Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: 4 September 2025

D. Li
Tsinghua University
K. Gao
S. Wang
L. Chen
Zhongguancun Laboratory
X. Geng
Huawei

3 March 2025

Fast Reroute based on Programmable Data Plane (PDP-FRR)
draft-li-fantel-pdp-frr-00

Abstract

This document introduces a fast reroute architecture within the programmable data plane (PDP-FRR) for enhancing network resilience through rapid failure detection and swift path migration, leveraging in-band network telemetry and source routing. Unlike traditional methods that rely on the control plane and face significant delays in rerouting, the proposed architecture utilizes a white-box modeling of the data plane to distinguish and analyze packet losses accurately, enabling immediate identification for link failures (including black-hole and gray failures). By utilizing in-band network telemetry and source routing, the proposed solution significantly reduces reroute times to a few milliseconds, offering a substantial improvement over existing practices and marking a pivotal advancement in failure tolerance.

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 4 September 2025.

Table of Contents

1. Introduction

In the rapidly evolving landscape of network technologies, ensuring the resilience and reliability of data transmission has become paramount. Traditional approaches to network failure detection and rerouting, heavily reliant on the control plane, often suffer from significant delays due to the inherent latency in failure notification, route learning, and route table updates. These delays can severely impact the performance of time-sensitive applications, making it crucial to explore more efficient methods for failure tolerance. Fast reroute based on programmable data plane (PDP-FRR) architecture leverages the capabilities of the programmable data plane to significantly reduce the time required to detect link

failures and reroute traffic, thereby enhancing the overall robustness of datacenter networks.

PDP-FRR architecture stands at the forefront of innovation by integrating in-band network telemetry (INT [RFC9232]) with source routing (SR [RFC8402]) to facilitate rapid path migration directly within the data plane. Unlike traditional schemes that treat the data plane as a "black box" and struggle to distinguish between different types of packet losses, PDP-FRR adopts a "white box" modeling of the data plane's packet processing logic. This allows for a precise analysis of packet loss types and the implementation of targeted statistical methods for failure detection. By deploying packet counters at both ends of a link and comparing them periodically, PDP-FRR can identify failure-induced packet losses with unprecedented speed and accuracy.

Furthermore, by pre-maintaining a path information table and utilizing SR (e.g., SRv6 [RFC8986] and SR-MPLS [RFC8660]), PDP-FRR architecture enables the sender to quickly switch traffic to alternative paths without the need for control plane intervention. This not only circumvents the delays associated with traditional control plane reroute but also overcomes the limitations of data plane reroute schemes that cannot pre-prepare for all failure scenarios. The integration of INT allows for real-time failure notification, making it possible to control traffic recovery times within a few milliseconds, significantly faster than conventional methods. This document details the principles, architecture, and operational mechanisms of PDP-FRR, aiming to contribute to the development of more resilient and efficient datacenter networks.

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

Packet Counters: The counter or data structure used to count the number of passed packets in a given time interval.

Path Information Table: The table maintained by the sender that contains information about the available paths and their associated metrics.

Upstream Meter (UM): The meter used to measure the number of packets

passing through the upstream egress port of a link.

Downstream Meter (DM): The meter used to measure the number of packets passing through the downstream ingress port of a link.

FDM-U: The FDM agent deployed on the upstream switch, it is used to generate probe packets to collect UM and DM.

FDM-D: The FDM agent deployed on the downstream switch, it is used to generate response packets to feedback UM and DM.

