802.1Qbp
Shared Tree (*.G) Algorithms
(for head end MCAST ECMP)

Peter Ashwood-Smith
peter.ashwoodsmith@huawei.com

bp-ashwood-shared-trees-1011-v3

Motivation

- 802.1Qbp is introducing new ECMP behavior in an 802.1aq network.
- There is a requirement to do ECMP (head end) over multicast trees.
- So far we have only discussed the (S,G) multicast trees (existing .1aq style and Ben's alternatives).
- I'd like to discuss some simple (*,G) options since state reduction without loss of functionality is possible especially in DC networks.

N.B (S,G) is source/group specific tree, i.e. <SpSourceID>||<SID> in the DA
(*.G) is shared by all sources but one group i.e. <Constant>||<ISID> in the DA

Considerations

- What we really want is a minimum spanning tree that covers just a subset of the nodes (those in the ISID).
- This is referred to as a Steiner Tree.
- A Steiner Tree computation is **NP-Complete**.
- "**N**on **P**olynomial" means its **>> O(N^c)** for any constant **c**.
- "**Complete**" is a way of saying we won't likely solve it here ..
- Basically its one of those problems that you have to enumerate all O(n!) solutions and pick the best.

Solutions

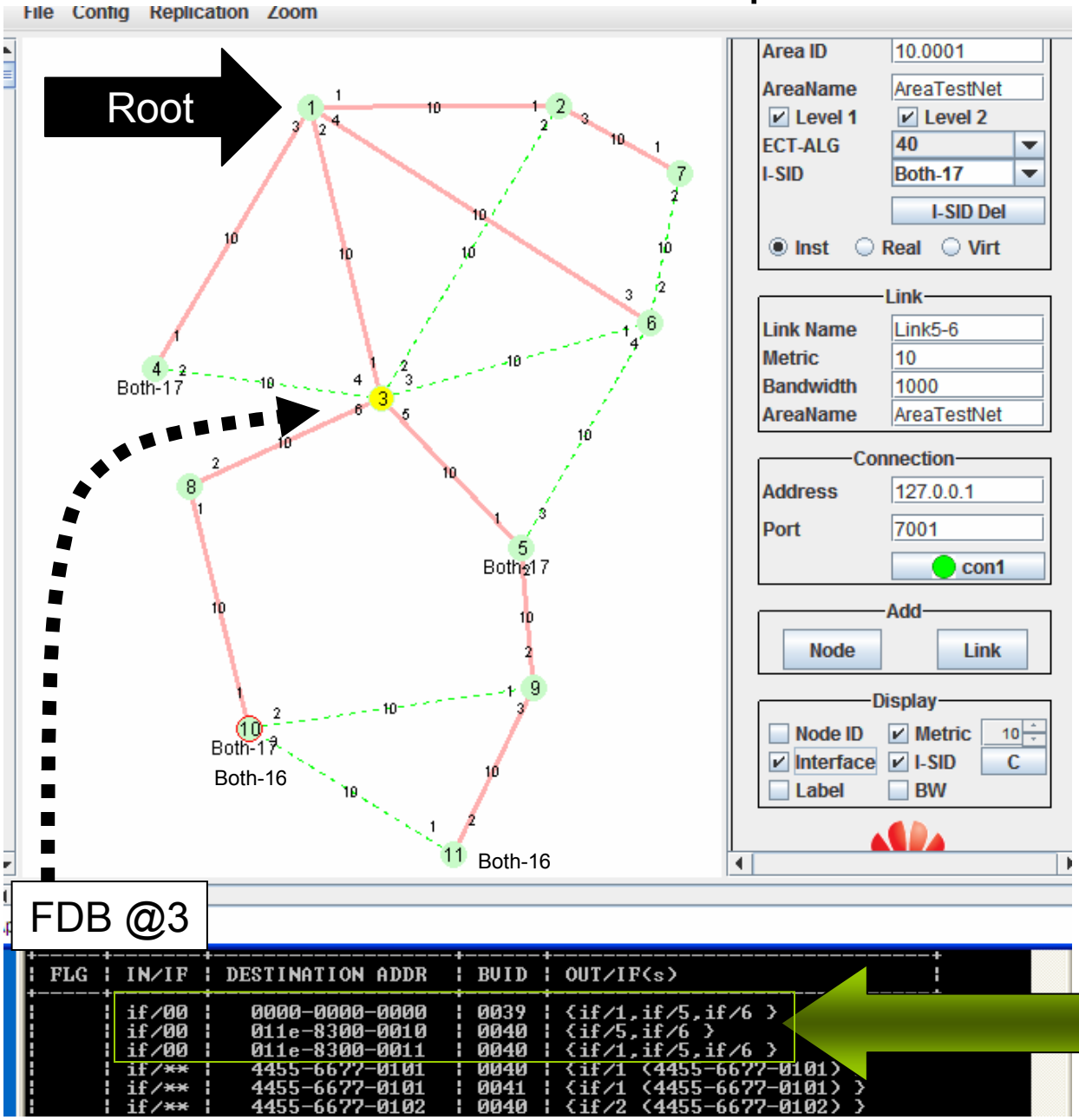
There are a few less optimal (*,G) solutions:

1. Pick some node as a root and use SPF from ‘it’ as the tree.
   - This is O(n*logN) but sends traffic everywhere!!!
   - So .. modify above by pruning per ISID (SPF is template).
   - This is O(N*logN + I*logI)
   - Still non shortest path routing but state is minimal
2. Other solutions aimed at reducing non shortest path routing issues but increase CPU.. these are FFS. (e.g. enforce root = member of ISID)

