Bridge Model for ECMP Operation

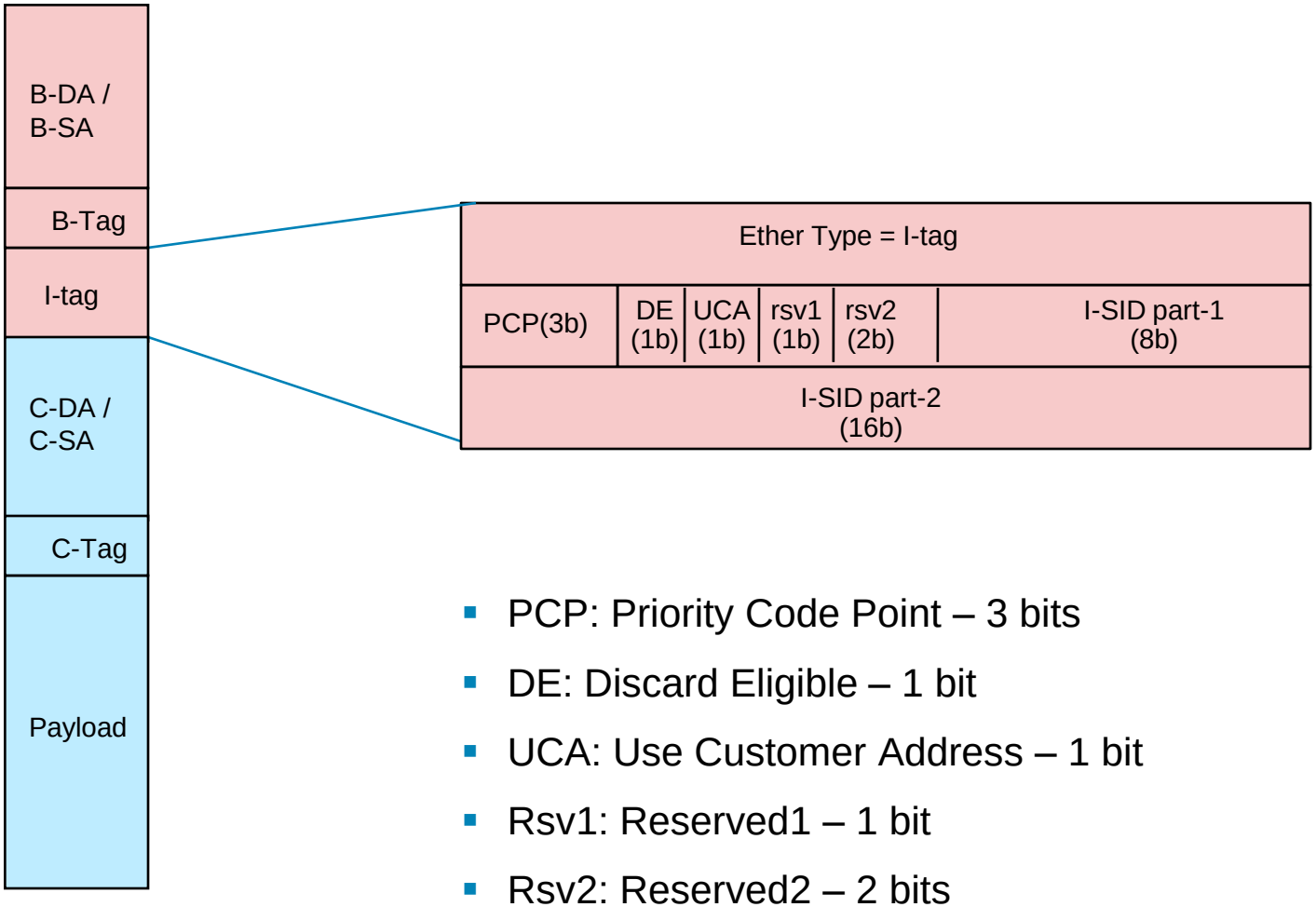
Ali Sajassi

February 24, 2011
802.1 Bi-Weekly ECMP Call

Requirements

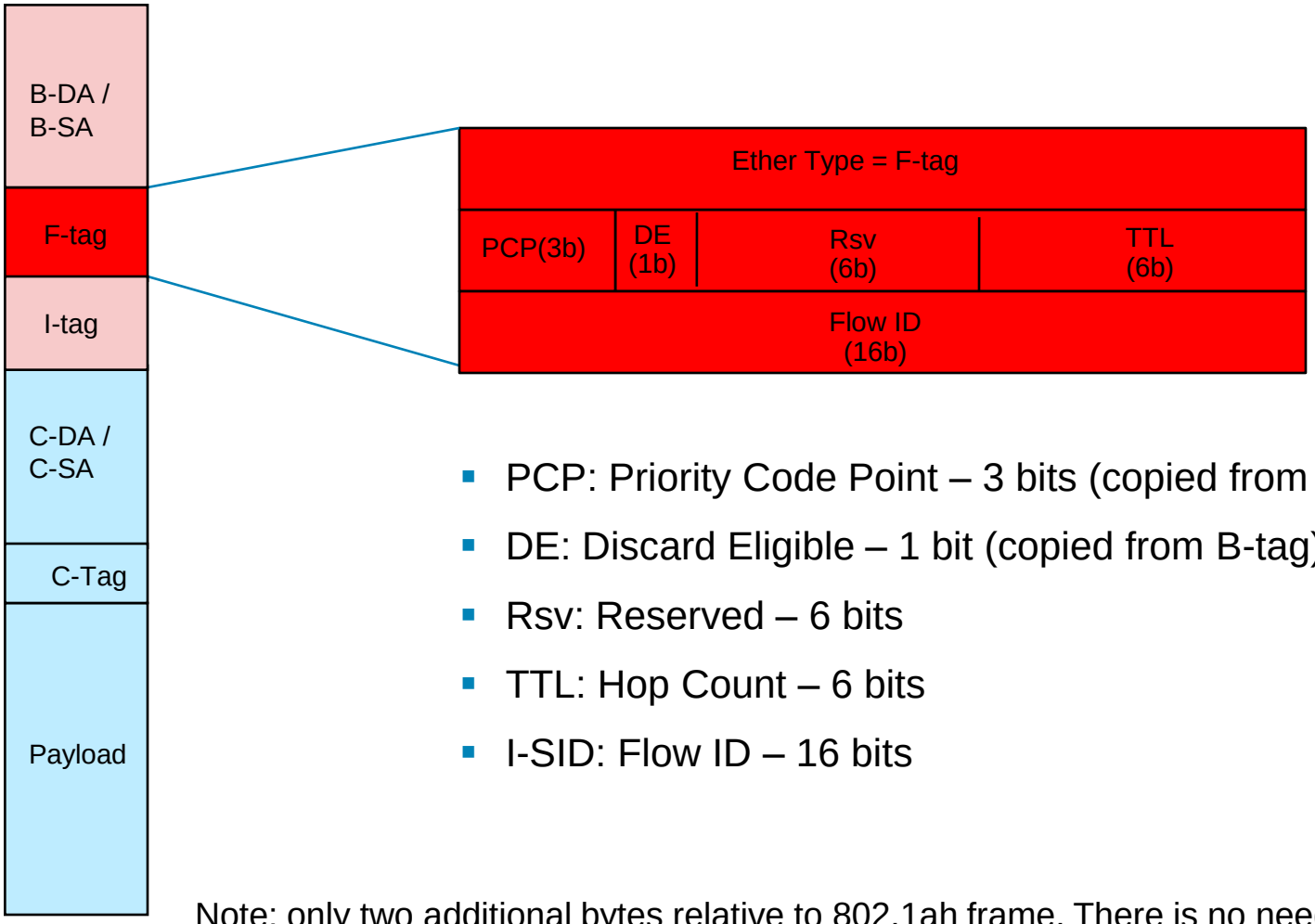
- Support of per-hop ECMP
- Support of TTL for loop mitigation
- Support of flow-id
  - To avoid deep packet inspection in the core
  - To provide proactive service-level monitoring
- No fixed n-tuple hash algorithm for flow-identification
  - Any node can use any set of n-tuples and any algorithm to derive a flow id
- Homogenous ECMP hashing algorithm network wide to support proactive service-level monitoring
  - For a given flow-id, the path for that flow through the network is deterministic
-

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



- PCP: Priority Code Point – 3 bits (copied from B-tag)
- DE: Discard Eligible – 1 bit (copied from B-tag)
- Rsv: Reserved – 6 bits
- TTL: Hop Count – 6 bits
- I-SID: Flow ID – 16 bits

Note: only two additional bytes relative to 802.1ah frame. There is no need to send B-tag in the frame because unicast ECMP frames don't need B-VID. PCP/DE portion of B-tag is reflected in the F-tag.

