# Shortest Path Bridging
# IEEE 802.1aq
# Overview

**APRICOT/Hong Kong/Feb 24th 2011**

Peter Ashwood-Smith

*peter.ashwoodsmith@huawei.com*

Fellow

HUAWEI

# Abstract

*802.1aq Shortest Path Bridging is being standardized by the IEEE as an evolution of the various spanning tree protocols. 802.1aq allows for true shortest path routing, multiple equal cost paths, much larger layer 2 topologies, faster convergence, vastly improved use of the mesh topology, single point provisioning for logical membership (E-LINE/E-LAN/E-TREE etc), abstraction of attached device MAC addresses from the transit devices, head end and/or transit multicast replication , all while supporting the full suit of 802.1 OA&M.*

*Applications consist of STP replacement, Data Center L2 fabric control, L2 Internet Distributed Exchange point fabric control, small to medium sized Metro Ethernet control planes. L2 wireless network backhaul….*

# Outline

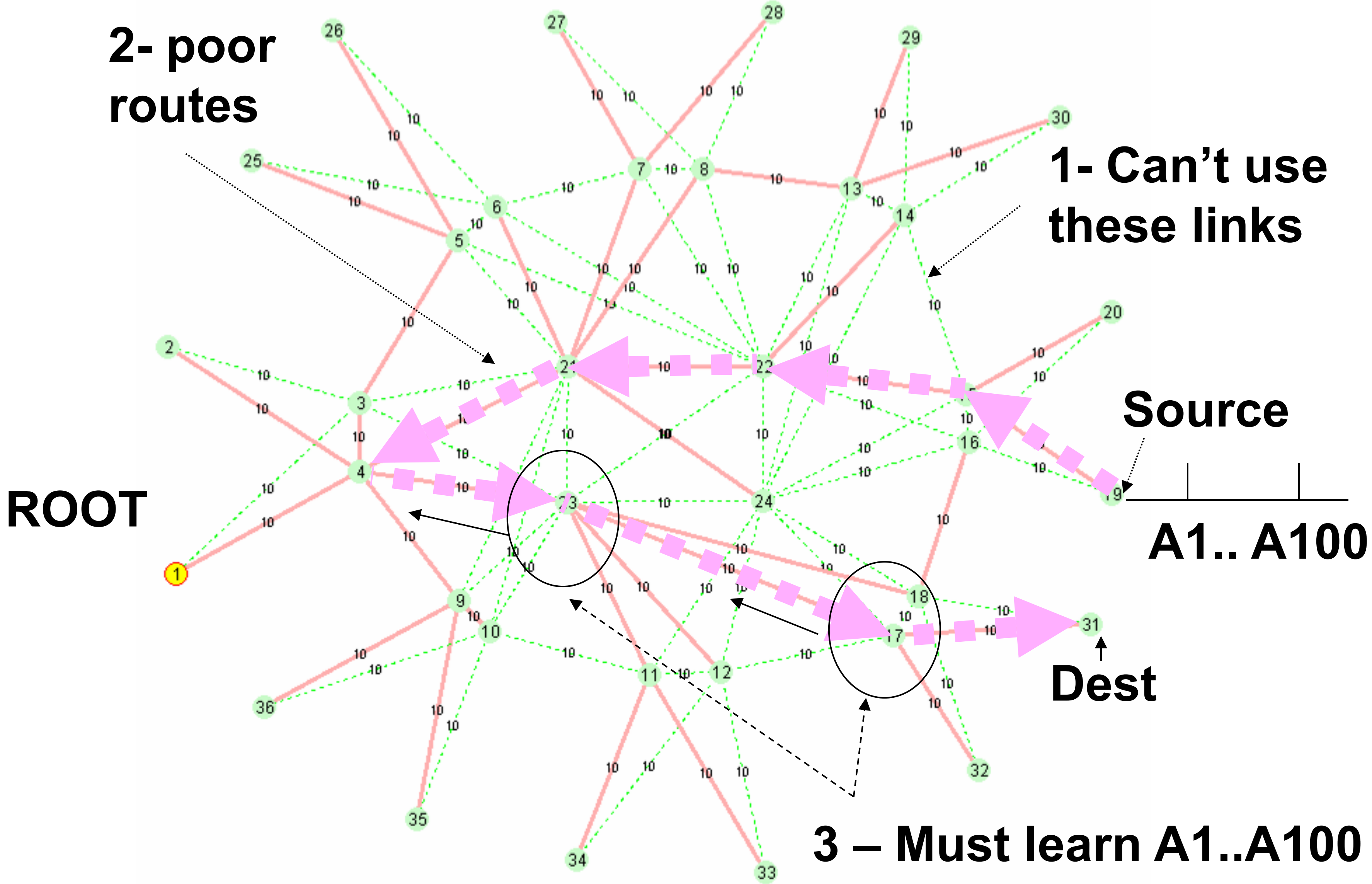
- **<u>Challenges</u>**
- What is 802.1aq/SPB
- Applications
- How does it work
- Status

# Challenges

- L2 networks that scale to ~1000 bridges.
- Use of arbitrary mesh topologies.
- Use of (multiple) shortest paths.
- Efficient broadcast/multicast routing and replication points.
- Avoid address learning by tandem devices.
- Get recovery times into 100's of millisecond range for larger topologies.
- Good scaling without loops.
- Allow creation of very many logical L2 topologies (subnets) of arbitrary span.
- Maintain all L2 properties within the logical L2 topologies (transparency, ordering, symmetry, congruence, shortest path etc).
- Reuse all existing Ethernet OA&M 802.1ag/Y.1731

**"Make a network of switches look like a single switch!"**

# Example problems of scaling up Native Ethernet



2- poor routes

1- Can't use these links

Source

ROOT

A1.. A100

Dest

3 – Must learn A1..A100

# Outline

- Challenges
- **What is 802.1aq/SPB**
- Applications
- How does it work
- Status

# What is 802.1aq/SPB

- **IEEE protocol builds on 802.1 standards**
- **A new control plane for Q-in-Q and M-in-M**
  - Leverage existing inexpensive ASICs
  - Q-in-Q mode called SPBV
  - M-in-M mode called SPBM
- **Backward compatible to 802.1**
  - 802.1ag, Y.1731, Data Center Bridging suite
- **Multiple loop free shortest paths routing**
  - Excellent use of mesh connectivity
  - Currently 16, path to 1000's including hashed per hop.
- **Optimum multicast**
  - head end or tandem replication

