RTGWG
Internet-Draft
Intended status: Standards Track
Expires: 2 September 2026

Z. Hu
Y. Zhu
J. Hu
T. Pi
China Telecom
1 March 2026

Fast Notification for tunnel-based lossless RDMA transmission in WAN
draft-hzh-fantel-wan-tunnel-02

Abstract

With the rapid development of Large Language Models (LLMs), many emerging AI services require lossless transmission of RDMA traffic over tunnels in Wide Area Network(WAN). Existing network mechanisms were not designed for the responsiveness and scale required by these dynamic services. WAN should support the real-time, lightweight network notification to enhance the responsiveness for traffic engineering, congestion mitigation, and failure protection.

This document analyzes typical scenarios where RDMA traffic need to be tunneled across WAN, and proposes fast network notification solutions based on ICMPv6 or UDP.

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 2 September 2026.

Copyright Notice

Table of Contents

1. Introduction 2
2. Conventions 3
2.1. Abbreviations 3
2.2. Requirements Language 4
3. Scenarios 4
3.1. Scenario 1: distributed model training across DCs 4
3.2. Scenario 2: distributed model inference between on-premise and third-party DC 4
3.3. Scenario abstraction 4
4. Process analyze 6
4.1. Failure protection 6
4.2. Congestion control 7
4.3. Load balancing for network state changes 8
5. Solutions 9
5.1. ICMPv6-based solution 9
5.2. UDP-based solution 11
6. Security Considerations 12
7. IANA Considerations 13
8. Acknowledgments 13
9. References 13
9.1. Normative References 13
9.2. Informative References 13
Authors' Addresses 14

1. Introduction

For modern AI services 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) operate over best-effort forwarding, where a small number of packet losses can result in a dramatic reduction in the effective throughput. Therefore, WAN requires the FAst Notification for Traffic Engineering and Load balancing (FANTEL) to ensure reliable and congestion-free data transfer.

[I-D.geng-fantel-fantel-gap-analysis] points existing TE mechanisms face limitations in responsiveness, coverage, and operational overhead, especially in high-speed, large-scale environments. 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. But it still relies on end-to-end signaling, making real-time feedback challenging in long-distance WAN. BFD[RFC5880] is designed for rapid fault detection by sending frequent control packets between peers, but higher probe frequency increases CPU and bandwidth usage, make it struggles to balance detection speed with system overhead.

[I-D.ietf-rtgwg-net-notif-ps] is an IETF Problem Statement for FANTEL, based on the analysis of gaps in current network mechanisms and the operational requirements of modern applications (e.g., AI/ML training), formally defines the scope and core requirements for fast network notifications. Moreover, it futher specifies what information such notifications carry, who the intended recipients are, how they should be delivered, and what kinds of timely actions they may enable.

To enable lossless data transmission, some drafts are proposed to support 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. [I-D.liu-rtgwg-adaptive-routing-notification] describes the mechanisms of delivering ARN message.

This document specifies the FANTEL mechanism for scenarios where service traffic is carried over tunnels in WAN. It first introduces the typical scenarios, then specifies the process of fast notification 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 training. Therefore, distributed model training across multiple DCs provides a more efficient and cost-effective solution to aggregate computing resources. In this scenario, TB-scale training parameters need to be rapidly synchronized over WAN.

3.2. Scenario 2: distributed model inference between on-premise 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 and third-party DC provides a more agile and cost-effective solution. In this scenario, data such as the KV cache and model parameters need to be rapidly synchronized over WAN.

