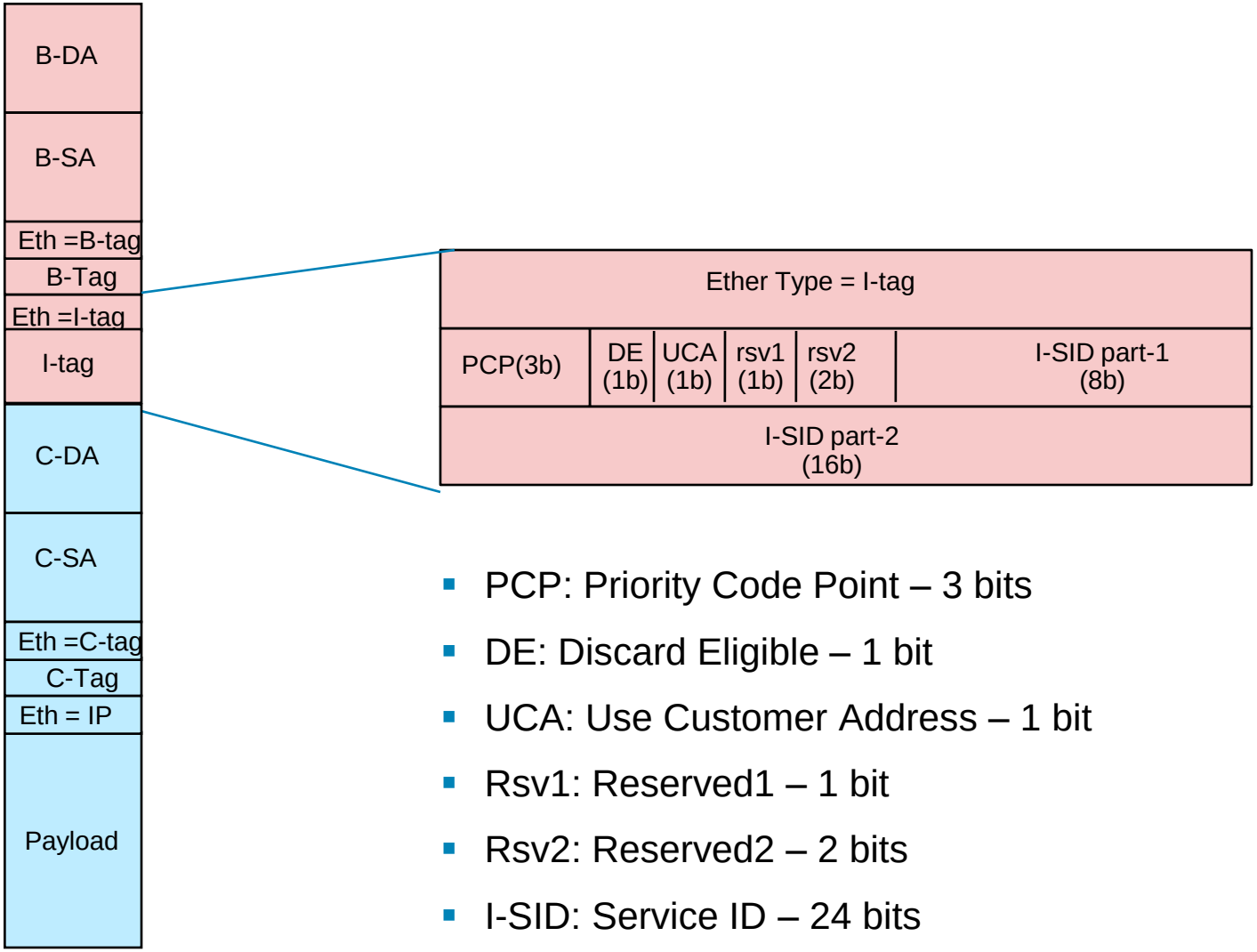
Bridge Model for ECMP Operation

Ali Sajassi, Peter Ashwood-Smith, Don Fedyk, Paul Unbehagen, Srikanth Keesara

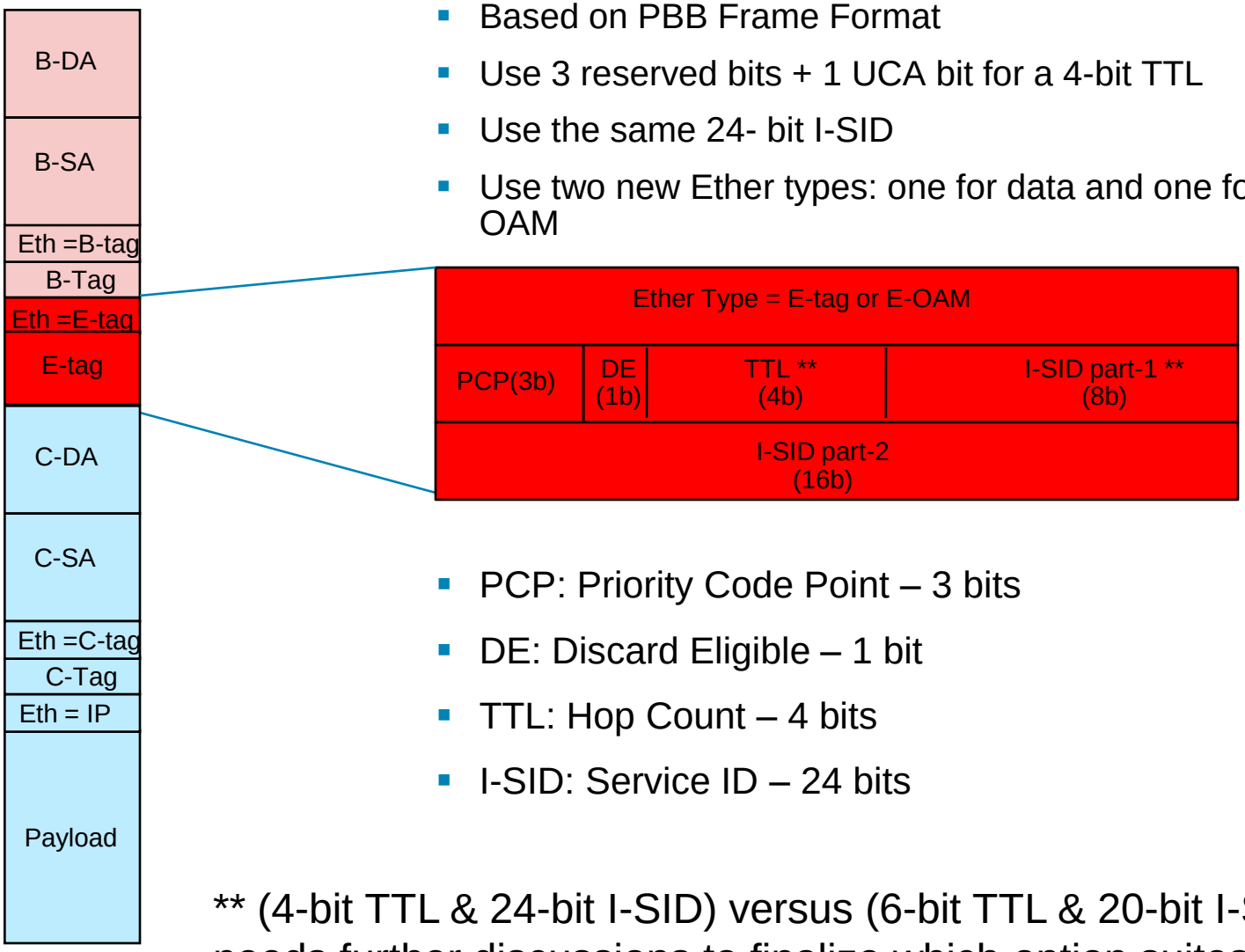
January 12, 2010
IEEE 802.1 Interim Meeting

Existing PBB Frame Format



- PCP: Priority Code Point – 3 bits
- DE: Discard Eligible – 1 bit
- UCA: Use Customer Address – 1 bit
- Rsv1: Reserved1 – 1 bit
- Rsv2: Reserved2 – 2 bits
- I-SID: Service ID – 24 bits

New ECMP Frame Format



- Based on PBB Frame Format
- Use 3 reserved bits + 1 UCA bit for a 4-bit TTL
- Use the same 24- bit I-SID
- Use two new Ether types: one for data and one for OAM

- PCP: Priority Code Point – 3 bits
- DE: Discard Eligible – 1 bit
- TTL: Hop Count – 4 bits
- I-SID: Service ID – 24 bits

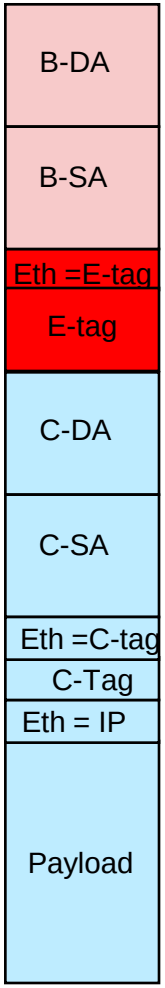
** (4-bit TTL & 24-bit I-SID) versus (6-bit TTL & 20-bit I-SID) needs further discussions to finalize which option suites our application the best

Explicit versus Implicit Flow ID

- Implicit Flow ID:
  - No explicit flow ID is passed in the frame
  - Every node along the path (except the egress BEB) computes the flow ID directly using the n-tuple flow parameters
  - Each node factors in a local parameter (such as system ID) in flow-id computation to avoid polarization
  - Each node uses this flow-id for its ECMP path selection
- Explicit Flow ID:
  - Ingress BEB computes a m-bits hash index (16, 20, or 24 bits) based on n-tuple flow parameters
  - Ingress BEB XORs this hash index with I-SID to get a flow-id
  - All core bridges (BCBs) use this flow-id in conjunction with a local parameter (e.g., system ID) for ECMP path selection and to avoid polarization
  - On the egress PE, demodulation is performed by first calculating the 20-bit hash index from the five-tuple and next XORing it with the flow-ID thus retrieving E-SID

