

INTERNET-DRAFT
Intended Status: Standards Track

Ali Sajassi
Samer Salam
Sami Boutros
Keyur Patel
Cisco
July 16, 2012

Expires: January 17, 2013

E-VPN Ethernet Segment Route
draft-sajassi-l2vpn-evpn-segment-route-01.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

INTERNET DRAFT [draft-sajassi-l2vpn-evpn-segment-route](#) July 16, 2012

Abstract

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

Table of Contents

1	Introduction	3
1.1	Terminology	3
2	Motivation and Usage	3
2.1	Support of Multi-Chassis Ethernet Bundles	3
2.2	Avoiding Relearning of Subscriber/Session State	4
2.3	Preventing Transient Loops and Packet Duplication	4
3	BGP Encoding	5
3.1	DF Election Attribute	5
3.2	Inter-chassis Communication Attribute	5
4	DF Election with Paxos Algorithm	6
5	LACP State Synchronization	7
6	Subscriber/Session State Synchronization	8
7	Security Considerations	8
8	IANA Considerations	8
9	References	9
9.1	Normative References	9
9.2	Informative References	9
	Author's Addresses	9

INTERNET DRAFT [draft-sajassi-l2vpn-evpn-segment-route](#) July 16, 2012

1 Introduction

[E-VPN] defines a solution and architecture for BGP MPLS-based Ethernet L2VPN services with advanced multi-homing capabilities. In this draft we define procedures and extensions that enhance 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
- Avoiding relearning of subscriber/session state

Two new BGP route attributes are defined: the DF Election attribute and the Inter-chassis Communication attribute.

[Section 2](#) discusses the motivation and usage of the new 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 2 BGP attributes, and describes their usage in E-VPN.

2.1 Support of Multi-Chassis Ethernet Bundles

When a CE is multi-homed to a set of PE nodes using the [802.1AX] Link Aggregation Control Protocol (LACP), the PEs 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 PEs 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 PE
- to detect and handle mis-wiring between CE and PE when LACP Port Key is configured on the PE

- 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 PEs, and the PE 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 PEs is performed using the Inter-chassis Communication attribute carried in the Ethernet Segment route, as described in the 'LACP State Synchronization' section below.

[2.2](#) Avoiding Relearning of Subscriber/Session State

For certain applications, the PE 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 PE nodes provides multi-homed connectivity for an Ethernet Segment, this 'soft' state is built on the active PE 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

PE (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 PE nodes via BGP, and that would enhance the convergence of the service after failure.

Synchronization of subscriber/session state between PE 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.

[2.3](#) Preventing Transient Loops and Packet Duplication

During routing transients, different PEs 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, an optional handshake mechanism is defined to ensure that the PE nodes connected to the same Ethernet Segment share a common view of the access network topology. This handshake is performed using the DF Election attribute carried in the Ethernet Segment route, as discussed in [Appendix I](#): 'DF Election with Paxos Algorithm'.

[3](#) BGP Encoding

This section defines the encoding of the BGP attributes.

[3.1](#) 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

```

[3.2](#) 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 PE 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 PE 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 PE then starts a timer to allow the reception of Ethernet Segment routes from other PE nodes in the same redundancy group.
3. When the timer expires, each PE builds an ordered list of the IP addresses of all the PE nodes connected to the Ethernet Segment (including itself), in increasing numeric value.
4. The first PE 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 PE node receives an Ethernet Segment route with the 'Proposal Pending' code, it takes one of the following options:
 - a. If the receiving PE ranks the transmitting PE'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 PE forfeits the AN role to another advertising PE with a numerically lower IP address.
 - b. If the receiving PE does not rank the transmitting PE's IP address as the top entry in its local ordered list, and the receiving PE had advertised an Ethernet Segment route with the 'Initializing' code or with the 'Proposal Pending' code, then the

PE takes no further action.

6. When the AN receives 'Promise Pending' from all of the PE 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 PE 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 PE 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 PE 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 PE 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 PE 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 PEs 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 PEs 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](#) Subscriber/Session State Synchronization

Synchronization of subscriber/session state between PE 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.

[7](#) Security Considerations

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

[8](#) IANA Considerations

To be added in a later revision.

[9](#) References

[9.1](#) Normative References

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

[9.2](#) 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

Sajassi et al.

Expires January 17, 2013

[Page 9]

INTERNET DRAFT

[draft-sajassi-l2vpn-evpn-segment-route](#)

July 16, 2012

Email: keyupate@cisco.com