One possible proposal

- The 802.1aq CIST algorithm (which is just the STP algorithm done as a computation), can be reused for per ISID (*,G) trees in .1Qbp
- The multicast address format can be the existing PBB format i.e: ***00-1e-83-xx-xx-xx*** (where ***xx-xx..*** Is the ISID)
- 16 different shared trees can be computed by finding the lowest BridgeIdentifier under the 16 802.1aq ECT masks i.e. 0x00.., 0xff.., 0x11.., 0x22.. ... 0xee...
- These shared trees produce almost symmetric congruent results to the .1aq (S,G) trees in fat tree networks.
- Root selection automatic based on algorithm, auto recovery to new root etc. No explicit encoding of root in DA required.
- Can use F-TAG with TTL, or can rely on digest for loop prevention, or both....
- Can use same B-VID as unicast (no SVL), or different (with SVL) or even no B-VID.

Example #1



A (*,G) is computed using the Lowest Bridge Identifier (node 1) CIST algorithm.

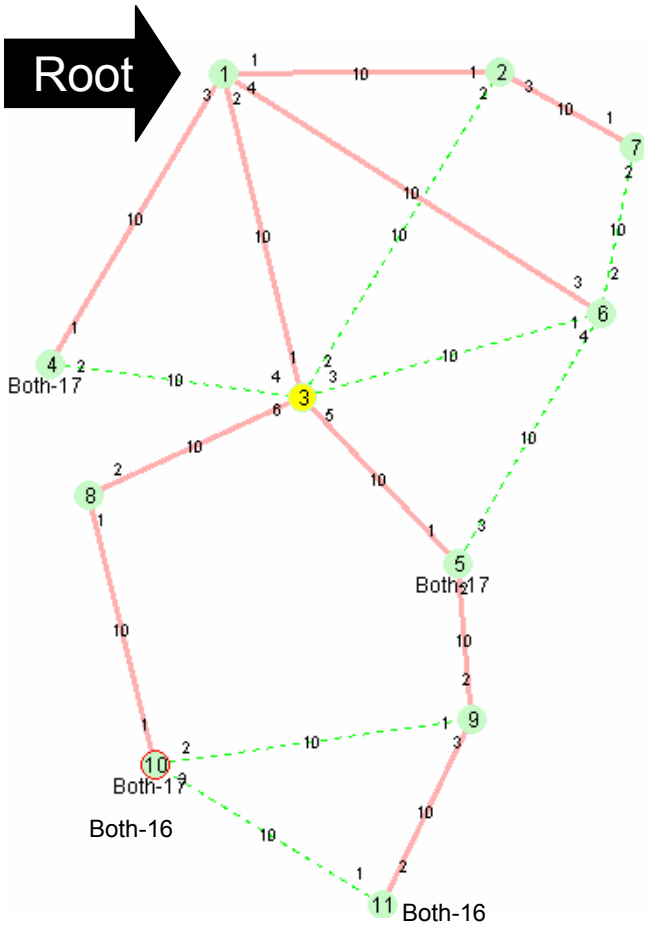
The full tree is shown in pink.

Two ISIDs are pruned against this tree for Multicast, sub trees below: ISID 17 and ISID 16.

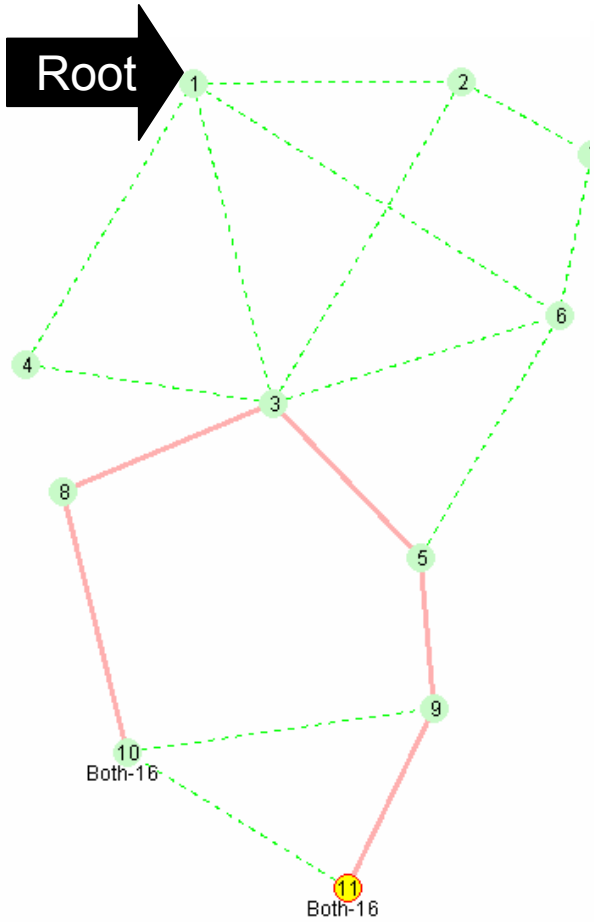
We show the Mcast state at node 3 for Each ISID.

