Path Selection for AV Bridges

Norman Finn

Cisco Systems

Rev. 1

Background and problem statement

Background: AVB Bridges

- **The Audio Visual Bridge projects in IEEE 802.1 are aimed at a home or studio connected via an arbitrary combination of wired and wireless links interconnected by AVB Bridges.**
  - **Redundant wired links among boxes in a stack, multiple endstations within a single box.**
  - **Multiple wireless links among stacks, "stacks" may roam.**
  - **AVB Bridge capability included in every box with more than one network connection.**
- **Utilizing all available links, rather than a single spanning tree, is necessary.**
  - **Otherwise, critical links could become saturated.**
  - **Wired and wireless links must be integrated into a single forwarding protocol.**
  - **A Wireless Access Point is a special case of an AVB Bridge.**
  - **P802.1aq Shortest Path Bridging is (will be) the forwarding protocol.**
- **Stations must reserve bandwidth for particular streams.**
  - **AVB Bridges provide priority levels that guarantee latency, jitter, and bandwidth.**
  - **Resources (buffer space, link bandwidth) must be allocated for each stream.**
- **The AVB Bridge solution must interoperate with and support one or more of the Layer 3 solutions used over longer distances.**

Background: AVB Bridges

- **The IEEE Std. 802.1Q-2005 tag (Q-tag) can mark a frame with a 3-bit "priority" that can be interpreted by a Q-bridge as a "class of service".**
  - **An IEEE Std. 802.1Q-2005 Bridge (Q-bridge) supports up to 8 output queues per port.**
  - **There are standard mappings of priority to queue for numbers of queues < 8.**
- **An AVB Bridge will be an extended Q-bridge.**
  - **A very few priority values (1-3) will be reserved for traffic requiring certain combinations of low latency and/or guaranteed bandwidth.**
  - **The remaining priority levels will be available for, at least, control traffic (highest priority, very low bandwidth) and two priority levels of best effort.**
  - **An AVB Bridge will require some number (TBD) of queues > 1 per port.**
- **No station may send traffic on a reserved priority value until it has successfully made an AVB reservation.**
  - **This is a Layer 2 reservation among AVB Bridges.**
  - **AVB reservations are not necessarily policed by the AVB Bridges; good will among stations is assumed.**
  - **AVB reservations may be triggered by RSVP.**
- **Plug and play operation is absolutely required.**
  - **The customer need not configure anything. Period.**
  - **Applications are required to issue reservations.**

Background: Recent developments

- **One basis of this proposal is IEEE P802.1aq Shortest Path Bridging**
  - **The link state version of this idea is assumed.**
  - **We will assume that VLANs, in the sense of "communities of interest", are not required in the AVB network.**
  - **A separate spanning tree instance is created for each AVB Bridge in the network, with that AVB Bridge as the Root.**
  - **A unique VLAN ID (VID) is assigned to each of those spanning tree instances.**
  - **Therefore, if there are *n* AVB Bridges in the network, there are *n* spanning trees and *n* VIDs in use.**
  - **These spanning trees are symmetrical, in the sense that the path from Bridge A to Bridge B is exactly the same as the path from Bridge B to Bridge A.**
- **The second basis of this proposal is IEEE P802.1?? Bandwidth reservation**
  - **This is bandwidth reservation protocol may be based on MRP (Project 802.1ak).**
  - **It may be an extension of the link state.**
  - **This protocol reserves multipoint-to-multipoint bandwidth for VLANs.**

First, the easy problem to be solved

- **AVB Bandwidth reservation.**

  Bandwidth must be reserved for two-way conversations at certain priority levels.

  There must be no configuration required; reservations are initiated by the applications, perhaps through IP and RSVP.

- **With MRP-based bandwidth reservation:**

  Reservations spread out from each (of the) end(s) of the conversation, each specifying both input and output bandwidth.

  Any port that sees reservations for the same conversation in both directions must allocate resources for that conversation.

  A port that sees reservations only in one direction knows about the reservation, but need not allocate resources.

- **With link state based bandwidth reservation:**

  Each bridge knows about all reservations.

The bigger problem to be solved

- **Alternate paths among stations, other than the shortest path, could be utilized when bandwidth is not available on the shortest path.**

  This is a new idea, one that is difficult to do in the big-I Internet, but is possible within the bounds of a bridged network.

