SPRING Working Group
Internet-Draft
Intended status: Standards Track
Expires: 2 September 2026

J. Yang
W. Cheng
M. Zhou
China Mobile
J. Wang
G. Zhang
Centec

1 March 2026

Flow-Level Precision Congestion Control for SRv6 Networks
draft-yang-srv6-precision-flow-control-00

Abstract

This document defines a flow-level precision congestion control mechanism for SRv6 networks. The mechanism specifies new congestion notification message formats that enable per-flow congestion information delivery and hop-by-hop backpressure control. Compared to traditional Priority-based Flow Control (PFC) which operates at the queue level, this mechanism provides finer-grained congestion control suitable for Wide-Area Network (WAN) environments, mitigating head-of-line blocking, congestion spreading, and deadlock issues. The document also describes interoperability models with traditional IEEE 802.1Qbb PFC.

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

Table of Contents

1. Introduction

With the exponential growth of intelligent computing services, scenarios such as distributed AI training, Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCEv2), and disaggregated storage-compute architectures require rigorous lossless transmission of large volumes of bursty traffic. As these services expand beyond data centers across Wide-Area Networks (WANs), maintaining zero-packet-loss guarantees becomes increasingly challenging.

Traditional Priority-based Flow Control (PFC), as defined in IEEE 802.1Qbb, is a Data Link Layer flow control mechanism primarily designed for intra-data center networks. When applied to WAN scenarios with higher Bandwidth-Delay Products (BDP), PFC faces severe structural limitations:

* High Propagation Latency: WAN transmission delays are orders of magnitude larger than those in data center networks. The propagation time required for a PFC PAUSE frame to reach the upstream node often results in severe buffer overflows at the congestion point.

* Coarse Control Granularity: PFC operates globally at the priority queue level. A congestion event triggered by a single micro-burst will cause all flows mapped to that Traffic Class (TC) to be paused, leading to the "collateral damage" known as Head-of-Line (HOL) blocking.

* Deadlock Vulnerability: In complex topologies involving cyclic routing or prolonged congestion, the hop-by-hop queue-level pause nature of PFC frequently leads to unrecoverable cyclic buffer dependencies, i.e., PFC Deadlocks.

To address these limitations, this document proposes a Flow-Level Precision Congestion Control mechanism. Operating within SRv6 networks, it allows network nodes to uniquely identify congested IP flows and explicitly signal upstream nodes to enforce granular rate reduction or pause actions exclusively on the offending flows.

1.1. 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. Terminology

PFC (Priority-based Flow Control): A Link Layer flow control mechanism defined in IEEE 802.1Qbb that pauses transmission of a specific priority queue on a link.

Stream ID: An identifier locally or globally allocated by network nodes to uniquely distinguish an upper-layer micro-flow within the SRv6 routing domain.

PFCM (Precision Flow Control Message): A newly defined IPv6 signaling message (either an ICMPv6 message or an IPv6 Extension Header) used to convey per-flow backpressure signals.

Precision Flow Control Time: The duration for which a targeted congestion control action (e.g., rate reduction or pause) MUST be maintained, measured in microseconds.

3. Protocol Operations

3.1. Architecture Overview

The mechanism operates within standard SRv6 data planes. To support Flow-Level Precision Congestion Control, participating routing nodes are REQUIRED to implement the following functional components:

* Flow Classification and Stream ID Management

* Per-flow state monitoring and buffer threshold management

* PFCM Generation (Downstream Node)

* PFCM Processing and Enforcement (Upstream Node)

3.2. Flow Classification and Stream ID Assignment

Forwarding nodes MUST perform flow classification to distinguish traffic streams. The default classification method SHOULD utilize the IPv6 Flow Label (as defined in [RFC6437]) combined with the Source and Destination IPv6 Addresses.

Alternatively, nodes MAY utilize a classic 5-tuple identifier (Source IP, Destination IP, Protocol, Source Port, Destination Port) where payload inspection is feasible. Implementation-specific classifications (such as Deep Packet Inspection for Layer-7 headers or traffic behavioral heuristics) MAY be used but are strictly outside the scope of this standard.

Upon detecting a stateful flow, the node allocates a unique Stream ID. The Stream ID management strategy can be localized (significant only between two adjacent hops) or globally coordinated (e.g., using an SDN controller across the SRv6 domain).

3.3. Congestion Detection and Forwarding Behavior

The lifecycle of precision congestion control is defined by the following state machine transitions:

1. Congestion Detection (Local State):

   A node actively monitors its egress buffer occupancy for each identified flow. When the instantaneous or average buffer depth for a specific Stream ID exceeds a pre-configured high-water mark threshold, the node transitions to the Congested state.

2. PFCM Generation (Signaling):

The congested node generates a Precision Flow Control Message (PFCM). The PFCM encapsulates the offending Stream ID, the local Queue ID, the requested Action (e.g., reduce rate by 50%), and the Precision Flow Control Time.

3. Reverse Path Transmission:

The PFCM is transmitted to the directly connected upstream node from which the congested flow was received. The PFCM SHOULD be routed to the upstream neighbor’s Link-Local IPv6 address.

