Lecture (06)
STP (I)

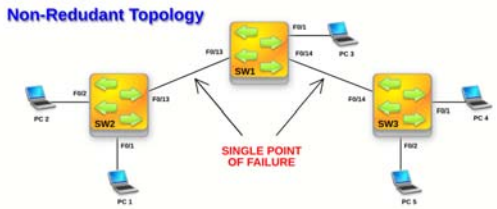
CS
IT

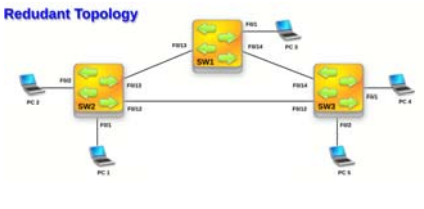
By:

Dr. Ahmed ElShafee

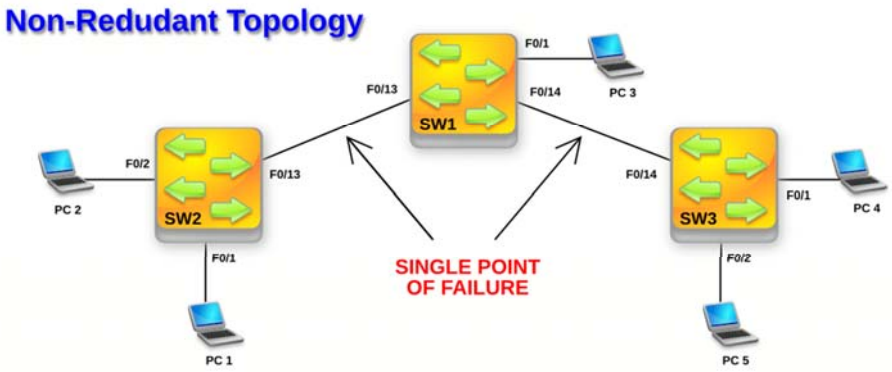
Problem statement

- If your network consists of layer 2 switches that allow computers connect and exchange data, you will need to consider the design that can withstand some types of failure: **Redundant Connections**

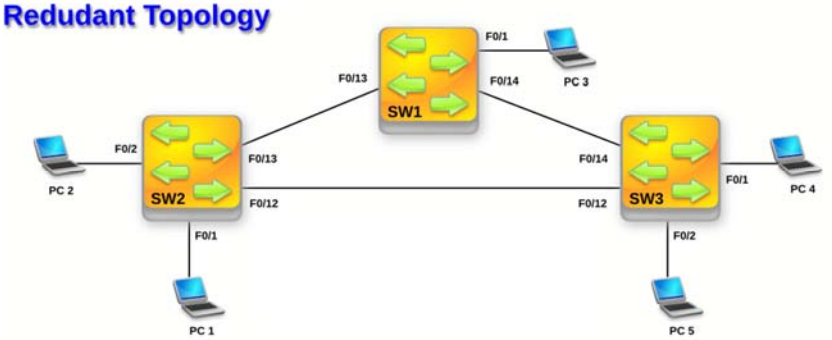




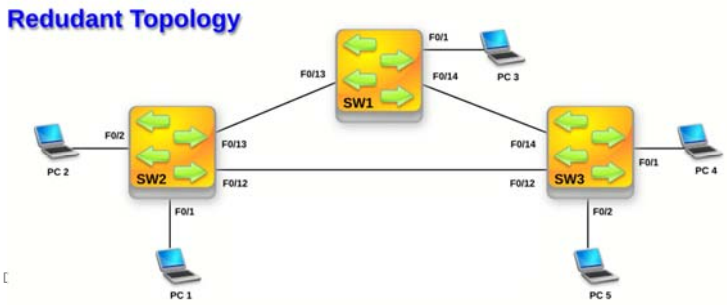
- Imagine that the **SW1**, **SW2** and **SW3** switches connect many devices and there is only a single connection between the switches



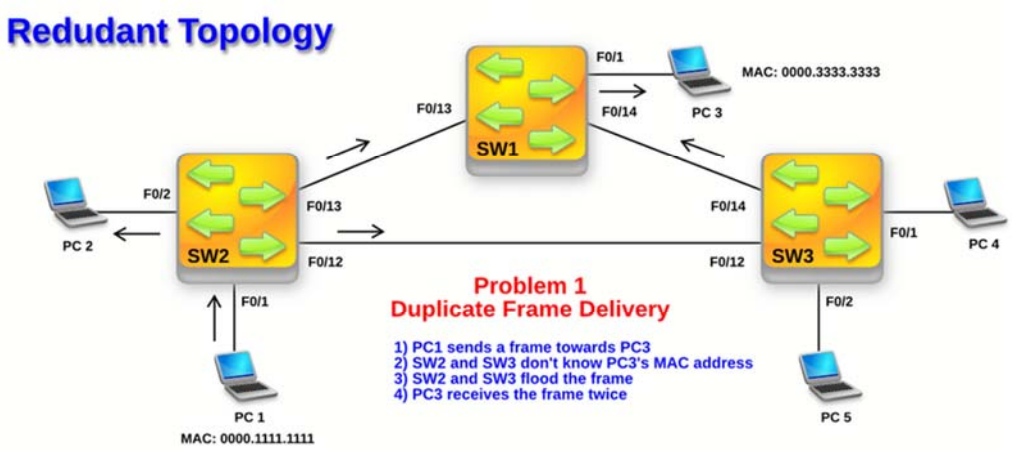
- links between the switches break, the communication between many devices fail.
- Such design creates a single point of failure.
- We could easily enhance this simple design to make it more resilient by adding an extra path between **SW2** and **SW3**