3.3. Scenario abstraction

In the above scenarios, a large volume of data between DCs need to be synchronized using RDMA protocol. RDMA traffic generated by LLM training or inference is highly concurrent, bursty, and extremely latency-sensitive. Therefore, operators typically encapsulate it in tunnels over the WAN to enable flexible steering and end-to-end service isolation. In these scenarios, the framework for RDMA traffic transmission over WAN tunnels is as follows:

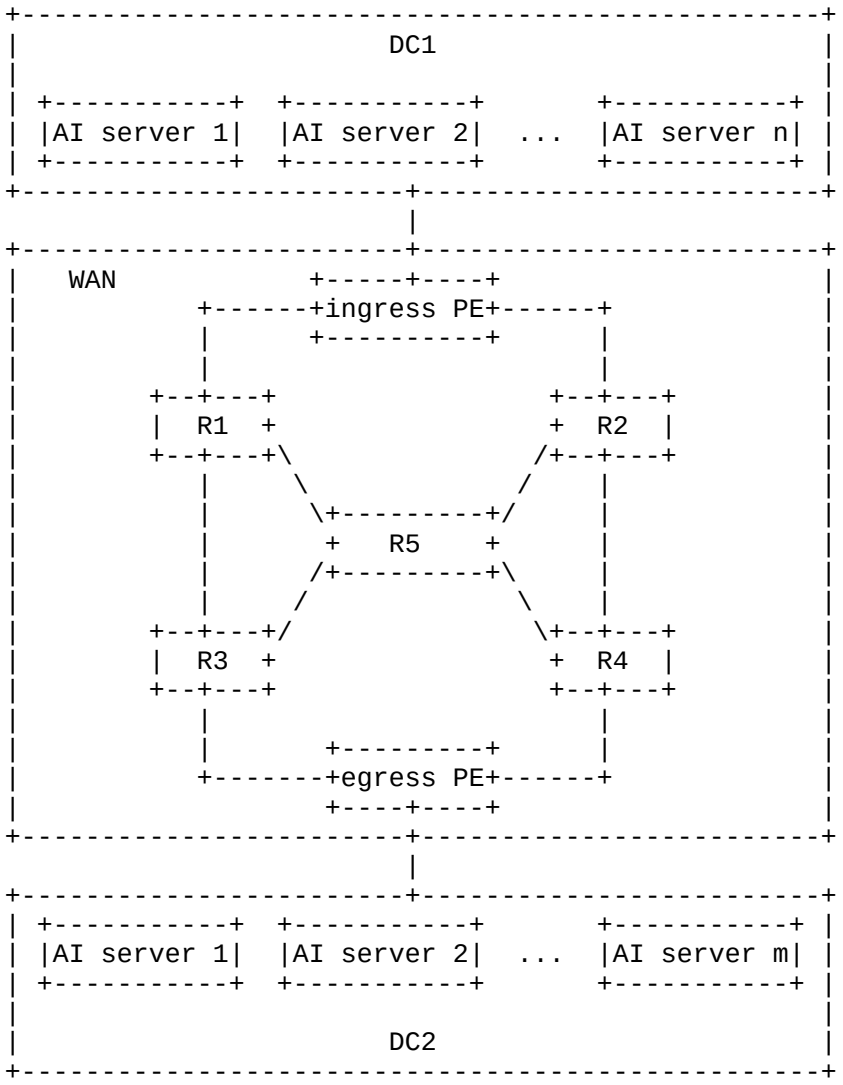

Figure 1: Network diagram

* 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. Moreover, AI workloads are highly sensitive to packet loss, latency and throughput. Network failures, congestion or underutilization can all lead to significant waste of compute resources. 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:

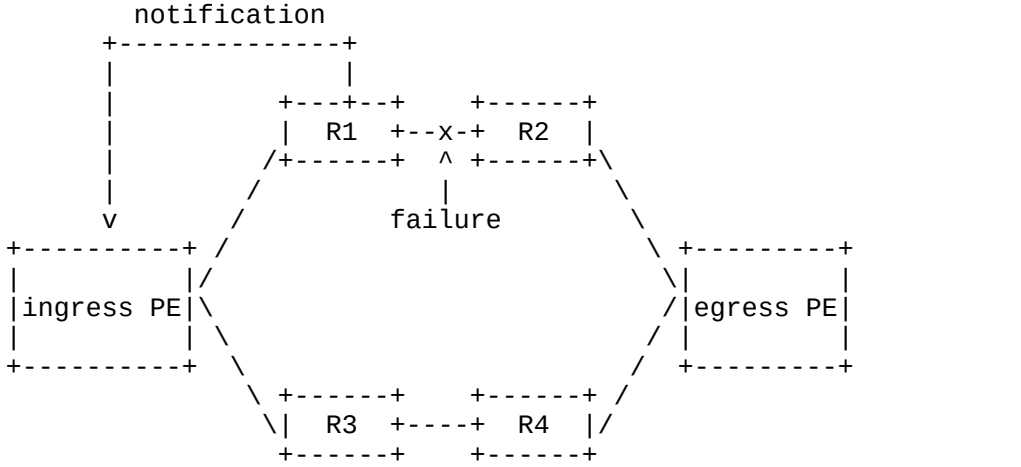

Figure 2: Failure protection procession

* 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:

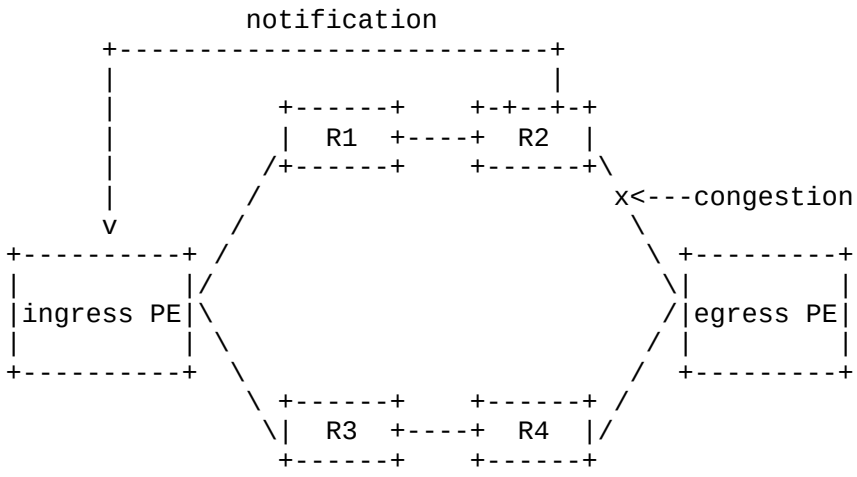

Figure 3: Congestion control procession

* 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:

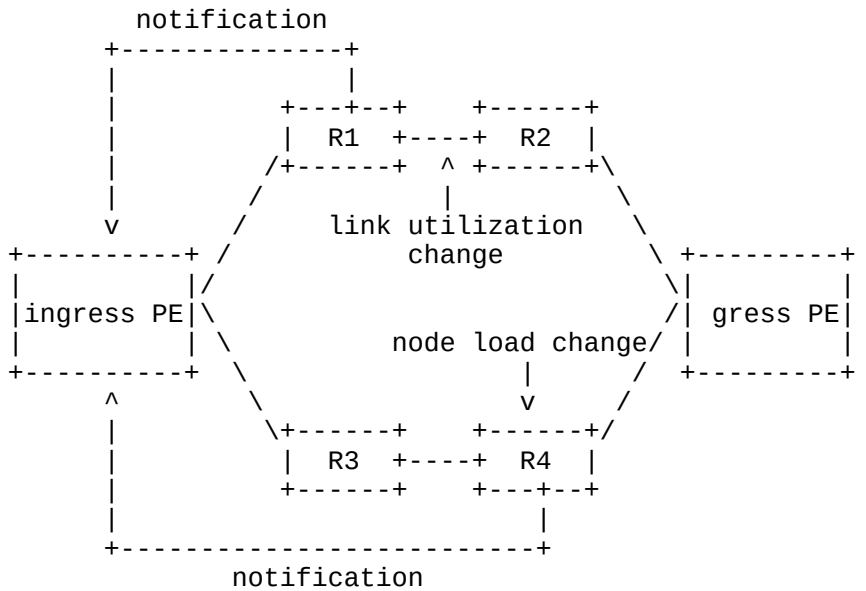

Figure 4: Load balancing for network state changes

* 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 ICMPv6 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:

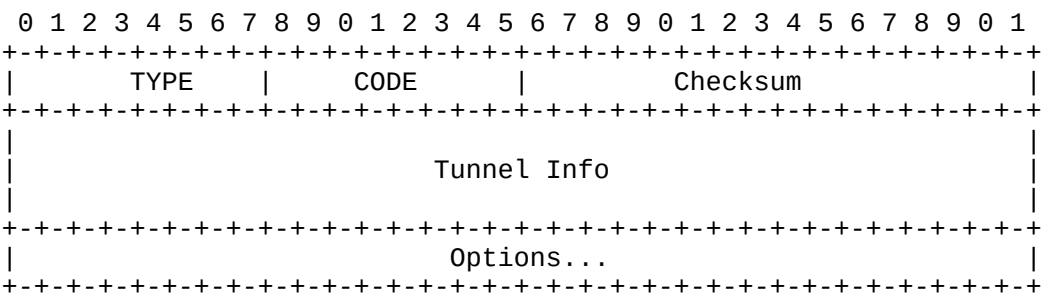

Figure 5: new ICMPv6 message for fast notification

TYPE:8-bit identifier for the purposes of notification. This document defines the following five TYPEs:

TYPE	description
TYPE1(TBA)	notification for failure
TYPE2(TBA)	notification for failure recovery
TYPE3(TBA)	notification for congestion
TYPE4(TBA)	notification for congestion elimination
TYPE5(TBA)	notification for load balancing

Table 1: description for TYPEs

CODE: This field is an 8-bit bitmap (bit 0 - 7), indicating which tunneling mechanism(s) are encoded in the Tunnel Info field. * If bit 0 is set to 1, the packet is forwarded based on IPv4 or IPv6 destination address lookup. This applies to IP-based tunnels such as GRE and IPsec. In this case, the Tunnel Info field contains the destination IPv4 or IPv6 address. * If bit 1 is set to 1, the packet is forwarded according to an SRv6 Policy. In this case, the Tunnel Info field contains a Segment Routing Header (SRH) as defined in RFC 8754. * If bit 2 is set to 1, the packet is forwarded using MPLS switching. In this case, the Tunnel Info field contains an MPLS shim header as defined in RFC 3032. Bits 3 through 7 are reserved for future use and MUST be set to 0 by the sender and ignored by the receiver.

Checksum: Used for error-checking the packet.

Tunnel Info: This field carries tunnel-specific information (e.g., destination address, SRH, or MPLS shim header) required by the recipient node to identify and divert traffic away from the affected path. Upon receiving a message containing this field, the recipient node SHALL cease forwarding traffic along the specified path.

options: A TLV-encoded optional field that conveys additional telemetry, traffic, or network state information to support fine-grained flow control. The TLV format is shown in Figure 6. It MAY include the following categories: * Event Location: Identifies where the triggering event occurred — e.g., the affected link or node. * Flow Information: Describes the traffic flow impacted by the event — e.g., its identity (such as a Flow ID) or transport 5-tuple. * State Metrics: Quantifiable network measurements — e.g., link utilization, queue length, one-way delay, or packet loss rate. The format of TLVs conveying link identifiers, node identifiers, and network metrics SHALL follow the corresponding definitions in BGP-LS [RFC7752] and the BGP-LS extensions [RFC8571].

