RTGWG
Internet-Draft
Intended status: Standards Track
Expires: 19 April 2026

Z. Hu
Y. Zhu
J. Hu
T. Pi
China Telecom

16 October 2025

# Fast Notification for tunnel-based lossless RDMA transmission in WAN
draft-hzh-fantel-wan-tunnel-01

## Abstract

With the rapid development of Large Language Models (LLMs), many emerging AI services require lossless transmission of RDMA packets over tunnels in Wide Area Network(WAN). To meet the stringent performance demands of these services, WAN should support the real-time network state notification to ensure high throughput, low latency, and zero packet loss. The current reactive notification mechanisms are limited by responsiveness, coverage, and operational efficiency. Therefore, a faster and proactive notification mechanism is needed to enable more responsive Traffic Engineering (TE) and Load Balancing (LB).

This draft describes typical scenarios for transmitting RDMA packets over WAN tunnels, specifies the fast notification framework to support key TE areas (e.g., congestion control, protection, and load balance), and defines the packet format for fast notification.

## 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 19 April 2026.

Table of Contents



## 1. Introduction

In use cases such as distributed LLMs training or inference, WAN needs to support the tunneling of RDMA traffic between data centers (DCs). RDMA is a widely used technology in high-performance computing and AI clusters, achieving low latency, reduced CPU overhead, and high network throughput. Currently, mainstream RDMA protocols (e.g., IB, RoCE) are based on the Go-Back-N mechanism,

where a small number of packet losses can result in a dramatic reduction in the effective throughput. Therefore, WAN requires a flow control mechanism that can timely awareness and adaptive response to network state changes.

[I-D.geng-fantel-fantel-gap-analysis] points existing mechanisms for flow control often lack responsiveness and scalability. ECN[RFC3168] is a widely deployed congestion control mechanism, which enables a forwarding element to notify the sender for congestion control without having to drop packets. When a router detects congestion, it marks the packets with an ECN code-point in the IP header. The receiver, upon receiving marked packets, sends a Congestion Notification Packet (CNP) to the sender, which then temporarily reduces its transmission rate until the path can accommodate higher traffic. ECN still relies on end-to-end signaling, making real-time feedback challenging in long-distance WAN.

To enable lossless data transmission, some drafts are proposed to support FAst Notification for Traffic Engineering and Load balancing (FANTEL). [I-D.wh-rtgwg-adaptive-routing-arn] proposes a proactive notification mechanism ARN for adaptive routing, and describes the information carried in ARN to notify remote nodes for re-routing. This draft proposes a unified mechanism for congestion notifications, link failure notifications, and even to convey other relevant network events for re-routing. [I-D.liu-rtgwg-adaptive-routing-notification] describes the mechanisms of delivering ARN message. This draft gives three options, each of which specifies the information carried in the ARN message and the mechanism of sending the message to specific network nodes. However, the mechanisms described in these drafts are not specific to tunnel-based WAN deployments.

This document specifies the FANTEL mechanism for scenarios where service traffic is carried over tunnels in WAN. It first introduces the typical scenarios of distributed lossless network, then specifies the mechanisms of FANTEL to achieve key TE areas such as congestion control, load balancing, and failure protection, and finally defines the protocol implementation.

## 2. Conventions

### 2.1. Abbreviations

CNP: Congestion Notification Packet

ECN: Explicit Congestion Notification

FANTEL: FAst Notification for Traffic Engineering and Load balancing

PFC: Priority-based Flow Control

RoCEv2: RDMA over Converged Ethernet version 2

WAN: Wide Area Network

## 2.2. 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.

# 3. Scenarios

## 3.1. Scenario 1: distributed model training across DCs

The growth of computing power of a single DC is limited by space and power supply, making it difficult to meet the fast-growing computing resources demands of LLMs. Therefore, distributed model training across multiple DCs provides a more efficient and cost-effective solution to aggregate computing resources. In this solution, a large volume of training parameter must be rapidly synchronized over WAN.

## 3.2. Scenario 2: distributed model inference between on-premise DC and third-party DC

Some customers deploy LLMs by building on-premises AI facilities, but as inference concurrency increases, scaling out these facilities requires significant investment. To address this, distributed model inference between customer on-premise DC and third-party DC provides a more agile and cost-effective solution to scale computing resource elasticly. In this solution, a large volume of inference parameter must be rapidly synchronized over WAN.

