Deterministic Ethernet:
Taking AVB to the next step

A program for further work by IEEE 802 and other SDOs

Norman Finn

Ver. 02

In a nutshell

The Audience

- This slide deck is directed at IEEE 802.1 Higher Layers Working Group in general, and the Audio Video Bridging Task Group, in particular.
- The author assumes that the reader is familiar with the current capabilities of the 802.1 AVB suite of protocols, and with the fact that 802.1 networks need to support time-critical mission control frames, constant bandwidth reserved resource streams, and best-effort traffic.
- The author assumes that the reader understands that AVB applications include industrial automation, process control, intra-vehicular networks, and home and studio audio/video networks.
- The author assumes that the reader understands that the MAC Layer is only one part of the problem space.

The Problem

- At present, there are a number of competing methods for building Deterministic Ethernet networks.
  - 802.1AB AVB: Some useful bits :) and Spanning Tree :(
  - IEEE 1588 and 802.1AS time sync.
  - SAE AS6802 Time-Triggered Ethernet
  - ODVA DLR: That is, Rings.
  - PROFINET: isynchronous real-time Ethernet
  - ISO/IEC 62439 and others: PRP, Rings, Traffic Engineering
  - ITU-T G.803x Protection Switching, including rings.
  - Whatever IETF may be up to, based on TRILL.
- Each standards suite has advantages and weaknesses.
- The market for Deterministic Ethernet is ready to grow very rapidly; in fact, to explode.
- **The biggest obstacle to market growth is the lack of a coherent standards story for vendors to build to.**

The Opportunity

- What is not obvious is that the above list of protocols can largely be brought together into a coherent plan by using the right kind of glue:
  - **Network cores** (VLAN tagging)
  - **IS-IS protocol**
- IEEE 802.1 should lead the way to open this market.

Fault Recovery Protocols

Problems or solutions?

Which protocol is better?

- **Spanning Tree**: IEEE 802.1Q Rapid Spanning Tree Protocol or Multiple Spanning Tree Protocol (MSTP). One or more trees, up to one per VLAN. All data on one VLAN follows the same tree.

  + Handles absolutely any topology.

  + Can be plug-and-play.

  – Often leads to very sub-optimal routing choices.

  – Worst-case convergence time is several seconds.

Which protocol is better?

- **Routing Technology**: SPB-V (Shortest Path Bridging V-mode) is defined by IEEE 802.1aq. Uses IS-IS to “route” at the MAC layer, rather like IETF TRILL, but without the encapsulation.

  + Handles absolutely any topology.

  + Can be plug-and-play (if 802.1 AVB defines a profile).

  + Unicasts and multicasts always routed along shortest path.

  0 Worst-case convergence time is sub-second.

  – (Number of VLANs) * (number of switches) < 4095

Which protocol is better?

- **Ring Protocols**: There are several candidates for protocols that assume that the network nodes are connected in a ring, and that the links are ordinary Ethernet links, including proprietary protocols, ITU-T’s G.8032, and ODVA DLR rings.

  + Fast (≈10 ms) response to a link or node failure.

  + Some protocols can be plug-and-play.

  0 Routing may or may not be optimal, but separate topologies for separate VLANs can be configured by some protocols.

  – Two or more failures lead to a loss of connectivity.

  – If the physical topology is not, in fact, as assumed by the configuration, then the network can melt down.

Which protocol is better?

- **Redundant delivery:** Either the end station or the edge switch replicates a frame and sends a separate copy along more than one path to the destination(s) **outside the control of any topology recovery protocol**. The receiver gets multiple copies.

  + Handles any suitably redundant topology.

  + Instant response to a single link or node failure.

  + Redundancy not dependent on a device taking an action.

  0 Paths and flows must be configured.

  – Bandwidth usage is at least doubled.

  – If the physical topology or the configuration have a serious mismatch, then the network can melt down.

Which protocol is better?

