RTGWG
Internet-Draft
Intended status: Informational
Expires: 16 September 2026

J. Zhang
W. Cheng
K. Liu
China Mobile

15 March 2026

Requirements and Gap Analysis of Multicast in AI Data Centers

draft-zhang-rtgwg-multicast-requirements-gaps-aidc-01

Abstract

Multicast has the potential to be applied in Artificial Intelligence Data Centers (AIDCs) to improve the efficiency of point-to-multipoint data transmission during large language model training and inference. This document identifies key requirements of multicast in AIDCs, and analyzes the gaps between these requirements and the capabilities of existing multicast technologies.

Table of Contents

1. Introduction

Artificial Intelligence (AI) Data Centers (AIDCs) serve as the key infrastructure for AI large language model (LLM) training and inference, where point-to-multipoint (P2MP) communication patterns are common and critical to overall system efficiency. Network multicast leverages in-network data replication to achieve efficient distribution of identical data, reducing processing overhead and network bandwidth consumption of the sender, thereby enhancing the efficiency of P2MP data transmission. Multicast is a promising technology for deployment in AIDCs.

Despite the potential opportunities, existing multicast technologies are not originally designed to address the specific characteristics of AIDC networks. AIDC networks are defined by ultra-high bandwidth (often 400 Gbps or greater), microsecond-level latency, and high reliability that demands near-zero packet loss. These core performance characteristics necessitate corresponding qualities in multicast technologies, including interactivity, reliability, and simplicity. Furthermore, emerging multicast use cases in AIDCs, such as token dispatch, also introduce specific requirements, including high dynamics and membership sparseness.

This document identifies the typical multicast use cases and key requirements for multicast in AIDCs, and analyzes the limitations of existing multicast technologies in meeting these requirements.

1.1. Requirements Language

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.

2. Multicast Use Cases

Tasks in AIDC include model loading and distribution, model training and inference, model saving, and other key operations. These tasks generate various traffic patterns, including communication between computing devices (e.g., GPUs), traffic among storage nodes, and data transmission between computing devices and storage nodes. Among these, many typical communication patterns exhibit P2MP characteristics, making multicast a critical enabling technology. The typical multicast use cases in AIDCs are as follows:

* Token dispatch in Mixture-of-Experts (MoE) models: MoE is a mainstream architecture for LLMs, adopted by many LLMs such as Llama4, Mixtral and DeepSeekV3. During execution, an input token is dispatched to multiple selected expert nodes based on the real-time routing decision. With expert parallelism, the token dispatch process manifests as an AlltoAll collective communication phase, where tokens are sent from source devices to multiple destination devices hosting the target experts. This token dispatch process naturally follows the P2MP traffic pattern [I-D.zzhang-bier-optimized-use-in-aidc].

* Broadcast phase in AllReduce operation: In distributed training of LLMs, AllReduce is a core collective communication operation for data parallelism and tensor parallelism. Although AllReduce can be implemented in several ways, decomposing it into Reduce and Broadcast phases is a basic approach, where the Broadcast phase exhibits a P2MP traffic pattern, natively matching the multicast semantics.

* Model distribution: The initial distribution of model parameters or checkpoints from storage nodes to all computing nodes is a critical step before training iterations begin. Given that the model data can reach terabytes in size, simultaneously transmitting this massive data from the storage node to many GPUs constitutes a typical P2MP traffic pattern.

* Multi-replica checkpoint storage: To avoid loss of training progress due to failures, training programs periodically save model states, i.e., checkpoints, to multiple storage nodes. Multicast is a promising technique in this scenario, which is supposed to improve the efficiency of transmitting massive data to multiple replicas [I-D.liu-multicast-for-computing-storage].

3. Multicast Requirements

3.1. Bidirectional Interactivity

AI workloads are highly sensitive to packet loss. In LLM training, packet loss without a reliability acknowledgment mechanism can corrupt model parameters, leading to degraded model quality or even training failures. Moreover, congestion control is required to actively avoid congestion and packet loss. Therefore, networks in AIDCs are required to support closed-loop control, such acknowledgment and congestion control, to meet the high-performance and high-reliability requirements of AI workloads.

Traditional IP multicast only supports best-effort P2MP data delivery, while multicast in AIDCs should support bidirectional interaction, including both efficient P2MP data forwarding and multipoint-to-point (MP2P) feedback forwarding. The core interactivity demands are as follows:

* P2MP forwarding and replication: Multicast should support efficient P2MP forwarding and in-network data replication, which is the fundamental requirement of multicast.