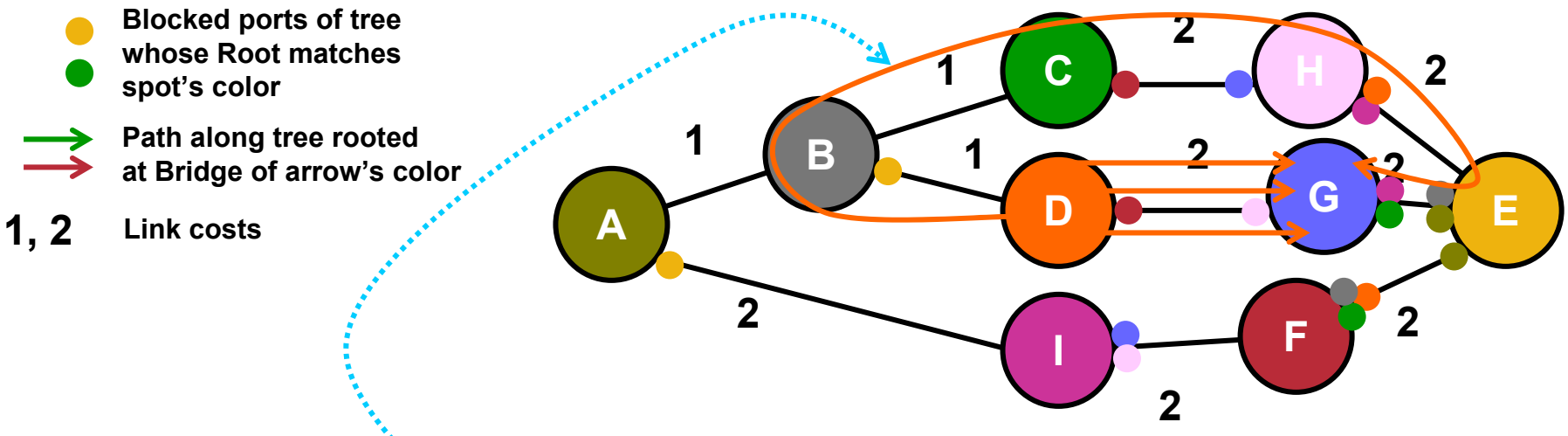
- **Simply using P802.1aq SPB is not sufficient:**

  SPB uses the spanning tree belonging to the Bridge to which the source station is attached for all traffic.

  Even if multiple topologies are generated using different metrics, all paths used are shortest paths, in some sense.

  So, a new scheme is presented, here, for alternate paths.

The hard problem illustrated



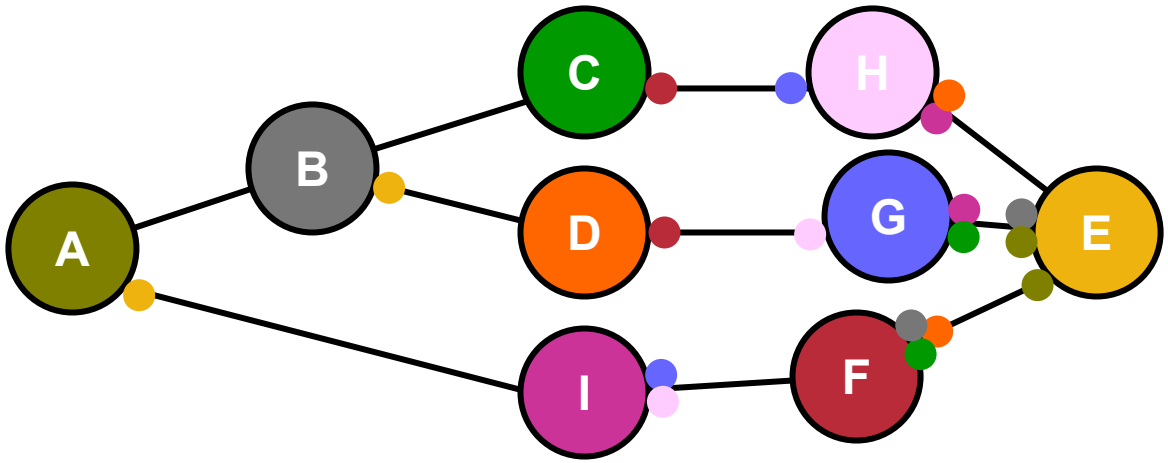
- Three streams from bridge D to bridge G have exhausted the capacity of the link between D and G.
- A fourth stream is needed from D to G.
- Wouldn't it be nice if that fourth stream could be sent the long way around, assuming of course that the resources are available and the path satisfies the latency requirements?