- The answer to “which one is better,” is, “**It depends** on your needs.” Each protocol has its advantages and disadvantages.
- In general, associated with each fault recovery protocol is a **suite of ancillary protocols** with capabilities relevant to an application. These additional capabilities can be **more important** than the features of the base protocols.
- As we will see in this presentation, it is possible to get **all of the benefits** of all of these fault recovery protocols **simultaneously**!

Network cores

KEY IDEA #1

- QoS features (SRP, Timed queues) operate on **priority**, and are **independent of forwarding choices** that are, in part, based on VLAN.

Priority and VLAN are independent

- This is not news to 802.1, but it is news to many others.
- The **VLAN and MAC addresses** determine on **which port(s)** a given frame is output.
- The **Priority** determines **when** the frame is output.

KEY IDEA #2

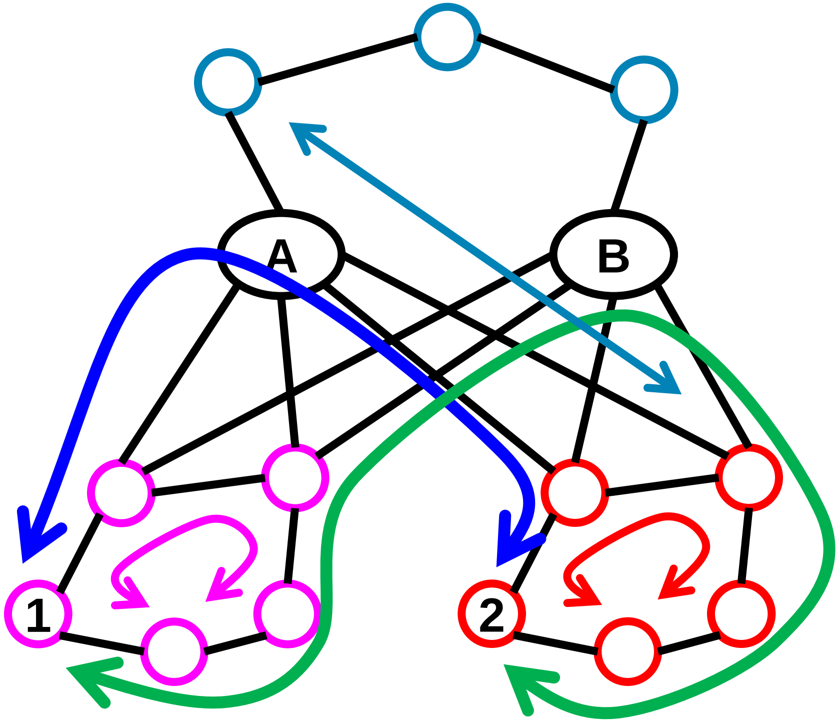
- If one is careful about VLAN usage, **one bridge can use multiple topology selection protocols at the same time** using its single Filtering Database.

Multiple topology protocols by VLAN

- This is not news to 802.1, but it is news to many others.
- **Standard bridge forwarding** using the 802.1 Filtering Information Database (FID) can be used with **many topology protocols simultaneously**, with the VLAN ID determining which topology a given frame follows:
  - o MSTP
  - o SPB-V
  - o SPB-M
  - o Protection Switching
  - o Duplicate delivery
  - o Various Ring Protocols

Which leads to: Network cores

- One can imagine an industrial network with **four cores**:



1. SPB-V protocol runs VLAN 1, that reaches everywhere.

2. Traffic engineered paths use VLAN 8 and VLAN 9.

3. Ring protocol runs VLAN 5 for local data.

4. Ring protocol runs VLAN 6 for local data.

Frames controlled by different topology control protocols can use the same Priority values, and hence the same queues.

Network cores: Separated by VLANs

- A “Network core” is a set of one or more **VLANs** that share a common **fault recovery mechanism** and a common **subset** (maybe all) **of the nodes and links** of a MAC Layer network.
- Network cores can **overlap** each other **arbitrarily**, since their traffic is separated by the VLAN ID carried in every frame.
- The preceding example shows a hierarchy of SPB-V above traffic engineering above ring, but hierarchy is not required.
    - o SPB-V allows management traffic to reach every node as long as any connectivity is present at all, with 1-second failover times.
    - o Traffic engineering supports 0-time failover for inter-ring data.
    - o Rings support very fast failover times for local data.

