rtgwg
Internet-Draft
Intended status: Standards Track
Expires: 23 April 2026

Z. Zhang
K. Wang
Juniper Networks
C. Lin
New H3C Technologies
N. Vaidya
Broadcom
J. Tantsura
Nvidia
Y. Liu
China Mobile
20 October 2025

Advertising Router Information
draft-zzhang-rtgwg-router-info-05

Abstract

This document specifies a generic mechanism for a router to advertise some information to its neighbors. One use case of this mechanism is to advertise link/path information so that a receiving router can better react to network changes.

Table of Contents

1. Introduction

[I-D.wang-idr-next-next-hop-nodes] describes a scenario where better load-balancing can be achieved in a CLOS network by considering the load information on the next hop router in addition to considering the local load information of the path to that next hop router.

[I-D.liu-rtgwg-path-aware-remote-protection] describes another scenario where link up/down information propagated via non-IGP mechanism can help with faster reroute.

[I-D.cheng-rtgwg-adaptive-routing-framework] describes a framework for Adaptive Routing which dynamically adjusts routing paths based on changes in global network conditions, thereby optimizing network performance and resource utilization.

To achieve that, a router needs to advertise some link/path information independently of IGP, often at very fast pace. [I-D.dong-fantel-problem-statement] describes the need for fast notification related solutions to support any high- throughput, low-latency and lossless application, and this document specifies a mechanism to do that, which can also be used to advertise any information.

As described in [I-D.wang-idr-next-next-hop-nodes], in a CLOS network the advertisement only needs to be "link-local", i.e., a receiving router does not need to re-advertise it further and the mechanism in this document does not consider re-advertisement. In an arbitrary topology, to achieve coordinated load-balancing the information may need to be advertised further but that is outside the scope of this document.

In some scenarios, a large amount of information may need to be advertised, and in some scenarios, the receiving router may not need to be directly connected.

While an IGP, if used for the CLOS network, may also be used to advertise the information using link-local flooding scope, it may not be a good fit when the information needs to be advertised and processed very rapidly not for routing purposes.

Therefore, UDP is chosen as the transport mechanism. An implementation may advertise and process the UDP messages in the forwarding path for timely responses.

This document does not suggest or restrict when and/or how frequently the information is advertised - it is an operational consideration on how frequent the advertisements need to be and whether the routers can handle that.

Fragmentation may be used if the delay related to the fragmentation/ reassembly is not a concern. Multiple UDP messages may be used to advertise pieces of all the information, whether fragmentation is used or not.

This document/revision only specifies the message format. How the information is maintained and used on the receiving router are outside the scope of this document/revision but may be added in future revisions.

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. Specification

The message format is defined as follows:

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |        UDP source port        |  UDP destination port TBD1    |
   +-------------------------------+-------------------------------+
   |           UDP length          |         UDP Checksum          |
   +-----------------------------+-+-------------------------------+
   |Version|R|R|R|R|    Flags    |A|
   +---------------+-------------+-+-------------------------------+
   |   Auth Type   |   Auth Len    |          Auth Data ...        ~
   +---------------+---------------+-------------------------------+
   |                       Source Router ID                        |
   +---------------------------------------------------------------+
   |                            Link ID                            |
   +---------------------------------------------------------------+
   |                        (Repeated) TLVs                        ~
   ~                                                               |
   +---------------------------------------------------------------+
```

The IP destination address in the outer IP header is typically an IPv4 "All Routers on this Subnet" multicast address (referred to as a link-local multicast address), an IPv6 Node-local All Routers Address (multicast) [RFC4291], a non-link/node-local multicast address, or a local/remote unicast address discovered by means outside the scope of this document.

The 4-bit Version field is for potential future extensions that cannot be achieved through additional TLV types. The current version is 0.

The four R-bits are reserved - they MUST be 0 upon transmission and MUST be ignored upon receiving.

The 1-octet Flags field currently has one A-flag defined. If it is set, the (Auth Type, Auth Len, Auth Data) tuple immediately follows the Flags field. If it is not set, the tuple is not present. The details of the tuple are as specified in [RFC5880] and not repeated here.

When the flooding happens on a local link, the Link ID field identified the flooding link. The value could be one of the following:

* An IPv4 interface address advertised by OSPFv2/ISIS [RFC2328] [RFC1195]

* An Interface ID advertised by OSPFv3 [RFC5340], or by OSPFv2 for an unnumbered interface

* A Link Local Identifier advertised by OSPFv2/ISIS for GMPLS [RFC4203] [RFC5307]

In this case, the destination address MUST be a link/node-local multicast address or a receiver’s unicast address on the local link and the TTL MUST be 1. The source address MUST be the local interface address for a numbered interface, or a loopback address in case of an unnumbered interface.

If the flooding is to one or more remote receivers, the Link ID MUST be set to 0, the destination address MUST be a remote unicast/multicast address, the source address MUST be a loopback address, and the TTL MUST be set to 255, following the Generalized TTL Security Mechanism (GTSM) [RFC5082].

The following TLVs are defined to allow maximum packing. If additional information needs to be advertised, new TLVs may be defined without using sub-TLVs to allow efficient encoding of additional information, or with sub-TLVs to allow flexibility but at the cost of processing complexity.

2.1. Neighbor Path Information

This TLV is used when the path information is at per-neighbor level.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   (Type) 1    |               Length          |S| Refresh Rate|
+-------------------------------+-------------------------------+
~                  repeated per-Neighbor Records                |
+---------------------------------------------------------------+
```

The Length field is two-octet. The Value part starts with a one-octet Refresh Rate field, which is followed by repeated per-Neighbor Records. The number of records is derived from the Length field.