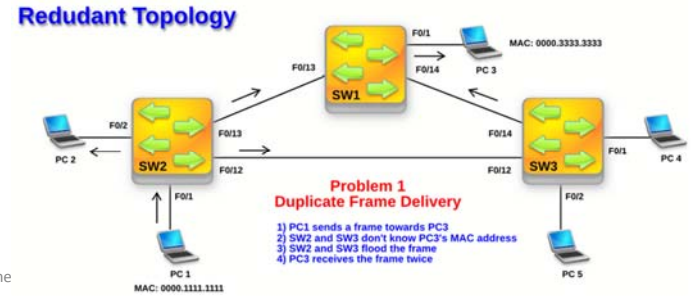
- creating the extra path here comes at a cost.
- The redundant connection between **SW2** and **SW3** creates a loop.
- The loop in turn, will create three serious problems.
- The last one in the list will eventually render our system unavailable.



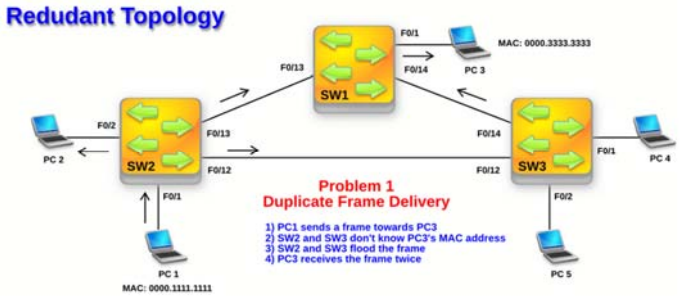
1. **Duplicate Frame Delivery**



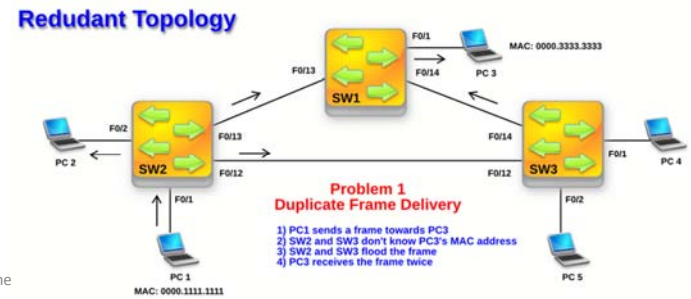
- **SW2** and **SW3** do not have the MAC address of **PC3** (0000.3333.3333) in their databases
- This can happen if the **PC3** doesn't speak for more than five minutes (This is the default time MAC address is kept in the database without refreshing it)
- Then, we have **PC1** sending frame towards **PC3**.



- As you recall, **SW2** will flood the frame out of its active ports if it does not know where **PC3** is located (unknown destination MAC address).
- The frame travels out **SW2**'s port F0/13 towards **SW1** and out the port F0/12 towards **SW3**.

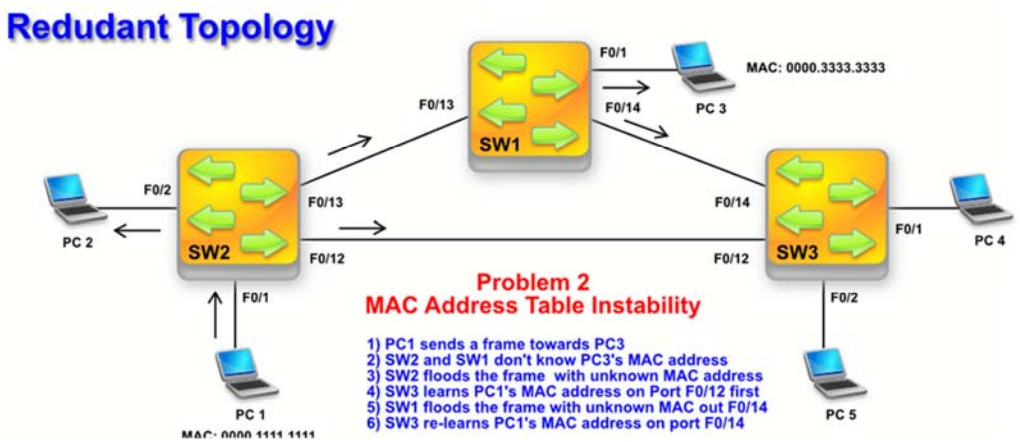


- **SW2** will deliver the frame to **PC3**.
- Since **SW3** floods the frame out as well, it will be sent towards **SW1** out of its port F0/14.
- Then, **SW1** obediently delivers the same copy of the frame to **PC3** again