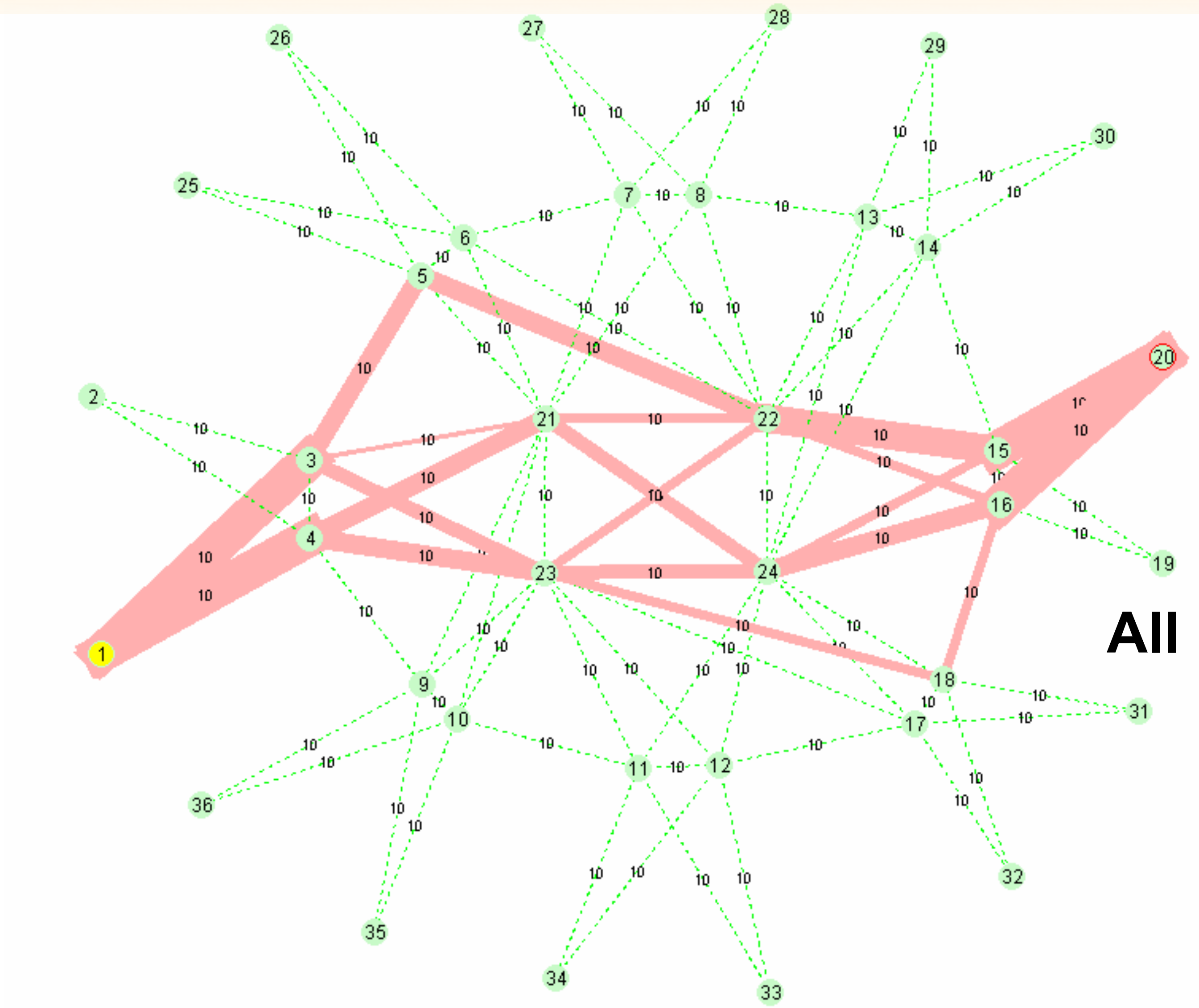
# What is 802.1aq/SPB (cont'd)

- **Light weight form of traffic engineering**
  - Head end assignment of traffic to 16 shortest paths.
  - Deterministic routing - offline tools predict exact routes.
- **Scales to ~1000 or so devices**
  - Uses IS-IS already proven well beyond 1000.
  - Huge improvement over the STP scales.
- **Good convergence with minimal fuss**
  - sub second (modern processor, well designed)
  - below 100ms (use of hardware multicast for updates)
  - Includes multicast flow when replication point dies. Pre-standard seeing 300ms recovery @ ~50 nodes.
- **IS-IS**
  - Operate as independent IS-IS instance, or within IS-IS/IP, supports Multi Topology to allow multiple instances efficiently.