NOTE: The use of homogenous ECMP algorithm (per B-VID) enables the egress PE of demodulating flow-ID and retrieving E-SID (another advantage of having homogenous ECMP)

ECMP Operation w/ only E-tag - Optional



If the network only requires a single ECMP algorithm, then the default B-VID can be used to identify this algorithm, thus avoiding sending B-tag with every frames and making the encapsulation the most optimal

This shortened frame format only applies to SPBM unicast frames because ECMP doesn't apply to multicast frames as ECT is used for the multicast frames

Pros/Cons

- Pros
  - Most efficient and compact encapsulation
  - E-SIDs and flow-ids are transported and used only where they are needed (e.g., E-SIDs used in BEBs and flow-ids in BCBs)
  - 4-bit (or 6-bit) TTL accommodates both SP and DC applications
  - It is based on 802.1ah frame format that many vendors and providers are familiar with
  - Use of flow ID avoids any deep packet inspection in the BCBs
  - ECMP frames can optionally be sent w/o B-tag resulting in the most efficient encapsulation possible
- Cons
  - Requires hash-index computation in disposition path of BEB
  - May reduce the Service ID field to 20 bits if 6-bit TTL is used (but this doesn't create any inter-op issue as we will see)

Explanation of Cons

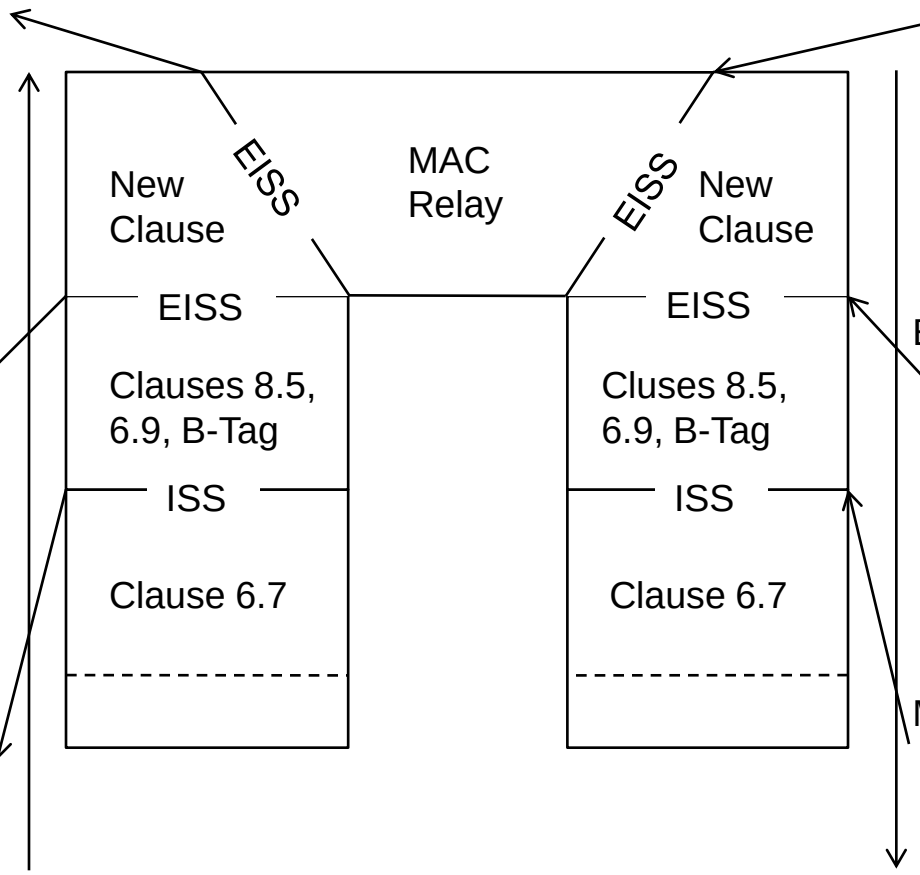
- Requires hash-index computation but
  - No additional logic is required in BEB because hash-index computation is needed on imposition path of BEB anyway
  - Hash-index computation on the disposition path makes processing symmetric in BEB
  - If there are LACP egress ports, then hash-index computation on disposition path is required anyway which is the typical case in the scenarios of interest for this project
- May reduce the Service ID field to 20 bits but
  - This doesn't create any inter-op issue because ENNI between ECMP and ECT domains will be provided via I-tag service interface – e.g., TTL don't cross domains and E-SID <-> I-SID translation is provided by ECMP-domain BEBs just like I-SID <-> I-SID translation in PBBN

Modified Baggy Pants Diagram for only TTL processing at Bridges

EM_UNITDATA.indication (
- destination_address,
- source_address,
- mac_service_DU,
- priority,
- drop_eligible,
- vlan_identifier,
- flow_identifier,
- etc.)

