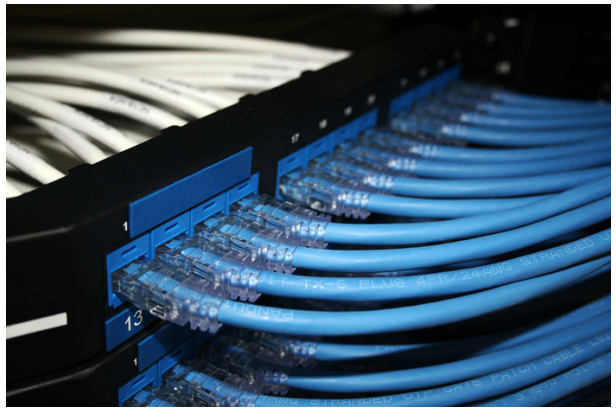
Data Center Ethernet



Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

These slides and audio/video recordings of this class lecture are at:

http://www.cse.wustl.edu/~jain/cse5700-25/

Student Questions

Announcements

- Videos of the first 3 lectures are on the course website
- Look out for an Email from WashU-IT about a web page space. Do not discard it.

©2025 Raj Jain



Overview

1. Residential vs. Data Center Ethernet
2. Review of Ethernet Addresses, devices, speeds, algorithms
3. Enhancements to Spanning Tree Protocol
4. Virtual LANs
5. Data Center Bridging Extensions

Student Questions

Quiz: True or False?

Which of the following statements is generally true?

T F

- ☐ ☐ Ethernet is a local area network (Local ≤ 2km)
- ☐ ☐ Token ring, Token Bus, and CSMA/CD (Carrier Sense Multiple Access with Collision Detection) are the three most common LAN access methods.
- ☐ ☐ Ethernet uses CSMA/CD.
- ☐ ☐ Ethernet bridges use spanning trees for packet forwarding.
- ☐ ☐ Ethernet frames are 1518 bytes.
- ☐ ☐ Ethernet does not provide any delay guarantees.
- ☐ ☐ Ethernet has no congestion control.
- ☐ ☐ Ethernet has strict priorities.

Student Questions

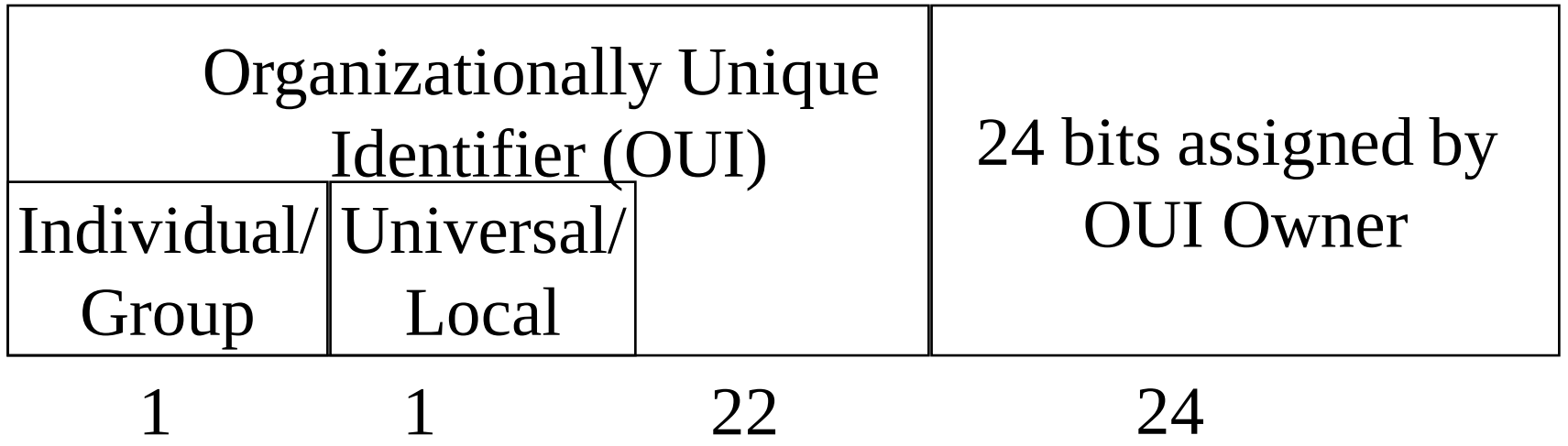
Residential vs. Data Center Ethernet

Residential	Data Center
❑ Distance: up to 200m	□ No limit
❑ Scale: ➢ Few MAC addresses ➢ 4096 VLANs	□ Millions of MAC Addresses □ Millions of VLANs Q-in-Q
❑ Protection: Spanning tree	□ Rapid spanning tree, … (Gives 1s, need 50ms)
❑ Path determined by spanning tree	□ Traffic-engineered path
❑ Simple service	□ Service Level Agreement. Rate Control.
❑ Priority ⇒ Aggregate QoS	□ Need per-flow/per-class QoS
❑ No performance/Error monitoring (OAM)	□ Need performance/BER

Student Questions

IEEE 802 Address Format

- 48-bit:1000 0000 : 0000 0001 : 0100 0011 : 0000 0000 : 1000 0000 : 0000 1100 = 80:01:43:00:80:0C

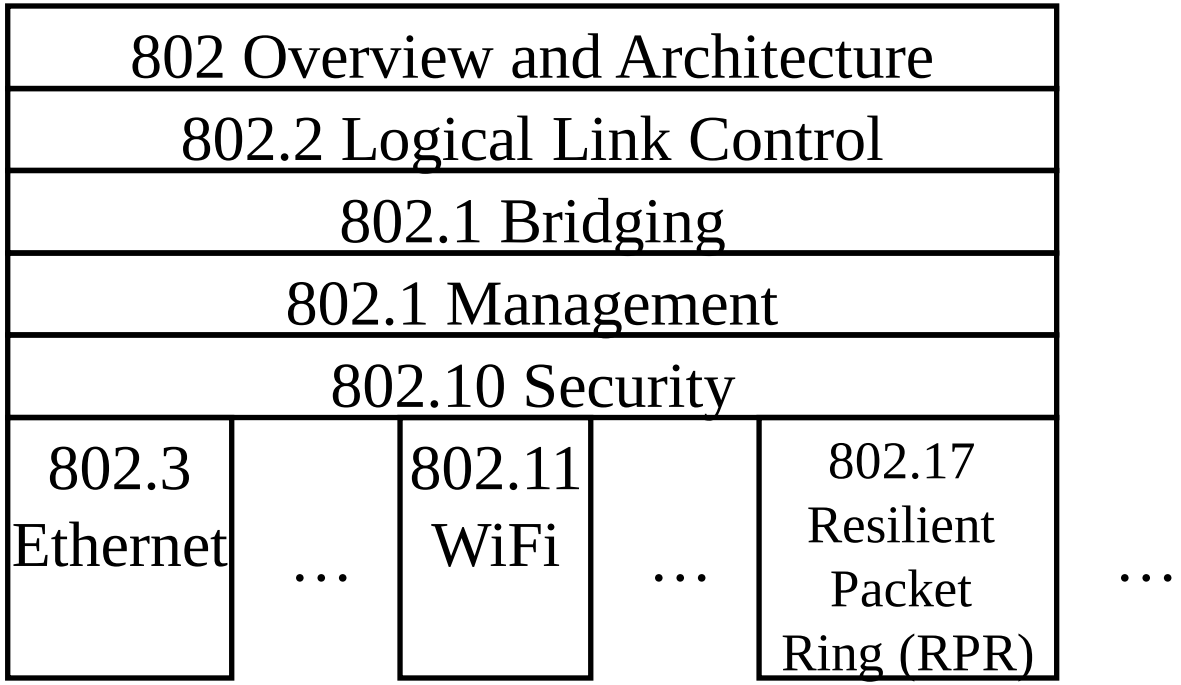


- Multicast = "To all bridges on this LAN"
- Broadcast = "To all stations" (Note: Local bit is set) = 111111....111 = FF : FF : FF : FF : FF : FF

