RTGWG
Internet-Draft
Intended status: Standards Track
Expires: 7 March 2026

Z. Hu
Y. Zhu
China Telecom
X. Geng
Huawei
J. Hu
T. Pi
China Telecom

3 September 2025

Inter-domain congestion notification for SRv6-based distributed RoCEv2 network
draft-hu-acn-rocev2-00

Abstract

Some AI services drive the need to transmit RMDA packets across wide area network (WAN) via SRv6 tunnels. RoCEv2 is the most popular open standard for achieving RDMA and network offloads over ethernet, with its congestion control based on the combination of PFC and ECN. Due to certain limitations of PFC and ECN, some drafts have been put forward to realize more faster congestion notification (FANTEL). Upon detection of congestion, these drafts proposals directly sending congestion notifications to relevant nodes, enabling near real-time congestion control. However, in SRv6-based WAN environments, congestion notifications cannot be directly delivered from WAN devices to intra-DC devices. This document specifies new SRv6 Segment Identifiers (SIDs) and the corresponding processing rules for device, supporting the forwarding of congestion notification across domains.

Table of Contents

1. Introduction

RDMA (Remote Direct Memory Access) enables direct access to memory locations on remote machines, bypassing the need for CPU involvement in data transfer processes. RDMA results in lower latency, reduced CPU overhead, and increased network throughput, making RDMA particularly beneficial for high-performance computing environments, cloud infrastructure, and storage networks.

RoCEv2 is an open standard enabling RDMA over ethernet, with its congestion control based on the combination of PFC and ECN. Priority-based Flow Control (PFC) is a data link level flow control mechanism, which can selectively pause traffic according to its priority and eliminate packet loss caused by network congestion. When using it in the WAN, the backpressure from PFC will cause head-of-line blocking and deadlocks, which degrade network throughput. Explicit Congestion Notification (ECN) is an extension to network layer protocol and transport layer protocol defined in [RFC3168], which enables the notification of network congestion. [I-D.geng-fantel-fantel-gap-analysis] points out that ECN still relies on end-to-end signaling and lacks precise real-time feedback.

[I-D.wh-rtgwg-adaptive-routing-arn] specifies a UDP-based Adaptive Routing Notification (ARN) mechanism to proactively disseminate congestion and failure notification. [I-D.hhz-fantel-sar-wan] defines a new ICMPv6 message to realize rapid notification in key traffic engineering areas including failure protection, congestion control, and load balancing. These solutions focus on the construction of notification messages, but give little consideration to their delivery process—especially in inter-domain scenarios, leading to difficulties in achieving end-to-end congestion control.

This document focuses on the scenario of transmitting RDMA packets across multiple DCs over WAN, where SRv6 tunnels are deployed between DCs. It introduces several typical scenarios, and then propuses new SRv6 SIDs and associated processing procedure to enable inter-domain delivery of notification messages.

2. Conventions used in this document

2.1. Abbreviations

DC: Data Center

CNP: Congestion Notification Packet

ECN: Explicit Congestion Notification

PFC: Priority-based Flow Control

RoCEv2: RDMA over Converged Ethernet version 2

WAN: wide area network

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. Scenarios for distributed RoCEv2 network

[I-D.hzh-fantel-wan-tunnel] introduces the main scenarios related to AI services in WAN, as well as the requirements for FANTEL(FAst Notification for Traffic Engineering and Load balancing) in these scenarios. Based on this, this document focuses several key scenarios and how they are carried over SRv6 tunnels in WAN.

3.1. Scenario 1: Directly transmitting sample data to AI servers

When leasing AI facilities in a third-party DC, customers directly upload sample data to the AI servers for model training, in order to prevent data leakage in the third-party DC storage.

