SR MPLS – Performance Monitoring

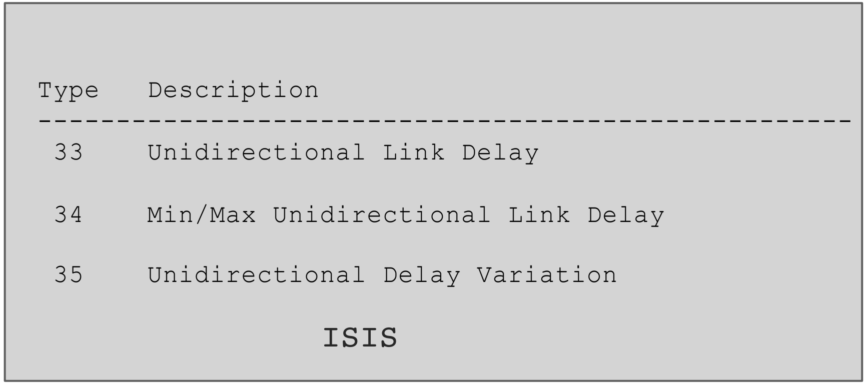
Clarence Filsfils
Cisco Fellow – cf@cisco.com

Disclaimer

“Many of the products and features described herein remain in varying stages of development and will be offered on a when-and-if-available basis. This roadmap is subject to change at the sole discretion of Cisco, and Cisco will have no liability for delay in the delivery or failure to deliver any of the products or features set forth in this document.”

Per-Link delay Measurement

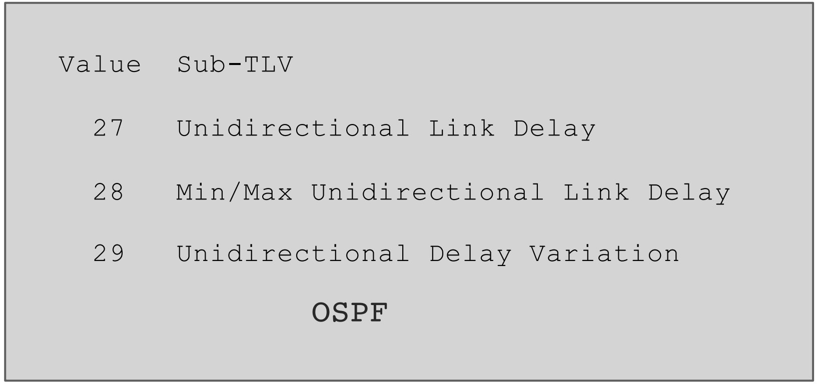
ISIS Signaling



```
Type   Description
----------------------------------------------------
 33    Unidirectional Link Delay

 34    Min/Max Unidirectional Link Delay

 35    Unidirectional Delay Variation

                 ISIS
```

- RFC 7810 (IS-IS Traffic Engineering (TE) Metric Extensions)
- Used to advertise extended TE metrics – e.g. link delay (in usec)

OSPF and BGP-LS



Value	Sub-TLV
27	Unidirectional Link Delay
28	Min/Max Unidirectional Link Delay
29	Unidirectional Delay Variation

OSPF

- RFC 7471 (OSPF Traffic Engineering (TE) Metric Extensions)
- Used to advertise extended TE metrics – e.g. link delay (in usec)
- BGP-LS: draft-ietf-idr-te-pm-bgp

Leveraged by SRTE – SR Policy

- SR Policy for min delay

```
segment-routing
  traffic-eng
    policy FOO
      color 20 end-point ipv4 1.1.1.3
      binding-sid mpls 1000
      candidate-paths
        preference 100
          dynamic mpls
            metric
              type delay
```

Leveraged by SRTE – IGP Flex Algo

- IGP SR Flex Algo for minimum delay

```
router isis 1
  flex-algo 128
    metric-type delay
```

