Adaptive Routing Framework

draft-cheng-rtgwg-adaptive-routing-framework-03

IETF 121

Weiqiang Cheng(CMCC)

Changwang Lin(H3C)

Kevin Wang(Juniper)

Jiaming Ye(CMCC)

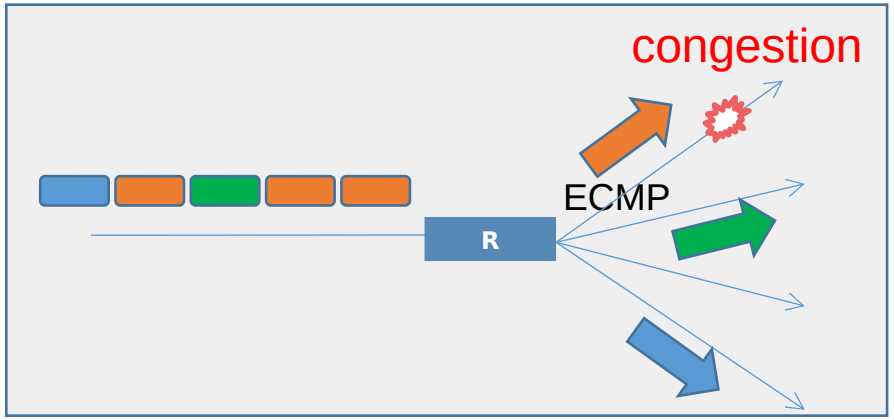
Rui Zhuang(CMCC)

Pengfei Huo(ByteDance)

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.



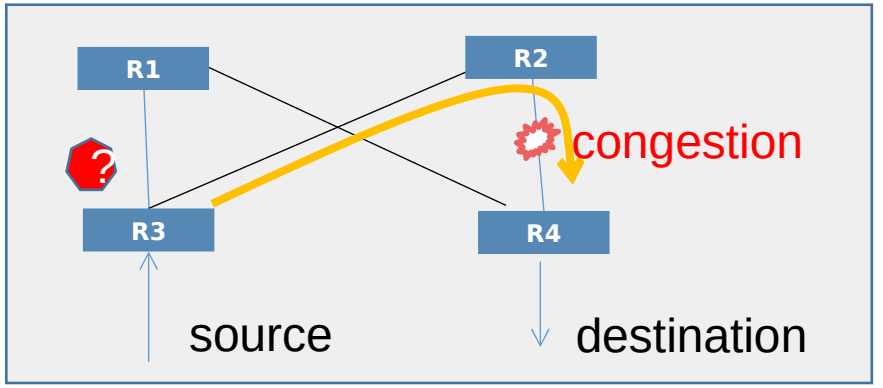
Possible Solutions

- 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

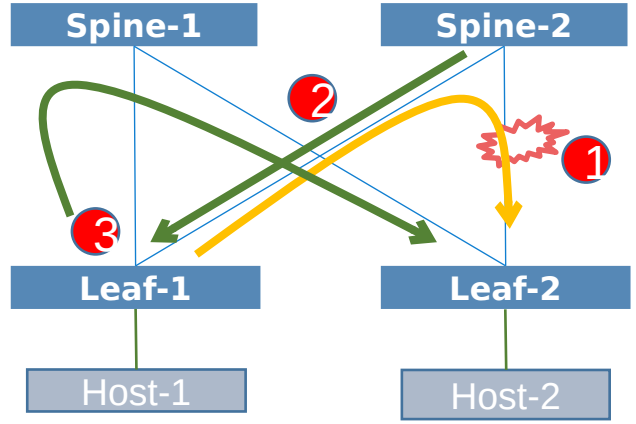
  Monitor real-time network conditions; Select the optimal path based on the network load and demand. Advantages: automatically adjust, better transmission performance and reliability, Qos.



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 perfomed by adjusting path weights, path loads or redirecting



① Spine-2 detects congestion.
② Spine-2 notifies Leaf1 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

Change of this draft

- Presented at IETF-120 and **made modifications based on comments**.
- **Updated the abstract and introduction sections.**
- **Clarified the scope of this document:** adaptive adjustments on multiple loop-free paths.
  - How to generate multiple loop-free paths is not within the scope of this document.
- **Added a definition of ``congestion''.**

Change of this draft

- Some Comments from IETF 120, responses are as follows:

  - Acee: There are drafts talking about per packet load balancing. This makes the assumption that if a very big flow, the problem is pushed to the application for packet re-ordering etc., then I'm not sure this works.

