Network Working Group
Internet-Draft
Intended status: Informational
Expires: August 21, 2015

M. Chen, Ed.
L. Zheng, Ed.
Huawei Technologies
G. Mirsky, Ed.
Ericsson
G. Fioccola, Ed.
Telecom Italia

February 17, 2015

IP Flow Performance Measurement Framework
draft-chen-ippm-coloring-based-ipfpm-framework-03

Abstract

This document specifies a measurement method, the IP flow performance measurement (IPFPM). With IPFPM, data packets are marked into different blocks of markers by changing one or more bits of packets. No additional delimiting packet is needed and the performance be measured in-service and in-band without the insertion of additional traffic.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction 2
   1.1. Author's List 4
2. Terminology 4
3. Overview and Concept 4
4. Reference Model and Functional Components 6
   4.1. Reference Model 6
   4.2. Measurement Control Point 6
   4.3. Measurement Agent 7
5. Period Number 7
6. Re-ordering Tolerance 8
7. Packet Loss Measurement 8
8. Packet Delay Measurement 10
9. Consideration on Marking Bits Selection 11
10. IANA Considerations 11
11. Security Considerations 11
12. Acknowledgements 12
13. References 12
   13.1. Normative References 12
   13.2. Informative References 12
Authors' Addresses 14

1. Introduction

Performance Measurement (PM) is an important tool for service provider for Service Level Agreement (SLA) verification, troubleshooting (e.g., fault localization or fault delimitation) and network visualization. Measurement methods could be roughly put into two categories - active measurement method and passive measurement method. Active method measures performance or reliability parameters by the examination of traffic (IP Packets) injected into the network, expressly for the purpose of measurement by intended measurement

point . On the contrary, passive method measures some performance or reliability parameter associated with the existing traffic (packets) on the network. Both passive and active methods have their strengths and should be regarded as complementary. There are certain scenarios where active measurements alone is not enough or applicable and passive measurements are desirable[I-D.deng-ippm-passive-wireless-usecase].

With active measurement method, the rate, numbers and interval between the injected packets may affect the accuracy of the results. And for injected test packets it may not be guaranteed to always be in-band with the data traffic in the pure IP network due to Equal Cost Multi-Path (ECMP).

The Multiprotocol Label Switching (MPLS) PM protocol [RFC6374] for packet loss could be considered an example of passive performance measurement method. By periodically inserting auxiliary Operations, Administration and Maintenance (OAM) packets, the traffic is delimited by the OAM packets into consecutive blocks, and the receivers count the packets and calculate the packets loss each block. However, solutions like [RFC6374] depend on the fixed positions of the delimiting OAM packets for packets counting, and thus are vulnerable to out-of-order arrival of packets. This could happen particularly with out-of-band OAM channels, but might also happen with in-band OAM because of the presence of multipath forwarding within the network. Out of order delivery of data and the delimiting OAM can give rise to inaccuracies in the performance measurement figures. The scale of these inaccuracies will depend on data speeds and the variation in delivery, but with out-of-band OAM, this could result in significant differences between real and reported performance.

This document specifies a different measurement method, the IP flow performance measurement (IPFPM). With IPFPM, data packets are marked into different blocks of markers by changing one or more bits of packets without altering normal processing in the network. No additional delimiting packet is needed and the performance can be measured in-service without the insertion of additional traffic. Furthermore, because marking based IP performance measurement does not require extra OAM packets for traffic delimitation, it can be used in situations where there is packets re-ordering. IP Flow Information eXport (IPFIX) [RFC7011] is used for reporting the measurement data of IPFPM to a central calculation element for performance metrics calculation. Several new Information Elements of IPFIX are defined for IPFPM. These are described in the companion document [I-D.chen-ippm-ipfpm-report].

1.1. Author's List

Hongming Liu, Huawei Technologies

Yuanbin Yin, Huawei Technologies

Rajiv Papneja, Huawei Technologies

Shailesh Abhyankar, Vodafone

Guangqing Deng, CNNIC

Yongliang Huang, China Unicom

2. Terminology

The acronyms used in this document will be listed here.