* MP2P forwarding: Multicast in AIDCs should natively support efficient MP2P forwarding, particularly for feedback signals generated from receivers to senders such as acknowledgments (ACKs) to confirm successful data reception and negative acknowledgments (NACKs) to report packet loss, as well as congestion notification signals.

* MP2P packet aggregation: In large-scale AIDCs with numerous receivers, if each receiver sends feedback packets to a sender independently, it can lead to excessive reverse traffic to the sender, resulting in network congestion, increased latency, and

amplified jitter. To address this issue, multicast in AIDCs should support MP2P packet aggregation. Network devices serving as rendezvous points should aggregate multiple feedback packets from different receivers into a single packet and send it to the sender.

3.2. High Reliability

AI workloads exhibit near-zero tolerance for packet loss. Even with acknowledgment a>nd retransmission, extremely low packet loss rates can trigger massive retransmissions. In multicast scenarios, loss at any receiver can force the sender to retransmit data to all receivers, causing significant redundant traffic and efficiency degradation. Furthermore, maintaining uninterrupted tasks for long periods is crucial for LLM training. However, hardware is prone to failures, and as the scale of training networks increases, the likelihood of network failures rises due to an increasing number of switches, network interface cards, and optical modules [I-D.cheng-rtgwg-ai-network-reliability-problem]. Therefore, multicast in AIDCs should provide high reliability to ensure service performance and continuity. The specific requirements are as follows:

* Lossless transmission: Under normal network conditions without network failures, multicast in AIDCs should satisfy the lossless requirement with no packet loss, to ensure both high reliability and high performance. To achieve lossless transmission, multicast should support reliability acknowledgment and effective flow/ congestion control, which need support of interactivity requirement.

* Fast failure detection: The multicast should support fast detection of link failures and node failures and efficient detection of gray failures, which are the prerequisite for any subsequent recovery action.

* Fast failure recovery: Upon failure detection, the multicast should support fast recovery mechanisms to restore multicast traffic rapidly. It is unacceptable to rely solely on global control-plane convergence and multicast tree reconstruction for slow recovery time.

* Minimized failure domain: The recovery mechanism should confine the impact of a failure to the smallest possible set of receivers. Local link or node failures should only affect the faulty segment, without spreading to the entire multicast tree or other service branches.

3.3. High Dynamics

AI workloads, especially those using sparse architectures like MoE, have highly dynamic communication patterns. MoE-based AI training and inference uses token dispatch, where gating networks select expert nodes per token at microsecond timescales, dynamically determining real-time multicast receiver sets with no fixed groups. This ultra-fast selection leaves no time for traditional multicast to establish, update, or tear down trees, leading to delays, packet loss, or AI task failure [I-D.zhang-rtgwg-llmmoe-multicast]. Therefore, multicast in AIDC should meet high dynamics requirements, and the key points are as follows:

* Fast change of multicast members: Multicast should be able to adapt to the dynamic change of multicast members in microsecond timescales.

* Low overhead for dynamic change: The dynamic change of multicast members should generate minimal overhead in both the control plane and data plane. Excessive signaling or processing overhead during dynamic change will increase transmission latency and reduce the efficiency of AI workloads.

3.4. Sparseness

Multicast in AIDCs frequently involves multicast groups where only a small fraction of the total nodes in the cluster are multicast members, a characteristic closely tied to the sparse activation mechanism of modern AI models such as MoE. For example, DeepSeekV3 uses 256 experts and activates 9 experts at a time. Multicast technologies that are designed for dense groups are inefficient for this sparse mode. The multicast should be efficient when the group size is small relative to the network size, and meet the following sparseness requirements:

* Efficient sparse member identification: Multicast technologies should support efficient identification of sparse multicast members. The methods for identifying multicast members should avoid unnecessary scanning or signaling of non-member nodes, and be efficient for forwarding.

* Low overhead for sparse state maintenance: The maintenance of multicast member state should be lightweight and low-overhead, adapting to the sparse characteristics of AIDC multicast groups. It should avoid maintaining redundant state information for non-member nodes, reducing state maintenance burden and ensuring that state updates do not introduce additional latency that affects AI task efficiency.

3.5. Simplicity

Simplicity is a foundational architectural principle for multicast in AIDCs, directly enabling the microsecond-timescale low-latency transmission in large-scale AIDC networks. Complexity in the control or data plane manifests as variable latency, unpredictable jitter, and an inability to meet the strict performance bounds of AI workloads. Therefore, multicast in AIDCs should be governed by the following overarching simplicity requirements:

* Control plane simplicity: The multicast control plane should be architecturally simple to maintain core functions like multicast routing, minimizing signaling interaction overhead and control processes. It should avoid complex state synchronization and protocol negotiation processes, to reduce network operation and maintenance complexity.