EM_UNITDATA.indication (
- destination_address,
- source_address,
- mac_service_DU,
- priority,
- drop_eligible,
- vlan_identifier, etc.)

M_UNITDATA.indication (
- destination_address,
- source_address,
- mac_service_DU,
- priority,
- drop_eligible, etc.)



EM_UNITDATA.request (
- destination_address,
- source_address,
- mac_service_DU,
- priority,
- drop_eligible,
- vlan_identifier,
- flow_identifier,
- etc.)

EM_UNITDATA.request (
- destination_address,
- source_address,
- mac_service_DU,
- priority,
- drop_eligible,
- vlan_identifier, etc.)

M_UNITDATA.request (
- destination_address,
- source_address,
- mac_service_DU,
- priority,
- drop_eligible, etc.)

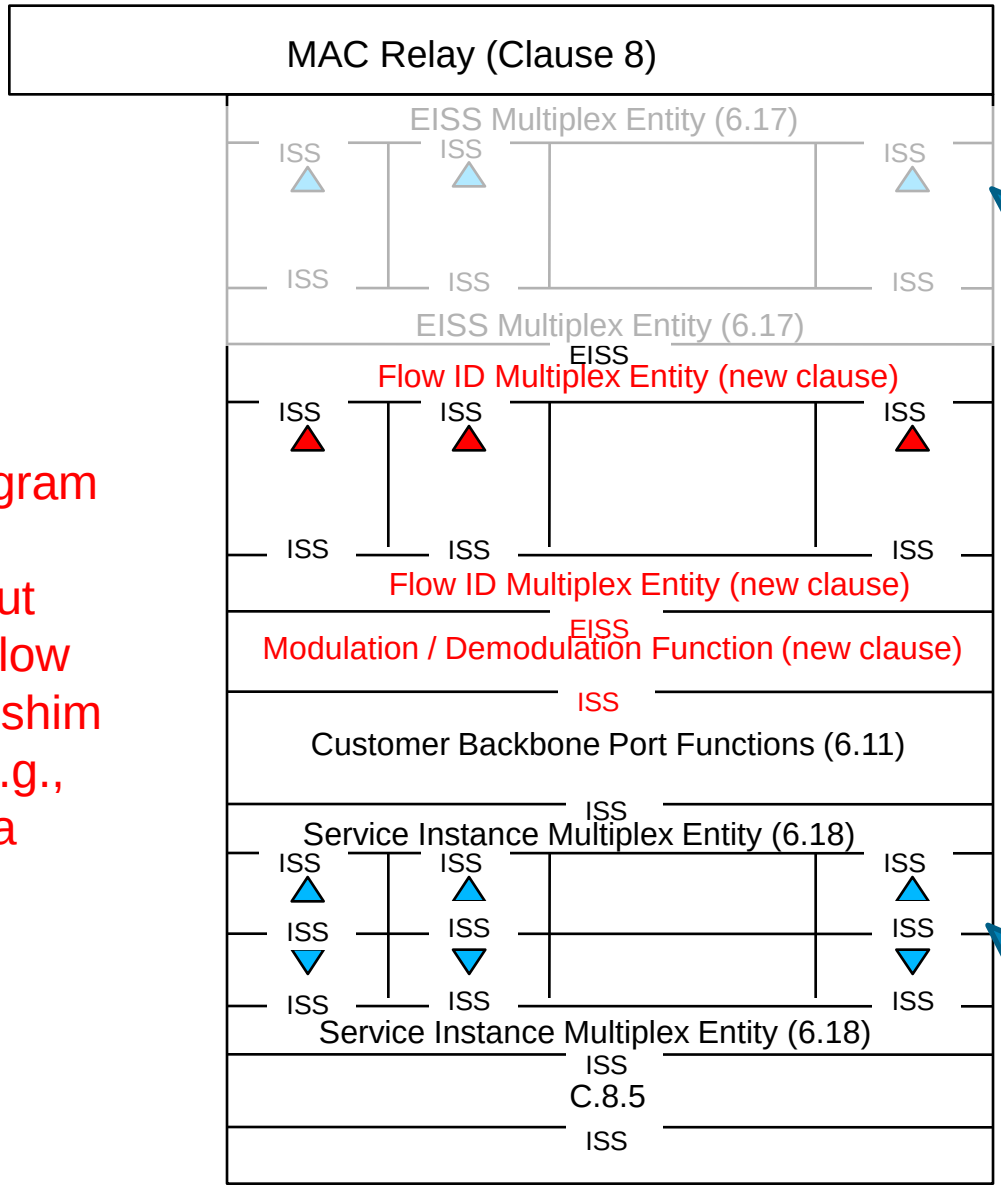
New Clause

- If next tag (after B-tag/S-tag) is E-tag, then extract TTL and flow_identifier and perform the following functions
- Use the flow_identifier in the MAC relay to select among ECMPs
- Use TTL to perform loop mitigation as follow:
  - Upon receiving TTL, if zero then discard the frame; otherwise, decrement TTL and process the frame
  - After decrementing TTL, if TTL==0 and EtherType = OAM (or UCA==0), then perform OAM processing
  - When setting TTL for unicast frames, it should be set to more than the min. required to accommodate re-forwarding during failure scenarios
  - When setting TTL for multicast frames, it should be set to the longest branch in the multicast tree plus a delta

New Clause – Cont.

