Path-aware Remote Protection Framework

draft-liu-rtgwg-path-aware-remote-protection-02

Presenter: Changwang Lin (H3C Technologies)

Co-authors: Yisong Liu (China Mobile)

Changwang Lin (H3C Technologies)

Zheng Zhang (ZTE Corporation)

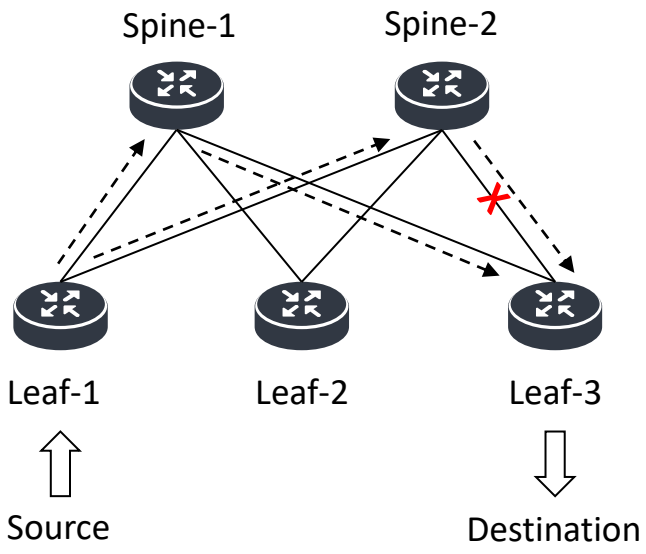
Kevin Wang (Juniper Network)

Zongyin He(Broadcom)

IETF-121, November 2024

Motivation

- Current IP network protection mechanisms can be mainly divided into local protection and end-to-end protection.
  - Local protection technologies, such as ECMP, LFA, and TI-LFA, can only perceive local failures and requires IGP.
  - End-to-end protection technologies are used for end-to-end TE paths. The head-end performs detection and switchover.
- There are some networks where current protection mechanisms cannot cover.



In a spine-leaf based DC network:

- Only BGP protocol is deployed, no IGP.
- IP-based BE forwarding, no TE.

When the failure occurs:

- Leaf-1 will continue to send traffic to both Spine-1 and Spine-2, until Leaf-1 receives BGP withdrawn routes from Spine-2.
- Waiting for control plane convergence would be quite long when there is a large number of BGP routes.

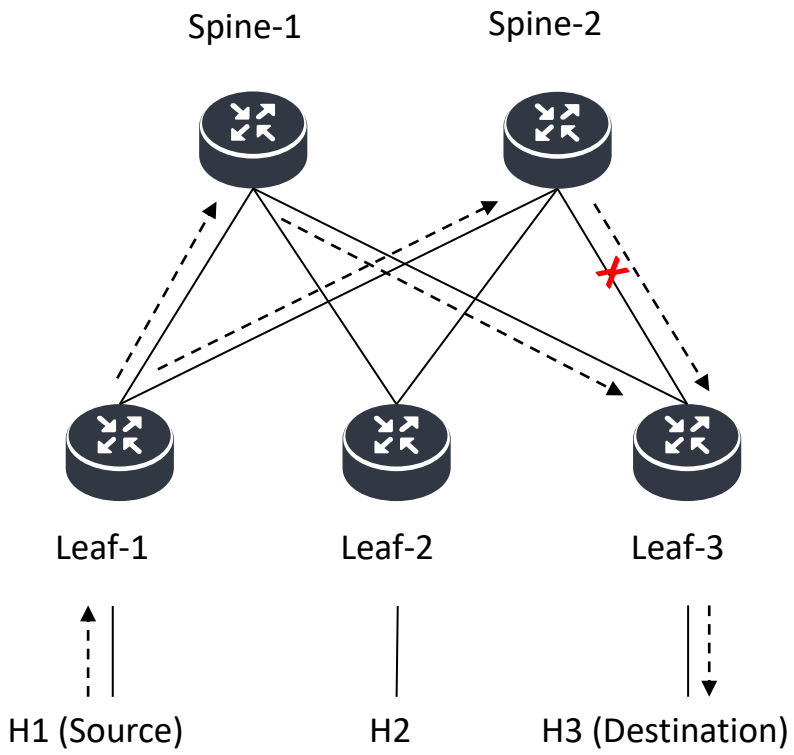
The idea is to allow Leaf-1 to detect the remote failure on the link between Spine-2 and Leaf-3, and then to invoke fast repairs!

