

RTGWG Working Group
Internet-Draft
Intended status: Standards Track
Expires: 22 March 2026

X. Min
ZTE Corp.
18 September 2025

Proxy for Congestion Notification
draft-xiao-rtgwg-proxy-congestion-notification-01

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 then 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.

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 22 March 2026.

Copyright Notice

Copyright (c) 2025 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	4
2.1. Abbreviations	4
2.2. Requirements Language	4
3. Congestion Notification Mechanisms	4
4. Congestion Notification to the Proxy Node	7
5. Advertising Proxy Node Capability Using IGP/BGP	8
5.1. Advertising Proxy Node Capability Using IS-IS	8
5.2. Advertising Proxy Node Capability Using OSPFv2	9
5.3. Advertising Proxy Node Capability Using OSPFv3	9
5.4. Advertising Proxy Node Capability Using BGP	10
6. Security Considerations	10
7. IANA Considerations	11
8. Acknowledgements	11
9. References	11
9.1. Normative References	11
9.2. Informative References	12
Author's Address	13

1. Introduction

[I-D.xiao-rtgwg-rocev2-fast-cnp] describes a congestion notification message called Fast Congestion Notification Packet (Fast 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 has already been widely deployed, and it 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 message supported by the sender must be used to alert the sender slowing down the transmission rate.

Considering there are multiple different congestion notification messages existing for the traffic sender, if a congested network node would send a congestion notification message to the traffic sender directly, there is a prerequisite for the congested network node to know what kind of congestion notification message is supported by each specific traffic sender; Secondly, when the congested network node is a VPN Provider (P) router, it's difficult for the congested network node to send a congestion notification message to the traffic sender directly, because there are different routing domains for the VPN P router and VPN Customer Edge (CE) router; Thirdly, when the traffic sender supports RoCEv2, it's difficult for the congested network node to construct a standard RoCEv2 CNP (please refer to Section 3 of [I-D.xiao-rtgwg-rocev2-fast-cnp]).

A proxy network node between the congested network node and the traffic sender can help to resolve the problems described above, being independent of the extension proposed in [I-D.xiao-rtgwg-rocev2-fast-cnp]. 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 using a congestion notification message known by the traffic sender (e.g., the standard RoCEv2 CNP). For the selection of the proxy network node, there are at least three rules. First one, the selected proxy network node must know what kind of congestion notification message is supported by the traffic sender; Second one, the selected proxy network node and the congested network node must be within the same routing domain; Third one, for RoCEv2 network, the selected proxy network node must be able to learn the mapping table between Source Queue Pair and Destination Queue Pair through data traffic, which means the selected proxy network node must be located where both the forward direction traffic and the backward direction traffic need to traverse. How to select a proxy network node for a specific traffic sender is deployment specific and beyond the scope of this document.

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 solution on how the congested network node knows the address of the proxy node is 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

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 three kinds of referenced congestion notification mechanisms. This document introduces the fourth congestion notification mechanism called "Fast Congestion Notification with Proxy".

The first congestion notification mechanism is referred to as classical congestion notification without dedicated packet, as shown in Figure 1.

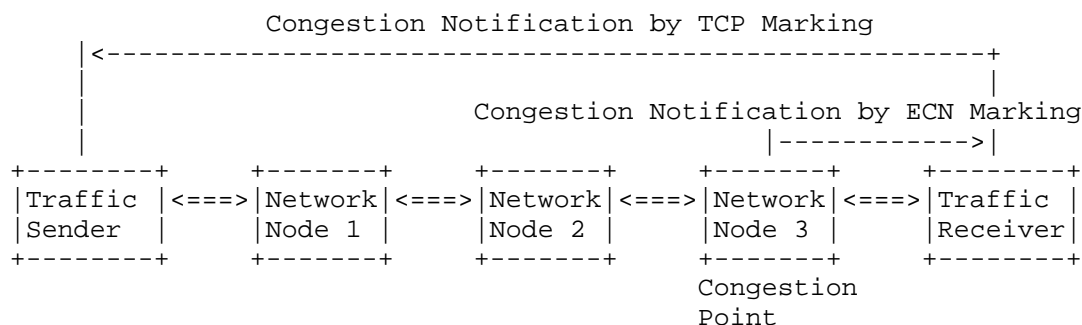


Figure 1: Classical Congestion Notification without Dedicated Packet

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 (Network Node 3 in Figure 1) 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 second congestion notification mechanism is referred to as classical congestion notification with dedicated packet, as shown in Figure 2.

