Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 19 November 2025

X. Xu
China Mobile
S. Hegde
Juniper
Z. He
Broadcom
J. Wang
Centec
H. Huang
Huawei
Q. Zhang
H3C
H. Wu
Ruijie Networks
Y. Liu
Y. Xia
Tencent
P. Wang
Baidu

18 May 2025

# Fully Adaptive Routing Ethernet using LSR
draft-xu-lsr-fare-04

## Abstract

Large language models (LLMs) like ChatGPT have become increasingly popular in recent years due to their impressive performance in various natural language processing tasks. These models are built by training deep neural networks on massive amounts of text data, as well as visual and video data, often consisting of billions or even trillions of parameters. However, the training process for these models can be extremely resource-intensive, requiring the deployment of thousands or even tens of thousands of GPUs in a single AI training cluster. Therefore, three-stage or even five-stage CLOS networks are commonly adopted for AI networks. The non-blocking nature of the network become increasingly critical for large-scale AI models. Therefore, adaptive routing is necessary to dynamically distribute traffic to the same destination over multiple equal-cost paths, based on network capacity and even congestion information along those paths.

## 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



Copyright Notice



Table of Contents

## 1. Introduction

Large language models (LLMs) like ChatGPT have become increasingly popular in recent years due to their impressive performance in various natural language processing tasks. These models are built by training deep neural networks on massive amounts of text data, as well as visual and video data, often consisting of billions or even trillions of parameters. However, the training process for these models can be extremely resource-intensive, requiring the deployment of thousands or even tens of thousands of GPUs in a single AI training cluster. Therefore, three-stage or even five-stage CLOS networks are commonly adopted for AI networks. Furthermore, In rail-optimized CLOS network topologies with standard GPU servers (HB domain of eight GPUs), the Nth GPUs of each server in a group of servers are connected to the Nth leaf switch, which provides higher bandwidth and non-blocking connectivity between the GPUs in the same rail. In rail-optimized network topology, most traffic between GPU servers would traverse the intra-rail networks rather than the inter-rail networks. In addition, whether in rail-optimal or rail-free networks, collective communication job schedulers always opt to schedule jobs with network topology awareness to minimize the amount of traffic going to the upper layers of the network.

The non-blocking nature of the network, particularly at the lower layers, is essential for large-scale AI training clusters. AI workloads are usually very bandwidth-hungry and often generate several large data flows simultaneously. If traditional hash-based ECMP load balancing is used without optimization, it can lead to serious congestion and high latency in the network when multiple large data flows are directed to the same link. This congestion can result in longer-than-expected model training times, as job completion time depends on worst-case performance. Therefore, adaptive routing is necessary to dynamically distribute traffic to the same destination across multiple equal-cost paths, taking into account network capacity and even congestion along these paths. In essence, adaptive routing is a capacity- and even congestion-aware dynamic path selection algorithm.

Furthermore, to reduce the congestion risk to the maximum extent, the routing should be more granular if possible. Flow-granular adaptive routing still has a certain statistical possibility of congestion. Therefore, packet-granular adaptive routing is more desirable although packet spray would cause out-of-order delivery issues. A flexible reordering mechanism must be put in place（e.g., egress ToRs or the receiving servers). Recent optimizations for RoCE and newly invented transport protocols as alternatives to RoCE no longer require handling out-of-order delivery at the network layer. Instead, the message processing layer is used to address it.

To enable adaptive routing, no matter whether flow-granular or packet-granular adaptive routing, it is necessary to propagate network topology information, including link capacity across the CLOS network. Therefore, it seems straightforward to use link-state protocols such as OSPF or ISIS as the underlay routing protocol in the CLOS network, instead of BGP.

Hence, this document defined a new prefix attribute sub-TLV referred to as Path Bandwidth sub-TLV, and describes how to use this sub-TLV together with the Maximum Bandwidth sub-TLV of the Link TLV as defined in OSPF or ISIS TE extensions [RFC3630] [RFC5329][RFC5305] to calculate end-to-end path bandwidth within the data center fabric so as to achieve adaptive routing.

