

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: 3 June 2022

I.V. Varlashkin
Google
R. Raszuk
NTT Network Innovations
K. Patel
Arrcus, Inc
M. Bhardwaj
S. Bayraktar
Cisco Systems
November 2021

Carrying next-hop cost information in BGP draft-ietf-idr-bgp-nh-cost-03

Abstract

BGP-LS provides a mechanism by which Link state and traffic engineering information can be collected from internal networks and shared with external network routers using BGP. BGP-LS defines a new Address Family to exchange this information using BGP.

BGP Optimal Route Reflection [BGP-ORR] [[RFC9107](#)] provides a mechanism for a centralized BGP Route Reflector to achieve requirements of a Hot Potato Routing as described in [Section 11 of \[RFC4456\]](#). Optimal Route Reflection requires BGP ORR to overwrite the default IGP location placement of the route reflector; which is used for determining cost to the nexthop contained in the path.

This draft augments BGP-LS and defines a new extensions to exchange cost information to next-hops for the purpose of calculating best path from a peer perspective rather than local BGP speaker own perspective.

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."

Internet-Draft

[draft-ietf-idr-bgp-nh-cost](#)

November 2021

This Internet-Draft will expire on 5 May 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the [Trust Legal Provisions](#) and are provided without warranty as described in the Revised BSD License.

Table of Contents

1.	Introduction	3
2.	NEXT-HOP INFORMATION BASE	3
3.	BGP Bestpath Selection Modification	4
4.	BGPLS Extensions	4
4.1.	RIB Metrics Prefix Descriptor	4
4.2.	RIB Protocol ID	5
4.3.	Information Exchange	5
4.4.	Termination of the session carrying next-hop cost	5
4.5.	Graceful Restart and Route-Refresh	5
5.	Security considerations	5
6.	IANA Considerations	6
7.	Acknowledgements	6
8.	References	6
8.1.	Normative References	6
8.2.	Informative References	7
Appendix A.	USAGE SCENARIOS	7
A.1.	Trivial case	7
A.2.	Non-IGP based cost	8
A.3.	Multiple route-reflectors	8
A.4.	Inter-AS MPLS VPN	9
A.5.	Corner case	9
	Authors' Addresses	9

1. Introduction

In a certain situation, route-reflector clients may not get optimum path to certain destinations. ADDPATH solves this problem by letting route-reflector to advertise multiple paths for a given prefix. If number of advertised paths are sufficiently big, route-reflector clients can choose same route as they would in case of full-mesh. This approach however places an additional burden on the control plane. Solutions proposed by [BGP-ORR] [[RFC9107](#)] use different approach - instead of calculating best path from the local speaker's own perspective the calculations are done using cost from the client to the next-hops. Although they eliminate need for transmitting redundant routing information between peers, there are scenarios where cost to the next-hop cannot be obtained accurately using these methods. For example, if next-hop information itself has been learned via BGP then simple SPF run on link-state database won't be sufficient to obtain cost information. There are also scenarios where while a Route Reflector can reach its clients, the client to client connectivity MAY be down.

BGPLS [[RFC7752](#)]. provides a mechanism by which Link state and traffic engineering information can be collected from internal networks and shared with external network routers using BGP. BGPLS defines a new Address Family to exchange this information using BGP.

To address such scenarios, this draft defines extensions to BGPLS to carry cost information of the next-hops. In particular, this draft defines a new Protocol ID to announce a Router's IGP routes, and a Prefix Descriptor to carry the cost information of the IGP routes used towards resolving next-hops.

2. NEXT-HOP INFORMATION BASE

To facilitate further description of the proposed solution we introduce a new table for all known next-hops and costs to it from various routers on the network.

Next-Hop Information Base (NHIB) stores cost to reach next-hop from an arbitrary router on the network. This information is essential for choosing best path from a peer perspective rather than BGP-speaker own perspective. In canonical form NHIB entry is triplet (router, next-hop, cost), however this specification does not impose any restriction on how BGP implementations store that information internally. The cost in NHIB is does not have to be an IGP cost, but all costs in NHIB MUST be comparable with each other.

NHIB can be populated from various sources including static routing and dynamic routing. However, this document focuses on populating NHIB using BGP.

An implementation implementing the BGP extension described in this draft MAY provide an operator-controlled configuration knob significant to an individual BGP speaker that treats next-hop cost information received from two or more clients as equivalent. For example a route-reflector could receive next-hop cost only from R1 but it will use it while calculating best-path also for R2, R3, Rn because it has been instructed to do so by locally-significant configuration. Multiple sources can be used for redundancy purpose.

[3.](#) BGP Bestpath Selection Modification

This section applies regardless of method used to populate NHIB.

When BGP speaker conforming to this specification selects routes to be advertised to a peer it SHOULD use cost information from NHIB rather than its own IGP cost to the next-hop after step (d) of 9.1.2.2 in [\[RFC4271\]](#).

