Network Working Group
Internet-Draft
Obsoletes: 8889 (if approved)
Intended status: Standards Track
Expires: 30 March 2023

G. Fioccola, Ed.
Huawei Technologies
M. Cociglio
Telecom Italia
A. Sapio
Intel Corporation
R. Sisto
Politecnico di Torino
T. Zhou
Huawei Technologies

26 September 2022

Clustered Alternate-Marking Method
draft-ietf-ippm-rfc8889bis-04

Abstract

This document generalizes and expands Alternate-Marking methodology to measure any kind of unicast flow whose packets can follow several different paths in the network that can result in a multipoint-to-multipoint network. The network clustering approach is presented and, for this reason, the technique here described is called "Clustered Alternate-Marking". This document obsoletes RFC 8889.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

The Alternate-Marking Method, as described in [I-D.ietf-ippm-rfc8321bis], is applicable to a point-to-point path. The extension proposed in this document applies to the most general case of a multipoint-to-multipoint path and enables flexible and adaptive performance measurements in a managed network.

The Alternate-Marking methodology consists in splitting the packet flow into marking blocks and the monitoring parameters are the packet counters and the timestamps for each marking period. In some applications of the Alternate-Marking method, a lot of flows and nodes are to be monitored. Multipoint Alternate-Marking aims to reduce these values and makes the performance monitoring more flexible in case a detailed analysis is not needed. For instance, by considering n measurement points and m monitored flows, the order of magnitude of the packet counters for each time interval is n*m*2 (1 per color). The number of measurement points and monitored flows may vary and depends on the portion of the network we are monitoring (core network, metro network, access network) and the granularity (for each service, each customer). So if both n and m are high values, the packet counters increase a lot, and Multipoint Alternate-Marking offers a tool to control these parameters.

The approach presented in this document is applied only to unicast flows and not to multicast. Broadcast, Unknown Unicast, and Multicast (BUM) traffic is not considered here, because traffic replication is not covered by the Multipoint Alternate-Marking method. Furthermore, it can be applicable to anycast flows, and Equal-Cost Multipath (ECMP) paths can also be easily monitored with this technique.

[I-D.ietf-ippm-rfc8321bis] applies to point-to-point unicast flows and BUM traffic. For BUM traffic, the basic method of [I-D.ietf-ippm-rfc8321bis] can easily be applied link by link and therefore split the multicast flow tree distribution into separate unicast point-to-point links. While, this document and its Clustered Alternate-Marking method apply to multipoint-to-multipoint unicast flows, anycast, and ECMP flows.

Therefore, the Alternate-Marking method can be extended to any kind of multipoint-to-multipoint paths, and the network-clustering approach presented in this document is the formalization of how to implement this property and allow a flexible and optimized performance measurement support for network management in every situation.

Without network clustering, it is possible to apply Alternate-Marking only for all the network or per single flow. Instead, with network clustering, it is possible to use the partition of the network into clusters at different levels in order to provide the needed degree of detail. In some circumstances, it is possible to monitor a multipoint network by monitoring the network clusters, without examining in depth. In case of problems (packet loss is measured, or the delay is too high), the filtering criteria could be enhanced in order to perform a detailed analysis by using a different combination of clusters up to a per-flow measurement as described in [I-D.ietf-ippm-rfc8321bis].

This approach fits very well with the Closed-Loop Network and Software-Defined Network (SDN) paradigm, where the SDN orchestrator and the SDN controllers are the brains of the network and can manage flow control to the switches and routers and, in the same way, can calibrate the performance measurements depending on the desired accuracy. An SDN controller application can orchestrate how accurately the network performance monitoring is set up by applying the Multipoint Alternate-Marking as described in this document.

It is important to underline that, as an extension of [I-D.ietf-ippm-rfc8321bis], this is a methodology document, so the mechanism that can be used to transmit the counters and the timestamps is out of scope here.

This document assumes that the blocks are created according to a fixed timer as per [I-D.ietf-ippm-rfc8321bis]. Switching after a fixed number of packets is possible but it is out of scope here.

Note that the fragmented packets' case can be managed with the Alternate-Marking methodology and the same guidance provided in section 6 of [I-D.ietf-ippm-rfc8321bis] apply also in the case of Multipoint Alternate-Marking.

1.1. Summary of Changes from RFC 8889

This document defines the Multipoint Alternate-Marking Method, addressing ambiguities and overtaking its experimental phase in the original specification [RFC8889].

The relevant changes are:

* Added the recommendations about the different deployments in case one or two flag bits are available for marking (Section 9).

* Changed the structure to improve the readability.

- Removed the wording about the experimentation of the method and considerations that no longer apply.

- Revised the description of detailed aspects of the methodology, e.g. synchronization and timing.

It is important to note that all the changes are totally backward compatible with [RFC8889] and no new additional technique has been introduced in this document compared to [RFC8889].

1.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Terminology

The definitions of the basic terms are identical to those found in Alternate-Marking [I-D.ietf-ippm-rfc8321bis]. It is to be remembered that [I-D.ietf-ippm-rfc8321bis] is valid for point-to-point unicast flows and BUM traffic.