Per-link delay Measurement

- Over a measurement internal
  - minimum → Used as metric for SRTE (Policy or Flex-Algo)
  - average
  - maximum
  - variance

  average, maximum, variance → Not used by SRTE
- One-way or Two-way
  - one-way requires clock synchronization

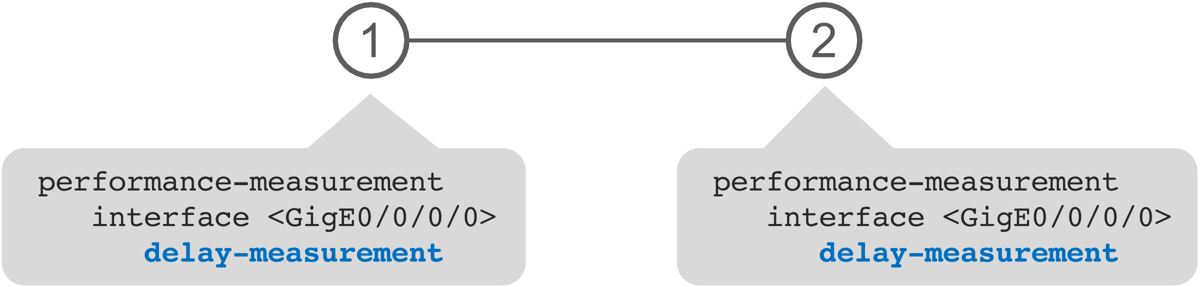
Minimum delay is of interest for SRTE

- Minimum delay provides the propagation delay
  - fiber length / speed of light
- A property of the topology
  - with awareness of DWDM circuit change
- SRTE (Policy or Flex-Algo) can optimize on min delay

Average, Max and Variance are dealt with by QoS

- Depends on congestion
  - (traffic burst over line rate) / line rate
- Highly variable at any time scale
- Not controlled by routing optimization
- Controller by QoS
  - Priority queue, WRR, WFQ…
  - Tail-Drop, RED…

Link Delay - Configuration

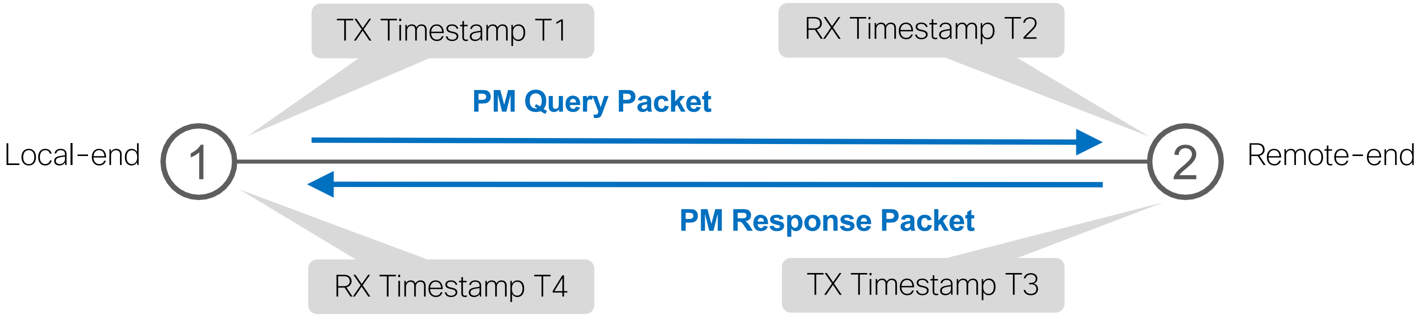


```
performance-measurement
   interface <GigE0/0/0/0>
      delay-measurement
```

```
performance-measurement
   interface <GigE0/0/0/0>
      delay-measurement
```

