ECMP Operation, Bridge Model, and OAM

Ali Sajassi

May 9, 2011
IEEE 802.1 Interim Meeting

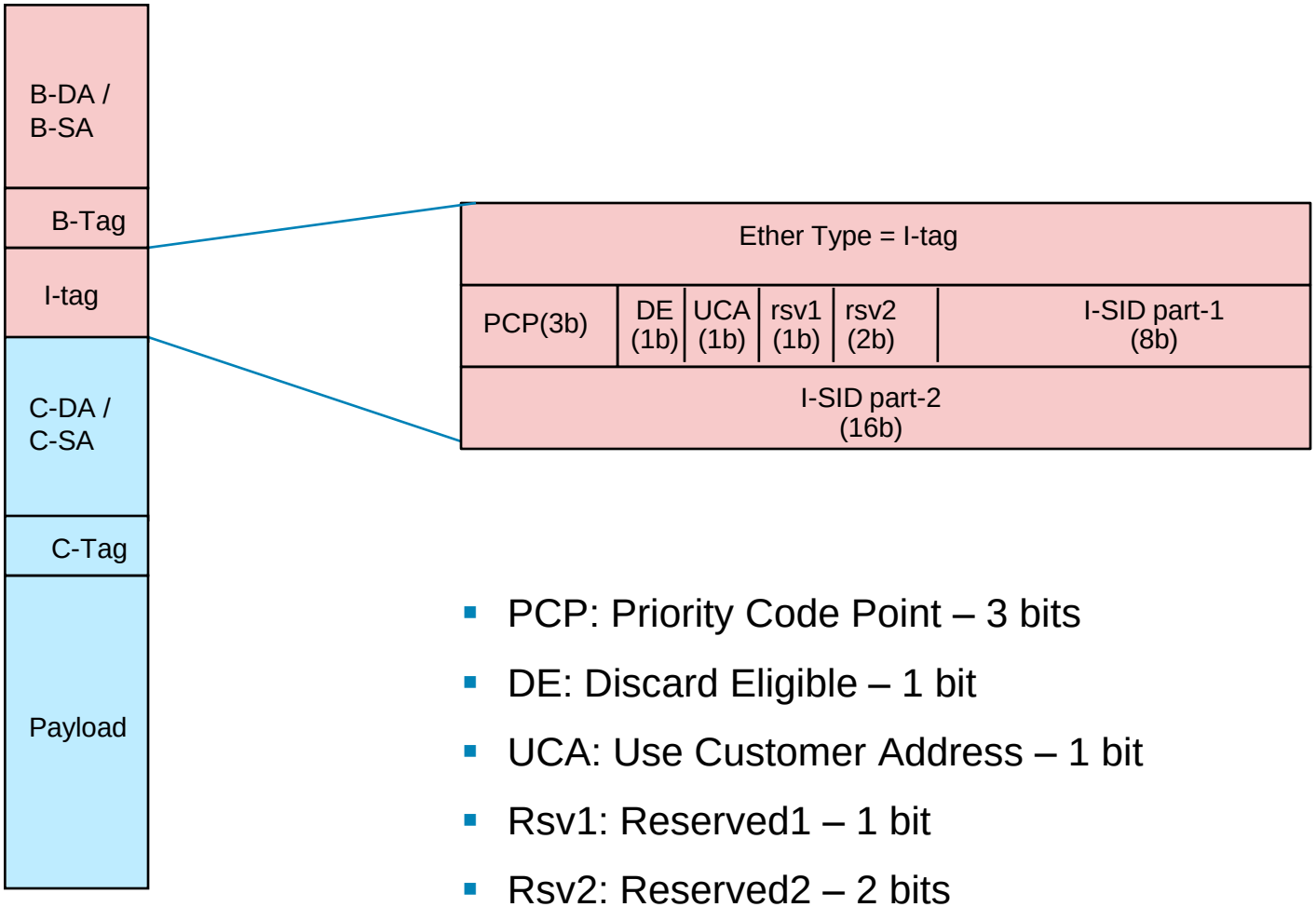
Agenda

- Frame Format
- General Operation
- Bridge Model
- OAM Operation
- Interoperability

Requirements

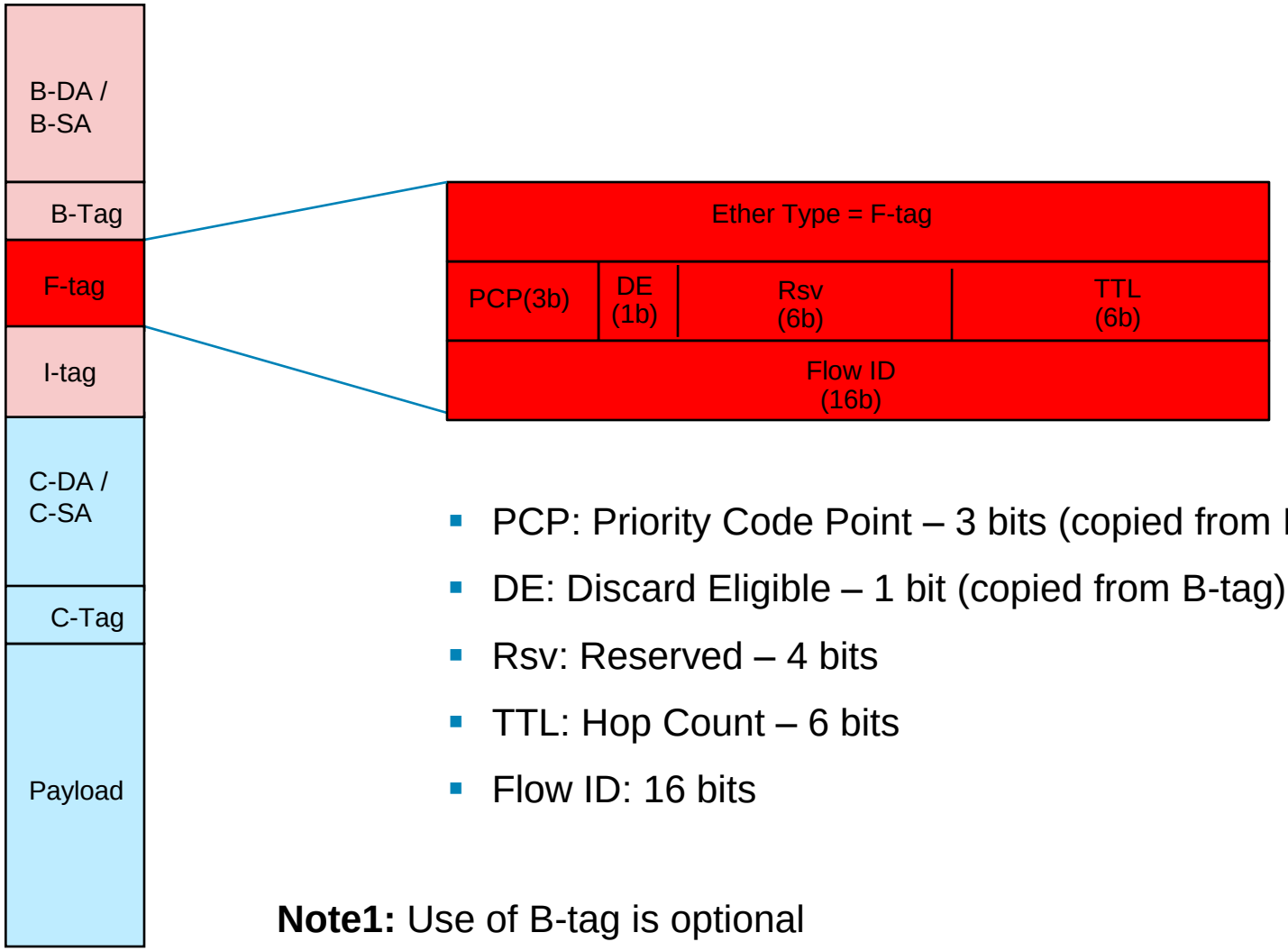
- Support of per-hop ECMP
- Support of TTL for loop mitigation
- Support of flow-id
  - To avoid deep packet inspection in the core
  - To provide proactive service-level monitoring
- Flexible n-tuple hash algorithm for flow-identification
  - Any edge node can choose any set of n-tuples and any hash algorithm to derive a flow id
- Support proactive service-level monitoring
  - For a given flow-id, the path for that flow through the network be 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 – 4 bits
- TTL: Hop Count – 6 bits
- Flow ID: 16 bits

