Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 7 March 2026

P. Mohapatra
Google LLC
R. Das, Ed.
Juniper Networks, Inc.
S. Mohanty, Ed.
Zscaler
S. Krier
Cisco Systems
R.J. Szarecki
Google LLC
A. Gattani
Arista Networks

3 September 2025

BGP Link Bandwidth Extended Community
draft-ietf-idr-link-bandwidth-16

Abstract

This document describes an application of BGP extended communities that allows a router to perform WECMP (Weighted Equal-Cost Multipath).

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].

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 March 2026.

Table of Contents

1. Introduction

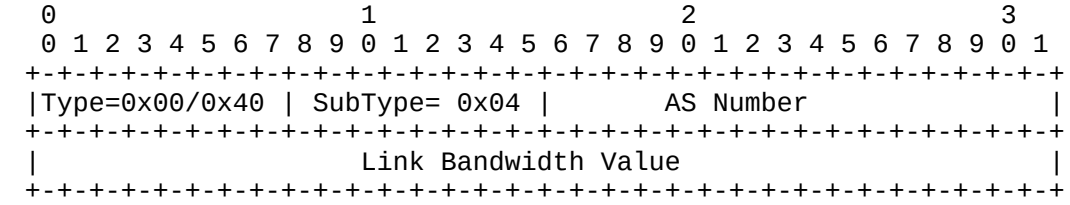
Load balancing is a critical aspect of network design, enabling efficient utilization of available bandwidth and improving overall network performance. Traditional equal-cost multi-path (ECMP) routing does not account for the varying capacities of different paths. This document suggests that the external link bandwidth be carried in the network using one of two new extended communities [RFC4360] - the transitive and non-transitive Link Bandwidth Extended Community. The Link Bandwidth Extended Community provides a mechanism for routers to advertise the bandwidth of their downstream

path(s), facilitating maximum utilization of network resources.

2. Link Bandwidth Extended Community

The Link Bandwidth Extended Community is defined as a BGP extended community that carries the bandwidth information of a router, represented by BGP Protocol Next Hop, connecting to remote network. This community can be used to inform other routers about the available bandwidth through a given route.

The Link Bandwidth Extended Community can be either transitive or non-transitive. Therefore the value of the high-order octet of the extended Type Field can be 0x00 or 0x40, respectively. The value of the low-order octet of the extended type field for this communities is 0x04. The value of the Global Administrator subfield in the Value Field SHOULD represent the Autonomous System of the router that attaches the Link Bandwidth Extended Community, but it can be set to any 2-byte value. If the Autonomous System number cannot be represented in two octets, as enabled by [RFC6793], AS_TRANS should be used in the Global Administrator subfield. The encoding of 4-octet ASN is out of scope of this document. The bandwidth of the link is expressed as 4 octets in [IEEE.754-2019] floating point format, units being bytes (not bits!) per second. It is carried in the Local Administrator subfield of the Value Field.



Type: 1-octet field MUST be set to 0x00 or 0x40 to indicate transitive/non-transitive.

SubType: 1-octet field MUST be set to 0x04 to indicate 'Link-Bandwidth'.

Global Administrator sub-field: 2-octet represent the Autonomous System.

Local Administrator sub-field: Bandwidth value (bytes per sec) encoded as 4 octets in IEEE floating point format.

Figure 1: Link Bandwidth Extended Community

3. Protocol Procedures

3.1. Sender (Originating Link Bandwidth Extended Community)

An originator of Link Bandwidth Extended Community SHOULD be able to originate either a transitive or a non-transitive Link Bandwidth Extended Community. Implementations SHOULD provide configuration to set the transitivity type of the Link Bandwidth Extended Community, as well as the Global Administrator and bandwidth values in (Local Administrator field), using local policy. For backward compatibility, different implementations MAY use different default values for the transitivity type of the Link Bandwidth Extended Community. The provided configuration SHOULD allow operators to override the default transitivity value as needed. An implementation MAY advertise a link bandwidth value as zero.

