ippm
Internet-Draft
Intended status: Experimental
Expires: 7 November 2022

R. Geib, Ed.
Deutsche Telekom

6 May 2022

A Connectivity Monitoring Metric for IPPM
draft-ietf-ippm-connectivity-monitoring-04

Abstract

Within a Segment Routing domain, segment routed measurement packets can be sent along pre-determined paths. This enables new kinds of measurements. Connectivity monitoring allows to supervise the state and performance of a connection or a (sub)path from one or a few central monitoring systems. This document specifies a suitable type-P connectivity monitoring metric.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 7 November 2022.

Copyright Notice

Copyright (c) 2022 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 (https://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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction

Within a Segment Routing domain, measurement packets can be sent along pre-determined segment routed paths [RFC8402]. A segment routed path may consist of pre-determined sub paths, specific router-interfaces or a combination of both. A measurement path may also consist of sub paths spanning multiple routers, given that all segments to address a desired path are available and known at the SR domain edge interface.

A Path Monitoring System (PMS, see [RFC8403]) is a dedicated central Segment Routing (SR) domain monitoring device (as compared to a distributed monitoring approach based on router-data and -functions only). Monitoring individual sub-paths or point-to-point connections is executed for different purposes. IGP exchanges hello messages between neighbors to keep alive routing and swiftly adapt routing to topology changes. Network Operators may be interested in monitoring connectivity and congestion of interfaces or sub-paths at a timescale of seconds, minutes or hours. In both cases, the periodicity is significantly smaller than commodity interface monitoring based on router counters, which may be collected on a minute timescale to keep the processing- or monitoring data-load low.

The IPPM architecture was a first step to that direction [RFC2330]. Commodity IPPM solutions require dedicated measurement systems, a large number of measurement agents and synchronised clocks. Monitoring a domain from edge to edge by commodity IPPM solutions increases scalability of the monitoring system. But localising the site of a detected network behaviour change may then require suitable network tomography methods.

The IPPM Metrics for Measuring Connectivity offer generic connectivity metrics [RFC2678]. These metrics allow to measure connectivity between end nodes without making any assumption on the paths between them. The metric and the type-p packet specified by this document follow a different approach: they are designed to monitor connectivity and performance of a specific single link or a path segment. The underlying definition of connectivity is partially the same: a packet not reaching a destination indicates a loss of connectivity. An IGP re-route may indicate a loss of a link, while it might not cause loss of connectivity between end systems. The metric specified here detects a loss of connectivity, defined by a complete absence of a path between two nodes in both directions of communication (whereas a re-routing will briefly disturb a path, but connectivity is restored by the network after a short disturbance).

A Segment Routing PMS is part of an SR domain. The PMS is IGP topology aware, covering the IP and (if present) the MPLS layer topology [RFC8402]. This allows to steer PMS measurement packets along arbitrary pre-determined concatenated sub-paths, identified by suitable Segment IDs. Basically, the SR connectivity metric as specified by this document requires set up of a number of constrained, overlaid measurement loops (or measurement paths). The delay of the packets sent along each of these measurement loops is measured. A single congested interface along a monitored sub-path adds latency along a unique subset of several measurement loops. If a monitored sub-path no longer provides IP/MPLS connectivity between two nodes, another unique subset of measurement loops will drop all traffic while connectivity is lost. The number of measurement loops required in total may be limited to one per sub-path (or connection) to be monitored, if a hub-and-spoke like sub-path topology as described below is monitored. In addition to information revealed by a commodity ICMP ping measurement, the metrics and methods specified here identify the location of a congested interface (or ingress of a congested sub-path, respectively). To do so, tomography assumptions and methods are combined to first plan the overlaid SR measurement loop set up and later on to evaluate the captured performance metrics.

There's another difference as compared with commodity ping: the measurement loop packets remain in the data plane of passed routers. These need to forward the measurement packets without any additional processing apart from that.

It is recommended to consider automated measurement loop set-up. The methods proposed here are error-prone if the topology and measurement loop design isn't followed properly. While details of an automated set-up are not within scope of this document, some formal defintions of constraints to be respected are given.

This document specifies type-p metrics determining properties of an SR path which allows to monitor connectivity and congestion of interfaces. The specified methods further allow to locate the path or interface which caused a change in the reported type-p metrics. This document is limited to the Segment Routing MPLS layer, but the methodology may be applied within SR domains or MPLS domains in general.

1.1. 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].

2. A brief segment routing connectivity monitoring framework

The Segment Routing IGP topology information consists of the IP and (if present) the MPLS layer topology. The minimum SR topology information consists of Node-Segment-Identifiers (Node-SID), identifying an SR router. The IGP exchange of Adjacency-SIDs [RFC8667], which identify local interfaces to adjacent nodes, is optional. It is RECOMMENDED to distribute Adj-SIDs in a domain operating a PMS to monitor connectivity as specified below. If Adj-SIDs aren't availbale, [RFC8287] provides methods how to steer packets along desired paths by the proper choice of an MPLS Echo-request IP-destination address. A detailed description of [RFC8287]

methods as a replacement of Adj-SIDs is out of scope of this document. Monitoring interfaces connecting nodes requires Adj-SIDs, if re-converged IP/MPLS layer connectivity would result in re-routing packets (and re-establishment of IP/MPLS layer connectivity) using Node-SIDs.

An active round trip measurement between two adjacent nodes is a simple method to monitor connectivity of a connecting link. If multiple links are operational between two adjacent nodes and only a single one fails, a single plain round trip measurement may fail to notice that or identify which link has failed. A round trip measurement further fails to identify which interface is congested, even if only a single link connects two adjacent nodes.

Segment Routing enables the set-up of extended measurement loops.