## 3.3. Scenario abstraction

In the above scenarios, parameter data between DCs need to be synchronized using RDMA protocol. Therefore, operators prefer to carry such RDMA traffic over tunnels across the WAN, ensuring efficient and lossless transmission. The framework for lossless RDMA data transmission over WAN tunnels is as follows:

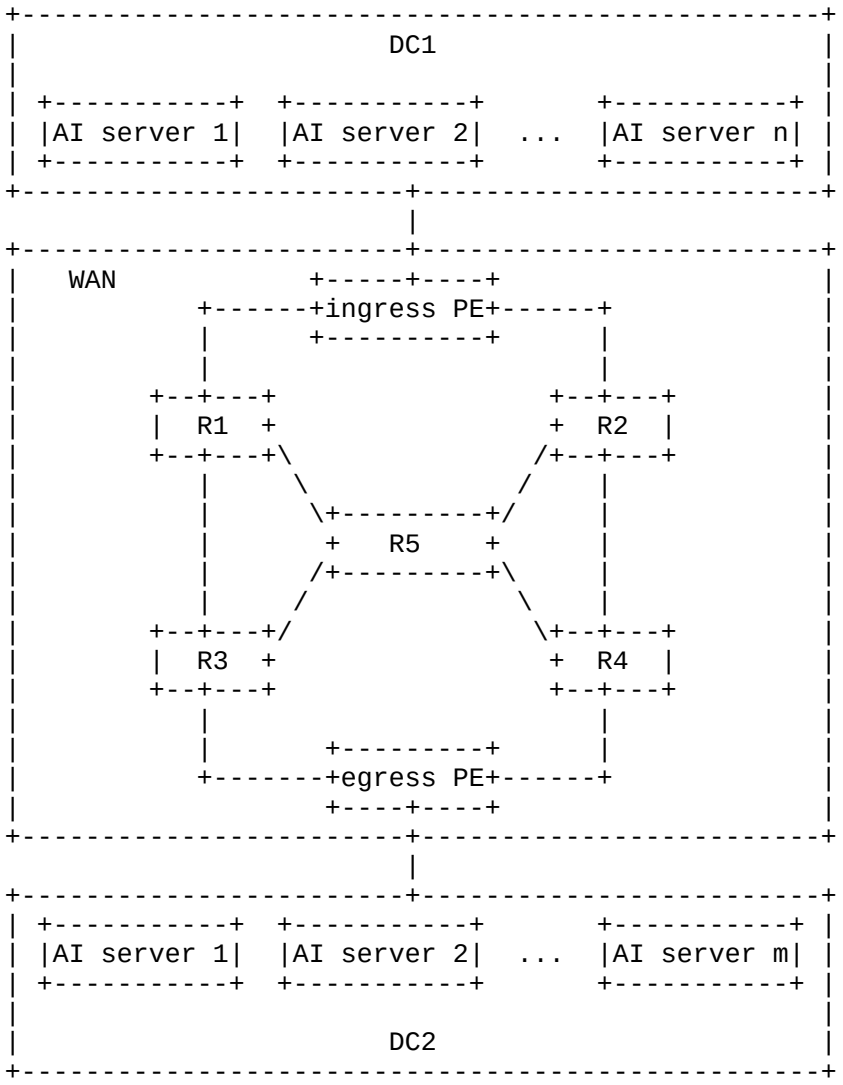


Figure 1: Network diagram

* The AI servers in DC1 sends RDMA traffic to WAN’s ingress PE.

* At the WAN’s ingress PE, the RDMA traffic is encapsulated according to the tunnel protocol and forwarded across WAN to egress PE.

* The WAN’s P node(R1-R5) transits the payload from ingress PE to egress PE via tunnels.

* At the WAN’s egress PE, the payload are decapsulated to RDMA packets and transmitted to the AI servers in DC2.

## 4. Process analyze

Tunneling technologies include various protocols, such as GRE, VXLAN, MPLS, and SRv6. AI traffic is characterized by high volume and high burstiness, making it prone to cause network congestion. Operators must adopt tunneling technologies that provide strict TE guarantees (process analyze herein is also based on the assumption of a strict TE environment). When transmittig RDMA traffic over tunnels, WAN should support FANTEL capability to realize rapid response to network conditions. Specifically, WAN devices should support fast notification mechanism to imporve three key TE scenarios: failure protection, flow control, and load balancing.