- Flow-id is calculated and passed as a parameter of EISS API to MAC relay
- The MAC relay filtering database is enhanced so that for MAC addresses that correspond to ECMPs, it maintains several interface IDs for each MAC address since different ECMPs can take different interfaces.
- The MAC relay uses the flow-id to hash among different interface IDs for a given MAC address and select one of them

Baggy Pants Model for OAM operation at BEB



The MEPs for B-VIDs are not used when doing ECMP

E-SID MEPs requires enhancement to perform round-robin of CCs among different flows for a given E-SID

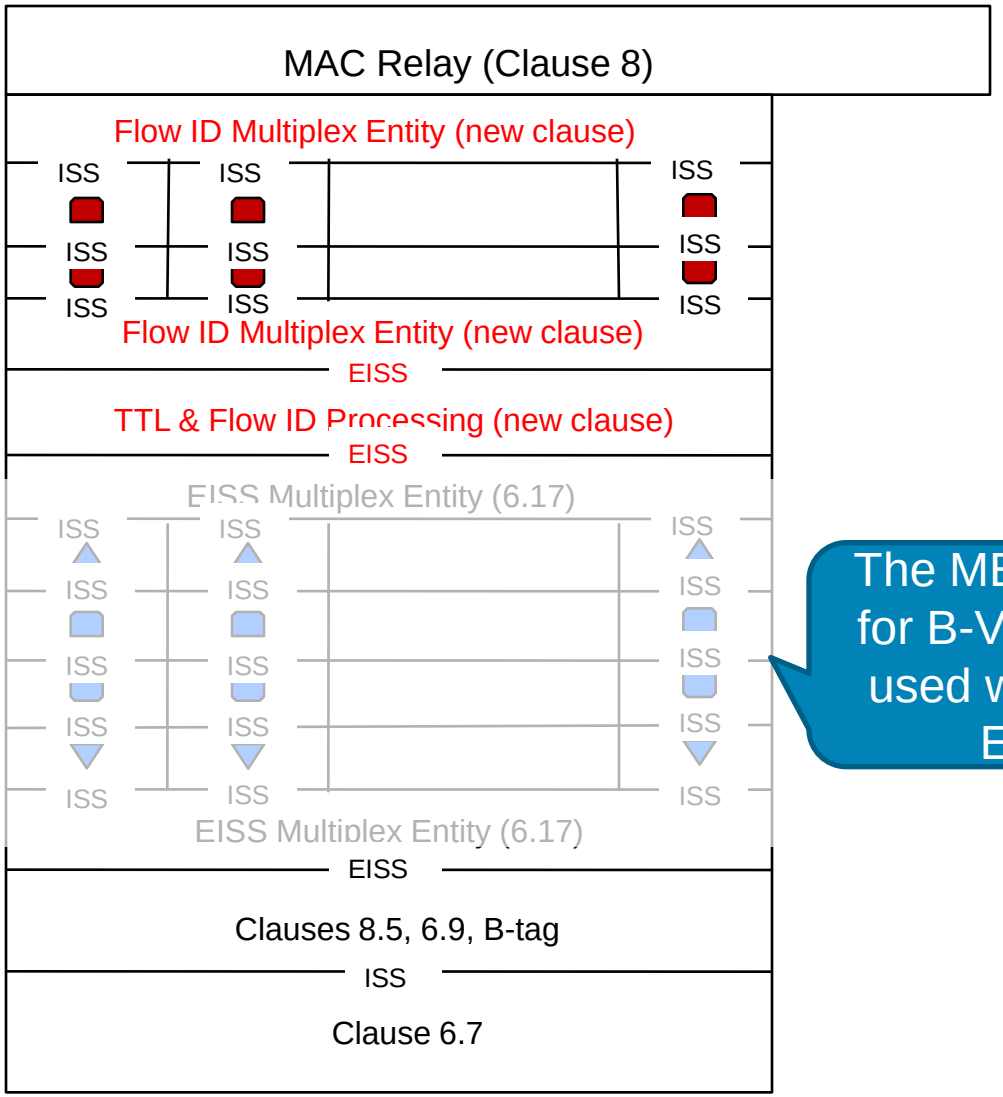
Note: This digram needs to be modified to put the shim for flow ID inside the shim for B-VID – e.g., to represent a nested shim.

Baggy Pants Model for OAM at BEB – Cont.

