RTGWG Working Group
Internet-Draft
Intended status: Standards Track
Expires: 5 June 2026

X. Min
H. Li
ZTE Corp.
K. Zhang
W. Cheng
J. Yang
China Mobile
X. He
China Telecom

2 December 2025

Fast Congestion Notification Packet (CNP) in RoCEv2 Networks
draft-xiao-rtgwg-rocev2-fast-cnp-04

Abstract

This document describes a Remote Direct Memory Access (RDMA) over Converged Ethernet version 2 (RoCEv2) congestion control mechanism, which is inspired by Really Explicit Congestion Notification (RECN) described in RFC 7514, also known as Fast Congestion Notification Packet (Fast CNP). By extending the RoCEv2 CNP, Fast CNP can be sent by the switch directly to the sender, advising the sender to reduce the transmission rate at which it sends the flow of RoCEv2 data traffic.

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 5 June 2026.

Copyright Notice

Table of Contents

1. Introduction

Remote Direct Memory Access (RDMA) is a method of accessing memory on a remote system without interrupting the processing of the Central Processing Unit (CPU) on that system. RDMA enables lower latency and higher throughput on the network and lower CPU utilization for the servers and storage systems. High Performance Computing (HPC) and Artificial Intelligence (AI) applications can be accelerated by RDMA.

InfiniBand is a lossless network optimized for HPC and AI. It typically supports RDMA enabling machines to communicate and share data without interrupting the host CPU.

RDMA over Converged Ethernet (RoCE) is an open standard enabling RDMA and network offloads over an Ethernet network. The current and most popular implementation is RDMA over Converged Ethernet version 2 (RoCEv2) [IBTA-Spec]. RoCEv2 runs the InfiniBand transport layer over UDP and IP protocols on an Ethernet network, bringing many of the advantages of InfiniBand to Ethernet networks.

The RoCEv2 networks often implement a proactive congestion control mechanism analogous to Explicit Congestion Notification (ECN) [RFC3168], in which the switch marks packets if congestion occurs in

the network. The marked packets alert the receiver that packet loss is imminent, and then the receiver alerts the sender by sending Congestion Notification Packet (CNP). After receiving the CNP, the sender knows to back off, slowing down the transmission rate temporarily until the flow path is ready to handle a higher rate of traffic.

This document describes a RoCEv2 congestion control mechanism, which is inspired by Really Explicit Congestion Notification (RECN) [RFC7514], also known as Fast CNP. By extending the RoCEv2 CNP, Fast CNP can be sent by the switch directly to the sender, advising the sender to reduce the transmission rate at which it sends the flow of RoCEv2 data traffic. The primary benefit of Fast CNP has been explicitly indicated by its name saying that it’s faster than the receiver-originated CNP.

2. Conventions Used in This Document

2.1. Abbreviations

AI: Artificial Intelligence

CNP: Congestion Notification Packet

CPU: Central Processing Unit

DoS: Denial-of-Service

ECN: Explicit Congestion Notification

ECMP: Equal-Cost Multipath

HPC: High Performance Computing

HPCC++: Enhanced High Precision Congestion Control

IBTA: InfiniBand Trade Association

IOAM: In situ Operations, Administration, and Maintenance

RDMA: Remote Direct Memory Access

RECN: Really Explicit Congestion Notification

RoCE: RDMA over Converged Ethernet

RoCEv2: RDMA over Converged Ethernet version 2

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. RoCEv2 Data Packet and CNP formats

RoCEv2 packets use a well-known UDP Destination Port number 4791 that unambiguously distinguishes them in a stateless manner. RoCEv2 data packet format and congestion notification packet format are shown in Figure 1 and Figure 2 respectively.

