

Shortest Path Bridging IEEE 802.1aq

NANOG49 June 13-16/2010

Peter Ashwood-Smith
Fellow



peter.ashwoodsmith@huawei.com

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.

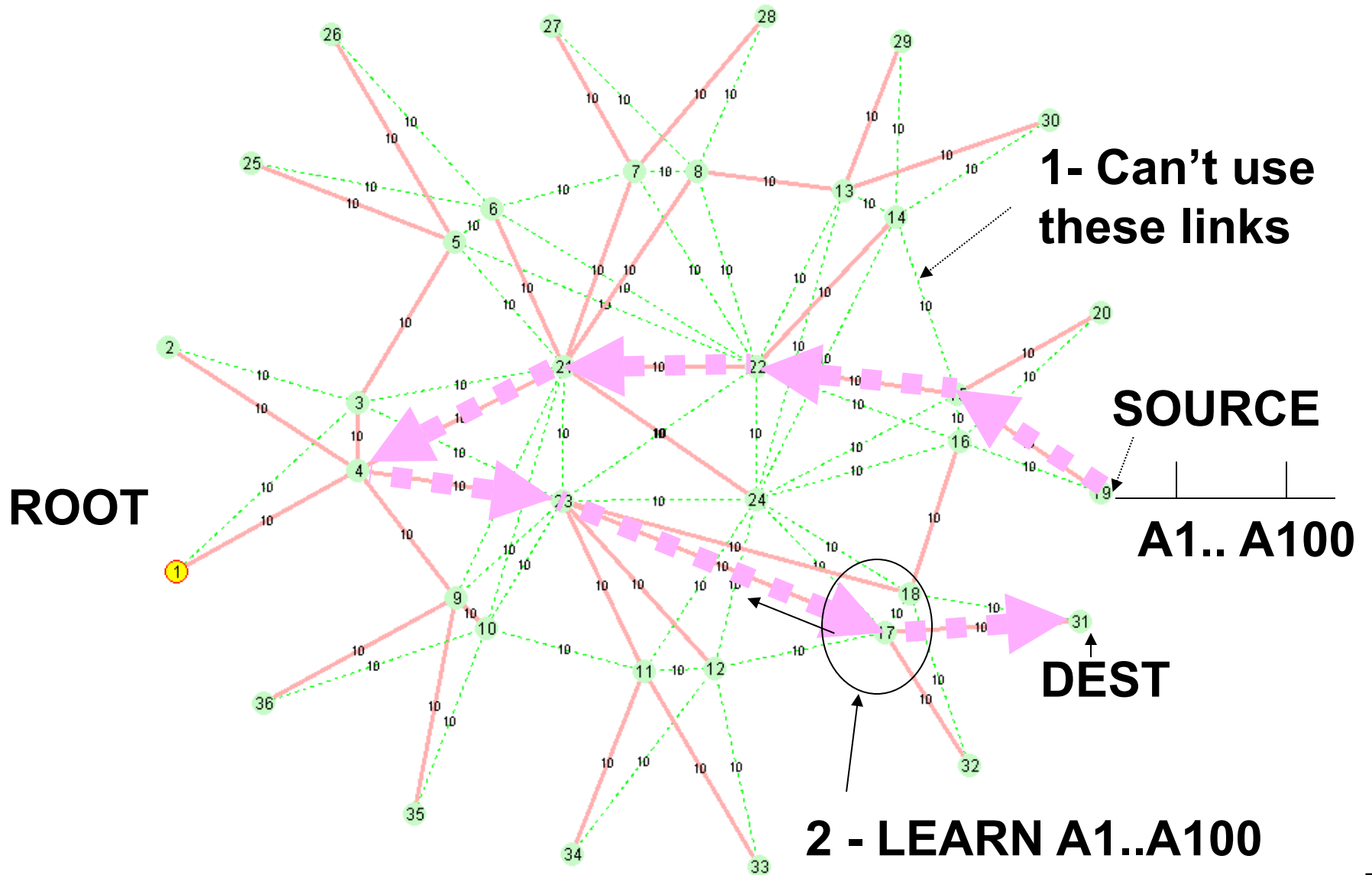
Outline

- Challenges
- What is 802.1aq/SPB
- Applications
- How does it work
- Example (won't cover but included here)

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

Example STP 36 nodes



Outline

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

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

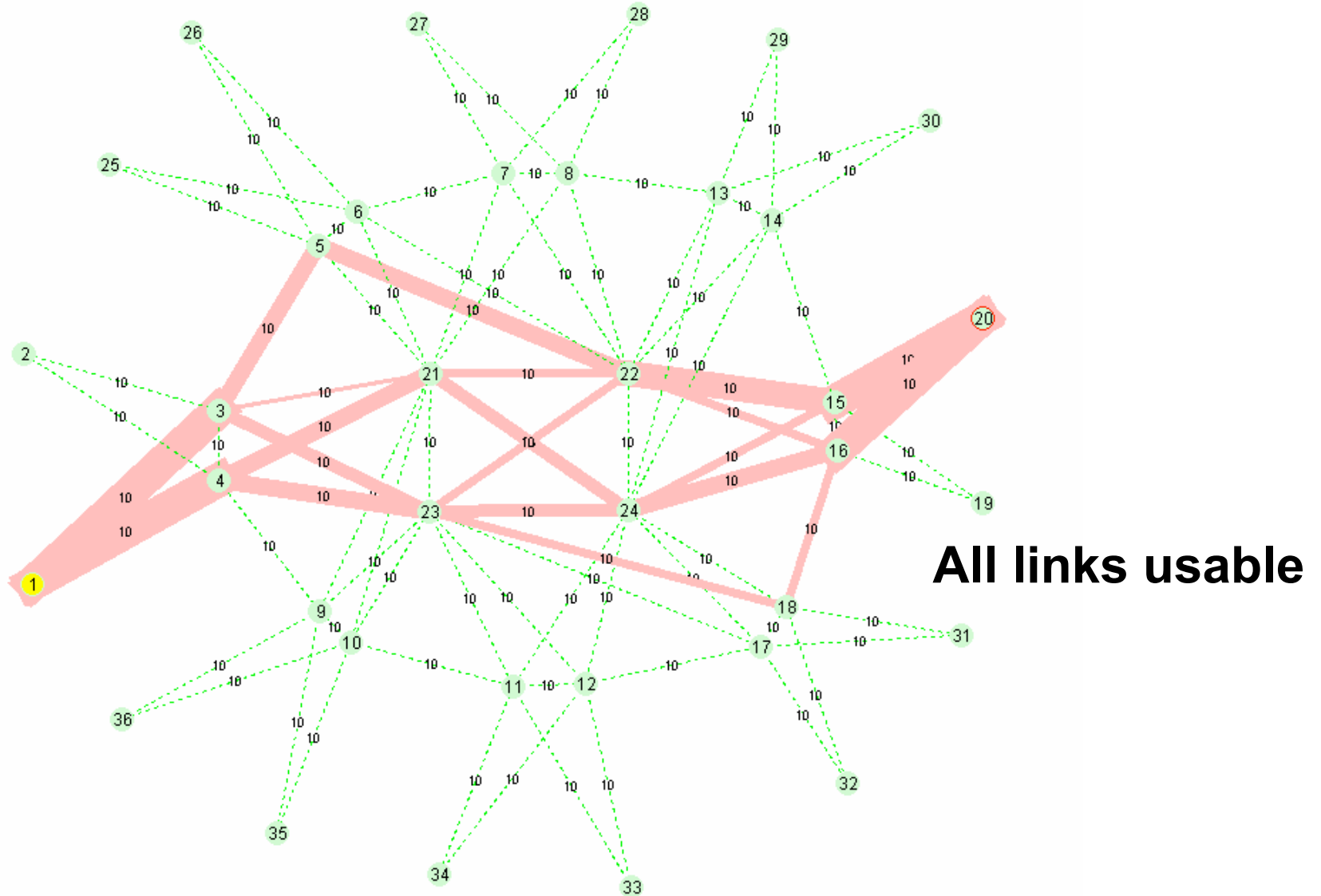
What is 802.1aq/SPB (cont'd)