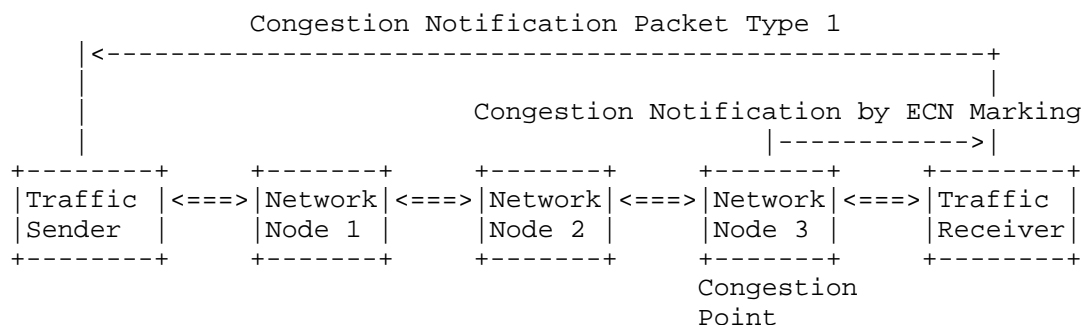


Figure 2: Classical Congestion Notification with Dedicated Packet

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 (Network 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 dedicated congestion notification packet. [IBTA-SPEC] details an example on how this kind of congestion notification mechanism works.

The third congestion notification mechanism is referred to as fast congestion notification without proxy, as shown in Figure 3.

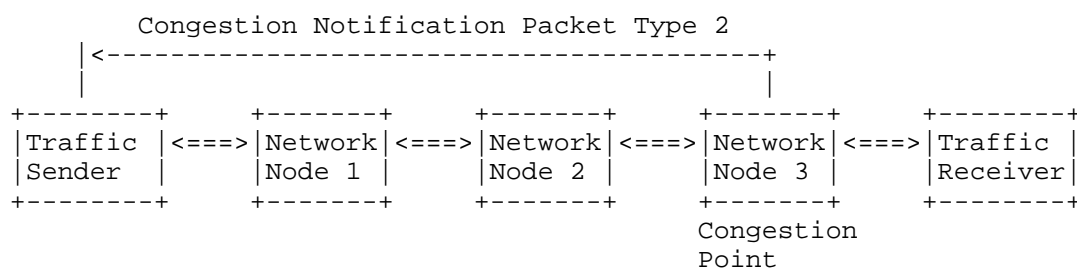


Figure 3: Fast Congestion Notification without Proxy

With this congestion notification mechanism, the congested network node (Network Node 3 in Figure 3) 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 fourth congestion notification mechanism is referred to as fast congestion notification with proxy, as shown in Figure 4.

