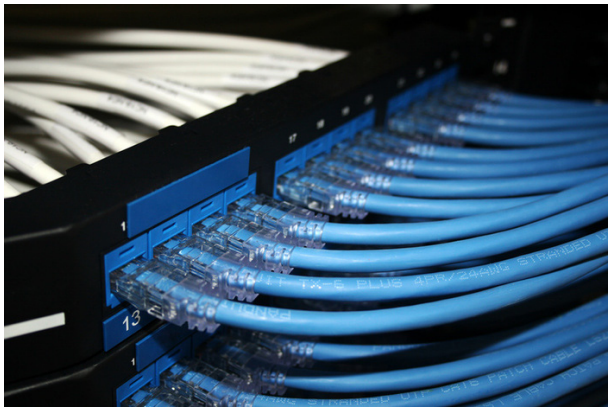
Data Center Ethernet



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/cse570-13/

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

Quiz: True or False?

Which of the following statements are generally true?

T F

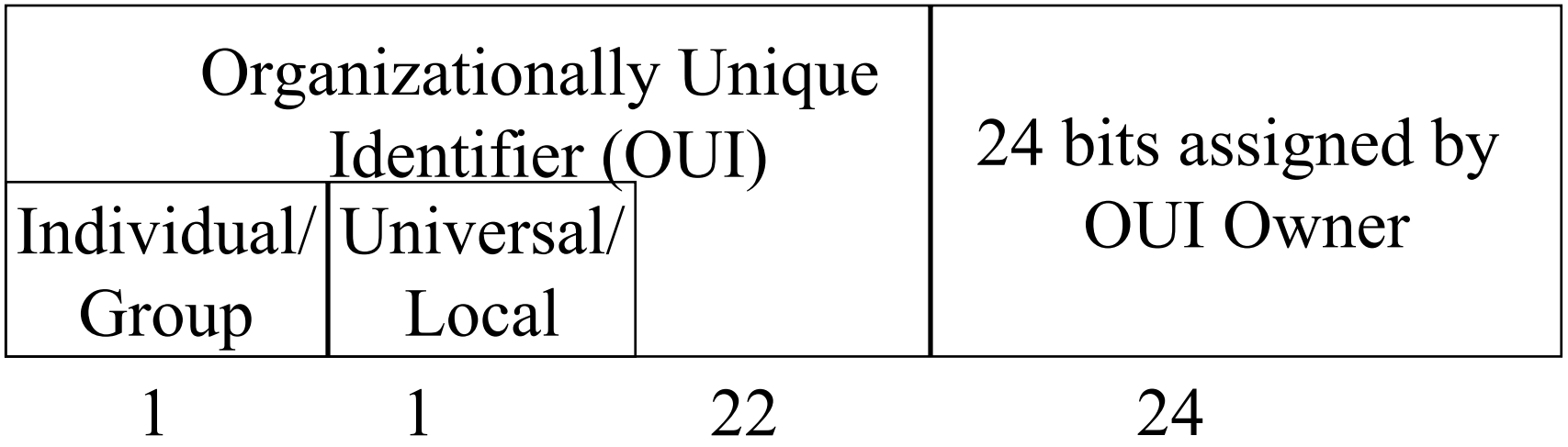
- ❑ ❑ Ethernet is a local area network (Local ≤ 2km)
- ❑ ❑ Token ring, Token Bus, and CSMA/CD are the three most common LAN access methods.
- ❑ ❑ Ethernet uses CSMA/CD.
- ❑ ❑ Ethernet bridges use spanning tree for packet forwarding.
- ❑ ❑ Ethernet frames are 1518 bytes.
- ❑ ❑ Ethernet does not provide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

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

IEEE 802 Address Format

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



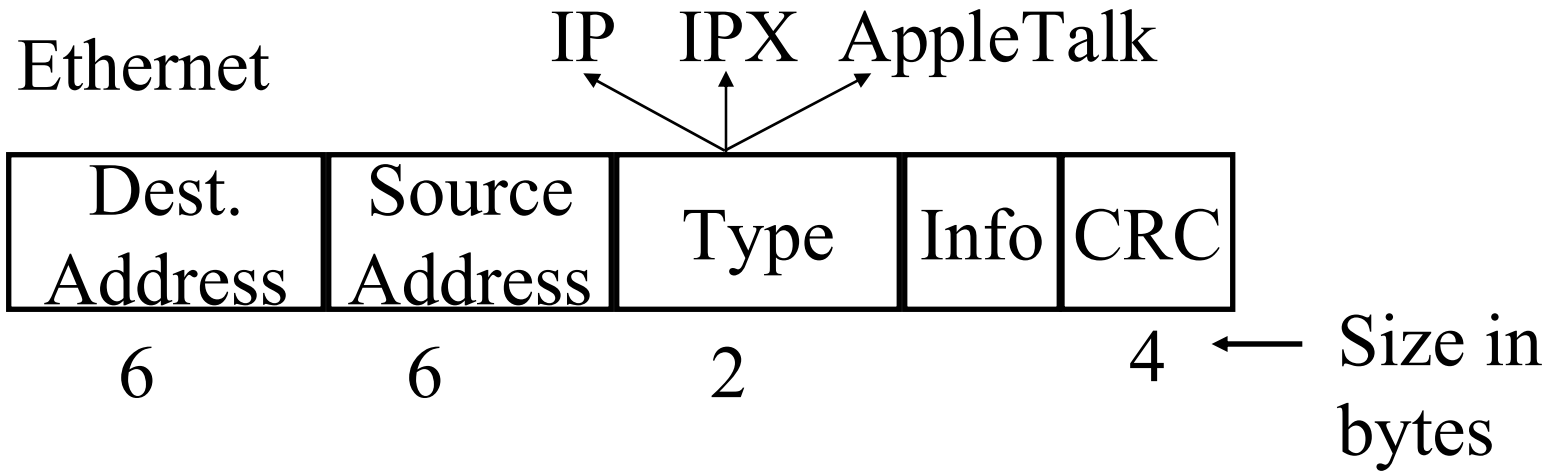
- Multicast = "To all bridges on this LAN"
- Broadcast = "To all stations"
  = 111111....111 = FF:FF:FF:FF:FF:FF