Cisco Systems

Solution overview

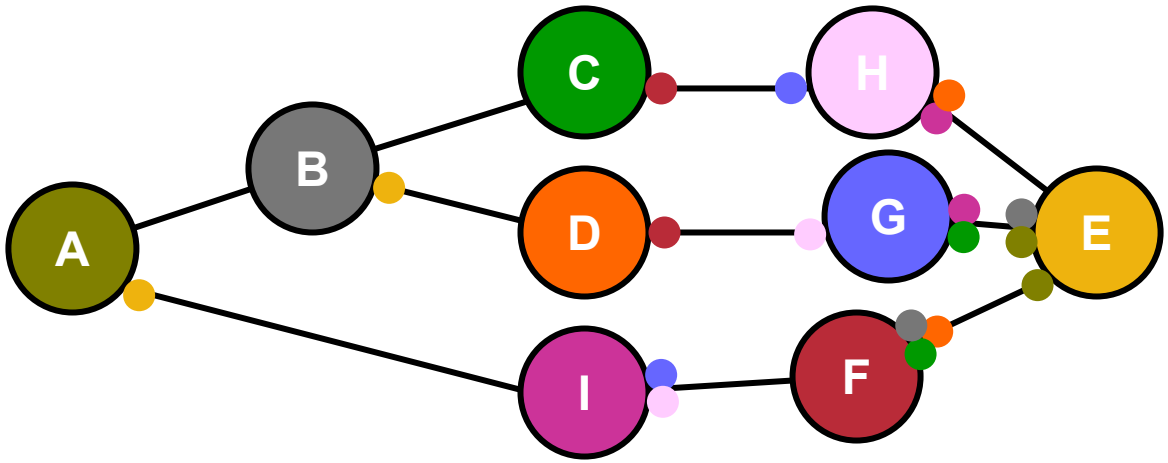
Plug and play VID assignment

- **SPB creates the "common set" of *n* symmetrical spanning trees for *n* bridges.**



