ippm
Internet-Draft
Intended status: Standards Track
Expires: June 26, 2021

R. Geib, Ed.
Deutsche Telekom

December 23, 2020

A Connectivity Monitoring Metric for IPPM
draft-ietf-ippm-connectivity-monitoring-00

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 will expire on June 26, 2021.

Copyright Notice

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 or 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 processor- 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 change in network behaviour may then require 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 link-loss, if the change in end-to-end delay along a new route is differing from that of the original path.

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 or a single loss of connectivity of a monitored sub-path cause a delay change on several measurement paths. Any single evnet of that type on one of the monitored sub-paths changes delays of a unique subset of measurement loops. The number of measurement loops 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. 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 delay measurements.

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 additional processing apart from that.

It is recommended to consider automated measurement loop set-ups. 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 respected are given.

This document specifies a type-p metric determining properties of an SR path which allows to monitor connectivity and congestion of interfaces and further allows to locate the path or interface which caused a change in the reported type-p metric. This document is limited on the 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, [RFC8029] 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 [RFC8029] methods as a replacement of Adj-SIDs is out of scope of this document.

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 also 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 here [CommodityTomography]. The approach pursued here uses a pre-specified measurement loop overlay design.

A centralised monitoring approach doesn't require report collection and result correlation from two (or more) receivers (the measured delays of different measurement loops still need to be correlated).

An additional property of the measurement path set-up specified below is that it allows to estimate the packet round trip and the one way delay of a monitored sub-path. The delay along a single link is not perfectly symmetric. Packet processing causes small delay differences per interface and direction. These cause an error, which can't be quantified or removed by the specified method. Quantifying this error requires a different measurement set-up. As this will introduce additional measurements loops, packets and evaluations, the cost in terms of reduced scalability is not felt to be worth the benefit in measurement accuracy. IPPM metrics prefer precision to accuracy and the mentioned processing differences are relatively stable, resulting in relatively precise delay estimates for each monitored 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 6 links (a very generic kind of sub-path) attaching the spoke-nodes L050, L060 and L070 to the hub-nodes L100 and L200.

```
      +---+   +----+     +----+
      |PMS|   |L100|-----|L050|
      +---+   +----+\   /+----+
        |    /    \  \_/_____
        |   /      \  /      \+----+
     +----+/        \/_  +----|L060|
     |L300|         /  |/     +----+
     +----+\       /   /\_
            \     /   /   \
             \+----+ /   +----+
              |L200|-----|L070|
              +----+     +----+
```

Hub and spoke connectivity verification with a PMS

Figure 1

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:

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

o 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., between L100 and L060 and then from L060 to L200)

o Every monitored link is passed by a single round trip measurement loop only once and further only once unidirectional by two other loops. These unidirectional mearurement 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., one measurement loop having a round trip L100 to L050 and back (M1, see below), a second loop passing L100 to L050 only (M3) and a third loop passing L050 to L100 only (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

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

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.

The Evaluation of the measurement loop Round Trip Delays M1 - M6 allow 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 (and changes the measured delay) of three loops. The traffic to the Node-SIDs is rerouted (in the case of a single links loss, no node is completely disconnected in the example network).

- If the loops are set up using Adj-SIDs only, any single complete 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.

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

- As an example, the formula for a single link (sub-path) Round Trip Delay (RTD) is shown 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 link of interest only in a single direction. From this sum subtract the RTD measured on all loops not passing the monitored link of interest to get four times the RTD of the monitored link of interest.

A closer look reveals that each single event of interest for the proposed metric, which are a loss of connectivity or a case of congestion, uniquely only impacts a single set of measurement loops which can be determined a-priori. If, e.g., connectivity is lost between L200 and L050, measurement loops (3), (4) and (6) indicate a change in the measured delay.

As a second example: if the interface L070 to L100 is congested, measurement loops (3) and (5) indicate a change in the measured delay. Without listing all events, all cases of single losses of connectivity or single events of congestion influence only delay measurements of a unique set of measurement loops.