Student Questions

IEEE Standards Numbering System

- IEEE 802.* and IEEE 802.1* standards (e.g., IEEE 802.1Q-2011) apply to all IEEE 802 technologies:
  - IEEE 802.3 Ethernet
  - IEEE 802.11 WiFi
  - IEEE 802.16 WiMAX



Student Questions

IEEE Standards Numbering (Cont)

- IEEE 802.3* standards apply only to Ethernet,
  e.g., IEEE802.3ba-2010
- Standards with all upper case letters are base standards
  E.g., IEEE 802.1AB-2009
- Standards with lowercase are additions/extensions/revisions.
  It is merged with the base standard in its next revision.
  e.g., IEEE 802.1w-2001 was merged with IEEE 802.1D-2004
- Standards used to be numbered sequentially, e.g., IEEE 802.1a, …, 802.1z, 802.1aa, 802.1ab, …
- Recently, they started showing base standards in the additions,
  e.g., IEEE 802.1Qau-2010

Student Questions

Names, IDs, Locators



Name: John Smith

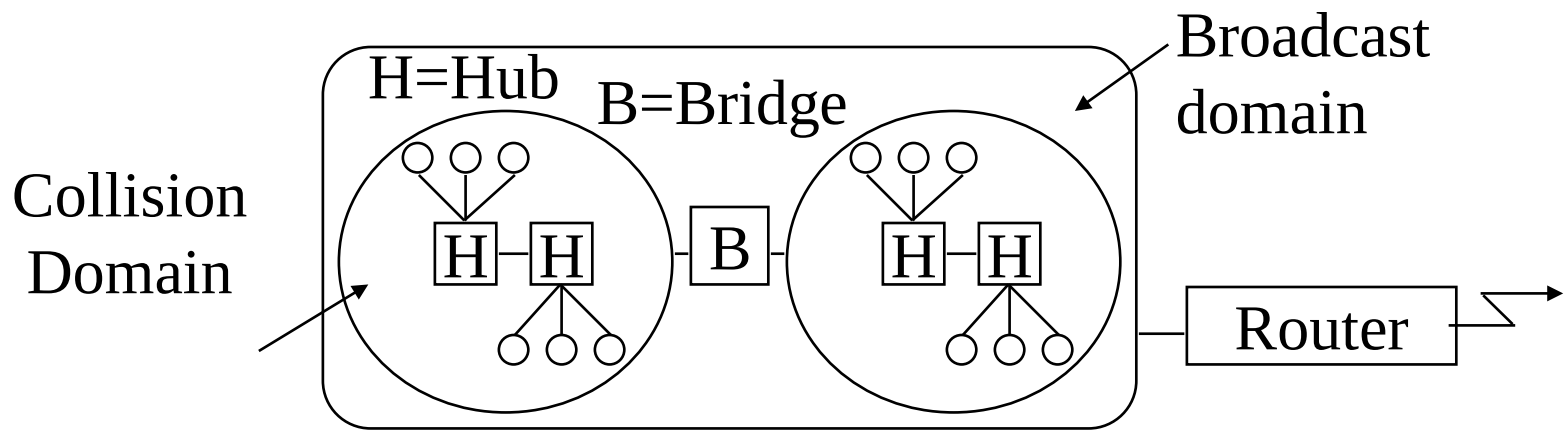
ID: 012-34-5678

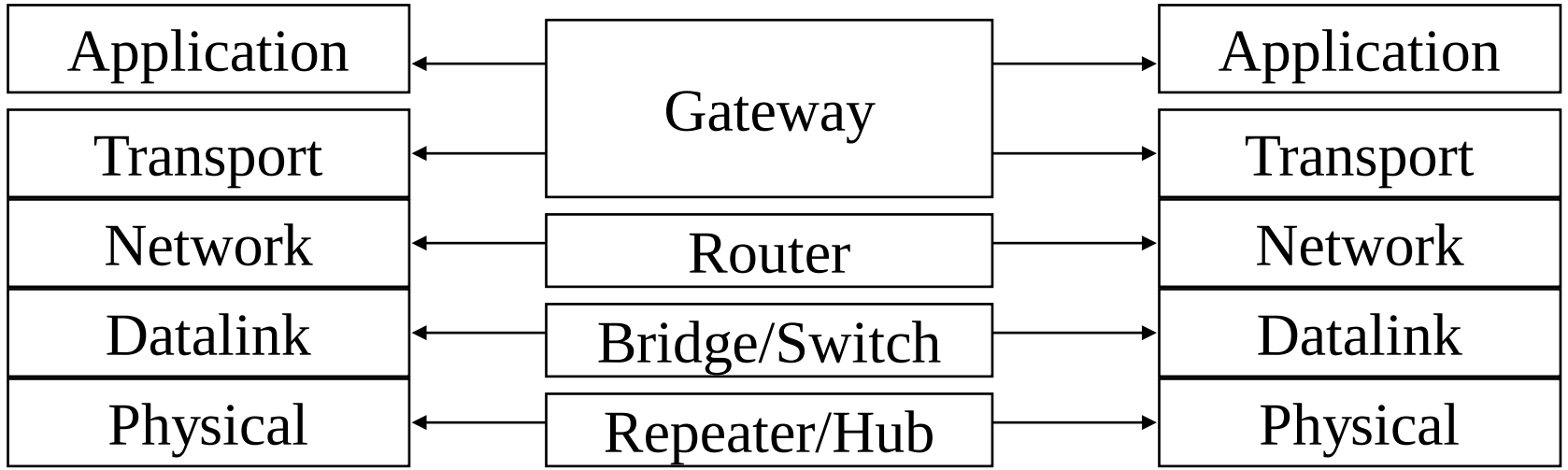
Locator:
1234 Main Street
Big City, MO 12345
USA

- The locator changes as you move, and ID and Names remain the same.
- **Examples**:
  - Names: Company names, DNS names (Microsoft.com)
  - IDs: Cell phone numbers, 800-numbers, Ethernet addresses, Skype ID, VOIP Phone number
  - Locators: Wired phone numbers, IP addresses

Student Questions

Interconnection Devices





Student Questions

Interconnection Devices (Cont)

- **Repeater**: PHY device that restores data and collision signals
- **Hub**: Multiport repeater + fault detection and recovery
- **Bridge**: Datalink layer device connecting two or more collision domains. MAC multicasts are propagated throughout the LAN.
- **Router**: Network layer device. IP, IPX, AppleTalk. Does not propagate MAC multicasts.
- **Switch**: Multiport bridge with parallel paths
- These are functions. Packaging varies.