Pros

- Modular tag design consistent with IEEE baggy pants diagram and shim addition/removal
- Keeps the I-tag intact so all the existing processing/procedure for I-tag can remain the same
- Easy processing at the disposition bridge - e.g., if there is an F-tag, then simply strip it and throw it away and process I-tag just as before
- Allows for F-tag to be used independently with other frame formats - e.g., SPBV
- PCP/DE bits in F-tag allows network operator to set CoS bits for ECMP packets independent from individual I-SIDs. This application is mostly relevant in MetroE as opposed to DC networks.

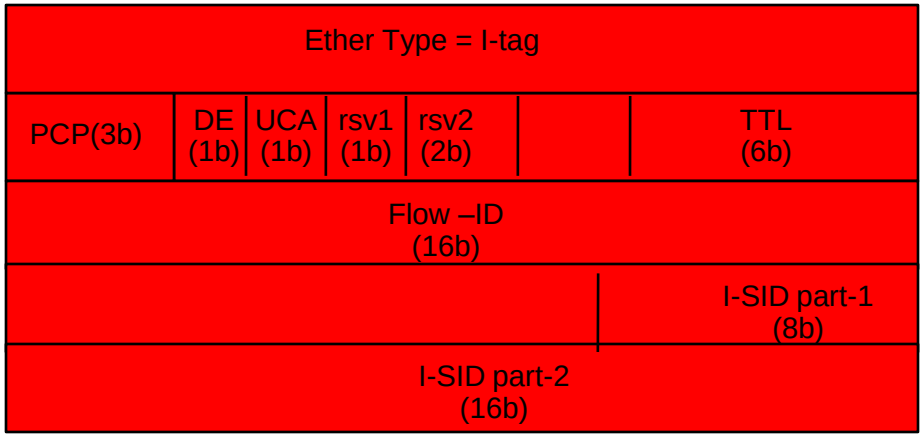
  Consider that there is a 802.1Qaq network at one end and 802.1Qbq (ECMP) at the other end and these two networks are getting connected via an I-tag NNI. The PCP/DE bits in F-tag and B-tag allows each respective networks (802.1Qbq and 802.1Qaq) to implement their CoS independent from individual I-SID CoS. In case of DC application, PCP/DE for F-tag can be simply set to the same one as I-tag.

