RTGWG
Internet-Draft
Intended status: Informational
Expires: 14 December 2025

Y. Liu
H. Li
W. Duan
ZTE

12 June 2025

Adaptive Routing Notification for Load-balancing
draft-liu-rtgwg-adaptive-routing-notification-02

Abstract

In this document, adaptive routing is referred to as a technology that makes dynamic traffic forwarding decisions based on changes in traffic load and network topology, devices with adaptive routing capabilities can dynamically select the outport in the forwarding table based on the congestion condition of the outport or downstream link. This document focuses on the information carried in (Adaptive Routing Notification)ARN messages and how they are delivered and processed in the network.

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 14 December 2025.

Copyright Notice

Table of Contents

1. Introduction

The term "Adaptive Routing" has different means under different circumstances. In this document, adaptive routing is referred to as a technology that makes dynamic traffic forwarding decisions based on changes in traffic load and network topology, devices with adaptive routing capabilities can dynamically select the outport in the forwarding table based on the congestion condition of the outport or downstream link.

When congestion is detected and there's no alternate local outport available, an adaptive routing notification (ARN) message would be generated by the device and sent to the upstream node which is able performance adaptive routing, so the traffic can be removed from the original path based on ARN to relief the congestion.

Generally, the goal of the congestion control mechanism is to prevent too much data from being injected into the network to relief the congestion, the sender(the host) would adjust the packet sending strategy based on the feedback from the network, while adaptive routing(in this document) aims to instant response to the changes in the network by adjusting traffic forwarding path, the change of the path may be temporary, other mechanism such as congestion control or global adjustment by the contoller may take place later.

The local adaptive routing mechanism on the device, e.g, how to determine congestion and locate the traffic that causes congestion, is implementation specific and out of the scope of the document.

This document focuses on the information carried in ARN messages and how they are delivered into the network, which involves the interaction between devices.

Generally, the ARN mechanism is more suitable for scenarios where bandwidth utilization in the network is uneven. For the packet-spray scenario, since the packets are evenly distributed on each equal-cost path, ARN may not be the most suitable mechanism in this case. But wether to deploy this function is implementation specific, and this document does not make any restrictions on the scenarios where the ARN mechanism can be used.

2. Terminology

AR: Adaptive Routing

ARN: Adaptive Routing Notification

Flowlet: A flowlet is defined as a burst of packets from the same flow followed by an idle interval.

Traffic: The set of packets with the same traffic identifier and take the same forwarding path within a certain period of time, e.g, a flow/flowlet.

3. Specification of Requirements

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.

4. Adaptive Routing Notification

On receiving the ARN message, the upstream node requires some necessary information to operate the path of the corresponding traffic, including traffic identifier(e.g, 5 tuples), the original outport of the traffic of the upstream node, and the ARN message triggering reason(e.g, congestion).

The traffic identifier is used to locate the corresponding forwarding table entry and the original outport is the port that needs to be blocked in the forwarding table.

In order to fulfill the above requirements, the following options are discussed in the following sub-sections.

4.1. ARN Option 1

After receiving a certain traffic, the node records the traffic identifier and the receiving port of the traffic. When congestion is detected due to this traffic, the ARN message is generated for it, the node SHOULD send the ARN message to its direct-connected upstream node via the original receiving port which is recorded locally. In other words, the ARN message is returned along the original forwarding path of the traffic.

After receiving the ARN, the direct-connected upstream node would treat the receiving port of the ARN message as the original outport of the traffic, so it can block this port from the forwarding table for this traffic to change the forwarding path. If there’s no other outport which meets the forwarding requirement on the this node, it SHOULD continue to send the ARN message to its direct-connected upstream router following the same procedure as mentioned above, which means this node SHOULD record the traffic identifier and the receiving port as well after receiving the traffic.

Figure 1 shows a three-tier clos topology. The port on S7 that connects to S4 is P7-4, port on S4 that connects to S2 is P4-2, and so on.