Note1: Use of B-tag is optional

Note2: If needed, OAM bit and sub-type can be incorporated in reserved field

Pros & Cons

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 for future application (if need arises)
- 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 802.1Qbp (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 (802.1Qbp 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.

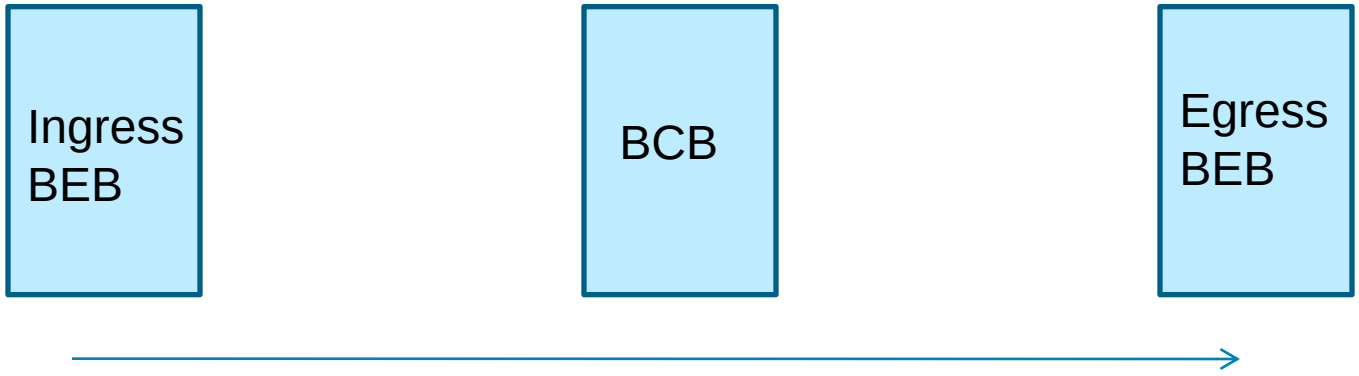
Cons

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

Agenda

- Frame Format
- General Operation
- Bridge Model
- OAM Operation
- Interoperability

Operation



- Compute flow-id based on n-tuple
- Add F-tag (flow-id & TTL) before I-tag
- Add B-tag if needed
- 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

Operation – Cont.

- For ECMP operation, use of F-tag is mandatory for both unicast and multicast frames
  - ➢ When F-tag is used for multicast frames, the TTL field of it is used for the purpose of loop mitigation –flow-id is not used because ECMP only applies to multicast frames
- Use of B-tag is optional for ECMP frames

Note-1: In most scenarios where a single IS-IS topology is used, then B-tag can be optional

Note-2: In special scenario where legacy protocols such as MVRP is run simultaneously in the network but 802.1Qbp bridges don't run them (e.g., they flood MVRP frames), then B-tag must be used with ECMP frames

Mixed Mode of Operation

- A network can simultaneously run a mix of .1ad, .1ah, .1aq, .1ay, and.1bp
- As always, B-VID is used to identify which frames belong to which mode of operation since each mode has its own set of dedicated B-VID(s)
- For .1ad, .1ah, .1aq, and .1ay, unicast data frames are tagged. Therefore, if an untagged unicast frame is received w/o F-tag, it shall be discarded
- For untagged 802.1Qbp operation, since default B-VID is used, we need to ensure that there is no ambiguity with other mode of operations
  - ➢ Legacy control frames such as BPDUs and MVRP are sent untagged and thus these received frames are subjected to default B-VID; however, these frames are multicast frames using reserved addresses and they get filtered and process accordingly
  - ➢ If MVRP is received by a 802.1Qbp bridge but it is flooded instead of processed, then VLAN filtering operation on the bridge ports may be impacted and in such scenarios, ECMP frames should be sent with valid B-tag.