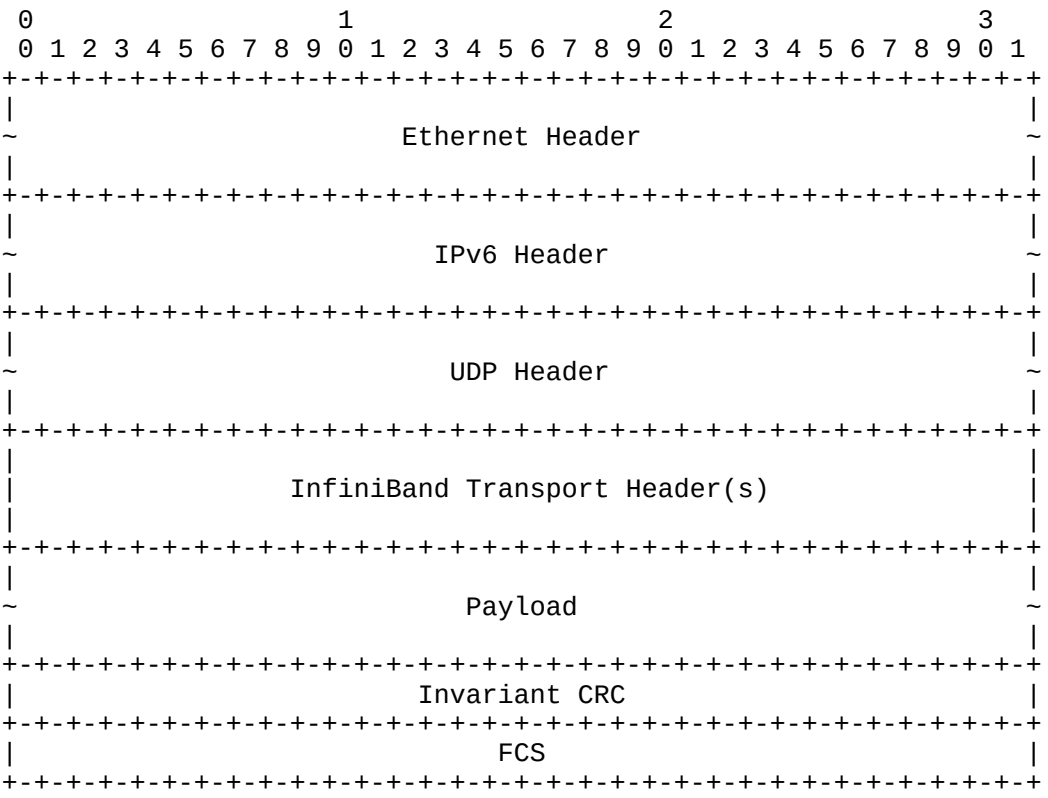

Figure 1: RoCEv2 Data Packet Format

In a RoCEv2 data packet, the InfiniBand Transport Header(s) must start with an InfiniBand Base Transport Header, followed by 0, 1, or multiple InfiniBand Extended Transport Header(s).

Within the InfiniBand Base Transport Header, there is a 24-bit field called Destination Queue Pair (Destination QP), indicating the Work Queue Pair Number at the destination. The QP is the virtual interface that the hardware provides to an InfiniBand architecture consumer, and it serves as a virtual communication port for the consumer. The operation on each QP is independent from the others.

In order to save the space, the Source QP indicating the Work Queue Pair at the source is not contained in the InfiniBand Base Transport Header. It’s assumed that the receiver is able to figure out the Source QP of a RoCEv2 data packet, because both the sender and the receiver know the mapping between the Source QP and the Destination QP.