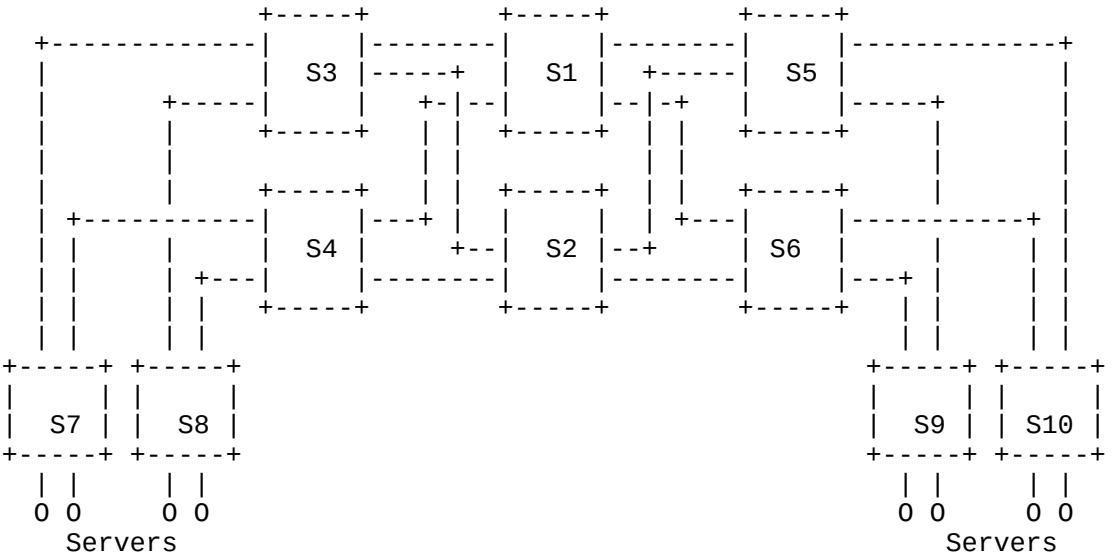

Figure 1: Three-tier Clos

Taking per-flow load-balancing as an example, at a certain time, for flow1, the forwarding path in the network is S7-S4-S2-S6-S10. After receiving flow1, S6 records the 5 tuples and the inport(i.e, P6-2) of it, and then S6 forwards flow1 to SW10 via P6-10. When congestion occurs on P6-10 due to flow1, S6 decides to change the forwarding path for it, but there is no other local port to S10 on S6, so S6 generates an ARN carrying the identifier of flow1 and sends it to S2 along the original forwarding path via P6-2. After S2 receives the ARN message from P2-6, S2 would block P2-6 from the forwarding table for flow1 and use P2-5 for forwarding. If P2-5 does not meet the forwarding requirements, S2 will generate an ARN for flow1 and send it to S4 via P2-4.

Overall, all the forwarding nodes except the headend node along the path MAY record its receiving port of the corresponding traffic in case there’s no alternate path for the traffic locally and the ARN message needs to be sent to the upstream node along the original path until there’s an upstream node can perform adaptive routing locally.

To fulfill the requirement, all the routers along the forwarding path of the traffic need to record the receiving port of the corresponding traffic. This method of storing traffic information locally on the node consumes additional local resources, especially when the amount of traffic that requires adaptive routing is not small.

4.2. ARN Option 2

Instead of storing the traffic forwarding information locally, another option is to embed the information into the packet.

After receiving the traffic, if there is more than one outport that meets the forwarding requirements(e.g, not congested), the node first choose an outport for traffic forwarding, then it embeds it’s own device identifier and the outport identifier into the packets of the traffic. The device identifier is the global identifier of the device in the network, a routable loopback address on the device is RECOMMENDED. The outport identifier is a local identifier that uniquely identifies a port on the this device.

If there’re already device identifier and outport identifier in the packet, the node SHOULD replace them with its own information, which insures that any node along the path can obtain the information of its nearest upstream node that can perform adaptive routing for the packet, as well as the outport through which the packet was originally forwarded at the this upstream node.

Once the congestion is detected by a node due to certain traffic, the node obtains identifiers from the packet belonging to the traffic to generate and send the ARN message. The outport identifier SHOULD be carried in the ARN along with the traffic identifier. And the ARN is sent directly to the device indicated by the device identifier in the original packet.

After receiving the ARN message carrying the node's own device identifier, e.g, the destination address is a local address on the node, the node would block the outport identified by the outport identifier in the ARN for the forwarding table of the corresponding traffic located by the traffic identifier.

4.3. ARN Option 3: Multicast

Sending ARN leveraging unicast like option 1 and option 2 means that when sending ARN, the sender needs to know exactly who is the expected receiver. Mulicast is another option, leaving out the requirement of the ARN receiver information. When ARN is generated, the simplest multicast mechanism is to send ARN(s) via all the active ports on the device. Afer receiving ARN, the node SHOULD check the local forwarding table for the traffic identified in the ARN, including:

* If the FIB for the traffic exists and there's other next-hop for the traffic available besides the node generating the ARN, the receiving node would switch the traffic for other outport, without further generating and sending ARN.

* If the receiving node is the headend of the traffic path, the node MUST NOT generate and send ARN.

* If the FIB for the traffic exists and there's no other next-hop, the node SHOULD further generate ARN and send it by multicast via all the ports available with the original ARN receiving port excluded.

* Otherwise, the node SHOULD ingore the ARN message without any process.

Some additional mechanism MAY be used to control the scale of the ARN messages, this would be discussed in the further version of this document.

5. ARN TAG

Regardless of the ARN option used, the implementation of ARN comes at a cost, requiring the devices to consume additional resources. In many cases, enabling the ARN mechanism for all traffic in the network is not the best solution. For example, when congestion occurs, rerouting mice flows has little effect on alleviating congestion compared with elephant flows. In addition, for some detection/ telemetry messages, their purpose is to detect the quality of the traffic path and/or find the blocking point. Therefore, the original path needs to be maintained and should not be changed even if there is congestion.

An ARN TAG is introduced in this document to control the enabling of ARN mechanism per traffic. This tag is carried in the data packet to indicate that adaptive routing is required for the traffic to which this tagged packet belongs. The specific field carrying the Tag in the packet is out of scope.

For option 1, the nodes only record the traffic identifier and receiving port of the tagged packet.

For option 2, only after receiving a tagged packet, the nodes will check whether there’s more than one outport for the packet and put the device identifier and outport identifier into it.

And in all the ARN options, the device SHOULD generate ARN only when the congestion is detected for the tagged packet.

6. IANA Considerations

This document has no IANA actions.

7. Security Considerations

TBA

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

Authors' Addresses

Yao Liu
ZTE
Nanjing
China
Email: liu.yao71@zte.com.cn

Hesong Li
ZTE
China
Email: li.hesong@zte.com.cn

Wei Duan
ZTE
China
Email: duan.wei1@zte.com.cn