Network Working Group
Internet Draft
Intended status: Informational
Expires: October 07, 2026

W. Cheng
China Mobile
C. Lin
New H3C Technologies
K. Wang
Juniper Networks
J. Ye
R. Zhuang
China Mobile
P. Huo
ByteDance

April 07,2026

Adaptive Routing Framework
draft-cheng-rtgwg-adaptive-routing-framework-05

Abstract

In many cases, ECMP (Equal-Cost Multi-Path) flow-based hashing leads to high congestion and variable flow completion time. This reduces applications performance. Load balancing based on local link quality is not always optimal, A global view of congestion, with information from remote links, is needed for optimal balancing. Adaptive routing is a technology that makes dynamic routing decision based on changes in traffic load and network topology.

This document describes a framework for Adaptive Routing. Specifically, it identifies a set of adaptive routing components, explains their interactions, and exemplifies the workflow mechanism.

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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on October 07, 2026.

Table of Contents

1. Introduction

In many cases, ECMP (Equal-Cost Multi-Path) flow-based hashing leads to high congestion and variable flow completion time. This reduces applications performance. Load balancing based on local link quality is not always optimal, A global view of congestion, with information from remote links, is needed for optimal balancing.

Adaptive routing is a network routing mechanism that dynamically adjusts routing paths based on changes in network conditions, thereby optimizing network performance and resource utilization.

This document defines a framework for Adaptive Routing. Specifically, it identifies adaptive routing components, explains their interactions, and illustrates the workflow mechanism. It focuses exclusively on dynamic load balancing for existing loop-free multiple paths, allowing adjustments based on remote link quality. The formation of loop-free paths is outside the scope of this document.

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. Problem Analysis

The current AI networks exhibit the following characteristics: a low number of flows, but each flow has a heavy load. The commonly used load balancing strategy employs an N-tuple hash algorithm to forward traffic on a per-flow basis. For current AI networks, this load balancing strategy can easily lead to load imbalances, causing network congestion.

When network congestion occurs, the current load balancing adjustment strategy typically involves nearby devices at the congestion point switching links based on the local link congestion state. However, this approach is inefficient because adjustments made by devices near the congestion point have limited impact. If load balancing adjustments could be initiated from the earliest routing devices, it would significantly improve the efficiency of load balancing.

2.1. Use Case 1

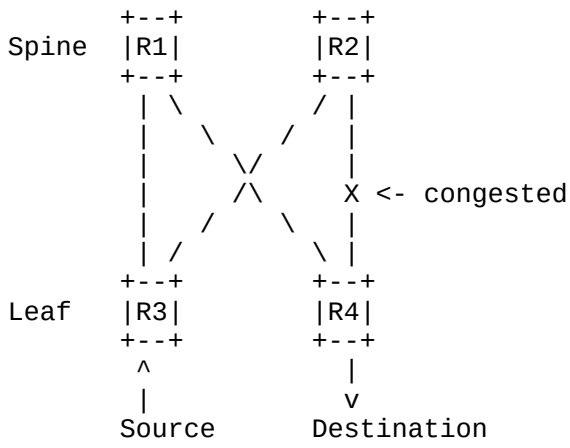

Figure 1 Spine-Leaf network

In the Spin-Leaf network shown in Figure 1, assuming that the R2-R4 link becomes congested, R3 will continue to send traffic to both R1 and R2. Due to the congestion, continuing to forward traffic at the current rate through R2 will exacerbate the link congestion, leading to the loss of some traffic.

2.2. Use Case 2

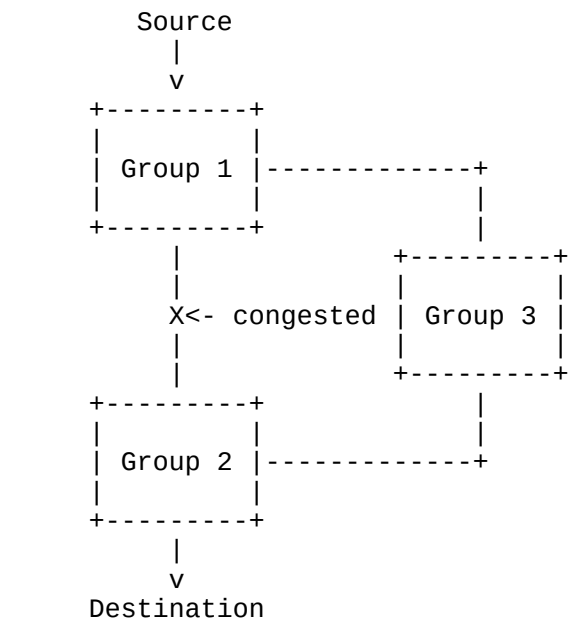

Figure 2 Dragon-fly network

In the dragon-fly network shown in Figure 2, the ECMP paths include Group1->Group2 and Group1->Group3->Group2 for load balancing. When the link between Group1 and Group2 becomes congested, Group1 continues to send traffic at the current rate through the Group1->Group2 link, exacerbating the congestion and causing the loss of some traffic.