### 4.1. Failure protection

For large-scale and dynamic networks, protection mechanisms need to ensure service continuity in case of failures. According to [I-D.geng-fantel-fantel-gap-analysis], existing failure handling methods, such as BFD and FRR, lack flexibility and responsiveness in complex typologies. Therefore, WAN should support fast notification for failures, allowing near-instantaneous and dynamic protection responses, minimizing failure impact.

Upon network failure, the ingress PE should immediately adapt its forwarding policy to steer traffic away from faulty links or nodes. Therefore, the fast-notification-based failure protection process is as follows:

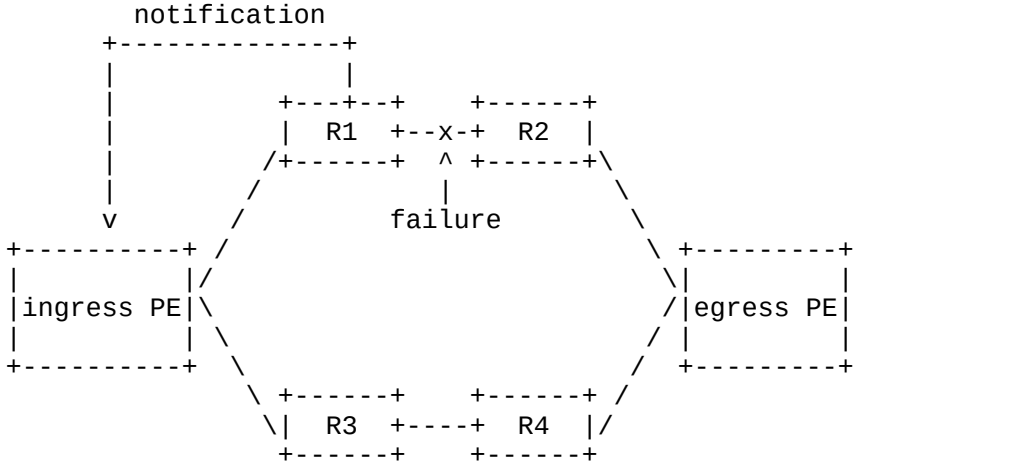


Figure 2: Failure protection procession

* When a P node detects a local link/node failure, it collects failure information about the affected link or flow.

* The P node sends notification to ingress PE with failure information (In addition to the identity of the failed link or node, the notification must also include information about the affected traffic).

* Ingress PE receives the notification and reroutes the traffic based on its content to exclude the failed link or node: *If backup path is available, ingress PE should switch the service traffic to the backup path. *If multiple feasible paths exist, ingress PE should updates its load-balancing policy to utilize all available paths. *If no feasible path is available, ingress PE should send a corresponding notification to the sender and controller.

## 4.2. Congestion control

RDMA traffic is bursty and highly sensitive to packet loss, and WAN require proactive congestion control mechanisms. [RFC6040] redefines how the explicit congestion notification (ECN) field of the IP header should be constructed on entry to and exit from any IP-in-IP tunnel, in order to achieve ECN-based congestion control across WANs between DCs. However, [I-D.geng-fantel-fantel-gap-analysis] analysis that ECN/TCP methods still relies on end-to-end signaling and lacks precise real-time feedback.

Currently, PFC is widely used in data centers to prevent data loss due to congestion. PFC uses a step-by-step back-pressure mechanism to control the upstream to stop or continue transmitting traffic. PFC achieves link-layer priority-based traffic control, but still faces problems such as queue head blocking and deadlock due to coarse control granularity.

When network congestion occurs, the ingress PE should immediately adapt its forwarding policy to reduce the traffic sent to congested nodes. Meanwhile, the upstream nodes to the congested node should reduce the transmission rate of corresponding traffic to minimize the likelihood of packet loss. Therefore, the fast-notification-based congestion control process is as follows:

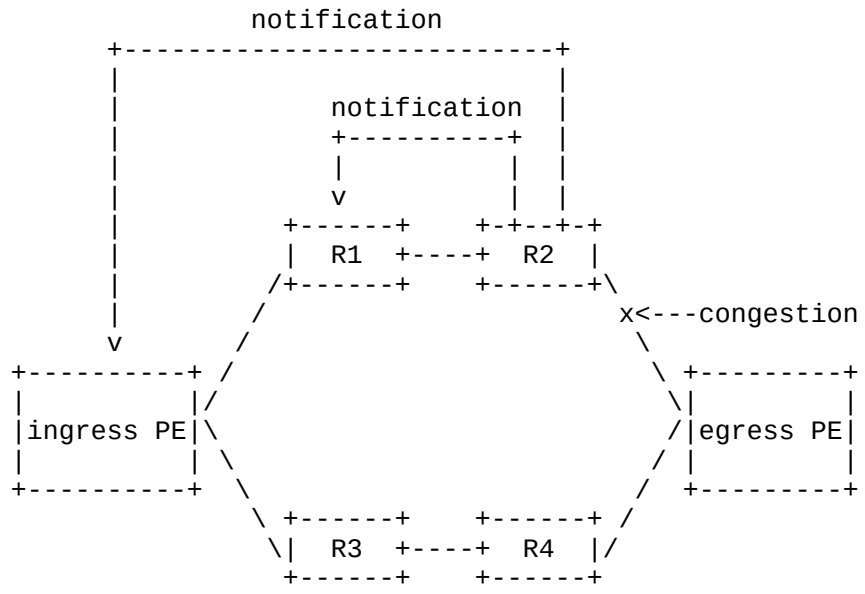


Figure 3: Congestion control procession

* when a P node detects congestion, it collects congestion information about the congested link or flow.

* The P node sends notification to ingress PE and upstream with congestion information.

* The upstream P node receives the notification and reduce the transmission rate of corresponding traffic.

* Ingress PE receives the notification and reroutes the traffic based on its content to exclude the congestion link: *If backup path is available, ingress PE should switch the service traffic to the backup path.  *If multiple feasible paths exist, ingress PE should updates its load-balancing policy to utilize all available paths.  *If no feasible path is available, ingress PE should reduce the transmission rate of corresponding traffic, and send notification to sender and controller.

## 4.3. Load balancing for network state changes

Devices and links in WAN often carry multiple services simultaneously.  In addition to failure and congestion, dynamic load balancing based on network state changes can effectively improve network resource utilization.

When significant changes occur in the network state, the ingress PE should dynamically adjust its forwarding strategy to maximize network resource utilization. Therefore, the fast-notification-based load balancing process is as follows:

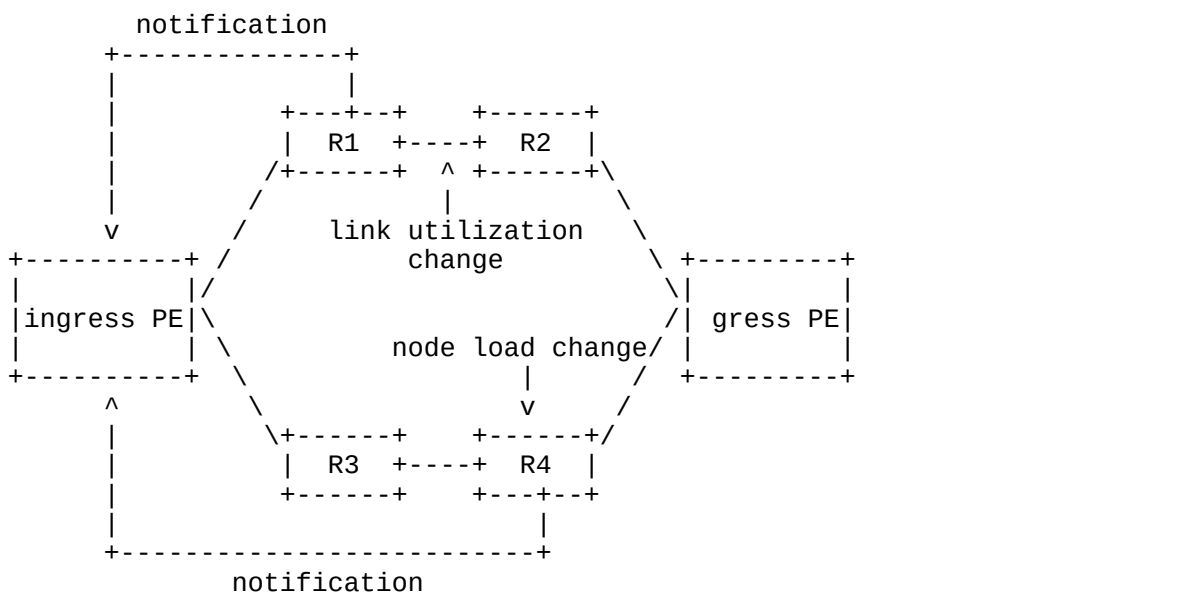


Figure 4: Load balancing for network state changes

* When a node detects the network state change, it collects the network state change information, such as link utilization, queue buildup.