[4.](#) BGPLS Extensions

[4.1.](#) RIB Metrics Prefix Descriptor

This draft defines a new Prefix Descriptor known as a Cost Prefix Descriptor with a TLV code point value to be assigned by IANA. The Cost descriptor looks like:

TLV Code Point	Description	Length	Value defined in:
TBD	Cost	4 bytes	Cost Value

Cost Value is a 4 byte Metric value computed by a Router's local RIB.

The Cost value is a cost associated with a prefix by a Router. The cost is typically computed by the routing protocols that owns a route.

[4.2.](#) RIB Protocol ID

This draft defines a new protocol ID for IPv4 and IPv6 Topology Prefix NLRI known as a RIB Protocol ID. The RIB Protocol ID has a value to be assigned by IANA. The Prefix NLRI with RIB Protocol ID is used to announce all the local and IGP computed routes that are installed in the RIB along with its Cost value.

[4.3.](#) Information Exchange

Typically BGPLS sessions will be established between route-reflectors and its internal peers (both clients and non-clients). As soon as the BGPLS session is ESTABLISHED, all the RIB routes used to resolve next-hop cost and information about next-hop costs MAY be sent immediately by clients to its route-reflector. Implementations are advised to announce BGP updates for this SAFI before any other SAFIs to facilitate faster convergence of other SAFIs on Route Reflectors.

Each internal neighbor of a route-reflector announces its IGP RIB Prefix information and its RIB metrics to the Route Reflector using a BGPLS session and a new NLRI Protocol ID and RIB metric Prefix Descriptor. Each neighbor updates Route Reflector with its IGP

prefix cost everytime a cost to an IGP route changes.

Upon a receipt of a BGP route and its associated cost, a Route Reflector stores the prefix, cost, and neighbor information in its local NHRIB database. It then uses the received cost towards calculation of bestpath from the respective clients perspective as opposed to its own IGP cost.

[4.4.](#) Termination of the session carrying next-hop cost

When the BGP session carrying next-hop cost terminates (for whatever reason), the BGP speaker SHOULD invalidate all the next-hop cost information (i.e same treatment that applies to the next-hop cost as to any other BGP learned information).

[4.5.](#) Graceful Restart and Route-Refresh

BGP sessions carrying next-hop cost could use Graceful Restart [[RFC4724](#)] and Route Refresh [[RFC7313](#)] mechanisms in the same way as it's used for IPv4 and IPv6 unicast.

[5.](#) Security considerations

This document does not introduce new security considerations above and beyond those already specified in [[RFC4271](#)], [[RFC9107](#)], [[RFC7752](#)].

[6.](#) IANA Considerations

This draft defines a new protocol id value for RIB Protocol ID. This draft requests IANA to allocate a value for a RIB Protocol ID from BGP Protocol ID Registry.

This draft defines a new RIB Metrics Prefix Descriptor value. This draft request IANA to allocate a TLV code value for the new descriptor from the Prefix Descriptor registry.

[7.](#) Acknowledgements

The authors would like to acknowledge David Ward, Anton Elita, Nagendra Kumar and Burjiz Pithawala for their critical reviews and feedback.

[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>>.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, [RFC 2328](#), DOI 10.17487/RFC2328, April 1998, <<https://www.rfc-editor.org/info/rfc2328>>.
- [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>>.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", [RFC 4456](#), DOI 10.17487/RFC4456, April 2006, <<https://www.rfc-editor.org/info/rfc4456>>.
- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", [RFC 4724](#), DOI 10.17487/RFC4724, January 2007, <<https://www.rfc-editor.org/info/rfc4724>>.
- [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>>.

- [RFC7313] Patel, K., Chen, E., and B. Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", [RFC 7313](#), DOI 10.17487/RFC7313, July 2014, <<https://www.rfc-editor.org/info/rfc7313>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", [RFC 7752](#),

DOI 10.17487/RFC7752, March 2016,
<<https://www.rfc-editor.org/info/rfc7752>>.

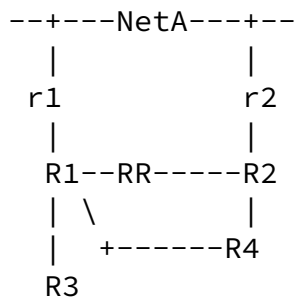
- [RFC9107] Raszuk, R., Ed., Decraene, B., Ed., Cassar, C., Åman, E., and K. Wang, "BGP Optimal Route Reflection (BGP ORR)", [RFC 9107](#), DOI 10.17487/RFC9107, August 2021, <<https://www.rfc-editor.org/info/rfc9107>>.

[8.2.](#) Informative References

- [RFC2918] Chen, E., "Route Refresh Capability for BGP-4", [RFC 2918](#), DOI 10.17487/RFC2918, September 2000, <<https://www.rfc-editor.org/info/rfc2918>>.

