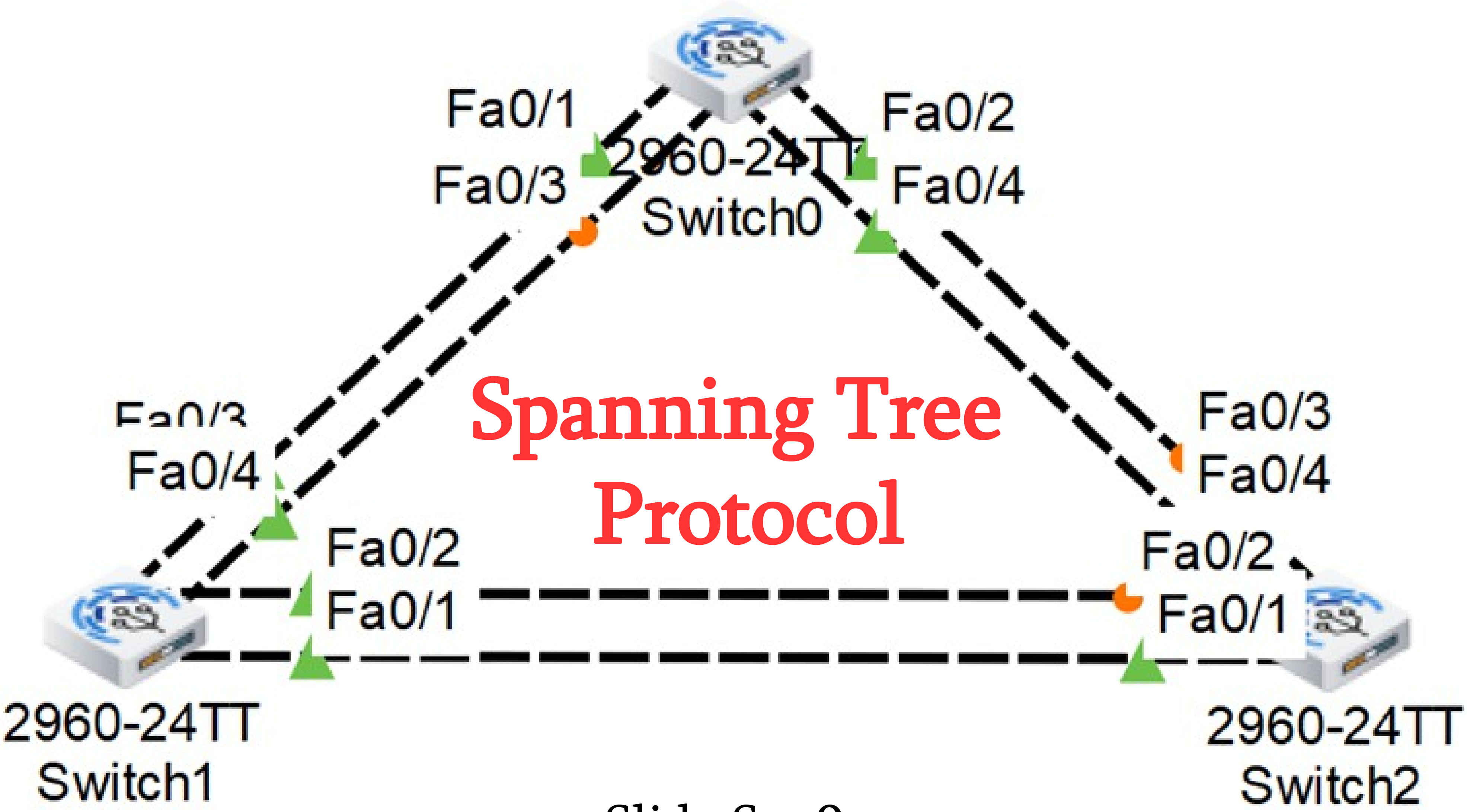
Spanning Tree Protocol



Slide Set 9

Overview

- Define redundancy and its importance in networking
- Describe the key elements of a redundant networking topology
- Define broadcast storms and describe their impact on switched networks
- Define multiple frame transmissions and describe their impact on switched networks
- Identify the benefits and risks of a redundant topology
- Describe the role of spanning tree in a redundant-path switched network
- Identify the key elements of spanning tree operation
- Describe the process for root bridge, root ports, designated ports election

Redundancy

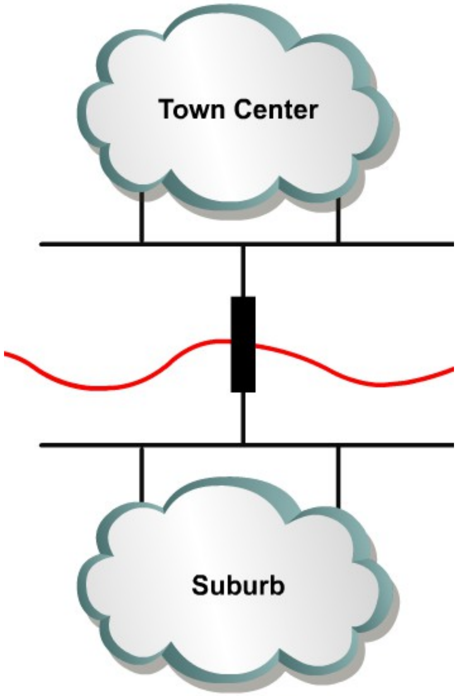
There is one car, can I drive to work?

There are two cars, can I drive to work?

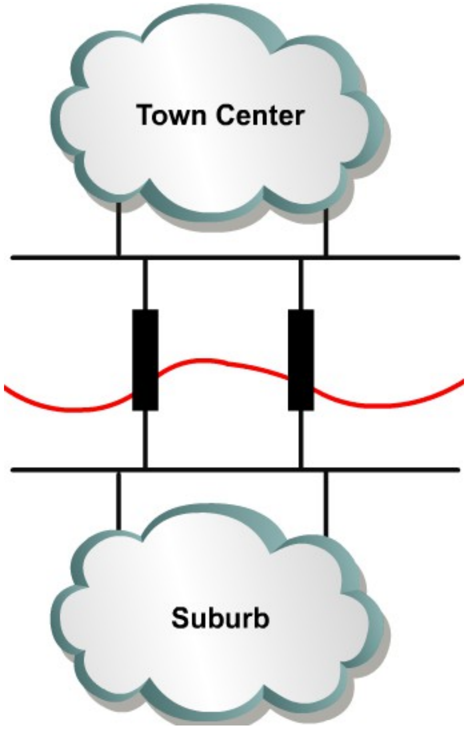
- Achieving such a goal requires extremely reliable networks.
- Reliability in networks is achieved by reliable equipment and by designing networks that are tolerant to failures and faults.
- The network is designed to reconverge rapidly so that the fault is bypassed.
- Fault tolerance is achieved by redundancy.
- Redundancy means to be in excess or exceeding what is usual and natural.

Redundant topologies

N 14th St
E Missouri Ave
Desert Storm
Park
N 18th Pl
20th St
22nd St
E Colter St
N 32nd St
E Camelback Rd
Town And
Country
Shopping
Center
Camelback
Esplanade
N 32nd St
N 30th Pl
N 32nd Pl
E Elm St
N 14th Pl
N 16th St
N 24th St
N 26th St
N 28th St
E Campbell Ave
E Roma Ave
E Turney Ave
N 13th Pl
E Glenrosa Ave
N 25th St
Los Olivos
Park
Grand Can
51
N 19th S
N 23rd



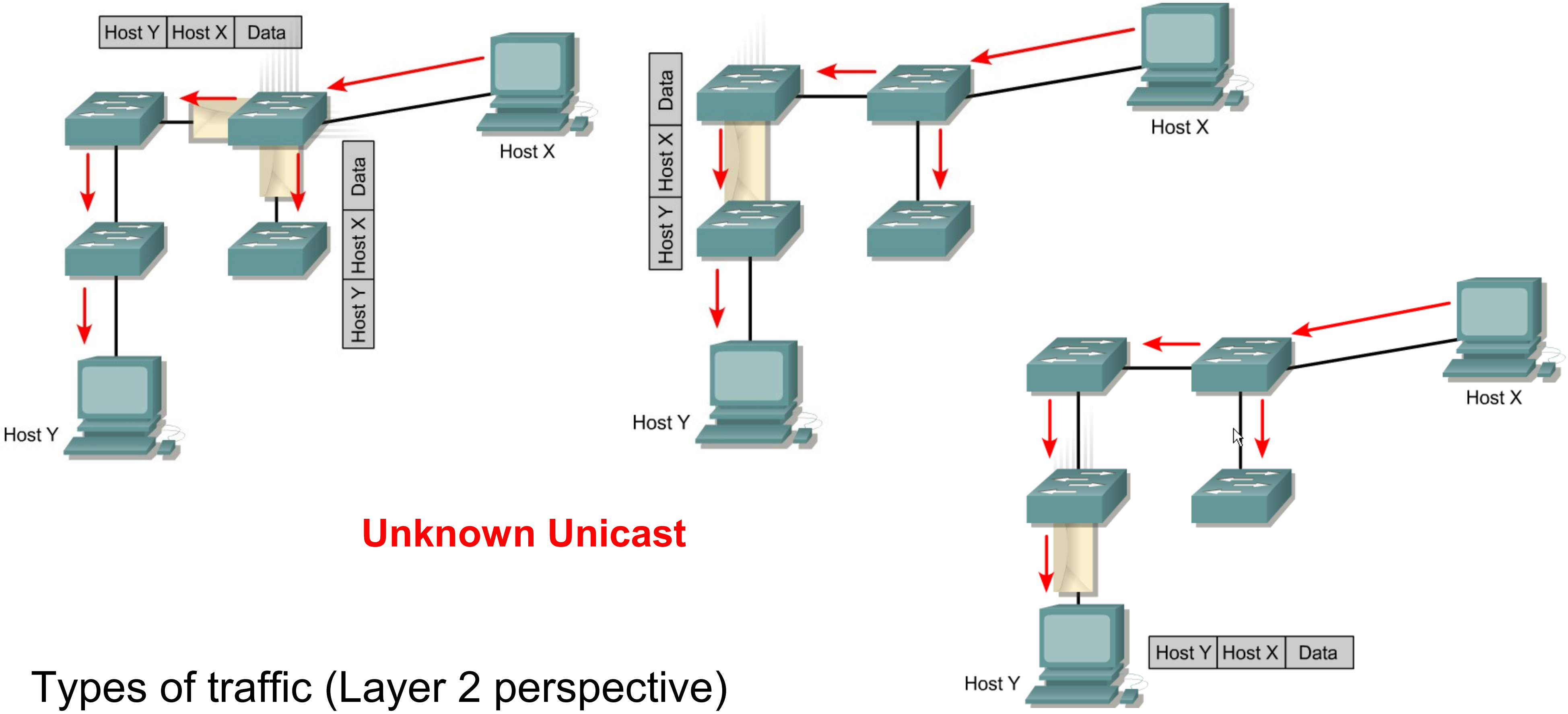
One Bridge



Redundant Bridges

- A network of roads is a global example of a redundant topology.
- If one road is closed for repair there is likely an alternate route to the destination

Types of Traffic



Unknown Unicast

Types of traffic (Layer 2 perspective)

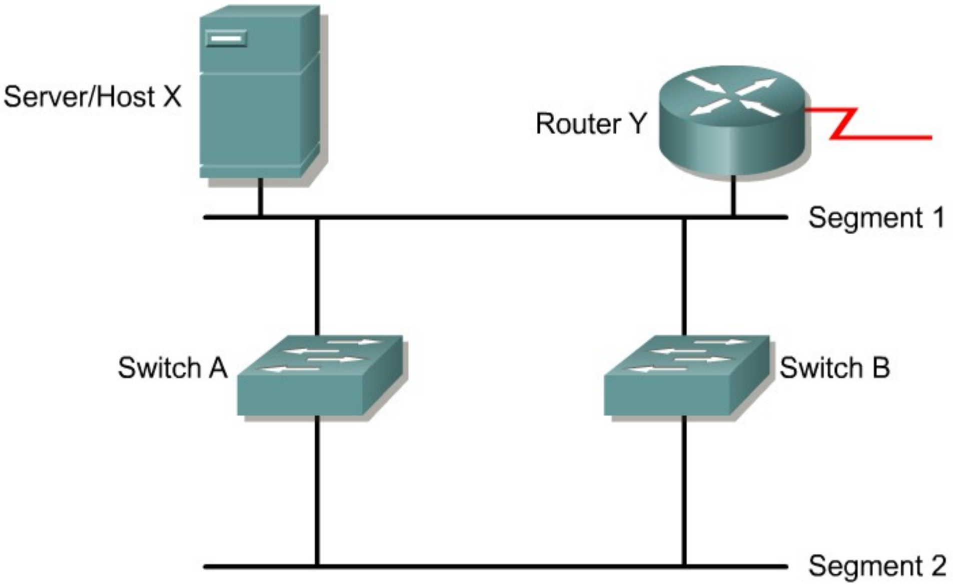
Known Unicast: Destination addresses are in Switch Tables

Unknown Unicast: Destination addresses are not in Switch Tables

Multicast: Traffic sent to a group of addresses

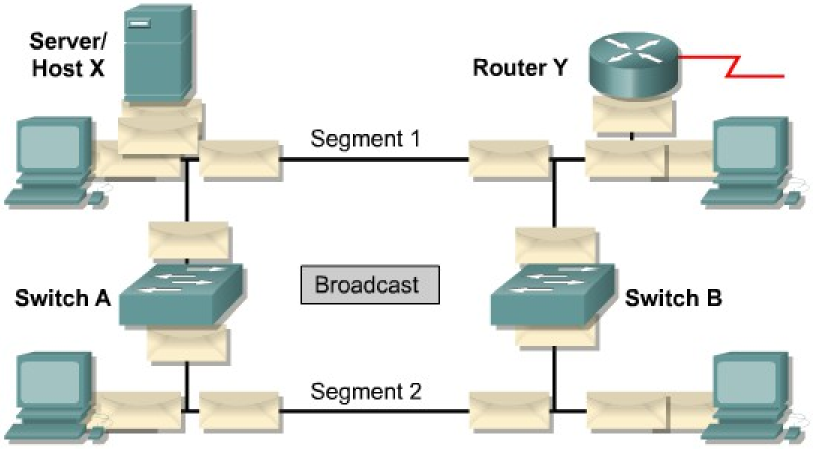
Broadcast: Traffic forwarded out all interfaces except incoming interface.