Several different measurement loops can be set up to form a partial overlay. If done properly, any network change impacts more than a single measurement loop's round trip delay or causes drops of packets of more than one loop. Randomly chosen measurement loop paths including the interfaces or paths to be monitored may fail to produce the desired unique result patterns, hence commodity network tomography methods aren't applicable [CommodityTomography]. The approach pursued here uses a pre-specified measurement loop overlay design to produce the desired results with a minimum effort.

A centralised monitoring approach doesn't require report collection and result correlation from two (or more) receivers. The metrics captured along different measurement loops however still need to be correlated.

An additional property of the measurement loop set-up specified below is that it allows to estimate the packet round trip delay of a monitored link or sub-path.

An example hub and spoke network, operated as SR domain, is shown below. The included PMS shown is supposed to monitor the connectivity of all the 6 links (a link is a simple and generic kind of sub-path) attaching the spoke-nodes L050, L060 and L070 to the hub-nodes L100 and L200. L300 only serves to connect the PMS to nodes L100 and L200.

```
      +----+  +-----+    +-----+
      |PMS|   |L100|-----|L050|
      +----+  +-----+\  /+-----+
        |    /    \  \_/_____
        |   /      \  /      \+-----+
     +-----+/       \/_  +----|L060|
     |L300|         /  |/     +-----+
     +-----+\      /   /\_
```

```
            \     /   /   \
             \+----+ /   +----+
              |L200|-----|L070|
              +----+     +----+
```

Figure 1

Example hub and spoke network allowing link connectivity verification with a PMS

The SID values are picked for convenient reading only. Node-SID: 100 identifies L100, Node-SID: 300 identifies L300 and so on. Adj-SID 10050: Adjacency L100 to L050, Adj-SID 10060: Adjacency L100 to L060, Adj-SID 60200: Adjacency L60 to L200 and so on (note that the Adj-SID are locally assigned per node interface, meaning two per link).

Monitoring the 6 links between hub nodes Ln00 (where n=1,2) and spoke nodes L0m0 (where m=5,6,7) requires 6 measurement loops, which have the following properties:

- Each measurement loop follows a single round trip from one hub Ln00 to one spoke L0m0 (e.g., from L100 and L050 and back to L100).

- Each measurement loop passes two more links: one between the same hub Ln00 and another spoke L0m0 and from there to the alternate hub Ln00 (e.g., from L100 to L060 and then from L060 to L200)

- Every monitored link is passed by a single round trip measurement loop only once and further only once unidirectional by two other loops. These latter, unidirectional measurement loop sections forward packets in opposing direction along the monitored link. In the end, three measurement loops pass each single monitored link (sub-path). In figure 1, e.g. the link between L100 and L050 is passed by one measurement loop following a round trip L100 to L050 (the measured delay is M1, see below), a second loop passes in direction L100 to L050 only (delay M3) and a third loop passes in direction L050 to L100 only (delay M6).

Note that any 6 links connecting two to five nodes can be monitored that way too. Further note that the measurement loop overlay chosen is optimised for 6 links and a hub and spoke topology of two to five nodes. The 'one measurement loop per measured sub-path' paradigm only works under these conditions.

The above overlay scheme results in 6 measurement loops for the given example. The start and end of each measurement loop is PMS to L300 to L100 or L200 and a similar sub-path on the return leg. These parts of the measurement loops are omitted here for brevity (some discussion may befound below). The following delays are measured along the SR paths of each measurement loop:

1. M1 is the delay along L100 -> L050 -> L100 -> L060 -> L200

2. M2 is the delay along L100 -> L060 -> L100 -> L070 -> L200

3. M3 is the delay along L100 -> L070 -> L100 -> L050 -> L200

4. M4 is the delay along L200 -> L050 -> L200 -> L060 -> L100

5. M5 is the delay along L200 -> L060 -> L200 -> L070 -> L100

6. M6 is the delay along L200 -> L070 -> L200 -> L050 -> L100

For brevity, in the following delay M1 also identifies the corresponding measurement loop number 1 and so on.

An example for a stack of Adj-SID segments the loop resulting in M1 is (top to bottom): 100 | 10050 | 50100 | 10060 | 60200 | PMS. As can be seen, the Node-SIDs 100 and PMS are present at top and bottom of the segment stack. Their purpose is to transport the packet from the PMS to the start of the measurement loop at L100 and return it to the PMS from its end. When connectivity is lost, a path determined by Adj-SIDs behaves deterministic: packets forwarded to an Adj-SID without connectivity to the neighboring node are dropped.

An example for a stack of a loop consisting of Node-SID segments allowing to capture M1 is (top to bottom): 100 | 050 | 100 | 060 | 200 | PMS.

The evaluation of the measurement loop round trip delays M1 - M6 allows to detect the follwing state-changes of the monitored sub-paths:

- If the loops are set up using Node-SIDs only, any single complete loss of connectivity caused by a failing single link between any Ln00 and any L0m0 node briefly disturbs three measurement loops and changes the delay measured along them. The traffic to the Node-SIDs is re-routed (in the case of a single link loss, no node is completely disconnected in the example network). In that case, a suitable metric characterising re-routing coupled with the loss of that single link is required. The change in propagation delay might be an approach for such a metric (if there is any delay change, as that depends on the resulting alternate route delay). A delay based connectiviy scheme may not work under all circumstances.

- If the measurement loops are set up using Adj-SIDs only, a loss of connectivity caused by a failing single link between any Ln00 and any L0m0 node terminates the traffic along three measurement loops. The packets of all three loops will be dropped, until the link gets back into service. Traffic to Adj-SIDs is not rerouted. Note that Node-SIDs may be used to foward the measurement packets from the PMS to the hub node, where the first sub-path to be monitored begins and from the hub node receiving the measurement from the last monitored sub path to the PMS.

- The simple example indicates superiority of Adj-SIDs over Node-SIDs only if links are monitored and the network architecture is similiar to the one shown in the figure. The generic advice is, that unambiguous connectivity monitoring is best based on packet loss, rather than on delay changes.