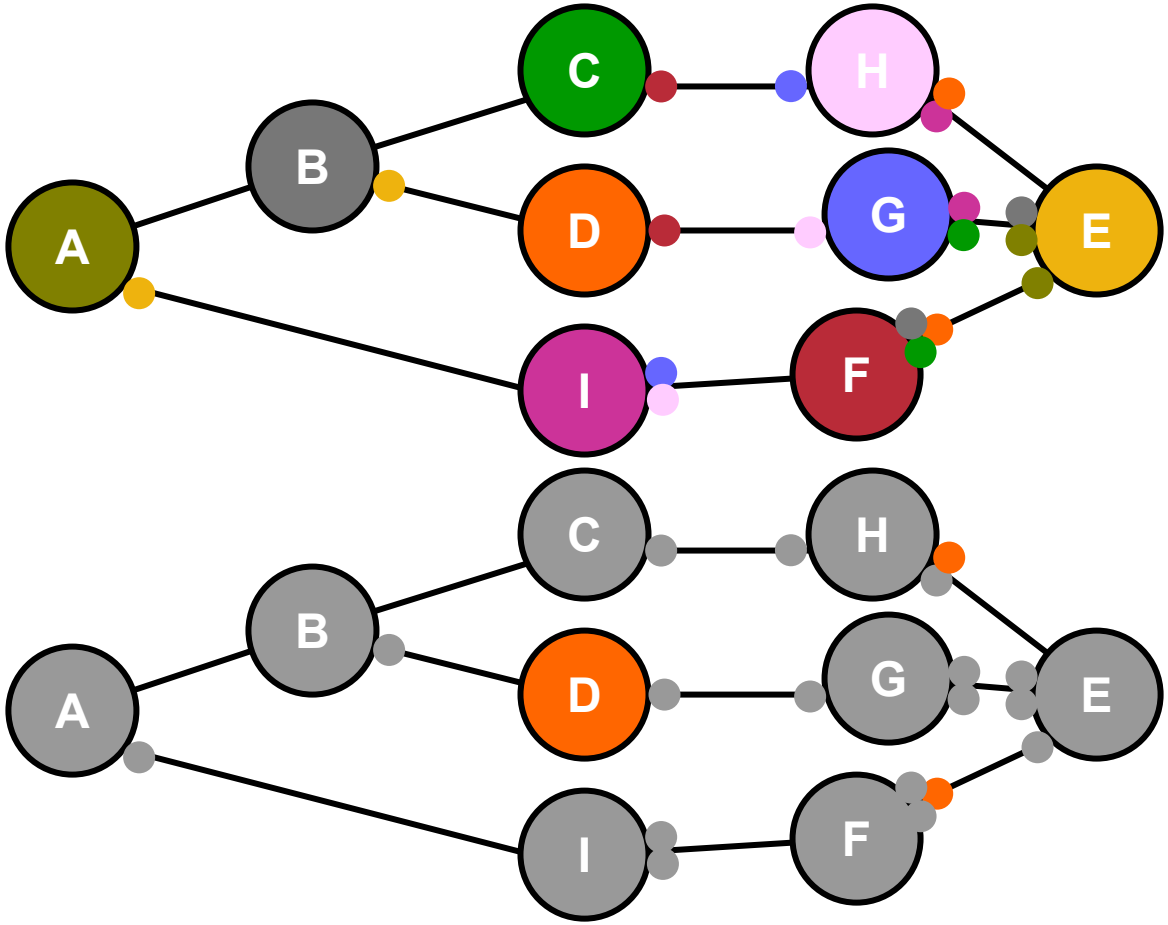
Plug and play VID assignment

- **One FDB in each bridge directs frames for the common set.**



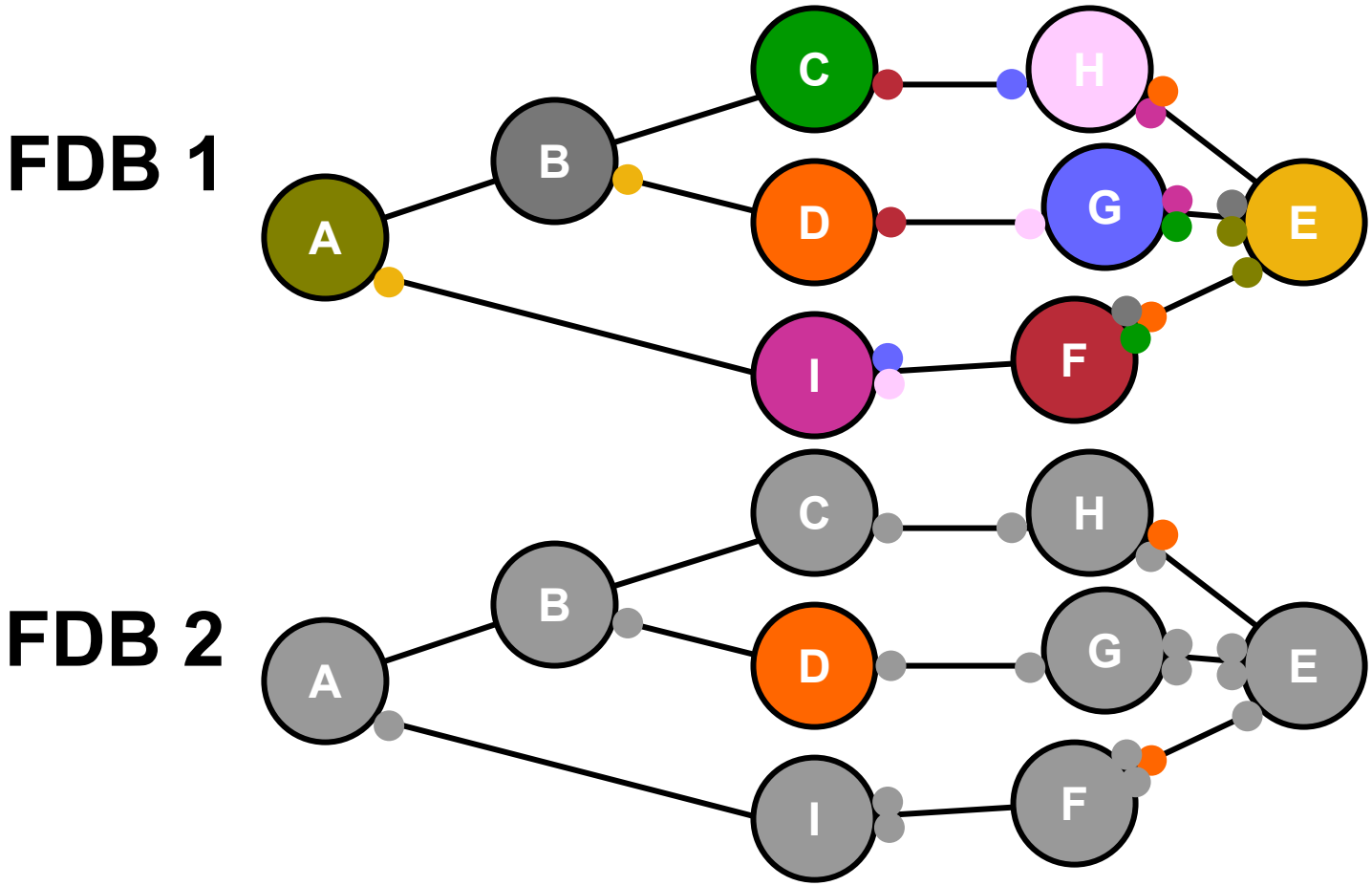
Plug and play VID assignment

- **Each Bridge gets a second tree (bridge D's is shown) that probably has (but doesn't necessarily have) the same topology as its common tree.**