* The node sends fast notification to the ingress PE with information about the network state change.

* Ingress PE receives the fast notification and updates its load-balancing policy to maximize the utilization of network resources.

## 5. Solutions

Based on the framework analysis of fast notification in key TE areas, a unified protocol implementation for fast notification should be established, with explicit forwarding procedures to realize tunnel-based lossless transmission of RDMA packets in WAN.

### 5.1. ICMPv6-based solution

The source quench mechanism has been deprecated in ICMPv6 because TCP’s built-in congestion avoidance algorithms are more efficient, and source quench may interfere with their normal operation. However, when transmitting RDMA data over WAN tunnels, the source quench notification is confined within the WAN domain (this message is used by WAN devices such as Ingress PE or transit node for traffic engineering) and does not affect transport layer congestion control.

This document specifies a new ICMP message to realize rapid notification in key traffic engineering areas including failure protection, congestion control, and load balancing. The message format is defined as follows:

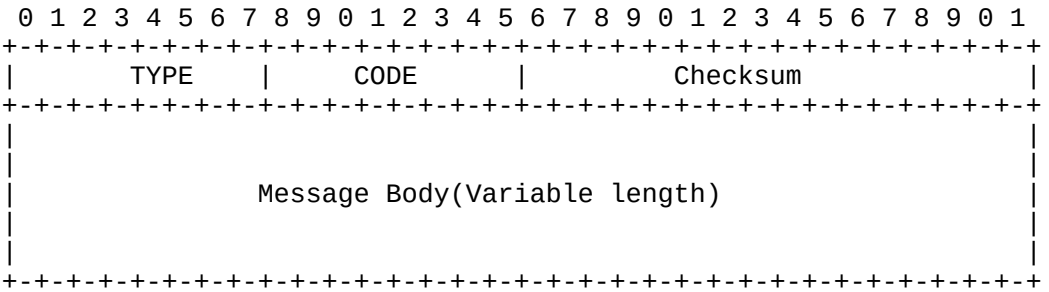


Figure 5: new ICMPv6 message for fast notification

*TYPE:8-bit identifier for the purposes of notification, When it is set to 1, it indicates the fast notification for failure protection; When it is set to 2, it indicates the fast notification for failure elimination; When it is set to 3, it indicates the fast notification for congestion control; When it is set to 4, it indicates the fast notification for congestion elimination; When it is set to 5, it indicates the fast notification for load balancing. Other bits are not defined.

*CODE: This field is an 8-bit bitmap that specifies which parameters are included in the message body of the packet.

*Checksum: Used for error-checking the packet.

*Message Body: It carries notification information specific to each areas: for failure protection, it includes path, five-tuple of flow, and failure cause; for congestion control, it contains path and buffer status; for load balancing, it comprises link utilization and device load. This field format need to be designed with extensibility, while subsequent refinements and specific packet forwarding mechanisms(TBD).

## 5.2. UDP-based solution

This document specifies a new UDP message to realize rapid notification in key traffic engineering areas including failure protection, congestion control, and load balancing. The message format is defined as follows:

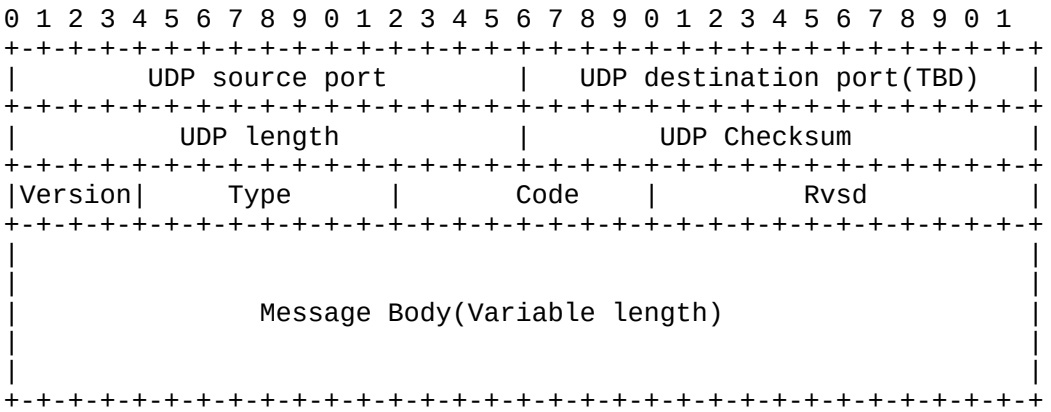


Figure 6: new UDP message for fast notification

Version: This field indicates the version number. The default value is 0.

