Washington University in St.Louis

Engineering

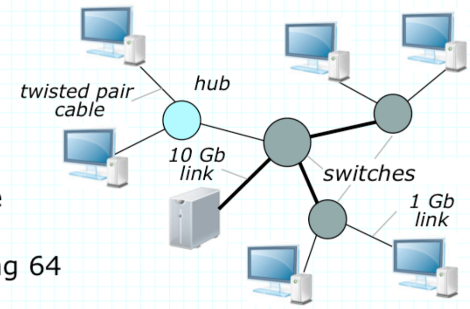
16. Switched Local Area Networks

- Routing in switched ethernet LANs
- Spanning tree algorithm
- Virtual LANs
- Layer 3 switching
- Datacenter networks

Jon Turner

Washington University in St. Louis | Engineering

Switched Ethernet LAN

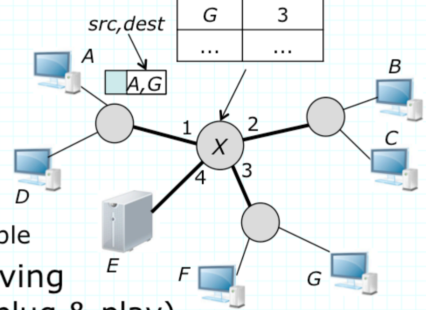


- Modern LANs are mostly switched
  - » even 1 Gb/s switches now quite inexpensive ($10/port)
  - » 10 Gb/s becoming reasonable (<$200/port)
  - » single chip switches supporting 64 10G ports now available
- Features extend far beyond classical Ethernet
  - » VLAN support
  - » priority-based queuing
  - » variety of router features
- Applications extend beyond traditional LAN
  - » used in WANs to configure router links
  - » used in data centers to connect many thousands of servers

2

Switched Ethernet Routing

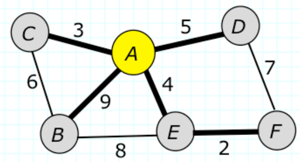
dest	output
A	1
G	3
…	…



- Links define a tree, so only one path between two hosts
  - » still, switches need to know on which link to forward packets
  - » switch routing tables map destination address to output
    - *flat* addresses: use CAM or hash table
- Switches learn routes by observing links on which packets arrive (plug & play)
  - » insert entry (*source address*, *arrival port*)
    - entries timeout eventually, to allow for moving hosts
  - » packets to "unknown" destinations are forwarded to *all* outputs
    - hosts discard if destination address does not match
  - » example: so, first packet from *A* to *G* adds entry (*A*,1) to switch *X* (and similar entries in other switches); reply from *G* to *A* is sent on direct path adding entry (*G*,3) to *X*

Washington University in St. Louis — Engineering

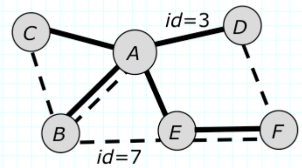
Spanning Tree Algorithm



- Switch links may form cycles
  - » either by accident or to provide protection against failures
  - » but routing depends on absence of cycles
  - » switches select a *spanning subtree* of network
- Basic approach
  - » select a *root switch* (based on smallest id)
  - » determine path lengths to root (link bandwidth determines cost)
  - » select switch port on shortest path to root as *root port*
    - peer of a root port is a *child port* – other ports get turned off
- Distributed algorithm (simplified)
  - » exchange configuration packets containing id of sending node, id of root (or best root candidate) and distance to root
  - » root node originates config packets, others propagate
  - » nodes act like root until they discover better candidate

4

Virtual LANs (VLAN)

