

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 3 January 2024

P. Pfister
I.J. Wijnands
S. Venaas
Cisco Systems
C. Wang

Z. Zhang
ZTE Corporation
M. Stenberg
2 July 2023

BIER Ingress Multicast Flow Overlay using Multicast Listener Discovery
Protocols
draft-ietf-bier-mld-08

Abstract

This document specifies the ingress part of a multicast flow overlay for BIER networks. Using existing multicast listener discovery protocols, it enables multicast membership information sharing from egress routers, acting as listeners, toward ingress routers, acting as queriers. Ingress routers keep per-egress-router state, used to construct the BIER bit mask associated with IP multicast packets entering the BIER domain.

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 3 January 2024.

Copyright Notice

Copyright (c) 2023 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 (<https://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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Overview	4
4. Applicability Statement	4
5. Querier and Listener Specifications	5
5.1. Configuration Parameters	5
5.2. MLDv2 instances.	6
5.2.1. Sending Queries	6
5.2.2. Sending Reports	7
5.2.3. Receiving Queries	8
5.2.4. Receiving Reports	8
5.3. Packet Forwarding	8
6. BIER MLD/IGMP Extension Type	9
7. Security Considerations	10
8. IANA Considerations	10
9. Acknowledgements	11
10. References	11
10.1. Normative References	11
10.2. Informative References	11
Appendix A. BIER Use Case in Data Centers	13
A.1. Convention and Terminology	14
A.2. BIER in data centers	15
A.3. A BIER MLD solution for Virtual Network information	15
Authors' Addresses	16

1. Introduction

The Bit Index Explicit Replication (BIER - [RFC8279]) forwarding technique enables IP multicast transport across a BIER domain. When receiving or originating a packet, ingress routers have to construct a bit mask indicating which BIER egress routers located within the same BIER domain will receive the packet. A stateless approach would consist of forwarding all incoming packets toward all egress routers, which would in turn make a forwarding decision based on local information. But any more efficient approach would require ingress routers to keep some state about egress routers multicast membership information, hence requiring state sharing from egress routers toward

ingress routers.

This document specifies how to use the Multicast Listener Discovery protocol version 2 [RFC3810] (resp. the Internet Group Management protocol version 3 [RFC3376]) as the ingress part of a BIER multicast flow overlay (BIER layering is described in [RFC8279]) for IPv6 (resp. IPv4). It enables multicast membership information sharing from egress routers, acting as listeners, toward ingress routers, acting as queriers. Ingress routers keep per-egress-router state, used to construct the BIER bit mask associated with IP multicast packets entering the BIER domain.

This document defines an MLDv2 and IGMPv3 extension type, using the extension scheme defined in [I-D.ietf-pim-igmp-mld-extension], that is used to provide BIER specific information about the message originator.

This specification is applicable to both IP version 4 and version 6. It therefore specifies two separate mechanisms operating independently. For the sake of simplicity, the rest of this document uses IPv6 terminology. It can be applied to IPv4 by replacing 'MLDv2' with 'IGMPv3', and following specific requirements when explicitly stated.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

The terms "Bit-Forwarding Router" (BFR), "Bit-Forwarding Egress Router" (BFER), "Bit-Forwarding Ingress Router" (BFIR), "BFR-id" and "BFR-Prefix" are to be interpreted as described in [RFC8279].

Additionally, the following definitions are used:

BIER Multicast Listener Discovery (BMLD): The modified version of MLD specified in this document.

BMLD Querier: A BFR implementing the Querier part of this specification. A BMLD Node MAY be both a Querier and a Listener.

BMLD Listener: A BFR implementing the Listener part of this specification. A BMLD Node MAY be both a Querier and a Listener.

