Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: 8 October 2026

H. Song, Ed.
Futurewei Technologies
Y. Wan
Southeast University
K. Zhu
Huawei Technologies

6 April 2026

Fast Latency and Congestion Notification
draft-song-rtgwg-falcon-00

Abstract

This document describes a standard-based method for fast latency and congestion notification. By combining in-network telemetry and source routing, it enables a source node to acquire a path's latency and congestion status in less than half baseline RTT. The more timely and accurate telemetry data allow the source node to apply more effective traffic steering and congestion control actions. The method is applicable to both WAN and DCN, and can be realized through existing IETF standards.

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 8 October 2026.

Copyright Notice

Table of Contents

1. Introduction

Many congestion control (CC) and load balancing (LB) schemes rely on timely path congestion status and/or the measurement of flow packet delay as the basis for path selection or rate limiting. The problem and requirements are articulated in [I-D.dong-fantel-problem-statement].

However, all conventional methods, either in-band (e.g., In-Network Telemetry or INT) or out-of-band (e.g., ping-mesh), can only return network status or measurements which are at least one RTT old. Unfortunately, RTT is proportional to the path congestion degree. The staleness is aggravated when the path is becoming more congested, which is exactly the moment the real-time network condition is needed the most. For example, a packet experiences serious congestion on its forwarding path and its ECN bit is set. However, due to the congestion, it takes longer than usual time for the signal to be fed back to the source node. When the source node reacts to the "outdated" signal, it might be too late. The problem is more severe in WAN because the RTT can be hundreds of milliseconds. The belated action would be either futile or even counterproductive.

Therefore, it is critical for the source node to know the most up-to-date network congestion status when making reaction decisions. The root cause of the staleness of the status is that the data is

collected on the probe’s forward direction. The lag between the data is collected by the probe and the data is received by the source node is determined by the physical distance as well as the path congestion status.

This document introduces a new method, FALCON (FAst Latency and COgestion Notification), to improve the freshness of the network congestion status sensed by the source node. Specifically, the latency and congestion data are collected on the reverse path toward the source node rather than the packet forwarding path, so the notification lag is reduced to less than half baseline RTT. The method combines the In-Network Telemetry (INT) and Source Routing (SR). A standard compliant implementation can take advantages of IETF IOAM [RFC9197] and SRv6 [RFC8754]. The basic approach is as follows: on the forward direction, the source node initiates INT to track the path by recording the nodes (and ports if necessary); the receiver node uses the INT tracked path to generate the reverse path toward the source node, and use SR plus INT to send a packet to the source node which collects the latency and congestion status along the path.

2. Terminology

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.

2.2. Definitions

This document uses the following terms:

INT:
: In-Network Telemetry. A normal data packet or a dedicated probe packet can carry an instruction header to collect network data on the network nodes along the packet forwarding path. The collected data can be added to the packet. IOAM trace mode [RFC9197] is an INT example.

SR:
: Source Routing. The sender node designates the forwarding path of a packet by listing the network nodes on the path. SRv6 [RFC8754] realizes SR in IPv6.

3. Problem Statement

A network node (source) needs to sense the current network status (e.g., path latency, congestion, etc.) to make timely reaction for traffic toward another network node (destination). Due to the limitation of physics, the lower bound of the data lag is determined by the physical distance between the source and the destination. In reality, the minimum lag also include the basic forwarding delay (excluding queuing delay) per network node on the path. This minimum lag is exactly the half of the minimum RTT (i.e., baseline RTT) and usually less than the half of the actually measured RTT due to queuing delay. We aims to achieve the minimum lag for the latency and congestion notifications.

The solution described in this document MUST satisfy the following requirements:

* Accurate: The notification reflects the true status of the intended path.

* Timely: The notification freshness approaches the theoretical limit.

* Lightweight: The solution requires low packet/bandwidth overhead and low implementation complexity.

* Standard compliant: The solution can be implemented with existing protocols.

4. Solution Overview

FALCON meets all the requirements listed above.