3. PDP-FRR Architecture Overview

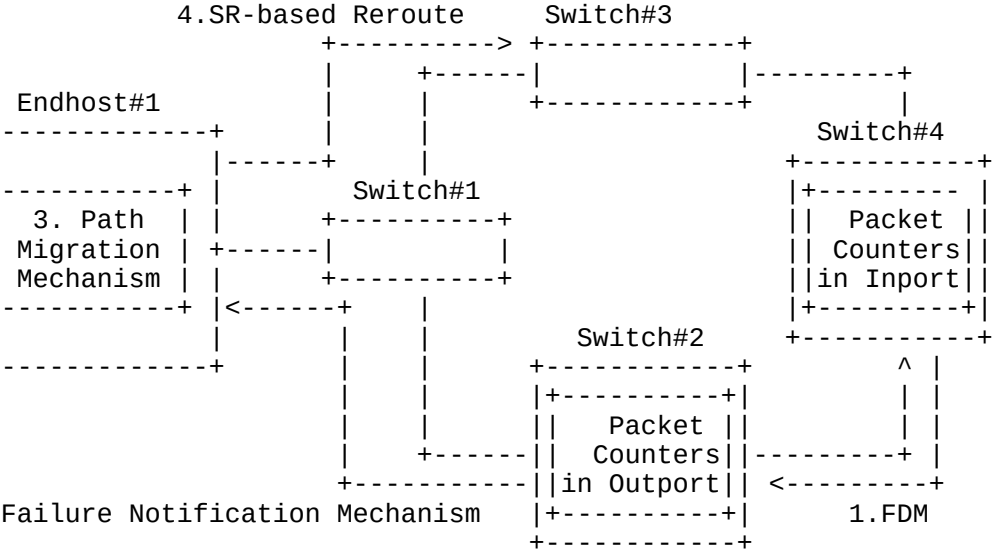

Figure 1: PDP-FRR Architecture.

Traditional network failure detection methods generate probe packets through the control plane (such as BFD [RFC5880]), treating the network data plane as a "black box". If there is no response to a probe, it is assumed that a link failure has occurred, without the ability to distinguish between fault-induced packet loss and non-fault packet loss (such as congestion loss, policy loss, etc.). PDP-FRR models the packet processing logic in the data plane as a white box, analyzing all types of packet loss and designing corresponding statistical methods. As shown in Figure 1, PDP-FRR deploys packet counters at both ends of a link, which tally the total number of packets passing through as well as the number of non-fault packet losses, periodically comparing the two sets of counters to precisely measure fault-induced packet loss. This method operates entirely in

the data plane, with probe packets directly generated by programmable network chips (e.g., P4), thus allowing for a higher frequency of probes and the ability to detect link failures within a millisecond.

After detecting a link failure, PDP-FRR enables fast path migration in data plane by combining INT with source routing. As shown in Figure 1, after a switch detects a link failure, it promptly notifies the sender of the failure information using INT technology; the sender then quickly reroutes the traffic to another available path using source routing, based on a path information table maintained in advance. All processes of this method are completed in the data plane, allowing traffic recovery time to be controlled within a few RTTs (on the order of milliseconds).

In summary, PDP-FRR architecture involves accurately detecting link failures within the network, distinguishing between packet losses caused by failures and normal packet losses, and then having switches convey failure information back to the end hosts via INT [RFC9232]. The end hosts, in turn, utilize SR (e.g., SRv6 [RFC8986] and SR-MPLS [RFC8660]) to change the paths used by the traffic. Therefore, the PDP-FRR architecture comprises three processes.

4. Failure Detection Mechanism

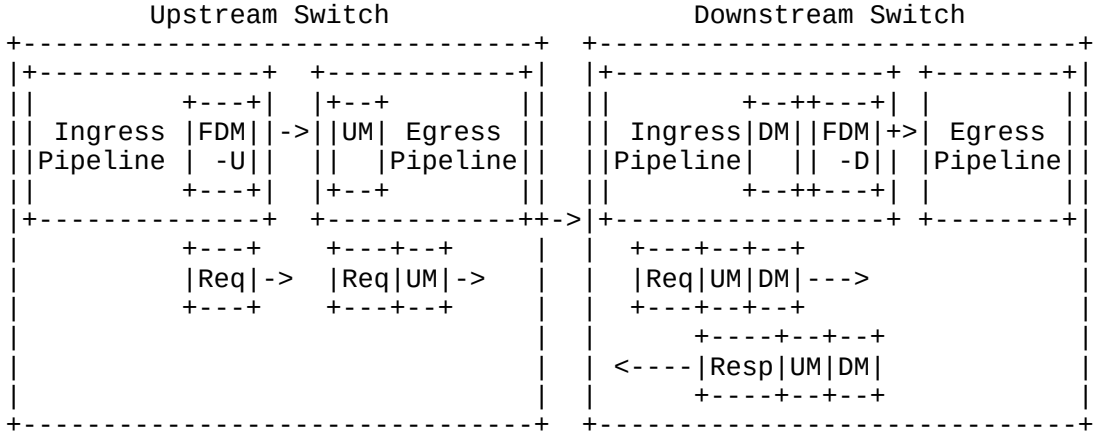

Figure 2: Failure Detection Mechanism: counter deployment locations and request packet generation.

This document designs a failure detection mechanism (FDM) based on packet counters, leveraging the programmable data plane. As shown in Figure 2, this mechanism employs counters at both ends of a link to tally packet losses. So adjacent switches can collaborate to detect failures of any type (including gray failures), and the mechanism is capable of accurately distinguishing non-failure packet losses, thus avoiding false positive.