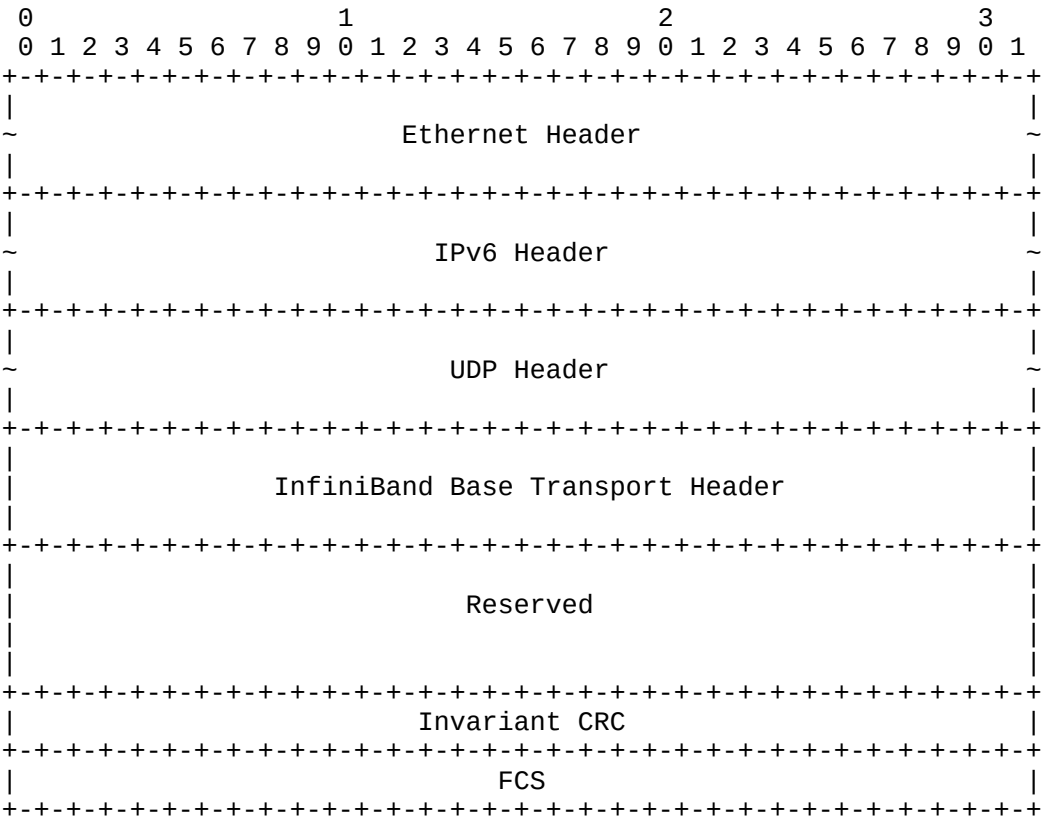

Figure 2: RoCEv2 Congestion Notification Packet Format

In a RoCEv2 CNP, following the IP/UDP headers, only the InfiniBand Base Transport Header but no any other InfiniBand Transport Header is present. In this document, only IPv6 is taken into account while IPv4 is beyond the scope.

The RoCEv2 CNP is generated by the receiver after receiving RoCEv2 data packet with ECN bits set. The Destination QP of the RoCEv2 CNP is set to the Work Queue Pair Number at the sender, corresponding to the Destination QP of the received RoCEv2 data packet.

After the sender receives the RoCEv2 CNP, the sender would reduce the transmission rate at which it sends the RoCEv2 data packets. The congestion control algorithm used by the sender to reduce the transmission rate is outside the scope of this document.

Fast CNP is an extended CNP generated by the switch at which congestion occurs, but not generated by the receiver. For a RoCEv2 CNP, it’s sent by a receiver and the Source QP of the sender is populated in the Destination QP field of the CNP, so it’s easy for the sender to know the Source QP after receiving the CNP. However, for a Fast CNP, it’s sent by a switch and the original Destination QP of the receiver is populated in the Destination QP field of the Fast CNP, which is not enough for the sender to know the Source QP. The reason is that the sender can communicate with multiple receivers, while different receivers may use the same Destination QP mapped to different Source QP at the sender. This document proposes to prepend an IPv6 extension header containing the original Destination Address to the RoCEv2 CNP, enabling the sender to figure out the Source QP by the combination of the original Destination QP and the original Destination Address (i.e., the receiver’s address).

4. IPv6 Destination Options for Fast CNP

The switch would send Fast CNP to the sender of RoCEv2 data packet causing congestion. If the switch doesn’t know about whether the sender is able to process the Fast CNP, then the switch MAY choose to mark the ECN bits of the RoCEv2 data packet at the same time of sending Fast CNP. The marked ECN bits of the RoCEv2 data packet would cause the receiver to send RoCEv2 CNP to the sender. In this case, the sender would receive both the Fast CNP and the receiver-originated CNP. If the switch knows that the sender is able to process the Fast CNP, then the switch MUST NOT mark the ECN bits of the RoCEv2 data packet at the same time of sending Fast CNP. How the switch can know the sender’s capability of processing Fast CNP is outside the scope of this document.

