ITUPublications

International Telecommunication Union
Standardization Sector

ITU-T Technical Report

(11/2025)

QSTR.MPLRA

Requirements and architecture for monitoring packet loss caused by network congestion

Technical Report ITU-T QSTR.MPLRA

Requirements and architecture for monitoring packet loss caused by network congestion

Summary

Real-time monitoring of network congestion status and trend analysis， as well as accurate congestion level assessment, serve as a critical foundation for network planning, capacity expansion, and optimization strategies.

This Technical Report studies the requirements and architecture for monitoring packet loss caused by network congestion. The scope of this technical report includes the following aspects: standardized work in the related SDOs, general requirements for monitoring packet loss caused by network congestion, architecture for monitoring packet loss caused by network congestion, interfaces and protocols for monitoring packet loss caused by network congestion and security considerations.

Keywords

Network congestion, packet loss monitoring, requirements and architecture.

Note

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Change Log

This document contains Version 1.0 of the ITU-T Technical Report QSTR.MPLRA "*Requirements and architecture for monitoring packet loss caused by network congestion*" approved at ITU-T SG11 meeting held in Geneva from 17 to 26 November 2025.

Editor:	Jinyou DAI	Email: djy@fiberhome.com
	China Information Communication Technologies Group	
Editor:	Xiaoming HE	Email: hexm4@chinatelecom.cn
	China Telecommunications Corporation	
Editor:	Yongsheng LIU	Email: liuys170@chinaunicom.cn
	China Unicom	
Editor:	Minrui SHI	Email: shimr@chinatelecom.cn
	China Telecommunications Corporation	

Table of Contents

Technical Report ITU-T QSTR.MPLRA

Requirements and architecture for monitoring packet loss caused by network congestion

1 Scope

This Technical Report studies the requirements and architecture for monitoring packet loss caused by network congestion.

The scope of this Technical Report includes the following aspects:

– Standardized activities in the related SDOs;

– Potential requirements for monitoring packet loss caused by network congestion;

– Architecture for monitoring packet loss caused by network congestion;

– Interfaces and protocols for monitoring packet loss caused by network congestion;

– Security considerations.

2 References

[ITU-T Y.1540] Recommendation ITU-T Y.1540 (2019), *Internet protocol data communication service – IP packet transfer and availability performance parameters.*

[gRPC] gRPC (2018), *gRPC Network Management Interface (gNMI) v0.6.0*.

[IETF RFC 0959] IETF RFC 0959 (1985), *File Transfer Protocol*.

[IETF RFC 1350] IETF RFC 1350 (1992), *The TFTP Protocol (Revision 2)*.

[IETF RFC 6241] IETF RFC 6241 (2011), *Network Configuration Protocol (NETCONF)*.

[IETF RFC 7799] IETF RFC 7799 (2016), *Active and Passive Metrics and Methods (with Hybrid Types In-Between)*.

[IETF RFC 8639] IETF RFC 8639 (2019), *Subscription to YANG Notifications*.

[IETF RFC 8640] IETF RFC 8640 (2019), *Dynamic Subscription to YANG Events and Datastores over NETCONF*.

3 Definitions

3.1 Terms defined elsewhere

This Technical Report uses the following terms defined elsewhere:

3.1.1 configured subscription [IETF RFC 8639]: A subscription installed via configuration into a configuration datastore.

3.1.2 dynamic subscription [IETF RFC 8639]: A subscription created dynamically by a subscriber via a remote procedure call (RPC).

3.2 Terms defined in this Technical Report

This Technical Report defines the following terms:

3.2.1 network telemetry: The process for acquiring and utilizing network data remotely for network monitoring and operation.

3.2.2 microburst: The phenomenon of device port receiving a large amount of burst data in a very short period of time (such as milliseconds duration), which results in instantaneous burst rates much higher than the average port rate, and even exceeding the port bandwidth.

4 Abbreviations and acronyms

This Technical Report uses the following abbreviations and acronyms:

ACL	Access Control List
ALR	Aggregate Loss Ratio
CPU	Central Processing Unit
DDoS	Distributed Denial of Service
eMMB	enhanced Mobile Broadband
EMS	Element Management System
ETH-LM	Ethernet Frame Loss Measurement
FIB	Forwarding Information Base
FTP	File Transfer Protocol
gNMI	gRPC Network Management Interface
gRPC	Google Remote Procedure Call
GRE	General Routing Encapsulation
HTTP	HyperText Transfer Protocol
IMT-2020	International Mobile Telecommunications for 2020
ID	Identifier
IOAM	In-situ Operation, Administration and Maintenance
IP	Internet Protocol
IPLR	IP Packet Loss Ratio
ITU-T	International Telecommunication Union – Telecommunication Standardization Sector
JSON	JavaScript Object Notation
MAC	Media Access Control
MEP	Measurement End Point
MP	Measurement Point
MPLS	Multi-Protocol Label Switching
MTU	Maximum Transmission Unit
NETCONF	Network Configuration Protocol
NMS	Network Management System
NTP	Network Time Protocol
OAM	Operation, Administration and Maintenance
OWAMP	One-Way Active Measurement Protocol
PLR	Packet Loss Ratio
QoE	Quality of Experience

QoS	Quality of Service
REST	Representational State Transfer
RPC	Remote Procedure Call
SNMP	Simple Network Management Protocol
SDO	Standards Developing Organization
SLA	Service Level Agreement
SLO	Service Level Objective
SR	Segment Routing
SR-MPLS	Segment Routing-Multi-Protocol Label Switching
SRv6	Segment Routing over IPv6
SRH	Segment Routing Header
STAMP	Simple Two-way Active Measurement Protocol
TWAMP	Two-Way Active Measurement Protocol
TCP	Transmission Control Protocol
TFTP	Trivial File Transfer Protocol
UDP	User Datagram Protocol
uRLLC	Ultra-Reliable Low Latency Communication
VPN	Virtual Private Network
VXLAN	Virtual extensible Local Area Network
WG	Working Group
XML	Extensible Markup Language

5 Conventions

None.