Redundant switched topologies



- Switches learn the MAC addresses of devices on their ports so that data can be properly forwarded to the destination.
- Switches will flood frames for unknown destinations until they learn the MAC addresses of the devices.
- Broadcasts and multicasts are also flooded. (Unless switch is doing Multicast Snooping or IGMP)
- A redundant switched topology ***may*** (STP disabled) cause broadcast storms, multiple frame copies, and MAC address table instability problems.

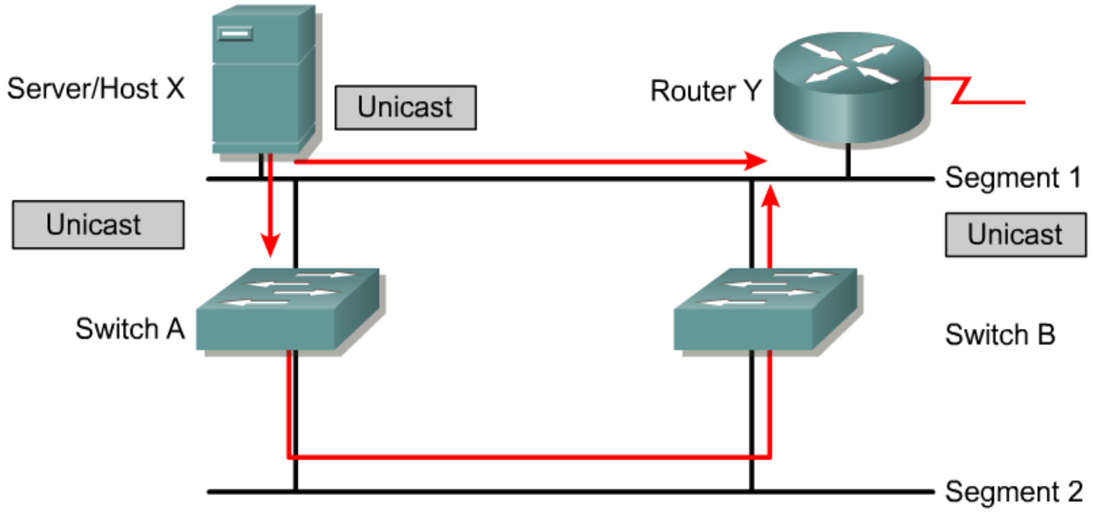
Broadcast Storm



A state in which a message that has been broadcast across a network results in even more responses, and each response results in still more responses in a snowball effect.
www.webopedia.com

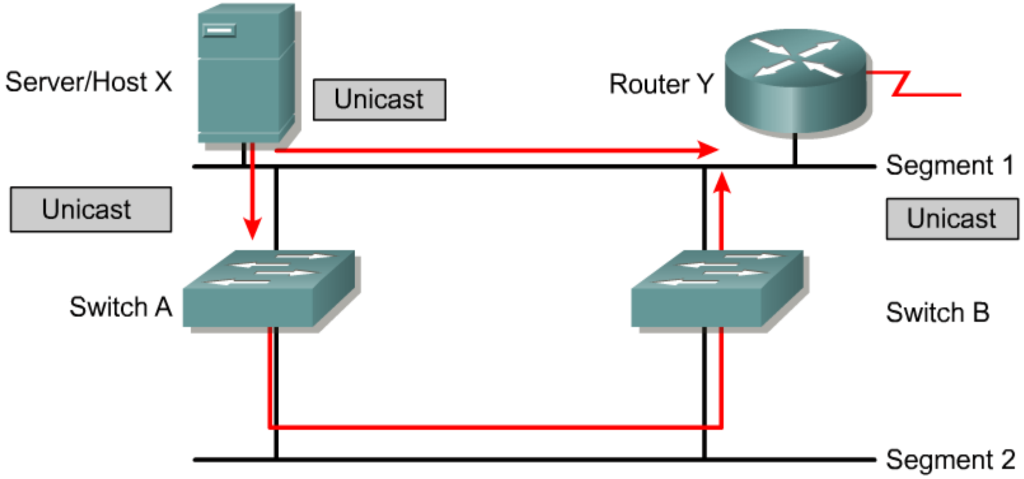
- Broadcasts and multicasts can cause problems in a switched network.
- If Host X sends a broadcast, like an ARP request for the Layer 2 address of the router, then Switch A will forward the broadcast out all ports.
- Switch B, being on the same segment, also forwards all broadcasts.
- Switch B sees all the broadcasts that Switch A forwarded and Switch A sees all the broadcasts that Switch B forwarded.
- Switch A sees the broadcasts and forwards them.
- Switch B sees the broadcasts and forwards them.
- The switches continue to propagate broadcast traffic over and over.
- This is called a broadcast storm.

Multiple frame transmissions



- In a redundant switched network it is possible for an end device to receive multiple frames.
- Assume that the MAC address of Router Y has been timed out by both switches.
- Also assume that Host X still has the MAC address of Router Y in its ARP cache and sends a unicast frame to Router Y.

Multiple frame transmissions



(Some changes to curriculum)

The router receives the frame because it is on the same segment as Host X.

Switch A does not have the MAC address of the Router Y and will therefore flood the frame out its ports. (Segment 2)

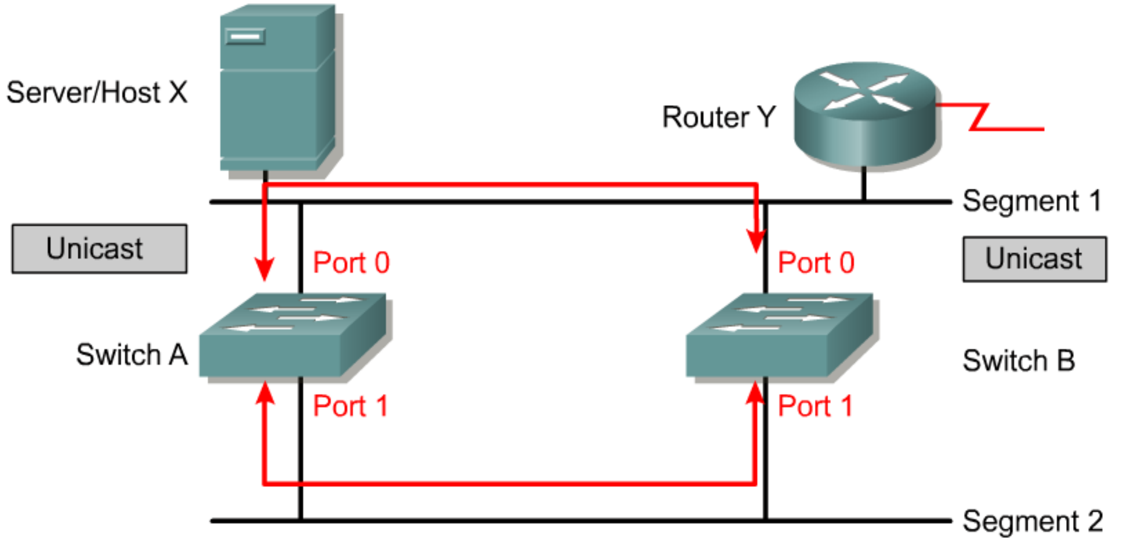
Switch B also does not know which port Router Y is on.

Note: Switch B will forward the the unicast onto Segment 2, creating multiple frames on that segment.

After Switch B receives the frame from Switch A , it then floods the frame it received causing Router Y to receive multiple copies of the same frame.

This is a causes of unnecessary processing in all devices.

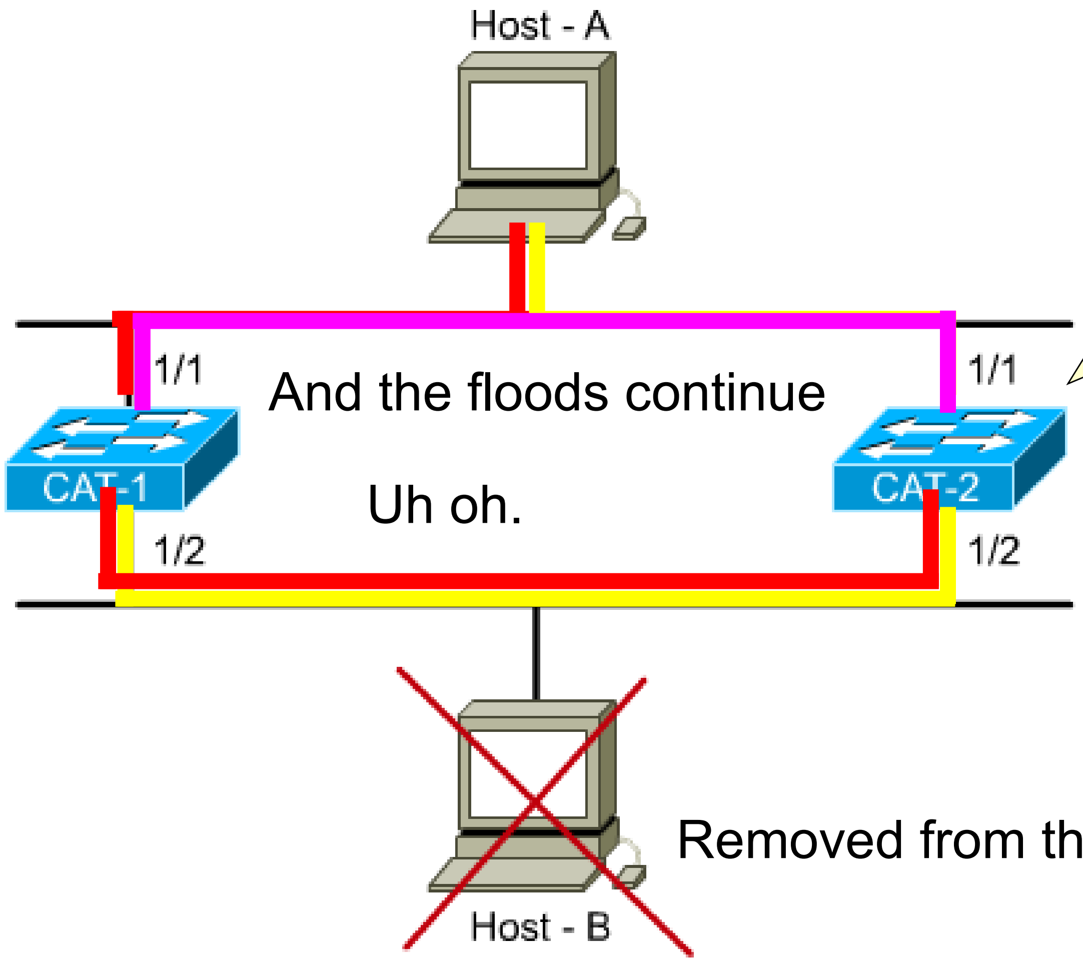
Media access control database instability



- In a redundant switched network it is possible for switches to learn the wrong information.
- A switch can incorrectly learn that a MAC address is on one port, when it is actually on a different port.
- Host X sends a frame directed to Router Y.
- Switches A and B learn the MAC address of Host X on port 0.
- The frame to Router Y is flooded on port 1 of both switches.
- Switches A and B see this information on port 1 and incorrectly learn the MAC address of Host X on port 1.

Layer 2 Loops - Flooded unicast frames

Where's Host B? FLOOD



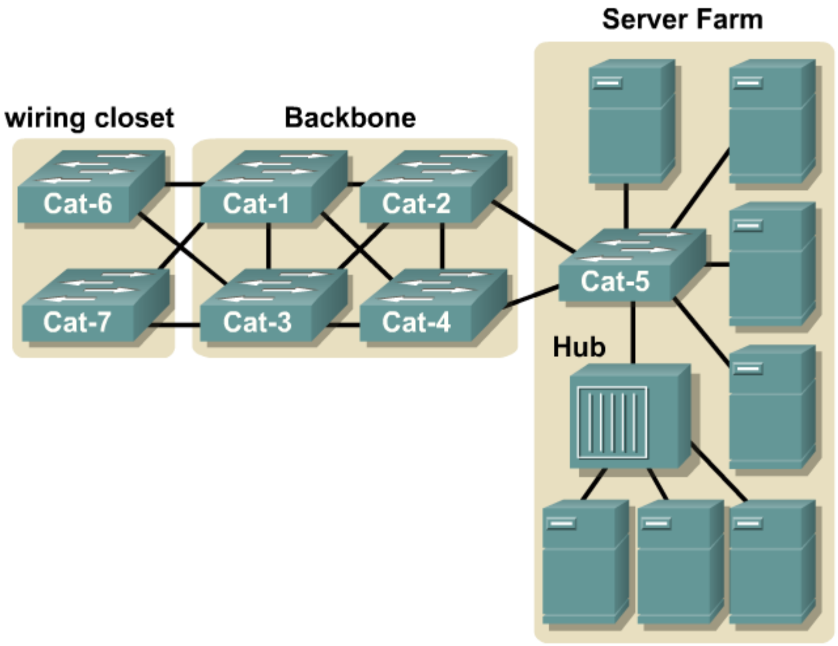
And the floods continue

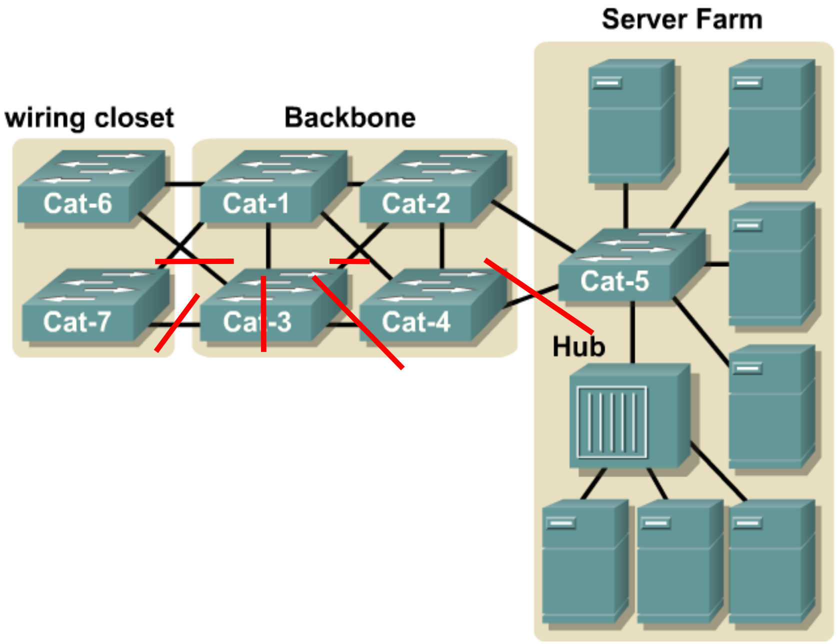
Uh oh.