- A single congested interface between any Ln00 and any L0m0 node always only impacts the measured delay of two measurement loops.

- As an example, the formula to calculate the (sub-path) Round Trip Delay (RTD) for link L100-L050 is given here

  4 * RTD_L100-L050-L100 = 3 * M1 + M3 + M6 - M2 - M4 - M5.

  This formula is reproducible for all other links: sum up 3*RTD measured along the loop passing the monitored link of interest in round trip fashion, and add the RTDs of the two measurement loops passing the evaluated monitored link only in a single direction. From this sum subtract the RTD captured for the measurement loops not passing the monitored link evaluated to get four times the RTD of the monitored link evaluated.

A closer look reveals that any single event of interest for the proposed metric, which are a single loss of connectivity or a single case of congestion, only impacts a unique set of measurement loops which can be determined a-priori. If, e.g., connectivity is lost between L200 and L050, measurement loops M3, M4 and M6 indicate packet loss (or a change of the measured delay, if a Node-SID based approach is preferred).

As a second example: if the interface L070 to L100 is congested, measurement loops M3 and M5 indicate a change in the measured delay. Without listing all events, it can be shown that all cases of single losses of connectivity or single events of congestion influence only delay measurements of a unique set of measurement loops.

The measurement loops are best set up while there's no congestion. In that case, the congestion free RTDs of all monitored links can be calculated as shown above which later allows to estimate the queue-depth under congestion. A single congestion event adds queuing delay to the RTD measured of two specific measurement loops. The two measurement loops impacted indicate the congested interface and enable estimation of the queue-depth (in terms of seconds based on comparing actual and prior delay measurements). The per link RTD can be calculated while the network is operating without congestion, say at interval T0. Then as an example, assume a queue of an average depth of 20 ms to build up at interface L200 to L070 at interval T1. The measurement loops M5 and M6 are the only ones passing the interface in that direction. Both indicate an added delay along M5 and M6 of + 20 ms during a measurement interval T1 with congestion on this interface, while M1-4 indicate unchanged delays. The location of the congested interface is determined by the combination of the two (and only two) measurement loops M5 and M6 showing a significant delay increase. The average queue depth [s] = (M5[T1] - M5[T0] + M6[T1] - M6[T0])/2.

As mentioned there's a constant delay added for each measurement loop, which is the delay of the path passed from PMS -> L100 + L200 -> PMS. Please note, that this added delay is appearing twice in the formula resulting in the monitored link delay estimate of the example network. Then it is the RTD PMS -> L100 + RTD L200 -> PMS. Both

RTDs can be directly measured by two additional measurements Cor1 = RTD (PMS -> L100 -> PMS) and Cor2 = RTD (PMS -> L200 -> PMS). The monitored link RTD formula was linkRTDuncor = 3*Mx + My + Mz - Ms - Mt - Mu. The correct 4*linkRTDx = 4*linkRTDxuncor - Cor1 - Cor2.

If the interface between PMS and L100/L200 is congested, all measurement loops M1-M6 as well as Cor1 and Cor2 will see a change. A congested interface of a monitored link doesn't impact the RTDs captured by Cor1 and Cor2.

The measurement loops may also be set up between hub nodes L100 and L200, if that's preferred and supported by the nodes. In that case, the above formulas apply without correction.

3. Topology and measurement loop set up requirements

3.1. General network topology requirements

The metric and methods specified below can be applied to monitor networks or sub-paths forming a hub and spoke topology. A single sub-path status change of type loss of connectivity or congestion can be detected. The nodes don't have to act as hubs or spokes, this terminology is only chosen to describe a topology requirement. In detail, the topology to be monitored MUST meet the following constraints:

* The SR domain sub-paths to be monitored create a hub and spoke topology with a PMS connected to all hub nodes. The PMS may reside in a hub.

* Exactly 6 (six) sub-paths are monitored.

* The monitored sub-paths connect at least two and no more than 5 nodes.

* Every spoke node MUST have at least one path to every hub node.

* Every spoke node MUST at least be connected to one (or more) hub node(s) by two monitored sub-paths.

* Sub-paths between spokes can't be monitored and therefore are out of scope (the overlay measurement loops can't be set up as

desired).

Shared resources, like a Shared Risk Link Group (e.g., a single fiber bundle) or a shared queue passed by several logical links need to be considered during set up. Shared resources may either be desired or to be avoided. As an example, if a set of logical links share one parental scheduler queue, it is sufficient to monitor a single logical connection to monitor the state of that parental scheduler.

3.2. Sub-path Monitoring measurement loop routing requirements

The methodologies sepcified by this document REQUIRE a measurement loop path overlay of all path delay measurement streams Fi, i in [1, 2...6] as defined in this section. In the follwing, a path delay measurement stream Fi is called measurement (loop) Fi for brevity.

- Define the segment routed Sub-paths SPi, i in [1, 2...6] to be monitored. The Sub-paths SPi SHOULD not share resources, if the operator isn't aware of the impact of the shared resources on the measurement loops Fi and the methodologies defined below. The Sub-path SPi topology SHOULD respect the general network topology requirements as specified above.

- Set up i = 1, 2...6 measurement loops Fi thus that measurement Fi passes SPi and only SPi bidirectional (or by a round-trip) from Hub to Spoke and back. Note that the correspondance of SPi and Fi isn't strictly required. Measurement Fi thus however appears in all methodologies calculating a metric related to SPi.

- Set up the SR path per measurement loops Fj and Fk thus that SPi is passed by exactly one other measurement loop Fj unidirectional in direction Hub to Spoke and by exactly one other measurement loop Fk unidirectional in the opposite direction (Spoke to Hub). The measurement loop Fi != Fj != Fk. As a description, one measurement loop Fj pass SPi in "downstream" direction from Hub to Spoke, whereas measurement loop Fk passes SPi in "upstream" direction from Spoke to Hub.

