Elements of Network Design

Jim Binkley

jrb@cs.pdx.edu

http://www.cs.pdx.edu/~jrb/netmgmt.html

Or everything you did NOT want to know about ethernet plumbing

aka layer-2 explosion

mostly LAN discussion

network design outline

- ethernet, past present and future
- building blocks:
  - hubs
  - bridges (spanning trees, adaptive learning),
  - ethernet switches
- vlans
- in summary
  - at least 3 povs
  - parting shots (promiscuous mode/QOS)!?

ethernet in 3 stages (4 soon)

- 10BASE - CSMA/CD - a bunch of hosts on a broadcast segment - 1982-on
  - collisions happen, shared link, bridges for unicast segmentation
- 100BASE - CSMA/CD BUT we have a star network and full-duplex - 92-4
  - full duplex (autonegotiation)-> collision-free/segmentation
- 1000BASE - death of CSMA/CD, Y2K
  - likely collision-free, star or pt. to pt.
- 10000BASE - on the way

3 kinds of Enet/MAC physical address

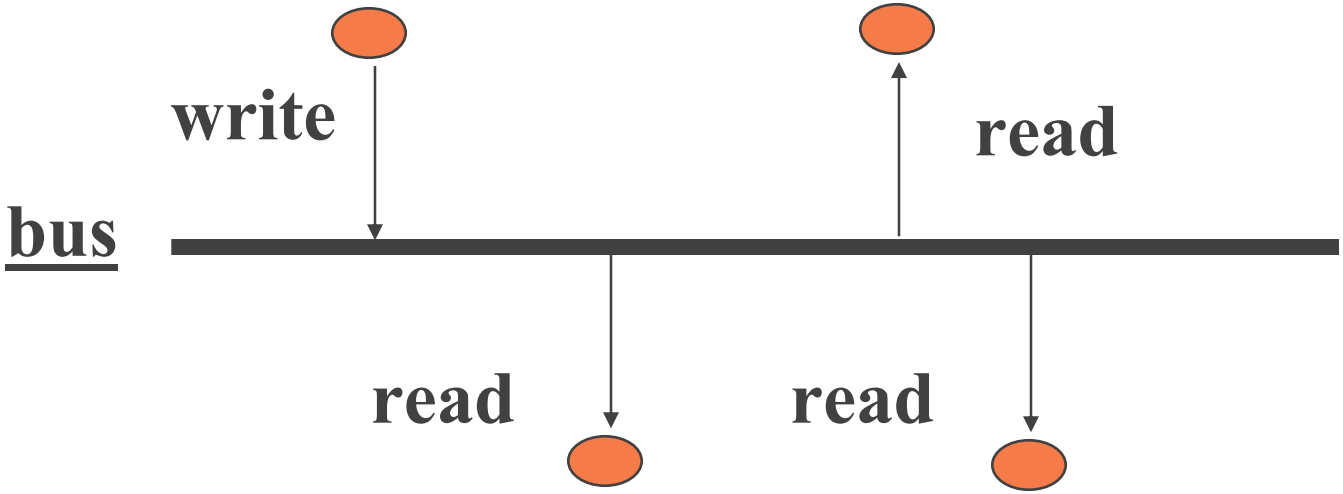
- **unicast** - physical address of controller
- **broadcast**: *ff:ff:ff:ff:ff:ff*
- **multicast**: *01:xx:xx:xx:xx:xx*
- IP multicast range: *[01:00:5E:00:00:00..01:00:5E:7f:ff:ff]*
- ip-enet mapping not 1-1, 32 ip addr to 1 enet/ip multicast address

10BASE Enet - properties

- original form: 10 mbps (10,000,000 bits)
  - (1.25 Mbytes per sec) (Mb rant here)
- broadcast bus
- distributed access control; i.e., no central "master" saying you may or may not
- hw gets every packet, may filter out
- CSMA/CD - carrier sense multiple access with collision detection

ethernet can broadcast (all)

where bcast goes == **broadcast domain** (think ARP)



1 write - many reads in parallel

fundamental broadcast idea/s

- includes one to one
- **broadcast** means 1 to all stations
- **multicast** means 1 to many, includes 1-1, 1-all (broadcast is subset of multicast), 1 - N, N < all
- Examples include ethernet, token-ring, radio
  - not pt. to pt. telephone links like ATM, ISDN
- questions include: can it do CSMA, CD?
- Collision means backoff and retry
  - and dead packets or packet shrapnel, CRC failures

collision detection/retransmission

- if collision, must send jam signal, random backoff and retransmit
- backoff is “binary exponential algorithm”
- wait 1, 2, 4, 8 time-slots, etc * a random delay, delay max 1023, 16 retransmits on collision max
- packets can be lost due to collision, especially if network is heavily used (note: **full-duplex** idea)
- modern network cards can saturate 10/100 link;
- best **utilization** put at %30 (over elapsed time) on 10BASE shared link

broadcast network attributes #1

- **broadcast domain** - “segment” over which broadcasts are forwarded and heard
  - with 10BASE/80s tech., this was a physical idea, now it is logical with multi-switch VLAN
- **collision domain** - “segment” over which collisions can occur
- have to ask ourselves what these mean in terms of routers/switches/VLANs/bridges/hubs/full-duplex?
- **broadcast isolation** - broadcast CANNOT cross there, and cannot meltdown network

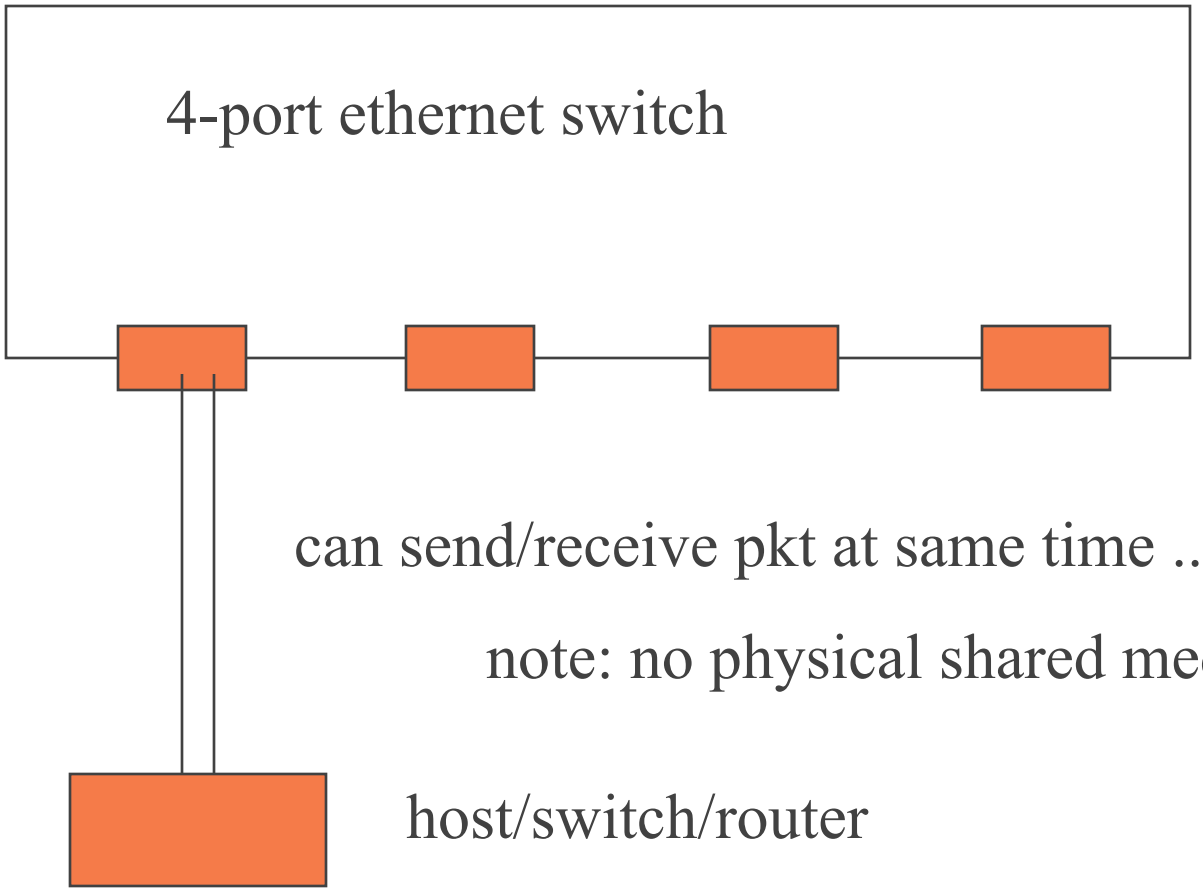
broadcast network attributes #2

- **segmentation**: typically meaning isolating hosts to a ratio of less hosts per collision domain (unicast mostly, broadcast too)
  - ideally: 1 host on 1 switched ethernet port
  - design goal: minimize collisions (none is good)
- **cut-thru** versus **store and forward**
  - meaning switch may try to trade-off fast forwarding of packets and lose **collision isolation (ethernet CRC verification)**

