Adaptive Routing Framework

draft-cheng-rtgwg-adaptive-routing-framework-01

IETF 120

Weiqiang Cheng(CMCC)

Changwang Lin(H3C)

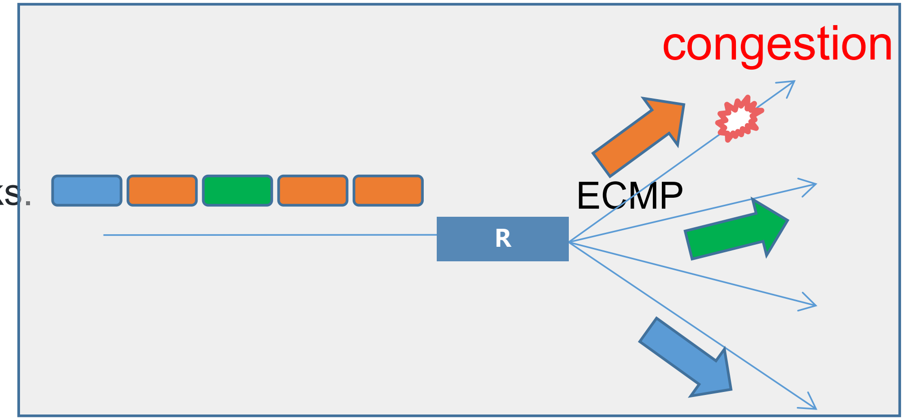
Kevin Wang(Juniper)

Jiaming Ye(CMCC)

Motivation

Problem in AI Network/Problem Statement

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



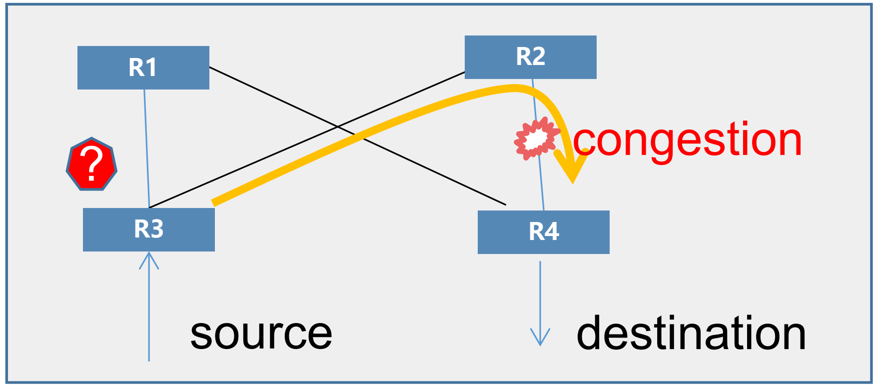
Possible solutions

- Increasing flow entropy by refining the granularity of load balancing algorithms

  e.g. cell-based, packet-based, and flowlet-based
- 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.
- Adaptive routing based on network state measurements

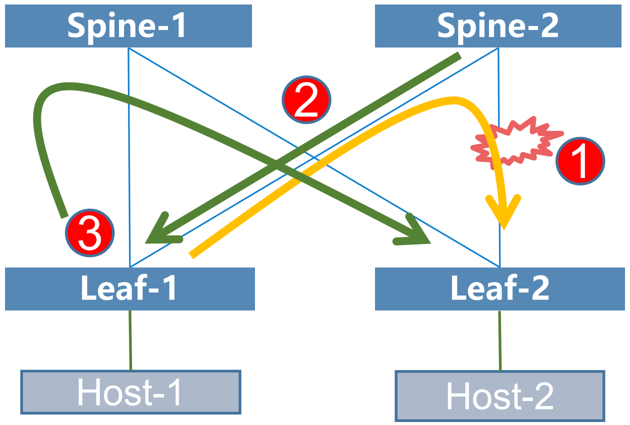
  Monitor real-time network conditions; Select the optimal path based on the network load and demand. Advantages: automatically adjust, better transmission performance and reliability, Qos.