For information on how to resolve the flooding issue caused by the use of link-state protocols in large-scale CLOS networks, please refer to the following document [I-D.xu-lsr-flooding-reduction-in-clos].

Note that while adaptive routing, especially at the packet-granular level can help reduce congestion between switches in the network, thereby achieving a non-blocking fabric, it does not address the incast congestion issue which is commonly experienced in last-hop switches that are connected to the receivers in many-to-one communication patterns. Therefore, a congestion control mechanism is always necessary between the sending and receiving servers to mitigate such congestion.

## 2. Terminology

This memo makes use of the terms defined in [RFC1195] [RFC2328] and [RFC5340].

## 3. Path Bandwidth Sub-TLV

When advertising IP reachability information across ISIS levels or OSPF areas, it needs to contain the path bandwidth associated with the advertised IP prefix which is used to indicate the minimum bandwidth of all links along the path towards that prefix.

For ISIS, an optional sub-TLV referred to as Path Bandwidth sub-TLV is to be defined. This sub-TLV is type of TBD, and is four octets in length. The value is filled with the path bandwidth associated with a given prefix in IEEE floating point format. The units are bytes per second. This sub-TLV COULD be conveyed in TLVs 135, 235, 236, or 237 , just like those prefix attribute-related sub-TLVs as defined in [RFC7794].

For OSPFv2, since The OSPFv2 Extended Prefix TLV [RFC7684] is used to advertise additional attributes associated with the prefix, an optional sub-TLV of the OSPFv2 Extended Prefix TLV referred to as Path Bandwidth sub-TLV is to be defined. This sub-TLV is type of TBD, and is four octets in length. The value is filled with the path bandwidth associated with a given prefix in IEEE floating point format. The units are bytes per second.

For OSPFv3, an optional sub-TLV of the Intra-Area-Prefix TLV, Inter-Area-Prefix TLV, and External-Prefix TLV [RFC8362] referred to as Path Bandwidth sub-TLV is to be defined. This sub-TLV is type of TBD, and is four octets in length. The value is filled with the path bandwidth associated with a given prefix in IEEE floating point format. The units are bytes per second.

# 4. Solution Description

## 4.1. Adaptive Routing in 3-stage CLOS

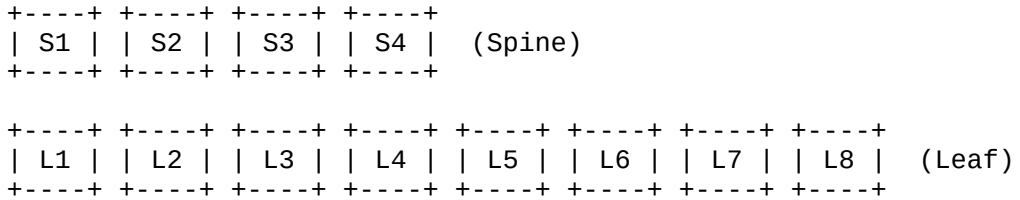


Figure 1

(Note that the diagram above does not include the connections between nodes. However, it can be assumed that leaf nodes are connected to every spine node.)

In a three-stage CLOS network as shown in Figure 1, also known as a leaf-spine network, all nodes MAY be in OSPF area zero or ISIS Level-2.

Leaf nodes and spine nodes are enabled for adaptive routing. As such, those nodes will advertise the link capacity by using the Maximum Bandwidth sub-TLV. In addition, leaf nodes will advertise the path bandwidth associated with each prefix originating from them by using the Path Bandwidth sub-TLV. The value of the Path Bandwidth sub-TLV is filled with a maximum bandwidth value by default.

When a leaf node, such as L1, calculates the shortest path to a particular IP prefix originated by another leaf node in the same OSPF area or ISIS Level-2 area, say L2, four equal-cost paths via four spine nodes (e.g., S1, S2, S3, and S4) respectively will be calculated. To achieve adaptive routing, the capacity associated with each path SHOULD be considered as a weight value of that path when performing weighted ECMP load-balancing. In particular, the minimum value among the capacity of the upstream link (e.g., L1->S1) , the capacity of the downstream link (S1->L2) of a given path (e.g., L1->S1->L2) and the path bandwidth associated with that prefix would be used as a weight value for that end-to-end path when performing weighted ECMP load-balancing.