Existing Mechanism for Route Convergence

Detection mechanism	Technology	Convergence Time	Influencing factors
Fast failure detection Detection	BFD	a few milliseconds	/
	CFM	milliseconds to seconds	/
Local Fast Failover	ECMP	a few milliseconds	convergence time primarily depends on the fault detection time.
	FRR	a few milliseconds	
Failure notification	IGP LinkState propagation	milliseconds to seconds	The notification time depends on the network size and the number of routes.
	BGP route updates	milliseconds to seconds	
Global Fast Failover	BGP PIC	milliseconds to seconds	The convergence time depends on both the fault notification time and the network size.
	IGP route calculation convergence	several hundred milliseconds to a few seconds	

- The convergence time for **local failure** handling includes local detection time and local fast switching time, usually taking tens of milliseconds.
- The convergence time for **remote failure** handling includes local detection time, fault notification time, and global fast switching time, typically taking a few seconds.

How Path-aware Remote Protection Work?



① Routing Control Protocol：

1. Routing protocol extension generates remote next-hop path information.
2. In the RIB routing table, in addition to the next hop, remote next-hop information should be added

② Forwarding table of Leaf-1:

Path-aware

H2:

Next-hop: Spine-1 -> Path: Spine-1, Leaf-2

Next-hop: Spine-2 -> Path: Spine-2, Leaf-2

H3:

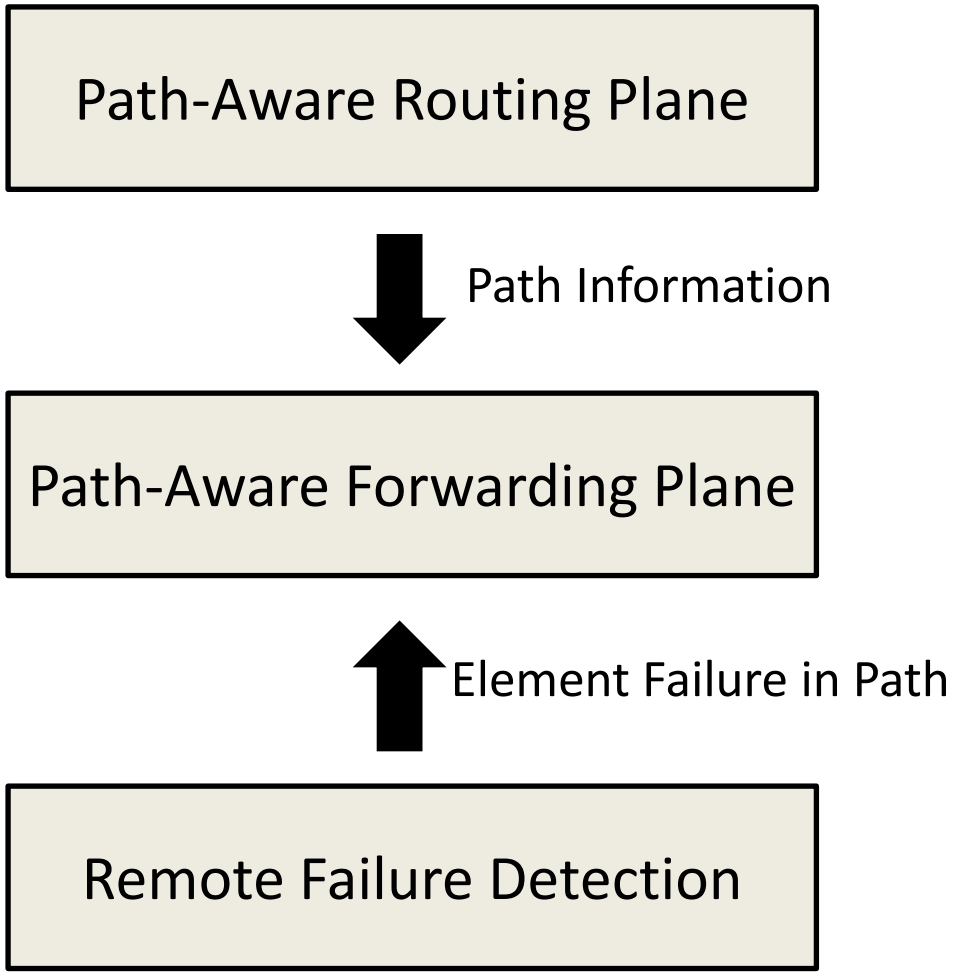
Next-hop: Spine-1 -> Path: Spine-1, Leaf-3

Next-hop: Spine-2 -> Path: Spine-2, Leaf-3

③ When the failure occurs on the link between Spine-2 and Leaf-3 :

1. Spine-2 detects the failure, and then notifies Leaf-1 of the failure on the link between Spine-2 and Leaf-3.
2. Leaf-1 finds the next-hop whose path has failure, and removes it from ECMP.