- Set up each segment routed measurement loop path Fi thus that it passes SPi bidirectional as specified above, SPj unidirectional from Hub to Spoke and SPk unidirectional from Spoke to Hub. The monitored Sub-path SPi MUST NOT be equal to SPj and MUST NOT be equal to SPk.

* The measurement loop set up to monitor all Sub-paths SPi is completed, if:

  + Each Sub-path SPi is passed by exactly three measurements loops Fi, Fj and Fk as specified above.

  + Each segment routed measurement loop path Fi passes exactly three concatenated Sub-paths SPi, SPm and SPn as specified above (indices m and n are chosen here only to avoid misconceptions which may result from picking indices j and k already appearing before - equality of j and k with either m and n is neither excluded nor required).

3.3. Path

This document specifies sub-path monitoring within a closed domain by a controlled and pre-designed measurement loop set-up. The path traversed by the packet SHOULD be reported, as detecting data plane forwarding in line with the desired measurement loop set-up is essential for the metric to enable and verify accurate evaluation. See [RFC8287] for SR MPLS OAM and [ID.draft-ietf-6man-spring-srv6-oam] for SRv6 OAM.

3.4. Sub-path Monitoring measurement loop packet spacing

Packets per measurement loop Fi are sent periodically by a temporal distance of IncT. For convenience, packets of the 6 measurement loops are assumed to be equally spaced at the sender too. Let's define the temporal distance IncF between two consecutive packets sent along to different measurement loops Fi and Fj at a single sender to be

IncF = IncT / 6

Further it seems useful to suggest IncF to be bigger than the largest measurement loop delay max (mi) under stable network operation (i.e., including some tolerance). Further assume the standard deviation of the measurement values mi to be much smaller than the delay mi, which is likely for a sub path being a regional or national link in many countries. Note that this definition isn't a strict requirement. Interpretation of results is however simplified by it. For the rest of the document assume

IncF > 2 * max (mi), i in [1...6], which results in

IncT > 12 * max (mi)

Discussion and reasoning for a reasonable smallest interval IncF in relation to max(mi) follows below.

4. Generic Type-P-SR-Path-Periodic-* metric

To reduce the redundant information presented in the detailed metrics sections that follow, this section presents the specifications that are common to two or more metrics. The section is organized using the same subsections as the individual metrics, to simplify comparisons.

4.1. Metric Name

All metrics use the Type-P convention as described in [RFC2330]. The rest of the name is unique to each metric.

4.2. Generic Metric Parameters

Refer to section 3.2. Metric Parameters: Type-P-* of [RFC6673]. The following parameters are added, enhanced or removed:

Dst SHOULD be a diagnostic IP address as specified by [RFC8287] and [RFC8029], if MPLS OAM is operated to capture the metric.

Fi, where i in [1, 2...6], a selection function defining unambiguously a packet of one particular stream i forming part of the monitoring overlay measurement loop set up.

L, a packet length in bits. The packets of all Type-P-SR-Path-Delay-Periodic-Streams Fi SHOULD all be of the same length.

MLAi, a stack of Segment IDs determining a monitoring loop Fi. The Segment-IDs MUST be chosen so that a singleton type-p packet of selection function Fi passes the sub-path i to be monitored.

No support: lambda (Poisson Streams remain ffs.)

4.3. Metric Units

Refer to section 3.4. Metric Units: Type-P-* of [RFC6673].

5. Singleton Definition for Type-P-SR-Path-Periodic-Delay

5.1. Metric Name

Type-P-SR-Path-Periodic-Delay

5.2. Metric Parameters

See section Section 4.2.

5.3. Delay Metric Units

A sequence of consecutive time values. The value of a Type-P-SR-Path-Periodic-Delay is either a real number or an undefined (informally, infinite) number of seconds per singleton of each stream Fi.

5.4. Definition

Section 3.4 of [RFC7679] applies per singleton of each stream Fi. The additional information related to singletons of section 4.2.4 of

[RFC3432] applies too.

5.5. Discussion

See section 3.5 of [RFC7679]. One generalisation seems appropriate: a global satellite navigation system affords one way to achieve synchronization within usec.

5.6. Methodologies

Section 3.6 of [RFC7679] applies per stream Fi with one exception: at the Src host, select Src and Dst IP addresses, if IP-routing is applied, or select the proper functional IP-destination address if an [RFC8287] SR MPLS OAM packet format is applied. Further add the appropriate stack of Segment IDs MLAi determining the monitoring loop Fi and form a test packet of Type-P with these addresses and the segment stack.

5.7. Errors and Uncertainties

See section 3.7 of [RFC7679] and section 4.6 of [RFC3432].

5.8. Reporting the metric

See section 3.8 of [RFC7679].

6. Singleton Definition for Type-P-SR-Path-Packet-Loss

Editors note: To be added based on existing loss metrics. A delay based approach indicating loss of a physical interface by detecting delay changes caused by re-routing can't be assumed to reliably cause unique delay change patterns under all circumstances (consider a shortest path routed multi-hop MPLS sub-path to be monitored rather than a link or a scenario where a bundle of 6 equivalent links is monitored connecting a single hub and spoke).

6.1. Metric Name

Type-P-SR-Path-Packet-Loss

6.2. Metric Parameters

See section Section 4.2.

6.3. Packet Loss Metric Units

The value of a Type-P-SR-Path-Packet-Loss is either a zero (signifying successful transmission of the packet) or a one (signifying loss) per singleton of each stream Fi.

6.4. Definition

Section 2.4 of [RFC7680] applies per singleton of each stream Fi.

6.5. Discussion

See section 3.5 of [RFC7680].