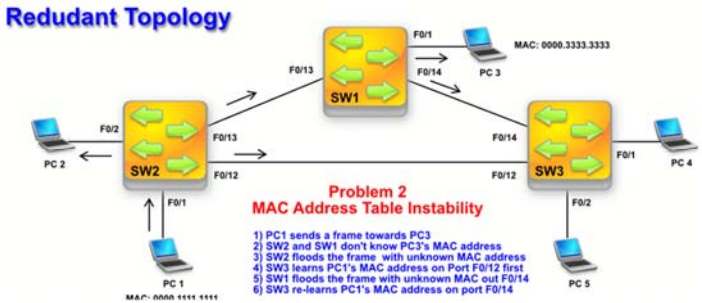


2. **MAC Address Table Instability**

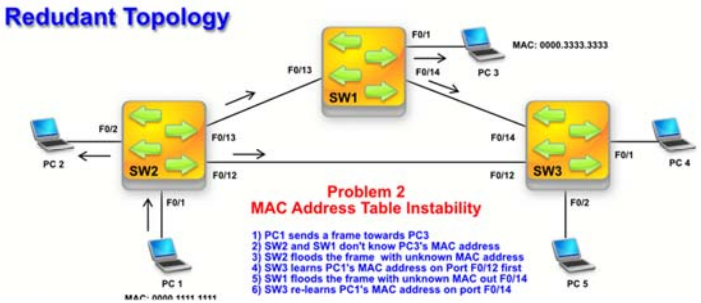
- switches change the MAC addresses depending on where they hear the sender.



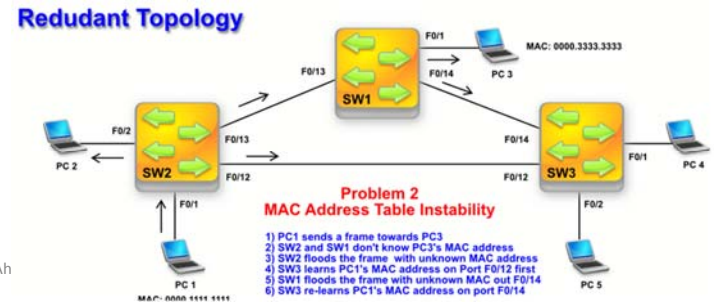
- **PC1** sends the frame to **PC3** (destination MAC: 0000.3333.3333).
- **SW2** floods the frame out F0/12 and F0/13 ports.
- **SW3** receives this frame sourced with 0000.1111.1111 MAC address (**PC1**). It learns the source MAC address and maps it to its F0/12 port where it arrived.



- **SW1** does not know where **PC3** is connected (at least right now) it will flood this frame out all active ports.
- This way, the frame is sent out **SW1**'s port F0/14 towards **SW3**. **SW3**, upon receiving the frame on its F0/14 port, reads the source MAC address (0000.1111.1111) and maps it to port F0/14 this time.

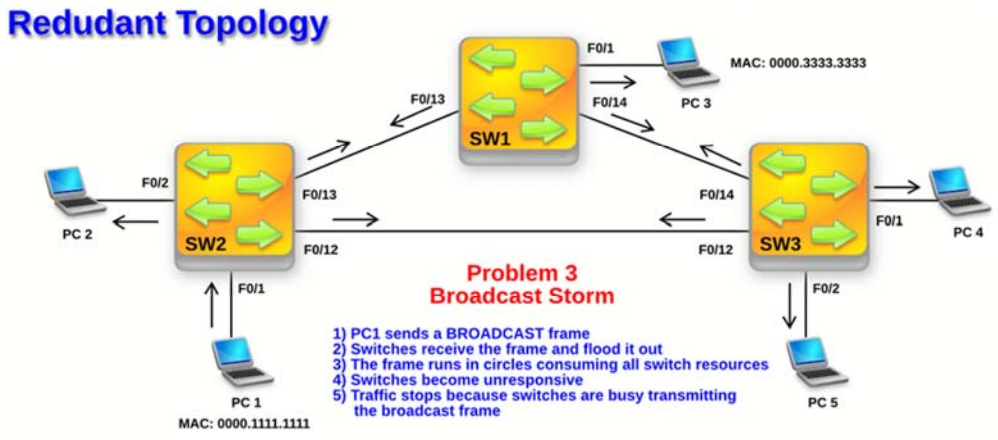


- This causes a little confusion as **SW3** learned it earlier on and it was port F0/12 before. Previous mapping is removed and F0/14 becomes the outbound port for 0000.1111.1111 now.

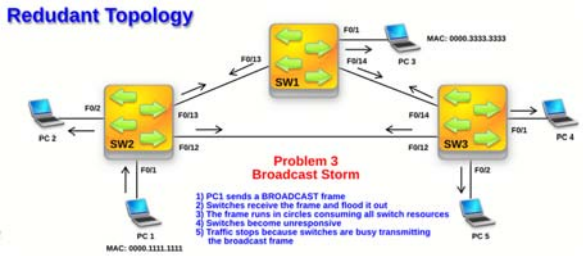


3. **Broadcast Storm**

- problem is really severe. It can bring our traffic to a halt

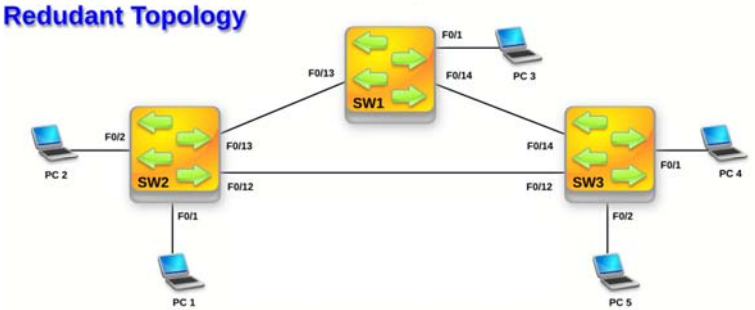


- **PC1** sends a broadcast frame. **SW2** upon receiving it, floods it out all its active ports. **SW1** receives it on port F0/13 and floods it out of other ports. **SW3** receives the broadcast frame on its F0/12 port and floods it.
- it receives this same broadcast frame from SW1 and again it floods it out all active ports except the port it arrived on