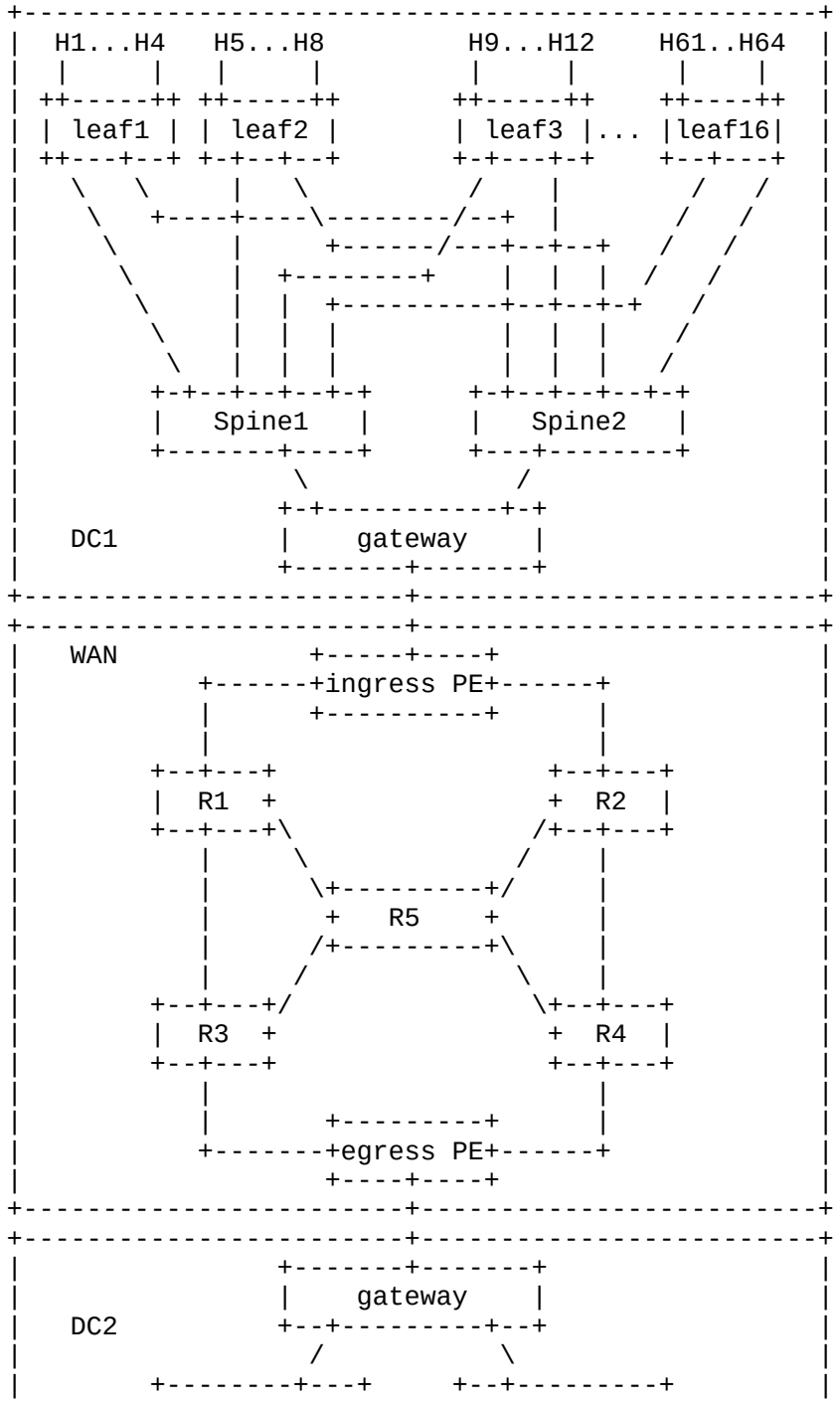
During separated storage and model training, WAN needs to carry sample data from storage clusters to compute clusters through tunnels, and this data is transmitted using RDMA protocols such as RoCEv2.