Plug and play VID assignment

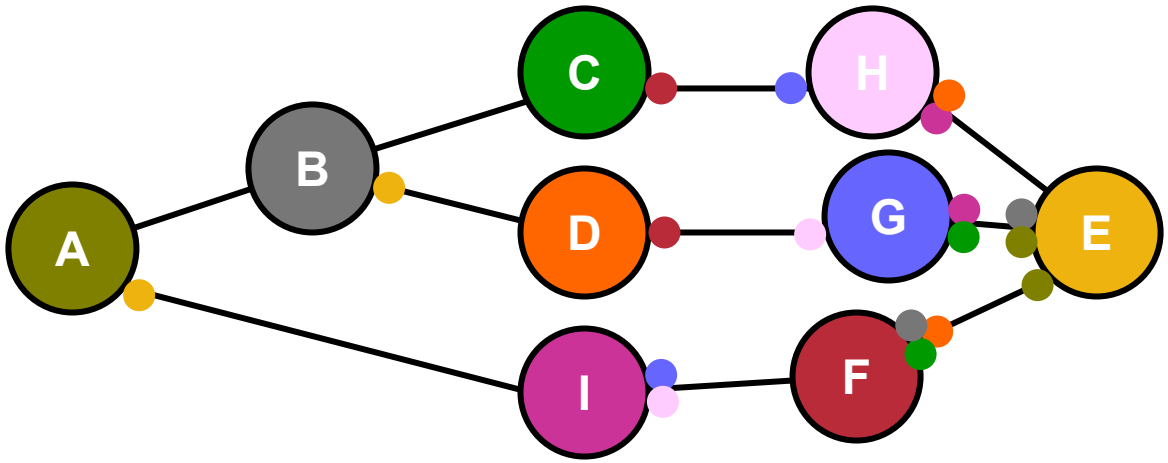
- **Each bridge's alternate tree gets its own FDB.**



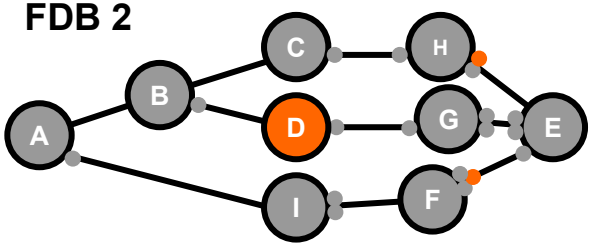
Plug and play VID assignment

- **And so on for each bridge**

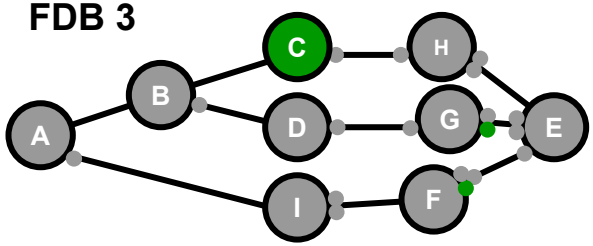
FDB 1



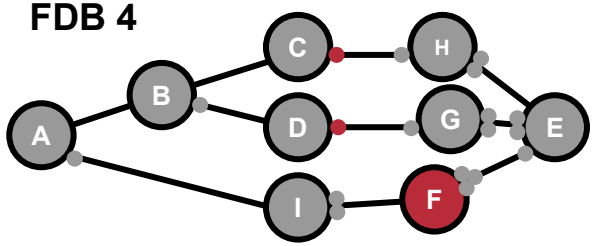
FDB 2



FDB 3



FDB 4



...

Cisco Systems

Multiple paths for Shortest Path Bridging

Plug and play VID assignment

- **To make Shortest Path Bridging plug-and-play, a means of assigning each Bridge to a particular Root Part (VID number) without configuration is necessary.**

