INTERNET-DRAFT
Intended Status: Standards Track

Ali Sajassi
Samer Salam
Sami Boutros
Keyur Patel
Cisco

Expires: January 4, 2011
July 4, 2011

E-VPN Ethernet Segment Route
draft-sajassi-l2vpn-evpn-segment-route-00.txt

Status of this Memo

Copyright and License Notice

Abstract

[E-VPN] defines a solution and architecture for BGP MPLS-based Ethernet VPNs. This document describes an additional BGP route and associated route attributes that enhance the multi-homing capabilities of the solution. These are: the Ethernet Segment Route, the ESI Import Extended Community, the DF Election Attribute and the Inter-chassis Communication Attribute. This draft describes their usage, advantages and encoding.

Table of Contents

1 Introduction

[E-VPN] defines a solution and architecture for BGP MPLS-based Ethernet L2VPN services with advanced multi-homing capabilities. To that end, [E-VPN] defines a new BGP NLRI with 5 route types:

1. Ethernet Auto-Discovery (A-D) route
2. MAC advertisement route
3. Inclusive Multicast Route
5. Selective Multicast Auto-Discovery (A-D) Route
6. Leaf Auto-Discovery (A-D) Route

In this draft, we define one additional route type:

4. Ethernet Segment Route

This route primarily enhances the multi-homing capabilities of the E-VPN solution in the following areas:

- Preventing transient loops and packet duplication
- Support of multi-chassis Ethernet bundles
- Designated Forwarder election with VLAN carving
- Avoiding relearning of subscriber/session state

In addition to the above route, 3 new BGP route attributes are defined: the ESI Import Extended Community attribute, the DF Election attribute and the Inter-chassis Communication attribute.

Section 2 discusses the motivation and usage of the new route and attributes. Section 3 describes the BGP encoding.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2 Motivation and Usage

This section focuses on the reasons for defining the Ethernet Segment route and its associated 3 BGP attributes, and describes its usage in E-VPN.

2.1 Preventing Transient Loops and Packet Duplication

The Designated Forwarder (DF) election procedures defined in [E-VPN] require that each MES constructs a candidate list of DFs from the received Ethernet A-D routes. By default, each MES then independently

chooses the MES with the highest IP address as the elected DF. There is no handshake mechanism between the MESes that are connected to the same Ethernet Segment. As a result of that, during routing transients, different MESes may end up electing different DFs for the same Ethernet Segment due to inconsistent views of the network. If the Ethernet Segment is a multi-homed device, this may lead to transient packet duplication. If the Ethernet Segment is a multi-homed network, the presence of multiple DFs may lead to transient forwarding loops in addition to potential packet duplication.

To eliminate these issues, a handshake mechanism is required between the MES nodes connected to the same Ethernet Segment, to ensure a common view of the network among them. This handshake is performed using the DF Election attribute carried in the Ethernet Segment route, as discussed in the 'DF Election with Paxos Algorithm' section.

2.2 Support of Multi-Chassis Ethernet Bundles

When a CE is multi-homed to a set of MES nodes using the [802.1AX] Link Aggregation Control Protocol (LACP), the MESes must act as if they were a single LACP speaker for the Ethernet links to form a bundle, and operate correctly as a Link Aggregation Group (LAG). To achieve this, the MESes connected to the same multi-homed CE must synchronize LACP configuration and operational data among them. The synchronization is required for the following reasons:

- to determine if the links in the Ethernet bundle are to operate in all-active or hot-standby resiliency mode

- to detect and handle CE mis-configuration when LACP Port Key is configured on the MES

- to detect and handle mis-wiring between CE and MES when LACP Port Key is configured on the MES

- to deterministically agree on which link(s) should join a bundle based on port and system priorities, especially when the number of links exceeds the aggregation capacity of the MESes, and the MES LACP System Priority is higher than the CE's

- to detect and react to actor/partner churn where the LACP speakers are not able to converge

Synchronization of LACP state between MESes is performed using the Inter-chassis Communication attribute carried in the Ethernet Segment route, as described in the 'LACP State Synchronization' section below.

2.3 Designated Forwarder (DF) Election with VLAN Carving

In the case where multiple MES nodes offer redundant connectivity for an Ethernet Segment, it is preferred to elect multiple DFs (one DF per VLAN) in order to distribute the traffic among the redundancy group members. This process of electing different DFs for different VLANs on an Ethernet Segment, for purpose of load-balancing, is referred to as 'VLAN Carving'.