I-SIDs to B-VID mapping

- 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 both share the same ECT tree
- For ECMP operation, unicast and multicast frames need to use different trees and load sharing algorithms
  - Unicast frames need to use ECMP algorithm
  - Multicast frames need to use ECT algorithm
- For ECMP operation, use of B-tag is optional
- Possible options for I-SIDs to B-VID mapping:
  a) Use multiple B-VIDs to designate different ECTs as before for multicast frames and use the same set of B-VIDs for unicast ECMP frames
  b) Use a single B-VID to designate different ECT algorithms for multicast frames and use this B-VID for unicast ECMP frames
  c) Use multiple B-VIDs for multicast frames and use a single B-VID for unicast frames

Option A) Multiple B-VIDs for Both

- In this option different B-VID represent different ECT algorithm but all these B-VIDs are mapped to the same ECMP algorithm
- All B-VIDs are mapped to the same bridge domain (VLAN) in the B domain
- Pros
  - No modification to either control or data planes are needed
- Cons
  - We lose the option of not using B-tag for unicast ECMP because for a given I-SID, we need to use the same B-tag for unicast data as for multicast data

Option B)

- In this option a single B-VID is used to represent different ECT algorithms for multicast frames as well as the ECMP algorithm for unicast frames
- Pros
  - Use of B-tag can be optional for both unicast and multicast data
  - No additional changes in data plane is required
- Cons
  - It requires changes to the control plane – e.g., it requires decoupling of B-VID to ECT algorithm in SPBM I-SID sub TLV so that I-SIDs can be directly associated with ECT algorithm

NOTE: Explicit versus automatic I-SID to ECT association

Option C)

- In this option multiple B-VIDs are used to designate different ECT algorithms for multicast frames but a single B-VID is used to designate ECMP for unicast frames
- Pros
  - No changes to control plane is required
  - Use of B-tag can be optional for unicast frames
- Cons
  - B-tag must always be used for multicast frames
  - It requires changes to data-plane to associate two B-VIDs for the same I-SID (one for unicast and the other for multicast)

Recommendation

- It seems like option (b) is a nice compromise because:
  - It enables us with all the functionality that we need including the option of efficient frame encapsulation w/o B-tag for both unicast and multicast frames
  - The price for it is low. It requires simple modification to control plane. More precisely it requires the following changes:

    SPBM ISID-ADDR TLV should be modified to replace base VID with ECT Algorithm ID so that I-SIDs are directly associated with the algorithm ID

Association of I-SIDs to a ECT Algorithm (indirectly via Base VID)

		Octet	Length
	Type (TBD)	1	1
	Length	2	1
	B-MAC Address	3-8	6
	reserved	9	4 bits
	Base VID	9-10	12 bits
I-SID Tuple 1	T	11	1 bit
	R	11	1 bit
	reserved	11	6 bits
	I-SID	12-14	3
	...		
I-SID Tuple n	T	(4n+7)	1 bit
	R	(4n+7)	1 bit
	reserved	(4n+7)	6 bits
	I-SID	(4n+8)-(4n+10)	3

Figure 28-11—SPBM Service Identifier and Unicast Address sub-TLV

ECT Algorithm sub-TLV

	Octet	Length
Type (TBD)	1	1
Length	2	1
ECT ALGORITHM	3-6	4
ECT Information	7-(Length+2)	variable

a) Type (8-bit) Value TBD

b) Length (8-bits)

Total number of bytes contained in the value field.

c) ECT-ALGORITHM (4-bytes)

ECT-ALGORITHM is advertised when the bridge supports a given ECT-ALGORITHM (by OUI/Index) on a given VID.

d) ECT Information (variable)

ECT-ALGORITHM Information of variable length.

Mapping between ECT Algorithm & Base VID

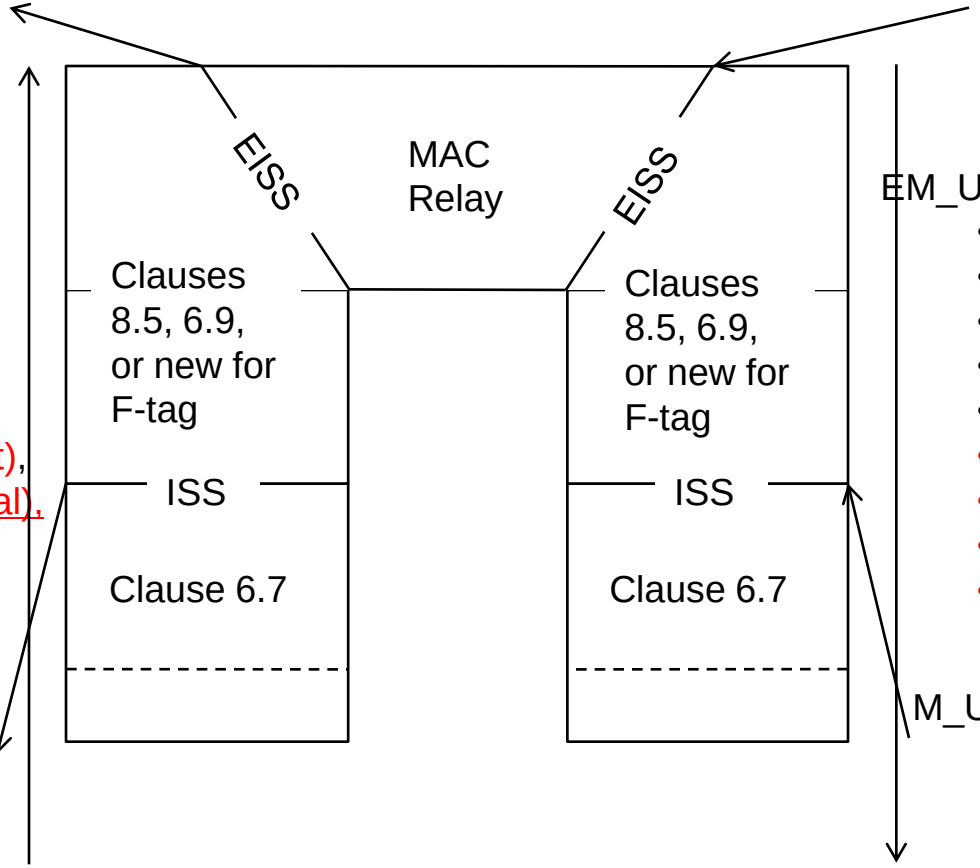
		Octet	Length
ECT-VID Tuple 1	Type (TBD)	1	1
	Length (6n)	2	1
	ECT Algorithm	3-6	4
	Base VID	7-8	12 bits
	U	8	1 bit
	M	8	1 bit
	reserved	8	2 bits
	...		
ECT-VID Tuple n	ECT Algorithm	(6n-3)-6n	4
	Base VID	(6n+1)-(6n+2)	12 bits
	U	6n+2	1 bit
	M	6n+2	1 bit
	reserved	6n+2	2 bits

Figure 28-5—SPB Base VLAN-Identifiers sub-TLV

Agenda

- Frame Format
- General Operation
- Bridge Model
- OAM Operation
- Interoperability

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 (optional),
- TTL (optional)
- 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, (optional)
- TTL (optional)
- etc.)