- The reason for having the flow MEPs at CBP instead of PIP is:
  a) CBP needs to receive un-modulated E-SID in order to perform functions of clause 6.11
  b) But if E-SID is un-modulated, then there is no flow-id and thus there can be no MIPs functions on interim BCBs
- There is no need to configure MEPs for B-VIDs because for ECMP operation, a given B-VID identifies a specific ECMP algorithm and not a broadcast domain!!
  a) A traditional MEP on B-VID can only monitor a single path among many possible paths for that B-VID at the presence of ECMP
  b) If using a single ECMP algorithm network wide, then the use of B-VID is optional
- E-SID MEPs require additional enhancement to transmit CC messages on a round robin among different flows for a given E-SID

Baggy Pants Diagram for OAM operation at BCB

Note: This digram needs to be modified to put the shim for flow ID inside the shim for B-VID – e.g., to represent a nested shim.



The MEPs & MIPs for B-VIDs are not used when doing ECMP

OAM Granularity: Network, Service & Flow

- Network OAM: OAM functions performed on a Test VLAN. Test Flows are chosen to exercise all ECMPs for the Test VLAN.
- Service OAM: OAM functions performed on the user VLAN itself. Test Flows are chosen to exercise all the ECMPs.
- Flow OAM: OAM functions performed on the user Flows.

Flow OAM (reactive)

- User supplies flow information, including one or more of:
  - •MAC SA and/or DA
  - •IP Src and/or Dst
  - •Src and/or Dst Port (TCP or UDP)
- Flow parameters are converted to a flow ID (e.g., NMS can query platform using flow parameters and get back flow ID)
- MEP monitors the flow by sending periodic CCMs for that flow.
  - •Monitoring of unicast flows uses unicast CCMs
  - •Monitoring of multicast flows uses multicast CCMs

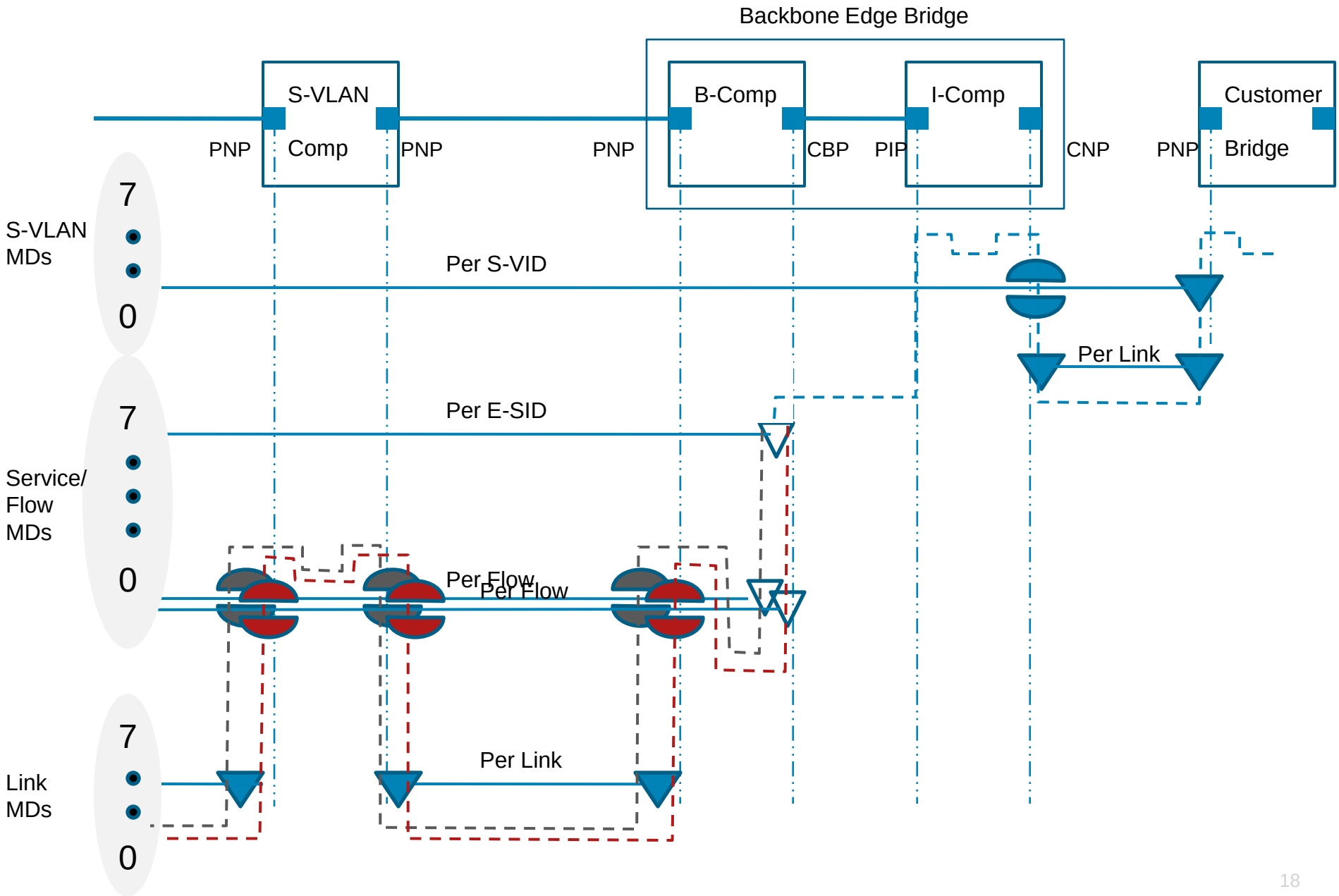
Service OAM (proactive)