* Data plane simplicity: The multicast data plane needs to be highly efficient and simple, including efficient member identification and forwarding adapting to sparse and dynamic multicast characteristics, and optimized packet processing mechanisms. These ensure minimal forwarding and processing overhead, meeting the low-latency transmission requirements of AI workloads.

4. Gap Analysis

To address the gaps between multicast requirements in AIDCs and existing technologies, typical multicast technologies are first introduced, followed by an analysis of their capabilities against key requirements.

4.1. Typical Multicast Technologies

Protocol Independent Multicast (PIM) is a widely deployed multicast routing protocol that operates independently of underlying unicast routing protocols. It supports dense mode (PIM-DM) [RFC3973] and sparse mode (PIM-SM) [RFC7761]. PIM-SM builds unidirectional shared trees rooted at a Rendezvous Point per group and it optionally creates shortest-path trees per source.

Multipoint extensions for Label Distribution Protocol (mLDP) [RFC6388] constructs the P2MP or multipoint-to-multipoint (MP2MP) Label Switched Paths (LSPs) in Multiprotocol Label Switching (MPLS) networks without interacting with or relying upon any other multicast tree construction protocol.

Segment Routing Point-to-Multipoint (SR-P2MP) [I-D.ietf-pim-sr-p2mp-policy] enables creation of P2MP trees for efficient multi-point packet delivery in a Segment Routing (SR) domain. It requires the routing module of the controller or ingress node to calculate and determine the path of the multicast traffic, and the data plane can reuse existing SR unicast forwarding mechanisms.

Bit Indexed Explicit Replication (BIER) [RFC8279] is a stateless multicast technology that eliminates the need for explicit tree construction. Instead, the set of intended receivers is encoded as a BitString within the packet header. Intermediate BIER Forwarding Routers (BFRs) replicate packets based on the BitString, without maintaining any per-flow or per-tree state.

4.2. Gap Analysis Against Requirements

The support of typical multicast technologies for multicast requirements in AIDCs is summarized in Table 1.

Technology	Interactivity	Reliability	Dynamics	Sparseness	Simplicity
PIM	Poor	Poor	Poor	Good	Poor
mLDP	Poor	Poor	Poor	Good	Poor
SR-P2MP	Poor	Moderate	Moderate	Good	Moderate
BIER	Poor	Moderate	Good	Poor	Good

Table 1: Gap Analysis

Interactivity: These multicast technologies can support best-effort P2MP data delivery, but none of them can natively support the reverse MP2P forwarding or aggregation to achieve bidirectional interactivity.

Reliability: These multicast technologies fail to meet the lossless requirement of AIDC networks. The reliability of PIM and mLDP basically relies on routing convergence and multicast tree reconstruction. Although some fast detection and recovery mechanisms [RFC9186][RFC9860][RFC7715] can be adopted to accelerate failure recovery, their tree-based architectures often keep the failure impact domain tree-level. In contrast, BIER and SR-P2MP can effectively reuse unicast’s reliability capabilities such as Fast ReRouting, and control the failure domain within the damaged receivers, demonstrating good reliability.

Dynamics: PIM and mLDP adjust multicast trees via control signals, leading to slow convergence that struggles to handle high-frequency member changes. SR-P2MP dynamically recalculates forwarding trees via a controller, which need global recalculating and result distribution. BIER only requires updating the BitString in packets, enabling faster responses to member changes and exhibiting good dynamics.

Sparseness: PIM, mLDP, and SR-P2MP can all adapt well to sparse scenarios, as they establish multicast trees or tunnels on demand, and multicast member identification is based on IP or other non-contiguous labels. In contrast, BIER encodes the receiver set as a BitString, whose length is proportional to the number of nodes in the domain. Even with sparse members, the full BitString must still be carried, leading to significant degradation in bandwidth overhead and forwarding efficiency. This limits BIER’s applicability in AIDC sparse multicast scenarios.

Simplicity: PIM and mLDP require the maintenance of complex multicast tree states and signaling mechanisms, resulting in high operational complexity and poor simplicity. SR-P2MP reuses the SR unicast forwarding plane, with the control plane relying on a controller, leading to moderate complexity but still requiring additional tree management logic. BIER, on the other hand, eliminates the need for explicit multicast tree construction, with no per-flow state at intermediate nodes, resulting in better simplicity. Moreover, simplicity still needs further optimization to meet the ultra-high performance requirements of AI networks.

