Network Working Group
Internet Draft
Intended status: Informational
Expires: February 05, 2026

Y. Liu
China Mobile
C. Lin
M. Chen
New H3C Technologies
Z. Zhang
ZTE Corporation
K. Wang
Juniper Network
Z. He
Broadcom

August 04, 2025

Path-aware Remote Protection Framework
draft-liu-rtgwg-path-aware-remote-protection-04

Abstract

This document describes the framework of path-aware remote protection.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on February 05, 2026.

Table of Contents

1. Introduction

Current IP network protection mechanisms can be mainly divided into local protection and end-to-end protection. Local protection technologies, such as ECMP, LFA [RFC5714], and TI-LFA [I-D.ietf-rtgwg-segment-routing-ti-lfa], can only perceive local failures and perform fast reroute. End-to-end protection technologies are usually targeted at end-to-end TE paths, where the head-end detects TE path failures and performs rapid switchover.

There is no mechanism to quickly detect remote failures and invoke repairs for non-TE paths. In addition, local protection such as TI-LFA technology relies on IGP deployment. For certain networks, current protection mechanisms may not meet the requirements. A typical scenario is the Spine-Leaf network, such as the AI-DC network, which is usually a two-layer architecture. Detecting remote failures and invoking fast repairs can provide protection against link or node failure and reduce the disruption time.

This paper proposes a path-aware remote protection mechanism and describes its framework.

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. Use Case

2.1. Spine-leaf Network

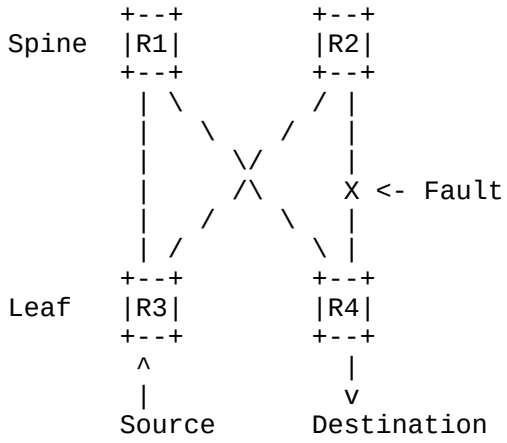


Figure 1

In the network shown in Figure 1, assuming that the R2-R4 link fails, R3 will continue to send traffic to both R1 and R2, and half of the traffic will be dropped by R2. It is not until R2 sends BGP withdrawn routes to R3 and the control plane converges that the traffic is fully restored. The convergence speed would be slow when there is a large number of BGP routes.

In some Spine-leaf networks, such as DC networks, only the BGP protocol is deployed without IGP, and thus TI-LFA cannot be applied. On the other hand, if TI-LFA is used, the traffic path during the protection period will be R3->R2->R3->R1->R4, which additionally increases the traffic in the direction of R2->R3 and may cause congestion.

The objective of path-aware remote protection is for R3 to detect R2-R4 link failure and then adjust ECMP quickly.

2.2. Dragonfly Network

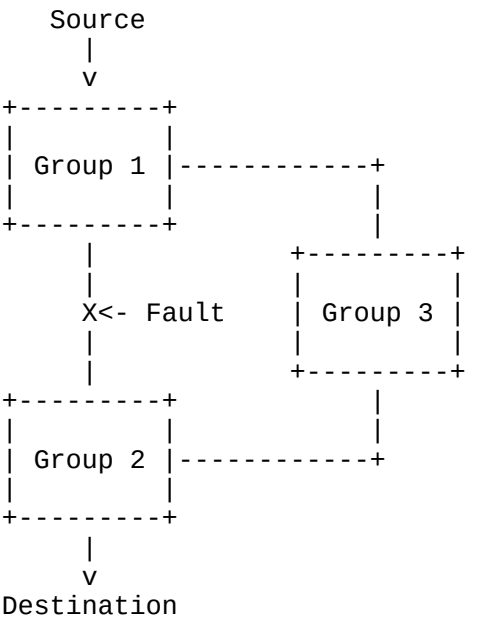


Figure 2

In the network shown in Figure 1, the primary path for the traffic is from Group 1 to Group 2, while the backup path detours from Group1 through Group3 and then to Group2.

The objective of path-aware remote protection is for the routers in Group 1 to detect the link failure between Group 1 and Group 2 and then switch to the backup path quickly.