Student Questions

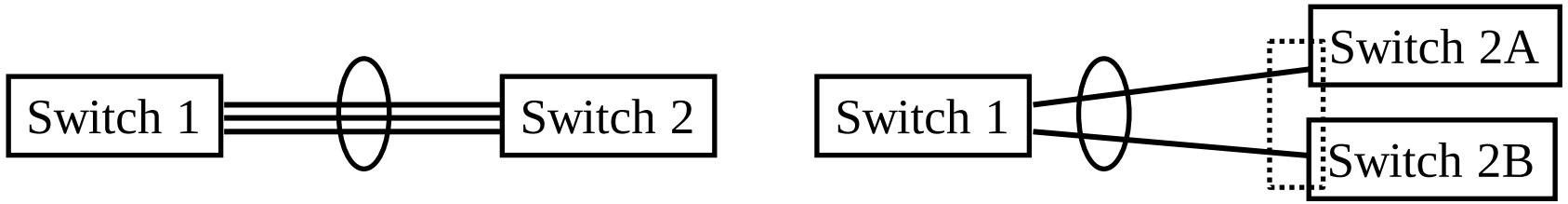
Ethernet Speeds

- IEEE 802.3ba-2010 (40G/100G) Standard
- 10Mbps, 100 Mbps, and 1 Gbps versions have both CSMA/CD and Full-duplex versions
- No CSMA/CD in 10G and up
- No CSMA/CD in practice now, even at home or at 10 Mbps
- 1 Gbps in residential, enterprise offices
- 1 Gbps in Data centers, moving to 10 Gbps and 40 Gbps
- 100G in some carrier core networks
  100G is still more expensive than 10×10G
- Note: Only decimal **bit** rates are used in networking
  No cheating like binary byte values used in storage
  1 Gbps = 10^9 b/s, Buy 256 GB Disk = 238.4 GB storage

Ref: http://en.wikipedia.org/wiki/100_Gigabit_Ethernet

Student Questions

Link Aggregation Control Protocol (LACP)



- IEEE 802.1AX-2008/IEEE 802.3ad-2000
- Allows several parallel links to be combined as one link
  3×1Gbps = 3 Gbps
- Allows any speed links to be formed
- Allows fault tolerance
  ⇒ The combined Link remains connected even if one of the member links fails
- Several proprietary extensions. E.g., aggregate links to two switches which act as one switch.

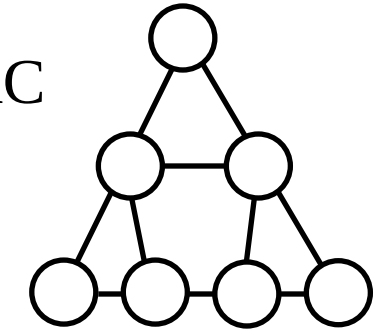
Ref: Enterasys, "Enterasys Design Center Networking – Connectivity and Topology Design Guide," 2013, http://*www.enterasys.com/company/literature/datacenter-design-guide-wp.pdf*

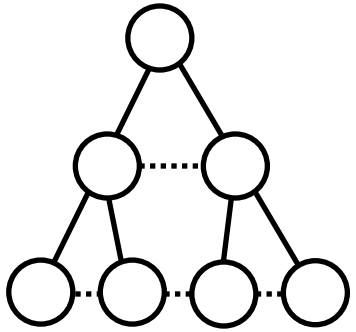
Student Questions

End of Part 1

Spanning Tree Algorithm

- Helps form a tree out of a mesh topology
- All bridges multicast to "All bridges"
  - My ID. 64-bit ID = 16-bit priority + 48-bit MAC address.
  - Root ID
  - My cost to root
- The bridges update their info using Dijkstra's algorithm and rebroadcast
- Initially, all bridges are roots but eventually converge to one root as they discover the lowest Bridge ID.
- On each LAN, the bridge with minimum cost to the root becomes the Designated bridge
- All ports of all non-designated bridges are blocked.





Student Questions

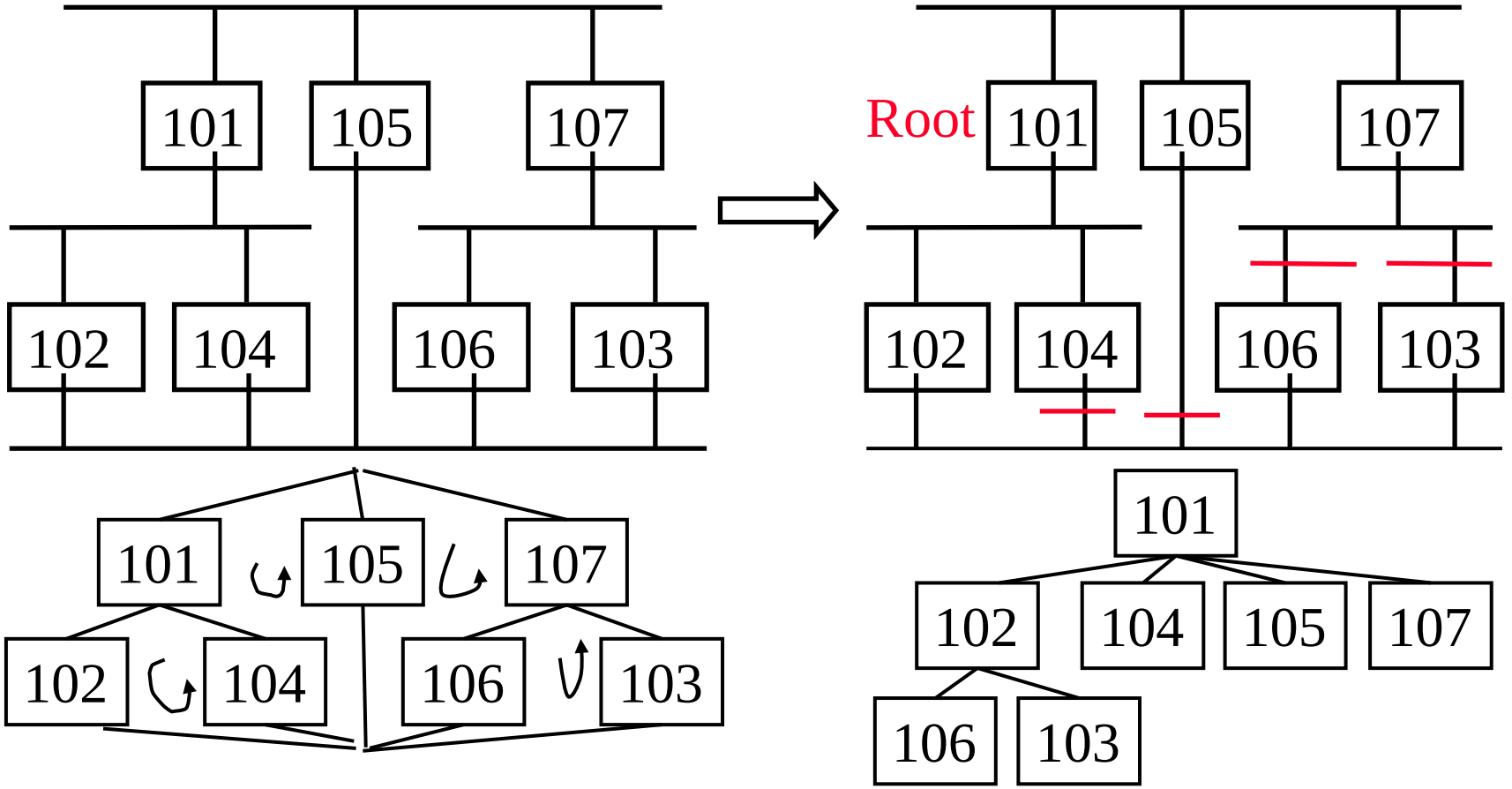
- How does a bridge decide which of its ports should be in a blocked state versus a forwarding state?

Only the shortest paths to the root are in the forwarding state.

