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

Y. Liu
China Mobile
Z. Zhang
ZTE Corporation
J. Zhang
China Mobile

12 February 2026

Multicast Use Cases for Large Language Model Synchronization
draft-liu-rtgwg-llmsync-multicast-00

Abstract

Large Language Models (LLMs) deployments are becoming increasingly widespread, with inference services being the most common application. This draft will discuss multicast use cases for inference cloud services.

Table of Contents

1. Introduction

With the rapid development of AI and the widespread application of large language models (LLMs), inference services are the most frequently used services. Different users may use different LLMs, and the same user may use multiple LLMs simultaneously for inference to obtain the optimal solution. AI inference cloud providers can provide large-scale real-time inference, fine-tuning, and model optimization services on GPU cloud platforms. However, the GPU infrastructure of AI inference cloud providers may cover multiple cloud platforms and regions, facing significant challenges in deployment and application, including highly concurrent model loading and severe cold start latency.

This draft will discuss multicast use cases for inference cloud services.

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. LLM Synchronization

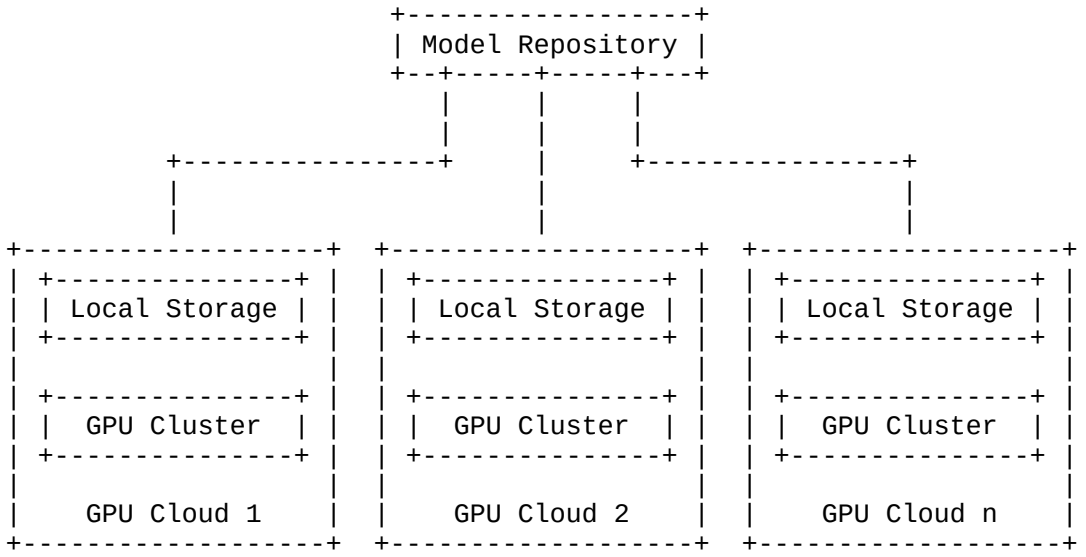

Figure 1

Highly concurrent model loading refers to the peak load and concurrency challenge of simultaneously downloading the same popular large model across dozens of GPU cloud platforms. A single deployment of a model may involve 10 to 100+ copies, each requiring a complete copy of the model file. Hundreds of GPU servers simultaneously downloading the same model (each ranging from 70GB to 1TB in size) within a single cluster generates enormous bandwidth demands and creates I/O bottlenecks.

Severe cold start latency refers to the high delay in initial model deployment caused by slow download speeds. Replica startup time is limited by inbound network bandwidth, which varies significantly between different cloud providers and is typically much lower than the cluster’s internal bandwidth. This significantly impacts the download efficiency of large models in practical applications.

Highly concurrent model downloading are typical multicast applications. Such multicast applications have the following characteristics:

* Large data volume: Due to the large size of the models, typically a single model can reach 70GB to 1TB, placing extremely high demands on network bandwidth.

* Transmission time: Due to cold start latency requirements, data transmission needs to be completed as quickly as possible. Transmission times exceeding tens of minutes will significantly impact user experience; therefore, the shorter the time, the better.

Therefore, when applying multicast technology to this scenario, it is necessary to consider ensuring high bandwidth and low latency.

3. Multicast technologies applying

Considering the need to conserve network bandwidth, ingress interface replication technology is not suitable for this scenario. PIM-SM, SR P2MP or BIER technologies should be considered instead.

Protocol Independent Multicast - Sparse Mode (PIM-SM) [RFC7761] is a traditional multicast technology. It is widely used in scenarios where the receivers are relatively fixed, such as IPTV systems. When the network topology of the multicast tree changes, a new multicast tree needs to be established for each multicast stream via PIM-SM signaling after the BGP/IGP protocol converges. The convergence time of the multicast tree is much longer than that of the IGP protocol.

SR-P2MP (Segment Routing Replication for Multipoint Service Delivery) [I-D.ietf-pim-sr-p2mp-policy] is a relatively new tunneling technology that uses SR-MPLS/SRv6 (Segment Routing over IPv6) tunneling technology for multicast traffic transmission. It requires the routing module of the controller or ingress node to calculate and determine the path of the multicast traffic. Then, the controller or ingress node issues a SID (Segment Identifier) for multicast operations to the replication point (i.e., the multicast replication point) in the network. When multicast traffic enters the tunnel, it is replicated and forwarded at the replication point according to the multicast operation SID. When the network topology changes, the controller or ingress node needs to recalculate and determine the replication point and issue the multicast operation SID to the changed replication point, so that subsequent multicast traffic will be forwarded through the new path.

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-SM and SR P2MP. When link failures or other issues occur on the multicast forwarding path, BIER can converge along with IGP convergence, a speed far exceeding that of PIM-SM and SR P2MP.

When considering applying multicast technology to large model synchronization scenarios, if the model is synchronized to the same destination GPU clouds each time, a multicast tree can be pre-established, or the SR replication path can be calculated using the controller, and PIM-SM or SR P2MP technologies can be used for model copying.

If the destination GPU clouds for each model synchronization is different, pre-establishing a multicast tree or multicast path each time using PIM-SM/SR P2MP technologies may be inefficient because multicast tree establishment takes time. In this case, using BIER technology is a better choice.

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>.

6.2. Informative References

[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>.

Authors' Addresses

Yisong Liu
China Mobile
China
Email: liuyisong@chinamobile.com

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

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