Where's Host B? FLOOD

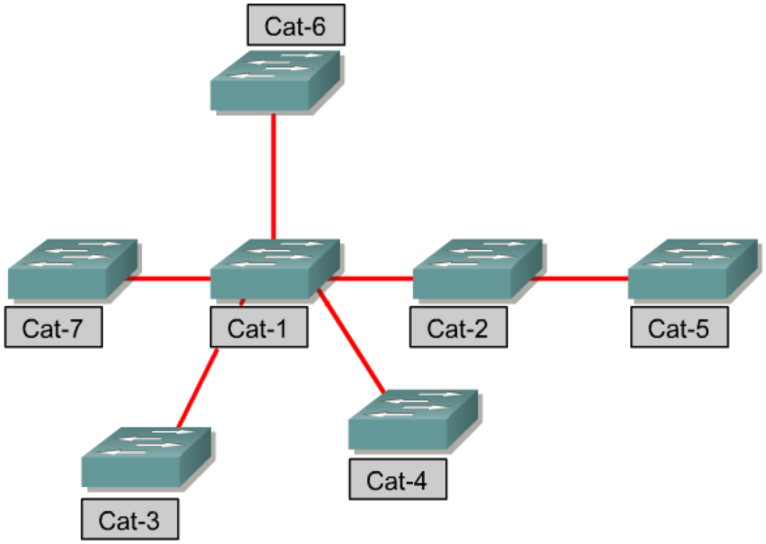
Removed from the network

Redundant topology and spanning tree

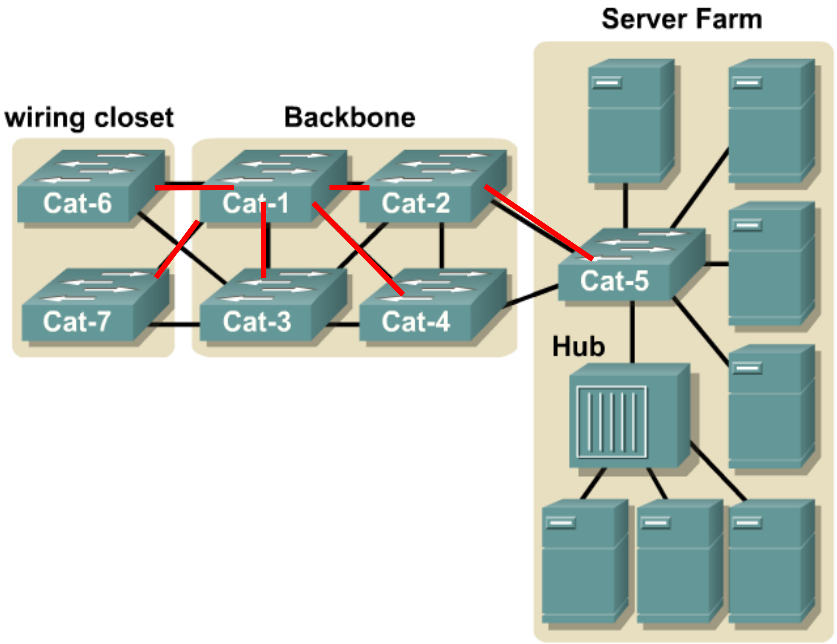


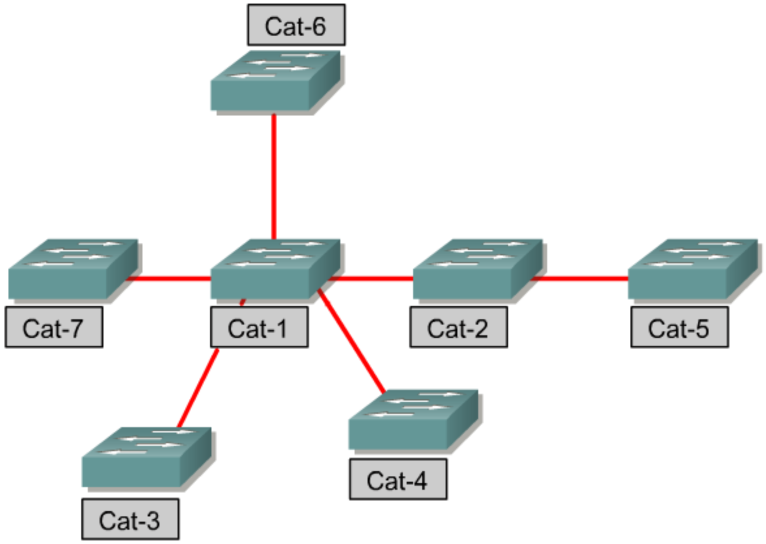


- Unlike IP, in the Layer 2 header there is no Time To Live (TTL).
- The solution is to allow physical loops, but create a loop free logical topology.
- The loop free logical topology created is called a tree.
- This topology is a star or extended star logical topology, the spanning tree of the network.



Redundant topology and spanning tree

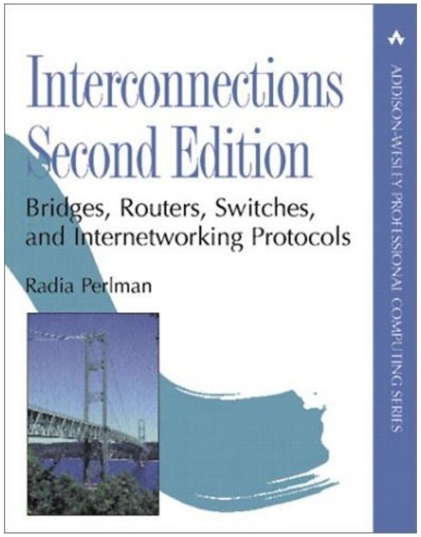


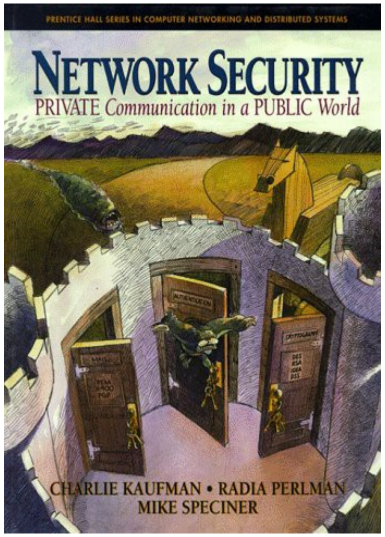


- It is a spanning tree because all devices in the network are reachable or spanned.
- The algorithm used to create this loop free logical topology is the **spanning-tree algorithm**.
- This algorithm can take a relatively long time to converge.

Spanning-Tree Protocol (STP)

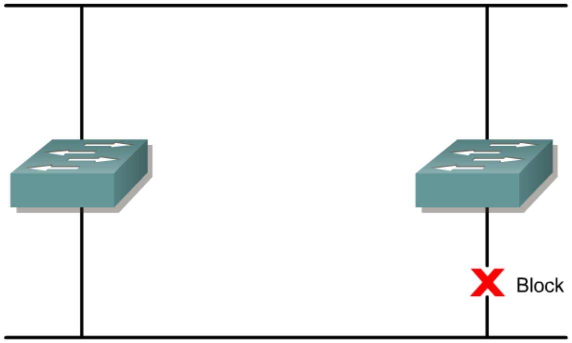
Radia Perlman, networking hero!


Interconnections
Second Edition
Bridges, Routers, Switches, and Internetworking Protocols
Radia Perlman
ADDISON-WESLEY PROFESSIONAL COMPUTING SERIES

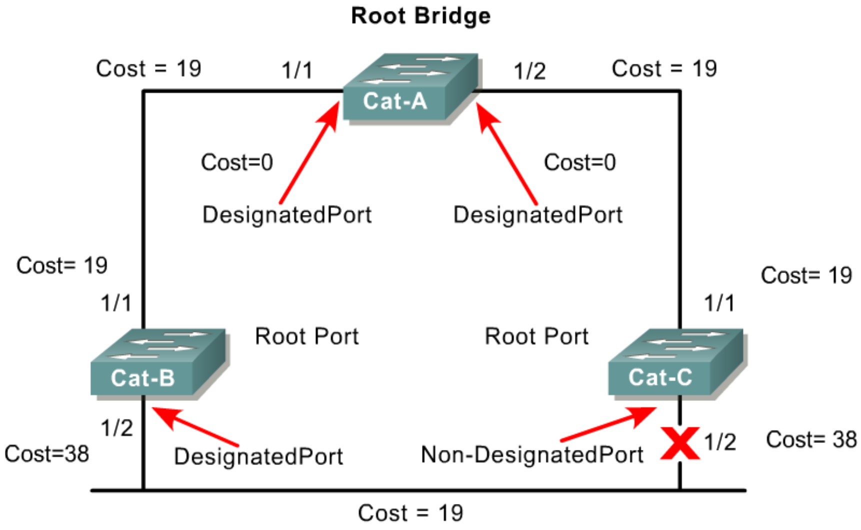

PRENTICE HALL SERIES IN COMPUTER NETWORKING AND DISTRIBUTED SYSTEMS
NETWORK SECURITY
PRIVATE Communication in a PUBLIC World
CHARLIE KAUFMAN • RADIA PERLMAN
MIKE SPECINER

- Ethernet bridges and switches can implement the **IEEE 802.1D Spanning-Tree Protocol** and use the spanning-tree algorithm to **construct a loop free shortest path network**.
- Radia Perlman “is the inventor of the spanning tree algorithm used by bridges (switches), and the mechanisms that make link state routing protocols such as IS-IS (which she designed) and OSPF (which adopted many of the ideas) stable and efficient. Her thesis on sabotage-proof networks is well-known in the security community.”
  http://www.equipecom.com/radia.html

Spanning-Tree Protocol (STP)



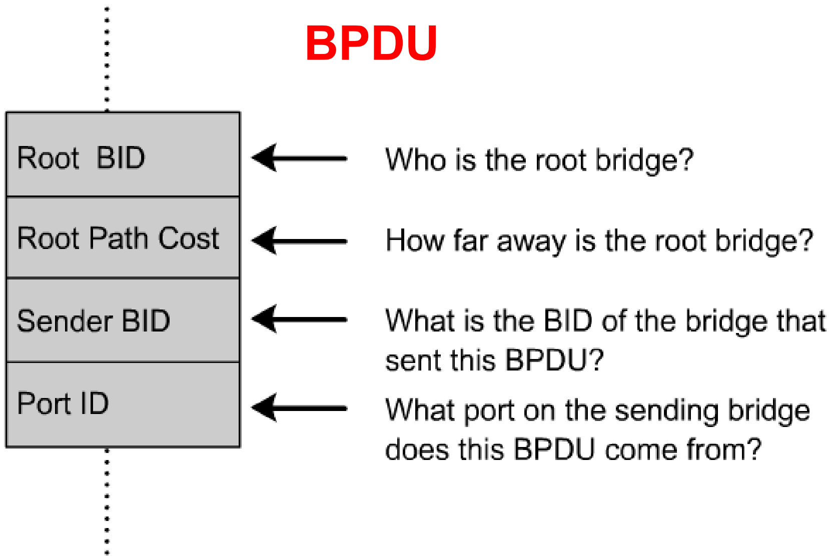
Link Speed	Cost(Revised IEEE Spec)	Cost (Previous IEEE Spec)
10 Gbps	2	1
1 Gbps	4	1
100 Mbps	19	10
10 Mbps	100	100



We will see how this works in a moment.

- Shortest path is based on cumulative link costs.
- Link costs are based on the speed of the link.
- The Spanning-Tree Protocol establishes a root node, called the root bridge.
- The Spanning-Tree Protocol constructs a topology that has one path for reaching every network node.
- The resulting tree originates from the **root bridge**.
- **Redundant links** that are not part of the shortest path tree are **blocked**.

Spanning-Tree Protocol (STP)



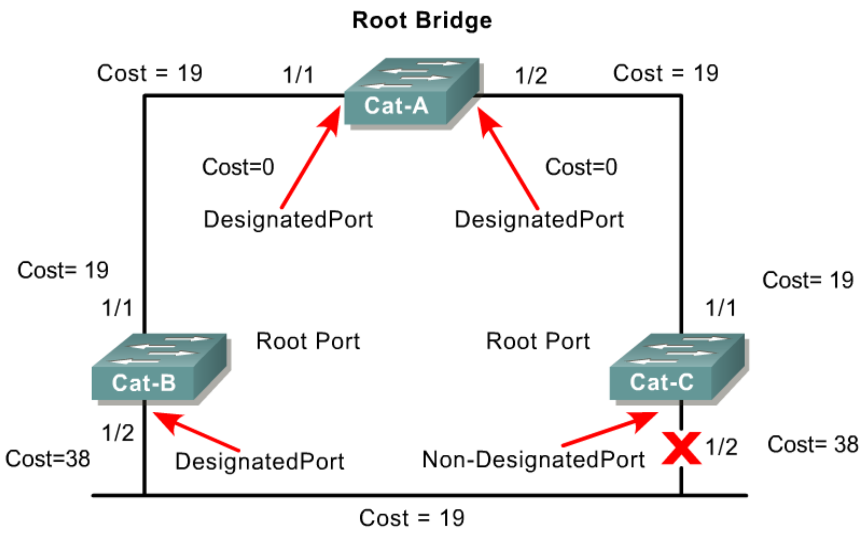
BPDU

Root BID ← Who is the root bridge?

Root Path Cost ← How far away is the root bridge?

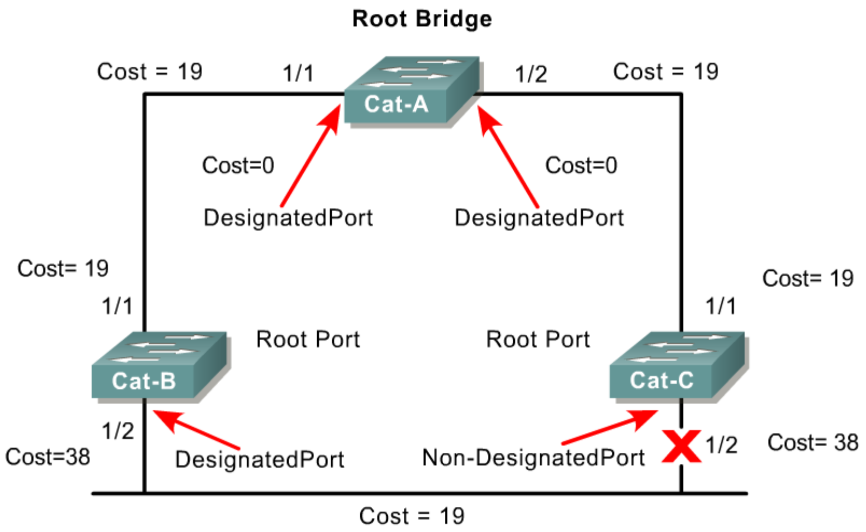
Sender BID ← What is the BID of the bridge that sent this BPDU?

