BIER WG
Internet-Draft
Intended status: Standards Track
Expires: April 13, 2016

C. Wang
Z. Zhang
F. Hu
ZTE Corporation

October 11, 2015

BIER Use Case in VxLAN
draft-wang-bier-vxlan-use-case-00

Abstract

Bit Index Explicit Replication (BIER) is an architecture that provides optimal multicast forwarding through a "BIER domain" without requiring intermediate routers to maintain any multicast related per-flow state. BIER also does not require any explicit tree-building protocol for its operation. A multicast data packet enters a BIER domain at a "Bit-Forwarding Ingress Router" (BFIR), and leaves the BIER domain at one or more "Bit-Forwarding Egress Routers" (BFERs). The BFIR router adds a BIER header to the packet. The BIER header contains a bit-string in which each bit represents exactly one BFER to forward the packet to. The set of BFERs to which the multicast packet needs to be forwarded is expressed by setting the bits that correspond to those routers in the BIER header.

This document tries to describe the drawbacks of how BUM services are deployed in current data centers, and proposes how to take full advantage of BIER to implement BUM services in data centers.

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 http://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 April 13, 2016.

Copyright Notice

Table of Contents

1. Introduction 3
2. Convention and Terminology 5
3. BIER in data centers 6
4. BIER IS-IS extension for VXLAN-specific information 7
5. BIER OSPF extension for VXLAN-specific information 9
6. BIER BGP extension for VXLAN-specific information 10
7. Considerations on BIER in data centers 11
8. Security Considerations 12
9. IANA Considerations 13
10. References 14
10.1. Normative References 14
10.2. Informative References 14
Authors' Addresses 16

1. Introduction

This document is motivated by [I-D.ietf-bier-use-cases].

In current data center virtualization, virtual eXtensible Local Area Network (VXLAN) [RFC7348] is a kind of network virtualization overlay technology which is overlaid between NVEs and is intended for multi-tenancy data center networks, whose reference architecture is illustrated as per Figure 1.