IS-IS under all

IS-IS

- The IS-IS protocol (Intermediate System to Intermediate System) is a link state protocol.
- **Every switch** (I use this term intentionally, as a node may be both a bridge and a router) **lists what links it has to neighbor switches**, along with additional information (e.g., its bridge number, what VLANs it needs to receive, and much more).
- Every switch puts all this information about itself into an "**advertisement**" that it sends to all of its neighbor switches.
- Every switch relays advertisements to and from its neighbors. The result of this flood behavior is that, eventually, **every switch receives every other switches' advertisements**.

SPB-V and IS-IS

- The IS-IS advertisements allow every Switch to construct a model of the topology of the **entire network**.
- Every switch uses that model to determine the path along which to forward any frame. **Every frame is forwarded along the least-cost path**, just like packets are forwarded by routers.
- The SPB-V uses the 12-bit VID field in the 802.1Q tag to encode the tree number == the source bridge ID. One VID value is required per VLAN per bridge. Practically speaking (though not technically), the number of bridges times the number of VLANs must be less than 4095.

SPB-V and IS-IS

- In an MSTP network, switches must exchange MVRPDUs and reconverge during and after a topology change, in order to propagate the VLAN requirements of each switch along the various spanning trees.
- In an SPB-V network, each switch’s VLAN needs are included in its advertisements. Therefore, after a topology change, every switch has the information it needs to recalculate what frames need to be pruned and what need to be forwarded **without the exchange of any MVRP PDUs**.
- That is, **SPB-V replaces PDU exchanges with additional computation**.

KEY IDEA #3

- **IS-IS** can perform a number of **protocol functions** at the same time, most (or even all!) of which can be **independent from** the **SPB-V** (or any other) topology control function.