The VLAN carving algorithm must ensure even distribution of VLANs among the MES nodes servicing the same Ethernet Segment. As new MES devices get commissioned or decommissioned, the VLANs must be redistributed over the available devices for even load-balancing. However, in the case of link, port or node failure, the VLAN carving algorithm should ensure that only the affected VLANs are reassigned to different MES(es), and none of the other active VLANs are shuffled. Otherwise, the fault decoupling capability of the redundancy group would be compromised.

VLAN carving requires exchange of information among the MES nodes connected to an Ethernet Segment in order to agree upon how the VLANs will be distributed. Since this information is only relevant to the MES nodes that are directly connected to a specific Ethernet Segment, the exchanges and associated processing should be localized to the redundancy group members.

DF Election with VLAN carving is performed using the DF Election attribute carried in the Ethernet Segment route, as described in the "VLAN Carving" section below.

2.4 Route Scalability with Granular DF Election

[E-VPN] allows for DF election to be performed at the granularity of either an Ethernet Segment or combination of Ethernet Segment and VLAN on that segment. In the latter case, an Ethernet A-D route per (ESI, VLAN) must be advertised by the MES regardless of whether the service interface is port-based, VLAN-based, VLAN bundling-based or VLAN aware bundling-based. In case of port-based and VLAN bundling-based services, these routes are only required for DF election and not for advertising forwarding labels. By using the Ethernet Segment route instead of the Ethernet A-D route for DF election, it is still possible to have per-VLAN DF granularity while significantly reducing the number of BGP routes advertised. For e.g., consider an Ethernet Segment ESI1 used for a port-based service. By using the Ethernet A-D route for per (ESI, VLAN) DF election, 4095 routes are needed. Whereas, using the Ethernet Segment route, only a single route is required.

2.5 Avoiding Relearning of Subscriber/Session State

For certain applications, the MES builds and maintains per subscriber or per session 'soft' state that is used for either optimizing the traffic forwarding or enforcing security. Examples of such per subscriber/session state includes:

- multicast state derived from IGMP or PIM snooping

- IP address to MAC address bindings gleaned from snooping ARP and/or DHCP packets, and used to prevent address spoofing or masquerading

When a set of MES nodes provides multi-homed connectivity for an Ethernet Segment, this 'soft' state is built on the active MES node that forwards and snoops the relevant protocol packets. In case of a link or node failure, the state must be reconstructed on the backup MES (e.g. by waiting for the next IGMP query or ARP message or by issuing unsolicited queries). This may cause traffic disruption and affect the availability of the service. Alternatively, the state can be synchronized among the MES nodes via BGP, and that would enhance the convergence of the service after failure.

Synchronization of subscriber/session state between MES nodes is performed using the Inter-chassis Communication attribute carried in the Ethernet Segment route, as described in the 'Subscriber/Session State Synchronization' section below.

3 BGP Encoding

This section defines the encoding of the BGP route and attributes.

3.1 Ethernet Segment Route

The Ethernet Segment Route is encoded in the E-VPN NLRI defined in [E-VPN] using the Route Type value of 4. The Route Type Specific field of the NLRI is formatted as follows:

```
             +---------------------------------------+
             |      RD   (8 octets)                  |
             +---------------------------------------+
             |Ethernet Segment Identifier (10 octets)|
             +---------------------------------------+
```

3.2 ES-Import Extended Community

This is a new transitive extended community carried with the Ethernet Segment route. When used, it enables all the MESes connected to the

same multi-homed site to import the Ethernet Segment routes. The value is derived automatically from the ESI by encoding the 6-byte MAC address portion of the ESI in the ES-Import Extended Community. The format of this extended community is as follows:

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| 0x44        |   Sub-Type    |          ES-Import              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     ES-Import Cont'd                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

3.3 DF Election Attribute

```
+---------------------------------------+
|      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 bytes)]|         |
+---------------------------------------+
```

State field can take one of the following values:

0x0000 Initializing
0x0001 Proposal Pending
0x0002 Promise Pending
0x0003 Active

Flags field is encoded as follows:

7 bits: reserved
Least significant bit: Protecting flag

3.4 Inter-chassis Communication Attribute

```
+---------------------------------------+
|       Type  (2 octets)                |
```

```
      +---------------------------------------+
      |       Length   (1 or 2 octets)        |
      +---------------------------------------+
      |       Opaque  (var)                   |
      +---------------------------------------+
```

4 DF Election with Paxos Algorithm

The procedures in this section guarantee that all MES nodes in a given redundancy group agree on a unique DF for a given Ethernet Segment. This eliminates the problem of transient forwarding loops and transient packet duplicates described above. The procedures can be broken down to the following steps:

1. When a MES discovers the ESI of the attached Ethernet Segment, it advertises an Ethernet Segment route with the associated ES-Import extended community attribute and with the 'Initializing' code in the State field of the DF Election attribute.

