RTGWG
Internet-Draft
Intended status: Informational
Expires: 7 January 2026

Z. Zhang
W. Duan
ZTE Corporation

6 July 2025

Multicast usage in LLM MoE
draft-zhang-rtgwg-llmmoe-multicast-00

Abstract

Large Language Models (LLMs) have been widely used in recent years. The Mixture of Experts (MoE) architecture is one of the features of LLMs that enables efficient inference and cost-effective training. With the MoE architecture, there are potential multicast use cases such as tokens dispatching. This draft attempts to analyze these use cases.

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 7 January 2026.

Copyright Notice

Copyright (c) 2025 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

In recent years, large language models (LLMs) have been widely used. Mixture of Experts Model (MoE) is one of the functions of LLM to achieve efficient inference and economical training. Many LLMs currently adopt the MoE architecture, such as DeepSeek-V2/V3, Google Gemini 1.5 Pro, xAI Grok-1, Mistral 8*22B, Qwen3, etc. During inference, MoE only activates a small number of parameters to determine each output token, which significantly reduces the amount of computation required by the processor, thereby reducing the overall computational requirements. Therefore, the fewer parameters are activated, the less computation the processor needs to perform. In the MoE architecture, one token needs to be sent to multiple experts, which is a typical multicast use case.

In most LLMs, two experts are activated during the computation: one is a routed expert and the other is a shared expert. In DeepSeekV3, one token activates eight routed experts and one shared expert.

When all activated experts are located on a node with multiple GPUs installed, only intra-node communication is required. When activated experts are located on different nodes, inter-node communication is required. Due to the bandwidth difference between intra-node and inter-node scenarios, communication across leaf switches and even spine switches is inevitable.

This draft analyzes the multicast use case of LLM in data centers.

1.1. Requirements Language

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].

2. Use case - tokens dispatching

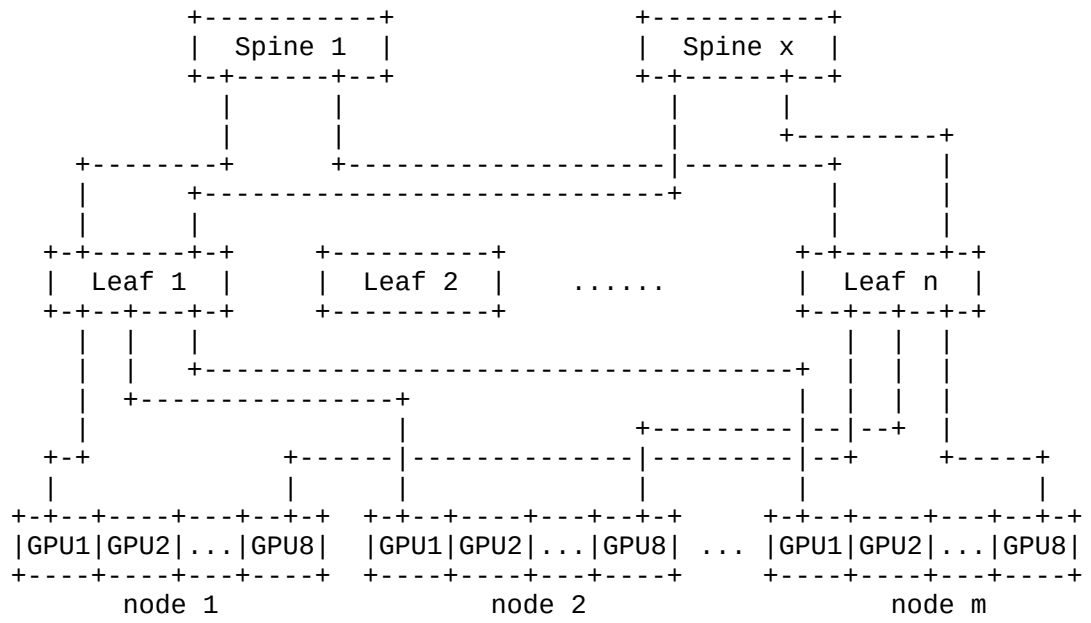

Figure 1

During the pre-filling and decoding phases, tokens need to be sent to all selected experts, including routed experts and shared experts. The tokens dispatching can be intra-node or inter-node. Different LLMs use different numbers of experts. For example, Mixtral uses 8 experts and activates 2 experts at a time; LlaMa 4 uses 16 experts (Scout) or 128 experts (Maverick), and activates 2 experts at a time; DeepSeekV3 uses 256 experts and activates 9 experts at a time. The more routed experts there are, the more distribution there is between nodes. In order to balance the experts, it is difficult to limit the number of experts to one node even if only two experts (one routed expert and one shared expert) are used.