6.6. Methodologies

Section 2.6 of [RFC7680] applies per stream Fi with one exception: at the Src host, select Src and Dst IP addresses, if IP-routing is applied, or select the proper functional IP-destination address if an [RFC8287] SR MPLS OAM packet format is applied. Further add the appropriate stack of Segment IDs MLAi determining the monitoring loop Fi and form a test packet of Type-P with these addresses and the segment stack.

6.7. Errors and Uncertainties

See section 2.7 of [RFC7680].

6.8. Reporting the metric

See section 2.8 of [RFC7680].

7. Definition of Samples for Type-P-SR-Path-Periodic-Delay

This sections defines metric samples and metrics derived from samples.

7.1. Generic Type-P-SR-Path-Periodic-Delay-* metric

To reduce the redundant information presented in the detailed metrics sections that follow, this section presents the specifications that are common to two or more metrics. The section is organized using the same subsections as the individual metrics, to simplify comparisons.

7.1.1. Metric Name

Type-P-SR-Path-Periodic-Delay-*

7.1.2. Metric Parameters

Src, the IP address of a host

Dst, the IP address of a host

MLAi, a stack of Segment IDs

Ti0, a time

Tif, a time

incT, a time

7.1.3. Metric Units

See section Section 5.3.

7.1.4. Metric Defintion

Given Ti0 and Tif and nominal inter-packet interval incT, those time values greater than or equal to Ti0 and less than or equal to Tif are then selected. At each of the selected times in this process, we obtain one value of Type-P-SR-Path-Periodic-Delay. The value of the sample is the sequence made up of the resulting [time, delay] pairs. If there are no such pairs, the sequence is of length zero and the sample is said to be empty.

7.1.5. Discussion

See section 4.4 of [RFC3432].

7.1.6. Errors and uncertainties

See section 4.6 of [RFC3432].

7.2. Definition of Type-P-SR-Path-Periodic-Delay-Stream

The only definition required for this metric is a unique metric name.

7.2.1. Metric Name

Type-P-SR-Path-Periodic-Delay-Stream

7.3. Definition of Type-P-SR-Path-Periodic-Delay-Variation

The smallest sample Type-P-SR-Path-Periodic-Delay-Stream is one of two consecutively received values. These may be used to calculate a Segment Routed Path Delay-Variation (SRDV) singleton, defined below.

7.3.1. Metric Name

Type-P-SR-Path-Periodic-Delay-Variation

7.3.2. Methodologies

SRDV[i,j], for each sample of packets j and j-1 of stream Fi, j > 1, the delay variation between successive packets is calculated as:

SRDV[i,j] = Delay[i,j] - Delay [i,j-1],

j in [2,3...N] and N the total number of packets received at Dst. If one or more of the M packets sent by Src are lost, they are ignored for the metric, as no reasonable metric value is defined here. If N > 1, the metric is calculated for every valid packet received and the preceding one.

7.3.3. Discussion of SRDV

Evaluation statistics of differential SRDV metric samples may help to identify issues.

7.3.4. Errors and uncertainties

See section 2.7 of [RFC3393].

7.4. Definition of Type-P-SR-Path-Periodic-Delay-Variation-Stream

The only definition required for this metric is a unique metric name.

7.4.1. Metric Name

Type-P-SR-Path-Periodic-Delay-Variation-Stream

7.4.2. Metric Defintion

Given Ti0 and Tif, those time values greater than or equal to Ti0 and less than or equal to Tif are then selected. At each of the selected times in this process, we obtain one value of Type-P-SR-Path-Periodic-Delay. The value of the sample is the sequence made up of the resulting [time, delay-variation] pairs with time being set to the Dst timestamp of the Delay-Variation singleton, for which a valid

singleton is calculated. If there are no such pairs, the sequence is of length zero and the sample is said to be empty. If N Delay singletons are captured and sampled N-1 Delay-Variation singletons are sampled during the same interval

8. Statistic Definitions for SR-Path-Periodic-*-Stream samples

Change point detection requires statistical defintions. These are provided below. The names of the statistics contain an "*" placeholder, which may be replaced by "Delay" or "Delay-Variation".

8.1. SR-Path-Periodic-*-Mean

For a type-p metric, the mean is specified by:

SR-*Mean = (1/N) * Sum(from a=1 to N, value[a])

- N sample size

- value sample value of a sampled [time, value] pair

8.2. SR-Path-Periodic-*-Std

For a type-p metric, the Standard-Deviation Std is specified by:

SR-*Std = [1/(N-1)] * Sum(from a=1 to N, [SR-*Mean - value[a]]^2)

- N sample size

- value sample value of a sampled [time, value] pair

- SR-*Mean sample mean of the same metric as defined above

The definition as given above requires a two-pass calculation per sample. Algorithms estimating the standard-deviation by one-pass calculation have been published and might be preferable, if metric singletons and samples aren't buffered or calculations need to be fast.

9. Statistic Definitions for Type-P-SR-Path-Packet-Loss

The packet loss ratio is a useful metric to characterise congestion.

9.1. SR-Path-Packet-Loss-Ratio

See section 4.1 of [RFC7680]

10. Sub-Path monitoring metrics derived from samples captured along the measurement loops

To produce meaningful sub-path monitoring values, the measurement loop metrics are captured during a phase with stable networking conditions. In a backbone network domain, the absence of congestion often is a sufficient condition (frequent traffic shifts due to changes in routing and traffic engineering aren't expected). This may be different in a network based on a shared medium. It may be outright difficult in networks with frequently changing traffic management- and routing-policies.

In the following, the index CS indicates a statistic captured during a mesurement interval with stable routing and no congestion.

10.1. Baseline measurement

Capture a sample of delay values Type-P-SR-Path-Periodic-Delay-Stream of sample size N for each measurment loop Fi. As a rule of thumb choose N in [30, 100].