Assume that the measurement loops are set up while there's no congestion. In that case, the congestion free RTDs of all monitored links can be calculated as shown above. A single congestion event adds queuing delay to the RTD measured by two specific measurement loops. The two measurement loops impacted allow to distinct the congested interface and calculation of the queue-depth in terms of seconds. As an example, assume a queue of an average depth of 20 ms to build up at interface L200 to L070 after the uncongested measurement interval T0. The measurement loops M5 and M6 are the only ones passing the interface in that direction. Both indicate a congestion M5 and M6 of + 20 ms during measurement interval T1, while M1-4 indicate no change. The location of the congested interface is determined by the combination of the two (and only two) measurement loops M5 and M6 showing an increased delay. The average queue depth = (M5[T1] - M5[T0] + M5[T1] - M5[T0])/2.

As mentioned there's a constant delay added for each measurement loop, which is the delay of the path transversed 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 measurements 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. Network topology requirements

The metric and methods specified below can be applied in networks with a hub and spoke topology. A single network change of type loss of connectivity or congestion can be detected. The nodes don't have to be hubs or spokes, this is just a topology requirement. In detail, the topology 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.

4. Singleton Definition for Type-P-SR-Path-Connectivity-and-Congestion

4.1. Metric Name

Type-P-SR-SubPath-Connectivity

4.2. Metric Parameters

- Src, the IP address of a source host
- Dst, the IP address of a destination host if IP routing is applicable; in the case of MPLS routing, a diagnostic address as specified by [RFC8029]
- T, a time
- L, a packet length in bits. The packets of a Type P packet stream from which the sample Path-Connectivity-and-Congestion metric is taken MUST all be of the same length.
- MLA, a stack of Segment IDs determining a Monitoring Loop. The Segment-IDs MUST be chosen so that a singleton type-p packet passes one single monitored sub-path_a bidirectional, one monitored sub-path_b unidirectional and one monitored sub-path_c unidirectional, where sub-path_a, -_b and -_c MUST NOT be identical and MUST NOT share properties to be monitored.
- P, the specification of the packet type, over and above the source and destination addresses
- DS, a constant time interval between two type-P packets in unit seconds

4.3. Metric Units

A sequence of consecutive time values.

4.4. Definition

A moving average of AV time values per measurement path is compared by a change point detection algorithm. The temporal packet spacing value DS represents the smallest period within which a change in connectivity or congestion may be detected.

A single loss of connectivity of a sub-path between two nodes affects three different measurement paths. Depending on the value chosen for DS, packet loss might occur (note that the moving average evaluation needs to span a longer period than convergence time; alternatively, packet-loss visible along the three measurement paths may serve as an evaluation criterium). After routing convergence the type-p packets along the three measurement paths show a change in delay.

A congestion of a single interface of a sub-path connecting two nodes affects two different measurement paths. The the type-p packets along the two congested measurement paths show an additional change in delay.

4.5. Discussion

Detection of a multiple losses of monitored sub-path connectivity or congestion of a multiple monitored sub-paths may be possible. These cases have not been investigated, but may occur in the case of Shared Risk Link Groups. Monitoring Shared Risk LinkGroups and sub-paths with multiple failures abd congestion is not within scope of this document.

4.6. Methodologies

For the given type-p, the methodology is as follows:

- The set of measurement paths MUST be routed in a way that each single loss of connectivity and each case of single interface congestion of one of the sub-paths passed by a type-p packet creates a unique pattern of type-p packets belonging to a subset of all configured measurement paths indicate a change in the measured delay. As a minimum, each sub-path to be monitored MUST be passed

-
  - by one measurement_path_1 and its type-p packet in bidirectional direction
  - by one measurement_path_2 and its type-p packet in "downlink" direction
  - by one measurement_path_3 and its type-p packet in "uplink" direction
- "Uplink" and "Downlink" have no architectural relevance. The terms are chosen to express, that the packets of measurement_path_2 and measuremnt_path_3 pass the monitored sub-path unidirectional in opposing direction. Measuremnt_path_1, measurement_path_2 and measurement_path_3 MUST NOT be identical.