6 Overview of real-time monitoring of packet loss

In the International Mobile Telecommunications for 2020 (IMT-2020) era, emerging services including enhanced mobile broadband (eMBB) and ultra-reliable low latency communication (uRLLC) have imposed stringent requirements on Internet protocol (IP) bearer network performance, demanding significantly reduced latency, minimized jitter, and near-zero packet loss rates. However, the inherent statistical multiplexing nature of transmission control protocol (TCP)/IP-based IP networks results in bursty traffic patterns, making network congestion an inevitable occurrence. Such congestion phenomena degrade network performance and introduce service delivery uncertainties. To mitigate these service quality risks, real-time monitoring of network congestion status and trend analysis become imperative. This enables accurate congestion level assessment, which serves as a critical foundation for network planning, capacity expansion, and optimization strategies.

Real-time monitoring of packet loss caused by congestion is also valuable to network operators in many aspects, including quickly locating the congestion nodes and links; optimizing routing path for loss-sensitive traffic flows once congestion loss is detected.

More importantly, based on real-time congestion loss statistics, the nature of congestion can be determined. For instance, if the number of dropped packets is persistently increasing for a longer time

(e.g., at the level of hours), a long-lived congestion event may occur; or, if packet loss happens in a very short time (e.g., milliseconds to seconds), a short-lived congestion event may occur.

Moreover, the characteristics of microbursts can be obtained from real-time congestion loss statistics, such as the frequency and duration of microburst occurrence. On the other hand, the measurement accuracy of packet loss caused by congestion is significant in evaluating congestion levels, which provide guidance for subsequent network expansion and optimization, as well as a reliable verification of user service level objective (SLO) for packet loss.

The real-time localization of congestion occurrence is also required. It can help operators carry out rapid troubleshooting, thus improving the efficiency of fault diagnosis and root cause analysis. In addition, to analyse the cause of congestion, it is necessary to identify which traffic flows are contained in discarded packets and further identify which traffic flows lead to the congestion so that rapid action can be taken against those culprits causing congestion.

From the perspective of adaptability of packet loss monitoring and measurement methods, today's networks need to provide different transport modes for different services. Accordingly, a packet loss monitoring method is also required to adapt to various transport modes. For instance, L2/3 virtual private network (VPN) is widely used by enterprises, and multi-protocol label switching (MPLS) technique has been deployed by network operators to deliver MPLS VPN services. With the evolution of the network and the emergence of new services, more transport protocols will emerge to adapt to the delivery of new services. An ideal packet loss monitoring scheme should be protocol-independent, that is, it should be applicable to all current and future transport protocols, including native IPv4/6, segment routing (SR) MPLS (SR-MPLS) [b-IETF RFC 8660], segment routing over IPv6 (SRv6) [b-IETF RFC 8986], virtual extensible local area network (VXLAN) [b-IETF RFC 7348] and general routing encapsulation (GRE) [b-IETF RFC 1701], etc., such that the data plane (e.g., forwarding chip) does not need to be upgraded and packet header encapsulation does not need to be modified to adapt to existing monitoring and measurement methods. Another praised feature for a loss monitoring scheme is to cause little or no interference to the network so that network load is less affected and hence the packet loss monitoring results can reflect the actual congestion state. In addition, for large-scale operators' networks, typically tens of thousands of user flows need to be measured simultaneously, and the scalability is also a vital factor in evaluating a good monitoring method, that is, the number of concurrent measurements should be less constrained by network resources (e.g., computing, storage or bandwidth).

In summary, the packet loss monitoring solution should include five aspects:

- Real-time: The monitoring system is required to collect and analyse congestion induced packet loss in real time to quickly pinpoint the congestion location and identify the cause of congestion.
- Accuracy: The monitoring system is required to provide accurate measurement results for packet loss caused by congestion so that the operators can accurately assess the status and tendency of network congestion and make appropriate decisions.
- Protocol-independent: The monitoring system is required to be independent of network transport protocols so that the data plane does not need to be upgraded and packet header encapsulation does not need to be modified.
- Little or even no interference to the network: The monitoring system is required to have less or even no interference with the network in order to reduce the impact on network traffic forwarding behaviour.
- Scalability: The monitoring system is required to have the capability to accommodate tens of thousands of measurement sessions simultaneously so that the number of concurrent measurements should be less limited by network resources.

In light of these considerations, the investigation into real-time congestion and packet loss monitoring techniques holds substantial significance. This Technical Report does not attempt to replace the

existing packet loss measurement techniques, but to make up for the limitations of the existing techniques in monitoring packet loss caused by congestion in real time. As a novel packet loss monitoring tool, it can help network operators to quickly localize the congested nodes and the affected traffic flows, thus improving the efficiency of fault diagnosis and root cause analysis.

7 Standardized activities in the related standards developing organizations (SDOs)

7.1 Standardization activities in ITU-T for monitoring packet loss

The following Recommendations developed by the International Telecommunication Union – Telecommunication Standardization Sector (ITU-T) are related to monitoring packet loss in IP networks.

[ITU-T Y.1540] defines the parameters that may be used to specify and assess the speed, accuracy, dependability, and availability of IP packet transfers of regional and international IP data communication services. The defined parameters are applied to end-to-end or point-to-point IP services and network portions. This Recommendation defines a lost IP packet outcome, that is, a lost packet outcome occurs when there is a single IP packet reference event at a permissible ingress measurement point 1 $(MP)_1$, and when some or all of the contents corresponding to that ingress packet do not result in an IP packet reference event at a permissible egress MP_n within the time T_{max}. It also defines the IP packet loss ratio (IPLR), which is the ratio of total lost IP packet outcomes to total transmitted IP packets in a population of interest. It does not involve the definition of packet loss caused by network congestion, nor does it involve how to monitor packet loss caused by congestion.

[b-ITU-T Y.1541] defines classes of network quality of service (QoS) with objectives for Internet protocol network performance parameters. Two of the classes contain provisional performance objectives. These classes are intended to be the basis for agreements among network providers, and between end users and their network providers. Appendix XI estimates the packet loss requirement for digital circuit emulation. This Recommendation does not involve monitoring packet loss caused by network congestion.

[b-ITU-T Y.1543] specifies a set of Internet protocol (IP) performance parameters and methods of measurement applicable when assessing the quality of packet transfer on inter-domain paths. The methods anticipate that there will be multiple measurement systems, each conducting measurements of a segment of the customer-to-customer path, and recommend configurations that should produce useful results in this cooperative scenario. The methods rely on existing parameter definitions and encompass both active and passive measurement techniques. This Recommendation gives the formula for calculating a multi-segment packet loss rate (PLR), called the aggregate loss ratio (ALR), which is expressed as follows:

$$\text{ALR} = 1 - (1 - \text{PLR for segment 1}) \times (1 - \text{PLR for segment 2}) \times (1 - \text{PLR for segment 3}) \quad (1)$$

However, it also does not refer to the real-time monitoring of packet loss caused by congestion.