Cons

- Two addition bytes are needed relative to 802.1ah frame format

  F-tag is two bytes longer than B-tag

  If a combined tag used for both F-tag and I-tag, then a saving of two additional bytes can be achieved because of single Eth-Type field. However, it will be difficult to compress further because of the need for even-word alignment



Rational for not including B-tag

- Currently in clause 6.11, I-SIDs are groups into different B-VIDs bins
- In 802.1aq, for a given I-SID, the same B-VID is used for both unicast and mcast frames because of congruency
- For ECMP operation, using the same B-VID for both mcast and unicast frames is both confusing and trouble some because B-VID identify different algorithms and thus the same B-VID cannot represent both ECT and ECMP algorithms
- For ECMP operation, the load balancing is performaced across the entire network (spanning across all the defined ECT). Therefore, ECMP should/can NOT use the same B-VID as ECT.
- Having a uniform algorithm for ECMP path selection based on flow-id, makes it possible to use a single default B-VID for ECMP I-SIDs
- Question: how do we indicate that an ECMP needs to be performed on an I-SID in CBP?
  - Should we simply add a 1-bit flag to the I-SID table to indicate if this I-SID to be subjected to ECMP processing.
  - What other options do we have if an I-tag frame is received over an I-tag interface ?
  - Adding a second column of B-VIDs for a given I-SID, would also solve this problem but it is too much of deviation/taxing
  - Using a different I-SID for ECMP unicast also solves the problem but creates some operational issues

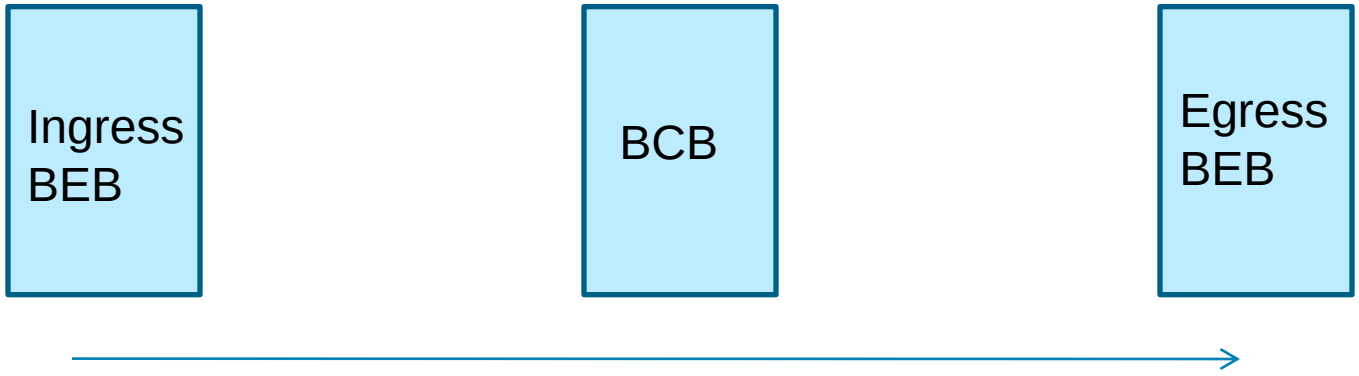
Hash () Requirements

- Homogenous hash() enables:
  - proactive service-level monitoring
  - validating performance of ECMP function – e.g., traffic is equally distributed among ECMPs
- Without homogenous hash(), only flow-level monitoring can be performed – similar to what is already done in MPLS and IP networks
- Without homogenous hash(), it is not possible to differentiate between a failure scenario and skewed hash() by a node

Hash Algorithm

- Break the hash algorithm into two parts:
  i) Use flow parameters (n-tuple) to generate a Flow ID
  ii) Use Flow ID and a local ID to generate a hash index
- Part-I is performed by only BEBs
- Part-II is performed by both BEBs and BCBs
- Only Part-II needs to be homogenous in order to meet the above requirements (which should be lot easier than mandating part-I to be homogenous)

