Interoperable Multi-Domain Delay-aware Provisioning using Segment Routing Monitoring and BGP-LS Advertisement

F. Paolucci(1), V. Uceda(2), A. Sgambelluri(3), F. Cugini(4), O. Gonzales De Dios(2), V. Lopez(2), L.M. Contreras(2), P. Monti(3), P. Iovanna (5), F. Ubaldi (5), T. Pepe (5), P. Castoldi(1)

(1) Scuola Superiore Sant'Anna, Pisa, Italy, fr.paolucci@sssup.it
(2) Telefonica I+D, Madrid, Spain, victor.lopezalvarez@telefonica.com
(3) KTH Royal Institute of Technology, Kista, Sweden, pmonti@kth.se
(4) CNIT, Pisa, Italy, filippo.cugini@cnit.it
(5) Ericsson, Pisa, Italy, paola.iovanna@ericsson.com

Abstract *This paper demonstrates a multi-domain SDN orchestrator using delay information to provision network services using BGP-LS and a novel monitoring system enabled by Segment Routing. Moreover, it is the first implementation and interoperability of the BGP-LS extensions for TE metrics.*

Introduction

Inter Data Center (DC) communication and advanced mobile services are pushing Operators to introduce efficient mechanisms to provision network services having strict constraints on end-to-end delay. This requires effective techniques to measure and monitor network delay performance as well as efficient mechanisms to handle collected measurements. This monitoring process is more complex in the case of multi-domain networks, where multiple controllers or Path Computation Elements (PCEs) have to provide their collected measurements to a network Orchestrator or Hierarchical PCE, which must correlate the parameters from different domains and finally enforces the provisioning of the end-to-end delay-constrained services.

Nowadays, network operator must deploy specific probes to measure the delay or to measure the delay using already established Label Switched Paths (LSPs). The probes solution is more costly, while the LSP approach requires to signal many LSPs just for measuring purposes, introducing scalability issues.

In addition, currently available controller-to-orchestrator communications do not include protocol extensions enabling the handling and the advertisement of the collected delay measurements.

In this paper, two main novelties are introduced. First, by relying on segment routing (SR[1,2]), we enable delay measurements over multiple candidate routes without requiring related LSP signalling sessions. Second, we extend the BGP-LS protocol (i.e., North-Bound Distribution of Link-State and TE Information using BGP[3]) to encompass retrieved delay parameters (as defined in[4]) within the controller to orchestrator communications. An experimental demonstration over a Pan-European testbed is presented, also including two different and fully interoperable BGP-LS protocol implementations[5].

Segment Routing Monitoring

SR is a TE technique compatible with traditional MPLS data plane and based on the source routing paradigm. In SR, a specifically computed stack of labels is enforced at the ingress node on data packets to define their routing. The stack of labels, called segment list in SR, is composed by an ordered list of segment identifiers (SIDs). A SID can represent an IP prefix, e.g., the loopback address of a node. During packet forwarding, only the top label in the segment list is processed and the packet is forwarded along the shortest path toward the network element represented by the top label.

Unlike traditional MPLS networks, SR maintains per-flow state only at the ingress node where the segment list is initialized. No state is maintained in transit nodes. Moreover, no signalling protocol is required, thus avoiding the time-consuming RSVP-TE signalling procedure while reducing the overall control plane load.

The SR technology is here exploited to implement a monitoring system which does not require signalled LSPs. The monitoring system relies on probes that are routed according to the enforced SR segment list.

Two types of probes can be considered. The first type is originated and terminated by the network nodes. Examples include MPLS Ping or Bidirectional Forwarding Detection (BFD) messages. The second type relies on external monitoring systems which inject and receive on different synchronized locations specifically designed timestamped probes. Typically, the former type has limited accuracy and it is able to retrieve only round trip delays. On the other hand, the latter type permits accurate unidirectional delay measurements subject that

