Independent Submission
Request for Comments: 8604
Category: Informational
ISSN: 2070-1721

C. Filsfils, Ed.
Cisco Systems, Inc.
S. Previdi
Huawei Technologies
G. Dawra, Ed.
LinkedIn
W. Henderickx
Nokia
D. Cooper
CenturyLink
June 2019

Interconnecting Millions of Endpoints with Segment Routing

Abstract

This document describes an application of Segment Routing to scale the network to support hundreds of thousands of network nodes, and tens of millions of physical underlay endpoints. This use case can be applied to the interconnection of massive-scale Data Centers (DCs) and/or large aggregation networks. Forwarding tables of midpoint and leaf nodes only require a few tens of thousands of entries. This may be achieved by the inherently scaleable nature of Segment Routing and the design proposed in this document.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This is a contribution to the RFC Series, independently of any other RFC stream. The RFC Editor has chosen to publish this document at its discretion and makes no statement about its value for implementation or deployment. Documents approved for publication by the RFC Editor are not candidates for any level of Internet Standard; see Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc8604.

Table of Contents

1. Introduction

This document describes how Segment Routing (SR) can be used to interconnect millions of endpoints.

2. Terminology

The following terms and abbreviations are used in this document:

Term	Definition
Agg	Aggregation
BGP	Border Gateway Protocol
DC	Data Center
DCI	Data Center Interconnect
ECMP	Equal-Cost Multipath
FIB	Forwarding Information Base
LDP	Label Distribution Protocol
LFIB	Label Forwarding Information Base
MPLS	Multiprotocol Label Switching
PCE	Path Computation Element
PCEP	Path Computation Element Communication Protocol
PW	Pseudowire
SLA	Service Level Agreement
SR	Segment Routing
SRTE Policy	Segment Routing Traffic Engineering Policy
TE	Traffic Engineering
TI-LFA	Topology Independent Loop-Free Alternate

3. Reference Design

The network diagram below illustrates the reference network topology used in this document:

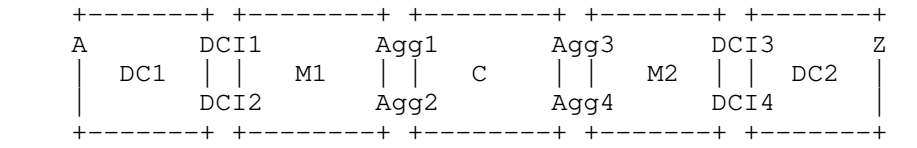

Figure 1: Reference Topology

The following apply to the reference topology above:

- Independent ISIS-OSPF/SR instance in core (C) region.
- Independent ISIS-OSPF/SR instance in Metro1 (M1) region.

- Independent ISIS-OSPF/SR instance in Metro2 (M2) region.
- BGP/SR in DC1.
- BGP/SR in DC2.
- Agg routes (Agg1, Agg2, Agg3, Agg4) are redistributed from C to M (M1 and M2) and from M to DC domains.
- No other route is advertised or redistributed between regions.
- The same homogeneous Segment Routing Global Block (SRGB) is used throughout the domains (e.g., 16000-23999).
- Unique SRGB sub-ranges are allocated to each metro (M) and core (C) domain:
  * The 16000-16999 range is allocated to the core (C) domain/region.
  * The 17000-17999 range is allocated to the M1 domain/region.
  * The 18000-18999 range is allocated to the M2 domain/region.
  * Specifically, the Agg1 router has Segment Identifier (SID) 16001 allocated, and the Agg2 router has SID 16002 allocated.
  * Specifically, the Agg3 router has SID 16003 allocated, and the anycast SID for Agg3 and Agg4 is 16006.
  * Specifically, the DCI3 router has SID 18003 allocated, and the anycast SID for DCI3 and DCI4 is 18006.
  * Specifically, at the Agg1 router, the binding SID 4001 leads to DCI pair (DCI3, DCI4) via a specific low-latency path {16002, 16003, 18006}.
- The same SRGB sub-range is reused within each DC (DC1 and DC2) region for each DC (e.g., 20000-23999). Specifically, nodes A and Z both have SID 20001 allocated to them.

4. Control Plane

This section provides a high-level description of how a control plane could be implemented using protocol components already defined in other RFCs.

The mechanism through which SRTE Policies are defined, computed, and programmed in the source nodes is outside the scope of this document.