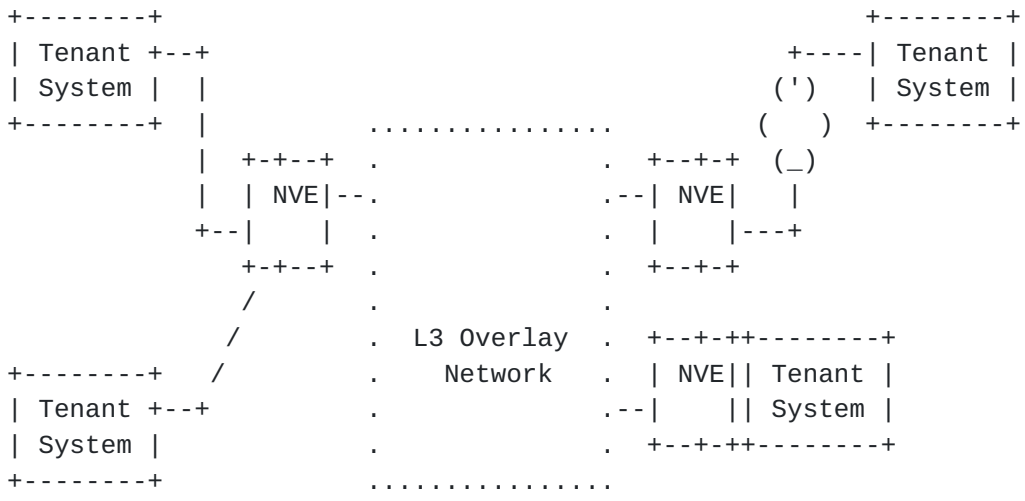

Figure 1: NVO3 Architecture

And there are two kinds of most common methods about how to forward BUM packets in this virtualization overlay network. One is using PIM as underlay multicast routing protocol to build explicit multicast distribution tree, such as PIM-SM[RFC4601] or PIM-BIDIR [RFC5015]multicast routing protocol. Then, when BUM packets arrive at NVE, it requires NVE to have a mapping between the VXLAN Virtual Network Instances (VNI) and the IP multicast group. According to the mapping, NVE can encapsulate BUM packets in a multicast packet which group address is the mapping IP multicast group address and steer them through explicit multicast distribution tree to the destination NVEs. This method has two serious drawbacks. It need the underlay network supports complicated multicast routing protocol and maintains multicast related per-flow state in every transit nodes. What!_s more, how to configure the ratio of the mapping between VNI and IP multicast group is also an issue. If the ratio is 1:1, there should be 16M multicast groups in the underlay network at maximum to map to the 16 M VNIs, which is really a significant challenge for the data center devices. If the ratio is n:1, it would result in inefficiency

bandwidth utilization which is not optimal in data center networks.

The other method is using ingress replication to require each NVE to create a mapping between the VXLAN Virtual Network Instances (VNI) and the remote NVEs!_ addresses which belong to the same virtual network. When NVE receives BUM traffic from the attached tenant, NVE can encapsulate these BUM packets in unicast packets and replicate them and tunnel them to different remote NVEs respectively. Although this method can eliminate the burden of running multicast protocol in the underlay network, it has a significant disadvantage: large waste of bandwidth, especially in big-sized data center where there are many receivers.

Bit Index Explicit Replication (BIER) [I-D.ietf-bier-architecture] is an architecture that provides optimal multicast forwarding through a "BIER domain" without requiring intermediate routers to maintain any multicast related per-flow state. BIER also does not require any explicit tree-building protocol for its operation. A multicast data packet enters a BIER domain at a "Bit-Forwarding Ingress Router" (BFIR), and leaves the BIER domain at one or more "Bit-Forwarding Egress Routers" (BFERs). The BFIR router adds a BIER header to the packet. The BIER header contains a bit-string in which each bit represents exactly one BFER to forward the packet to. The set of BFERs to which the multicast packet needs to be forwarded is expressed by setting the bits that correspond to those routers in the BIER header.

The following section tries to proposes how to take full advantage of BIER to implement BUM services in data centers.

2. Convention and 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 [RFC2119].

The terms about BIER are defined in [I-D.ietf-bier-architecture].

The terms about NVO3 are defined in [RFC7365].

Here tries to list the most common terminology mentioned in this draft.

BIER: Bit Index Explicit Replication(Bit Index Explicit Replication (The overall architecture of forwarding multicast using a Bit Position).

NVE: Network Virtualization Edge, which is the entity that implements the overlay functionality. An NVE resides at the boundary between a Tenant System and the overlay network.

VXLAN: Virtual eXtensible Local Area Network

VNI: VXLAN Network Identifier

3. BIER in data centers

This section tries to describe how to use BIER as an optimal scheme to forward the broadcast, unknown and multicast (BUM) packets when they arrive at the NVE.

The principle of using BIER to forward BUM traffic is that: it requires each NVE to have a mapping between the VXLAN Virtual Network Instances (VNI) and the bit-string in which each bit represents exactly one remote NVE to forward the packet to. On other words, this requires the underlay network to support BIER which already be elaborated in [I-D.ietf-bier-architecture].

Already mentioned above, BIER requires no explicit tree-building protocols and maintains no multicast related per-flow state on the end nodes and intermediate nodes, just extends the IGP protocol or BGP protocol to advertise BIER-specific information to form BIER forwarding table in the BIER forwarding routers, such as NVEs and intermediate nodes in the data centers.

More importantly, as for how each NVE knows the other remote NVEs that belong to the same virtual network can also be discovered by additional BIER extensions. The following sections describe how to extend IGP protocol and BGP protocol to advertise VXLAN-specific information to tell each NVE where the other NVEs are in the same virtual network. As a result of this advertisement, each NVE creates the mapping between the VXLAN Virtual Network Instances (VNI) and the bit-string in which each bit represents exactly one remote NVE to forward the packet to.