2. The MES then starts a timer to allow the reception of Ethernet Segment routes from other MES nodes in the same redundancy group.

3. When the timer expires, each MES builds an ordered list of the IP addresses of all the MES nodes connected to the Ethernet Segment (including itself), in increasing numeric value.

4. The first MES in the ordered list then elects itself as the Arbiter Node (AN). It initiates the handshake by sending an Ethernet Segment route with 'Proposal Pending' code in the State field of the DF Election attribute.

5. When a MES node receives an Ethernet Segment route with the 'Proposal Pending' code, it takes one of the following options:

   a. If the receiving MES ranks the transmitting MES's IP address as the top entry in its local ordered list, it acknowledges the handshake by responding with an Ethernet Segment route with the 'Promise Pending' code in the State field of the DF Election attribute. This includes the scenario where the receiving MES forfeits the AN role to another advertising MES with a numerically lower IP address.

   b. If the receiving MES does not rank the transmitting MES's IP address as the top entry in its local ordered list, and the receiving MES had advertised an Ethernet Segment route with the 'Initializing' code or with the 'Proposal Pending' code, then the MES takes no further action.

6. When the AN receives 'Promise Pending' from all of the MES nodes in the ordered list, it sends an updated Ethernet Segment route with the 'Active' code in the DF Election attribute.

7. When the other MES nodes in the redundancy group receive the 'Active' code from the AN, they respond with an updated Ethernet Segment route with the 'Active' code in the DF Election attribute. This concludes the handshake.

In the case where the DF election is performed at the granularity of an Ethernet Segment, i.e. there is a single DF for all VLANs on the segment, the Arbiter Node is effectively the Designated Forwarder for the segment. All the MES nodes start off with their ports, that are connected to the segment, blocked in Step 1 (for multi-destination traffic from core). And in Step 6, the MES confirmed as the AN (i.e. DF) unblocks its port towards the Ethernet Segment. DF election at the granularity of (Ethernet Segment, VLAN) is discussed in the "VLAN Carving" section below.

5 LACP State Synchronization

To support CE multi-homing with multi-chassis Ethernet bundles, the MES nodes connected to a given CE should synchronize [802.1AX] LACP state amongst each other. This includes the following LACP specific configuration parameters:

- System Identifier (MAC Address): uniquely identifies a LACP speaker.
- System Priority: determines which LACP speaker's port priorities are used in the Selection logic.
- Aggregator Identifier: uniquely identifies a bundle within a LACP speaker.
- Aggregator MAC Address: identifies the MAC address of the bundle.
- Aggregator Key: used to determine which ports can join an Aggregator.
- Port Number: uniquely identifies an interface within a LACP speaker.
- Port Key: determines the set of ports that can be bundled.
- Port Priority: determines a port's precedence level to join a bundle in case the number of eligible ports exceeds the maximum number of links allowed in a bundle.

The above information must be synchronized between the MES nodes wishing to form a multi-chassis bundle with a given CE, in order for the former to convey a single LACP peer to that CE. This is required for initial system bring-up and upon any configuration change. Furthermore, the MESes must also synchronize operational (run-time) data, in order for the LACP Selection logic state-machines to

execute. This operational data includes the following LACP operational parameters, on a per port basis:

- Partner System Identifier: this is the CE System MAC address.
- Partner System Priority: the CE LACP System Priority
- Partner Port Number: CE's AC port number.
- Partner Port Priority: CE's AC Port Priority.
- Partner Key: CE's key for this AC.
- Partner State: CE's LACP State for the AC.
- Actor State: PE's LACP State for the AC.
- Port State: PE's AC port status.

The above state needs to be communicated between MESes forming a multi-chassis bundle during LACP initial bring-up, upon any configuration change and upon the occurrence of a failure.

It should be noted that the above configuration and operational state is localized in scope and is only relevant to PEs within a given Redundancy Group, i.e. which connect to the same Ethernet Segment over a given Ethernet bundle. Furthermore, the communication of state changes, upon failures, must occur with minimal latency, in order to minimize the switchover time and consequent service disruption.

Without synchronization of the above parameters, the system is subject to the issues outlined in section 2.2 above.

6 VLAN Carving

It is possible to elect multiple DFs per Ethernet Segment (one per VLAN) by using a slightly modified version of the procedures described in the "DF Election with Paxos Algorithm" section above.

