BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 21, 2016

S. Mohanty
K. Patel
A. Sajassi
Cisco Systems, Inc.
J. Drake
Juniper Networks, Inc.
A. Przygienda
Ericsson

October 19, 2015

A new Designated Forwarder Election for the EVPN
draft-mohanty-bess-evpn-df-election-02

Abstract

This document describes an improved EVPN Designated Forwarder Election (DF) algorithm which can be used to enhance operational experience in terms of convergence speed and robustness over a WAN deploying EVPN

Table of Contents

1. Introduction

Ethernet MPLS VPN (EVPN) [RFC7432] is an emerging technology that is gaining prominence in Internet Service Provider IP/MPLS networks. In EVPN, mac addresses are disseminated as routes across the geographical area via the Border Gateway Protocol, BGP [RFC4271] using the familiar L3VPN model [RFC4364]. An EVPN instance that spans across PEs is defined as an EVI. Constrained Route Distribution [RFC4684] can be used in conjunction to selectively advertise the routes to where they are needed. One of the major advantages of EVPN over VPLS [RFC4761],[RFC6624] is that it provides a solution for minimizing flooding of unknown traffic and also provides all Active mode of operation so that the traffic can truly be multi-homed. In technologies such as EVPN or VPLS, managing Broadcast, Unknown Unicast and multicast traffic (BUM) is a key requirement. In the case where the customer edge (CE) router is multi-homed to one or more Provider Edge (PE) Routers, it is necessary that one and only one of the PE routers should forward BUM traffic into the core or towards the CE as and when appropriate.

Specifically, quoting Section 8.5, [RFC7432], Consider a CE that is a host or a router that is multi-homed directly to more than one PE in

an EVPN instance on a given Ethernet segment. One or more Ethernet Tags may be configured on the Ethernet segment. In this scenario only one of the PEs, referred to as the Designated Forwarder (DF), is responsible for certain actions:

a. Sending multicast and broadcast traffic, on a given Ethernet Tag on a particular Ethernet segment, to the CE.

b. Flooding unknown unicast traffic (i.e. traffic for which an PE does not know the destination MAC address), on a given Ethernet Tag on a particular Ethernet segment to the CE, if the environment requires flooding of unknown unicast traffic.

```
                                +---------------+
                                |   IP/MPLS     |
                                |   CORE        |
                  +----+ ES1 +----+           +----+
                  | CE1|-----|    |-----------|    |____ES2
                  +----+     | PE1|           | PE2|    \
                             |    |--------   +----+     \+----+
                             +----+        |    |         | CE2|
                                |          |  +----+     /+----+
                                |          |__|    |____/   |
                                |             | PE3|    ES2 /
                                |             +----+       /
                                |               |         /
                                +-------------+----+     /
                                              | PE4|____/ES2
                                              |    |
                                              +----+
```

Figure 1 Multi-homing Network of E-VPN

Figure 1

Figure 1 illustrates a case where there are two Ethernet Segments, ES1 and ES2. PE1 is attached to CE1 via Ethernet Segment ES1 whereas PE2, PE3 and PE4 are attached to CE2 via ES2 i.e. PE2, PE3 and PE4 form a redundancy group. Since CE2 is multi-homed to different PEs on the same Ethernet Segment, it is necessary for PE2, PE3 and PE4 to agree on a DF to satisfy the above mentioned requirements.

Layer2 devices are particularly susceptible to forwarding loops because of the broadcast nature of the Ethernet traffic. Therefore

it is very important that in case of multi-homing, only one of the links be used to direct traffic to/from the core.

One of the pre-requisites for this support is that participating PEs must agree amongst themselves as to who would act as the Designated Forwarder. This needs to be achieved through a distributed algorithm in which each participating PE independently and unambiguously selects one of the participating PEs as the DF, and the result should be unanimously in agreement.

The DF election algorithm as described in [RFC7432] has some undesirable properties and in some cases can be somewhat disruptive and unfair. This document describes those issues and proposes a mechanism for dealing with those issues. These mechanisms do involve changes to the DF Election algorithm , but do not require any protocol changes to the EVPN Route exchange and have minimal changes to their content per se.

1.1. Finite State Machine

Since the specification in EVPN RFC [RFC7432] does leave several questions open as to the precise final state machine behavior of the DF election, the document also includes a section describing precisely the intended behavior. The finite state machine is presented in Section 7.1

1.2. Requirements Language

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 [RFC2119].