Operation

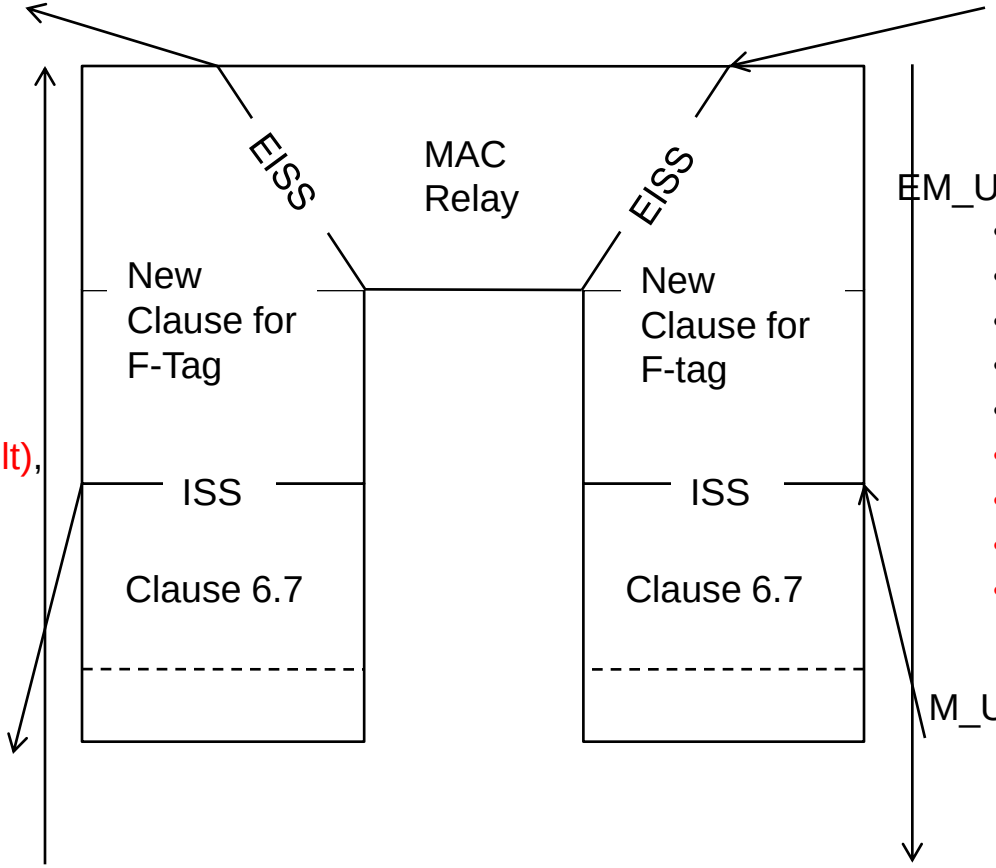


- Compute flow-id based on n-tuple
- Add F-tag in lieu of B-tag (w/ Flow-ID, TTL, PCP) to I-tag frame
- Perform ECMP using Flow-ID

- Perform per-hop ECMP using Flow-ID

- Simply strip and discard F-tag
- Proceed w/ I-tag processing as before

Modified Baggy Pants Diagram for TTL & Flow-ID processing at Bridges



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

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

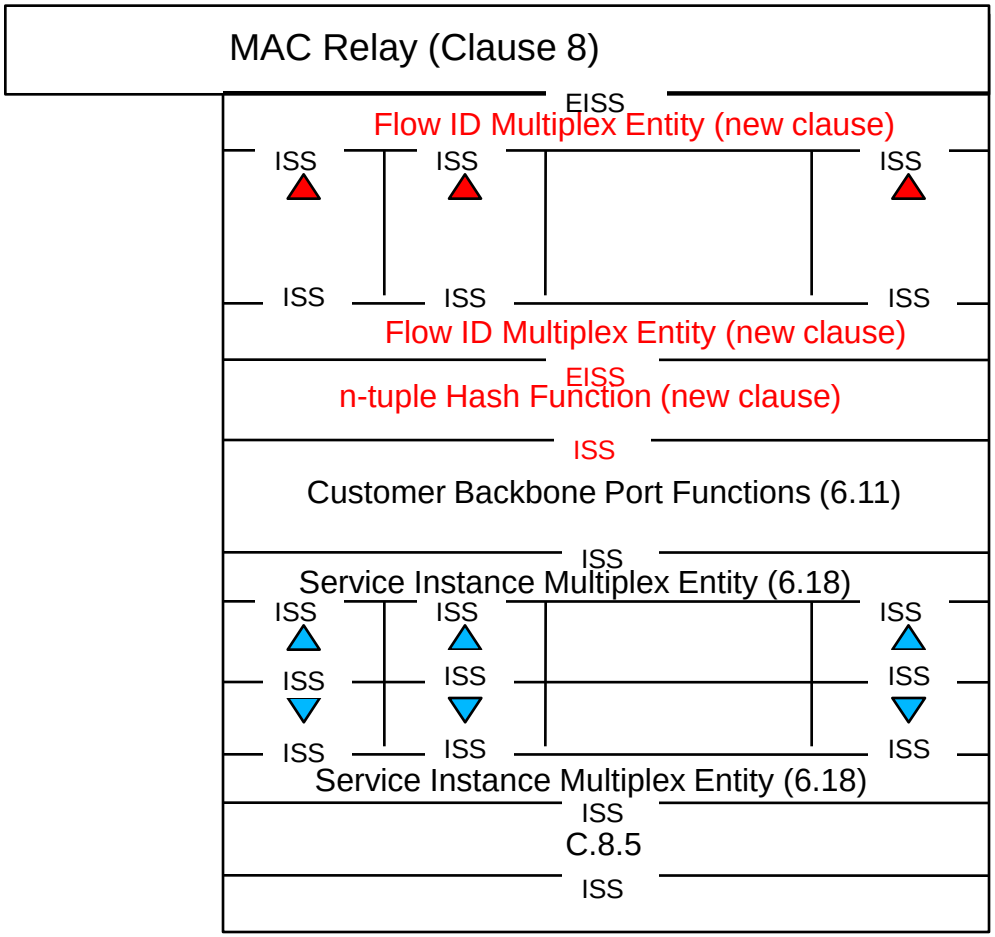
New Clause

- If the tag is F-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 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