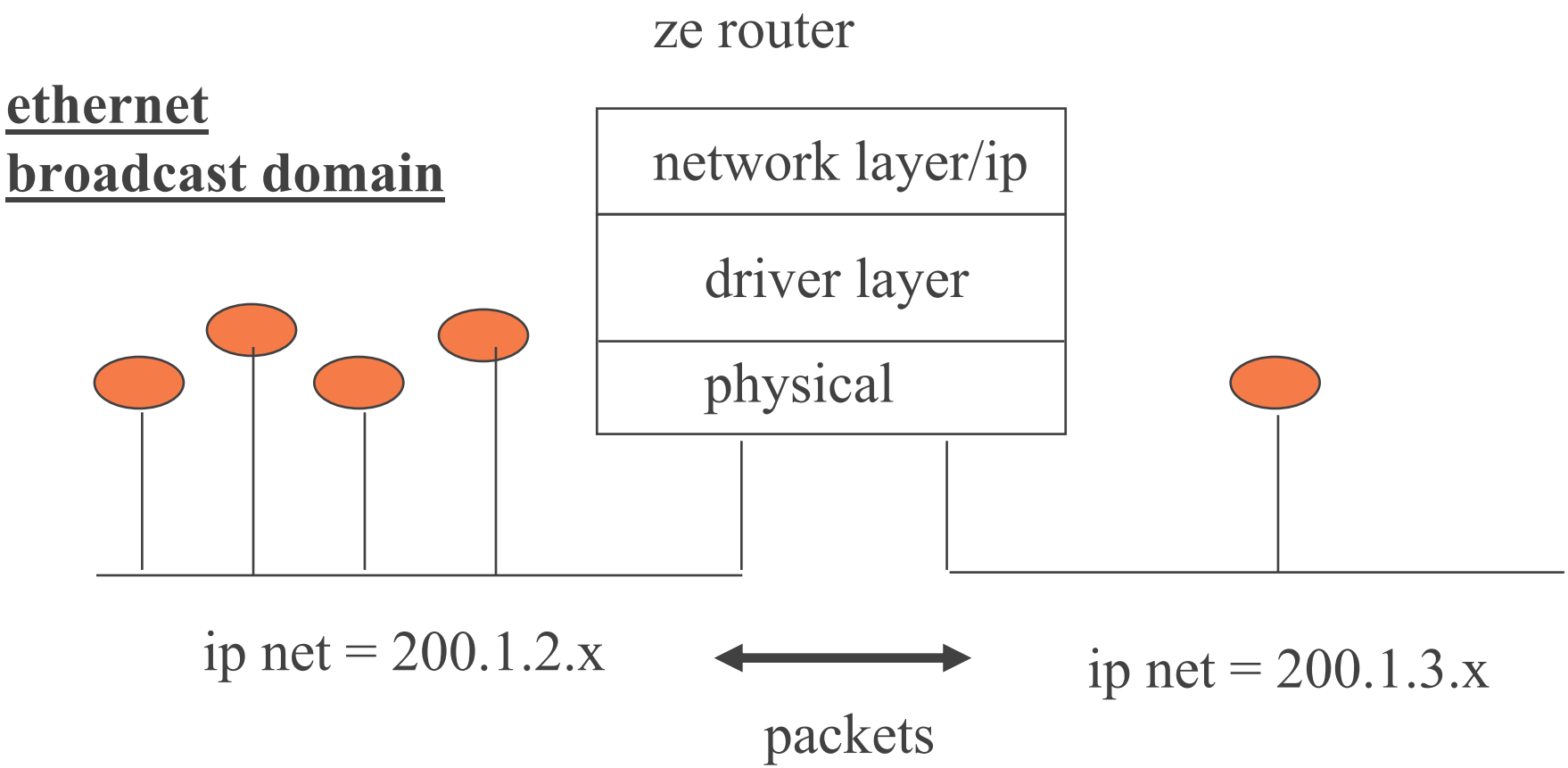
full-duplex (assume 100/1000)

- feature introduced mostly about the time of 100BASE (found on 10BASE though)
- **full-duplex** vs **half-duplex**
- full-duplex, only two hosts on physical wire
- both can send in parallel; i.e.,
- **COLLISION-FREE**
- **100BASE likely**
- **1000BASE requirement**
- **no CSMA/CD likely with 1000BASE**

full-duplex


4-port ethernet switch
can send/receive pkt at same time ....
note: no physical shared media here
host/switch/router

IP/subnet and 10BASE network



10BASE network

- segmentation came from adding a router
- still might have MANY hosts on one wire
- which was ok when they were slow
- now they can destroy each other
  - TCP can use up 10/100BASE for a web page
  - too many collisions
- you might also want to extend the network for reasons of convenience

consider these boxes

- repeater/hub (repeaters are rare), L1
- bridge (classic), L2
  - mixed-media, or same-media (all ethernet)
- switch (ethernet), L2 (L2/L3 possible)
- router, L3
- and consider them in terms of previous stated attributes
- and new forms of ethernet (100/1000)

hub

- mostly but not entirely operates at physical layer
- extends broadcast domain and segment size
- may or may not extend collision domain
  - if limits collision domain, done by store and forward
- hence gives weak form of segmentation
  - suppress collisions, no unicast segmentation
- does NOT enable more throughput
- should be MANAGED hub (speaks snmp)
  - collects ethernet error statistics (see SNMP dot one MIB)
- does not understand network layer (how does it ping?)

timeout for question ...

- you own a managed hub (80211 AP too)
- it speaks SNMP
- it has a default route
- it has a static IP address
- layer 1 device with layer 7 application (SNMP)
- assume no routing table, how can it be pingable, implementation-wise?

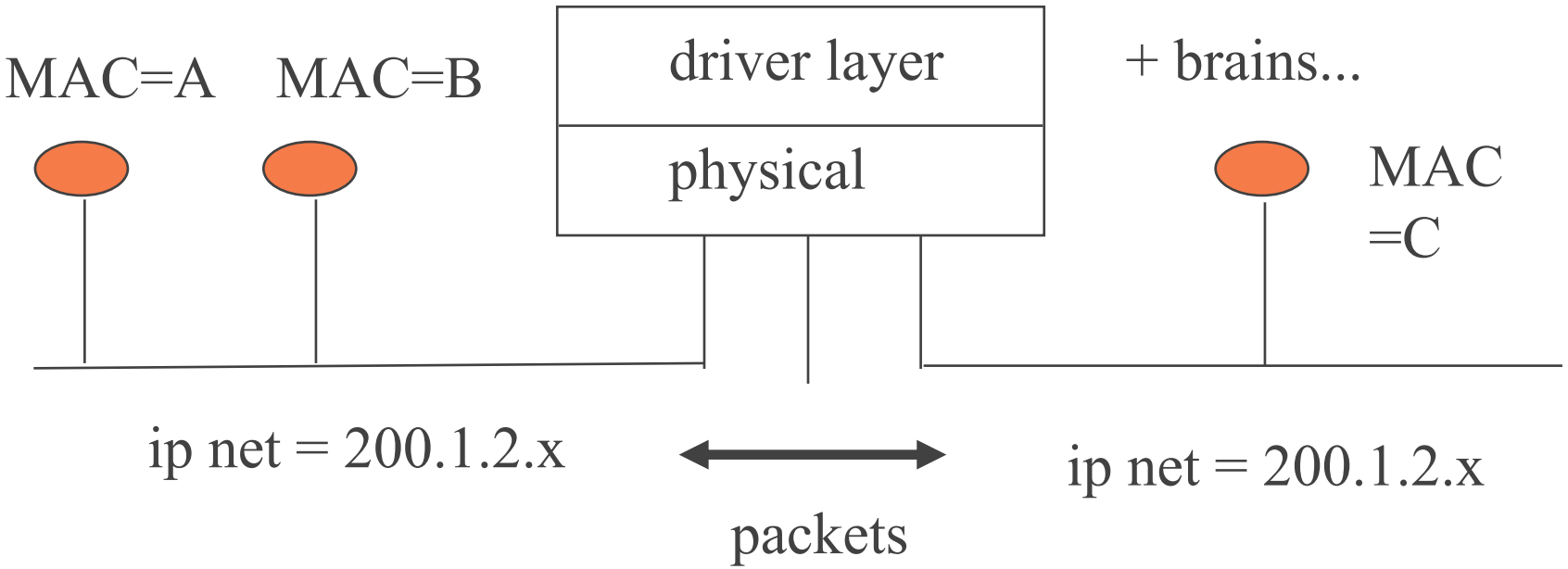
introducing the bridge

- **more than a hub; less than a router**
- **learning/adaptive bridge**
  - allows SOME (unicast) segmentation as can learn and not forward across itself
- ethernet bridge vs mixed-media bridge
  - nevermind mixed-media bridges - BAD IDEA
    - » ATM in, ethernet out (put in a router)
- bridges flood by definition and learn to optimize; hence give **unicast segmentation**

bridge (adaptive/learning)

src A to dst B learns to not forward
src A to dst C must always forward

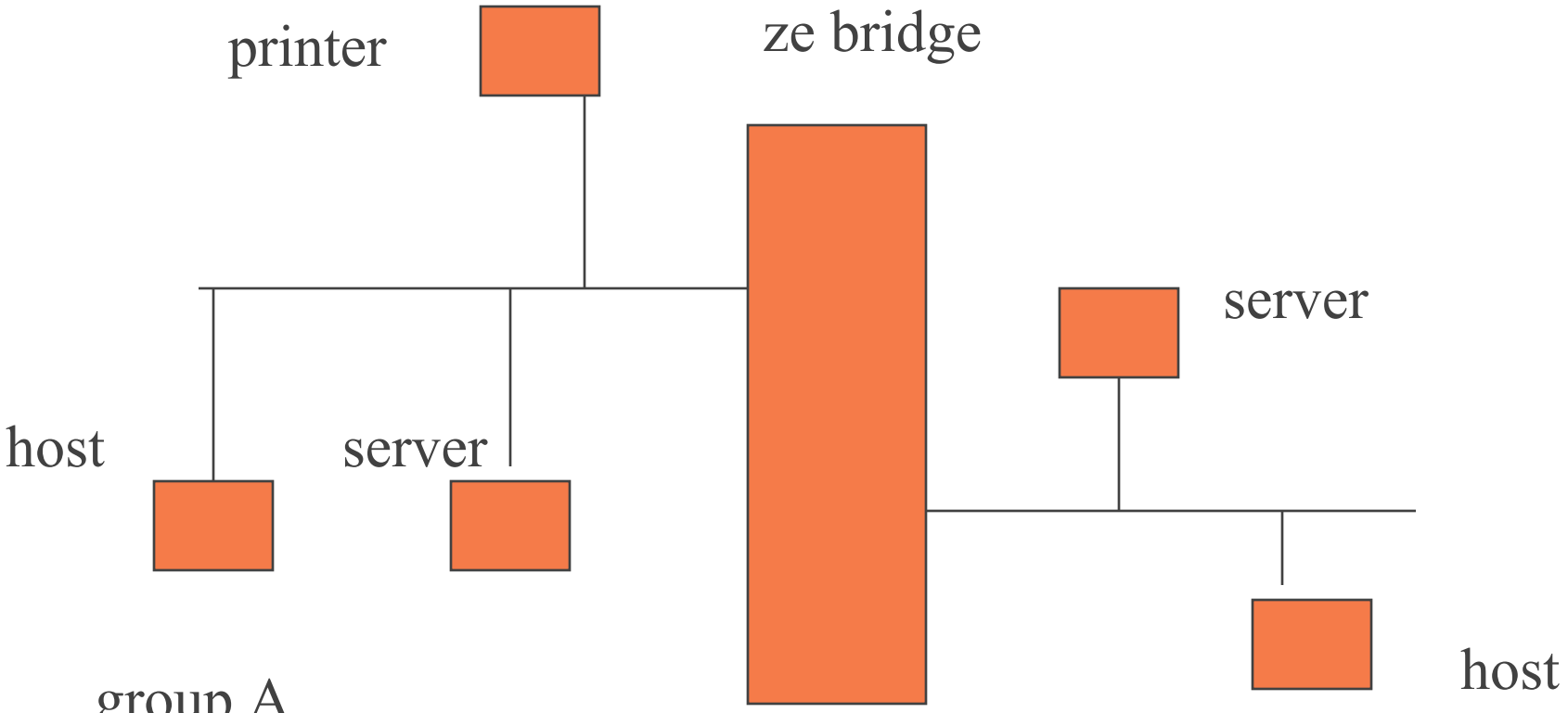
link layer



must still

- flood broadcast/multicast
- there exist ways to optimize multicast flooding
- note that unicast still leaks ...
  - flood when can't map dst to a port
- broadcast domain still on all sides
- collision domain MAY/may not be limited
- some segmentation/but not per-host
  - might put local server/local host on same side of segment