- All measurement paths SHOULD terminate between identical sender and receiver interfaces. It is recommended to connect the sender and receiver as closely to the paths to be monitored as possible. Each intermediate sub-path between sender and receiver one one hand and sub-paths to be monitored is an additional source of errors requiring separate monitoring.
- Segment Routed domains supporting Node- and Adj-SIDs should enable the monitoring path set-up as specified. Other routing protocols may be used as well, but the monitoring path set up might be complex or impossible.
- Pre-compute how the two and three measurement path delay changes correlate to sub-path connectivity and congestion patterns. Absolute change valaues aren't required, a simultaneous change of two or three particular measurement paths is.
- Ensure that the temporal resolution of the measurement clock allows to reliably capture a unique delay value for each configured measurement path while sub-path connectivity is complete and no congestion is present.

- Synchronised clocks are not strictly required, as the metric is evaluating differences in delay. Changes in clock synchronisation SHOULD NOT be close to the time interval within which changes in connectivity or congestion should be monitored.

- At the Src host, select Src and Dst IP addresses, and address information to route the type-p packet along one of the configured measurement path. Form a test packet of Type-P with these addresses.

- Configure the Dst host access to receive the packet.

- At the Src host, place a timestamp, a sequence number and a unique identifier of the measurement path in the prepared Type-P packet, and send it towards Dst.

- Capture the one-way delay and determine packet-loss by the metrics specified by [RFC7679] and [RFC7680] respectively and store the result for the path.

- If two or three subpaths indicate a change in delay, report a change in connectivity or congestion status as pre-computed above.

- If two or three sub paths indicate a change in delay, report a change in connectivity or congestion status as pre-computed above.

Note that monitoring 6 sub paths requires setting up 6 monitoring paths as shown in the figure above.

4.7. Errors and Uncertainties

Sources of error are:

- Measurement paths whose delays don't indicate a change after sub-path connectivity changed.

- A timestamps whose resolution is missing or inacurrate at the delays measured for the different monitoring paths.

- Multiple occurrences of sub path connectivity and congestion.

- o Loss of connectivity and congestion along sub-paths connecting the measurement device(s) with the sub-paths to be monitored.

4.8. Reporting the Metric

The metric reports loss of connectivity of monitored sub-path or congestion of an interface and identifies the sub-path and the direction of traffic in the case of congestion.

The temporal resolution of the detected events depends on the spacing interval of packets transmitted per measurement path. An identical sending interval is chosen for every measurement path. As a rule of thumb, an event is reliably detected if a sample consists of at least 5 probes indicating the same underlying change in behavior. Depending on the underlying event either two or three measurement paths are impacted. At least two consecutively received measurement packets per measurement path should suffice to indicate a change. The values chosen for an operational network will have to reflect scalability constraints of a PMS measurement interface. As an example, a PMS may work reliable if no more than one measurement packet is transmitted per millisecond. Further, measurement is configured so that the measurement packets return to the sender interface. Assume always groups of 6 links to be monitored as described above by 6 measurements paths. If one packet is sent per measurement path within 500 ms, up to 498 links can be monitored with a reliable temporal resolution of roughly one second per detected event.

Note that per group measurement packet spacing, measurement loop delay difference and latency caused by congestion impact the reporting interval. If each measurement path of a single 6 link monitoring group is addressed in consecutive milliseconds (within the 500 ms interval) and the sum of maximum physical delay of the per group measurement paths and latency possibly added by congestion is below 490 ms, the one second reports reliably capture 4 packets of two different measurement paths, if two measurement paths are congested, or 6 packets of three different measurement paths, if a link is lost.

A variety of reporting options exist, if scalability issues and network properties are respected.

5. Singleton Definition for Type-P-SR-Path-Round-Trip-Delay-Estimate

This section will be added in a later version, if there's interest in picking up this work.

6. IANA Considerations

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

7. 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.

8. References

8.1. 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>.

[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>.

[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>.

8.2. 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>.

[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
Darmstadt 64295
Germany

Phone: +49 6151 5812747
Email: Ruediger.Geib@telekom.de