# What is 802.1aq/SPB (cont'd)

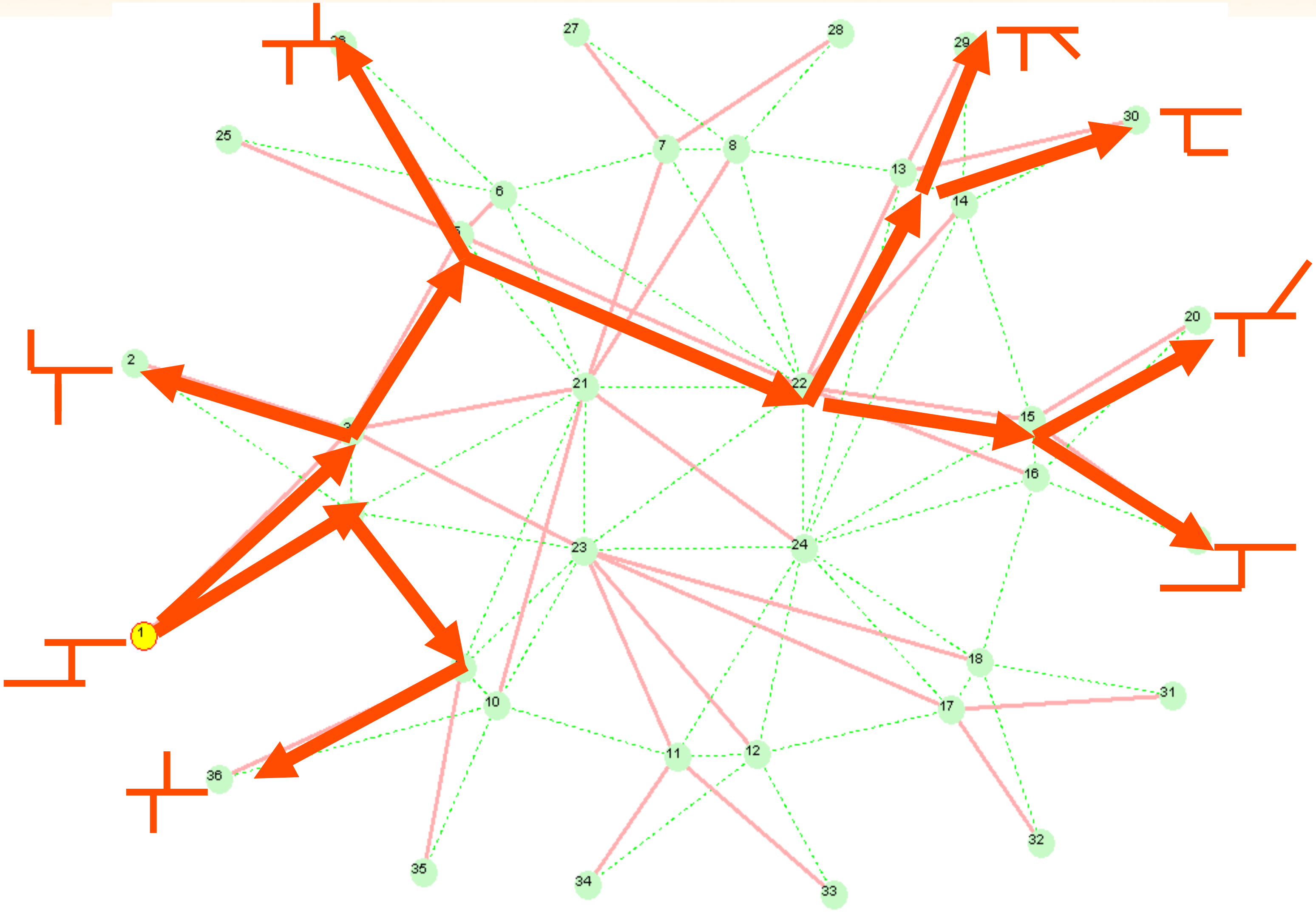
- **Membership advertised in same protocol as topology.**
  - Minimizes complexity, near plug-and-play
  - Support E-LINE/E-LAN/E-TREE
  - All just variations on membership attributes.
- **Address learning restricted to edge (M-in-M)**
  - FDB is computed and populated just like a router.
  - Unicast and Multicast handled at same time.
  - Nodal or Card/Port addressing for dual homing.
- **Computations guarantee ucast/mcast…**
  - Symmetry (same in both directions)
  - Congruence (unicast/multicast follow same route)
  - Tune-ability (currently 16 equal costs paths – opaque allows more)

# End result - Visually



All links usable

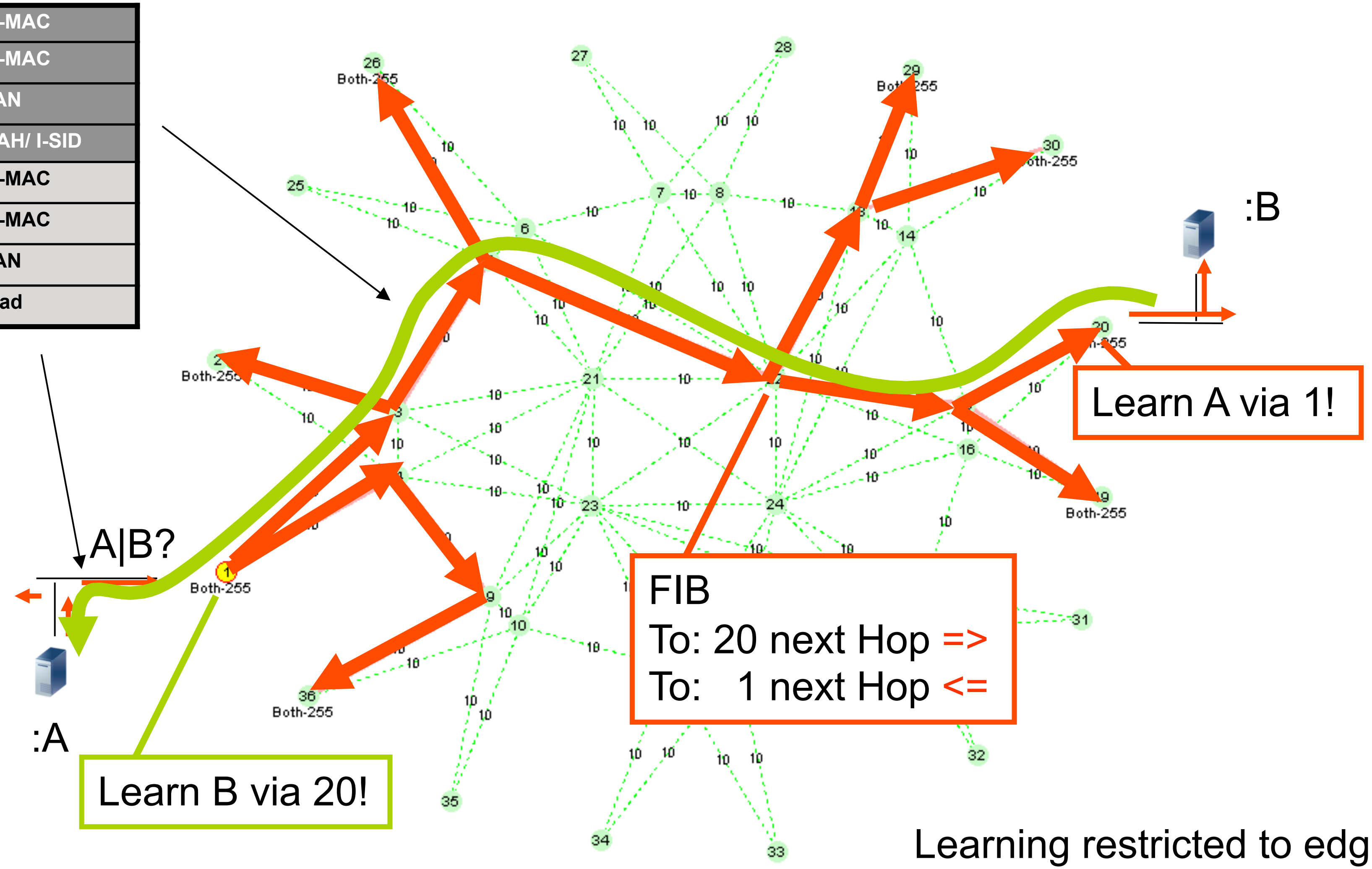
# SPF trees form multicast template



Shortest Path First Tree becomes template for multicast tree and is pruned automatically to proper membership.