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

Add New or Modify Sub-Clauses 6.9 & 6.10

- If the tag is F-tag, then extract TTL and flow_identifier from the F-tag 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

Sub-Clause – Cont.

- Flow-id is 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

Advantages of Homogenous Hash()

- Enables proactive service-level monitoring

  Without homogenous hash(), only flow-level monitoring can be performed – similar to what is already done in MPLS and IP networks

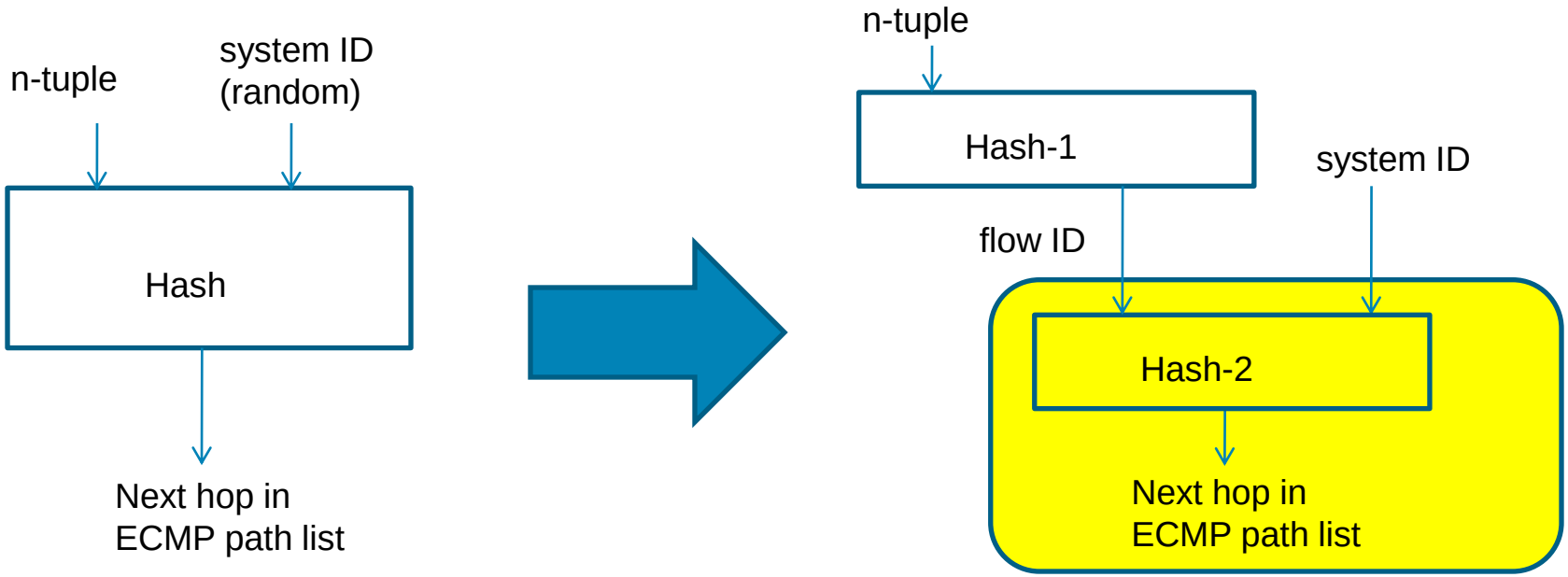
- Validates performance of ECMP function – e.g., traffic is equally distributed among ECMPs

  Without homogenous hash(), it is not possible to differentiate between a failure scenario and skewed hash() by a node

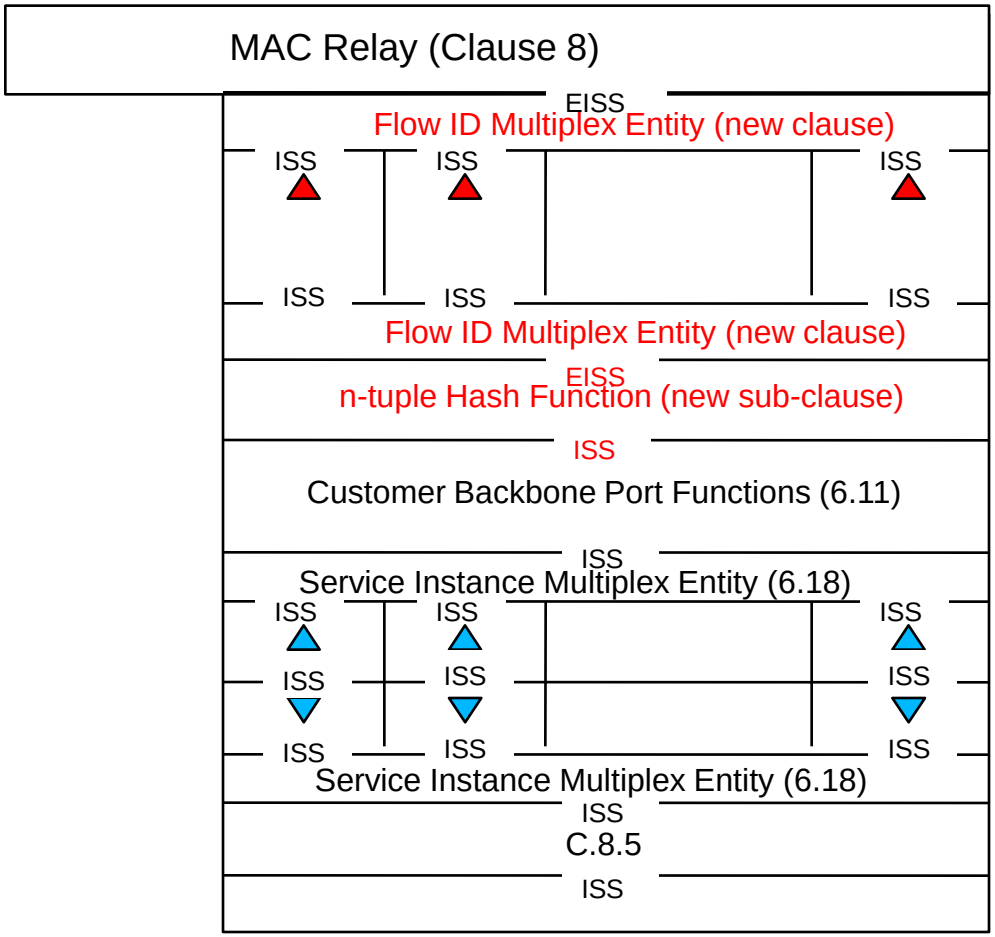
Breaking Hash() into two Parts

- 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 is lot easier than mandating part-I to be homogenous)