What is Adaptive Routing?

Adaptive Routing

- To achieve even load distribution, per-flow load balancing and per-packet load-balancing could be performed.
- Each device performs congestion detection, including link-based detection and flow-based congestion detection.
- Upon detecting congestion, notification should be sent to the remote devices to perceive congestion at earlier nodes.
- Respond to congestion notifications, congestion adjustments could be performed by adjusting path weights or path loads or by redirecting flows.



①Spine-2 detects congestion.
②Spine-2 notifies Leaf1 of congestion.
③Leaf-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

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

Routing Plane

$\downarrow$ Remote Path Info

Forwarding Plane

- Responsible for path adjustments based on the policies of Adaptive Routing and remote link congestion information.

Adaptive Routing Policy $\rightarrow$ Forwarding Plane

- Responsible for remote link congestion information or flow information
- Dynamically adjusting routing accordingly and updating the Forwarding Plane.

Adaptive Routing Policy

$\uparrow$ Congestion Notify

Remote Congestion Detection

- Responsible for detecting link congestion and
- sending Congestion Notification to neighboring devices.

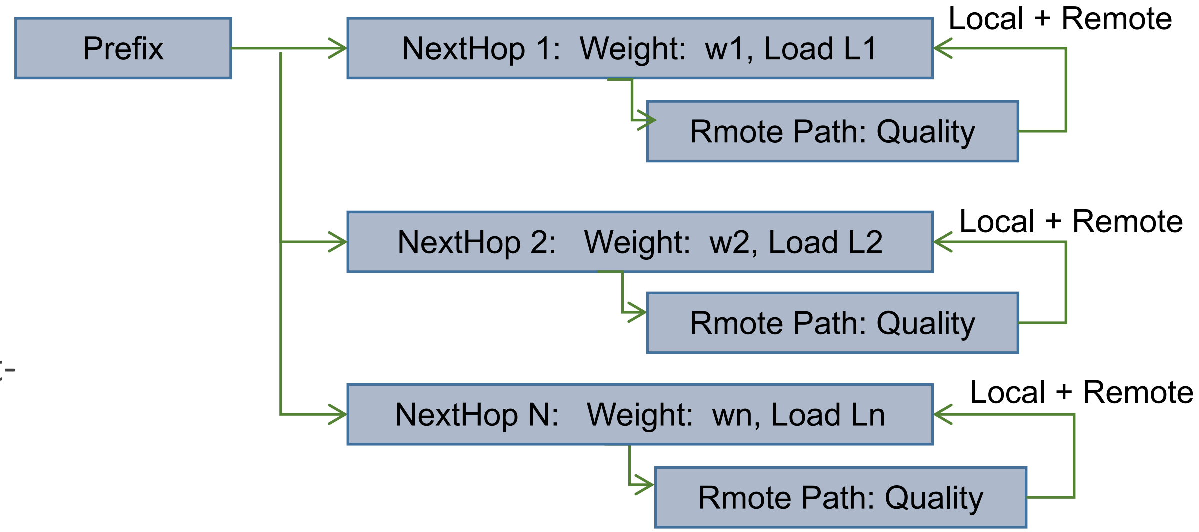
Framework Components

Routing Plane

- When calculating routes, the path needs to be perceived, and the path information will be attached to the next hop.
- For BGP-based networks: Remote path info can be the BGP identifier corresponding to the next-next-hop, as described in [I-D.wang-idr- next-next-hop-nodes]. It can also be the BGP AS-PATH information or BGP router-id, which is not detailed in this document.
- For IGP-based networks: Remote path info can be the interface information from the next-hop neighbor device to the next-hop device, which could be the interface index, or the interface's local address.

Forwarding Plane

- The forwarding plane maintains the forwarding table based on the routing table provided by the routing layer.
- It 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.
- Forwarding table is used to generate the appropriate flow table for flow-based forwarding and to perform load balancing during packet-based forwarding.