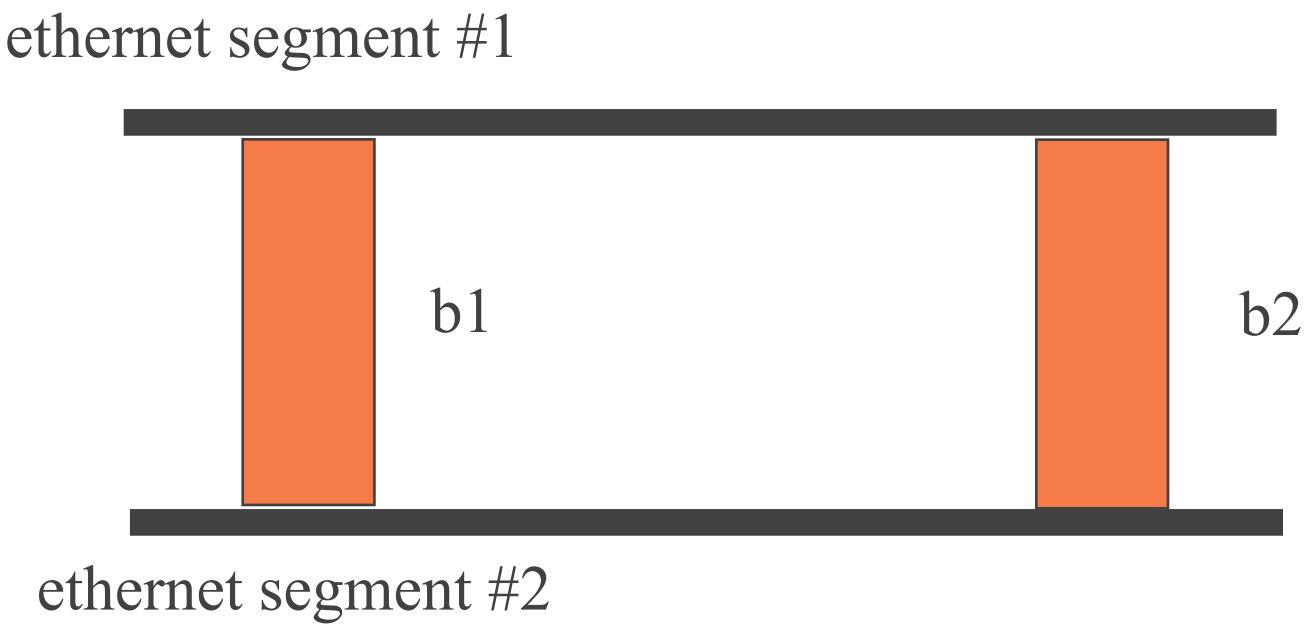
traditional bridge segmentation scheme (unicast, not broadcast)



can you centralize the servers?

Jim Binkley

what happens if a host broadcasts? 2 segs/2bridges



assume 2 bridges hook 2 ethernet segments together to make 1 big segment.

no problem, right?

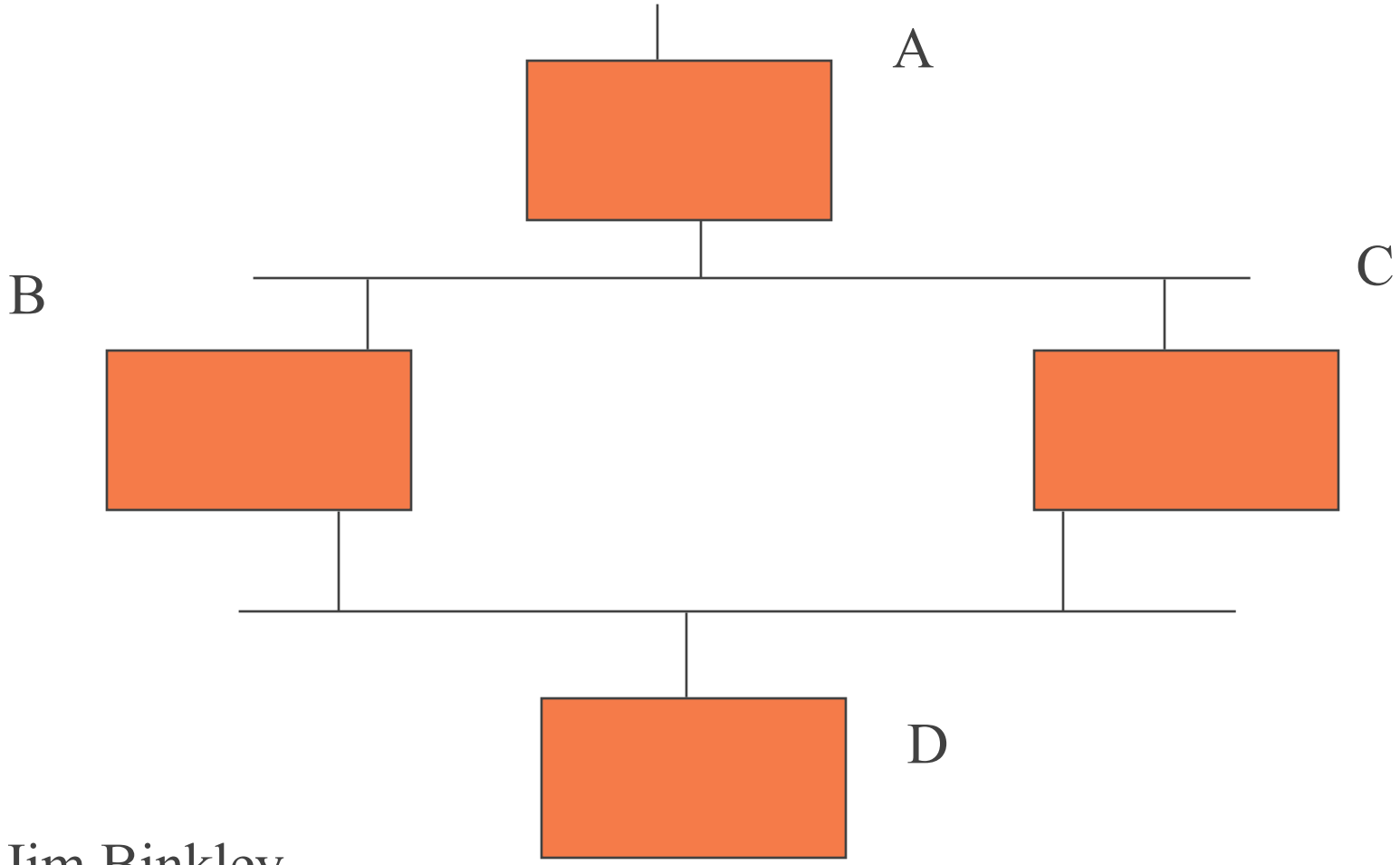
not a good thing

- 1 broadcast may cause network to
- **meltdown**

802.1d – spanning tree

- see Stallings, Local and Metropolitan Area Networks, for more info
- IEEE 802 standard (802.1D)
- bridge protocol at link layer
- bridges form rooted spanning tree, **no cycles**
  - **aka no loops**
- ports ultimately in {forwarding, blocked} state
  - on or off
- done with simple L2 flooding protocol

4 bridges, what happens?



operation

- each bridge has ID based on 1 mac address
- each port has MAC address (port ID)
- root bridge is top of tree
- root chosen by Spanning Tree Algorithm
- (low) path cost may be associated with bridges by manager in order to influence choice
- may also set **PRIORITY** to influence root

more operation

- **designated bridge/port**, bridge on LAN that is chosen to forward packets to/from another lan
- **root port** - each bridge discovers 1st hop on minimum-cost path to root bridge. if two ports on a LAN, then use lower port number.

basic idea:

- 1. determine root bridge
- 2. determine root port on other bridges
- 3. determine designated port on each LAN
- consequence: if two bridges connect same two LANs, one is left out
- timers used so that if designated port fails, another may be chosen; i.e.,
- **at boot, or at change, STA recalculated**

BPDU/s

- BPDU - bridge protocol data unit
- sent out on all ports to ALL BRIDGES multicast group address
- in general, BPDU from one bridge flooded out the other ports, and used in both
- send whilst maintaining (periodic resend) or
- rooted tree STA recalculation

BPDU cont.

- 2 packet types, config, topological (start over)
- configuration BPDU is 35 bytes, root resends at hello time interval, hello time default is 2 seconds (root sends)
- root id field in BPDU (5 bytes in), 8 bytes
  - 2 bytes of root priority, 6 bytes of MAC
- config sent during STA, stable state, election time
- topo packet only 4 bytes
- topological change sent when bridge believes configuration change occurred, therefore redo STA
- stable state: root issues configuration/everybody else forwards it

BPDU encapsulation

dst - group multicast address
src - unique per port
SAP/s - 01000010 (palindrome)

tcpdump -e -n -i <ifname>

11:32:41.457906 0:a0:c9:47:cb:21 > 1:80:c2:0:0:0 **802.1d**
ui/C len=43
0000 0000 0080 0000 a0c9 47cb 2000 0000
0080 0000 a0c9 47cb 2080 0200 0014 0002
000f 0000 0000 0000 0000 00

note: mac dst is 1:80:c2:0.0.0 - ALL-BRIDGES mcast
note 0:a0:c9:47:cb:20 in data portion, part of root ID

better:
tcpdump –vvv –e –i xl1

- src mac 1:80:c2:0:0:0 0026 64: 802.1d config 8000.00:d0:58:3a:9b:42.8019 root 8000.00:d0:58:3a:9b:42 pathcost 0 age 0 max 20 hello 2 fdelay 15

port state machine

- **listening** - STA algorithm used, but bridge does not learn, on timer elapse can become
- **learning** - in addition, bridge can learn, timer elapse can become
- **forwarding** - bridge port root/designated
- **blocking** - bridge learns that this port is not part of ST, therefore blocks port
  - any change puts in listening state
- listening/learning/forwarding on timer elapse done to prevent loops - downside is can be slow

STA operation

- everybody assumes root to start with
- flooding clues them in to who actually has the lowest root ID
- root announces I AM ROOT
- directly connected bridges, send BPDU to say one hop away out other port
- closest bridge becomes path
  - if more than one, smaller bridge MAC wins

election algorithm (cost, priority, MAC):

