Gap Analysis from the Perspective of Protection and Load Balancing

IETF 122 side meeting

Rui Zhuang (CMCC)

Yisong Liu (CMCC)

Weiqiang Cheng (CMCC)

Standardized fast notification mechanism to support critical network functions

Main Challenges

Adaptive Routing

- Dependent on real-time link status (e.g., latency, congestion, failures ...).
- Traditional protocols (e.g., BGP) converge slowly, cannot meet the sub-second level path switching requirements.

Traffic Management

- Need global traffic distribution and priority information.
- Static strategies are unable to handle burst traffic.

Remote Protection

- Fast fault detection and recovery require cross-domain collaboration.
- Existing mechanisms (e.g., BFD) are limited to single-hop detection and lack a whole network view.

Fast notification is needed to improve network function effectiveness and reduce failure recovery time.

Main Requirements

- Real-time: milliseconds state synchronization (e.g., link congestion, node failures).
- Compatibility: unified devices and protocols across manufacturers to avoid fragmented private protocols.
- Scalability: support large-scale nodes concatenation (e.g., data center network).

Adaptive Routing Framework: draft-cheng-rtgwg-adaptive-routing-framework

(1) Motivation

Problem in AI Network/Problem Statement

- ECMP flow-based hash leads to high congestion and variable flow completion time.
- Not distinguishing between large and small flows, leading to Load imbalance.
- The lack of congestion awareness exacerbates the increased load on already congested links.

Possible Solutions

a) Increasing flow entropy by refining the granularity of load balancing algorithms

e.g. cell-based, packet-based, and flowlet-based

b) Prompting re-hash or re-route by modifying flow characteristics

e.g. Congestion Control: RTT, ECN, etc.

Flow characteristics: 5-tuple, IPv6 Flow Label, etc.

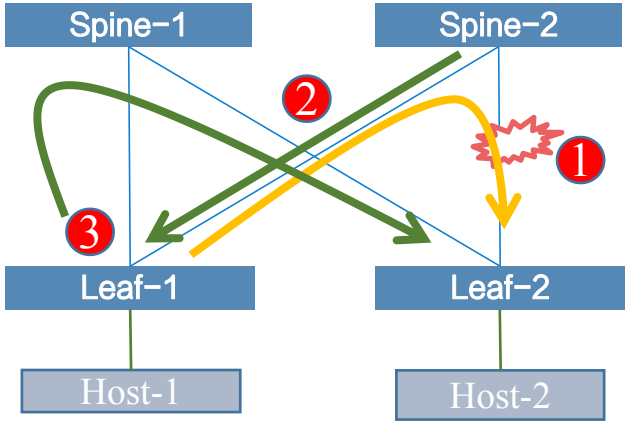
c) Adaptive routing based on network state measurements

Monitor real-time network conditions; Select the optimal path based on the network load and demand

Adaptive Routing Framework: draft-cheng-rtgwg-adaptive-routing-framework

(2) What is Adaptive Routing?

- Each device performs congestion detection, including link-based detection and flow-based congestion detection.
- Upon detecting congestion, notifications should be sent to the remote devices to perceive congestion at earlier nodes.
- Respond to congestion notifications, congestion adjustments could be perfomed by adjusting path weights, path loads or redirecting.



① Spine-2 detects congestion.

② Spine-2 notifies Leaf-1 of congestion.

③ Leaf-1 adjusts paths in response to congestion.

Adjust

Dest	NextHop	Remote Path
Leaf-2	Spine-1	Spine-1->Leaf-2
Leaf-2	Spine-2	Spine-2->Leaf-2

Adaptive Routing Framework: draft-cheng-rtgwg-adaptive-routing-framework

(3) Overall Framework

- Dynamically adjusts the Weight and Load values of forwarding table according to local and remote link quality as well as the payload of forwarded data packets.

- Dynamically adjusts routing and updating the Forwarding Plane.

- Responsible for the transmission and calculation of routes.
- The calculated routes should include remote path information.
- The routes and remote path information should be correlated and updated to the Forwarding Plane.

- Detects link congestion and sends Congestion Notification to neighboring devices.

Adaptive Routing Framework: draft-cheng-rtgwg-adaptive-routing-framework

(4) Remote Congestion Detection

Congestion Detection

- Bandwidth, buffer utilization, queue occupation, ...
- Sending test traffic to identify network performance and congestion points

Congestion Definition

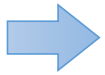
- Based on interface bandwidth or forwarding buffer utilization, measured using a quality level

Congestion Notification

- Communicating link congestion status or congested flow information to remote devices in order to adjust traffic scheduling from the source.

