Chapter 4
Spanning Tree *in Depth*
NET3011 – 17W

(Further material provided in additional slide decks)

STP - Basics

- STP is *only* needed with L2 redundant links
- STP is the *only* Protocol controlling L2 topology
- STP does *not* ensure optimal forwarding path, but simply a loop-free topology (to/away from the root)
- 3 flavours of STP: original, Rapid, and Multiple 802.1D*, 802.1w, 802.1s
- NET3011 covers STP *in depth*, including features and details for **stability** and **performance**

* "802.1D" *may* mean 802.1D-1990 (original), 802.1D-1998, or 802.1D-2004 (really RSTP!)

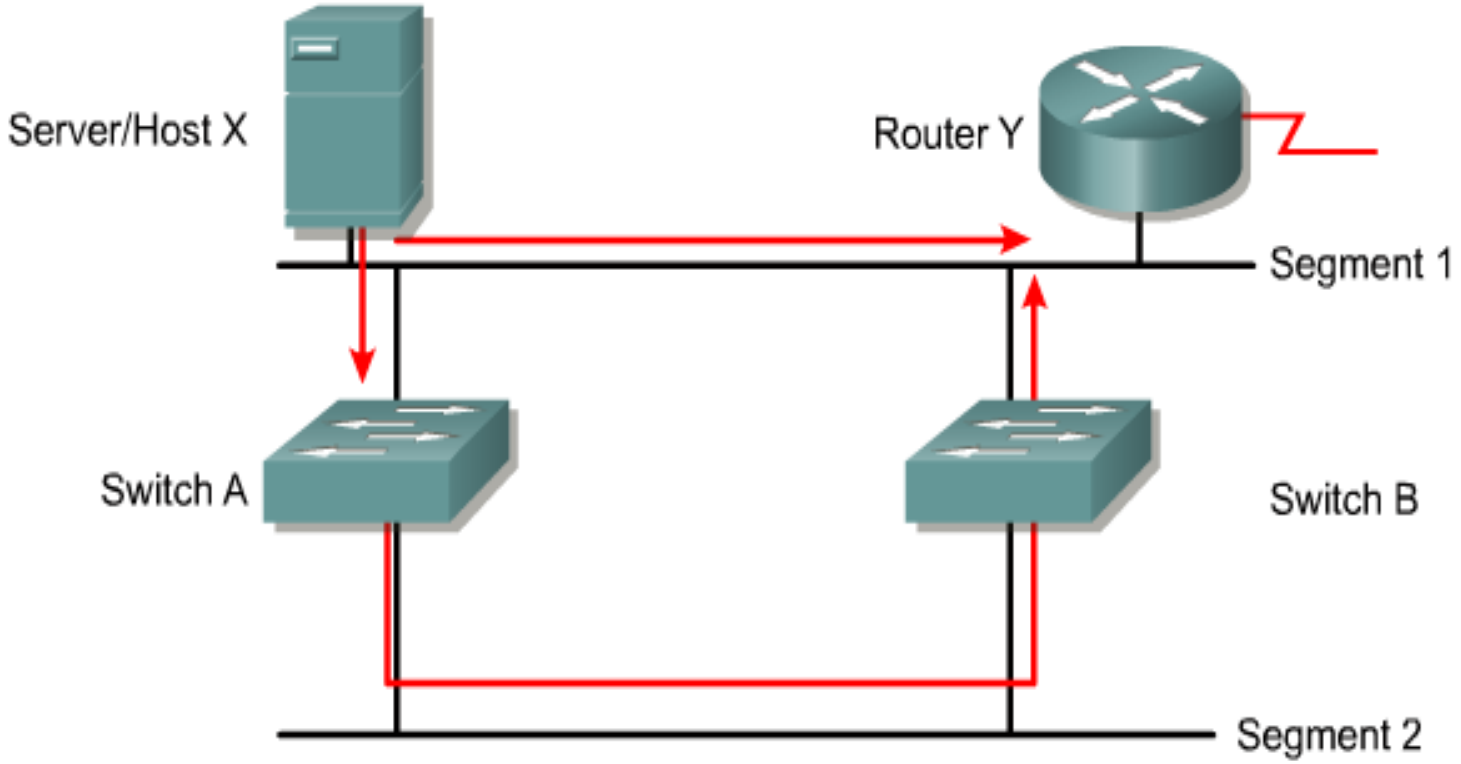
Who Invented STP?

- Radia Perlman

  -Internet Hall of Fame (2014)
  -SIGCOMM Award (2010)
  -USENIX Lifetime Achievement Award (2006)

STP: Why?

- Redundancy at Layer 2 is desirable for fault tolerance, but may cause several problems:
  - Broadcast Storms
  - Reception of Duplicate Frames
  - MAC Table Instability/Corruption



STP: Why?

- STP is a NECESSITY! It manages & protects:
  - intentional bridging loops for redundancy
  - accidental bridging loops (NB: not a L3 problem!)