Port ID ← What port on the sending bridge does this BPDU come from?



- It is because certain paths are blocked that a loop free topology is possible.
- Data frames received on blocked links are dropped.
- The Spanning-Tree Protocol requires network devices to exchange messages to detect bridging loops.
- Links that will cause a loop are put into a blocking state.
- topology, is called a **Bridge Protocol Data Unit (BPDU).**
- BPDUs continue to be received on blocked ports.
- This ensures that if an active path or device fails, a new spanning tree can be calculated.

Spanning-Tree Protocol (STP)



BPDUs contain enough information so that all switches can do the following:

Select a **single switch that will act as the root** of the spanning tree

Calculate the **shortest path from itself to the root switch**

Designate one of the switches as the closest one to the root, for each LAN segment. This bridge is called the “**designated switch**”.

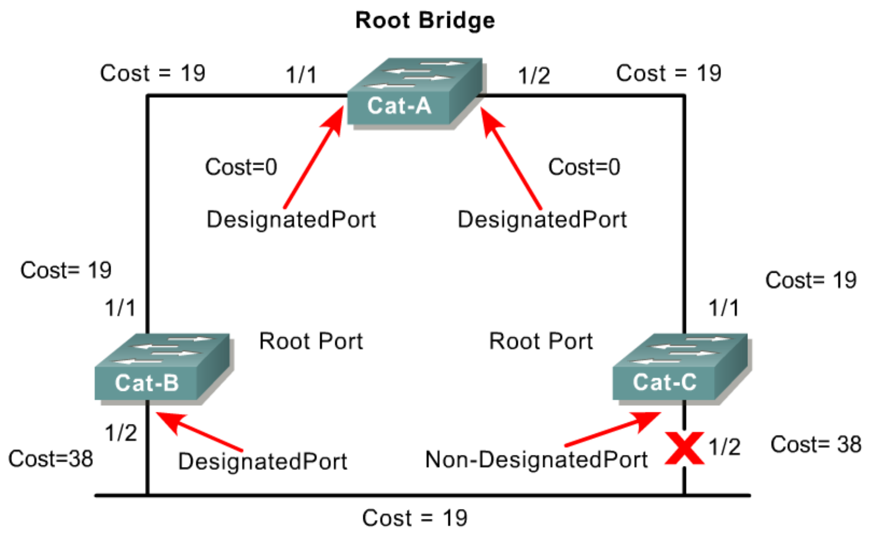
> The designated switch handles all communication from that LAN towards the root bridge.

Choose one of its ports as its root port, for each non-root switch.

> This is the interface that gives the best path to the root switch.

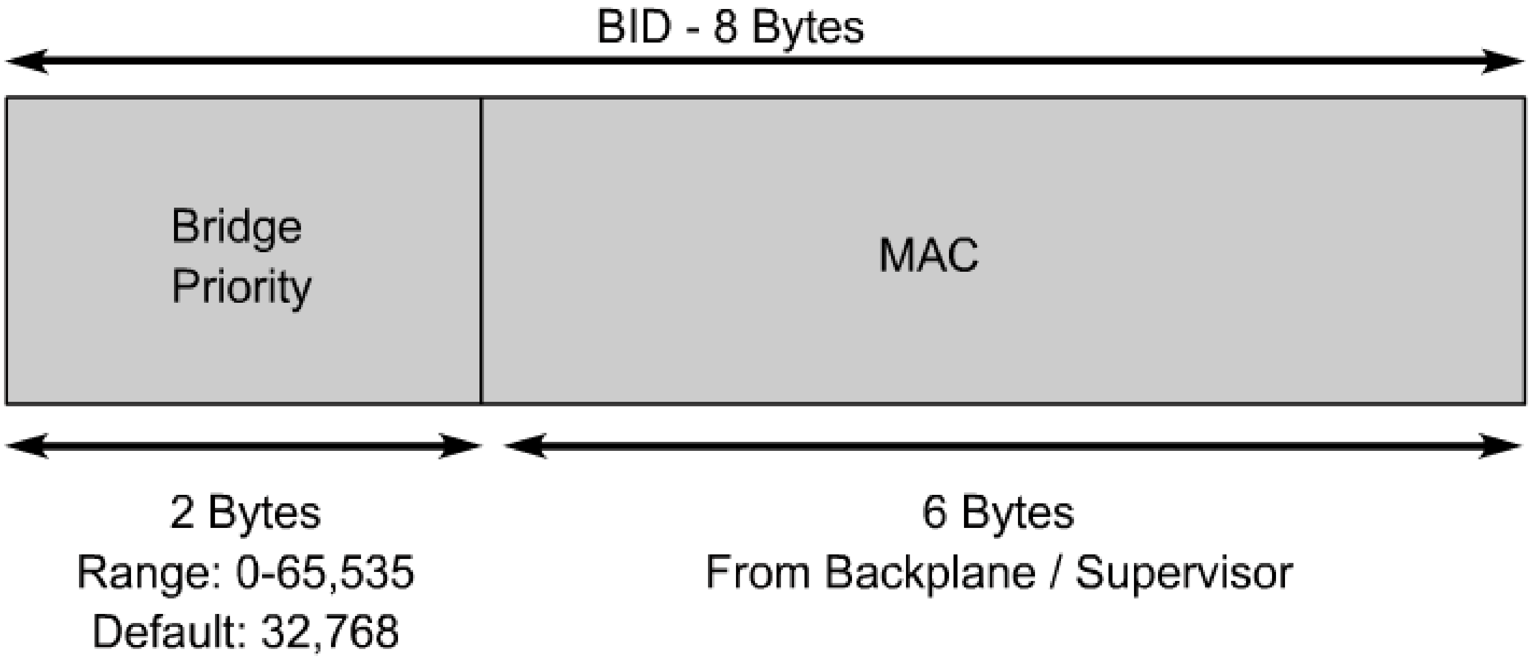
Select ports that are part of the spanning tree, the designated ports. Non-designated ports are blocked.

Two Key Concepts: BID and Path Cost



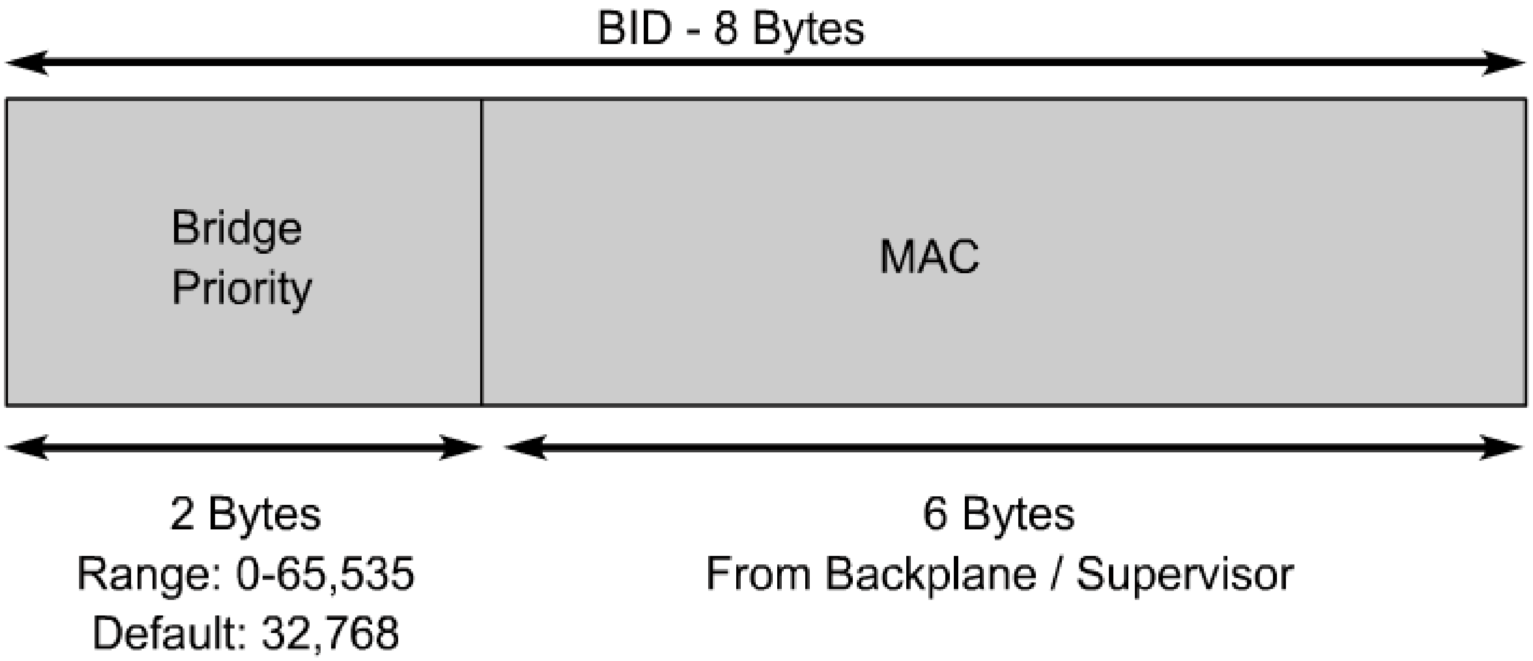
- STP executes an algorithm called Spanning Tree Algorithm (STA).
- STA chooses a reference point, called a root bridge, and then determines the available paths to that reference point.
  - If more than two paths exists, STA picks the best path and blocks the rest
- STP calculations make extensive use of two key concepts in creating a loop-free topology:
  - **Bridge ID**
  - **Path Cost**

Bridge ID (BID)



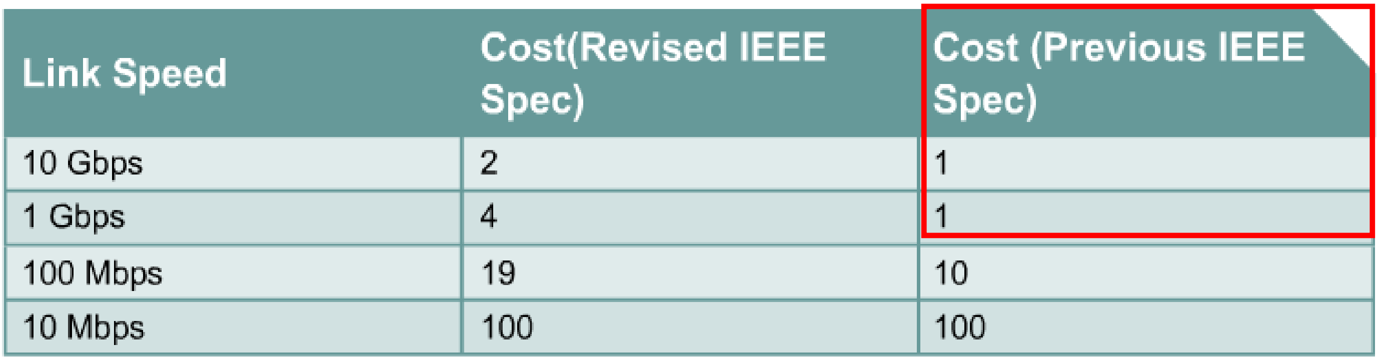
- **Bridge ID (BID)** is used to identify each bridge/switch.
- The BID is used in determining the center of the network, in respect to STP, known as the root bridge.
- Consists of two components:
  - **A 2-byte Bridge Priority**: Cisco switch defaults to **32,768** or 0x8000.
  - **A 6-byte MAC address**

Bridge ID (BID)



- **Bridge Priority** is usually expressed in **decimal format** and the **MAC address** in the BID is usually expressed in **hexadecimal format**.
- BID is used to elect a root bridge (coming)
- **Lowest Bridge ID is the root.**
- If all devices have the same priority, the bridge with the lowest MAC address becomes the root bridge. (Yikes!)

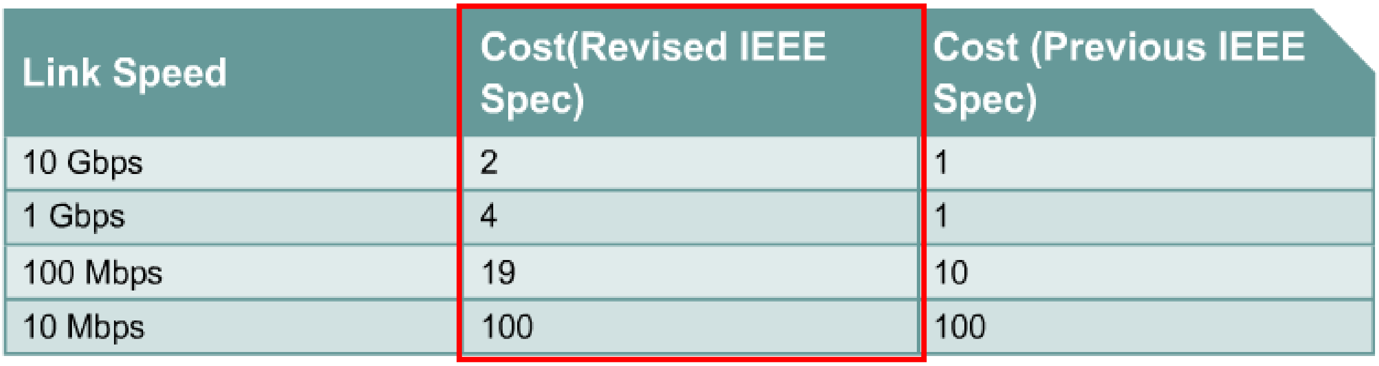
Path Cost



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

- Bridges use the concept of cost to evaluate how close they are to other bridges.
- This will be used in the STP development of a loop-free topology .
- **Originally, 802.1d** defined cost as 1000/bandwidth of the link in Mbps.
  - Cost of 10Mbps link = 100 or 1000/10
  - Cost of 100Mbps link = 10 or 1000/100
  - Cost of 1Gbps link = 1 or 1000/1000
- Running out of room for faster switches including 10 Gbps Ethernet.