3. Overview and Concept

The concept of marking IP packets for performance measurement is described in [I-D.tempia-opsawg-p3m]. Marking of packets in a specific IP flow to different markings divides the flows into different consecutive blocks. Packets in a block have same marking and consecutive blocks will have different markings. This enables the measuring node to count and calculate packet loss and/or delay based for each block of markers without any additional auxiliary OAM packets. The following figure (Figure 1) is an example that illustrates the different markings in a single IP flow in alternate 0 and 1 blocks.

```
    |  0  Block  |   1  Block |  0  Block  |   1  Block |
     000000000000 111111111111 000000000000 111111111111
```

Figure 1: Packet Marking

For packet loss measurement, there are two ways to mark packets: fixed packet numbers or fixed time period for each block of markers. This document considers only fixed time period method. The sender and receiver nodes count the transmitted and received packets/octets based on each block of markers. By counting and comparing the transmitted and received packets/octets, the packet loss can be detected.

For packet delay measurement, there are three solutions. One is similar to the packet loss, that it still marks the IP flows to different blocks of markers and uses the time of the marking change

as the reference time for delay calculations. This solution requires that there must not be any out-of-order packets; otherwise, the result will not be accurate. Because it uses the first packet of each block of markers for delay measurement, if there is packet reordering, the first packet of each block at the sender will be probably different from the first packet of the block at the receiver. An alternate way is to periodically mark a single packet in the IP flow. Within a given time period, there is only one packet that can be marked. The sender records the timestamp when the marked packet is transmitted, the receiver records the timestamp when detecting the marked packet. With the two timestamps, the packet delay can be computed. An additional method consists in taking into account the average arrival time of the packets within a single block (i.e. the same block of markers used for packet loss measurement). The network device locally sums all the timestamps and divides by the total number of packets received, so the average arrival time for that block of packets can be calculated. By subtracting the average arrival times of two adjacent devices it is possible to calculate the average delay between those nodes. This method is robust to out of order packets and also to packet loss (only an error is introduced dependent from the number of lost packets).

A centralized calculation element Measurement Control Point (MCP) is introduced in Section 4.2 of this document, to collect the packet counts and timestamps from the senders and receivers for metrics calculation. The IP Flow Information eXport (IPFIX) [RFC7011] protocol is used for collecting the performance measurement statistic information [draft-chen-ippm-ipfpm-report]. For the statistic information collected, the MCP has to know exactly what packet pair counts (one from the sender and the other is from the receiver) are based on the same block of markers and a pair of timestamps (one from the sender and the other is from the receiver) are based on the same marked packet. In case of average delay calculation the MCP has to know in addition to the packet pair counters also the pair of average timestamps for the same block of markers. The "Period Number" based solution Section 5 is introduced to achieve this.

For a specific IP flow to be measured, there may be one or more upstream and downstream Measurement Agents (MAs)(Section 4.3). An IP flow can be identified by the Source IP (SIP) and Destination IP (DIP) addresses, and it may combine the SIP and DIP with any or all of the Protocol number, the Source port, the Destination port, and the Type of Service (TOS) to identify an IP flow. For each flow, there will be a flow identifier that is unique within a certain administrative domain. To simplify the process description, the flows discussed in this document are all unidirectional. A bidirectional flow can be seen as two unidirectional flows.

IFPFM supports the measurement of Multipoint-to-Multipoint (MP2MP) flow, which satisfy all the scenarios that include Point-to-Point (P2P), Point-to-Multipoint (P2MP), Multipoint-to-Point (MP2P), and MP2MP. P2P scenario is obvious and can be used anywhere. P2MP and MP2P are very common in mobile backhaul networks. For example, a Cell Site Gateway (CSG) multi-homing to two Radio Network Controller (RNC) Site Gateways (RSGs) is a typical network design. When there is a failure, there is a requirement to monitor the flows between the CSG and the two RSGs hence to determine whether the fault is in the transport network or in the wireless network (typically called "fault delimitation"). This is especially useful in the situation where the transport network belongs to one service provider and wireless network belongs to other service providers.

4. Reference Model and Functional Components

4.1. Reference Model

The outline of the measurement system of large-scale measurement platforms (LMAP) network is introduced in [I-D.ietf-lmap-framework]. It describes the main functional components of the LMAP measurement system, and the interactions between the components. The Measurement Agent (MA) of IPFPM could be considered equivalent to the MA of LMAP. The Measurement Control Point (MCP) of IPFPM could be considered as the combined function of Controller and Collector. IP Flow Information eXport (IPFIX) [RFC7011] protocol is used for collecting the performance measurement data on the MAs and reported to MCP. The details are specified in the companion document [draft-chen-ippm-ipfpm-report]. The Control between MCP and MAs are left for future study. Figure 1 gives the reference model of IPFPM.

```
                                     +-----+
                              +------| MCP |------+
                              |      +-----+      |
                    +-----+   |  +---/     \---+  |   +-----+
                    | MA1 |---+  |             |  +---| MA3 |
                    +-----+      |             |      +-----+
                    +-----+      |             |      +-----+
                    | MA2 |------+             +------| MA4 |
                    +-----+                           +-----+
```