3. Solution

To address the problem of load imbalance mentioned above, solutions can be classified into two types: flow-based adjustments and packet-based adjustments. In flow-based adjustment, each flow is forwarded along a single path, and the dynamic adjustment is in the load distribution across multiple paths. In packet-based adjustment, packets can be forwarded across multiple paths on a per-packet basis, and if ordering is required, the receiving end must handle the reordering.

3.1. Flow-based solution

Flow-based load balancing adjustments can be further categorized into weight-based dynamic ECMP flow mode and redirecting congested flows mode. The weight-based dynamic ECMP flow mode adjusts the forwarding ratio across multiple paths in real-time to prevent

further congestion on remote paths. In contrast, the redirecting congested flows mode reroutes actual congested flows from the congested link to other links, suitable for scenarios with relatively stable flows.

3.1.1. Weight-Based Dynamic ECMP Flow Mode

When congestion occurs and nearby devices detect it, congestion information is sent to remote devices. The remote devices then dynamically adjust the forwarding weights of the ECMP paths for this link according to the congestion status. This reduces traffic through the congested link and alleviates the load. Using a weighted load balancing strategy instead of a hash-based strategy can more effectively utilize the bandwidth resources of multiple links. By assigning forwarding weights based on the status of each link, the load can be more evenly balanced.

The disadvantage of weight-based dynamic ECMP flow Mode is that it cannot adjust existing flows, only new ones.

Examples of weight-based dynamic ECMP flow mode are as follows:

Example 1:

In Figure 1, when R2 detects congestion on the R2->R4 link, R2 sends the congestion information to R3 via the control plane. R3 then dynamically adjusts the forwarding weights of the ECMP paths based on the congestion status, reducing the forwarding weight for the congested link. This decreases the traffic on that link and alleviates its load. Once the congestion is cleared, R2 sends a congestion clearance message to R3 via the control plane, and R3 restores the original forwarding weight for that link.

Example 2:

In Figure 2, when the egress router in Group 1 detects inter-group link congestion, it sends a congestion message to the ingress router via the control plane. The ingress router dynamically adjusts the forwarding weights of the ECMP paths based on the congestion status, reducing the traffic through the Group1->Group2 link to alleviate the load on the congested link. Once the congestion is cleared, the egress router in Group 1 notifies the ingress router in Group 1 of the congestion-clearance message, and the ingress router restores the ECMP link weights.

3.1.2. Flow Redirection Mode

When congestion occurs and nearby devices detect congestion in a specific flow, they send congestion information to remote devices. The remote devices then recompute the load balancing for the

congested flow and select a less loaded ECMP link for forwarding the congested flow.

The advantage of the flow redirection mode is that it specifically adjusts the congested flows, which can quickly alleviate congestion.

Examples of flow redirection mode are as follows:

Example 1:

In Figure 1, when R2 detects congestion in a specific flow S on the R2->R4 link, R2 sends congestion information about flow S to R3 via the control plane. R3 then redirects flow S by selecting a less loaded ECMP path to forward flow S.

Example 2:

In Figure 2, when the egress router in Group 1 detects congestion in a specific flow S on the inter-group link, it sends a congestion message about flow S to the ingress router via the control plane. The ingress router then redirects flow S by selecting a less loaded link to forward flow S.

3.2. Packet-based solution

The packet-based adjustment method does not select load-balancing paths based on flows; instead, it chooses the load-balancing path for each individual packet. This approach helps prevent congestion caused by imbalanced load distribution from extremely large flows.

For multiple ECMP links, link load status is monitored based on link quality. When selecting the forwarding path for each packet, the ECMP link with the lowest load is chosen. After forwarding a packet, the recorded load for that link is correspondingly increased.

When responding to congestion notifications from remote devices, the load on the ECMP links is adjusted to influence the path selection for subsequent packets. This dynamic adjustment of link load helps achieve load balancing. Upon receiving a packet, per-packet forwarding mode can be used based on the packet’s markers or the current per-packet forwarding mode. Each time, the least loaded link is selected for sending, in order to achieve the effect of overall load balancing.

