Routing Area Working Group
Internet-Draft
Intended status: Informational
Expires: 23 April 2026

Jiayuan. Hu, Ed.
China Telecom

20 October 2025

Precise ECN in WAN
draft-hu-rtgwg-pre-ecn-wan-00

Abstract

This draft defines the precise ECN during used in WAN. With the growing demand for AI computing power, the computational capacity of a single Artificial Intelligence Data Center (AIDC) can no longer meet the requirements of large-scale model training. This has led to the emergence of cross-AIDC distributed model training, driving the need for transmitting RoCEv2 packets over WAN networks. AI training is highly sensitive to network packet loss, where even minimal packet loss can significantly degrade training efficiency. Additionally, elephant flows and extreme concurrent traffic impose higher demands on network performance.

ECN achieves active feedback of network congestion by setting ECN flag bits in the header of IP packets, which is an effective traffic control method. RFC6040 introduces the application of ECN in WAN. However, due to the much higher end-to-end delay in WAN than in DC, and the frequent occurrence of instantaneous traffic bursts in WAN, it is easy to trigger ECN at the wrong time. This draft focuses on the precise use of ECN in WAN, by introducing different reactions of ECN in different WAN transmission scenarios

Table of Contents

1. Introduction

The rapid growth of AI computing power, particularly for large-scale model training, has necessitated distributed training across multiple Artificial Intelligence Data Centers (AIDCs). This shift has increased the demand for reliable and high-performance transmission of RoCEv2 (RDMA over Converged Ethernet version 2) traffic over the WAN. However, AI workloads are highly sensitive to network congestion and packet loss, even minor packet drops can significantly degrade training efficiency. Due to the long links and significant end-to-end latency in wide area networks, traditional congestion control mechanisms may not be effective in a timely manner. They are insufficient for AI workloads due to their reactive nature and inability to guarantee zero packet loss.

To address these challenges, this draft explores the precise utilization of Explicit Congestion Notification (ECN) in WAN environments, particularly for RoCEv2 over IP tunnels. ECN enables

proactive congestion signaling by marking packets instead of dropping them, allowing endpoints to adjust transmission rates before congestion escalates. However, traditional ECN implementations face challenges in WAN scenarios, including inconsistent ECN propagation across tunnel boundaries and inefficient congestion response mechanisms. This work focuses on optimizing ECN for lossless RoCEv2 transmission in WANs by:

1. Ensuring Accurate ECN Propagation: Defining rules for consistent ECN field handling across IP-in-IP tunnels to prevent packet loss.

2. Enhancing Congestion Feedback: Adjust the sending rate within a small range of the wide area network to reduce the impact of latency on end-to-end communication.

3. Supporting Multi-Level Congestion Signaling: Extending ECN to differentiate between varying congestion severities, improving responsiveness for AI traffic.

By refining ECN mechanisms for WAN environments, this approach enhances network efficiency for distributed AI training while maintaining backward compatibility with existing protocols. The proposed framework provides a scalable and reliable solution for future large-scale distributed computing applications.

2. Conventions Used in This Document

2.1. 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.

2.2. Abbreviations

AIDC: Artificial Intelligence Data Center

RoCEv2: RDMA over Converged Ethernet version 2

ECN: Explicit Congestion Notification

CNP: Congestion Notification Packet

3. ECN for WAN

3.1. ECN Mechanism for WANs

In WANs, tunneling is a fundamental technique used to encapsulate and transport data packets across different network domains while maintaining security, performance, and compatibility. Tunneling works by embedding an original packet (the inner payload) within a new packet (the outer header), allowing it to traverse intermediate networks that may not natively support the original protocol.

ECN, as a traditional congestion notification mechanism, has also been extended from DC to WAN. [RFC6040] introduces how to label and use ECN mechanisms in tunnels, which are divided into tunnel ingress behavior and tunnel egress behavior. each behavior contain two encapsulation modes: a "compatibility mode," which is for backward compatibility with tunnel decapsulators that do not comprehend ECN, and a REQUIRED "normal mode." The detail of ingress behavior is shown below:

Incoming Header (also equal to departing Inner Header)	Departing Outer Header	
	Compatibility Mode	Normal Mode
Not-ECT	Not-ECT	Not-ECT
ECT(0)	Not-ECT	ECT(0)
ECT(1)	Not-ECT	ECT(1)
CE	Not-ECT	CE

Figure 1: New IP in IP Encapsulation Behaviours

For the decapsulation behaviour, detail is shown below:

Arriving Inner Header	Arriving Outer Header			
	Not-ECT	ECT(0)	ECT(1)	CE
Not-ECT	Not-ECT	Not-ECT(!!!)	Not-ECT(!!!)	drop (!!!)
drop	ECT(0)	ECT(0)	light CE	CE
drop	ECT(1)	ECT(1) (!)	light CE	CE
CE	CE	CE	CE(!!!)	CE

Figure 2: New IP in IP Decapsulation Behaviour

ECT(0) and ECT(1) can both indicate the same degree of congestion marking (such as "not congestion marked") according to the reasoning above. However, it also makes it possible to construct future schemes in which ECT(1) can represent other situation in WAN scenario.

3.2. Two-Threshold ECN Mechanism for WAM

To address the issue of delayed congestion transmission caused by high notification latency in wide area networks, this draft proposes the Two Threshold ECN Mechanism. Devices that support ECN in WANs will set two thresholds, with different thresholds representing different queue congestion situations. The supported devices will respond differently when different thresholds are reached. Here, the outer IP packet encapsulation behavior and decapsulation behavior have no change, the meaning of the ECT(1) codepoint has change from indicate ECN enable to indicate light congestion happen, detail procedure is as follows:

1. When queue occupancy reaches T1 (lower threshold): devices mark packets with ECT(1) codepoint, marking probability increases linearly with queue length and intended as early warning signal, then send a CNP packet to the PE which is tunnel ingress point. When the ingress PE receive the CNP packet, it will reduce the transmission rate or reroute the packet to other path. In this situation, ingress PE will not copy the ECN code to the inner packet header.

2. When queue occupancy reaches T2 (higher threshold): devices mark packets with CE codepoint, marking probability follows RED-like curve and need indicates immediate congestion requiring rate reduction. then send a CNP packet to the PE which is tunnel ingress point. When the ingress PE receive the CNP packet, it will copy the ECN code to the inner packet header and send the packet to the sender. When the sender receive the notification, it will reduce the transmission rate.

4. IANA Considerations

TBC

5. Security Considerations

TBC

6. References

6.1. 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>.

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

Contributors

Thanks to all the contributors.

Author’s Address

Jiayuan Hu (editor)
China Telecom
109, West Zhongshan Road, Tianhe District
Guangzhou
Guangzhou, 510000
China
Email: hujy5@chinatelecom.cn