3. Framework

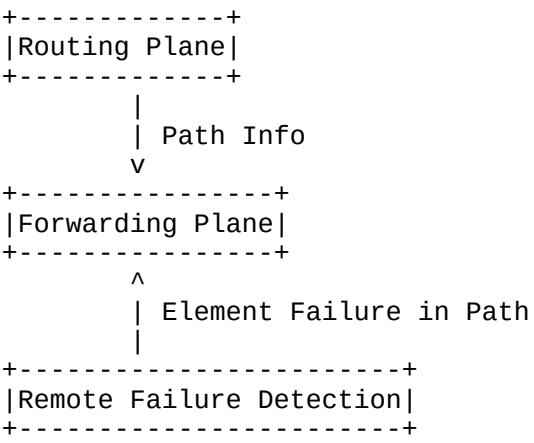


Figure 3

The framework of path-aware remote protection is shown in Figure 3.

On the routing plane, the route calculation is not limited to the next hop, but requires path awareness. And then the path information is downloaded to the forwarding plane. When a failure occurs in any component along the path, it is required to quickly detect the failure and invoke repairs.

3.1. Remote Failure Detection

When a failure occurs, it is first detected by the router adjacent to it. The local failure detection may be based on existing techniques such as BFD. Then, that router notifies its neighbors of the failure, especially the upstream neighbors. After the remote repairing router receives the failure notification, the remote protection is invoked.

The failure notification between neighboring routers has the following requirements:

o Independent of routing protocols.

o Avoiding broadcast flooding.

For one example, in a two-level spine-leaf network, a spine router can use BFD to monitor the adjacent links. When a link fails, the spine router can use a BGP-independent protocol to notify

neighboring leaf routers. The failure notification is limited in one hop.

For another example, a flow-based mechanism can be used to detect failure. When the traffic packets are dropped, a notification is triggered and sent to neighbors in the direction of the incoming traffic. The failure notification is limited in the upstream direction.

The design of the failure notification protocol may consider different rates for fault and normal conditions. In normal conditions, the status of path information may be refreshed at a low rate. When a fault occurs, the notification would be repeated at a high rate. In addition, acknowledgments by receivers may be used in the fault condition to improve reliability and efficiency.

The detailed mechanisms are out of the scope of this document.

3.2. Path-Aware Forwarding Plane

In the forwarding table, each next-hop is associated with a path. When detecting any failure in the path, the protection for the corresponding next-hop will be invoked.

Figure 4 shows the forwarding entries for ECMP next-hops.

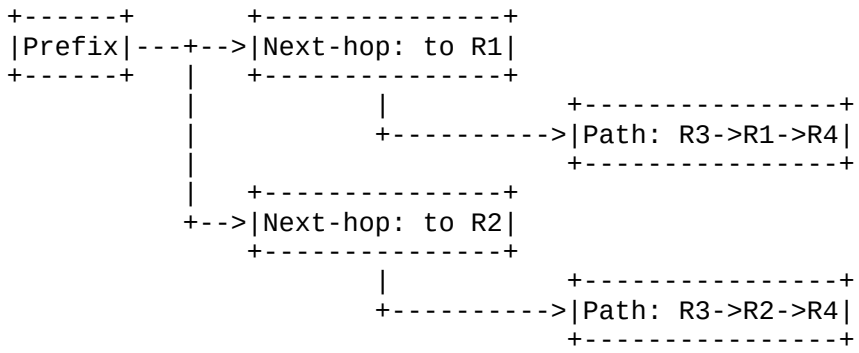


Figure 4

Figure 5 shows the forwarding entries for primary and backup next-hops.