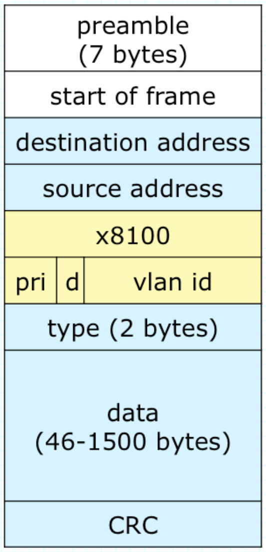


- Allows hosts to be divided among different VLANs
  - » Ethernet packets do not propagate beyond VLAN boundaries
  - » to go between VLANs, packet must pass through a router
    - but many switches support router-like functions that handle this
    - VLANs often correspond to IP subnets, but need not
- VLANs can increase network's traffic capacity
- Packet's VLAN identified by VLAN id carried in packet
  - » 32 bit VLAN "tag" inserted just before the "ethertype" field
  - » packets with VLAN id *X* are sent only on ports that belong to *X*
    - "host ports" typically belong to one VLAN
    - VLAN tag is typically added/removed by switches at "host ports"

5

Washington University in St. Louis — Engineering

Ethernet Frame With VLAN Tag

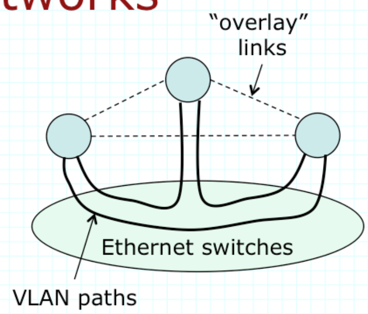


- *Tag* starts with two byte value x8100
  - » takes place of type field, allowing packet to be identified as tagged packet
- *Priority* field (3 bits)
  - » 0 for best-effort, 7 for highest priority
- *Drop Eligible Bit*
  - » indicates packet that can be preferentially discarded during congestion
- *VLAN Identifier* (12 bits)
  - » value of 0 means "no vlan"
- *Double tagging*
  - » a vlan tag that starts with 0x9100, identifies the first in a pair of tags
  - » allows ISPs to use VLAN tags while carrying "customer-tagged" packets

6

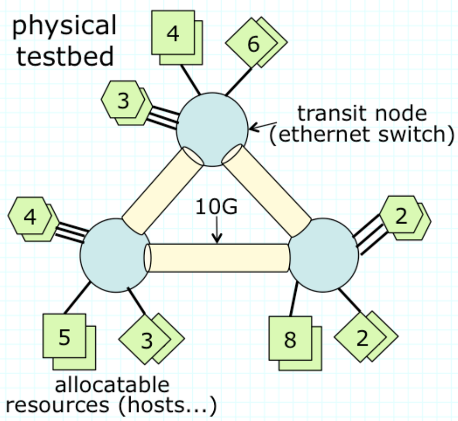
VLANs and Overlay Networks

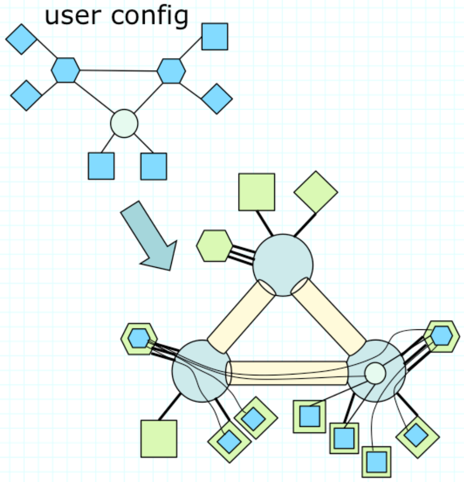
- VLANs can be used to implement virtual links joining routers
  - » configured by network managers
  - » can be "provisioned" to provide guaranteed bandwidth
  - » support for "private WANs"
- Routers treat these much like physical links
  - » link rates can be configured so several overlay links can share physical links – no constraint on individual link rates
    - and several overlay links can share a single router port
  - » makes it easy to add new links between routers, as needed
  - » overlay link rates can be changed in response to traffic
    - may require re-routing of some overlay links
  - » effectively replaces router ports with cheaper switch ports



