IDR
Internet-Draft
Intended status: Standards Track
Expires: 16 September 2026

J. Zhang
R. Zhuang
China Mobile
Z. Zhang
D. Yuan
ZTE Corporation

15 March 2026

BGP PORT EC for AIDC
draft-zhang-idr-portid-ec-01

Abstract

This document introduces a new BGP extended community attribute for use in AI computing, which announces the port ID between Leaf switches and servers as preparation for sending large-scale traffic before initiating AI tasks.

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 16 September 2026.

Copyright Notice

Table of Contents

1. Introduction

With the rapid development of Artificial Intelligence (AI) and Machine Learning (ML), AI tasks often generate large traffic due to the characteristics of large language model computation (LLM). If the link bandwidth is insufficient, packet loss may occur. AI computation has very high reliability requirements and extremely low tolerance for packet loss and latency. When there is link congestion in the network that leads to packet loss or excessive latency, it will have a significant impact on the computational efficiency of AI tasks.

In data centers used for AI and machine learning, BGP is often used as the routing protocol [RFC7938]. In some implementations, sufficient bandwidth between the destination server and its connected leaf switches must be ensured before sending traffic for AI tasks. On the network side, specifically the area comprised of the Leaf and Spine switches in Figure 1, there are numerous ECMP links. Techniques such as Packet Spray can be used to minimize congestion and packet loss. However, on the computing side, specifically the last hop between the Leaf switches and the server, congestion can easily lead to packet loss, significantly reducing the efficiency of AI tasks. To minimize or eliminate packet loss on the last hop, BGP needs to be extended to include port information on the destination leaf switch. This allows the sender to negotiate based on this information before sending traffic, ensuring sufficient bandwidth is

available in the last hop and preventing congestion and packet loss due to insufficient bandwidth. To reduce the stress caused by full-mesh connections, Leaf switches do not establish neighbors with each other.

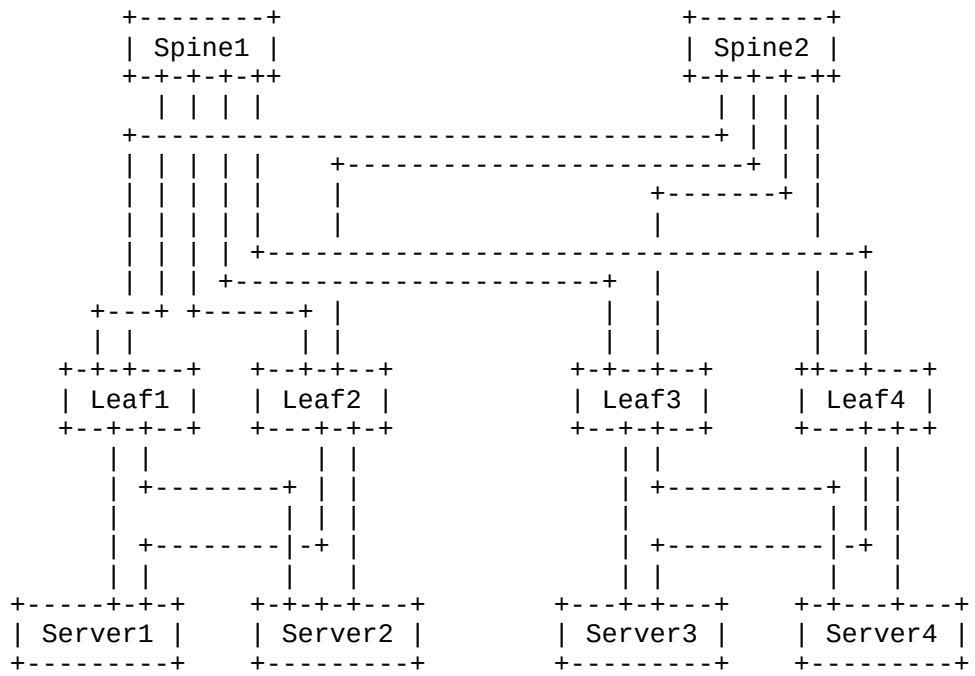

Figure 1

Figure 1 shows a typical data center used for AI computing. In this network, when Server2, 3, or 4 sends traffic to Server1 through leaf1, a common incast congestion problem may occur. That is, the link 1 between Leaf1 switch and Server1 may be congested due to insufficient bandwidth, resulting in packet loss.

Currently, some implementations negotiate before sending traffic from devices like Server2 and Server3 to Server1. The AI task traffic is only sent if the link bandwidth between the destination server and its connected Leaf switch (referred to as the destination switch) is sufficient. This negotiation method is outside the scope of this draft. However, before negotiation, the port information connecting the destination switch to the server needs to be obtained. This information will be sent via the newly added extended community "Route Port ID" in BGP.

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. Format

When announcing the route to the connected server, the BGP protocol on the Leaf switch carries the switch’s address and the port ID information connected to the destination server.

Transitive IPv4-Address-Specific Extended Community defined in [RFC7153] and [I-D.ietf-idr-rfc4360-bis] with new sub-type "Route Port ID" is used for carry the IPv4 address of Leaf switch and the related port ID to the destination server.

Transitive IPv6-Address-Specific Extended Community defined in [RFC5701] with new sub-type "Route Port ID" is used for carry the IPv6 address of the leaf switch and the related port ID to the destination server.