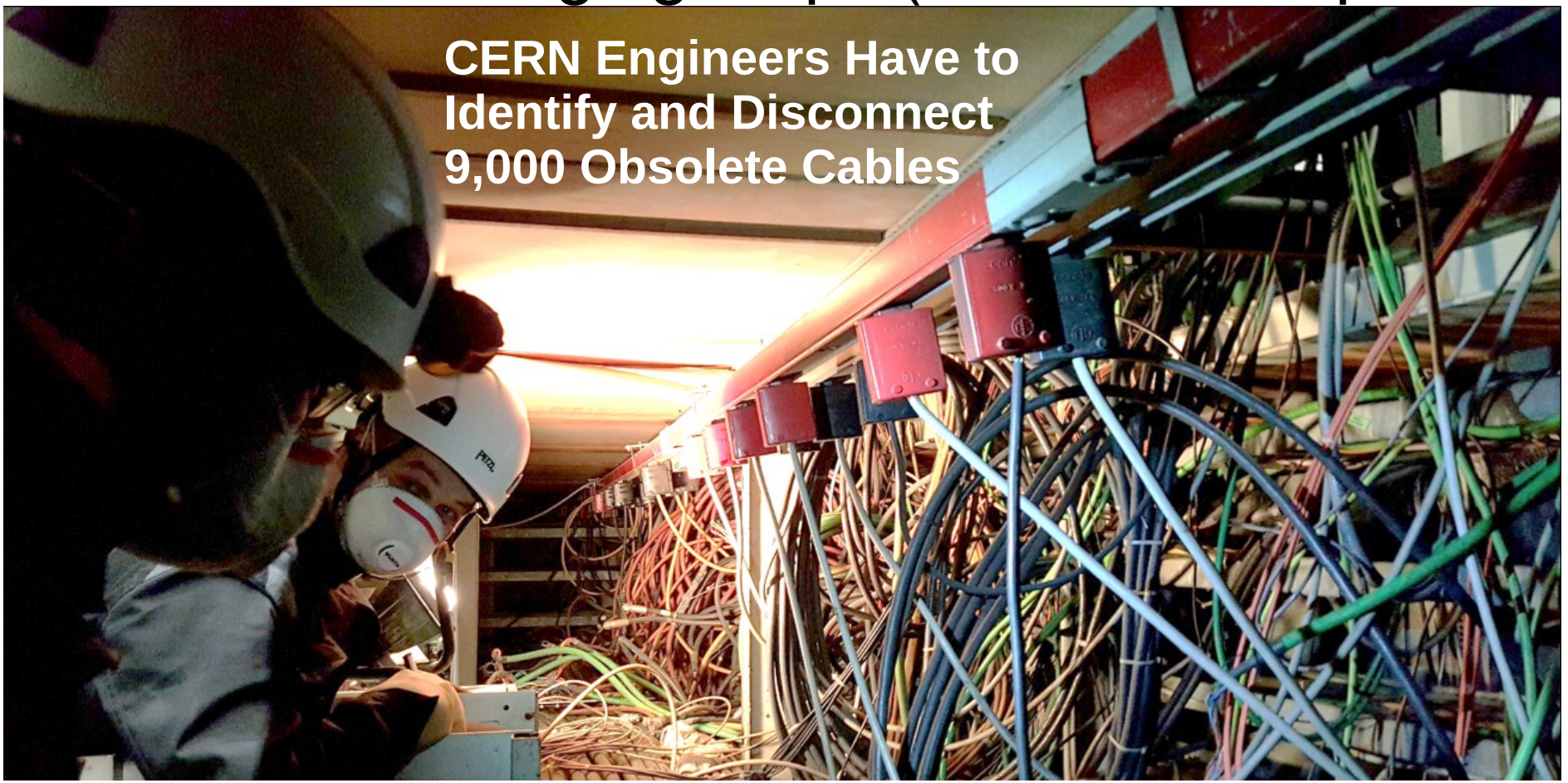


Image: Cern; retrieved from: https://motherboard-images.vice.com/content-images/article/29959/1453831578179193.png

STP Characteristics

- Topology determined by 3 elections, *lowest wins* !

A. Root Bridge

1. one criteria: Bridge ID = priority + MAC addr

B. Root Port (for a bridge)

1. Accumulated **cost** to root
2. Bridge ID (of sender)

3A. Port ID (of sender) = priority + port #

3B. Port ID (local) = priority + port #

C. Designated Bridge with Desg Port (for segment)

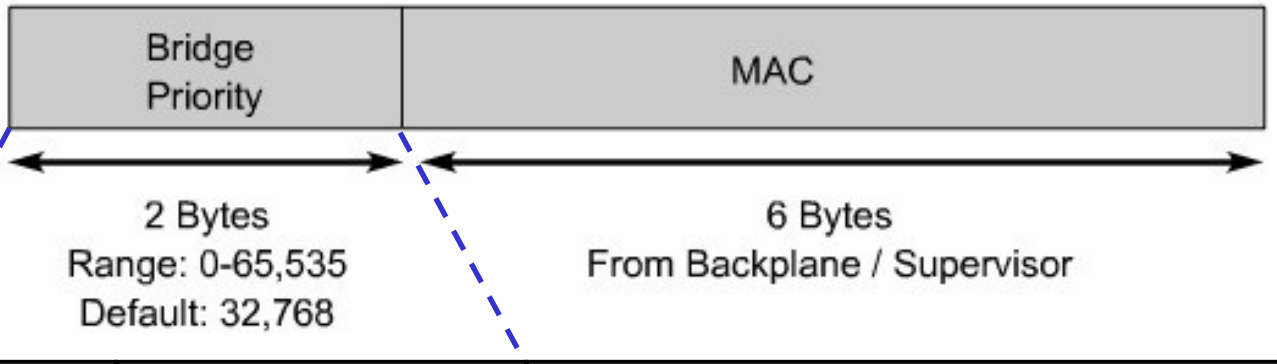
1. Accumulated **cost** to root
2. Bridge ID (of sender)
3. Port ID (on entire segment) = priority + port #