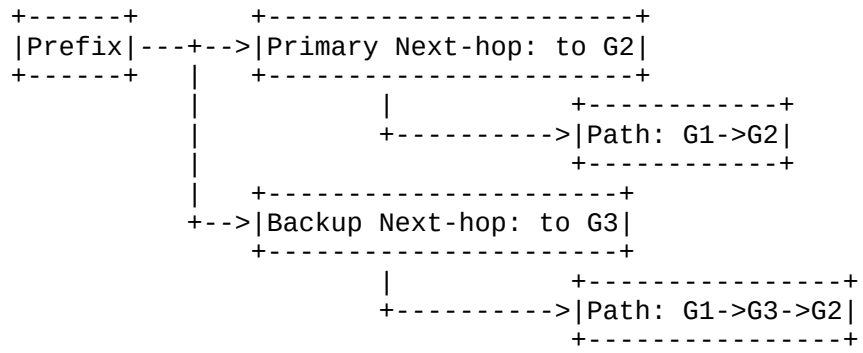


Figure 5

When receiving failure notification from a neighbor, the next-hop entries corresponding to that neighbor will be checked to determine whether the associated path information contains the failed component. If detecting any failure in the path, the corresponding next-hop is regarded as failed. For a failed ECMP next-hop, it will be removed from the ECMP, and the traffic will be switched to the other ECMP next-hops. For a failed primary next-hop, the traffic will be switched to the backup next-hop.

3.3. Path-Aware Routing Plane

When calculating routes, the path needs to be perceived and the path information will be attached to the next hop.

In a BGP-based network, a BGP route may carry the router-id of the peer from which that route is received, and the router-id will be added into the path information when calculating that route. The BGP protocol may needs some extensions to support such feature.

For an EBGP-based DC network, a router may use the AS-PATH attribute (with SEQUENCE type) in the BGP route as the path information, without any protocol extensions.

In an IGP-based network, a router may compute the path information based on the SPF tree and attach it into the next hop.

The detailed mechanisms are out of the scope of this document.

4. Role Types

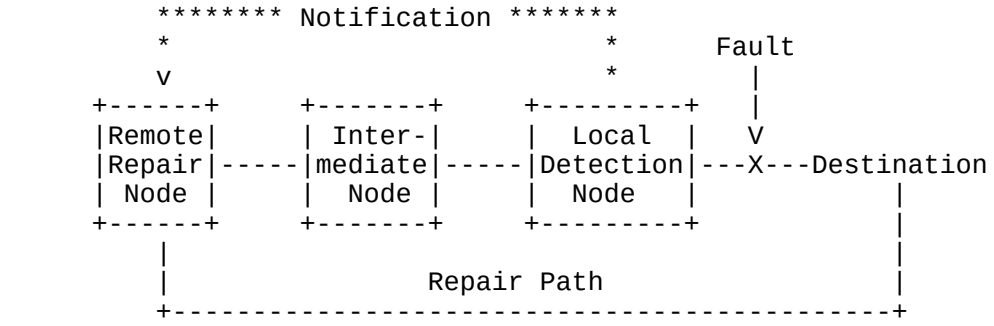


Figure 6

In the path-aware remote protection, there are three types of roles for a router:

- Remote repair node: It has the repair path(s) and provides the remote protection function.

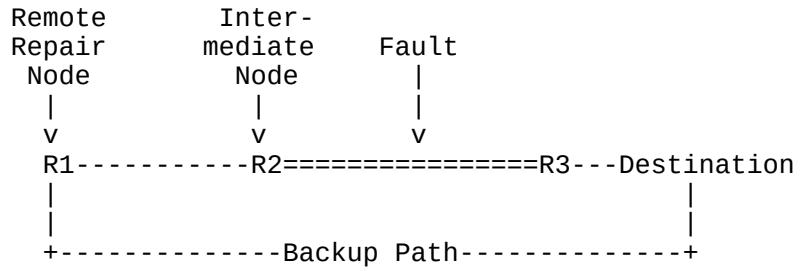
- Local detection node: It is adjacent to the failure and detects the failure first. Then, it sends failure notification messages to the remote repair node.

- Intermediate node: It exists only if there are multiple hops between the remote repair node and the local detection node. It helps deliver the failure notification messages from the local detection node to the remote repair node.

5. Path Information

Path information is at either per-neighbor level or per-link level. The level of path information in routing plane, forwarding plane and failure notifications needs to be consistent accordingly.

Take the following network as an example. Assume that the fault occurs on the links between R2 and R3, or on the node R3. R1 is the remote repair node, and R2 is the local detection node.



5.1. Per-neighbor Level

In the routing distribution and calculation, the neighbor ID of R3 is included to indicate the path from R2 to R3. The router ID may be used as the neighbor ID.

When both two links between R2 and R3 fail (or the node R3 fails), R2 will notify R1 of the path failure with R3’s neighbor ID.