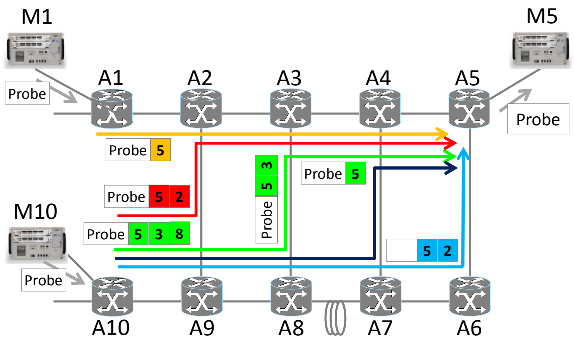


Fig. 1: Network testbed for SR monitoring on domain A

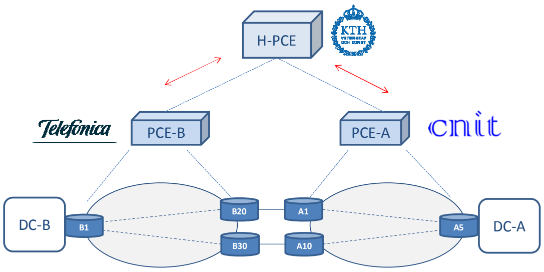


Fig. 2: Multi-domain network at the H-PCE

external monitoring systems (i.e., probe generators/analysers) are available. Both types are supported by the considered SR monitoring technique. Here, the second type is exploited. Moreover, in agreement with[4], delay values do not vary significantly based upon the offered traffic load. Indeed, the measured values refer only to a traffic class for delay-sensitive service that experiences minimal queuing delay.

In this implementation, the network domain A shown in Fig. 1 is considered. It consists of ten SR-capable nodes. Router A1 and A10 are domain border nodes, A5 is connected to the client network (e.g., a Data Center). Three synchronized monitoring systems M1, M10 and M5 are attached to nodes A1, A10, and A5 respectively. The SR monitoring architecture is used to evaluate the delay performance of the candidate routes A1-A5 and A10-A5.

The domain controller, by operating on ingress nodes A1 and A10 configures the segment lists to be enforced on the probes.

On A1, a single label with SID 5 is sufficient to route the probe generated by M1 along the unique shortest path towards A5, where the SR label is popped and the probe delivered to M5.

On A10, four equal cost candidate routes are available towards A5. That is, four segment lists need to be configured in A10. For example, segment list with three SIDs 8(top)-3-5(bottom) allows the probe to follow the green route, i.e., to be forwarded first to R8, where the top label is popped, then to R3 when also label SID 3 is popped, and finally to A5 where the last SR label is removed and the probe delivered to M5.

Besides shortest routes, additional routes may need to be considered, e.g., to evaluate potential service degradations in the case of link/node failures. In this implementation, we measure path statistics in order to describe a path with two delay values: a min value (on the route having minimum delay) and a max delay value (on the disjoint route). For example, to retrieve statistics for the disjoint A1-A5 route, the segment list 10-6-5 is also applied.

All the segment list enforcements do not trigger any RSVP-TE signalling session in the network.

Extended BGP-LS Implementation

The unidirectional delay statistics measured by the monitoring system are elaborated by the Domain Controller (PCE-A and -B) to define the delay values to be associated to the set of intra-domain link descriptions advertised to the multi-domain Orchestrator (H-PCE).

In this implementation, the multi-domain network shown in Fig. 2 is considered, including domain A of Fig. 1. Two bidirectional inter-domain links are present: B20-A1 and B30-A10. Each domain controller is responsible to provide to the domain Orchestrator both outgoing inter-domain link and intra-domain network information. Intra-domain links can refer to two cases: the full internal topology or an abstracted border-to-border topology. Both cases can be considered, providing different scalability and path computation capability performance at the Orchestrator. In this study, we consider the latter case where, for example, Controller A advertises the virtual intra-domain links A1-A5 and A10-A5, i.e., those evaluated through the SR monitoring system.

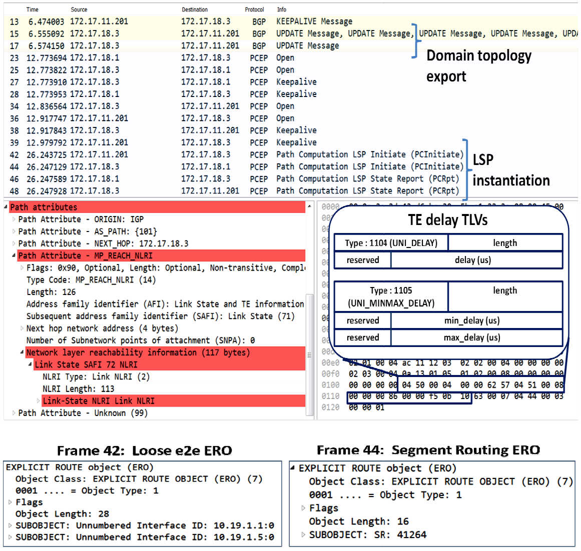
BGP-LS is used for this purpose, here extended to encompass monitoring information according to[3,4]. In particular, both Unidirectional Link Delay TLV and Min/Max Unidirectional Link Delay TLV are implemented to carry the measured border-to-border delay values. To prevent oscillations which may trigger excessive protocol updates, the min and max values are here configured as the delay of the shortest border-to-border path and of its disjoint path, respectively.