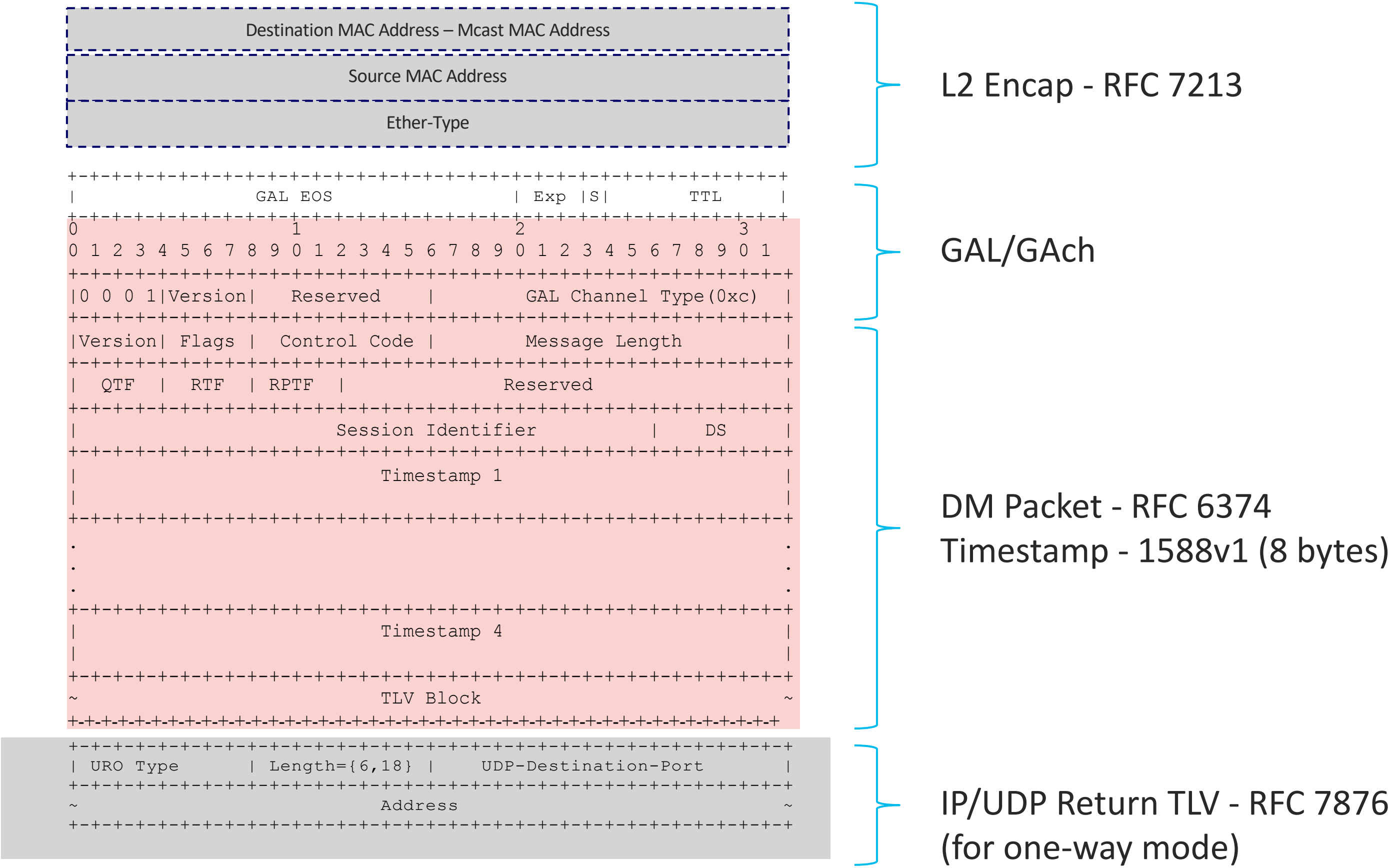
- If the link is enabled for an IGP, then this IGP automatically includes the delay TLV in its LSP/LSA

Link Delay – Probe Measurement



- One Way Delay = (T2 – T1)
- Two-Way Delay = (T2 – T1) + (T4 – T3)