2. The modulus based DF Election Algorithm

The default procedure for DF election at the granularity of (ESI,EVI) is referred to as "service carving". With service carving, it is possible to elect multiple DFs per Ethernet Segment (one per EVI) in order to perform load-balancing of multi-destination traffic destined to a given Segment. The objective is that the load-balancing procedures should carve up the EVI space among the redundant PE nodes evenly, in such a way that every PE is the DF for a disjoint set of EVIs.

The existing DF algorithm as described in the EVPN RFC(Section 8.5 [RFC7432]) is based on a modulus operation. The PEs to which the ES (for which DF election is to be carried out per vlan) is multi-homed from an ordered (ordinal) list in ascending order of the PE ip address values. Say, there are N PEs, P0, P1, ... PN-1 ranked as per

increasing IP addresses in the ordinal list; then for each vlan with ethernet tag v, configured on the ethernet segment ES1, PEx is the DF for vlan v on ES ES1 when x equals (v mod N). In the case of VLAN bundle only the lowest VLAN is used. In the case when the vlan density is high meaning there are significant number of vlans and the vlan-id or ethernet-tag is uniformly distributed, the thinking is that the DF election will be spread across the PEs hosting that ethernet segment and good service carving can be achieved.

3. Problems with the modulus based DF Election Algorithm

There are three fundamental problems with the current DF Election.

First, the algorithm will not perform well when the ethernet tag follows a non-uniform distribution, for instance when the ethernet tags are all even or all odd. In such a case let us assume that the ES is multi-homed to two PEs; all the vlans will only pick one of the PEs as the DF. This is very sub-optimal. It defeats the purpose of service carving as the DFs are not really evenly spread across. In this particular case, in fact one of the PEs does not get elected all as the DF, so it does not participate in the DF responsibilities at all. Consider another example where referring to Figure 1, lets assume that PE2, PE3, PE4 are in ascending order of the IP address; and each vlan configured on ES2 is associated with an Ethernet Tag of of the form (3x+1), where x is an integer. This will result in PE3 always be selected as the DF.

Even in the case when the ethernet tag distribution is uniform the instance of a PE being up or down results in re-computation ((v mod N-1) or (v mod N+1) as is the case); The resulting modulus value need not be uniformly distributed but subject to the primality of N-1 or N+1 as may be the case.

The third problem is one of disruption. Consider a case when the same Ethernet Segment is multi homed to a set of PEs. When the ES is down in one of the PEs, say PE1, or PE1 itself reboots, or the BGP process goes down or the connectivity between PE1 and an RR goes down, the effective number of PEs in the system now becomes N-1 and DFs are computed for all the vlans that are configured on that ethernet segment. In general, if the DF for a vlan v happens not to be PE1, but some other PE, say PE2, it is likely that some other PE will become the new DF. This is not desirable. Similarly when a new PE hosts the same Ethernet segment, the mapping again changes because of the mod operation. This results in needless churn. Again referring to Figure 1, say v1, v2 and v3 are vlans configured on ES2 with associated ethernet tags of value 999, 1000 and 10001 respectively. So PE1, PE2 and PE3 are also

the DFs for v1, v2 and v3 respectively. Now when PE3 goes down, PE2 will become the DF for v1 and PE1 will become the DF for v2.

One point to note is that the current DF election algorithm assumes that all the PEs who are multi-homed to the same Ethernet Segment and interested in the DF Election by exchanging EVPN routes have a V4 peering with each other or via a Route Reflector. This need not be the case as there can be a v6 peering and supporting the EVPN address-family.

Mathematically, a conventional hash function maps a key k to a number i representing one of m hash buckets through a function h(k) i.e. i=h(k). In the EVPN case, h is simply a modulo-m hash function viz. h(v) = v mod N, where N is the number of PEs that are multi-homed to the Ethernet Segment in discussion. It is well-known that for good hash distribution using the modulus operation, the modulus N should be a prime-number not too close to a power of 2 [CLRS2009]. When the effective number of PEs changes from N to N-1 (or vice versa); all the objects (vlan v) will be remapped except those for which v mod N and v mod (N-1) refer to the same PE in the previous and subsequent ordinal rankings respectively.