Baggy Pants Model for OAM operation at BEB

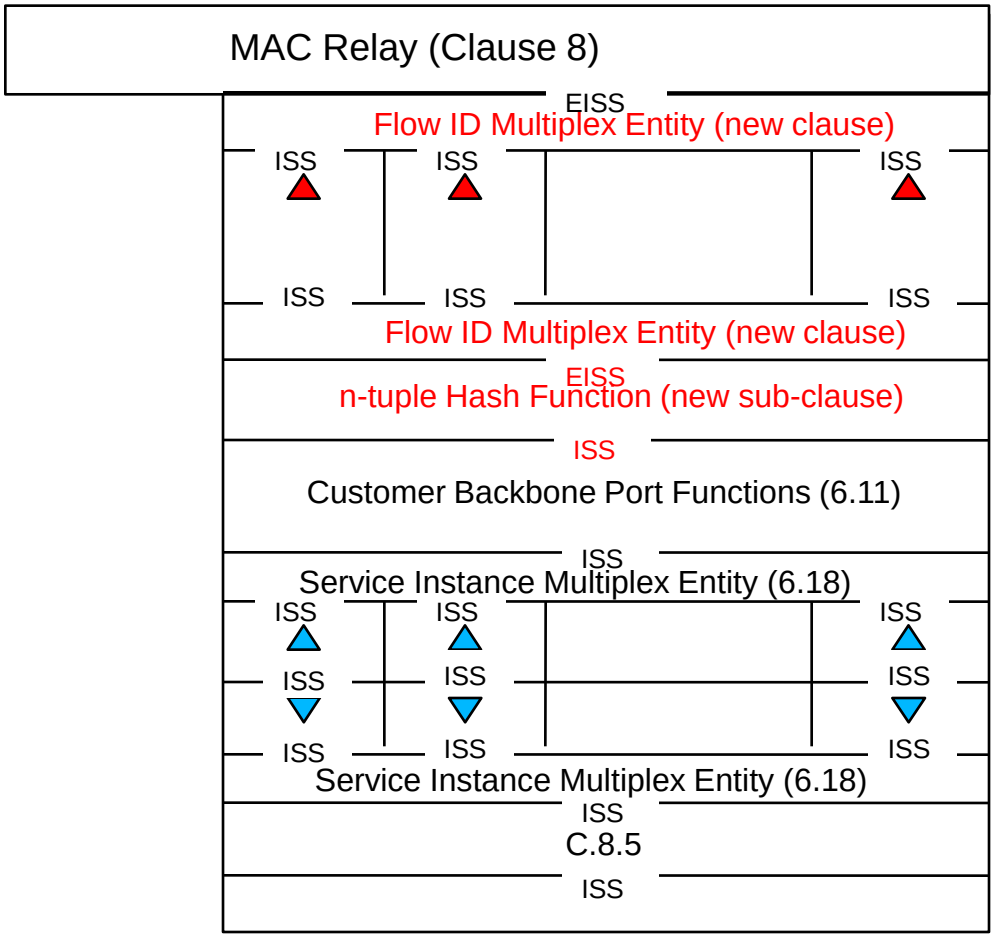


NOTE: Clause 6.11 needs to be modified to indicate that all ECMP I-SIDs are mapped to a single default B-VID

Agenda

- Frame Format
- Bridge Model & Operation
- OAM Operation
- Interoperability

Baggy Pants Model for OAM operation at BEB

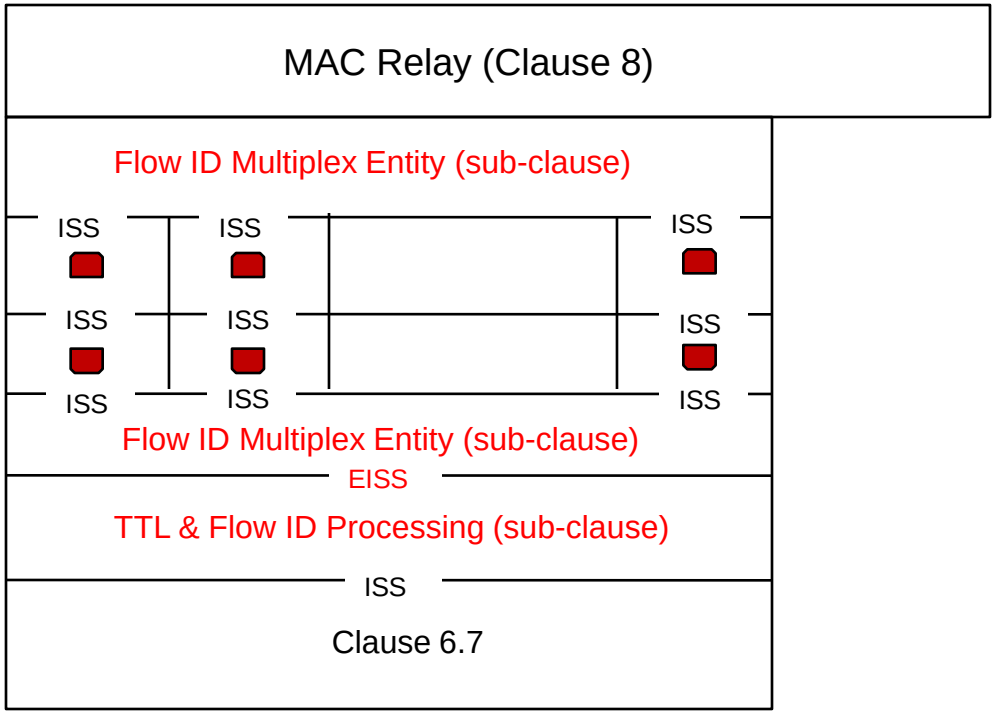


NOTE: Clause 6.11 needs to be modified to indicate that all ECMP I-SIDs are mapped to a single default B-VID