Example:

In Figure 1, when R2 detects congestion on the R2->R4 link, it sends congestion information to R3. R3 then dynamically adjusts the load status of the ECMP paths according to the congestion status, increasing the load for the congested link. As a result, subsequent

packets will choose other links due to the higher load on the congested link, thus relieving congestion. Once the congestion is cleared, R2 sends a congestion clearance message to R3 via the control plane, and R3 decreases the load for that link.

4. Framework

4.1. Framework Overview

A high-level view of the Adaptive Routing framework, without expanding the functional entities in the network, is illustrated in Figure 3.

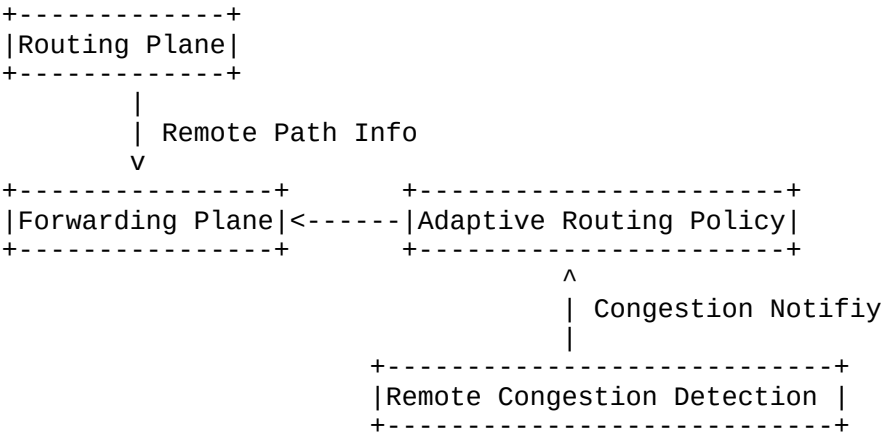

Figure 3 Adaptive Routing Framwork Overview

As shown in Figure 3, the following planes are defined:

* Routing Plane: 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. Control plane protocols can statically specify remote paths or, through protocol extensions, calculate remote path information and update it to the forwarding plane.

* Forwarding Plane: Responsible for path adjustments based on Adaptive Routing policies and remote link congestion information, implementing adjusted forwarding strategies for traffic. In addition to the traditional next-hop in routing, extra information is needed to describe remote link details. Beyond the direct next-hop, an additional remote next-hop is required to convey remote path information.

* Adaptive Routing Policy: Handles remote link congestion or flow information, dynamically adjusting routing and updating the Forwarding Plane. There are two adjustment modes: flow-based and packet-based. Congestion information and adjustment modes update local forwarding table information.

* Remote Congestion Detection: Detects link congestion and sends Congestion Notifications to neighboring devices. It dynamically senses congestion information, defines it consistently, and promptly informs surrounding devices.

4.2. Remote Path Info

Currently, the forwarding table contains information about the route destination, next hop, and exit interface. Local dynamic load balancing can dynamically adjust the weight of load distribution based on the link metric of local interfaces, such as interface traffic load and queue size.

Load balancing based on local link quality is not always optimal. Global congestion awareness, with information from remote links, is needed for optimal balancing. Therefore, the forwarding table needs to contain not only local exit interface information but also remote path info and remote link congestion information.

Remote path info can be remote links or remote nodes, specifically as follows:

* 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.

By using remote path info, routes can be associated with remote paths.

4.3. 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 need some extensions to support such a feature. The

specific extensions can refer to [I-D.wang-idr-next-next-hop-nodes] or other extensions, which are not detailed in this document.

In an IGP-based network, a router may compute the path information based on the SPF tree and attach it to the next hop. Path info can be a link-local address, interface ID, or Link Local Identifier, or other extensions. The detailed mechanisms are out of the scope of this document.

4.4. Forwarding Plane

Taking Figure 1 as an example, the forwarding table on R3 is illustrated in Figure 4. Below is an explanation of the table’s structure.

For each prefix, the next hop and the weight corresponding to each path are recorded. If per-packet load balancing is supported, the load on the path is also recorded.

The next hop for the prefix is constructed from the local next hop and remote path information. The forwarding weight and load are determined by the quality of the local next-hop interface (local(q)) and the quality of the remote link in the remote path (remote(q)). Additionally, the load varies with the total traffic forwarded to this path during per-packet forwarding.

When responding to local congestion events, the next-hop address in the congestion event is used to find the corresponding ECMP entry, and the weight and load of this ECMP entry are modified according to the congestion level.

