Routing Area Working Group (RTGWG)
Internet-Draft
Intended status: Standards Track
Expires: 9 November 2026

H. Zhang
Alibaba

8 May 2026

Lossless ECMP Convergence Based on LD/RD Degradation Signal Propagation
draft-zhang-rtgwg-ecmp-lossless-convergence-00

Abstract

This document describes a mechanism for achieving lossless Equal-Cost Multi-Path (ECMP) convergence by utilizing the propagation of Local Degrade (LD) and Remote Degrade (RD) signals defined in the Physical Coding Sublayer (PCS) layer of IEEE 802.3. The mechanism enables proactive ECMP path switching upon detection of link degradation, before an actual link failure occurs, thereby preventing packet loss during routing convergence.

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 9 November 2026.

Copyright Notice

Table of Contents

1. Introduction

In current network implementations, packet loss occurs during the window between a transport-layer link failure and the completion of ECMP convergence at the IP/routing layer. Even with transport-layer protection mechanisms such as OCHP (Optical Channel Protection) or SNCP (Sub-Network Connection Protection), packet loss prior to and during the switchover is unavoidable.

This document introduces a method that utilizes the propagation of Local Degrade (LD) and Remote Degrade (RD) signals to achieve lossless convergence. By detecting signal degradation at the PCS layer and proactively triggering ECMP rerouting before a hard failure occurs, the mechanism eliminates the packet loss window entirely.

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. Problem Statement

2.1. Packet Loss During ECMP Convergence

When a transport link fails or degrades beyond a threshold, the IP/routing layer requires time to detect the failure and reconverge. Failure detection typically relies on BFD session timeouts or underlying link state changes (Link Down). During this convergence window, hashing algorithms may still forward traffic to failed ECMP member paths, resulting in packet loss.

Typical convergence times include:

* BFD detection: 50-900 ms (depending on timer configuration)

* Link Down detection: Depends on hardware interrupt latency 1~5ms and link-down delay configuration 100ms~1s, resulting in a total convergence time of roughly 1s.

* Total packet loss window: 50 ms to 1s

2.2. Limitations of Existing Approaches

Existing fast-reroute mechanisms are triggered only after a failure is detected. They cannot prevent packet loss that occurs between the onset of degradation and the detection event.

BFD-based detection, while effective for hard failures, cannot detect gradual link degradation that causes increasing BER before a complete link failure.

3. Mechanism Overview

3.1. LD/RD Signals in PCS Layer

The IEEE 802.3 [IEEE802.3] standard defines the LD (Local Degrade) and RD (Remote Degrade) indicator bits within the AM_SF (Alignment Marker - Signal Fail) field of the PCS 64/66 encoding. These bits are used at the PCS layer to propagate information about local or remote signal degradation.

* Local Degrade (LD): Indicates that the local port has detected signal degradation (e.g., BER exceeding a pre-configured threshold).

* Remote Degrade (RD): Indicates that the remote end has detected degradation and is notifying the local end via the PCS encoding.

3.2. Soft Reroute Procedure

The lossless convergence mechanism operates as follows:

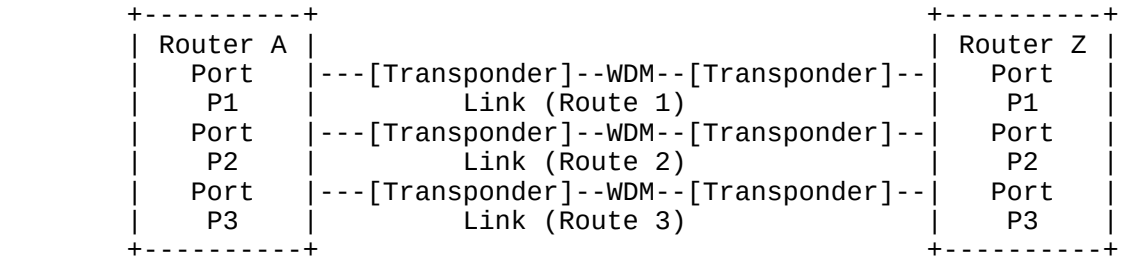
1. On a data communication device (switch/router) port, if the received Bit Error Rate (BER) exceeds a configured threshold, or if a port receives a PCS frame containing an LD indicator bit, the port SHALL report an LD status.