Experimental demonstration

The proposed SR-based monitoring systems and extended BGP-LS advertisements have been validated by reproducing the networks shown in Fig. 1 and 2 on the Pan European Sandbox testbed created within the 5G Exchange project[6].

Domain A is implemented at CNIT lab in Pisa, Italy. It includes ten SR-capable nodes derived from commercially-available routers enhanced with ad-hoc SR middleware software acting as

Path Computation Client with instantiation capabilities. In particular, a geographic Pisa-Stockholm-Pisa link (120ms RTT), realized by means of an IPSec tunnel, connects nodes A7 and A8. The domain is controlled by a SR Stateful child PCE derived from[2]. A co-located BGP speaker, derived from[5] and developed in C++, exports intra-domain (either physical or virtual) topology with delay parameters. The delay monitoring probe system is implemented by means of the Spirent SPTN4U generator and analyzer, capable of performing accurate delay measurements. Such values are averaged and provided to the BGP speaker. Domain B is implemented at Telefonica Lab premises in Spain. It is based on the Open Source Netphony ABNO architecture[5]. For this scenario, a child PCE is deployed, as well as a BGP Peer supporting BGP-LS extensions. The multi-domain Orchestrator is implemented at KTH premises in Sweden. It consists of a parent PCE, derived from[5]. Two dedicated IPSec tunnels are used to connect the child PCEs of domain A and B (average latency of 59ms and 57ms RTT, respectively). The orchestrator is equipped with an extended version of a BGP peer, supporting the exchange of TE delay metrics. The orchestrator is also connected to an external client to receive multi-domain LSP requests with minimum delay constraint.


```
    Time      Source                       Destination    Protocol  Info
13   6.474003 172.17.11.201                172.17.18.3     BGP      KEEPALIVE Message
15   6.555092 172.17.18.3                  172.17.11.201   BGP      UPDATE Message, UPDATE Message, UPDATE Message, UPDATE Message, UPDA
17   6.574150 172.17.18.3                  172.17.11.201   BGP      UPDATE Message
23  12.773694 172.17.18.1                  172.17.18.3     PCEP     Open
25  12.773822 172.17.18.3                  172.17.18.1     PCEP     Open
27  12.773910 172.17.18.3                  172.17.18.1     PCEP     Keepalive
28  12.773953 172.17.18.1                  172.17.18.3     PCEP     Keepalive
34  12.836564 172.17.18.3                  172.17.11.201   PCEP     Open
36  12.917747 172.17.11.201                172.17.18.3     PCEP     Open
38  12.917843 172.17.18.3                  172.17.11.201   PCEP     Keepalive
39  12.979792 172.17.11.201                172.17.18.3     PCEP     Keepalive
42  26.243725 172.17.11.201                172.17.18.3     PCEP     Path Computation LSP Initiate (PCInitiate)
44  26.247129 172.17.18.3                  172.17.18.1     PCEP     Path Computation LSP Initiate (PCInitiate)
46  26.247589 172.17.18.1                  172.17.18.3     PCEP     Path Computation LSP State Report (PCRpt)
48  26.247928 172.17.18.3                  172.17.11.201   PCEP     Path Computation LSP State Report (PCRpt)
```

```
◢ Path attributes
  ▷ Path Attribute - ORIGIN: IGP
  ▷ Path Attribute - AS_PATH: {101}
  ▷ Path Attribute - NEXT_HOP: 172.17.18.3
  ◢ Path Attribute - MP_REACH_NLRI
    ▷ Flags: 0x90, Optional, Length: Optional, Non-transitive, Compl
      Type Code: MP_REACH_NLRI (14)
      Length: 126
      Address family identifier (AFI): Link State and TE information
      Subsequent address family identifier (SAFI): Link State (71)
    ▷ Next hop network address (4 bytes)
      Number of Subnetwork points of attachment (SNPA): 0
    ◢ Network layer reachability information (117 bytes)
      ◢ Link State SAFI 72 NLRI
          NLRI Type: Link NLRI (2)
          NLRI Length: 113
        ▷ Link-State NLRI Link NLRI
  ▷ Path Attribute - Unknown (99)
```

```
00e0  02 01 00 04 ac 11 12 03  02 02 00 04 00 00 00 00
00f0  02 03 00 04 0a 13 01 05  01 02 00 08 00 00 00 00
0100  00 00 00 00 04 50 00 04  00 00 62 57 04 51 00 08
0110  00 00 00 86 00 00 f5 0b  10 63 00 07 04 44 00 03
0120  00 00 01
```