For each measurement loop Fi, calculate the following metrics characterising the monitored Sub-Paths during stable and congestion free network conditions:

* SR-Path-Delay-MeanCSi, the mean delay captured along measurement loop Fi

* SR-Path-Delay-StdCSi, the standard-deviation of the delay captured along measurement loop Fi

* SR-Path-Delay-Variation-MeanCSi, the mean delay variation captured along measurement loop Fi

* SR-Path-Delay-Variation-StdCSi, the standard-deviation of the delay variation captured along measurement loop Fi

A stable and uncongested network should produce rather constant delays, resulting in low standard-deviation values and almost zero mean delay variation. [Editors note: Add text to select the median of a small set of stream mean captures, like 5 samples captured consecutively.]

Example data was captured in a lightly loaded Gigabit network. 11 routers are passed per measurement loop. The sample size is 30 packets, more than 200 samples were captured per measurement loop. The loops are set up for a different purpose than specified here, they are picked due to a high number of passed routers. Note that SR-DV-Mean here refers to an abs(SR-DV-Mean) sample, thus small, positive, non-zero means result. The time unit is microseconds.

Metric	Quantile	SR-D-Mean	SR-D-Std	SR-DV-Mean	SR-DV-Std
Loop1	95%	34507	62	41	84

Loop2	95%	35104	45	34	49
Loop1	50%	34496	19	19	17
Loop2	50%	35088	15	14	12
Loop1	5%	34491	14	20	12
Loop2	5%	35080	13	12	9

Figure 2

Example baseline metrics for an 11 hop measurement loop (quantiles refer to SR-D-Mean)

10.2. Discussion of the baseline measurement

Delay outliers may occur at any time in any communication network, and the measurement system packet processing itself may also produce some. It is fair to expect only single outliers in a stable, not congested network. It may be worth to capture several consecutive SR-Path-Periodic-*-Stream samples and compare their statistics, before picking reasonable baseline metric values. Samples showing higher standard deviations (compare the 95% quantile values in the above figure to the 50% quantile values) may benefit from removing the maximum singleton value from the sample. This will smooth the mean and standard-deviation, and if the result then is closer to those of the majority of the samples, foster confidence in determining the baseline metrics. Depending on the preferred method of data-processing and storing, this may require capturing the sample maximum as a separate metric.

10.3. Definition of SR-Path-Sub-Path-RTD-Estimate

Within a single evaluation interval of identical Time T0 and Tf, SR-Path-Delay-MeanCSi(from now on DMeanCSi)is the mean delay of the measurement loop passing the monitored Sub-Path SPi by a round trip.

Let's keep the indexig applied above, then Fj and Fk with captured mean delays DMeanCSj and DMeanCSk pass SPi uniderictional. Further, 3 measurement loops Fx, Fy and Fz don't pass Sub-Path SPi at all. The corresponding mean delays are DMeanCSs, DMeanCSt and DMeanCSu.

The the SR-Path-Sub-Path-RTD-Estimate of the Round Trip Delay along the monitored Sub-Path Fi, RTD_Fi, is

RTD_Fi=(3*DMeanCSi+DMeanCSj+DMeanCSk-DMeanCSx-DMeanCSy-DMeanCSz)/4

10.4. Definition of SR-Path-Sub-Path-*-Changepoint

The asterisk stands for "Interface" as well as "Connectivity". If connectivity is lost and no path is available between two nodes, any packets to be transmitted will are dropped. A change in sub-path routes with a change in measurement loop delay indicitates a re-routimg event (a temporal loss in connectivity), not a long lasting loss of connectivity. Hence a change in measurement loop delays caused by a re-routed monitored sub isn't useful to derive a metric indicating connectivity loss on a monitored sub path (a sub-path-route-change metric might be of interest, but isn't within scope of this document).

Network changes like congestion or re-routing are often characterised by a change in the mean delay of a monitoring measurement. CUSUM (cumulative sum) charts have been shown to be efficient in detecting shifts in the mean of a process [NIST]. The upper bound CUSUM is defined as:

Sup(t)-Fi-Delay = max(0,Sup(t-1) + xt - SR-Path-*-MeanCSi - ki)

with Sup(0) = 0, ki = Delta * SR-Path-*-StdCSi (Delta is a dimensionless integer number), xt = Type-P-SR-Path-Periodic-* singleton for measurement loop Fi at time t.

The actual SR-Path-Delay-Mean of Measurement Loop Fi is decided to be significantly above SR-Path-*-MeanCSi, if:

Sup(t)-Fi-Delay > h_SP, with h_SP = d*ki (d is a dimensionless integer number).

An analogus CUSUM controls changes to a lower mean delay (which may be caused by a re-routing event):

Slo(t)-Fi-Delay = max(0,Slo(t-1) + SR-Path-*-MeanCSi - xj - k)

The actual SR-Path-Delay-Mean of Fi is decided to be significantly below SR-Path-*-MeanCSi, if:

Slo(t)-Fi-Delay > h_SP

10.5. Discussion of SR-Path-Sub-Path-*-Changepoint

CUSUM chart based changepoint detection is sensible even to small changes in the mean. CUSUM charts offer a limited protection against single, isolated outliers. A cumulated sum only grows, if the controled process consistenly changes its mean (or standard deviation, respectively). Assuming constant physical minimum delays to characterise wireline communication networks, a change in standard deviation not affecting the mean delay doesn't seem to be caused by a change in networking conditions.

The measured delays will change once a Sub-Path route has changed, or once persistent congestion starts to fill a queue. Both indicate changes in the network. As the Sub-Pathes SPi form an overlay with designed properties, every network change affecting a sub-path creates correlated SR-Path-* metric changes. As the correspondance of network changes to Sub-Path metrics is known a-priory, detecting correlated SR-Path-* metric changes allows to locate the change.

In the absence of packet re-routing, packet loss is characterising a loss of connectivity. Packet loss requires a time threshold when to decide that an active measurement packet was lost, and consecutive loss requires receiver awareness, that packets have been sent (this argues for the sender to be the receiver, unless both comminicate fast and reliable out of band).

