Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 30 August 2026

X. Xu
China Mobile
Z. He
Broadcom
N. Wang
Intel
N. Wang
Hygon
H. Wang
Moore Threads
J. Guo
Biren Technology
X. Li
Enflame Technology
T. Zhou
Resnics Technology
Y. Yang
Centec
Y. Xia
W. Zhang
Tencent
P. Wang
Baidu
Y. Zhuang
Huawei Technologies
F. Yang
Cloudnine Information Technologies
C. Li
Metanet Networking Technology
X. Wang
Ruijie Networks

26 February 2026

Fully Adaptive Routing Ethernet in Scale-Up Networks
draft-xu-rtgwg-fare-in-sun-02

Abstract

The Mixture of Experts (MoE) has become a dominant paradigm in transformer-based artificial intelligence (AI) large language models (LLMs). It is widely adopted in both distributed training and distributed inference. To enable efficient expert parallelization and even tensor parallelization across dozens or even hundreds of Graphics Processing Units (GPUs) in MoE architectures, an ultra-high-throughput, ultra-low-latency AI scale-up network (SUN) is critical. This document describes how to extend the Weighted Equal-Cost Multi-Path (WECMP) load-balancing mechanism, referred to as Fully Adaptive Routing Ethernet (FARE), which was originally designed for scale-out networks, to scale-up networks.

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 RFC 2119 [RFC2119].

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 30 August 2026.

Copyright Notice

Table of Contents

1. Introduction

The Mixture of Experts (MoE) has become a dominant paradigm in transformer-based artificial intelligence (AI) large language models (LLMs). It is widely adopted in both distributed training and distributed inference. To enable efficient expert parallelization and even tensor parallelization across dozens or even hundreds of Graphics Processing Units (GPUs) in MoE architectures, an ultra-high-throughput, ultra-low-latency AI scale-up network (SUN) is indispensable. This network serves as the interconnection fabric, allowing GPUs to function as a unified super GPU, referred to as a SuperPoD. The scale-up network is fundamental for efficiently transporting substantial volumes of communication traffic within the SuperPoD. It includes but not limited to: 1) all-to-all traffic for Expert Parallelism (EP) communication, and 2) all-reduce traffic for Tensor Parallelism (TP) communication, ensuring consistent tensor values across GPUs during training and inference.

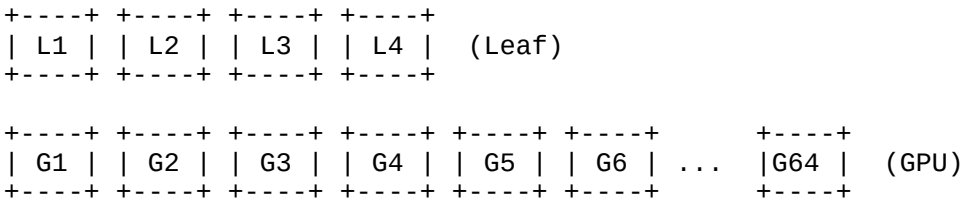

Figure 1

(Note that the diagram above does not include the connections between GPUs and leaf switches. However, it can be assumed that GPUs are connected to every leaf switch in the above scale-up network topology.)

As shown in Figure 1, it’s a 64-GPU SuperPoD that consists of 64 GPUs and four leaf switches with high radix (e.g., 128 400G ports). To achieve inter-GPU bandwidths of several terabits per second (Tbps) or higher, each GPU is typically equipped with multiple scale-up network ports (e.g., four 800 Gbps ports). Each port connects to a separate scale-up leaf switch via a Y-cable, forming four distinct network planes.

In such multi-plane scale-up networks, achieving ultra-high bandwidth and ultra-low latency requires two key strategies. First, efficiently distributing data across all network planes is critical. For instance, if an 800G port on a GPU fails, traffic destined for that GPU over the faulty plane must immediately cease. If only one 400G sub-cable of a given 800G Y-cable malfunctions, halving the bandwidth of the affected network plane, traffic on that network plane between the relevant GPU pair should be proportionally reduced. Second, incast traffic patterns inherent to all-to-all communication may cause congestion on the egress ports of a last-hop switch; therefore, a more efficient congestion management mechanism is required.

This document describes how to extend the Fully Adaptive Routing Ethernet (FARE) using BGP (FARE-BGP in short) as described in [I-D.xu-idr-fare], which was originally designed for scale-out netowrks, to scale-up networks.

2. Terminology

This memo makes use of the terms defined in [RFC2119].

3. Solution Description

Each pair of GPUs establishes multiple Remote Direct Memory Access (RDMA) Queue Pairs (QPs) for data transmission using the loopback addresses of the GPU servers. It is recommended that each loopback address be bound to a single GPU. While the use of port-level or sub-port-level physical addresses for QP establishment is technically supported, this approach is not recommended.

Additionally, upper-layer adaptations (e.g., transaction layer) can facilitate memory semantic operations (load/store/atomic) based on RDMA message semantics. However, implementation details are beyond the scope of this document.

Acting as stub BGP speakers, servers exchange BGP routes with connected switches across different planes, advertising the reachability of their loopback addresses and learning the reachability of remote GPUs. Additionally, by extending FARE-BGP from switches to servers, they can obtain path bandwidth information related to ECMP routes for other GPUs. This capability enables GPUs to perform WECMP load balancing across all available network planes of a scale-up network.

When the path bandwidth of a route through a specific network plane to a destination GPU degrades due to events such as network plane failures or partial link outages, existing Queue Pairs (QPs) traversing unaffected planes maintain their established forwarding paths. Meanwhile, the source GPU must adjust the traffic load allocated to the affected network plane based on updated weight values. Conversely, when the path bandwidth through a previously degraded network plane recovers—such as after failed links or planes are restored—the source GPU should increase the traffic load allocated to that plane. This approach ensures optimal traffic distribution across all operational network planes.