  - Jeff Tantsura: By the time you get to recognizing the flow problems, a great deal of the flow may be lost. I do not know of any mechanism that can quickly readjust those flows.

  [Reply]: You can set a congestion threshold, and adjustments will begin once a specific threshold is reached.

  There are three operating modes for handling congestion and remote link failures:

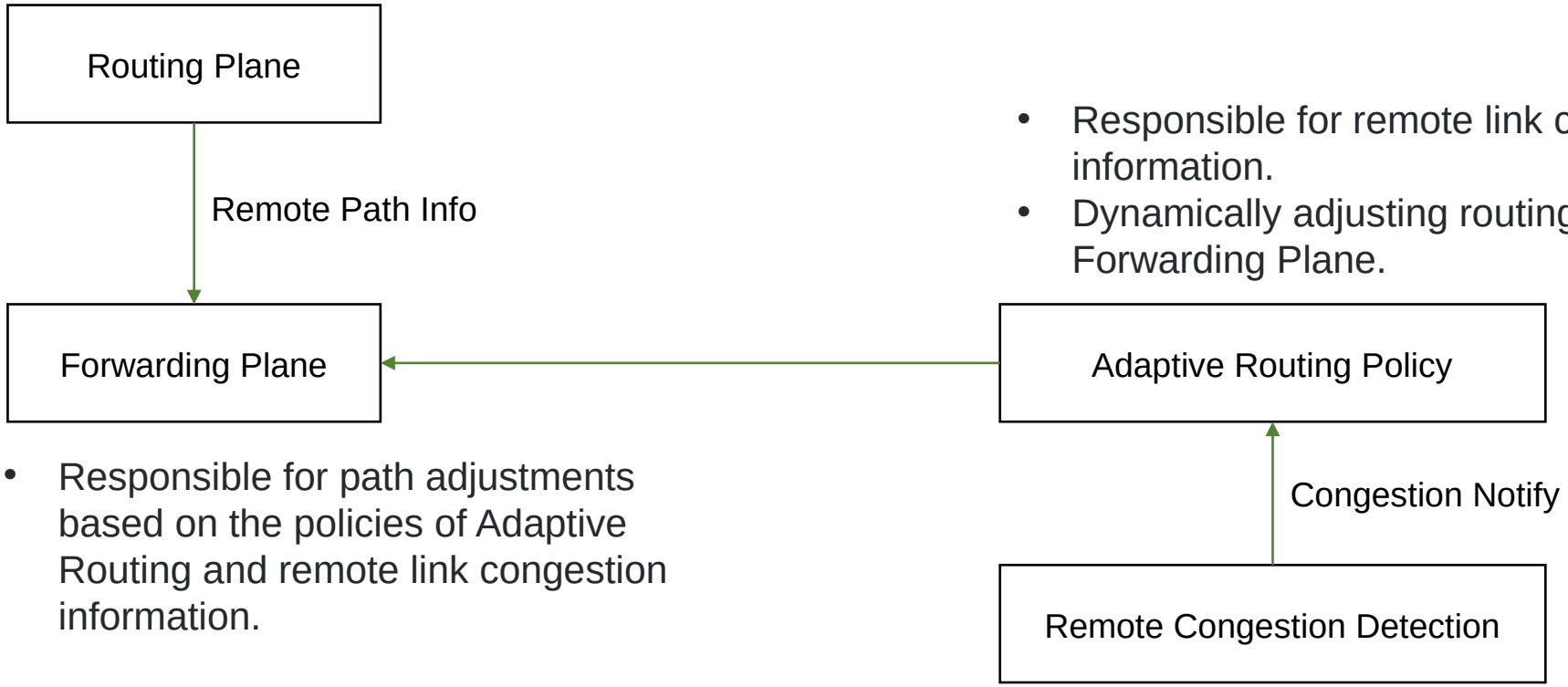
  a) **Weight-Based Dynamic ECMP Flow Adjustment Mode:** Periodically receives congestion status from the remote link and dynamically adjusts the load distribution weight.

  b) **Flow Redirection Mode:** Redirects part of the congested flow. Suitable for scenarios like AI training where the flow is regular.

  c) **Packet-Based Adjustment Mode:** Requires endpoint support for ordering, each packet is forwarded via the link with the least load.

