BESS WG
Internet-Draft
Intended status: Standards Track
Expires: 31 July 2024

S. Dikshit
Aruba, HPE
V. Joshi
Oracle India Pvt Ltd

28 January 2024

All PEs as DF
draft-saumvinayak-bess-all-df-bum-07

Abstract

The Designated forwarder concept is leveraged to prevent looping of BUM traffic into tenant network sourced across NVO fabric for multihoming deployments. [RFC7432] defines a preliminary approach to select the DF for an ES,VLAN or ES,Vlan Group, panning across multiple NVE’s. [RFC8584] makes the election logic more robust and fine grained by inculcating fair election of DF handling most of the prevalent use-cases. This document presents a deployment problem and a corresponding solution which cannot be easily resolve by rules mentioned in [RFC7432] and [RFC8584]. It involves redundant firewall deployment on disparate overlay sites connected over WAN. The requirement is to allow reachability, ONLY, to the local firewall, unless there is an outage. In case of outage the reachability can be extended to remote site’s firewall over WAN.

Status of This Memo

Copyright Notice

Table of Contents

1. Important Terms

DF: Designated Forwarder as defined in [RFC7432].

VTEP: Virtual Tunnel End Point or Vxlan Tunnel End Point

2. Introduction

The Designated forwarder concept is leveraged to prevent looping of BUM traffic into tenant network sourced across NVO fabric for multihoming deployments. [RFC7432] defines a preliminary approach to select the DF for an ES,VLAN or ES,Vlan Group, panning across multiple NVE's. [RFC8584] makes the election logic more robust and fine grained by inculcating fair election of DF handling most of the prevalent use-cases. This document presents a deployment problem and a corresponding solution which cannot be easily resolve by rules mentioned in [RFC7432] and [RFC8584]. It involves redundant firewall deployment on disparate overlay sites connected over WAN. The requirement is to allow reachability, ONLY, to the local firewall, unless there is an outage. In case of outage the reachability can be extended to remote site’s firewall over WAN.

3. 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].

When used in lowercase, these words convey their typical use in common language, and they are not to be interpreted as described in [RFC2119].

4. Problem Description

It’s a typical deployment case of Firewall devices, also configured as default gateway for NVO fabric. The default gateways inturn redirects traffic to firewalls over a shared vlan. This example, for simplicity, assumes the former case wherein firewall is also configured as default gateway for all VLANs on the site (SITE-1 and SITE-2).

All PEs(Vtep1 and Vtep2 in below example) in the diagram are attached to same ES and both intend to act act as DF for the broadcast domain (BD-1) for their respective sites. As already mentioned, this is a typical case of firewall-gateways (active/active) across fabrics (sites). The preferred firewall-gateway is the one local to the site. All ARP broadcast request generated for the gateway are directed to the local firewall and NOT to the remote one.

Whereas, upon failure of the local firewall, all packets orginating from the affected site, including broadcast packets like ARP requests, need to be redirected (over WAN, via DCI/VPN) towards the remote site firewall. The firewall-device is connected to it’s first-hop vtep over the same bridge-domain and same ESI across all sites.

All in all, it’s an emulated multi-homing scenario. This is a scenario of firewall devices hosting same(IP and MAC) credentials.

4.1. Problem Example

The following details out the problem further. There are two sites, SITE-1 and SITE-2 in the below diagram. Traffic (including BUM) generated by Host1 (in SITE-1) (for a bridge-domain) should run through site-local firewall instance (firewall_1) preferably and should not be leaked to the remote sites.

Only in case of local-outage, the traffic should be send across over WAN to the remote firewall (firewall_2). Same should apply to traffic generated by Host2 (in SITE-2), wherein, it should ONLY run through the local firewall (firewall_2), unless there is local-firewall. In that case should go over the WAN towards remote sites firewall, firewall_1.

Vtep1 and/or Vtep2 learn the firewall MAC (MAC_F) as a local host learning and also from the remote vteps, Vtep2 and Vtep1, respectively. But since both the learnings are over the same ESI, it should not lead to MAC move. Cometh the local firewall failure, Vtep1 or Vtep2 should start redirecting the traffic to remote SITE firewall, firewall_2 and firewall_1, respectively. Any ARP request (BUM traffic) for firewall credentials landing at either Vtep1 or Vtep2 from the remote fabric, should then be flooded to network or LAN towards the locally connected firewall.

```
         SITE-1          |          SITE-2
------------------------------------------------------
      Host1                             Host2
       \                                  /
        \                               /
Vtep_host1                             Vtep_host2
         |                             |
         |       [ EVPN-fabric ]       |
         |                             |
     Vtep1============WAN==============Vtep2
        /                                 \
       /                                   \
Firewall _1                           Firewall_2
  (MAC_F)                              (MAC_F)
```

Figure 1: Figure 1: Active-Active Firewall Across Sites

5. Solution(s)

The control plane part of the solution can be leveraged from the 'DF Election Extended Community' described in [RFC8584]. Since the requirement is to ensure all the PEs attached to ESI, forward the BUM traffic arriving from hosts connected to local NVO fabric towards the Attachment circuits (ACs), that are configured over the ES for a BD (broadcast or bridge domain) mapped to Vlan or bundle of Vlans. As explained in the above section, that this is a case, where PEs are in disparate networks and the ACs behind them are not connected to common physical device, even though they are part of the same ES. The diagram gives an overview of the network or deployment in contention.