When responding to remote congestion events, the path info in the congestion message is used to find the corresponding ECMP entry. The link quality of the remote path is updated, and a new weight value and new load value are calculated based on the local and remote link quality. Then the weight and load of this ECMP entry are modified according to the congestion level.

In per-packet load balancing mode, each time the path with the smallest load is selected for forwarding. Each time a packet is forwarded, the load for the corresponding path is increased accordingly. When a link congestion event is received, the load for that path is increased according to the level of congestion.

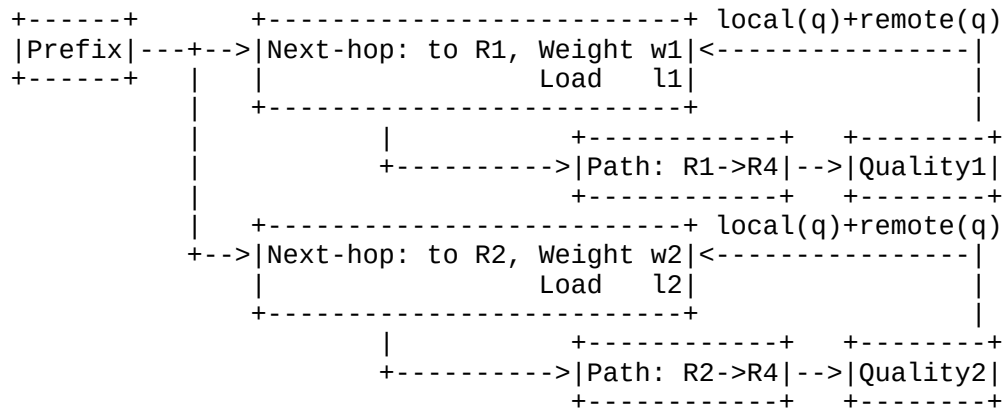

Figure 4 Forwarding table for Adaptive Routing

When the number of flows is small or when there are elephant flows, adaptive routing needs to be performed through flow redirection. The following figure 5 is a schematic of the forwarding layer flow table maintenance. The flow tables are maintained according to the five-tuple of the traffic, recording the path information corresponding to this flow.

When responding to remote flow congestion events as described in section 4.7, the flow will be rehashed to choose an ECMP path, and this flow is redirected to the least loaded ECMP path.

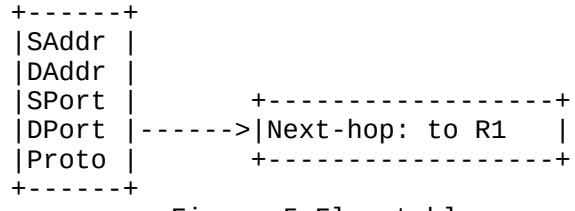

Figure 5 Flow table

4.5. Adaptive Routing Mode

Adaptive routing adjustment modes can be categorized into flow-based adjustment modes and packet-based adjustment modes. Flow-based adjustments can be further divided into link congestion-based load balancing weight adjustments and flow redirection-based adjustments. In packet-based adjustment modes, each time a packet is forwarded, the link with the lightest load is chosen for forwarding. This can lead to out-of-order reception at the receiving end, requiring the application side to handle out-of-order issues.

4.5.1. Flow-Based Adjustment Mode

In flow-based adjustment modes, the load balancing weight of links can be adjusted, or specific flows can be redirected.

Weight-Based Dynamic ECMP Flow Mode For link-level congestion events, based on the congestion status of remote links and combined with the local link congestion status, the forwarding weights of the corresponding ECMP links in the forwarding table are dynamically adjusted. This helps to achieve global link load balancing by reducing the flow weights on congested links. The forwarding weight is calculated based on the quality of the local and remote links.

Flow Redirection Mode For specific flow congestion events, the congested flow is redirected to ECMP links with lighter loads. During flow redirection, the quality, such as the remaining bandwidth, of these links must be considered to avoid congestion on the redirected link.

4.5.2. Packet-Based Adjustment Mode

Based on the congestion status of local and remote links, the load on corresponding ECMP links is dynamically adjusted. During data forwarding, using a per-packet forwarding model, each data packet is sent through the link with the lightest load. The application side must be capable of handling out-of-order packets when receiving them.

4.6. Congestion Detection

Congestion detection is generally performed by devices near the congestion point, including the detection of link congestion and congestion clearance. Network performance and congestion points can be identified by sending test traffic. A queue exceeds a threshold depth may send congestion notification. Congestion can also be inferred by monitoring the packet loss rate to determine if a link is congested. Congestion detection also includes flow-based congestion detection. Congestion Specific detection methods are beyond the scope of this document.