Path Cost

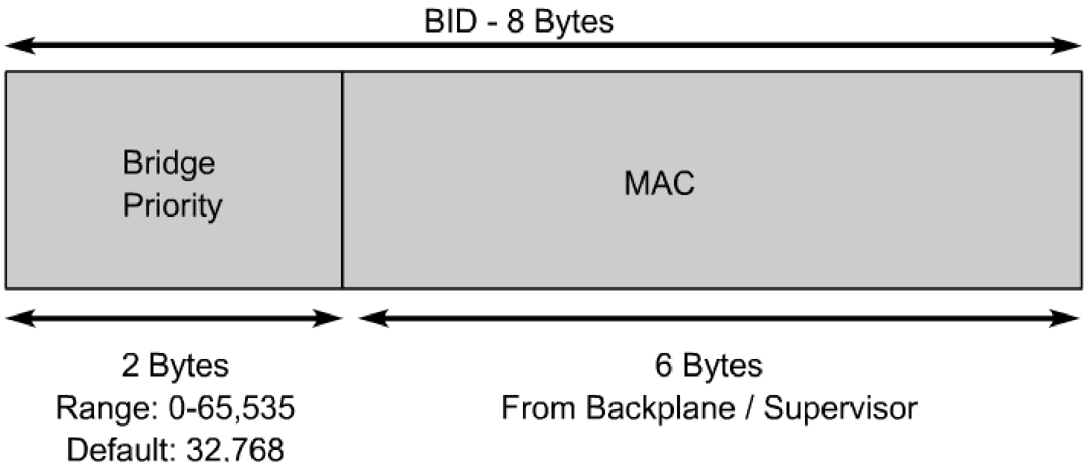


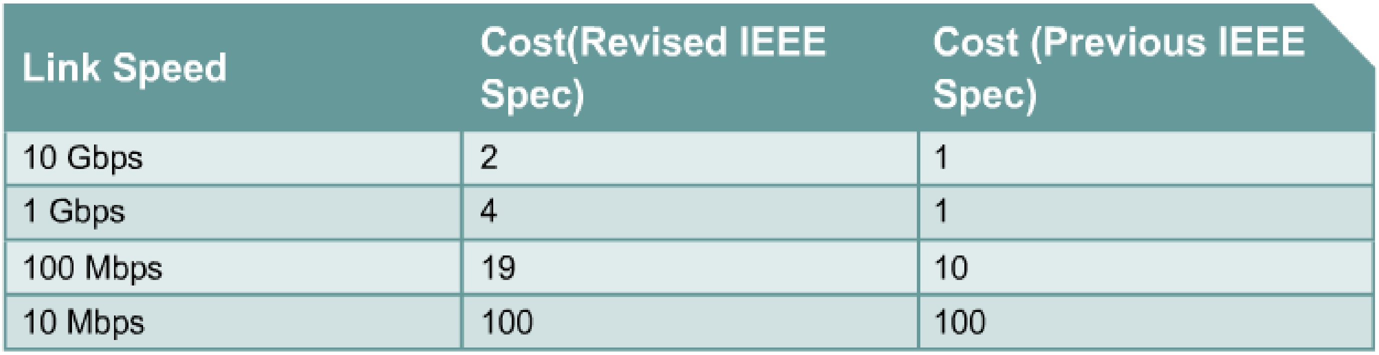
Link Speed	Cost(Revised IEEE Spec)	Cost (Previous IEEE Spec)
10 Gbps	2	1
1 Gbps	4	1
100 Mbps	19	10
10 Mbps	100	100

- IEEE modified the most to use a non-linear scale with the new values of:

–	4 Mbps	250	(cost)
–	10 Mbps	100	(cost)
–	16 Mbps	62	(cost)
–	45 Mbps	39	(cost)
–	100 Mbps	19	(cost)
–	155 Mbps	14	(cost)
–	622 Mbps	6	(cost)
–	1 Gbps	4	(cost)
–	10 Gbps	2	(cost)

Path Cost





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

- You can modify the path cost by modifying the cost of a port.
  - Exercise caution when you do this!
- BID and Path Cost are used to develop a loop-free topology .
- But first the **3 STP Decision Sequence**

Three-Step STP Decision Sequence

- When creating a loop-free topology, STP always uses the same three-step decision sequence whenever it has to assign **Root Ports (RP) or Designated Ports (DP)** :

  Three-Step decision Sequence

  Step 1 - Lowest Path Cost to Root Bridge

  Step 2 - Lowest Sender BID

  Step 3 - Lowest Sender Port ID

- Bridges use Configuration BPDUs during this four-step process.
  - There is another type of BPDU known as Topology Change Notification (TCN) BPDU.

Three-Step STP Decision Sequence

BPDU key concepts:

Bridges save a copy of only the best BPDU seen on every port.

When making this evaluation, it considers all of the BPDUs received on the port, as well as the BPDU that would be sent on that port.

As every BPDU arrives, it is checked against this three-step sequence to see if it is more attractive (lower in value) than the existing BPDU saved for that port.

Only the lowest value BPDU is saved.

Bridges send configuration BPDUs until a more attractive BPDU is received.

Okay, lets see how this is used...

Three Steps of Initial STP Convergence

- The STP algorithm uses three simple steps to converge on a loop-free topology.
- Switches go through three steps for their initial convergence:

STP Convergence

Step 1 Elect one Root Bridge
Step 2 Elect Root Ports
Step 3 Elect Designated Ports

- All STP decisions are based on a the following predetermined sequence:

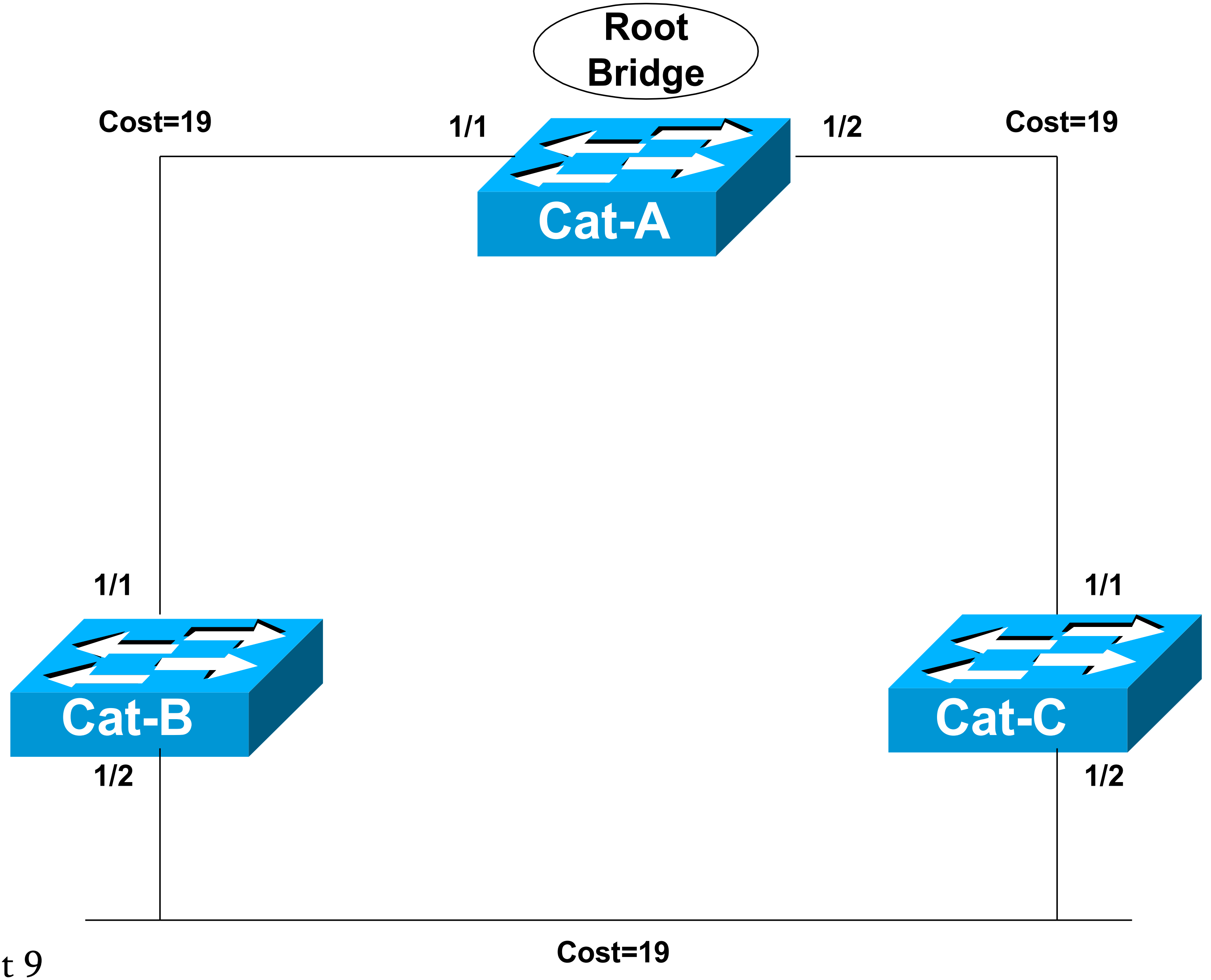
Three-Step decision Sequence

Step 1 - Lowest Path Cost to Root Bridge

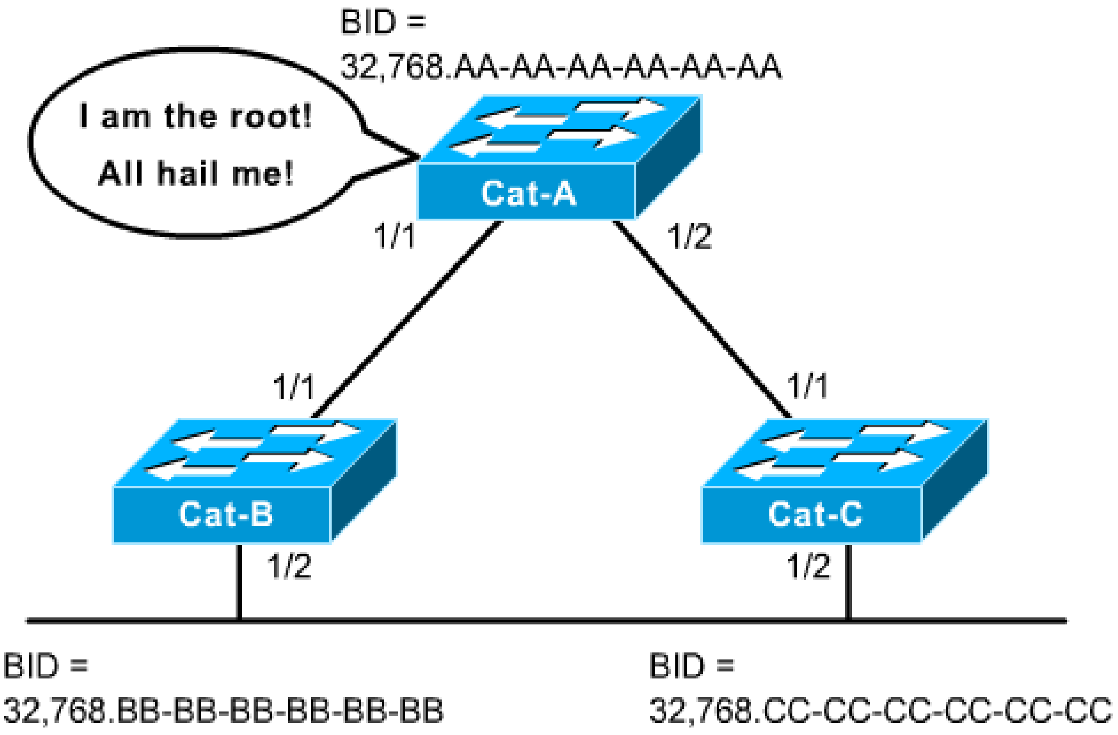
Step 2 - Lowest Sender BID

Step 3 - Lowest Sender Port ID

Step 1   Elect one Root Bridge



Step 1 Elect one Root Bridge



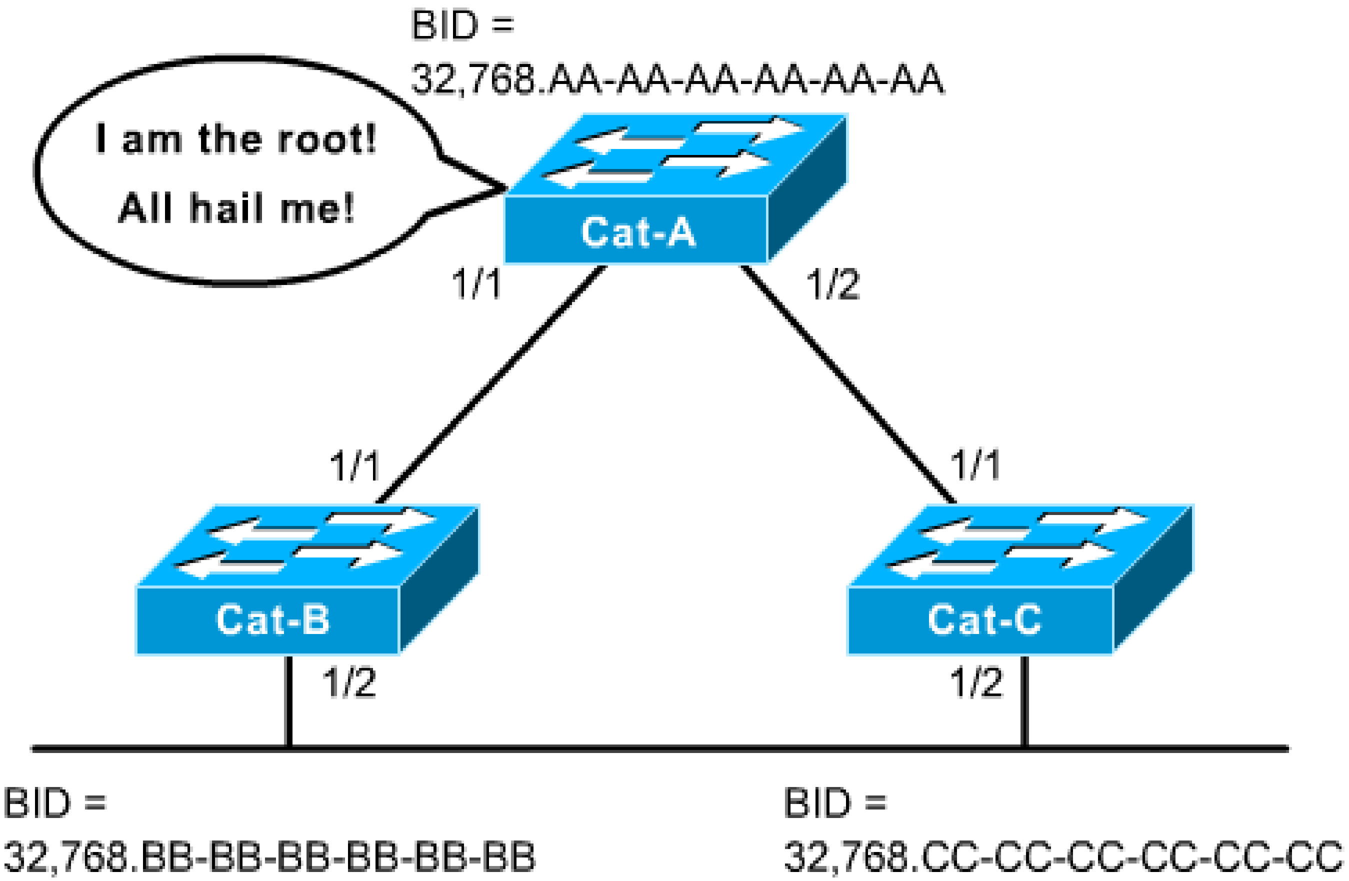
- When the network first starts, all bridges are announcing a chaotic mix of BPDUs.
- All bridges immediately begin applying the four-step sequence decision process.
- Switches need to elect a single Root Bridge.
- Switch with the **lowest BID** wins!