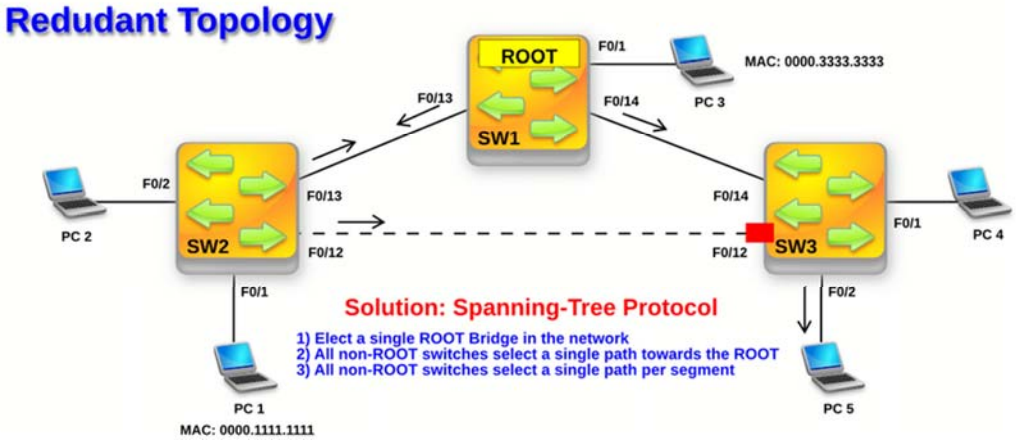
Spanning-Tree Protocol Overview

- STP is a layer 2 loop prevention mechanism. Switches running this protocol use special frames called **Bridge Protocol Data Unit** (BPDU).
- These frames contain enough information to allow the switches to create a loop free topology.
- This is accomplished using three distinct phases:

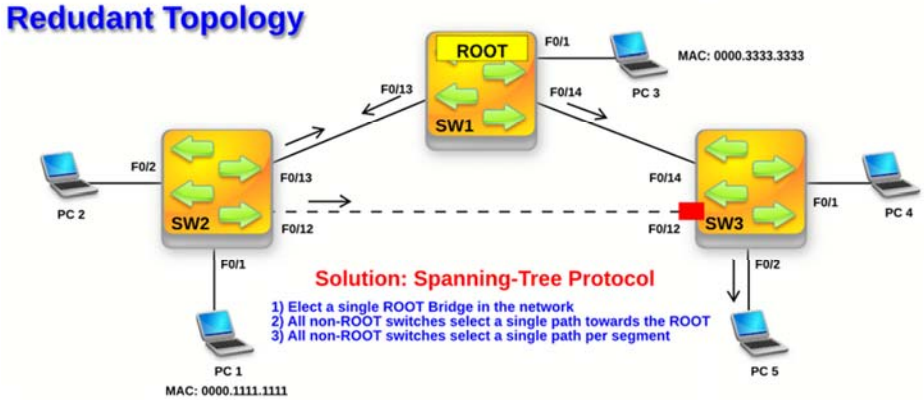


1. Elect **a single switch** to be the **root bridge** machine which is the central device in the layer 2 network. This machine will have all its ports in the forwarding state (**designated port** role).
2. All other switches (non-root switches), will select **a single path** towards the root bridge. That port is called the '**root port**' and will be forwarding traffic that is destined out of the switch through the root bridge. This path is the least cost (best) path towards the root.
3. All other switches will select **a single path per segment** in order to block stop the loop. The port that is forwarding traffic is called **designated port**. The port that is blocking traffic to stop the loop is called **non-designated port**.

- **SW1** has been elected as the root bridge. **SW2** uses port F0/13 as its root port (the best, or the least cost path towards the root).



- **SW3** uses it port F0/14 as the root port. **SW3** blocks the port F0/12 to stop the loop. SW2 keeps sending BPDU frames originated by the root bridge (**SW1**) out its F0/12 port towards SW3.

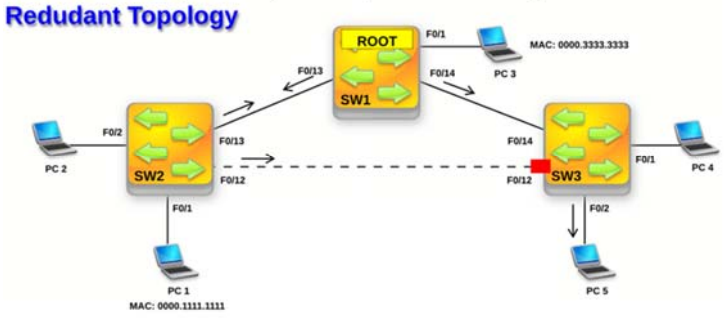


Spanning-Tree Protocol Terminology

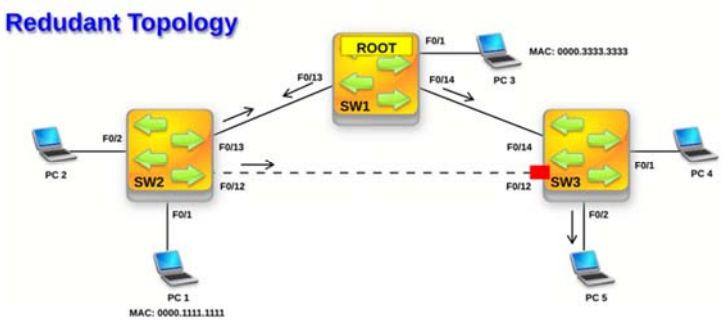
- The ports participating in STP play different **roles** and those roles use different **states** of operation:

Spanning-Tree Port Roles

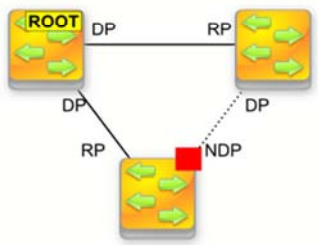
- **Root Port** (RP) - It is a port on a non-root switch, which is the shortest (the best) path towards the root bridge. Root bridge does NOT have any root ports. (no shortest path to itself .



- **Designated Port** (DP) - It is a port that is in the forwarding state. All ports of the root bridge are designated ports (they are never in a blocking state). BPDU frames our sent out this port.
- **Non-Designated Port** (NDP) - It is a port that is in a blocking state in the STP topology.



STP Terminology



STP Port Roles:

- Root Port (Forward upstream)
- Designated Port (Forward downstream)
- Non-Designated Port (Block)

- **Spanning-Tree Port States**
- **Disabled** - The port in this state does not participate in the STP operation (it is shut down).
- **Blocking** - The port does NOT forward any Ethernet frames, does NOT accept any Ethernet frames (discards arriving frames), does NOT learn any MAC addresses.
- However, the port **DOES** process **BPDU frames** received from a neighboring switch.
- If the port transitions to this state (blocking), it can stay blocked for **20 seconds by default (max_age)**.

- **Listening** - The port in this state **CAN send and receive the BPDU frames**.
- However, the port in this state does NOT learn any MAC addresses, and does NOT forward or process incoming frames either.
- All Ethernet frames are being discarded.
- The computation of loop free topology takes place in this state.
- If the port transitions to this state (listening), it can stay in this state for **15 seconds by default (forward_delay)**.

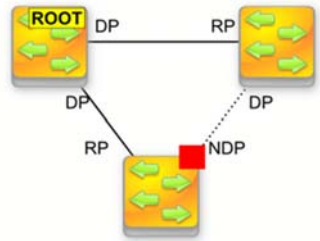
- **Learning** - The port in this state already knows its role (root port or designated port) in the STP domain.
- However, the port will not forward any Ethernet frames yet.
- It will be learning MAC addresses from the frames arriving at the port in order to populate MAC address table.
- This helps avoid too much flooding when the port transition to the forwarding state.
- If the port transitions to this state (learning), it can stay in this state for **15 seconds by default (forward_delay)**.

- **Forwarding** - The port in this state will forward all Ethernet frames as per switch operation.
- Also, the port will process all incoming Ethernet frames and will actively learn MAC addresses from the arriving traffic.

<u>Step one, root bridge</u>

- **Root Bridge**
  Root bridge is the switch that has all ports working in the designated role.
- It will be the reference point
- Root bridge will impose the timers that other switches will use such as:
  hello time - how often BPDUs are going to be sent/relayed (default timer=2 seconds),
- **max age** - how long the configuration is valid (default timer=20 seconds),
- **forward delay** - how long a port should be in listening/learning state (default timer=15 seconds).

- Root bridge will be announcing its presence by sending BPDU frames.
- Other switches will relay those frames out their designated port given the hello time.
- Also, the root bridge has all its ports in the designated role (forwarding).

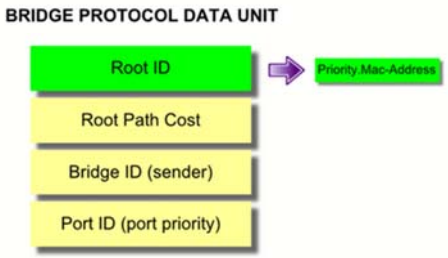


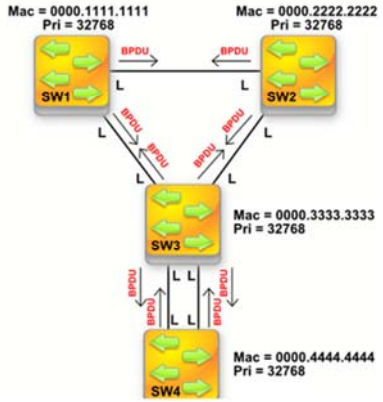
- Root election is based on a single parameter that is found in the BPDU frame called: Bridge ID.
- **The switch with the lowest Bridge ID becomes the root**. Bridge ID has the following format:

priority.base-mac-address

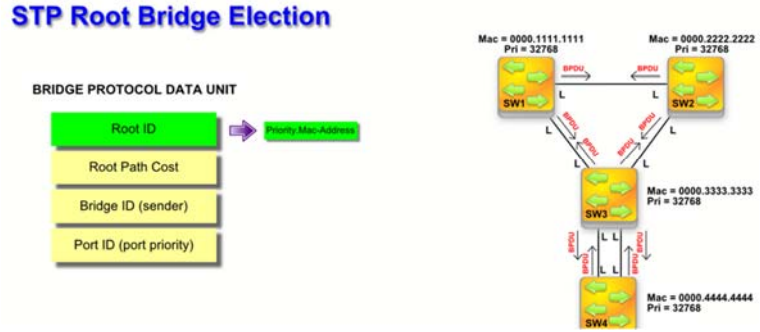
- **Priority** is configurable parameter that is used to elect the root bridge a device you want to be the root.
- The default value is: **32768**. The lower the value is the more likely for a switch to become a root.
- **Base Mac Address** is the unique mac address every switch has been given by the manufacturer.
- It is a main parameter in case the priority on all switches is identical.