## 4.2. Adaptive Routing in 5-stage CLOS

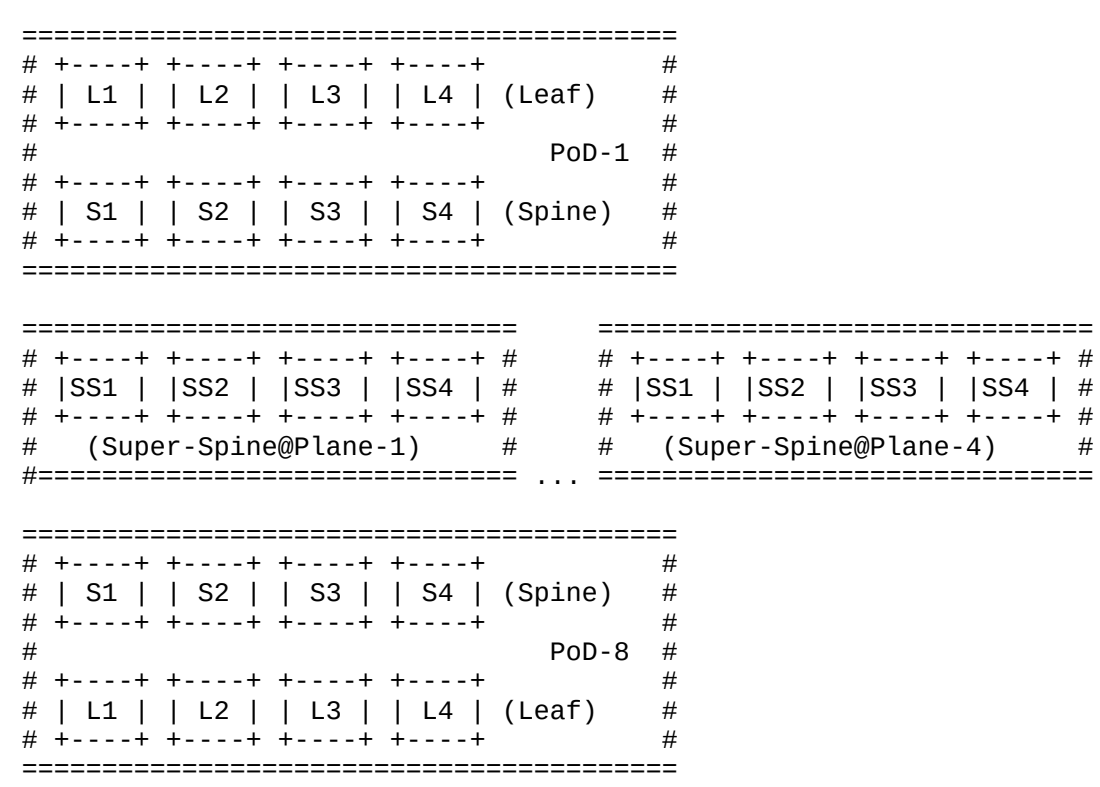


Figure 2

(Note that the diagram above does not include the connections between nodes. However, it can be assumed that the leaf nodes in a given PoD are connected to every spine node in that PoD. Similarly, each spine node (e.g., S1) is connected to all super-spine nodes in the corresponding PoD-interconnect plane (e.g., Plane-1).)

For a five-stage CLOS network as illustrated in Figure 2, each Pod consisting of leaf and spine nodes is configured as an OSPF non-zero area or an ISIS Level-1 area. The PoD-interconnect plane consisting of spine nodes and super-spine nodes is configured as an OSPF area zero or an ISIS Level-2 area. Therefore, spine nodes play the role of OSPF area border routers or ISIS Level-1-2 routers.

All nodes are enabled for adaptive routing. As such, those nodes will advertise the link capacity by using the Maximum Bandwidth sub-TLV. In addition, leaf nodes will advertise the path bandwidth associated with each prefix originating from itself by using the Path Bandwidth sub-TLV. The value of the Path Bandwidth sub-TLV SHOULD be filled with a maximum bandwidth value by default.