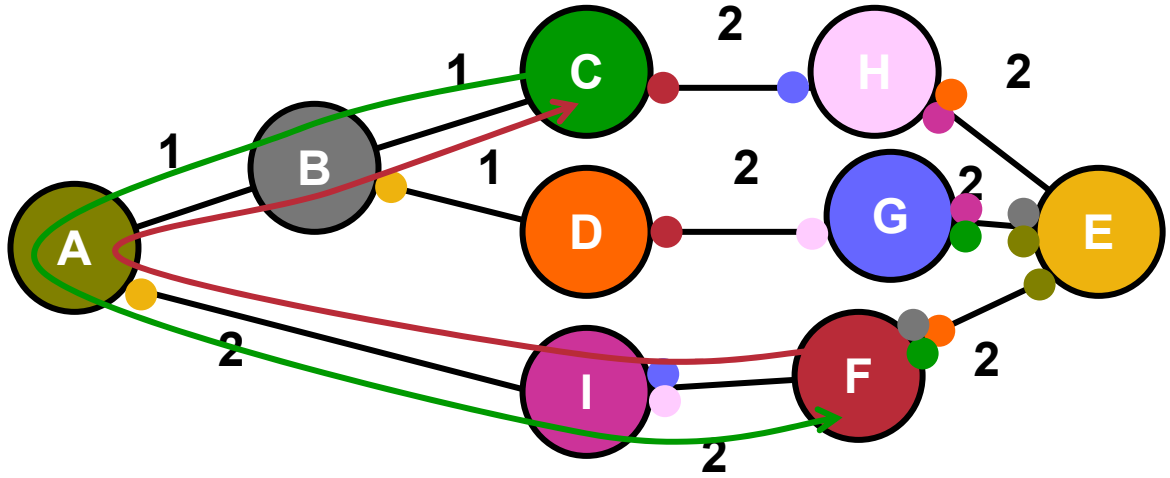
  We will take this problem as solved by P802.1aq.

Multiple paths in Shortest Path Bridging

Blocked ports of tree whose Root matches spot's color

Path along tree rooted at Bridge of arrow's color

1, 2 Link costs



- **In Shortest Path Bridging:**

  Every frame is forwarded using a VID that identifies a spanning tree rooted at the entry Bridge.

  Spanning trees are symmetrical, in that the two trees rooted at the two Bridges A and B take exactly the same path between A and B.

  All of the VIDs corresponding to a single community of interest are in the same Filtering Database (FDB, identified by the FID, derived from the VID).

  Tree symmetry + same FID means that source MAC addresses learned on the tree rooted at A can be used to forward frames with that destination MAC address on the tree rooted at B.

Multiple paths in Shortest Path Bridging

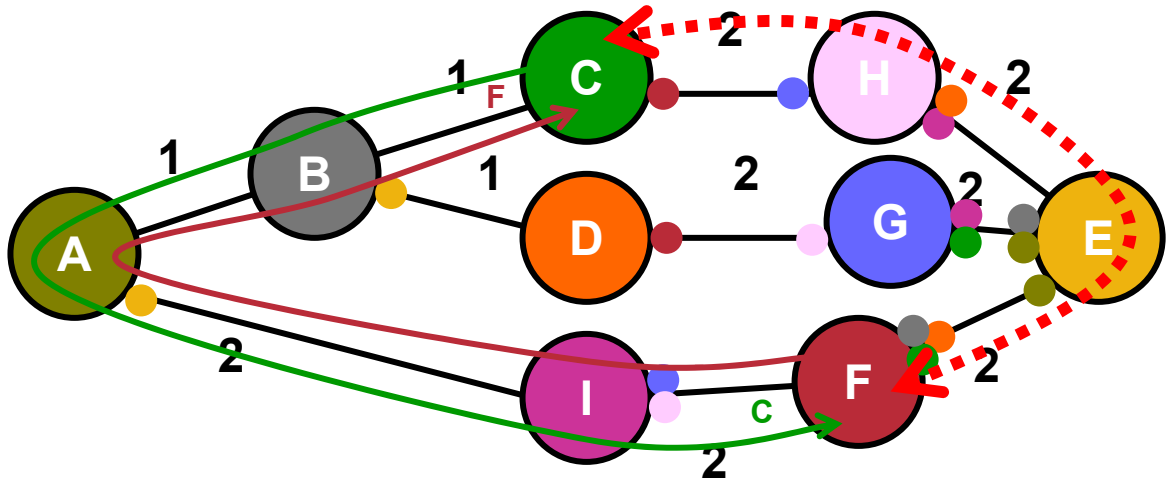
Blocked ports of tree whose Root matches spot's color

Path along tree rooted at Bridge of arrow's color

1 2 Link costs

Proposed new path between C and F

C F Learned MAC addresses



- If the resources along the normal path between AVB Bridges C and F are exhausted, they might want to use E's spanning tree for a new conversation.

  But, this would confuse Bridge F's learning process; F has learned C's MAC address on its port that faces Bridge I.

  C has the same problem learning F's MAC address.