# Edge Learning - Visually

- Dst.B-MAC
- Src.B-MAC
- B-VLAN
- 801.1AH/ I-SID
- Dst.C-MAC
- Src.C-MAC
- C-VLAN
- Payload



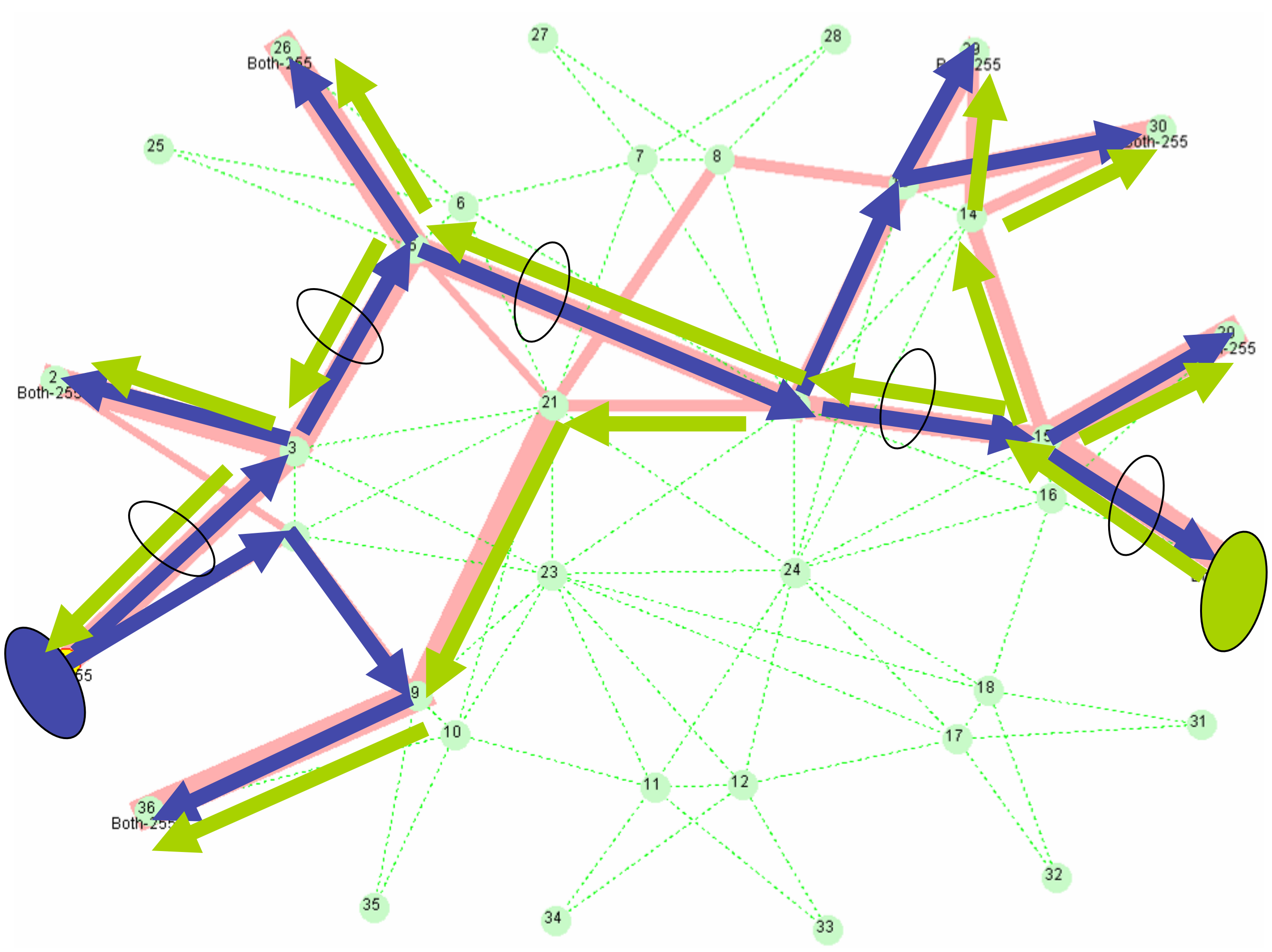
Learn A via 1!

FIB
To: 20 next Hop =>
To: 1 next Hop <=

Learn B via 20!

Learning restricted to edges and only where I-SID tree reaches. Mac-in-Mac encap.

# Animation for 8 member E-LAN ‘255’



I-SID 255 has 8 members

Shown are all routes used by this I-SID in pink.

Two trees shown blue/green.

Note symmetry of trees between source/dest

If transit multicast selected fork points in trees are replication points.

# Outline

- Challenges
- What is 802.1aq/SPB
- **<u>Applications</u>**
- How does it work
- Status

# Applications

- Anywhere that Spanning Tree is being used.

  Take existing STP/MSTP based network and migrate to Shortest Path Routing.

- Ethernet Exchange Points

  Big distributed switch to interconnect hundreds of different customers cheaply with L2VPNs.