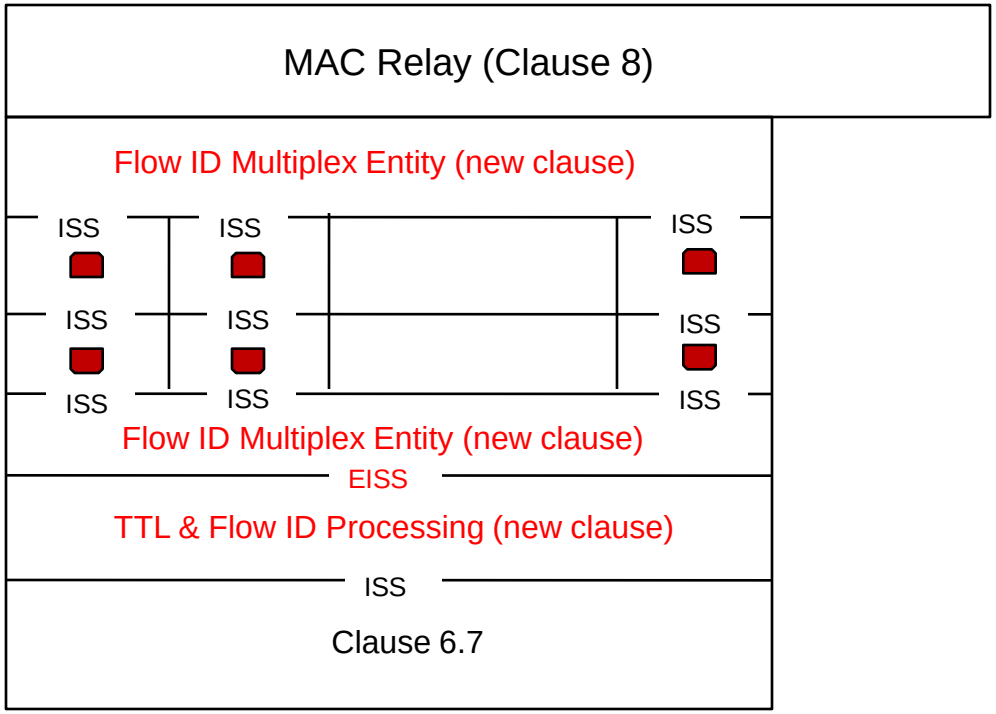


NOTE: Clause 6.11 needs to be modified to indicate that all ECMP I-SIDs are mapped to a single default B-VID

Baggy Pants Model for OAM at BEB – Cont.

- The reason for having the flow MEPs at CBP instead of PIP is to have a consistent model and operation for both BEB with B component and BEB with IB components
- I-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