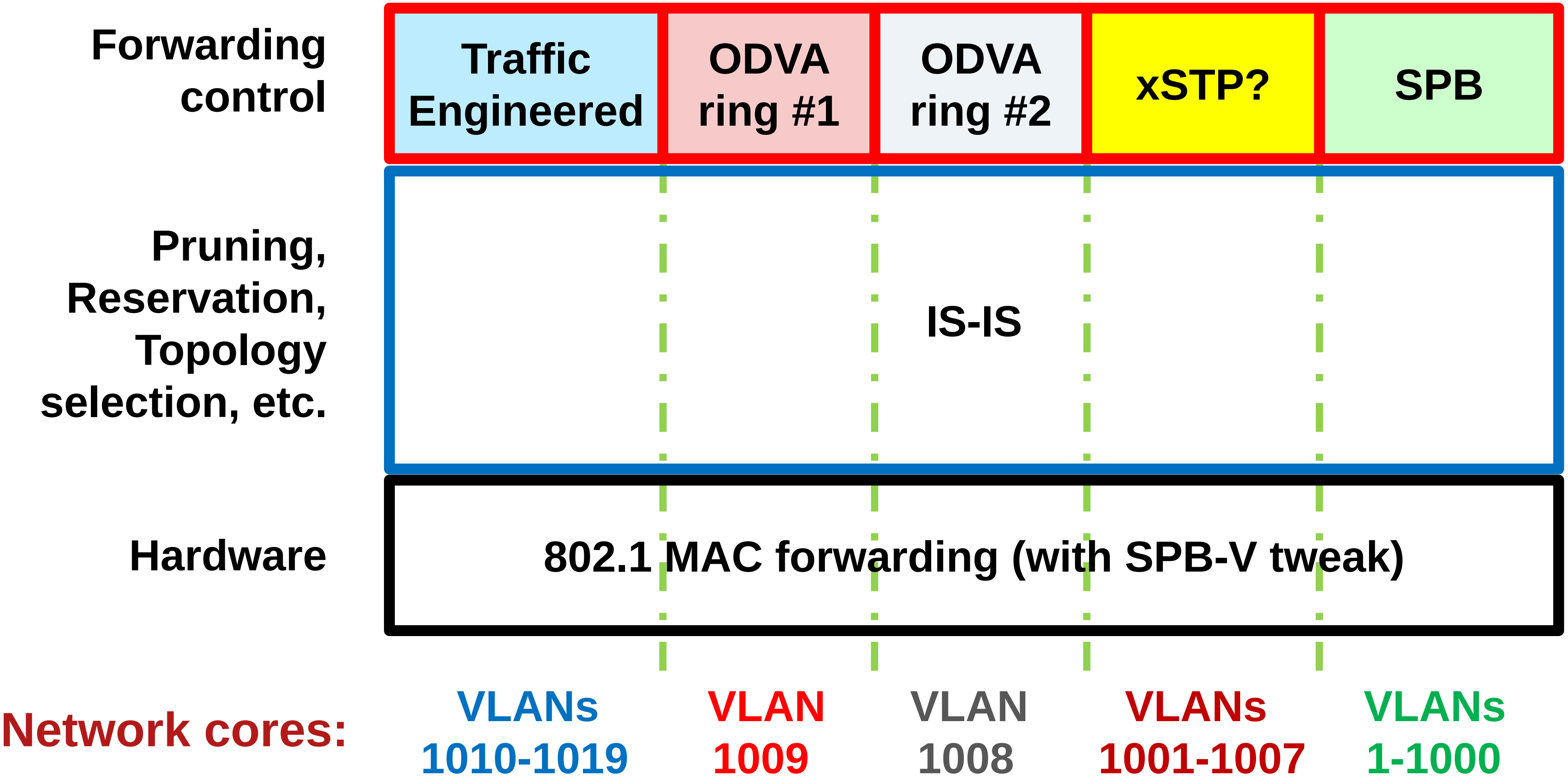
Orthogonality example: VLAN pruning

- Suppose VIDs 1-1000 control 500 SPB-V bridges that implement two VLANs. That is, the forwarding of any frame carrying VIDs 1-1000 is governed by the SPB-V port states.
- Suppose VIDs 1001-1010 are governed by ten G.8032 rings, using tagged Ether OAM.
- Suppose VIDs 2000-2019 control by MSTP, using ten “private VLANs”, each an even-odd pair of VIDs.
- There are, effectively, 22 VLANs: 1, 501, 1001-1010, and 2000, 2002, 2004, ... 2018.
- **Every VLAN could be pruned based on information passed by IS-IS. MVRP need not be used, and G.8032 does not need a pruning protocol!**

IS-IS outside the fault recovery context

- Thus, IS-IS could carry information about **SRP Talkers and Listeners**, so that the switches can allocate port bandwidth for any VLAN using **any topology control protocol, without** running any **SRPDUs**.
- IS-IS could carry information about 802.1AS **Master Clocks**, so that the clock distribution topology could reconverge more quickly after a failure.
- IS-IS topology information enables the members of a **ring** to verify that their actual connectivity matches their configured connectivity, **without** running a separate ring **topology verification** protocol.
- IS-IS topology information enables the switches on a **traffic-engineered path** to verify that the path is **valid**.

Brick Wall diagram



Forwarding control

Traffic Engineered

ODVA ring #1

ODVA ring #2

xSTP?

SPB

Pruning, Reservation, Topology selection, etc.

IS-IS

Hardware

802.1 MAC forwarding (with SPB-V tweak)

Network cores:

VLANs 1010-1019

VLAN 1009

VLAN 1008

VLANs 1001-1007

VLANs 1-1000

- IS-IS and normal MAC underlie all network cores.