3. Overview

This document proposes to use the mechanisms described in MLDv2 in order to enable multicast membership information sharing from BFERs toward BFIRs within a given BIER domain. BMLD queries (resp. reports) are sent over BIER toward all BMLD Nodes (resp. BMLD Queriers) using modified MLDv2 messages which IP destination is set to a configured 'all BMLD Nodes' (resp. 'all BMLD Queriers') IP multicast address.

By running MLDv2 instances with per-listener explicit tracking, BMLD Queriers are able to map BMLD Listeners with MLDv2 membership states. This state is then used to construct the set of BFERs associated with each incoming IP multicast data packet.

4. Applicability Statement

BMLD runs on top of a BIER Layer and provides the ingress part of a BIER multicast flow overlay, i.e, it specifies how BFIRs construct the set of BFERs for each ingress IP multicast data packet. The BFER part of the Multicast Flow Overlay is out of scope of this document.

The BIER Layer MUST be able to transport BMLD messages toward all BMLD Queriers and Listeners. Such packets are IP multicast packets with a BFR-Prefix as source address, a multicast destination address, and containing a MLDv2 message.

BMLD only requires state to be kept by Queriers, and is therefore more scalable than PIMv2 [RFC7761] in terms of overall state, but is also likely to be less scalable than PIMv2 in terms of the amount of control traffic and the size of the state that is kept by individual routers.

This specification is applicable to both IP version 4 and version 6. It therefore specifies two separate mechanisms operating independently. For the sake of simplicity, this document uses IPv6 terminology. It can be applied to IPv4 by replacing 'MLDv2' with 'IGMPv3', and following specific requirements when explicitly stated.

If multiple BFIRs have connectivity to the same source, a mechanism is needed to determine which BFIR should be the forwarder, that is not specified in this document. As a special case, if BIER is used end-to-end such that sources would be directly connected to the BFIRs, then an election mechanism is needed if there are multiple BFIRs on the same link as the source. One option is to utilize PIM DR Election where the DR is the BIER forwarder, but other election mechanisms could be used. In order to allow quick failover, the BFIRs that are not forwarders should still track BFER interest so that they have the correct state in case they become forwarders.

5. Querier and Listener Specifications

Routers desiring to receive IP multicast traffic (e.g., for their own use, or for forwarding) MUST behave as BMLD Listeners. Routers receiving IP multicast traffic from outside the BIER domain, or originating multicast traffic, MUST behave as BMLD Queriers.

BMLD Queriers (resp. BMLD Listeners) MUST act as MLDv2 Queriers (resp. MLDv2 Listeners) as specified in [RFC3810] unless stated otherwise in this section.

5.1. Configuration Parameters

Both Queriers and Listeners MUST operate as BFIRs and BFERs within the BIER domain in order to send and receive BMLD messages. They MUST therefore be configured accordingly, as specified in [RFC8279].

All Listeners MUST be configured with an 'all BMLD Queriers' multicast address and the BFR-ids of all the BMLD Queriers. This is used by Listeners to send BMLD reports over BIER toward all Queriers. All Queriers MUST be configured to accept BMLD reports sent to this address.

All Queriers MUST be configured with an 'all BMLD Nodes' multicast address and the BFR-ids of all the Queriers and Listeners. This information is used by Queriers to send BMLD queries over BIER toward all BMLD Nodes. All BMLD Nodes MUST be configured to accept BMLD queries sent to this address.

It may be cumbersome to configure the exact set of BFR-ids for Queriers and Listeners. One MAY configure the set of BFR-ids to contain any potentially used BFR-id, perhaps having all bit positions set. There is no harm in configuring unused BFR-ids. Configuring the BFR-ids of additional routers would in most cases cause no harm, as a router would drop the BMLD message unless it is configured as a Querier or a Listener.

Note that BMLD (unlike MLDv2) makes use of per-instance configured multicast group addresses rather than well-known addresses so that multiple instances of BMLD (using different group addresses) can be run simultaneously within the same BIER domain. Configured group addresses MAY be obtained from allocated IP prefixes using [RFC3306]. One MAY choose to use the well-known MLDv2 addresses in one instance, but different instances MUST use different addresses.