[b-ITU-T Y.1731] provides mechanisms for user-plane operation, administration and maintenance OAM functionality in Ethernet networks. The Ethernet frame loss measurement (ETH-LM) function is used to collect counter values applicable for ingress and egress service frames where the counters maintain a count of transmitted and received data frames between a pair of measurement end points (MEPs). The ETH-LM function is performed by sending frames with ETH-LM information to a peer MEP and similarly receiving frames with ETH-LM information from the peer MEP. Though this Recommendation provides a method of measuring two-way Ethernet frame loss, it does not specify the requirements for monitoring packet loss caused by congestion.

7.2 Standardization activities in IETF for monitoring packet loss

The IP performance measurement working group (WG) develops and maintains standard metrics that can be applied to the quality, performance, and reliability of Internet data delivery services and

applications running over IP networks. It also develops and maintains methodologies and protocols for the measurement of these metrics. These metrics, protocols, and methodologies are designed such that they can be used by network operators, end users, or independent testing groups.

Up to now, 72 RFCs and 32 active Internet drafts including more than 10 working group documents have been developed, among which the two-way active measurement protocol (TWAMP) [b-IETF RFC 5357], the simple two-way active measurement protocol (STAMP) [b-IETF RFC 8762], the alternate-marking method [b-IETF RFC 9341], and the in-situ operations, administration, and maintenance (IOAM) [b-IETF RFC 9197] have been widely supported by network device vendors and deployed in large-scale operator networks. According to [IETF RFC 7799], active methods generate packet streams. Commonly, the packet stream of interest is generated as the basis of measurement. Therefore, the TWAMP and STAMP methods can be classified as active methods.

Passive methods of measurement are based solely on observations of an undisturbed and unmodified packet stream of interest. In other words, the passive methods of measurement must not add, change, or remove packets or fields, nor change field values anywhere along the path.

Hybrid methods are methods of measurement that use a combination of active methods and passive methods, and also make augmentation or modification of the stream of interest. The alternate-marking and IOAM methods can be classified as hybrid methods.

The one-way active measurement protocol (OWAMP) [b-IETF RFC 4656]) measures unidirectional characteristics such as one-way delay and one-way loss. OWAMP actually consists of two protocols: OWAMP-control and OWAMP-test. OWAMP-control is used to initiate, start, and stop test sessions and to fetch their results, whereas OWAMP-test is used to exchange test packets between two measurement nodes. The TWAMP, based on OWAMP, adds two-way or round-trip measurement capabilities. Similarly to OWAMP, TWAMP also consists of two protocols: TWAMP-control and TWAMP-Test.

The simple two-way active measurement protocol (STAMP) enables measurement of both one-way and round-trip performance metrics, such as delay, delay variation, and packet loss. STAMP adopts a lightweight architecture of TWAMP, and its configuration is simpler; the deployment process is further simplified, and usage is more convenient.

The alternate-marking method performs packet loss, delay, and jitter measurements on live traffic. It allows the synchronization of the measurements at different points by dividing the packet flow into batches. Thus, it is possible to get coherent counters and timestamps in every marking period and therefore measure the performance metrics for the monitored flow.

In situ IOAM collects operational and telemetry information in the packet while the packet traverses a path between two points in the network domain. The term "in situ" refers to the fact that the OAM data are added to the data packets rather than being sent within packets specifically dedicated to OAM. The IOAM method can be used in the following scenarios:

- Proving that a certain traffic flow takes a predefined path
- Verifying the service level agreement (SLA) for the live data traffic
- Providing detailed statistics on traffic distribution paths in networks that distribute traffic across multiple paths, or
- Providing scenarios in which probe traffic is potentially handled differently from regular data traffic by the network devices.

7.3 The limitations of existing monitoring and measurement techniques for monitoring packet loss caused by congestion

Packet loss measurement is an important means of evaluating network performance and user QoS. The active measurement methods mentioned above can obtain one-way or two-way packet loss by means of the external probes or the device's built-in measurement module. Active methods have some

limitations. Firstly, they can only indirectly measure the monitored traffic by sending probe packets to simulate real traffic, thus resulting in deviations from the actual loss results. Secondly, they may interfere with network traffic forwarding behaviour when adding excessive probe traffic to the network for measurement. Thirdly, active methods can only detect packet loss between the source node and the destination node. When a loss event occurs, they have no ability to localize packet loss to a specific network device or interface. Lastly, the active methods commonly set the measurement period on the order of seconds or minutes, which cannot meet the need of real-time detection.

Compared to active methods, passive methods only depend on the presence of the measured traffic flows at one or more observation points. They do not generate additional traffic that disturbs the network. Some existing research on detecting packet loss was conducted by using TCP packets, which is time-consuming and leads to the inability to detect packet loss in real time. Other research has been conducted using a packet loss detection algorithm based on probability theory, where some specific traffic models are assumed and this leads to inaccurate packet loss results for real traffic. More current studies have focused on sampled traffic data provided by NetFlow [b-IETF RFC 3954], where the accuracy of loss detection results depends on traffic sampling schemes and the real-time performance is poor.

In recent years, the emerging on-path detection techniques, including the alternate-marking method [b-IETF RFC 9341] and IOAM [b-IETF RFC 9197], have aroused great interest from both industry and academia. These types of measurement and monitoring techniques are classified as hybrid methods.

The greatest advantage of the alternate-marking method lies in its high efficiency and credibility, as it can directly measure the real traffic with a single-marking bit. But there are some deficiencies. Firstly, two counters (one for the marking bit toggled to "1" and another for that toggled to "0") need to be employed for each measured flow at every measurement point, and many more counters will be needed when the number of concurrent measured flows and measurement points are large. For example, in hop-by-hop mode, considering m monitored flows traversing n nodes, then n*m*2 packet counters are needed. Secondly, it is required to define different packet header encapsulations to adapt to different transport protocols for the data plane, and the forwarding chip of the data plane also needs to be upgraded accordingly.