3.1. Per-Flow Weighted Load Balancing

Per-flow weighted load balancing is recommended when ordered packet delivery is essential.

For per-flow weighted load balancing, at least one Queue Pair (QP) per sub-port must be established between a pair of GPUs. When QPs are configured using the loopback address assigned to each GPU, each QP should be assigned a unique UDP source port to differentiate traffic flows across all network planes between the GPU pair. If QPs are configured using the physical addresses assigned to ports, each QP should be assigned a unique UDP source port to differentiate traffic flows across the same network plane. If QPs are configured using the physical addresses assigned to sub-ports, there is no need for assigning unique UDP source port for each QP anymore.

The traffic allocated to a given network plane is evenly distributed among all available QPs traversing that plane.

The switch within each network plane SHOULD perform per-flow load balancing as well to ensure ordered packet delivery for all QPs.

3.2. Per-Packet Weighted Load Balancing

Rer-packet weighted load balancing is recommended in the case where disordered packet delivery is acceptable.

For per-packet weighted load balancing, all QPs established between a pair of GPUs must support disordered packet delivery (e.g., through the Direct Data Placement mechanism [RFC7306]). In this mode, a single QP per network plane between a given GPU pair is sufficient, with the traffic of that QP evenly distributed across all available routes within that network plane.

The switch within each network plane SHOULD perform per-packet weighted load balancing since disordered packet delivery is acceptable for all QPs.

4. Considerations on Memory Semantic Operations

When implementing memory semantics, the ordering guarantees for network transmission can be categorized as follows:

a. Weak Ordering Guarantee for Network Transmission: The network adopts full packet spraying, and the GPUs rely entirely on the Reorder Buffer (ROB) to maintain ordering. This results in a significant increase in implementation complexity on the GPU side.

b. Partial Ordering Constraint for Network Transmission: For transactions with strict ordering requirements (e.g., fence and barrier operations), sequential execution is mandatory. These transactions are marked with a "strong ordering" flag, and the endpoint side uses a blocking mechanism to wait and satisfy the ordering requirement. For transactions that allow out-of-order transmission, the network provides a baseline hash-based ordering guarantee mechanism. When the GPU generates transactions with the same hash key, in-order delivery is enforced between these transactions. This approach grants the GPU ample flexibility while enabling fine-grained local control over ordering.

c. Strong Ordering Guarantee for Network Transmission: To simplify the implementation of memory semantic transactions, some GPUs require that the same transaction stream be transmitted strictly in order along the entire network path, with out-of-order transmission completely prohibited. This achieves a highly simplified implementation on the GPU side.

When implementing native Load/Store memory semantics directly on top of RDMA QPs, additional purpose-built mechanisms are required to guarantee the sequential consistency of memory transactions—particularly for GPUs built on weak-order memory models. Specifically, for weak-order memory models, transactions of the same type targeting the same memory address must maintain consistent ordering throughout their entire network transmission and transaction processing pipeline. To achieve this, transactions should be routed

to the same QP via a hash-based strategy: all transactions targeting the same memory address are hashed to the same QP. Furthermore, each QP enforces strict in-order transmission and completion along its dedicated network path when operating in per-flow weighted load-balancing mode.

5. Acknowledgements

TBD.

6. IANA Considerations

TBD.

7. Security Considerations

TBD.

8. References

8.1. 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>.

8.2. Informative References

[I-D.xu-idr-fare] Xu, X., Hegde, S., Patel, K., He, Z., Wang, J., Huang, H., Zhang, Q., Wu, H., Liu, Y., Xia, Y., Wang, P., and Tiezheng, "Fully Adaptive Routing Ethernet using BGP", Work in Progress, Internet-Draft, draft-xu-idr-fare-04, 18 December 2025, <https://datatracker.ietf.org/doc/html/draft-xu-idr-fare-04>.

[RFC7306] Shah, H., Marti, F., Noureddine, W., Eiriksson, A., and R. Sharp, "Remote Direct Memory Access (RDMA) Protocol Extensions", RFC 7306, DOI 10.17487/RFC7306, June 2014, <https://www.rfc-editor.org/info/rfc7306>.

Authors' Addresses

Xiaohu Xu
China Mobile
Email: xuxiaohu_ietf@hotmail.com

Zongying He
Broadcom
Email: zongying.he@broadcom.com

Nan Wang
Intel
Email: nan.wang@intel.com

Nan Wang
Hygon
Email: wangn@hygon.cn

Hua Wang
Moore Threads
Email: wh@mthreads.com

Jian Guo
Biren Technology
Email: jguo@birentech.com

Xiang Li
Enflame Technology
Email: xiang.li@enflame-tech.com

Tianyou Zhou
Resnics Technology
Email: tzhou@resnics.com

Yongtao Yang
Centec
Email: yangyt@centec.com

Yinben Xia
Tencent
Email: forestxia@tencent.com

Weifeng Zhang
Tencent
Email: wikkizhang@tencent.com

Peilong Wang
Baidu
Email: wangpeilong01@baidu.com

Yan Zhuang
Huawei Technologies
Email: zhuangyan.zhuang@huawei.com

Fajie Yang
Cloudnine Information Technologies
Email: yangfajie@cloudnineinfo.com

Chao Li
Metanet Networking Technology
Email: lichao22@ieisystem.com

Wang Xiaojun
Ruijie Networks
Email: wxj@ruijie.com.cn