3.2. Scenario 2: coordinated model training/inference

The computational power growth of a single DC is limited by multiple factors such as space and power consumption. Therefore, customers are inclined to support large AI model training/inference by coordinating distributed model training/inference across multiple AIDCs.

During distributed model training, WAN needs to carry parameter synchronization data between multiple AIDCs through tunnels, and this data is transmitted using RDMA protocols such as RoCEv2.

3.3. Scenario abstraction


H1...H4
H5...H8
H9...H12
H61..H64
leaf1
leaf2
leaf3
leaf16
Spine1
Spine2
gateway
DC1
WAN
ingress PE
R1
R2
R5
R3
R4
egress PE
gateway
DC2

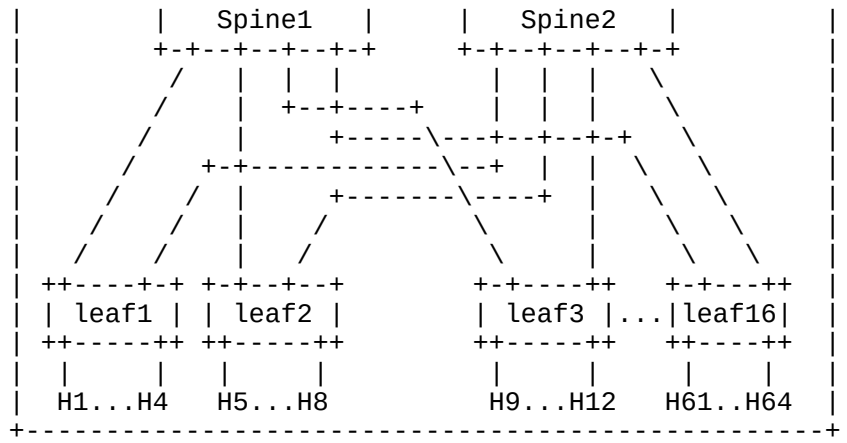


Figure 1: Network diagram

The scenarios mentioned in Chapter 3 can be summarized as the network diagram shown in Figure 1. In these scenarios, the sender in DC needs to transmit RoCEv2 packets to the receiver in another DC through SRv6 tunnels in WAN, the process is as follows:

* The sender (H1-H64) in DC1 sends RoCEv2 packets to WAN’s ingress PE through the leaf, spine, and gateway devices.

* At the WAN’s ingress PE, the RoCEv2 packets are encapsulated according to the SRv6 tunnel protocol.

* The WAN’s transit nodes (R1-R5) transmit the payload from ingress PE to egress PE through SRv6 tunnels.

* At the WAN’s egress PE, the payload are decapsulated to RoCEv2 packets and transmitted to the receiver (H1-H64) in DC2.

4. Inter-domain congestion notification

4.1. Precise flow control

PFC is the feature used in ethernet to prevent data loss due to congestion. In PFC, traffic is classified into 8 priorities according to the 802.1Q protocol, and each priority maintains its queue. When the queue length of the router’s receive port exceeds a certain threshold, the router sends a PAUSE frame to upstream to stop transmitting traffic. When the queue length of the router’s receive port drops below a certain threshold, the router sends a RESUME frame to upstream to continue transmitting traffic.

Since PFC operates at the data link layer, its use in WAN can cause significant impact on other services if deadlock or storms occur. Many recent works aim to define new mechanisms at the network layer or transport layer to address these issues, enabling fine-grained flow control at the service or path level. These new mechanisms typically rely on IPv6 to convey backpressure information to upstream devices.

Since standard PFC is still used within DCs, this document defines an SRv6 SID (END.C) to instruct the ingress PE to perform interoperability between PFC and the new mechanisms in WAN. END.C is a 128-bit value configured on PE and disseminated to other routers via IGP. The endpoint action associated with END.C is to query the priority mapping between congestion notification and PFC, and send the PAUSE or RESUME frame with the corresponding link priority to upstream device.