This document proposes a new mode of DF-election called 'ALL-PEs-DF', where-in, all the pariticipating PEs, intend to play DF role for subset of vlan(s) enabled on an ESI. This requires "DF Election Extended Community" to carry this information with the ES route to indicate it to remote PEs. This ensures all PEs receiving BUM traffic over NVO fabric destined to ESI, BD, SHOULD flood it on the associated ES on the access/tenant side. A PE device MAY be explicitly configured to choose the ALL-PEs-DF mode.

5.1. Sending All PEs are DF mode

The All-PEs-DF mode is used as follows:

(1) PEs configured to use ALL-PEs-DF mode SHOULD set "DF Alg" algorithm field in 'DF Election Extended Community' to appropriate value.

(2) This document proposes value TBD for All-PEs-DF mode, as values '0', '1' and '2' are already reserved for usage.

(3) This algorithm is agnostic to the values carried in 'Bitmap' but does not discounts any use-case(s) in future which may need extra information carried in 'Bitmap' along with All-PEs-DF mode.

5.2. Receive All PEs are DF mode

When a PE receives the ES routes from all the other PEs, for the ES in question, carrying the ALL-PEs-DF mode set, in 'DF Election Extended Community', it SHOULD, check to see if all the advertisements have the Extended Community with 'All-DF-mode' set as 'DF Algo'. If yes, then PE SHOULD ignore the 'Bitmap' and 'Rsvd' field in the extended community. As also mentioned in [RFC8584] , even if, a single advertisement for Route Type 4 is received without the locally configured DF Alg and capability, the default DF election algorithm MUST be used as mandated in [RFC7432].

5.3. Example of algorithm

The BGP-EVPN control plane extension, as mentioned in this document, helps in resolving the problem decsribed in Section 4. If PEs, Vtep1 and Vtep2 are configured to use ALL-PEs-DF mode, then any BUM traffic from respective local hosts Host1/Host2 connected to the EVPN fabric, SHOULD get redirected towards the AC for the ESI,Vlan to which the firewall_1/firewall_2 (respectively) is attached. For example the arp-request for the Firewall IP will be honored by the Firewall_1 behind the Vtep1 which receives the ARP-request. Whereas, when Vtep2 receives the arp-request it will be honored by Firewall_2. Vtep1 and Vtep2 will publish the arp-request in their respective ACs attached to the firewall on which Vlan,ESI is configured and enabled

6. Interoperability with other Algos

Since All-DF-algo is special mode and not exactly an algorithm, which requires the participation of all PEs for an ESI, VLAN. Hence, even if one PE publishes an algo which is NOT "All-DF-mode", other PEs SHOULD revert back to default algorithm. The reason being that, if there are PE1, PE2, PE3 and PE4 in contention. PE1 and PE2 publishes DF Algo 'ALL-PEs-DF', PE3 publishes '0' and PE4 publishes '1'. Once this mismatch is perceived, all PEs SHOULD try and converge towards the default mode. An admin intervention may be required to achieve the same or to converge on any other supported 'DF Algo'.

7. Backward Compatibility

As prescribed in [RFC8584], PEs not supporting (hence not publishing) 'ALL-PEs-DF', SHOULD ignore the processing of the 'DF Election Extended Community' and SHOULD indulge in DF-election using the default aglorithm mentioned in [RFC7432]. The PEs configured with this new alogrithm (hence publishing it), if receive Route Type 4 without 'DF Election Extended Community', SHOULD also revert back to default algorithm. If PEs receive Route Type 4 with another algorithm published in 'DF Election Extended Community', then it should follow procedures prescribed in Section 6.

8. Impact on Local Bias

There is no impact on the local-bias handling, as the PE receiving the BUM from access side over {ESI, VLAN} and relays it to other PEs that published {ESI, VLAN} in Route Type 4; the receiving side PEs will not relay it to EVPN fabric nor will they redirect it to same ESI configured with same VLAN on the access/tenant side.

9. Security Considerations

This document inherits all the security considerations discussed in [RFC7432] and [RFC8584].

10. IANA Considerations

IANA considerations yet to be concluded as the value of mode proposed here is still under discussion.

11. Acknowledgements

The authors want to thank Jorge Rabadan and Luc Andre for their valuable comment. They also advised on other potential solution while helping in paraphrasing the problem statement.

12. References

12.1. Normative References

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

12.2. Informative References

[RFC7348] Mahalingam, M., "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, August 2014, <http://www.rfc-editor.org/rfc/rfc7348.txt>.

[RFC7432] Sajassi, A., "BGP MPLS-Based Ethernet VPN", RFC 7432, February 2015, <http://www.rfc-editor.org/rfc/rfc7432.txt>.

[RFC8584] Rabadan, J., "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, April 2019, <http://www.rfc-editor.org/rfc/rfc8584.txt>.

[RFC9014] Rabadan, J., "Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks", RFC 9014, May 2021, <http://www.rfc-editor.org/rfc/rfc9014.txt>.

Authors’ Addresses

Saumya Dikshit
Aruba Networks, HPE
Mahadevpura

Bangalore 560 048
Karnataka
India
Email: saumya.dikshit@hpe.com

Vinayak Joshi
Oracle India Pvt Ltd
Bangalore
Karnataka
India
Email: vinayak.j.joshi@oracle.com