D. All other ports are Blocked (in 802.1D-1990)

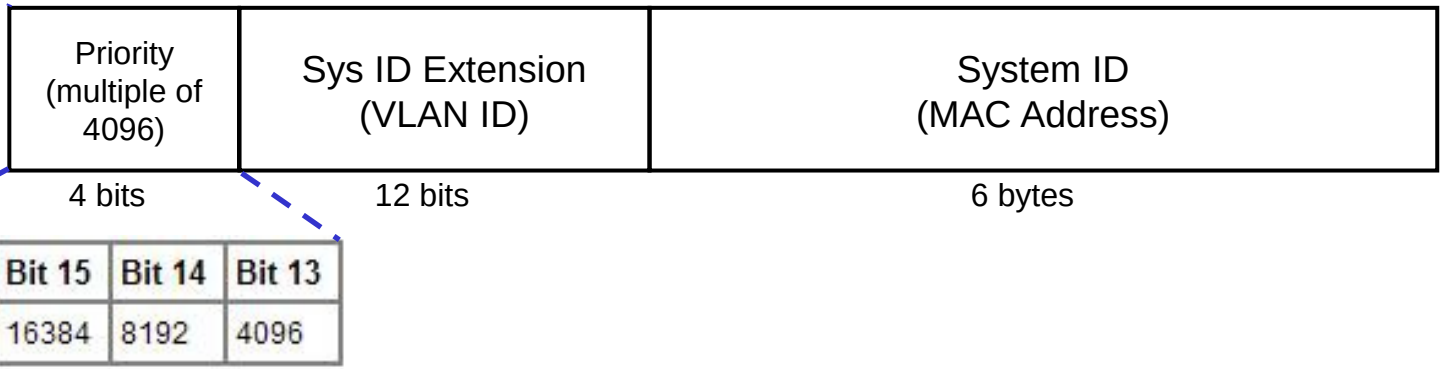
Bridge ID and Port ID

- Bridge ID (BID) = bridge priority + MAC address

Original Bridge ID



Revised Bridge ID

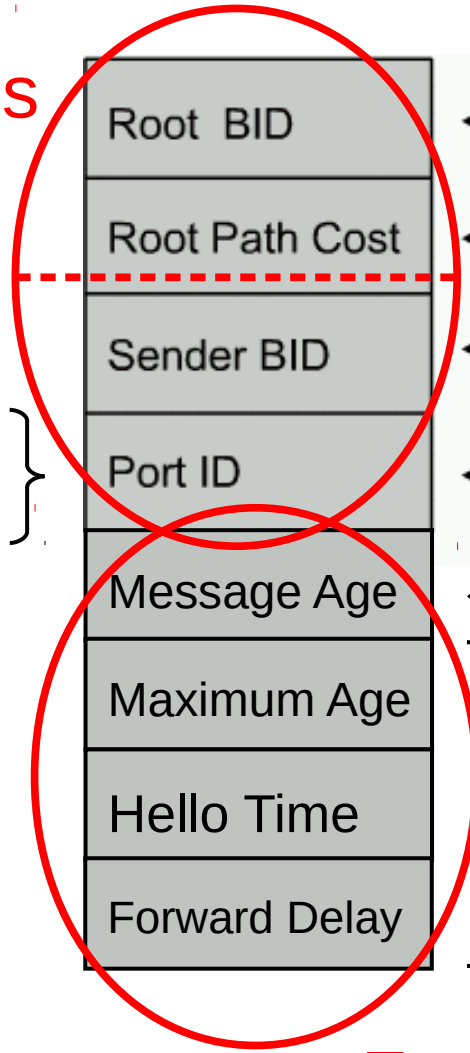


Bridge Priority

Bit 16	Bit 15	Bit 14	Bit 13
32768	16384	8192	4096

- Similarly, Port ID (PID) = port priority + port #

Bridge PDU – Key Fields



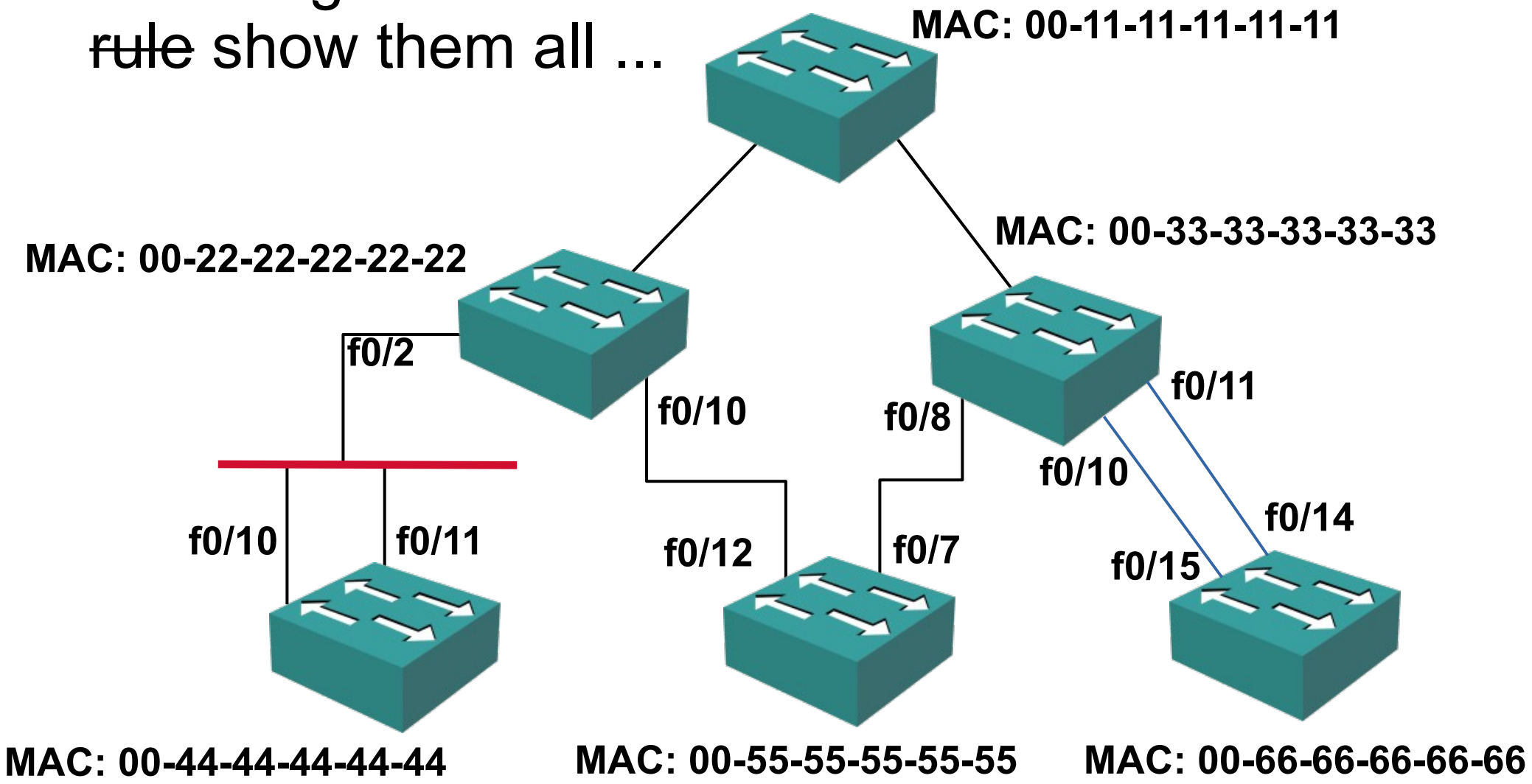
Root Bridge and BPDUs

- Bridges communicate using (IEEE 802.3) multicast frames (Dst MAC=01.80.C2.00.00.00* for VLAN1) called Bridge Protocol Data Units (BPDU)
- Upon initialization, each bridge sources Configuration BPDUs at the Hello interval (default: every 2 sec), with the Root BID declared as itself (!)
  - As all bridges do this, it is called a "root war".
- Rules for sending BPDUs:
  - If a bridge receives a BPDU with a better (ie. lower) Root BID value than its own, it changes its own BPDUs to propagate the better Root BID.
  - Eventually, the device with the lowest BID wins the war, once it's BID has propagated throughout the L2 domain.