4.7. Congestion definition

The definition of congestion can be 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 based on current bandwidth and buffer usage, using a specific ratio.

For instance, with 16 quality levels, on a 400G interface, level 0 could represent 25G and level 15 could represent 400G.

The exact method for calculating the quality level is beyond this document's scope, but the rules must be consistent among routers exchanging this information.

4.8. Congestion Notify

When a change in congestion status is detected, it needs to be communicated to remote devices in order to adjust traffic scheduling from the source.

Congestion messages can be of two types:

1) The first type includes Path information, which helps in identifying the corresponding route for adjustments. It also includes the congestion information of the link corresponding to the Path. With this information, global congestion calculation can be performed to derive the weight information for the forwarding table. For details, refer to section 4.4.

2) The second type includes the five-tuple information of the congested flow. By using this congested flow information, congestion flow redirection can be implemented. For details, refer to sections 4.4 and 4.5.

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. However, transmitting congestion information through traditional routing protocols presents performance challenges. On one hand, congestion notifications need to be sent more frequently than routing information. On the other hand, processing congestion messages should ideally occur in the forwarding plane. Therefore, to improve the performance of adaptive routing, new protocols can be designed specifically for this purpose. Specific extensions can refer to [I-D. draft-zzhang-rtgwg-router-info] or other extensions, which are not detailed in this document. Congestion messages can be transmitted in-band or out-of-band. For high-performance solutions, additional protocols may be needed for efficient out-of-band message transmission. Specific methods are beyond the scope of this document.

5. Work Flow

Taking the network shown in Figure 1 as an example, the neighbor ID of R4 is included in the routing distribution and calculation to indicate the path from R2 to R4. The router ID can be used as the

neighbor ID. Using BGP routing as an example, explain the entire working process.

When congestion occurs on the link between R2 and R4, R2 will use R4’s neighbor ID to notify R1 of the link congestion.

Below is a description of the workflow for handling congestion information in various modes:

5.1. Weight-Based Dynamic ECMP Flow Adjustment Mode

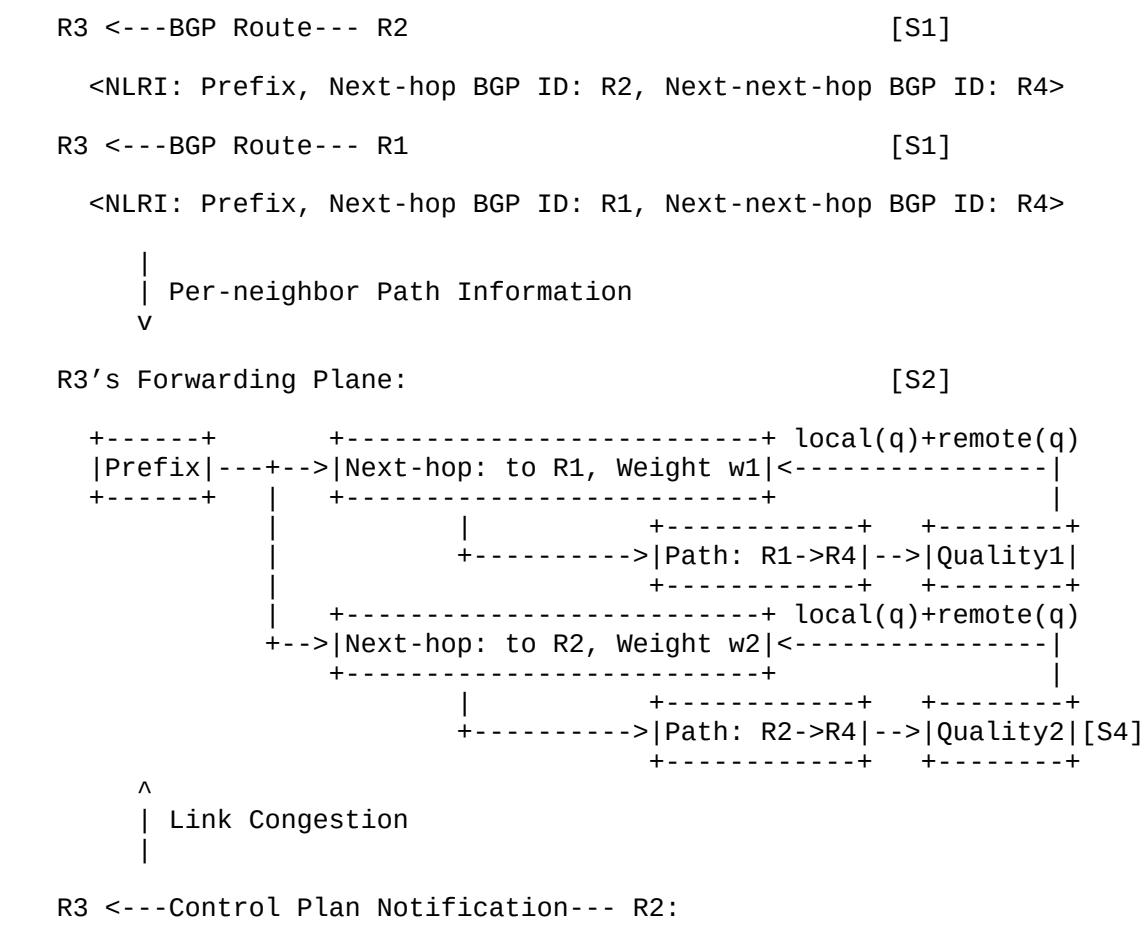

Figure 6

As shown in Figure 6, the workflow for handling remote link congestion by stream mode is as follows:

[S1]:There are two paths from R3 to R4: R3->R1->R4 and R3->R2->R4. Link information is advertised by R2 through the BGP.

When R2 delivers BGP routes to R3, 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 R4.

When R1 delivers BGP routes to R3, the NNHN Capability TLV is carried in the attributes [I-D.wang-idr-next-next-hop-nodes], indicating that the next-hop is R1 and the next-next-hop is R4.

[S2]:R3 learns the routes and maintains two ECMP paths, both with initial equal weights set to 1.

[S3]:R2 detects a change in congestion on the R2->R4 link using congestion detection methods and classifies the congestion into levels according to severity. This information, including the detecting node, the congested link, and the congestion level, is notified to R1 through the control plane [I-D. draft-zzhang-rtgwg-router-info].

[S4]:R3 receives the remote notification and, based on the congested node (R2) and next-hop information (R4), looks up its local forwarding table. It then adjusts the forwarding weights of the corresponding ECMP entries according to the congestion level, assuming the weight is adjusted to 10, as shown in Figure 7.