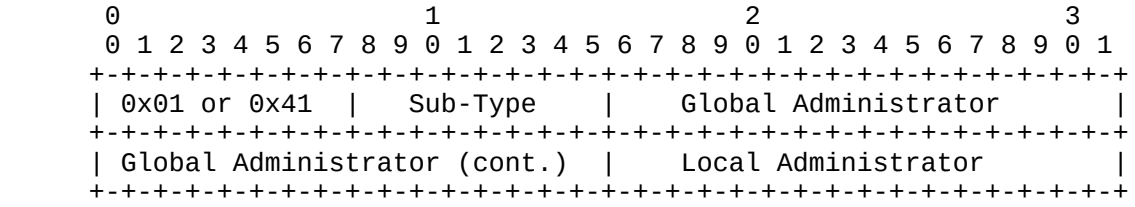

Figure 2

Figure 2 shows the format of IPv4-Address-Specific Extended Community, where:

* Sub-Type: TBD. This indicates that this is the Route Port ID extended community;

* Global Administrator: 4 octets, set to the IPv4 address of the switch that advertises the server route. This address can be the loopback address for establishing the BGP connection;

* Local Administrator: 2 octets, set to the ID of the port connecting the switch and the server, with a value range of 1 to 65535.

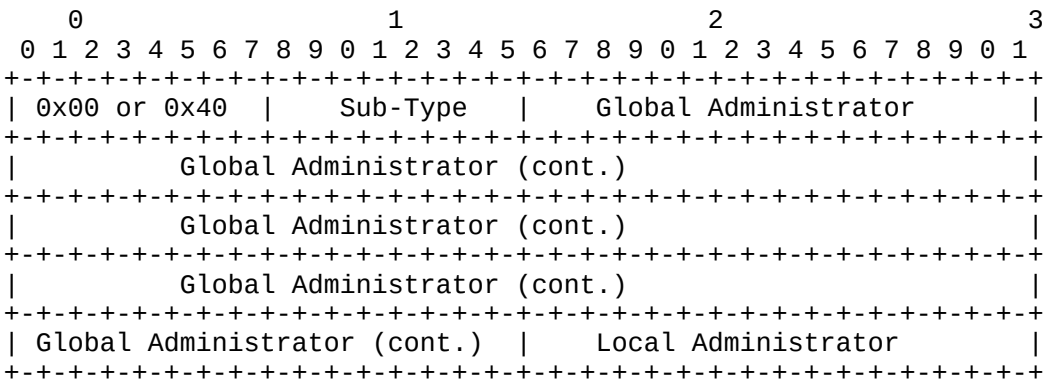

Figure 3

Figure 3 shows the format of IPv6-Address-Specific Extended Community, where:

* Sub-Type: TBD. This indicates that this is the Route Port ID extended community;

* Global Administrator: 16 octets, set to the IPv6 address of the switch that advertises the server route. This address can be the loopback address for establishing the BGP connection;

* Local Administrator: 2 octets, set to the ID of the port connecting the switch and the server, with a value range of 1 to 65535.

3. Specification

When the Leaf switch advertises routes to the server, the advertisement includes the Route Port ID extended community, which is transmitted along with the route advertisement.

In the example shown in Figure 1, Leaf1, when advertising routes to the Spine switch, includes the Route Port ID extended community, which contains the Loopback address used to establish the BGP connection and the port ID connected to the server. The Leaf2 is the same.

Upon receiving the route carrying the Route Port ID extended community, the leaf switch checks if the address is reachable. If unreachable, the extended community is ignored. If reachable, the address and port information are stored locally or sent to the server. This storing or sending process is outside the scope of this draft.

Because data centers used for AI computing have a large number of ECMP paths, deploying this feature requires enabling the ADD-PATH advertisement function defined in [RFC7911], to ensure the propagation of extended community attributes. Spine or higher-level switches do not need to generate entries based on this extended community attribute. To avoid a large number of route advertisements that may result from enabling the ADD-PATH function, this advertisement SHOULD be limited to a single PoD.

When a server wants to send large traffic for AI tasks, it will negotiate bandwidth based on the destination switch and port information obtained from BGP. Traffic will only be sent after successful negotiation, thus avoiding packet loss caused by congestion. Traffic will be sent to the server via the successfully negotiated Leaf switch. This negotiation process is outside the scope of this draft.

In the example shown in Figure 1, the routes advertised by Leaf1 and Leaf2 to Server1 will carry the Route Port ID extended community. When Server3 wants to send AI task traffic to Server1, it can first negotiate with Leaf1. If the negotiation fails, it may negotiate with Leaf2. Only after the negotiation succeeds will the traffic be sent. In this example, assuming Leaf1 is successfully negotiated, traffic will be sent to Server1 through Leaf1.

4. IANA Considerations

IANA is requested to allocate two new code points from the "Transitive IPv4-Address-Specific Extended Community Sub-Types" and the "Transitive IPv6-Address-Specific Extended Community Sub-Types" registry.

Type	Description	Reference
TBD	Route Port ID	This Document

Table 1: TABLE_1

5. Security Considerations

This extension to BGP has similar security implications as BGP Extended Communities [RFC7153], [RFC5701] and [I-D.ietf-idr-rfc4360-bis].

6. References

6.1. Normative References

[I-D.ietf-idr-rfc4360-bis]
Sangli, S. R. and N. Kao, "BGP Extended Communities Attribute", Work in Progress, Internet-Draft, draft-ietf-idr-rfc4360-bis-02, 6 December 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-rfc4360-bis-02>.

[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>.

[RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, DOI 10.17487/RFC5701, November 2009, <https://www.rfc-editor.org/info/rfc5701>.

[RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, DOI 10.17487/RFC7153, March 2014, <https://www.rfc-editor.org/info/rfc7153>.

6.2. Informative References

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

Authors' Addresses

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

Rui Zhuang
China Mobile
China
Email: zhuangruiyjy@chinamobile.com

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

Dongyu Yuan
ZTE Corporation
China
Email: yuan.dongyu@zte.com.cn