Framework Components

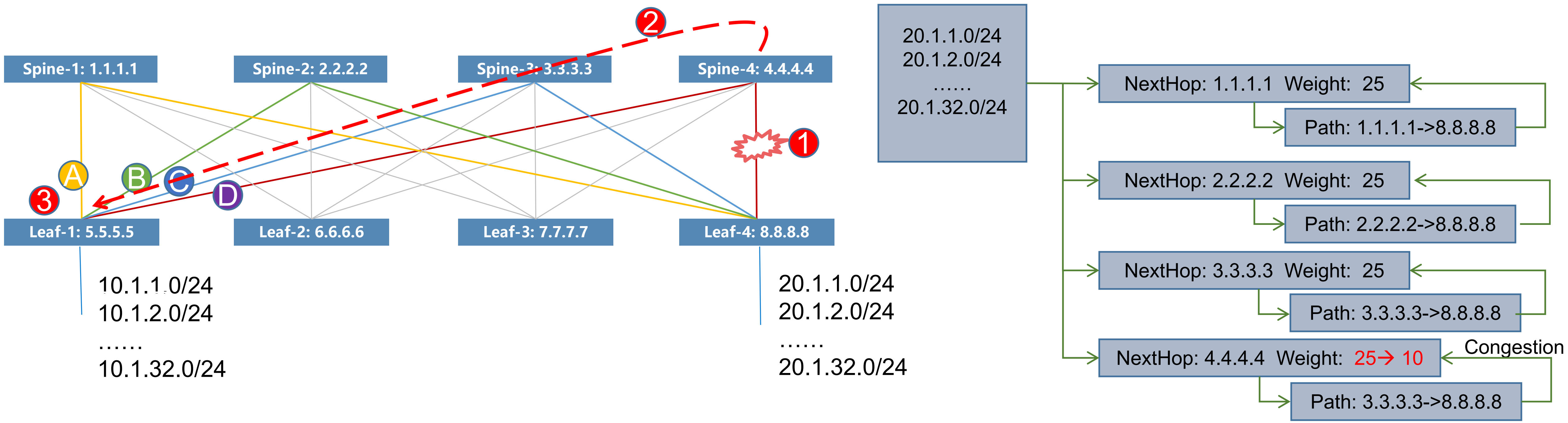
Adaptive Routing Policy

- Flow-Based Adjustment Mode
  - Weight-Based Dynamic ECMP Flow Mode
    - Load balance per flow with each ECMP path assigned a weight.
    - Adjust path weights to avoid congested paths for subsequent flows.
  - Flow Redirection Mode
    - Load balance per flow with each path dynamically maintaining a Load value.
    - Adjust the Load value for congested paths
    - Redirect congested flows to paths with lower Load values.
- Packet-Based Adjustment Mode
  - Load balance per packet with each path dynamically maintaining a Load value.
  - Forward each packet through the path with the lowest Load value.
  - To resolve potential out-of-order issues within the same flow on the application receiving side.

Congestion Control

- Congestion Detection
  - Link's buffer, bandwidth, and queue congestion
  - The corresponding flow experiences congestion
- Congestion Notify
  - To notify link congestion or inform about the congested flow information
  - Refer to [I-D. draft-zzhang-rtgwg-router-info] for implementation

WorkFlow 1: Weight-Based Dynamic ECMP Flow Mode



Before congestion occurs

- When Leaf-4 advertises the 20.1.1.0/24 to 20.1.32.0/24 network segments via BGP to all spines, the spines will then advertise these network segments with the next-next-hop set to 8.8.8.8.
- On Leaf-1, a forwarding table is created, as shown in the top-right diagram, and it contains the remote path information.