Minimal switches

Minimal switches

- One of the biggest advantages possessed by designers of Deterministic Ethernet networks is that they are, so far, at least, relatively small (max ≈ 100s of switches, 1000s of stations) compared to a Data Center. **This is a good thing.** One can make a very good argument that Data-Center-sized networks are fundamentally inappropriate for Deterministic Ethernet applications.
- But, even in a small network, 100s of switches can result it non-trivial memory and computational requirements for every switch (100s of kbytes).
- It is worth investigating whether there are methods by which SPB-V **convergence time** or **pruning accuracy** could be **traded** against **memory** / computation requirements in switches with only **two network connections**.

End stations

End stations

- End stations are most flexible if they are VLAN-aware, with one virtual Ethernet port per VLAN.
- End stations need to tag frames with MAC priority, as well.
- Stations using only one VLAN can be VLAN-unaware; the switch can be configured to provide the VLAN.
- A switch can also, within the 802.1Q standard, assign different VLANs to different protocols from VLAN-unaware stations.
- Documenting the idea that VID != VLAN would enable improved use of shared media with mixtures of bridges and end stations.

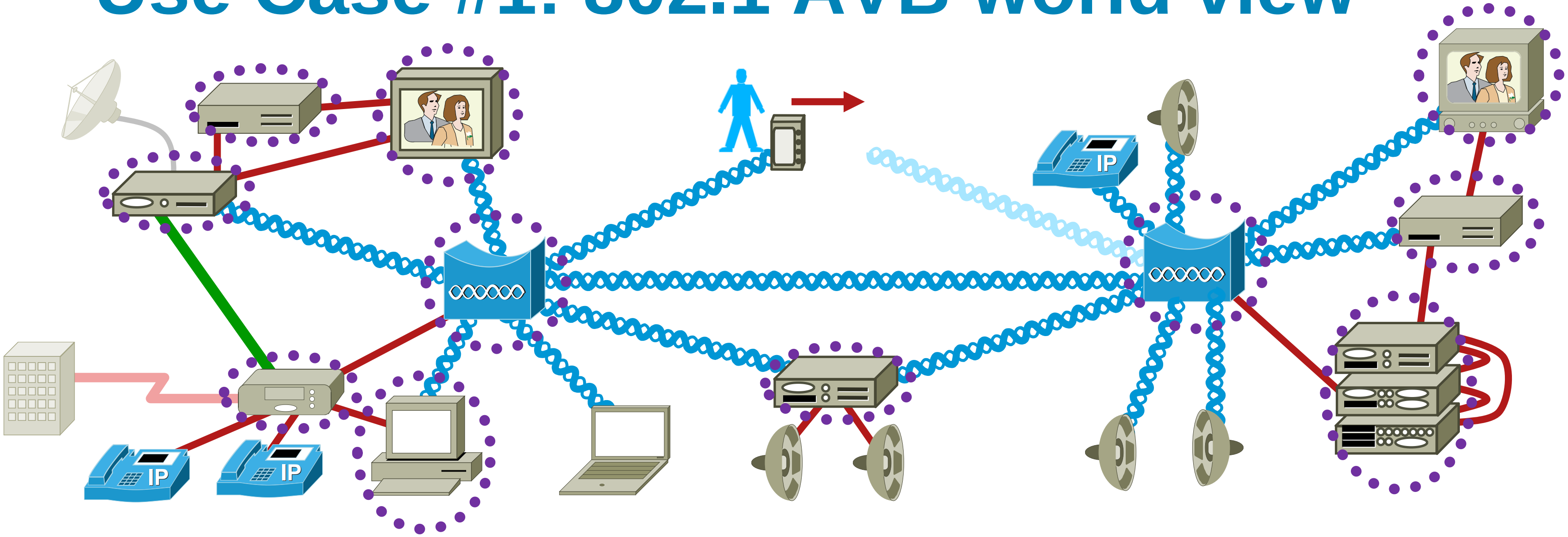
802.1AS and bridges

802.1AS enhancement

- 802.1AS went to a lot of trouble to eliminate passing packets through as data and modifying them in transit.
- At present, 802.1AS speaks of the system in which the time synchronization protocol runs as being either an end station or a bridge.
- In fact, there is no function of a bridge that is required for an 802.1AS “thing that is not an end station”.
- For this reason, **802.1AS needs to be divorced from**:
  - Whether the PDUs carry **MAC Layer or IP Layer** encapsulations and addresses.
  - Whether the device that relays time sync information is a **bridge**, a **router**, or a **multi-homed end station**.