- How does classic STP pick a root and shut down alternate paths, and what tradeoffs does that create?

The node with the highest root priority and lowest MAC address becomes the root. Alternate paths would create loops. Some traffic will never get out.

Spanning Tree Example



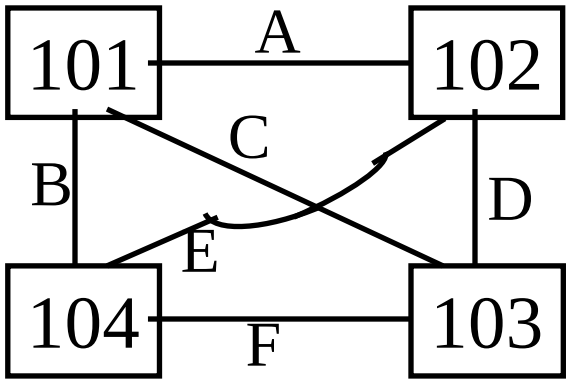
Ref: Huawei, "Understanding Spanning-Tree Protocol Topology Changes,"
https://support.huawei.com/enterprise/en/doc/EDOC1100092144

Student Questions

©2025 Raj Jain

Homework 4

- Which links in the following diagram will be blocked by the spanning tree? Justify your answer.



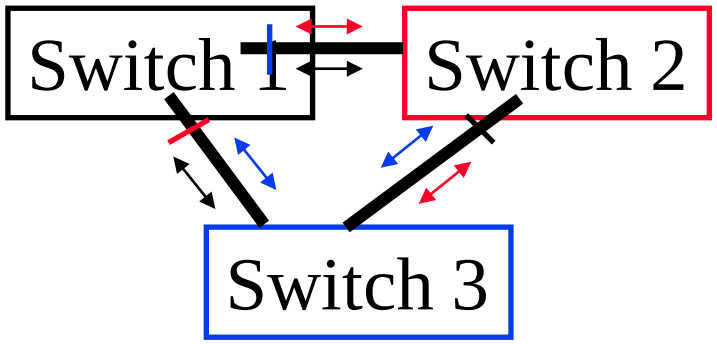
Student Questions

Enhancements to STP

- A topology change can result in a 1-minute traffic loss with STP ⇒ All TCP connections break
- Rapid Spanning Tree Protocol (RSTP)
  IEEE 802.1w-2001 incorporated in IEEE 802.1D-2004
- One tree for all VLANs ⇒ Common spanning tree
- Many trees ⇒ Multiple spanning tree (MST) protocol
  IEEE 802.1s-2002 incorporated in IEEE 802.1Q-2005
- One or more VLANs per tree.

Student Questions

MSTP (Multiple Spanning Tree)



- MSTP (Multiple STP)
  IEEE 802.1s-2002 incorporated in IEEE 802.1Q-2005
- Each tree serves a group of VLANs.
- A bridge port could be in a forwarding state for some VLANs and a blocked state for others.

Student Questions

IS-IS Protocol

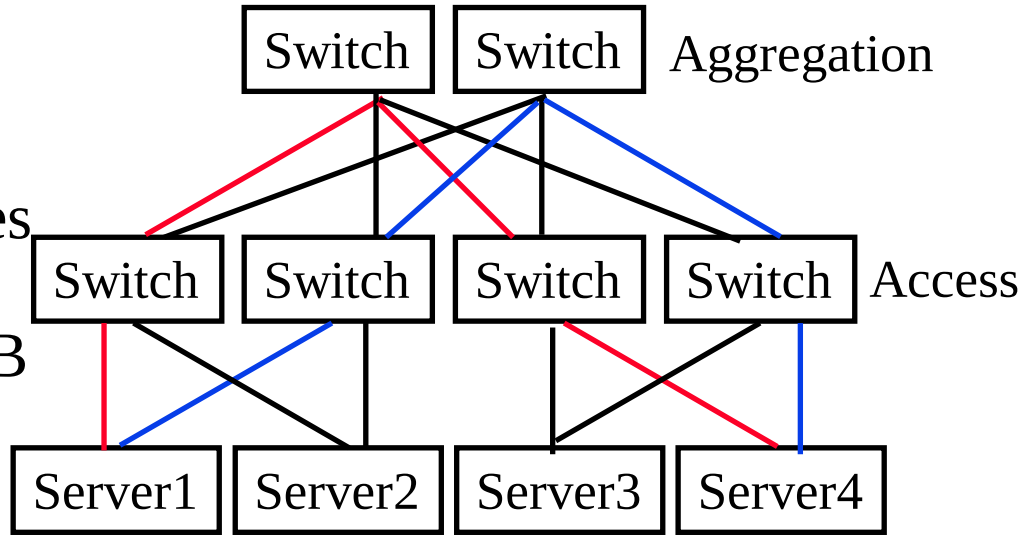
- Intermediate System to Intermediate System (IS-IS) is a protocol for building routing tables. Link-State routing protocol
  ⇒ Each node sends its connectivity (link state) information to all nodes in the network
- Each node then uses Dijkstra's algorithm to build its routing table.
- Similar to OSPF (Open Shortest Path First).
- OSPF is designed for IPv4 and then extended for IPv6.
  IS-IS is general enough to be used with any addresses
- OSPF is designed to run on top of IP
  IS-IS is general enough to be used on any transport
  ⇒ Adopted by Ethernet

Student Questions

Ref: http://en.wikipedia.org/wiki/IS-IS

Shortest Path Bridging

- IEEE 802.1aq-2012 (later incorporated in 802.1Q-2014)
- Allows all links to be used ⇒ Better CapEx
- IS-IS link-state protocol (similar to OSPF) is used to build shortest-path trees for each node to every other node within the SPB domain
- Equal-cost multi-path (ECMP) is used to distribute the load



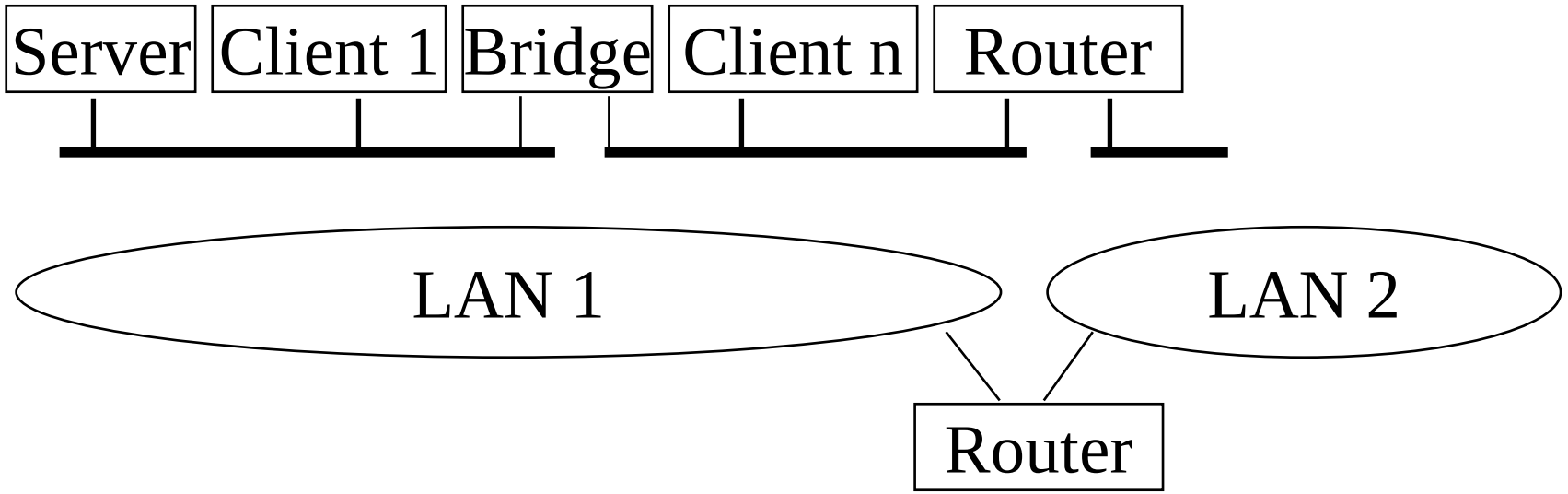
Ref: http://en.wikipedia.org/wiki/Shortest_Path_Bridging

Student Questions

- How does SPB use the IS-IS link state protocol to build the shortest path tree?

