IDR
Internet-Draft
Intended status: Standards Track
Expires: 22 April 2026

K. Wang, Ed.
J. Haas, Ed.
HPE
C. Lin
New H3C Technologies
J. Tantsura
Nvidia

19 October 2025

# BGP Next-next Hop Nodes
# draft-ietf-idr-next-next-hop-nodes-00

## Abstract

BGP speakers learn their next hop addresses for NLRI in RFC 4271 in the NEXT_HOP field and in RFC 4760 in the "Network Address of Next Hop" field. Under certain circumstances, it might be desirable for a BGP speaker to know both the next hops and the next-next hops of NLRI to make optimal forwarding decisions. One such example is global load balancing (GLB) in a Clos network.

Draft-ietf-idr-entropy-label defines the "Next Hop Dependent Characteristics Attribute" (NHC) which allows a BGP speaker to signal the forwarding characteristics associated with a given next hop.

This document defines a new NHC characteristic, the Next-next Hop Nodes (NNHN) characteristic, which can be used to advertise the next-next hop nodes associated with a given next hop.

Table of Contents



## 1. Introduction

BGP speakers learn their next hop addresses for NLRI in [RFC4271] in the NEXT_HOP field and in [RFC4760] in the "Network Address of Next Hop" field. Under certain circumstances, it might be desirable for a BGP speaker to know both the next hops and the next-next hops of NLRI to make optimal forwarding decisions. One such example is the global load balancing (GLB) in a Clos network [I-D.cheng-rtgwg-adaptive-routing-framework].

When a route's ECMP has multiple next hops, packets forwarded using that ECMP are hashed to the member next hops for load balancing purposes. If one of the member next hop links is congested due to

uneven hashing, dynamic load balancing (DLB) allows the node to adjust the hashing so that the congestion on that link can be mitigated. When all next hop link(s) are congested, DLB on the local node will not help to mitigate the congestion. Such nodes will require help from the previous hop(s) to shift the traffic towards alternative nodes to mitigate such congestion. This process is called global load balancing.

In a Clos network, a congested link will affect the load balancing decisions of the previous layer nodes equally. Because of this, the previous-previous layer nodes do not need to change their load balancing decisions towards the previous layer nodes to mitigate this link congestion. This means we only need to know the link congestion status of the next-next hops of given BGP route in order to make GLB decisions. The combined link quality of each next hop and its corresponding next-next hops can be used as the feedback for DLB.

The purpose of this document is to provide a method for BGP to learn the next-next hops - or more specifically, the next-next hop nodes. When a next hop node has more than one next-next hops towards a next-next hop node, DLB helps to balance the load between the multiple next-next hops by locally adjusting the volume of traffic hashed over a given ECMP member link. Thus, only the overall link congestion between the next hop node and the next-next hop node is important for GLB.

Note that the mechanism for detecting link congestion and communicating them to the previous hop nodes is out of the scope of this document. [I-D.zzhang-rtgwg-router-info] defined one approach to notify link quality (congestion/failure) to directly connected nodes.

This document defines a new NHC characteristic, the Next-next Hop Nodes (NNHN) characteristic, for the BGP Next Hop Dependent Characteristics Attribute (NHC) defined in [I-D.ietf-idr-entropy-label]. A downstream BGP speaker can use the NNHN to advertise the next-next hop nodes corresponding to the next hop of an NLRI. This allows the upstream BGP speaker to learn the next-next hop nodes corresponding to each of its next hop nodes.

## 1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 2. BGP Next-next Hop Nodes (NNHN) Characteristic

[I-D.ietf-idr-entropy-label] defines NHC as a container for characteristic TLVs. Next-next Hop Nodes is one such characteristic. It specifies the next-next hop nodes corresponding to the next hop field in the NHC.