4.1.1. Process analysis

Taking Figure 1 as an example, within DC1 and DC2, PFC is still used to implement the flow control based on port priority. Within WAN, IPv6-based solutions are used to achieve congestion control at service level or path level, effectively avoiding impact on other services. The congestion handling process is as follows:

* When WAN devices (R1, R2) detect congestion, they encapsulate an outer IP header (destination address is set to END.C) outside the notification and send it back to the ingress PE.

* When the ingress PE detects that the destination address of the received packet hits END.C, and it queries the priority mapping between congestion notification and PFC, and send the PAUSE or RESUME frame with corresponding link priority to gateway.

* When gateway receives the PAUSE or RESUME frame, it suspend or resume traffic transmission for the corresponding link priority.

4.2. Explicit congestion notification

ECN enables a forwarding element (e.g., a router) to notify the sender for congestion control without having to drop packets [RFC3168]. When a router detects congestion, instead of dropping packets as a signal of congestion, it marks the packets with an ECN codepoint in the IP header. 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 slow down the transmission rate temporarily until the flow path is ready to handle a higher rate of traffic.

[RFC7514] defines Really Explicit Congestion Notification (RECN), also known as Fast Congestion Notification Packet (Fast CNP). By extending the RoCEv2 CNP, Fast CNP can be sent by the intermediate router directly to the sender, advising the sender to reduce the transmission rate at which it sends the flow of RoCEv2 data traffic. When transporting RoCEv2 packets through SRv6 tunnels in WAN, intermediate devices are unable to send Fast CNP back to the ingress PE based on the information in RoCEv2 packets. Additionally, the ingress PE is unable to recognize the Fast CNP and therefore cannot forward them to the sender.

Therefore, this document proposes a new SRv6 SID (END.E) to address aforementioned issues. END.E is a 128-bit value configured on PE and disseminated to other routers via IGP. The endpoint action associated with END.E is to decapsulate the packet and then look up the routing table for traffic forwarding.

4.2.1. Process analysis

Taking Figure 1 as an example, when a WAN device detects congestion, it directly sends Fast CNP to the ingress PE for processing, and then the ingress PE forwards it to the sender. The congestion handling process is as follows:

* When WAN devices detect congestion, they encapsulate an outer IP header (destination address is set to END.E) outside the Fast CNP and send it back to the ingress PE.

* When the ingress PE detects that the destination address of the received packet hits END.E, it decapsulates the Fast CNP and pass it to the sender in DC.

* When the sender receives Fast CNP, it reduces the transmission rate of the corresponding flow.

5. Advertising new SRv6 SIDs using IGP

5.1. Advertising new SRv6 SIDs Using IS-IS

Before advertising END.E and END.C, PE nodes should first inform other nodes whether they have the corresponding congestion handling capabilities. This can be achieved by defining a 1-bit flag (i.e., the C-flag) in the Flags field of the SRv6 Capabilities Sub-TLV.

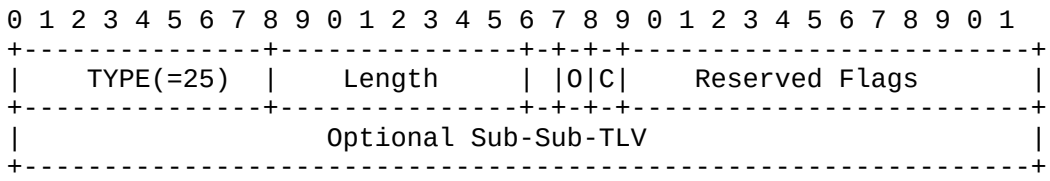


Figure 2: C-Flag in SRv6 Capabilities Sub-TLV

Since both END.C and END.E require specific processing at the node, they can be advertised to other neighbors via the SRv6 END SID Sub-TLV.