STP Root Bridge Election

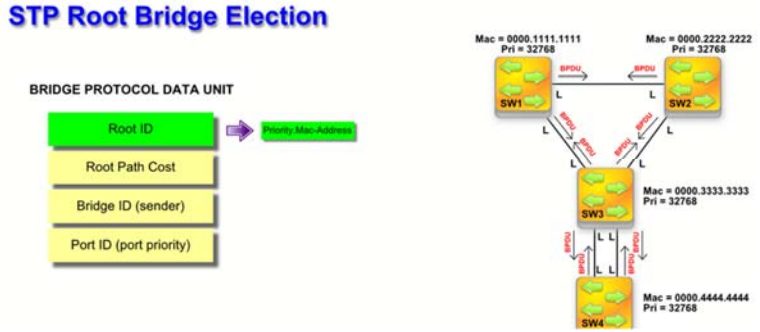




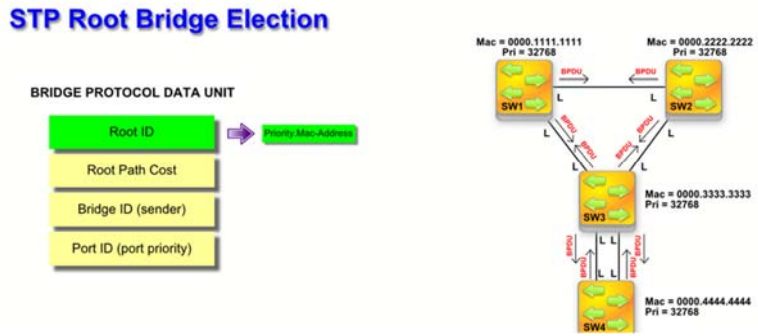
- start up all the switches and as soon as their ports transition to LISTENING state, they begin to send BPDU frames out of all active ports.
- In those frames both **Bridge ID** and **Root ID** parameters point to their own priority.base-mac-address value.



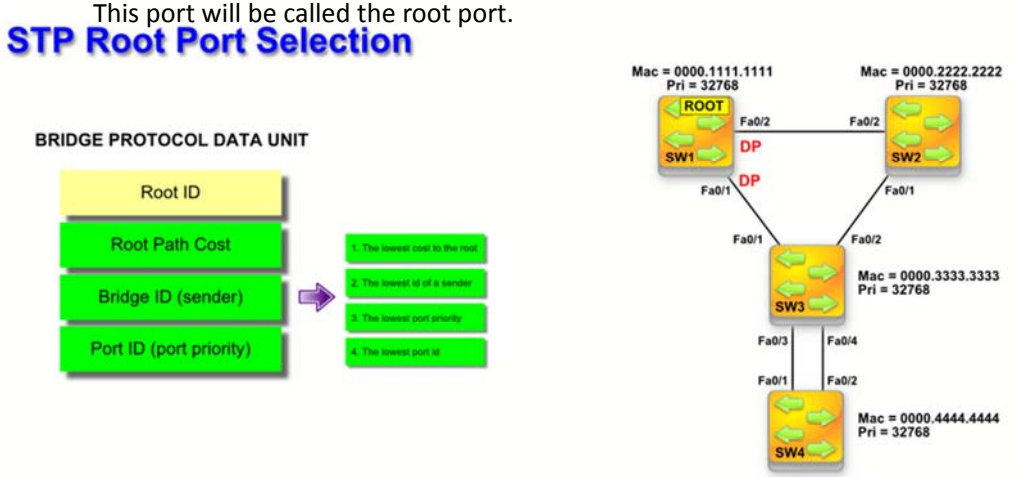
- Since they are processing the incoming BPDU's from the neighbors, **SW2** and **SW3** realize that **SW1**'s Bridge ID is lower than theirs.
- From that point onwards, they begin to relay BPDU frames saying that **SW1** as the root bridge.



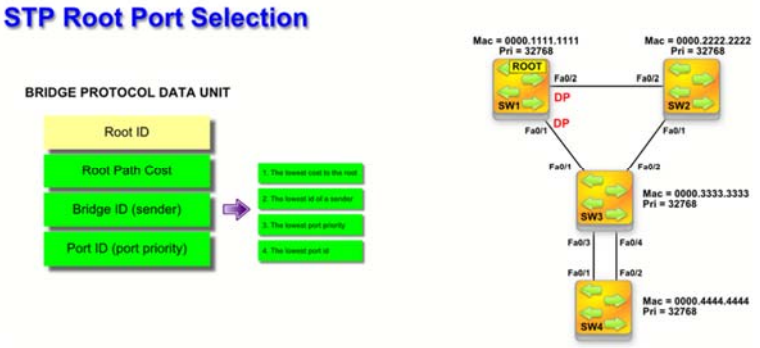
- example,
- **For SW3**
- **Bridge ID = 32768.0000.3333.3333**
- **Root ID = 32768.0000.1111.1111**



- Step 2: **Root Port Selection**
- As soon as the root has been elected, all non-root switches begin to calculate which port is the best (the least cost) towards the root bridge. This port will be called the root port.