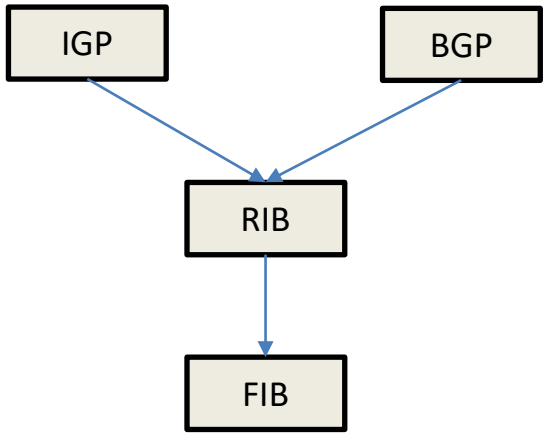
Overview



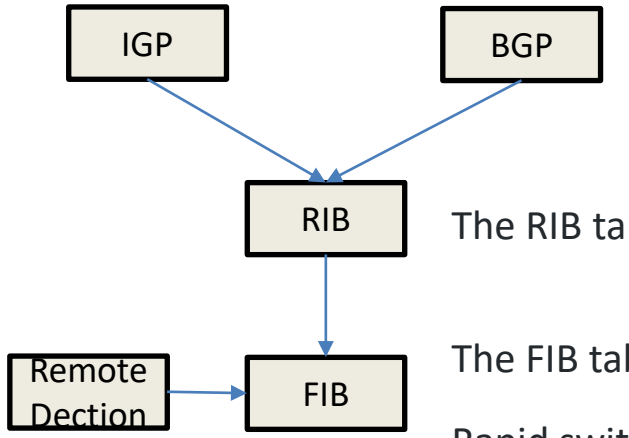
- Routing calculations on the routing plane are not limited to the next hop but require path awareness. The control plane's routing protocol needs to be extended to support calculating remote path information.
- The RIB table of the control plane and the FIB table of the forwarding plane need to include information on remote next hops. The path information is downloaded to the forwarding plane.
- When a failure occurs in any component along the path, it must be quickly detected and repairs should be invoked.

Path-Aware Routing and forwarding Plane

Current routing and forwarding Plane



Path-aware routing and forwarding Plane



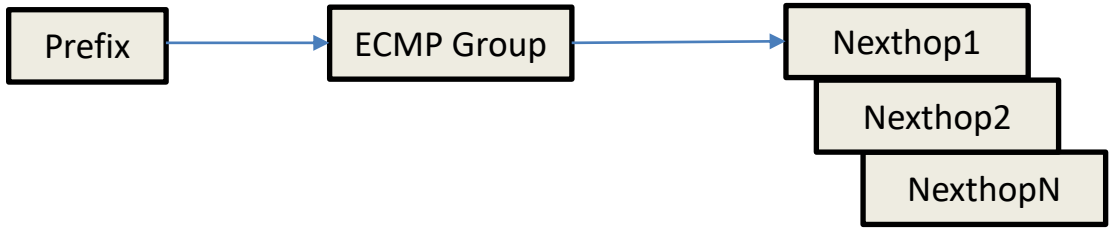
Routing protocol needs to be extended to compute remote Path information.

The RIB table needs to add a Remote Path.