IEEE Standards Numbering System

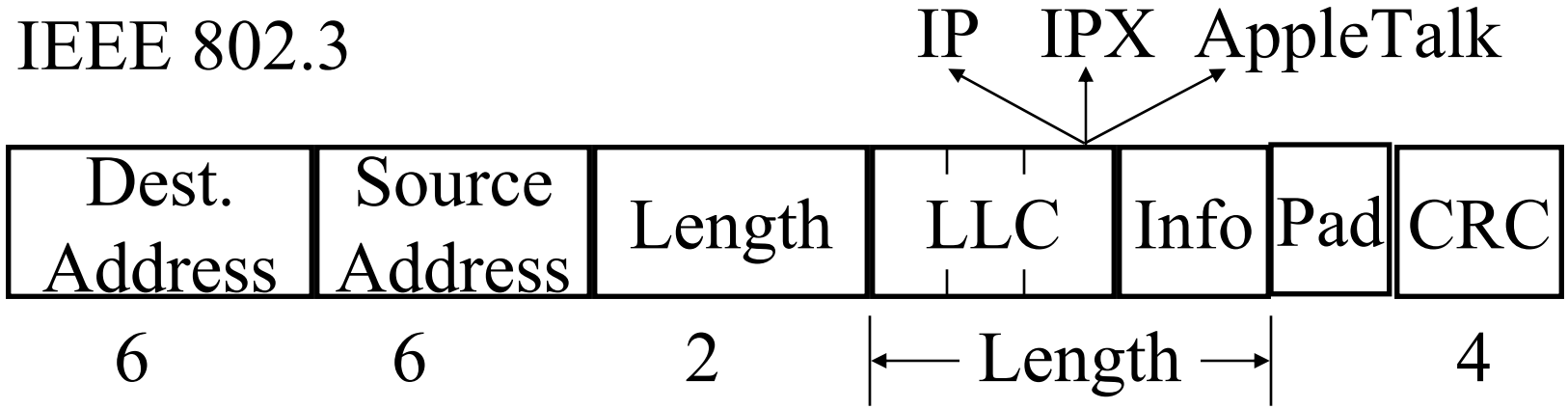
- IEEE 802.* and IEEE 802.1* standards (e.g., IEEE 802.1Q-2011) apply to all IEEE 802 technologies:
  - IEEE 802.3 Ethernet
  - IEEE 802.11 WiFi
  - IEEE 802.16 WiMAX
- IEEE 802.3* standards apply only to Ethernet, e.g., IEEE802.3ba-2010
- Standards with all upper case letters are base standards E.g., IEEE 802.1AB-2009
- Standards with lower case are additions/extensions/revisions. 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

Ethernet vs IEEE 802.3

- Ethernet



- IEEE 802.3



Names, IDs, Locators

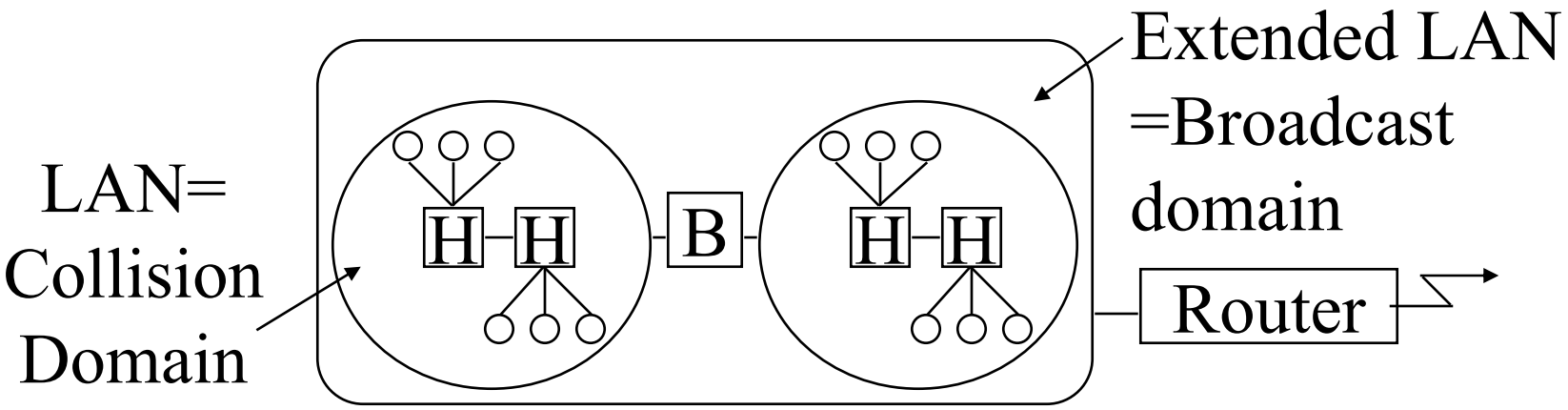
Name: John Smith

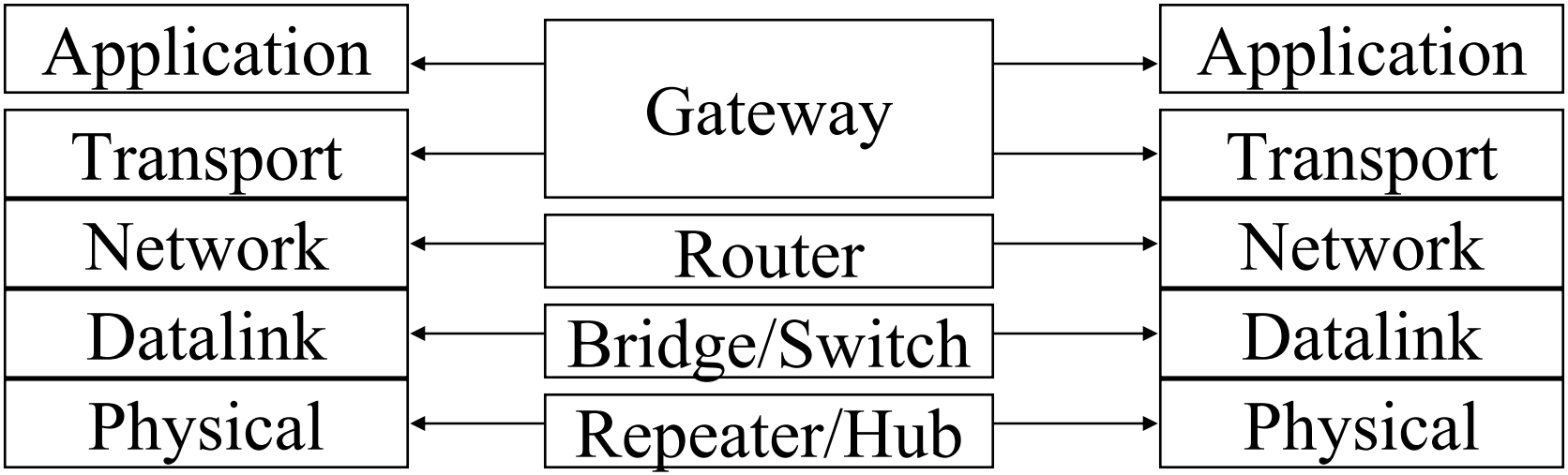
ID: 012-34-5678

Locator:
1234 Main Street
Big City, MO 12345
USA

- Locator changes as you move, 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