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

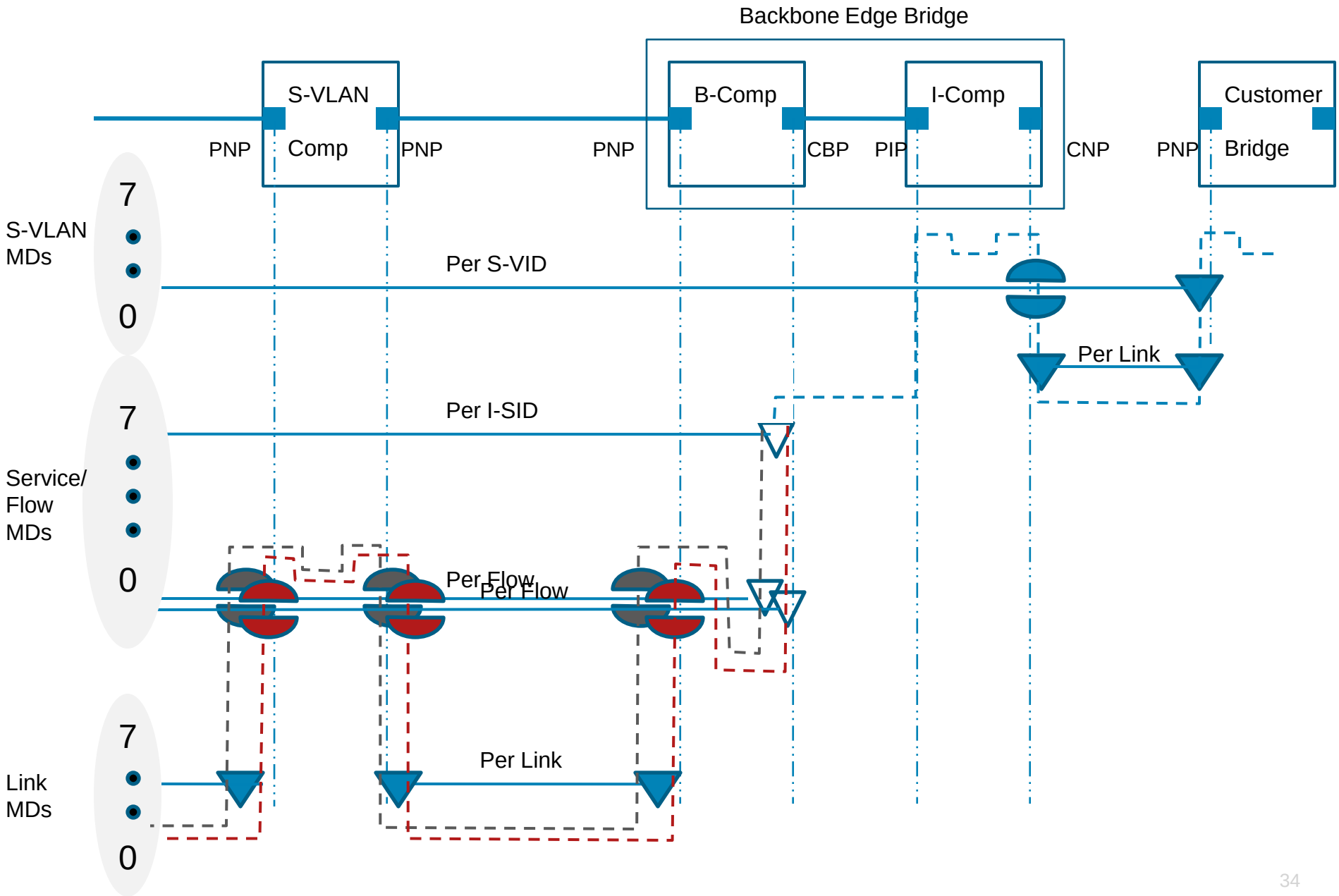
Service OAM (proactive)