IP packets coming from outside of the BIER domain and having a destination address set to the configured 'all BMLD Queriers' or the 'all BMLD Nodes' group address MUST be dropped. It is RECOMMENDED that these configured addresses have a limited scope, enforcing this behavior by scope-based filtering on BIER domain's egress interfaces.

5.2. MLDv2 instances.

BMLD Queriers MUST run a MLDv2 Querier instance with per-host tracking, which means they keep track of the MLDv2 state associated with each BMLD Listener. For that purpose, Listeners are identified by their respective BFR-Prefix, used as IP source address in all BMLD reports.

BMLD Listeners MUST run a MLDv2 Listener instance expressing their interest in the multicast traffic they are supposed to receive for local use or forwarding.

BMLD Listeners and Queriers MUST NOT run the MLDv1 (IGMPv2 and IGMPv1 for IPv4) backward compatibility procedures.

5.2.1. Sending Queries

BMLD Queries are IP packets sent over BIER by BMLD Queriers:

- * Toward all BMLD Nodes (i.e., providing to the BIER Layer the BFR-ids of all BMLD Nodes).
- * Without the IPv6 router alert option [RFC2711] in the hop-by-hop extension header [RFC8200] (or the IPv4 router alert option [RFC2113] for IPv4).
- * With the IP destination address set to the 'all BMLD Nodes' group address.
- * With a deterministic IP source address. It is RECOMMENDED that the address is a BFR-Prefix of the sender, but it MAY be another value. This address is only used for querier election.

- * With a TTL value large enough such that the packet can be received by all BMLD Nodes, depending on the underlying BIER layer (whether it decrements the IP TTL or not) and the size of the network. The default value is 64.
- * The extension type defined in Section 6 MUST be included once, specifying the Sub-domain-id, BFR-id and BFR-Prefix of the sender. This information may be useful for logging and debugging.

5.2.2. Sending Reports

BMLD Reports are IP packets sent over BIER by BMLD Listeners:

- * Toward all BMLD Queriers (i.e., providing to the BIER layer the BFR-ids of all BMLD Queriers).
- * Without the IPv6 router alert option [RFC2711] in the hop-by-hop extension header [RFC8200] (or the IPv4 router alert option [RFC2113] for IPv4).
- * With the IP destination address set to the 'all BMLD Queriers' group address.
- * With a deterministic IP source address. It is RECOMMENDED that the address is a BFR-Prefix of the sender.
- * With a TTL value large enough such that the packet can be received by all BMLD Queriers, depending on the underlying BIER layer (whether it decrements the IP TTL or not) and the size of the network. The default value is 64.
- * The extension type defined in Section 6 MUST be included once, specifying the Sub-domain-id, BFR-id and BFR-Prefix of the sender. This information is used to create the necessary forwarding state for requested flows, and may be useful for logging and debugging.

Since the reports may contain a large number of records, they may become larger than the maximum BIER payload that can be delivered to all the BMLD Queriers. Hence an implementation will need to either use a small default maximum size, allow configuration of a maximum size, or rely on MTU discovery. MTU discovery may be done for a sub-domain using BIER MTU Discovery [I-D.ietf-bier-mtud] or for the set of BMLD Queriers using Path MTU Discovery [I-D.ietf-bier-path-mtu-discovery].

5.2.3. Receiving Queries

BMLD Queriers and Listeners MUST check the destination address of all the IP packets that are received or forwarded over BIER whenever their own BIER bit is set in the packet. If the destination address is equal to the 'all BMLD Nodes' group address the packet is processed as specified in this section.

If the IPv6 (resp. IPv4) packet contains an ICMPv6 (resp. IGMP) message of type 'Multicast Listener Query' (resp. of type 'Membership Query'), and include the extension defined in Section 6), it is processed by the MLDv2 (resp. IGMPv3) instance run by the BMLD Querier. It MUST be dropped otherwise.