Interconnection Devices





Interconnection Devices

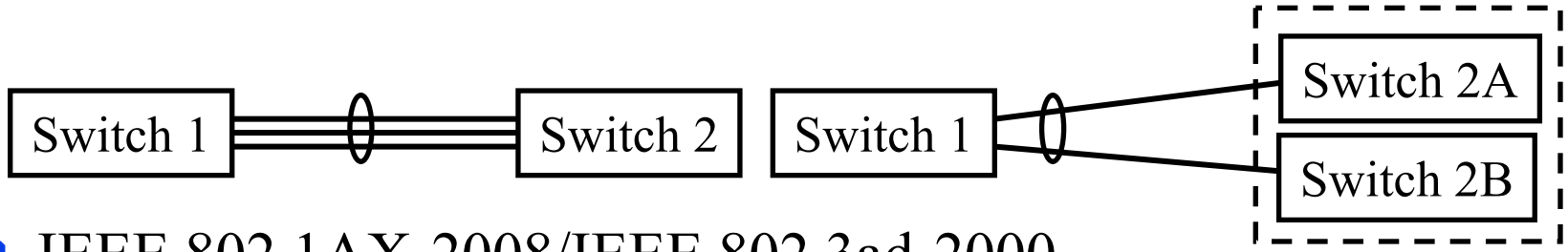
- **Repeater**: PHY device that restores data and collision signals
- **Hub**: Multiport repeater + fault detection and recovery
- **Bridge**: Datalink layer device connecting two or more collision domains. MAC multicasts are propagated throughout "extended LAN."
- **Router**: Network layer device. IP,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

Ethernet Speeds

- IEEE 802.3ba-2010 (40G/100G) standard
- 10Mbps, 100 Mbps, 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

Link Aggregation Control Protocol (LACP)

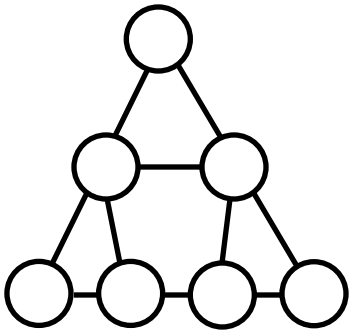


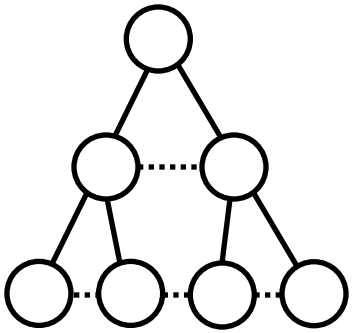
- IEEE 802.1AX-2008/IEEE 802.3ad-2000
- Allows several parallel links to be combined as one link
  3×1Gbps = 3 Gbps
- Allows any speed links to be formed
- Allows fault tolerance
  ⇒ Combined Link remains connected even if one of the member links fails
- Several proprietary extensions. E.g., aggregate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

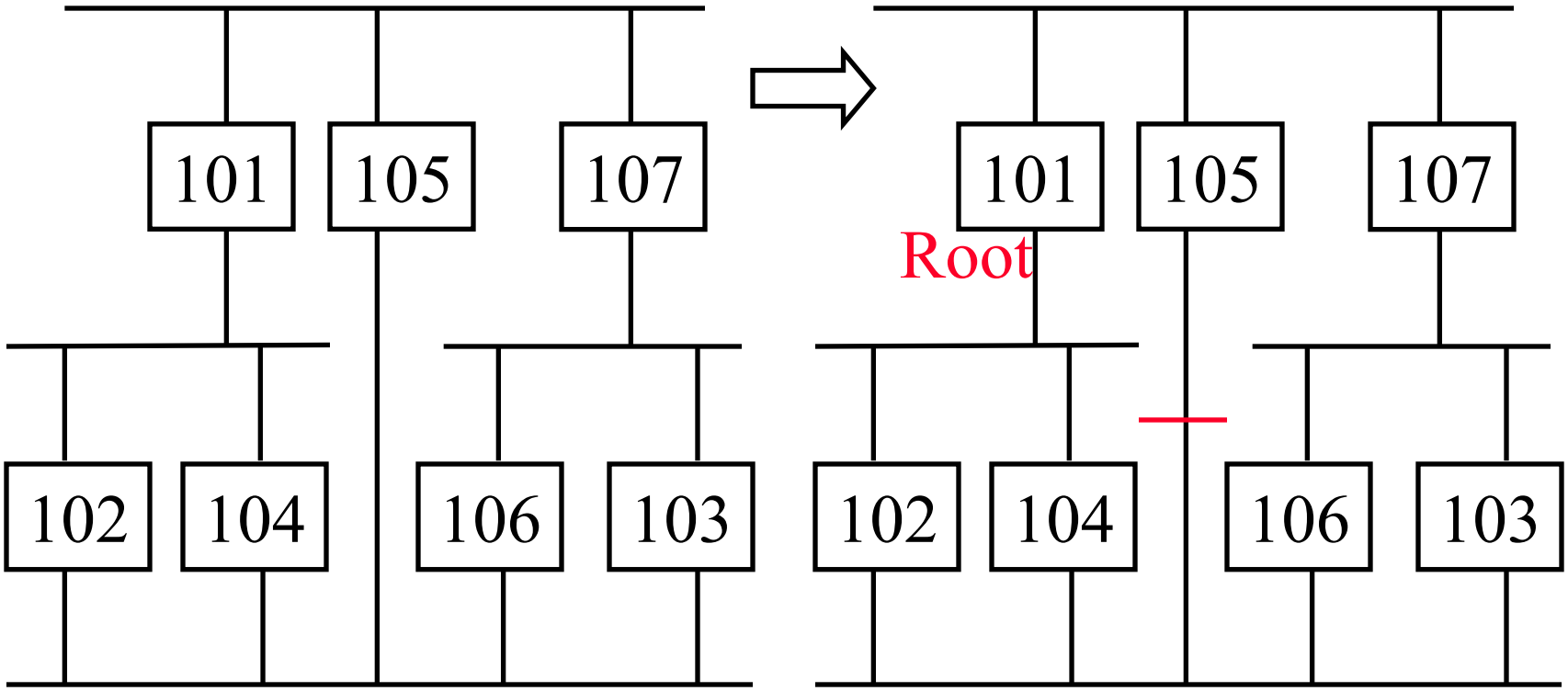
Spanning Tree Algorithm

- 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 find out 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.