In summary, the most critical common gap is the lack of native support for efficient, scalable bidirectional interactivity, which is the cornerstone for implementing closed-loop acknowledgement and congestion control. Furthermore, no single multicast technology excels in all dimensions: some lack reliability, dynamics or simplicity (PIM, mLDP, SR-P2MP), others are inefficient for sparse groups (BIER). Consequently, merely deploying or combining these existing technologies is insufficient to meet the stringent demands of AIDC workloads. This gap analysis underscores the need for either a new architecture designed from the ground up for AIDCs or significant extensions to existing technologies.

5. IANA Considerations

TBD.

6. Security Considerations

TBD.

7. References

7.1. 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, <https://www.rfc-editor.org/info/rfc2119>.

[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>.

7.2. Informative References

[RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, DOI 10.17487/RFC3973, January 2005, <https://www.rfc-editor.org/info/rfc3973>.

[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>.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <https://www.rfc-editor.org/info/rfc6388>.

[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>.

[RFC9186] Mirsky, G. and X. Ji, "Fast Failover in Protocol Independent Multicast - Sparse Mode (PIM-SM) Using Bidirectional Forwarding Detection (BFD) for Multipoint Networks", RFC 9186, DOI 10.17487/RFC9186, January 2022, <https://www.rfc-editor.org/info/rfc9186>.

[RFC9860] Liu, Y., McBride, M., Zhang, Z., Xie, J., and C. Lin, "Multicast-Only Fast Reroute (MoFRR) Based on Topology Independent Loop-Free Alternate (TI-LFA) Fast Reroute", RFC 9860, DOI 10.17487/RFC9860, October 2025, <https://www.rfc-editor.org/info/rfc9860>.

[RFC7715] Wijnands, IJ., Ed., Raza, K., Atlas, A., Tantsura, J., and Q. Zhao, "Multipoint LDP (mLDP) Node Protection", RFC 7715, DOI 10.17487/RFC7715, January 2016, <https://www.rfc-editor.org/info/rfc7715>.

[I-D.ietf-pim-sr-p2mp-policy]
Parekh, R., Voyer, D., Filsfils, C., Bidgoli, H., and Z. J. Zhang, "Segment Routing Point-to-Multipoint Policy", Work in Progress, Internet-Draft, draft-ietf-pim-sr-p2mp-policy-22, 4 September 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-pim-sr-p2mp-policy-22>.

[I-D.zzhang-bier-optimized-use-in-aidc]
Zhang, Z. J., Xu, X., Zhang, Z., Tantsura, J., and A. Mahale, "Optimized Use of BIER in AIML Data Centers", Work in Progress, Internet-Draft, draft-zzhang-bier-optimized-use-in-aidc-00, 20 October 2025, <https://datatracker.ietf.org/doc/html/draft-zzhang-bier-optimized-use-in-aidc-00>.

[I-D.zhang-rtgwg-llmmoe-multicast]
Zhang, Z., Duan, W., and X. Xu, "Multicast usage in LLM MoE", Work in Progress, Internet-Draft, draft-zhang-rtgwg-llmmoe-multicast-01, 20 October 2025, <https://datatracker.ietf.org/doc/html/draft-zhang-rtgwg-llmmoe-multicast-01>.

[I-D.liu-multicast-for-computing-storage]
Liu, Y. and X. Geng, "Multicast for Computing and Storage", Work in Progress, Internet-Draft, draft-liu-multicast-for-computing-storage-00, 10 July 2023, <https://datatracker.ietf.org/doc/html/draft-liu-multicast-for-computing-storage-00>.

[I-D.cheng-rtgwg-ai-network-reliability-problem]
Cheng, W., Lin, C., wangwenxuan, and B. Xu, "Reliability in AI Networks Gap Analysis, Problem Statement, and Requirements", Work in Progress, Internet-Draft, draft-cheng-rtgwg-ai-network-reliability-problem-03, 6 June 2025, <https://datatracker.ietf.org/doc/html/draft-cheng-rtgwg-ai-network-reliability-problem-03>.

Authors’ Addresses

Junye Zhang
China Mobile
China
Email: zhangjunye@chinamobile.com

Weiqiang Cheng
China Mobile
China
Email: chengweiqiang@chinamobile.com

Kefei Liu
China Mobile
China
Email: liukefei@chinamobile.com