- **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

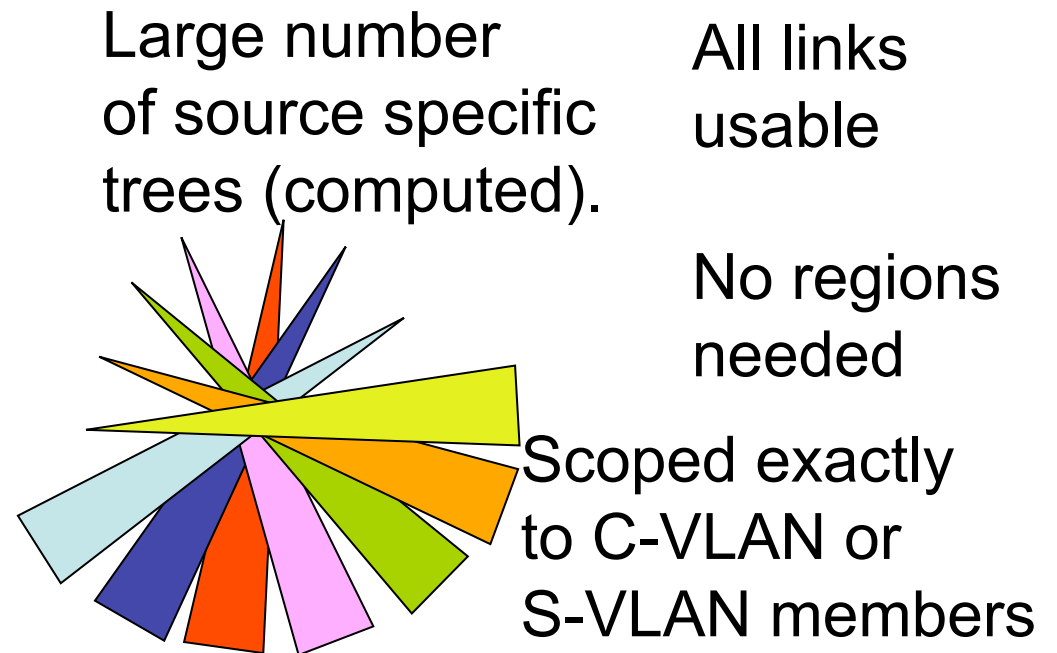
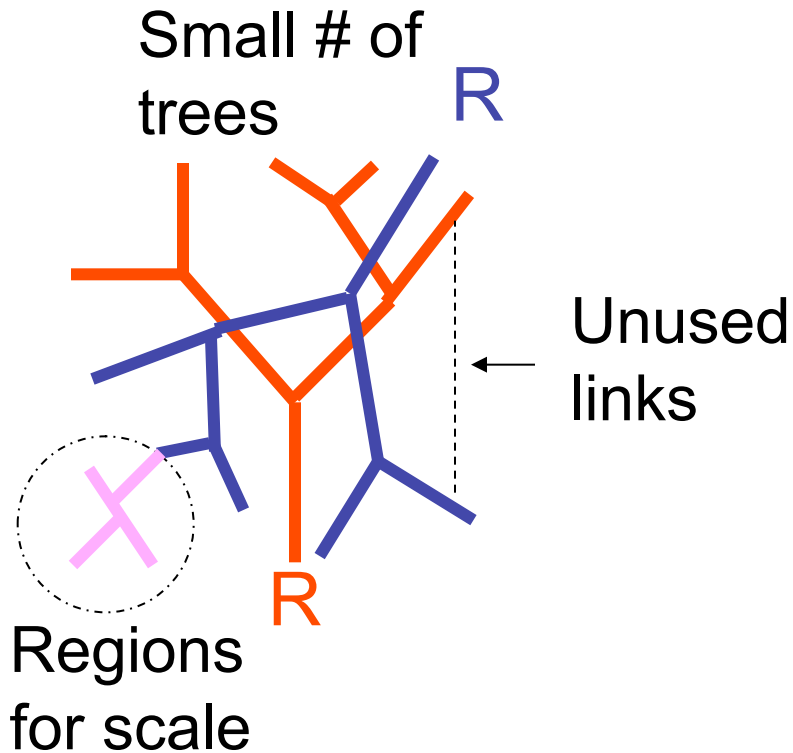


Multiple Shortest Path routing + Ethernet OA&M

Outline

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

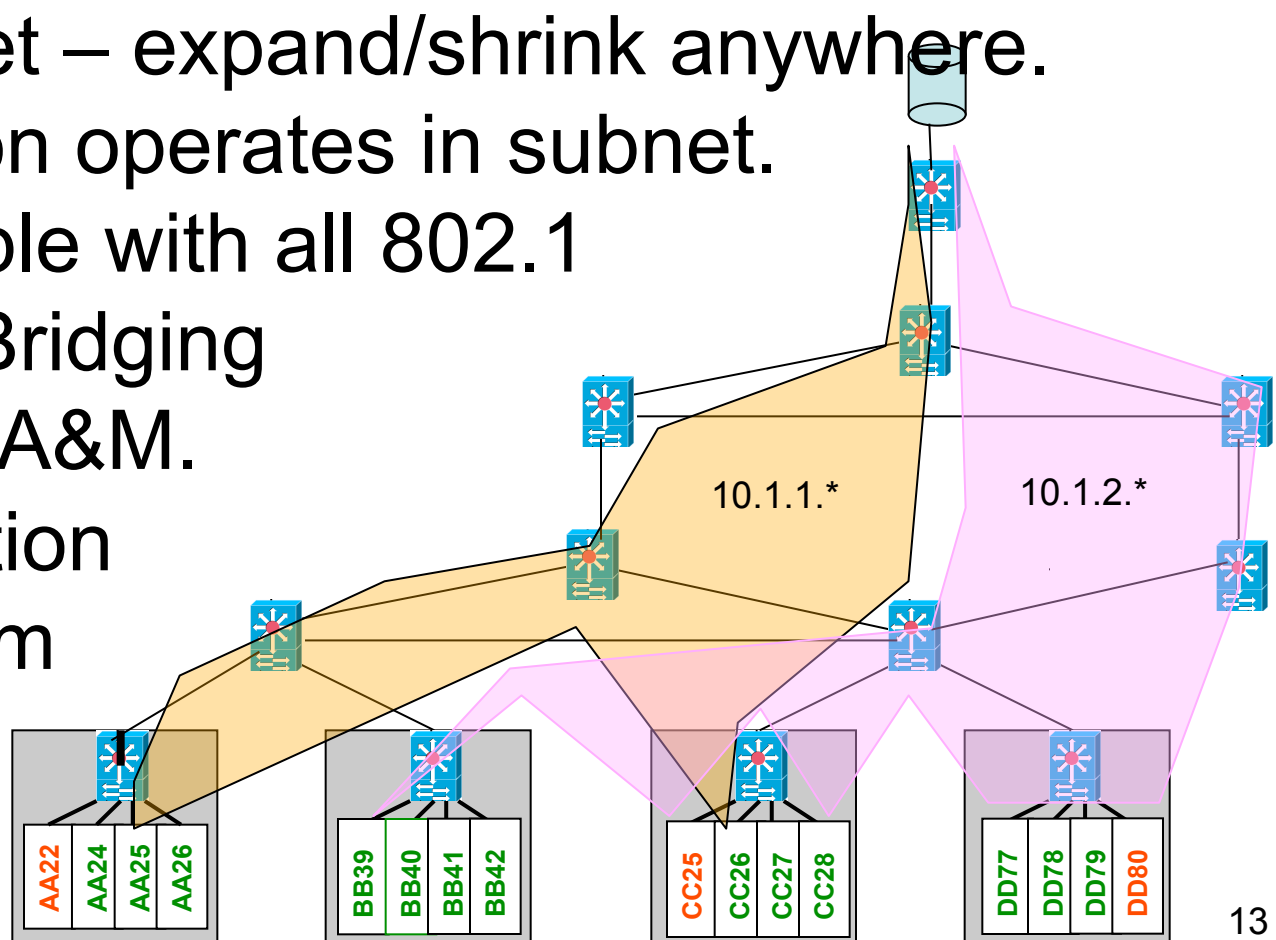
Application (M|R)STP replacement



- Many more nodes without regions
- Low effort to get good routing
- Fast convergence – link state v.s. distance vec
- Address isolation m-in-m.

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

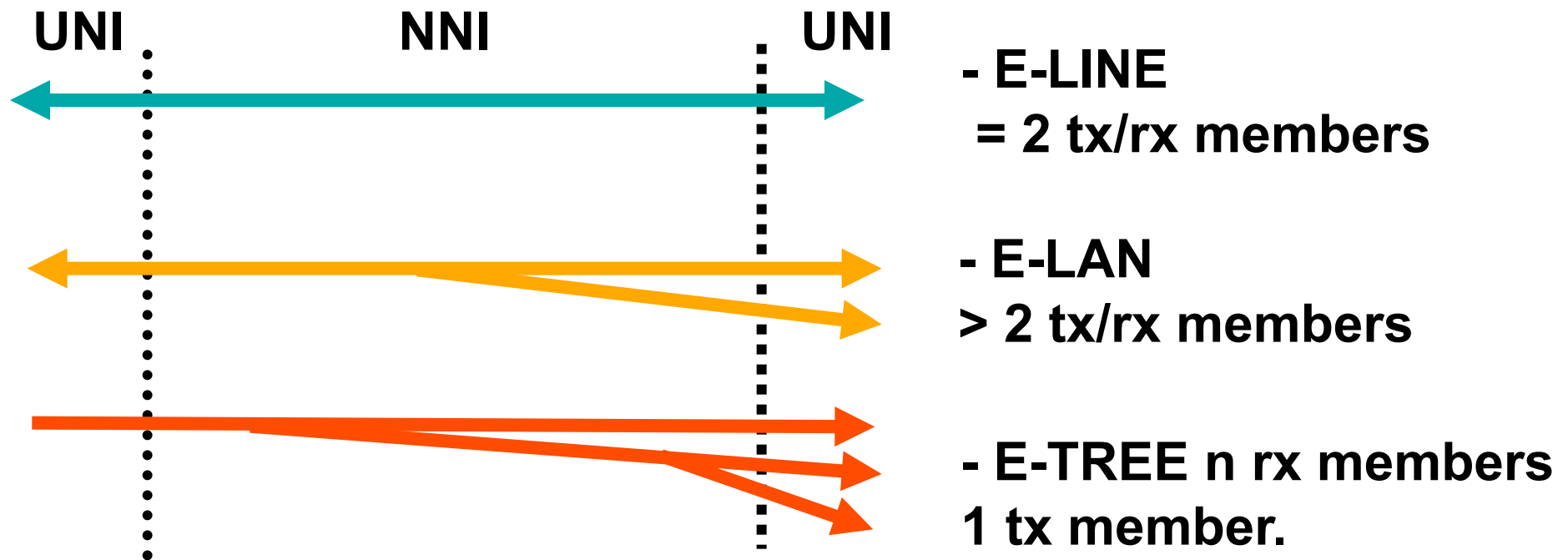


Application Data Center (cont'd)

- Totally compatible with Vmware server functions:
 - OA&M, motion, backup etc.
 - Apps that sit on Vmware 'just work'.
- Totally compatible with Microsoft load balancing (multicast over the L2)
- VRRP transparent.
- It just makes the L2 part of the DC larger and better utilized.
- Compatible with emerging Inter DC overlay work.