The tokens dispatching can be optimized. For example, in DeepSeekV3, LLM first selects the node group and then selects the expert from the node. By implementing the node restricted routing function, a maximum of four nodes are selected to reduce the inter-node consumption of tokens dispatching. In addition, in order to maximize the usage of the high intra-node bandwidth, after the switch or GPU in the node receives the tokens, the switch or GPU needs to distribute the tokens to the experts in the same node. This optimization aims to reduce the inter-node distribution, but it cannot avoid multicast between nodes.

Therefore, the use of multicast may be intra-node or inter-node. The existing multicast implementation methods are different in intra-node and inter-node scenarios, and multicast management is more difficult.

2.1. Intra-node multicast

When tokens need to be sent to multiple GPUs in the same node, the GPU or the switch connected to the GPU may send the tokens in a multicast manner. This requires the switch or GPU to support the multicast function. This function can reduce the computational burden of the source GPU and reduce the bandwidth consumption between nodes.

2.2. Inter-node multicast

When tokens need to be sent to multiple nodes, Leaf switches and even Spine switches need to forward tokens. Due to the limitation of inter-node bandwidth, the more packets there are, the greater the risk of congestion. Using multicast technology can reduce the burden on the source GPU and reduce the risk of congestion.

3. Multicast technologies analysis

Protocol Independent Multicast - Sparse Mode (PIM-SM) [RFC7761] is a traditional multicast technology. It relies on PIM signaling to build the multicast tree. When the receivers change, the multicast tree may need to be rebuilt. When PIM is used for intra-node or inter-node multicast, the stability of the multicast tree is more important. It may not be applicable when the expert combination is flexible. Even in the intra-node scenario, the number of potential multicast trees may be large despite the limited number of GPUs in a single node.

BIER (Bit-Indexed Explicit Replication) [RFC8279] is an architecture that provides optimal multicast forwarding through a "multicast domain", without requiring intermediate routers to maintain any per-flow state or to engage in an explicit tree-building protocol. BIER is more flexible than PIM. Experts can be numbered and can act as ingress or egress BFRs in BIER. BIER header encapsulation can be a function defined in [RFC8296], [I-D.ietf-bier-bierin6], or [I-D.zzhang-bier-unmasked-bier]. By using the BIER function, Leaf and Spine switches, and even GPUs or switches connected to GPUs, can pre-build expert-based forwarding tables. tokens can be sent to any selected expert.

Other multicast methods, such as PIM DM (dense mode) and ingress replication, may consume more bandwidth and may not be a good choice for multicast scenarios such as LLM tokens dispatching.

While the network layer can provide multicast capabilities for multicast scenarios such as tokens dispatching, the multicast approach needs to work in conjunction with the LLM software. It may work in conjunction with the implementation of collective communication and NIC (network interface card).

4. IANA Considerations

There are no IANA consideration introduced by this draft.

5. Security Considerations

There are no security issues introduced by this draft.

6. References

6.1. 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>.

[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>.

[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>.

[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>.

6.2. Informative References

[I-D.ietf-bier-bierin6]
Zhang, Z., Zhang, Z. J., Wijnands, I., Mishra, M. P., Bidgoli, H., and G. S. Mishra, "Supporting BIER in IPv6 Networks (BIERin6)", Work in Progress, Internet-Draft, draft-ietf-bier-bierin6-11, 2 March 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-bier-bierin6-11>.

[I-D.zzhang-bier-unmasked-bier]
Przygienda, T., Zhang, Z. J., Bigdoli, H., and I. Wijnands, "Unmasked BIER Mode", Work in Progress, Internet-Draft, draft-zzhang-bier-unmasked-bier-01, 23 February 2025, <https://datatracker.ietf.org/doc/html/draft-zzhang-bier-unmasked-bier-01>.

Authors' Addresses

Zheng Zhang
ZTE Corporation
China
Email: zhang.zheng@zte.com.cn

Wei Duan
ZTE Corporation
China
Email: duan.wei1@zte.com.cn