Query Packet using RFC 6374 Packet Format



```
Destination MAC Address – Mcast MAC Address
Source MAC Address
Ether-Type

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               GAL EOS                 | Exp |S|      TTL      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0 0 0 1|Version|   Reserved    |         GAL Channel Type(0xc) |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Flags |  Control Code |         Message Length        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  QTF  |  RTF  | RPTF  |               Reserved                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Session Identifier         |      DS       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Timestamp 1                          |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                                                               .
.                                                               .
.                                                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Timestamp 4                          |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                          TLV Block                            ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| URO Type      | Length={6,18} |    UDP-Destination-Port       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                          Address                              ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

L2 Encap - RFC 7213

GAL/GAch

DM Packet - RFC 6374
Timestamp - 1588v1 (8 bytes)

IP/UDP Return TLV - RFC 7876
(for one-way mode)

Default

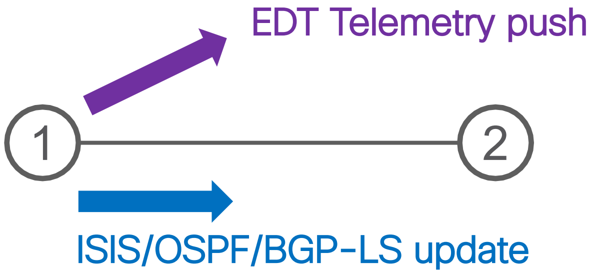
- Every 3 second, a query
  - a two-way query is sent
- Every 30 seconds, a probe
  - min, avg, max, var are computed over the last 10 queries
  - Last-Probe EDT trigger with (min, avg, max, var)
- Every 120 seconds, an aggregation
  - min, avg, max, var over the last 4 probes are computed
  - Last-Aggregation EDT trigger with (min, avg, max, var)
  - IF [abs(min-F.min)/F.min >= 10%] and [abs(min-F.min)>=1000usec]
    THEN an LSDB change is triggered to flood the new link delay values
    a last-advertisement EDT is triggered with these values

F.min is the last flooded value of min-delay. This is what the rest of the network thinks of this link min delay.

Routing stability – Telemetry accuracy

Every 30sec

Every 120sec
IF significant min change
THEN trigger an ISIS/OSPF flood



EDT Telemetry push

ISIS/OSPF/BGP-LS update

- SRTE optimization only needs minimum delay
  - IGP to only flood/update if the meaningful parameter changes (min)
- Use telemetry to collect the evolution of other delay components at finer time scale (min, max, avg at probe period)

Default

- Automated discovery of the per-link propagation delay
- Automated signaling in ISIS/OSPF/BGP-LS
- Automated churn protection
  - a change is advertised if > 10% and > 1000usec (200km of fiber)
- Automated detection of optical path change
  - Worst-case 240sec for a degradation
  - 60sec if accelerated mode enabled

Customization

- Ample ability to customize the measurement behavior

If we had more time

- Bundle
- Per SR Policy delay measurement
  - ECMP support
- Per SR Policy loss measurement
- Per-Link Loss measurement

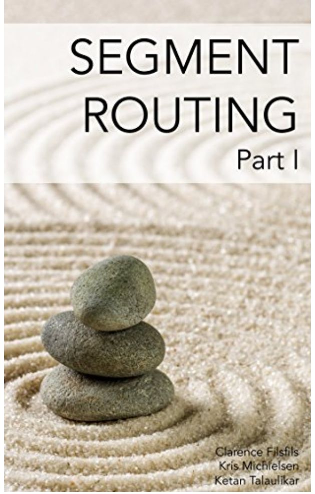
Conclusion

Conclusion

- SRTE integrated framework for SLA delivery
- Per-link delay measurement
  - Automation
  - Simplicity
  - Scale
  - Functionality

Stay uptodate

amzn.com/B01I58LSUO



segment-routing.net

linkedin.com/groups/8266623

twitter.com/SegmentRouting

facebook.com/SegmentRouting/

Contributors

- Rakesh Gandhi
- Sagar Soni
- Patrick Khordoc
- Kris Michielsen

Appendix

Customization- Burst