The Refresh Rate field’s leftmost bit S denotes the unit of the rate. If it is set, the rate is in deciseconds (100ms). If it is not set, the rate is in milliseconds. If the refresh rate is 0, it means that the information will not be refreshed. This can be used for one-time notifications when the consequence of loss is tolerable.

The per-Neighbor record has the following format:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Neighbor ID                                  |
+---------------+-----------------------------------------------+
| encoded info  |
+---------------+
```

Neighbor ID: The 4-octet Neighbor ID identifies (the path to) a neighbor that is discovered by means outside the scope of this document. It MAY be a BGP-ID described in [I-D.wang-idr-next-next-hop-nodes] or some other identifiers that are unique in the domain where the signaling is used. The neighbor can be reached via ECMP, e.g., a set of links but the nature is outside the scope of this document.

encoded info: The 1-octet field following the 4-octet Neighbor ID field which encodes the information of the path to the neighbor.The following encoded info are defined:

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|R|R|R|U|Quality|
+---------------+
```

where:

U-Flag: State Flag. If it is set, the path to the neighbor is UP. If it is not set, the path to the neighbor is DOWN.

The three R-bits: Reserved. They MUST be 0 upon transmission and MUST be ignored upon receiving.

Quality Level: The 4-bit Quality field is used for path quality. The value range is from 0 to 15.
The quality level can be customized, with the lower the level, the poorer the path quality. The quality level can be calculated based on the current bandwidth and the utilization of the forwarding buffer. Bandwidth and buffer use a certain ratio to calculate the quality level. The exact method for calculating the quality level is beyond the scope of this document, but must ensure that the calculation rules are consistent among the routers the information is flooded to/from.

For instance, a 400G interface can be divided into sixteen quality levels based on bandwidth utilization, with each level representing 25G of bandwidth usage. When the Quality level is 0, the available bandwidth is up to 25G. When the Quality Level is 15, the available bandwidth is up to 400G.

2.2. Link Information

This TLV is used when the information is at per-link level.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   (Type) 3    |                  Length       |S| Refresh Rate|
+---------------+-------------------------------+-+-------------+
|              repeated per-link Records                        |
+---------------------------------------------------------------+
```

The value part is repeated records of the following. The number of records is derived from the Length field.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Link ID                            |
+---------------+-----------------------------------------------+
| encoded info  |
+---------------+
```

The Link ID is as described earlier.

encoded info: The 1-octet field following the 4-octet Link ID field which encodes the information of the link. It is encoded exactly the same as in the neighbor case.

2.3. Refreshing and Aging

The sender MUST re-advertise the information when there is a change, and MUST re-fresh previous advertisements at the advertised Refresh Rate even when there is no change.

The receiver MUST age out the received information if it does not receive a refresh within a period of three times of the refresh rate. Each time an advertisement for a neighbor is received, the aging timer is reset according to the latest Refresh Rate.

The sender MAY adjust the Refresh Rate on its own or based on notification from a receiver (Section 2.4). For example, if the information does not change often, a sender may move to a larger (slower) Refresh Rate.

The aging, refreshing and adjustment of the refresh rate are all at the per-neighbor/link level. Neighbors/links whose Refresh Rates are the same SHOULD be packed in the same TLV, but MAY be put into different TLVs and messages due to MTU limitations and/or fragmentation concerns. Neighbors/links whose Refresh Rates are different MUST be put into different TLVs.

The same information MAY be sent out of different links or to different set of receivers with different rates.