### 2.1. Encoding NNHN

The NNHN TLV has the NHC characteristic code 2, as assigned in Section 5 of [I-D.ietf-idr-entropy-label]. The NHC characteristic length specifies the remaining number of octets in the NNHN TLV. The NNHN characteristic format is shown in Figure 1:

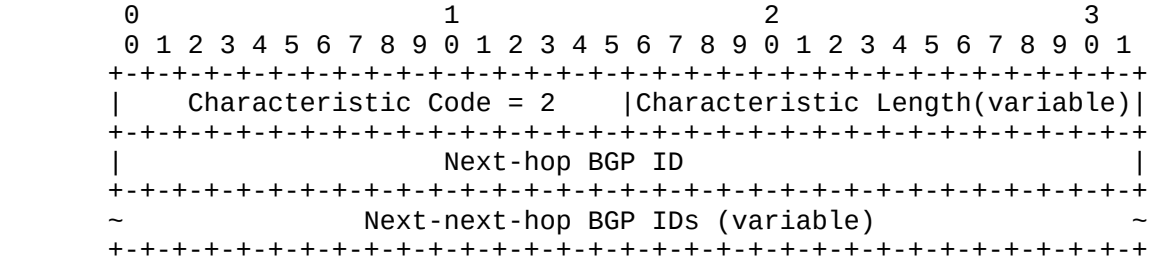


Figure 1: NNHN Characteristic TLV Format

Next-hop BGP ID:
32-bit BGP Identifier of the next hop node attaching this NHC characteristic.

Next-next-hop BGP IDs:
One or more 32-bit BGP Identifiers, each representing a next-next hop node used by the next hop node for ECMP forwarding for the NLRI in the BGP Update.

### 2.2. Sending NNHN

All procedures from Section 2.2 of [I-D.ietf-idr-entropy-label] apply.

When a BGP speaker S has a BGP route R it wishes to advertise with next hop self to its peer, it MAY choose to originate an NNHN characteristic. The "Next-hop BGP ID" field MUST be set to the BGP Identifier this BGP speaker uses with the peer.

For all the ECMP paths of route R which are used for forwarding, the BGP Identifiers of those BGP peers MUST be encoded as the "Next-next-hop BGP IDs". When more than one paths are from the same BGP peer, the characteristic MUST have only one BGP Identifier of that peer.

When there are more than one "Next-next-hop BGP IDs" in the characteristic, they MUST be encoded in the numerically ascending order treating the BGP Identifier as a network byte order encoded 32-bit unsigned integer.

An NNHN with no "Next-next-hop BGP IDs" MUST NOT be sent.

When a BGP speaker S has a BGP route R it wishes to advertise with next hop self to its peer, it MUST NOT forward the NNHN characteristic received from downstream peers. It either originates its own NNHN characteristic as described above or does not send one.

When a BGP speaker S has a BGP route R it wishes to advertise with the next hop that has not been set to self, it MUST NOT originate an NNHN characteristic. However, if a NNHN characteristic has been received for route R and passed the NHC validation as defined in [I-D.ietf-idr-entropy-label], the NNHN characteristic SHOULD be forwarded.

A BGP speaker MUST NOT include more than one instance of NNHN in an NHC.

## 2.3. Receiving NNHN

All procedures from Section 2.3 of [I-D.ietf-idr-entropy-label] apply.

When a BGP speaker wishes to enforce hop-by-hop eBGP propagation of the NNHN, if the received NNHN characteristic’s Next-hop BGP Identifier does not match the BGP Identifier of the BGP speaker the UPDATE was received from, it MUST be ignored and discarded.

The receiver of the NNHN characteristic MUST be able to handle any order of the "Next-next-hop BGP IDs".

Duplicate BGP Identifiers in the "Next-next-hop BGP IDs" MUST be silently ignored.

The receiver of the NNHN characteristic will use it together with the next-hop to determine the first two hops of a route. One example of using NNHN characteristic is for global load balancing: When link quality information is received from the next-hop node, through approaches like [I-D.zzhang-rtgwg-router-info], NNHN ID(s) of the

route can be matched against the node IDs from the link quality to determine the next-next-hop link quality reported for this route. The next-next-hop link quality will be combined with the next-hop link quality for global load balancing. Another example of using NNHN characteristic is for remote protection, as described in [I-D.liu-rtgwg-path-aware-remote-protection].