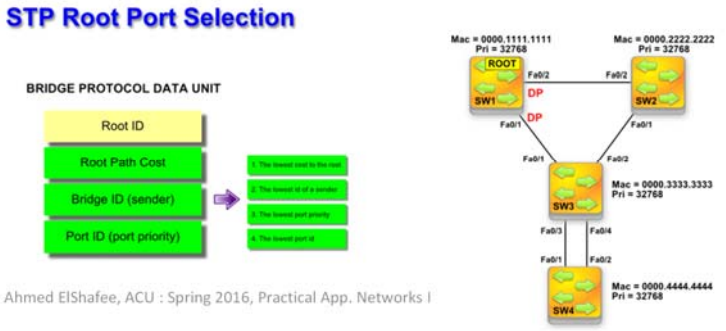
- **SW2**, **SW3** and **SW4** receive BPDUs from different directions.
- For instance, **SW2** will receive them on its port F0/1 and F0/2
- The accumulative cost (the sum of the cost in the path towards the root), is taken into consideration. **The lowest cost to reach the root becomes the root port**.



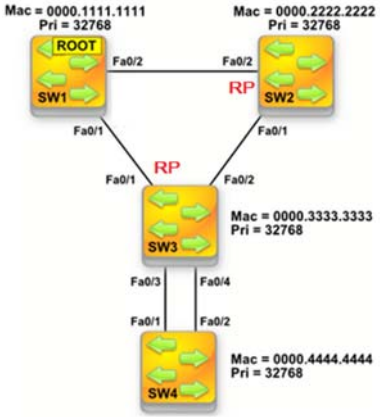
cost of path calculation:

- Each speed has its arbitrarily assigned cost which is configurable. A few examples are below:

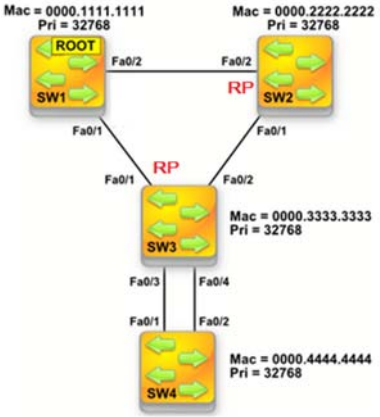
10 Mbps = 100
100 Mbps = 19
1 Gbps = 4
10 Gbps = 2



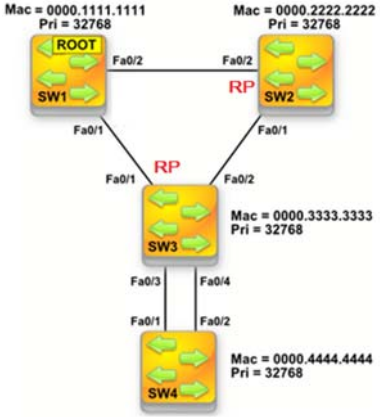
- The root bridge (here **SW1**) is sending its BPDU frame every 2 seconds.
- It uses the parameter called: **Root Path Cost** in BPDU to advertise the cost to the root.
- It puts the value of '0' in it, as it is the root bridge and has no cost to itself.
- The frame is sent out its port F0/1 towards **SW3** and F0/2 towards **SW2**.
- **SW2**, upon receiving it, adds the cost used to reach the sender of BPDU based on the predefined speed-to-cost value (all ports in our topology are FastEthernet=19).



- Root Path Cost = 0 + 19 = **19** via F0/2
- **SW2** is going to advertise its best (as of now) cost out of F0/1 port towards **SW3**.
- **SW3** will receive BPDU from **SW1** with the **Root Path Cost=0** on its F0/1 port.
- It will also receive BPDU from **SW2** on its F0/2 interface with the **Root Path Cost=19**.
- **SW3**; As both ports have the cost of 19 towards those BPDU senders, the following math is done to choose the least cost path towards the root bridge
- Root Path Cost = 0 + 19 = **19** via F0/1
  Root Path Cost = 19 + 19 = **38** via F0/2

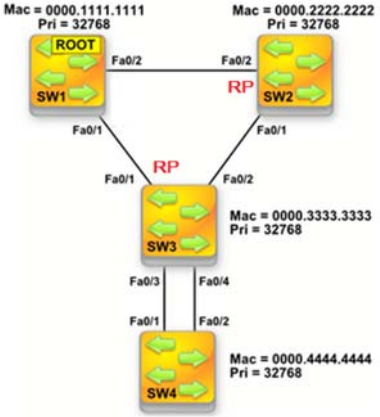


- It is clear that the direct connection towards root bridge via F0/1 is going to be selected as the root port.
- **SW3** has the least cost towards equal 19 (via F0/1 port).
- This cost is going to be added to Root Path Cost while it sends the BPDUs out F0/2, F0/3 and F0/4.
- Of course, **SW2** also chooses its F0/2 port as the root port since the cost is smaller.