4.1. Counter Deployment

FDM places a pair of counter arrays on two directly connected programmable switches to achieve rapid and accurate failure detection. Figure 2 illustrates the deployment locations of these counters, which include two types of meter arrays: (1) the Upstream Meter (UM) is positioned at the beginning of the egress pipeline of the upstream switch; (2) the Downstream Meter (DM) is located at the end of the ingress pipeline of the downstream switch. Each meter records the number of packets passing through. With this arrangement, the difference between UM and DM represents the number of packets lost on the link. It is important to note that packets dropped due to congestion in the switch buffers are not counted, as the counters do not cover the buffer areas.

Furthermore, to exclude packet losses caused by non-failure reasons, each meter array includes some counters to tally the number of non-failure packet losses (e.g., TTL expiry). Therefore, FDM is capable of accurately measuring the total number of failure-induced packet losses occurring between UM and DM, including losses due to physical device failures (e.g., cable dust or link jitter) and control plane oscillations (e.g., route lookup misses).

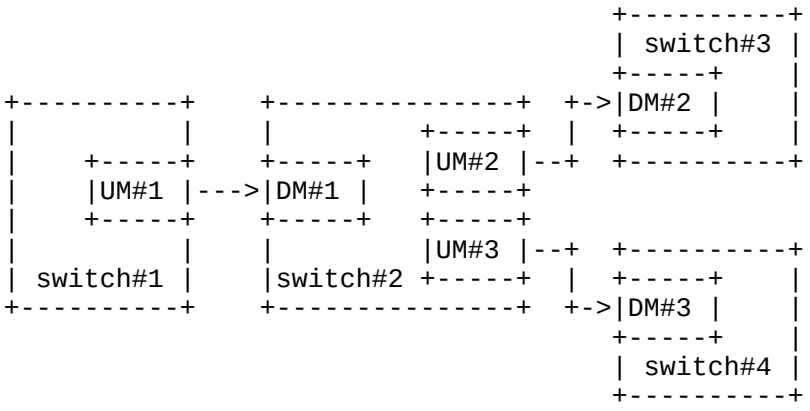

Figure 3: FDM (UM and DM) deployment on all network links.

Figure 3 illustrates the deployment method of FDM across the entire datacenter network. Similar to the BFD mechanism, FDM needs to cover every link in the network. Therefore, each link in the network requires the deployment of a pair of UM and DM. It is important to note that although only the unidirectional deployment from Switch#1 to Switch#2 is depicted in Figure 3, Switch#2 also sends traffic to Switch#1. To monitor the link from Switch#2 to Switch#1, FDM deploys a UM on the egress port of Switch#2 and a DM on the ingress port of Switch#1. Consequently, FDM utilizes two pairs of UM and DM to monitor a bidirectional link.

4.2. Counter Comparison

As shown in Figure 2, the FDM agent in the upstream switch (FDM-U) periodically sends request packets to the link's opposite end. These request packets record specific data of UM and DM along the path through the INT mechanism. Upon detecting the request packets, the FDM agent in the downstream switch (FDM-D) immediately modifies them as response packets and bounces them back, allowing the packets containing UM and DM data to return to the FDM-U. Subsequently, the FDM-U processes the response packets and calculates the packet loss rate of the link over the past period. If FDM-U continuously fails to receive a response packet, indicating that either the response or request packets are lost, then FDM-U considers the packet loss rate of that link to be 100%. This can be used to detect black-hole failure in the link. In other scenarios, if the packet loss rate exceeds a threshold (e.g., 5%) for an extended period, FDM-U will mark that outgoing link as failure.

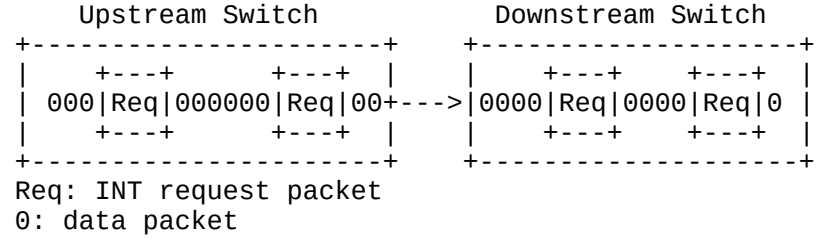

Figure 4: An example for illustrating the batch synchronization provided by request packets.

To ensure the correctness of packet loss rate statistics, FDM must ensure that the packets recorded by UM and DM belong to the same batch. Upon closer analysis, it's found that request packets provide native batch synchronization, and FDM only needs to reset the counters upon receiving a request packet and then start counting the new batch. Specifically, since packets between two directly connected ports do not get out of order, the sequence of packets

passing through UM and DM is consistent. As shown in Figure 4, the request packets serve to isolate different intervals and record the number of packets in the right interval. When such a request packet reaches the downstream switch, the DM records the number of packets for the same interval. Thus, UM and DM count the same batch of packets. However, the loss of request packets would disrupt FDM's batch synchronization. To avoid this, FDM configures active queue management to prevent the dropping of request packets during buffer congestion. If a request packet is still lost, it must be due to a failure.

