Congestion Control Working Group
Internet-Draft
Intended status: Informational
Expires: 2 September 2026

Y. Tian
W. Cheng
J. Yang
China Mobile
J. Wang
G. Zhang
Centec
K. Zhang
China Mobile

1 March 2026

Datapath Processing Architecture for In-Band Congestion Signaling (IBCS)
draft-tian-ccwg-ibcs-datapath-processing-00

Abstract

In-band congestion signaling protocols, such as Congestion Signaling (CSIG) and High Precision Congestion Control (HPCC++), require intermediate Network Elements (NEs) to actively parse scalar congestion metrics from packet headers, evaluate them against local link states, and conditionally rewrite these fields before transmission. To ensure end-to-end algorithmic consistency and avoid unintended interactions with routing topologies (e.g., packet reordering), the datapath of these NEs must adhere to a standardized logical processing model.

This document defines the normative datapath processing architecture for Network Elements participating in In-Band Congestion Signaling (IBCS). By establishing abstract topological roles (Edge vs. Transit NEs) and standardizing the "Compare-and-Replace" operational paradigm, this specification abstracts the signal update logic from hardware-specific pipelines. It guarantees strict orthogonality between congestion signaling and Equal-Cost Multi-Path (ECMP) routing invariants, supporting diverse congestion metrics across multi-vendor deployments.

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/.

Copyright Notice

Table of Contents

1. Introduction

Modern high-speed data center networks increasingly rely on fine-grained, In-Band Congestion Signaling (IBCS) to achieve ultra-low latency and high throughput. Protocols being discussed in the IETF, such as CSIG [I-D.ravi-ippm-csig] and HPCC++ [I-D.miao-ccwg-hpcc], utilize packet headers to convey link-level congestion telemetry directly to end-hosts. A fundamental paradigm of these proposals is the "Compare-and-Replace" operation: as a packet traverses the network, each transit Network Element (NE) compares the congestion signal carried in the packet against its own local congestion metric. If the local NE represents a more severe bottleneck, it overwrites the signal field with its local metric.

Unlike traditional stacking-based telemetry (such as IOAM [RFC9197]) where metadata is appended hop-by-hop, the Compare-and-Replace paradigm maintains a constant header size, avoiding Maximum Transmission Unit (MTU) exhaustion. However, updating a packet header on-the-fly introduces significant architectural challenges for datapath pipelines. If the processing behavior is not rigorously defined, modifying packet fields can inadvertently alter hash-based load balancing (ECMP), leading to micro-burst flow reordering. Furthermore, inconsistent state handling at domain boundaries can result in spoofed or corrupted signals reaching the congestion control algorithm.

This document specifies the normative datapath behavior and abstract processing model required to support IBCS safely and efficiently. It introduces a role-based architecture (differentiating edge initialization from transit evaluation) and specifies a protocol-agnostic extremum evaluation model (e.g., evaluating minimum available bandwidth or maximum queue delay). By establishing this unified architectural framework, this document aims to ensure operational interoperability and robust signal delivery across heterogeneous network infrastructures.

2. Terminology

IBCS (In-Band Congestion Signaling): A general mechanism where congestion state metrics are embedded within the data packet header and dynamically updated by Network Elements along the forwarding path.

P_Metric (Packet Metric): The congestion signal value currently carried within the packet header. It represents the most severe bottleneck encountered so far on the packet’s path.

L_Metric (Local Metric): The locally computed congestion metric at

the transit NE's egress port (e.g., residual bandwidth, queue utilization, or link delay).

SUF (Signal Update Function): The abstract logical entity within a Network Element's datapath responsible for evaluating P_Metric against L_Metric and executing the conditional header rewrite.

Extremum Operator: The mathematical comparison operator (MIN or MAX) dictated by the specific signaling protocol's semantics to determine the tightest bottleneck.

2.1. 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.

3. Applicability Statement

This document defines abstract datapath behaviors applicable to single administrative domains (e.g., autonomous data center fabrics) deploying end-to-end congestion control loops based on fixed-length, mutable in-band signals.

This specification explicitly differentiates itself from In-situ Operations, Administration, and Maintenance (IOAM) [RFC9197] and INT [P4-INT]. While IOAM focuses on comprehensive visibility through metadata stacking (Trace Option), the behavior described herein strictly addresses fixed-length "Compare-and-Replace" updates designed specifically for fast-path congestion control algorithms, where per-hop state history is discarded in favor of the path's bottleneck state.