In the forward direction from a sender node S to a receiver node R, a packet P (which can be either a normal packet of a flow F or a dedicated probing packet) is added an INT header which instructs the packet to record the sequence of routers or switches it traverses in order (e.g., S_1, S_2, ...S_n). When R receives P, it constructs a responding packet P' for it. An SR header is added to P' which dictates the path of P' as S_n -> ... -> S_2 -> S_1 -> S (i.e., reverse the path P takes). An INT instruction is also added to P' to ask it to collect the congestion and/or delay information on each router or switch. P' is given a high forwarding priority, so it experiences zero or negligible queuing delay and can be considered to only experience the baseline propagation delay (i.e., the minimum lag).

Specifically, if P' reaches a switch S_i through its port r, we know P in the forward direction leave S_i through the same port r. So the egress queue depth of port r can be used to acquire the more up-to-date congestion status and the queuing delay for the flows on this path.

If the queue depth exceeds the ECN threshold, ECN congestion bit can be set. As a result, S get the ECN status with the minimum lag which is shorter than half RTT. The queuing delay on S_i can be calculated by dividing the queue size in bytes by the egress port bandwidth. If P' acquires and accumulates the queuing delay for each S_i on its path, then the queuing delay of the full path can be accurately approximated at S, and its staleness is also less than half RTT.

5. Detail Description

The path delay for P contains two parts: the baseline propagation delay and the queuing delay. The queuing delay is the good indicator to the congestion degree. It can be acquired throughput the back-tracing packet P'. The propagation delay can also be measured by P'. Since P' suffers no queuing delay, so the delay experienced by P' is exactly the propagation delay which is also identical to the propagation delay of P due to the path symmetry. Thus, we can add the propagation delay of P' to the calculated queuing delay to get the full path delay for P.

If the switches/routers have the computing capability, they can directly calculate the local queuing delay and add it to the accumulated delay from previous switches/routers. Thus, P' needs only carry an accumulated queuing delay value with a smaller packet overhead. When S receives P', it can directly retrieve the path queuing delay. If the congestion control scheme only needs the information (e.g., queue length) on the bottleneck node, the switches/routers can also easily support it by simple compare-and-swap operations.

If the switches/routers have limited computing power, on-path accumulation and aggregation can be infeasible. In this case, P' just needs to collect the queue sizes on each network node (or accumulate them into a single value if all the links have the same bandwidth) and present the data to S to let S calculate the actual queuing delay with the knowledge of the link bandwidth.

6. Implementation and Gap Analysis

INT can use IETF IOAM trace mode [RFC9197] to collect node ID and possibly egress interface. In IPv6-based network, SR can use IETF SRv6 SRH [RFC8754] to backtrack the path.

The switches and routers on path are addressed through IP addresses. Each L2 switch only has a single IP address, so it is enough for P to just record its IP address and use it to construct the SRH for P'. In contrast, each port on a router has an IP address. In this case, P records the egress port's address on each router which is used to construct the SRH for P'. In a managed network, it also possible for P to just record the unique device ID, and the receiver node R, when constructing the SRH, will translate device ID to IP address based on a pre-configured mapping table.

Standard is needed to support IOAM encapsulation in SRv6 packet. A possible solution is given in [I-D.song-spring-siam].

IOAM may need to be extended to support new data types (TBD)

7. Security Considerations

The implementation follows the security considerations for IOAM and SRv6.

8. IANA Considerations

TBD.

9. References

9.1. Normative References

[I-D.dong-fantel-problem-statement]
Dong, J., McBride, M., Clad, F., Zhang, Z. J., Zhu, Y., Xu, X., Zhuang, R., Pang, R., Lu, H., Liu, Y., Contreras, L. M., Mehmet, D., and R. Rahman, "Fast Network Notifications Problem Statement", Work in Progress, Internet-Draft, draft-dong-fantel-problem-statement-05, 2 February 2026, <https://datatracker.ietf.org/doc/html/draft-dong-fantel-problem-statement-05>.

[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>.

[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>.

[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>.

9.2. Informative References

[I-D.song-spring-siam] Song, H., Mishra, G. S., and T. Pan, "SRv6 In-situ Active Measurement", Work in Progress, Internet-Draft, draft-song-spring-siam-07, 4 March 2024, <https://datatracker.ietf.org/doc/html/draft-song-spring-siam-07>.

Authors' Addresses

Haoyu Song (editor)
Futurewei Technologies
United States of America
Email: haoyu.song@futurewei.com

Ying Wan
Southeast University
China
Email: wy25@seu.edu.cn

Keyi Zhu
Huawei Technologies
China
Email: zhukeyi@huawei.com