Typically, a controller or a service orchestration system programs node A with a PW to a remote next-hop node Z with a given SLA contract (e.g., low-latency path, disjointness from a specific core plane, disjointness from a different PW service).

Node A automatically detects that node Z is not reachable. It then automatically sends a PCEP request to an SR PCE for an SRTE policy that provides reachability information for node Z with the requested SLA.

The SR PCE [RFC4655] is made of two components: a multi-domain topology and a computation engine. The multi-domain topology is continuously refreshed through BGP - Link State (BGP-LS) feeds [RFC7752] from each domain. The computation engine is designed to implement TE algorithms and provide output in SR Path format. Upon receiving the PCEP request [RFC5440], the SR PCE computes the requested path. The path is expressed through a list of segments (e.g., {16003, 18006, 20001}) and provided to node A.

The SR PCE logs the request as a stateful query and hence is able to recompute the path at each network topology change.

Node A receives the PCEP reply with the path (expressed as a segment list). Node A installs the received SRTE policy in the data plane. Node A then automatically steers the PW into that SRTE policy.

5. Illustration of the Scale

According to the reference topology shown in Figure 1, the following assumptions are made:

- There is one core domain, and there are 100 leaf (metro) domains.
- The core domain includes 200 nodes.
- Two nodes connect each leaf (metro) domain. Each node connecting a leaf domain has a SID allocated. Each pair of nodes connecting a leaf domain also has a common anycast SID. This yields up to 300 prefix segments in total.

- A core node connects only one leaf domain.

- Each leaf domain has 6,000 leaf-node segments. Each leaf node has 500 endpoints attached and thus 500 adjacency segments. This yields a total of 3 million endpoints for a leaf domain.

Based on the above, the network scaling numbers are as follows:

- 6,000 leaf-node segments multiplied by 100 leaf domains: 600,000 nodes.

- 600,000 nodes multiplied by 500 endpoints: 300 million endpoints.

The node scaling numbers are as follows:

- Leaf-node segment scale: 6,000 leaf-node segments + 300 core-node segments + 500 adjacency segments = 6,800 segments.

- Core-node segment scale: 6,000 leaf-domain segments + 300 core-domain segments = 6,300 segments.

In the above calculations, the link-adjacency segments are not taken into account. These are local segments and, typically, less than 100 per node.

It has to be noted that, depending on leaf-node FIB capabilities, leaf domains could be split into multiple smaller domains. In the above example, the leaf domains could be split into six smaller domains so that each leaf node only needs to learn 1,000 leaf-node segments + 300 core-node segments + 500 adjacency segments, yielding a total of 1,800 segments.

6. Design Options

This section describes multiple design options to illustrate scale as described in the previous section.

6.1. Segment Routing Global Block (SRGB) Size

In the simplified illustrations in this document, we picked a small homogeneous SRGB range of 16000-23999. In practice, a large-scale design would use a bigger range, such as 16000-80000 or even larger. A larger range provides allocations for various TE applications within a given domain.

6.2. Redistribution of Routes for Agg Nodes

The operator might choose to not redistribute the routes for Agg nodes into the Metro/DC domains. In that case, more segments are required in order to express an inter-domain path.

For example, node A would use an SRTE Policy {DCI1, Agg1, Agg3, DCI3, Z} in order to reach Z instead of {Agg3, DCI3, Z} in the reference design.

6.3. Sizing and Hierarchy

The operator is free to choose among a small number of larger leaf domains, a large number of small leaf domains, or a mix of small and large core/leaf domains.

The operator is free to use a two-tier (Core/Metro) or three-tier (Core/Metro/DC) design.

6.4. Local Segments to Hosts/Servers

Local segments can be programmed at any leaf node (e.g., node Z) in order to identify locally attached hosts (or Virtual Machines (VMs)). For example, if node Z has bound a local segment 40001 to a local host ZH1, then node A uses the following SRTE Policy in order to reach that host: {16006, 18006, 20001, 40001}. Such a local segment could represent the NID (Network Interface Device) in the context of the service provider access network, or a VM in the context of the DC network.

6.5. Compressed SRTE Policies

As an example and according to Section 3, we assume that node A can reach node Z (e.g., with a low-latency SLA contract) via the SRTE policy that consists of the path Agg1, Agg2, Agg3, DCI3/4(anycast), Z. The path is represented by the segment list {16001, 16002, 16003, 18006, 20001}.

