Workgroup: BESS WG
Published: 30 July 2021
Intended Status: Standards Track
Expires: 31 January 2022
Authors: Y. Wang R. Chen
ZTE Corporation ZTE Corporation

Egress Protection for EVPN BUM

Abstract

When the procedures per [I-D.ietf-rtgwg-srv6-egress-protection] are applied to EVPN services, the egress-protection on PLR is typically more faster than the new DF node of the affected ESI is re-elected. As a result of that, replicated BUM packets will be sent to the CEs of the affected ES. This draft describes a "NDF-bias" mechanism that is used to avoid such excessive BUM packets.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on 31 January 2022.

Table of Contents

1. Introduction

A principal feature of EVPN is the ability to support multi-homing from a customer equipment (CE) to multiple PE with Ethernet segment (ES) links. This draft leverages the egress protection mechanism per [I-D.ietf-rtgwg-srv6-egress-protection] in the multi-homed cases and enhance EVPN convergency on the egress PE node failures, especially for the convergency of BUM packets (Section 3.3).

1.1. Terminology and Acronyms

Most of the acronyms and terms used in this documents comes from [RFC7432], [RFC8679], [I-D.ietf-bess-srv6-services] and [I-D.ietf-rtgwg-srv6-egress-protection] except for the following:

*Mirrored SID - A mirrored SID is a VPN SID which will be installed in the context of a mirror SID (as per [I-D.ietf-rtgwg-srv6-egress-protection]).

*bypassed BUM packet - A BUM packet that is received via a mirrored End.DT2M SID.

*EVPN SID - SRv6 SID for EVPN Instances, e.g. End.DT2M SID, End.DT2U SID, End.DX2 SID, End.DX2V SID.

*NDF - non-DF, non Designated-Forwarder. Note that Backup DF is a special NDF.

*NDF-Bias - An exception for filtering bypassed BUM packets. It says that when an outgoing AC is a DF on its ES, the bypassed BUM

packets will be dropped but when an outgoing AC is a NDF on its ES, the bypassed BUM packets will not be dropped.

*Pairing Route - When IMET_PE3 and IMET_PE4 are two IMET routes of same Bridge Domain, and IMET_PE4 is a local IMET route, while IMET_PE3 is a remote IMET route whose End.DT2M SID is allocated from a remote locator that is protected by a local Mirror SID, we say that IMET_PE4 are paired with IMET_PE3 by the Mirror SID, or just say that IMET_PE4 is IMET_PE3's pairing route. Note that we don't say IMET_PE3 is IMET_PE4's pairing route, because that a local route will have many pairing routes, one per local mirror SID.

*ORIP - Originating Router's IP Address, or Originator's IP Address.

2. Problem Statement

Figure 1 shows an example of protecting egress PE3 of a SR path, which is from ingress PE1 to egress PE3.

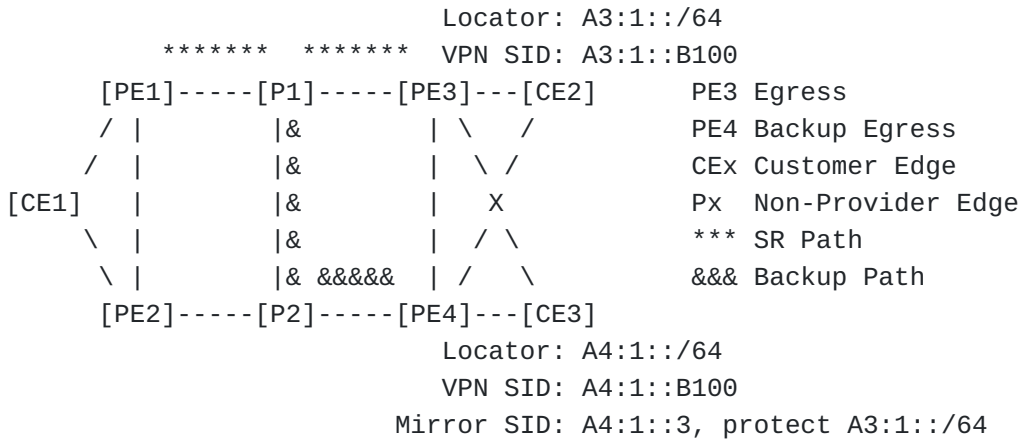

Figure 1: Egress Protection Scenario

In Figure 2, All PEs are EVPN PEs per [RFC7432], and Both CE2 and CE3 are dual homed to PE3 and PE4. PE3 has a locator A3:1::/64 and a End.DT2M SID A3:1::B100. PE4 has a locator A4:1::/64 and a End.DT2M SID A4:1::B100. A Mirror SID A4:1::3 is configured on PE4 for protecting a remote locator (PE3's locator) A3:1::/64. P1 has a locator A1:1::/64. CE2 is on Ethernet Segment ES2 whose DF is PE4, non-DF is PE3. CE3 is on Ethernet Segment ES3 whose DF is PE3, non-DF is PE4. Both ES2 and ES3 are all-active mode.