The important new terms that need to be explained are listed below:

Multipoint Alternate-Marking: Extension to [I-D.ietf-ippm-rfc8321bis], valid for multipoint-to-multipoint unicast flows, anycast, and ECMP flows. It can also be referred to as Clustered Alternate-Marking.

Flow definition: The concept of flow is generalized in this document. The identification fields are selected without any constraints and, in general, the flow can be a multipoint-to-multipoint flow, as a result of aggregate point-to-point flows.

Monitoring Network: Identified with the nodes of the network that are the measurement points (MPs) and the links that are the connections between MPs. The monitoring network graph depends on the flow definition, so it can represent a specific flow or the entire network topology as aggregate of all the flows. Each node of the monitoring network cannot be both a source and a destination of the flow.

Cluster: Smallest identifiable non-trivial subnetwork of the entire monitoring network graph that still satisfies the condition that the number of packets that go in is the same as the number

that go out. A cluster partition algorithm, such as that found in Section 5.1, can be applied to split the monitoring network into clusters.

Multipoint metrics: Packet loss, delay and delay variation are extended to the case of multipoint flows. It is possible to compute these metrics on the basis of multipoint paths in order to associate the measurements to a cluster, a combination of clusters, or the entire monitored network. For delay and delay variation, it is also possible to define the metrics on a single-packet basis, and it means that the multipoint path is used to easily couple packets between input and output nodes of a multipoint path.

The next section highlights the correlation with the terms used in RFC 5644 [RFC5644].

2.1. Correlation with RFC 5644

RFC 5644 [RFC5644] is limited to active measurements using a single source packet or stream. Its scope is also limited to observations of corresponding packets along the path (spatial metric) and at one or more destinations (one-to-group) along the path.

Instead, the scope of this memo is to define multiparty metrics for passive and hybrid measurements in a group-to-group topology with multiple sources and destinations.

RFC 5644 [RFC5644] introduces metric names that can be reused here but have to be extended and rephrased to be applied to the Alternate-Marking schema:

a. the multiparty metrics are not only one-to-group metrics but can be also group-to-group metrics;

b. the spatial metrics, used for measuring the performance of segments of a source to destination path, are applied here to clusters.

3. Flow Classification

A unicast flow is identified by all the packets having a set of common characteristics. This definition is inspired by RFC 7011 [RFC7011].

As an example, by considering a flow as all the packets sharing the same source IP address or the same destination IP address, it is easy to understand that the resulting pattern will not be a point-to-point connection, but a point-to-multipoint or multipoint-to-point connection.

In general, a flow can be defined by a set of selection rules used to match a subset of the packets processed by the network device. These rules specify a set of Layer 3 and Layer 4 header fields (identification fields) and the relative values that must be found in matching packets.

The choice of the identification fields directly affects the type of paths that the flow would follow in the network. In fact, it is possible to relate a set of identification fields with the pattern of the resulting graphs, as listed in Figure 1.

A TCP 5-tuple usually identifies flows following either a single path or a point-to-point multipath (in the case of load balancing). On the contrary, a single source address selects aggregate flows following a point-to-multipoint, while a multipoint-to-point can be the result of a matching on a single destination address. In the case where a selection rule and its reverse are used for bidirectional measurements, they can correspond to a point-to-multipoint in one direction and a multipoint-to-point in the opposite direction.

So the flows to be monitored are selected into the monitoring points using packet selection rules, which can also change the pattern of the monitored network.

Note that, more generally, the flow can be defined at different levels based on the potential encapsulation, and additional conditions that are not in the packet header can also be included as part of matching criteria.

The Alternate-Marking method is applicable only to a single path (and partially to a one-to-one multipath), so the extension proposed in this document is suitable also for the most general case of multipoint-to-multipoint, which embraces all the other patterns of Figure 1.

```
       point-to-point single path
           +------+      +------+      +------+
       ---<>  R1  <>----<>  R2  <>----<>  R3  <>---
           +------+      +------+      +------+
```

```
          point-to-point multipath
                           +------+
                          <>  R2  <>
                         / +------+ \
                        /            \
              +------+ /              \ +------+
          ---<>  R1  <>                <>  R4  <>---
              +------+ \              / +------+
                        \            /
                         \ +------+ /
                          <>  R3  <>
                           +------+
```

```
          point-to-multipoint
                                      +------+
                                     <>  R4  <>---
                                    / +------+
                          +------+ /
                         <>  R2  <>
                        / +------+ \
              +------+ /            \ +------+
          ---<>  R1  <>              <>  R5  <>---
              +------+ \              +------+
                        \ +------+
                         <>  R3  <>
                          +------+ \
                                    \ +------+
                                     <>  R6  <>---
                                      +------+
```

```
          multipoint-to-point
              +------+
          ---<>  R1  <>
              +------+ \
                        \ +------+
                        <>  R4  <>
                        / +------+ \
              +------+ /            \ +------+
          ---<>  R2  <>              <>  R6  <>---
              +------+              / +------+
                          +------+ /
                         <>  R5  <>
                        / +------+
              +------+ /
          ---<>  R3  <>
              +------+
```

```
          multipoint-to-multipoint
              +------+                +------+
          ---<>  R1  <>              <>  R6  <>---
              +------+ \            / +------+
                        \ +------+ /
                         <>  R4  <>
                          +------+ \
              +------+              \ +------+
          ---<>  R2  <>             <>  R7  <>---
              +------+ \            / +------+
                        \ +------+ /
                         <>  R5  <>
                        / +------+ \
              +------+ /            \ +------+
          ---<>  R3  <>              <>  R8  <>---
              +------+                +------+
```

Figure 1: Flow Classification

The case of unicast flow is considered in Figure 1. The anycast flow is also covered, since it is only a special case of a unicast flow if routing is stable throughout the measurement period. Furthermore, an ECMP flow is in scope by definition, since it is a point-to-multipoint unicast flow.