WiFi is not at the edge!

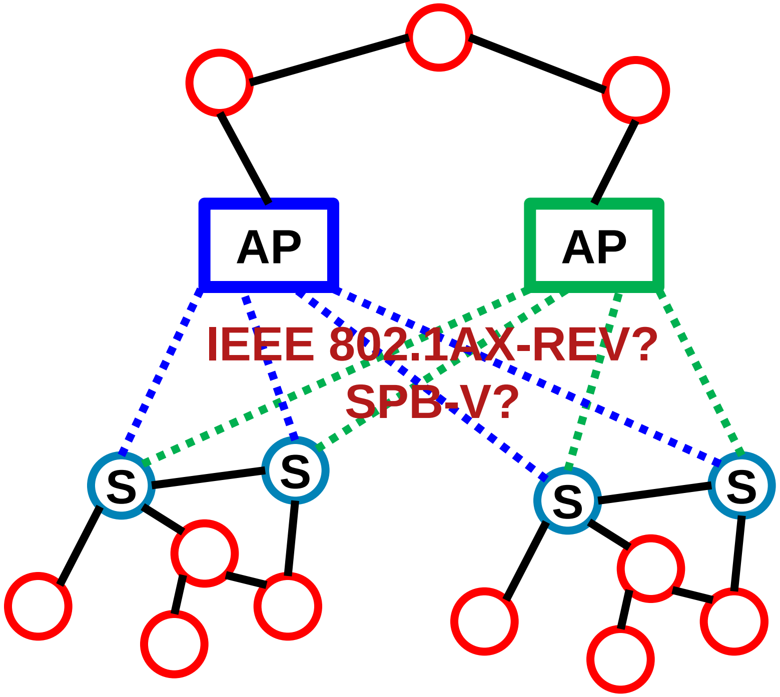
Use Case #1: 802.1 AVB world view



- In a home or small studio, there may be many Ethernet-like links: 802.3, 802.11, MoCA, Ether/DSL, etc.
- You expect wired stacks connected via wireless.
- To ensure connectivity, every device with multiple ports is an **802.1 bridge**, and **stations are bridges, too**!

Use Case #2: Factory floor

- Consider a possible industrial control network:



Backbone network

Access Points

Wireless “links”

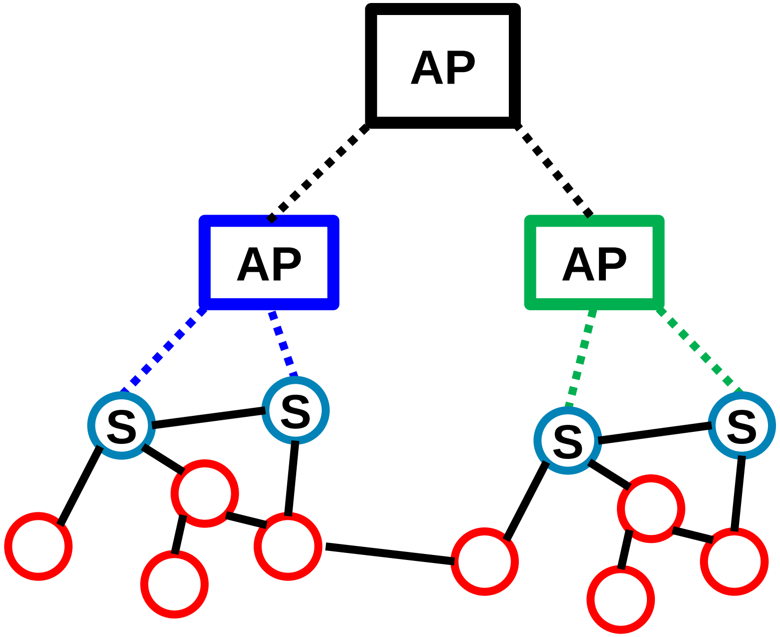
Non-AP Station/switches

Semi-mobile wired networks

- Either Link Aggregation or SPB-V bridging technologies can make this a **single MAC Layer network**.