After the configuration, PE4 advertises this information through an IGP LS (i.e., LSA in OSPF or LSP in IS-IS), which includes PE3's locator and Mirror SID A4:1::3. Every node in the SR domain will

receive this IGP LS, which indicates that PE4 wants to protect PE3's locator with Mirror SID A4:1::3.

When PE4 (e.g., BGP on PE4) receives a IMET route IMET_PE3 whose VPN SID belongs to PE3 that is protected by PE4 through Mirror SID A4:1::3, it finds IMET_PE3's VPN SID corresponding to a remote locator (A3:1::/64) which is protected by a local Mirror SID (A4:1::3). Then PE4 find out the corresponding local IMET route (say IMET_PE4) which are paired with IMET_PE3 by the Mirror SID.

Note that although PE4 don't have a local IMET route of IMET_PE3, It can know that the IMET_PE3 is for the same EVPN VPLS of a local IMET route IMET_PE4. Because that the IMET_PE3 will be imported into the EVPN VPLS of the IMET_PE4. So the IMET_PE3 is the corresponding route of IMET_PE4 from the Mirror SID's point of view. In other words, we say that IMET_PE4 are paired with IMET_PE3 by the Mirror SID, or just say that IMET_PE4 is IMET_PE3's pairing route. This procedure is different from [I-D.ietf-rtgwg-srv6-egress-protection].

The forwarding behaviors for the VPN SID (PE3's EVPN SID A3:1::B100 of End.DT2M Function) of IMET_PE3 are the same as the VPN SID (A4:1::B100) of the local IMET_PE4 from function's point of view. If the behavior for PE3's VPN SID in PE3 forwards the packet with it to CE2, then the behavior for PE4's VPN SID in PE4 forwards the packet to the same CE2; and vice versa. PE4 creates a forwarding entry for PE3's VPN SID A3:1::B100 in the FIB table identified by Mirror SID A4:1::3 according to the forwarding behavior for PE4's VPN SID A4:1::B100.

Note that there are two FIB entries for A3:1:B100, one(say FIB_1) is on PE3 , the other(say FIB_2) is on PE4. We call the FIB entry FIB_2 as the FIB entry FIB_1's mirrored FIB entry. Because that the FIB_2 is in the FIB table identified by a Mirror SID. The SID instance corresponding to a mirrored FIB entry is called as a Mirrored SID.

Node P1's pre-computed backup path for destination PE3`s locator is from P1 to PE4 having mirror SID A4:1::3. When P1 receives a packet destined to PE3's VPN SID A3:1::B100, in normal operations, it forwards the packet with source A1:1:: and destination PE3's VPN SID A3:1::B100 according to the FIB using the destination PE3's VPN SID A3:1::B100.

When PE3 fails, P1 as PLR sends the packet to PE4 via the backup path pre-computed. P1 encapsulates the packet using H.Encaps before sending it to PE4.

When PE4 receives the re-routed packet, it decapsulates the packet and forwards the decapsulated packet by executing End.DT6 behavior for an End.DT6 SID instance. The SID instance is End.M, the Mirror

SID that is associated with the IPv6 FIB table for PE3. The packet received by PE4 is (T, Mirror SID A4:1::3) (A1:1::, PE3's VPN SID A3:1::B100)Pkt0.

PE4 obtains Mirror SID A4:1::3 in the outer IPv6 header of the packet, removes this outer IPv6 header, and then processes the inner IPv6 packet (A1:1::, A3:1::B100)Pkt0. It finds the FIB table for PE3 using Mirror SID A4:1::3 as the context ID, gets the forwarding entry for PE3's VPN SID A3:1::B100 from the table, and forwards the packet to CE2 using the entry.

When the SID instance A3:1::B100 is End.DT2M, the Pkt0 is a BUM packet. The Pkt0 will be replicated to CE2 as long as PE4 is the DF of <ES2,EVI1> The Pkt0 will be replicated to CE3 as long as PE4 is the DF of <ES3,EVI1>

Typically, the local-repair on P1 is more faster than the new DFs of ES2 and ES3 are re-elected. So PE4 will still be the DF of <ES2, EVI1> untill the DF re-election finishes, and PE4 will still be the non-DF of <ES3, EVI1> untill DF the re-election finishes,

Note that PE1 will broadcast a BUM packet (say BUM0) to PE3 and PE4 separately. One copy of BUM0 which is replicated by PE1 for PE3 is called as BUM3. The other copy of BUM0 which is replicated by PE1 for PE4 is called as BUM4.