4.3. Failure Recovery Detection

To ensure stable network operation after failure recovery, FDM also periodically monitors the recovery status of links. This requires the FDM-U to send a batch of test packets, triggering UM and DM to count. Then, the FDM-U sends request packets to collect data from UM and DM. If the link's packet loss rate remains below the threshold for an extended period, FDM-U will mark the link as healthy. To reduce the bandwidth overhead of FDM, considering that the detection of failure recovery is not as urgent as failure detection, FDM can use a lower recovery detection frequency, such as once every second.

4.4. An Example

This section presents an example of how FDM calculates the packet loss rate of a link. Assume that 100 packets pass through the upstream switch UM, which records [100,0], with 0 representing no non-fault-related packet loss. Suppose 8 packets are dropped on the physical link and 2 packets are dropped at the ingress pipeline of the downstream switch due to ACL rules. Then, the DM records [90,2], where 90 represents the number of packets that passed through DM, and 2 represents the number of packets dropped due to non-fault reasons. Finally, by comparing the UM with DM, FDM calculates the packet loss rate of the link as 8% ((100-90-2)/100), rather than 10%.

5. Failure Notification Mechanism

Traditional control plane reroute schemes require several steps after detecting a failure, including failure notification, route learning, and routing table updates, which can take several seconds to modify traffic paths. Data plane reroute schemes, on the other hand, cannot prepare alternative routes for all possible failure scenarios in advance. To achieve fast reroute in the data plane, PDP-FRR combines INT with source routing to quickly reroute traffic.

Assume that the sender periodically sends INT probe packets along the path of the traffic to collect fine-grained network information, such as port rates, queue lengths, etc.. After a switch detects a link failure, it promptly notifies the sender of the failure information within the INT probe. Specifically, when a probe emitted by an end host is about to be forwarded to an egress link that has failed, PDP-FRR will immediately bounce the probe back within the data plane and mark the failure status in the probe. Finally, the probe with the failure status will return to the sender.

6. Path Migration Mechanism

To enable sender-driven fast reroute within data plane, the sender needs to maintain a path information table in advance so that it can quickly switch to another available path upon detecting network failure. Specifically, within the transport layer protocol stack of the sender, this document designs a Path Migration Mechanism (PMM), which periodically probes all available paths to other destinations. Of course, this information can also be obtained through other means, such as from an SDN controller. Then, for a new flow, the sender will randomly select an optimal available path from the path information table and use source routing (e.g., SRv6 [RFC8986] and SR-MPLS [RFC8660]) to control the path of this flow. Similarly, the sender also controls the path of the INT probes using source routing, allowing them to probe the path taken by the traffic flow. The fine-grained network information brought back by these probes can be used for congestion control, such as HPCC [hpcc].

When the above FDM mechanism is effective, and the INT information makes the sender aware of a failure on the path, the sender will immediately mark this path as faulty in the path information table and choose other available paths, accordingly modifying the source routing headers of both the data packets and the INT probes. To promptly understand the availability of other paths, PMM will periodically probe other paths and update the path information table, including failure entering and recovering.

7. Security Considerations

TBD.

8. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

Acknowledgements

TBD.

References

Normative References

[RFC9232] Song, H., Qin, F., Martinez-Julia, P., Ciavaglia, L., and A. Wang, "Network Telemetry Framework", RFC 9232, DOI 10.17487/RFC9232, May 2022, <https://www.rfc-editor.org/rfc/rfc9232>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/rfc/rfc8986>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/rfc/rfc8660>.

[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/rfc/rfc8402>.

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

[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/rfc/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/rfc/rfc8174>.

Informative References

[hpcc] "Inband Telemetry for HPCC++", 2024, <https://datatracker.ietf.org/doc/draft-miao-ccwg-hpcc-info/>.

Authors' Addresses

Dan Li
Tsinghua University
Beijing
China
Email: tolidan@tsinghua.edu.cn

Kaihui Gao
Zhongguancun Laboratory
Beijing
China
Email: gaokh@zgclab.edu.cn

Shuai Wang
Zhongguancun Laboratory
Beijing
China
Email: wangshuai@zgclab.edu.cn

Li Chen
Zhongguancun Laboratory
Beijing
China
Email: lichen@zgclab.edu.cn

Xuesong Geng
Huawei
Beijing
China
Email: gengxuesong@huawei.com