- 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

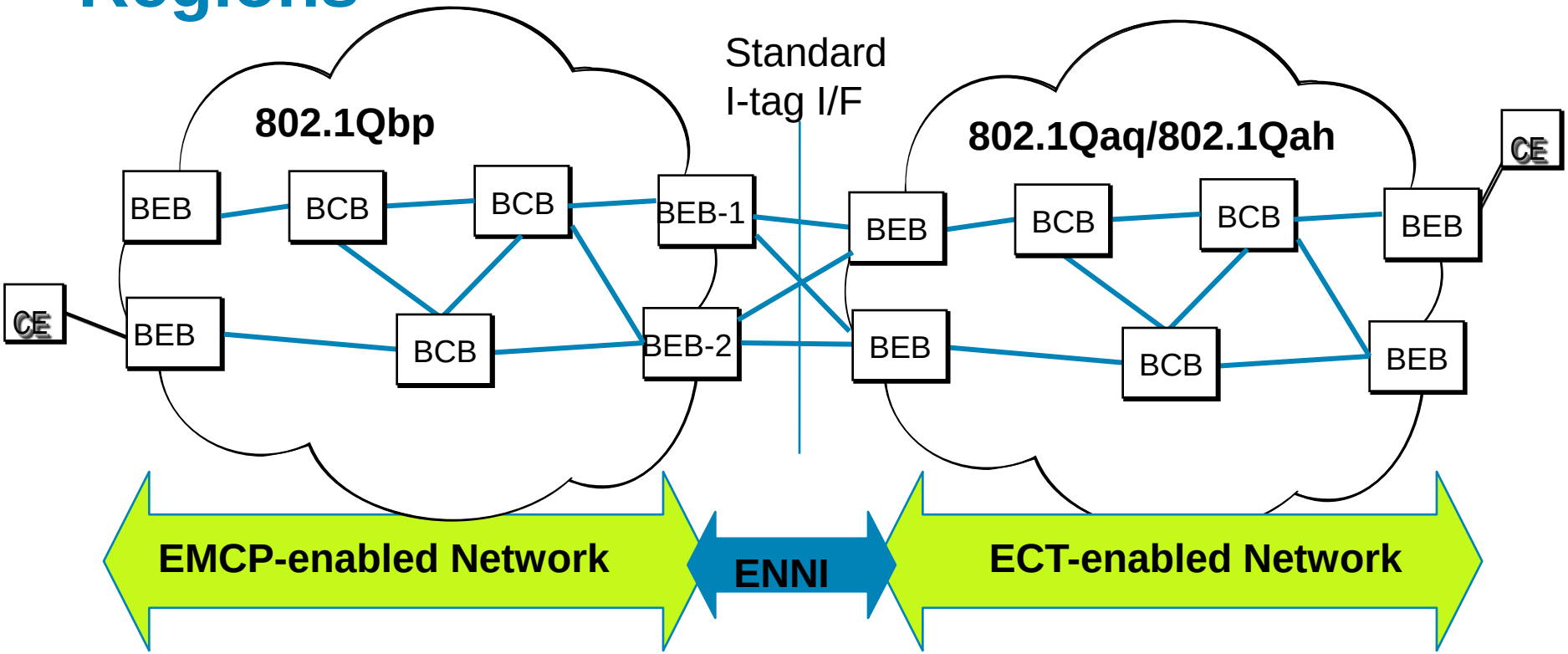
CFM Flow


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 I-SID
Per Flow
Per Flow
Link
MDs
7
0
Per Link

Agenda

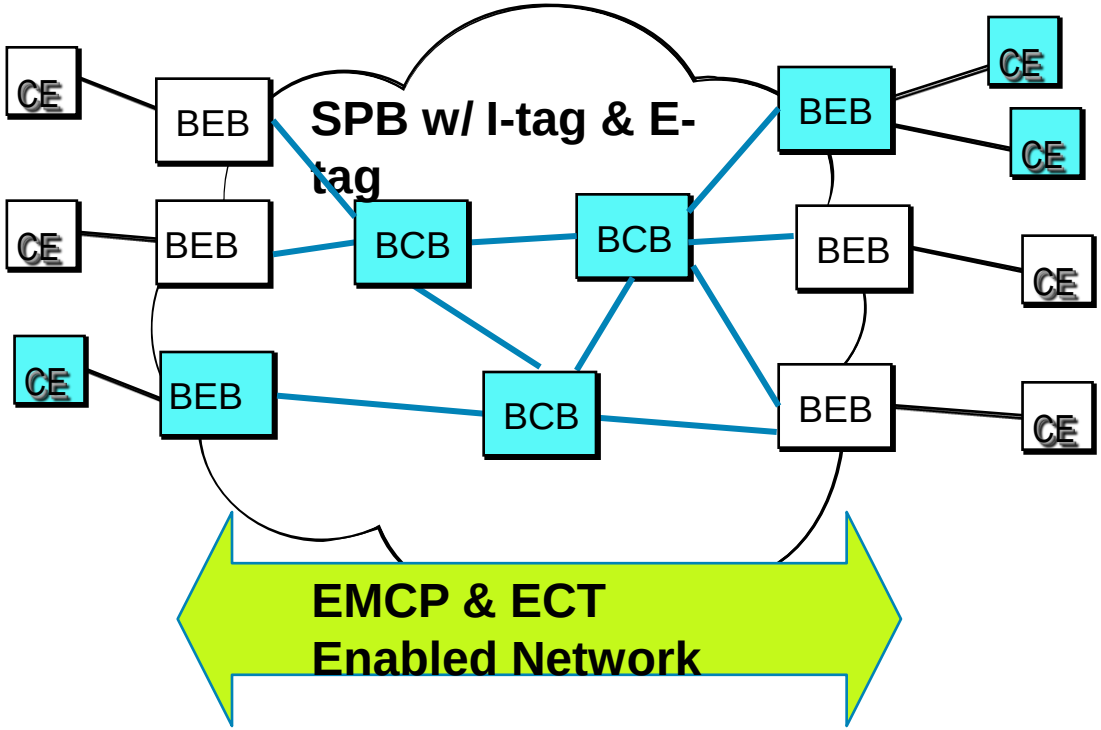
- Frame Format
- Bridge Model & Operation
- OAM Operation
- Interoperability

Inter-operability: Between Different Regions



- Tx: BEBs in the ECMP-enabled network (e.g., BEB-1 and BEB-2) will strip F-tag and pass and I-tag frame to the other region. If any I-SID translation is required, then it is done per clause 6.11
- Rx: BEBs in the ECMP-enabled network upon receiving an I-tag frame, check to see if it is ECMP-enabled. If so, then perform hashing function and add an F-tag to the frame
- No changes is needed on the ECT-enabled network (both BEBs and BCBs)

Inter-operability: Within one Region



- Use default B-VID to identify ECMP frames just like
  - B-VIDs used for 802.1aq (one per ECT)
  - B-VIDs used for PBB-TE
  - B-VIDs used for PBB with MSTP
- To support ECMP
  - ECMP-enabled bridges form a sub-graph using IS-IS and exchange F-tag frames only among themselves
  - Non-ECMP-enabled bridges will never receive F-tag frames

Backward Compatibility

- 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, Interoperability is easily provided using I-tag service interface.
- IS-IS can support both ECMP and non-ECMP BCBs in the same network and ensure gradual migration