Application Metro/L2VPN

- Very light weight L2VPNs (2^{24} data path) of:
E-LAN, E-LINE, E-TREE flavors (a very cheap VPLS)
- Can do VPLS style head end replication
- Can do p2mp style transit replication (just one tx flag).
- Can support receive only membership (E-TREE)



Outline

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

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..

- **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.

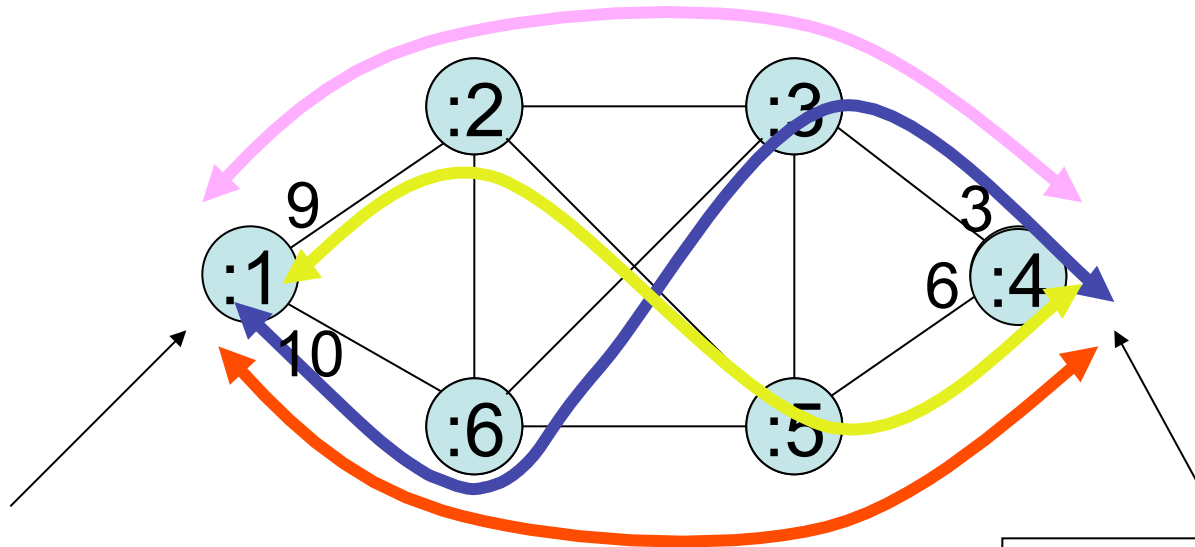
Data Path (M-in-M mode)

- 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.
- B-MAC 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 .. Yawn..

Else (mp2mp like but without signaling)

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

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

How does it work – transit multicast format (n/a for head replication)

42
43
44
45
46
47
48
49
50
51
52
53
54

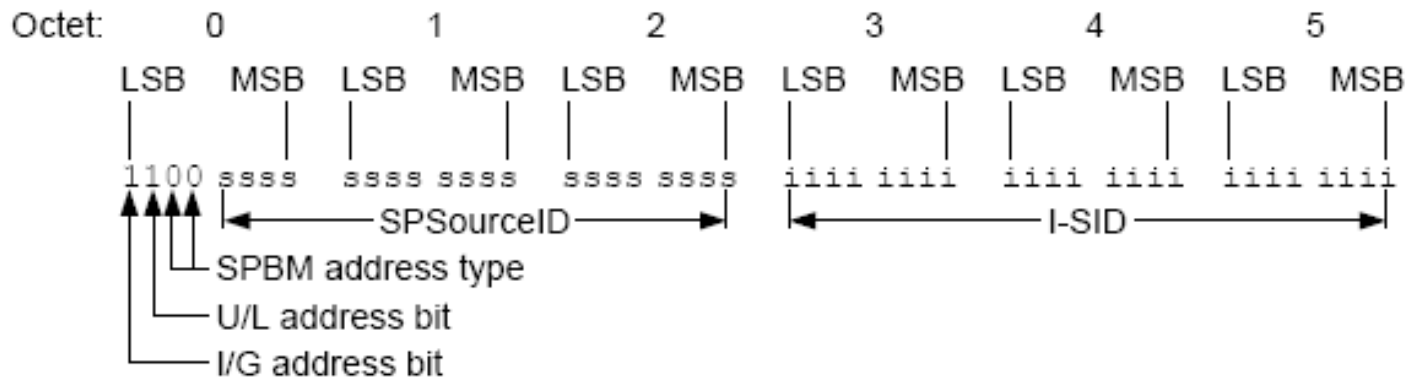


Figure 27-3—SPBM Group MAC Addresses used by this standard

Copyright © 2010 IEEE. All rights reserved.

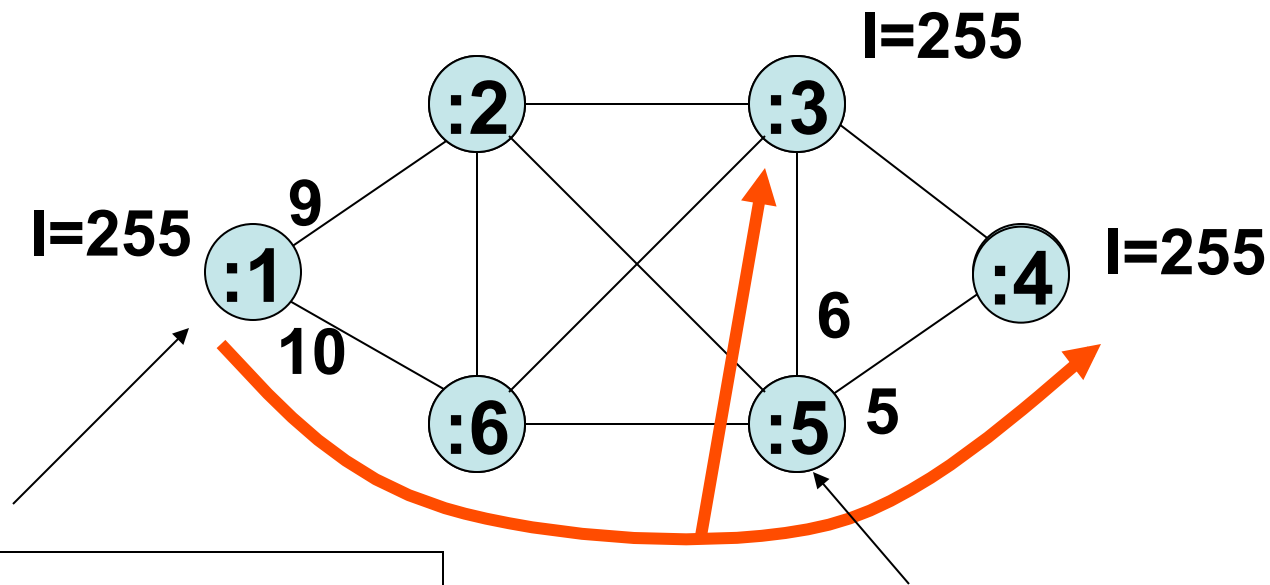
178

Example: { SOURCE: 0A-BC-DE / ISID: fe-dc-ba }

MMAC-DA: **A3-BC-DE-FE-DC-BA**

0011

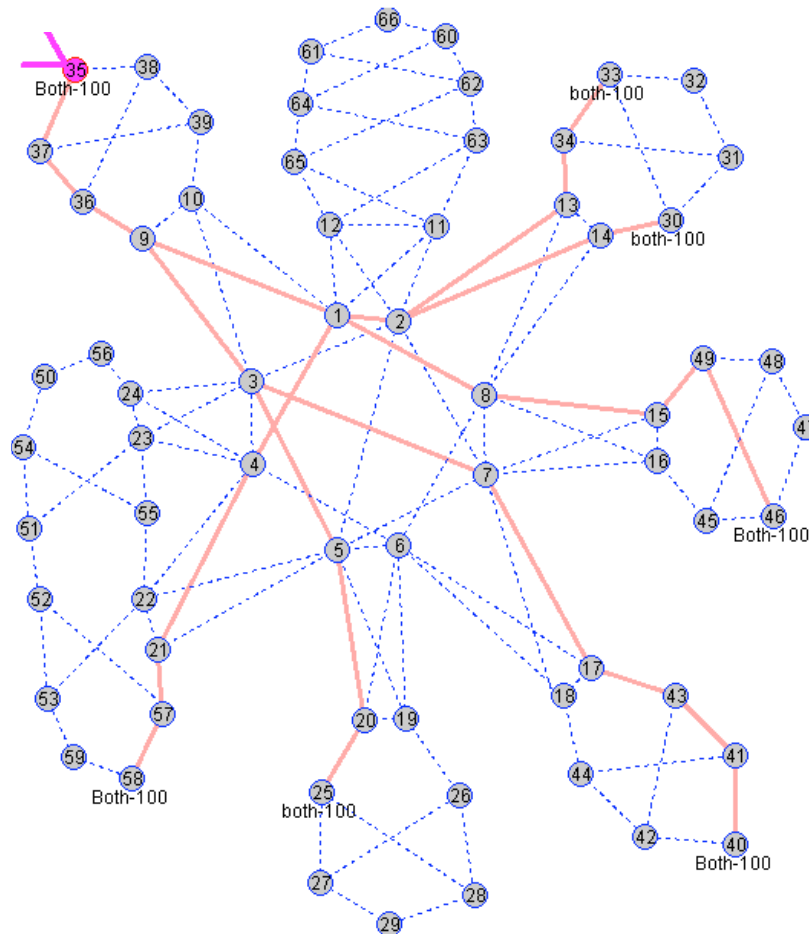
FDB visually: mcast m-in-m mode



MMAC	BVID	IF
{ :1/255 }	4	10

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

ANIMATION FOR E-LAN '100' WITH 7 MEMBERS



Highlighted is the routing from each member to all others.