The following figure shows an example of path information at per-neighbor level:

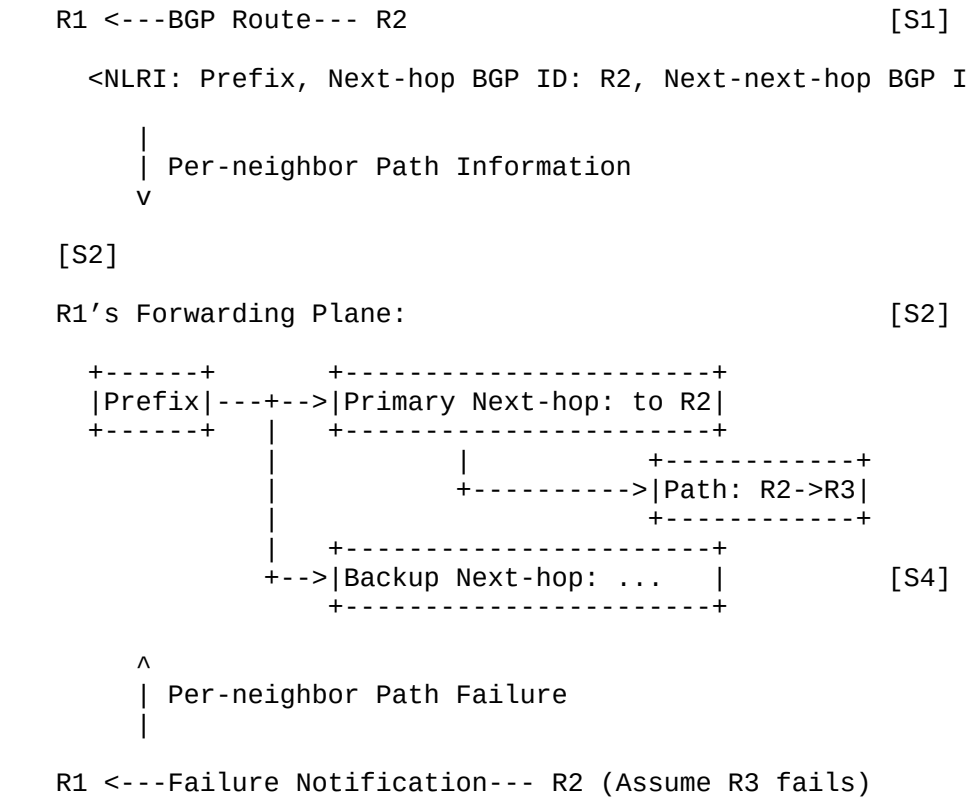


Figure 7

[S1]: When R2 delivers BGP routes to R1, the NNHN Capability TLV is carried in the attributes [I-D.wang-idr-next-next-hop-nodes], indicating that the next-hop is R2 and the next-next-hop is R3.

[S2]: R1 receives the BGP routes containing per-neighbor path information, performs the routing calculation, and installs the path-aware forwarding entries.

[S3]: Assume that R3 fails (or both the links between R2 and R3 fail), R2 sends the failure notification to R1, indicating its neighbor R3 fails.

[S4]: R1 receives the failure notification, it checks the next-hop entries corresponding to R2 and finds the associated path information contains the failed neighbor R3. Then, R1 invokes switchover to the backup next-hop.

5.2. Per-link Level

In the routing distribution and calculation, the link IDs of both the two links between R2 and R3 are included to indicate the path from R2 to R3. The interface identifier may be used as the link ID.

When either of the two links between R2 and R3 fail, R2 will notify R1 of the path failure with the ID of the failed link.

The following figure shows an example of path information at per-link level:

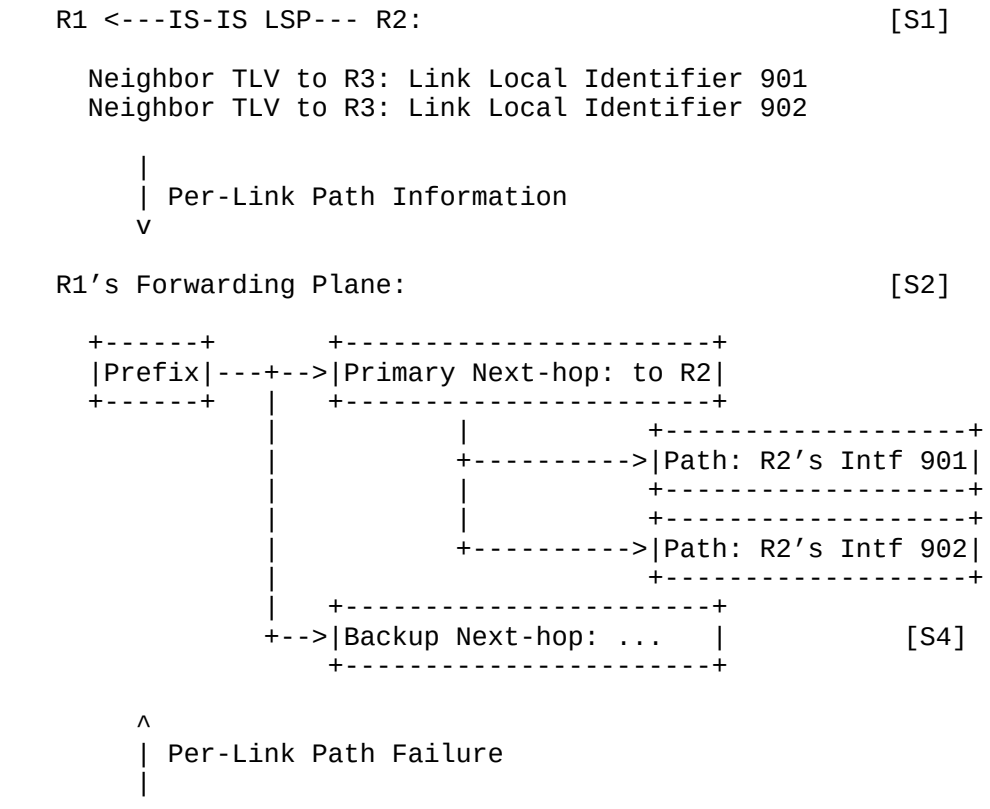


Figure 8

[S1]: When R2 generates IS-IS LSP, the link local identifiers (interface ID 901 and 902) of the links to R3 is carried in the neighbor TLV.

[S2]: R1 receives the IGP routes containing per-link path information, performs the routing calculation, and installs the path-aware forwarding entries.

[S3]: Assume that one link between R2 and R3 fails, R2 sends the failure notification to R1, indicating its interface 901 fails.

[S4]: R1 receives the failure notification, it checks the next-hop entries corresponding to R2 and finds the associated path information contains the failed link 901. Note that, the traffics

can still be transmitted over the non-failed link, so R1 may choose not to invoke switchover until both two links on the path fail.

6. Protection Scope

The scope of remote protection covers at least two hops from the remote repair node to the failure.

As the protection scope increases, the number of intermediate nodes increases, which may slower the speed and wider the propagation of fault notification. So, it would bring benefits to limit the scope of remote protection to a reasonable range.

One recommendation is that, the node closest to the failure and with a repair path should provide the protection function.

For example, in a spine-leaf network with multiple levels, usually there are ECMP paths on every two levels. Remote protection only needs to cover two hops.

7. Security Considerations

TBD.

8. IANA Considerations

TBD.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, May 2017

9.2. Informational References

[RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, DOI 10.17487/RFC5714, January 2010, <https://www.rfc-editor.org/info/rfc5714>.

[I-D.ietf-rtgwg-segment-routing-ti-lfa] Litkowski, S., Bashandy, A., Filsfils, C., Francois, P., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", draft-ietf-rtgwg-segment-routing-ti-lfa-21 (work in progress), February 2025.

[I-D.wang-idr-next-next-hop-nodes] Wang, K. and J. Haas, "BGP Next-next Hop Nodes", Work in Progress, Internet-Draft, draft-wang-idr-next-next-hop-nodes-03, 18 June 2025, <https://datatracker.ietf.org/doc/html/draft-wang-idr-next-next-hop-nodes-03>.

Authors' Addresses

Yisong Liu
China Mobile
China
Email: liuyisong@chinamobile.com

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

Mengxiao Chen
New H3C Technologies
China
Email: chen.mengxiao@h3c.com

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

Kevin Wang
Juniper Networks
10 Technology Park Dr
Westford, MA 01886
United States of America
Email: kfwang@juniper.net

Zongying He
Broadcom
Email: Zongying.he@broadcom.com