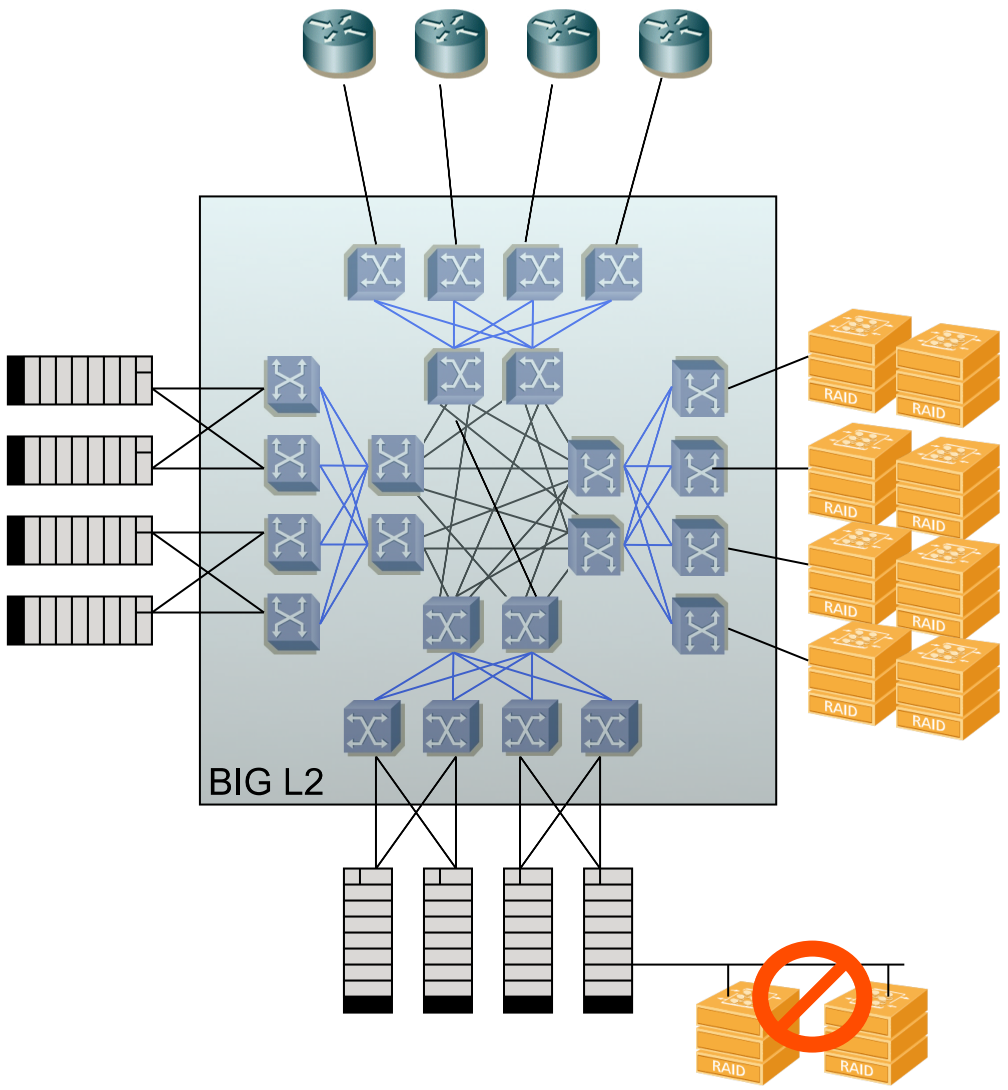
- Metro Ethernet

  Light weight metro protocol, L2VPN solution simpler than VPLS with lower capex/opex.

- Wireless backhaul

  Use of L2VPN for LTE backhaul

# Application Data Center



Treat DC network as one big L2 switch by combining 100's of smaller switches in 'non blocking' topology – why?

- Any server anywhere.
- Any router anywhere.
- Any appliance anywhere.
- Any VM anywhere.
  - Any IP address anywhere.
  - Any subnet anywhere.
- Any storage anywhere.
- Minimal congestion issues.
- Total flexibility for power use

# Application Data Center

- Multiple shortest path routing
  - inter server traffic

- Deterministic traffic flows.

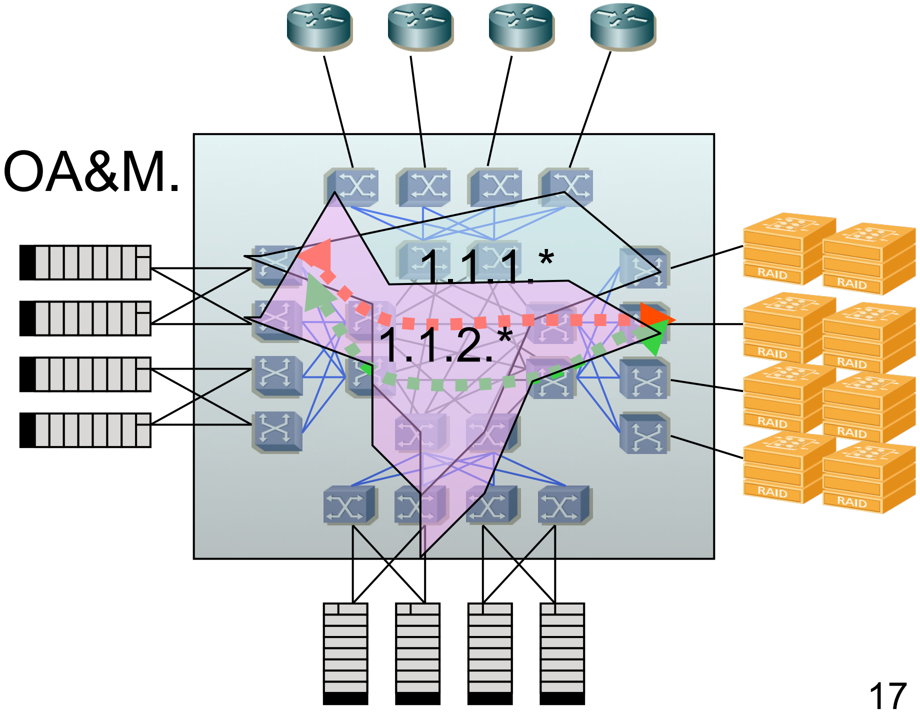
- Flexible subnet – expand/shrink anywhere.
  - Virtualization operates in subnet.

- Fully compatible with all 802.1
  Data Center Bridging protocols & OA&M.