From a forwarding perspective, this is a churn, as it results in programming the CE and PE side ports as blocking or non-blocking at potentially all PEs when the DF changes either because (i) a new PE is added or (ii) another one goes down or loses connectivity or else cannot take part in the DF election process for whatever reason. This draft addresses this problem and furnishes a solution to this undesirable behavior.

4. Highest Random Weight

Highest Random Weight (HRW) as defined in [HRW1999] is originally proposed in the context of Internet Caching and proxy Server load balancing. Given an object name and a set of servers, HRW maps a request to a server using the object-name (object-id) and server-name (server-id) rather than the state of the server states. HRW forms a hash out of the server-id and the object-id and forms an ordered list of the servers for the particular object-id. The server for which the hash value is highest, serves as the primary responsible for that particular object, and the server with the next highest value in that hash serves as the backup server. HRW always maps a given object object name to the same server within a given cluster; consequently it can be used at client sites to achieve global consensus on object-server mappings. When that server goes down, the backup server becomes the responsible designate.

Choosing an appropriate hash function that is statistically oblivious to the key distribution and imparts a good uniform distribution of the hash output is an important aspect of the algorithm,. Fortunately many such hash functions exist. [HRW1999] provides pseudorandom functions based on Unix utilities rand and srand and easily constructed XOR functions that perform considerably well. This imparts very good properties in the load balancing context. Also each server independently and unambiguously arrives at the primary server selection. HRW already finds use in multicast and ECMP [RFC2991],[RFC2992].

In the existing DF algorithm Section 2, whenever a new PE comes up or an existing PE goes down, there is a significant interval before the change is noticed by all peer PEs as it has to be conveyed by the BGP update message involving the type-4 route. There is a timer to batch all the messages before triggering the service carving procedures. When the timer expires, each PE will build the ordered list and follow the procedures for DF Election. In the proposed method which we will describe shortly this "jittered" behavior is retained.

5. HRW and Consistent Hashing

HRW is not the only algorithm that addresses the object to server mapping problem with goals of fair load distribution, redundancy and fast access. There is another family of algorithms that also addresses this problem; these fall under the umbrella of the Consistent Hashing Algorithms [CHASH]. These will not be considered here.

6. HRW Algorithm for EVPN DF Election

The applicability of HRW to DF Election can be described here. Let DF(v) denote the Designated Forwarder and BDF(v) the Backup Designated forwarder for the ethernet tag V, where v is the vlan, Si is the IP address of server i and weight is a pseudorandom function of v and Si. In case of a vlan bundle service, v denotes the lowest vlan similar to the 'lowest vlan in bundle' logic of [RFC7432].

1. DF(v) = Si: Weight(v, Si) >= Weight(V, Sj) , for all j. In case of a tie, choose the PE whose IP address is numerically the least.

2. BDF(v) = Sk: Weight(v, Si) >= Weight(V, Sk) and Weight(v, Sk) >= Weight(v, Sj). in case of tie choose the PE whose IP address is numerically the least.

Since the Weight is a Pseudorandom function with domain as a concatenation of (v, S), it is an efficient deterministic algorithm

which is independent of the Ethernet Tag V sample space distribution.

Choosing a good hash function for the pseudorandom function is an important consideration for this algorithm to perform provably better than the existing algorithm. As mentioned previously, such functions are described in the HRW paper. We take as candidate hash functions two of the ones that are preferred in [HRW1999].

1. Wrand(v, Si) = (1103515245((1103515245.Si+12345)XOR D(v))+12345)(mod 2^31) and

2. Wrand2(v, Si) = (1103515245((1103515245.D(v)+12345)XOR Si)+12345)(mod 2^31)

Here D(v) is the 31-bit digest of the ethernet-tag v and Si is address of the ith server. The server's IP address length does not matter as only the low-order 31 bits are modulo significant. Eventually we plan to choose one of the two candidate hash functions as the preferred one.

A point to note is that the the domain of the Weight function is a concatenation of the ethernet-tag and the PE IP-address, and the actual length of the server IP address (whether V4 or V6) is not really relevant, so long as the actual hash algorithm takes into consideration the concatenated string. The existing algorithm in [RFC7432] as is cannot employ both V4 and V6 neighbor peering address.