4. BIER IS-IS extension for VXLAN-specific information

Specifically, in [I-D.ietf-bier-isis-extensions], there defines a new BIER Info sub-TLV which is illustrated in Figure 2. Here, extending a VXLAN-specific sub-sub-TLV to current BIER Info sub-TLV for IS-IS, a reference format is illustrated in Figure 3.

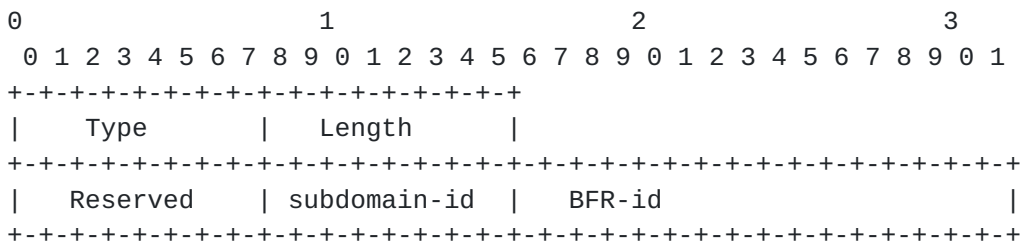

Figure 2: IS-IS BIER Info sub-TLV extensions for BIER-specific information

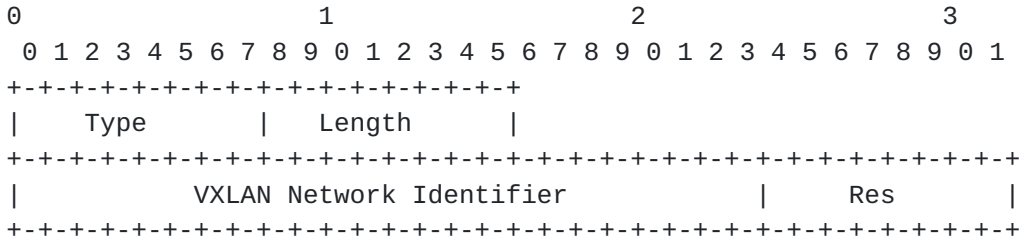

Figure 3: IS-IS VXLAN sub-sub-TLV extensions for VXLAN-specific information

Type:

indicates VXLAN sub-sub-TLV

Length: 1 cotet.

VXLAN Network Idenfifier:

indicates a virtual subnet

Then NVEs and intermediate nodes flood this VXLAN-specific sub-sub-TLV together with BIER Info sub-TLV through IS-IS in overlay network. When one NVE receives this IS-IS advertisement, this NVE builds a mapping between the receiving VNI in the VXLAN-specific sub-sub-TLV and the bit-string which represents the sending NVE and can extract

from the BIER Info sub-TLV. Once this NVE receives some other IS-IS advertisements which include the same VXLAN-specific sub-sub-TLV, it updates the bit-string in the mapping and adds the corresponding sending NVEs to the updated bit-string.

After finishing the above IS-IS flooding, each NVE knows where are the remote NVEs in the same virtual network. When receiving BUM traffic from the attached tenant, each NVE knows exactly how to forward this traffic to.

This can be used in both IPv4 network and IPv6 network.

5. BIER OSPF extension for VXLAN-specific information

Specifically, in [I-D.ietf-bier-ospf-bier-extensions], there defines a new BIER Info sub-TLV as well. Here, extending a VXLAN-specific sub-sub-TLV to current BIER Info sub-TLV for OSPF, a reference format is also illustrated in Figure 3.

Then NVEs and intermediate nodes flood this VXLAN-specific sub-sub-TLV together with BIER Info sub-TLV through OSPF in overlay network. When one NVE receives this OSPF advertisement, this NVE builds a mapping between the receiving VNI in the VXLAN-specific sub-sub-TLV and the bit-string which represents the sending NVE and can extract from the BIER Info sub-TLV. Once this NVE receives some other OSPF advertisements which include the same VXLAN-specific sub-sub-TLV, it updates the bit-string in the mapping and adds the corresponding sending NVEs to the updated bit-string.