7

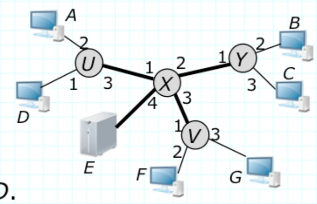
VLANS in ONL



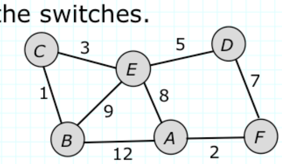


- ONL maps user configurations onto physical testbed using “hidden” ethernet switches
  - » point-to-point user links map to point-to-point vlans
  - » ethernet switches in user config also map to vlans

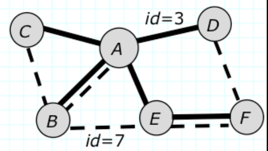
Exercises



1. The diagram at right represents a switched Ethernet network. Suppose that initially, all the switch routing tables are empty. Now, suppose *B* sends a packet to *E* and *G* sends a packet to *D*. Show the final contents of the routing tables at all the switches.



2. Assume that the diagram at right represents a switched Ethernet network, and that the letters represent the switch ids, with *A* being the smallest. What links would be included in the spanning tree?



3. In the diagram at right, suppose that host *d* is on vlan 7 at switch *D*, host *f* is on vlan 3 at switch *F* and router *c* is on switch *C* and has connections to both VLANs. What sequence of links is used by a packet going from *d* to *f* assuming no other routers?
4. In the diagram from problem 2, assume the links are 10 Gb/s and show how we could configure 3 "virtual links" between hosts on *C* and *F* with capacities of 6 Gb/s, 3 Gb/s and 5 Gb/s. Identify the ports that are used by each VLAN at switches *A* and *E.*

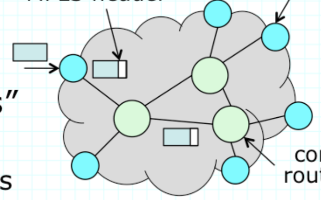
Washington University in St. Louis — Engineering

Newer Developments

- Faster response to topology changes
  - » original protocol can take nearly a minute to converge on new spanning tree after a link fails
  - » Rapid Spanning Tree Protocol cuts time to under 10 seconds
- Computing multiple spanning trees
  - » when VLANs were first introduced, all VLANs used the same spanning tree
    - manual configuration required to use all available links
  - » Multiple Spanning Tree Protocol allows automatic configuration of multiple *subtrees* (that is, may restrict a tree to a *region*)
    - VLANs are mapped to trees (so several VLANs may share a tree)
- Shortest Path Bridging (standardized in 2012)
  - » link-state protocol (like OSPF, but different)
    - switches distribute topology information, compute shortest-path-trees and configure local routing tables

10

MPLS



- Multiprotocol Label Switching extends IP to provide "virtual links" within IP networks
  - » MPLS-only switches are potentially less expensive than routers
  - » MPLS features often added to standard routers
  - » allows finer-grained management of traffic than IP routing alone
- MPLS headers added by "edge routers" (before IP header)
- Core routers switch packets using "labels" in MPLS headers
  - » labels used to select entries in MPLS routing tables
  - » packets may contain multiple "stacked" headers
  - » routing table entries can be configured to select output based on label, replace label value with another, push/pop headers

11

Washington University in St. Louis — Engineering

Layer 3 Switching

- Many switches have extensive support for IP routing
  - » routing features first added to connect subnets in different VLANs
  - » feature sets have expanded as way to "add value" to products
- Example features
  - » IP forwarding, ARP, ICMP, DHCP, RIP, ...
  - » Diffserv QoS with 8 queues per link
  - » IGMP support (snooping and querier functions)
  - » Access Control lists (firewall functions)
- Getting harder to distinguish routers and switches
  - » routers support different kinds of layer 2 links (not just Ethernet) and support multiple L3 protocols
  - » routers have more extensive feature sets, more configurable
  - » routers have larger routing tables & buffers, flexible queueing
  - » switches generally far less expensive