The IOAM method can directly monitor OAM information by embedding the monitoring instructions and carrying the OAM data of the forwarding path into the monitored packet. For instance, IOAM tracing data, including node ID, transit delay, queue depth, and so on, are expected to be collected at every IOAM transit node that a packet traverses to capture the visibility of the entire path. However, this method generates some additional telemetry data which may consume significant bandwidth, even causing path maximum transmission unit (MTU) issues, since every node along the path needs to add tens of bytes of OAM data to the monitored packet, and the whole forwarding path might accumulate telemetry data equal to or even greater than the original packet. Similarly, like the alternate-marking method, it is also protocol-dependent, and it needs to define different packet header encapsulations for different transport protocols, thus increasing the complexity of forwarding chips. More importantly, IOAM lacks the ability to detect packet loss. Packet loss may occur due to various factors, such as link quality degradation, improper configuration, access control list (ACL), mismatched forwarding information base (FIB), and loss caused by congestion is only one of many loss events. The above-mentioned monitoring and measurement methods have not been specially designed to monitor packet loss caused by congestion. These methods will face challenges when they are employed for monitoring network congestion-induced packet loss.

8 Considerations of monitoring packet loss caused by network congestion

8.1 Considerations of network device for monitoring packet loss caused by congestion

Consideration for network device monitoring packet loss caused by network congestion includes:

– Network device supports dynamic subscriptions, where a subscriber initiates a subscription negotiation with a publisher via a remote procedure call (RPC).

NOTE 1 – The lifetime of a dynamic subscription is bound by the transport session used to establish it, and the loss of the transport session will result in the immediate termination of any associated dynamic subscriptions.

– Network device supports configured subscriptions, which allow the management of subscriptions via a configuration.

NOTE 2 – Configured subscriptions can be configured to persist across reboots and even when their publisher is fully disconnected from any network.

– Network device supports the capability to subscribe to periodic updates. The subscription period shall be configurable as part of the subscription request.

– For periodic subscription, network device is recommended to support the ability for redundant suppression, where a telemetry update should not be generated unless the value of the subscribed data objects has changed.

– Network device supports the capability to subscribe to updates on change, i.e., whenever values of the subscribed data objects change.

– For on-change subscription, network device supports a dampening period that needs to pass before subsequent on-change updates are sent. The dampening period should be configurable as part of the subscription request.

– Network device supports detecting packet loss caused by congestion in real time (i.e., millisecond interval) by dedicated hardware.

– Network device reports packet loss events in a real-time manner, and the reported telemetry data is required to carry the time of packet loss occurrence, the number of discarded packets, and the localization of packet loss such as device ID, port ID, and queue ID.

– Network device reports packet loss events periodically.

– Network device reports packet loss event on change.

– Network device caches all discarded packets caused by queue overflow.

– Network device uploads all discarded packets as a file or compressed file in real-time manner.

– Network device supports time synchronization for measuring packet loss ratio caused by congestion, and time synchronization precision is less than 50 ms.

8.2 Considerations of the collection and analysis system for monitoring packet loss caused by congestion

The requirements of the collection and analysis system for monitoring packet loss caused by network congestion include:

– It supports configured subscriptions as a server, accepting subscription data.

– It supports dynamic subscriptions as a client, initiating subscription requests and accepting subscription data.

– It collects packet loss events for statistics and analysis.

– It supports data repository for storing all discarded packets uploaded.

– It supports parsing the header of all discarded packets so as to determine the flow attribute of every discarded packet.

– It analyses the service types of discarded packets, and counts the number of discarded packets of each traffic flow in a real-time manner.

– It supports periodic measurement of packet loss ratio (PLR) based on the total number of discarded packets divided by the total number of sent packets.

– It supports measurement of PLR according to the number of the discarded packets divided by the number of the sent packets for the specified user traffic.
– It localizes the location of packet loss in real time based on the reported loss events.
– It supports data analysis methods (e.g., statistical analysis method, visualization analytical method, correlation analytical method) to process the discarded packets.
– It supports visualization of data analysis for the discarded packets in the form of tables and figures, which are easily understandable by users.

9 Architecture of monitoring packet loss caused by network congestion

In order to meet the requirements for monitoring packet loss caused by network congestion described in clause 8, the architecture for monitoring packet loss caused by congestion is presented in Figure 9-1.

The architecture is mainly composed of network devices and a collection and analysis system. A telemetry interface (e.g., network configuration protocol (NETCONF) [IETF RFC 6241], Google remote procedure call (gRPC) [gRPC]) with a subscription mechanism serves as the interface between the collection and analysis system and network devices to collect real-time packet loss data. File transfer protocol (FTP) [IETF RFC 0959], [IETF RFC 1350] is used for the real-time upload of packet loss file. RESTful is appropriate for the interface between the collection and analysis system and network management system (NMS) / element management system (EMS), as well as the network controller.

All network devices of an IP network need to report the packet loss events caused by congestion to the collection and analysis system in a real-time manner, and also cache the discarded packets overflowed by the port queue and upload them to the collection and analysis system. The collection and analysis system counts the total number of discarded packets reported, analyses the service types of discarded packets, and counts the number of discarded packets for every traffic flow, and so on. The telemetry interface is used to push loss data immediately when a loss event occurs, avoiding the inefficiency of the traditional simple network management protocol (SNMP) polling mode. At the same time, the real-time visibility of packet loss obtained from the collection and analysis system can feed into NMS/EMS so that network operators can quickly pinpoint the congested nodes and the affected traffic flows. Also, with the injection of such real-time visibility of packet loss, the network controller can make timely path optimization for the affected traffic flows with the sensitivity of latency and loss to improve user quality of experience (QoE).This monitoring mode, with "little or even no interference to network", can not only accurately locate packet loss caused by congestion, determine the time of packet loss occurrence, and accurately count the number of the discarded packets and identify the types of traffic flows to which they belong, but also be used as a candidate method for packet loss measurement to achieve more accurate measurement results. Therefore, it is of significance for real-time network congestion monitoring.