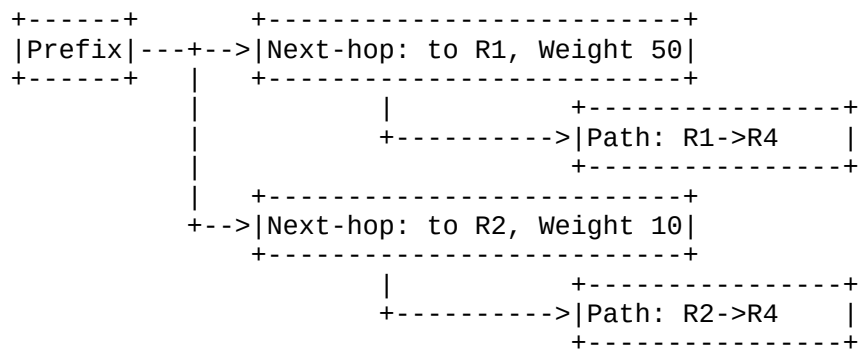

Figure 7 Adaptive forwarding table

5.2. Flow Redirect Mode

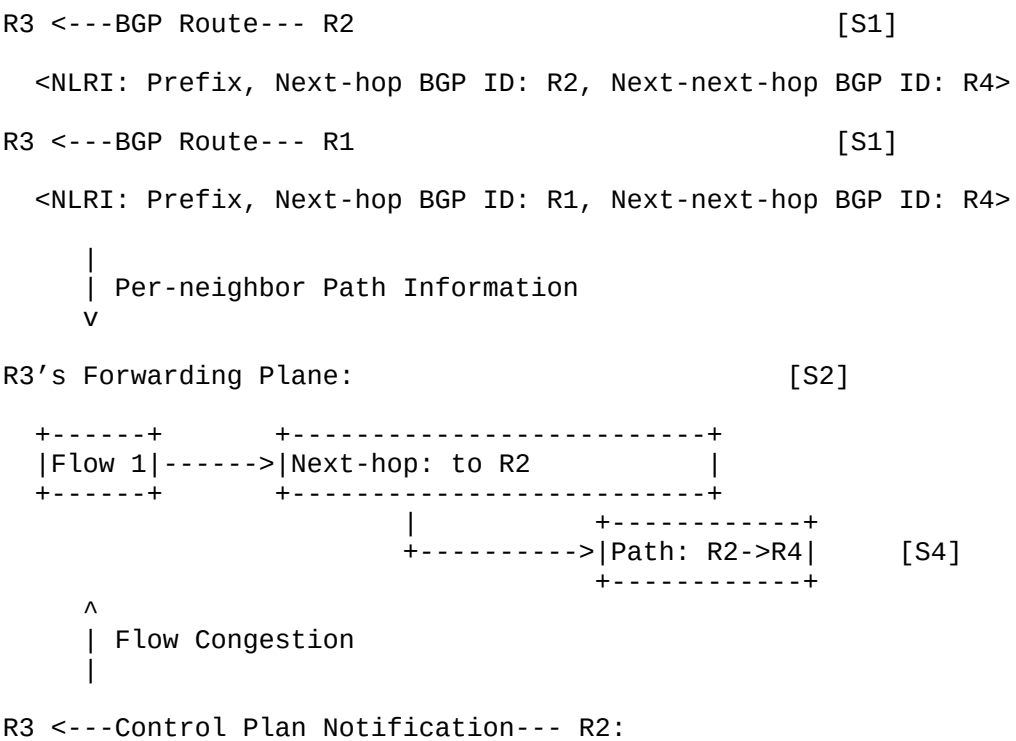

Figure 8

As shown in Figure 8, the workflow for handling remote flow congestion is as follows:

[S1]:There are two paths from R3 to R4: R3->R1->R4 and R3->R2->R4. Link information is advertised by R2 through the BGP.

When R2 delivers BGP routes to R3, 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 R4.

When R1 delivers BGP routes to R3, the NNHN Capability TLV is carried in the attributes [I-D.wang-idr-next-next-hop-nodes], indicating that the next-hop is R1 and the next-next-hop is R4.

[S2]:The initial Flow 1 selects the path R2->R4 for forwarding and establishes a flow table.

[S3]:R2 detects Flow congestion on a specific flow passing through the R3->R4 link using congestion detection methods; R2 notifies the remote device R3 of the congestion change event, including the congested path info and flow information [I-D. draft-zzhang-rtgwg-router-info];

[S4]:R3 receives the flow congestion event and looks up the flow table based on the flow information, redirecting the flow to the least loaded link among the ECMP links; Subsequently, the flow is forwarded according to the new flow table.