When leaking an IP prefix reachability from an OSPF non-zero area to area zero or from ISIS level-1 to level-2 (e.g., an IP prefix attached to a leaf node, such as L1@PoD-1), the path bandwidth value associated with the prefix would be readvertised or updated by OSPF border routers or ISIS level-1-2 routers (e.g., S1@PoD-1), and the value is filled with the minimum value between the bandwidth of the link towards the originating router (e.g., L1@PoD-1) and the original path bandwidth value associated with the prefix.

When leaking the above IP prefix reachability from the OSPF area zero to a non-zero area or from ISIS level-2 to level-1, the path bandwidth value associated with the prefix would be readvertised or updated by OSPF border routers or ISIS level-1-2 routers (e.g., S1@PoD-8) and the value is filled with the minimum value between the original bandwidth value associated with the prefix and the total bandwidth of all paths towards the advertising router of that prefix (e.g., S1@PoD-1).

When a leaf node within PoD-8, calculates the shortest path to the above IP prefix, four equal-cost paths will be created via four spine nodes: S1, S2, S3, and S4 in PoD-8. To enable adaptive routing, the capacity of each path SHOULD be considered as a weight value for weighted ECMP load-balancing. In particular, the minimum value between the capacity of the upstream link (e.g., L1@Pod-8->S1@Pod-8) of each path (e.g., L1@Pod-8->S1@PoD-8) and the path bandwidth associated with that prefix is used as a weight value of that path when performing weighted ECMP load-balancing.

## 5. Modifications to SPF Computation Behavior

Once an OSPF or ISIS router is enabled for adaptive routing, the capacity of each SPF path SHOULD be calculated as a weight value of that path for weighted ECMP load-balancing purposes.

## 6. Acknowledgements

TBD.

## 7. IANA Considerations

TBD.

## 8. Security Considerations

TBD.

## 9. References

### 9.1. Normative References

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, DOI 10.17487/RFC1195, December 1990, <https://www.rfc-editor.org/info/rfc1195>.

[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>.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <https://www.rfc-editor.org/info/rfc3630>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC5329] Ishiguro, K., Manral, V., Davey, A., and A. Lindem, Ed., "Traffic Engineering Extensions to OSPF Version 3", RFC 5329, DOI 10.17487/RFC5329, September 2008, <https://www.rfc-editor.org/info/rfc5329>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <https://www.rfc-editor.org/info/rfc7684>.

[RFC8362] Lindem, A., Roy, A., Goethals, D., Reddy Vallem, V., and F. Baker, "OSPFv3 Link State Advertisement (LSA) Extensibility", RFC 8362, DOI 10.17487/RFC8362, April 2018, <https://www.rfc-editor.org/info/rfc8362>.

## 9.2. Informative References

[I-D.xu-lsr-flooding-reduction-in-clos] Xu, X., "Flooding Reduction in CLOS Networks", Work in Progress, Internet-Draft, draft-xu-lsr-flooding-reduction-in-clos-01, 21 November 2023, <https://datatracker.ietf.org/doc/html/draft-xu-lsr-flooding-reduction-in-clos-01>.

[RFC7794] Ginsberg, L., Ed., Decraene, B., Previdi, S., Xu, X., and U. Chunduri, "IS-IS Prefix Attributes for Extended IPv4 and IPv6 Reachability", RFC 7794, DOI 10.17487/RFC7794, March 2016, <https://www.rfc-editor.org/info/rfc7794>.

## Authors' Addresses

Xiaohu Xu
China Mobile
Email: xuxiaohu_ietf@hotmail.com

Shraddha Hegde
Juniper
Email: shraddha@juniper.net

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

Junjie Wang
Centec
Email: wangjj@centec.com

Hongyi Huang
Huawei
Email: hongyi.huang@huawei.com

Qingliang Zhang
H3C
Email: zhangqingliang@h3c.com

Hang Wu
Ruijie Networks
Email: wuhang@ruijie.com.cn

Yadong Liu
Tencent
Email: zeepliu@tencent.com

Yinben Xia
Tencent
Email: forestxia@tencent.com

Peilong Wang
Baidu
Email: wangpeilong01@baidu.com