No more than one Link Bandwidth Extended Community SHOULD be attached to a route. For purpose of backward compatibility during transition, a BGP speaker MAY attach one Link Bandwidth Extended Community per transitivity (transitive/non-transitive) both having the same 'Link Bandwidth Value' field.

A Link Bandwidth Extended Community MAY be attached or updated for a BGP route upon receipt during Adj-RIB-In processing. The Link Bandwidth Extended Community MAY be attached or updated for a BGP route's Adj-RIB-Out entry while being advertised to a neighboring BGP speaker.

Note: Implementations MAY provide a configuration option to send non-transitive Link Bandwidth Extended Communities on external BGP sessions.

3.2. Receiver (Receiving Link Bandwidth Extended Community)

A BGP receiver MUST be able to process Link Bandwidth Extended Community of both transitive and non-transitive types. The receiver MUST NOT flap or treat the route as malformed based on the transitivity of the Link Bandwidth Extended Community and/or BGP session type (internal vs. external).

Note: Implementations MAY provide configuration to accept non-transitive Link Bandwidth Extended Communities from external BGP sessions.

Implementations MUST be able to process and accept a Link Bandwidth Extended Community where the bandwidth value is set to zero. WECMP can be utilized when all contributing paths have a non-zero value in the Link Bandwidth Extended Community.

In case some paths have a zero value but others have non-zero value, or all paths have Link Bandwidth with zero value, the behavior is determined by local policy. For example, an implementation may exclude the paths with zero value from WECMP formation or an implementation may fallback to ECMP.

3.3. Re-advertisement Procedures

3.3.1. Re-advertisement with Next hop Self

When a BGP speaker re-advertises a route with Link Bandwidth Extended Community and sets the next hop to itself, it SHOULD follow the same procedures as outlined in Section 3.1.

In the absence of any import or export policies that alter the Link Bandwidth Extended Community, any received Link Bandwidth Extended Community on the route will be re-advertised unchanged, in accordance with standard BGP procedures.

3.3.2. Re-advertisement with Next Hop Unchanged

A BGP speaker that receives a route with a Link Bandwidth Extended Community, re-advertises or reflects the same without changing its next hop, SHOULD NOT change the Link Bandwidth Extended Community in any way.

3.4. Link Bandwidth Extended Community Arithmetic and BGP Multipath

In a BGP multipath ECMP environment, the link bandwidth value that is sent or re-advertised may be calculated based on the Link Bandwidth Extended Community of the routes contributing to multipath in the Local Routing Information Base (Local-RIB). This topic is beyond the scope of this document.

4. Error Handling

If a BGP speaker receives a route with more than one Link Bandwidth Extended Communities and uses the route to compute WECMP, it SHOULD use the extended community with the lowest "Link Bandwidth Value", ignoring the transitivity. Implementations MAY provide configuration to change the above preference.

Between transitive and non-transitive types of Link Bandwidth Extended Communities that have the same ’Link Bandwidth Value’, the transitivity doesn’t matter for purpose of computing WECMP or programming to FIB (Forwarding Information Base).

Note that these procedures mean that a BGP speaker reflecting a route with next hop unchanged (e.g. RR) will re-advertise the Link Bandwidth Extended Communities received on the route as-is without any modification, while following the extended community transitivity rules.

Link Bandwidth Extended Communities with a negative value SHALL be ignored and MUST NOT be originated.

If any of the paths lack a valid Link Bandwidth Extended Community, ECMP (Equal-Cost Multi-Path) MUST be used instead.

5. Document History

BGP Link Bandwidth Extended Community has evolved over several versions of the IETF draft. In the earlier versions up to draft-ietf-idr-link-bandwidth-08, only the non-transitive version of Link Bandwidth Extended Community was supported. However, starting from draft-ietf-idr-link-bandwidth-09, both transitive and non-transitive versions of Link Bandwidth Extended Community are supported.

An old sender/receiver is a BGP speaker that uses procedures up to draft (https://datatracker.ietf.org/doc/html/draft-ietf-idr-link-bandwidth-08) or any undocumented behavior for Link Bandwidth Extended Community.

A new sender/receiver is a BGP speaker that implements procedures specified in this document.