2.4. Refresh Rate Negotiation

A receiver SHOULD send notifications to the sender if it can not keep up with the sender, using a Notification TLV of Type 4:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   (Type) 4    |             Length            |     Value     ~
+---------------+-------------------------------+---------------+
~                        (continued) Value                      |
+---------------------------------------------------------------+
```

The Value part includes sub-TLVs, whose Types share the same space as the top level TLVs.

When sending a notification to a remote node or from an unnumbered interface, a loopback address MUST be used as the source address. Otherwise, the local interface address SHOULD be used as the source address. The destination address MUST be set to the source address in the received flooding packet for which the notification is.

To notify the sender the desired Refresh Rate for a certain advertisement, the corresponding TLV (e.g., the Neighbor Path Information TLV) is included as a sub-TLV, and no per-Neighbor/Link records are included. The Refresh Rate field along with the S-flag are set to indicate the desired rate. The Length of the sub-TLV is set accordingly. Other types of TLVs, e.g., this type-4 "Refresh Rate Notification" TLV itself, MUST NOT be included as sub-TLVs.

While a typical physical link is point-to-point even in the Ethernet case, there may be multiple receivers of an advertisement sent out of a link (e.g., in the case of LAN) or sent to a group of remote receivers via multicast. If multiple notifications of Refresh Rate are received for an advertisement, the largest requested rate MUST be used by the sender.

Consider that a router advertises to a link the information about some neighbor/link set S1 at rate R1 and the information about some other neighbor/link set S2 at rate R2 where R1 < R2, i.e., the S2 information is advertised less frequently. A receiver on the link sends back a notification with rate R3 where R1 < R3 < R2. Then this router MUST use R3 as the refresh rate for S1 and R2 as the refresh rate for S2.

2.5. Flow Redirection

It may be desired for a router to request its upstream to redirect a specific flow away from it. This is done via Flow Redirection TLV:

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |   (Type) 6/7  |              Length           |S| Refresh Rate|
    +-------------------------------+-------------------------------+
    |    Protocol   |
    +-------------------------------+-------------------------------+
    |           Source Address (4 or 16 octets)                     ~
    +-------------------------------+-------------------------------+
    |           Destination Address (4 or 16 octets)                ~
    +-------------------------------+-------------------------------+
    |          Source Port          |      Destination Port         |
    +-------------------------------+-------------------------------+
    |                Repeated 5-Tuple Information                   ~
    +-------------------------------+-------------------------------+
```

The Type is 6 for IPv4 flows or 7 for IPv6 flows. The Value field encodes one or more 5-tuple records that identify flows by (Protocol, Source Address, Destination Address, Source Port, Destination Port). The number of records is derived from the Type and Length fields.

3. Security Considerations

Because the information may be flooded rapidly, potential Denial Of Service (DOS) attack may happen. In the case of local flooding only, the attack needs to happen from within the network, and in that case, other attacks may happen as well and may be worse. In the case of remote flooding, GTSM [RFC5082] and ingress filtering at the network boundary are used to reduce the risk. Overall, this is expected to be used in a secure walled-garden network.

4. IANA Considerations

This document requests IANA to allocate a UDP port number TBD1 from the User Ports range of the Service Name and Transport Protocol Port Number Registry.

This document requests IANA to create a Router Information TLV Type registry with the following initial allocations:

Value	Description	Reference
0	Reserved	This Document
1	Neighbor Path Information	This Document
3	Link Information	This Document
4	Notification	This Document
6	IPv4 Flow Redirection	This Document
7	IPv6 Flow Redirection	This Document

The registration procedure is Standards Action for value range 0-127 and First Come First Served for range 128-255.

5. Acknowledgments

The authors thank Jeffrey Haas and Michal Styszynski for their review, comments and suggestions to make this document and solution more complete.

6. References

6.1. 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>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <https://www.rfc-editor.org/info/rfc4291>.

[RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., Ed., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, DOI 10.17487/RFC5082, October 2007, <https://www.rfc-editor.org/info/rfc5082>.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <https://www.rfc-editor.org/info/rfc5880>.

[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>.

6.2. Informative References

[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.dong-fantel-problem-statement] (editor), J. D., (editor), M. M., (editor), F. C., Zhang, J., Zhu, Y., Xu, X., Pang, R., Lu, H., Liu, Y., Contreras,

L. M., and M. Durmus, "Fast Notification Problem Statement", Work in Progress, Internet-Draft, draft-dong-fantel-problem-statement-00, 20 October 2025, <https://datatracker.ietf.org/doc/html/draft-dong-fantel-problem-statement-00>.

[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.wang-idr-next-next-hop-nodes]
Wang, K., Haas, J., Lin, C., and J. Tantsura, "BGP Next-next Hop Nodes", Work in Progress, Internet-Draft, draft-wang-idr-next-next-hop-nodes-04, 26 August 2025, <https://datatracker.ietf.org/doc/html/draft-wang-idr-next-next-hop-nodes-04>.

[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>.

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

[RFC4203] Kompella, K., Ed. and Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, DOI 10.17487/RFC4203, October 2005, <https://www.rfc-editor.org/info/rfc4203>.

[RFC5307] Kompella, K., Ed. and Y. Rekhter, Ed., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008, <https://www.rfc-editor.org/info/rfc5307>.

[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>.

Authors’ Addresses

Zhaohui Zhang
Juniper Networks
Email: zzhang@juniper.net

Kevin F. Wang
Juniper Networks
Email: kfwang@juniper.net

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

Niranjan Vaidya
Broadcom
Email: niranjan.vaidya@broadcom.com

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

Yisong Liu
China Mobile
Email: liuyisong@chinamobile.com