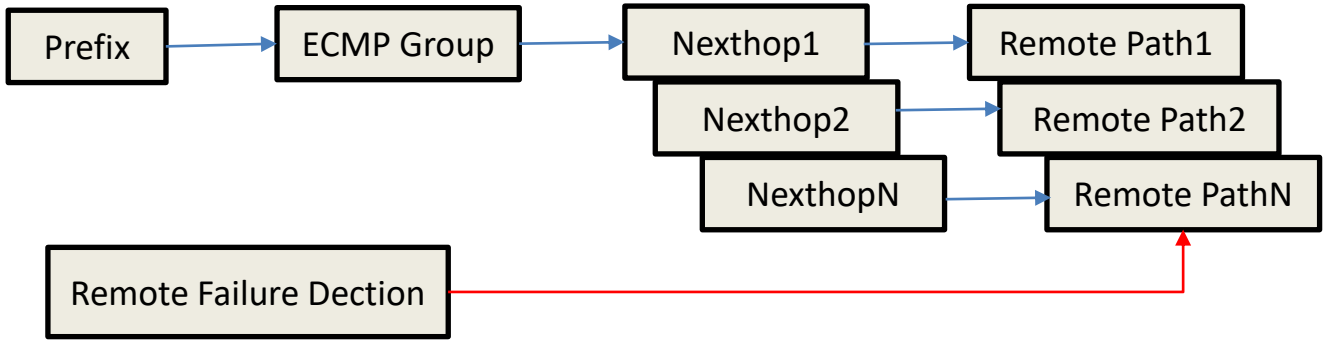
The FIB table needs to add a Remote Path.

Rapid switchover upon detecting remote failures.

Forwarding table for ECMP routes:



Forwarding table for ECMP routes (Add Remote Path Info):



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.

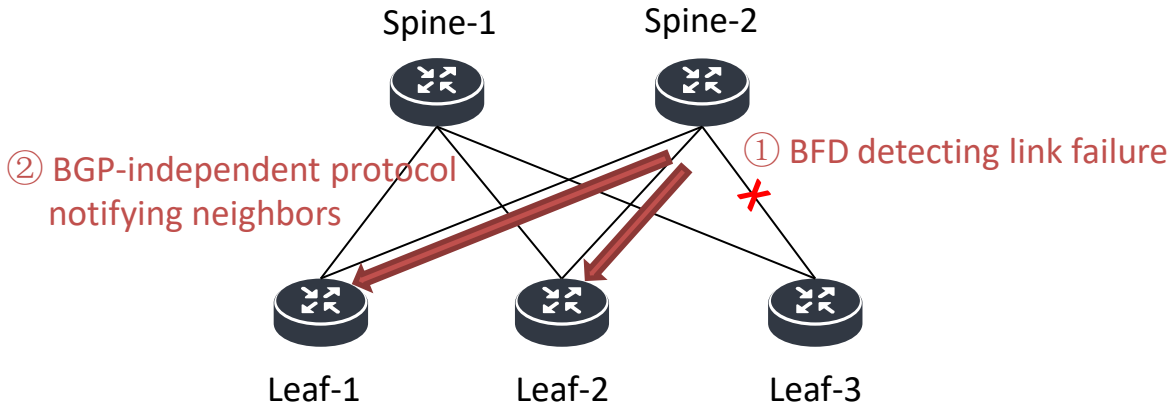
Why not use BFD for Remote detection at the head-end?

Reason: For scenarios with multiple ecmp paths, BFD is difficult to probe a specified path.

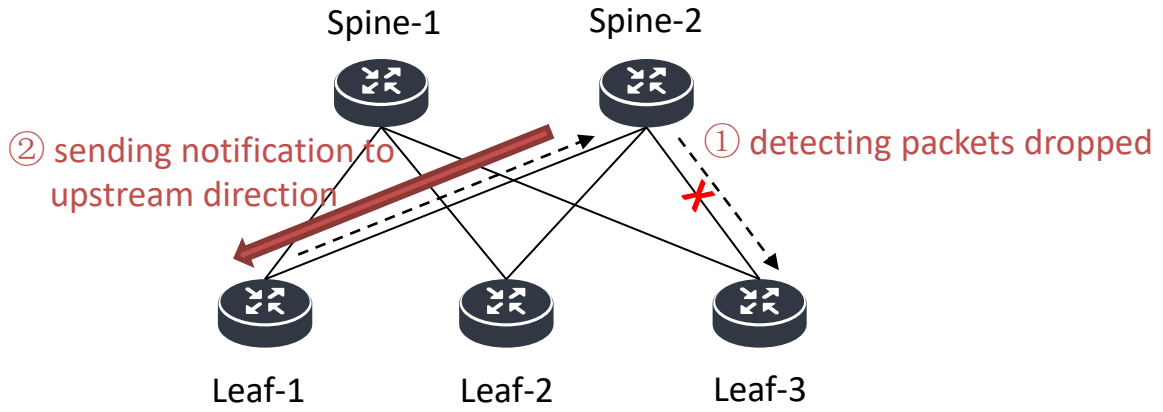
The failure notification between neighboring routers has the following requirements:

- Independent of routing protocols.
- Easy to implement on hardware, achieving fault notification in milliseconds or even microseconds.