-CIST
-Pruned for ISID 16
-Pruned for ISID 17

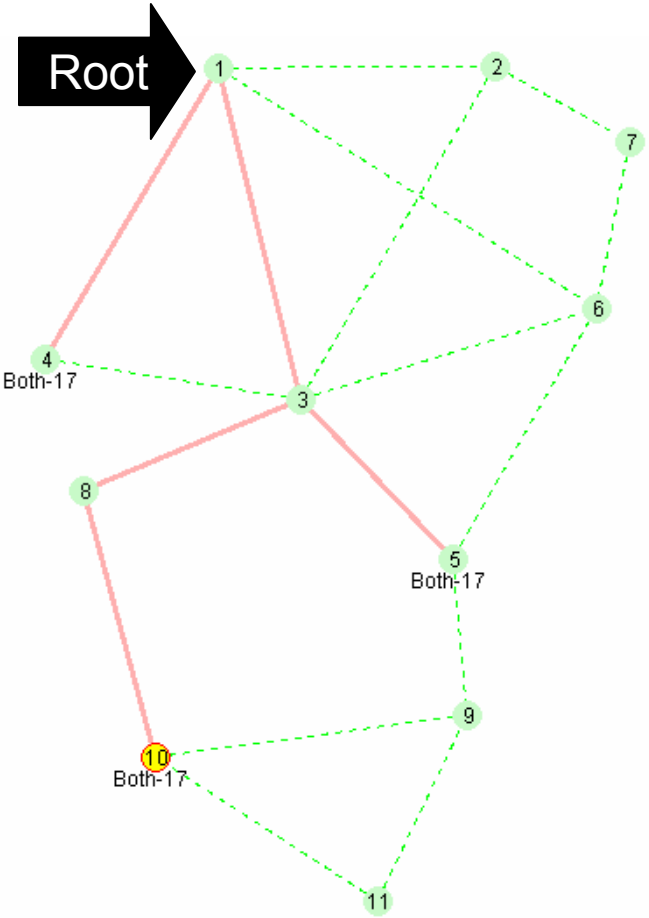
Example #1 – pruning



FULL MASK0x00
(ROOT=1) TREE

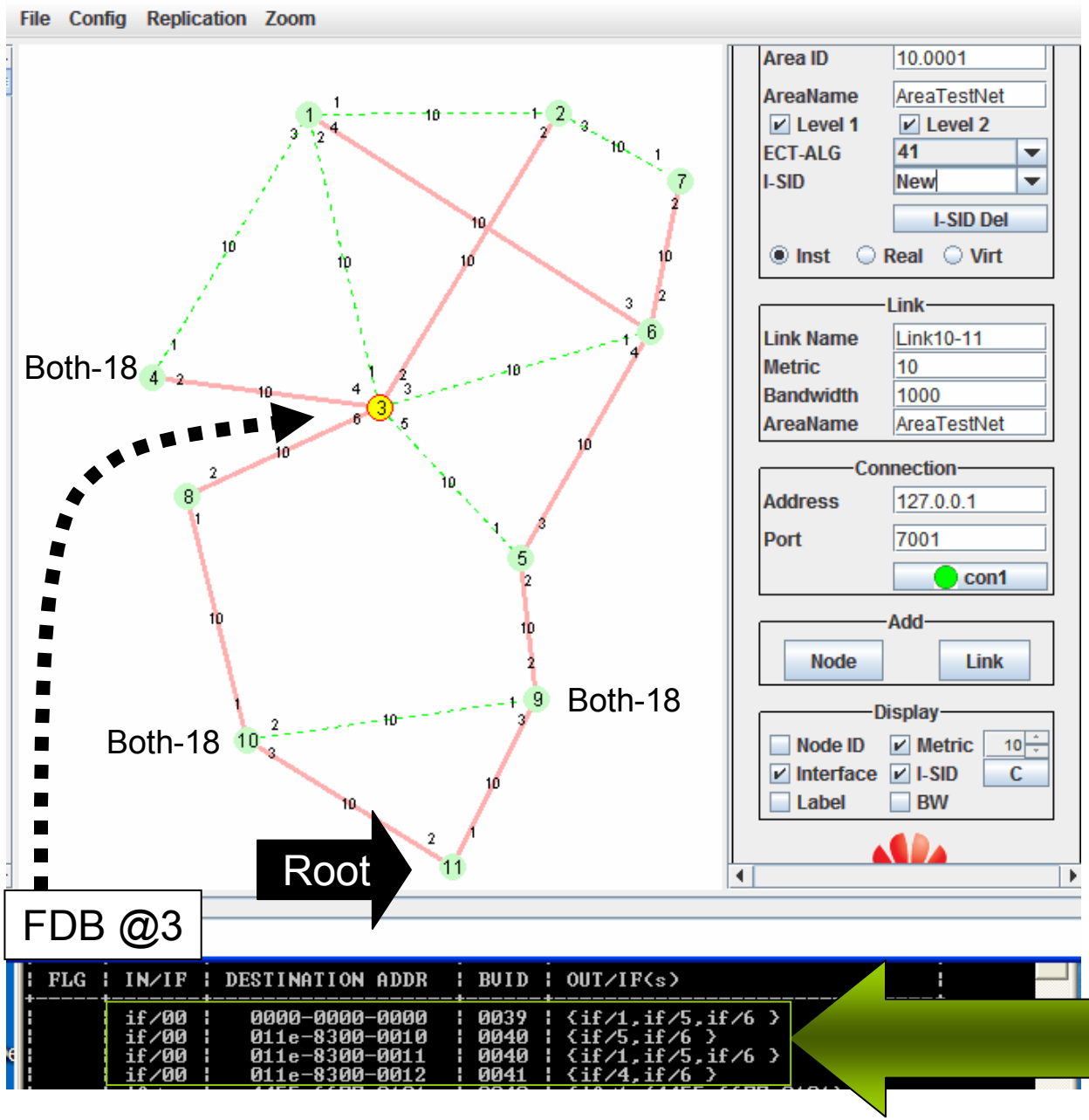


ISID 16
PRUNED



ISID 17
PRUNED

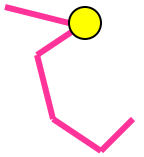
Example#2



A (*,G) is computed using the **highest** Bridge Identifier (node 11) i.e. CIST algorithm XOR 0xff.