- Address isolation through m-in-m

- Fast recovery

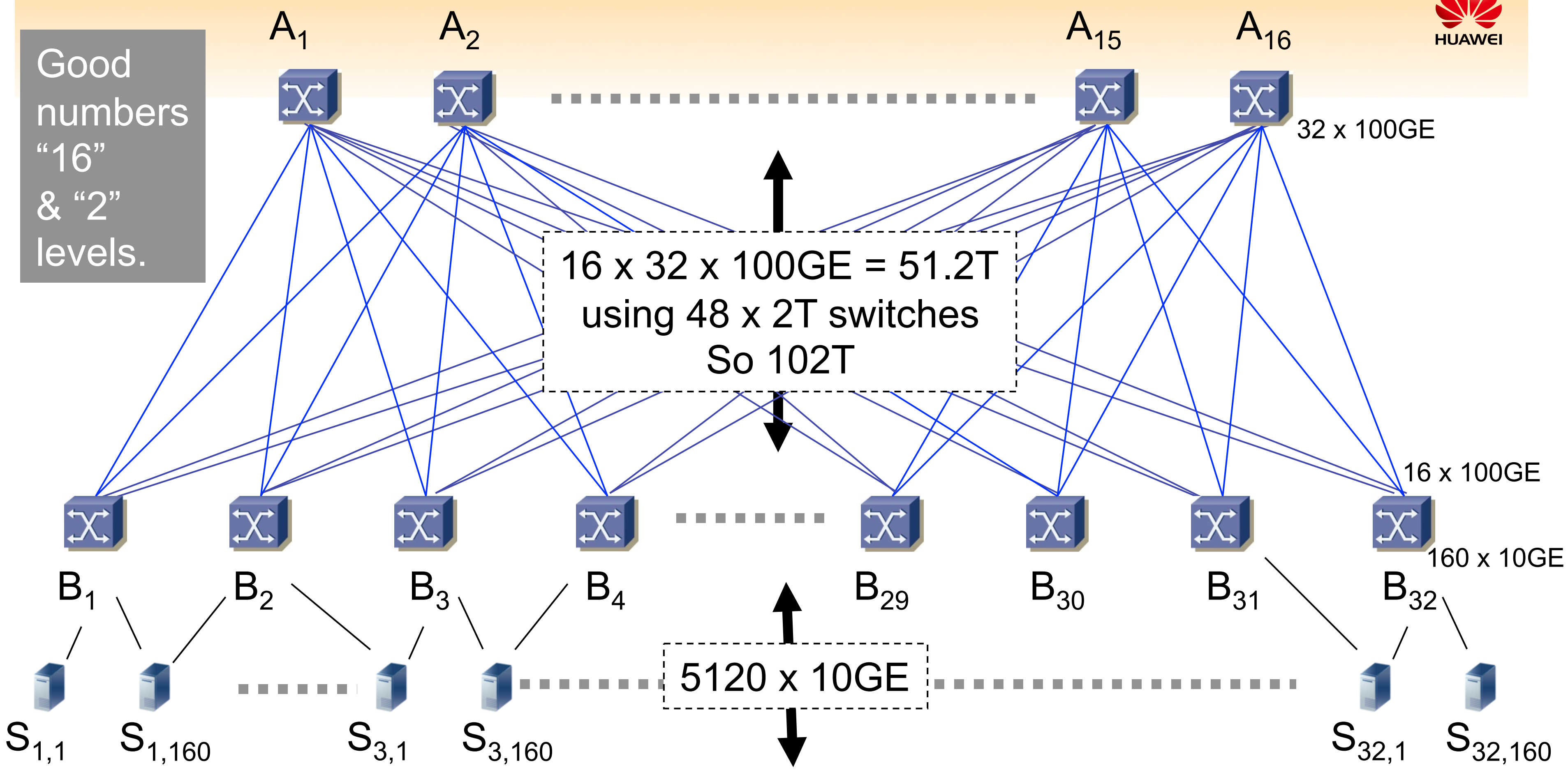
- No loops



# **Application Data Center** (cont'd)

- Totally compatible with Vmware server functions:
  - OA&M, motion, backup etc.
  - Apps that sit on Vmware 'just work'.
- Fully compatible with all load balancer ADC appliances.
- VRRP transparent (primary/stdby rtr per subnet) or proprietary variations on same protocol.
- Compatible with emerging Inter DC overlay work or Inter DC L2 tunnels.

# Non Blocking Switching Cluster



Good numbers “16” & “2” levels.

16 x 32 x 100GE = 51.2T using 48 x 2T switches So 102T

5120 x 10GE

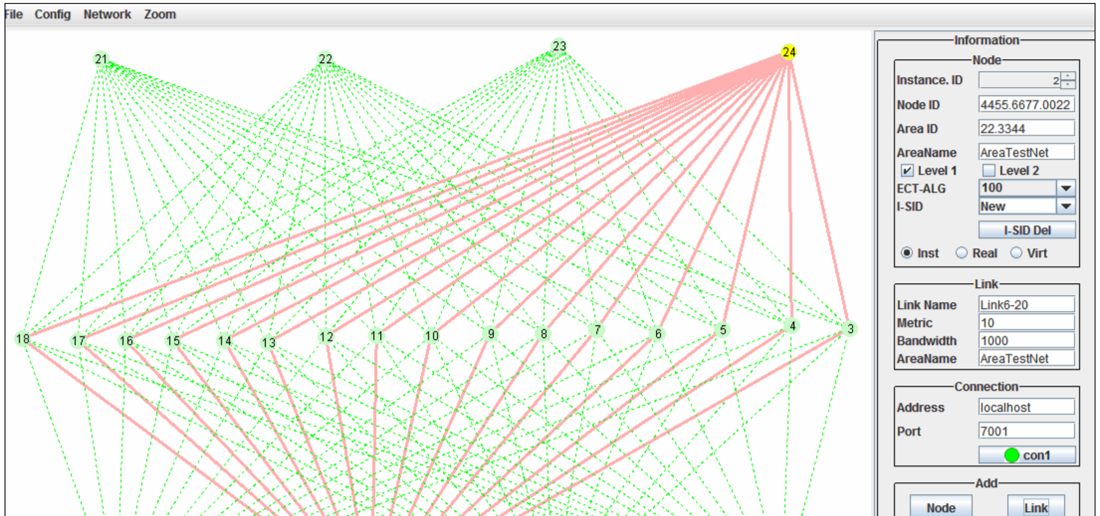
- 48 switch non blocking 2 layer L2 fabric
- 16 at “upper” layer $\mathbf{A_1..A_{16}}$
- 32 at “lower” layer $\mathbf{B_1.. B_{32}}$
- 16 uplinks per $\mathbf{B_{n,}}$ & 160 UNI links per $\mathbf{B_n}$
- 32 downlinks per $\mathbf{A_n}$

- (16 x 100GE per $\mathbf{B_n}$ )x32 = 512x100GE = 51.2T
- 160 x 10GE server links (UNI) per $\mathbf{B_n}$
- (32 x 160)/2 = **2560 servers @ 2x10GE** per

100+ Terra non blocking interconnection fabric (if switches non blocking)

# ECMP in DC

Can get perfect balance down spine of a two layer 16 ECT L2 Fabric. Shown Are all 16 SPF's from 2<->24