- if 2 paths to root, we choose the one with the “lowest cost”
- path cost first (choose faster link) then
- choose between priority+MAC “id”
- smallest value wins for that 3 tuple
- is this a good idea or a bad idea in terms
  - of root selection?
- remember Murphy is watching ...

spanning tree algorithm - summary

- 50 seconds or so to resettle network possible upon failure (default is 30, reality can take longer)
- you can “feel it” (net is down)
- pro: **redundancy**, and somewhat idiot-proof
  - function is anti L2 loop after all
- con: ports not in use, downtime is con too
- may wish to use root **priority** to decide who is ROOT, but usually not tweaked
  - set priority LOWER to win
  - make sure implementation supports redundancy

Spanning tree design thoughts

- KISS design - keep from you making a loop and taking a net down
  - do not turn it off unless you know what you are doing (typically on by default)
  - consider wiring morass, especially inter-building on campus
- LARGE scale (e.g., campus-wide) tree probably a BAD idea
  - flaky switch on DMZ could cause 50-sec. outages?
- SPT 1-1 with VLAN, IP subnet (bcast domain)

it's not dead yet, Jim: but

- IEEE 802.1w – Rapid Spanning Tree protocol has been introduced
- goal is to NOT WAIT 60 seconds to reform spanning tree
- not a good thing in an exchange for example
- ironically: OSPF may converge faster than 802.1d

basic idea:

- decouple port state (blocked, forwarding) from role (root, designated port)
- 3 states in RSTP:
  - learning, forwarding, discarding
- 4 port roles in RSTP:
  - root port, designated port, alternate, backup
  - root port – port closest to root bridge
  - designated port – port not root port, that is best port for forwarding pkts (downstream port)

more

- port roles
  - alternate – blocked better better BPDUs come from some other bridge
  - backup – port blocked by better BPDUs from same bridge it is on
  - alternate + blocking more or less == old blocked

BPDU format overview

- type 2, version 2 – therefore older switches will ignore it
- every BPDU issued has port role and state marked in flags
  - therefore recv. can figure out what to do
- BPDUs are sent per port
  - not “flooded” from root anymore
  - must reflect sender’s state

BPDU protocol changes

- BPDU is now hello
  - must hear from neighbor with 6 seconds
  - 3 retries at 2 times per sec.
  - else begin election
  - can be sure problem between you and neighbor
    - » not somewhere between you and root
  - fault is now local, not global
  - this allows faster aging to occur

BPDU protocol changes

- accepting inferior (less good path) information
  - if we hear less good news from the root
  - we believe it immediately
  - e.g., B talks to root and C
  - B loses root, tells C B is root
  - C tells B, nope … I have path to root
  - B believes C

BPDU protocol changes

- fast transition to forwarding state
  - don't need to wait for slow timers due to port info and bridge feedback about convergence
  - 2 new variables: 1. edge ports, 2. link type
  - edge port: if port is connected to workstation, it cannot create a bridging loop
    - » if link toggles does not generate topo change
  - link type: if edge port or full-duplex can make rapid transition, otherwise cannot

feedback mechanism

- an inferior bridge can tell superior to start forwarding
  - and it blocks downstream ports to prevent a loop
- this recursively works to create a loop-free tree
- and make convergence much faster

new topo change mechanism

- in 802.1d when topo change is detected
  - any non-root bridge notifies in direction of root bridge
  - root advertises TC for max-age+forward delay
- in RSTP
  - TC sent by forwarding state change, not edge port
  - very different from 802.1d

topo change in RSTP

- if bridge detects TC
  - 1. starts TC while timer for 2*hello time on non-edge designated and root ports
    - » BPDUs have TC bit set
  - 2. flushes mac addresses associated with those ports
- so any bridge can do this, not just root
- takes a few seconds
- clears MAC forwarding tables (VLAN CAM tables in Cisco speak)

trad. bridge function summary

- adaptive learning - unicast isolation as long as MAC src location can be learned
  - else unicast is flooded
- same broadcast domain on both sides - forward multicast/broadcast
- store and forward, therefore collision detection (based on ethernet CRC)
- spanning tree - prevent link loops

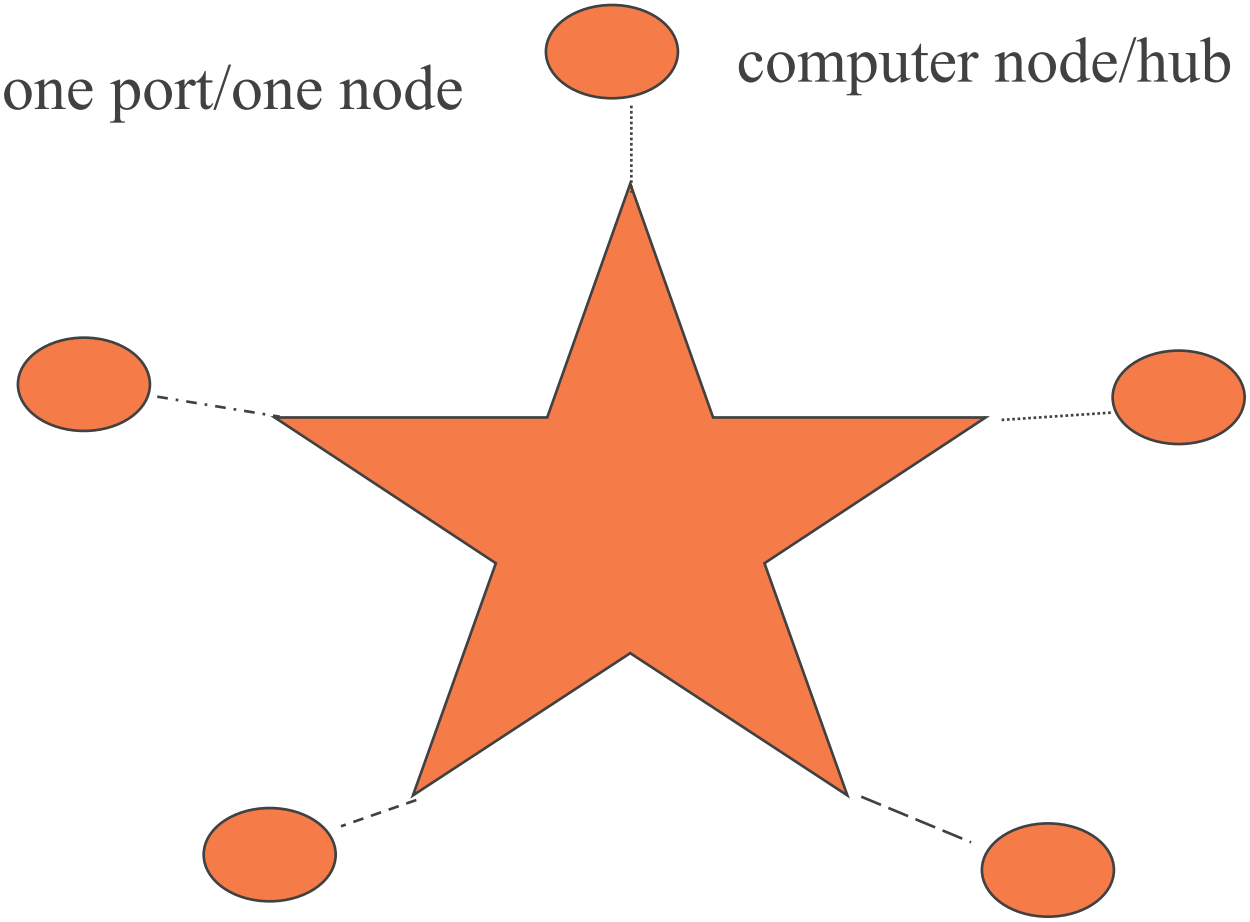
bridges now **switches**

- in a switch, packets forwarded from port A to port B are forwarded in parallel
- in a hub, not so
- 10BASE switches created,
- then 100BASE, then (now) 1000BASE
- traditional shared broadcast link replaced by
- **1 port - 1 host (2 macs per link) switched network is goal (100BASE nics are cheap)**
- **STAR network, with parallel backplane**