Washington University in St. Louis — Engineering

Data Center Networks

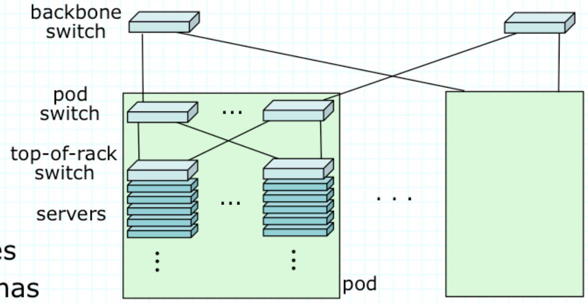
- 10's to 100's of thousands of hosts, often closely coupled, in close proximity:
  - » e-business (e.g. Amazon)
  - » content-servers (e.g., YouTube, Akamai, Apple, Microsoft)
  - » search engines, data mining (e.g., Google)
- Challenges
  - » multiple applications, each serving many clients
  - » managing/balancing load
  - » multiple "tenants"
    - must keep tenants isolated
    - actions of one tenant must not interfere with another

Inside a 40-ft Microsoft container, Chicago data center

13

Scaling Up

- Example
  - » Each rack has
    - 32 servers each with 32 cores
    - TOR switch with 1G and/or 10G links
  - » 32 rack pod has 1024 servers, and 32K cores
  - » 64 pod configuration has 32K servers, 1M cores



- Networking challenges
  - » addressing – too many L2 addresses for typical switches
  - » must balance load across many paths
    - Ethernet routing too restrictive (even with advanced routing features)
    - basic options: per packet load-balancing, flow-based load balancing
  - » requires new class of *data center switches*

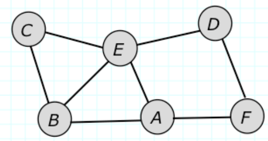
Washington University in St. Louis | Engineering

Blurring the L2/L3 Boundary

- Market forces are pushing Ethernet beyond the LAN
  - » traditionally, large volume of Ethernet switch market has kept prices low
  - » smaller sales volume for routers kept prices high
- Technology improvements support greater capability
  - » early switches were simple and cheap (no VLANs, a few thousand routing table entries)
  - » modern switches can have >100K routing table entries and support thousands of VLANs, extensive IP features, ...
  - » market expectations have kept prices moderate, even as feature sets have ballooned
- Not clear how successful some advanced features will be
  - » big attraction of Ethernet is simple operation, but advanced features compromise simplicity
  - » challenge for new switches to interoperate with "legacy" switches

15

Exercises



1. The diagram at right represents a *core network* for some ISP. Assume all the nodes are MPLS switches and that each connects to one or more *edge routers.* Describe how MPLS can be used to distribute traffic between switches *C* and *F* to use two different paths. Show MPLS routing table entries for all the switches along these paths, using different labels on each hop. Can you spread the load like this if the nodes were all conventional routers, using OSPF-routing?
2. Estimate the number of "end-user hosts" in the world (laptops, smart phones, tablets,...). Estimate the number of servers Google needs to respond to search requests from these users (think first about the number of search requests each end-user host generates). It's been reported that Google has close to a million servers. Are your estimates consistent with this?
3. In the example data center network on page 12, how many distinct Ethernet addresses are needed, assuming each server has one 10G interface? How many IP addresses are needed, assuming we want to support a "virtual host" on each core? How big must the the Ethernet switch tables be to support this network? How might this change if the switches were L3 switches and routed using IP?
4. Discuss the pros and cons of switches with advanced L3 features. Do you think it would be better to maintain strict separation between L2 and L3 functions? Is the whole concept of protocol layering inevitably undermined by market forces?