IS-IS uses Dijkstra's Link State Algorithm to find the shortest path. It is not a tree. Every node pair has a shortest path between them.

What is a LAN?

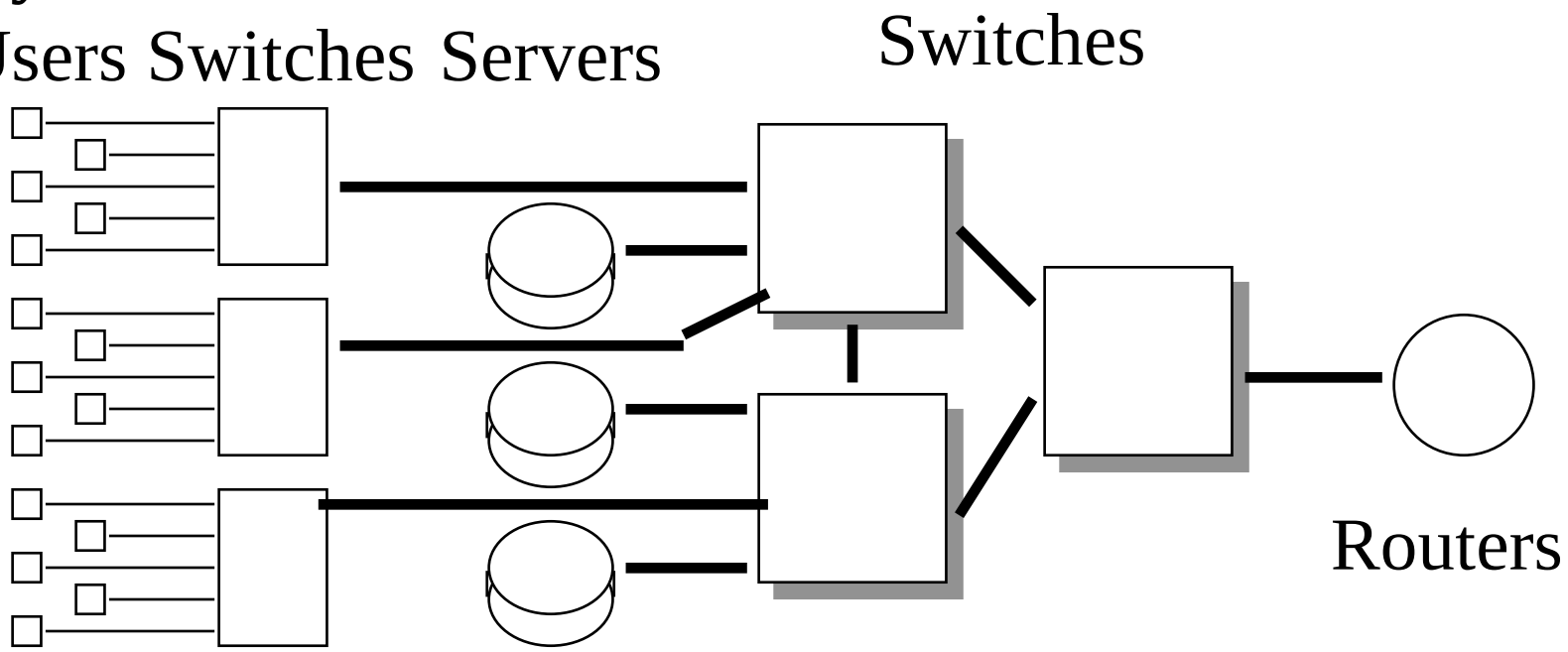


- LAN = Single broadcast domain = Subnet
- No routing between members of a LAN
- Routing required between LANs

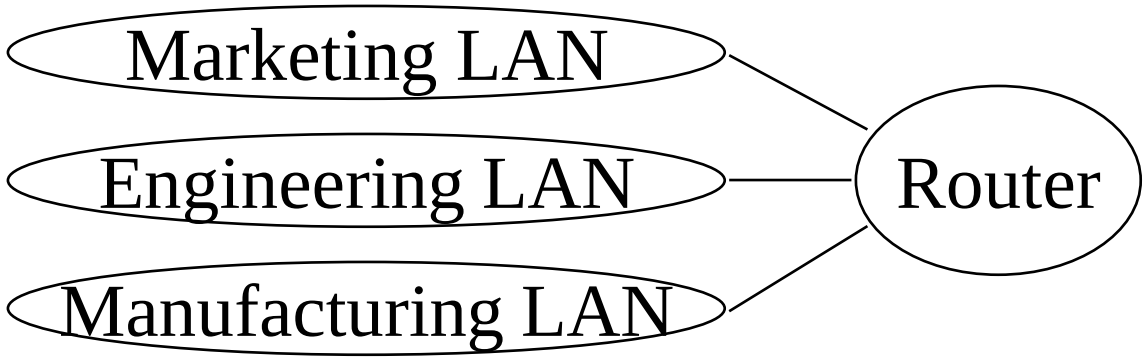
Student Questions

What is a Virtual LAN?

- **Physical View**

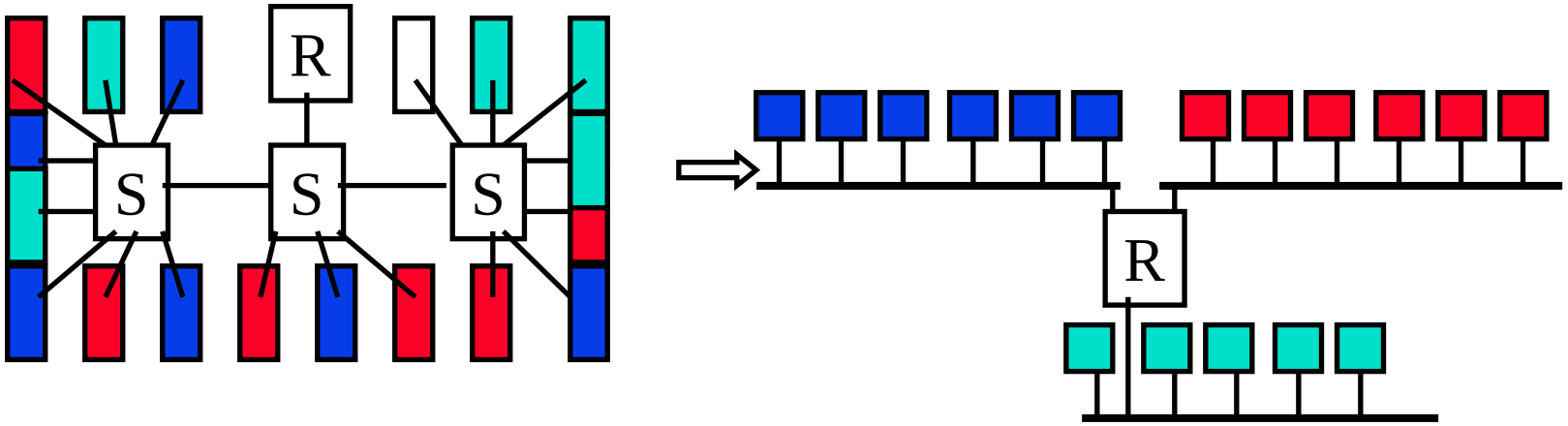


- **Logical View**



Student Questions

Virtual LAN



- Virtual LAN = Broadcasts and multicast goes only to the nodes in the virtual LAN
- LAN membership is defined by the network manager ⇒ Virtual