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.



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

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

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

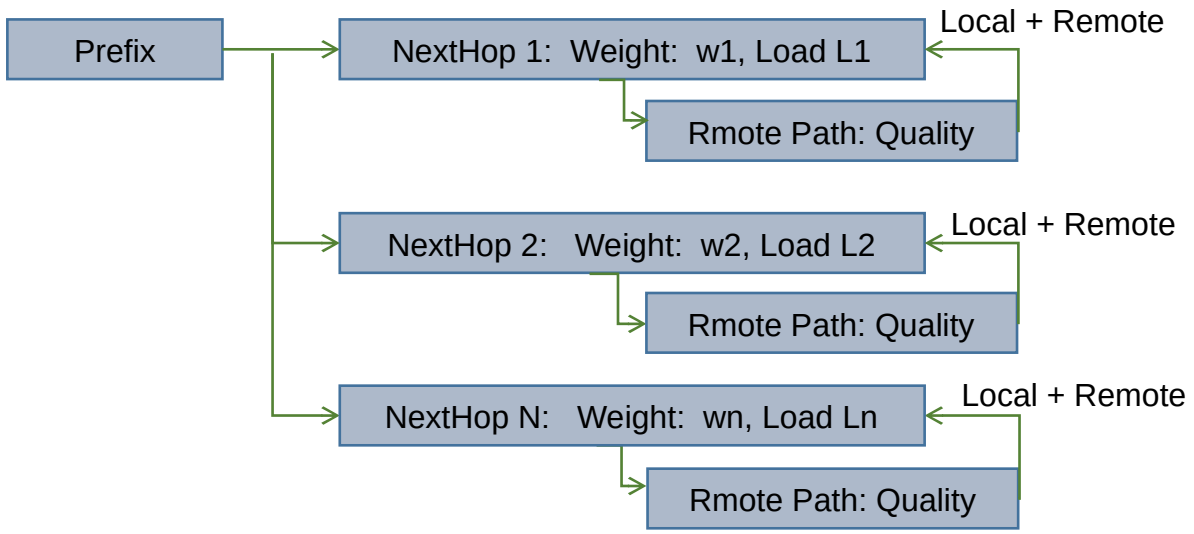
Framework Components

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 plan 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.

Framework Components

Remote Congestion Control

Congestion Detection

- Link's buffer, bandwidth, and queue congestion
- Network performance and congestion points can be identified by sending test traffic
- Specific congestion detection methods are beyond the scope of this document

Congestion Definition

- Based on interface bandwidth or forwarding buffer utilization, measured using a quality level
- This level can be tailored so that lower levels indicate poorer path quality and can be calculated

Congestion Notify

- To notify link congestion or inform about the congested flow information
- Refer to [I-D. draft-zzhang-rtgwg-router-info] for implementation

Remote Congestion Control

Congestion Definition

- Congestion can be defined based on interface bandwidth or forwarding buffer utilization, measured using a quality level.
- Quality Level:
  - can be calculated based on current bandwidth and buffer usage.
  - can be tailored so that lower levels indicate poorer path quality.
- For instance, with 16 quality levels, on a 400G interface, level 0 could represent 25G-available and level 15 could represent 400G-available.

Congestion Notify

- Communicate congestion status 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 can be performed with this information.
  - The second type includes the **five-tuple information of the congested flow**. Congestion flow redirection can be implemented with this information.
- This can be achieved by extending the IGP protocol to transmit link state information within the IGP domain, or by extending the BGP protocol and setting up BGP reflectors to facilitate communication between BGP neighbors.

Running Code

Implemented chips or vendors:

- **Broadcom TH5 GLB:** Weight-Based Dynamic ECMP Flow Mode, Packet-Based Adjustment Mode
- **Juniper:** Junos OS Evolved Release 23.4R2(QFX5240-64OD/QFX5240-64QD) Weight-Based Dynamic ECMP Flow Mode, Packet-Based Adjustment Mode
- **H3C:**

  - Comware V9 B58 (S9827-128DH/S9827-64EP) : Weight-Based Dynamic ECMP Flow Mode, Packet-Based Adjustment Mode

  - Comware V7 B70D064 (7800XPG/7500G) : Flow Redirection Mode

Next Steps

- Any questions or comments are Welcomed.

Thanks