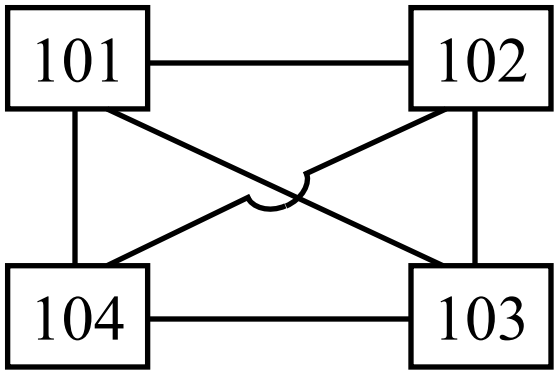
Spanning Tree Example



Ref: Cisco, "Understanding Spanning-Tree Protocol Topology Changes,"
http://www.cisco.com/en/US/tech/tk389/tk621/technologies_tech_note09186a0080094797.shtml

Homework 4

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



Enhancements to STP

- A topology change can result in 1 minute of 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.

Rapid Spanning Tree

- IEEE 802.1w-2001 incorporated in IEEE 802.1D-2004
- Normal spanning tree takes a few minutes to stabilize after a topology change ⇒ All traffic interrupted for this time
- RSTP fixes this by:
  1. Being time + event driven instead of just event driven
     - Once converged, STP sends BPDUs only on change
     - RSTP sends Hellos every 2 seconds. Quick failure detection.
  2. Differentiating between edge ports (servers) and non-edge ports (switches). No loops ever on edge ports
  3. Differentiating between point-to-point links (full duplex) and shared links (half-duplex). RSTP only on full-duplex

Ref: Cisco, Understanding Rapid Spanning Tree Protocol (802.1w), http://www.cisco.com/en/US/tech/tk389/tk621/technologies_white_paper09186a0080094cfa.shtml

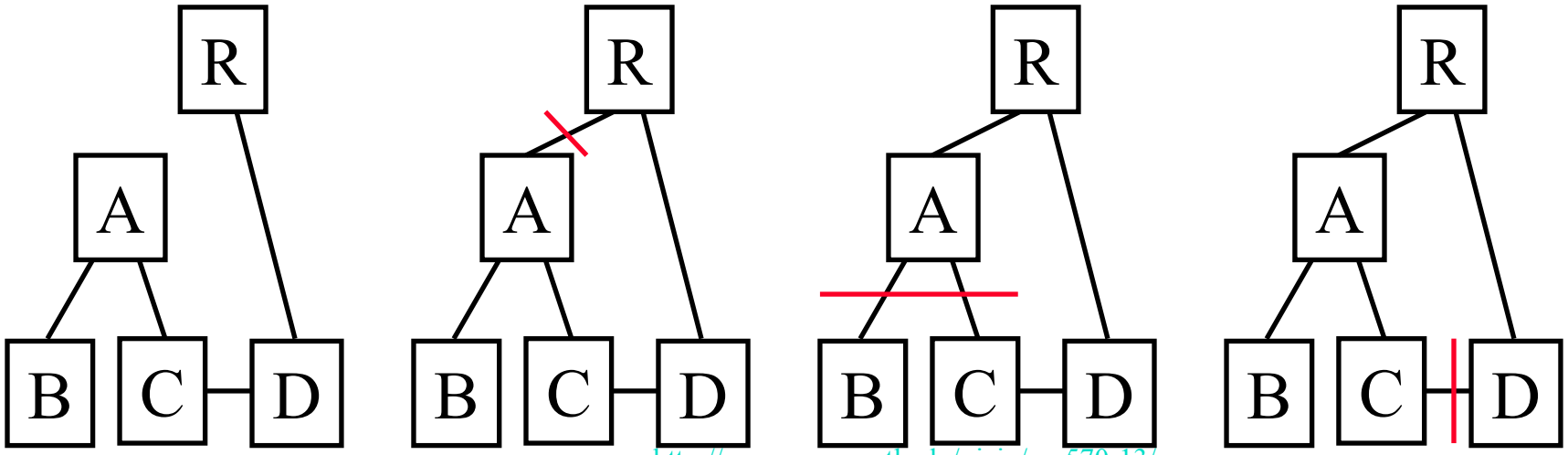
RSTP (Cont)

4. Merging three port states (Disabled, blocking, listening) in to one (discarding).
5. Adding 4 new flags in BPDU, that allow sending a proposal and accepting or not accepting the received proposal

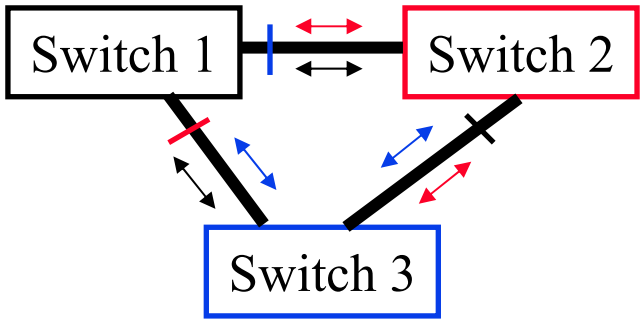
- RSTP is backward compatible with STP.
RSTP-unaware bridge drop RSTP and RSTP is not used.

RSTP Example

- A new link is added between R and A.
  The link comes up in discarding state.
- R and A exchange proposal.
  A realizes that it has a shorter path to the root.
- A unblocks the R-A link and blocks A-B, A-C links and sends proposal.
- B is edge port, it always accepts. C accepts and blocks C-D



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 forwarding state for some VLANs and blocked state for others.