HRW solves the disadvantage pointed out in Section 3 and ensures

- with very high probability that the task of DF election for respective vlans is more or less equally distributed among the PEs even for the 2 PE case

- If a PE, hosting some vlans on given ES, but is neither the DF nor the BDF for that vlan, goes down or its connection to the ES goes down, it does not result in a DF and BDF reassignment the other PEs. This saves computation, especially in the case when the connection flaps.

- More importantly it avoids the needless disruption case (c) that are inherent in the existing modulus based algorithm

- In addition to the DF, the algorithm also furnishes the BDF, which would be the DF if the current DF fails.

7. Protocol Considerations

Note that for the DF election procedures to be globally convergent and unanimous, it is necessary that all the participating PEs agree on the DF Election algorithm to be used. It is not possible that some PEs continue to use the existing modulus based DF election and some newer PEs use the HRW. For brownfield deployments and for interoperability with legacy boxes, its is important that all PEs need to have the capability to fall back on the modulus algorithm. A PE (one with a newer version of the software) can indicate its willingness to support HRW by signaling a new extended community along with the Ethernet-Segment Route (Type-4). This extended community is explained in the next paragraph. When a PE receives the Ethernet-Segment Routes from all the other PEs for the ethernet segment in question, it checks to see if all the advertisements have the extended community attached; in the case that they do, this particular PE, and by induction all the other PEs proceed to do DF Election as per the HRW Algorithm. Otherwise if even a single advertisement for the type-4 route is not received with the extended community or the received DF types (including locally configured type) do not ALL match a single value, the default modulus algorithm is used as before. Also, the HRW algorithm needs to be executed after the "batching" time.

A new BGP extended community attribute [RFC4360] needs to be defined to identify the DF election procedure to be used for the Ethernet Segment. We propose to name this extended community as the DF Election Extended Community. It is a new transitive extended community where the Type field is 0x06, and the Sub-Type is to be defined. It may be advertised along with Ethernet Segment routes.

Each DF Election Extended Community is encoded as a 8-octet value 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Type=0x06     | Sub-Type(TBD) | DF Type(One Octet) |Reserved=0  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Reserved = 0                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2

The DF Type state is encoded as one octet. A value of 0 means that the default (the mod based) DF election procedures are used and a value of 1 means that the HRW algorithm will be employed. A request

needs to registered with the IETF authority for the subtype [I-D.ietf-idr-extcomm-iana]

7.1. Finite State Machine

Per [RFC7432], the FSM described in Figure 3 is executed per ESI/VLAN in case of VLAN aware service or ESI/[VLANs in VLAN Bundle] in case of VLAN Bundle on each participating PE.

Observe that currently the VLANs are derived from local configuration and the FSM does not provide any protection against misconfiguration where same EVI,ESI combination has different set of VLANs on different participating PEs or one of the PEs elects to consider VLANs as VLAN bundle and another as separate VLANs for election purposes (service type mismatch).

The FSM is normative in the sense that any design or implementation MUST behave towards external peers and as observable external behavior (DF) in a manner equivalent to this FSM.

```
                                           LOST_ES
                 RCVD_ES                   RCVD_ES
                 LOST_ES                   +-----+
                 +-----+                   |     v
                 |     |                  ++----++   RCVD_ES
                 |   +-+----+    ES_UP    |  DF  +<---------+
                 +->+ INIT +--------------> WAIT |          |
                    ++-----+              +----+-+          |
                     ^                         |            |
+------------+       |                         |DF_TIMER    |
| ANY STATE +--------+          VLAN_CHANGE    |            |
+------------+ ES_DOWN    +------------------+    |            ^
                          |    LOST_ES       v    v            |
                    +-----++              ++----+-+            |
                    |  DF  |              |  DF  +----------+
                    | DONE +<-------------+ CALC +v-+       |
                    +-+----+   CALCULATED  +----+-+   |       |
```

```
                   |                        |    |       |
                   |                        +----+       |
                   |                        LOST_ES      |
                   |                        VLAN_CHANGE  |
                   |                                     |
                   +-------------------------------------+
```

Figure 3

States:

1. INIT: Initial State

2. DF WAIT: State in which the participants waits for enough information to perform the DF election for the EVI/ESI/VLAN combination.

3. DF CALC: State in which the new DF is recomputed.

4. DF DONE: State in which the according DF for the EVI/ESI/VLAN combination has been elected.

Events:

1. ES_UP: The ESI has been locally configured as 'up'.

2. ES_DOWN: The ESI has been locally configured as 'down'.

3. VLAN_CHANGE: The VLANs configured in a bundle that uses the ESI changed. This event is necessary for VLAN bundles only.