During the MLDv2 processing, the packet MUST NOT be checked against the MLDv2 consistency conditions (i.e., the presence of the router alert option, the TTL equaling 1 and, for IPv6 only, the source address being link-local).

5.2.4. Receiving Reports

BMLD Queriers MUST check the destination address of all the IP packets that are received or forwarded over BIER whenever their own BIER bit is set. If the destination address is equal to the 'all BMLD Queriers' the packet is processed as specified in this section.

If the IPv6 (resp. IPv4) packet contains an ICMPv6 (resp. IGMP) message of type 'Multicast Listener Report Message v2' (resp. 'Version 3 Membership Report'), and include the extension defined in Section 6), it is processed by the MLDv2 (resp. IGMPv3) instance run by the BMLD Querier. It MUST be dropped otherwise.

During the MLDv2 processing, the packet MUST NOT be checked against the MLDv2 consistency conditions (i.e., the presence of the router alert option, the TTL equaling 1 and, for IPv6 only, the source address being link-local).

5.3. Packet Forwarding

BMLD Queriers configure the BIER Layer using the information obtained using BMLD, and the extension Section 6), to track membership state, including the Sub-domain-id, BFR-id and BFR-Prefix of the members.

More specifically, the membership state associated with each BMLD Listener is provided to the BIER layer such that whenever a multicast packet enters the BIER domain, if that packet matches the membership information from a BMLD Listener, its Sub-domain-id and BFR-id is added to the set of Sub-domains and BFR-ids the packet should be forwarded to by the BIER-Layer.

6. BIER MLD/IGMP Extension Type

A new MLD/IGMP extension type adds BIER specific information to IGMP/MLD messages, using the extension scheme defined in [I-D.ietf-pim-igmp-mld-extension]). The BIER specific information is the same as the PTA tunnel identifier in [RFC8556] and is shown in Figure 1. Note that, as defined in the MLD (resp. IGMP), existing implementations are supposed to ignore this additional data.

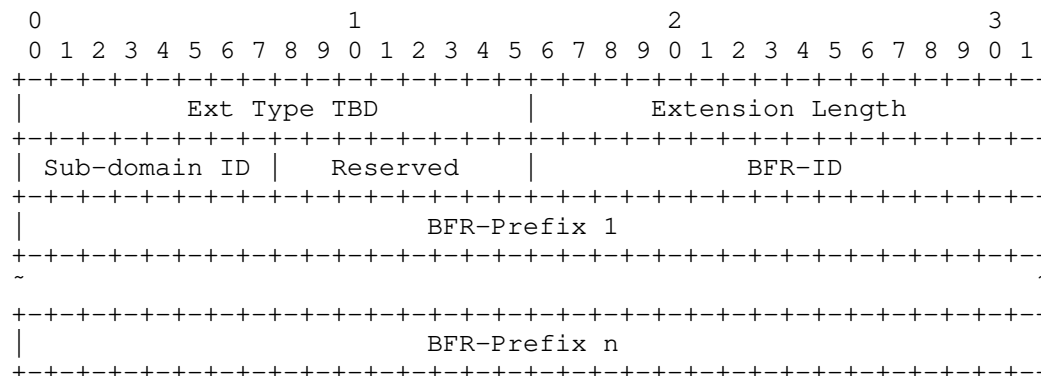


Figure 1: MLD/IGMP Extension Type for BIER

- * Ext Type: Assigned by IANA, identifying this BIER extension.
- * Extension Length: The length in octets of the data after this field. If there are n IPv4 prefixes, the length would be $4 + 4 * n$, if there are n IPv6 prefixes, the length would be $4 + 16 * n$.
- * Sub-domain-id: A single octet containing a BIER sub-domain-id (see [[RFC8279]]). This indicates the BIER sub-domain of the router originating the message.
- * Reserved: A single octet, MUST be set to 0 when sending and ignored when receiving.