Figure 1: IPFPM Reference Model

4.2. Measurement Control Point

The Measurement Control Point (MCP) is responsible for collecting the measurement data from the Measurement Agents (MAs) and calculating the performance metrics according to the collected measurement data. For packet loss, based on each block of markers, the difference

between the total counts received from all upstream MAs and the total counts received from all downstream MAs are the lost packet numbers. The MCP must make sure that the counts from the upstream MAs and downstream MAs are related to the same marking/packets block. For packet delay (e.g., one way delay), the difference between the timestamps from the downstream MA and upstream MA is the packet delay. Similarly to packet loss, the MCP must make sure the two timestamps are based on the same marked packet. This document introduces a Period Number (PN) based synchronization mechanism which is discussed in details in Section 5.

4.3. Measurement Agent

The Measurement Agent (MA) executes the measurement actions (e.g., marks the packets, counts the packets, records the timestamps, etc.), and reports the data to the Measurement Control Point (MCP). Each MA maintains two timers, one (C-timer, used at upstream MA) is for marking change, the other (R-timer, used at downstream MA) is for reading the packet counts and timestamps. The two timers have the same time interval but are started at different time. A MA can be either an upstream or a downstream MA: the role is specific to an IP flow to be measured. For a specific IP flow, the upstream MA will change the marking and read the packet counts and timestamps when the C-timer expires, the downstream MA just reads the packets counts and timestamps when the R-timer expires. The MA may delay the reading for certain time when R-timer expires, in order to tolerant certain degree of packet re-ordering. Section 6 describes this in details.

For each Measurement Task (corresponding to an IP flow) [I-D.ietf-lmap-framework], a MA maintains a pairs of packet counters and a timestamp counter for each block of markers. As for the pair of packet counters, one is for counting packets and the other is for counting octets.

5. Period Number

When data are collected on the upstream MA and downstream MA, e.g. packet counts or timestamps, and periodically reported to the MCP, certain synchronization mechanism is required to ensure the data collected are correlated. This document introduces the Period Number (PN) to help the MCP to determine whether any two or more packet counts (from distributed MAs) are related to the same block of markers or any two timestamps are related to the same marked packet.

Period Number assures the data correlation by literally split the packets into different measurement period. The PN is generated each time a MA reads the packet counts or timestamps, and is associated with Each packet count and timestamp reported to the MCP. When the

MCP see e.g. two PN associated with two packet counts from a upstream and downstream MA, it consider this two packet counts are for the same measurement period by the same PN, i.e. this two packet counts are related to the same block of markers. The assumption is that the upstream and downstream MAs are time synchronized. This requires that the upstream and downstream MAs having a certain time synchronization capability (e.g., supporting the Network Time Protocol (NTP) [RFC5905], or the IEEE 1588 Precision Time Protocol (PTP) [IEEE1588].) The PN is calculated as the modulo of the local time (when the counts or timestamps are read) and the interval of the marking time period.

6. Re-ordering Tolerance

In order to allow for a certain degree of packets re-ordering, the R-timer on downstream MA should be started delta-t (&#916;t) later after the C-timer is started. Delta-t is a defined period of time and should satisfies the following conditions:

(Time-L - Time-MRO) < &#916;t < (Time-L + Time-MRO)

Where

Time-L: the link delay time between the sender and receiver;

Time-MRO: the maximum re-ordering time difference; if a packet is expected to arrive at t1 but actually arrives at t2, then the Time-MRO = | t2 - t1|.

So, the R-timer should be started at "t + &#916;t" (where the t is the time when C-timer started).

For simplicity, the C-timer should be started at the beginning of each time period. This document recommends the implementation to support at least these time periods (1s, 10s, 1min, 10min and 1h). So, if the time period is 10s, the C-timer should be started at the time of any multiples of 10 in seconds (e.g., 0s, 10s, 20s, etc.), then the R-timer should be started (e.g., 0s+&#916;t, 10s+&#916;t, 20s+&#916;t, etc.). With this method, each MA can independently start its C-timer and R-timer given that the clocks have been synchronized.