4. DF_TIMER: DF Wait timer has expired.

5. RCVD_ES: A new or changed Ethernet Segment Route is received in a BGP REACH UPDATE. Receiving an unchanged UPDATE MUST NOT trigger this event.

6. LOST_ES: A BGP UNREACH UPDATE for a previously received Ethernet Segment route has been received. If an UNREACH is seen for a

route that has not been advertised previously, the event MUST NOT be triggered.

7. CALCULATED: DF has been succesfully calculated.

According actions when transitions are performed or states entered/ exited:

1. ANY STATE on ES_DOWN: (i)stop DF timer (ii) assume non-DF for local PE

2. INIT on ES_UP: (i)do nothing

3. INIT on RCVD_ES, LOST_ES: (i)do nothing

4. DF_WAIT on entering the state: (i) start DF timer if not started already or expired (ii) assume non-DF for local PE

5. DF_WAIT on RCVD_ES, LOST_ES: do nothing

6. DF_WAIT on DF_TIMER: do nothing

7. DF_CALC on entering or re-entering the state: (i) rebuild according list and hashes and perform election (ii) FSM generates CALCULATED event against itself

8. DF_CALC on LOST_ES or VLAN_CHANGE: do nothing

9. DF_CALC on RCVD_ES: do nothing

10. DF_CALC on CALCULATED: (i) mark election result for VLAN or bundle

11. DF_DONE on exiting the state: (i)if RFC7432 election or new election and lost primary DF then assume non-DF for local PE for VLAN or VLAN bundle.

12. DF_DONE on VLAN_CHANGE or LOST_ES: do nothing

8. Auto-Derivation of ES-Import Route Target

Section 7.6 of RFC7432 describes how the value of the ES-Import Route Target for ESI types 1, 2, and 3 can be auto-derived by using the high-order six bytes of the nine byte ESI value. This document extends the same auto-derivation procedure to ESI types 0, 4, and 5.

9. Operational Considerations

TBD.

10. Security Considerations

This document raises no new security issues for EVPN.

11. Acknowledgements

The authors would like to thank Tamas Mondal, Sami Boutros, Jakob Heitz, Jorge Rabadan and Patrice Brissette for useful feedback and discussions.

12. References

12.1. Normative References

[HRW1999] Thaler, D. and C. Ravishankar, "Using Name-Based Mappings to Increase Hit Rates", IEEE/ACM Transactions in networking Volume 6 Issue 1, February 1998.

[I-D.ietf-idr-extcomm-iana] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", draft-ietf-idr-extcomm-iana-02 (work in progress), December 2013.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <http://www.rfc-editor.org/info/rfc4360>.

[RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <http://www.rfc-editor.org/info/rfc4761>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

12.2. Informative References

[CHASH] Karger, D., Lehman, E., Leighton, T., Panigrahy, R., Levine, M., and D. Lewin, "Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web", ACM Symposium on Theory of Computing ACM Press New York, May 1997.

[CLRS2009] Cormen, T., Leiserson, C., Rivest, R., and C. Stein, "Introduction to Algorithms (3rd ed.)", MIT Press and McGraw-Hill ISBN 0-262-03384-4., February 2009.

[RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, DOI 10.17487/RFC2991, November 2000, <http://www.rfc-editor.org/info/rfc2991>.

[RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path

Algorithm", RFC 2992, DOI 10.17487/RFC2992, November 2000, <http://www.rfc-editor.org/info/rfc2992>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <http://www.rfc-editor.org/info/rfc4364>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <http://www.rfc-editor.org/info/rfc4684>.

[RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, DOI 10.17487/RFC6624, May 2012, <http://www.rfc-editor.org/info/rfc6624>.

Authors' Addresses

Satya Ranjan Mohanty
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
USA

Email: satyamoh@cisco.com

Keyur Patel
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
USA

Email: keyupate@cisco.com

Ali Sajassi
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134

USA

Email: sajassi@cisco.com

John Drake
Juniper Networks, Inc.
1194 N. Mathilda Drive
Sunnyvale, CA 95134
USA

Email: jdrake@juniper.com

Antoni Przygienda
Ericsson
300 Holger Way
San Jose, CA 95134
USA

Email: antoni.przygienda@ericsson.com