Note the symmetry.

Unicast and multicast Follows exactly these Routes.

Multicast can be replicated at fork points or head end replicated to the uni-cast paths by configuration at edge.

The Control Plane (m-in-m mode)

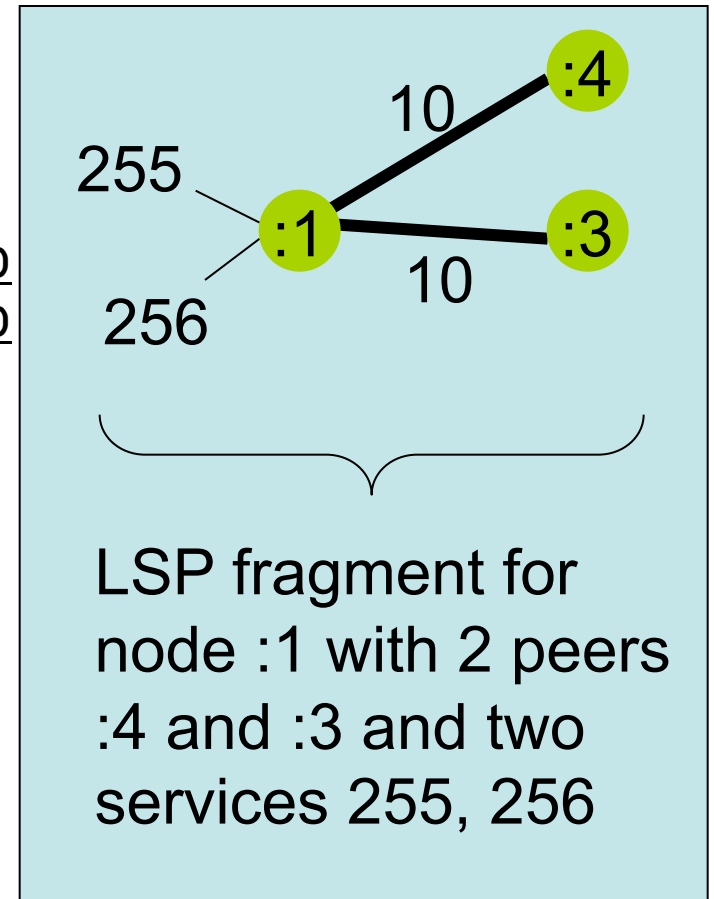
- Industry standard IS-IS Link State Protocol is basis for 802.1aq.
- Does not require any IP to operate.
- Does not preclude IPV4 or IPV6 being present in same IS-IS instance.
- SYSID carries B-MAC address
- Introduces no new PDU's to IS-IS.
- Hello TLVs augmented to pass Equal Cost Algorithm / Vid information and new NLPID.
- Update TLV's augmented to advertise SPB specific link costs.
- Update TLVs augmented to advertise ISID information.
- Update TLVs augmented to advertise nodal 'short form' name SPSOURCEID (transit mcast only).

802.1aq ISIS LSP extensions at a glance

```

LSPID                               Seq Num       Checksum ....
-----
4455.6677.0001.00-00  0x0000007a    0xc01f  ....
SOURCE          4455.6677.0001.00 (1)
HOST NAME       Instance_1
NLPID           SPB (0xC1) (2)
AREA ADDR      22.3344 (3)
NBR ID         4455.6677.0004.00  COST: 10
NBR ID         4455.6677.0003.00  COST: 10
SPSOURCEID     07-00-01
SPB  ECT-ALGORITHM 1  ECT-VID 101 (4)
SPB  ECT-ALGORITHM 0  ECT-VID 100
SPB  BMAC  44-55-66-77-00-01
          ECT-VID 100
          SPB  ISID  255T&R (5)
SPB  BMAC  44-55-66-77-00-01
          ECT-VID 101
          SPB  ISID  256T&R

```



Loop Suppression & Avoidance

Suppression

- done on the data path using an SA check.
- prevents 99.99% loops if FDB's create one.
- no impact on convergence rates.
- exploits symmetric/congruence properties of routing.
- uses reverse learning options of most h/w to discard.

Avoidance

- done by the control path
- ensures no loops are ever configured in FDBs.
- hellos augmented with topology 'digests'
- mismatched digests => some forwarding entries unsafe.
- blocks only 'unsafe' entries.
- works for ALL forwarding modes current and planned.

802.1aq OA&M (inherited *by design*)

Service/Network Layer – 802.1ag Connectivity Fault mgmt

- Hierarchy (honors maintenance levels/abstraction)
 - Continuity Check
 - L2 traceroute
 - L2 ping

Link Layer – 802.3ah

- Link Monitoring (logical/physical)
- Remote Failure Indication
- Remote Loopback

Service Layer - Y.1731

- Multicast Loopback – depends on congruency/symmetry
- Performance Measurements (Loss/Delay etc.)
- One way/two way delays – symmetry important

Recovery

- ISIS augmented with multicast LSP flood to all 802.1aq nodes.
- Every 802.1aq node joins default service ISID 0xfffff.
- This E-LAN is just for control plane.
- LSPs can be advertised over this E-LAN
- Very fast distribution protocol (h/w multicast).
- On failure each end of link advertises over this 'default' E-LAN (in addition to normal updates).
- Reaches all 802.1aq participants at h/w multicast speed with no CPU involvement transit.
- Conceptually like having a shared LAN joining all nodes with a physical port but no DR election etc. is done, only used as unreliable very fast distribution mechanism backed up by normal IS-IS hop by hop LSPs.

Outline

- Challenges
- What is 802.1aq/SPB
- Applications
- How does it work
- **Example (included in this deck - enjoy)**
- Q&A

Outline

- Challenges
- What is 802.1aq/SPB
- Applications
- How does it work
- Example (backup slides)
- Q&A (avail anytime)



References

“**IEEE 802.1aq**” : www.wikipedia.org:
http://en.wikipedia.org/wiki/IEEE_802.1aq

“**IEEE 802.1aq**” www.ieee802.org/1/802-1aq-d2-6.pdf

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

“**Provider Link State Bridging**” :
IEEE Communications Magazine V46/N9– Sept 2008
[http://locuhome.com/wp-content/uploads/2009/02/
ieeecommunicationsmagazinevol46no9sep2008-carrierscaleethernet.pdf](http://locuhome.com/wp-content/uploads/2009/02/ieeecommunicationsmagazinevol46no9sep2008-carrierscaleethernet.pdf)

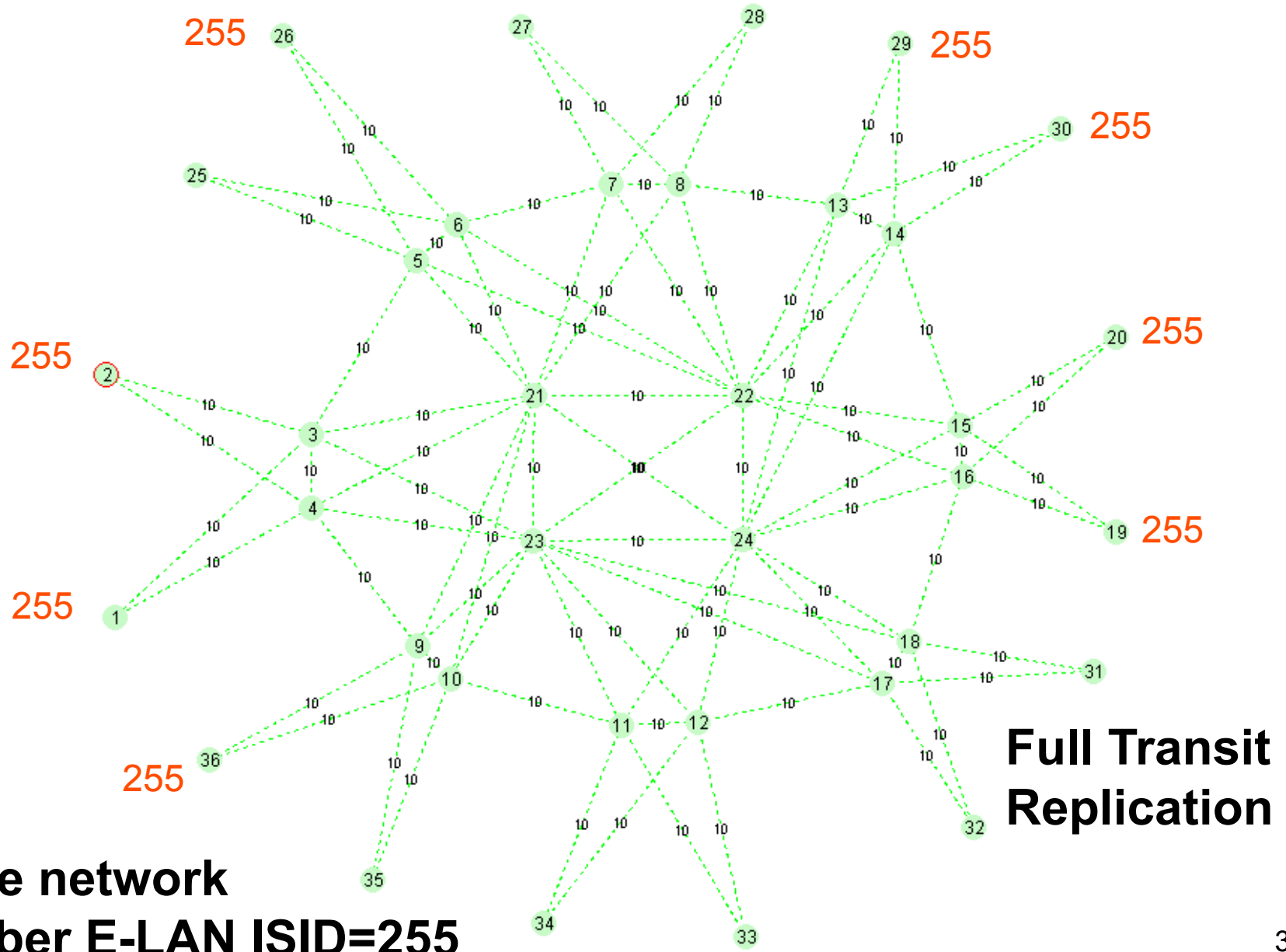
See also the worked **example** – in backup slides in this deck