7. Packet Loss Measurement

To simplify the process description, the flows discussed in this document are all unidirectional. A bidirectional flow can be seen as two unidirectional flows. For a specific flow, there will be

upstream and downstream MAs and upstream and downstream packet counts/timestamp accordingly.

For packet loss measurement, this document defines the following counters and quantities:

U-CountP[n][m]: U-CountP is a two-dimensional array that stores the number of packets transmitted by each upstream MA in each marking time period. Specifically, parameter "n" is the "period number" of measured blocks of markers while parameter "m" refers to the m-th MA of the upstream MAs.

D-CountP[n][m]: D-CountP is a two-dimensional array that stores the number of packets received by each downstream MA in each marking time period. Specifically, parameter "n" is the "period number" of measured blocks of markers while parameter "m" refers to the m-th MA of the downstream MAs.

U-CountO[n][m]: U-CountO is a two-dimensional array that stores the number of octets transmitted by each upstream MA in each marking time period. Specifically, parameter "n" is the "period number" of measured blocks of markers while parameter "m" refers to the m-th MA of the upstream MAs.

D-CountO[n][m]: D-CountO is a two-dimensional array that stores the number of octets received by each downstream MA in each marking time period. Specifically, parameter "n" is the "period number" of measured blocks of markers while parameter "m" refers to the m-th MA of the downstream MAs.

LossP: the number of packets transmitted by the upstream MAs but not received at the downstream MAs.

LossO: the total octets transmitted by the upstream MAs but not received at the downstream MAs.

The total packet loss of a flow can be computed as follows:

LossP = U-CountP[1][1] + U-CountP[1][2] + ... + U-CountP[n][m] -
D-CountP[1][1] - D-CountP[1][2] - .... - D-CountP[n][m'].

LossO = U-CountO[1][1] + U-CountO[1][2] + ... + U-CountO[n][m] -
D-CountO[1][1] - D-CountO[1][2] - .... - D-CountO[n][m'].

Where the m and m' are the number of upstream MAs and downstream MAs, respectively.

8. Packet Delay Measurement

For packet delay measurement, there will be only one upstream MA and may be one or more (P2MP) downstream MAs. Although the marking based IPFPM supports P2MP model, this document only discusses P2P model, the P2MP model is left for future study. This document defines the following timestamps and quantities:

U-Time[n]: U-Time is a one-dimension array that stores the time when marked packets are sent; in case the "average delay" method is being used, U-Time stores the average of the time when the packets of the same block are sent; parameter "n" is the "period number" of marked packets.

D-Time[n]: D-Time is a one-dimension array that stores the time when marked packets are received; in case the "average delay" method is being used, D-Time stores the average of the time when the packets of the same block are received; parameter "n" is the "period number" of marked packets. This is only for P2P model.

D-Time[n][m]: D-Time a two-dimension array that stores the time when the marked packet is received by downstream MAs at each marking time period; in case the "average delay" method is being used, D-Time stores the average of the times when the packets of the same block are received by downstream MAs at each marking time period. Here, parameter "n" is the "period number" of marked packets while parameter "m" refers to the m-th MA of the downstream MAs. This is for P2MP model which is left for future study.

One-way Delay[n]: The one-way delay metric for packet networks is described in [RFC2679]. The "n" identifies the "period number" of the marked packet.

One-way Delay[1] = D-Time[1] - U-Time[1].

One-way Delay[2] = D-Time[2] - U-Time[2].

...

One-way Delay[n] = D-Time[n] - U-Time[n].

In the case of two-way delay, the delay is the sum of the two one-way delays of the two flows that have the same MAs but have opposite directions.

Two-way Delay[1] = (D-Time[1] - U-Time[1]) + (D-Time'[1] - U-Time'[1]).

Two-way Delay[2] = (D-Time[2] - U-Time[2]) + (D-Time'[2] - U-Time'[2]).

...

Two-way Delay[n] = (D-Time[n] - U-Time[n]) + (D-Time'[n] - U-Time'[n]).

Where the D-Time and U-Time are for one forward flow, the D-Time' and U-Time' are for reverse flow.