The preferred CUSUM parametrisation will depend on the kind of events to detected and on the outlier characteristics.

ki = Delta * SR-Path-*-StdCSi may be set to a value relevant high enough to exclude single outliers to trigger an alert, but low enough to indicate persistent changes in delay. The same holds for the to be picked for d.

A broader discussion on CUSUM parametrisation may be found in literature. Networking skills are required to parametrise CUSUM, as well as to interprete the results (notably to differ re-routing from congestion).

10.6. Definition of SR-Path-Sub-Path-Congestion-Location

An interface along a single monitored Sub-Path SPi whose queue is persistently filled adds latency to measurement loop Fi and one of the two unidirectional measurement loops Fj and Fk passing Sub-Path SPi. Fj has been defined to pass SPi from Hub to Spoke and Fk pass SPI in opposite direction. Then SR-Path-Sub-Path-Congestion-Location metric for the traffic directed from "Hub to Spoke" along Sub-Path SPi is:

SPi_ConLoc_ij = Sup(t)_SPi_Periodic-Delay + Sup(t)_SPj_Periodic-Delay

And for the opposite traffic direction, from "Spoke to Hub":

SPi_ConLoc_ik = Sup(t)_SPi_Periodic-Delay + Sup(t)_SPk_Periodic-Delay

Note that another 10 SR-Path-Sub-Path-Congestion-Location metrics are calculated, one per monitored Sub Path and traffic direction. The evaluation can be simplified as follows:

```
IF SPi_ConLoc_ij > h_SP

AND h_SP > Sup(t)_SPk_Periodic-Delay

AND h_SP > Sup(t)_SPx_Periodic-Delay

AND h_SP > Sup(t)_SPy_Periodic-Delay

AND h_SP > Sup(t)_SPz_Periodic-Delay
```

Then Sub-Path SPi faces congestion in direction "Hub to Spoke".

```
IF SPi_ConLoc_ik > h_SP

AND h_SP > Sup(t)_SPj_Periodic-Delay

AND h_SP > Sup(t)_SPx_Periodic-Delay

AND h_SP > Sup(t)_SPy_Periodic-Delay

AND h_SP > Sup(t)_SPz_Periodic-Delay
```

Then Sub-Path SPi faces congestion in direction "Spoke to Hub".

Here, h_SP is a universal threshold in unit time to indicate a filling queue or a significant change in delay due to a Sub-Path reroute or another persistent change in topology (like e.g. automated Layer 1 / Layer 2 topology changes). Packets following SPx, SPy and SPz don't pass the congested interface of Sub-Path SPi.

10.7. Definition of SR-Path-Sub-Path-Disconnected

The idea of this document is to monitor a set of sub-paths for a single case of congestion or a single loss of connectivity. If a single sub-path SPi looses connectivity, i.e., all packets are dropped in both sub-path forwarding directions, then three measurement loops mi, mj and mk fail to receive any traffic. A single interface congestion will add latency to mi and one of mj or mk, respectively. Still, if it is congestion of a single sub-path SPi interface causing additional latency, either mj or mk face no congestion and the one measured delay mj or mk should be within the expected range of values. Rather than basing a loss of connectivity metric on a "reliable" indication SR-Path-Packet-Loss on each measurement loop mi, mj and mk by waiting for Tmax to receive any of the missed packets, this allows for a reaction independant of a conservative packet loss threshold like Tmax. The idea is to judge on disconnectivity if no packet is received on all three measurement loops mi, mj and mk after the time interval the last single packet was expected to be received, if there was no prior indication of congestion.

If the spacing of packets along consecutive measurement loops Fi is IncF as defined within section Section 3.4, then under stable network conditions every measurement packet sent along measurement loop Fi is received, before the next measurement packet is sent along measurement loop Fj. If a measurement interval starts at T1 and none of the three measurement loops Fi, Fj and Fk received a packet within T1 + incT = T1 + 6 * incF, monitored Sub-Path i is disconnected. It doesn't matter, along which of the three measurement loops the first not received packet was sent (there's no order here).

incF > max (SR-Path-Delay-MeanCSi+ d * Delta * SR-Path-Delay-StdCSi), i in [1...6]

With d and Delta being integer numbers as specified in section Section 10.4. If Fi and Fi+1 are measurement loops along which measurement packets are sent in consecutive order, this definition of incF ensures that the measurement packet sent along measurement loop Fi is received prior to sending the next measurement packet along measurement loop Fi+1 (under stable network conditions). The product d * Delta * SR-Path-Delay-StdCSi allows to set the preferred tolerance for outliers. It impacts the tradeoff between speed of

detection and false positive ratio. With this parameterisation, the metric indicationg a loss of bidirectional connectivity along Sub-Path i is defined as

either zero or one (or some logical equivalent), where LofCi=1 indicates loss of continuity along monitored Sub-Path Fi and LofCi=0 indicates successful arrival of at least one packet sent along measurement-loop Fi, Fj or Fk within incT.

Under conditions of section Section 3.4, if at any sliding interval incT no singleton was received along measurement-loops Fi, Fj and Fk, no more packets are forwarded in any direction of monitored sub-path SPi.

Faster detection of disconnectivity is likely possible by a different metric definition, which likely will depend on the measurement-loop delay Mi, Mj and Mk. The metric chosen above allows for a simple parametrisation. Metrics allowing for a faster determination of disconnection are not within scope of this document.