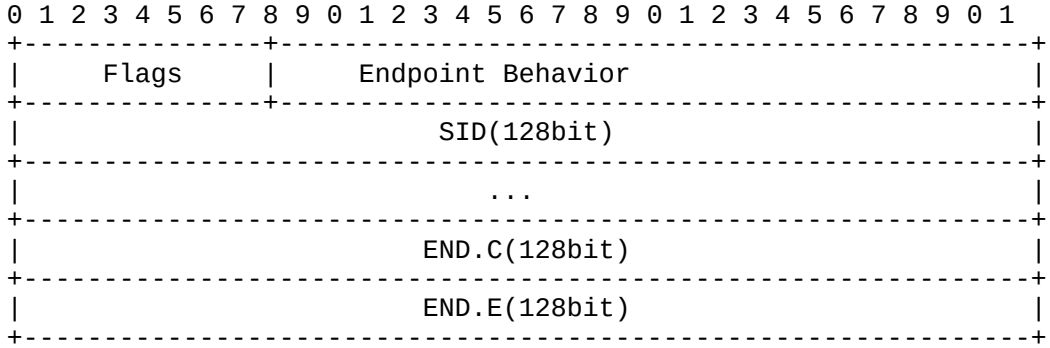


Figure 3: new SRv6 SIDs in SRv6 END SID Sub-TLV

5.2. Advertising new SRv6 SIDs Using OSPFv3

TBD

6. Security Considerations

TBD

7. IANA Considerations

TBD

8. Acknowledgments

TBD

9. References

9.1. 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>.

[RFC3688] Mealling, M., "The IETF XML Registry", BCP 81, RFC 3688, DOI 10.17487/RFC3688, January 2004, <https://www.rfc-editor.org/info/rfc3688>.

[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>.

9.2. Informative References

[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>.

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

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

[I-D.wh-rtgwg-adaptive-routing-arn] Wang, H., Huang, H., Geng, X., Xu, X., and Y. Xia, "Adaptive Routing Notification", Work in Progress, Internet-Draft, draft-wh-rtgwg-adaptive-routing-arn-03, 13 September 2024, <https://datatracker.ietf.org/doc/html/draft-wh-rtgwg-adaptive-routing-arn-03>.

[I-D.geng-fantel-fantel-gap-analysis] Geng, X., Huo, P., Cheng, W., Li, D., Zhu, Y., and H. Zhengxin, "Gap Analysis of Fast Notification for Traffic Engineering and Load Balancing", Work in Progress, Internet-Draft, draft-geng-fantel-fantel-gap-analysis-01, 7 July 2025, <https://datatracker.ietf.org/doc/html/draft-geng-fantel-fantel-gap-analysis-01>.

[I-D.hhz-fantel-sar-wan]
Hu, J., Hu, Z., and Y. Zhu, "FANTEL scenarios and requirements in Wide Area Network", Work in Progress, Internet-Draft, draft-hhz-fantel-sar-wan-00, 6 July 2025, <https://datatracker.ietf.org/doc/html/draft-hhz-fantel-sar-wan-00>.

[I-D.hzh-fantel-wan-tunnel]
Hu, Z., Zhu, Y., Hu, J., and T. Pi, "Fast Notification for Traffic Engineering and Load Balancing for tunnel-based lossless transmission in WAN", Work in Progress, Internet-Draft, draft-hzh-fantel-wan-tunnel-00, 6 July 2025, <https://datatracker.ietf.org/doc/html/draft-hzh-fantel-wan-tunnel-00>.

Authors' Addresses

Zehua Hu
China Telecom
Guangzhou
China
Email: huzh2@chinatelecom.cn

Yongqing Zhu
China Telecom
Guangzhou
China
Email: zhuyq8@chinatelecom.cn

Xuesong Geng
Huawei
China
Email: gengxuesong@huawei.com

Jiayuan Hu
China Telecom
Guangzhou
China
Email: hujy5@chinatelecom.cn

Tanxin Pi
China Telecom
Guangzhou
China

Email: pitx1@chinatelecom.cn