The full tree is shown in pink.

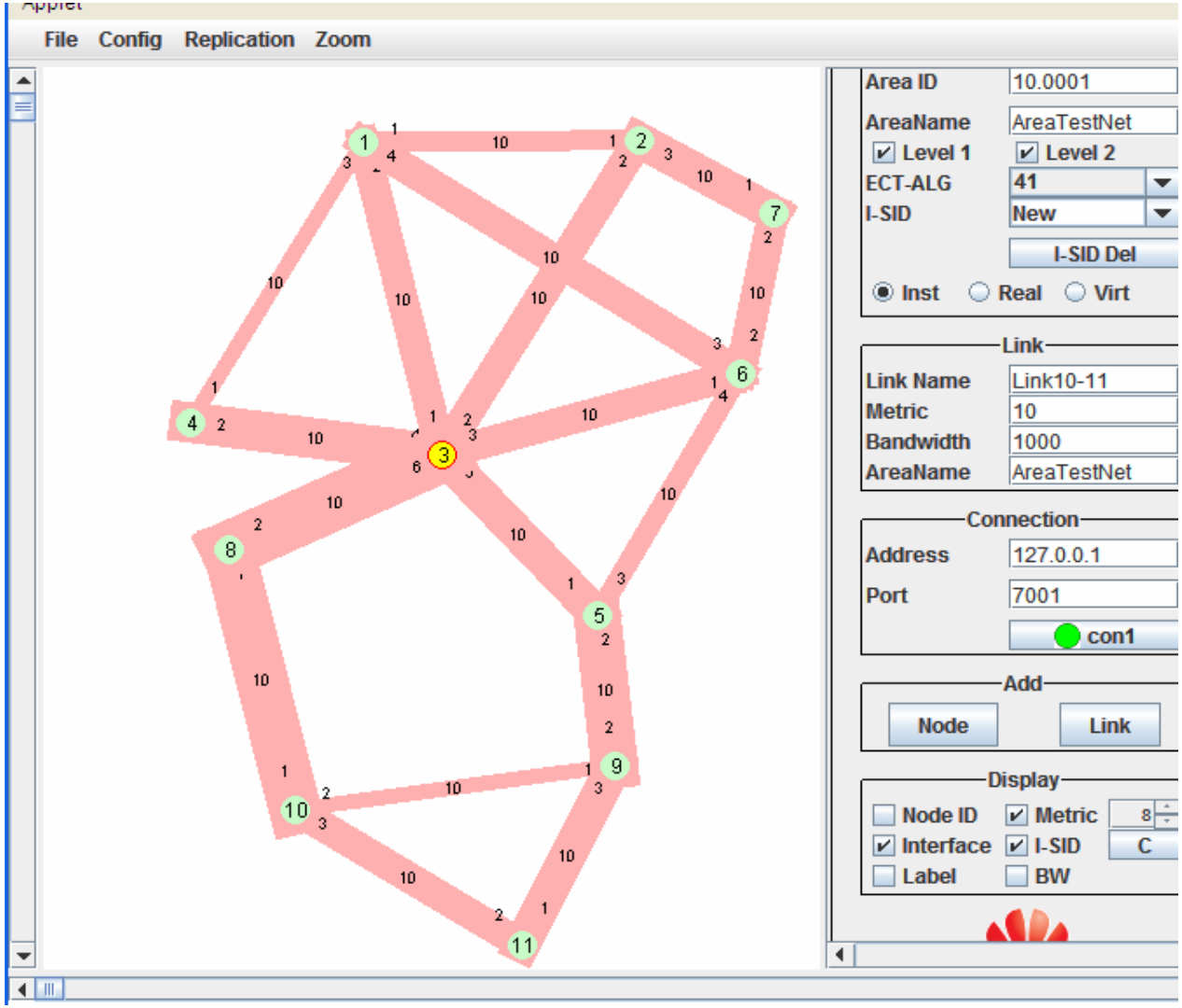
One ISIDs is pruned against this tree for Multicast, sub trees below: ISID 18



We show the Mcast state at node 3 for Each ISID.