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

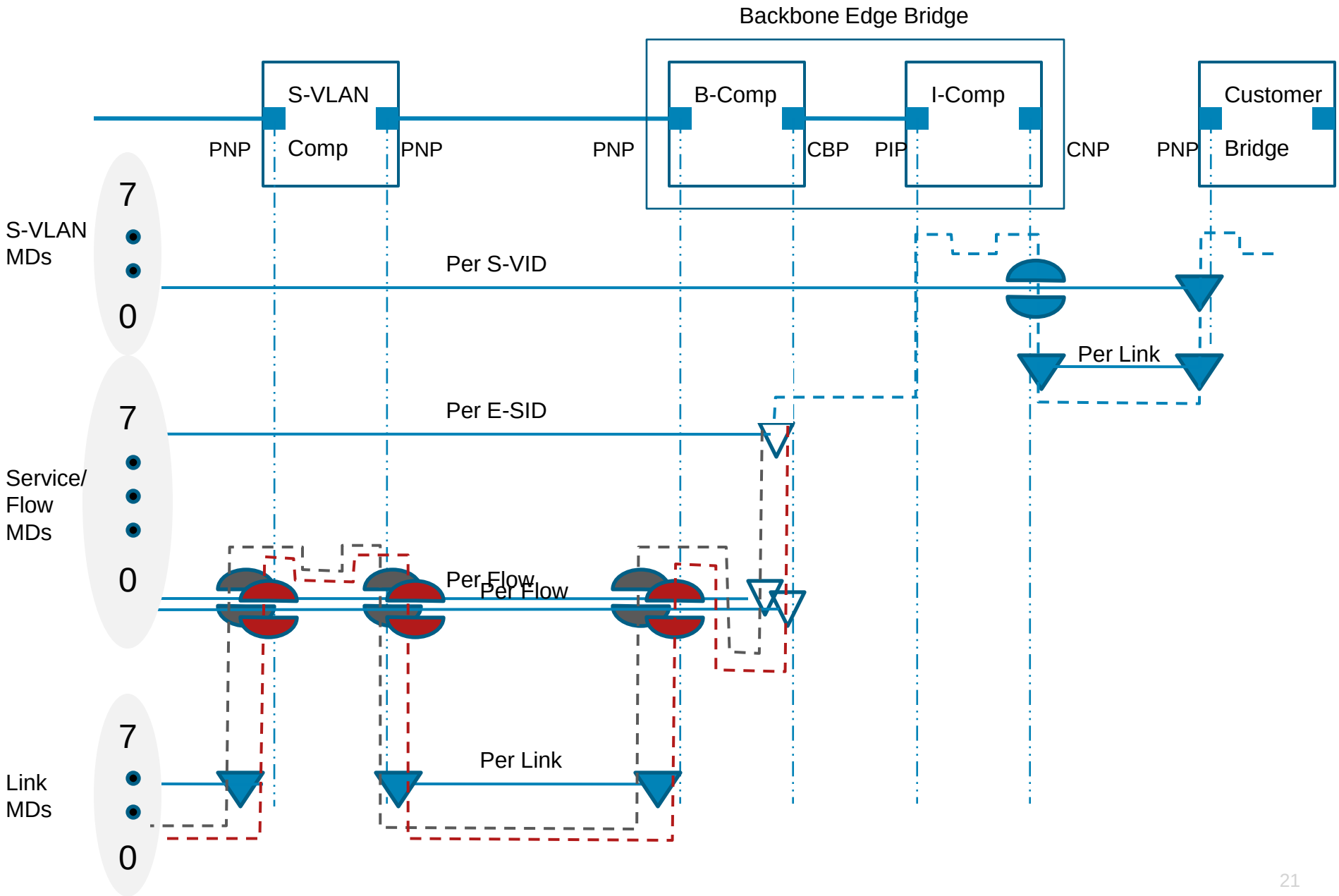
Service OAM (proactive)

- 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)

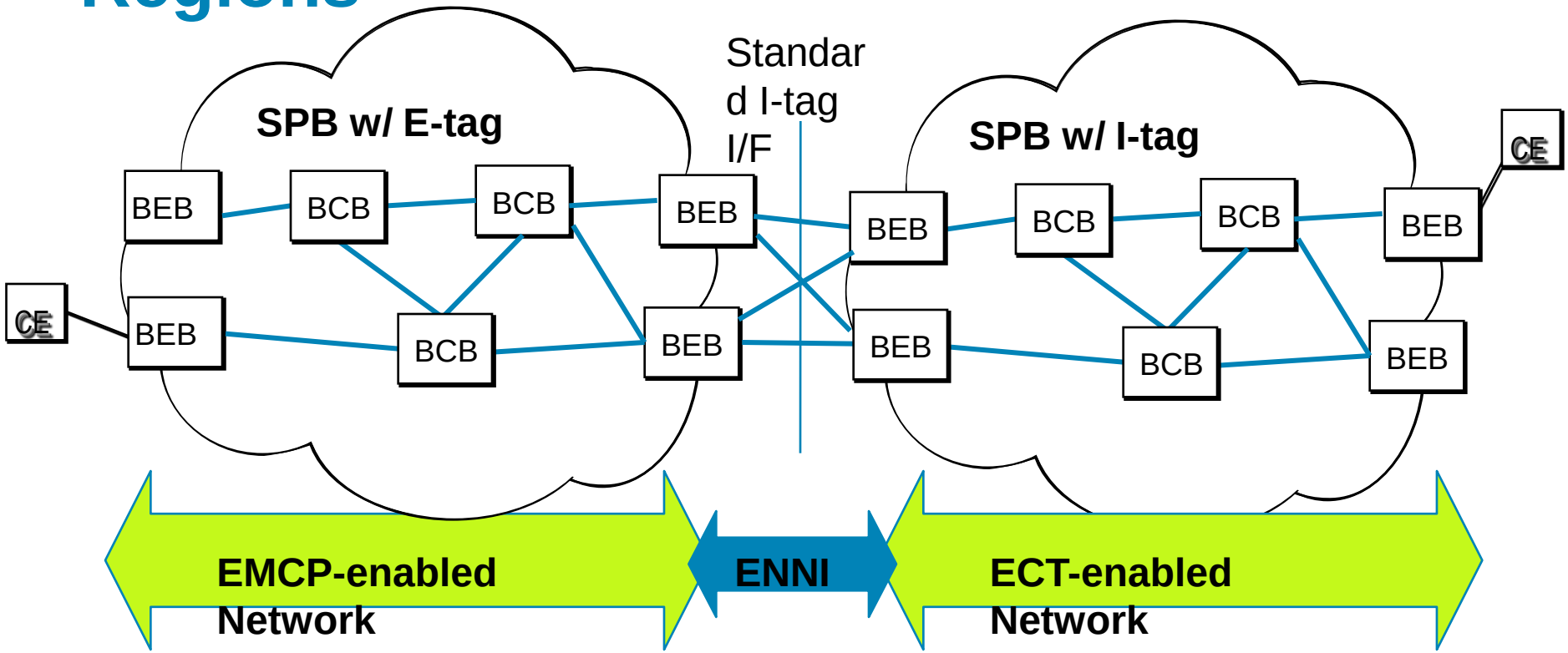
- Network OAM is a degenerate case of service OAM where a single default E-SID can be configured on all BEBs and the CFM is performed for that default E-SID just as described above for service-level OAM
  - This default E-SID is per B-VID – e.g., per ECMP algorithm. If there are multiple ECMP algorithms 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