Thank-You



EXAMPLE

EXAMPLE NETWORK :



36 node network
8 member E-LAN ISID=255

EXAMPLE – ISIS PEERS AT NODE :3

```
<ottawa-9300-3>d spb
The current global spb information is :
Device HMAC is 44-55-66-77-00-03
Spsid is 07-00-03
Ect vlan amount is 2
Ect vlan sequence number [1] is: vlan 100 !
Ect vlan sequence number [2] is: vlan 101 !
<ottawa-9300-3>
```

```
<ottawa-9300-3>d isis peer
```

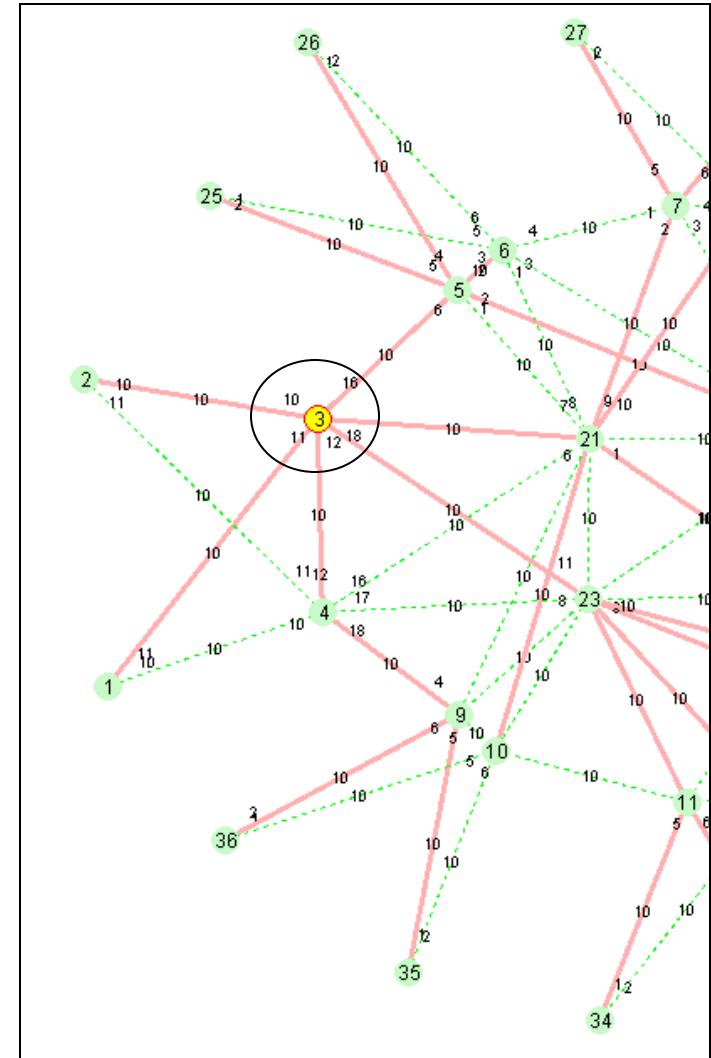
Peer information for ISIS(1)

System Id	Interface	Circuit Id	State	HoldTime	Type
4455.6677.0001	Vlanif211	0000000002	Up	26s	L1
4455.6677.0004	Vlanif212	0000000003	Up	23s	L1
4455.6677.0005	Vlanif216	004	Up	27s	L1
4455.6677.0015	Vlanif217	005	Up	27s	L1
4455.6677.0017	Vlanif218	006	Up	25s	L1

```
Total Peer(s) : 5
```

```
<ottawa-9300-3>
```

Logging on to node :3
We can see the basic SPB info
and the ISIS peers....



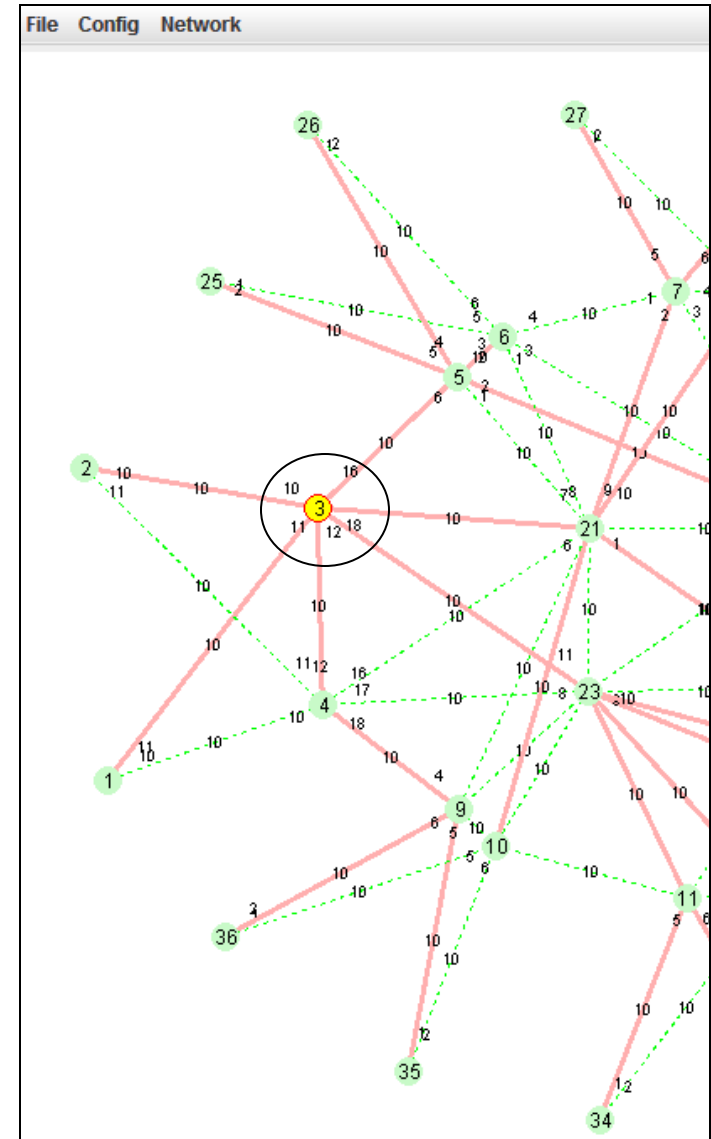
EXAMPLE – LSDB at node :3

Database information for ISIS(1)

 Level-1 Link State Database

LSPID	Seq Num	Checksum	Holdtime	Length	ATT/P/OL
4455.6677.0001.00-00	0x0000fd2	0x1cea	1044	236	0/0/0
4455.6677.0003.00-00*	0x00001448	0x3d27	683	323	0/0/0
4455.6677.0004.00-00	0x0000ff8	0xd1d9	1090	323	0/0/0
4455.6677.0005.00-00	0x0000b3b	0x9ba7	586	317	0/0/0
4455.6677.0006.00-00	0x0000b3b	0xbc31	819	293	0/0/0
4455.6677.0007.00-00	0x0000b3b	0xce10	1075	293	0/0/0
4455.6677.0008.00-00	0x0000b3e	0xebe0	288	293	0/0/0
4455.6677.0009.00-00	0x0000b3b	0x8b77	355	317	0/0/0
4455.6677.000a.00-00	0x0000b3b	0x57a	840	294	0/0/0
4455.6677.000b.00-00	0x0000b3a	0x1665	608	294	0/0/0
4455.6677.000c.00-00	0x0000b3a	0x6903	764	294	0/0/0
4455.6677.000d.00-00	0x0000b3a	0x89fd	431	294	0/0/0
4455.6677.000e.00-00	0x0000b39	0x3445	611	294	0/0/0
4455.6677.000f.00-00	0x0000b3a	0xdabe	616	294	0/0/0
4455.6677.0010.00-00	0x0000b3b	0x810e	452	294	0/0/0
4455.6677.0011.00-00	0x0000b3a	0x1b46	645	294	0/0/0
4455.6677.0012.00-00	0x0000b39	0x1b3e	447	294	0/0/0
4455.6677.0013.00-00	0x0000b3a	0x943a	419	176	0/0/0
4455.6677.0014.00-00	0x0000b3b	0xdff2	693	176	0/0/0
4455.6677.0015.00-00	0x0000b41	0xdade	1141	508	0/0/0
4455.6677.0016.00-00	0x0000b3e	0xa832	1011	464	0/0/0
4455.6677.0017.00-00	0x0000b40	0x1563	640	508	0/0/0
4455.6677.0018.00-00	0x0000b3a	0xadee	417	464	0/0/0
4455.6677.0019.00-00	0x0000b3a	0xcff	291	158	0/0/0
4455.6677.001a.00-00	0x0000b3b	0xb131	794	176	0/0/0
4455.6677.001b.00-00	0x0000b3b	0x7062	822	176	0/0/0
4455.6677.001c.00-00	0x0000b3b	0x5876	463	176	0/0/0
4455.6677.001d.00-00	0x0000b3b	0xa610	460	176	0/0/0
4455.6677.001e.00-00	0x0000b3a	0xf5c6	627	176	0/0/0
4455.6677.001f.00-00	0x0000b3b	0x8a19	825	176	0/0/0
4455.6677.0020.00-00	0x0000b3a	0x960a	584	176	0/0/0
4455.6677.0021.00-00	0x0000b3b	0x5b58	1033	176	0/0/0
4455.6677.0022.00-00	0x0000b3a	0x8927	693	176	0/0/0
4455.6677.0023.00-00	0x0000b3b	0x9352	664	158	0/0/0
4455.6677.0024.00-00	0x0000b39	0xc8ec	736	176	0/0/0

*(In TLV)-Leaking Route, *(By LSPID)-Self LSP, +-Self LSP(Extended),
 ATT-Attached, P-Partition, OL-Overload



EXAMPLE LSP VERBOSE OF NODE :1 at NODE :3

```
<ottawa-9300-3>d isis lsdb 4455.6677.0001.00-00 verbose
```

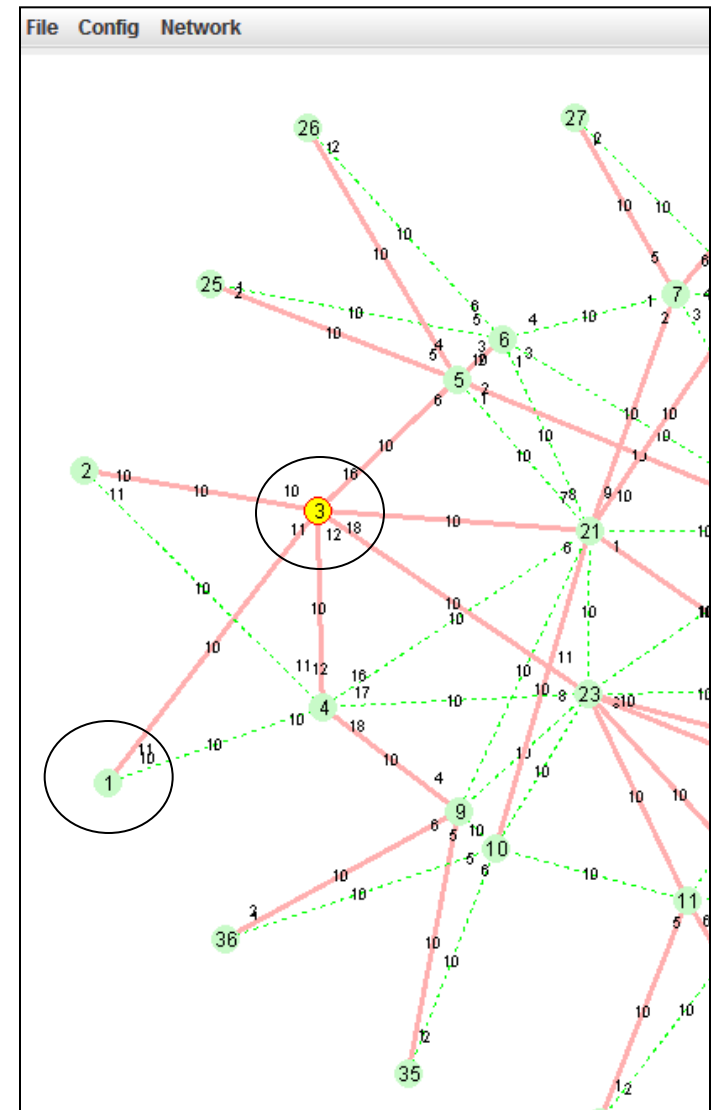
```
Database information for ISIS(1)
```

```
-----
```

Level-1 Link State Database

LSPID	Seq Num	Checksum	Holdtime
4455.6677.0001.00-00	0x00000fd3	0x1aeb	1194
SOURCE	4455.6677.0001.00		
NLPID	SPB(0xC1)		
AREA ADDR	22.3344		
+NBR ID	4455.6677.0003.00	COST: 10	
+NBR ID	4455.6677.0004.00	COST: 10	
SPB ECT-ALGORITHM 0	ECT-VID 100		
SPB ECT-ALGORITHM 1	ECT-VID 101		
SPB ECT-ALGORITHM 2	ECT-VID 0		
.....			
SPB ECT-ALGORITHM 15	ECT-VID 0		
SPSID	07-00-01		
SPB BMAC	44-55-66-77-00-01		
ECT-VID	100		
SPB ISID	255T&R		

```
<ottawa-9300-3>
```



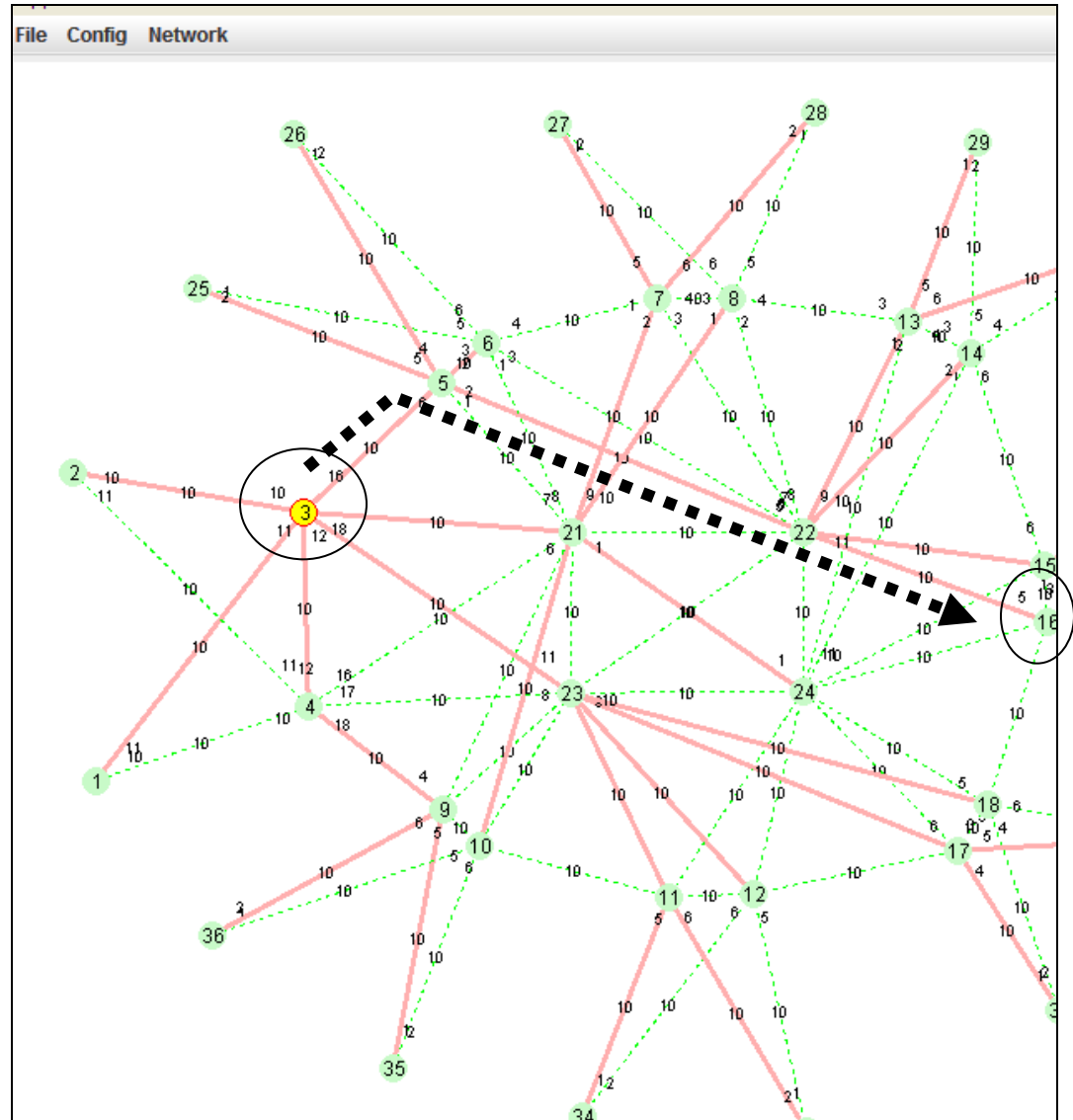
EXAMPLE – NODE :3 ROUTE TO :10 (first equal cost path)

<ottawa-9300-3>d spb umac

BMAC	BVLAN	IF NAME
4455-6677-0001	100	GE2/0/11
4455-6677-0001	101	GE2/0/11
4455-6677-0004	100	GE2/0/12
4455-6677-0004	101	GE2/0/12
4455-6677-0005	100	GE2/0/16
4455-6677-0005	101	GE2/0/16
4455-6677-0006	100	GE2/0/16
4455-6677-0006	101	GE2/0/17
4455-6677-0007	100	GE2/0/17
4455-6677-0007	101	GE2/0/17
4455-6677-0008	100	GE2/0/17
4455-6677-0008	101	GE2/0/17
4455-6677-0009	100	GE2/0/12
4455-6677-0009	101	GE2/0/18
4455-6677-000a	100	GE2/0/17
4455-6677-000a	101	GE2/0/18
4455-6677-000b	100	GE2/0/18
4455-6677-000b	101	GE2/0/18
4455-6677-000c	100	GE2/0/18
4455-6677-000c	101	GE2/0/18
4455-6677-000d	100	GE2/0/16
4455-6677-000d	101	GE2/0/18
4455-6677-000e	100	GE2/0/16
4455-6677-000e	101	GE2/0/18
4455-6677-000f	100	GE2/0/16
4455-6677-000f	101	GE2/0/18
4455-6677-0010	100	GE2/0/16
4455-6677-0010	101	GE2/0/18
...		
4455-6677-0024	100	GE2/0/12
4455-6677-0024	101	GE2/0/18