*TYPE:8-bit identifier for the purposes of notification, When it is set to 1, it indicates the fast notification for failure protection; When it is set to 2, it indicates the fast notification for failure elimination; When it is set to 3, it indicates the fast notification for congestion control; When it is set to 4, it indicates the fast notification for congestion elimination; When it is set to 5, it indicates the fast notification for load balancing. Other bits are not defined.

*CODE: This field is an 8-bit bitmap that specifies which parameters are included in the message body of the packet.

Rvsd:Reserved

*Message Body: It carries notification information specific to each areas: for failure protection, it includes path, five-tuple of flow, and failure cause; for congestion control, it contains path and buffer status; for load balancing, it comprises link utilization and device load. This field format need to be designed with extensibility, while subsequent refinements and specific packet forwarding mechanisms(TBD).

## 6. Security Considerations

TBD

## 7. IANA Considerations

TBD

## 8. Acknowledgments

TBD

## 9. References

### 9.1. 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>.

[RFC3688] Mealling, M., "The IETF XML Registry", BCP 81, RFC 3688, DOI 10.17487/RFC3688, January 2004, <https://www.rfc-editor.org/info/rfc3688>.

[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>.

### 9.2. Informative References

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <https://www.rfc-editor.org/info/rfc3168>.

[RFC6040] Briscoe, B., "Tunnelling of Explicit Congestion Notification", RFC 6040, DOI 10.17487/RFC6040, November 2010, <https://www.rfc-editor.org/info/rfc6040>.

[RFC7514] Luckie, M., "Really Explicit Congestion Notification (RECN)", RFC 7514, DOI 10.17487/RFC7514, April 2015, <https://www.rfc-editor.org/info/rfc7514>.

[I-D.wh-rtgwg-adaptive-routing-arn]
Wang, H., Huang, H., Geng, X., Xu, X., and Y. Xia, "Adaptive Routing Notification", Work in Progress, Internet-Draft, draft-wh-rtgwg-adaptive-routing-arn-03, 13 September 2024, <https://datatracker.ietf.org/doc/html/draft-wh-rtgwg-adaptive-routing-arn-03>.

[I-D.liu-rtgwg-adaptive-routing-notification]
Liu, Y., lihesong, and W. Duan, "Adaptive Routing Notification for Load-balancing", Work in Progress, Internet-Draft, draft-liu-rtgwg-adaptive-routing-notification-02, 12 June 2025, <https://datatracker.ietf.org/doc/html/draft-liu-rtgwg-adaptive-routing-notification-02>.

[I-D.xiao-rtgwg-rocev2-fast-cnp]
Min, X. and lihesong, "Fast Congestion Notification Packet (CNP) in RoCEv2 Networks", Work in Progress, Internet-Draft, draft-xiao-rtgwg-rocev2-fast-cnp-03, 9 June 2025, <https://datatracker.ietf.org/doc/html/draft-xiao-rtgwg-rocev2-fast-cnp-03>.

[I-D.geng-fantel-fantel-gap-analysis]
Geng, X., Huo, P., Cheng, W., Li, D., Zhu, Y., and H. Zhengxin, "Gap Analysis of Fast Notification for Traffic Engineering and Load Balancing", Work in Progress, Internet-Draft, draft-geng-fantel-fantel-gap-analysis-01, 7 July 2025, <https://datatracker.ietf.org/doc/html/draft-geng-fantel-fantel-gap-analysis-01>.

## Authors' Addresses

Zehua Hu
China Telecom
Guangzhou
China
Email: huzh2@chinatelecom.cn

Yongqing Zhu
China Telecom
Guangzhou
China
Email: zhuyq8@chinatelecom.cn

Jiayuan Hu
China Telecom
Guangzhou
China
Email: hujy5@chinatelecom.cn

Tanxin Pi
China Telecom
Guangzhou
China
Email: pitx1@chinatelecom.cn