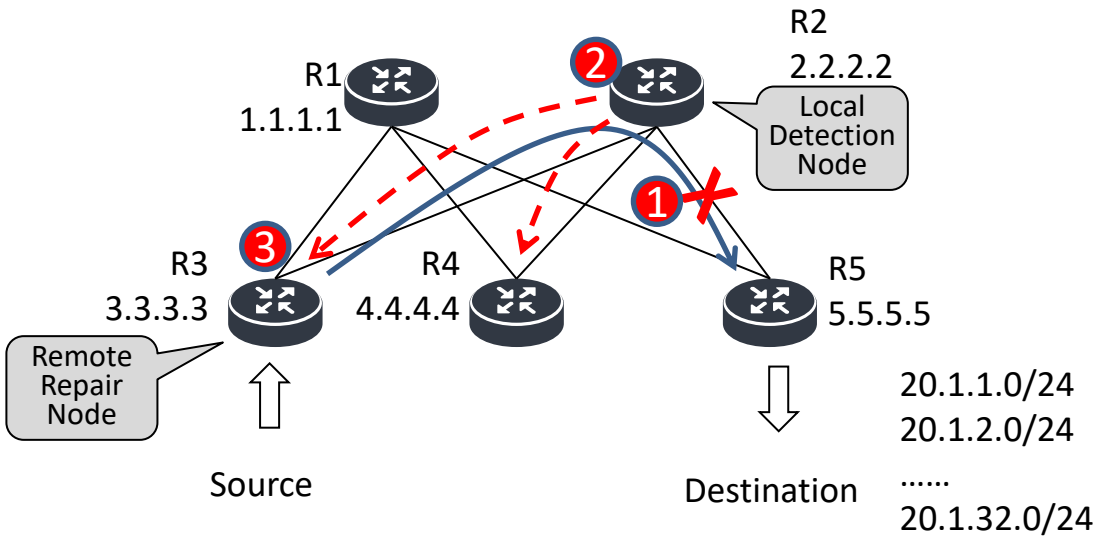
Example-1:

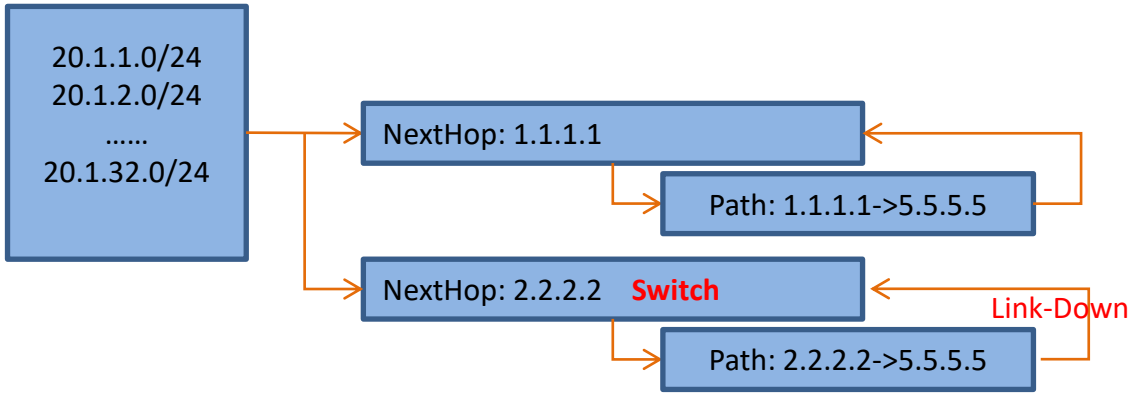


Example-2:



Usecase





Forwarding table on R3

R1 and R2 establish BGP neighbors with R3, R4, and R5 respectively.

① R2 receives routes sent by R5 via BGP, and when R2 forwards these routes to R3 and R4, it adds an extension to carry the next hop information to R5. [draft-wang-idr-next-next-hop-node]

② On R3, after the BGP route selection, the resulting forwarding table is shown above. In addition to the directly connected next hop, there is also remote next hop information.

③ R2 detects link down on R2-to-R5.R2 notifies the remote R3 and R4 about the link-down event. This link-down notification message includes the path information: 2.2.2.2->5.5.5.5. [draft-zzhang-rtgwg-router-info]

④ R3 received a remote link-down notification message and quickly switched the route to another available path based on the path information.

Next Steps

- Any questions or comments are Welcomed.

Thanks