Information

Node

Instance. ID 2

Node ID 4455.6677.0022

Area ID 22.3344

AreaName AreaTestNet

☑ Level 1 □ Level 2

ECT-ALG 100

I-SID New

I-SID Del

● Inst ○ Real ○ Virt

Link

Link Name Link6-20

Metric 10

Bandwidth 1000

AreaName AreaTestNet

Connection

Address localhost

Port 7001

● con1

Add

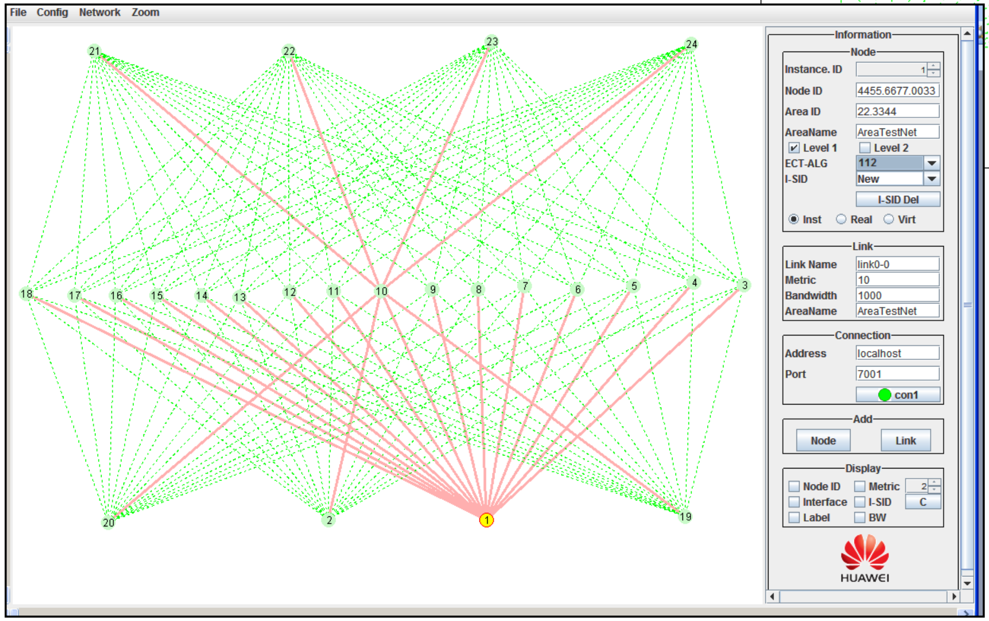
Node Link

Display

□ Node ID □ Metric 3

□ Interface □ I-SID C

□ Label □ BW



Information

Node

Instance. ID 1

Node ID 4455.6677.0033

Area ID 22.3344

AreaName AreaTestNet

☑ Level 1 □ Level 2

ECT-ALG 112

I-SID New

I-SID Del

● Inst ○ Real ○ Virt

Link

Link Name link0-0

Metric 10

Bandwidth 1000

AreaName AreaTestNet

Connection

Address localhost

Port 7001

● con1

Add

Node Link

Display

□ Node ID □ Metric 2

□ Interface □ I-SID C

□ Label □ BW

16 different SPF trees Each use different spine as replication point. Shown is one of the 16 SPF's from/to node 1.

# Outline

- Challenges
- What is 802.1aq/SPB
- Applications
- **How does it work**
- Status

# How does it work?

- **From Operators Perspective**
  - Plug NNI's together
  - Group ports/c-vlan/s-vlan at UNIs that you want to bridge ($2^{24}$ groups='services' m-in-m mode.)
  - Assign an I-SID to each group..
  - Use your .1ag OA&M
- **Internally**
  - IS-IS reads box MAC, forms NNI adjacencies
  - IS-IS advertises box MACs (so no config).
  - IS-IS reads UNI port services and advertises.
  - Computations produce FIBs that bridge service members.

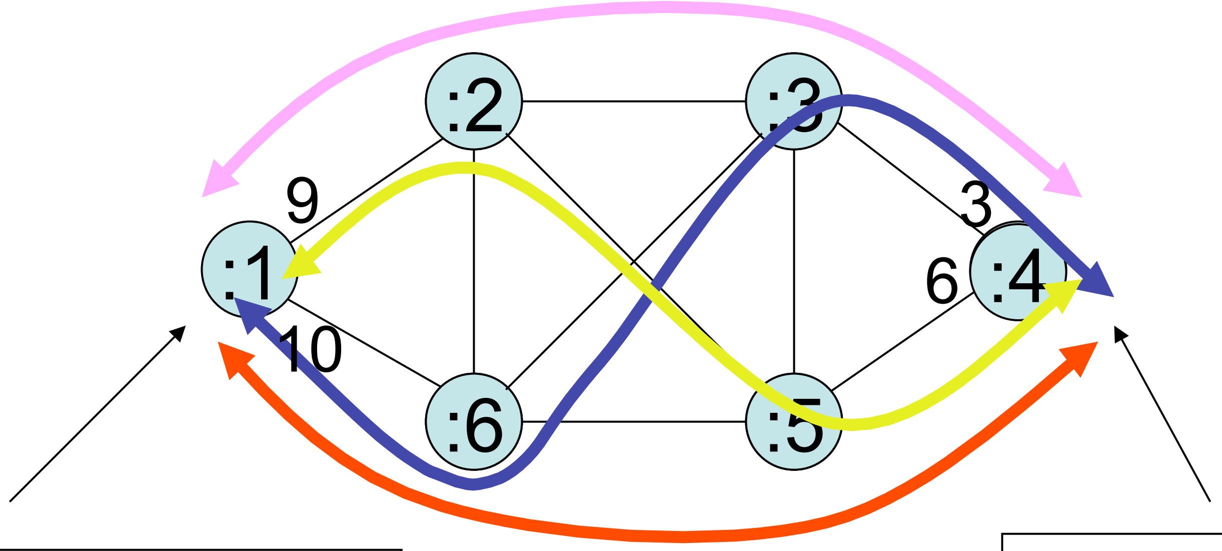
# Data Path (M-in-M mode)

- C-vlan/S-vlan or untagged traffic arrives at UNI
- Its encapsulated with B-SA of bridge
- Its encapsulated with I-SID configured for group
- Its encapsulated with B-VID chosen for route
- C-DA is looked up, if found B-DA is set
- C-DA not found, B-DA is multicast that says:
  - Multicast to all other members of this I-SID group from ‘me’. Or can head-end replicate over unicast.
  - C addresses to B address association learned at UNI only.

# FDB (unicast M-in-M mode)

- A unique shortest path from node to all others is computed.
- BMAC of other nodes installed in FIB pointing to appropriate out interface.
- Above is repeated for 16+ shortest paths each causes a different B-VID to be used.
- Symmetry is assured through special tie-breaking logic. 16+ different tie-breaking algorithms permit 16+ different shortest paths.