4. Extension of the Method to Multipoint Flows

By using the Alternate-Marking method, only point-to-point paths can be monitored. To have an IP (TCP/UDP) flow that follows a point-to-point path, in general we have to define, with a specific value, 5 identification fields (IP Source, IP Destination, Transport Protocol, Source Port, Destination Port).

Multipoint Alternate-Marking enables the performance measurement for multipoint flows selected by identification fields without any constraints (even the entire network production traffic). It is also possible to use multiple marking points for the same monitored flow.

4.1. Monitoring Network

The monitoring network is deduced from the production network by identifying the nodes of the graph that are the measurement points, and the links that are the connections between measurement points. It can be modeled as a set of nodes and a set of directed arcs which connect pairs of nodes.

There are some techniques that can help with the building of the monitoring network (as an example, see [I-D.ietf-ippm-route]). In general, there are different options: the monitoring network can be obtained by considering all the possible paths for the traffic or periodically checking the traffic (e.g. daily, weekly, monthly) and updating the graph as appropriate, but this is up to the Network Management System (NMS) configuration.

So a graph model of the monitoring network can be built according to the Alternate-Marking method: the monitored interfaces and links are identified. Only the measurement points and links where the traffic has flowed have to be represented in the graph.

A simple example of a monitoring network graph is showed in Appendix A.

Each monitoring point is characterized by the packet counter that refers only to a marking period of the monitored flow. Also, it is assumed that there be a monitoring point at all possible egress points of the multipoint monitored network.

The same is also applicable for the delay, but it will be described in the following sections.

The rest of the document assumes that the traffic is going from left to right in order to simplify the explanation. But the analysis done for one direction applies equally to all directions.

4.2. Network Packet Loss

Since all the packets of the considered flow leaving the network have previously entered the network, the number of packets counted by all the input nodes is always greater than, or equal to, the number of packets counted by all the output nodes. It is assumed that routing is stable during the measurement period while packet fragmentation must be handled as described in [I-D.ietf-ippm-rfc8321bis].

In the case of no packet loss occurring in the marking period, if all the input and output points of the network domain to be monitored are measurement points, the sum of the number of packets on all the ingress interfaces equals the number on egress interfaces for the monitored flow. In this circumstance, if no packet loss occurs, the intermediate measurement points only have the task of splitting the measurement.

It is possible to define the Network Packet Loss of one monitored flow for a single period. In a packet network, the number of lost packets is the number of packets counted by the input nodes minus the number of packets counted by the output nodes. This is true for every packet flow in each marking period.

The monitored network packet loss with n input nodes and m output nodes is given by:

PL = (PI1 + PI2 +...+ PIn) - (PO1 + PO2 +...+ POm)

where:

PL is the network packet loss (number of lost packets)

PIi is the number of packets flowed through the i-th input node in this period

POj is the number of packets flowed through the j-th output node in this period

The equation is applied on a per-time-interval basis and a per-flow basis:

> The reference interval is the Alternate-Marking period, as defined in [I-D.ietf-ippm-rfc8321bis].
>
> The flow definition is generalized here. Indeed, as described before, a multipoint packet flow is considered, and the identification fields can be selected without any constraints.

5. Network Clustering

The previous equation of Section 4.2 can determine the number of packets lost globally in the monitored network, exploiting only the data provided by the counters in the input and output nodes.

In addition, it is possible to leverage the data provided by the other counters in the network to converge on the smallest identifiable subnetworks where the losses occur.

As defined in Section 2, a cluster is a non-trivial subnetwork of the entire monitoring network graph that still satisfies the condition that the number of packets that go in is the same as the number that go out, if no packet loss occurs. According to this definition, a cluster should contain all the arcs emanating from its input nodes and all the arcs terminating at its output nodes. This ensures that we can count all the packets (and only those) exiting an input node again at the output node, whatever path they follow.

As for the entire monitoring network graph, the cluster is defined on a per-flow basis. In a completely monitored network (a network where every network interface is monitored), each network device corresponds to a cluster, and each physical link corresponds to two clusters (one for each device).

Clusters can have different sizes depending on the flow-filtering criteria adopted.

Moreover, sometimes clusters can be optionally simplified. For example, when two monitored interfaces are divided by a single router (one is the input interface, the other is the output interface, and the router has only these two interfaces), instead of counting exactly twice, upon entering and leaving, it is possible to consider a single measurement point. In this case, we do not care about the internal packet loss of the router.

It is worth highlighting that it might also be convenient to define clusters based on the topological information so that they are applicable to all the possible flows in the monitored network.

Note that, in case of translation or encapsulation, the cluster properties must also be invariant.