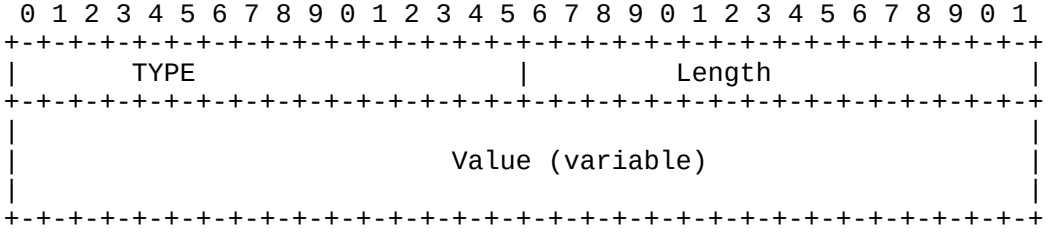

Figure 6: TLV Format

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:

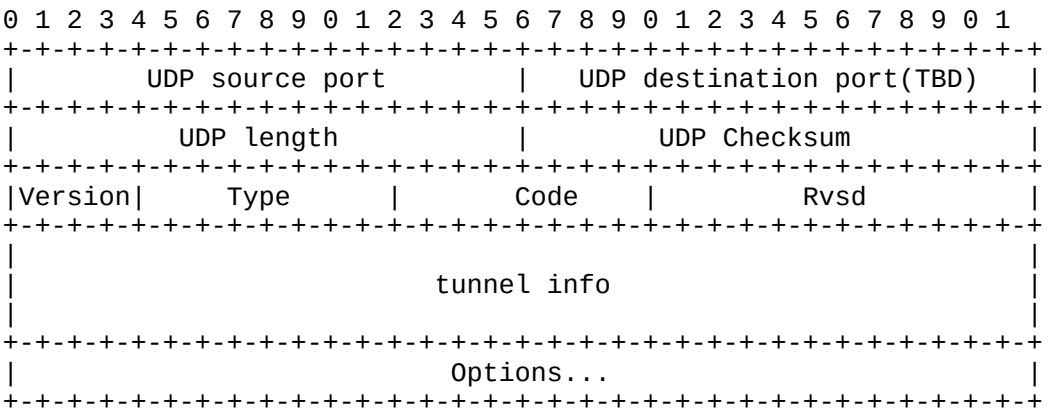

Figure 6: new UDP message for fast notification

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

The definitions of the TYPE, CODE, tunnel info, and Options fields are the same as those in Section 5.1.

Rvsd:Reserved

6. Security Considerations

This document specifies Fast Notification (FANTEL) mechanisms for tunnel-based lossless RDMA transmission in WAN, using ICMPv6 and UDP as transport protocols. While these protocols are widely deployed and well-understood, extending them with new notification semantics introduces potential security considerations that must be addressed.

Implementations MUST enforce the rate limiting behavior specified in RFC 4443 [RFC4443] §2.4 for all ICMPv6 messages carrying FANTEL information.

The TLV parser MUST validate that the sum of all TLV Length fields does not exceed the total ICMPv6 payload length. Any packet failing this check MUST be silently discarded.

All FANTEL notifications MUST be sent from a control-plane interface of the originating node (e.g., a loopback interface configured for management), and MUST NOT originate from data-plane forwarding interfaces (e.g., physical ports carrying customer traffic). This ensures that FANTEL traffic cannot be injected by compromised customer devices.

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>.

[RFC4443] Gupta, Mukesh., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", RFC 4443, DOI 10.17487/RFC4443, March 2006, <https://www.rfc-editor.org/info/rfc4443>.

[RFC5880] Katz, Dave., "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, January 2010, <https://www.rfc-editor.org/info/rfc5880>.

[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>.

[I-D.ietf-rtgwg-net-notif-ps]
Dong, J., McBride, M., Clad, F., Zhang, Z. J., Zhu, Y., Xu, X., Zhuang, R., Pang, R., Lu, H., Liu, Y., Contreras, L. M., Mehmet, D., and R. Rahman, "Fast Network Notifications Problem Statement", Work in Progress, Internet-Draft, draft-ietf-rtgwg-net-notif-ps-00, 11 February 2026, <https://datatracker.ietf.org/doc/html/draft-ietf-rtgwg-net-notif-ps-00>.

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