Note: Many texts refer to the term “highest priority” which is the “lowest” BID value.

- This is known as the “Root War.”

Step 1 Elect one Root Bridge

Cat-A has the lowest Bridge MAC Address, so it wins the Root War!



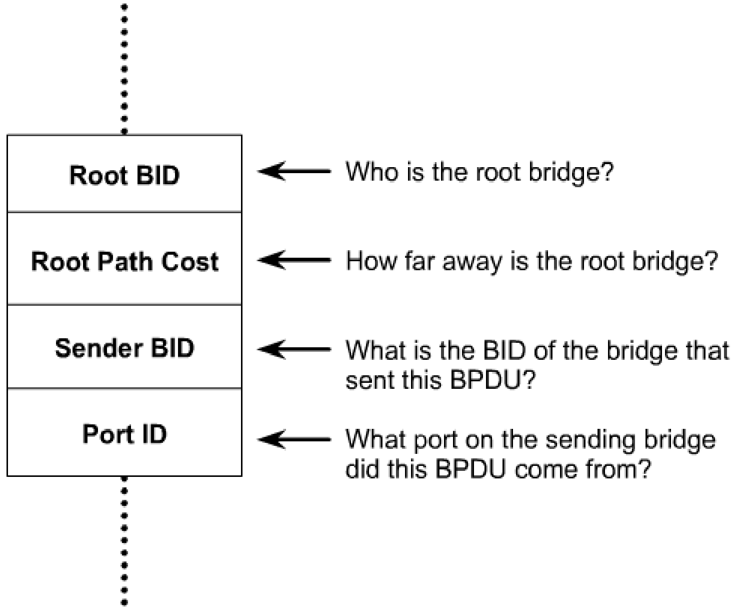
All 3 switches have the same default Bridge Priority value of 32,768

Step 1   Elect one Root Bridge

BPDU

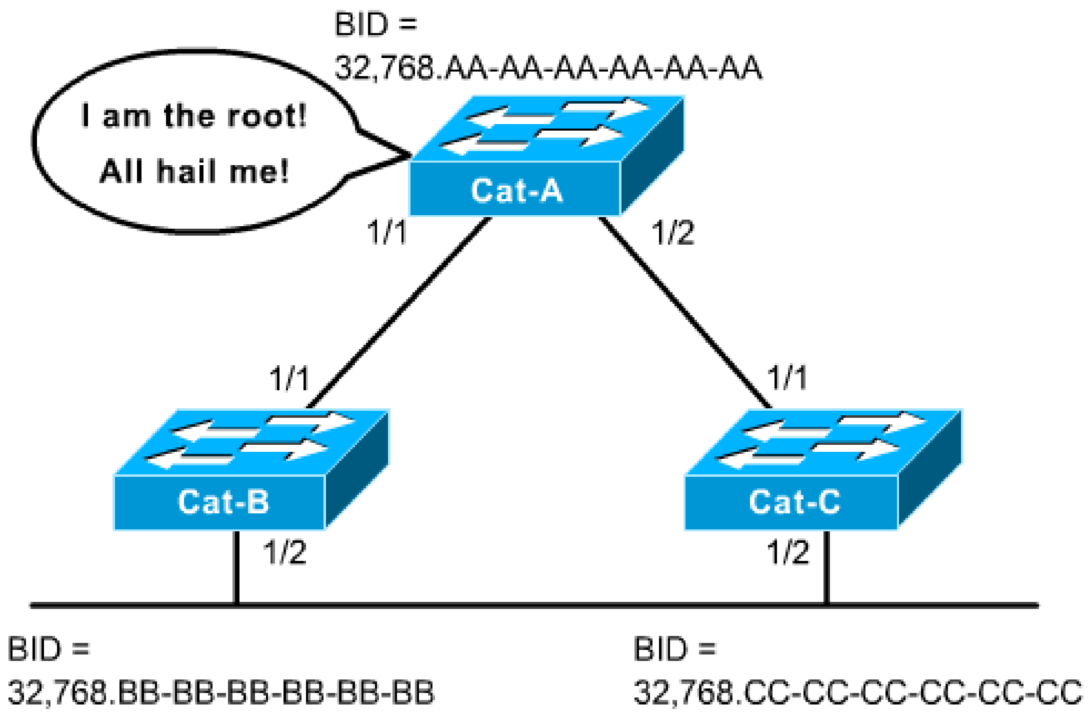
```
802.3 Header
   Destination:   01:80:C2:00:00:00   Mcast 802.1d Bridge
     group
   Source:        00:D0:C0:F5:18:D1
   LLC Length:    38
802.2 Logical Link Control (LLC) Header
   Dest. SAP:     0x42   802.1 Bridge Spanning Tree
   Source SAP:    0x42   802.1 Bridge Spanning Tree
   Command:       0x03   Unnumbered Information
802.1 - Bridge Spanning Tree
   Protocol Identifier:  0
   Protocol Version ID:  0
   Message Type:         0  Configuration Message
   Flags:                %00000000
   Root Priority/ID:     0x8000/ 00:D0:C0:F5:18:C0
   Cost Of Path To Root: 0x00000000  (0)
   Bridge Priority/ID:   0x8000/ 00:D0:C0:F5:18:C0
   Port Priority/ID:     0x80/ 0x1D
   Message Age:          0/256 seconds  (exactly 0  seconds)
   Maximum Age:          5120/256 seconds  (exactly 20
     seconds)
   Hello Time:           512/256 seconds  (exactly 2
     seconds)
   Forward Delay:        3840/256 seconds  (exactly 15
     seconds)
```

Its all done with BPDUs!


Root BID
Who is the root bridge?
Root Path Cost
How far away is the root bridge?
Sender BID
What is the BID of the bridge that sent this BPDU?
Port ID
What port on the sending bridge did this BPDU come from?

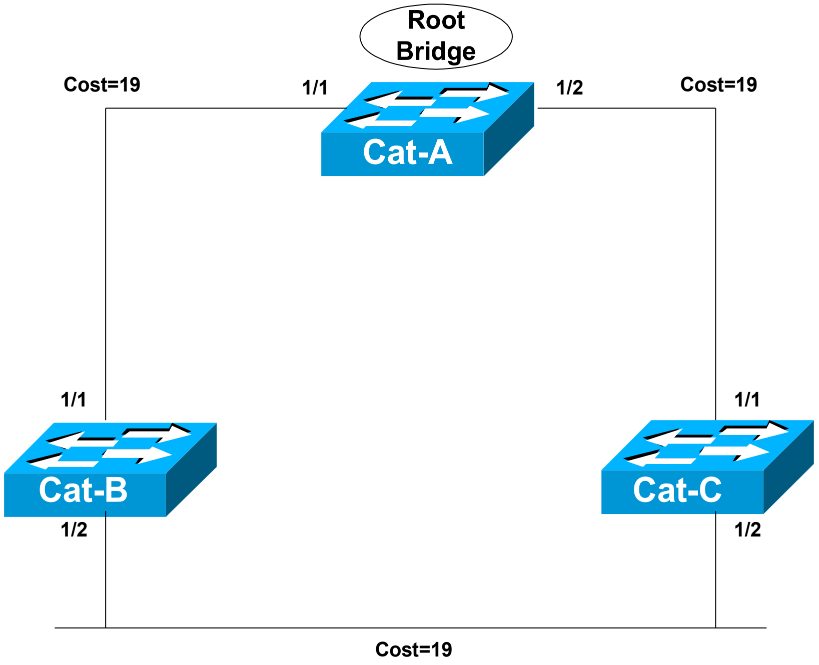
Configuration BPDUs are sent every 2 seconds by default.

Step 1   Elect one Root Bridge



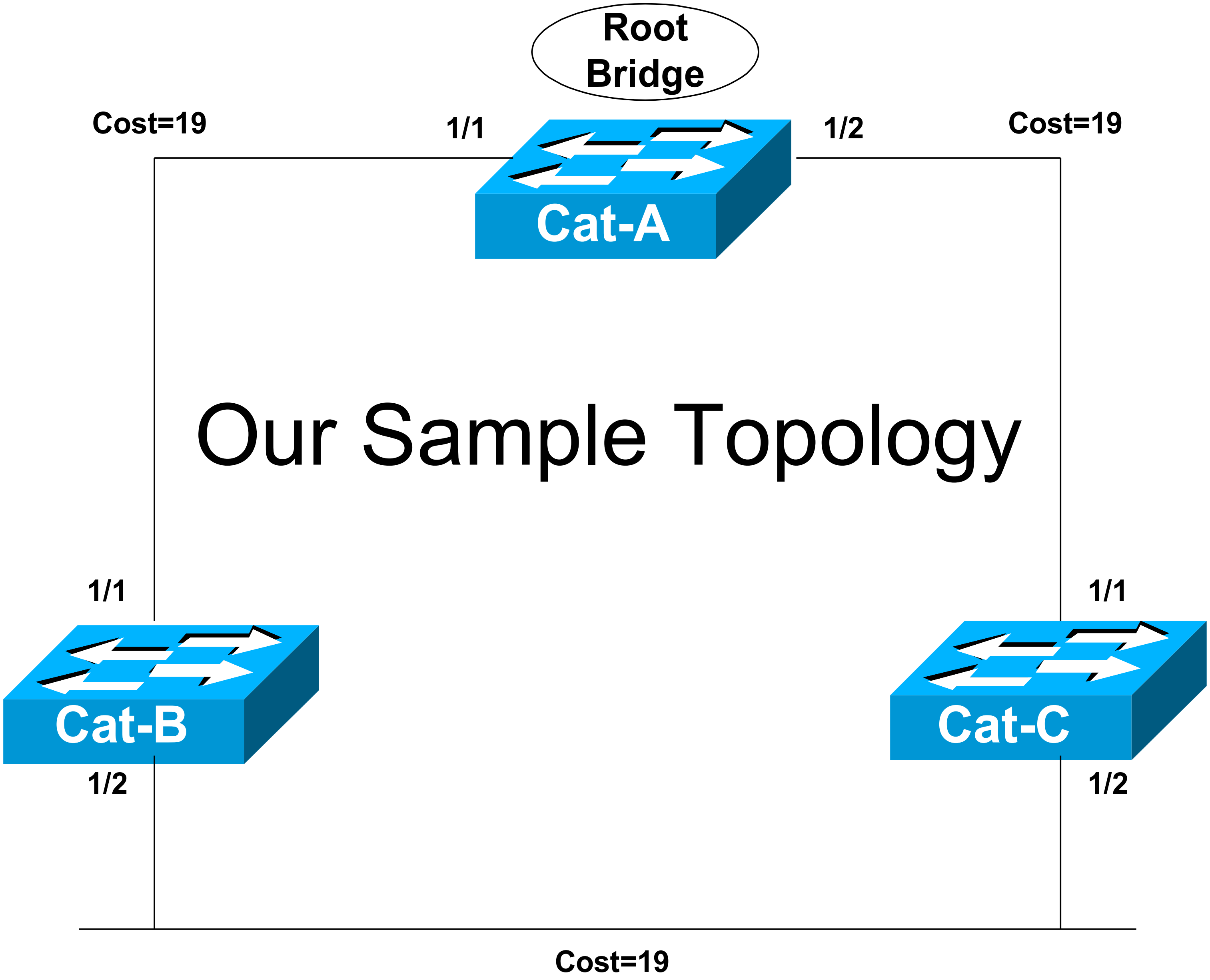
- At the beginning, all bridges assume they are the center of the universe and declare themselves as the Root Bridge, by placing its own BID in the Root BID field of the BPDU.
- Once all of the switches see that Cat-A has the lowest BID, they are all in agreement that Cat-A is the Root Bridge.

Step 2 Elect Root Ports

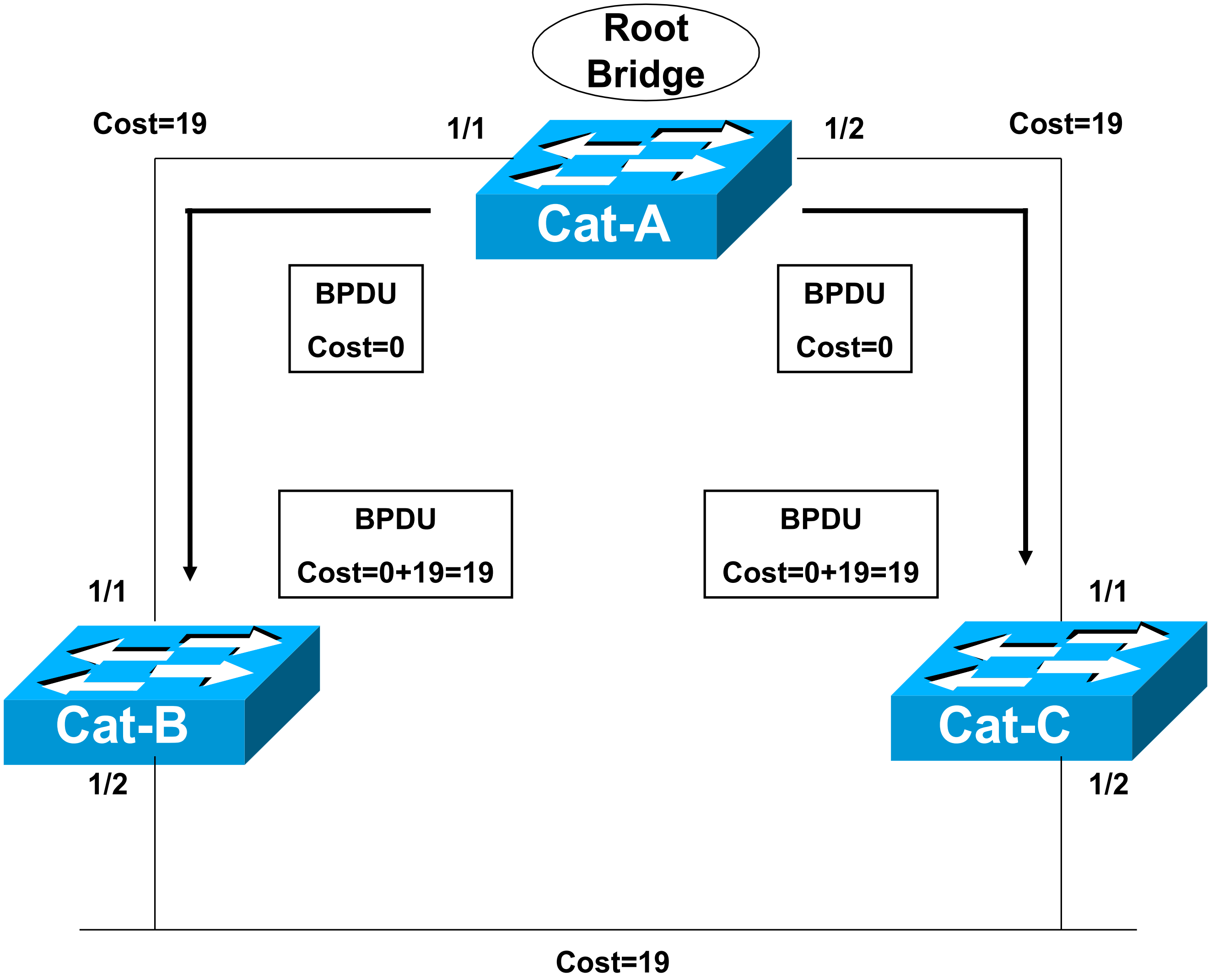


- Now that the Root War has been won, switches move on to selecting **Root Ports**.
- A bridge's **Root Port** is the ***port closest to the Root Bridge***.
- Bridges use the **cost** to determine closeness.
- **Every non-Root Bridge will select only one Root Port!**
- Specifically, bridges track the **Root Path Cost**, the cumulative cost of all links to the Root Bridge.

