

IPPM Working Group
Internet Draft
Intended status: Standards Track
Expires: July 5, 2024

H. Wang
S. Weng
China Mobile
C. Lin
New H3C Technologies
X. Min
ZTE Corporation
G. Mirsky

Ericsson

January 6, 2024

Distributed Flow Measurement in IPv6
draft-wang-ippm-ipv6-distributed-flow-measurement-04

Abstract

In IPv6 networks, performance measurements such as packet loss, delay and jitter of real traffic can be carried out based on the Alternate-Marking method. Usually, the controller needs to collect statistical data on network devices, calculate and present the measurement results. This document proposes a distributed method for on-path flow measurement, which is independent of the controller.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/1id-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

This Internet-Draft will expire on July 5 2024.

Copyright Notice

Copyright (c) 2024 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the [Trust Legal Provisions](#) and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1.	Introduction	3
1.1.	Requirements Language	3
2.	Requirement scenarios	3
3.	End-to-end measurement	5
4.	Hop-by-hop measurement	5
5.	Extension to the Flow Monitor Option	6
5.1.	Previous period count bit (bit1)	6
5.2.	Packet timestamp bit (bit2)	6
6.	Measurement information and result notification	7
6.1.	Data fields	8
6.1.1.	Base data structure	8
6.1.2.	Packet count TLV	8
6.1.3.	Time Stamp TLV	9
6.1.4.	Packet loss TLV	10
6.1.5.	Packet delay TLV	11
6.1.6.	Average Packet loss TLV	11
6.1.7.	Average Packet delay TLV	12
6.2.	Transport channel	13
6.2.1.	Independent controller protocol	13
6.2.2.	Extending BGP Protocol	14
6.2.3.	Reverse traffic	14
7.	Application of measurement results	14
8.	IANA Considerations	15
9.	Security Considerations	15
10.	References	16
10.1.	Normative References	16
	Authors' Addresses	17

1. Introduction

[[draft-wang-ippm-ipv6-flow-measurement](#)] describes how to measure the network by carrying the detection data in the traffic in the IPv6 network based on Alternate-Marking.

The nodes participating in the measurement need to collect information such as message statistics and processing time stamps, and transport the collected information to the controller through telemetry technology or other methods. The controller calculates the packet loss and delay of each flow according to the telemetry data.

Based on the basic method of [[draft-wang-ippm-ipv6-flow-measurement](#)], this document proposes a flow measurement without the participation of the controller. The nodes involved in the measurement calculate the network metrics such as packet loss and delay Distributed.

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. Requirement scenarios

The method described in [[draft-wang-ippm-ipv6-flow-measurement](#)] requires the controller to summarize the data collected by each node and then calculate the final measurement result. In some specific scenarios, this method could not meet the requirements of measurement well.

o Scenario1:

For the customers who have high requirements for SLA such as banks and finance, this method cannot meet the customers well. Firstly, each participating measuring node reports to the controller, and then the centralized controller calculates the path quality, and then the controller notifies the source node to schedule the path of traffic. The whole processing path is too long and it is difficult to guarantee the SLA requirements of customers in this way.

o Scenario2:

For the transport network with multiple AS domains and multi-level controllers, one inter-AS controller is deployed and one intra-AS controller is deployed for each AS typically.

Inter-AS controller programs end-to-end paths, but do not manage network devices. Each intra-AS controller only manages devices in its own AS and is not aware of the entire end-to-end path.

Therefore, the measurement data will be reported to the intra-AS controller by the measurement node, but the final data needs to be summarized, calculated and presented on the centralized inter-AS controller. This will cause the interaction between different levels of controllers to be too complex.

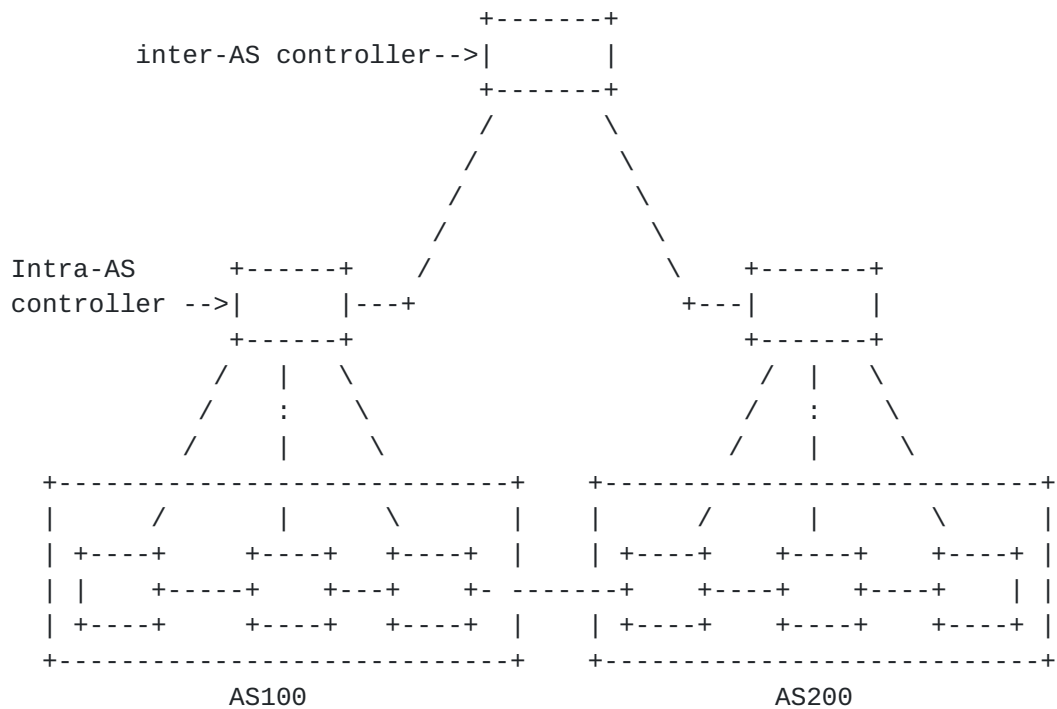


Figure 1: reference topology of multiple level controller

o Scenario3:

For some networks may not have the conditions or requirements to deploy controllers, but they also hope to use the technology of flow measurement to measure and present the quality of traffic forwarding path.

In order to meet the requirement of these scenarios, this document proposes a distributed flow measurement, which does not depend on the controller. All the nodes participating in the measurement complete the measurement, and finally the measurement results can be

used on the source node for fast intelligent routing, simplifying operation and maintenance, and optimize the experience.

3. End-to-end measurement

For end-to-end measurement, there are two working models, which are suitable for different scenarios.

o Source node model:

The source node completes the summary and calculation of statistical data.

The source node inserts the required flow measurement indicators into the specified traffic, and marks the traffic according to [[draft-wang-ippm-ipv6-flow-measurement](#)]. The end node collects the statistical data and time stamp, the collected information is periodically notified to the source node, which completes the calculation of the measurement results.

In this model, the source node undertakes the work of the controller and can count the data measured by the traffic through source node.

o End node model:

The end node is responsible for calculating measurement result. In addition to marking the traffic, the source node also needs to carry additional information through the flow monitor option. For example, in order to measure packet loss, the traffic count of the source node in the previous period need to be carried in the flow monitor option, packet delay measurement requires the source node to carry a timestamp when marking the D bit.

Through this information, the end node could calculate the packet loss and delay stream on the flow. Furthermore, the average packet loss and delay could be calculated. All the result could be send to the corresponding source node.

This model is suitable for the scenario of one source node vs multiple end nodes, such as multicast. It can reduce the calculation pressure of the source node and transfer the workload to each end node.

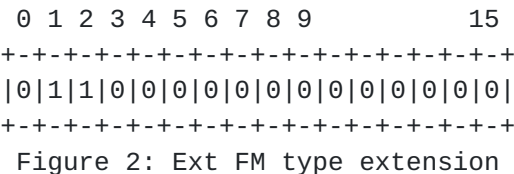
4. Hop-by-hop measurement

Hop-by-hop measurement requires that intermediate nodes also participate in data collection, so only the source node model should be used.

5. Extension to the Flow Monitor Option

Refer to [[draft-wang-ippm-ipv6-flow-measurement](#)], the additional information required by the end node model can be carried by extending Ext FM type.

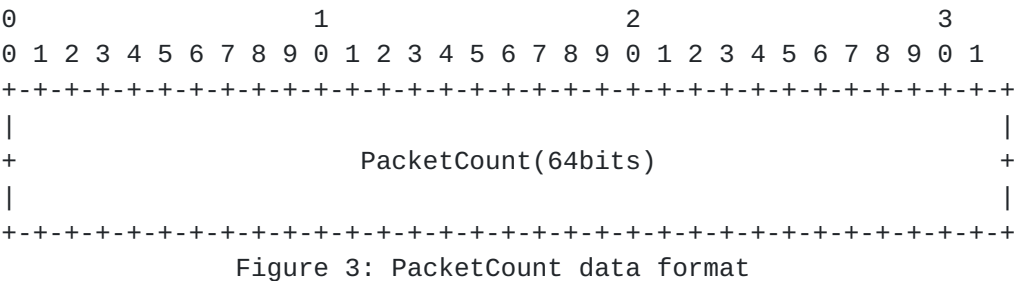
Define the corresponding bit and data format for the packet count of previous period and time stamp.



5.1. Previous period count bit (bit1)

This bit indicates the flow monitor option carries the packet count of the source node in the previous period. The end node can calculate the packet loss data according to this value in combination with the locally recorded count.

The data format is shown below:



o PacketCount 64bits Packet count of the previous period of the source node.

5.2. Packet timestamp bit (bit2)

This bit indicates the flow monitor option carries the timestamp set by the source node, which is the time when the source node receives the packet. The end node could calculate the packet delay according to this value in combination with the packet receiving time of end node.

The data format is shown below:

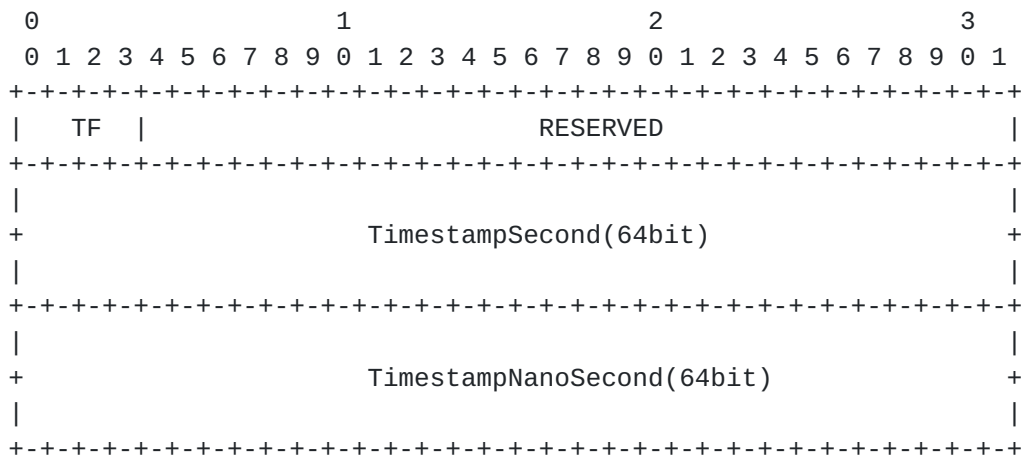


Figure 4: timestamp data format

o TF: 4bits, Timestamp format. The values are as follows:

1: PTP (see [[RFC8877](#)])

2: NTP (see [[RFC5905](#)])

3: POSIX

o TimestampSecond: 64bits, Integer value of the second part from 1970 to the time when the message is received in the timestamp format specified by NF field.

o TimestampNanoSecond: 64bits, Integer value of the nanosecond part from 1970 to the time of receiving the message in the timestamp format specified by NF field.

6. Measurement information and result notification

For the source node model, the measurement data of the intermediate node and the end node need to be sent to the corresponding source node.

For the end node model, the end node needs to send the calculated measurement results to the corresponding source node.

The address of the original node is obtained through the outer encapsulation source address of the packet carrying the monitor data. The notification period of collection data or results can be according to the measurement period or the configured period.

6.1. Data fields

The notification data structure includes a base data structure and multiple data structures defined through TLV.

6.1.1. Base data structure

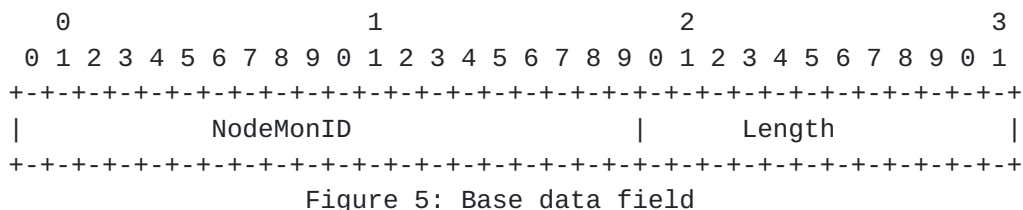


Figure 5: Base data field

The fields are defined as following:

- o NodeMonID: A 20 bits field, which is consistent with the definition in flow monitor option.
- o Length: A 12 bits field, Length of the notification data in 4-octet units, not including the first 4 octets.

6.1.2. Packet count TLV

This TLV is used to notify packet count to source node and is used in the source node model. The TLV is defined as follows:

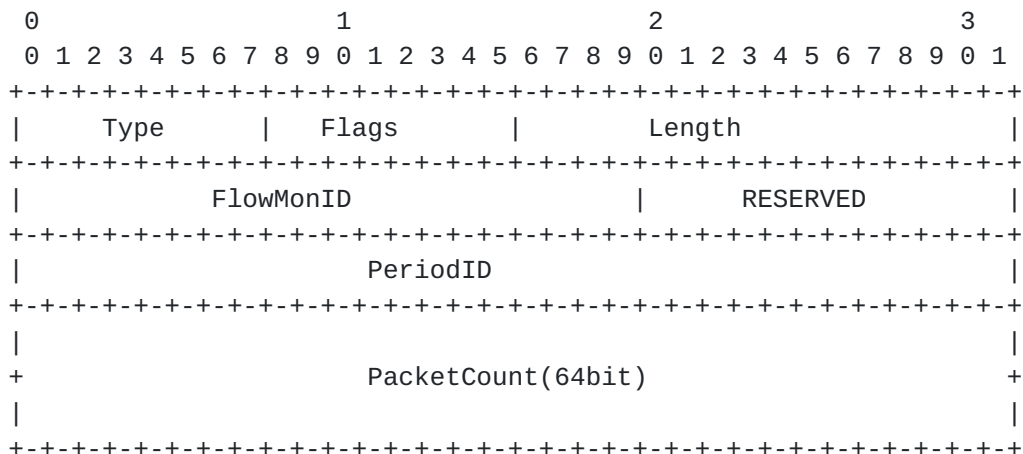


Figure 6: packet count TLV

- o Type: A one-octet field. Value 1 will be register in IANA.
- o Flags: A one-octet field.
- o Length: A two-octet field equal to the length of the Value field in octets.

- o FlowMonID: A 20 bits field, which is consistent with the definition in flow monitor option.
- o PeriodID: A 4 Octets period ID of the packet count.
- o PacketCount: A 8 Octets packet count in the period received by node.

6.1.3. Time Stamp TLV

This TLV is used to notify time stamp to source node and is used in the source node model. The tlv is defined as follows:

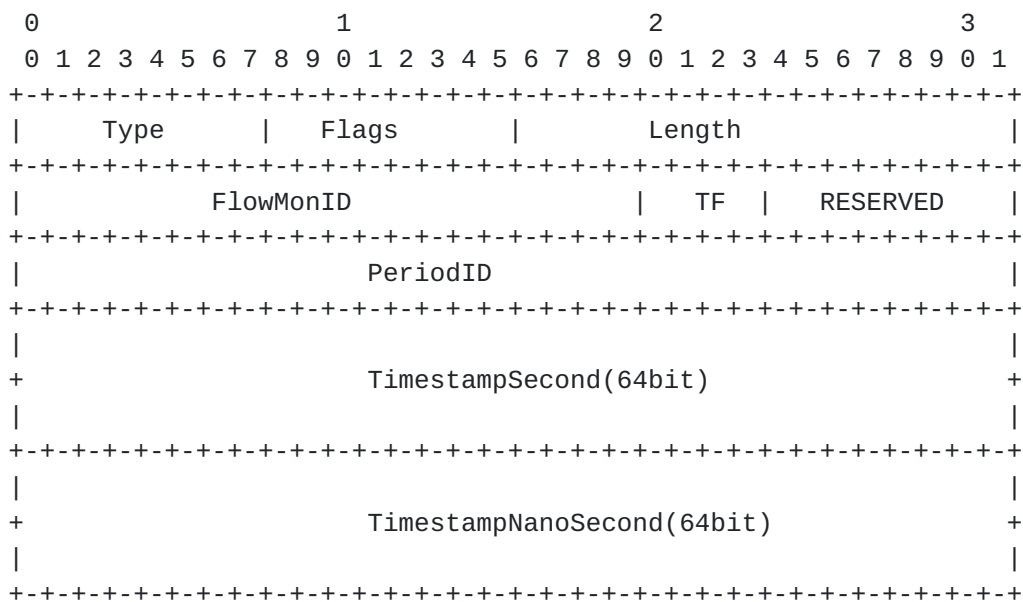


Figure 7: time stamp TLV

- o Type: A one-octet field. Value 2 will be register in IANA.
- o Flags: A one-octet field.
- o Length: A two-octet field equal to the length of the Value field in octets.
- o FlowMonID: A 20 bits field, which is consistent with the definition in flow monitor option.
- o TF: A 4 bits format of Timestamp. The values are as follows:

1: PTP (see [[RFC8877](#)])

2: NTP (see [[RFC5905](#)])

3: POSIX

- o PeriodID: A 4 Octets period ID of the packet count.
- o TimestampSecond: 64bits, Integer value of the second part from 1970 to the time when the message is received.
- o TimestampNanoSecond:64bits, Integer value of the nanosecond part from 1970 to the time of receiving the message.

Note that if the clocks of nodes participating in flow measurement are unstable, clock synchronization between nodes is required. The clock synchronization mechanism used is outside the scope of this document. For example, NTP clock [RFC5905] or PTP clock [RFC8877] can be used.

6.1.4. Packet loss TLV

This TLV is used to notify measurement of packet loss to source node and is used in the end node model. The tlv is defined as follows:

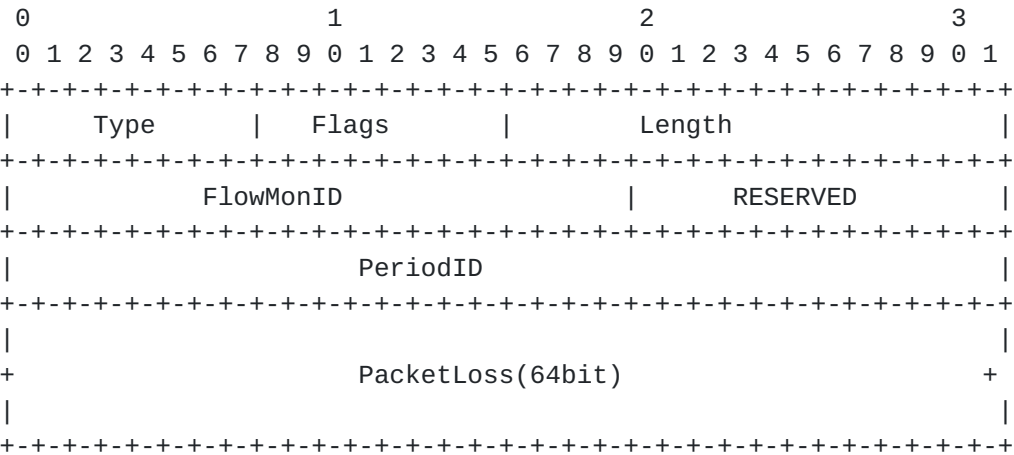


Figure 8: packet count TLV

- o Type: A one-octet field. Value 3 will be register in IANA.
- o Flags: A one-octet field.
- o Length: A two-octet field equal to the length of the Value field in octets.
- o FlowMonID: A 20 bits field, which is consistent with the definition in flow monitor option.

- o PeriodID: A 4 Octets period ID of the packet count.
- o PacketLoss: A 8 Octets count of packet loss in the period specified by periodID.

6.1.5. Packet delay TLV

This TLV is used to notify measurement of packet delay to source node and is used in the end node model. The tlv is defined as follows:

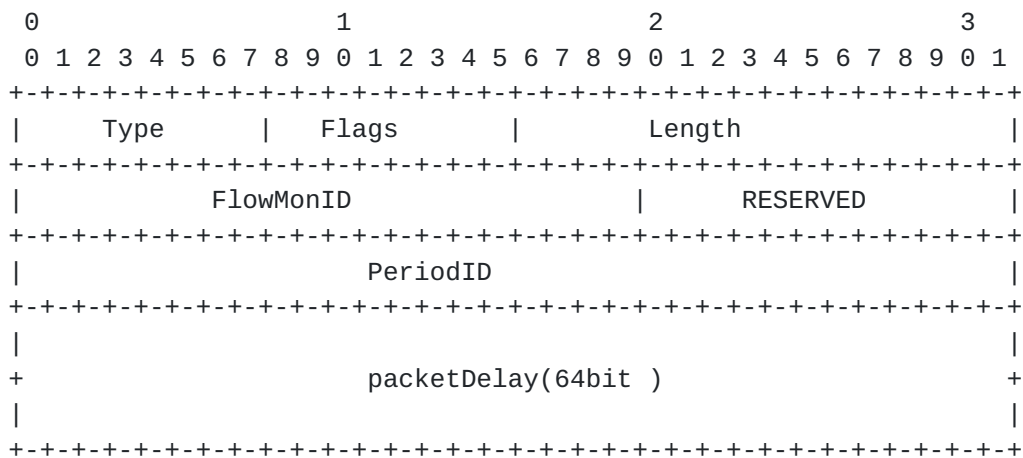


Figure 9: time stamp TLV

- o Type: A one-octet field. Value 4 will be register in IANA.
- o Flags: A one-octet field.
- o Length: A two-octet field equal to the length of the Value field in octets.
- o FlowMonID: A 20 bits field, which is consistent with the definition in flow monitor option.
- o PeriodID: A 4 Octets period ID of the packet count.
- o packetDelay: 64bits field of nanosecond, which is the packet delay in the period specified by PeroidID.

6.1.6. Average Packet loss TLV

This TLV is used to notify measurement of average packet loss to source node and is used in the end node model. The TLV is defined as follows:

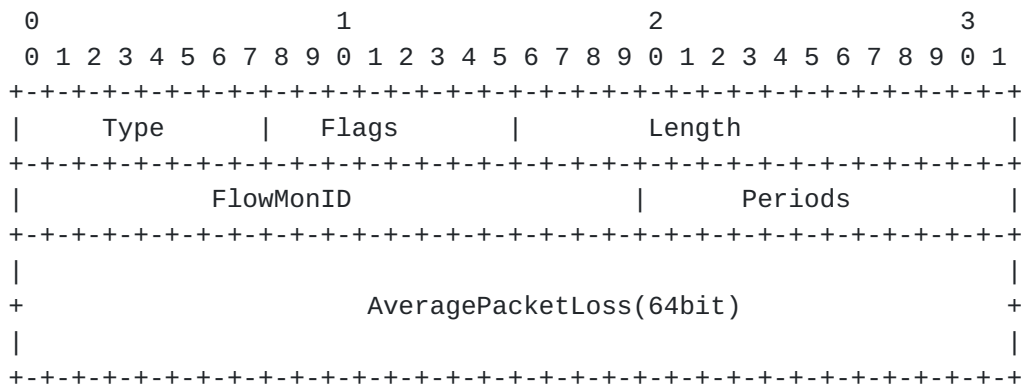


Figure 10: Average packet loss TLV

- o Type: A one-octet field. Value 3 will be register in IANA.
- o Flags: A one-octet field.
- o Length: A two-octet field equal to the length of the Value field in octets.
- o FlowMonID: A 20 bits field, which is consistent with the definition in flow monitor option.
- o Periods A 12 bits field, which identifies the number of periods used to calculate the average packet loss. It can be determined based on the capacity or configuration of the end node.
- o AveragePacketLoss: A 8 Octets count of packet loss, which is the Average Packet loss in the past periods.

6.1.7. Average Packet delay TLV

This TLV is used to notify measurement of average packet delay to source node and is used in the end node model. The TLV is defined as follows:

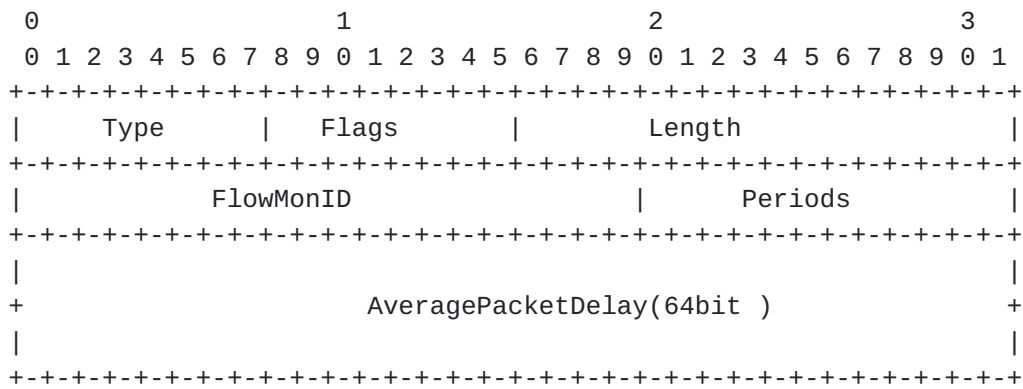


Figure 11: Average packet delay TLV

- o Type: A one-octet field. Value 5 will be register in IANA.
- o Flags: A one-octet field.
- o Length: A two-octet field equal to the length of the Value field in octets.
- o FlowMonID: A 20 bits field, which is consistent with the definition in flow monitor option.
- o Periods: A 12 bits field, which identifies the number of periods used to calculate the average packet delay. The number of periods used to calculate the average value could base on the capacity or configuration of the end node.
- o AveragePacketDelay: 64bits field of nanosecond, which is the Average Packet delay in the past periods.

6.2. Transport channel

The methods used by the data and result notification channels are out of the scope of this draft, and the following methods can be considered.

6.2.1. Independent controller protocol

Notify the statistical results or collection data of the source node through an independent controller protocol. UDP can be considered as the transport protocol.

A specific UDP port will be registered in IANA in the future for distributed flow measurement, or the UDP port number can be

designated on each node through configuration, such as CLI and NETCONF.

6.2.2. Extending BGP Protocol

For end-to-end measurement type, only source and end nodes are involved. In the scenario where BGP is deployed, the collection data or result can be carried by extending BGP protocol.

This method requires a new definition of BGP measurement address family, which is used to publish collection data and results.

6.2.3. Reverse traffic

This method is only applicable to end-to-end measurement type too. The end node could carry the collection data and results to the source node through reverse data flow.

7. Application of measurement results

Using the distributed flow measurement method described in this document, the source node can obtain the quality results of the actual traffic forwarding path faster. According to different actual needs, the source node could present the measurement results and optimize the path based on the measurement results, and more other application.

As illustrated in the figure below, in the SRv6 scenario, the traffic from CE1 to CE2 requires the SLA of low delay. There are two paths on PE1 to form a primary-slave relationship.

Path1: PE1->P1->P2->PE2->CE2

Path2: PE1->P3->P4->PE3->CE2

Path1 is the primary path.

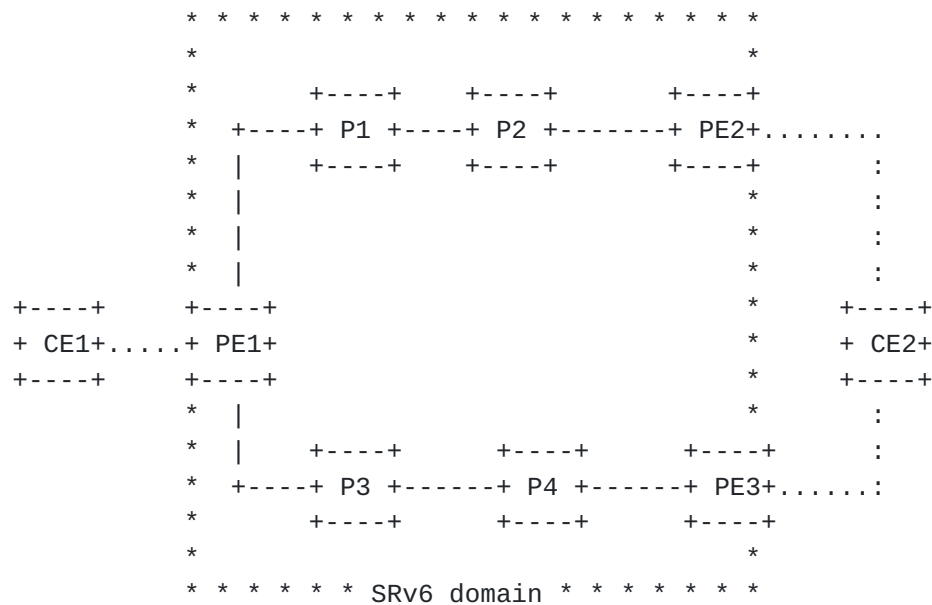


Figure 12: reference topology

The distributed flow measurement function can be deployed to measure the quality of the path. PE1, as the source node of the measurement, adopts the tail node mode. The end nodes PE2 of primary path complete the calculation of the measurement results and notify PE1.

When PE1 finds out that the delay of path 1 exceeds the threshold, it can immediately start the switching between the primary and standby paths, switch the traffic to the standby path, and send an alarm message. Then, the end node PE3 of standby path continues to measure the flow and notifies PE1 of the measurement results.

More kinds of applications based on measure results on source nodes are not in the scope of this document.

8. IANA Considerations

TBD

9. Security Considerations

The potential security threats of Alternate-Marking method have been described in detail in [Section 10](#) of [I-D.draft-ietf-ippm-rfc8321bis]. The performance measurement method described in this document does not introduce additional new security issues.

10. References

10.1. Normative References

- [I-D.wang-ippm-ipv6-flow-measurement]Wang, H.,Liu, Y., Lin, C., Xiao, M., "Flow Measurement in IPv6 Network", [draft-wang-ippm-ipv6-flow-measurement-02](#)(work in progress), August 2022.
- [I-D.ietf-6man-ipv6-alt-mark]Fioccola, G., Zhou, T., Cociglio, M., Qin, F., and R.Pang, "IPv6 Application of the Alternate Marking Method",[draft-ietf-6man-ipv6-alt-mark-16](#) (work in progress), July 2022.
- [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>>.
- [I-D.[draft-ietf-ippm-rfc8321bis](#)] Fioccola, G., Ed., Cociglio, M., Mirsky, G., Mizrahi, T., Zhou, T., "Alternate-Marking Method", [draft-ietf-ippm-rfc8321bis-03](#) (work in progress), July 2022.
- [RFC8877] Mizrahi, T., Fabini, J., and A. Morton, "Guidelines for Defining Packet Timestamps", [RFC 8877](#), DOI 10.17487/RFC8877, September 2020, <<https://www.rfc-editor.org/info/rfc8877>>.
- [RFC5905] Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", [RFC 5905](#), DOI 10.17487/RFC5905, June 2010, <<https://www.rfc-editor.org/info/rfc5905>>.

Authors' Addresses

Haojie Wang
China Mobile
Beijing
China

Email: wanghaojie@chinamobile.com

Sijun Weng
China Mobile
Beijing
China

Email: wengsijun@chinamobile.com

Changwang Lin
New H3C Corporation
Beijing
China

Email: linchangwang.04414@h3c.com

Xiao Min
ZTE Corporation
Beijing
China

Email: xiao.min2@zte.com.cn

Greg Mirsky
Ericsson

Email: gregimirsky@gmail.com

Wang, et al.

Expires July, 2024

[Page 17]