5.1. Algorithm for Clusters Partition

A simple algorithm can be applied in order to split the monitoring network into clusters. This can be done for each direction separately, indeed a node cannot be both a source and a destination. The clusters partition is based on the monitoring network graph, which can be valid for a specific flow or can also be general and valid for the entire network topology.

It is a two-step algorithm:

* Group the links where there is the same starting node;

* Join the grouped links with at least one ending node in common.

Considering that the links are unidirectional, the first step implies listing all the links as connections between two nodes and grouping the different links if they have the same starting node. Note that it is possible to start from any link, and the procedure will work. Following this classification, the second step implies eventually joining the groups classified in the first step by looking at the ending nodes. If different groups have at least one common ending node, they are put together and belong to the same set. After the application of the two steps of the algorithm, each one of the composed sets of links, together with the endpoint nodes, constitutes a cluster.

A simple application of the clusters partition is showed in Appendix A.

The algorithm, as applied in the example of a point-to-multipoint network, works for the more general case of multipoint-to-multipoint network in the same way. It should be highlighted that for a multipoint-to-multipoint network the multiple sources MUST mark coherently the traffic and MUST be synchronized with all the other nodes according to the timing requirements detailed in Section 8.

When the clusters partition is done, the calculation of packet loss, delay and delay variation can be made on a cluster basis. Note that the packet counters for each marking period permit calculating the packet rate on a cluster basis, so Committed Information Rate (CIR) and Excess Information Rate (EIR) could also be deduced on a cluster basis.

Obviously, by combining some clusters in a new connected subnetwork the packet-loss rule is still true. So it is also possible to consider combinations of clusters if and where it suits.

In this way, in a very large network, there is no need to configure detailed filter criteria to inspect the traffic. It is possible to check a multipoint network and, in case of problems, go deep with a step-by-step cluster analysis, but only for the cluster or combination of clusters where the problem happens.

In summary, once a flow is defined, the algorithm to build the clusters partition is based on topological information; therefore, it considers all the possible links and nodes that could potentially be crossed by the given flow, even if there is no traffic. So, if the flow does not enter or traverse all the nodes, the counters have a non-zero value for the involved nodes and a zero value for the other nodes without traffic; but in the end, all the formulas are still valid.

The algorithm described above is an iterative clustering algorithm since it executes steps in iterations, but it is also possible to apply a recursive clustering algorithm as detailed in [IEEE-ACM-ToN-MPNPM].

The complete and mathematical analysis of the possible algorithms for clusters partition, including the considerations in terms of efficiency and a comparison between the different methods, is in the paper [IEEE-ACM-ToN-MPNPM].

6. Multipoint Packet Loss Measurement

The Network Packet Loss, defined in Section 4.2, valid for the entire monitored flow, can easily be extended to each multipoint path (e.g., the whole multipoint network, a cluster, or a combination of clusters). In this way it is possible to calculate Multipoint Packet Loss that is representative of a multipoint path.

The same equation of Section 4.2 can be applied to a generic multipoint path like a cluster or a combination of clusters, where the number of packets are those entering and leaving the multipoint path.

By applying the algorithm described in Section 5.1, it is possible to split the monitoring network into clusters. Then, packet loss can be measured on a cluster basis for each single period by considering the counters of the input and output nodes that belong to the specific cluster. This can be done for every packet flow in each marking period.

7. Multipoint Delay and Delay Variation

The same line of reasoning can be applied to delay and delay variation. The delay measurement methods defined in [I-D.ietf-ippm-rfc8321bis] can be extended to the case of multipoint flows. It is important to highlight that both delay and delay-variation measurements make sense in a multipoint path. The delay variation is calculated by considering the same packets selected for measuring the delay.

In general, it is possible to perform delay and delay-variation measurements on the basis of multipoint paths or single packets:

- Delay measurements on the basis of multipoint paths mean that the delay value is representative of an entire multipoint path (e.g., the whole multipoint network, a cluster, or a combination of clusters).

- Delay measurements on a single-packet basis mean that it is possible to use a multipoint path just to easily couple packets between input and output nodes of a multipoint path, as described in the following sections.

7.1. Delay Measurements on a Multipoint-Paths Basis

7.1.1. Single-Marking Measurement

Mean delay and mean delay-variation measurements can also be generalized to the case of multipoint flows. It is possible to compute the average one-way delay of packets in one block, a cluster, or the entire monitored network.

The average latency can be measured as the difference between the weighted averages of the mean timestamps of the sets of output and input nodes. This means that, in the calculation, it is possible to weigh the timestamps with the number of packets for each endpoint.

Note that, since the one-way delay value is representative of a multipoint path, it is possible to calculate the two-way delay of a multipoint path by summing the one-way delays of the two directions, similarly to [I-D.ietf-ippm-rfc8321bis].