Our Sample Topology



Step 2
Elect Root Ports



Step 1

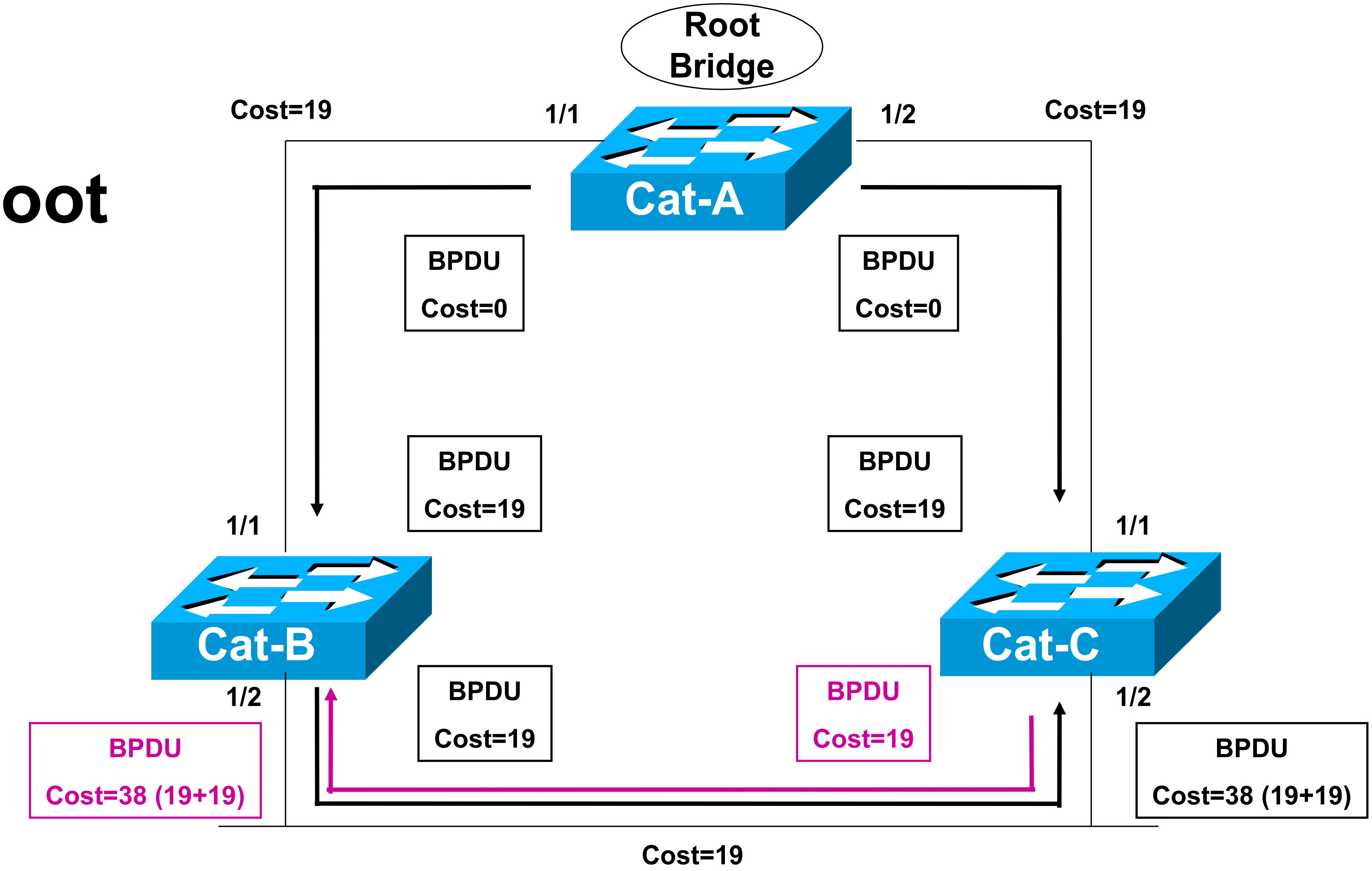
Cat-A sends out BPDUs, containing a Root Path Cost of 0.

Cat-B receives these BPDUs and adds the Path Cost of Port 1/1 to the Root Path Cost contained in the BPDU.

Step 2

Cat-B adds Root Path Cost 0 PLUS its Port 1/1 cost of 19 = 19

Step 2
Elect Root Ports



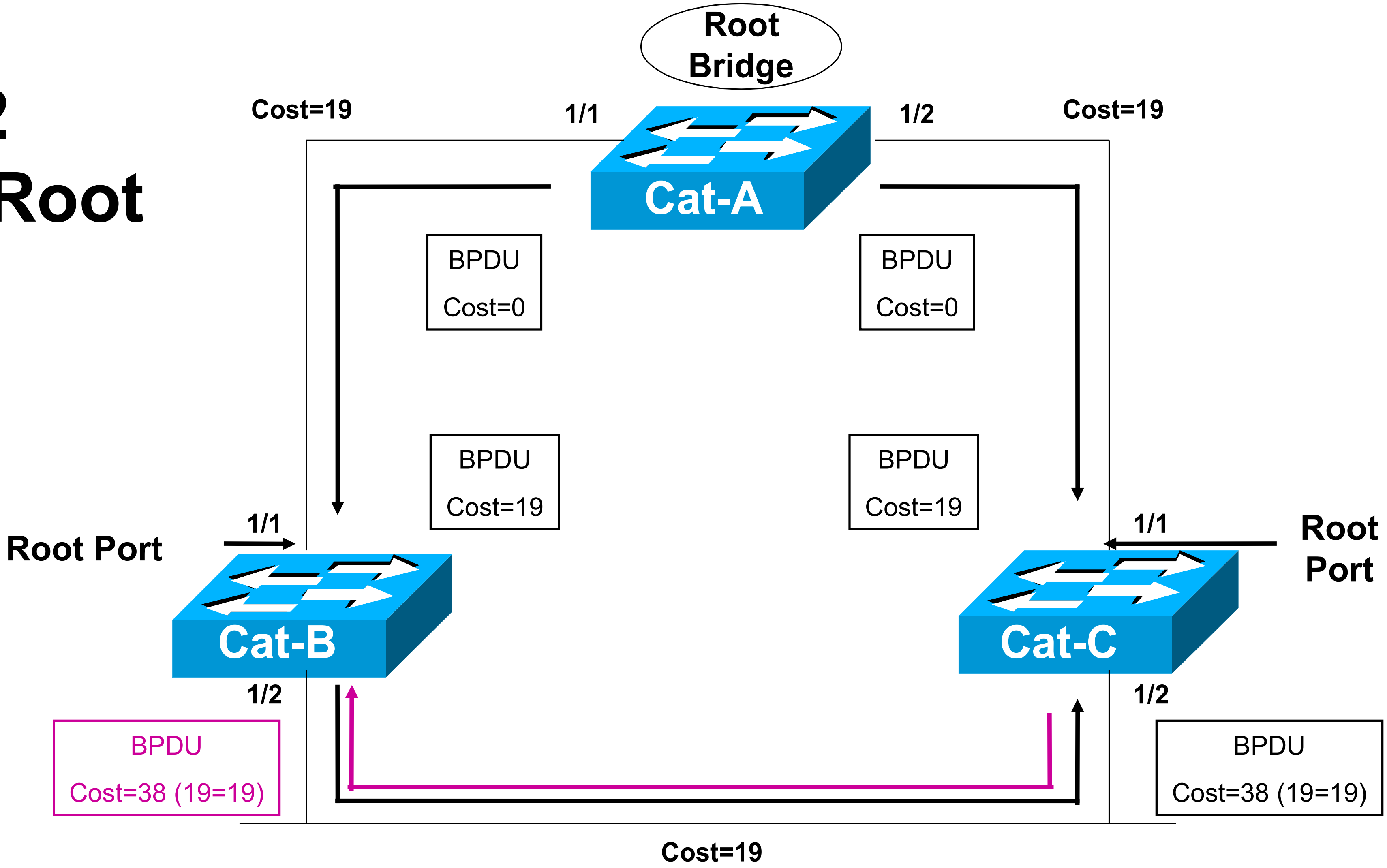
Step 3

Cat-B uses this value of 19 internally and sends BPDUs with a Root Path Cost of 19 out Port 1/2.

Step 4

Cat-C receives the BPDU from Cat-B, and increased the Root Path Cost to 38 (19+19). (Same with Cat-C sending to Cat-B.)

Step 2 Elect Root Ports



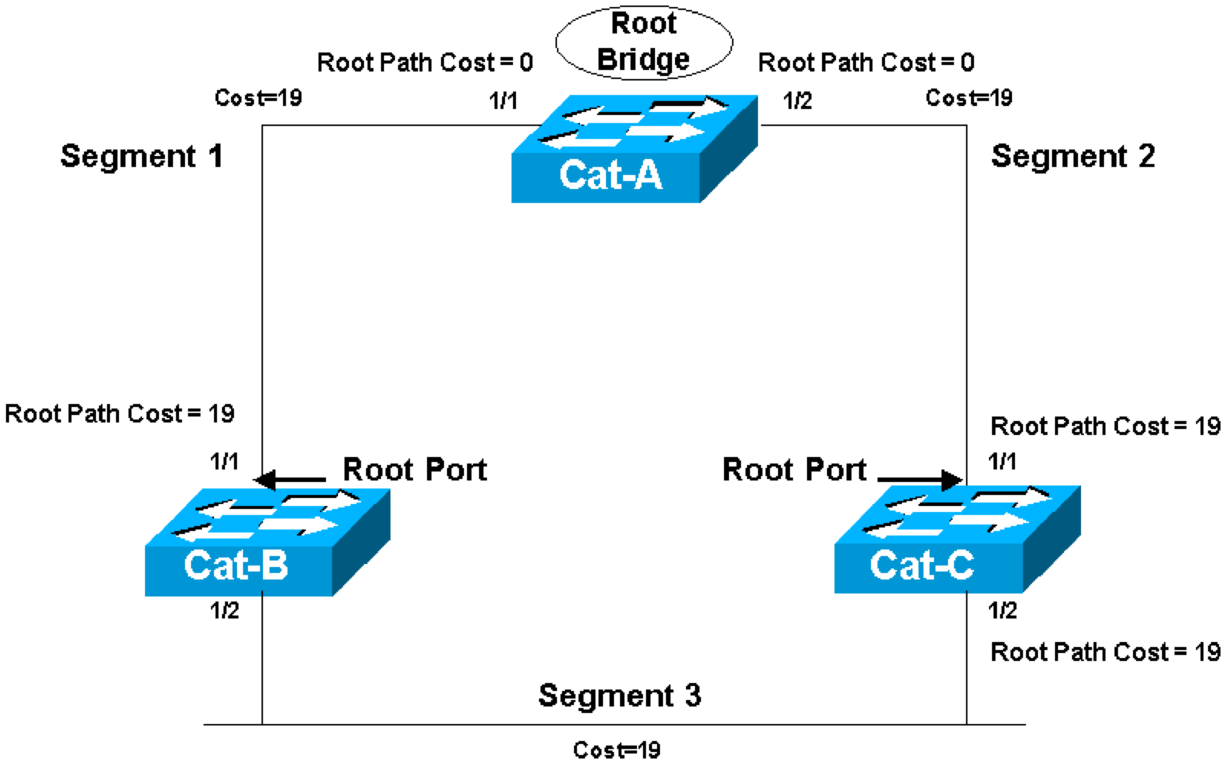
Step 5

Cat-B calculates that it can reach the Root Bridge at a cost of 19 via Port 1/1 as opposed to a cost of 38 via Port 1/2.

Port 1/1 becomes the Root Port for Cat-B, the port closest to the Root Bridge.