4. Topological Roles and Boundary Behaviors

To guarantee the integrity of the IBCS loop, Network Elements MUST apply different processing rules depending on their topological placement relative to the signaling domain. This document defines three distinct abstract NE roles:

4.1. IBCS Ingress Edge NE

The Ingress Edge NE operates at the boundary where traffic enters the trusted IBCS domain (e.g., a Top-of-Rack switch receiving traffic from a bare-metal server or an untrusted tenant VM).

The Ingress Edge NE MUST inspect arriving packets for existing IBCS fields. To prevent signal spoofing attacks, it MUST act as a signal scrubber: any recognized IBCS field arriving from an untrusted interface MUST be reset to its protocol-defined UNINITIALIZED state before further processing. Only after initialization may the packet be passed to the SUF for its first local evaluation.

4.2. IBCS Transit NE

Transit NEs operate entirely within the trusted boundaries of the IBCS domain (e.g., Spine or Core switches). Transit NEs implicitly trust the P_Metric carried in the packet header.

A Transit NE MUST NOT unconditionally reset or scrub the P_Metric. Its sole responsibility regarding the IBCS field is to execute the strict Compare-and-Replace logic defined in Section 6, ensuring that the metric is only overwritten if the local datapath represents a tighter bottleneck.

4.3. IBCS Egress Edge NE

The Egress Edge NE operates at the boundary where traffic exits the trusted IBCS domain. If the destination is outside the administrative domain and no explicit IBCS peering agreement exists, the Egress Edge NE SHOULD strip or zero-out the IBCS field to prevent internal telemetry leakage to external observers.

5. Abstract Datapath Processing Model

Regardless of the physical hardware pipeline architecture (e.g., run-to-completion, multi-stage ASIC, or programmable switch), the externally observable behavior of any IBCS-enabled Network Element MUST conform to the following abstract sequence. This model ensures that routing invariants are preserved.

5.1. Phase 1: Header Resolution

The datapath parses the designated IBCS header field to extract the current P_Metric. If the NE does not recognize the protocol or the IBCS field is absent, the packet MUST bypass all subsequent IBCS update logic and be forwarded opaquely.

5.2. Phase 2: Strict Forwarding Orthogonality

The packet undergoes routing, access control list (ACL) application, and Equal-Cost Multi-Path (ECMP) or Link Aggregation Group (LAG) path selection.

CRITICAL REQUIREMENT: The IBCS signal update process MUST be strictly orthogonal to path selection. The datapath MUST NOT mutate the P_Metric or any related congestion header fields prior to or during the hash computation phase. Altering header values before ECMP hashing violates fundamental flow invariance, causing packets within the same microflow to traverse asymmetric paths, resulting in TCP/ transport reordering degradation.

5.3. Phase 3: The Signal Update Function (SUF)

Once the deterministic egress port is resolved, the Signal Update Function (SUF) retrieves the real-time L_Metric specifically associated with that port. The SUF evaluates P_Metric against L_Metric and conditionally commits the update to the packet header. This mutation MUST be treated as an atomic transaction applied immediately prior to serialization on the wire.

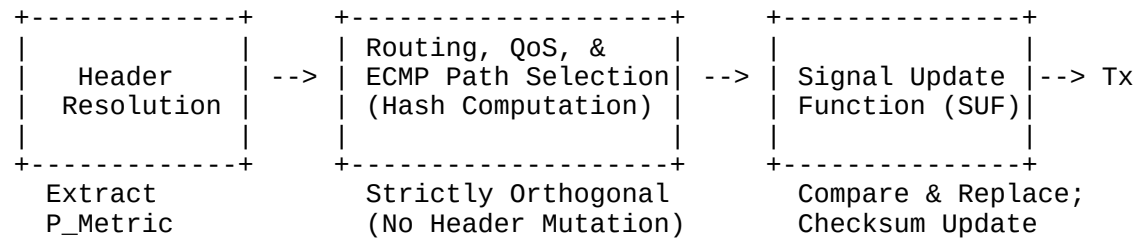

Figure 1: Abstract Datapath Processing Model for IBCS

6. Normative Evaluation Rules (Compare-and-Replace)

6.1. Abstract Extremum Evaluation

Different IBCS protocols characterize congestion semantics differently. For instance, CSIG signals Minimum Available Bandwidth (requiring a MIN operator), whereas HPCC++ may signal Maximum Queue Depth (requiring a MAX operator). The SUF MUST implement a configurable Extremum_Operator to accommodate the semantics of the deployed protocol.