Fast CNP's Source IPv6 address is set to the IPv6 loopback address of the switch which sends the Fast CNP, and the Destination IPv6 address of the Fast CNP is copied from the Source IPv6 address of the RoCEv2 data packet causing congestion. After the sender receives a Fast CNP, the sender can optionally use the Source IPv6 address to check whether it's a Fast CNP as defined in this document. To this end it will compare the source address of the received Fast CNP with the intended destination address of the original packet, which can be found in the IPv6 extension header appended in the packet. If these two addresses are the same, then it is a receiver-originated Fast CNP, otherwise, if the two addresses differ, it means that the sender of the Fast CNP is a switch along the path. Whether it's necessary for the receiver to send Fast CNP is outside the scope of this document, while that's not excluded by this document. Furthermore, if the sender knows how to detour the congested switch (e.g., by changing the ECMP field(s) of the flow of RoCEv2 data packets that were subject to forward congestion), then the sender can also use the Source IPv6 address of the Fast CNP (sent by a switch) as a signal to detour the congested switch.

Fast CNP's Destination QP within the InfiniBand Base Transport Header is copied from the Destination QP within the InfiniBand Base Transport Header of the RoCEv2 data packet causing congestion.

Fast CNP adds an IPv6 extension header [RFC8200] to the RoCEv2 CNP, specifically, an IPv6 Destination Options header with one IPv6 destination option is added. There are two types of IPv6 destination option which can be added.

When the RoCEv2 data packet causing congestion doesn't carry an IPv6 In situ OAM (IOAM) Hop-by-Hop Trace Option [RFC9486], the following IPv6 destination option is carried in the Fast CNP.

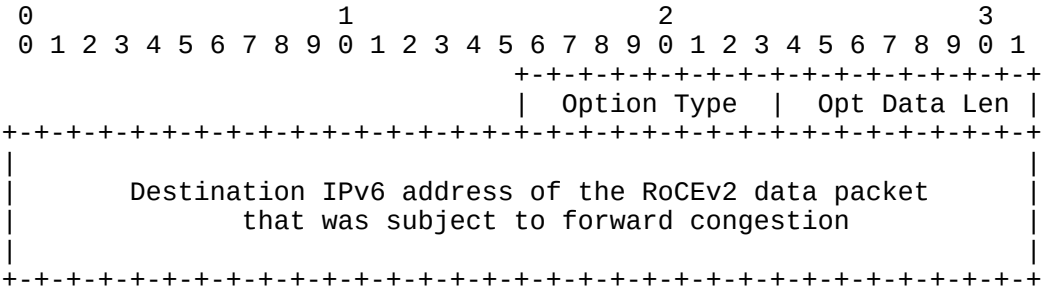

Figure 3: IPv6 Destination Option Format for Carrying Destination IPv6 address of the Congested RoCEv2 Data Packet

Option Type: 8-bit identifier of the type of Option that needs to be allocated. [RFC8200] defines how to encode the three high-order bits of the Option Type field. The two high-order bits specify the action that must be taken if the processing IPv6 node does not recognize the Option Type; for this Option, these two bits MUST be set to 10 (discard the packet and, regardless of whether or not the packet's Destination Address was a multicast address, send an ICMP Parameter Problem, Code 2, message to the packet's Source Address, pointing to the unrecognized Option Type). The third-highest-order bit specifies whether the Option Data can change en route to the packet's final destination; for this Option, the value of this bit MUST be set to 0 (Option Data does not change en route).

Opt Data Len: 16. It is the length of the Option Data Field of this Option in bytes.

Option Data: Destination IPv6 address of the RoCEv2 data packet that was subject to forward congestion. The Option Data, combined with the Destination QP within the InfiniBand Base Transport Header, are used by the sender to obtain the Work Queue Pair Number for which the transmission rate would be reduced.

When the RoCEv2 data packet causing congestion carries an IPv6 IOAM Hop-by-Hop Trace Option, the following IPv6 destination option is carried in the Fast CNP.