A BGP speaker (Sender or Receiver) needs to be upgraded to support the procedures defined in this document to provide full interoperability for both transitive and non-transitive versions of Link Bandwidth Extended Community. In order to simplify implementations, it is not a goal to provide interoperability by upgrading only the RR.

6. IANA Considerations

This document defines a specific application of the two-octet AS specific extended community.

IANA is requested to update the Transitive Two-Octet AS-Specific Extended Community Sub-Types registry (Type 0x00) and Sub-Type 0x04 to:

Name
transitive Link Bandwidth Extended Community

IANA is requested to update the Non-Transitive Two-Octet AS-Specific Extended Community Sub-Types registry (Type 0x40) and Sub-Type 0x04 to:

Name
non-transitive Link Bandwidth Extended Community

Both updates are to Reference this document.

7. Security Considerations

There are no additional security risks introduced by this design.

8. Operational Considerations

8.1. Inconsistent Deployment

Prior deployments of the feature specified in this document have involved implementations that only understood one of the two extended community transitivity types. As a result, such implementations would treat the use of the other transitivity type in a "ships in the night" fashion. The procedures in this document govern how multiple transitivity types for link bandwith should operate.

In circumstances where networks have deployed a mixture of implementations supporting this document’s current procedures for both transitivity types, and older implementations that only understand one transitivity type, inconsistent behavior could result. A primary example is when a route received by a BGP speaker contains both a transitive and a non-transitive Link Bandwidth Extended Community and that BGP speaker performs an operation that updates only one of the Link Bandwidth Extended Communities, the other community may be have an inconsistent value. As a result, downstream BGP speakers that may receive such routes may perform inappropriate ECMP load balancing.

To mitigate such issues, when operators are aware that older implementations are in present in their networks, they may wish to take actions to address such inconsistencies. One example would be to filter either at advertisement time on the older BGP speaker the unsupported transitivity type of Link Bandwidth Extended Community - if the implementation is capable of such filtering. Alternatively, a receiving BGP speaker, knowing that the sending speaker is incapable of doing such operations, could strip the Link Bandwidth Extended Community type that is unsupported by the sender.

Ideally this operational consideration is short-lived until the network has been upgraded to implementations that consistently support the procedures in this draft.

9. Contributors

Kaliraj Vairavakkalai
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: kaliraj@juniper.net

Natrajan Venkataraman
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: natv@juniper.net

Rex Fernando
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: rex@cisco.com

10. Acknowledgments

The authors would like to thank Yakov Rekhter, Srihari Sangli and Dan Tappan for proposing unequal cost load balancing as one possible application of the extended community attribute. The authors would like to thank Jeff Haas for all the discussions and providing text for operational considerations.

The authors would like to thank Bruno Decraene, Robert Raszuk, Joel Halpern, Aleksi Suhonen, Randy Bush, Stephane Litkowski, Mankamana Mishra, Moshiko Nayman, Yingzhen Qu, Anoop Ghanwani, Dongjie (Jimmy) and John Scudder for their comments and contributions.

11. Normative References

[IEEE.754-2019]
IEEE, "IEEE Standard for Floating-Point Arithmetic", 22 July 2019, <https://ieeexplore.ieee.org/document/8766229>.

[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>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, DOI 10.17487/RFC6793, December 2012, <https://www.rfc-editor.org/info/rfc6793>.

Authors' Addresses

Pradosh Mohapatra
Google LLC
Email: pradosh@google.com

Reshma Das (editor)
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: dreshma@juniper.net

Satya Mohanty (editor)
Zscaler
120 Holger Way,
San Jose, CA 95134
United States of America
Email: smohanty@zscaler.com

Serge Krier
Cisco Systems
Pegasus Parc, De Kleetlaan 6a
Belgium
Email: sekrier@cisco.com

Rafal Jan Szarecki
Google LLC
1160 N Mathilda Ave,
Sunnyvale, CA 94089
United States of America
Email: rszarecki@gmail.com

Akshay Gattani
Arista Networks
5453 Great America Parkway
Santa Clara, CA 95054
United States of America
Email: akshay@arista.com