5.3. Packet-Based Adjustment Mode

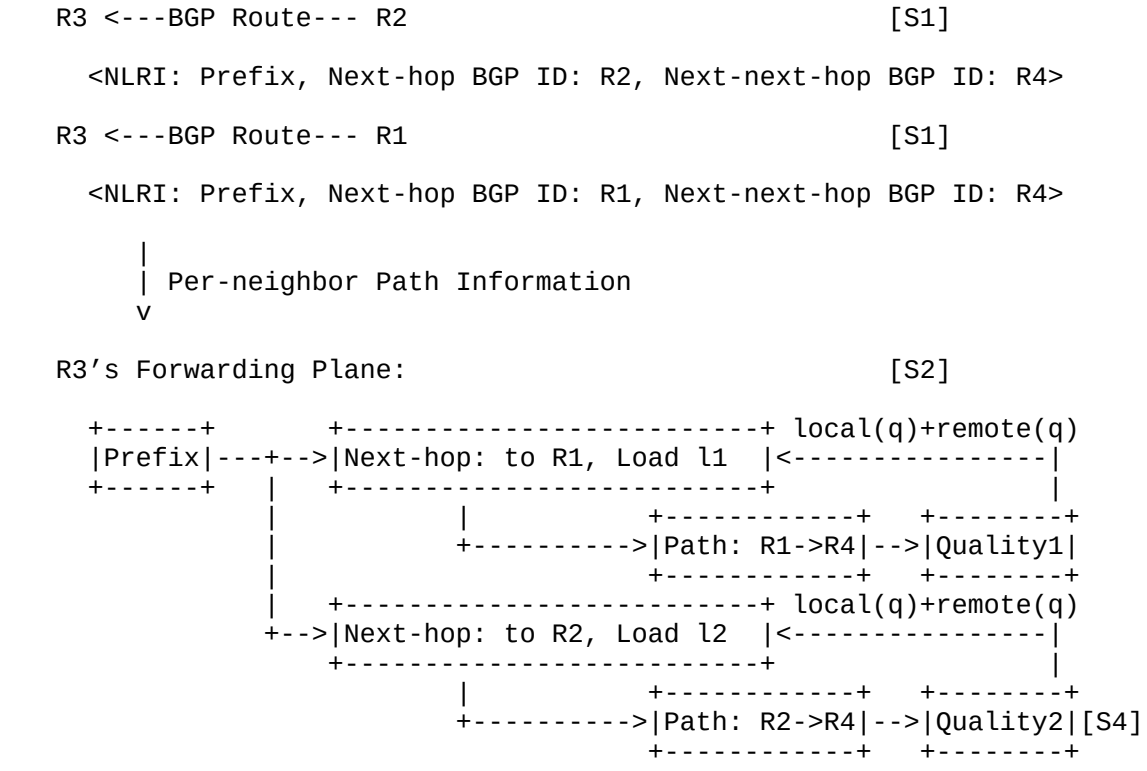

Figure 9

As shown in Figure 9, the workflow for handling remote link congestion by packet mode is as follows:

[S1]:There are two paths from R3 to R4: R3->R1->R4 and R3->R2->R4. Link information is advertised by R2 through the BGP.

When R2 delivers BGP routes to R3, 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 R4.

When R1 delivers BGP routes to R3, the NNHN Capability TLV is carried in the attributes [I-D.wang-idr-next-next-hop-nodes], indicating that the next-hop is R1 and the next-next-hop is R4.

[S2]:R3 learns the routes and maintains two ECMP paths, both with initial equal load set to 50.

[S3]:R2 detects a change in congestion on the R2->R4 link using congestion detection methods and classifies the congestion into levels according to severity. This information, including the detecting node, the congested link, and the congestion level, is notified to R1 through the control plane [I-D. draft-zzhang-rtgwg-router-info].

[S4]:R3 receives the remote notification and, based on the congested node (R2) and next-hop information (R4), looks up its local forwarding table. It then adjusts the forwarding load of the corresponding ECMP entries according to the congestion level, assuming the load is adjusted to 100, as shown in Figure 9. Afterwards, when selecting a path for per-packet forwarding, the lower load path R1->R4 will be chosen for forwarding, until the load on the path R1->R4 becomes higher than that of R2->R4.

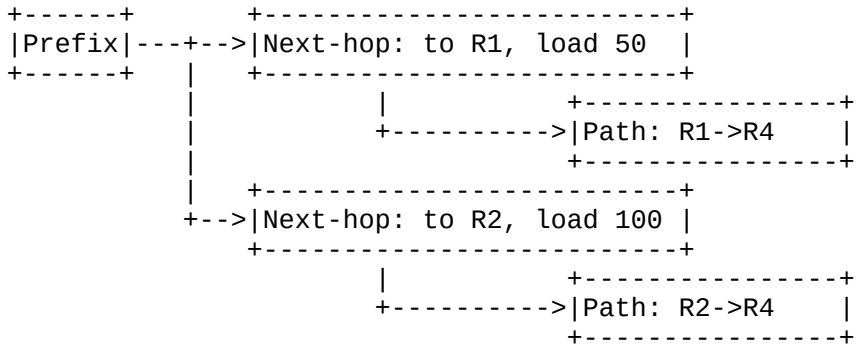

Figure 10

6. Security Considerations

TBD.

7. IANA Considerations

TBD.

8. References

8.1. 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

8.2. Informative References

[I-D.wang-idr-next-next-hop-nodes] Wang, K. , J. Haas. And Lin, C. , "BGP Next-next Hop Nodes", Work in Progress, Internet-Draft, draft-wang-idr-next-next-hop-nodes-01, 4 September 2024, <https://datatracker.ietf.org/doc/html/draft-wang-idr-next-next-hop-nodes-01>.

[I-D. draft-zzhang-rtgwg-router-info] Zhang, Z. , Wang, K. and Lin, C., " Advertising Router Information", Work in Progress, Internet-Draft, draft-zzhang-rtgwg-router-info-01, 18 September 2024, <https://datatracker.ietf.org/doc/html/draft-zzhang-rtgwg-router-info-01>.

9. Acknowledgments

TBD.

Authors' Addresses

Weiqiang Cheng
China Mobile
China
Email: chengweiqiang@chinamobile.com

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

Kevin F. Wang
Juniper Networks
Email: kfwang@juniper.net

Jiaming Ye
China Mobile
China
Email: yejiaming@chinamobile.com

Rui Zhuang
China Mobile
China
Email: zhuangruiyjy@chinamobile.com

PengFei Huo
ByteDance
China
Email: huopengfei@bytedance.com