* or Dst=01.00.0C.CC.CC.CD for VLANs other than 1 when using Cisco's PVST+
https://www.arista.com/assets/data/pdf/Whitepapers/STPInteroperabilitywithCisco.pdf

STP Elections

- One diagram to ~~rule~~ show them all ...



Optimal Root Bridge

- The device chosen as Root Bridge should:
  - have a central position in the topology (network radius should be 7 switches or less)
  - be a high-end device;
  - with lots of processing power,
  - and ample memory;
  - have good fault tolerance (e.g. redundant PSUs)
  - be connected with high bandwidth links,
  - be located where it will be noticed (LEDs visible, should things go badly)

STP BID Information

```
Switch# show span
VLAN0001
  Spanning tree enabled protocol ieee
  Root ID Priority  32769
        Address     1c17.d3d2.df00
        Cost        19
        Port        11 (FastEthernet0/9)
        Hello Time  2sec  Max Age 20sec Forward Delay 15sec

  Bridge ID Priority 32769  (priority 32768 sys-id-ext 1)
        Address     7010.5c16.2580
        Hello Time  2sec  Max Age 20sec Forward Delay 15sec
        Aging Time  300 sec

Interface              Role Sts Cost      Prio.Nbr Type
-------------------- ---- --- --------- -------- ----------
Fa0/1                Desg FWD 19        128.3    P2p
Fa0/2                Desg FWD 19        128.4    P2p
Fa0/9                Root FWD 19        128.11   P2p
```

= 802.1D-1990

BID = Priority + MAC

received in BPDU

Root Port

local values

Assigned Priority + VLAN #

1. Controlling Root Bridge Election

- Configure bridge priority:

```
SW(config)#spanning-tree vlan {vlan-id} priority {prio}
```

  - if “sys ID extension” is enabled, only multiples of 4096
- Configure as primary root bridge:

```
SW(config)#spanning-tree vlan {vlan-id} root primary
```

  - sets priority to *just enough* to become root (usually)
- Configure as secondary root bridge:

```
SW(config)#spanning-tree vlan {vlan-id} root secondary
```

  - sets bridge priority to 28672 (always)

2. Controlling Root Port Election

- Configure port/link **cost** to *determine* the **local** choice of root port:

```
SW(config-if)# spanning-tree [vlan {n}] cost {value}
```

… or revert to the default cost:

```
SW(config-if)# no spanning-tree [vlan {n}] cost
```

- Configure **port priority** to *influence* **downstream** choice of root port (only if costs are equal!)

```
SW(config-if)# span [vlan {n}] port-priority {0-240}
```

  - in increments of 16; default is 128

- Did you notice: port priority is **advertised** in a BPDU but link cost is **not**!

Costs, in detail

- To date, there have been three different sets of default values for link cost

802.1D-1998 802.1D-1990

- Original link cost limit was 16 bits (max 65535); Root Path cost was 32 bits

Link Speed	Cost(Revised IEEE Spec)	Cost (Previous IEEE Spec)
10 Gbps	2	1
1 Gbps	4	1
100 Mbps	19	10
10 Mbps	100	100

$$\frac{1000}{\text{Bandwidth}}$$

- As port speeds increased, link cost became 32-bits with new defaults
- Current 802.1D-2004 std: 20x10**9 / speed (in kpbs)

Table 17-3—Port Path Cost values 802.1D-2004

Link Speed	Recommended value	Recommended range	Range
<=100 Kb/s	200 000 000*	20 000 000–200 000 000	1–200 000 000
1 Mb/s	20 000 000[a]	2 000 000–200 000 000	1–200 000 000
10 Mb/s	2 000 000[a]	200 000–20 000 000	1–200 000 000
100 Mb/s	200 000[a]	20 000–2 000 000	1–200 000 000
1 Gb/s	20 000	2 000–200 000	1–200 000 000
10 Gb/s	2 000	200–20 000	1–200 000 000
100 Gb/s	200	20–2 000	1–200 000 000
1 Tb/s	20	2–200	1–200 000 000
10 Tb/s	2	1–20	1–200 000 000

*Bridges conformant to IEEE Std 802.1D, 1998 Edition, i.e., that support only 16-bit values for Path Cost, should use 65 535 as the Path Cost for these link speeds when used in conjunction with Bridges that support 32-bit Path Cost values.

Source: http://standards.ieee.org/getieee802/portfolio.html

Costs, in detail

- The Root bridge always sends out it's BPDUs with a Root Path Cost of 0 (zero).
- As a BPDU is *received*, the receiving bridge adds the cost of the link on which the BPDU arrived to the Root Path Cost.
- The value of Root Path Cost, from the perspective of a given device, is always the aggregate cost to reach the Root Bridge from that device.

STP Port Information

```
Switch# show spanning-tree
VLAN0001
  Spanning tree enabled protocol ieee
  Root ID Priority  32769
        Address     1c17.d3d2.df00
        Cost        19
        Port        11 (FastEthernet0/9)
        Hello Time  2sec  Max Age 20sec Forward Delay 15sec

  Bridge ID Priority 32769  (priority 32768 sys-id-ext 1)
        Address     7010.5c16.2580
        Hello Time  2sec  Max Age 20sec Forward Delay 15sec
        Aging Time  300 sec

Interface              Role Sts Cost      Prio.Nbr Type
-------------------- ---- --- --------- -------- ---------
Fa0/1                Desg FWD 19        128.3    P2p
Fa0/2                Desg FWD 19        128.4    P2p
Fa0/9                Root FWD 19        128.11   P2p
```

Received Root Path Cost + link cost

Configured port cost (this VLAN)

Conf'd & Advertised port ID = priority + port #

Blocked Ports

- All inter-switch ports that are neither elected root nor designated are blocked:

```
SW# show spanning-tree blocked
```

- The LED on blocked switchports will be amber (unless they are trunks, in which case it shows trunk status)
- RSTP/MST have other roles instead of blocked:
  - Backup (seen on Cisco switches even in STP)
  - Alternate (also on Cisco switches even in STP)

STP Support for VLANs

- Cisco runs a separate instance of STP for each and every VLAN (!)
- Cisco BPDUs are tagged (via an extra appended field) so STP can run on a per-VLAN basis
- TRUNK links can thus be in STP Forwarding state for some VLAN(s) while Blocked for other VLANs
- The "system ID extension" attribute indicates that VLAN # is added to assigned priority in order to generate the Bridge Priority **for each VLAN**
- Configure port cost and priority values per VLAN:

```
(config-if)#spanning-tree vlan {#} cost {value}
(config-if)#spanning-tree cost {value}
```
!for **all** VLANs

Communication in Each STP **State**

Disabled (not listed)	• Inoperative (administratively shut down, failure, or error (not *exactly* an STP state)
Blocking "**BLK**"	• Receive only; process received BPDUs, **no** forwarding of user data
Listening "**LIS**"	• receive (Root) / send (Desg) BPDUs while waiting for topology discovery to complete (= forward delay); **no** data forwarding
Learning "**LRN**"	• Spend time (= forward delay) learning user MAC addresses so no need to flood later; receive (Root) and send (Desg) BPDUs; **no** forwarding of user data
Forwarding "**FWD**"	• Normal operation: forward user data, learn MAC addresses; propagate BPDUs

Can you draw a correct state diagram, showing all transitions?