4. Upstream Enforcement (Backpressure):

Upon reception of a PFCM, the upstream node parses the Stream ID and maps it to its local forwarding state. It MUST immediately apply the specified Action for the duration of the Precision Flow Control Time. If the upstream node cannot absorb the backpressure locally, it MAY recursively generate a new PFCM to its own upstream node.

3.4. Interoperability with Legacy L2 PFC

Heterogeneous networks may contain legacy devices incapable of L3 per-flow control. To ensure seamless backward compatibility, a border node receiving a PFCM MAY translate the L3 signaling into an IEEE 802.1Qbb L2 PFC frame.

In such translation operations:

* The Queue ID field in the PFCM MUST be directly mapped to the corresponding Class of Service (CoS) priority enable vector in the PFC frame.

* The Precision Flow Control Time (microseconds) MUST be quantized and converted into the standard PFC PAUSE quanta value.

4. Packet Formats

4.1. IPv6 Extension Header Format

Precision flow control telemetry MAY be carried in an IPv6 Hop-by-Hop Options header or Destination Options header ([RFC8200]). This is highly optimal for in-band telemetry or when piggybacked on reverse-path traffic.

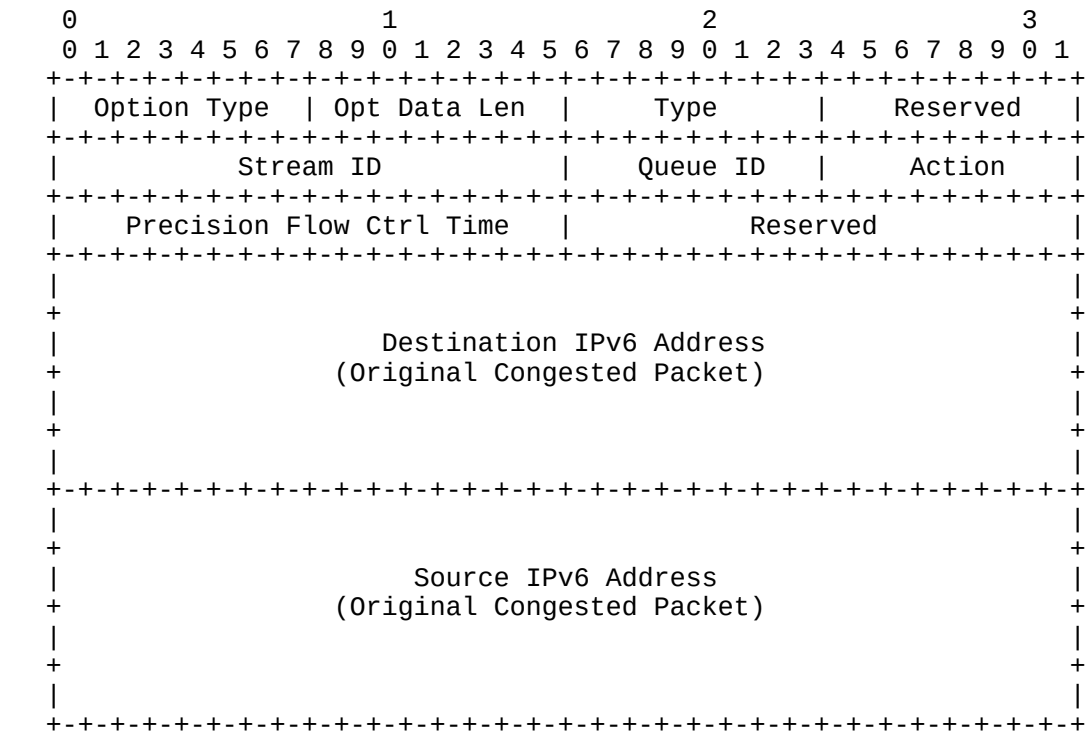

Figure 1: IPv6 Option Format for Precision Flow Control

The fields are defined as follows:

Option Type (8 bits): Identifies the precision flow control option. Value TBA by IANA. The highest-order 2 bits SHOULD be set to '00' (skip over if not recognized).

Opt Data Len (8 bits): Length of the option data in octets, excluding the Option Type and Opt Data Len fields.

Type (8 bits): Sub-type for precision flow control. MUST be set to 0 and reserved for future versioning.

Stream ID (16 bits): The flow identifier causing congestion.

Queue ID (8 bits): The physical or logical priority queue experiencing congestion.

Action (8 bits): Specifies the congestion mitigation directive.

Bits [0:1] specify the action type: 00 = No Backpressure, 01 = Pause Flow, 10 = Reduce Rate. Bits [2:7] represent the rate reduction ratio as an absolute percentage (0-100) when the action type is 10.

Precision Flow Ctrl Time (16 bits): The temporal duration for the specified action, represented in microseconds.

Destination & Source IPv6 Addresses (128 bits each): The IP addresses extracted from the data packet that triggered the congestion event. This allows the upstream node to precisely correlate the telemetry with its local forwarding cache.

4.2. ICMPv6 Message Format

Out-of-band signaling utilizes ICMPv6 messages. This mechanism guarantees delivery independent of reverse-path data traffic availability.