The normative state-machine logic executed by the SUF is defined as follows:

```
Function SUF_Evaluate(P_Metric, L_Metric, Extremum_Operator):
    // Rule 1: Initialization Handling
    IF P_Metric == UNINITIALIZED:
        Rewrite packet header: P_Metric = L_Metric
        RETURN

    // Rule 2: Protocol-Specific Bottleneck Evaluation
    IF Extremum_Operator == MIN:
        IF L_Metric < P_Metric:
            Rewrite packet header: P_Metric = L_Metric

    ELSE IF Extremum_Operator == MAX:
        IF L_Metric > P_Metric:
            Rewrite packet header: P_Metric = L_Metric

    // Rule 3: Preservation
    // If local state is NOT the tighter bottleneck,
    // the header MUST NOT be modified.
```

Atomicity: The rewrite operation MUST be robust. Partial byte updates or malformed header emissions MUST NOT occur, even under extreme internal buffer exhaustion or exception path processing.

6.2. Checksum and Integrity Implications

If the IBCS field is encapsulated within an IPv4 or UDP header, the SUF MUST update the corresponding Layer 3 / Layer 4 checksums. To achieve line-rate processing without introducing significant latency jitter, incremental checksum calculation [RFC1141] is highly RECOMMENDED.

If the IBCS field is embedded in a Layer 2 extension or a custom tag (as commonly deployed in closed data center fabrics), IP/UDP checksum modifications are bypassed, substantially reducing the silicon processing overhead.

7. Operational Considerations

7.1. L_Metric Stability and Sampling Frequencies

The stability of the congestion control loop is inherently tied to how L_Metric is generated. While the specific hardware counter implementation is out of scope for this document, the NE MUST guarantee that L_Metric is relatively stable and decoupled from instantaneous micro-burst noise.

Network Elements SHOULD provide a configurable moving average or sampling window for L_Metric. The optimal sampling interval typically corresponds to the baseline Round-Trip Time (RTT) of the network domain (e.g., 5 to 50 microseconds). Sampling too frequently causes signal oscillation; sampling too slowly creates stale telemetry that dampens transport responsiveness.

7.2. Fail-Open Capability

If a Network Element experiences an internal architectural fault where the real-time L_Metric from the egress port becomes temporarily unavailable to the SUF, the NE MUST NOT drop the packet. Instead, it MUST execute a fail-open behavior, forwarding the packet with the existing P_Metric completely unmodified. This guarantees that transient local datapath faults do not sever the end-to-end signaling loop.

8. Security Considerations

TBD

9. IANA Considerations

This document has no IANA actions.

10. 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>.

11. Informative References

[I-D.ravi-ippm-csig] Ravi, A., Dukkipati, N., Mehta, N., and J. Kumar, "Congestion Signaling (CSIG)", Work in Progress, Internet-Draft, draft-ravi-ippm-csig-01, February 2024, <https://datatracker.ietf.org/doc/draft-ravi-ippm-csig/>.

[I-D.miao-ccwg-hpcc]
Miao, R., "HPCC++: Enhanced High Precision Congestion Control", Work in Progress, Internet-Draft, draft-miao-ccwg-hpcc-03, January 2025, <https://datatracker.ietf.org/doc/draft-miao-ccwg-hpcc/>.

[P4-INT] P4.org, "In-band Network Telemetry (INT) Dataplane Specification, v2.0", February 2020, <https://github.com/p4lang/p4-applications/blob/master/docs/INT_v2_0.pdf>.

[RFC1141] Mallory, T. and A. Kullberg, "Incremental updating of the Internet checksum", RFC 1141, DOI 10.17487/RFC1141, January 1990, <https://www.rfc-editor.org/info/rfc1141>.

[RFC4302] Kent, S., "IP Authentication Header", RFC 4302, DOI 10.17487/RFC4302, December 2005, <https://www.rfc-editor.org/info/rfc4302>.

[RFC9197] Brockners, F., Ed., Bhandari, S., Ed., and T. Mizrahi, Ed., "Data Fields for In Situ Operations, Administration, and Maintenance (IOAM)", RFC 9197, DOI 10.17487/RFC9197, May 2022, <https://www.rfc-editor.org/info/rfc9197>.

Authors' Addresses

Yuchi Tian
China Mobile
Beijing
100053
China
Email: tianyuchi@chinamobile.com

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

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

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

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

Kan Zhang
China Mobile
Beijing
100053
China
Email: zhangkan@chinamobile.com