### 2.4. NNHN Error Handling

The NNHN characteristic length MUST be at least 8 and MUST be a multiple of 4, otherwise it is malformed. Malformed NNHN characteristics MUST be discarded and SHOULD be logged.

If more than one instance of NNHN is included in an NHC, instances beyond the first MUST be discarded and SHOULD be logged.

## 3. Operational Considerations

Since BGP Identifiers are used to identify the next-next hop nodes, we need to make sure they are unique across the network where NNHN characteristic is sent.

## 4. IANA Considerations

NHC Characteristic Code 2, has been assigned in Section 5 of [I-D.ietf-idr-entropy-label], for the NNHN characteristic defined in this document.

## 5. Implementation Status

[The RFC-Editor should remove this section before publishing this document as an RFC]

Global load balancing using NNHN characteristic has been implemented by HPE Networking Juniper and New H3C Technologies. Interoperation tests between both implementations were successfully conducted in 2024.

## 6. Security Considerations

Insertion of a syntactically valid but bogus NNHN characteristic by an attacker could potentially make the forwarding behavior of the route non-optimal.

## 7. References

### 7.1. 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>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[I-D.ietf-idr-entropy-label] Decraene, B., Scudder, J., Kompella, K., Satya, M. R., Wen, B., Wang, K., and S. Krier, "BGP Next Hop Dependent Characteristics Attribute", Work in Progress, Internet-Draft, draft-ietf-idr-entropy-label-18, 20 July 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-entropy-label-18>.

## 7.2. Informative References

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[I-D.cheng-rtgwg-adaptive-routing-framework] Cheng, W., Lin, C., Wang, K., Ye, J., Zhuang, R., and P. Huo, "Adaptive Routing Framework", Work in Progress, Internet-Draft, draft-cheng-rtgwg-adaptive-routing-framework-04, 24 April 2025, <https://datatracker.ietf.org/doc/html/draft-cheng-rtgwg-adaptive-routing-framework-04>.

[I-D.liu-rtgwg-path-aware-remote-protection] Liu, Y., Lin, C., Chen, M., Zhang, Z., Wang, K., and Z. He, "Path-aware Remote Protection Framework", Work in Progress, Internet-Draft, draft-liu-rtgwg-path-aware-remote-protection-04, 4 August 2025, <https://datatracker.ietf.org/doc/html/draft-liu-rtgwg-path-aware-remote-protection-04>.

[I-D.zzhang-rtgwg-router-info] Zhang, Z. J., Wang, K., Lin, C., Vaidya, N., Tantsura, J., and Y. Liu, "Advertising Router Information", Work in

Progress, Internet-Draft, draft-zzhang-rtgwg-router-info-03, 22 April 2025, <https://datatracker.ietf.org/doc/html/draft-zzhang-rtgwg-router-info-03>.

## Appendix A. Alternative Solutions

An alternative way to carry next-next hops is via a separate path attribute. We evaluated both approaches and choose the NNHN characteristic approach for several reasons:

* Next-next hops depend on next hops, this makes it naturally fit into the existing NHC attribute.

* The next hop carried in the existing NHC attribute can help to validate that the next-next hop nodes are indeed for the next hop of the NLRI.

* Carrying next-next hop nodes via a seperate path attribute will cost an additional attribute code, which is supposed to be allocated for more generally used attributes.

## Acknowledgements

The author would like to thank John Scudder, Jie Dong, Robert Raszuk for their reviews, comments, and suggestions.

## Contributors

TBD.

## Authors' Addresses

Kevin Wang (editor)
HPE
Email: kfwang@juniper.net

Jeff Haas (editor)
HPE
Email: jhaas@juniper.net

Changwang Lin
New H3C Technologies
Email: linchangwang.04414@h3c.com

Jeff Tantsura
Nvidia
Email: jefftant.ietf@gmail.com