The sub-path SPi is judged to be disconnected from the earliest time, when a packet was sent but not received on any of the three sub-paths Fi, Fj or Fk. The sub-path SPi is judged to be connected, whenever a measurement packet sent along one or more of the measurement-loops Fi, Fj and Fk is received again.

```
             Fi = send time of a packet along measurement-loop Fi
                  i in [1...6]
             Mi = receive time of a packet sent along Fi
             incT interval between two packets sent along Fi
             incF > max (Mi)

               IncF                       IncT = 6 * IncF
          __/\__         ___________________/\__________________
         /      \       /                                       \
         +------+------+------+------+------+------+------+------+
         t=0    1   |  2      3      4  |   5      6   |  7   |  8
                F1  |  F2     F3     F4 |   F5     F6  |  F1  |  F2
                    M1                  M4             M6     M1 |
                                                                     |
         At time 8, next packet should be sent along F2.         |
         No packets were received along F2, F3 and F5 yet.       |
             Indicates discontinuity along SP3 at time 8.  <------+
```

Figure 3

Illustration of the sub-path disconnectivity metric; sub-path SP3 is link L100 <-> L070 of the example network Figure 1.

Note, if F2 sent at time 2 was received at time 2 + M2, but no more packet passing SP3 afterwards, discontinuity of SP3 is indicated at time 9, when F3 is to send the next packet. Also note that discontinuity of SP3 could be indicated as early as time 6 in the example. That requires a different metric. Basing the metric definition on incT however covers all potential intervals between relevant Fi, Fj and Fk.

11. Discussion of Temporal Resolution

A loss of connectivity is detected after a temporal distance of IncT, the time period between two packets beeing sent along the same measurement-loop Fi. IncT is specified as 6*IncF, where IncF is 2 times the largest measurement-loop delay in the absence of congestion. Hence a loss of connectivity is indicated after 12 * the largest measurement-loop delay.

Reliable indications of lost connectivity may be possible also at smaller timescales. The specification chosen seems to be simple as well as reliable and thus defines a starting point for advanced designs offering faster reaction.

12. IANA Considerations

If standardised, the metric will require an entry in the IPPM metric registry.

13. Security Considerations

This draft specifies how to use methods specified or described within [RFC8402] and [RFC8403]. It does not introduce new or additional SR features. The security considerations of both references apply here too.

14. References

14.1. Normative References

[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>.

[RFC2678] Mahdavi, J. and V. Paxson, "IPPM Metrics for Measuring Connectivity", RFC 2678, DOI 10.17487/RFC2678, September 1999, <https://www.rfc-editor.org/info/rfc2678>.

[RFC3393] Demichelis, C. and P. Chimento, "IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)", RFC 3393, DOI 10.17487/RFC3393, November 2002, <https://www.rfc-editor.org/info/rfc3393>.

[RFC3432] Raisanen, V., Grotefeld, G., and A. Morton, "Network performance measurement with periodic streams", RFC 3432, DOI 10.17487/RFC3432, November 2002, <https://www.rfc-editor.org/info/rfc3432>.

[RFC6673] Morton, A., "Round-Trip Packet Loss Metrics", RFC 6673, DOI 10.17487/RFC6673, August 2012, <https://www.rfc-editor.org/info/rfc6673>.

[RFC7679] Almes, G., Kalidindi, S., Zekauskas, M., and A. Morton, Ed., "A One-Way Delay Metric for IP Performance Metrics (IPPM)", STD 81, RFC 7679, DOI 10.17487/RFC7679, January 2016, <https://www.rfc-editor.org/info/rfc7679>.

[RFC7680] Almes, G., Kalidindi, S., Zekauskas, M., and A. Morton, Ed., "A One-Way Loss Metric for IP Performance Metrics (IPPM)", STD 82, RFC 7680, DOI 10.17487/RFC7680, January 2016, <https://www.rfc-editor.org/info/rfc7680>.

[RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <https://www.rfc-editor.org/info/rfc8029>.

[RFC8287] Kumar, N., Ed., Pignataro, C., Ed., Swallow, G., Akiya, N., Kini, S., and M. Chen, "Label Switched Path (LSP) Ping/Traceroute for Segment Routing (SR) IGP-Prefix and IGP-Adjacency Segment Identifiers (SIDs) with MPLS Data Planes", RFC 8287, DOI 10.17487/RFC8287, December 2017, <https://www.rfc-editor.org/info/rfc8287>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8667] Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", RFC 8667, DOI 10.17487/RFC8667, December 2019, <https://www.rfc-editor.org/info/rfc8667>.

14.2. Informative References

[CommodityTomography]
Lakhina, A., Papagiannaki, K., Crovella, M., Diot, C., Kolaczyk, ED., and N. Taft, "Structural analysis of network traffic flows", 2004, <https://www.cc.gatech.edu/classes/AY2007/cs7260_spring/papers/odflows-sigm04.pdf>.

[ID.draft-ietf-6man-spring-srv6-oam]
Zafar, A., Filsfils, C., Matsushima, S., Voyer, D., and M. Chen, "Operations, Administration, and Maintenance (OAM) in Segment Routing Networks with IPv6 Data plane (SRv6)", 2021.

[NIST] NIST, "NIST/SEMATECH e-Handbook of Statistical Methods, section CUSUM Control Charts", 2021, <http://www.itl.nist.gov/div898/handbook/>.

[RFC2330] Paxson, V., Almes, G., Mahdavi, J., and M. Mathis, "Framework for IP Performance Metrics", RFC 2330, DOI 10.17487/RFC2330, May 1998, <https://www.rfc-editor.org/info/rfc2330>.

[RFC8403] Geib, R., Ed., Filsfils, C., Pignataro, C., Ed., and N. Kumar, "A Scalable and Topology-Aware MPLS Data-Plane Monitoring System", RFC 8403, DOI 10.17487/RFC8403, July 2018, <https://www.rfc-editor.org/info/rfc8403>.

Author's Address

Ruediger Geib (editor)
Deutsche Telekom
Heinrich Hertz Str. 3-7
64295 Darmstadt
Germany
Phone: +49 6151 5812747
Email: Ruediger.Geib@telekom.de