- A MEP, knowing the topology and how to exercise the ECMPs, first calculates the necessary Test Flows for full coverage of all paths in a given service instance.
- On a per service instance basis, MEPs perform monitoring of all unicast and multicast paths using the Test Flows.
- MEPs follow a ‘round-robin of Test Flows’ scheme to verify connectivity over all ECMP paths (unicast) and shared trees (multicast).
  - –Round-robin scheduling reduces processing burden on nodes, and modulates the volume of OAM messaging over the network.
  - – Comes at the expense of relatively longer fault detection time
  - – For critical flows, it is possible to schedule their connectivity check continuously.
  - – MEP CCDB will track every flow independently (timer per flow per remote MEP rather than per remote MEP in CFM)

Network OAM (Proactive)

- Network OAM is a degenerate case of service OAM where a single default E-SID can be configured on all BEBs and the CFM is performed for that default E-SID just as described above for service-level OAM
  - This default E-SID is per B-VID – e.g., per ECMP algorithm. If there are multiple ECMP algorithms in the network and the E-SIDs are divided among these algorithms, then one default E-SID is needed per E-SID group (e.g., per B-VID).
  - Typically there is only a single ECMP algorithm

CFM Flow

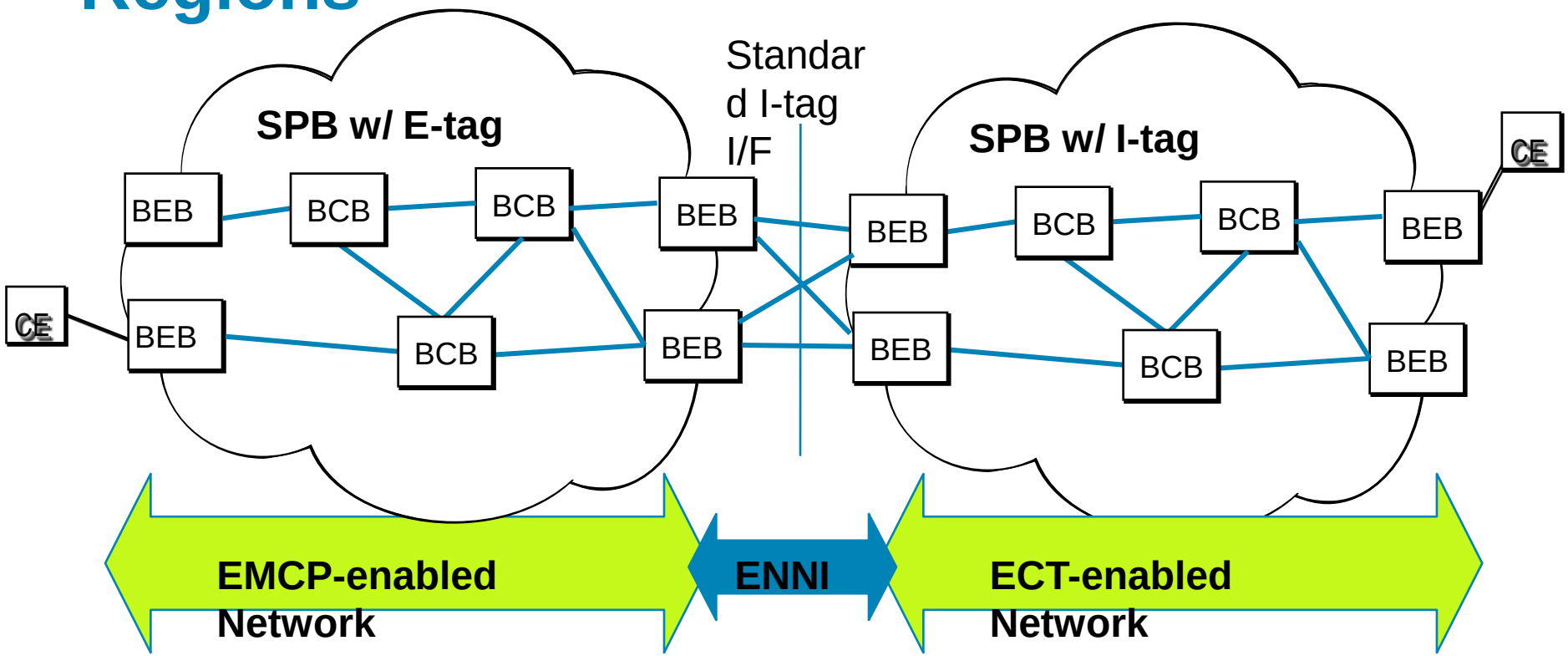

Backbone Edge Bridge
S-VLAN
Comp
B-Comp
I-Comp
Customer
Bridge
PNP
PNP
PNP
CBP
PIP
CNP
PNP
S-VLAN
MDs
7
0
Per S-VID
Per Link
Service/
Flow
MDs
7
0
Per E-SID
Per Flow
Per Flow
Link
MDs
7
0
Per Link

