BIER
Internet-Draft
Updates: 4604 (if approved)
Intended status: Standards Track
Expires: 23 April 2026

Z. Zhang
Juniper Networks
X. Xu
China Mobile
Z. Zhang
ZTE
J. Tantsura
Nvidia
A. Mahale
Cerebras

20 October 2025

Optimized Use of BIER in AIML Data Centers
draft-zzhang-bier-optimized-use-in-aidc-00

Abstract

Use of multicast in AI Data Centers (AIDC) is getting attention for efficient large-scale distribution of the same data to many receivers, given the collectives like All2All and AllReduce. Emerging technologies like In Network Compute(INC) can also benefit from using in-network-distribution of the packets by offloading the distribution of the flows to the network. Given the bursty nature of short-lived all2all flows, BIER is a very good multicast technology for AIDC because it does not need the per-establishment of the multicast tree states. This document discusses further optimization of BIER use in AIDC or similar deployment scenarios, and updates RFC4604 by specifying an IGMP/MLD extension for sources to report receiver information to the First Hop Routers. The extension can be useful and is only needed when the source cannot impose the BIER encapsulation.

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 23 April 2026.

Table of Contents

1. Introduction

Use of multicast in AI Data Centers (AIDC) is getting attention for efficient large-scale distribution of the same data to many receivers. Given the bursty nature of short-lived all2all flows, BIER is a very good multicast technology here because it does not need per-establishment of the multicast-tree states.

This document first discusses the typical way of using BIER in general networks, then discusses some optimizations applicable in this AIDC scenario (or any scenario with similar properties), and specifies an IGMP/MLD [RFC4604] extension that facilitates the optimization.

1.1. Typical Use of Multicast with BIER

Typically, a multicast source selects an IP multicast group address for a certain data flow and advertises the group address. The receivers typically are not chosen by the source but rather decide themselves if they want to join the group. If yes, they send IGMP/MLD membership reports to the Last Hop Routers (LHRs), who in turn send PIM join messages towards the source, establishing a multicast tree along the way.

BIER can serve as the underlay for (part of) the IP multicast trees (the overlay). Instead of establishing the entire IP multicast tree hop-by-hop across the BIER domain, the BIER Forwarding Egress Router (BFER) sends BIER Flow Overlay Signaling messages to the BIER Forwarding Ingress Router (BFIR) based on their IGMP/MLD/PIM states to indicate that they need to receive certain IP multicast (overlay on top of the BIER underlay). The BFIR then encapsulates incoming IP multicast traffic with a BIER header, in which the BitString is based on the flow overlay signaling (which could be MVPN via BGP, or PIM/IGMP/MLD through BIER itself, or controller-provisioned on the host).

1.2. Optimizations and Deployment Considerations

In the AI DC scenario, typically the source chooses the receivers for multicast flows (e.g., a token to be forwarded to a group of experts). Therefore, the above typical use of BIER can be optimized.

Instead of having the receivers discover the groups and join via IGMP/MLD and flow overlay signaling, the sources and receivers become BFIRs/BFERs in the control plane, and are assigned with BFR-IDs. However, they do not have to be involved in the BIER signaling or encapsulation/decapsulation - the LHR performs Penultimate Hop Popping [I-D.ietf-bier-php], and the source notifies the FHR of the receiver identities via a new IGMP/MLD message, referred to as Receiver Proxy Report, so that the FHR knows what BitString it needs to use when it puts on the BIER header. Note that existing IGMP/MLD Membership Report are used for receivers to notify the LHRs that there are receivers on an interface, while the new Receiver Proxy Report are used for sources to notify the FHRs which receivers need to receive the traffic.

Note that if the sources/receivers can handle BIER encapsulation/decapsulation, then PHP or the IGMP/MLD extension for notifying the FHR of the receivers (so that the FHR can impose the BIER encapsulation) is not needed. The rest of this document focuses on the case where the source/receivers are not capable of BIER encapsulation/decapsulation.

Because the LHR knows exactly which receivers need to receive the traffic (because of the BitString), a single common multicast group address can be used, e.g., the well-known all-host multicast address 224.0.0.2, so that a receiver only needs to prepare to receive packets addressed to that single multicast group address (e.g., setting up its NIC to accept Ethernet frames with a multicast MAC address corresponding to the IP group address and deliver up the stack), no matter how many flows (of different sets of receivers) it may need to receive.

The source does need to choose a different group address for different flows though, so that the FHR knows what BitString to use. When the FHR receives an IGMP/MLD Receiver Proxy Report for a <source, group>, it sets up forwarding state for the <source, group> with the following forwarding behaviors:

* Replace the destination address with the common address

* Encapsulate the traffic with a BIER header with a BitString corresponding to the reported receivers and forward the BIER packets

It is expected that the source will choose receivers in clusters (i.e., with the BFR-IDs in close proximity) so that they can fit into as few "BIER Sets" as possible and as few replications as possible are performed.

Note that this optimization is applicable to any deployment scenario where sources know the receivers inherently (w/o relying on the flow overlay signaling) - not just in the AIML DC.