Appendix – A

AI

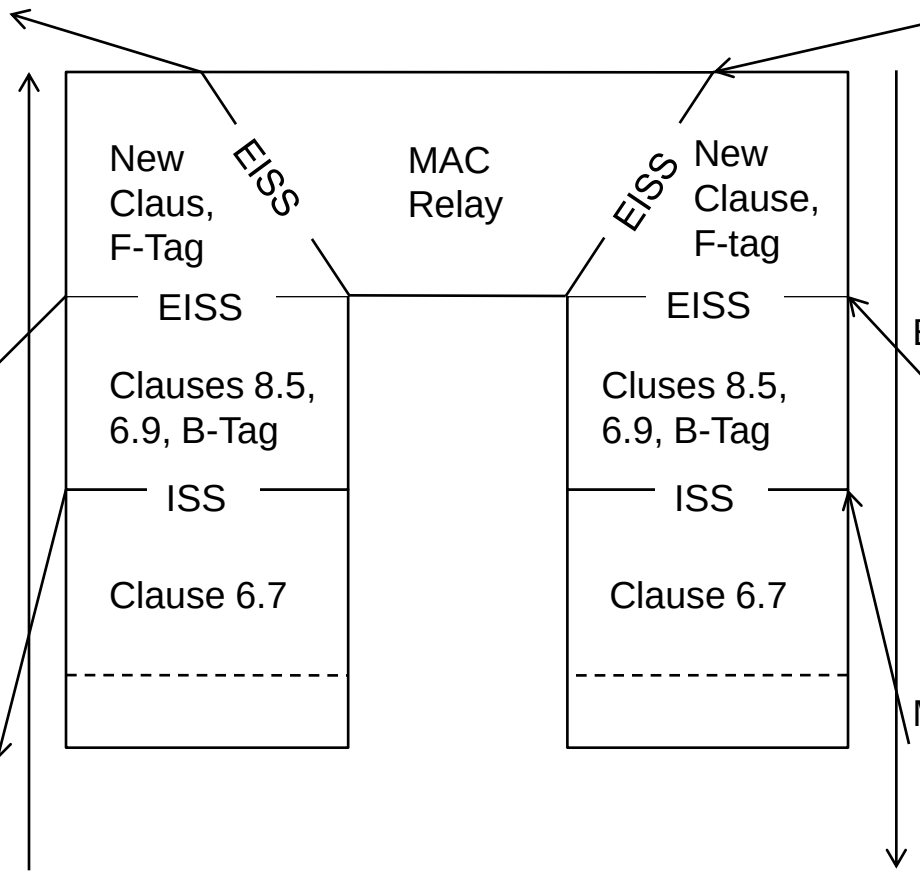
- How to make the adjacency list consistent across all nodes ?
- OAM for proactive service-level monitoring needs to be expanded with step by step procedures for fault detection & verification

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.)

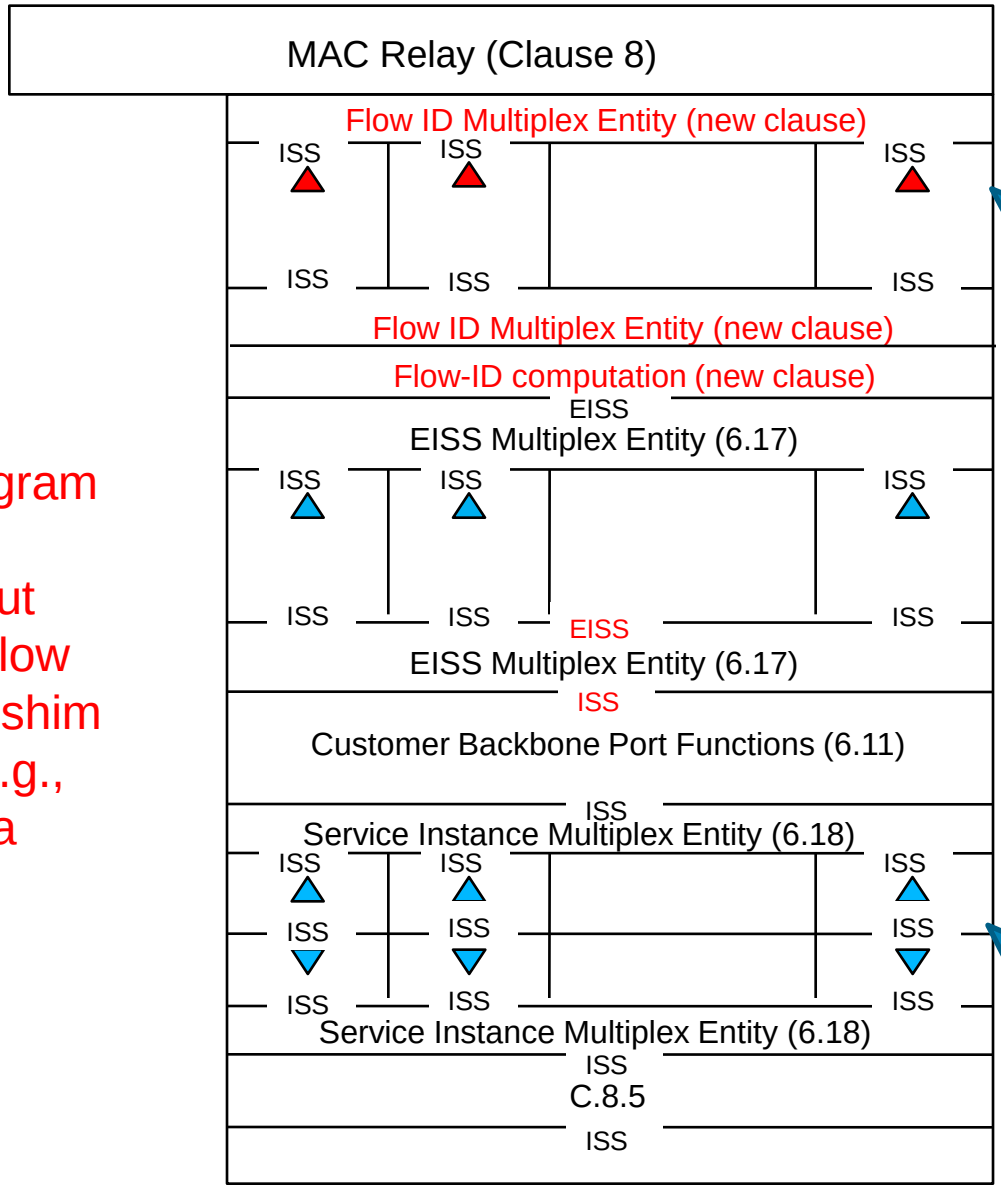


EM_UNITDATA.request (
- destination_address,
- source_address,
- mac_service_DU,
- priority,
- drop_eligible,
- vlan_identifier,
- flow_identifier,
- etc.)

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.

Pros & Cons for Single B-VID

- Pros:
  - A single VID to identify both unicast and mcast traffic consistent w/ other IEEE projects
  - changes are limited to control plane modifications
  - using default B-VID, it allows for mcast traffic to be sent w/o B-tag (I like this one :-)
- Cons:
  - A single B-VID is used to represent multiple ECTs inconsistent w/ 802.1aq
  - I-SID to algorithm mapping will be implicit (and thus may weaken the check)
  - There are now two different ways to specify ECTs

Pros & Cons for Multiple B-VIDs

- Pros:
  - Handling of ECTs and mapping to B-VIDs are per 802.1aq
  - No modifications to IS-IS and existing TLVs
- Cons:
  - Requires data-plane changes (e.g., modification to I-SID/B-VID mapping table of 6.11)
  - Uses different B-VIDs for unicast & mcast that doesn't have precedence in 802.1


CISCO