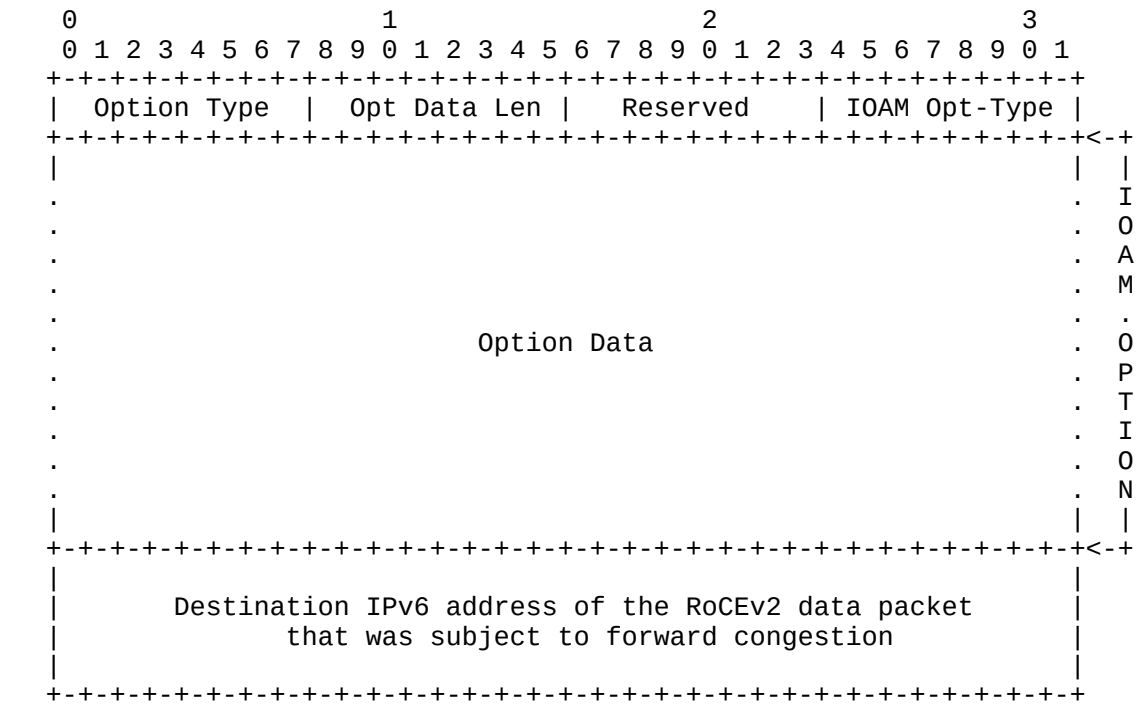

Figure 4: IPv6 Destination Option Format for Carrying IOAM Option and Destination IPv6 address of the Congested RoCEv2 Data Packet

Option Type: 8-bit identifier of the type of Option that needs to be allocated. For this Option, the two high-order bits MUST be set to 10 (discard the packet and, regardless of whether or not the packet's Destination Address was a multicast address, send an ICMP Parameter Problem, Code 2, message to the packet's Source Address, pointing to the unrecognized Option Type). The third-highest-order bit MUST be set to 0 (Option Data does not change en route).

Opt Data Len: 8-bit unsigned integer. It is the length of the Option Data Field of this Option in bytes.

Option Data: IOAM Trace Option Data and Destination IPv6 address of the RoCEv2 data packet that was subject to forward congestion. IOAM Trace Option Data is copied from the IPv6 Hop-by-Hop Options header of the RoCEv2 data packet. The Destination IPv6 address of the RoCEv2 data packet, combined with the Destination QP within the InfiniBand Base Transport Header, are used by the sender to obtain the Work Queue Pair Number for which the transmission rate would be reduced. The IOAM Trace Option Data is used by the sender to decide how to reduce the transmission rate, based on a congestion control

algorithm. One example of the IOAM Trace Option Data and the congestion control algorithm is Enhanced High Precision Congestion Control (HPCC++) [I-D.miao-ccwg-hpcc] [I-D.miao-ccwg-hpcc-info].

5. Security Considerations

The Fast CNP MUST be applied in a specific controlled domain. A limited administrative domain provides the network administrator with the means to select, monitor, and control the access to the network, making it a trusted domain.

To avoid potential Denial-of-Service (DoS) attacks, it is RECOMMENDED that implementations apply rate-limiting policies when generating Fast CNPs.

To protect against unauthorized sources sending Fast CNP to the host, implementations MUST provide a means of checking the source addresses of Fast CNP against an access list before accepting the packet. For instance using [I-D.ietf-savnet-intra-domain-architecture].

A deployment MUST ensure that border-filtering drops inbound Fast CNP from outside of the domain and that drops outbound Fast CNP leaving the domain.

A deployment MUST support the configuration option to enable or disable the Fast CNP feature defined in this document. By default, the Fast CNP feature MUST be disabled.

As this document describes new options for IPv6, containing IOAM data or not, the security considerations of [RFC8200], [RFC9098], and [RFC9486] apply.

6. IANA Considerations

This document requests the following IPv6 Option Type assignments from the Destination Options and Hop-by-Hop Options sub-registry of Internet Protocol Version 6 (IPv6) Parameters (https://www.iana.org/assignments/ipv6-parameters/).

Hex Value	Binary Value act	chg	rest	Description	Reference
TBD1	10	0	tbd1	Fast CNP Destination Option1	[This draft]
TBD2	10	0	tbd2	Fast CNP Destination Option2	[This draft]

Table 1

7. Acknowledgements

The authors would like to acknowledge Luigi Iannone for his careful review and valuable discussion.

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>.

[RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <https://www.rfc-editor.org/info/rfc8200>.

[RFC9486] Bhandari, S., Ed. and F. Brockners, Ed., "IPv6 Options for In Situ Operations, Administration, and Maintenance (IOAM)", RFC 9486, DOI 10.17487/RFC9486, September 2023, <https://www.rfc-editor.org/info/rfc9486>.

8.2. Informative References

[I-D.ietf-savnet-intra-domain-architecture] Li, D., Wu, J., Qin, L., Geng, N., and L. Chen, "Intra-domain Source Address Validation (SAVNET) Architecture", Work in Progress, Internet-Draft, draft-ietf-savnet-intra-domain-architecture-03, 13 October 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-savnet-intra-domain-architecture-03>.

[I-D.miao-ccwg-hpcc] Miao, R., Anubolu, S., Pan, R., Lee, J., Gafni, B., Tantsura, J., Alemania, A., and Y. Shpigelman, "HPCC++: Enhanced High Precision Congestion Control", Work in Progress, Internet-Draft, draft-miao-ccwg-hpcc-03, 6 January 2025, <https://datatracker.ietf.org/doc/html/draft-miao-ccwg-hpcc-03>.

[I-D.miao-ccwg-hpcc-info]
Miao, R., Anubolu, S., Pan, R., Lee, J., Gafni, B., Tantsura, J., Alemania, A., and Y. Shpigelman, "Inband Telemetry for HPCC++", Work in Progress, Internet-Draft, draft-miao-ccwg-hpcc-info-04, 6 January 2025, <https://datatracker.ietf.org/doc/html/draft-miao-ccwg-hpcc-info-04>.

[IBTA-Spec]
InfiniBand Trade Association, "InfiniBand Architecture Specification Volume 1, Release 1.4", 2020, <https://www.infinibandta.org/ibta-specification/>.

[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>.

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

[RFC9098] Gont, F., Hilliard, N., Doering, G., Kumari, W., Huston, G., and W. Liu, "Operational Implications of IPv6 Packets with Extension Headers", RFC 9098, DOI 10.17487/RFC9098, September 2021, <https://www.rfc-editor.org/info/rfc9098>.

Authors' Addresses

Xiao Min
ZTE Corp.
Nanjing
China
Phone: +86 18061680168
Email: xiao.min2@zte.com.cn

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

Kan Zhang
China Mobile
Beijing
China

Email: zhangkan@chinamobile.com

Weiqiang Cheng
China Mobile
Beijing
China
Email: chengweiqiang@chinamobile.com

Jin Yang
China Mobile
Beijing
China
Email: yangjinwl@chinamobile.com

Xiaoming He
China Telecom
Guangzhou
China
Email: hexm4@chinatelecom.cn