9. Consideration on Marking Bits Selection

This document does not specify which bits should be used for marking; it primarily introduces options that the operator can choose for packet marking. This document introduces the following options:

1. For IPv4, there is only one bit (the last reserved bit of the Flag field of the IPv4 header) in IP header that can be used for marking. With one bit, at any time it can only be used for loss or delay measurement and cannot be used for packet loss and delay measurement simultaneously. In case of average delay coupled to packet loss measurement, a single bit is enough for both metrics together;

2. For IPv6, it can leverage the IPv6 extension header for marking, for example, adding a new option to the Hop-by-Hop Options header[RFC2460] for marking. More detail will be added in a future version or in a separate document.

For the above options, the operators should carefully think of the marking bits selection to make sure that the setting or changing of the marking bits SHOULD NOT affect the normal packet forwarding and process.

The implementations SHOULD provide knobs for operators to configure and change the marking bits according to their network design and policies.

10. IANA Considerations

This document makes no request to IANA.

11. Security Considerations

This document specifies a passive mechanism for measuring packet loss and delay within a Service Provider's network where the IP packets are marked with the unused bits in IP head field, and then inserting

additional OAM packets during the measurement is avoided. Obviously, such mechanism does not directly affect other applications running on the Internet but may lead to potential affects to the measurement itself.

First, the measurement itself may be affected by routers (or other network devices) along the path of IP packets intentionally altering the value of marking bits of packets. Just as mentioned before, the mechanism specified in this document is just in the context of one Service Provider's network, so the routers (or other network devices) are controllable and thus this kind of attack can be omitted.

Second, the measurement can be harmed by attackers injecting artificial traffic. Then authentication techniques, like digital signatures, may be used to guard against such kind of attack.

12. Acknowledgements

The authors would like to thank Adrian Farrel for his review, suggestion and comments to this document.

13. References

13.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

13.2. Informative References

[I-D.chen-ippm-ipfpm-report]
Chen, M., Zheng, L., Liu, H., Yin, Y., Papneja, R., Abhyankar, S., Deng, G., and Y. Huang, "IP Flow Performance Measurement Report", draft-chen-ippm-ipfpm-report-00 (work in progress), July 2014.

[I-D.deng-ippm-passive-wireless-usecase]
Lingli, D., Zheng, L., and G. Mirsky, "Use-cases for Passive Measurement in Wireless Networks", draft-deng-ippm-passive-wireless-usecase-01 (work in progress), January 2015.

[I-D.ietf-lmap-framework]
Eardley, P., Morton, A., Bagnulo, M., Burbridge, T., Aitken, P., and A. Akhter, "A framework for large-scale measurement platforms (LMAP)", draft-ietf-lmap-framework-10 (work in progress), January 2015.

[I-D.tempia-opsawg-p3m]
Capello, A., Cociglio, M., Castaldelli, L., and A. Bonda, "A packet based method for passive performance monitoring", draft-tempia-opsawg-p3m-04 (work in progress), February 2014.

[IEEE1588]
IEEE, "1588-2008 IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems", March 2008.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[RFC2679] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999.

[RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.

[RFC4656] Shalunov, S., Teitelbaum, B., Karp, A., Boote, J., and M. Zekauskas, "A One-way Active Measurement Protocol (OWAMP)", RFC 4656, September 2006.

[RFC5357] Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)", RFC 5357, October 2008.

[RFC5905] Mills, D., Martin, J., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

[RFC7011] Claise, B., Trammell, B., and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information", STD 77, RFC 7011, September 2013.

Authors' Addresses

Mach(Guoyi) Chen (editor)
Huawei Technologies

Email: mach.chen@huawei.com

Lianshu Zheng (editor)
Huawei Technologies

Email: vero.zheng@huawei.com

Greg Mirsky (editor)
Ericsson
USA

Email: gregory.mirsky@ericsson.com

Giuseppe Fioccola (editor)
Telecom Italia
Via Reiss Romoli, 274
Torino 10148
Italy

Email: giuseppe.fioccola@telecomitalia.it

Hongming Liu
Huawei Technologies

Email: liuhongming@huawei.com

Yuanbin Yin
Huawei Technologies

Email: yinyuanbin@huawei.com

Rajiv Papneja
Huawei Technologies

Email: Rajiv.Papneja@huawei.com

Shailesh Abhyankar
Vodafone
Vodafone House, Ganpat Rao kadam Marg Lower Parel
Mumbai 40003
India

Email: shailesh.abhyankar@vodafone.com

Guangqing Deng
CNNIC
4 South 4th Street, Zhongguancun, Haidian District
Beijing
China

Email: dengguangqing@cnnic.cn

Yongliang Huang
China Unicom

Email: huangyl@dipmt.com