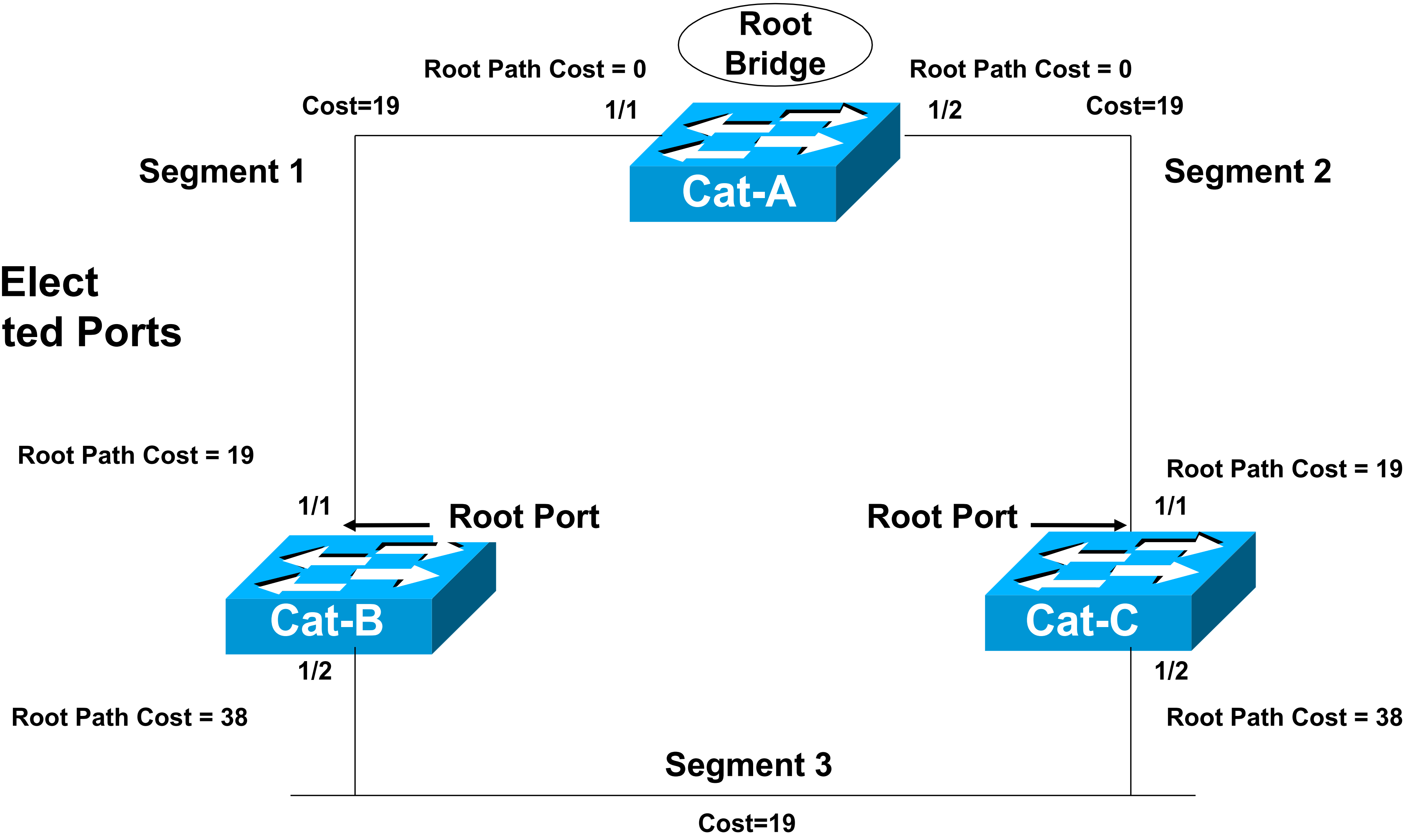
Cat-C goes through a similar calculation. Note: Both Cat-B:1/2 and Cat-C:1/2 save the best BPDU of 19 (its own).

Step 3 Elect Designated Ports

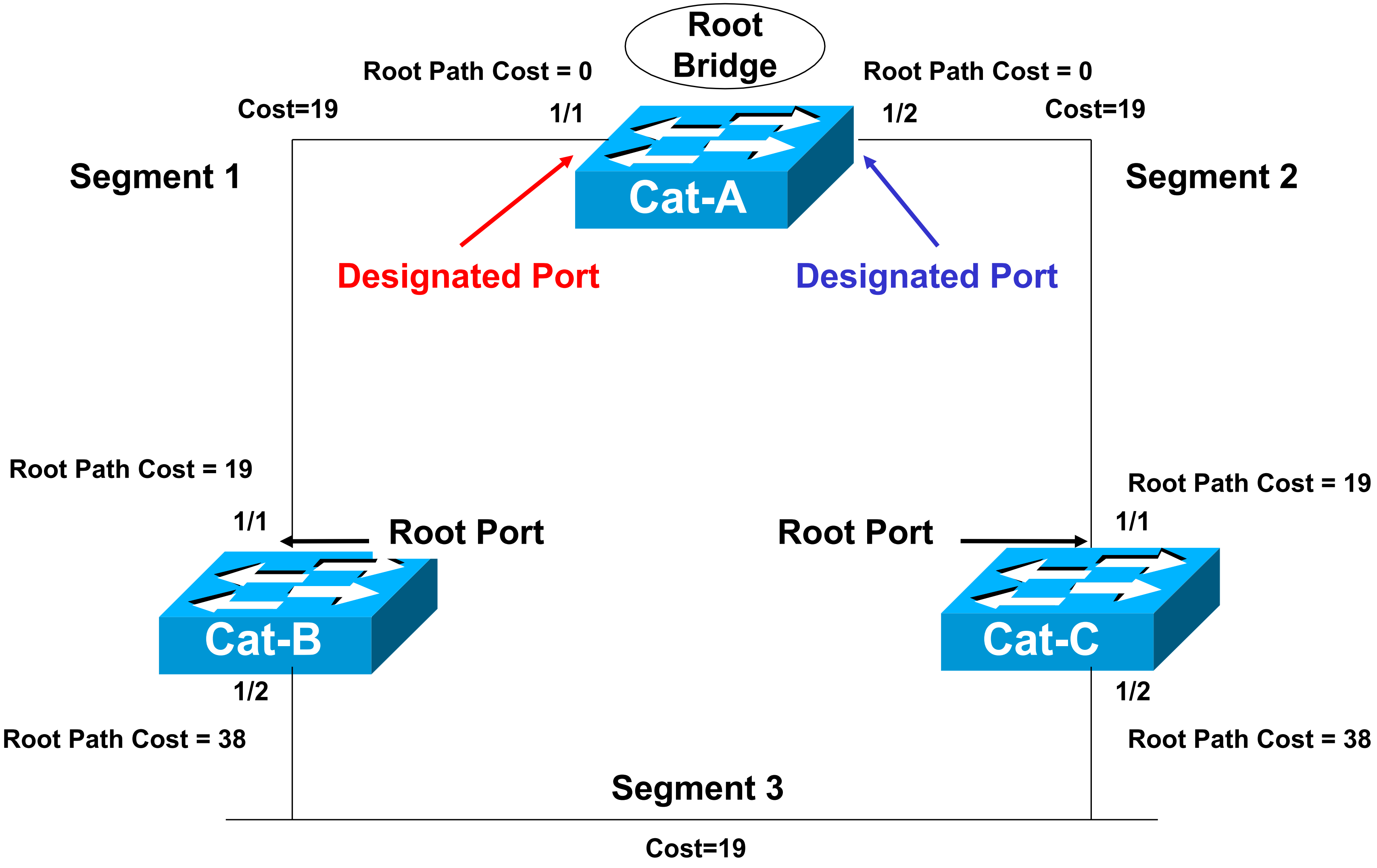


- The loop prevention part of STP becomes evident during this step, electing designated ports.
- A **Designated Port** functions as ***the single bridge port that both sends and receives traffic to and from that segment and the Root Bridge.***
- **Each segment in a bridged network has one Designated Port, chosen based on cumulative Root Path Cost to the Root Bridge.**
- The switch containing the Designated Port is referred to as the **Designated Bridge** for that segment.
- To locate Designated Ports, lets take a look at each segment.
- **Root Path Cost**, the cumulative cost of all links to the Root Bridge.

Step 3 Elect Designated Ports



- **Segment 1**: Cat-A:1/1 has a Root Path Cost = 0 (after all it has the Root Bridge) and Cat-B:1/1 has a Root Path Cost = 19.
- **Segment 2**: Cat-A:1/2 has a Root Path Cost = 0 (after all it has the Root Bridge) and Cat-C:1/1 has a Root Path Cost = 19.
- **Segment 3**: **Cat-B:1/2** has a **Root Path Cost = 38** and **Cat-C:1/2** has a **Root Path Cost = 38**. ***It's a tie!***

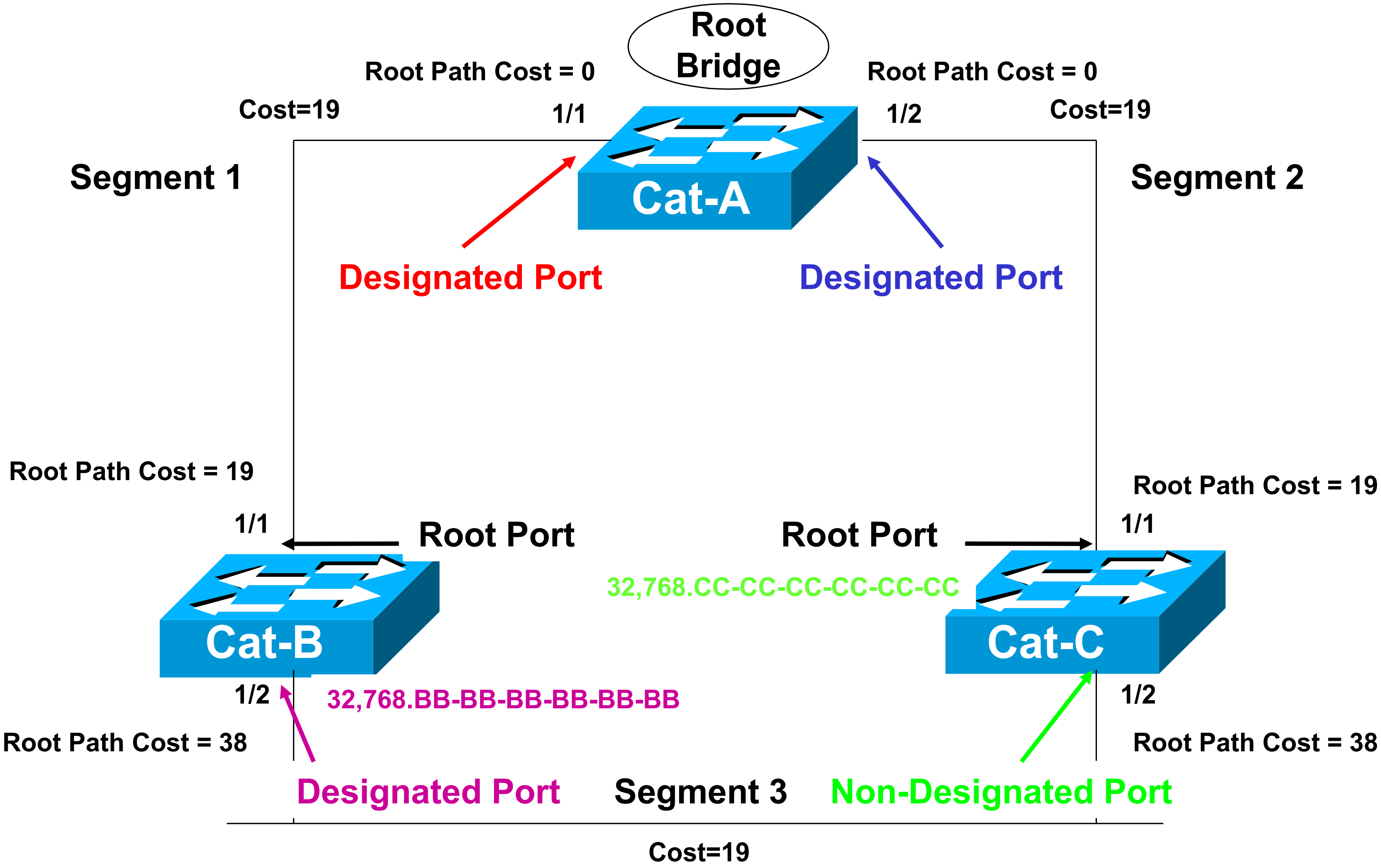


Segment 3

Both Cat-B and Cat-C have a **Root Path Cost of 38**, a tie!

When faced with a tie (or any other determination) STP always uses the three-step decision process:

1. Lowest Path Cost to Root Bridge; 2. Lowest Sender BID; 3. Lowest Sender Port ID



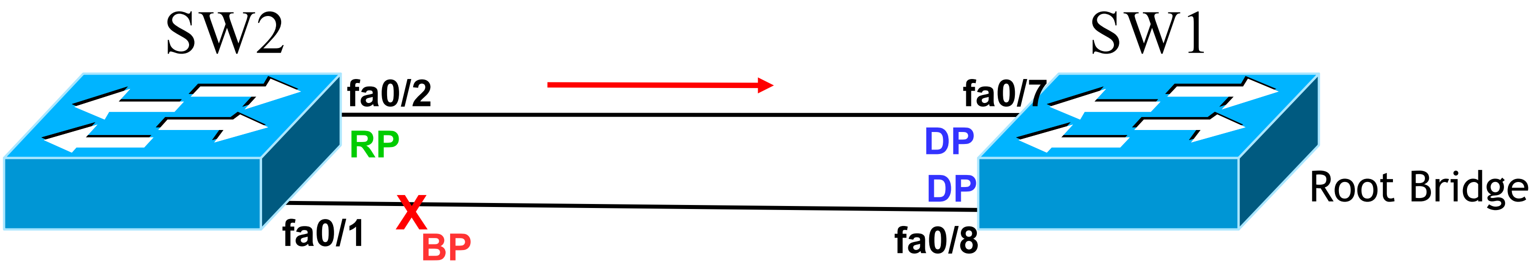
Segment 3 (continued)

1) Root Path Cost for both is 39, so a tie.

3) The sender's BID is lower on Cat-B than Cat-C, so Cat-B:1/2 becomes the **Designated Port for Segment 3**.

Cat-C:1/2 therefore becomes the **non-Designated (Blocking) Port for Segment 3.**

Sender Port ID tie breaker



Assume that SW1 is the Root Bridge. The path cost are the same on both links (say 19). So it's a tie!. Next, both ports on SW2 share the same BridgeID, so that is also a tie!

So, which Port on SW2 will be the Root Port? We might be tempted to say port fa0/1 as it has the lower portID. **WRONG !!!**

In fact we should look at the **Sender** Port ID as last tie-breaker!

The sender is the **Root Bridge** as he is the one **sending BPDUs** !!!

Therefore **fa0/2 on SW2** will be the **Root Port** as it is connected to PortID fa0/7 on the Root bridge which has a lower PortID than fa0/8.

STP Recap

- All switches go through three steps for their initial convergence:

STP Convergence

Step 1 Elect one Root Bridge
Step 2 Elect Root Ports
Step 3 Elect Designated Ports

- All STP decisions are based on a the following predetermined sequence:

Three-Step decision Sequence

Step 1 - Lowest Path Cost to Root Bridge

Step 2 - Lowest Sender BID

Step 3 - Lowest Sender Port ID

- **All ports of Root Bridge** are by **default Designated Ports**.
- **If one end of a link is a Root Port** then the other end **must be a Designated Port.**
- **If a link is unlabeled,** then one end **must be a Designated Port** and the other end **must be a non-Designated (Blocking) Port.**