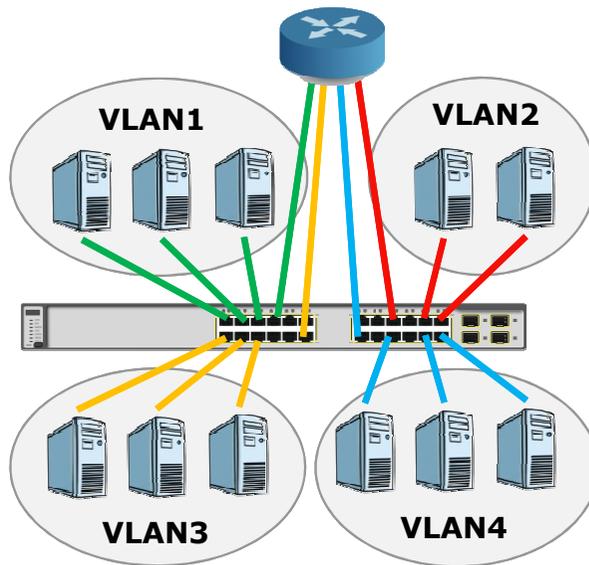


Inter-VLAN routing



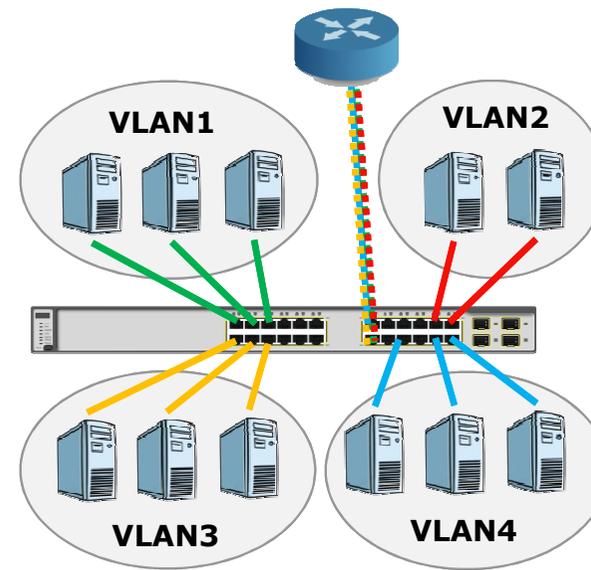
- ▶ When a node in one VLAN needs to communicate with a node in another VLAN, a router is necessary to route the traffic between VLANs
 - ▶ Without the routing device, inter-VLAN traffic would not be possible
- ▶ The routing function may be external or internal to the switch
 - ▶ In the latter case, the switch itself acts as a router (so called *multilayer switches* or L3 switches)
- ▶ External router
 - ▶ Approach #1: the router is connected to the switch by one link per VLAN
 - ▶ Approach #2: the router is connected to the switch by one trunk link for all the VLANs
 - ▶ Also known as “router on a stick”
 - ▶ Possible only if the router supports sub-interfaces to divide a single physical interface into multiple logical interfaces

Router connected by as many links as the VLANs to be connected



Router-on-a-stick

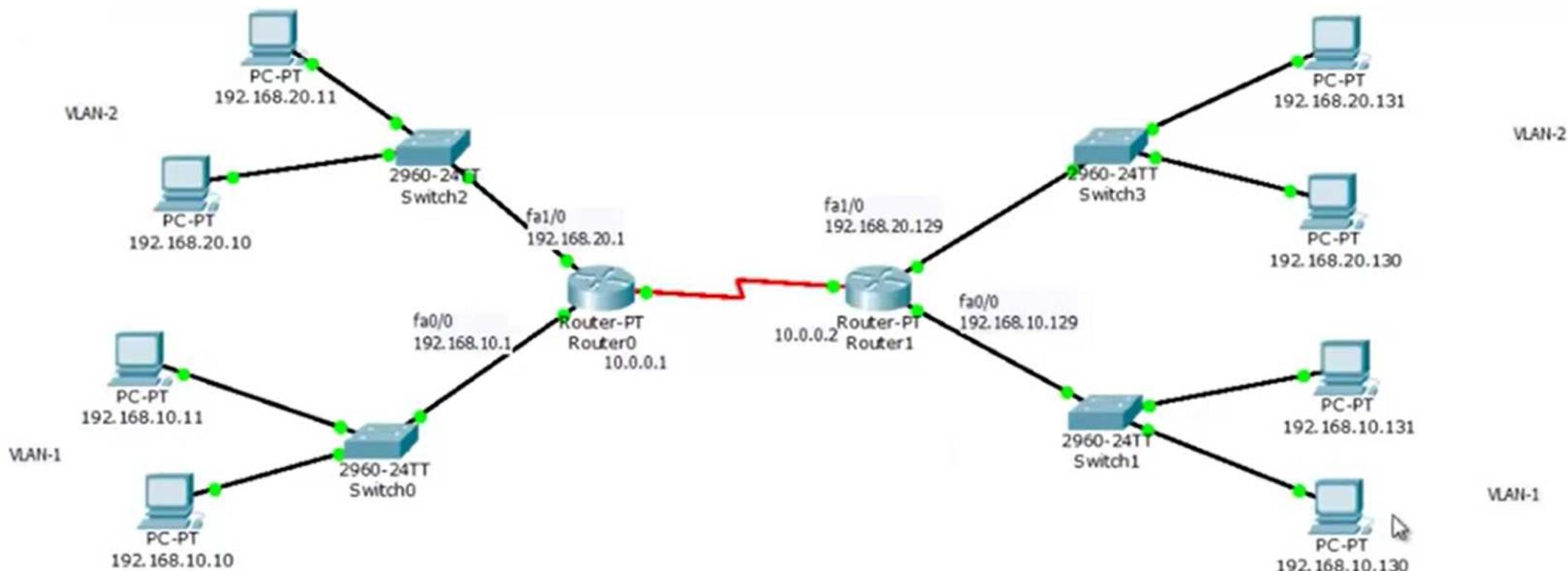
Router connected by one trunk link for all the VLANs



Inter-VLAN routing across different switches



- ▶ This scenario is an enterprise network, does not fit a datacenter
- ▶ Two VLANs, spread across two distinct switches connected by routers
- ▶ In fact, these are four VLANs, each associated to a /25 subnet
- ▶ Communication between host 192.168.20.10 (on the left) and 192.168.10.10 (on the left) is routed by Router0
- ▶ Communication between host 192.168.10.11 (on the left) and 192.168.10.130 (on the right) is routed by Router0 and Router1



Multilayer switches in a datacenter



- ▶ A multilayer switch is able to perform both kinds of packet forwarding: *bridging* at Layer 2 and *routing* at Layer 3
- ▶ Layer 3 routing in an aggregation switch can be used to route traffic among different VLANs without the need for an external router by means of so-called “Virtual Switch Interfaces” (SVIs)
 - ▶ An SVI should be configured with the VLAN’s default gateway IP address
- ▶ In a typical datacenter networks, aggregation layer switches are multilayer switches
- ▶ If one needs to exchange traffic among 2 servers (or 2 VMs) associated to 2 different VLANs, this machine-to-machine traffic would traverse the network hierarchy up to the aggregation switch even though the communicating hosts (or VMs) are physically located in the same rack