Timer Values

- Each bridge adopts the timer values propagated from the Root Bridge.
  - Timer values may be set on each individual device, but they're **not** used unless the device becomes Root!
- Default MaxAge of 20 secs is based on 7 segment hops (of up to 2 secs each) and 3 lost Hellos
  - Root Bridge plus 6 more devices to any leaf switch
- Upon failure, transition back to forwarding could take up to 50 secs:
  - 20 sec (aging) + 15 secs (listen) + 15 secs (learn)

STP Timer Information

```
Switch# show span
VLAN0001
  Spanning tree enabled protocol ieee
  Root ID Priority  32769
        Address     1c17.d3d2.df00
        Cost        19
        Port        11 (FastEthernet0/9)
        Hello Time  2sec  Max Age 20sec  Forward Delay 15sec

  Bridge ID Priority 32769  (priority 32768 sys-id-ext 1)
        Address     7010.5c16.2580
        Hello Time  2sec  Max Age 20sec Forward Delay 15sec
        Aging Time  300 sec

Interface            Role Sts Cost      Prio.Nbr Type
-------------------- ---- --- --------- -------- ---------
Fa0/1                Desg FWD 19        128.3    P2p
Fa0/2                Desg FWD 19        128.4    P2p
Fa0/9                Root FWD 19        128.11   P2p
```

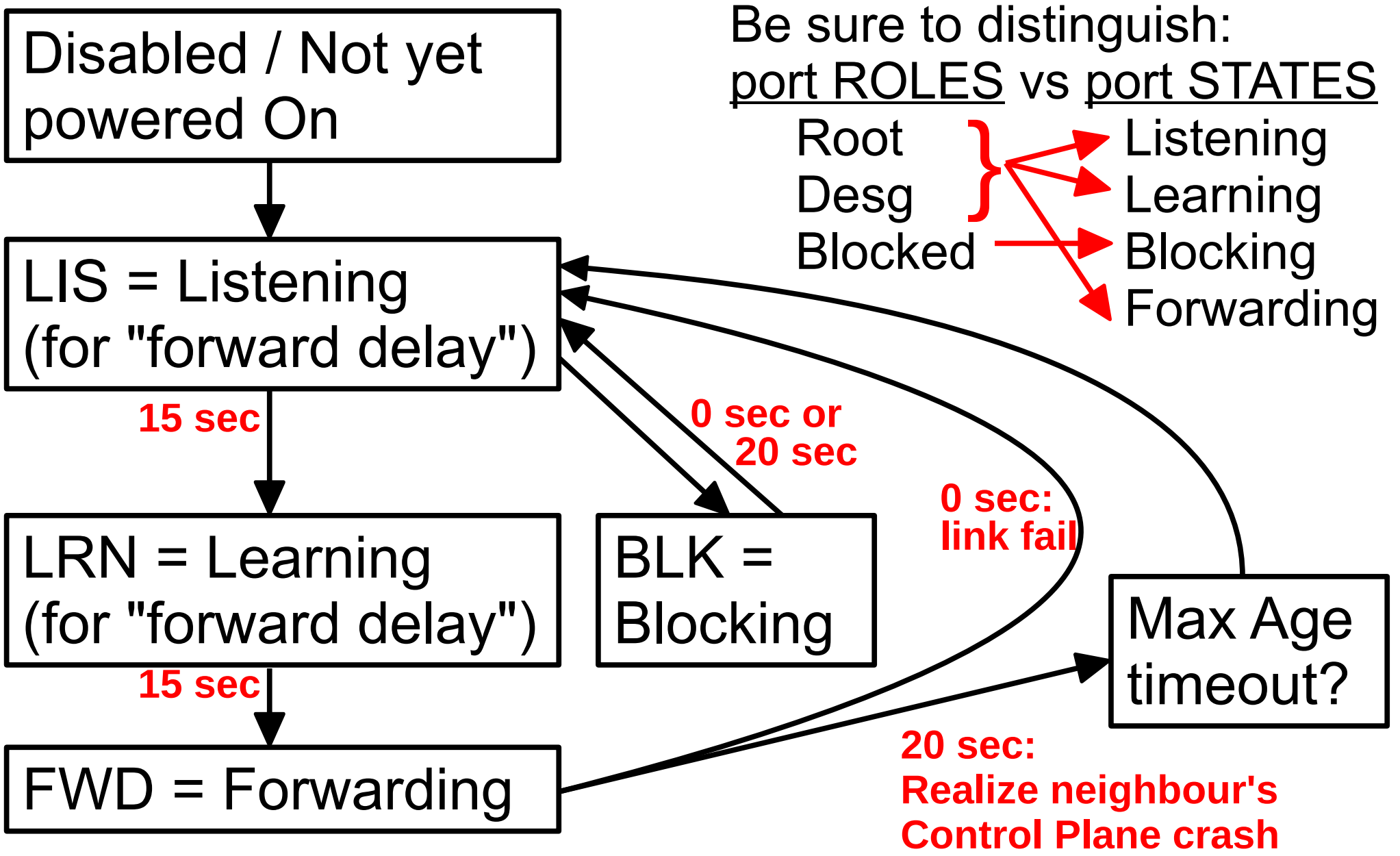
All three timer values are received in BPDUs

for TCN

(over-ruled by Root Bridge)