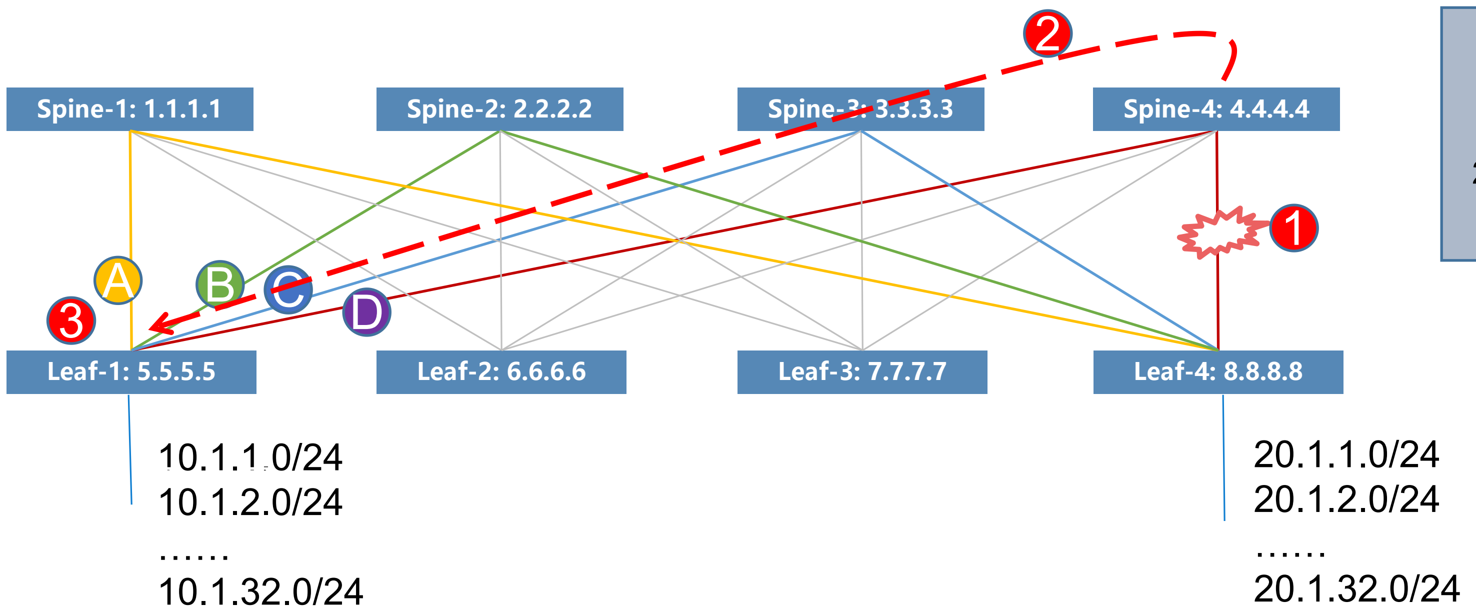
Adaptive adjustment based on weight

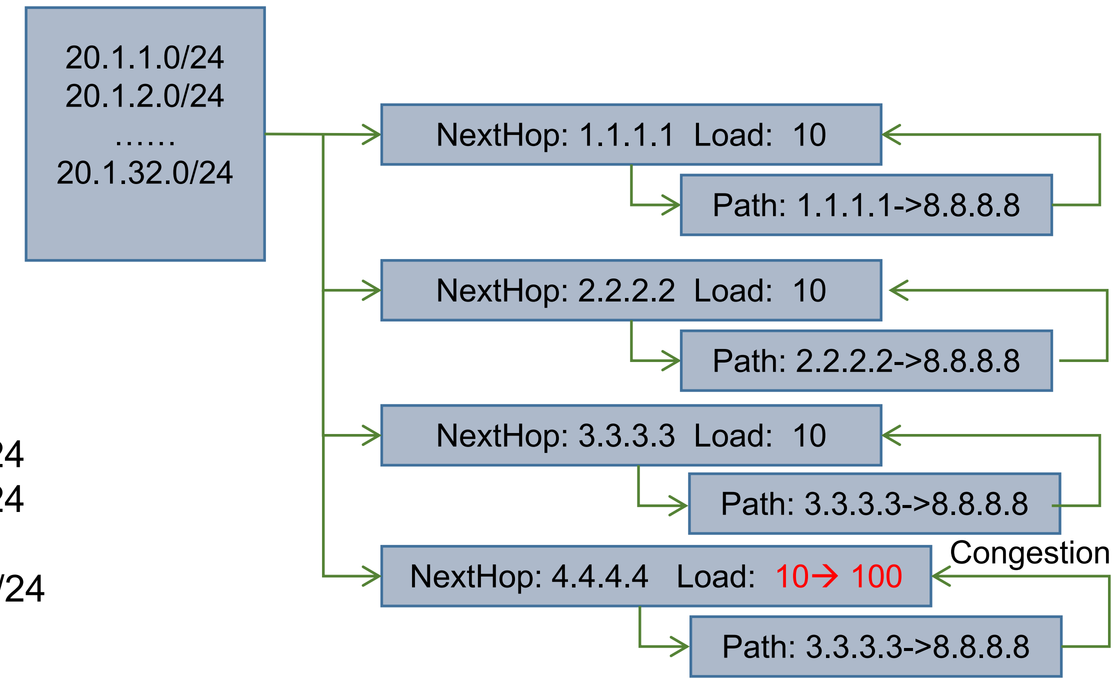
① Spine-4 detects congestion on the Spine4-to-Leaf4 link and notifies the remote Leaf-1 about the congestion event. This congestion notification message includes the path information: 4.4.4.4 -> 8.8.8.8.

② Spine-1 receives the congestion notification message, identifies the corresponding forwarding entry based on the path information, and reduces the weight from 25 to 10.

③ As a result of this congestion adjustment, new traffic will be forwarded based on the updated proportions, reducing the load on the congested link.

WorkFlow 2: Flow Redirection Mode





Before congestion occurs

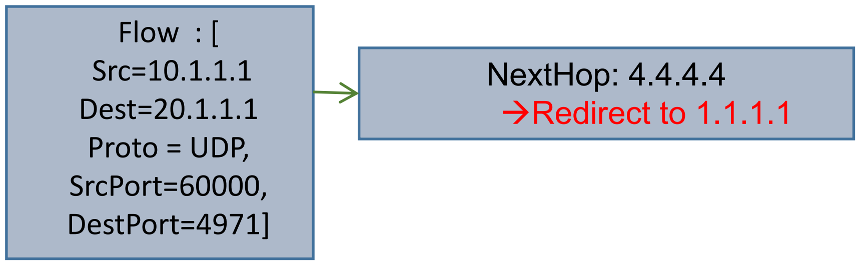
- Flow: src=10.1.1.1, dest=20.1.1.1, proto=UDP, SrcPort=60000, DestPort=4971, forwarding path is Leaf-1, Spine-4, Leaf-4.