bridge as switch

ideal: one port/one node

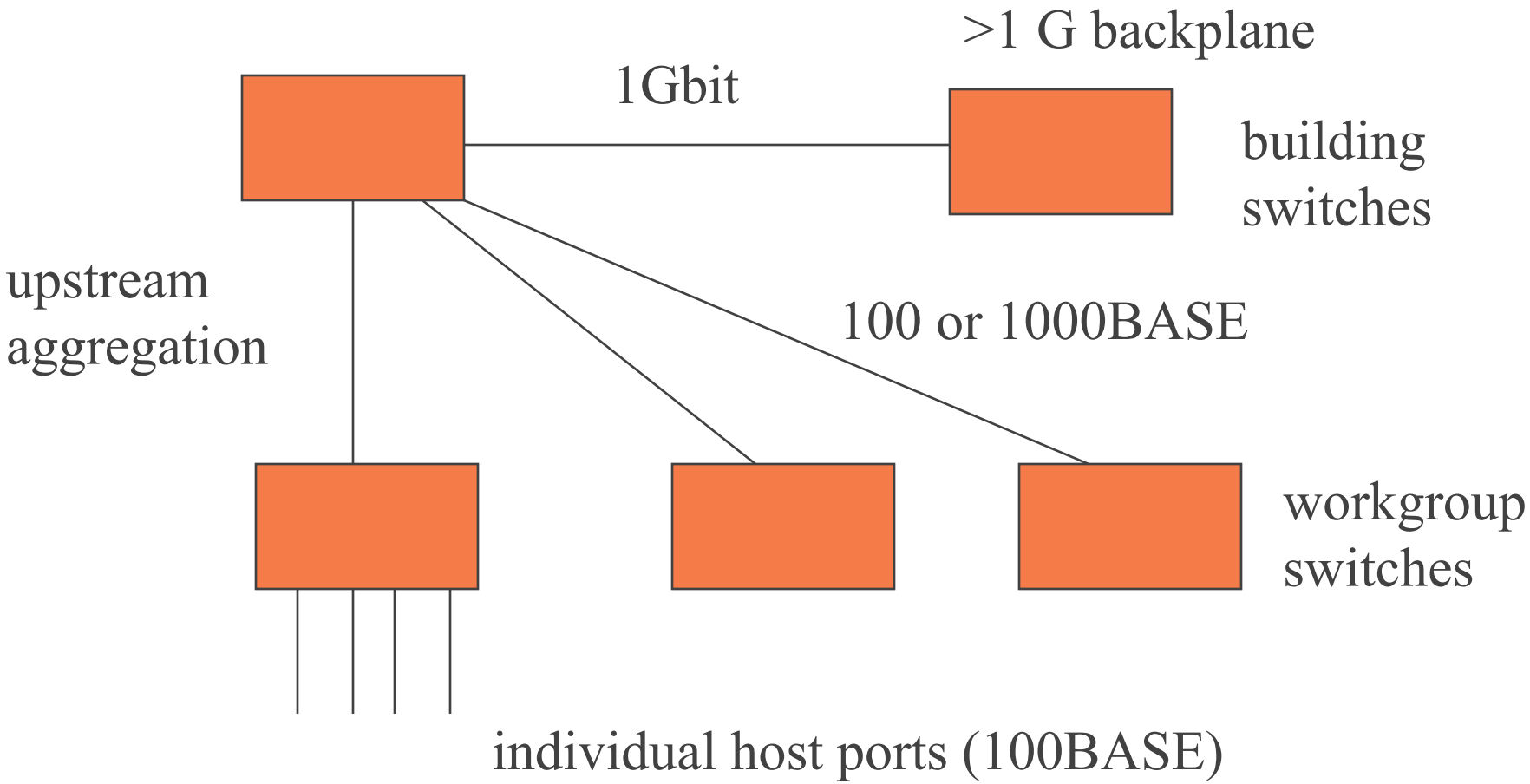
computer node/hub



Jim Binkley

10/100mbit enet: bridge backplane N * 10/100

current complex site net model



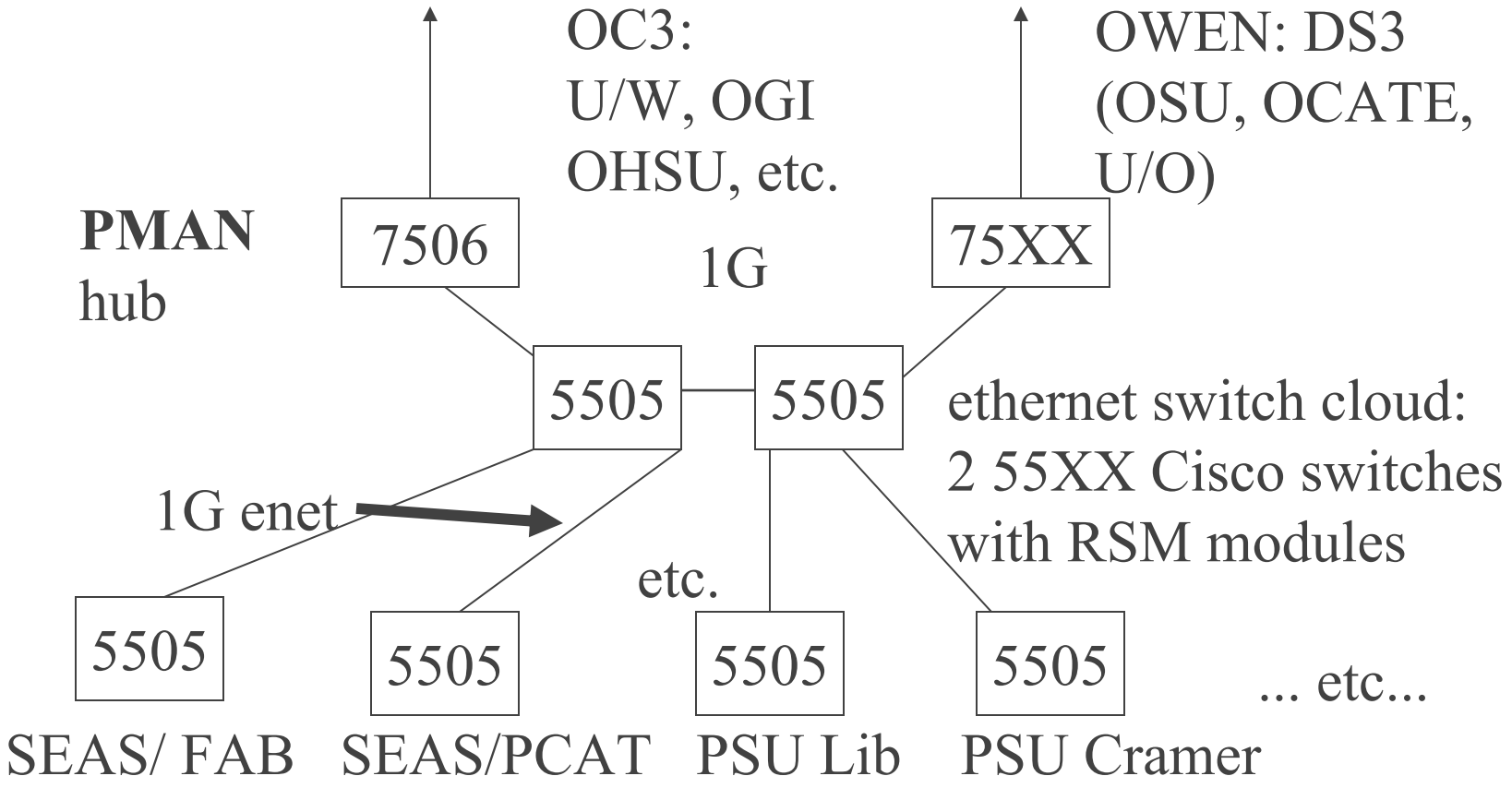
design ideas:

- minimize port/host ratio, but sharing may still occur
  - especially if 10BASE systems
  - but remember collisions can be problem
- server ports should be isolated, not-shared
- may use **ether-channel** (port aggregation), 1 + 1 == 2
- some expensive switches may have port failover
- **big switches offer SNMP manageability centralization over lots of little switches**
- redundancy courtesy of spanning tree is EASY
  - load-sharing may require level-3 thinking (e.g., OSPF multipath)

more design ideas

- you may not always want a spanning-tree
  - if you can't take the outage time;
  - e.g., L2 switched exchange
- full-duplex is important efficiency consideration
  - auto-negotiation can fail however
- upstream switch ports ideally bigger than downstream ports for aggregation

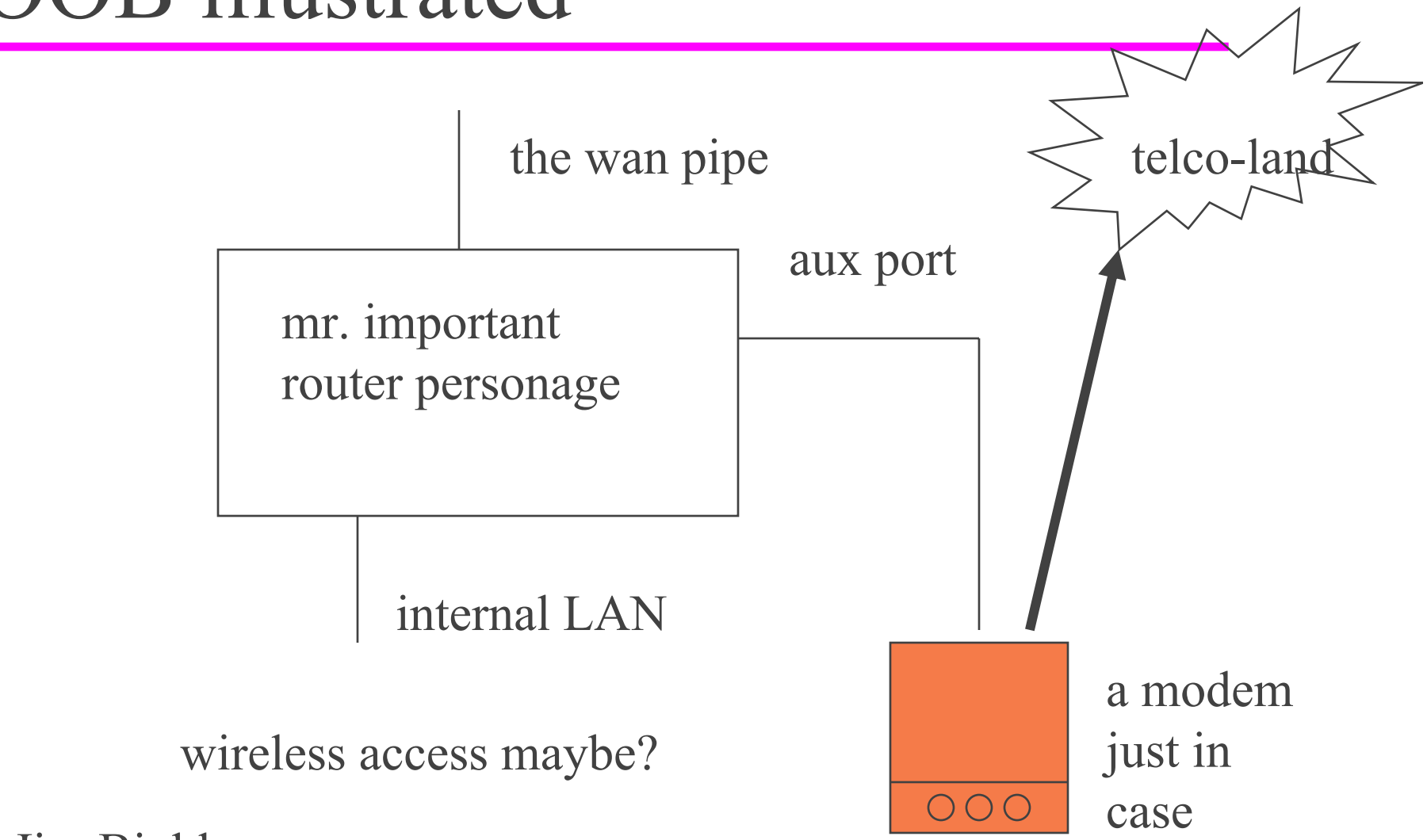
PSU Previous DMZ (L3 POV)



out of band access

- **OOB access is important network-design idea**
- we may want two ways that are completely different to get to an important net component (important security/redundancy idea)
- 2nd access path to router/switch/network console, etc.
- may use aux port for modem/POTS access
  - dialup access to router
- useful if network appears down - who you gonna call?

OOB illustrated



some near-current cisco switches

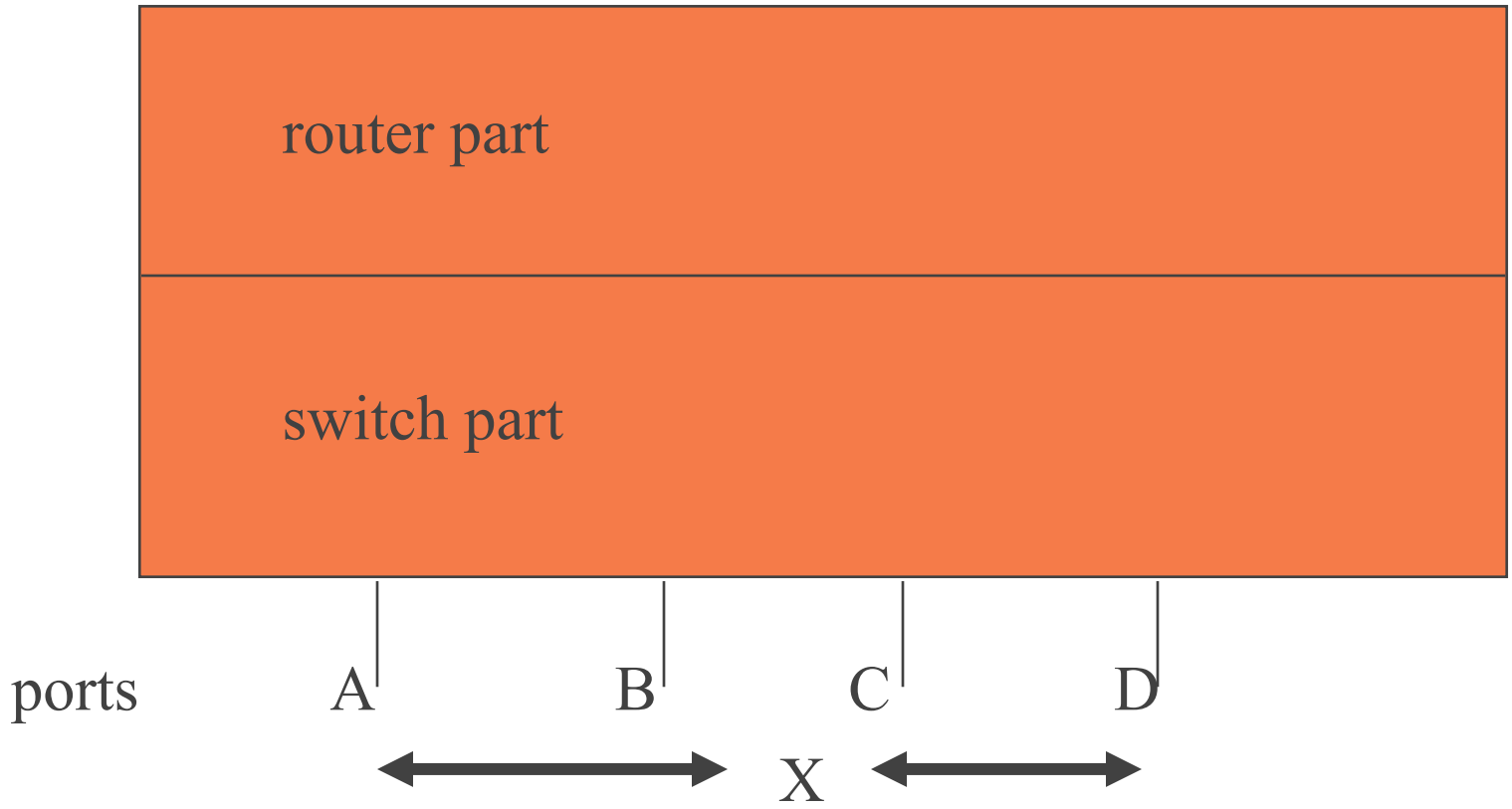
- 5505, 5513 (last 2 digits, # of slots)
  - one card slot used by supervisor
  - might have card with 24 10/100 ports
  - 9 1G ports with so-called GBICS
- 2924 smaller switch (fixed chassis)
  - 24 10/100 base ports
  - add small 2 port module with 100BASE-FX
  - or 1000BASE port for uplink (SX/LX)

VLAN

- VLAN - virtual lan (broadcast group)
- VLAN means we have ability in switch to logically group segments
- VLAN X on port Y/Z, means Y/Z have shared broadcast domain.
  - logical ethernet segment, not necessarily physical
- on router/switch, thus if pkt crosses from VLAN Y to X, then only is routed

VLAN picture - combined router/switch

note: may be two separate boxes or one integrated box



Jim Binkley

vlan X = ports A/D, pkts to B routed

vlans and switches and subnets

- assume IP subnet 1 to 1 with vlan
- logical vlan connectivity MAY exist (IEEE 802.1Q) between switches
- means -- intra and inter switch vlans
- port i, j on switch I, and port X on switch Y all in same vlan V (same bcast domain)
- cisco tag switching is one proprietary example (ISL) or IEEE 802.1Q

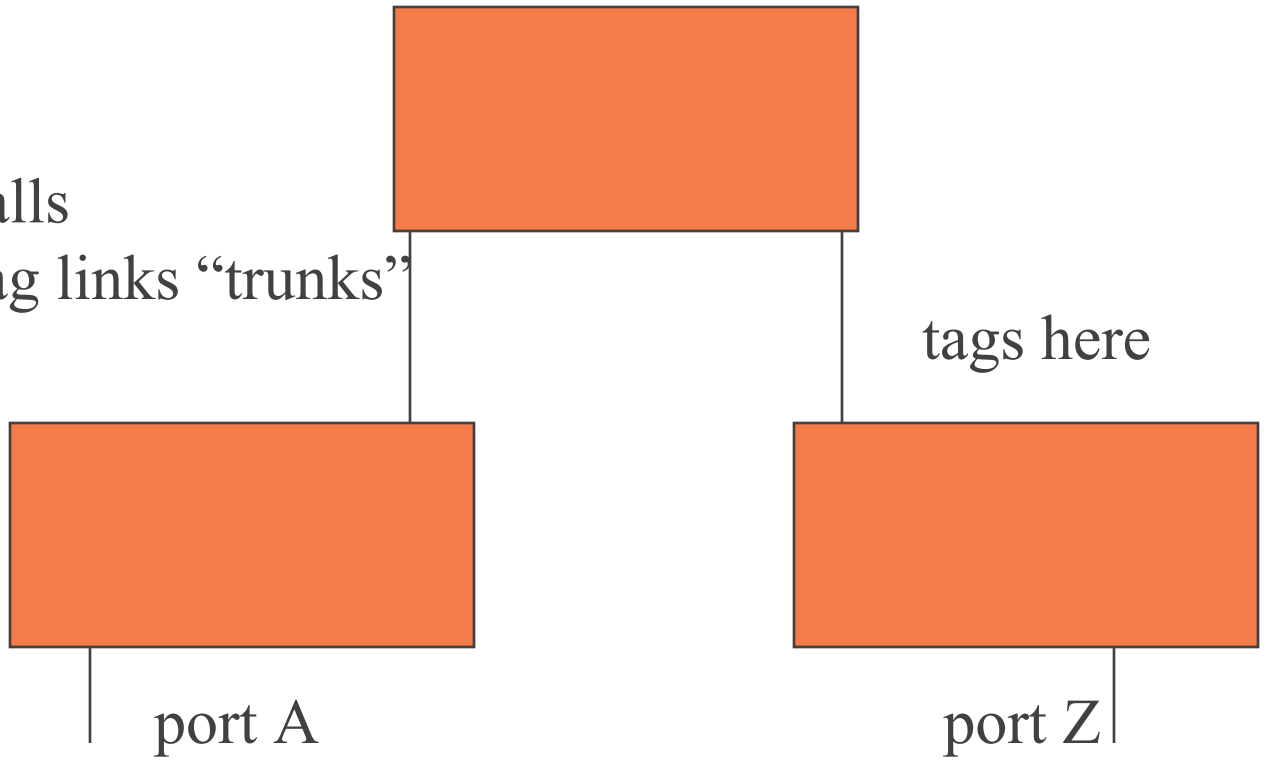
and how is it done?

- tags; i.e., inter-switch packets must contain VLAN identifier
- Cisco ISL - Inter Switch Link
  - tag is prepended in ISL header on ethernet (or other link type)
- IEEE 802.1Q - VLAN tag follows ethernet dst/mac/type, before network portion

inter-switch VLANS

3 switches

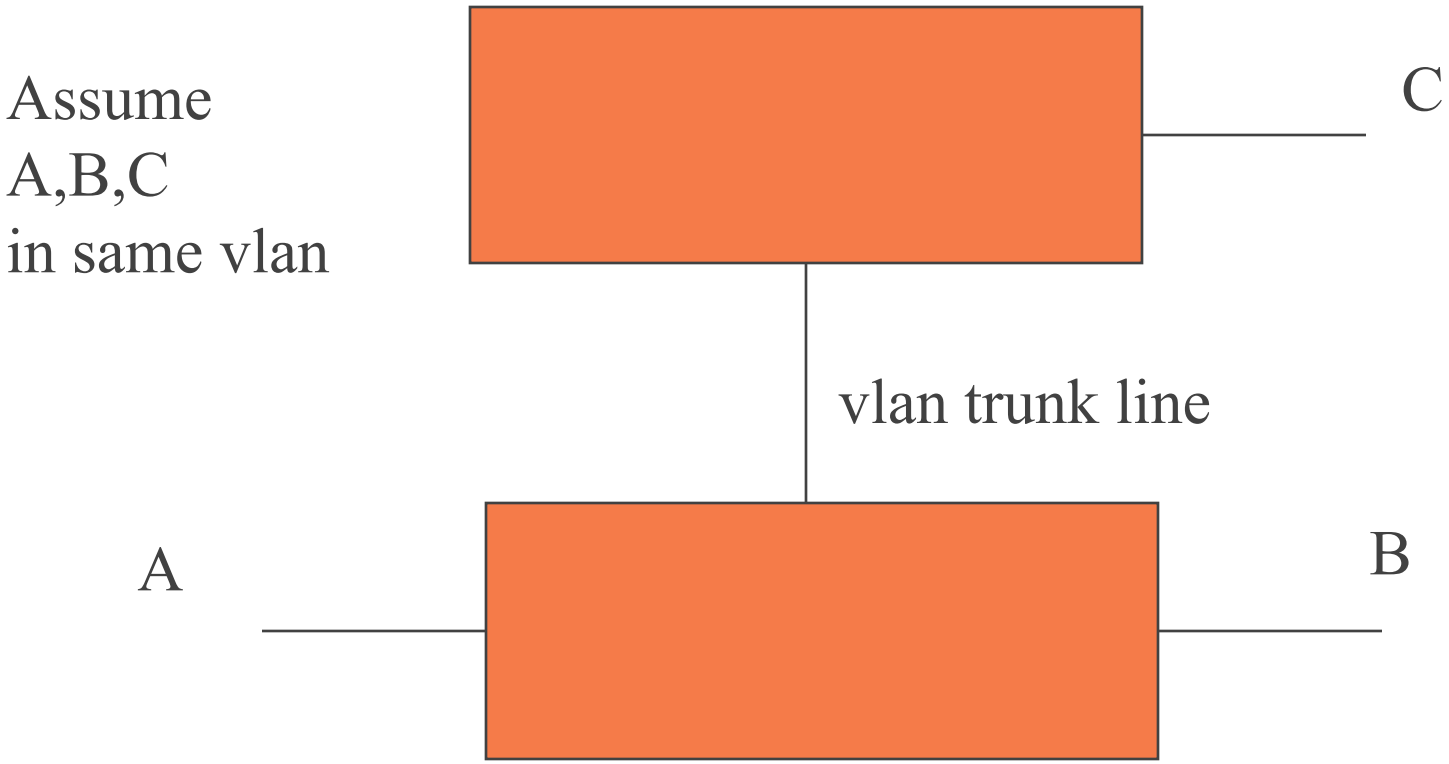
note: 1 vlan == 1 spanning tree



Jim Binkley

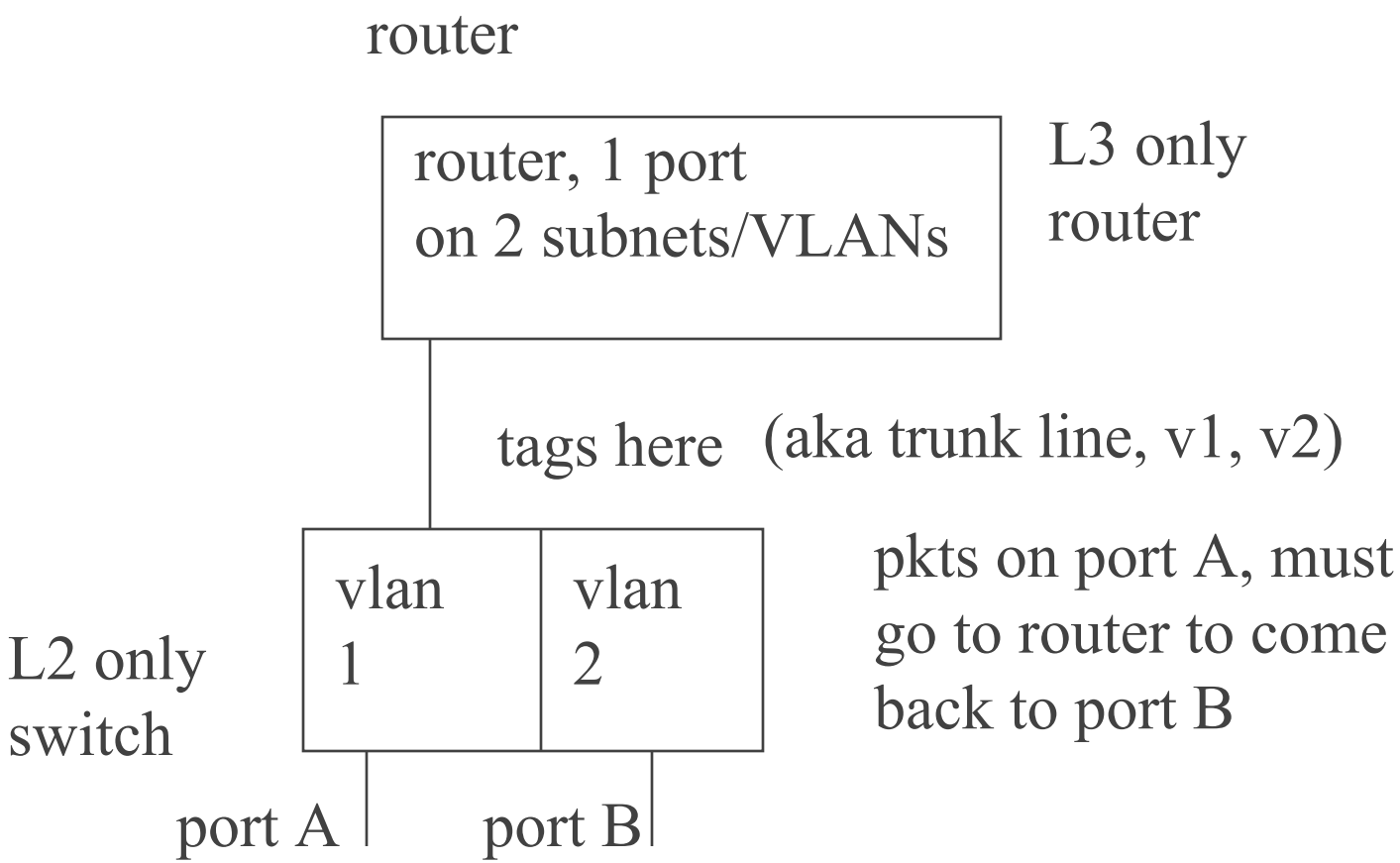
port {A,Z} == one broadcast domain

vlan and adaptive learning?



how does learning work here?

Cisco calls this: router on a stick



assume, A, B, in different VLANs

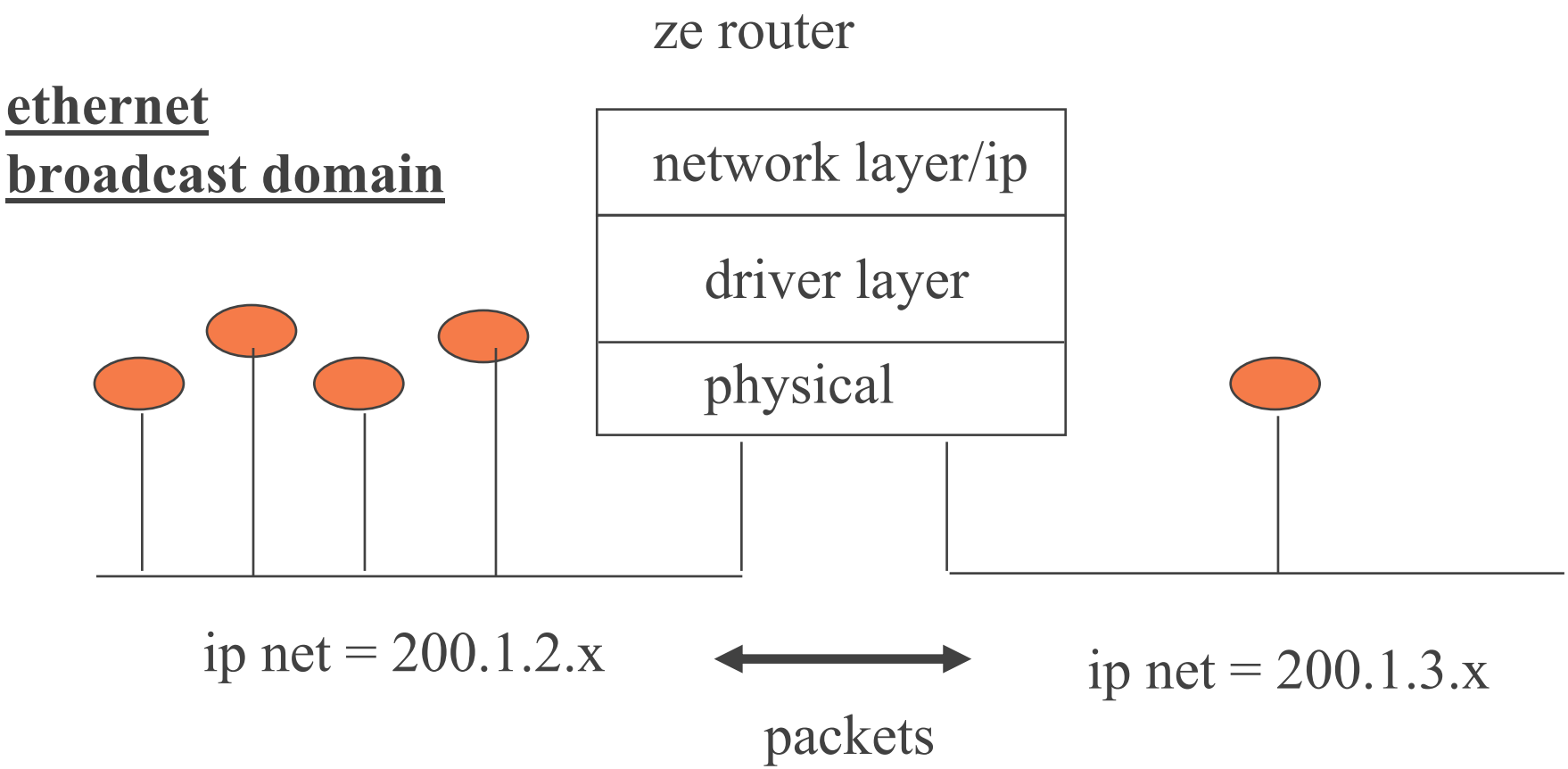
how does router affect collision/bcast domain?

- broadcasts are NOT usually forwarded
  - exceptions exist: e.g., DHCP/BOOTP request
- multicast the SAME, (barring multicast routing)
- collision domain limited as well
- **routers may be viewed as absolute sanity firewalls for ethernet segment disasters**
  - broadcast meltdown ...

summary - 2.5 (3) points of view

- talking about net design over time
- 1. router ip/subnet and strict segmentation for broadcast domains (traditional)
- 2. switched layer 2 ethernet
  - no segmentation for broadcast though (all hosts on link affected)
  - IEEE spanning tree still there (so is learning)
- 3. VLAN hybrid (1 subnet/1 spanning tree) possible both inter/intra-switch ip

IP subnet/router POV



router function (bubble up to top or close to outside)?

- traditional function: WAN interfaces
  - security ACCESS lists imposed on external to Internet points of contact
  - mixed media exchange
- must still tie together packets crossing subnets (or VLANs)
- must still be used to limit broadcast domains ABSOLUTELY (the spanning tree stops here)

switch function

- lives down below
- try and directly inter-connect (or switch connect)
  - hosts and their servers of use (minimize routing)
    - » e.g., file sy stem
    - » printer
    - » web server

switch to switch

- and router to switch
- should be as fast as possible
- remember packets here may be aggregated from many leaf hosts
- e.g., 100BASE (now 1000) switch to switch
- 10BASE (now 100) switch to host

router IP subnet POV cons

- switches live inside routers - somewhat invisible
  - if no IP address, totally invisible
- but SNMP doesn't support switches/multi-vlans (can see L3, not L2)
- can't see physical topology, only logical topology with SNMP
- switch MIB from IETF (and vlan specs) theoretically on the way (no standards)
- Cisco has proprietary SNMP MIBS
  - both CDP and SNMP hack for VLANs/SPT

Cisco CDP - Cisco Discovery Protocol

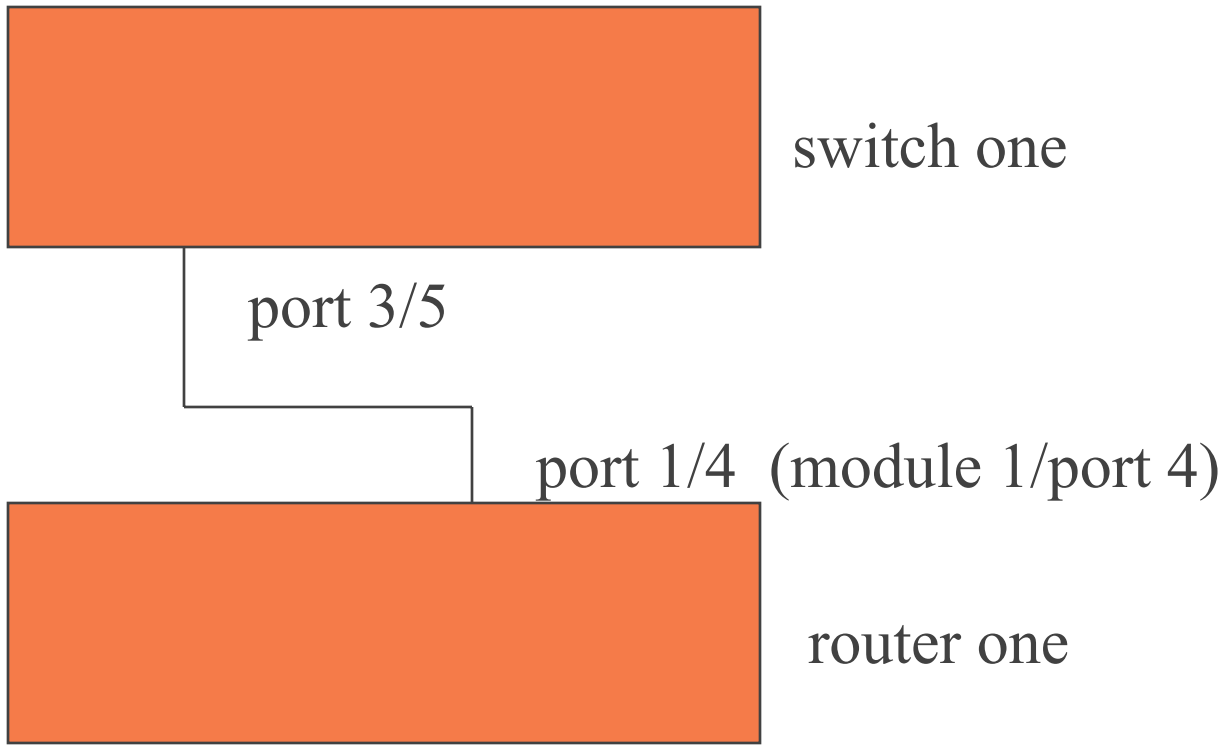
- switches/routers periodically multicast discovery packets out ports
- info includes: equipment type, port label
- SNMP MIB so can be fetched via SNMP
- high-level tool like ciscoview can show link-layer switch mesh including labels of ports on both sides of segment
- low-level telnet access useful too

cisco> show cdp neighbors

Capability Codes: R - Router, T - Trans Bridge, B - Source Route Bridge S - Switch, H - Host, I - IGMP, r - Repeater

Device ID	Local Intrfce	Holdtme	Capability	Platform	Port ID
Wintermute.ee.pdx.ed	Eth 1/5	136	R	RSP1	Eth 3/4
nanomite00C01D818526	Eth 1/7	130	T S	1900	16
nanomite00C01D818526	Eth 1/6	130	T S	1900	8
pcat142a	Fas 3/0	136	S	WS-C2924-X	Fas 0/1
pcat142b	Fas 0/0	128	S	WS-C2916M-	Fas 0/12
160a	Fas 2/0	159	S	WS-C2916M-	Fas 1/2

cdp point being:



“small” but powerful

- can logically see *physical* connections, port to port
  - you can’t tell where the wires are of course
- you can use telnet to see if someone has disappeared (crashed)
- you can use higher level mapping tools in Ciscoworks
  - to learn switch infrastructure

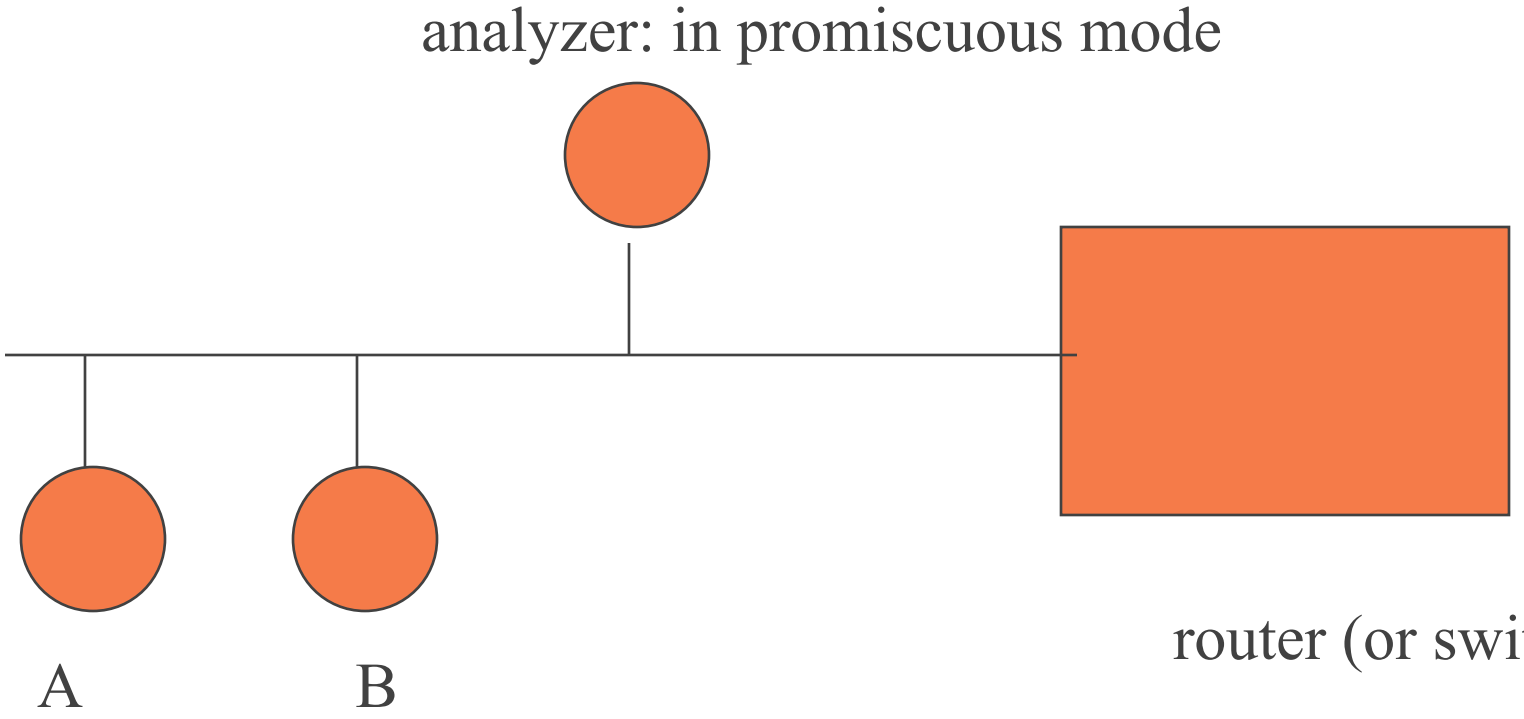
redundancy considerations

- spanning-tree can still give redundancy upon failure, but not 2X bandwidth
  - unless multiple vlan? or ether-channel
- network-layer IGPs like OSPF, EIGRP can take advantage of equal-cost paths between hosts (round-robin packets)
  - some switch produces can do that for local ports
- dynamic routing can provide traditional fallback if > 1 interface/path between networks

what of promiscuous mode ?

- traditionally hook up sniffer or RMON analysis tool
- one port on one host sucks down all packets
- and displays them in order (network analysis or “sniffing” or protocol debug)
  - sniffing is NG trademark name
- or categorizes (top N src/dst, which protocols in use acc. to percent, etc)

network analysis picture (trad)



analyzer: can hear A,B, to/from router traffic
on traditional 10BASE shared link

problem is switches

- 802.1D would not forward unicast traffic to another port if analyzer is on another port
  - or on another switch
- don't want analyzer for 1 link - 2 NIC card, full-duplex model
- may have too many switches anyway and too many ports
- University has hundreds of switches, 10's of routers (and not all easy to get to)

one fixup:

- Cisco has “SPAN”, called elsewhere
  - port mirroring
- one port on switch may be told to suck down all traffic on:
  - another port on same switch
  - range of ports on that switch
  - VLAN
- but traffic does not magically cross switches for inter-switch VLAN

pros/cons

- makes promiscuous mode hard if lots of switches/ports
- hurts RMON (too costly other than to centralize in network center)
- needs to be available per switch
- pro: makes network sniffing to get passwords less likely to succeed
  - host A can't see host B/host C traffic

maybe re promiscuous mode

- “he’s dead, Jim” ... (not really)
- shared 10BASE ports still exist though
- study question: assume you can run tcpdump on that there linux host
  - you have to install the package ...
  - how can you tell if you are on a switched port or not?

RSN: network interior QOS

- IP type of service combined with network-based packet queuing scheduling coming back (not end to end, just switch mesh)
- IEEE 802.1P - combined with tags to say layer 2 priority
- IETF diff-serv, use IPv4 traditional priority fields
- just a few priorities (say control/best/average)

QOS crudely considered

- of course, we can glue two pipes together to make one logical link
- **ether-channel** can logically glue two switch inter-connections into one logical port
  - with 2X (or more) speed
- velly interesting if 1 G ethernet pipes
- fatter pipes will always help

summary

- ethernet cheap, and faster, and changing
- point to point/star focus in switches
  - ideas include death of csma/cd collisions
  - port segmentation, full-duplex
- switches still have spanning tree, adaptive learning + VLAN
- POVS include network and link-layer
- routers still important for subnet forwarding and link-layer mayhem limitation

security considerations

- Cisco IOS images have bugs too
  - DOS attack against your switch
  - block access from outside world to net boxes
- switch/VLAN segmentation
  - can reduce damage by local link hacker sniffer searching for passwords
  - but bugs/flooding can lead to disaster
  - don't count on this for security against sniffers
- redundancy is important
  - L3 broadcast domain limitation is a good idea
  - spanning tree, more L3 domains

security (more)

- L3 has ACLs
  - use it to protect your border router
  - entire subnets
  - individual hosts
  - or the expensive firewall that sits right behind
    - » the router