Use Case #3: Tiered network access

- 802.11ac, gigabit WiFi, makes this even more imperative. **A Gb/s link is not always at the edge of the network!**



.11ac Access Point

1 Gb/s Wireless “links”

.11g Access Points

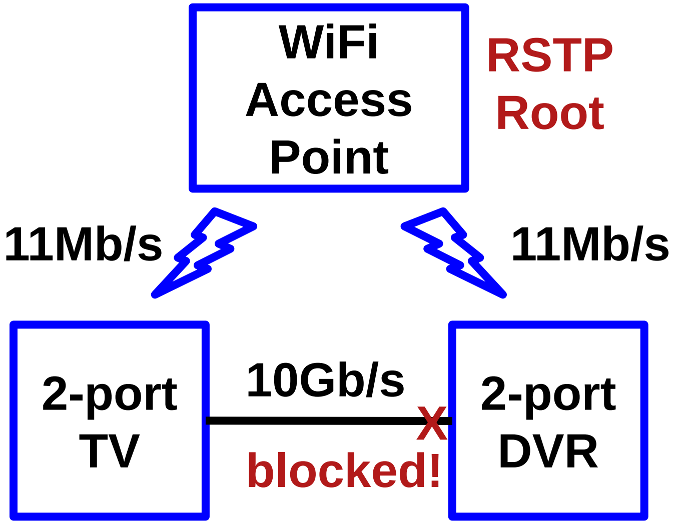
54 Mb/s Wireless “links”

Wired bits of a single company in a multi-tenant building.

- **Don’t build layers of ad-hoc solutions!** Simply make these devices ordinary switches.

WiFi is not at the edge of the network!

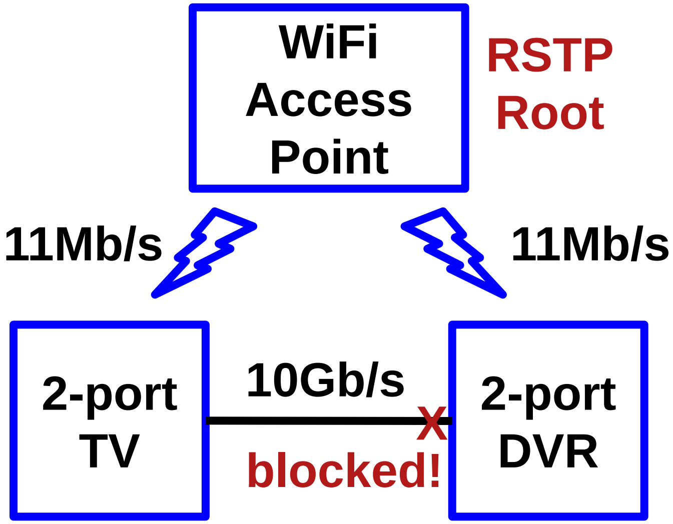
- As has been known for a long time, spanning tree has issues in simple networks with links of widely disparate data rates.



- This diagram illustrates the problem in the home.

WiFi is not at the edge of the network!

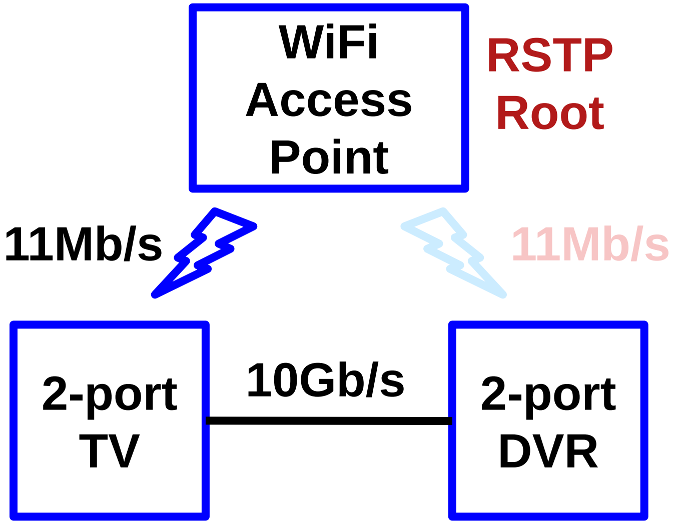
- But, the issue is more fundamental than “STP sucks”. **All of these use cases are illegal** according to the 802.11 architecture unless all wired/wireless boxes are **Access Points**. (Just what an apartment building needs!)