```
performance-measurement
        delay-profile interfaces
            probe
                interval < 30-3600 SEC >
                burst
                    count < 1-30 COUNT >
                    interval < 30-15000 msec >
```

- Probe interval
  - By default, a probe packet is sent every 30 seconds.
- Burst
  - By default, burst is enabled with 10 packets sent per probe.
  - Fastest burst interval is 30 msec
  - Default burst interval is 3000 msec (when burst-count is > 1).
  - Burst count x burst interval cannot be > probe interval

Customization- One-way

```
performance-measurement
  delay-profile interfaces
     probe
       one-way
```

- Default: Two-way
    - By default, two-way delay measurement is enabled. All four time-stamps (T1-T4) are used defined in the RFC 6374 packet format.
    - Querier requests for in-band PM replies.
    - Probes and replies, both are sent as RFC 6374 MPLS GAL packets.
    - One-way delay is computed as two-way delay divided by 2.
    - Hardware clock synchronization not required between querier and responder nodes.
- One-way
    - When one-way delay is enabled, IP/UDP TLV (defined in RFC 7876) is added in the query packet, to receive PM reply via IP/UDP.
    - Only two time-stamps (T1 and T2) are used in the RFC 6374 packets.
    - Hardware clocks must be synchronized between querier and responder nodes (using PTP).

Customization– Periodic Advertisement

```
performance-measurement
     delay-profile interfaces
          advertisement
                periodic
                    disabled                  (default enabled)
                    interval < 30-3600 SEC >  (default 120)
                    threshold < 0-100% >      (default 10%)
                    minimum < 0-100000 usec > (default 1000 usec ~ 200km optical fiber)
```

- Periodic advertisement is enabled by default. It can be disabled by adding disabled config.
- At the end of the periodic interval, if the change in a measured value (min/max/average/variance) compared to the last advertised value is,
  above the periodic threshold (%), AND above the minimum (VALUE)
- then, all delay values (average/min/max/variance) are advertised for that link.
- Advertisement interval is rounded up to the next multiple of probe interval internally to avoid advertisement in the middle of a probe (e.g. advertisement interval of 45 with probe interval 30 will round up to 60 (2*30)).
- Advertisement interval less than the probe interval is rounded up to the same value as the probe interval.

Customization - Accelerated Advertisement

```
config# performance-measurement

        delay-profile interfaces
            advertisement
                accelerated                         (default disabled)
           threshold < 0-100% >                  (default 20%)
                    minimum < 0-100000 usec >       (default 1000 usec)
```

- Accelerated advertisement is disabled by default.
- When accelerated advertisement is enabled,
- if the change in the measured minimum link metric compared to the last advertised minimum link metric is, above the accelerated threshold (%), AND, above the minimum (VALUE)
- then, all delay values (average/min/max/variance) are advertised for that link.
- Accelerated advertisements will occur at least one probe interval apart.

Customization – Telemetry only

```
performance-measurement
    delay-profile interfaces
        advertisement
            periodic
               disabled
```

- Used for monitoring the link delay metrics with streaming telemetry without flooding them in the network
- This is achieved by adding disabled configuration under periodic advertisement
- The link delay metrics will not be flooded in the network by the IGPs or advertised by the BGP-LS

Show CLIs

- **Querier side show CLIs**

```
# show performance-measurement summary [ detail ]                                     [ location <> ]
# show performance-measurement interfaces  [ <name> ]    [ detail ]            [ location <> ]
# show performance-measurement history probe interfaces  [ <name> ]          [ location <> ]
# show performance-measurement history aggregated interfaces  [ <name> ]  [ location <> ]
# show performance-measurement sessions [ <session-id> ] [ detail]              [ location <> ]
# show performance-measurement counters   [ interface <name> ]                 [ location <> ]
```

- **Responder side show CLIs**