[Appendix A.](#) USAGE SCENARIOS

[A.1.](#) Trivial case



In this scenario r1 and r3 along with NetA are part of AS1; and R1-R4 along with RR are in AS2.

If RR implements non-optimized route-reflection, then it will choose path to NetA via R1 and advertise it to both R3 and R4. Such choice is good from R3 perspective, but it results in suboptimal traffic flow from R4 to NetA.

Using the proposed BGPLS extensions, the route-reflector will learn

that cost from R4 to R1 is 8 whereas to R2 it's only 1. RR will announce NetA to R4 with next-hop set to R2, while its announce to R3 will still have R1 as next-hop. Both R3 and R4 now will send traffic to NetA via closest exit, achieving same behaviour as if full iBGP mesh would have been configured.

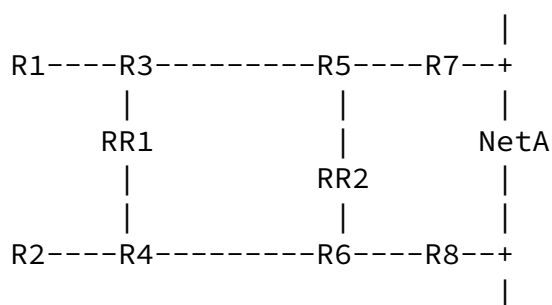
A.2. Non-IGP based cost

When it's desirable to direct traffic over an exit other than the one with smallest IGP cost, BGP extensions can be used to convey cost which is not based on IGP. For example, network operator may arrange exit points in order of administrative preference and configure routers to send this instead of IGP cost. Route reflector then will then calculate best path based on administrative preference rather than IGP metrics.

Network operators should exercise care to ensure that all routers up to and including exit point do not divert packets on to a different path, otherwise routing loops may occur. One way to achieve this is to have consistent administrative preference among all routers. Another option is to use a tunneling mechanism (e.g. MPLS-TE tunnel) between source and the exit point, provided that the router serving as exit point will send packets out of the network rather than diverting them to another exit point.

A.3. Multiple route-reflectors

This example demonstrates that BGP/LS extensions are necessary only between routers that already exchange other AFI/SAFI.



In the above network the routers R1-R4 are clients of RR1, and R5-R8 are clients of RR2. RR1 and RR2 also peer with each other and use ADDPATH.

RR2 learns about NetA from R7 and R8. Since it sends not just best-path but all prefixes to RR1, there is no need for RR2 to learn cost information from R1 and R2 towards R7 and R8. On the other hand RR1 does exchange cost information using BGPLS with R1 and R2 so that each of them can receive routes, which are best from their perspective.

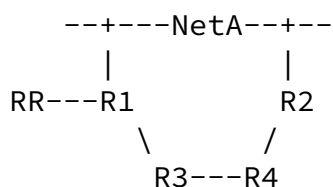
As addition to ADDPATH a mechanism could be devised that would allow RR2 to learn how many alternative routes does it need to send to RR1. For example, if NetA would also be connected to R9 (not shown) but all clients of RR1 prefer R7 as exit point and R9 as next-best, then there is no need for RR2 to send NetA routes with next-hop R8 to RR1.

Discussion: authors would like to solicit discussion whether there is sufficient interest in such mechanism.

[A.4.](#) Inter-AS MPLS VPN

Previous example could be transposed to Inter-AS MPLS VPN Option C scenario. In this case route reflectors RR1 and RR2 can be from different autonomous system. Essentially the behaviour of routers remains as already described.

[A.5.](#) Corner case



In the above network cost from R3 to R1 is 10, all other costs are 1. If RR advertises NetA to R3 based on cost information received from R3, but uses its own cost when advertising NetA to R4, there will be a loop formed. This is the reason why section "BGP best path selection modification" requires RR to have next-hop cost information for every next-hop and every peer.

Note that the problem is the same as if RR would not use extensions described in this document and R3 would peer directly with R1 and R2, while R4 would peer only with RR.

Authors' Addresses

Ilya Varlashkin
Google

Email: ilya@nobulus.com

Varlashkin, et al.

Expires 3 June 2022

[Page 9]

Internet-Draft

[draft-ietf-idr-bgp-nh-cost](#)

November 2021

Robert Raszuk
NTT Network Innovations
940 Stewart Dr
Sunnyvale, CA 94085
United States of America

Email: robert@raszuk.net

Keyur Patel
Arrcus, Inc
2077 Gateway Pl
San Jose, CA 95110, 95110
United States of America

Email: keyur@arrcus.com

Manish Bhardwaj
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95124, 95134
United States of America

Email: manbhard@cisco.com

Serpil Bayraktar
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95124, 95134
United States of America

Email: serpil@cisco.com