- This notion that 802.11 is always at the edge of the network has prevented any consideration of a standard means for accomplishing it, except the unsuccessful 802.11s.

WiFi is not at the edge of the network!

- There are ad hoc non-standard solutions for certain wired/wireless cases, such as when the second box does not have a wireless link.



WiFi Access Point

RSTP Root

11Mb/s

11Mb/s ← If this link not here,

2-port TV

10Gb/s

2-port DVR

← Wired link can be used (by non-standard means)

- But, the general case of an arbitrary network of wired and wireless links has **no viable MAC Layer standard**.

So, just make each station a bridge!

- **Duh!**
- Long answer: I have presentations on this issue going back to 2005 (e.g. http://www.ieee802.org/1/files/public/docs2005/new-nfinn-generalized-lan-emulation-ieee-0305.pdf).
- Short answer: The AP reflects my (a station/bridge's) own broadcasts back to me, and that breaks MAC address learning.
- **802.11 and 802.1 have to fix this!**
- **Fix 1**: Bridges use the 802.11n headers to suppress their own reflections. (Applies to any protocol.)
- **Fix 2**: Use host MAC address routing instead of learning. (SPB-V can do this.)

IS-IS is too haaaaard!

IS-IS is too haaaaard!

- We can work this problem piecemeal, one issue and one fix at a time, or we can solve the whole problem at once.
- Divide and conquer: Apply each fault recovery mechanism in its sweet spot, using all in parallel, with IS-IS as the underlying and unifying starting point.
- The argument that “IS-IS is too complex” is short-sighted. What is “too complex” is trying to replicate the same capabilities across five suites of competing protocols. What is “too complex” is trying to force the one protocol to do two incompatible jobs, because the two protocols that can do those jobs are incompatible.
- And, we may be able to simplify IS-IS for the smallest devices.

Work plan and Summary

Work Plan

- **IEEE 802** needs to work on:
  - Our **current efforts** (802.1ASbt, 802.1Qbu, 802.1Qbv, 802.3 new 1 Gb/s Ethernet, 802.3 preemption?).
  - **Extending IS-IS** to:
    - Incorporate the Stream Reservation Protocol (and support multicasts and multiple talkers).
    - Control Master Clock Distribution (part of .1ASbt?).
    - Distribute Traffic Engineered Paths.
    - Protect the network from Rings and/or Traffic Engineered paths, when catastrophic topology errors are present.
    - Be useable by minimally-intelligent bridges.
  - A useable standard for **end stations**.
  - Amend **802.1AS** to not speak in terms of bridges.
  - Correcting the 802.11 illusion that **WiFi** is only at the edge.
- **Multiple standards groups** need to work together to:
  - Support different topology control protocols on different **VLANs**.

Summary

- The work plan for IEEE 802 and for other Standards Development Organizations should be centered on:
  - Using **VLANs** to separate **Network cores** that have separate addressing and separate forwarding topology control protocols.
  - Basing QoS features on the **Q-tag Priority field**, orthogonal to VLANs and forwarding topology control protocols.
  - Using **IS-IS** as a **unifying basis** for ancillary protocols for all forwarding topology control protocols.
- The result is that:
  - A **maze of incompatible standards** becomes a **menu of customer choices**.
  - But, **IEEE 802.1 must keep moving** to continue to provide a sound base for all these efforts.