```
# show performance-measurement responder summary                         [ location <> ]
# show performance-measurement responder interface    [ <name> ]         [ location <> ]
# show performance-measurement responder sessions [ <session-id> ]        [ location <> ]
# show performance-measurement responder counters [ interface <name> ]  [ location <> ]
```

Show performance-measurement summary

```
# show performance-measurement summary  [detail]  [ location <> ]
-------------------------------------------------------------------------------
0/0/CPU0
-------------------------------------------------------------------------------
Delay-Measurement:
  Profile configuration:
    Probe interval                              : 30 seconds
    Burst interval                               : 3000 mSec
    Burst count                                  : 10 packets
    Periodic advertisement               : Enabled
      Interval                                        : 120 (effective: 120) sec
      Threshold                                 : 10%
      Minimum                                  : 1000 uSec
    Advertisement accelerated        : Disabled
  Counters:
    Total interfaces                          : 2
    Total sessions                            : 2
    Packets:
      Total sent                                      : 855220
      Total received                          : 855220
      Total sent errors                      : 0
      Total received errors               : 0
    Probes:
      Total started                            : 85522
      Total completed                      : 85522
      Total incomplete                     : 0
    Total advertisements                : 63
```

- By default, counters from all LCs and active RP are returned when location is not specified.
- Total counters are per location (RP or LC).

Show performance-measurement interfaces

```
# show performance-measurement interfaces [ <name> ] [ location <> ]

--------------------------------------------------------------------------------------
0/0/CPU0
--------------------------------------------------------------------------------------
Interface Name: Bundle-Ether1 (ifh: 0x1000060)
  Delay-Measurement      : Enabled
  Local IPV4 Address     : 15.15.15.2
  Local IPV6 Address     : 15:15:15::2
  Local MAC Address      : 02f1.175b.a9ec
  Primary VLAN Tag       : None
  Secondary VLAN Tag     : None
  State                  : Up

  Delay Measurement session:
    Session ID           : 1

    Last advertisement:
        Advertised at: 11:40:45 Wed 12 Apr 2017 (1890 seconds ago)
        Advertised reason: periodic timer | accelerated threshold crossed
        Advertised delays (uSec): avg: 5456, min: 5200, max: 5601, variance: 1234

    Current advertisement:
        Scheduled in 1 more probe (roughly every 120 seconds)
        Current delays (uSec): avg: 5345, min: 5190, max: 5543, variance: 1230
```

Show performance-measurement interfaces detail

```
# show performance-measurement interfaces [ <name> ] detail [ location <> ]

--------------------------------------------------------------------------------------
0/0/CPU0
--------------------------------------------------------------------------------------
Interface Name: Bundle-Ether1 (ifh: 0x1000060)

  <snip>

  Delay-Measurement:
    Session ID          : 1

    <snip>

    Current Probe:
       Started at: 11:40:45 Wed 12 Apr 2017 (10 seconds ago, runs every 30 seconds)
       Packets sent: 4, received: 4
       Measured delays (uSec): avg: 5711, min: 5497, max: 5927, variance: 1230

       Probe samples:
           Packet Tx Timestamp                     Measured delay (nSec)
           11:40:45.100 Wed 12 Apr 2017            5954010
           11:40:48.200 Wed 12 Apr 2017            5786011
           11:40:45.300 Wed 12 Apr 2017            5669230
           11:40:45.300 Wed 12 Apr 2017            5702000

       Next probe scheduled at 11:41:15 Wed 12 Apr 2017 (in 20 seconds)
       Next burst packet scheduled for send in 72 uSec | burst completed
```

Show performance-measurement history interfaces

