

Routing Area Working Group  
Internet-Draft  
Intended status: Informational  
Expires: 2 September 2026

Jiayuan. Hu, Ed.  
China Telecom  
1 March 2026

Precise ECN in WAN  
draft-hu-rtgwg-pre-ecn-wan-01

## 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

## 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

Copyright (c) 2026 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Conventions Used in This Document . . . . .	3
2.1. Requirements Language . . . . .	3
2.2. Abbreviations . . . . .	3
3. ECN for WAN . . . . .	4
3.1. ECN Mechanism for WANs . . . . .	4
3.2. Two-Threshold ECN Mechanism for WAN . . . . .	5
3.2.1. Mechanism Overview . . . . .	5
3.2.2. Congestion Notification Packet (CNP) and Ingress PE Action . . . . .	6
3.2.3. Deployment and Compatibility Considerations . . . . .	7
4. IANA Considerations . . . . .	7
5. Security Considerations . . . . .	7
5.1. Threats Related to the Two-Threshold Mechanism . . . . .	8
5.2. Covert Channel Considerations . . . . .	8
6. References . . . . .	8
6.1. Normative References . . . . .	8
Contributors . . . . .	9
Author's Address . . . . .	9

## 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 behavior, detail is shown below:

Arriving		Arriving Outer Header			
Inner	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

The high latency and bursty nature of WAN links introduce significant challenges for timely and accurate congestion management. Traditional single-threshold ECN or packet-drop mechanisms often result in delayed feedback, causing over-correction (global synchronization) or under-correction (persistent congestion). To address this, this draft proposes a Two-Threshold ECN Mechanism specifically designed for WAN environments carrying latency-sensitive, loss-averse traffic like RoCEv2 for AI training.

#### 3.2.1. Mechanism Overview

This mechanism redefines the use of the ECT(1) codepoint within a controlled domain (e.g., a provider's WAN core), leveraging the flexibility permitted by RFC 8311. Network devices (e.g., routers, switches) supporting this mechanism are configured with two queue occupancy thresholds: a Lower Threshold (T1) and a Higher Threshold (T2).

ECT(1) as "Pre-Congestion" or "Early Warning" Signal: When the average queue length exceeds T1, the device interprets this as incipient or light congestion. Packets with an outer IP header ECN field of ECT(0) or ECT(1) are remarked to ECT(1) with a probability that increases linearly with the queue length. This ECT(1) marking is defined within this WAN domain as a pre-congestion notification (PCN). Its purpose is to signal an impending congestion condition before queues build to a level that would impact latency or cause loss.

CE as "Severe Congestion" Signal: When the average queue length exceeds T2, the device interprets this as severe congestion requiring immediate action. Packets are marked with the CE codepoint following a standard RED-like algorithm. This signal mandates a direct and measurable reduction in the data sender's transmission rate.

### 3.2.2. Congestion Notification Packet (CNP) and Ingress PE Action

A critical component of this mechanism is the generation and processing of a Congestion Notification Packet (CNP). This is a control packet generated by the congested device (or a network controller monitoring it) and sent to the tunnel ingress Provider Edge (PE) device—the point where the traffic entered the WAN domain.

#### 3.2.2.1. Upon reaching T1 (ECT(1) marking):

The congested device generates and sends a CNP to the ingress PE. This CNP identifies the affected flow (e.g., via 5-tuple) and indicates a light congestion event.

Ingress PE Action: Upon receiving this CNP, the ingress PE MAY take proactive measures to alleviate the impending congestion without involving the end host. This can include:

Local Rate Adjustment: Slightly reducing the transmission rate for the identified flow into the tunnel.

Traffic Rerouting: Dynamically steering the flow to an alternative, less congested path within the WAN if available (e.g., using SRv6 policy).

ECN Propagation: Crucially, at this stage, the ingress PE does NOT copy the outer ECT(1) marking to the inner IP header during decapsulation (following a modified "pipe model" logic for this codepoint). The end host remains unaware of this early warning, preventing an over-reaction from a distant sender whose feedback loop is delayed by the WAN RTT.

#### 3.2.2.2. Upon reaching T2 (CE marking):

The congested device generates and sends a CNP to the ingress PE, now indicating a severe congestion event.

Ingress PE Action: The ingress PE MUST take action to ensure the end host's congestion control is engaged. It performs standard RFC 6040 "normal mode" decapsulation: the CE codepoint from the outer header is copied to the inner IP header.

The packet, now with CE set in the inner header, is forwarded to the receiver. The receiver's transport (e.g., RoCEv2) then feeds this congestion signal back to the original sender, which MUST reduce its transmission rate according to its congestion control algorithm.

#### 3.2.3. Deployment and Compatibility Considerations

Backward Compatibility: Devices not implementing this two-threshold mechanism will treat ECT(1) as equivalent to ECT(0) per [RFC3168], and will process CE normally. This ensures safe co-existence and incremental deployment.

Domain of Application: This mechanism is intended for deployment within a managed WAN domain (e.g., a single provider's core). The re-semantic of ECT(1) is a local policy. At the egress PE leaving this domain, standard RFC 6040 rules apply for forwarding packets into external networks.

Threshold Tuning: The values of T1 and T2 are critical and should be set based on link capacity, typical traffic profiles, and the desired latency-loss trade-off for the target applications (e.g., AI training). T1 should be set low enough to provide meaningful early warning but high enough to avoid triggering on transient micro-bursts.

#### 4. IANA Considerations

TBC

#### 5. Security Considerations

The proposed enhancements introduce new mechanisms that must be evaluated for potential security vulnerabilities. This section expands upon the security considerations of [RFC3168] and [RFC6040] within the context of this draft.

### 5.1. Threats Related to the Two-Threshold Mechanism

CNP Spoofing and Forgery: An attacker could generate malicious CNPs and send them to an ingress PE, falsely indicating congestion. This could trigger unnecessary rate reduction or rerouting, leading to denial-of-service (performance degradation) for legitimate flows or manipulation of traffic paths for interception.

Threshold Manipulation: An on-path attacker with access to network device configuration could alter the T1 or T2 thresholds. Lowering T1 excessively would cause frequent ECT(1) marking and CNP generation, leading to under-utilization of the link. Raising T2 excessively could suppress legitimate CE signals, leading to bufferbloat and packet loss.

ECN Field Tampering within the Tunnel: As noted in [RFC3168] and [RFC6040], the outer ECN field is mutable. An attacker within the WAN could erase CE marks to hide congestion from the sender, or could set false CE/ECT(1) marks to artificially throttle flows. The two-threshold mechanism's use of ECT(1) as a significant signal creates a new vector for manipulation.

### 5.2. Covert Channel Considerations

[RFC6040] explicitly relaxed earlier restrictions on the covert channel bandwidth across tunnels, deeming a 2-bit per packet channel manageable. This mechanism does not alter that fundamental assessment. However, the specific semantics where the ingress PE does not propagate ECT(1) outwards but does act on a CNP could theoretically be exploited by a colluding ingress and egress point to encode information. This is considered a manageable risk within a single administrative domain.

## 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>>.
- [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>>.

#### 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](mailto:hujy5@chinatelecom.cn)