After finishing the above OSPF flooding, each NVE knows where are the remote NVEs in the same virtual network. When receiving BUM traffic from the attached tenant, each NVE knows exactly how to forward this traffic to.

This can be used in both IPv4 network and IPv6 network.

6. BIER BGP extension for VXLAN-specific information

Specifically, in [I-D.ietf-bier-idr-extensions], there defines a new BGP path attribute referred to as the BIER attribute. Here, extending a VXLAN-specific sub-TLV to current BIER attribute TLV for BGP, a reference format is also illustrated in Figure 3.

Then NVEs and intermediate nodes flood this VXLAN-specific sub-TLV together with BIER attribute TLV through BGP in overlay network. When one NVE receives this BGP attribute, this NVE builds a mapping between the receiving VNI in the VXLAN-specific sub-TLV and the bit-string which represents the sending NVE and can extract from the BIER attribute TLV. Once this NVE receives some other BIER attribute TLV which include the same VXLAN-specific sub-TLV, it updates the bit-string in the mapping and adds the corresponding sending NVEs to the updated bit-string.

After finishing the above BGP advertisement, each NVE knows where are the remote NVEs in the same virtual network. When receiving BUM traffic from the attached tenant, each NVE knows exactly how to forward this traffic to.

This can be used in both IPv4 network and IPv6 network.

7. Considerations on BIER in data centers

TBD

8. Security Considerations

It will be considered in a future revision.

9. IANA Considerations

There need a new Type for VXLAN sub-sub-TLV.

10. References

10.1. Normative References

[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>.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <http://www.rfc-editor.org/info/rfc4601>.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <http://www.rfc-editor.org/info/rfc5015>.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <http://www.rfc-editor.org/info/rfc7348>.

[RFC7365] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for Data Center (DC) Network Virtualization", RFC 7365, DOI 10.17487/RFC7365, October 2014, <http://www.rfc-editor.org/info/rfc7365>.

10.2. Informative References

[I-D.ietf-bier-architecture] Wijnands, I., Rosen, E., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast using Bit Index Explicit Replication", draft-ietf-bier-architecture-02 (work in progress), July 2015.

[I-D.ietf-bier-idr-extensions] Xu, X., Chen, M., Patel, K., Wijnands, I., and T. Przygienda, "BGP Extensions for BIER", draft-ietf-bier-idr-extensions-00 (work in progress), September 2015.

[I-D.ietf-bier-isis-extensions] Ginsberg, L., Aldrin, S., Zhang, J., and T. Przygienda,

"BIER support via ISIS",
draft-ietf-bier-isis-extensions-00 (work in progress),
April 2015.

[I-D.ietf-bier-ospf-bier-extensions]
Psenak, P., Kumar, N., Wijnands, I., Dolganow, A.,
Przygienda, T., Zhang, J., and S. Aldrin, "OSPF Extensions
For BIER", draft-ietf-bier-ospf-bier-extensions-00 (work
in progress), April 2015.

[I-D.ietf-bier-use-cases]
Kumar, N., Asati, R., Chen, M., Xu, X., Dolganow, A.,
Przygienda, T., arkadiy.gulko@thomsonreuters.com, a.,
Robinson, D., and V. Arya, "BIER Use Cases",
draft-ietf-bier-use-cases-01 (work in progress),
August 2015.

Authors' Addresses

Cui Wang
ZTE Corporation
No.50 Software Avenue, Yuhuatai District
Nanjing
China

Email: wang.cui1@zte.com.cn

Zheng Zhang
ZTE Corporation
No.50 Software Avenue, Yuhuatai District
Nanjing
China

Email: zhang.zheng@zte.com.cn

Fangwei Hu
ZTE Corporation

Email: hu.fangwei@zte.com.cn