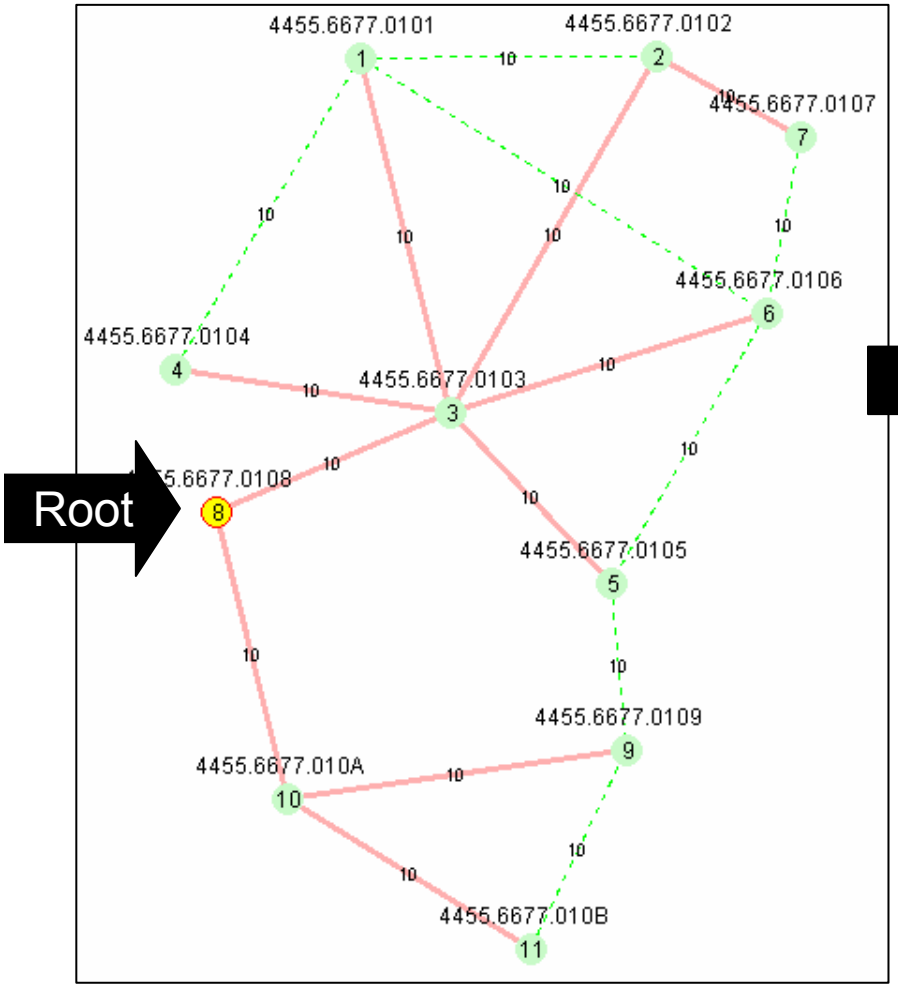
-CIST
-Pruned for ISID 16
-Pruned for ISID 17
-Pruned for ISID 18

Example#3- Coverage is not bad

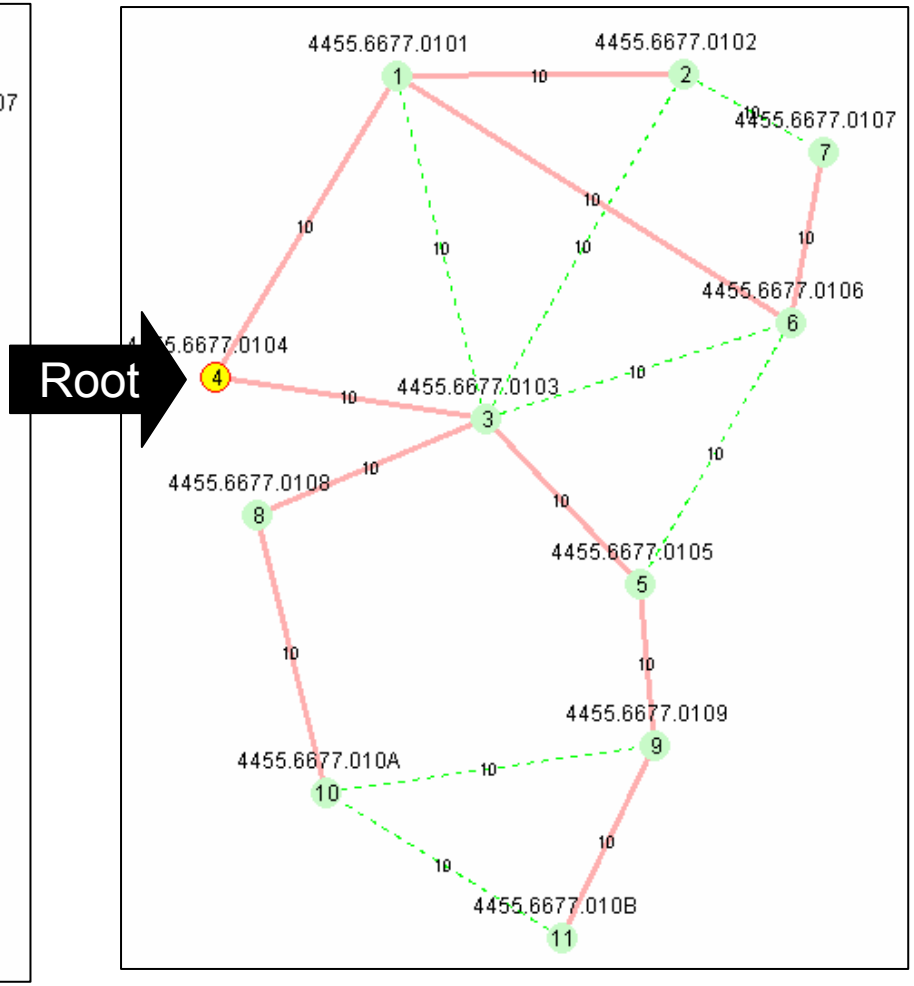


ALL 16 (*,G) Trees shown superimposed. Basically the CIST algorithm
16 times but with different root choices based on BridgeIdentifier XOR Mask[i]

Example#3- Some of the individual trees

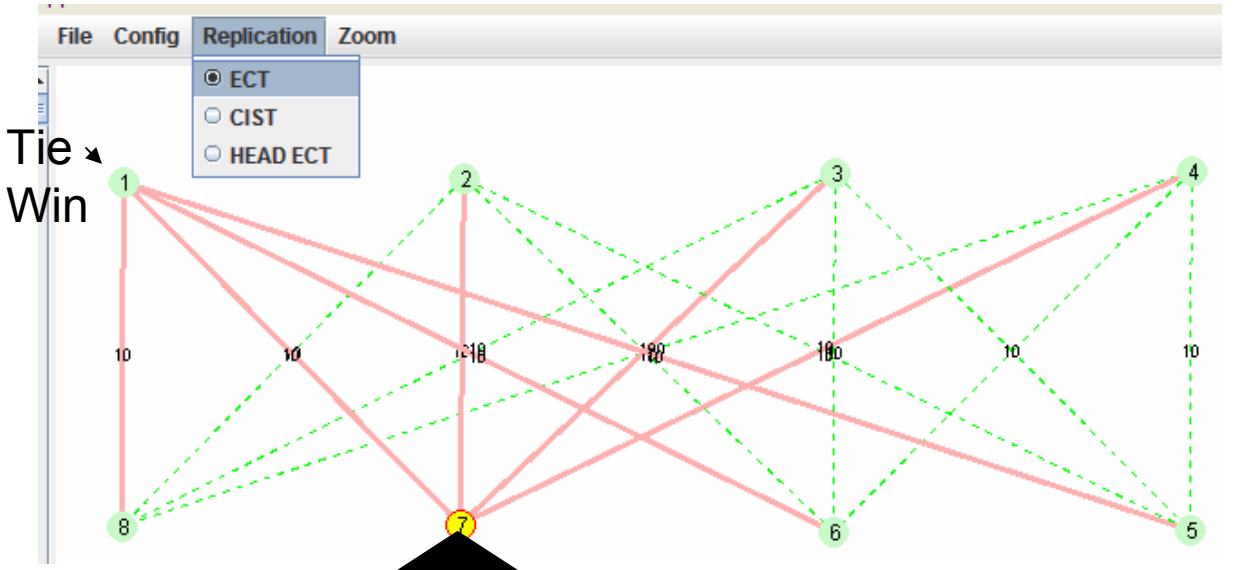


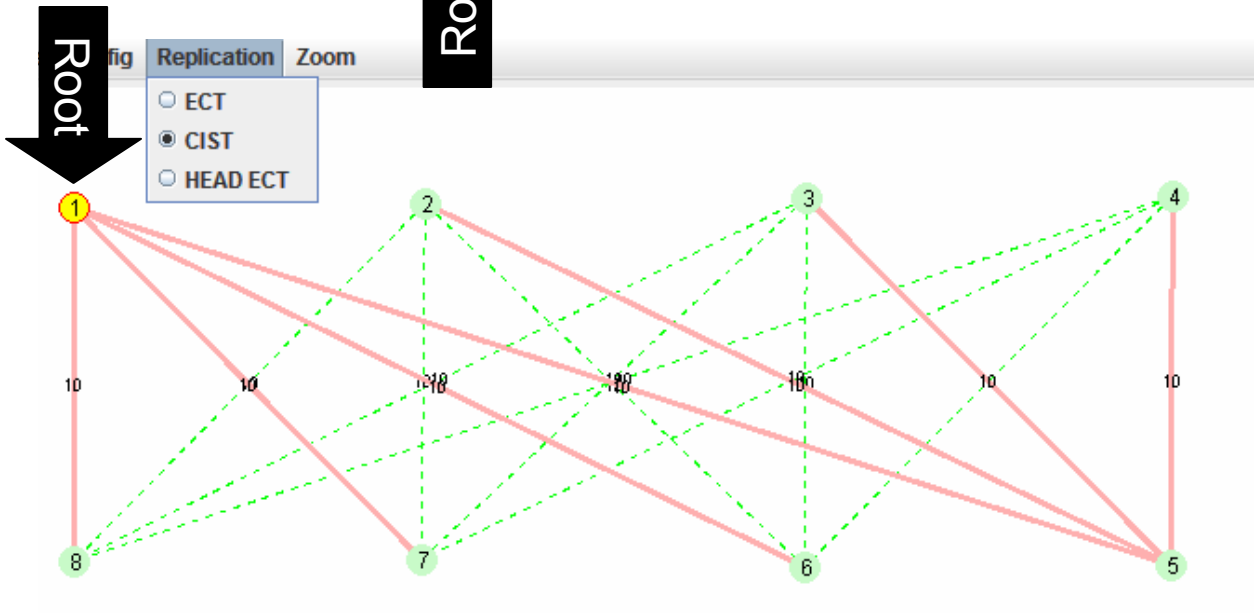
ALG MASK=0x8888.
So node .. 108 is root.



ALG MASK=0x444444.
so node 104 is root.

Example#3- Comparison to ECT source tree in Fat Tree





(S,G) Tree

Unicast and Mcast Routes from Node 7 to all other nodes.

(*,G) TREE

Multicast Shared Tree Routes from node 7.

Note routes to all other leaves 8,6,5 is identical To (S,G) tree above.

Basic Algorithm

Compute Shared Tree (alg, self) { // alg==0 => .1aq CIST

root = find lowest BridgeIdentifier **XOR Mask[alg]***

run SPF from root where

tie break on equal cost winner =

lowestBridgeIdentifier **XOR Mask[alg]***

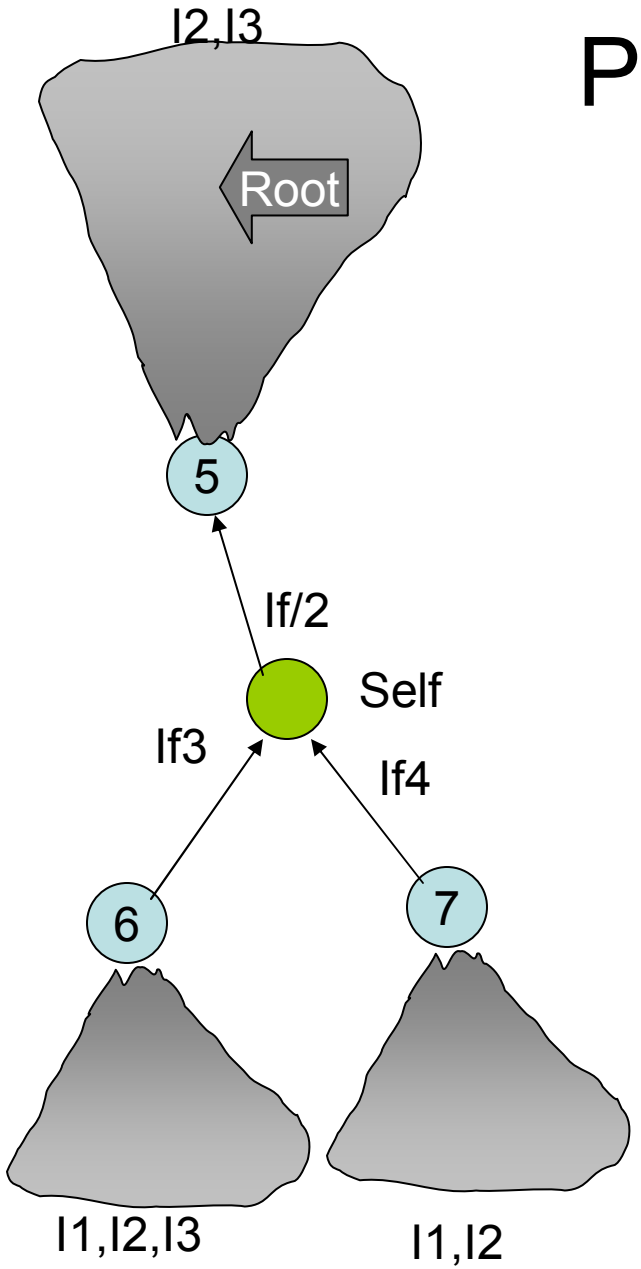
}

Multicast DA per ISID can then easily be generated by sorting the set of all ISIDs and the interface to reach that that ISID... Next slide..

So total run time is O(16 x [(NxLog(N)) + (I x Log(I))])

** Recall Mask[] = {0x00.., 0xff., 0x11.., 0x22.., 0x33..., 0xee.. }*

Pruning - One Possibility



1. At self do the SPF from selected root.
   Result is upward pointing parent pointers to root.

2. For each node in network assign it the local interface that reaches it. Eg: 5 and everything above it via if/2; 7 and everything below it by if/4 etc.

3. Then traverse network and generate a list of.
   <ISID, IF/#> records ..will have lots of duplicates.

I2,if/2
I3,if/2
I1,if/3
I2,if/3
I3,If/3
I1,if/4
I2,if/4

4
SORT →

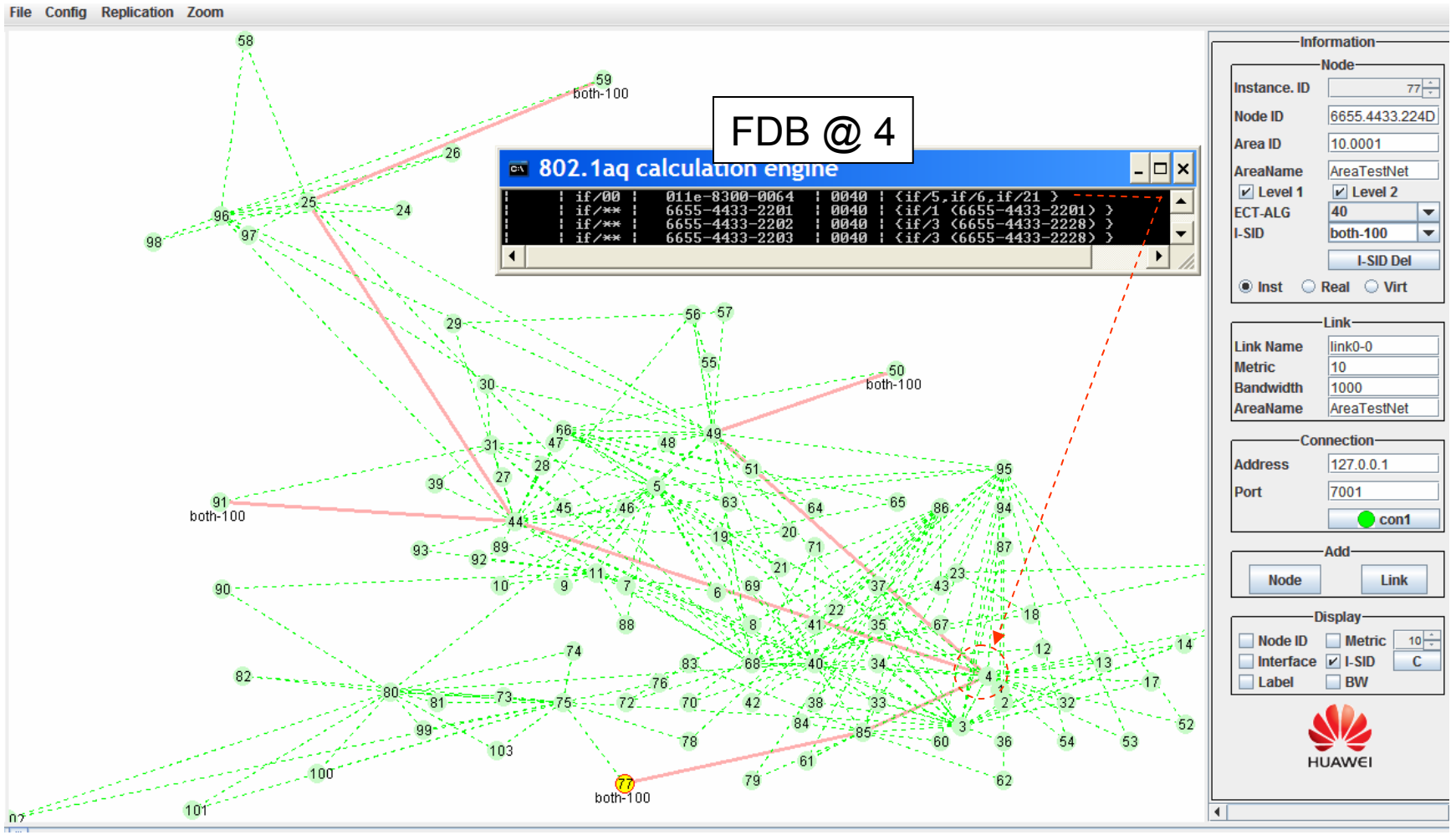
I1,if/3
I1,if/4
i2,If/2
i2,if/3
i2,If/4
i3,if/2
i3,if/3

Traverse and generate Mcast

DA=00-1e-83-00-00-02
OutIF = {if/2, if/3, if/4]

Ignore if only reachable via one interface ..

100+ node example – ISID 100 with 4 attachment points



Notes : Addressing Options

1. ***DA = 01-1e-83-xx-xx-xx & VID = F(ALG)***
   - ***Tree is identified by the VID so overlapping ISIDs (as used for ECMP) requires the VID to differentiate.***
   - ***Local bit NOT set so can co-exist with (S,G) .1aq trees.***
2. ***DA = F(ALG)-xx-xx-xx & VID = Const | Absent***
   - ***Tree is identified by the DA so overlapping ISIDs do not require VID to differentiate and in fact VID can be absent even with overlapping ISIDs.***

Note that encoding root of (,G) tree in address appears unnecessary as root is a function of the Algorithm used to pick tree.*

Note 24 bit ISID value represented as xx-xx-xx

Addressing Option 2 Cont'd

802.1aq Group MAC format

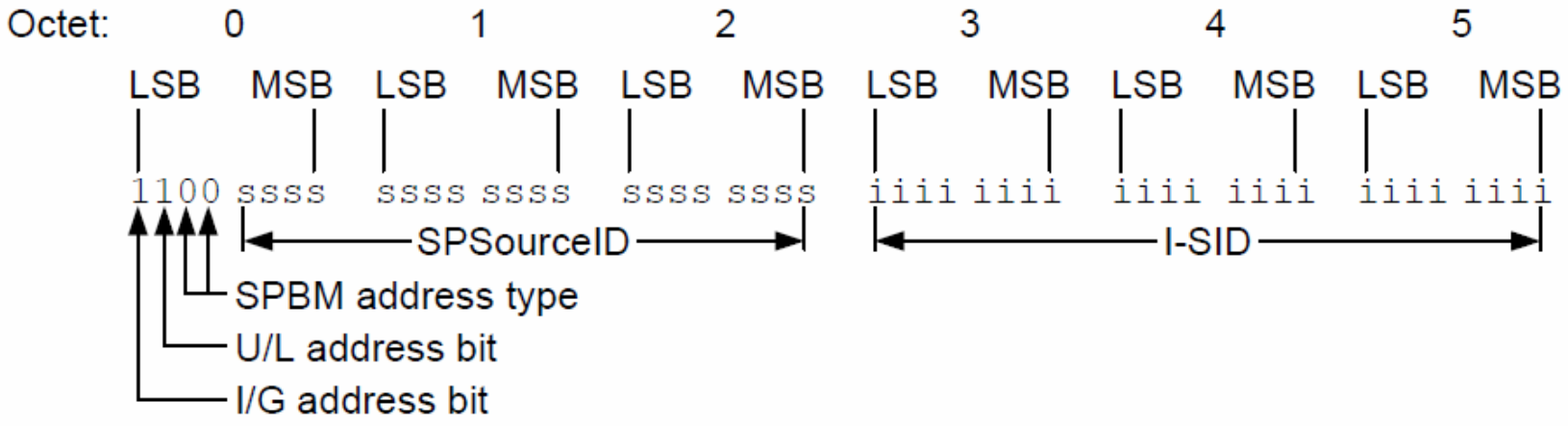


Figure 27-3—SPBM Group MAC Addresses used by this standard

SPBM uses only address type 00, therefore we could use address type 01 to implement:

DA = 1101-F(ALG):20-xxxxxxxxxx.....xxxxxxx

So this gives up to 2^{20} shared trees .... more than we can do with B-VID!

Questions

- Do we need new *BridgePriority* for (*,G) root selection?
  - We could allow greater flexibility with separate ECT tie breaking *BridgePriority* and ECMP root selection *BridgePriority* but it adds more complexity...
  - We have opaque TLV that can be used to carry new ‘things’ without involving ISIS-wg.