Frame 42: Loose e2e ERO

```
EXPLICIT ROUTE object (ERO)
  Object Class: EXPLICIT ROUTE OBJECT (ERO) (7)
  0001 .... = Object Type: 1
▷ Flags
  Object Length: 28
▷ SUBOBJECT: Unnumbered Interface ID: 10.19.1.1:0
▷ SUBOBJECT: Unnumbered Interface ID: 10.19.1.5:0
```

Frame 44: Segment Routing ERO

```
◢ EXPLICIT ROUTE object (ERO)
    Object Class: EXPLICIT ROUTE OBJECT (ERO) (7)
    0001 .... = Object Type: 1
  ▷ Flags
    Object Length: 16
  ▷ SUBOBJECT: SR: 41264
```

Fig. 3 : Experimental Demonstration. Wireshark capture

Fig. 3 reports a Wireshark capture (collected at CNIT child PCE, address 172.17.18.3) of the demonstration including BGP-LS topology export and PCEP instantiation of an inter-domain path. First, domain topology is exported (frames 15-17). BGP-LS Update Frame 17 is expanded, showing the Link State NLRI Path Attribute enclosing the UNI_DELAY (type 1104) and the UNI_MINMAX (type 1105) TLVs. When a new inter-domain LSP request is submitted to the Orchestrator with minimum e2e delay objective function (OF=3000), it runs a minimum delay e2e path computation algorithm. According to the delay parameters stored in the TEDB for both intra-domain abstracted links and inter-domain links (as shown in Fig. 2), the algorithm computes the e2e delay-based shortest path across the two domains and triggers the setup of the path by sending two PCEP Initiate messages to the child PCEs. In particular considering the domain A, due to the geographic link, link A10-A5 advertised delay is around 25ms, whereas A1-A5 only 0.1ms. Thus, the Orchestrator selects A1-A5 path by sending a PCEP Initiate message with loose ERO (i.e., source and destination, frame 42). Child PCE performs intra-domain expansion and segment list computation, then sends an Initiate message to ingress router A1 with strict SR-ERO (frame 44). PCEP Report messages are provided upwards to acknowledge the path activation. Once received Report messages from both the child PCEs, the Orchestrator informs the client of the activation of the e2e multi-domain path. The overall procedure takes 142ms. The main contribution is due to the latency of the IPSec tunnels between the Orchestrator and the child PCEs. The e2e algorithm contribution is less than 2ms while the SR-PCE segment computation takes less than 4ms.

Conclusions

This paper demonstrates the utilization of Segment Routing to measure unidirectional delay statistics on a real testbed. Moreover, it presents the first implementation and interoperability demonstration of the BGP-LS TE metrics. Finally, this work validates an orchestrator architecture to deal with services considering the delay metrics.

Acknowledgements

This work has been partially supported by the EU H2020 project 5GEx (Grant Agreement no. 671636).

References

[1] R. Geib et al., draft-ietf-spring-oam-usecase-01, 2015

[2] A Sgambelluri et al., “Experimental demonstration of segment routing”, J. of Lightweight Technology, Jan ‘16

[3] S. Previdi et al., draft-previdi-idr-bgpls-te-metric-extensions-00, Feb 2016

[4] S. Giacalone, OSPF TE Metric Extension, RFC7471, 2015

[5] O. G. De Dios et al., Multipartner Demonstration of BGP-LS-Enabled Multidomain EON Control and Instantiation with H-PCE", JOCN, vol.7, n.12, 2015

[6] C. J. Bernardos et al., ”5G Exchange (5GEx) – Multi-domain Orchestration for Software Defined Infrastructures”, EUCNC 2015.