7.2. Delay Measurements on a Single-Packet Basis

7.2.1. Single- and Double-Marking Measurement

Delay and delay-variation measurements associated with only one picked packet per period, both single and double marked, cannot be easily performed in a multipoint scenario since there are some limitations:

> Single marking based on the first/last packet of the interval does not work properly. Indeed, by considering a point-to-multipoint scenario, it is not possible to recognize which path the first packet of each block takes over the multipoint flow in order to correlate it. This is also true for the general case of the multipoint-to-multipoint scenario.
>
> Double marking or multiplexed marking works but only through statistical means. In a point-to-multipoint scenario, by selecting only a single packet with the second marking for each block, it is possible to follow and calculate the delay for that picked packet. But the measurement can only be done for a single path in each marking period. To traverse all the paths of the multipoint flow, it can theoretically be done by continuing the measurement for the following marking periods and expect to span

all the paths. In the general case of a multipoint-to-multipoint path, it is also needed to take into account the multiple source nodes which complicate the correlation of the samples. In this case, it can be possible to select the second marked packet only for a source node at a time for each block and cover the remaining source nodes one by one in the next marking periods.

Note that, since the one-way delay measurement is done on a single-packet basis, it is always possible to calculate the two-way delay but it is not immediate since it is necessary to couple the measurement on each single path with the opposite direction. In this case the NMS can do the calculation.

If a delay measurement is performed for more than one picked packet and for all the paths of the multipoint flow in the same marking period, neither the single- nor the double-marking method are applicable in the multipoint scenario. The packets follow different paths and it becomes very difficult to correlate marked packets in a multipoint-to-multipoint path if there are more than one per period.

A desirable option is to monitor simultaneously all the paths of a multipoint path in the same marking period. For this purpose, hashing can be used, as reported in the next section.

7.2.2. Hashing Selection Method

RFCs 5474 [RFC5474] and 5475 [RFC5475] introduce sampling and filtering techniques for IP packet selection.

The hash-based selection methodologies for delay measurement can work in a multipoint-to-multipoint path and can be used either coupled to mean delay or stand-alone.

[IEEE-Network-PNPM] introduces how to use the hash method (RFC 5474 [RFC5474] and RFC 5475 [RFC5475]) combined with the Alternate-Marking method for point-to-point flows. It is also called Mixed Hashed Marking because it refers to the conjunction of the marking method and the hashing technique. It involves only the single marking, indeed it is supposed that double marking is not used with hashing. The coupling of the single marking with the hashing selection allows choosing a simplified hash function since the alternation of blocks gives temporal boundaries for the hashing samples. The marking batches anchor the samples selected with hashing and this eases the correlation of the hashing packets along the path. For example, in case a hashed sample is lost, it is confined to the considered block without affecting the identification of the samples for the following blocks.

Using the hash-based sampling, the number of samples in each block may vary a lot because it depends on the packet rate that is variable. A dynamic approach can help to have an almost fixed number of samples for each marking period, and this is a better option for making regular measurements over time. In the hash-based sampling, Alternate-Marking is used to create periods, so that hash-based samples are divided into batches, which allows anchoring the selected samples to their period. Moreover, in a dynamic hash-based sampling, it can be possible to dynamically adapt the length of the hash value to meet the current packet rate, so that the number of samples is bounded in each marking period.

In a multipoint environment, the hashing selection may be the solution for performing delay measurements on specific packets and overcoming the single- and double-marking limitations.

8. Synchronization and Timing

It is important to consider the timing aspects, since out-of-order packets happen and have to be handled as well, as described in [I-D.ietf-ippm-rfc8321bis].

However, in a multisource situation, an additional issue has to be considered. With multipoint path, the egress nodes will receive alternate marked packets in random order from different ingress nodes, and this must not affect the measurement.

So, if we analyze a multipoint-to-multipoint path with more than one marking node, it is important to recognize the reference measurement interval. In general, the measurement interval for describing the results is the interval of the marking node that is more aligned with the start of the measurement, as reported in Figure 2.

Note that the mark switching approach based on a fixed timer is considered in this document.

```
        time -> start         stop
        T(R1)   |-------------|
        T(R2)     |-------------|
        T(R3)        |------------|
```

Figure 2: Measurement Interval

In Figure 2, it is assumed that the node with the earliest clock (R1) identifies the right starting and ending times of the measurement, but it is just an assumption, and other possibilities could occur. So, in this case, T(R1) is the measurement interval, and its recognition is essential in order to make comparisons with other active/passive/hybrid Packet Loss metrics.

Regarding the timing constraints of the methodology, [I-D.ietf-ippm-rfc8321bis] already describes two contributions that are taken into account: the clock error between network devices and the network delay between the measurement points.

When we expand to a multipoint environment, we have to consider that there are more marking nodes that mark the traffic based on synchronized clock time. But, due to different synchronization issues that may happen, the marking batches can be of different lengths and with different offsets when they get mixed in a multipoint flow. According to [I-D.ietf-ippm-rfc8321bis], the maximum clock skew between the network devices is A. Therefore, the additional gap that results between the multiple sources can be incorporated into A.

```
...BBBBBBBBB | AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA | BBBBBBBBB...
             |<======================================>|
             |                   L                    |
...=========>|<==================><==================>|<==========...
             |         L/2                 L/2        |
             |<====>|                          |<====>|
                d   |                          |   d
                    |<========================>|
                    available counting interval
```

Figure 3: Timing Aspects

Moreover, it is assumed that the multipoint path can be modeled with a normal distribution, otherwise it is necessary to reformulate based on the type of distribution. Under this assumption, the definition of the guard band d is still applicable as defined in [I-D.ietf-ippm-rfc8321bis] and is given by:

d = A + D_avg + 3*D_stddev,

where A is the clock accuracy, D_avg is the average value of the network delay, and D_stddev is the standard deviation of the delay.