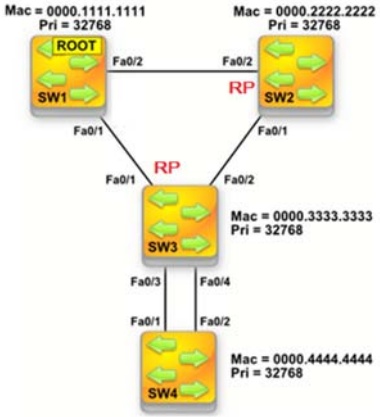


- if the Root Cost Path is identical!
- hat situation on **SW4**.
- It receives BPDUs on its ports F0/1 and F0/2 with the following parameters:

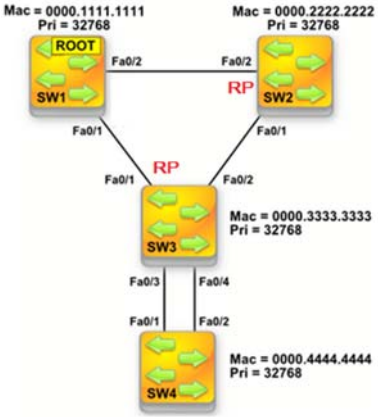
  Bridge ID = **32768.0000.3333.3333**
  Root ID = **32768.0000.1111.1111**
  Root Path Cost = **19**



- refer the **lowest port priority** of the sender.
- That parameter has a default value 128 and is configurable.
- The designated switch (**SW3**), is the same switch i.e. the same Bridge ID (32768.0000.3333.3333).
- The designated switch (**SW3**) sends BPDUs out of its F0/3 and F0/4 ports with the same priority = 128



- lowest Port ID where BPDU frames arrive on **SW4**.
- Port f0/3 becomes the root port since F0/**3** is lower than F0/**4 on SW3**.



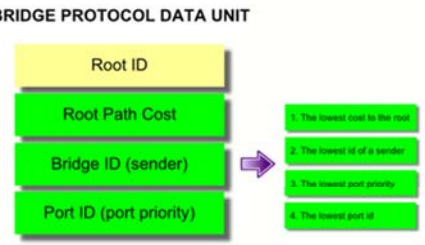
The following algorithm is used to determine the root port or designated port (in order):

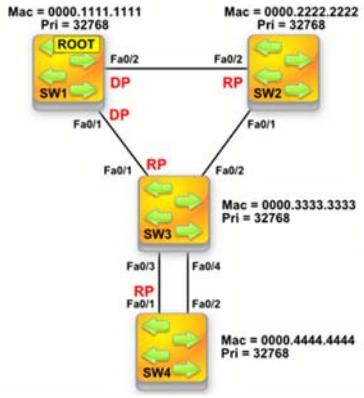
1. Prefer the **lowest Root Path Cost**.
2. In case of the same Root Path Cost, prefer the **lowest Bridge ID** of the designated switch (the neighbor that sends BPDUs).
3. In case of receiving BPDUs on multiple ports from the same designated switch (BPDU sender), prefer the **lowest port priority** of the sender. That parameter has a default value 128 and is configurable.
4. In case of all above are did not resolve the problem, prefer the **lowest Port ID of the BPDU sender**.

Step 3: **Designated Port Selection.**

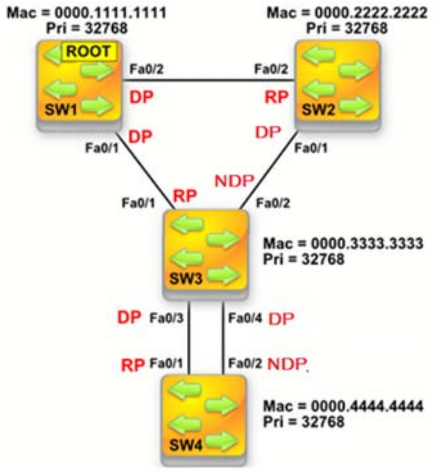
This procedure follows exactly the same algorithm used for root port selection.

STP Designated Port Selection

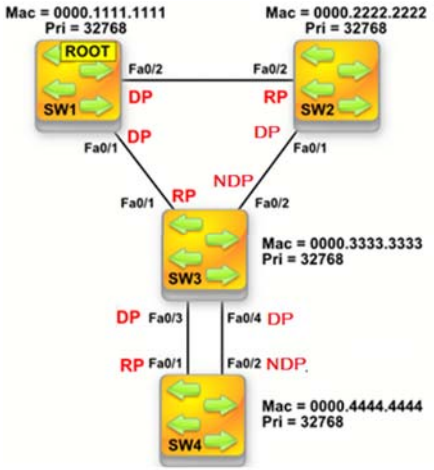




- Since root port is the best port towards the root bridge it is going to be in the forwarding state
- Root Path Cost advertised by **SW2** is 19 and so is the cost advertised by **SW3**.
- **SW2** has lower Bridge ID (32768.0000.2222.2222) than **SW3** (32768.0000.3333.3333). **SW3** must block its F0/2.
- .



- And last selection is going to happen between **SW3** (port F0/4) and **SW4** (port F0/2).
- Root Path Cost Advertised by **SW3** is 19, but **SW4** advertises its cost as 38 (two hops via F0/1).
- **SW4** blocks its port F0/2 (non-designated), the **SW3** promotes its port F0/4 to designated role (forwarding).



- This process happens in the LISTENING state of all ports. Since the topology has been computed and does not have loops (blocking appropriate ports), it is safe to move to next states: learning and finally forwarding.