As a result of current DF filtering rules, CE2 will receive both BUM3 and BUM4 untill the DF re-election finishes. At the same time, CE3 will receive niether BUM3 nor BUM4.

3. Solutions

3.1. Solution 1: Separate Locators for End.DT2M SIDs

The End.DT2M SIDs are not allocated from the same locator with the End.DT2U/End.DX2/End.DX2V SIDs. The locator from where the End.DT2M SIDs are allocated are called as DT2M_LOCATOR. The DT2M_LOCATOR must not be protected by any Mirror SID.

This solution can prevent CE2 from receiving excessive BUM packets, but can not help to accelerate the convergence of CE3's BUM packets.

3.2. Solution 2: The mirrored End.DT2M SIDs are not installed

When the SID instance A3:1::B100 is not installed on PE4, BUM4 will be dropped after PLR's egress protection procedures.

Note that a mirrored SID is very different from a Mirror SID. A mirrored SID is a VPN SID which will be installed in the context of

a mirror SID as per [I-D.ietf-rtgwg-srv6-egress-protection]. But it must not be installed in this solution.

This solution can prevent CE2 from receiving excessive BUM packets, but can not help CE3 to receive one copy of BUM0 as soon as possible.

3.3. Solution 3: NDF-Bias Mechanism

In order to accelerate the convergence of bypass-BUM packets, some specific rules are defined in the following6:

The mirrored End.DT2M SIDs need to be installed, and those BUM packets which is received via a mirrored End.DT2M SID are called as bypass BUM packets. Those bypass BUM packets will not be dropped when they are about to be forwarded to an AC whose DF-role is NDF. Note that the single-homing ACs are always considered as DF-role, so those ACs will filter all bypass BUM packets. Accordingly, those bypass BUM packets will be dropped when they are about to be forwarded to an AC whose DF-role is DF.

These rules are called as "NDF-Bias" rules in this draft.

This solution can prevent CE2 from receiving excessive BUM packets, at the same time, this solution can accelerate the convergence of CE3's BUM packets.

4. Condisderations on EVPN Mulicast and Other EVPNs

4.1. Condisderations on EVPN Mulicast

The pairing routes for EVPN multicast are two EVPN routes of the same <Multicast source,Multicast Group,BD>. And they are of the same EVPN route type, while they have different ORIPs.

4.2. Condisderations on Other EVPNs

Just the recognition methods to pick the bypassed BUM packets out are different.

*MPLS EVPN - The bypassed BUM packets will be received over an ILM entry in a context-specific label-space.

*VXLAN EVPN - see [I-D.wang-bess-evpn-egress-protection].

5. IANA Considerations

no IANA Considerations.

6. Security Considerations

TBD.

7. References

7.1. Normative References

[I-D.ietf-rtgwg-srv6-egress-protection]
Hu, Z., Chen, H., Chen, H., Wu, P., Toy, M., Cao, C., He, T., Liu, L., and X. Liu, "SRv6 Path Egress Protection", Work in Progress, Internet-Draft, draft-ietf-rtgwg-srv6-egress-protection-03, 28 May 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-rtgwg-srv6-egress-protection-03>.

[I-D.ietf-bess-srv6-services]
Dawra, G., Filsfils, C., Talaulikar, K., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "SRv6 BGP based Overlay Services", Work in Progress, Internet-Draft, draft-ietf-bess-srv6-services-07, 11 April 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-srv6-services-07>.

[RFC8679] Shen, Y., Jeganathan, M., Decraene, B., Gredler, H., Michel, C., and H. Chen, "MPLS Egress Protection Framework", RFC 8679, DOI 10.17487/RFC8679, December 2019, <https://www.rfc-editor.org/info/rfc8679>.

[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, <https://www.rfc-editor.org/info/rfc7432>.

7.2. Informative References

[I-D.wang-bess-evpn-egress-protection] Wang, Y. and R. Chen, "EVPN Egress Protection", Work in Progress, Internet-Draft, draft-wang-bess-evpn-egress-protection-04, 29 October 2020, <https://datatracker.ietf.org/doc/html/draft-wang-bess-evpn-egress-protection-04>.

Authors' Addresses

Yubao Wang
ZTE Corporation
No.68 of Zijinghua Road, Yuhuatai Distinct
Nanjing
China

Email: wang.yubao2@zte.com.cn

Ran Chen
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: chen.ran@zte.com.cn