As shown in Figure 3 and according to [I-D.ietf-ippm-rfc8321bis], the condition that must be satisfied to enable the method to function properly is that the available counting interval must be > 0, and that means:

L - 2d > 0.

This formula needs to be verified for each measurement point on the multipoint path.

Note that the timing considerations are valid for both packet loss and delay measurements.

9. Recommendations for Deployment

The methodology described in the previous sections can be applied to various performance measurement problems, as also explained in [I-D.ietf-ippm-rfc8321bis]. [RFC8889] reports experimental examples and [IEEE-Network-PNPM] also includes some information about the deployment experience.

Different deployments are possible using one flag bit, two flag bits or hashing selection:

One flag: packet loss measurement MUST be done as described in Section 6 by applying the network clustering partition described in Section 5. Delay measurement MUST be done according to the Mean delay calculation representative of the multipoint path, as described in Section 7.1.1. Single-marking method based on the first/last packet of the interval cannot be applied, as mentioned in Section 7.2.1.

Two flags: packet loss measurement MUST be done as described in Section 6 by applying the network clustering partition described in Section 5. Delay measurement SHOULD be done on a single packet basis according to double-marking method Section 7.2.1. In this case the Mean delay calculation (Section 7.1.1) MAY also be used as a representative value of a multipoint path. The choice depends on the kind of information that is needed, as further detailed below.

One flag with hash-based selection: packet loss measurement MUST be done as described in Section 6 by applying the network clustering partition described in Section 5. Hash-based selection methodologies, introduced in Section 7.2.2, MUST be used for delay measurement.

Similarly to [I-D.ietf-ippm-rfc8321bis], there are some operational guidelines to consider for the purpose of deciding to follow the recommendations above and use one or two flags or one flag with hash-based selection.

> The Multipoint Alternate-Marking method utilizes specific flags in the packet header, so an important factor is the number of flags available for the implementation. Indeed, if there is only one flag available there is no other way, while if two flags are available the option with two flags can be considered in comparison with the option of one flag with hash-based selection.
>
> The duration of the Alternate-Marking period affects the frequency of the measurement and this is a parameter that can be decided on the basis of the required temporal sampling. But it cannot be freely chosen, as explained in Section 8.
>
> The Multipoint Alternate-Marking methodologies enable packet loss, delay and delay variation calculation, but in accordance with the method used (e.g. single-marking or double-marking or hashing selection), there is a different kind of information that can be derived. For example, to get measurements on a multipoint-paths basis, one flag can be used. To get measurements on a single-packet basis, two flags are preferred. For this reason, the type of data needed in the specific scenario is an additional element to take into account.
>
> The Multipoint Alternate-Marking methods imply different computational load depending on the method employed. Therefore, the available computational resources on the measurement points can also influence the choice. As an example, mean delay calculation may require more processing and it may not be the best option to minimize the computational load.

The experiment with Multipoint Alternate-Marking methodologies confirmed the benefits of the Alternate-Marking methodology ([I-D.ietf-ippm-rfc8321bis]), as its extension to the general case of multipoint-to-multipoint scenarios.

The Multipoint Alternate-Marking Method MUST only be applied to controlled domains, as per [I-D.ietf-ippm-rfc8321bis].

10. A Closed-Loop Performance-Management Approach

The Multipoint Alternate-Marking framework that is introduced in this document adds flexibility to Performance Management (PM), because it can reduce the order of magnitude of the packet counters. This allows an SDN orchestrator to supervise, control, and manage PM in large networks.

The monitoring network can be considered as a whole or split into clusters that are the smallest subnetworks (group-to-group segments), maintaining the packet-loss property for each subnetwork. The clusters can also be combined in new, connected subnetworks at different levels, depending on the detail we want to achieve.

An SDN controller or a Network Management System (NMS) can calibrate performance measurements, since they are aware of the network topology. They can start without examining in depth. In case of necessity (packet loss is measured, or the delay is too high), the filtering criteria could be immediately reconfigured in order to perform a partition of the network by using clusters and/or different combinations of clusters. In this way, the problem can be localized in a specific cluster or a single combination of clusters, and a more detailed analysis can be performed step by step by successive approximation up to a point-to-point flow detailed analysis. This is the so-called "closed loop".

This approach can be called "network zooming" and can be performed in two different ways:

1) change the traffic filter and select more detailed flows;

2) activate new measurement points by defining more specified clusters.

The network-zooming approach implies that some filters or rules are changed and that therefore there is a transient time to wait once the new network configuration takes effect. This time can be determined by the Network Orchestrator/Controller, based on the network conditions.

For example, if the network zooming identifies the performance problem for the traffic coming from a specific source, we need to recognize the marked signal from this specific source node and its relative path. For this purpose, we can activate all the available measurement points and better specify the flow filter criteria (i.e., 5-tuple). As an alternative, it can be enough to select packets from the specific source for delay measurements; in this case, it is possible to apply the hashing technique, as mentioned in the previous sections.

[I-D.song-opsawg-ifit-framework] defines an architecture where the centralized Data Collector and Network Management can apply the intelligent and flexible Alternate-Marking algorithm as previously described.