- * BFR-id: A two-octet field containing the BFR-id, in the specified sub-domain, of the router originating the message.
- * BFR-prefix: The BFR-prefix (see [[RFC8279]]) of the router that is originating the message. The BFR-prefix will either be a /32 IPv4 address or a /128 IPv6 address.

This extension type MUST be present once in all IGMP and MLD messages when originated with a BIER header to identify the BIER originator. It is expected that any BIER router originating IGMP/MLD messages in BIER supports this specification. Any IGMP/MLD messages that do not contain the extension Section 6) MUST be dropped by the decapsulating router with no processing other than potentially logging or debugging. It is expected that any BIER router processing IGMP/MLD messages with BIER encapsulation supports this specification. If they do not, they will likely ignore the report since they cannot identify the BIER receiver, but they may be able to derive some of the receiver information from the BIER header.

7. Security Considerations

BMLD makes use of IGMPv3/MLDv2 messages transported over BIER in order to configure the BIER Layer of BFIRs. BMLD messages MUST be secured, either by relying on physical or link-layer security, by securing the IP packets (e.g., using IPsec [RFC4301]), or by relying on security features provided by the BIER Layer.

By spoofing the IP source address, an attacker could become the IGMP/MLD querier. Once one becomes the querier, several attack vectors are possible. This is similar to regular IGMP/MLD without BIER encapsulation.

An attacker could send reports with the BIER IGMP/MLD extension Section 6) specifying a BFR-ID and BIER prefix identifying another router. This would allow the attacker to:

- * Redirect undesired traffic toward the spoofed router by subscribing to undesired multicast traffic.
- * Prevent desired multicast traffic from reaching the spoofed router by unsubscribing to some desired multicast traffic.

8. IANA Considerations

This document requests that IANA assigns a new type called BIER information in the registry defined in [I-D.ietf-pim-igmp-mld-extension].

9. Acknowledgements

Comments concerning this document are very welcome.

10. References

10.1. Normative References

- [RFC2113] Katz, D., "IP Router Alert Option", RFC 2113, DOI 10.17487/RFC2113, February 1997, <<https://www.rfc-editor.org/info/rfc2113>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, DOI 10.17487/RFC3376, October 2002, <<https://www.rfc-editor.org/info/rfc3376>>.
- [RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <<https://www.rfc-editor.org/info/rfc3810>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <<https://www.rfc-editor.org/info/rfc8279>>.
- [I-D.ietf-pim-igmp-mld-extension] Sivakumar, M., Venaas, S., Zhang, Z., and H. Asaeda, "Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Version 2 (MLDv2) Message Extension", Work in Progress, Internet-Draft, draft-ietf-pim-igmp-mld-extension-08, 3 June 2022, <<https://datatracker.ietf.org/doc/html/draft-ietf-pim-igmp-mld-extension-08>>.

10.2. Informative References

- [RFC2711] Partridge, C. and A. Jackson, "IPv6 Router Alert Option", RFC 2711, DOI 10.17487/RFC2711, October 1999, <<https://www.rfc-editor.org/info/rfc2711>>.
- [RFC3306] Haberman, B. and D. Thaler, "Unicast-Prefix-based IPv6 Multicast Addresses", RFC 3306, DOI 10.17487/RFC3306, August 2002, <<https://www.rfc-editor.org/info/rfc3306>>.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <<https://www.rfc-editor.org/info/rfc4301>>.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <<https://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, <<https://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, <<https://www.rfc-editor.org/info/rfc7365>>.
- [RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <<https://www.rfc-editor.org/info/rfc7761>>.
- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.
- [RFC8556] Rosen, E., Ed., Sivakumar, M., Przygienda, T., Aldrin, S., and A. Dolganow, "Multicast VPN Using Bit Index Explicit Replication (BIER)", RFC 8556, DOI 10.17487/RFC8556, April 2019, <<https://www.rfc-editor.org/info/rfc8556>>.
- [I-D.ietf-bier-mtud]
Venaas, S., Wijnands, I., Ginsberg, L., and M. Sivakumar, "BIER MTU Discovery", Work in Progress, Internet-Draft,

draft-ietf-bier-mtud-00, 27 February 2019,
<https://datatracker.ietf.org/doc/html/draft-ietf-bier-mtud-00>.