```
# show performance-measurement history probe interfaces [ <name> ]
--------------------------------------------------------------------------------------
0/0/CPU0
--------------------------------------------------------------------------------------
Interface Name: Bundle-Ether1 (ifh: 0x1000060)
  Delay-Measurement history (uSec):
    Probe Start Timestamp      Pkt(TX/RX)  Average        Min        Max
    11:40:45 Wed 12 Apr 2017          4/4     5711       5497       5927
    11:41:15 Wed 12 Apr 2017          4/4     5594       5219       5871
    11:41:45 Wed 12 Apr 2017          4/4     5541       5149       5796
    11:42:15 Wed 12 Apr 2017          4/4     5621       5379       5921
    11:42:45 Wed 12 Apr 2017          4/4     5564       5034       5987
    11:43:15 Wed 12 Apr 2017          4/4     5643       5432       5936
    11:43:45 Wed 12 Apr 2017          4/4     5350       5029       5858
    11:44:15 Wed 12 Apr 2017          4/4     5616       5404       5928
    11:44:45 Wed 12 Apr 2017          4/4     5581       5128       5904
    11:45:15 Wed 12 Apr 2017          4/4     5482       5183       5772
```

Show performance-measurement counters

```
 # show performance-measurement counters interfaces [ <name> ] [ location <> ]

--------------------------------------------------------------------------------------
0/0/CPU0
--------------------------------------------------------------------------------------
Interface: Bundle-Ether1
 Delay-Measurement:
    Advertisements               : 8101
    Probes Started                : 85563
    Probes Complete            : 85563
    Probes Incomplete          : 0
    Query packets sent         : 427815
    Reply packets received  : 427815
    Query packets errored    : 0
    Reply packets errored    : 0
```

Show performance-measurement responder summary

```
 # show performance-measurement responder summary [ location <> ]

 --------------------------------------------------------------------------------------
 0/0/CPU0
 --------------------------------------------------------------------------------------
Delay-Measurement:
  Total interfaces                                                  : 0

  Total query packets received                            : 0
  Total reply packets sent                                    : 0
  Total reply packets sent errors                         : 0
  Total URO TLV not present errors                   : 0
  Total invalid port number errors                       : 0
  Total no source address errors                        : 0
  Total no retrun path errors                                : 0
  Total unsupported querier control code errors : 0
  Total unsupported timestamp format errors     : 0
  Total timestamp not available errors                : 0
  Total unsupported mandatory TLV errors        : 0
  Total invalid packet errors                                : 0
  Current rate                                                       : 0 pkts/sec
  Rate high water mark                                         : 0 pkts/sec
```

Show performance-measurement responder interfaces

```
 # show performance-measurement responder interfaces [ <name> ] [ location <> ]

--------------------------------------------------------------------------------
0/2/CPU0
--------------------------------------------------------------------------------

Interface Name: GigabitEthernet0/2/0/2
  Interface Handle        : 0x10000a0
  Local IPV4 Address      : 13.13.13.3
  Local IPV6 Address      : 13:13:13::3
  Current rate            : 0 pkts/sec
  Rate high water mark    : 1 pkts/sec
  Cleanup time remaining  : 3580 sec
```

Show performance-measurement responder counters

```
 # show performance-measurement responder counters interfaces [ <name> ] [ location <> ]

 --------------------------------------------------------------------------------------
 0/0/CPU0
 --------------------------------------------------------------------------------------
Interface Name: GigabitEthernet0/2/0/4
  Delay-Measurement:
    Query packets received        : 428615
    Reply packets sent                : 428615
    Query packets errored          : 0
    Reply packets errored           : 0
```

Action CLIs

- Querier side action CLIs

```
# clear performance-measurement delay interfaces           [ <name> ] [ location <> ]
```

- ➢ Clear the data for PM delay-measurement session(s) and history on the given interface or location on querier side.
- ➢ Clear the advertised metrics.

```
# clear performance-measurement counters           [ interfaces <name> ] [ location <> ]
```

- ➢ Clear the counters for the given interface or location on querier side.

- Responder side action CLIs

```
# clear performance-measurement responder counters      [ <name>] [ location <> ]
```

- ➢ Clear the counters for the given interface or location on responder side.