Two types of congestion message:

- The first type includes path information, which includes the congestion information of links corresponding to the Path → Global congestion calculation
- The second type includes the five-tuple information of the congested flow → Congestion flow redirection



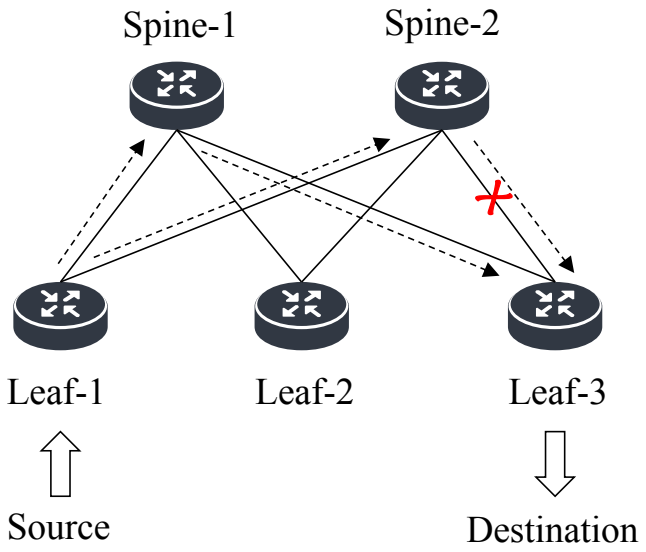
How to achieve fast congestion notification?

- Extending the IGP protocol to transmit link state information within the IGP domain
- Extending the BGP protocol and setting up BGP reflectors to facilitate communication between BGP neighbors
- .....

Path aware remote protection: draft-liu-rtgwg-path-aware-remote-protection

(1) 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.

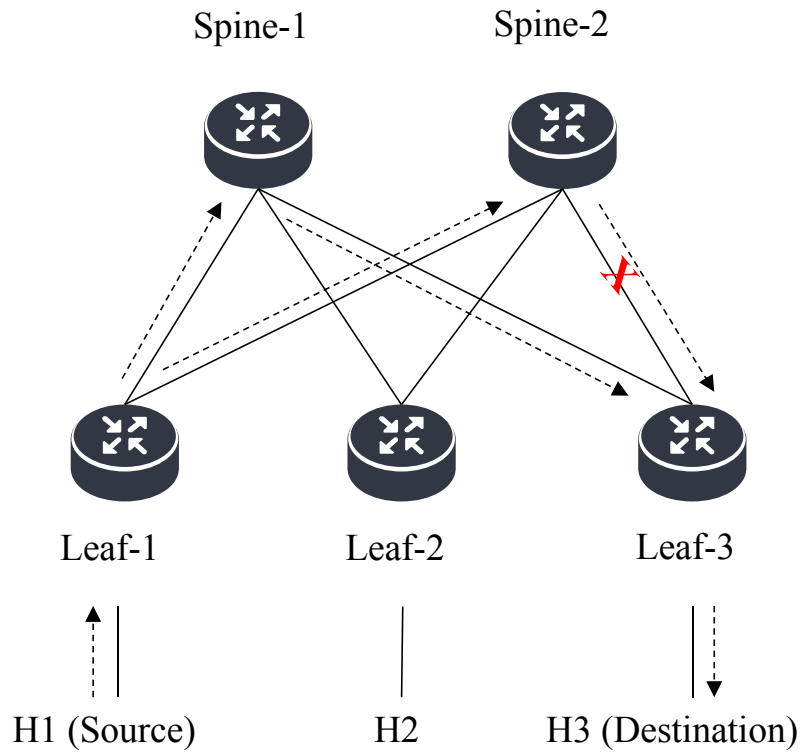
When the failure occurs:

- 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!

Path aware remote protection: draft-liu-rtgwg-path-aware-remote-protection

(2) Running Process



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

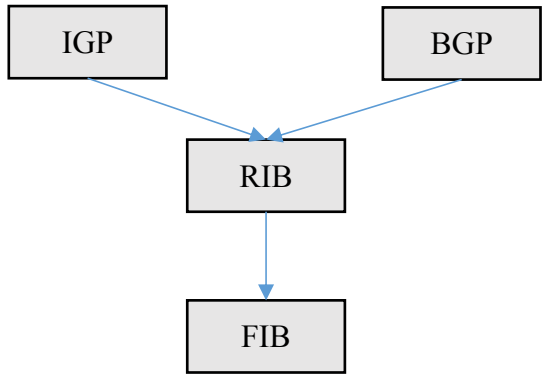
③ When the failure occurs on the link between Spine-2 and Leaf-3:

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.