Frame marking for alternate paths

- **In order to allow normal MAC address learning to take place on the set of symmetrical spanning trees, we must separate traffic using alternate paths from traffic using symmetrical paths.**
  - **Source MAC addresses of frames using the symmetrical paths must be learned in a common FDB (Shared VLAN Learning SVL).**
  - **Source MAC addresses of frames using the alternate paths must each be learned using a different FDB (Independent VLAN Learning IVL).**
  - **The frames must be marked, somehow, as being different, so the Bridges know which FDB to use, and whether to learn.**
- **There are three obvious possibilities for marking these frames:**
  1. **Use the Priority; that identifies the frame as "AVB reserved".**
  2. **Use the CFI/DEI bit and a new tag Ethertype.**
  3. **Use a bit from the VID.**

Frame marking for alternate paths

- **Use the Priority field.**

  This is possible, but it is a significant change to existing Bridge designs*, and raises an architectural objection: priority is orthogonal to VID; priority is for queue selection, not for filtering and forwarding decisions.

- **Use the CEI/DEI bit with a new Ethertype.**

  This is possible, but it is a significant change to existing Bridge designs*, and makes the DEI bit unavailable in the AVB network.

- **Use a bit from the VID.** ← **(the right answer)**

  This is perfectly consistent with the use of VIDs in Bridges, but reduces the maximum size of the network from 4094 Bridges to 2046 Bridges.

*** It increases the VID-to-FID table size from 4k×12 bits to 8k×12, or even 8k×13 bits.**

Address learning on alternate paths

- **Alternate paths are only used for AVB-reserved traffic, then:**
  - **The multicast MAC address FDB entries are supplied in the AVB reservation, which among other things, serves as a GMRP/MGRP/IGMP registration.**
  - **The unicast MAC addresses can be taken from the payload in the reservations.**
  - **The unicast MAC addresses can be installed in the FDB for the duration of the reservations.**
  - **Source MAC addresses of frames using that AVB reservation need not be learned.**
- **Frames in AVB-reserved streams should never be flooded.**
  - **This would disrupt others' bandwidth reservations.**
  - **Learning is not done on the fly – the FDB entries are created at reservation time.**
  - **Frames with unknown destination addresses are filtered.**
  - **This is a change to current Bridge operation.**

Changes to Shortest Path Bridging

- **The link state information must include the parameters needed for an AV Bridge to decide whether a given far-away link can or cannot support a given AVB Registration, e.g. latency or jitter characteristics such as those encountered with wireless links.**

  This does not include the amount of non-AV traffic being carried, or stability will suffer.

- **Each AVB Bridge acquires two VIDs, *v* and 2048+*v*, 1 ≤ *v* ≤ 2047.**

  This bit identifies the alternate path VIDs, on which learning and/or forwarding operates slightly differently.

  This use of the high-order bit (the "AV Part") is new, but perfectly compatible with current practice.

- **Both VIDs are attached to that AVB Bridge's spanning tree instance.**

Changes to Shortest Path Bridging

- **As before, each Bridge is the root of a single spanning tree instance, symmetrical with other Bridges' spanning tree instances.**

  Multiple topologies could be configured, further subdividing the VID space, but this capability is of little utility in AV Bridges' plug-and-play situation.

  Separate communities of interest (VLANs) could be configured, but, since the number of communities times the number AVB Bridges (times the number of configured topologies) must be < 2047 (was 4095), the number of communities (or Bridges) is rather restricted.

  Of course, any configuration of multiple topologies or communities of interest eliminates plug-and-play.

- **Traffic reserved via AVB Reservations could use the ordinary VIDs, or could be confined to the alternate VIDs only.**

  The requirement to not flood traffic in AVB reserved flows means that all AVB reserved traffic must be marked as such.

  It is awkward for the Priority field to affect forwarding.

  Therefore the alternate VIDs and only alternate VIDs carry the AVB reserved traffic, so the alternate VIDs can be called, "AV VIDs".

Cisco Systems

The AVB Reservation Protocol

AVB Registration

- **The station making the RSVP reservation, or the entry AV Bridge that detects the RSVP reservation, generates an AVB Reservation.**
- **The AVB Reservation includes :**

  The source MAC address of the data flow (payload).

  The destination MAC address (may be a Group address) of the data flow (payload).

  The bandwidth and latency requirements for the flow (payload).

  The Priority of the flow (key).

  The Identity of the flow (key).

  The connectivity priority of the flow (payload).

  A “failure” flag (payload, only if MRP based).

  The AV VID on which the flow is to be carried (key, implied by the Q-tag of the AVB Reservation frame, itself).