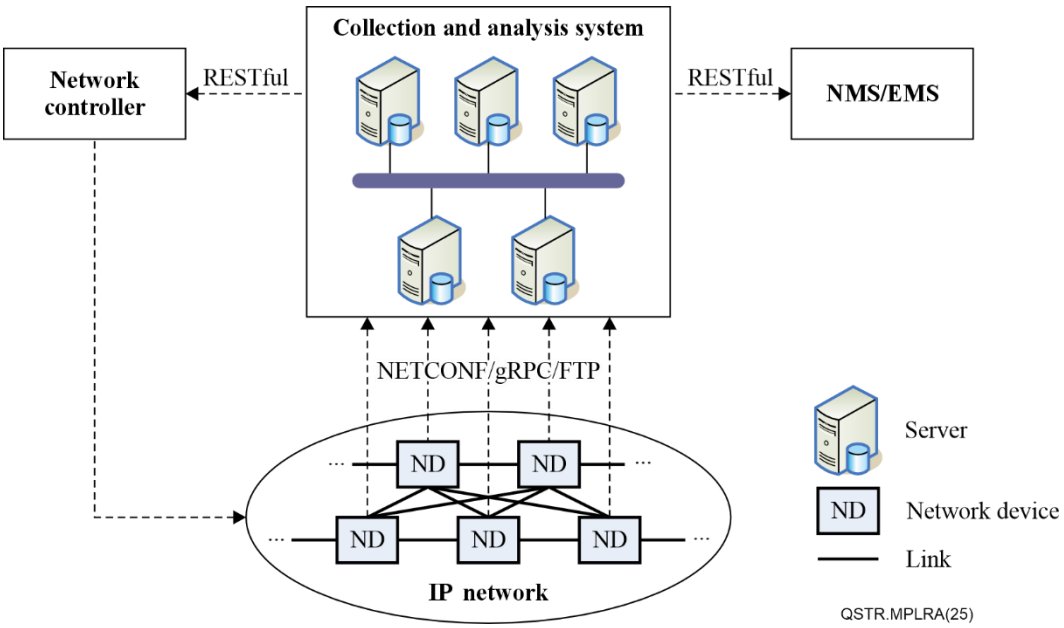

Figure 9-1 – Architecture of monitoring packet loss caused by congestion

9.1 Network devices

In IP networks, network devices such as routers and switches are mainly used to perform packet forwarding. Traditional network devices can only record the number of discarded packets of port or queue but cannot record the time of packet loss occurrence. Network operators can only log on the device (e.g., through CLI) to query packet loss, but cannot detect congestion and packet loss in real time. Network devices are required to have the ability for real-time awareness of congestion and packet loss. The traditional query method of using the central processing unit (CPU) on the main control board is resource-intensive, and the network device must possess built-in dedicated hardware to detect packet loss in real time. On the other hand, existing forwarding devices do not cache the packets overflowed by queue, but simply drop them, making it unclear which packets are discarded, and which traffic flows contribute to congestion or microburst. In order to capture what types of services are related to the packet loss, the cache for the discarded packets is required. The forwarding structure of network device is shown in Figure 9-2.

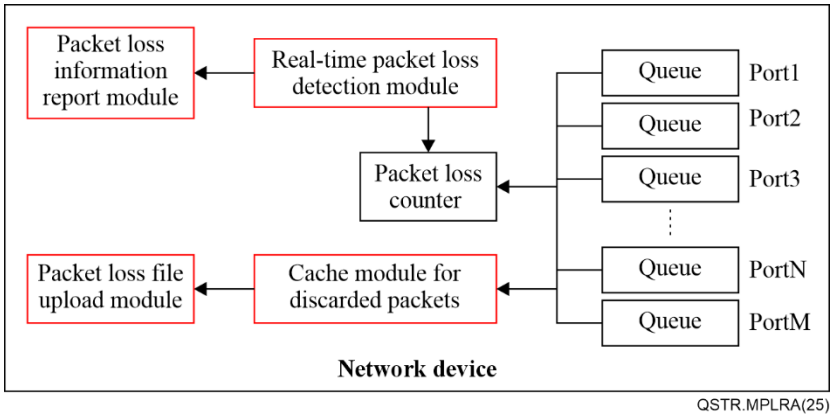

Figure 9-2 – The forwarding structure of network device

The forwarding structure of network device is required to add four new functional modules: real-time packet loss detection module, packet loss information report module, cache module for discarded packet and packet loss file upload module, which are described as follows.

- Real-time packet loss detection module: Leverages the built-in dedicated hardware to read the packet loss counter of every queue at millisecond intervals; at the same time, records the location and time of packet loss occurrence.
- Packet loss information report module: Sends packet loss information according to subscription request, including the number of discarded packets, the time of packet loss occurrence, the localization of packet loss such as device identifier (ID), port ID and queue ID. It is required to support periodic updates. Also, in order to improve the real-time awareness of packet loss, it is also required to support updates on change, where a dampening period should be configurable to reduce data volume to the maximum extent.
- Cache module for discarded packet: Caches packets dropped by overflow of each queue, records the number of discarded packets, the time of packet loss occurrence, the localization of packet loss such as device ID, port ID, and queue ID. In order to save storage space, the discarded packets should be cleaned as soon as they are uploaded.
- Packet loss file upload module: Packages the cached discarded packets as a file or compressed file and uploads it to the collection and analysis system according to the specified rule.

9.1.1 Cache for discarded packets

To analyse packet drops caused by queue overflow, implementing a cache mechanism is essential for capturing discarded packets. However, since packet parsing and statistical analysis consume significant local resources (such as memory and computing power), these tasks are better suited for a remote central processing entity. Given the constraints of circuit board size and power consumption, the cache capacity must be minimized. Fortunately, since packet headers typically contain all necessary service type and flow attribute information, truncating discarded packets to a fixed length (e.g., the first 64 bytes) provides sufficient data for analysis while dramatically reducing cache requirements.

During the uploading the packet loss file and cleaning the discarded packets, any loss event may occur, potentially leading to no buffer available for the subsequent dropped packets. In order to avoid this situation, the cache should be divided into two distinct spaces in appropriate proportion: primary space and spare space. The primary space is used to cache the discarded packets for uploading each time, and the spare space is used to cache subsequent discarded packets during the current packet packaging and uploading operation. Figure 9-3 illustrates the process of the discarded packet being truncated and loaded into the cache.

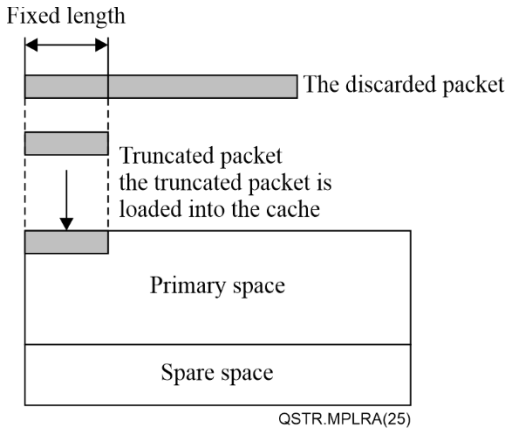

Figure 9-3 – The process of the discarded packet being truncated and loaded into the cache

9.1.2 Packet loss file upload

In order to support the real-time uploading of packet loss files, file transfer protocol should be used for transferring the file immediately when the loss file is available. File transfer protocol (FTP)

[IETF RFC 0959] is widely used to upload and download files. But it needs to use two transmission control protocol (TCP) connections: one for control command, and another for data transfer. Although TCP connection for control command can be persistent, the one for data transfer is short-lived, and it shuts down the connection after finishing the current file transfer. So when the next file needs to upload, a new TCP connection must be re-established, delaying the file transfer.

Hypertext transfer protocol (HTTP) is of long-connection characteristic and can satisfy the requirement of the real-time file transfer. However, HTTP is mainly used in web environment, and needs more memory and processing overhead, which is not suitable for network device.

Compared to FTP, trivial file transfer protocol (TFTP) [IETF RFC 1350], which uses user datagram protocol (UDP), results in better timeliness of file transfer. TFTP requires less memory and processing overhead, providing simple and efficient file transfer services. Therefore, TFTP is more preferable to serve as File Transfer Protocol for the real-time upload of packet loss file.

9.1.3 Telemetry data collection

The local device is also required to collect real-time loss data caused by congestion. In order to capture loss events in real time, the network device needs to leverage the built-in dedicated hardware to read the packet loss counter of each port or queue at millisecond interval, and send telemetry data about loss information according to subscription request. In order to improve the real-time awareness of packet loss in some scenarios such as traffic optimization and congestion discovery, the on-change update (compared to periodic update) is more preferable, that is, a telemetry update is sent immediately when packet loss counter value changes. While supporting on-change update, a dampening period should be configurable to minimize the amount of data sent.

On the other hand, in order to measure PLR caused by congestion, the network device is required to collect the statistical data of the monitored traffic flows and send the corresponding telemetry data to the collection and analysis system periodically. The ingress device, such as access router and Provider Edge router (PE), is required to configure the receiving packet counter for the monitored traffic. The specified traffic flows may be identified by (but not limited to) the following features:

- Physical interface or sub-interface
- Layer 2 flows, e.g., based on source or/and destination media access control (MAC) address, Virtual Local Area Network Identifier (VLAN ID), Virtual extensible local area network identifier (VXLAN VNI)
- Layer 3 flows, e.g., identified by N-tuple, flow label field of IPv6 packet header
- Layer 2/3 VPN ID carried in SR-MPLS label stack or IPv6 segment routing header (SRH).

9.1.4 Time synchronization

The global time synchronization is also required for the accurate calculation of PLR measurement. For instance, when the ingress device periodically reports the received VPN traffic statistical data (packet counter value) with the timestamp in telemetry data, and during some report period, this specified VPN traffic has happened to encounter packet loss caused by a microburst, and the loss information is immediately reported with the timestamp of loss occurrence. Based on their respective timestamps (e.g., the timestamp of loss occurrence falls between the timestamps carried by the two consecutive traffic telemetry data), the collection and analysis system can correctly calculate the PLR of the specified VPN traffic at that exact period. Figure 9-4 depicts the timing relationship between the time of telemetry data of the specified traffic reported and that of loss occurrence reported. The network device is required to support time synchronization techniques such as network time protocol (NTP) or precision time protocol (PTP) [b-IEEE 1588], which are widely deployed in operator's networks. Generally, NTP can meet a precision of 50 ms and PTP can meet a precision of microseconds. Time synchronization precision depends on measurement period. For normal measurement period of tens of seconds or even minutes, synchronization precision of 50 ms (easy to implement) is enough to satisfy the measurement requirement.

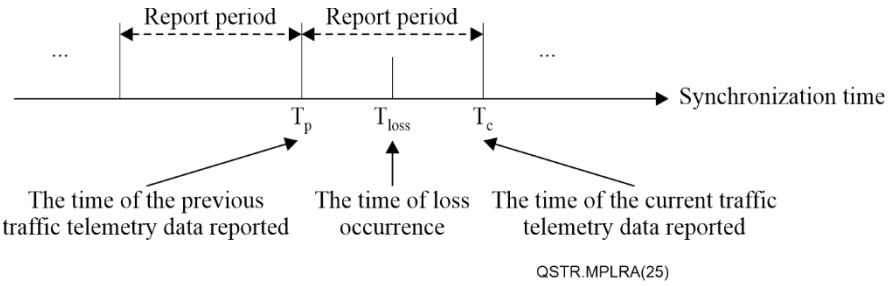

Figure 9-4 – The timing relationship between the time of telemetry data of the specified traffic reported and that of loss occurrence

9.2 Collection and analysis system

The framework is required to handle packet loss information and demands higher real-time requirements. Therefore, an independent collection and analysis system is more suitable for monitoring the real-time packet loss caused by congestion. The structure of the collection and analysis system is shown in Figure 9-5.

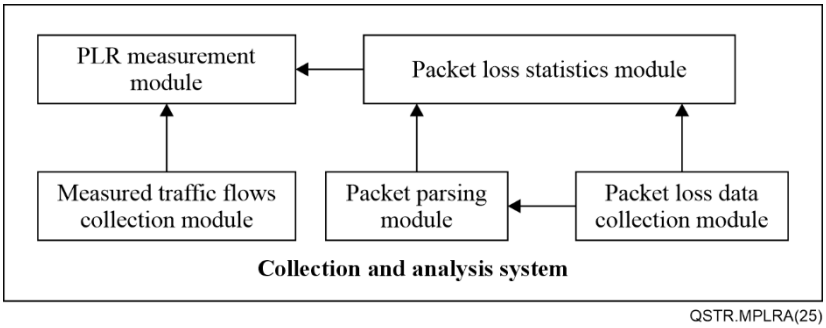

Figure 9-5 – The structure of collection and analysis system

The structure of collection and analysis system mainly consists of five functional modules: packet loss data collection module, measured traffic flows collection module, packet parsing module, packet loss statistics module and PLR measurement module, which are described as follows:

- Packet loss data collection module: Accepts the packet loss data from network devices, including the telemetry data of packet loss information reported and packet loss files uploaded, and stores them for a specified time; records the number of discarded packets, the timestamp and location ID carried in the telemetry data every time it is reported.
- Measured traffic flows collection module: Accepts the telemetry data of the measured traffic flows from network devices, and stores them for a specified time; records the number of packets and the timestamp carried in the telemetry data every time it is reported.
- Packet parsing module: Employs the professional packet parsing tools (e.g., Wireshark) to perform the real-time resolution of the discarded packets from packet loss files uploaded.
- Packet loss statistics module: Based on packet parsing results, counts the number of the discarded packets of each traffic flow; based on packet loss information reported, counts the total number of the discarded packets of each device, each port and each queue, and also records the time and location of packet loss occurrence.
- PLR measurement module: Based on the statistics of the measured traffic flows reported periodically and the number of the discarded packets of the measured traffic flows from the packet loss statistics module, calculates PLR of the measured traffic flows according to the requirements of network operators (e.g., measurement duration).

9.2.1 Packet parsing

The discarded packets should be parsed as soon as possible to meet the real-time requirement of packet loss statistics and measurement. For the purpose of the real-time visibility of packet loss statistics as well as online PLR measurement, packet parsing time for the currently uploaded packet loss file should be as little as possible, say, within 100 ms. The packet flow parsing of the discarded packets should at least include the following features:

- Layer 2 flows, e.g., based on source or/and destination MAC address, VLAN ID, VXLAN VNI.
- Layer 3 flows, e.g., identified by N-tuple, Flow Label field of IPv6 packet header.
- Layer 2/3 VPN ID carried in SR-MPLS label stack or SRH.

9.2.2 Packet loss statistics

In order to quickly pinpoint the congested nodes and the affected traffic flows, improving the efficiency of fault diagnosis and root cause analysis, the real-time visibility of packet loss is necessary. Firstly, on receiving packet loss information reported, it is required to count the total number of the discarded packets of each device, each port and each queue, record the time of packet loss occurrence, and localize the location of packet loss occurrence in real time. Secondly, after capturing the discarded packet parsing results, it is required to count the number of packets related to different traffic flows in real time. In order to intuitively exhibit packet loss statistical analysis, the results should be visualized in the form of charts, tables, or graphs.

9.2.3 PLR measurement

PLR measurement module can obtain the number of packets and timestamp carried in the telemetry data of the measured traffic flow from the measured traffic flows collection module. Meanwhile, it also can obtain the number of the discarded packets of the measured traffic flow and the timestamp carried in the loss information or the packet loss file from packet loss statistics module. Therefore, based on the timing relationship between the timestamp carried in the telemetry data of the measured traffic flow and that of loss occurrence, as well as the number of received packets carried in the telemetry data of the measured traffic flow and the number of the discarded packets of the measured traffic flow, the PLR measurement module can calculate the PLR of the measured traffic flow during a specified measurement period. For example, the collection and analysis system receives the previous telemetry data of the measured traffic flow carrying the number N1 of received packets and the timestamp T1, and the current telemetry data carrying the number N2 of received packets and the timestamp T2. Meanwhile, it also obtains the number N3 of the discarded packets of the measured traffic flow and the timestamp T3 carried in the packet loss file. If the timestamp T3 is between timestamps T1 and T2, then the PLR of the measured traffic flow for the current measurement period (T2-T1) is precisely calculated as:

$$\mathrm{PLR} = \mathrm{N3}/（\mathrm{N2} - \mathrm{N1}） \quad (2)$$

PLR measurement is required to support (but not limited to) the following traffic flows:

- Layer 2 flows, e.g., based on source or/and destination MAC address, VLAN ID, VXLAN VNI.
- Layer 3 flows, e.g., identified by N-tuple, flow label field of IPv6 packet header).
- Layer 2/3 VPN flows, e.g., identified by VPN ID carried in SR-MPLS label stack or IPv6 SRH.

10 Interface and protocol

According to the framework for monitoring packet loss caused by congestion, the collection and analysis system is required to support the interfaces with network devices for collecting the telemetry

data of packet loss information reported and packet loss files uploaded, as well as for collecting the telemetry data of the measured traffic flows periodically. Additionally, the collection and analysis system is required to communicate with NMS/EMS to notify the real-time visibility of packet loss caused by network congestion. Also, it is required to communicate with the network controller for timely path optimization for those key traffic flows based on such real-time visibility of packet loss. The interface requirements are shown in Figure 9-1.

10.1 NETCONF

The NETCONF [IETF RFC 6241], [IETF RFC 8639], [IETF RFC 8640] protocol defines a simple mechanism through which a network device can be managed, configuration data information can be retrieved, and notification data information can be subscribed to.

The collection and analysis system and network device are required to support NETCONF for subscription configuration, collecting the telemetry data of packet loss information, as well as the telemetry data of the measured traffic flows. NETCONF supports the following requirements, including:

– Dynamic subscriptions.

– Configured subscriptions.

– The ability to subscribe to periodic updates. The subscription period shall be configurable as part of the subscription request.

– The ability to subscribe to updates on-change, i.e., whenever values of subscribed data objects change.

– The termination of a subscription when requested by the subscriber.

– The ability to suspend and resume a subscription on request of a client.

10.2 gRPC network management interface (gNMI)

The gNMI supports modification and retrieval of configuration, as well as control and transmission of telemetry streams from a network device to a data collection system. gNMI derives a number of benefits from being built on gRPC and HTTP/2, including modern security mechanisms, bidirectional streaming, binary framing, and a wide variety of language bindings to simplify integration with management applications. With protobuf encoding, it also provides significant efficiency advantages over XML serialization with a 3 to 10 times reduction in data volume. It is preferable to use gNMI to interface between the collection and analysis system and network devices for collecting telemetry data of the measured traffic flows and packet loss information.

The conceptual layers and requirements for gNMI serving as the collection and encapsulation of telemetry data are depicted in Table 1.

Table 1 – Conceptual layers and requirements for gNMI serving as the collection and encapsulation of telemetry data

Layer		Requirements
Data model	Service data	MUST include the service data from the specified paths
	Telemetry header	MUST include the timestamp, the time when the device collects the service data, and the paths from which the service data originates
gRPC		gRPC provides the following RPCs: • Capabilities, used by the client and target as an initial handshake to exchange capability information. • Get, used to retrieve snapshots of the data on the target by the client. • Set, used by the client to modify the state of the target. • Subscribe, used to control subscriptions to data on the target by the client.
HTTP/2		HTTP/2 supports header field compression and binary framing, allowing multiple concurrent exchanges on the same connection.
Secure transport based on TCP		gNMI connection provides authentication, data integrity, confidentiality, and replay protection (e.g., transport layer security (TLS)

The collection and analysis system and network device are required to support gNMI for subscription configuration, collecting the telemetry data of packet loss information, as well as the telemetry data of the measured traffic flows. The gNMI supports the following features, including:

– STREAM subscription, a long-lived subscription which continues to transmit updates. STREAM Subscription includes one of the following modes:
  • On change - data updates are only sent when the value of the data item changes.
  • Sampled - the value of the data item(s) MUST be sent once per sample interval to the client.
– Structured data sent by the client or the target in an update message, which MUST be serialized according to one of the supported encodings.
– Protobuf and JavaScript object notation (JSON) encoding.

Service data model for the structured data when supporting Protobuf encoding includes the following aspects:

(1) Packet loss information

The structured data for packet loss information when supporting Protobuf encoding is defined as follows:

string name = 1 //device ID

string name = 2 //port ID

string name = 3 //queue ID

uint64 the number of dropped packets = 4 //the number of dropped packets overflowed by queue

uint64 timestamp = 5//the current sampling time

(2) The measured traffic flow

The structured data for the measured traffic flow when supporting Protobuf encoding is defined as follows:

string name = 1 //device ID

string name = 2 //port ID

string name = 3 //queue ID

string name = 4 // traffic flow ID

uint64 the number of packets sent = 5 //the number of packets sent from the ingress device

uint64 timestamp = 6//the current sampling time

10.3 File transfer protocol

In order to support the real-time upload of packet loss file, a file transfer protocol should transfer the file immediately as soon as the packet loss file is available. FTP [IETF RFC 0959] is widely used to upload and download files. However, FTP needs to use two transmission control protocol (TCP) connections: one for control command, and another for data transfer. Although TCP connection for control command can be persistent, the TCP connection for data transfer is short-lived, and it shuts down the connection after finishing the current file transfer. When the next file needs to be uploaded, a new TCP connection for data transfer must be re-established. Another file transfer protocol similar to FTP is TFTP [IETF RFC 1350], which uses UDP, resulting in better timeliness of file transfer. TFTP requires less memory and processing overhead, providing simple and efficient file transfer services. Therefore, TFTP is more preferable to serve as file transfer protocol for the real-time upload of packet loss file. A packet loss file may consist of several blocks of the discarded packets, and each packet loss block is required to include the following fields:

- The field for the number of the discarded packets.
- The field for the location of packet loss such as device ID, port ID, queue ID.
- The field for the timestamp of packet loss occurrence.

10.4 RESTful

Representational state transfer (REST), based on web service architecture, is a widely used lightweight interface with the characteristics of platform-independence and language-independence. It is a stateless protocol in that the client request must contain all state information, and the client uses GET, POST, PUT, and DELETE methods to operate the resources of server. So, RESTful is appropriate for the interface between the collection and analysis system and NMS/EMS as well as the network controller.

11 Security considerations

According to the framework, all network devices need to report the packet loss events caused by congestion to the collection and analysis system in real-time manner, and also cache the dropped packets overflowed by the port queue and upload them to the collection and analysis system.

There might be some circumstances where malicious users launch a persistent distributed denial of service (DDoS) attack by flooding the specified interfaces with a great amount of traffic, as a result, the affected device has to frequently upload the packet loss files that package a large number of dropped packets, which may exhaust the resources of control plane (i.e., CPU). An appropriate threshold is required to be set for CPU utilization, and when the CPU is overloaded, the device should refuse to upload these packet loss files.

On the other hand, the network devices are required to send telemetry data about packet loss information according to subscription requests. When on-change subscription is configured to report the packet loss information, if the counter values for the dropped packets are always changing, the device has to report packet loss information continuously, which may exhaust the resources of control plane. Thus, the network devices are required to support a dampening period that needs to be passed before the subsequent on-change updates are sent, so as to avoid this situation.

Other security considerations also include interface security, such as NETCONF and gNMI, providing authentication, data integrity, confidentiality, and replay protection.

Bibliography

[b-ITU-T Y.1541] Recommendation ITU-T Y.1541 (2011), *Network performance objectives for IP-based services.*

[b-ITU-T Y.1543] Recommendation ITU-T Y.1543 (2018), *Measurements in Internet protocol networks for inter-domain performance assessment.*

[b-ITU-T Y.1731] Recommendation ITU-T G.8013/Y.1731 (2023), *Operation, administration and maintenance (OAM) functions and mechanisms for Ethernet-based networks.*

[b-IEEE 1588] IEEE 1588 (PTPv2) (2008), *IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems*.

[b-IETF RFC 1701] IETF RFC 1701 (1994), *Generic Routing Encapsulation (GRE)*.

[b-IETF RFC 3954] IETF RFC 3954 (2004), *Cisco Systems NetFlow Services Export Version 9*.

[b-IETF RFC 5357] IETF RFC 5357 (2008), *A Two-Way Active Measurement Protocol (TWAMP)*.

[b-IETF RFC 7348] IETF RFC 7348 (2014), *Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks*.

[b-IETF RFC 8660] IETF RFC 8660 (2019), *Segment Routing with the MPLS Data Plane*.

[b-IETF RFC 8762] IETF RFC 8762 (2020), *Simple Two-Way Active Measurement Protocol*.

[b-IETF RFC 8986] IETF RFC 8986 (2021), *Segment Routing over IPv6 (SRv6) Network Programming*.

[b-IETF RFC 9197] IETF RFC 9197 (2022), *Data Fields for In Situ Operations, Administration, and Maintenance (IOAM)*.

[b-IETF RFC 9341] IETF RFC 9341 (2022), *Alternate-Marking Method*.