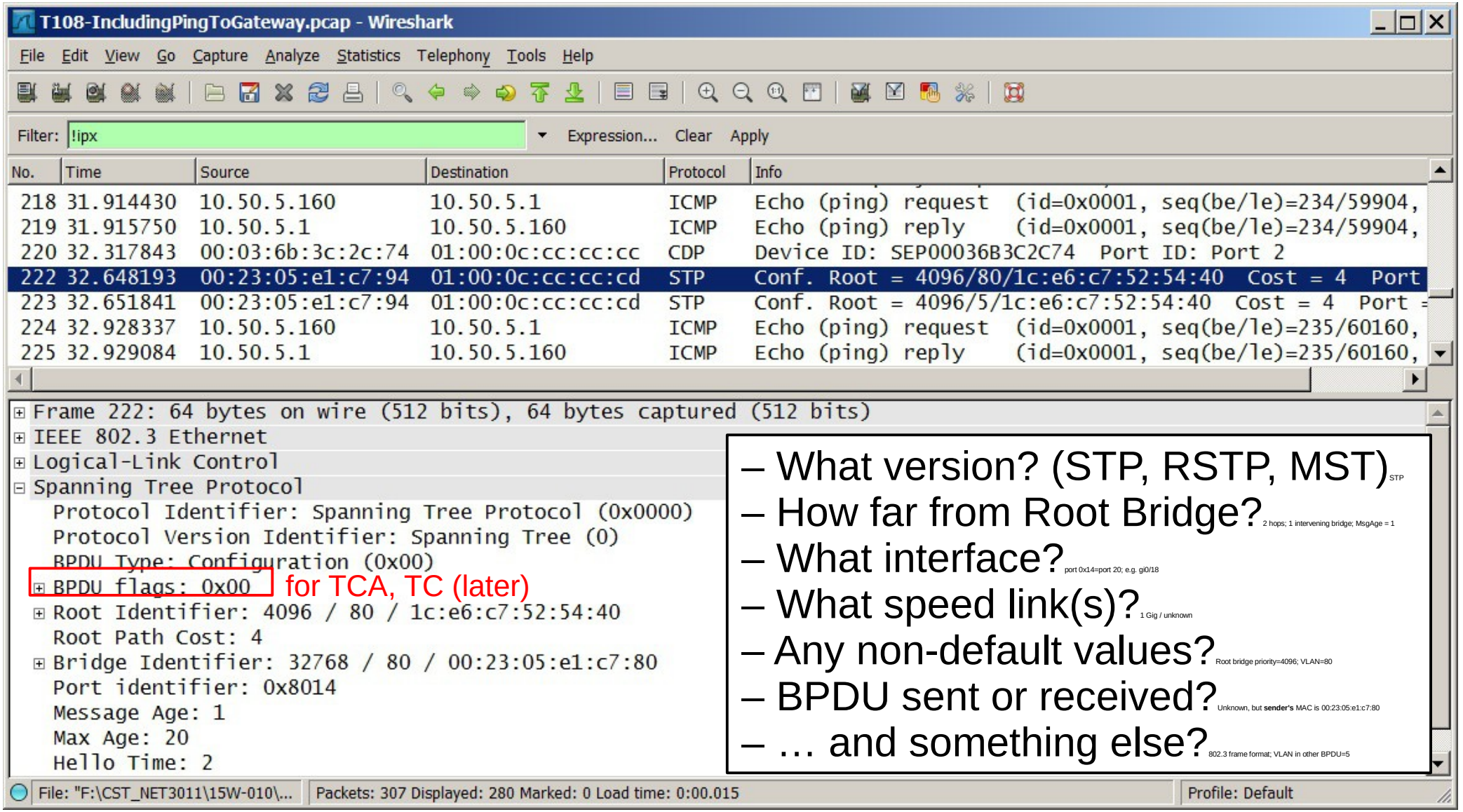
MAC addr. aging-time (local value)