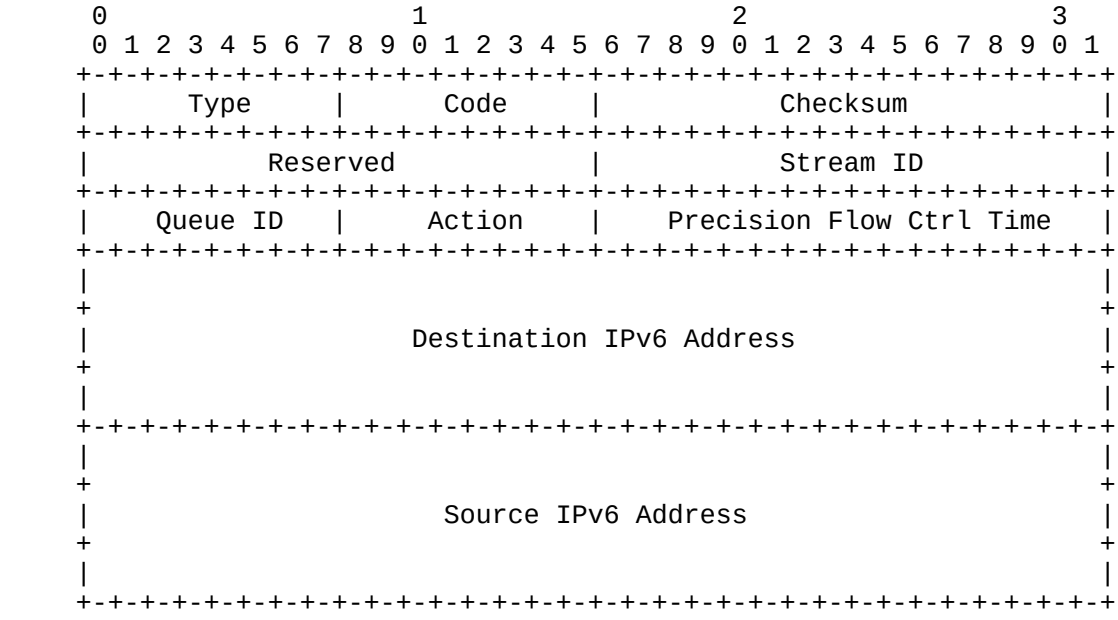

Figure 2: ICMPv6 Message Format for Precision Flow Control

The ICMPv6 header fields are strictly defined as:

Type (8 bits): A new ICMPv6 message type assigned by IANA indicating Precision Flow Control Notification.

Code (8 bits): ICMPv6 message sub-type (0x00 default).

Checksum (16 bits): The standard ICMPv6 checksum ([RFC4443]).

5. Security Considerations

The introduction of L3/L4 flow-level pause and backpressure signaling inherently expands the attack surface of the network architecture. Malicious actors could spoof PFCM packets to arbitrarily pause critical infrastructure flows, leading to a severe Denial of Service (DoS) attack.

To mitigate these threats, the following security constraints MUST be enforced by compliant implementations:

* Hop Limit Verification:

  When processing an ICMPv6 PFCM, a node MUST verify that the IP Hop Limit is exactly 255. Packets arriving with a smaller Hop Limit MUST be silently discarded, guaranteeing that the signal originated from an immediate neighbor.

* Cryptographic Authentication:

  In untrusted or multi-tenant transport domains, the precision flow control messages SHOULD be secured using the IPsec Authentication Header (AH) or Encapsulating Security Payload (ESP) to ensure data integrity and neighbor origin authentication.

* Rate Limiting:

  Nodes MUST implement strict control-plane policing (CoPP) and rate limiting for PFCM processing to prevent CPU resource exhaustion attacks.

6. IANA Considerations

This document requests the following allocations from IANA:

1. A new Option Type in the "Destination Options and Hop-by-Hop Options" registry for the Precision Flow Control Congestion Notification.

2. A new Type value in the "ICMPv6 Type Numbers" registry for the Precision Flow Control Congestion Notification messages.

7. References

7.1. 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>.

[RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <https://www.rfc-editor.org/info/rfc4443>.

[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>.

7.2. Informative References

[RFC6437] Amante, S., Carpenter, B., Jiang, S., and J. Rajahalme, "IPv6 Flow Label Specification", RFC 6437, DOI 10.17487/RFC6437, November 2011, <https://www.rfc-editor.org/info/rfc6437>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

Acknowledgements

The authors would like to thank the contributors and reviewers who provided valuable feedback on this document.

Authors' Addresses

Jin Yang
China Mobile
Beijing
100053
China
Email: yangjinwl@chinamobile.com

Weiqiang Cheng
China Mobile
Beijing
100053
China
Email: chengweiqiang@chinamobile.com

Ming Zhou
China Mobile
Beijing
100053
China
Email: zhoumingyjy@chinamobile.com

Junjie Wang
Centec
Suzhou
215000
China
Email: wangjj@centec.com

Guoying Zhang
Centec
Suzhou
215000
China
Email: zhanggy@centec.com