draft-sajassi-l2vpn-evpn-segment-route-00.txt

Ali Sajassi , Samer Salam, Sami Boutros,
Keyur Patel

July 28th, 2011

IETF Quebec City

Objective

- Enhance baseline E-VPN with one additional route type (Ethernet Segment Route).
- The goal is to enhance multi-homing capabilities of E-VPN by:-

1. Preventing transient loops and packet duplication.
2. Adding support of multi-chassis Ethernet bundles.
3. Improving Designated Forwarder election with VLAN carving.
4. Avoiding relearning of subscriber/session states.

1. Preventing Transient Loop & Packet Duplication

- In E-VPN there is no DF election handshaking:-
  - Each MES constructs a candidate list of DFs from the received Ethernet AD routes, and choose the elected DF independently.
  - During routing transients, MESes may end up electing different DFs for the same ES, leading to packet duplication or forwarding loops.

2. Adding Support of Multi-chassis Ethernet Bundles

- When a CE is multi-homed to a set of MESes using the LACP, the MESes must act as a single LACP speaker

- For MESes to act as a single LACP speaker, they must synchronize
  - LACP configuration
  - Operational data among themselves.

3. Improving Designated Forwarder Election with VLAN Carving

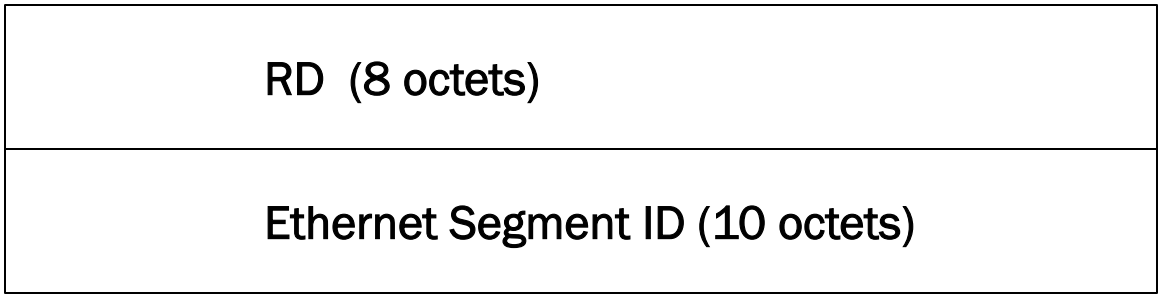
- Use ES route to:-
  - Distribute VLANs across all multihomed MESes and redistribute VLANS among available MESes as they are commissioned or decommissioned. (all VLANs are affected)
  - When a multihomed port or MES fails, the affected VLANs are just reassigned to other MESes (existing VLANs on other MESes are not affected)
  - Improve scale as lot fewer AD routes need to be distributed for port-based and vlan-bundling services

4. Avoiding relearning of subscriber/ session states

- For some applications, the active MES builds and maintains per subscriber (or per session) ‘soft’ state
- In case of link or node failure, this ‘soft’ state must be rebuilt on the backup MES and this may lead to traffic disruption affecting service availability.
- If the states are synchronized between active and backup MESes prior to the failure, then the traffic disruption or service impact will be minimal.

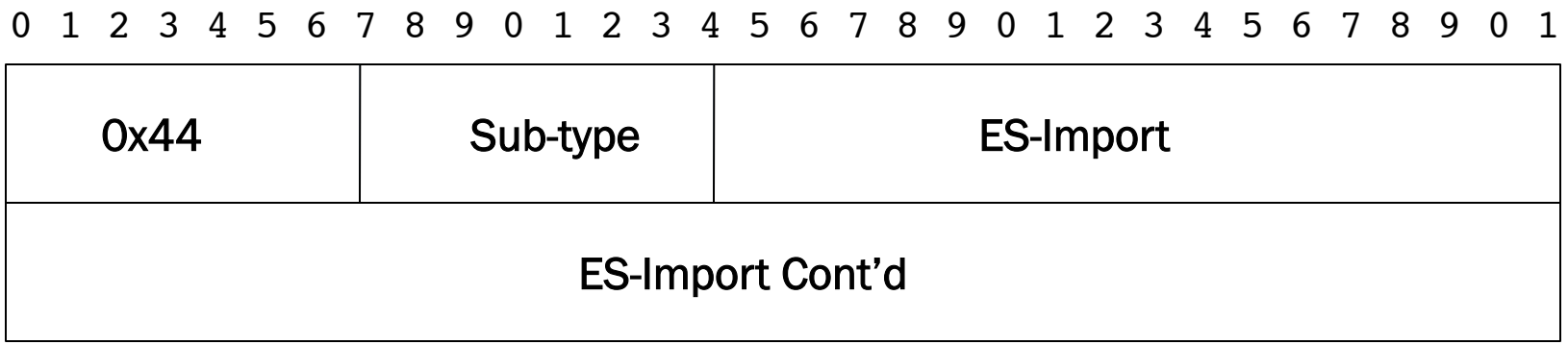
BGP Encoding: Ethernet Segment Route

- The Ethernet Segment Route is encoded in the E-VPN NLRI using the Route Type value of 4