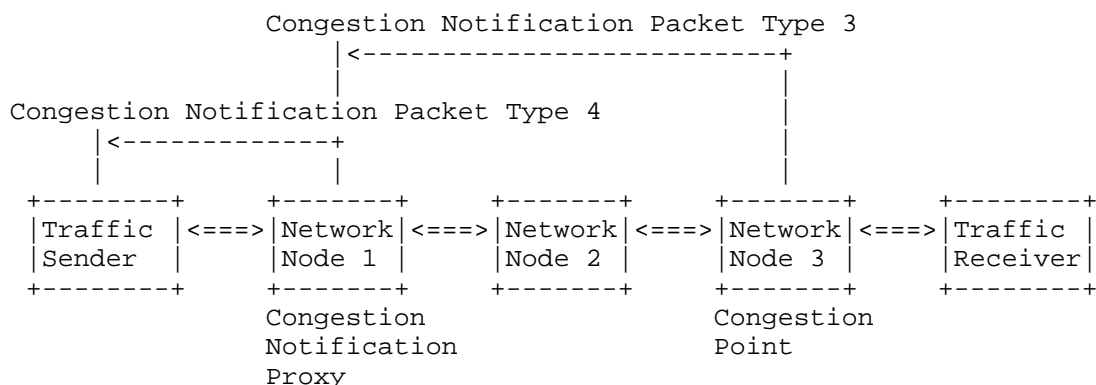


Figure 4: Fast Congestion Notification with Proxy

With this congestion notification mechanism, the congested network node (Network Node 3 in Figure 4) 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 message supported by the traffic sender. This document details how this kind of congestion notification mechanism works, except that the specific congestion notification message 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 can be a UDP message or an ICMP message, if a UDP message it's formatted as follows:

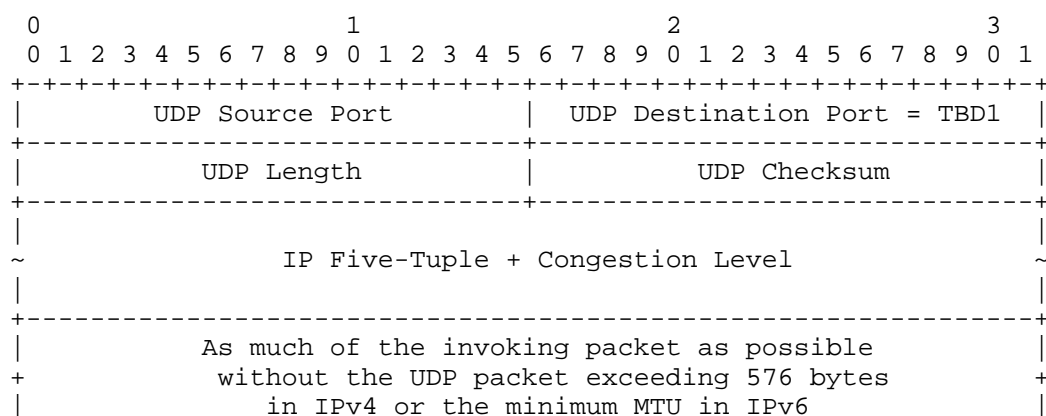


Figure 5: Congestion Notification Message Format

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. When the congested network node is a VPN P router, the IP five-tuple is carried below the VPN encapsulation.

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. Advertising Proxy Node Capability Using IGP/BGP

Before the congested network node can send the congestion notification message to the proxy network node, the congested network node has to know about the IP address of the proxy network node. The proxy network node can notify the congested network node of its IP address by advertising its proxy capability in advance.

Even though the 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. 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 6.

```

      0 1 2 3 4 5 6 7...
    +---+---+---+---+---+---+...
    |X|R|N|E|A|U|U|P|...
    | | | | | | |P|...
    +---+---+---+---+---+---+...

```


Figure 6: 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].

5.2. 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 7.

```

      0 1 2 3 4...
+---+---+---+...
|U|U|P| | |...
| |P| | | |...
+---+---+---+...

```

Figure 7: 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.

5.3. 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 8.

```

      0 1 2 3 4...
+---+---+---+...
|U|U|P| | |...
| |P| | | |...
+---+---+---+...

```

Figure 8: 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.

5.4. 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 9.

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
Characteristic Code = TBD2										Characteristic Length = 0																													

Figure 9: 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.

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.

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.

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-18, 20 July 2025, <<https://datatracker.ietf.org/doc/html/draft-ietf-idr-entropy-label-18>>.

- [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>>.
- [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>>.

[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>>.

Author's Address

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