As for [I-D.ietf-ippm-rfc8321bis], it is possible to classify the traffic and mark a portion of the total traffic. For each period, the packet rate and bandwidth are calculated from the number of packets. In this way, the network orchestrator becomes aware if the traffic rate surpasses limits. In addition, more precision can be obtained by reducing the marking period; indeed, some implementations use a marking period of 1 sec or less.

In addition, an SDN controller could also collect the measurement history.

It is important to mention that the Multipoint Alternate-Marking framework also helps Traffic Visualization. Indeed, this methodology is very useful for identifying which path or cluster is crossed by the flow.

11. Security Considerations

This document specifies a method of performing measurements that does not directly affect Internet security or applications that run on the Internet. However, implementation of this method must be mindful of security and privacy concerns, as explained in [I-D.ietf-ippm-rfc8321bis].

12. IANA Considerations

This document has no IANA actions.

13. Contributors

Greg Mirsky Ericsson Email: gregimirsky@gmail.com

Tal Mizrahi Huawei Technologies Email: tal.mizrahi.phd@gmail.com

Xiao Min ZTE Corp. Email: xiao.min2@zte.com.cn

14. Acknowledgements

The authors would like to thank Martin Duke and Tommy Pauly for their assistance and their detailed and precious reviews.

15. References

15.1. Normative References

[I-D.ietf-ippm-rfc8321bis]
Fioccola, G., Cociglio, M., Mirsky, G., Mizrahi, T., and T. Zhou, "Alternate-Marking Method", Work in Progress, Internet-Draft, draft-ietf-ippm-rfc8321bis-03, 25 July 2022, <https://www.ietf.org/archive/id/draft-ietf-ippm-rfc8321bis-03.txt>.

[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>.

[RFC5475] Zseby, T., Molina, M., Duffield, N., Niccolini, S., and F. Raspall, "Sampling and Filtering Techniques for IP Packet Selection", RFC 5475, DOI 10.17487/RFC5475, March 2009, <https://www.rfc-editor.org/info/rfc5475>.

[RFC5644] Stephan, E., Liang, L., and A. Morton, "IP Performance Metrics (IPPM): Spatial and Multicast", RFC 5644, DOI 10.17487/RFC5644, October 2009, <https://www.rfc-editor.org/info/rfc5644>.

[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>.

15.2. Informative References

[I-D.ietf-ippm-route]
Alvarez-Hamelin, J. I., Morton, A., Fabini, J., Pignataro, C., and R. Geib, "Advanced Unidirectional Route Assessment (AURA)", Work in Progress, Internet-Draft, draft-ietf-ippm-route-10, 13 August 2020, <https://www.ietf.org/archive/id/draft-ietf-ippm-route-10.txt>.

[I-D.song-opsawg-ifit-framework]
Song, H., Qin, F., Chen, H., Jin, J., and J. Shin, "A Framework for In-situ Flow Information Telemetry", Work in Progress, Internet-Draft, draft-song-opsawg-ifit-framework-18, 6 September 2022, <https://www.ietf.org/archive/id/draft-song-opsawg-ifit-framework-18.txt>.

[IEEE-ACM-ToN-MPNPM]
IEEE/ACM TRANSACTION ON NETWORKING, "Multipoint Passive Monitoring in Packet Networks", DOI 10.1109/TNET.2019.2950157, 2019, <https://doi.org/10.1109/TNET.2019.2950157>.

[IEEE-Network-PNPM]
IEEE Network, "AM-PM: Efficient Network Telemetry using Alternate Marking", DOI 10.1109/MNET.2019.1800152, 2019, <https://doi.org/10.1109/MNET.2019.1800152>.

[RFC5474] Duffield, N., Ed., Chiou, D., Claise, B., Greenberg, A., Grossglauser, M., and J. Rexford, "A Framework for Packet Selection and Reporting", RFC 5474, DOI 10.17487/RFC5474, March 2009, <https://www.rfc-editor.org/info/rfc5474>.

[RFC7011] Claise, B., Ed., Trammell, B., Ed., and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information", STD 77, RFC 7011, DOI 10.17487/RFC7011, September 2013, <https://www.rfc-editor.org/info/rfc7011>.

[RFC8889] Fioccola, G., Ed., Cociglio, M., Sapio, A., and R. Sisto, "Multipoint Alternate-Marking Method for Passive and Hybrid Performance Monitoring", RFC 8889, DOI 10.17487/RFC8889, August 2020, <https://www.rfc-editor.org/info/rfc8889>.

Appendix A. Example of Monitoring Network and Clusters Partition

Figure 4 shows a simple example of a monitoring network graph:

```
                                                    +------+
                                                   <>  R6  <>---
                                                  / +------+
                           +------+     +------+ /
                          <>  R2  <>---<>  R4  <>
                         / +------+ \   +------+ \
                        /            \            \ +------+
              +------+ /   +------+   \ +------+   <>  R7  <>---
          ---<>  R1  <>---<>  R3  <>---<>  R5  <>   +------+
              +------+ \   +------+ \   +------+ \
                        \            \            \ +------+
                         \            \            <>  R8  <>---
                          \            \            +------+
                           \            \
                            \            \ +------+
                             \            <>  R9  <>---
                              \            +------+
                               \
                                \ +------+
                                 <>  R10 <>---
                                  +------+
```

Figure 4: Monitoring Network Graph

In the monitoring network graph example, it is possible to identify the clusters partition by applying this two-step algorithm described in Section 5.1.