2. Simultaneously, the port SHALL insert an RD indicator into the PCS encoding of its transmitted signals.

3. Upon receiving an RD indicator bit from the remote end, a data communication device port SHALL proactively initiate an ECMP path switch, thereby avoiding packet loss that would otherwise occur after a link failure.

3.3. End-to-End Signal Propagation

The LD/RD mechanism operates across the entire path between two routers, including intermediate transport equipment:



When the Z->A direction of Route 1 degrades:

1. The transport reception port at the degraded segment detects FEC degradation and generates an LD signal.

2. The LD signal propagates via transponder equipment and relays to Port P1 of Router A.

3. Port P1's PCS detects the LD signal. Port P1 of Router A then inserts an RD signal into its transmitted PCS.

4. The RD signal propagates back through the transponder equipment and relays to Port P1 of Router Z.

5. When Port P1 of Router Z's PCS detects the RD signal, the switching ASIC detects the RD-induced interruption reported by P1.

6. The router control plane removes the P1 path from the ECMP group, and the switching ASIC redistributes the transmitted traffic to Route 2 and Route 3.

7. Following this process, the ECMP switchover for the degraded path is complete. Route 1 carries no traffic.

8. When Route 1 recovers, the LD and RD conditions are cleared, and the control plane re-adds Route 1 to the ECMP group.

3.4. LD/RD Propagation in OTN Networks

Router vendors require the LD/RD feature to be applied to OTN [G709] networks interconnecting two routers. Specifically:

* On the line side of transport equipment, the overhead of the line signal (e.g., OTN ODU overhead) can be used to carry the LD and RD signals for each client signal.

* When the line side of the transport equipment detects a BER exceedance, it SHOULD insert an LD indicator into the signal sent out from the client side.

* When the client side of the transport equipment detects a BER exceedance, it SHALL mark the local ingress signal as LD, and the Ethernet PCS transmitted to the remote end SHALL also carry the LD indicator.

3.5. LD/RD Propagation in ZR Applications

When a 400ZR [OIF400ZR] or 800ZR coherent optics module is directly plugged into a router or switch port, the transport-layer LD/RD functionality is implemented within the ZR module itself, eliminating the need for external transponder equipment.

In this scenario, the ZR module performs the same role as a transport device:

* When the ZR module detects line-side FEC degradation (BER exceeding the configured threshold), it SHALL generate an LD signal and propagate it to the host router/switch via the client-side PCS interface.

* Upon receiving an RD signal from the host router/switch via the client-side PCS, the ZR module SHALL insert the corresponding degradation indicator into the line-side signal toward the remote end.

The implementation within ZR modules is analogous to that of OTN transport equipment, with the key difference being that the ZR module is co-located with the router/switch, reducing the propagation delay between degradation detection and ECMP switchover action.

4. Performance Requirements

4.1. Timing Requirements

Lossless convergence based on LD/RD signal propagation requires that the total "detection-propagation-action" latency be within milliseconds. This includes:

* BER/LD detection time at data communication ports

* RD insertion latency

* LD/RD generation and propagation by intermediate transport equipment

* ECMP hardware switchover time

4.2. Hardware Switchover Performance

Two switchover scenarios are considered based on equipment capabilities:

In the first scenario, considering extreme cases in production networks where a link may experience instantaneous interruption, the ECMP switchover must be completed within a very short time to avoid packet loss. This requires hardware-level ECMP switchover capability, with a target completion time of 100 microseconds.

In the second scenario, for data communication equipment that lacks hardware-based ECMP switchover capability, the ECMP switchover can be performed by the CPU control plane. This approach can cover slower interruptions in production networks, which represent the majority of interruption scenarios. In this case, the CPU control plane switchover time should be controlled to approximately 10 milliseconds.