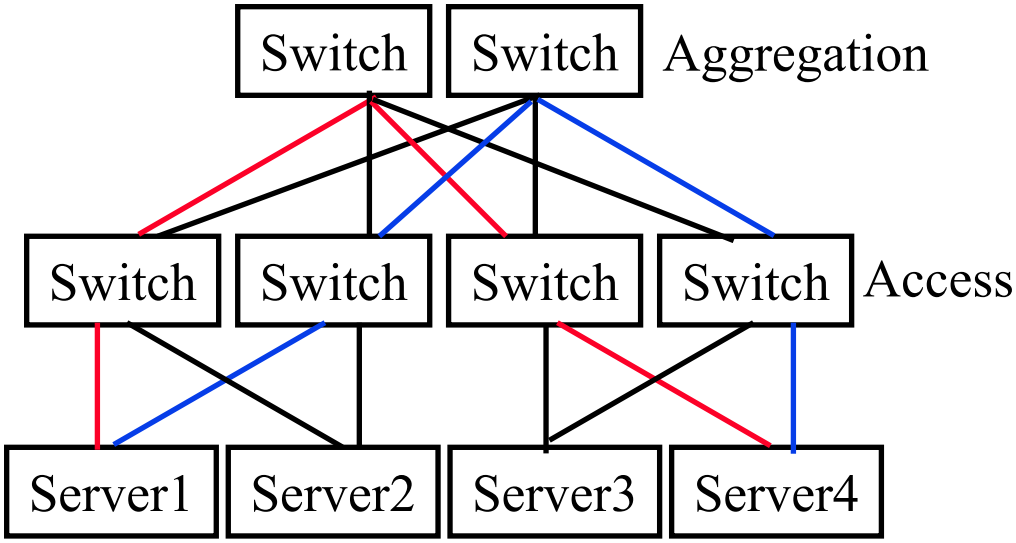
IS-IS Protocol

- Intermediate System to Intermediate System (IS-IS) is a protocol to build routing tables. Link-State routing protocol => Each nodes sends its connectivity (link state) information to all nodes in the network
- Dijkstra's algorithm is then used by each node 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 type of addresses
- OSPF is designed to run on the top of IP
  IS-IS is general enough to be used on any transport
  ⇒ Adopted by Ethernet

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

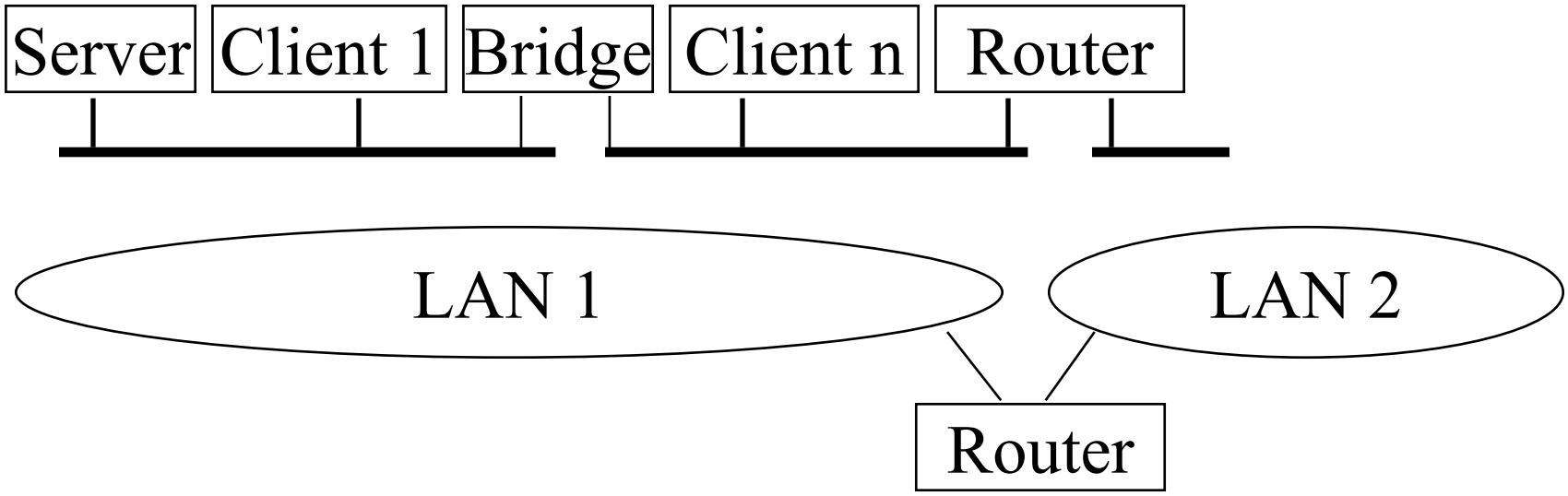
Shortest Path Bridging

- IEEE 802.1aq-2012
- 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) used to distribute load



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

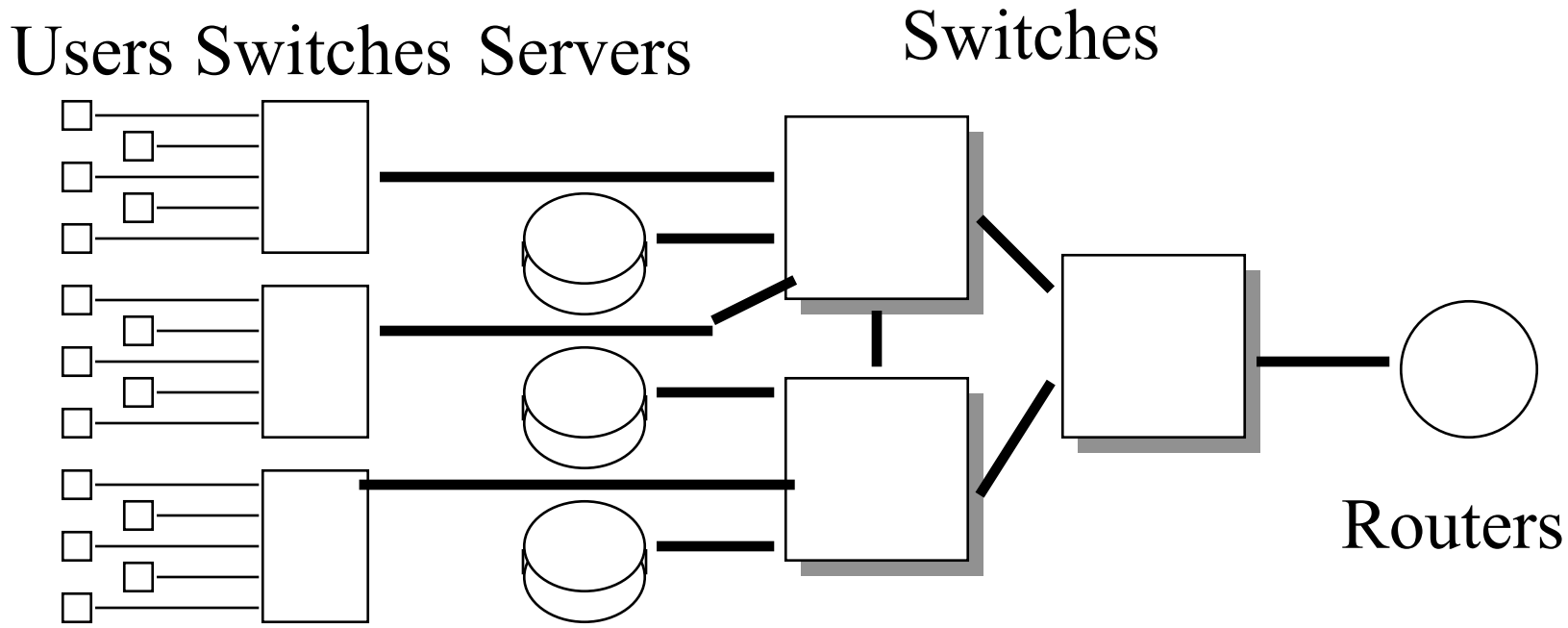
What is a LAN?



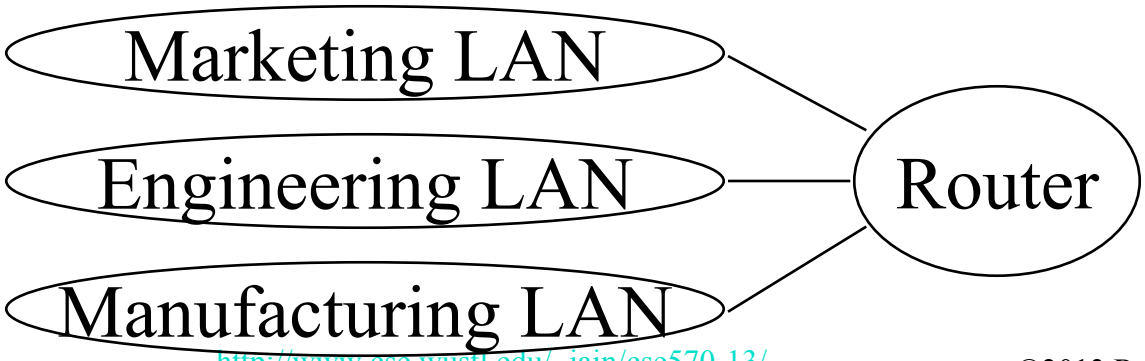
- LAN = Single broadcast domain = Subnet
- No routing between members of a LAN
- Routing required between LANs

What is a Virtual LAN

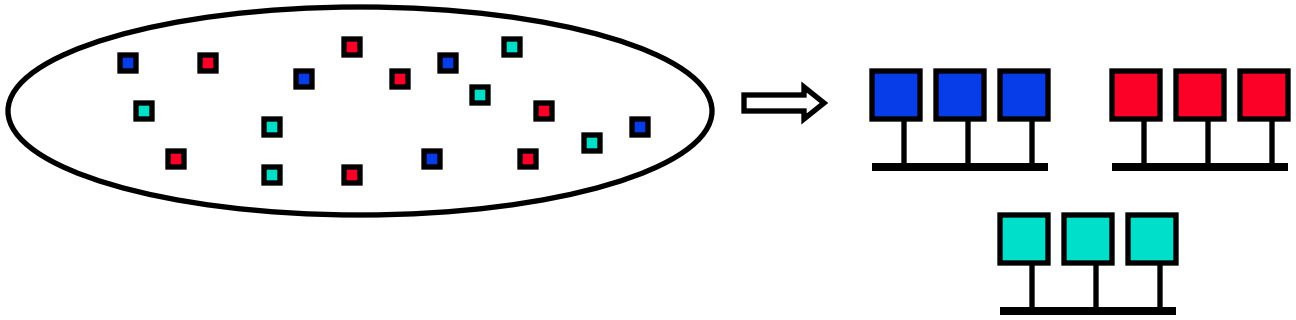
- **Physical View**



- **Logical View**



Virtual LAN



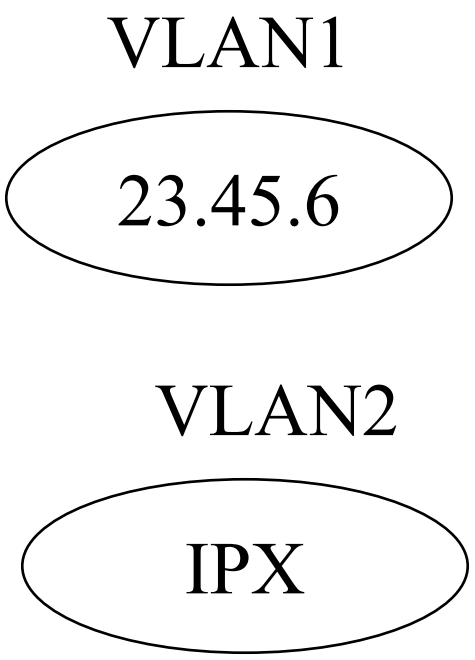
- Virtual LAN = Broadcasts and multicast goes only to the nodes in the virtual LAN
- LAN membership defined by the network manager ⇒ Virtual

Types of Virtual LANs

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

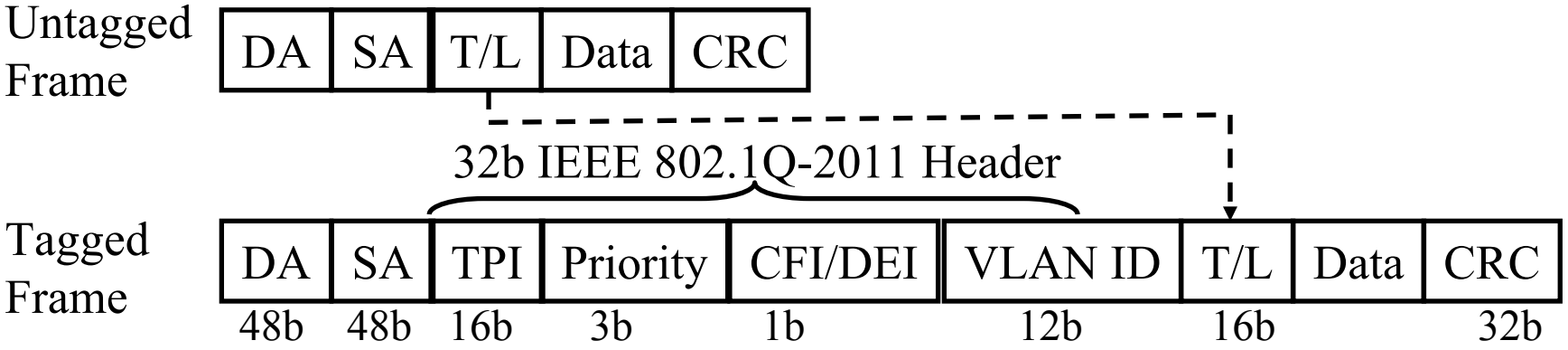
Switch Port	VLAN 1	VLAN 2
A1	√	
A2		√
A3	√	
B1		√
B2	√	

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



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 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. MSB Addresses, http://support.lexmark.com/index?page=content&id=HO1299&locale=en&userlocale=EN_US

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

Link Layer Discovery Protocol (LLDP)

- IEEE 802.1AB-2009
- Neighbor discovery by periodic advertisements
- Every minute a 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, …

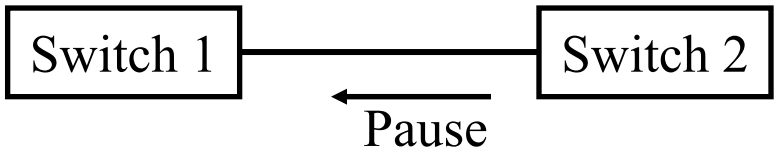
Ref: Extreme Networks, "Link Layer Discovery Protocol (LLDP)," http://*www.extremenetworks.com/libraries/products/**LLDP**_TB.pdf*
Ref: M. Srinivasan, "Tutorial on LLDP," http://www.eetimes.com/document.asp?doc_id=1272069
Ref: http://en.wikipedia.org/wiki/Link_Layer_Discovery_Protocol

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)

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

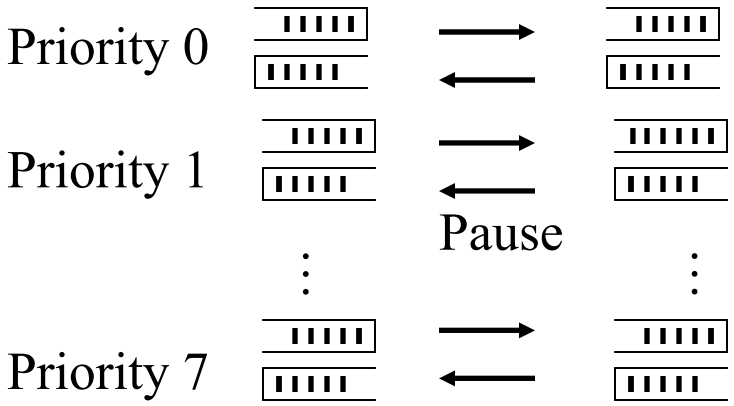
Ethernet Flow Control: Pause Frame



- 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 any 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

Priority-based Flow Control (PFC)

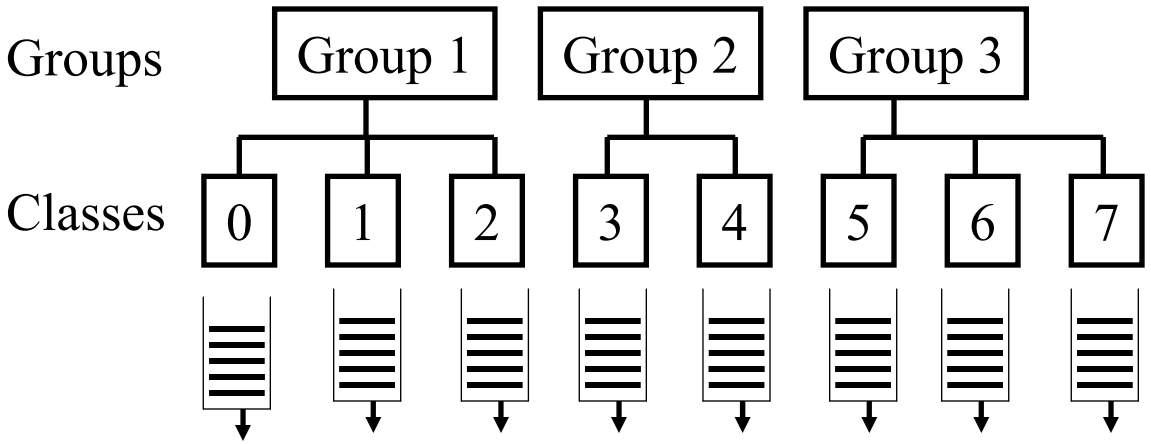


- IEEE 802.1Qbb-2011
- IEEE 802.1Qbb-2011 allows any single priority to be stopped. Others keep sending

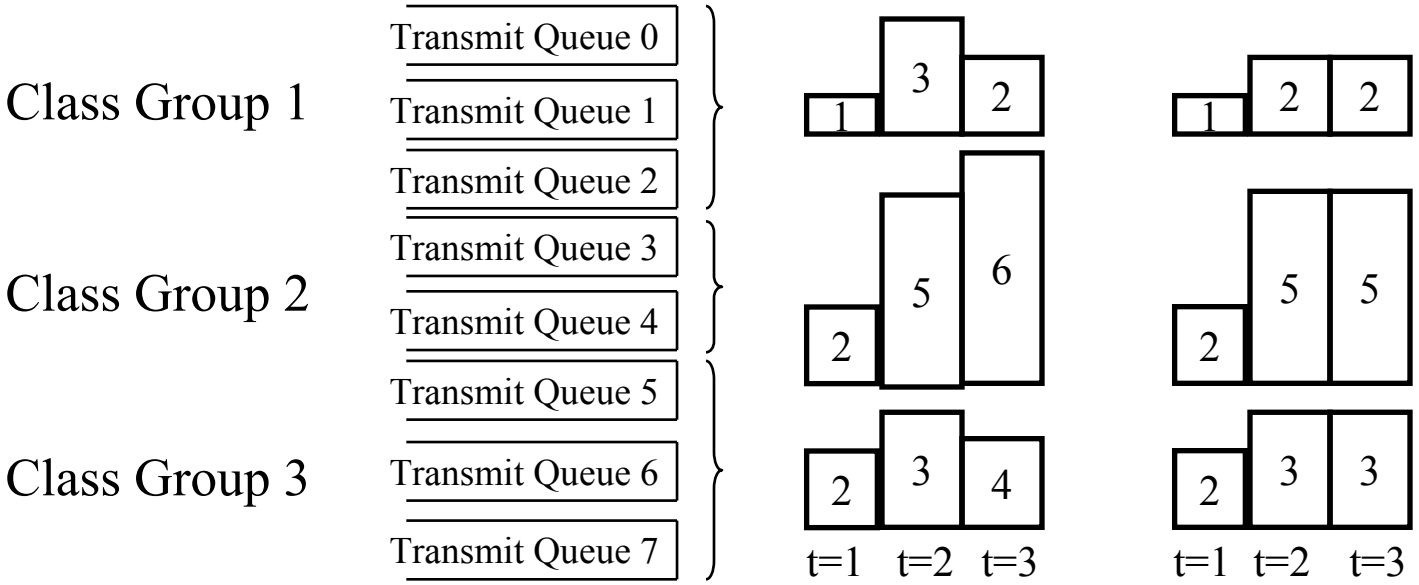
Ref: J. L. White, "Technical Overview of Data Center Networks," SNIA, 2013, http://www.snia.org/sites/default/education/tutorials/2012/fall/networking/JosephWhite_Technical%20Overview%20of%20Data%20Center%20Networks.pdf