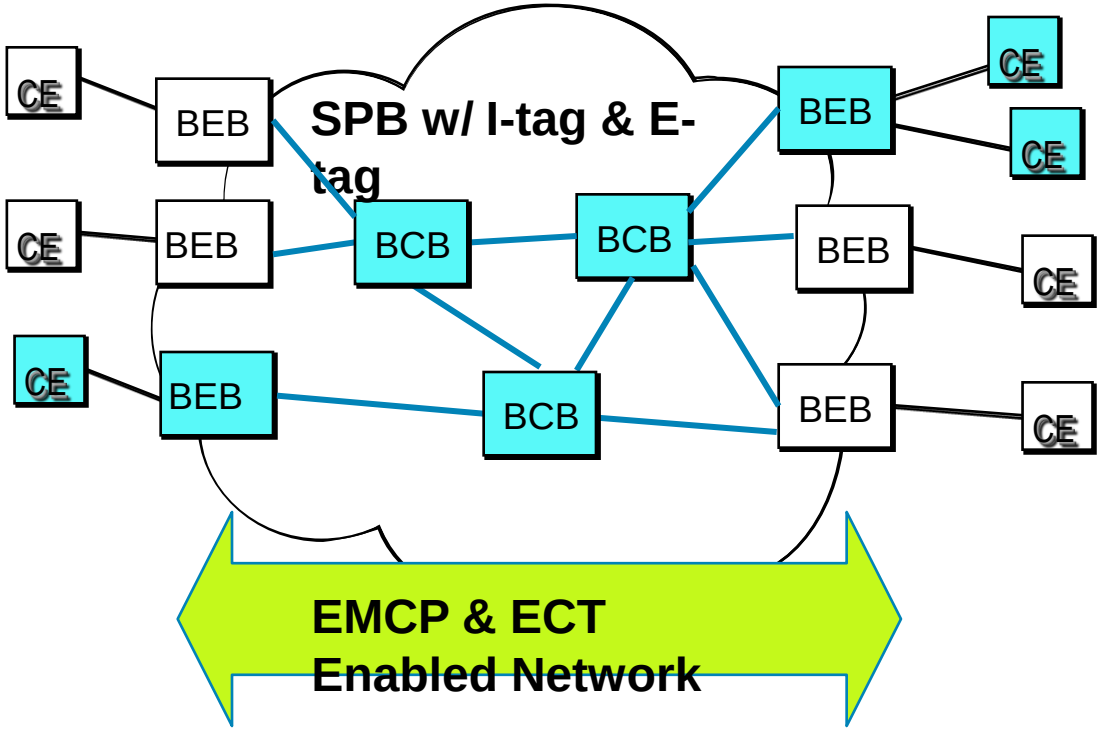
Appendix – A Interoperability

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

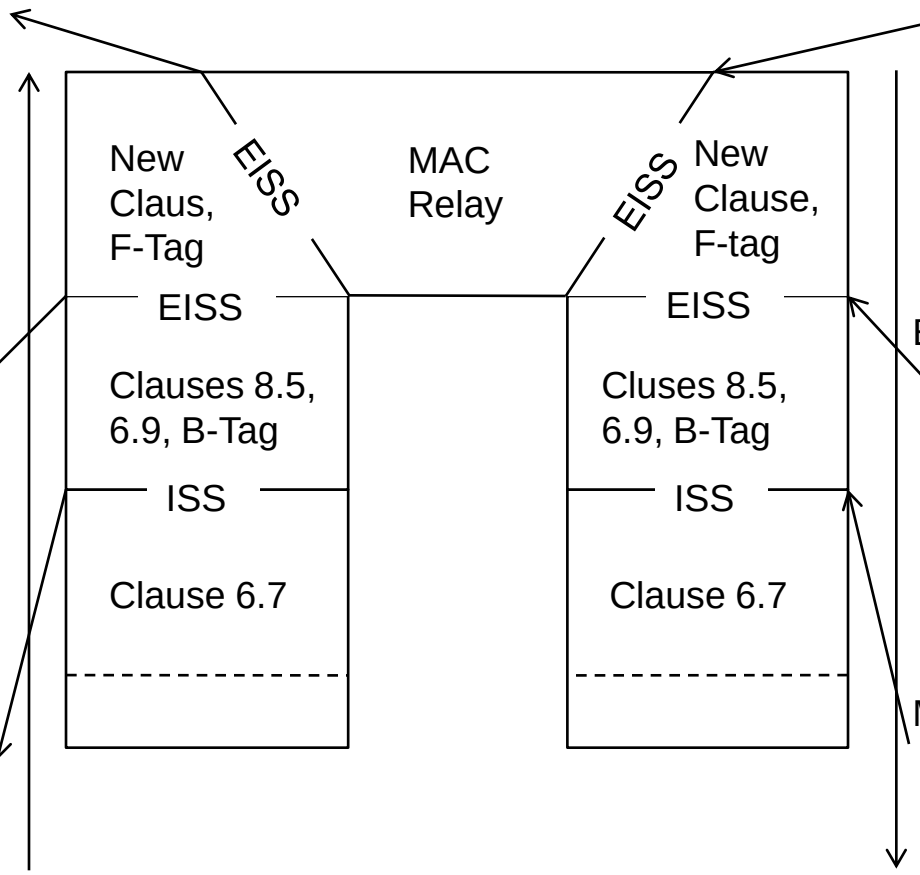
- Any per-hop ECMP (whether TTL is used or not) requires additional new processing anyway:
  - Hashing based on user data flow headers to determine egress interface or
  - using the pre-calculated hash-index to determine egress interface
- A network can be configured to simultaneously support ECMP and 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