Total unicast fib entries is 68

<ottawa-9300-3>



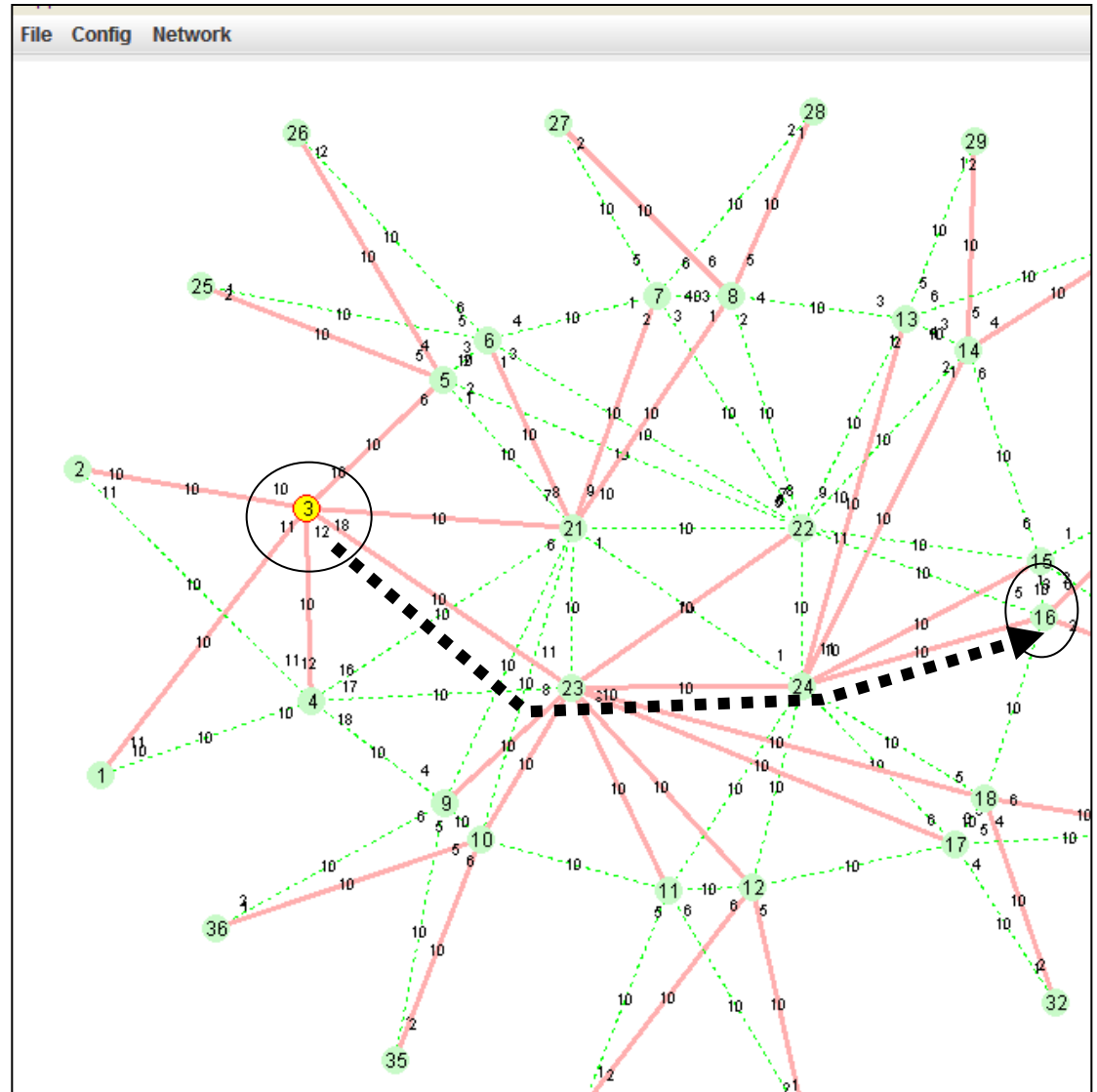
EXAMPLE – NODE :3 ROUTE TO :10 (second equal cost path)

```
<ottawa-9300-3>d spb umac
```

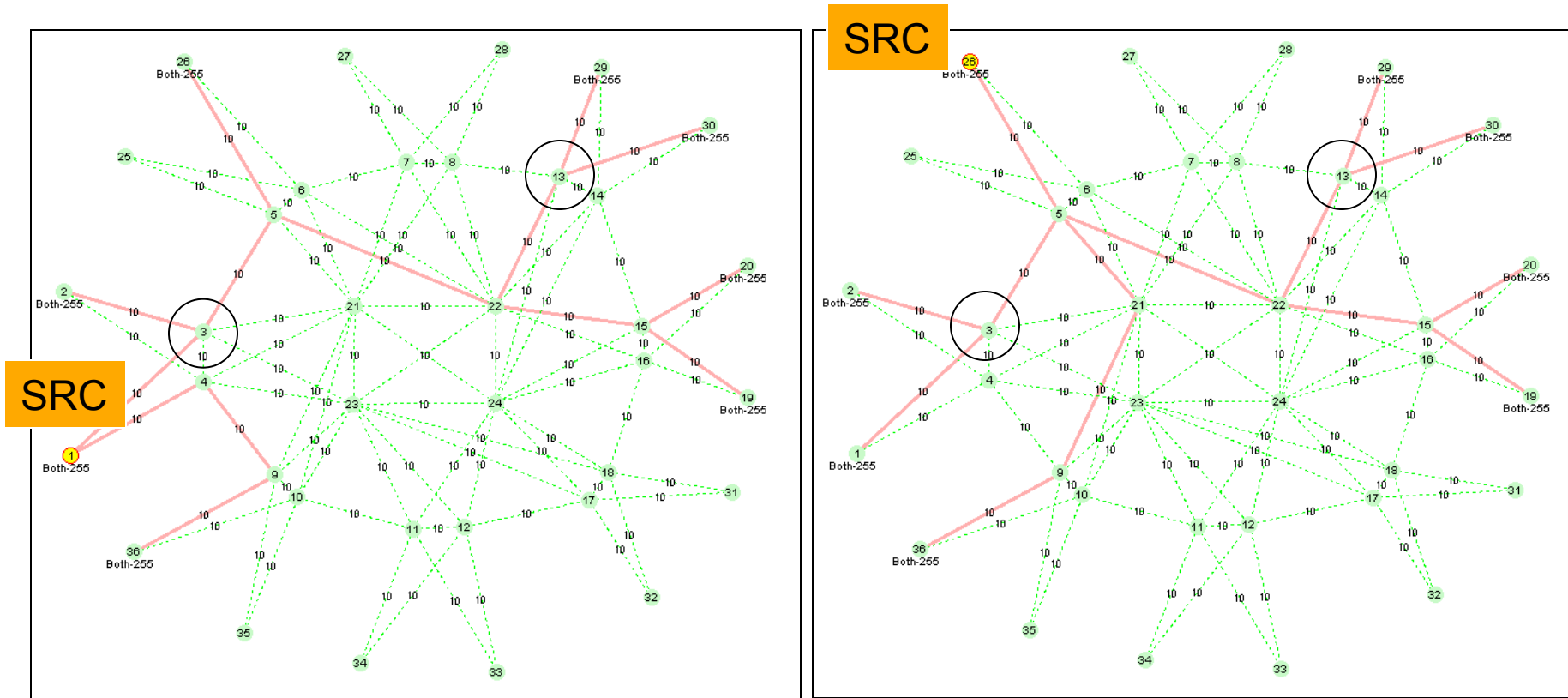
BMAC	BVLAN	IF NAME
4455-6677-0001	100	GE2/0/11
4455-6677-0001	101	GE2/0/11
4455-6677-0004	100	GE2/0/12
4455-6677-0004	101	GE2/0/12
4455-6677-0005	100	GE2/0/16
4455-6677-0005	101	GE2/0/16
4455-6677-0006	100	GE2/0/16
4455-6677-0006	101	GE2/0/17
4455-6677-0007	100	GE2/0/17
4455-6677-0007	101	GE2/0/17
4455-6677-0008	100	GE2/0/17
4455-6677-0008	101	GE2/0/17
4455-6677-0009	100	GE2/0/12
4455-6677-0009	101	GE2/0/18
4455-6677-000a	100	GE2/0/17
4455-6677-000a	101	GE2/0/18
4455-6677-000b	100	GE2/0/18
4455-6677-000b	101	GE2/0/18
4455-6677-000c	100	GE2/0/18
4455-6677-000c	101	GE2/0/18
4455-6677-000d	100	GE2/0/16
4455-6677-000d	101	GE2/0/18
4455-6677-000e	100	GE2/0/16
4455-6677-000e	101	GE2/0/18
4455-6677-000f	100	GE2/0/16
4455-6677-000f	101	GE2/0/18
4455-6677-0010	100	GE2/0/16
4455-6677-0010	101	GE2/0/18
...		
4455-6677-0024	100	GE2/0/12
4455-6677-0024	101	GE2/0/18