The first step identifies the following groups:

1. Group 1: (R1-R2), (R1-R3), (R1-R10)

2. Group 2: (R2-R4), (R2-R5)

3. Group 3: (R3-R5), (R3-R9)

4. Group 4: (R4-R6), (R4-R7)

5. Group 5: (R5-R8)

And then, the second step builds the clusters partition (in particular, we can underline that Groups 2 and 3 connect together, since R5 is in common):

1. Cluster 1: (R1-R2), (R1-R3), (R1-R10)

2. Cluster 2: (R2-R4), (R2-R5), (R3-R5), (R3-R9)

3. Cluster 3: (R4-R6), (R4-R7)

4. Cluster 4: (R5-R8)

The flow direction here considered is from left to right. For the opposite direction, the same reasoning can be applied, and in this example, you get the same clusters partition.

In the end, the following 4 clusters are obtained:

```
          Cluster 1
                           +------+
                          <>  R2  <>---
                         / +------+
                        /
              +------+ /   +------+
          ---<>  R1  <>---<>  R3  <>---
              +------+ \   +------+
                        \
                         \
                          \
                           \
                            \
                             \
                              \
                               \
                                \ +------+
                                 <>  R10 <>---
                                  +------+


          Cluster 2
              +------+     +------+
          ---<>  R2  <>---<>  R4  <>---
              +------+ \   +------+
                        \
              +------+   \ +------+
          ---<>  R3  <>---<>  R5  <>---
              +------+ \   +------+
                        \
                         \
                          \
                           \
                            \ +------+
                             <>  R9  <>---
                              +------+
```

```
          Cluster 3
                          +------+
                         <>  R6  <>---
                        / +------+
              +------+ /
          ---<>  R4  <>
              +------+ \
                        \ +------+
                         <>  R7  <>---
                          +------+


          Cluster 4
              +------+
          ---<>  R5  <>
              +------+ \
                        \ +------+
                         <>  R8  <>---
                          +------+
```

Figure 5: Clusters Example

There are clusters with more than two nodes as well as two-node clusters. In the two-node clusters, the loss is on the link (Cluster 4). In more-than-two-node clusters, the loss is on the cluster, but we cannot know in which link (Cluster 1, 2, or 3).

Appendix B. Changes Log

Changes from RFC 8889 in draft-fioccola-rfc8889bis-00 include:

* Minor editorial changes

* Removed section on "Examples of application"

Changes in draft-fioccola-rfc8889bis-01 include:

* Considerations on BUM traffic

* Reference to RFC8321bis for the fragmentation part

* Revised section on "Delay Measurements on a Single-Packet Basis"

* Revised section on "Timing Aspects"

Changes in draft-fioccola-rfc8889bis-02 include:

- Clarified the formula in the section on "Timing Aspects" to be aligned with RFC 8321

- Considerations on two-way delay measurements in both sections 8.1 and 8.2 on delay measurements

- Clarified in section 4.1 on "Monitoring Network" that the description is done for one direction but it can easily be extended to all direction

- New section on "Results of the Multipoint Alternate Marking Experiment"

Changes in draft-fioccola-rfc8889bis-03 include:

- Moved and renamed section on "Timing Aspects" as "Synchronization and Timing"

- Renamed old section on "Multipoint Packet Loss" as "Network Packet Loss"

- New section on "Multipoint Packet Loss Measurement"

- Renamed section on "Multipoint Performance Measurement" as "Extension of the Method to Multipoint Flows"

Changes in draft-fioccola-rfc8889bis-04/draft-ietf-ippm-rfc8889bis-00 include:

- Revised section 5.1 on "Algorithm for Clusters Partition"

Changes in draft-ietf-ippm-rfc8889bis-01 include:

- New section on "Summary of Changes from RFC 8889"

Changes in draft-ietf-ippm-rfc8889bis-02 include:

- Revised sections on "Single- and Double-Marking Measurement", "Hashing Selection Method" and "Synchronization and Timing"

- Revised references

Changes in draft-ietf-ippm-rfc8889bis-03 include:

- Comments addressed from Last Call review

- Renamed section 9 as "Recommendations for Deployment"

Changes in draft-ietf-ippm-rfc8889bis-04 include:

- Comments addressed from Last Call review

Authors' Addresses

Giuseppe Fioccola (editor)
Huawei Technologies
Riesstrasse, 25
80992 Munich
Germany
Email: giuseppe.fioccola@huawei.com

Mauro Cociglio
Telecom Italia
Email: mauro.cociglio@outlook.com

Amedeo Sapio
Intel Corporation
4750 Patrick Henry Dr.
Santa Clara, CA 95054
United States of America
Email: amedeo.sapio@intel.com

Riccardo Sisto
Politecnico di Torino
Corso Duca degli Abruzzi, 24
10129 Torino
Italy
Email: riccardo.sisto@polito.it

Tianran Zhou
Huawei Technologies
156 Beiqing Rd.
Beijing
100095
China
Email: zhoutianran@huawei.com