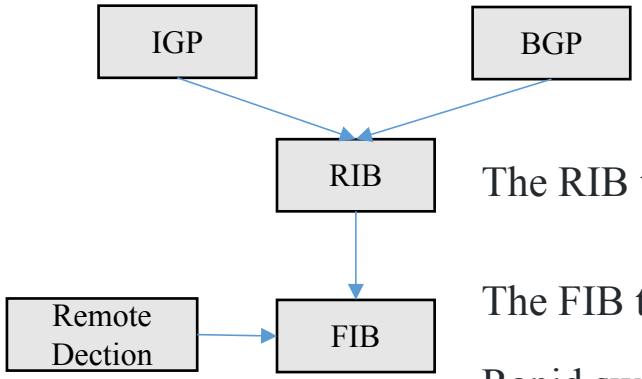
Path aware remote protection: draft-liu-rtgwg-path-aware-remote-protection

(3) Routing and Forwarding Plane

Current routing and forwarding Plane



Path-aware routing and forwarding Plane



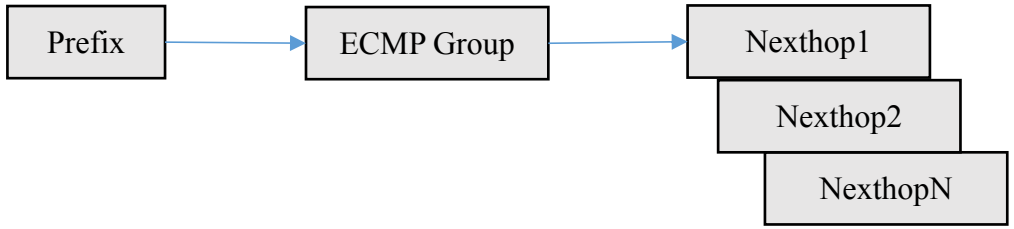
Routing protocol needs to be extended to compute remote Path information.

The RIB table needs to add a Remote Path.

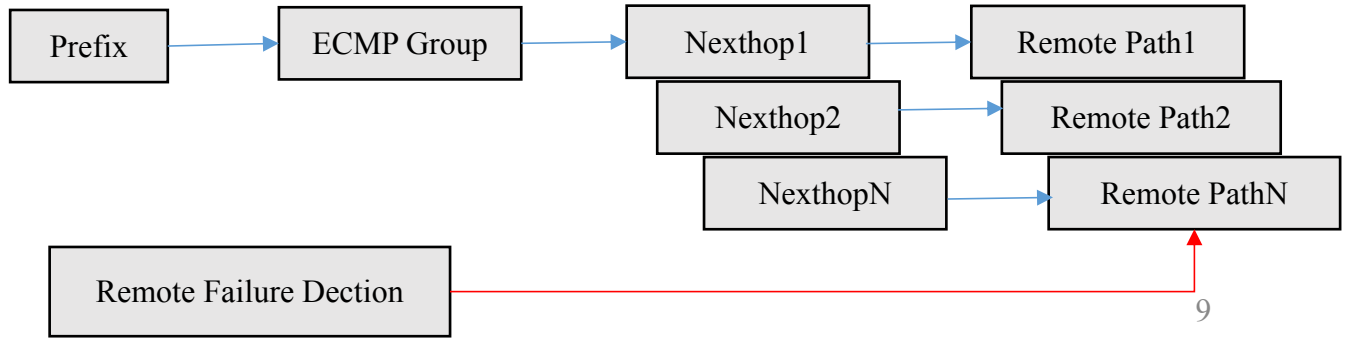
The FIB table needs to add a Remote Path.

Rapid switch-over upon detecting remote failures.

Forwarding table for ECMP routes:



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



Path aware remote protection: draft-liu-rtgwg-path-aware-remote-protection

(4) 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?

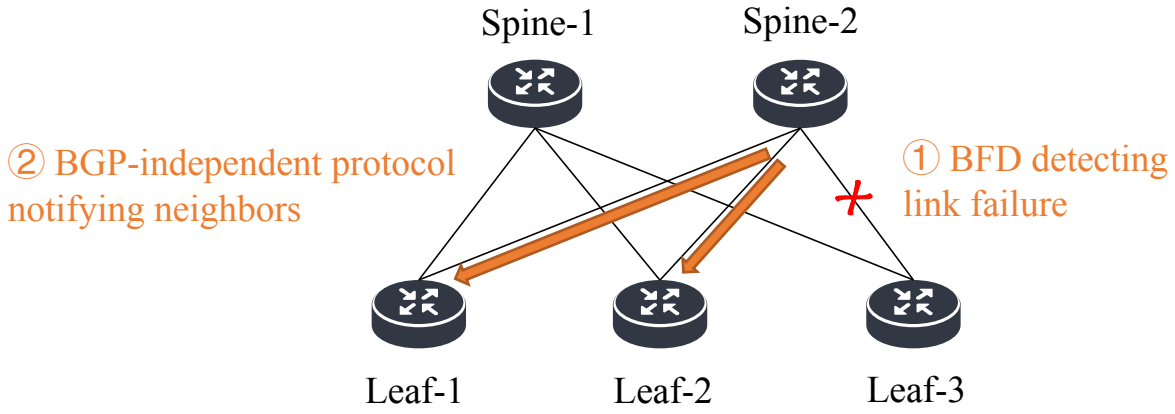
Reason: For scenarios with multiple ecmp paths, BFD is difficult to probe a specified path.

A rapid detection and notification mechanism is needed (independent of routing protocols), [draft-zzhang-rtgwg-router-info] provides a possible solution

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: