RTGWG Working Group
Internet-Draft
Intended status: Standards Track
Expires: 18 December 2025

X. Min
ZTE Corp.

16 June 2025

Proxy for Congestion Notification
draft-xiao-rtgwg-proxy-congestion-notification-00

Abstract

This document describes the necessity and feasibility to introduce a proxy network node between the congested network node and the traffic sender. The proxy network node is used to translate the congestion notification. The congested network node sends the congestion notification to the proxy network node in a format defined in this document, and the proxy network node translates the received congestion notification to a format known by the traffic sender and resends the translated congestion notification to the traffic sender.

Table of Contents

1. Introduction

[I-D.xiao-rtgwg-rocev2-fast-cnp] describes a congestion notification message called Fast Congestion Notification Packet (CNP), which can be sent by a congested network node to the traffic sender directly. Fast CNP extends the CNP [IBTA-SPEC] consumed by the traffic sender supporting Remote Direct Memory Access (RDMA) over Converged Ethernet version 2 (RoCEv2).

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. For a traffic sender supporting RoCEv2, congestion control is important, and the RoCEv2 CNP or RoCEv2 Fast CNP must be used to alert the sender slowing down the transmission rate. For a traffic sender not supporting RoCEv2, congestion control

is still important, and the corresponding congestion notification mechanism supported by the sender must be used to alert the sender slowing down the transmission rate.

Considering there are multiple different congestion notification mechanisms existing for the traffic sender, if a congested network node would send a congestion notification message to the traffic sender directly, it's not easy for the congested network node to know what kind of congestion notification mechanism is supported by the traffic sender. In the case that the congested network node doesn't know exactly what kind of congestion notification mechanism is supported by each specific traffic sender within the network, a proxy network node nearer to the specific traffic sender can help to do the transition as a translator. That is to say, the congested network node sends a congestion notification message to a proxy network node first, and then the proxy network node notifies the traffic sender about the congestion, as long as the proxy network node knows what kind of congestion notification mechanism is supported by the traffic sender.

This document describes the necessity and feasibility to introduce a proxy network node between the congested network node and the traffic sender. Specifically, the problem statement is described in Sections 1 and 3, and the format of the congestion notification message sent from the congested network node to the proxy network node is defined in Section 4, and the potential solutions on how the congested network node knows the address of the proxy node are defined in Section 5.

2. Conventions Used in This Document

2.1. Abbreviations

ABR: Area Border Router

CNP: Congestion Notification Packet

DoS: Denial-of-Service

ECN: Explicit Congestion Notification

ELC: Entropy Label Capability

ELCv3: Entropy Label Characteristic

IBTA: InfiniBand Trade Association

PFC: Priority-based Flow Control

PNC: Proxy Node Capability

RDMA: Remote Direct Memory Access

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. Congestion Notification Mechanisms

In the field of congestion control, there are at least four kinds of referenced congestion notification mechanisms. This document introduces the fifth congestion notification mechanism called fast congestion notification with proxy.

The first congestion notification mechanism is referred to as classical stepwise back pressure with dedicated Ethernet pause frame, as shown in Figure 1.

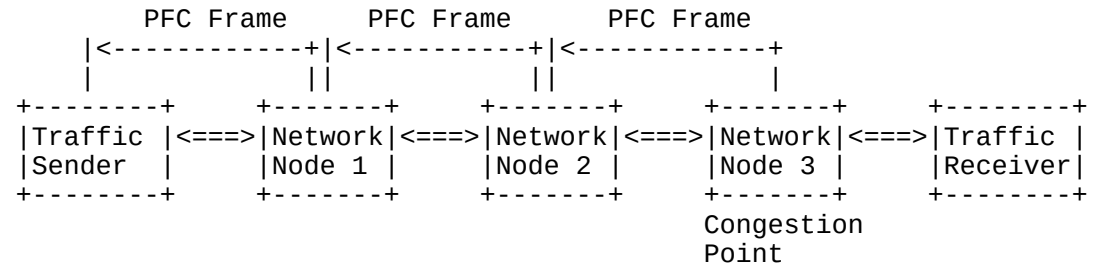

Figure 1: Classical Stepwise Back Pressure with Dedicated Ethernet Pause Frame

With this congestion notification mechanism, the congested network node (Netwok Node 3 in Figure 1) notifies the directly connected upstream network node (Network Node 2 in Figure 1) about the congestion by a dedicated Ethernet pause frame called Priority-based Flow Control (PFC) frame, and then the upstream network node may stepwise notify its directly connected upstream network node about the congestion by a PFC frame, until the most upstream network node (Network Node 1 in Figure 1) may notify the directly connected traffic sender about the congestion by a PFC frame. [IEEE8021Q-2022] details how this kind of congestion notification mechanism works.

The second congestion notification mechanism is referred to as classical congestion notification without dedicated packet, as shown in Figure 2.

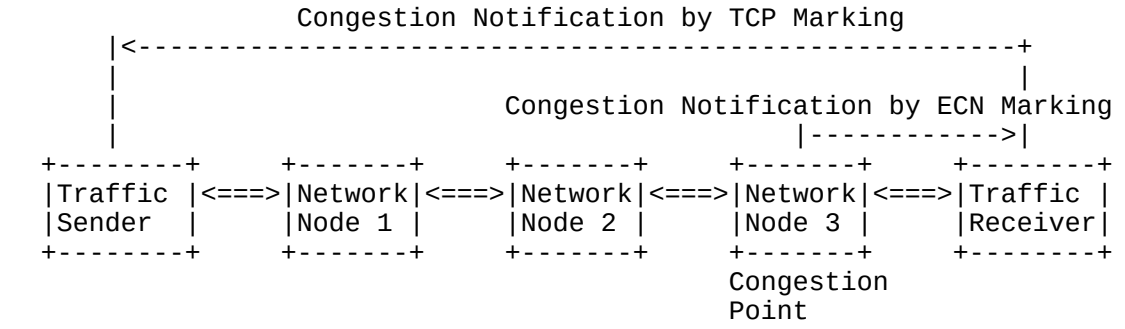

Figure 2: Classical Congestion Notification without Dedicated Packet

With this congestion notification mechanism, the traffic sender indicates that it supports the congestion notification from the traffic receiver by a specific Explicit Congestion Notification (ECN) marking within the IP header of the data packet, and the congested network node (Netwok Node 3 in Figure 2) notifies the traffic receiver about the congestion by a specific ECN marking. After receiving a data packet with the specific ECN marking, the traffic receiver would notify congestion to the traffic sender by a specific TCP marking within the TCP header of the data packet. [RFC3168] details how this kind of congestion notification mechanism works.

The third congestion notification mechanism is referred to as classical congestion notification with dedicated packet, as shown in Figure 3.

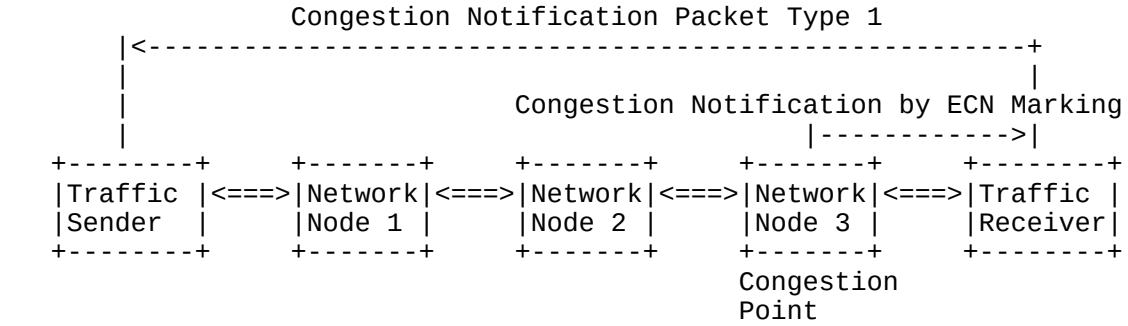

Figure 3: Classical Congestion Notification with Dedicated Packet

With this congestion notification mechanism, the traffic sender indicates that it supports the congestion notification from the traffic receiver by a specific ECN marking within the IP header of the data packet, and the congested network node (Netwok Node 3 in Figure 3) notifies the traffic receiver about the congestion by a specific ECN marking. After receiving a data packet with the specific ECN marking, the traffic receiver would notify congestion to the traffic sender by a dedicated congestion notification packet. [IBTA-SPEC] details an example on how this kind of congestion notification mechanism works.

The fourth congestion notification mechanism is referred to as fast congestion notification without proxy, as shown in Figure 4.

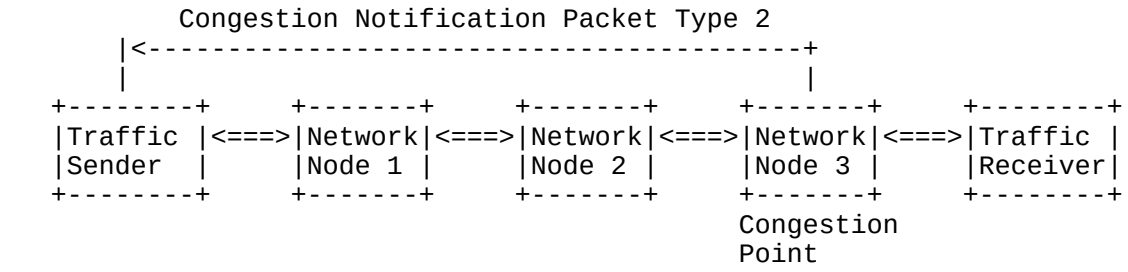

Figure 4: Fast Congestion Notification without Proxy

With this congestion notification mechanism, the congested network node (Netwok Node 3 in Figure 4) notifies the traffic sender about the congestion directly by a dedicated congestion notification packet. [I-D.xiao-rtgwg-rocev2-fast-cnp] details an example on how this kind of congestion notification mechanism works.

The fifth congestion notification mechanism is referred to as fast congestion notification with proxy, as shown in Figure 5.

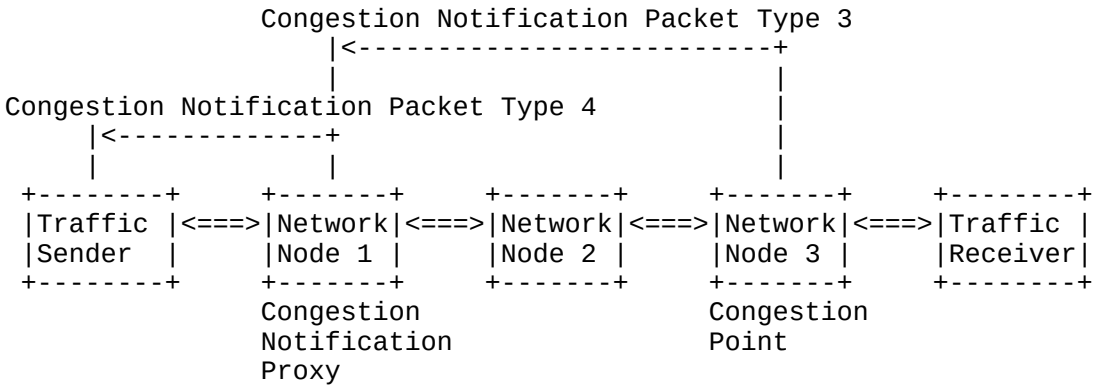

Figure 5: Fast Congestion Notification with Proxy

With this congestion notification mechanism, the congested network node (Netwok Node 3 in Figure 5) notifies the proxy network node about the congestion by a dedicated congestion notification packet, and then the proxy network node notifies the traffic sender about the congestion by a congestion notification mechanism supported by the traffic sender. This document details how this kind of congestion notification mechanism works, except that the congestion notification mechanism between the proxy network node and the traffic sender is beyond the scope of this document.

4. Congestion Notification to the Proxy Node

The congestion notification message sent from the congested network node to the proxy network node is a UDP message which is formatted as follows:

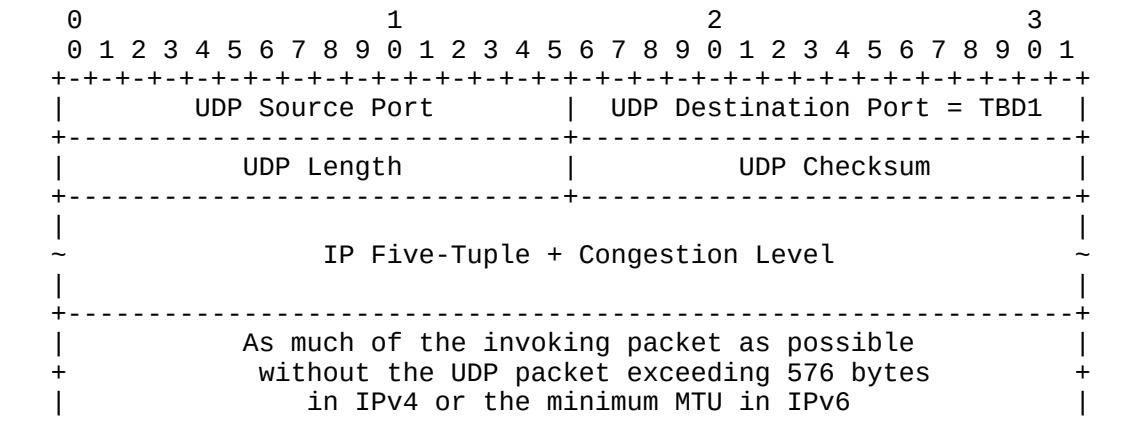

Figure 6: Congestion Notification Message Format

UDP Header: The UDP header as specified in [RFC768] includes the UDP source port, UDP destination port, UDP length, and UDP checksum. A well-known UDP destination port needs to be allocated for this Congestion Notification Message.

IP Five-Tuple: The IP five-tuple as described in [RFC6438] includes the source IP address, destination IP address, protocol number, source port number, and destination port number. The IP five-tuple is copied from the data packet causing congestion, and it’s used to identify a flow for which the transmission rate needs to be reduced by the traffic sender.

Congestion Level: This 3-bit field indicates the congestion level. Value 0 of this field represents the lowest congestion level and value 7 of this field represents the highest congestion level.

5. Proxy Notification to the Congestion Node

Before the congested network node can send the congestion notification message to the proxy network node, the congested network node has to be notified about the IP address of the proxy network node. There are two kinds of notification mechanisms. One kind of notification mechanism is through control plane advertisement, which is specified in Section 5.1. Another kind of notification mechanism is through data plane dissemination, which is specified in Section 5.2.

5.1. Advertising Proxy Node Capability Using IGP/BGP

Even though Proxy Node Capability (PNC) is a property of the node, in some cases it is advantageous to associate and advertise the PNC with a prefix. When PNC is advertised with a prefix, that means the congested network node should send the congestion notification packet to the proxy network node but not the traffic sender associated with that prefix.

5.1.1. Advertising Proxy Node Capability Using IS-IS

Analogous to the Entropy Label Capability (ELC) Flag (E-flag) defined in Section 3 of [RFC9088], a new bit PNC Flag (P-flag) is defined, which is Bit 7 in the Prefix Attribute Flags [RFC7794], as shown in Figure 7.

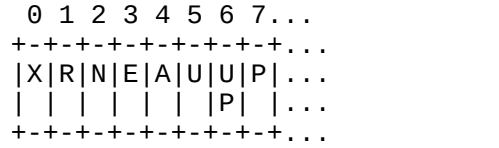

Figure 7: IS-IS Prefix Attribute Flags

P-Flag: PNC Flag (Bit 7)

Set for the local host prefix of the originating node if it’s used as a congestion notification proxy node for the prefix.

The PNC signaling MUST be preserved when a router propagates a prefix between ISIS levels [RFC5302].

5.1.2. Advertising Proxy Node Capability Using OSPFv2

Analogous to the ELC Flag (E-flag) defined in Section 3.1 of [RFC9089], a new bit PNC Flag (P-flag) is defined, which is Bit 2 in OSPFv2 Prefix Attribute Flags field [RFC9792], as shown in Figure 8.

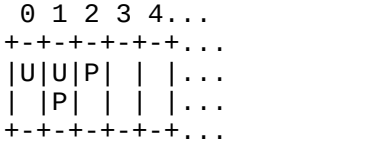

Figure 8: OSPFv2 Prefix Attribute Flags

P-Flag: PNC Flag (Bit 2)

Set for the local host prefix of the originating node if it’s used as a congestion notification proxy node for the prefix.

The PNC signaling MUST be preserved when an OSPFv2 Area Border Router (ABR) distributes information between areas. To do so, an ABR MUST originate an OSPFv2 Extended Prefix Opaque LSA [RFC7684] including the received PNC setting.

5.1.3. Advertising Proxy Node Capability Using OSPFv3

Analogous to the ELC Flag (E-flag) defined in Section 3.2 of [RFC9089], a new bit PNC Flag (P-flag) is defined, which is Bit 2 in OSPFv3 Prefix Attribute Flags field [RFC9792], as shown in Figure 9.

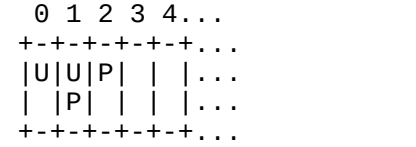

Figure 9: OSPFv3 Prefix Attribute Flags

P-Flag: PNC Flag (Bit 2)

Set for the local host prefix of the originating node if it’s used as a congestion notification proxy node for the prefix.

The PNC signaling MUST be preserved when an OSPFv3 Area Border Router (ABR) distributes information between areas. The setting of the PNC Flag in the Inter-Area-Prefix-LSA [RFC5340] or in the Inter-Area-Prefix TLV [RFC8362], generated by an ABR, MUST be the same as the value the PNC Flag associated with the prefix in the source area.

5.1.4. Advertising Proxy Node Capability Using BGP

Analogous to the Entropy Label Characteristic (ELCv3) TLV defined in Section 3.1 of [I-D.ietf-idr-entropy-label], a new PNC characteristic TLV is defined, which uses code value TBD2 in "BGP Next Hop Dependent Characteristic Codes" registry requested by [I-D.ietf-idr-entropy-label], as shown in Figure 10.

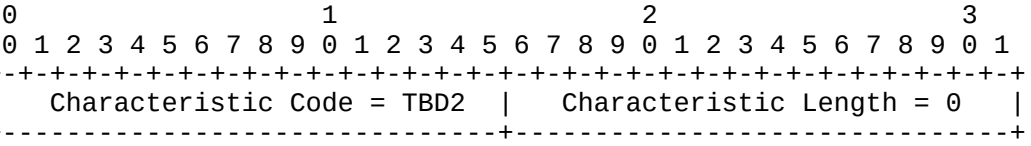

Figure 10: BGP Next Hop Dependent Characteristic PNC TLV Format

PNC TLV: code TBD2, length 0, and carries no value

Carried for the local host prefix of the originating node if it's used as a congestion notification proxy node for the prefix.

5.2. Notifying Proxy Node Address Using IPv6 Destination Option

This section specifies an extension to the data packet when sent from the traffic sender. Specifically, an IPv6 Destination Options header with one IPv6 destination option carrying the IPv6 address of the proxy network node is added to the data packet. The congested network node would copy the IPv6 address of the proxy network node from the IPv6 destination option of data packet causing congestion to the destination address field of the congestion notification packet.

The Proxy Network Node Address Destination Option has the following format:

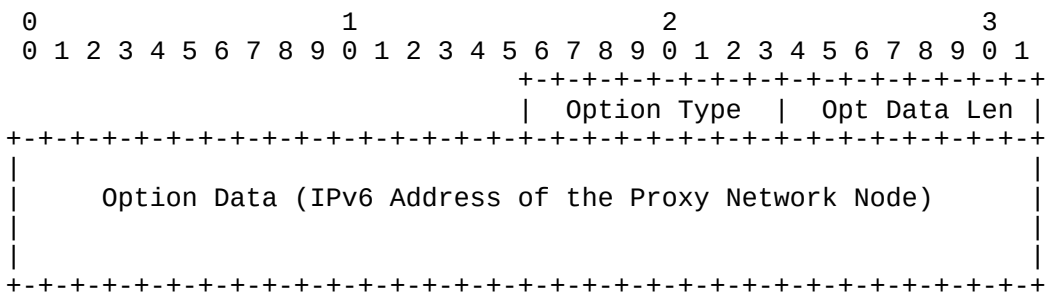

Figure 11: Format of the Proxy Network Node Address Destination Option

Option Type: This 8-bit field identifies 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 00 (skip over this option and continue processing the header). 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: This 8-bit field indicates the length of the Option Data field of this Option in bytes. This field MUST be set to 16.

Option Data: This 16-octet field indicates the proxy network node’s IPv6 address to which the congested network node would send the congestion notification message defined in Section 4.

6. Security Considerations

The congestion notification from congested network node to the proxy network node 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 and receiving congestion notification messages.

A deployment MUST ensure that border-filtering drops inbound congestion notification message from outside of the domain and that drops outbound congestion notification message leaving the domain.

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

As this document describes new option for IPv6, the security considerations of [RFC8200] and [RFC9098] apply.

7. IANA Considerations

This document requests the following allocations from IANA:

- A well-known UDP port number TBD1 in the "Service Name and Transport Protocol Port Number" registry is requested to be assigned to the Congestion Notification Message.

- Bit 7 in the "IS-IS Bit Values for Prefix Attribute Flags Sub-TLV" registry is requested to be assigned to the PNC Flag (P-Flag).

- Bit 2 in the "OSPFv2 Prefix Attribute Flags" registry is requested to be assigned to the PNC Flag (P-Flag).

- Bit 2 in the "OSPFv3 Prefix Attribute Flags" registry is requested to be assigned to the PNC Flag (P-Flag).

- Code value TBD2 in the "BGP Next Hop Dependent Characteristic Codes" registry is requested to be assigned to the PNC characteristic TLV.

- An IPv6 Option Type in the "Destination Options and Hop-by-Hop Options" registry is requested to be assigned to the Proxy Node Address Destination Option.

8. Acknowledgements

The author would like to acknowledge Jinghai Yu and Shaofu Peng for the very helpful discussion.

9. References

9.1. Normative References

[I-D.ietf-idr-entropy-label]
Decraene, B., Scudder, J., Kompella, K., Satya, M. R., Wen, B., Wang, K., and S. Krier, "BGP Next Hop Dependent Characteristics Attribute", Work in Progress, Internet-Draft, draft-ietf-idr-entropy-label-17, 30 March 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-entropy-label-17>.

[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>.

[RFC5302] Li, T., Smit, H., and T. Przygienda, "Domain-Wide Prefix Distribution with Two-Level IS-IS", RFC 5302, DOI 10.17487/RFC5302, October 2008, <https://www.rfc-editor.org/info/rfc5302>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC768] Postel, J., "User Datagram Protocol", STD 6, RFC 768, DOI 10.17487/RFC0768, August 1980, <https://www.rfc-editor.org/info/rfc768>.

[RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <https://www.rfc-editor.org/info/rfc7684>.

[RFC7794] Ginsberg, L., Ed., Decraene, B., Previdi, S., Xu, X., and U. Chunduri, "IS-IS Prefix Attributes for Extended IPv4 and IPv6 Reachability", RFC 7794, DOI 10.17487/RFC7794, March 2016, <https://www.rfc-editor.org/info/rfc7794>.

[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>.

[RFC8362] Lindem, A., Roy, A., Goethals, D., Reddy Vallem, V., and F. Baker, "OSPFv3 Link State Advertisement (LSA) Extensibility", RFC 8362, DOI 10.17487/RFC8362, April 2018, <https://www.rfc-editor.org/info/rfc8362>.

[RFC9792] Chen, R., Zhao, D., Psenak, P., Talaulikar, K., and L. Gong, "Prefix Flag Extension for OSPFv2 and OSPFv3", RFC 9792, DOI 10.17487/RFC9792, June 2025, <https://www.rfc-editor.org/info/rfc9792>.

9.2. Informative References

[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>.

[IBTA-SPEC] InfiniBand Trade Association, "InfiniBand Architecture Specification Volume 1, Release 1.8", July 2024, <https://www.infinibandta.org/ibta-specification>.

[IEEE8021Q-2022]
IEEE, "IEEE Standard for Local and Metropolitan Area Networks--Bridges and Bridged Networks", DOI 10.1109/IEEESTD.2022.10004498, IEEE Std 802.1Q-2022, December 2022, <https://ieeexplore.ieee.org/document/10004498>.

[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>.

[RFC6438] Carpenter, B. and S. Amante, "Using the IPv6 Flow Label for Equal Cost Multipath Routing and Link Aggregation in Tunnels", RFC 6438, DOI 10.17487/RFC6438, November 2011, <https://www.rfc-editor.org/info/rfc6438>.

[RFC9088] Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S., and M. Bocci, "Signaling Entropy Label Capability and Entropy Readable Label Depth Using IS-IS", RFC 9088, DOI 10.17487/RFC9088, August 2021, <https://www.rfc-editor.org/info/rfc9088>.

[RFC9089] Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S., and M. Bocci, "Signaling Entropy Label Capability and Entropy Readable Label Depth Using OSPF", RFC 9089, DOI 10.17487/RFC9089, August 2021, <https://www.rfc-editor.org/info/rfc9089>.

[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>.

Author’s Address

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