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.)

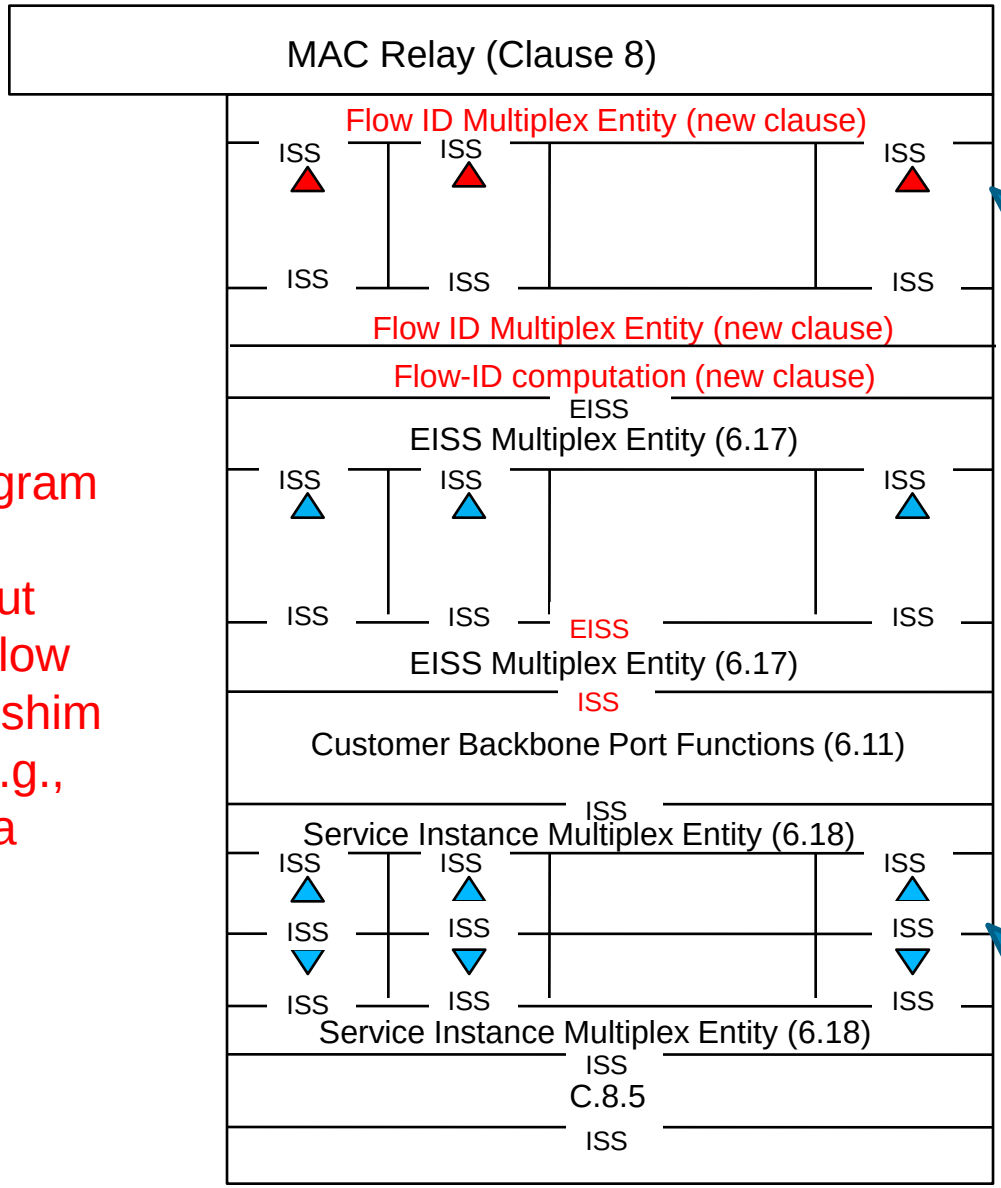


EM_UNITDATA.request (
- destination_address,
- source_address,
- mac_service_DU,
- priority,
- drop_eligible,
- vlan_identifier,
- flow_identifier,
- etc.)

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.)

Baggy Pants Model for OAM operation at BEB



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

Optional because B-tag is optional

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.


CISCO