Congestion Notification in Pre-defined Network Domains

Mehmet Toy, Ph.D

February, 2014

COMCAST

Problem Statement

- Ethernet is the dominant technology for access networks and becoming an alternative technology for backbone networks as well.
- Ethernet Virtual Connection (EVC) between two or more User-Network Interfaces (UNIs) serves as a transport path for service frames of MEF Metro Ethernet Services.
- Applications such as voice and video riding over Ethernet use MEF services employing EVCs. These applications may use TCP/IP or other L3 and L4 protocols.
- Bursty nature of some applications such as video and internet access and/or possible overbooking of network resources is likely to drive network consisting of EVCs into a temporary or sustain congestion state.
- Congestion Control in layer 2 and layer 3 networks have been extensively studied. However, none of them addresses congestion in EVCs.
- A congestion notification technique for EVCs to inform traffic source to slow down before the EVC becomes congested is necessary.

DE Bit for Congestion Notification

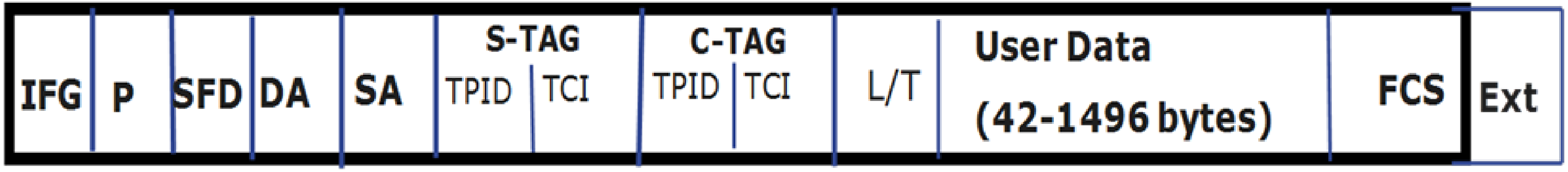
- EVC may cross one or multiple operator networks providing end-to-end connectivity between users or between a user and application servers.
- DE bits in S-Tag of service frames can be used to indicate congestion in EVC end-to-end or in a pre-defined domain on the EVC path.
- When the congestion control is end-to-end, the traffic source causing congestion and destination experiencing congestion are EVC termination points. When the congestion control is between domain boundaries, the source causing congestion and the destination experiencing congestion may not be EVC termination points, but intermediate points on the EVC path. In all cases, we will call them domain boundaries, which will be defined in another contribution.
- When the EVC resources on the destination domain boundary node is congested, the node sets DE bits of service frames belongs to the EVC to 1 and sends the frame to upstream node on the EVC to indicate it is congested.

DE Bit for Congestion Notification (Cont.)

- **The upstream node transmits the frames with DE=1 to the next upstream node without changing DE bit and does not change its rate for downstream, unless it is one of the domain boundary nodes. If the receiving node is a boundary node, it reduces its rate for downstream, processes the received user frames, and originates new user frames with DE=0.**

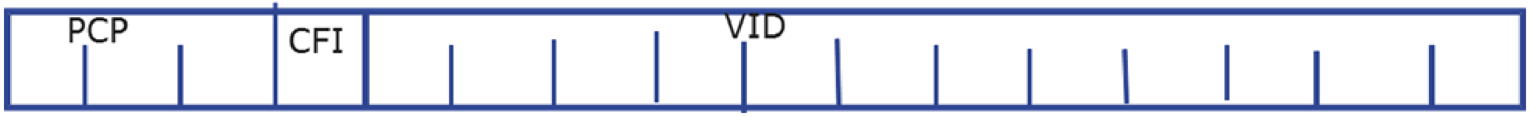
Ethernet Frames

IEEE 802.3-2005 Frame Format

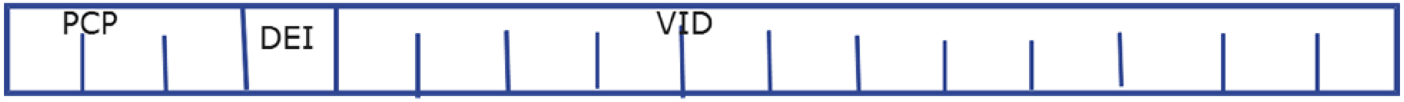


IFG	P	SFD	DA	SA	S-TAG TPID TCI	C-TAG TPID TCI	L/T	User Data (42-1496 bytes)	FCS	Ext