Flow Redirection Mode

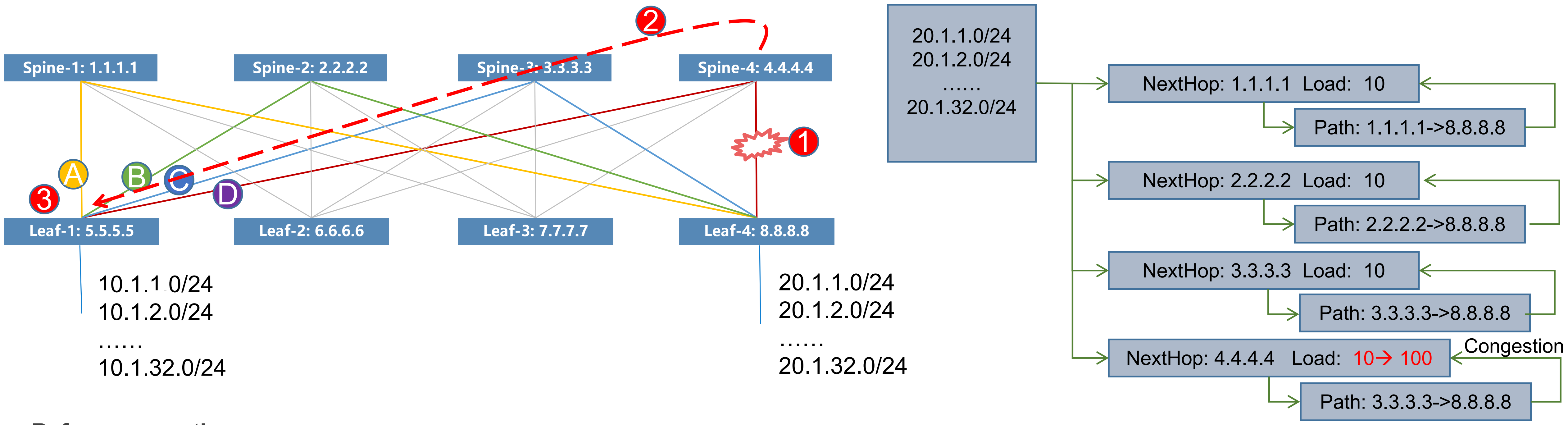
① Spine-4 detects congestion on the Spine-4 to Leaf-4 link and notifies the remote Leaf-1 of the congestion event. The congestion notification message includes the path information: 4.4.4.4 -> 8.8.8.8, and the congested flow information [10.1.1.1, 20.1.1.1, UDP, 60000, 4971].

② Leaf-1 responds to the congestion notification message, locates the corresponding forwarding entry based on the path information, and increases the weight value from 10 to 100. Additionally, Leaf-1 locates the corresponding flow table based on the congestion flow information and redirects the congested flow to the least-loaded path.

③ Congestion adjustment result:
The congested flow is redirected to non-congested paths for forwarding.



WorkFlow 3: Packet-Based Adjustment Mode



Before congestion occurs

- When Leaf-4 advertises the 20.1.1.0/24 to 20.1.32.0/24 network segments via BGP to all spines, the spines will then advertise these network segments with the next-next-hop set to 8.8.8.8.
- On Leaf-1, a forwarding table is created, as shown in the top-right diagram, and it contains the remote path information.

Adaptive adjustment based on

① Spin4 detects congestion on the Spin4->Leaf4 link and notifies the remote Leaf1 of the congestion event. The congestion notification message includes the path information: 4.4.4.4->8.8.8.8.

② Leaf1 responds to the congestion notification message, locates the corresponding forwarding entry based on the path information, and increases the load value 10 -> 100.

③ Congestion adjustment result: Subsequent packet will preferentially choose non-congested paths.

Next Steps

- Any questions or comments are Welcomed.

Thanks

Question

- **Problem 1: How to ensure timely adjustments**
  Congestion notifications can be sent via software or through dedicated hardware similar to BFD (Bidirectional Forwarding Detection), allowing high-frequency transmissions. This can enable adjustments at the millisecond or even microsecond level.

- **Problem 2: When adjusting weight modes, multiple devices might make adjustments. How to ensure adjustments are accurate and effective?**
  Adjustments should be gradual and not be too large in magnitude. If there is a clearly identified elephant flow, consider using flow-based or packet-based adjustments.

- **Problem 3: When adjusting flow mode, how can we identify the congested flow and ensure that moving the congested flow does not cause new congestion elsewhere?**
  A global perspective is needed to calculate the approximate load on the links to which the traffic will be moved.

- **Problem 4: In per-packet forwarding mode, what is the difference from the spray link mechanism?**
  In per-packet forwarding, sorting is done at the receiving end rather than on the device. Adaptive routing does not make excessive demand on the device's chip, making it suitable for a wider range of scenarios.