New AIs

- Need to define a homogeous ECMP algorithm to do both hash-index computation and ECMP selection

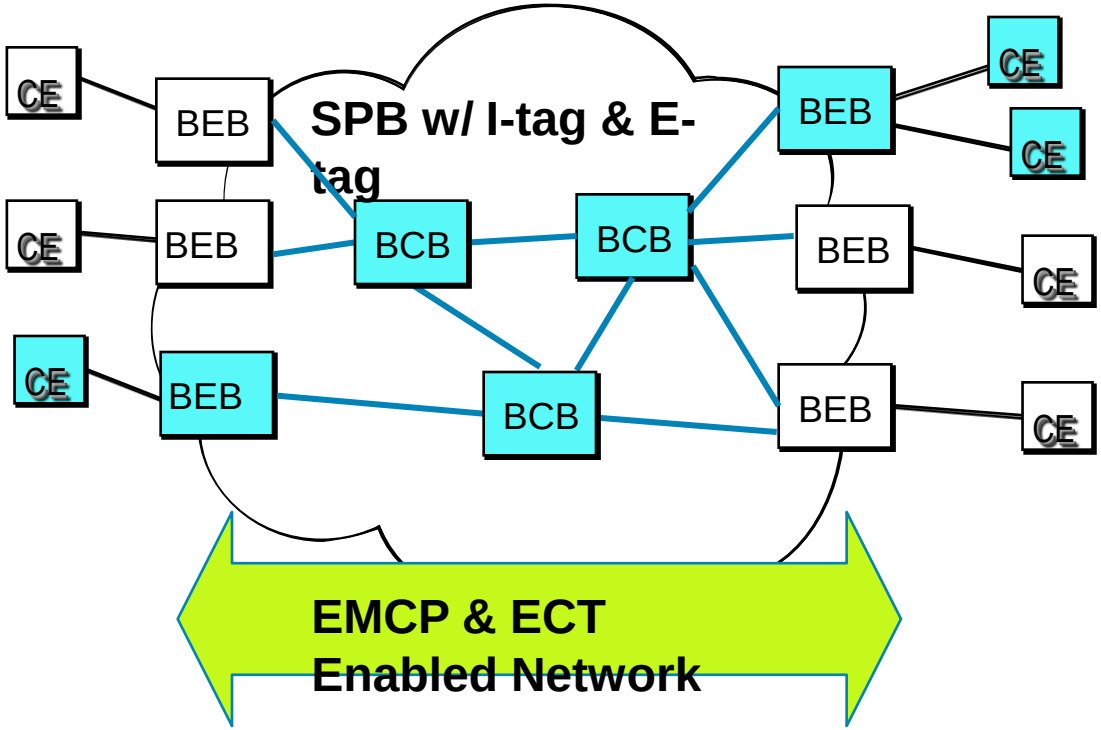
Appendix – A Interoperability

Inter-operability: Between Different Regions



- BEB in the ECMP-enabled network will perform E-SID <-> I-SID mapping per existing 802.1ah functionality (per clause 6.11)
- BEB in the ECMP-enabled network will encode the derived I-SID into its corresponding I-tag and then send it to ECT-enabled network
- No changes is needed on the ECT-enabled network (both BEBs and BCBs)
- Number of I-SIDs supported over ENNI (using I-tag service interface) will be limited to 1 million instead of 16 millions (still much larger than any practical requirements) !!
- If needed to support 16 millions or more (upto 4 billions), then we can limit the scope of E-SID to B-VLAN

Inter-operability: Within one Region



- Use designated B-VID(s) for ECMP just like
  - A set of B-VIDs for 802.1aq (one per ECT)
  - A set of B-VIDs for PBB-TE
  - A set of B-VIDs for PBB with MSTP
- To support ECMP
  - Some BCBs must support TTL
  - Only BEBs that are configured for E-SIDs, need to support TTL

Backward Compatibility

- Any per-hop ECMP (whether TTL is used or not) requires additional new processing anyway:
  - Hashing based on user data flow headers to determine egress interface or
  - using the pre-calculated hash-index to determine egress interface
- A network can be configured to simultaneously support ECMP and ECT modes
- In a single network, we cannot mixed ECMP service points with non-ECMP because it doesn't make sense
- In multiple networks where an ECMP service in one network needs to interoperate with non-ECMP service in another network, I-tag mapping capability of BEB can be used to ensure such interoperablity
- Multiple topology configuration can be used to support both ECMP and non-ECMP BCBs in the same network and ensure gradual migration


CISCO