AVB Registration

- **The choice of which AV VID to use for the registration is made by the entry AVB Bridge, using its knowledge of the topology of the network and the current path registrations, on the basis of (in priority order from high to low):**
  - **Whether the VID can or cannot support the flow.**
  - **The VID with the least total cost.**
  - **The spanning tree with the best Bridge ID.**

AVB Registration: The MRP application

- **MRP modifications:**

  Registration must carry payload, as well as key.

  The key identifies which registration is being made or withdrawn. It is the VLAN, Priority, and Flow Identity.

  The data is in several parts, and the flow of data between Bridge Ports is a key part of AVB Registration.

  The bandwidth requirement is unidirectional, and allows the separate allocation of a link's resources in each direction.

- **Link state modifications:**

  Flow registrations can simply be added to the link state advertisements.

  Same key + payload rules apply.

Flow Identity and Connectivity Priority

- **A Flow Identity is required to match the two (or more) ends of the flow to detect which ports are affected.**

  The RSVP flow identifier would suffice for this purpose.

  A purely MAC-layer identifier can be constructed from the source and destination MAC addresses and an integer, but creating that integer implies either an asymmetry (e.g. a server/client relationship), or the use of 802 LSAPs, neither of which is particularly attractive.

- **A Connectivity Priority is helpful when turning down reservations.**

  Just using the MAC addresses of the flows seems arbitrary.

  The high-order field in the Connectivity Priority is a configured value that overrides all other considerations.

  Since AV Bridges have synchronized clocks, the time at which the AVB Registration was issued seems like a great choice for the Connectivity Priority; the oldest connection wins.

AVB Registration: flow summation

- **Source MAC address.**

  The source MAC address list for key *k* registered (output) on Port *a* to the neighbor AV Bridge is the union of all of the source MAC address lists received (inputs) in registrations on all other Ports for that same key *k*.

- **The destination MAC address (may be a Group address) of the data flow (as payload).**

  The destination MAC address lists are collected as for source addresses.

- **Bandwidth requirement.**

  The output bandwidth requirement is the sum of the input requirements.

- **Latency requirement.**

  The output latency requirement is the minimum of the input requirements.

- **Connectivity priority.**

  The output connectivity priority is the best of the input connectivity priorities.

- **Failure flag.**

  The failure flag output on Port *a* is (the logical OR of all of the registered failure flags from all other input ports) AND (the logical OR of all of the input failure flags registered on Port *a*) AND (Port *a* can support the flow) AND (at least one other port can support the flow).

AVB Registration: endpoints

- The endpoints engage in some communication (presumably non-AVB) to determine that they want to reserve bandwidth for an AVB circuit.
- The entry AV Bridges determine, using link state information, whether the circuit is possible, and if so, which AV VID to use.
- Each endpoint of the conversation issues two AVB registrations, one for each direction, defining the flow.

  These may be combined into a single frame.

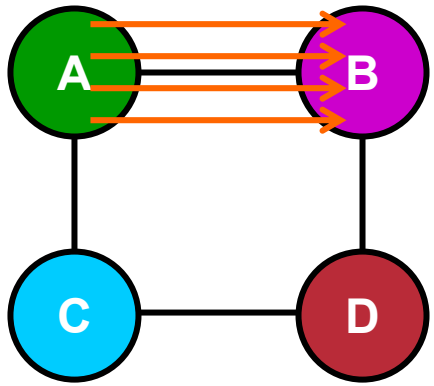
  If link state registration is used, this registration flow is simulated.
- If the registration(s) from the other end(s) of the circuit are received without failure flags, then the circuit is ready for use.
- If failures are received, then the registration is withdrawn.
- The registration might be retried on another AV VID, but this is for future study.

Cisco Systems

The bottom line

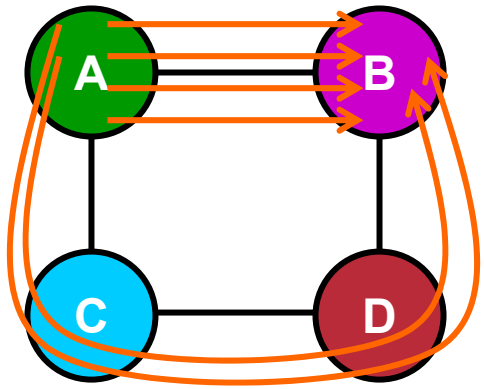
The bottom line

- **If you've filled up the bandwidth between bridge A and bridge B and still have more traffic you want to put between them, ...**



The bottom line

- **... you can take the long way around!**




CISCO SYSTEMS