BGP Encoding: ES-Import Extended Community

- It enables all the MESes connected to the same multi-homed site to import the Ethernet Segment routes
- Value is derived automatically from the ESI by encoding the 6-byte MAC address portion of the ESI in the ES-Import Extended Community.



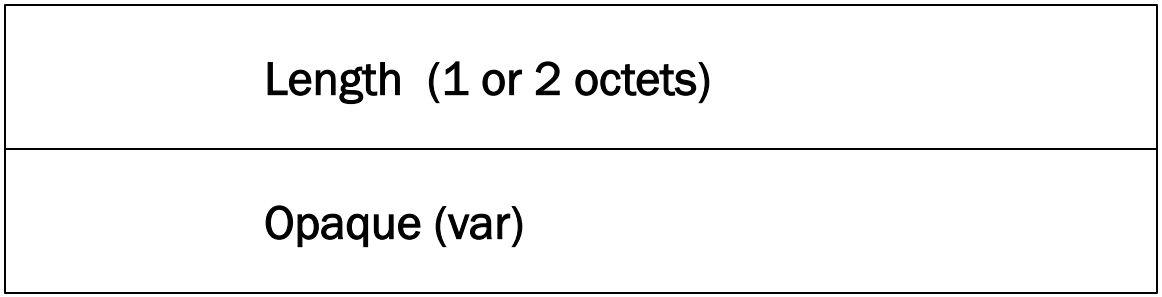
BGP Encoding: DF Election Attribute

- Used for handshaking among MESes in a given ES

State (2 octets)
Sequence No. (4 octets)
Local No. of Links (2 octets)
Total No. of Links (2 octets)
Flags (1 octet)
No. of IP addresses (1 octet)
Ordered list of tuples: [IP address length (1 octet), IP address (4 or 16 octets)]

BGP Encoding: ICC Attribute

- Used for synchronizations of configuration and status information between two halves of the LACP speaker



1. DF Election with Paxos Algorithm

1. When a MES discovers an ESI on an ES, it advertises ES route with associated ES-Import attribute and w/ state=initialization in DF Election attribute.
2. Each MES then starts a timer to allow reception of other ES routes from other MESes connected to the ES, upon timer expiry, each MES builds an ordered list and starts the following handshake:-
    1. The first MES in the list selects itself as the Arbiter Node and initiates the handshake by sending ES route w/ state = ‘proposal pending’ in DF election attribute
    2. When a MES node receives an ES route w/ proposal pending, it either acknowledges it or does nothing.
    3. When the Arbiter node receives ‘promise pending’ from all of the MES nodes in the ordered list, it sends an ES route with ‘Active’ code
    4. When other MES nodes in the ES receive the ‘Active’ code, they respond with ‘Active’ to conclude the handshake

LACP State Synchronization

- MESes belonging to a group representing a single LACP speaker must synchronize following config parameters:
  - System ID & system priority
  - Aggregator ID, MAC address, and Key
  - Port number, key, and priority
- MESes belonging to a group representing a single LACP speaker must synchronize following operational parameters:
  - Partner system ID & system priority
  - Partner port number & port priority
  - Partner key & state
  - Actor state & port state

VLAN Carving

- VLAN carving is achieved by using a slightly modified version of the Paxos procedures for DF election
- Each MES nodes assigns an ordinal for itself based on its position in the ordered list (lowest IP address uses ordinal of 0)
- A given MES selects itself as a DF for a given VLAN using:
  - Assuming N MESes in an ES, the MES with ordinal I is the DF for VLAN V when (V mod N)=i
- Each MES node unblocks only the VLANs for which it is a DF for the ES

Subscriber/Session Synchronization

- Synchronization is performed using ICC Attribute of ES route
- Various applications are responsible for the encoding and decoding of relevant data
- The exact encoding/decoding is outside of the scope of this draft
- BGP is used as a reliable transport service