Student Questions

Types of Virtual LANs

- Layer-1 VLAN = Group of Physical ports
- Layer-2 VLAN = Group of MAC addresses
- Layer-3 VLAN = IP subnet

Switch Port	VLAN 1	VLAN 2
A1	√	
A2		√
A3	√	
B1		√
B2	√	

VLAN1	VLAN2
A1B234565600	21B234565600
D34578923434	634578923434
1345678903333	8345678903333
3438473450555	9438473450555
4387434304343	5387434304343
4780357056135	6780357056135
4153953470641	9153953470641
3473436374133	0473436374133
3403847333412	8403847333412
3483434343143	8483434343143
4343134134234	0343134134234

VLAN1

23.45.6

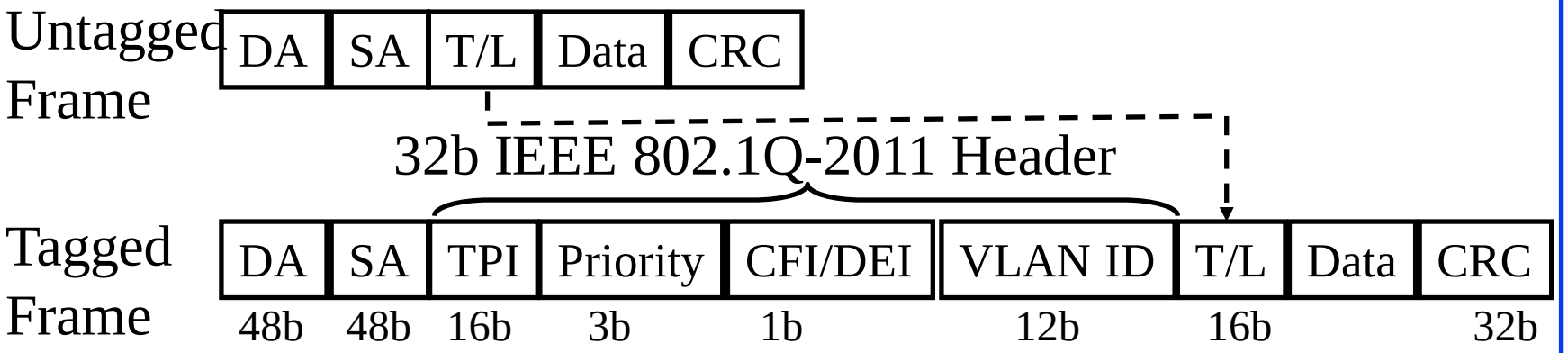
VLAN2

IPX

Student Questions

IEEE 802.1Q-2011 Tag

- Tag Protocol Identifier (TPI)
- Priority Code Point (PCP): 3 bits = 8 priorities 0..7 (High)
- Canonical Format Indicator (CFI): 0 ⇒ Standard Ethernet, 1 ⇒ IBM Token Ring format (non-canonical or non-standard)
- CFI is now replaced by Drop Eligibility Indicator (DEI)
- VLAN Identifier (12 bits ⇒ 4095 VLANs)
- Switches forward based on MAC address + VLAN ID Unknown addresses are flooded.



Ref: Canonical vs. noncanonical format, https://www.ibm.com/docs/en/zos/2.2.0?topic=address-canonical-versus-noncanonical-format

Ref: G. Santana, "Data Center Virtualization Fundamentals," Cisco Press, 2014, ISBN:1587143240

Student Questions

©2025 Raj Jain

Link Layer Discovery Protocol (LLDP)

- IEEE 802.1AB-2009
- Neighbor discovery by periodic advertisements
- Every minute the LLC frame is sent on every port to neighbors
- LLDP frame contains information in the form of Type-Length-Value (TLV)
- Types: My Chassis ID, My Port ID, Time-to-live, Port description (Manufacturer, product name, version), Administratively assigned system name, capabilities, MAC address, IP Address, Power-via-MDI, Link aggregation, maximum frame size, …



Ref: M. Srinivasan, "Tutorial on LLDP," http://www.eetimes.com/document.asp?doc_id=1272069
Ref: http://en.wikipedia.org/wiki/Link_Layer_Discovery_Protocol

Student Questions

End of Part 1

Data Center Bridging

- Goal: To enable storage traffic over Ethernet
- Four Standards:
  - Priority-based Flow Control (IEEE 802.1Qbb-2011)
  - Enhanced Transmission Selection (IEEE 802.1Qaz-2011)
  - Congestion Control (IEEE 802.1Qau-2010)
  - Data Center Bridging Exchange (IEEE 802.1Qaz-2011)
- All of these are now incorporated in IEEE 802.1Q-2014

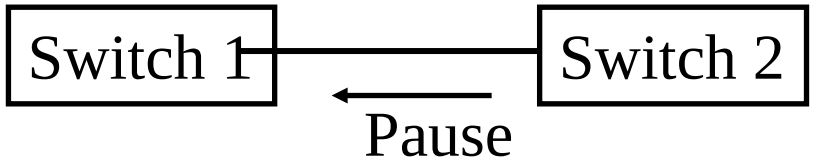
Student Questions

- Why does DCB reduce packet loss for storage/RDMA—what specific features matter, and what are their limits?

The three features are listed first on this slide. The fourth is required to use these three. They reduce the loss and allow higher-priority traffic first.

Ref: M. Hagen, "Data Center Bridging Tutorial," http://www.iol.unh.edu/services/testing/dcb/training/DCB-Tutorial.pdf

Ethernet Flow Control: Pause Frame



- Defined in IEEE 802.3x-1997. A form of on-off flow control.
- A receiving switch can stop the adjoining sending switch by sending a "Pause" frame.
  Stops the sender from sending further information for a time specified in the pause frame.
- The frame is addressed to a standard (well-known) multicast address. This address is acted upon but not forwarded.
- Stops all traffic. Causes congestion backup.

Ref: http://en.wikipedia.org/wiki/Ethernet_flow_control

Student Questions

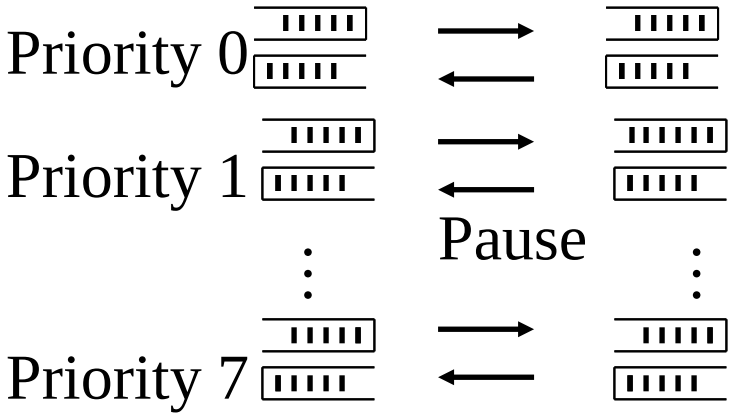
- Why does a Pause Frame stop all traffic on the link rather than just a single flow?