STP State Diagram



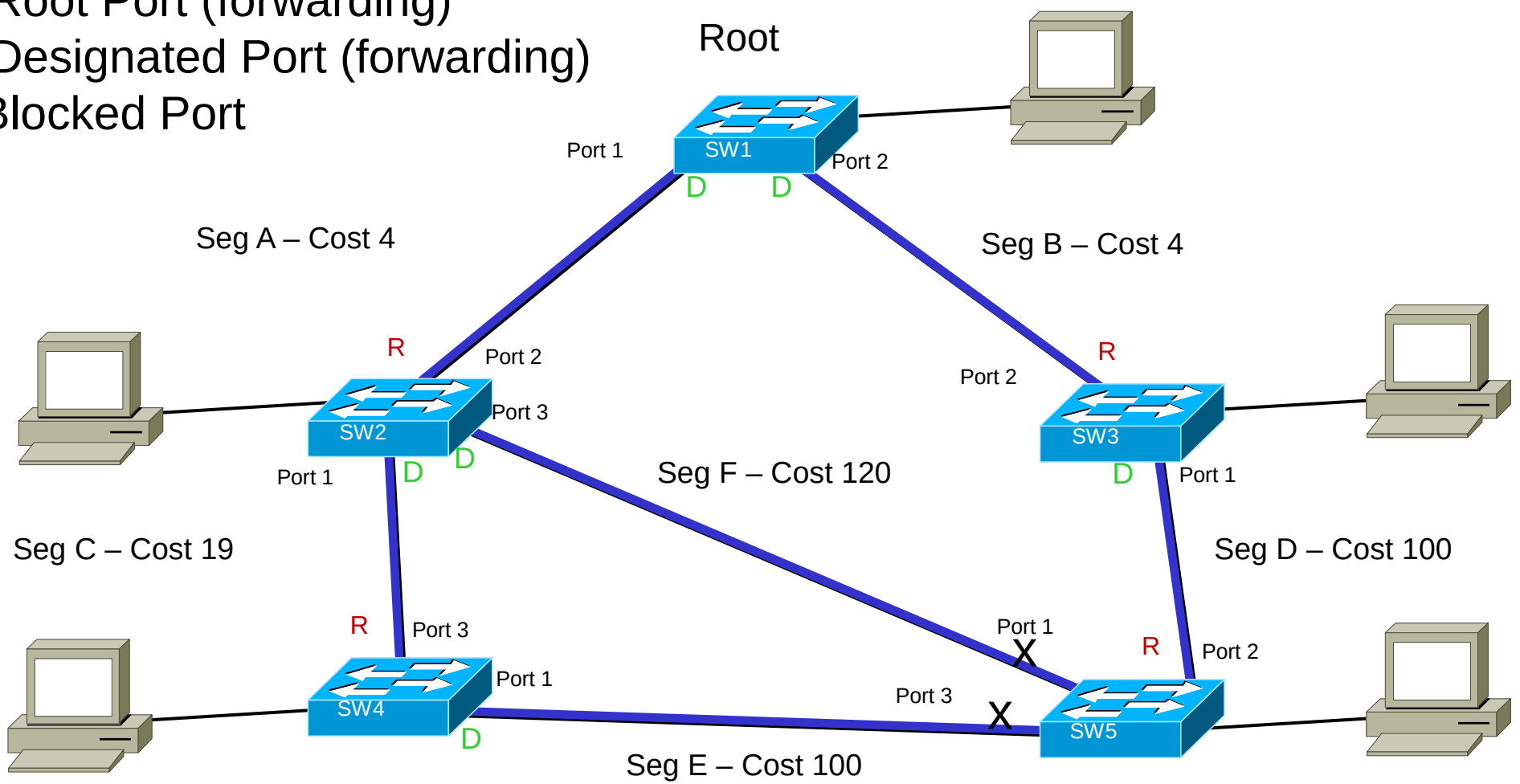
Wireshark BPDU Example

- It's amazing what can you tell from a capture!



STP Topology 1 – All links P-to-P

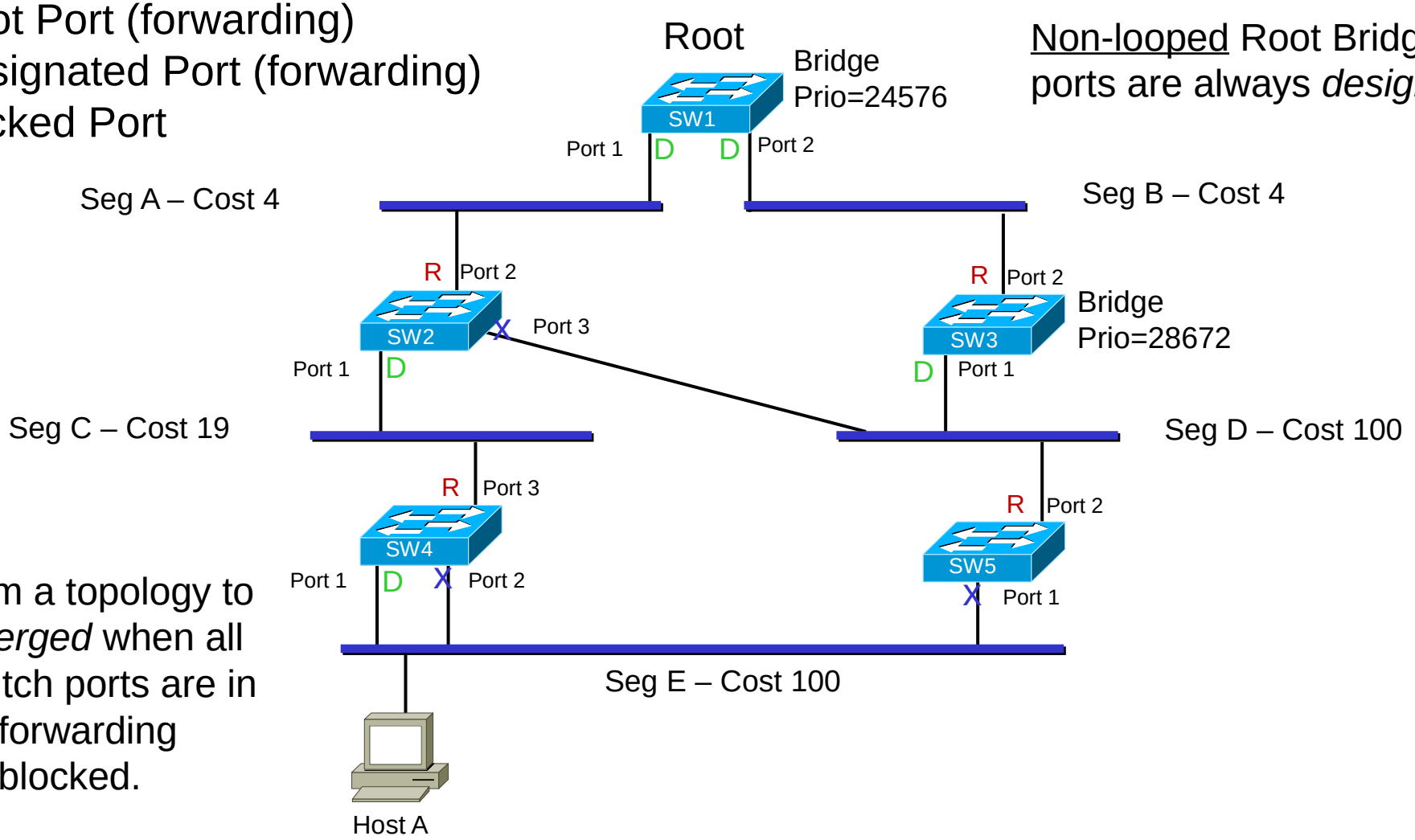
R – Root Port (forwarding)
D – Designated Port (forwarding)
x – Blocked Port



NB: In this example, assume the MAC ID of SW(n) is less than that of SW(n+1) and that all priorities are at their defaults, unless otherwise shown.