Total unicast fib entries is 68

```
<ottawa-9300-3>
```

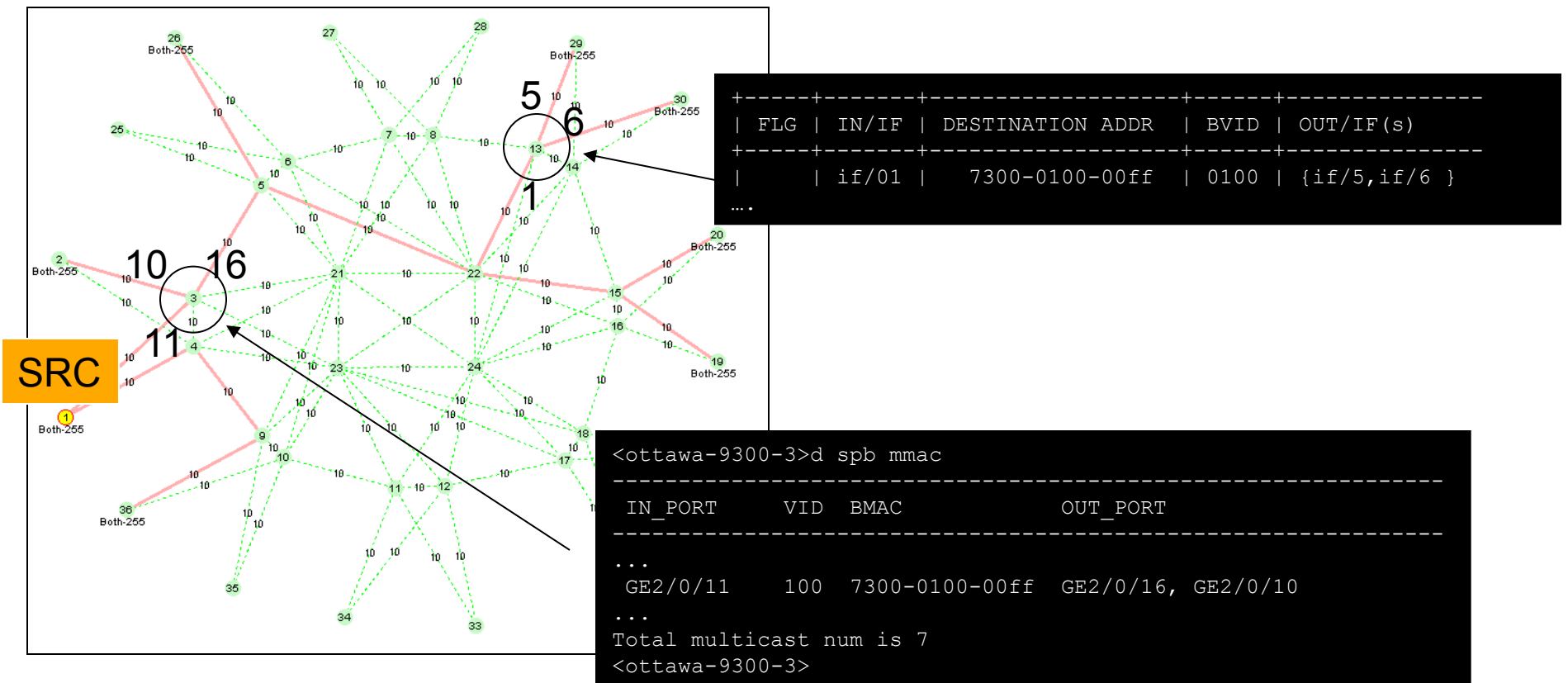


EXAMPLE: E-LAN MCAST ROUTES FROM :1 (left) and :26 (right)



Here are the multicast routes from node 1 for service 255 and also from node 26 for service 255. Note the symmetry in the route between the two multicast trees. The unicast route between :1 and :26 is also along that same path for the chosen B-VID. Since we've asked for transit replication for all members of the E-LAN we install MCAST ...

EXAMPLE: E-LAN MCAST ROUTES FROM :1 (left) and :26 (right)



MULTICAST ADDRESS IS: [SOURCE = 07-00-01 | ISID=00-00-ff]

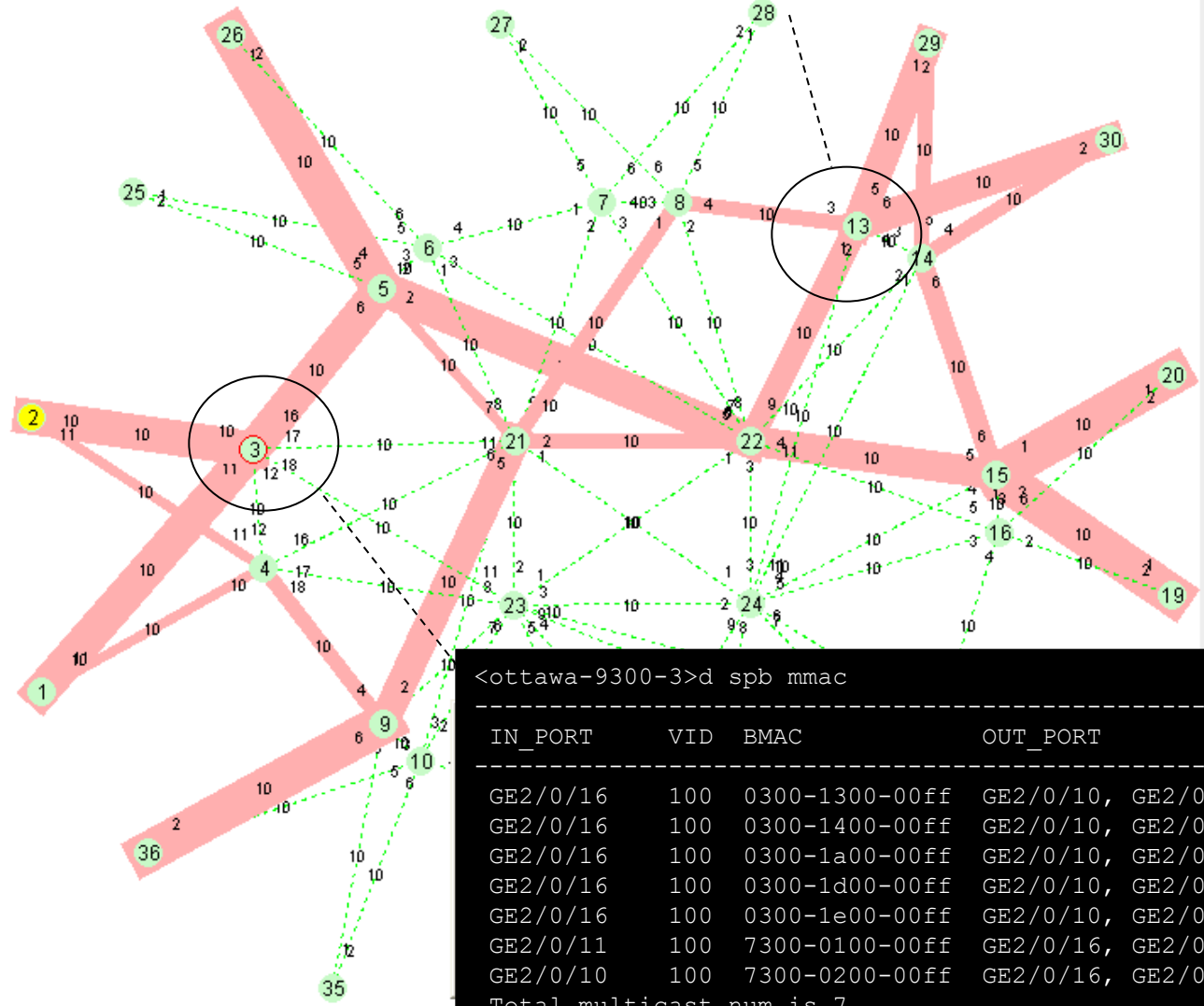
We only get this state if we configure transmit membership in the E-LAN.
 Transmit still possible without multicast state but uses serial replication at head end.
 Operator chooses trade-off between state/bandwidth usage.

```

+-----+-----+-----+-----+-----+
| FLG | IN/IF | DESTINATION ADDR | BVID | OUT/IF(s) |
+-----+-----+-----+-----+-----+
|     | if/01 | 7300-0100-00ff  | 0100 | {if/5,if/6 } |
|     | if/01 | 7300-0200-00ff  | 0100 | {if/5,if/6 } |
|     | if/01 | 0300-1a00-00ff  | 0100 | {if/5,if/6 } |
|     | if/05 | 0300-1d00-00ff  | 0100 | {if/1,if/3,if/6 } |
|     | if/06 | 0300-1e00-00ff  | 0100 | {if/1,if/3,if/5 } |
|     | if/03 | 0300-2400-00ff  | 0100 | {if/5,if/6 } |

```

Here are all mFIBs on nodes :3 and :13 related to this E-LAN.



Information

Node

Instance. ID: 3

Node ID: 4455.6677.0003

Area ID: 22.3344

AreaName: AreaTestNet

Level 1 Level 2

VLAN ID: 100

I-SID: New

Inst Real Virt

Link

Link Name: link0-0

Metric: 10

Bandwidth: 1000

AreaName: AreaTestNet

Connection

Address: localhost

Port: 7001

con1

```

<ottawa-9300-3>d spb mmac
-----
IN_PORT      VID  BMAC                                OUT_PORT
-----
GE2/0/16    100  0300-1300-00ff  GE2/0/10, GE2/0/11
GE2/0/16    100  0300-1400-00ff  GE2/0/10, GE2/0/11
GE2/0/16    100  0300-1a00-00ff  GE2/0/10, GE2/0/11
GE2/0/16    100  0300-1d00-00ff  GE2/0/10, GE2/0/11
GE2/0/16    100  0300-1e00-00ff  GE2/0/10, GE2/0/11
GE2/0/11    100  7300-0100-00ff  GE2/0/16, GE2/0/10
GE2/0/10    100  7300-0200-00ff  GE2/0/16, GE2/0/11
Total multicast num is 7
<ottawa-9300-3>

```