* ECMP hardware switchover: MUST complete within 100 microseconds

* CPU control plane switchover: SHOULD be controlled to approximately 10 milliseconds

4.3. Comparison with BFD-based and Link down-based Detection

Metric	BFD-based	Link down-based	LD/RD-based
Detection granularity	Link failure	Link failure	Link degradation
Detection time	50-900 ms	< 1 s	< 1 ms
Protocol overhead	BFD packets	LF/RF packets	None (in-band)
Hardware dependency	CPU/ software	CPU	PCS hardware
Packet loss	Yes	Yes	No (proactive)

Table 1

5. Implementation Considerations

5.1. Data Communication Equipment Requirements

Data communication equipment (routers/switches) MUST support:

* LD detection based on configurable BER threshold
* RD signal insertion in PCS encoding upon LD detection
* ECMP path removal triggered by RD reception
* ECMP path re-addition upon LD/RD clearance

5.2. Transport Equipment Requirements

Transport equipment MUST support:

* LD/RD signal passthrough on client-side PCS
* LD generation based on line-side and client-side FEC degradation detection
* LD/RD propagation via OTN overhead on line-side
* End-to-end LD/RD signal relay across multiple spans

5.3. Interoperability

The mechanism requires coordinated support from both data communication equipment and transport equipment. Multi-vendor interoperability testing is RECOMMENDED before deployment.

6. Security Considerations

The LD/RD signals are carried within the PCS layer encoding and are not accessible to higher-layer protocols. However, the following security aspects should be considered:

* A malicious device could inject false LD/RD signals to trigger unnecessary ECMP rerouting, potentially causing traffic oscillation.
* Implementations SHOULD include hysteresis mechanisms for the assertion and clearance of LD triggered by BER, to prevent rapid ECMP path flapping caused by intermittent LD/RD signals

7. IANA Considerations

This document has no IANA actions.

8. Acknowledgements

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>.

[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>.

[IEEE802.3] IEEE, "IEEE Standard for Ethernet", 2022.

[G709] ITU-T, "Interfaces for the optical transport network", ITU-T Recommendation G.709, June 2020.

[OIF400ZR] OIF, "Implementation Agreement 400ZR", November 2022.

9.2. Informative References

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <https://www.rfc-editor.org/info/rfc5880>.

[RFC7880] Pignataro, C., Ward, D., Akiya, N., Bhatia, M., and S. Pallagatti, "Seamless Bidirectional Forwarding Detection (S-BFD)", RFC 7880, DOI 10.17487/RFC7880, July 2016, <https://www.rfc-editor.org/info/rfc7880>.

Appendix A. Implementation Status

[Note to RFC Editor: Please remove this section before publication.]

Multiple vendors' various models of data communication and transport equipment already support the key capabilities described in this document.

Appendix B. Relationship to IEEE 802.3

The LD and RD indicator bits are defined in the IEEE 802.3 [IEEE802.3] standard within the PCS 64/66 encoding. This document does not propose any modifications to the IEEE 802.3 standard. Rather, it defines how existing LD/RD signals should be utilized by the IP/routing layer to achieve lossless ECMP convergence.

Appendix C. Relationship to ITU-T G.709

ITU-T G.709 [G709] defines the OTN (Optical Transport Network) frame structure and overhead, including the mechanisms for transporting client-side LD and RD signals over the line side of transport equipment. This document does not propose any modifications to the definitions in ITU-T G.709 [G709]. Rather, it leverages these existing definitions to convey transport link degradation information to data communication equipment, enabling the data communication equipment to perform fast link convergence based on the received degradation signals.

Appendix D. Relationship to BFD

BFD RFC5880 [RFC7880] provides bidirectional forwarding detection but operates at a higher layer and with slower detection times. The LD/RD mechanism described in this document complements BFD by providing faster, hardware-based degradation detection. Both mechanisms MAY be deployed simultaneously for defense-in-depth.

Author’s Address

Huan Zhang
Alibaba
Email: yuanjing.zh@alibaba-inc.com