[I-D.ietf-bier-path-mtu-discovery]

Mirsky, G., Przygienda, T., and A. Dolganow, "Path Maximum Transmission Unit Discovery (PMTUD) for Bit Index Explicit Replication (BIER) Layer", Work in Progress, Internet-Draft, draft-ietf-bier-path-mtu-discovery-14, 26 March 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-bier-path-mtu-discovery-14>.

Appendix A. BIER Use Case in Data Centers

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 2.

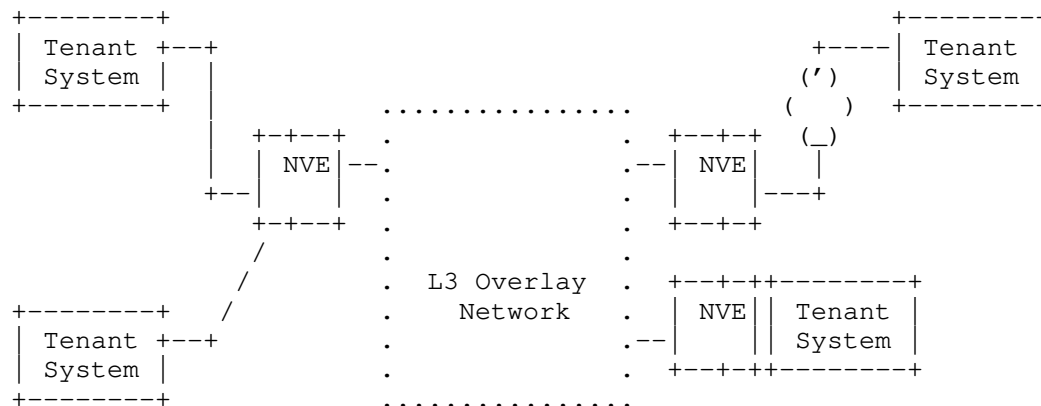


Figure 2: NVO3 Architecture

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 [RFC7761] or PIM-BIDIR [RFC5015] multicast routing protocol. Then, when BUM packets arrive at NVE, it requires NVE to have a mapping between the VXLAN Network Identifier 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 is 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 16M 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 Network Identifier and the remote addresses of NVEs 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.

BIER [RFC8279] 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. Specifically, for BIER-TE, the BIER header may also contain a bit-string in which each bit indicates the link the flow passes through.

The following sub-sections try to propose how to take full advantage of overlay multicast protocol to carry virtual network information, and create a mapping between the virtual network information and the bit-string to implement BUM services in data centers.

A.1. Convention and Terminology

The terms about NVO3 are defined in [RFC7365]. The most common terminology used in this appendix is listed below.

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

Virtual Network Context Identifier: Field in an overlay encapsulation header that identifies the specific VN the packet belongs to.

A.2. 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 ingress NVE in data centers.

The principle of using BIER to forward BUM traffic is that: firstly, it requires each ingress NVE to have a mapping between the Virtual Network Context Identifier and the bit-string in which each bit represents exactly one egress NVE to forward the packet to. And then, when receiving the BUM traffic, the BFIR/Ingress NVE maps the receiving BUM traffic to the mapping bit-string, encapsulates the BIER header, and forwards the encapsulated BUM traffic into the BIER domain to the other BFERs/Egress NVEs indicated by the bit-string.