# FDB visually: ucast m-in-m mode



| MAC | BVID | IF |
|---|---|---|
| :4 | 1 | 9 |
| :4 | 2 | 9 |
| :4 | 3 | 10 |
| :4 | 4 | 10 |

| MAC | BVID | IF |
|---|---|---|
| :1 | 1 | 3 |
| :1 | 2 | 6 |
| :1 | 3 | 3 |
| :1 | 4 | 6 |

# FDB (mcast M-in-M mode)

***If*** no services require tandem replication
there is no tandem FDB:

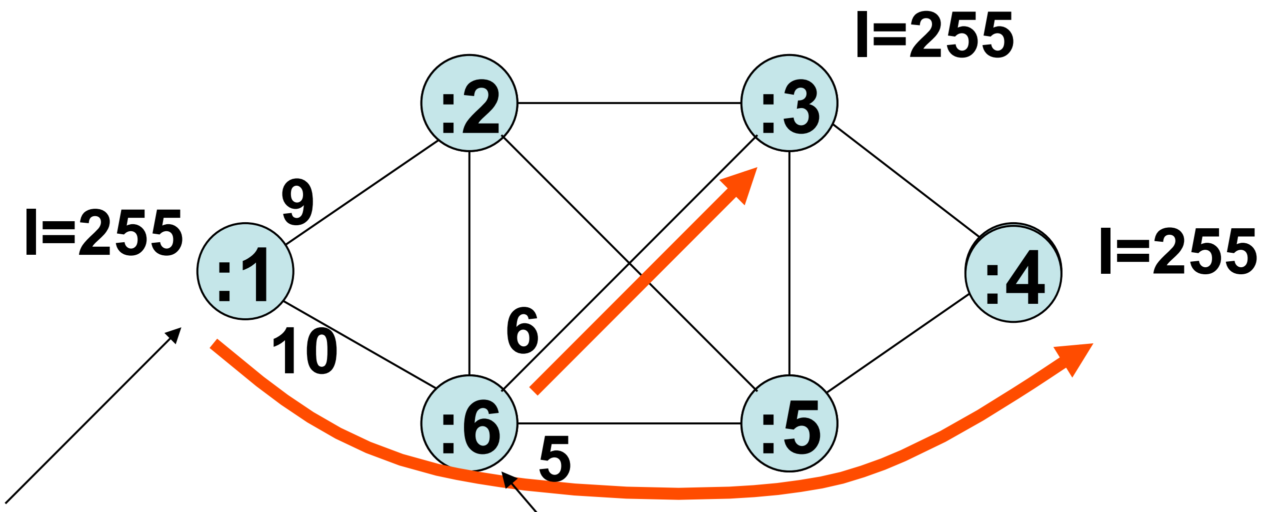
  Very VPLS like .. Pretty boring….head replication over unicast paths

***Else*** (mp2mp)

  ***If*** my node is on a unique shortest path between node **A** , (which transmits for a group **I**) and node **B**
  (which receives on the same group **I)**, then:

    merge into the FDB an entry for traffic from
    DA ={ **A**/Group **I**} to the interface towards **B**.

# FDB visually: mcast m-in-m mode

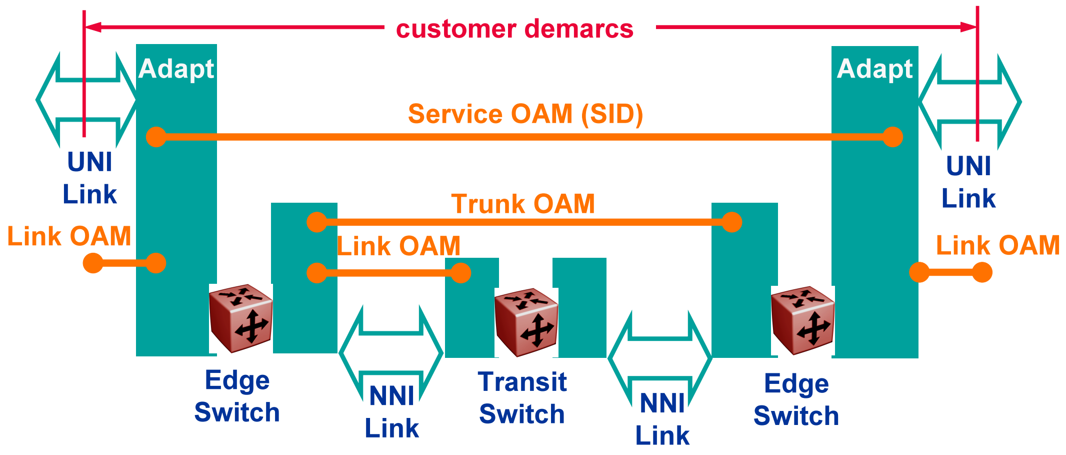


| MMAC | BVID | IF |
|---|---|---|
| {:1/255} | 4 | 10 |

| MMAC | BVID | IF |
|---|---|---|
| {:1/255} | 4 | 5,6 |

# 802.1aq OAM capabilities

1. **Continuity Check (CC)**
   a) Multicast/unidirectional heartbeat
   b) Usage: Fault detection
2. **Loopback – Connectivity Check**
   a) Unicast bi-directional request/response
   b) Usage: Fault verification
3. **Traceroute (i.e., Link trace)**
   a) Trace nodes in path to a specified target node
   b) Usage: Fault Isolation



4. **Discovery** (not specifically supported by .1ag however Y.1731 and 802.1ab support it)
   a) Service (e.g. discover all nodes supporting common service instance)
   b) Network (e.g. discover all devices common to a domain)
5. **Performance Monitoring** (MEF10 and 12 - Y.1731 for pt-pt now extending to pt-mpt and mpt-mpt)
   a) Frame Delay, Frame Loss, Frame Delay Variation (derived)
   b) Usage: Capacity planning, SLA reporting

# Outline

- Challenges
- What is 802.1aq/SPB
- Applications
- How does it work
- **Status**

# Status

- DEPLOYMENTS:
  - Pre-standard SPBM live customer networks:
    - 3 carrier (20+nodes)
    - 5 enterprise
    - 3 dc deployments
  - SPBM Data path (PBB) and OA&M of course has large number of deployments world wide.
- INTERWORKING:
  - Avaya (ERS 8800) + Huawei (S9300) successful Inter-working including full line rate data paths + L2 ping x 5 physical 32 logical nodes
- IETF:
  - In IESG last call, RFC imminent ~1Q 11
- IEEE:
  - Expected completion ~3Q 11.

# References

"***IEEE 802.1aq***" : www.wikipedia.org:
http://en.wikipedia.org/wiki/IEEE_802.1aq
Good overview, up to date with lots of references / tutorial videos all linked.

http://www.ietf.org/internet-drafts/draft-ietf-isis-ieee-aq-04.txt
The IETF IS-IS draft soon to be RFC.

"***Shortest Path Bridging*** – Efficient Control of Larger Ethernet Networks" :
IEEE Communications Magazine – Oct 2010

"***Provider Link State Bridging***" :
IEEE Communications Magazine V46/N9– Sept 2008

## Thank-You