Enhanced Transmission Selection

- IEEE 802.1Qaz-2011
- Goal: Guarantee bandwidth for applications sharing a link
- Traffic is divided in to 8 classes (not priorities)
- The classes are grouped

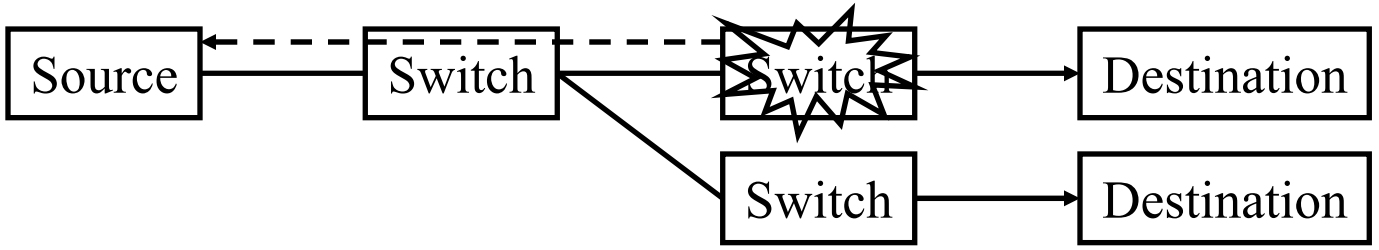


ETS (Cont)



- Fairness within a group
- 3-8 classes.
  At least 3: 1 with PFC, 1 W/O PFC, 1 Strict Priority
- Bandwidth allocated per class group
- Bandwidth unused by a class group is consumed by others
- Example: Group 3=3, Group 2=5

Quantized Congestion Notification (QCN)



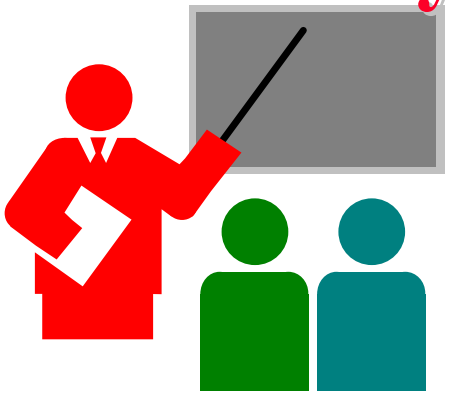
- IEEE 802.1Qau-2010 Dynamic Congestion Notification
- A source quench message is sent by the congested switch direct to the source. The source reduces its rate for that flow.
- Sources need to keep per-flow states and control mechanisms
- Easy for switch manufacturers but complex for hosts.
  Implemented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

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

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

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 the packet loss by enhanced transmission selection and Priority-based flow control

List of Acronyms

- BPDU Bridge Protocol Data Unit
- CFI Canonical Format Indicator
- CSMA Carrier Sense Multiple Access with Collision Detection
- 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
- ID Identifier
- IS-IS Intermediate System to Intermediate System
- iSCSI Internet Small Computer System Interface

List of Acronyms (Cont)

- 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
- MSTP Multiple Spanning Tree
- OAM Operations, Administration, and Management
- OSPF Open Shortest Path First
- 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
- SPB Shortest Path Bridging

List of Acronyms (Cont)

- STP Spanning Tree Protocol
- TLV Type-Length-Value
- TPI Tag Protocol Identifier
- VLAN Virtual Local Area Network
- VM Virtual machine
- VOIP Voice over IP
- WAN Wide Area Network

Reading List

- G. Santana, "Data Center Virtualization Fundamentals," Cisco Press, 2014, ISBN:1587143240
- Enterasys, "Enterasys Design Center Networking - Connectivity and Topology Design Guide," 2013, http://www.enterasys.com/company/literature/datacenter-design-guide-wp.pdf
- Cisco, "Understanding Spanning-Tree Protocol Topology Changes," http://www.cisco.com/en/US/tech/tk389/tk621/technologies_tech_note09186a0080094797.shtml
- Cisco, Understanding Rapid Spanning Tree Protocol (802.1w), http://www.cisco.com/en/US/tech/tk389/tk621/technologies_white_paper09186a0080094cfa.shtml
- Canonical vs. MSB Addresses, http://support.lexmark.com/index?page=3Dcontent&id=3DHO1299&locale=3Den&userlocale=3DEN_US
- Extreme Networks, "Link Layer Discovery Protocol (LLDP)," http://www.extremenetworks.com/libraries/products/LLDP_TB.pdf

Reading List (Cont)

- M. Srinivasan, "Tutorial on LLDP," http://www.eetimes.com/document.asp?doc_id=3D1272069
- P. Thaler, et al, "IEEE 802.1Q" IETF 86 tutorial, March 10, 2013, http://www.ietf.org/meeting/86/tutorials/86-IEEE-8021-Thaler.pdf
- 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, http://www.snia.org/sites/default/education/tutorials/2012/fall/networking/JosephWhite_Technical%20Overview%20of%20Data%20Center%20Networks.pdf
- I. Pepelnjak, "DCB Congestion Notification (802.1Qau)," http://blog.ipspace.net/2010/11/data-center-bridging-dcb-congestion.html

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

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

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