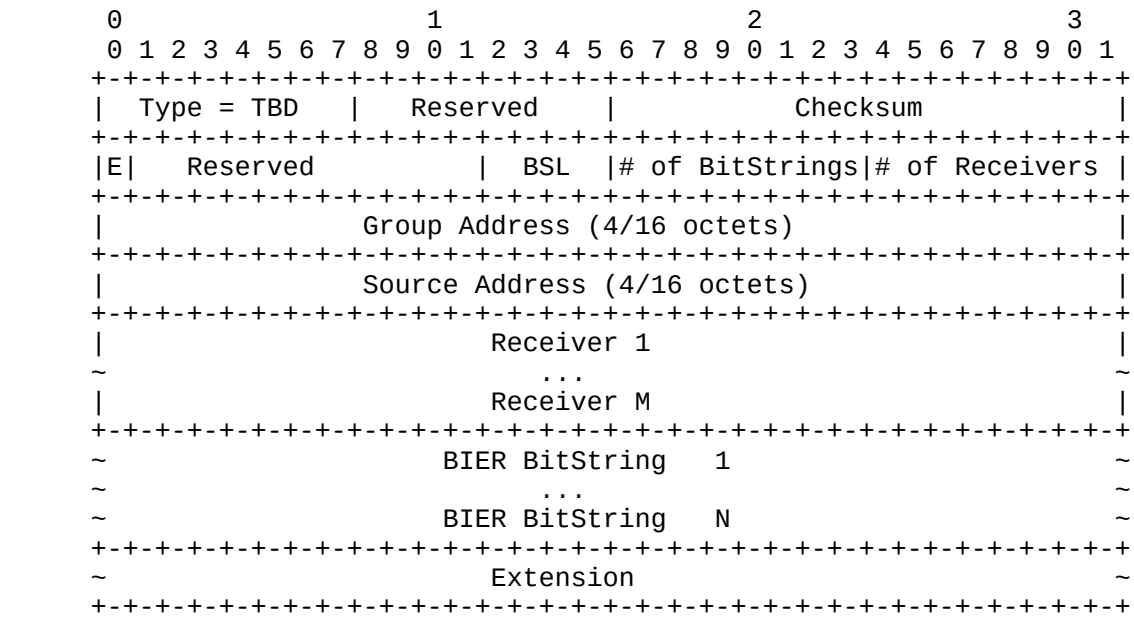
2. Specification

2.1. Advertising BFR-IDs

While the sources/receivers are considered BFIRs/BFERs with assigned BFR-IDs, they do not participate in BIER signaling or forwarding. Their BFR-IDs MUST be advertised by the routers connected to them (the FHRs/LHRs) in the BIER proxy range sub-TLV [I-D.ietf-bier-prefix-redistribute].

2.2. IGMP/MLD Receiver Proxy Report

The Receiver Proxy Report is a new IGMP/MLD message with type TBD and the following format:



* E-bit and Extension: for extensions, as specified in [RFC9279].

* Group Address: The group address selected by the sender.

* Source Address: The sender’s address.

* # of Receivers: number of IPv4/IPv6 addresses. Could be 0.

* # of BitStrings: number of BIER BitStrings. Could be 0.

* BSL: BIER BitStringLength, encoded as specified in [RFC8296]. Significant only if the number of BitString is non-zero.

* BIER BitString: the receivers encoded in a BIER BitString.

The receivers are encoded either as IP addresses or in BitStrings. The Number of Receivers and Number of BitStrings MUST NOT be non-zero at the same time.

3. Security Considerations

To be added.

4. IANA Considerations

This document requests IANA to allocate ... To be added.

5. Acknowledgments

6. Normative References

[I-D.ietf-bier-php]
Zhang, Z. J., "BIER Penultimate Hop Popping", Work in Progress, Internet-Draft, draft-ietf-bier-php-16, 4 December 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-bier-php-16>.

[I-D.ietf-bier-prefix-redistribute]
Zhang, Z., Wu, B., Zhang, Z. J., Wijnands, I., Liu, Y., and H. Bidgoli, "BIER Prefix Redistribute", Work in Progress, Internet-Draft, draft-ietf-bier-prefix-redistribute-09, 21 August 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-bier-prefix-redistribute-09>.

[RFC4604] Holbrook, H., Cain, B., and B. Haberman, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast", RFC 4604, DOI 10.17487/RFC4604, August 2006, <https://www.rfc-editor.org/info/rfc4604>.

[RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <https://www.rfc-editor.org/info/rfc8296>.

[RFC9279] 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", RFC 9279, DOI 10.17487/RFC9279, August 2022, <https://www.rfc-editor.org/info/rfc9279>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks
Email: zzhang@juniper.net

Xiaohu Xu
China Mobile
Email: xuxiaohu_ietf@hotmail.com

Zheng Zhang
ZTE
Email: zhang.zhen@zte.com.cn

Jeff Tantsura
Nvidia
Email: jefftant.ietf@gmail.com

Aditya Mahale
Cerebras
Email: aditya.ietf@gmail.com