Furthermore, as for how each ingress NVE knows the other egress NVEs that belong to the same virtual network and creates the mapping is the main issue discussed below. Basically, BIER Multicast Listener Discovery is an overlay solution to support ingress routers to keep per-egress-router state to construct the BIER bit-string associated with IP multicast packets entering the BIER domain. The following section tries to extend BIER MLD to carry virtual network information (such as Virtual Network Context identifier), and advertise them between NVEs. When each NVE receives these information, they create the mapping between the virtual network information and the bit-string representing the other NVEs belonged to the same virtual network.

A.3. A BIER MLD solution for Virtual Network information

The BIER MLD solution allows having multiple MLD instances by having unique pairs of BMLD Nodes and BMLD Querier addresses for each instance. Assume for now that we have a unique instance per VNI and that all BMLD routers are using the same mapping between VNIs and BMLD address pairs. Also for each VNI there is a multicast group used for encapsulation of BUM traffic over BIER. This group may potentially be shared by some or all of the VNIs.

Each NVE acquires the Virtual Network information, and advertises this Virtual Network information to other NVEs through the MLD messages. For a given VNI it sends BMLD reports to the BMLD nodes

address used for that VNI, for the group used for delivering BUM traffic for that VNI. This allows all NVE routers to know which other NVE routers have interest in BUM traffic for a particular VNI. If one attached virtual network is migrated, the NVE will withdraw the Virtual Network information by sending an unsolicited BMLD report. Note that NVEs also respond to periodic queries to BMLD Nodes addresses corresponding to VNIs for which they have interest.

When ingress NVE receives the Virtual Network information advertisement message, it builds a mapping between the receiving Virtual Network Context Identifier in this message and the bit-string in which each bit represents one egress NVE who sends the same Virtual Network information. Subsequently, once this ingress NVE receives some other MLD advertisements which include the same Virtual Network information from some other NVEs, it updates the bit-string in the mapping and adds the corresponding sending NVE to the updated bit-string. Once the ingress NVE removes one virtual network, it will delete the mapping corresponding to this virtual network as well as send withdraw message to other NVEs.

After finishing the above interaction of MLD messages, each ingress NVE knows where the other egress NVEs are in the same virtual network. When receiving BUM traffic from the attached virtual network, each ingress NVE knows exactly how to encapsulate this traffic and where to forward them to.

This can be used in both IPv4 network and IPv6 network. In IPv4, IGMP protocol does the similar extension for carrying Virtual Network information TLV in Version 2 membership report message.

Note that it is possible to have multiple VNIs map to the same pair of BMLD addresses. Provided VNIs that map to the same BMLD address uses different multicast groups for encapsulation, this is not a problem, because each instance is tracking interest for each multicast group separately. If multiple VNIs map to the same pair and the multicast group used is not unique, some NVEs may receive BUM traffic for which they are not interested. An NVE would drop packets for an unknown VNI, but it means wasting some bandwidth and processing. This is similar to the non-BIER case where there is not a unique multicast group for encapsulation. The improvement offered by using BMLD is by using multiple instance, hence reducing the problems caused by using the same transport group for multiple VNIs.

Authors' Addresses

Pierre Pfister
Cisco Systems
Paris
France
Email: pierre.pfister@darou.fr

IJsbrand Wijnands
Cisco Systems
De Kleetlaan 6a
1831 Diegem
Belgium
Email: ice@cisco.com

Stig Venaas
Cisco Systems
Tasman Drive
San Jose, CA 95134
United States of America
Email: stig@cisco.com

Cui(Linda) Wang
Email: lindawangjoy@gmail.com

Zheng(Sandy) Zhang
ZTE Corporation
No.50 Software Avenue, Yuhuatai District
Nanjing
CA,
China
Email: zhang.zheng@zte.com.cn

Markus Stenberg
FI-00930 Helsinki
Finland
Email: markus.stenberg@iki.fi