C-TAG Format



S-TAG Format



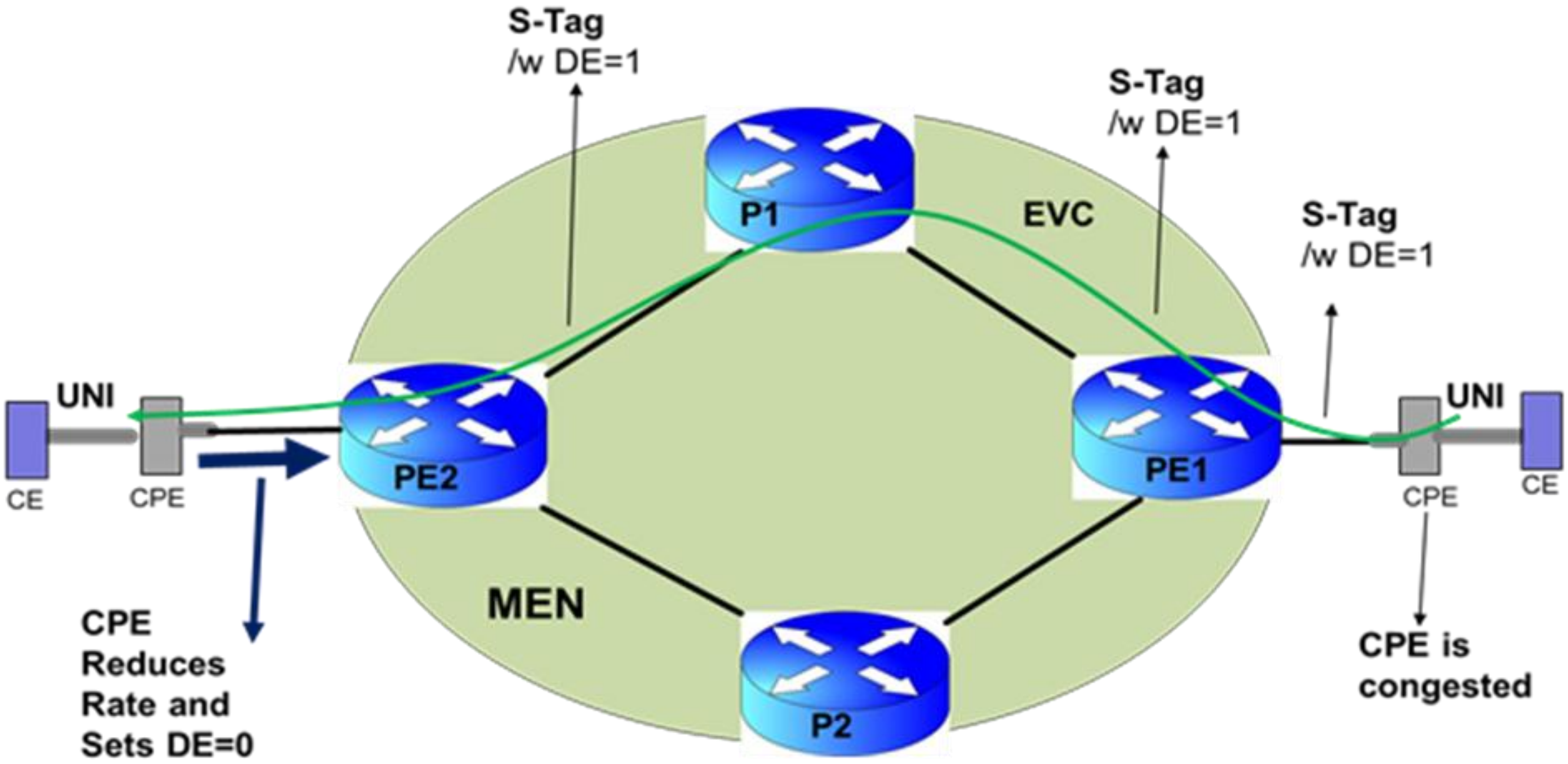
PCP-Priority Code Point, 3 bits

CFI-Canonical Format Indicator , 1 bit

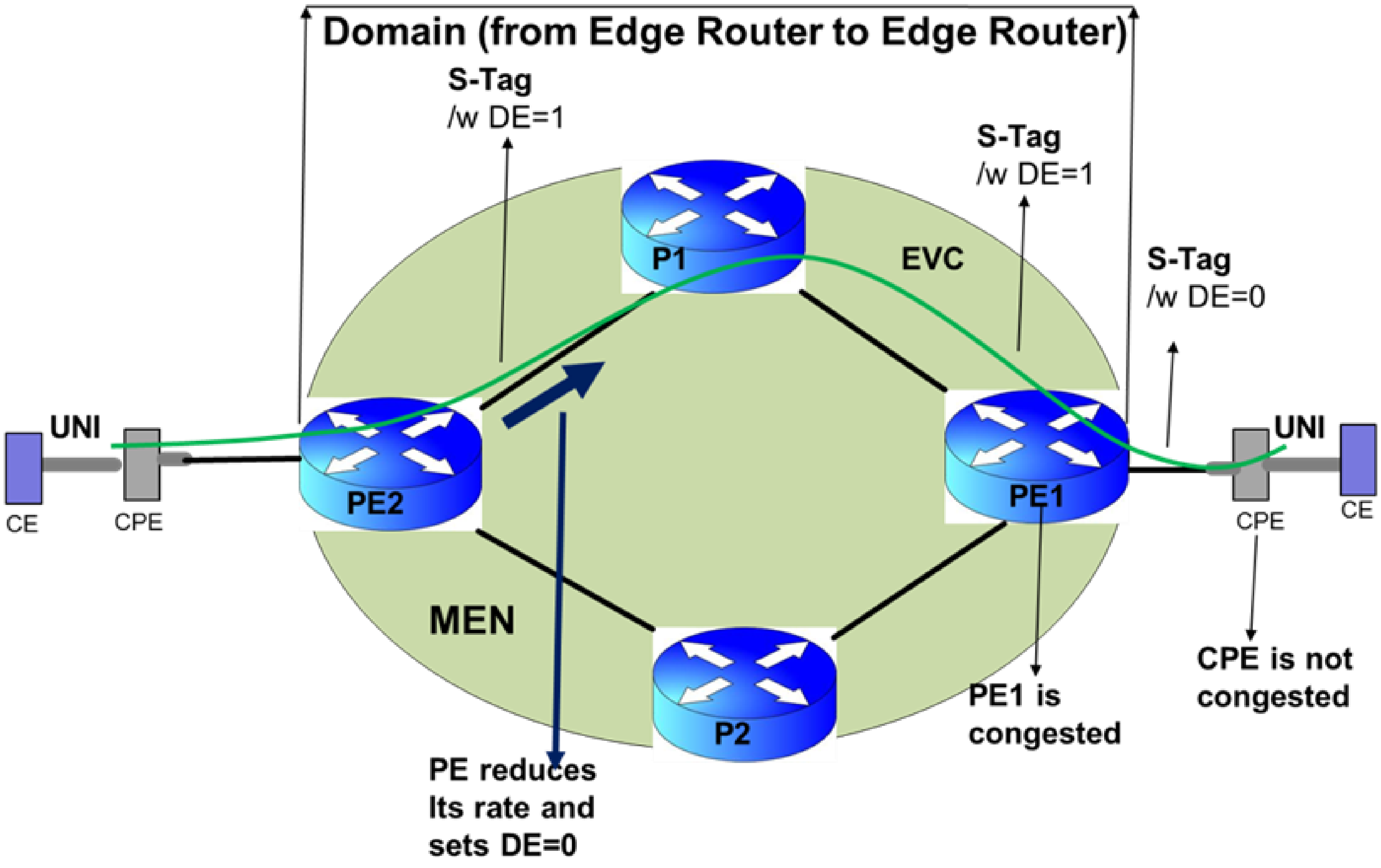
VID-VLAN Identifier, 12 bits (0 -4094)

DEI-Drop Eligibility Bit

Example I: Congestion Control with DE Bit Between CPEs



Example II: Congestion Control Between PEs Using DE Bit



DE Bit Usage

- DE bit can be used to indicate congestion in a port, a CoS flow and a set of CoS flows that are represented by an EVC, where CoS and frame color are represented by DSCP or PCP.
- DE bit can be used between CPEs, between NEs that are at the edge of the network, and even between NEs in the backbone, as long as these devices are able to look at S-Tag within Ethernet frames and interpret DE bit as an indicator for congestion state
- DE bit may be applied to DOMAINs using MD levels (via filtering or manually configuring boundary nodes of domain).
- Since the notification method is independent from higher layer protocols, it is expected to work with any L3 and L4 protocols such as TCP, UDP and IP.
- DE bit can be used in shared links such as EPON.
- Service providers can choose to implement their own algorithms for rate regulations within domain boundaries in conjunction with this notification technique.