STP Topology 2: Multi-access links

R – Root Port (forwarding)
D – Designated Port (forwarding)
x – Blocked Port

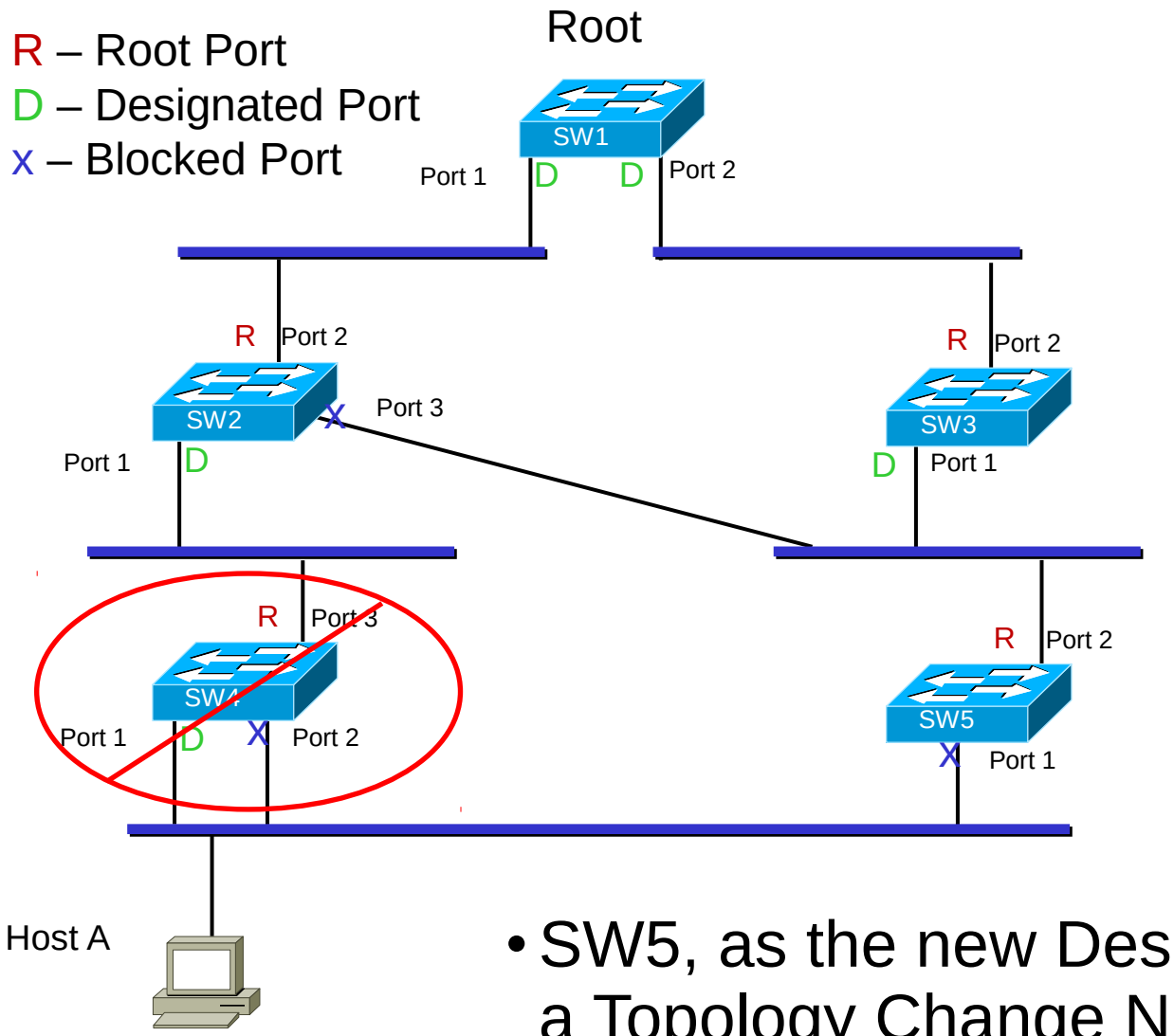


Non-looped Root Bridge ports are always *designated*

We deem a topology to be *converged* when all inter-switch ports are in either a forwarding state or blocked.

NB: In this example, assume the MAC ID of SW(n) is less than that of SW(n+1) and that all priorities are at their defaults, unless otherwise shown.

Redundancy – Failure Recovery



- SW5 stops receiving BPDUs on Port 1 (from SW4 Port 1)
- Seg E suffers a temporary loss of connectivity while the topology reconverges (20 + 15 + 15 = 50 secs)
- How does the rest of the network reach Host A?

- SW5, as the new Designated Bridge, generates a Topology Change Notification (TCN) BPDU to help reconvergence!

Failure Recovery – Designated Bridge

- When a Designated Port serving a segment fails, (whether port, link, or Designated Bridge), there are 2 possibilities:
  1. If no other switch ports are eligible to serve that segment, connectivity is lost until the fault is corrected
  2. If one or more other service-eligible switch ports exist, they must be Blocking (why?) …
     - all blocking ports will cease receiving the Config-BPDUs normally sent by Designated Bridge, leading to a timeout after MaxAge
     - all such ports re-run the election to become Designated Bridge:
       i. Ports move to Listening state, and propagate at least one BPDU onto the segment (for *Forward Delay* secs)
       ii. The port having the superior BPDU will become Designated, move to Learning state (Forward Delay secs), and then to Forwarding state
       iii. The new Designated Bridge originates a TCN to the Root Bridge

Topology Change Negotiation ("3-way")

- During **normal operation**, the root bridge generates Configuration BPDUs (message type 0x00) every Hello interval; all other switches relay the BPDUs out their Designated Ports (Root BID + own BID info)
- For any **change in topology** (eg. link down or new Designated Bridge), the relevant switch will source a **Topology Change Notification BPDU** (TCN) (**message type 0x80**) out it's root port (unless that port has failed); any upstream switches (except the root) propagate the TCN out their root port
- Switches receiving a TCN (including the root) are required to acknowledge it by returning a Configuration BPDU with the Topology Change Acknowledge (TCA) flag set (flags = 0x80).
- Any switch sending a TCN will repeat the TCN every Hello interval until an acknowledgment is received from its upstream neighbour.
- A root bridge receiving a TCN will set the Topology Change (TC) bit in its Configuration BPDUs (flags = 0x01) for the next MaxAge+FwdDelay secs (default 20+15). All switches then shorten their MAC Table aging time to FwdDelay (from default of 300 sec), thereby flushing any learned MAC addresses not in active use.

Failure Recovery – Root Path

Upon failure of a device's Root Path (whether the port, link or upstream device):

1. It will no longer be receiving the Config-BPDUs originated by the Root Bridge, leading to a timeout after MaxAge
   - If the device can directly sense root port failure, it can transition to the next step without waiting for the MaxAge timeout period
2. The device discards its saved best BPDU
3. It must now choose another root port, which it does in the same fashion as during initial startup (BPDU with least root path cost)
   - Once the new root port is determined, a new best BPDU is saved
   - Note: If the failure prevents all reachability to the Root Bridge, a new Root Bridge would be elected as a result.
4. As the new root port transitions to Forwarding, the device sources a TCN towards the Root Bridge

Reminder

- *LOTS* of details do not appear in these slides
- You are responsible for *reading* the textbook to gain the knowledge (memorization) and understanding (apply the knowledge)
- Ensure you have a detailed understanding of what happens in each case:
  - the sequence and timing of the events
  - which device issues a TCN and the consequence of that throughout the topology
  - what changes (if any) to the BPDUs propagated in the topology (and for what period of time)
  - how the topology re-converges