It is how it was originally standardized.

- The pause frame technically stops all traffic when congestion is detected. Is this in practice in the modern world?

Yes, if you have an old switch.

Priority-based Flow Control (PFC)



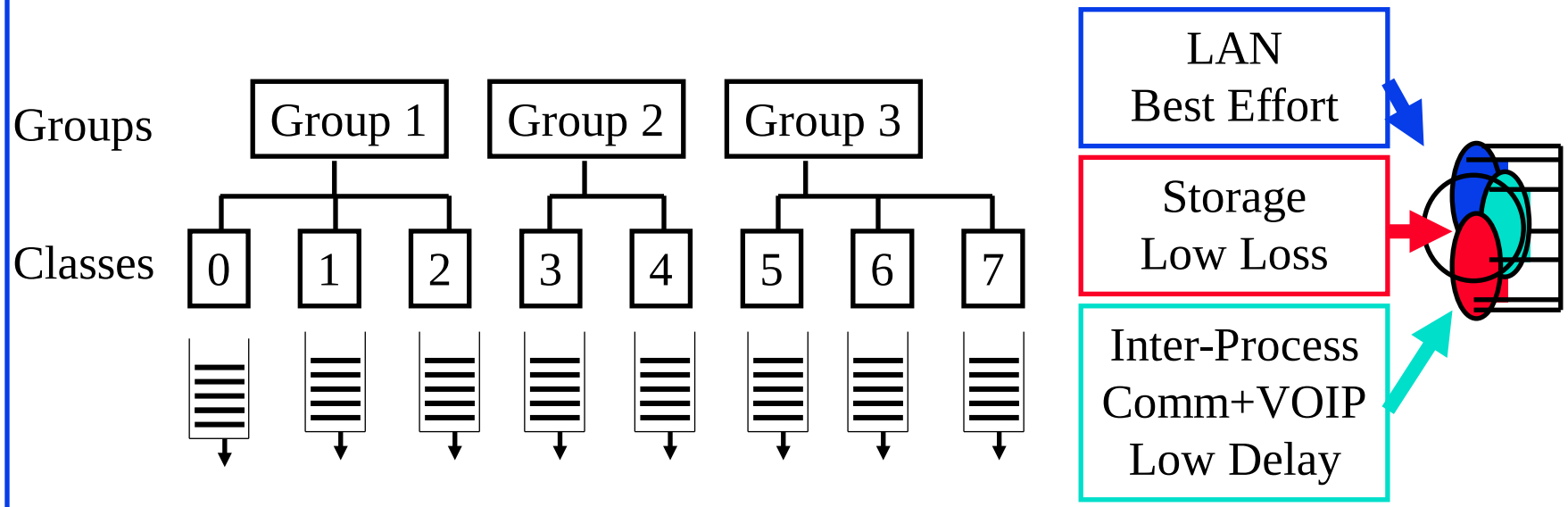
- IEEE 802.1Qbb-2011
- IEEE 802.1Qbb-2011 allows any single priority to be stopped. Others keep sending

Ref: J. L. White, "Technical Overview of Data Center Networks," SNIA, 2013, http://www.snia.org/sites/default/education/tutorials/2012/fall/networking/JosephWhite_Technical%20Overview%20of%20Data%20Center%20Networks.pdf

Student Questions

Enhanced Transmission Selection

- IEEE 802.1Qaz-2011
- Goal: Guarantee bandwidth for applications sharing a link
- Traffic is divided into eight classes (not priorities)
- The classes are grouped.
- The standard requires min 3 groups: 1 with PFC (Storage with low loss), 1 W/O PFC (LAN), 1 with Strict Priority (Inter-process communication and VOIP with low latency)

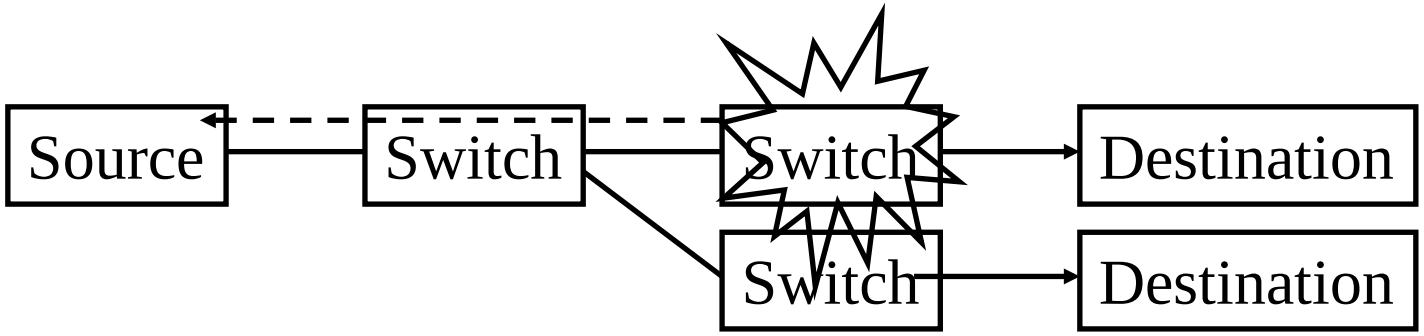


Student Questions

- Can you review this slide again?

Sure.

Quantized Congestion Notification (QCN)



- IEEE 802.1Qau-2010 Dynamic Congestion Notification
- A source quench message is sent by the congested switch directly to the source. The source reduces its rate for that flow.
- Sources need to keep per-flow states and control mechanisms
- Easy for switch manufacturers but complex for hosts. Implemented in switches but not in hosts ⇒ Not effective.
- The source may be a router in a subnet and not the real source ⇒ The router will drop the traffic. QCN does not help.

Student Questions

Ref: *I. Pepelnjak*, "DCB Congestion Notification (802.1Qau)," http://blog.ipspace.net/2010/11/data-center-bridging-dcb-congestion.html

©2025 Raj Jain

DCBX

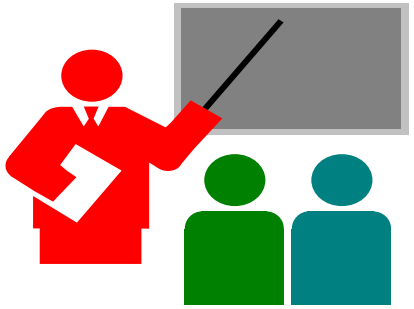
- Data Center Bridging eXchange, IEEE 802.1Qaz-2011
- Uses LLDP to negotiate quality metrics and capabilities for Priority-based Flow Control, Enhanced Transmission Selection, and Quantized Congestion Notification
- New TLV's
  - Priority group definition
  - Group bandwidth allocation
  - PFC enablement per priority
  - QCN enablement
  - DCB protocol profiles
  - FCoE and iSCSI profiles

Student Questions

Recent Advances in Data Center Ethernet

- IEEE 802.3df-2024 standard for 800 GbE.
  IEEE 802.dj is next for 1.6 TbE.
- Ethernet is being tailored for AI/ML needs.
- 10BASE-T1L Single-Pair Ethernet for low-speed long reach applications in Industrial IoT
- IEEE 802.3at for Power over Ethernet+ (PoE+) provides 25.5W compared to 82.3af PoE standard with 15.4W limit.