In step 3, each of the MES nodes assigns an ordinal for itself based on the order of its IP address in the list. The first MES in the list (the one with the numerically lowest IP address) is given an ordinal of 0. The ordinals are used to determine which MES node will be the DF for a given VLAN on the Ethernet Segment using the following rule:

Assuming a redundancy group of N MES nodes, the MES with ordinal i is the DF for VLAN V when (V MOD N) = i.

In step 6, the AN unblocks only the VLANs for which it is a DF for the Ethernet Segment.

In step 7, each MES node unblocks only the VLANs for which it is a DF for the Ethernet Segment.

In the case of a port, link or node failure, the AN takes over the forwarding for the affected VLANs on the segment and advertises an updated Ethernet Segment route with the 'Active' code and 'Protecting' flag set in the DF Election attribute. Therefore, when VLAN carving is used, the AN acts as the Backup DF (BDF) for the Ethernet Segment. This ensures that only the affected VLANs are failed over, and none of the other VLANs are shuffled.

When the fault clears, the following procedure is followed to revert the VLANs to the recovering MES:

1. The recovering MES advertises an Ethernet Segment route with the 'Initializing' code in the State field of the DF Election attribute.

2. The recovering MES receives from the other MES nodes Ethernet Segment routes with the 'Active' code in the DF Election attribute. The MES can, then, build its ordered list.

3. The recovering MES advertises an Ethernet Segment route with the 'Proposal-Pending' code in the DF Election attribute. This is meant to indicate to the AN that the recovering MES is ready to take over its VLANs.

4. Upon receiving the route with the 'Proposal Pending' code, the AN blocks all the VLANs that belong to the recovering MES. The AN then advertises an updated Ethernet Segment route with the 'Protecting' flag cleared.

5. Upon receiving the above route from the AN, the recovering MES unblocks the VLANs for which it is the DF. The recovering MES then transmits an Ethernet Segment route with the 'Active' code. This completes the reversion.

If the failed MES is the AN, then the MES node with the second best claim to be AN (i.e. whose IP address is the second in the ordered list) takes over the failed VLANs and advertises an updated Ethernet Segment route with the 'Active' code and 'Protecting' flag set in the DF Election attribute. The procedures for reversion, in this case, are as follows:

1. The recovering AN advertises an Ethernet Segment route with the 'Initializing' code in the State field of the DF Election attribute.

2. The recovering AN receives from the other MES nodes Ethernet Segment routes with the 'Active' code in the DF Election attribute.

3. The recovering AN advertises an Ethernet Segment route with the 'Proposal-Pending' code in the DF Election attribute.

4. The other MES nodes respond to that advertisement with Ethernet Segment routes with the 'Promise-Pending' code in the DF Election attribute. At this point, the BDF blocks all the VLANs that belong to the recovering AN before advertising its Ethernet Segment route, with the 'Promise-Pending' code and 'Protecting' flag cleared.

5. The recovering AN unblocks the VLANs for which it is the DF upon receiving the 'Promise-Pending' advertisements from the BDF. The AN then advertises an Ethernet Segment route with the 'Active' code once it receives the Ethernet Segment route with 'Promise-Pending' code from all of the MES nodes in the redundancy group.

6. The other MES nodes respond with Ethernet Segment routes with the 'Active' code. This marks the end of the reversion.

7 Subscriber/Session State Synchronization

Synchronization of subscriber/session state between MES nodes is performed using the Inter-chassis Communication attribute carried in the Ethernet Segment route. The various applications are responsible for the encoding and decoding of the relevant data, and this is outside the scope of this draft. BGP provides a reliable transport service in this case.

8 Security Considerations

There are no additional security aspects beyond those of VPLS/H-VPLS that need to be considered.

9 IANA Considerations

To be added in a later revision.

10 References

10.1 Normative References

[RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

10.2 Informative References

[E-VPN] Aggarwal et al., "BGP MPLS Based Ethernet VPN", draft-raggarwa-sajassi-l2vpn-evpn-02.txt, work in progress,

March, 2011.

[EVPN-REQ] Sajassi et al., "Requirements for Ethernet VPN (E-VPN)", draft-sajassi-raggarwa-l2vpn-evpn-req-00.txt, work in progress, October, 2010.

Author's Addresses

Ali Sajassi
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sajassi@cisco.com

Samer Salam
Cisco
595 Burrard Street, Suite 2123
Vancouver, BC V7X 1J1, Canada
Email: ssalam@cisco.com

Sami Boutros
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sboutros@cisco.com

Keyur Patel
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: keyupate@cisco.com