NE Behavior for Congestion Control at Port Level

- When a frame with DE=1 due to congestion from downstream is received by the boundary node on its port (i.e. Case a) or the port (i.e. access port or network port) buffer exceeds a pre-set threshold (T_B) (i.e. Case b) or combination of both (i.e. Case c), then the port is in congested state.
- In case (a), although the domain boundary port itself is not congested, the port must be in congested state and activate congestion control algorithm to reduce its rate in the reverse direction for the EVC. Although the EVC is bidirectional, it may not make sense to reduce rate in both directions, unless frames of both directions do share the same buffer.
- In case (b), the boundary node port itself is congested, therefore, the device activates its congestion control algorithm to reduce its rate in both directions for the EVC, and set DE=0.
- For case (c), the boundary node behavior will be the same as that for case (b).

NE Behavior for Congestion Control at EVC Level

Congestion control can be per EVC if multiple EVCs are multiplexed in a boundary port (i.e. UNI):

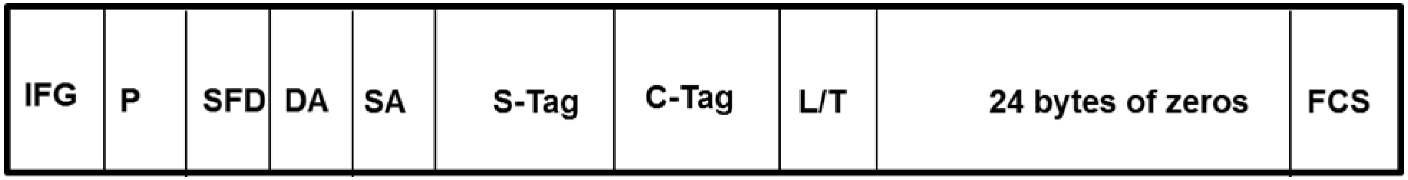
- **When a frame with DE=1 for a given EVC from downstream is received by the boundary node port (i.e. Case a) or the boundary node's EVC buffer exceeds a pre-set threshold (t_B) (i.e. Case b) or combination of both (i.e. Case c), then the EVC is congested.**
- **In case (a), although the boundary node EVC buffer is not congested, the device EVC buffer must be in congested state and activate congestion control algorithm to reduce its rate in the reverse direction for the given EVC. Although EVC is bidirectional, it may not make sense to reduce rate in both directions, unless frames of both directions do share the same buffer.**
- **In case (b), the boundary node EVC buffer is congested, therefore, the device activates its congestion control algorithm to reduce its rate in both directions for the EVC transported on that port, and set DE=0.**
- **For case (c), the boundary node behavior will be the same as that for case (b).**

NE Behavior for Congestion Control at CoS Level

- **The EVC behavior described before is applicable to a CoS flow if the EVC carries traffic that belongs to one CoS.**

Congestion Notification

- **In applications such as broadcast where the receiving boundary node(s) do not send user frames to the source(s) contributing to the congestion, there is no way to communicate congestion indication to the transmitting boundary node(s) via user frames when DE bit is used for congestion indication.**
- **In non-broadcast applications, when receiving boundary node does not have a user frame to send or very slow in sending the frames to the source(s) when DE bit is used for congestion indication, congestion indication is either not communicated or communicated untimely.**
- **A solution is to have the congested boundary node to send a special frame, Congestion Notification (CI) Frame, once or twice with no data and DE=1 at the rate of congested entity. It is possible to repeat sending CI frames after a time-out period if the rate is not reduced by the sender.**



IFG	P	SFD	DA	SA	S-Tag	C-Tag	L/T	24 bytes of zeros	FCS

Congestion Notification (CI) Frame with no data

Pros and Cons of Using DE Bit

- **DE bits in S-Tag can be used to identify frame colors at ENNI, if PCP and DSCP are not used for it.**
- **PCP is adequate to represent CoS and colors, unless there are more than 6 CoS.**
- **DSCP is adequate to represent both CoS and color at ENNI.**
- **If there are 7 or 8 CoS, DE bit may need to be used at ENNI to represent colors. However, currently MEF standards have only 3 CoS.**
- **For EVCs with bandwidth profiles that do not contain EIR and EBS, there is no need for color representation.**
- **The procedure will not be used where EVC uses only C-Tag for user frames. However using S-Tag for EVC is a common practice in the industry.**

Thanks!