Student Questions



Summary

1. Ethernet's use of IDs as addresses makes it very easy to move systems in the data center ⇒ Keep traffic on the same Ethernet
2. Spanning tree is wasteful of resources and slow.
   Ethernet now uses shortest-path bridging (similar to OSPF)
3. VLANs allow different non-trusting entities to share an Ethernet network
4. Data center bridging extensions reduce packet loss by enhanced transmission selection and Priority-based flow control
5. High (1.6 Tbps) and Single-Pair Ethernets are being designed.
   PoE+ can deliver up to 25.4 W.

Student Questions

List of Acronyms

- BER Bit Error Rate
- BPDU Bridge Protocol Data Unit
- CD Collision Detection
- CFI Canonical Format Indicator
- CRC Cyclic Redundancy Check
- CSMA/CD Carrier Sense Multiple Access with Collision Detection
- DA Destination Address
- DCB Data Center Bridging
- DCBX Data Center Bridging eXtension
- DEI Drop Eligibility Indicator
- DNS Domain Name System
- ECMP Equal-cost multi-path
- ETS Enhanced Transmission Selection
- GB Giga Byte

Student Questions

List of Acronyms (Cont)

ID	Identifier
IP	Internet Protocol
IEEE	Institution of Electrical and Electronics Engineers
IS-IS	Intermediate System to Intermediate System
iSCSI	Internet Small Computer System Interface
LACP	Link Aggregation Control Protocol
LAN	Local Area Network
LLC	Logical Link Control
LLDP	Link Layer Discovery Protocol
MAC	Media Access Control
MDI	Medium Dependent Interface
MSB	Most significant byte first
MST	Multiple Spanning Tree
MSTP	Multiple Spanning Tree Protocol
OAM	Operations, Administration, and Management

Student Questions

List of Acronyms (Cont)

- OSPF Open Shortest Path First
- OUI Organizationally Unique Identifier
- PCP Priority Code Point
- PFC Priority-based Flow Control
- PHY Physical layer
- QCN Quantized Congestion Notification
- QoS Quality of Service
- RSTP Rapid Spanning Tree Protocol
- SA Source Address
- SNIA Storage Networking Industries Association
- SPB Shortest Path Bridging
- STP Spanning Tree Protocol
- TCP Transmission Control Protocol
- TLV Type-Length-Value
- TPI Tag Protocol Identifier
- VLAN Virtual Local Area Network
- VM Virtual machine

Student Questions

List of Acronyms (Cont)

- VOIP Voice over IP
- WAN Wide Area Network
- WiFi Wireless Fidelity
- WiMAX Wireless Interoperability for Microwave Access

Student Questions

Reading List

- G. Santana, "Data Center Virtualization Fundamentals," Cisco Press, 2013, Safari Book, ISBN:1587143240
- Huawei, "Understanding Spanning-Tree Protocol Topology Changes," https://support.huawei.com/enterprise/en/doc/EDOC1100092144
- Cisco, Understanding Rapid Spanning Tree Protocol (802.1w), http://www.cisco.com/en/US/tech/tk389/tk621/technologies_white_paper09186a0080094cfa.shtml
- Canonical vs. noncanonical format, https://www.ibm.com/docs/en/zos/2.2.0?topic=address-canonical-versus-noncanonical-format
- M. Hagen, "Data Center Bridging Tutorial," http://www.iol.unh.edu/services/testing/dcb/training/DCB-Tutorial.pdf
- J. L. White, "Technical Overview of Data Center Networks," SNIA, 2013, https://www.snia.org/educational-library/overview-data-center-networks-fall-2013-2013

Student Questions

Reading List (Cont)

- I. Pepelnjak, "DCB Congestion Notification (802.1Qau)," http://blog.ipspace.net/2010/11/data-center-bridging-dcb-congestion.html

Student Questions

Wikipedia Links

- http://en.wikipedia.org/wiki/10-gigabit_Ethernet
- http://en.wikipedia.org/wiki/100_Gigabit_Ethernet
- http://en.wikipedia.org/wiki/Data_center
- http://en.wikipedia.org/wiki/Data_center_bridging
- http://en.wikipedia.org/wiki/Data_link_layer
- http://en.wikipedia.org/wiki/EtherChannel
- http://en.wikipedia.org/wiki/Ethernet
- http://en.wikipedia.org/wiki/Ethernet_flow_control
- http://en.wikipedia.org/wiki/Ethernet_frame
- http://en.wikipedia.org/wiki/Ethernet_physical_layer
- http://en.wikipedia.org/wiki/EtherType
- http://en.wikipedia.org/wiki/Fast_Ethernet
- http://en.wikipedia.org/wiki/Gigabit_Ethernet

Student Questions

Wikipedia Links (Cont)

- http://en.wikipedia.org/wiki/IEEE_802.1aq
- http://en.wikipedia.org/wiki/IEEE_802.1D
- http://en.wikipedia.org/wiki/IEEE_802.1Q
- http://en.wikipedia.org/wiki/IEEE_802.3
- http://en.wikipedia.org/wiki/IEEE_P802.1p
- http://en.wikipedia.org/wiki/IS-IS
- http://en.wikipedia.org/wiki/Link_Aggregation
- http://en.wikipedia.org/wiki/Link_Aggregation_Control_Protocol
- http://en.wikipedia.org/wiki/Link_layer
- http://en.wikipedia.org/wiki/Link_Layer_Discovery_Protocol
- http://en.wikipedia.org/wiki/Logical_link_control
- http://en.wikipedia.org/wiki/MAC_address
- http://en.wikipedia.org/wiki/MC-LAG

Student Questions

©2025 Raj Jain

Wikipedia Links (Cont)

- http://en.wikipedia.org/wiki/Media_Independent_Interface
- http://en.wikipedia.org/wiki/Minimum_spanning_tree
- http://en.wikipedia.org/wiki/Network_switch
- http://en.wikipedia.org/wiki/Organizationally_unique_identifier
- http://en.wikipedia.org/wiki/Port_Aggregation_Protocol
- http://en.wikipedia.org/wiki/Priority-based_flow_control
- http://en.wikipedia.org/wiki/RSTP
- http://en.wikipedia.org/wiki/Shortest_Path_Bridging
- http://en.wikipedia.org/wiki/Spanning_tree
- http://en.wikipedia.org/wiki/Spanning_Tree_Protocol
- http://en.wikipedia.org/wiki/Subnetwork_Access_Protocol
- http://en.wikipedia.org/wiki/Virtual_LAN

Student Questions

Scan This to Download These Slides



Raj Jain
http://rajjain.com

http://www.cse.wustl.edu/~jain/cse5700-25/m_04dce.htm

Student Questions

Related Modules



CSE567M: Computer Systems Analysis (Spring 2013),
https://www.youtube.com/playlist?list=PLjGG94etKypJEKjNAa1n_1X0bWWNyZcof

CSE473S: Introduction to Computer Networks (Fall 2011),



https://www.youtube.com/playlist?list=PLjGG94etKypJWOSPMh8Azcgy5e_10TiDw



Wireless and Mobile Networking (Spring 2016),
https://www.youtube.com/playlist?list=PLjGG94etKypKeb0nzyN9tSs_HCd5c4wXF

CSE571S: Network Security (Fall 2011),



https://www.youtube.com/playlist?list=PLjGG94etKypKvzfVtutHcPFJXumyyg93u



Video Podcasts of Prof. Raj Jain's Lectures,
https://www.youtube.com/channel/UCN4-5wzNP9-ruOzQMs-8NUw

Student Questions