It is clear that the control-plane solution can install an SRTE Policy {16002, 16003, 18006} at Agg1, collect the binding SID allocated by Agg1 to that policy (e.g., 4001), and hence program node A with the compressed SRTE Policy {16001, 4001, 20001}.

From node A, 16001 leads to Agg1. Once at Agg1, 4001 leads to the DCI pair (DCI3, DCI4) via a specific low-latency path {16002, 16003, 18006}. Once at that DCI pair, 20001 leads to Z.

Binding SIDs allocated to "intermediate" SRTE Policies achieve the compression of end-to-end SRTE Policies.

The segment list {16001, 4001, 20001} expresses the same path as {16001, 16002, 16003, 18006, 20001} but with two less segments.

The binding SID also provides for inherent churn protection.

When the core topology changes, the control plane can update the low-latency SRTE Policy from Agg1 to the DCI pair to DC2 without updating the SRTE Policy from A to Z.

7. Deployment Model

It is expected that this design will be used in "green field" deployments as well as interworking ("brown field") deployments with an MPLS design across multiple domains.

8. Benefits

The design options illustrated in this document allow interconnections on a very large scale. Millions of endpoints across different domains can be interconnected.

8.1. Simplified Operations

Two control-plane protocols not needed in this design are LDP and RSVP-TE. No new protocol has been introduced. The design leverages the core IP protocols ISIS, OSPF, BGP, and PCEP with straightforward SR extensions.

8.2. Inter-domain SLAs

Fast reroute and resiliency are provided by TI-LFA with sub-50-ms fast reroute upon failure of a link, node, or Shared Risk Link Group (SRLG). TI-LFA is described in [SR-TI-LFA].

The use of anycast SIDs also provides improved availability and resiliency.

Inter-domain SLAs can be delivered (e.g., latency vs. cost-optimized paths, disjointness from backbone planes, disjointness from other services, disjointness between primary and backup paths).

Existing inter-domain solutions do not provide any support for SLA contracts. They just provide best-effort reachability across domains.

8.3. Scale

In addition to having eliminated the need for LDP and RSVP-TE, per-service midpoint states have also been removed from the network.

8.4. ECMP

Each policy (intra-domain or inter-domain, with or without TE) is expressed as a list of segments. Since each segment is optimized for ECMP, the entire policy is optimized for ECMP. The benefit of an anycast prefix segment optimized for ECMP should also be considered (e.g., 16001 load-shares across any gateway from the M1 leaf domain to the Core and 16002 load-shares across any gateway from the Core to the M1 leaf domain).

9. IANA Considerations

This document has no IANA actions.

10. Manageability Considerations

This document describes an application of SR over the MPLS data plane. SR does not introduce any changes in the MPLS data plane. The manageability considerations described in [RFC8402] apply to the MPLS data plane when used with SR.

11. Security Considerations

This document does not introduce additional security requirements and mechanisms other than those described in [RFC8402].

12. Informative References

[RFC4655] Farrel, A., Vasseur, J.-P., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <https://www.rfc-editor.org/info/rfc4655>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <https://www.rfc-editor.org/info/rfc5440>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[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>.

[SR-TI-LFA] Litkowski, S., Bashandy, A., Filsfils, C., Decraene, B., Francois, P., Voyer, D., Clad, F., and P. Camarillo, "Topology Independent Fast Reroute using Segment Routing", Work in Progress, draft-ietf-rtgwg-segment-routing-ti-lfa-01, March 2019.

Acknowledgements

We would like to thank Giles Heron, Alexander Preusche, Steve Braaten, and Francis Ferguson for their contributions to the content of this document.

Contributors

The following people substantially contributed to the editing of this document:

Dennis Cai
Individual

Tim Laberge
Individual

Steven Lin
Google Inc.

Bruno Decraene
Orange

Luay Jalil
Verizon

Jeff Tantsura
Individual

Rob Shakir
Google Inc.

Authors' Addresses

Clarence Filsfils (editor)
Cisco Systems, Inc.
Brussels
Belgium

Email: cfilsfil@cisco.com

Stefano Previdi
Huawei Technologies

Email: stefano@previdi.net

Gaurav Dawra (editor)
LinkedIn
United States of America

Email: gdawra.ietf@gmail.com

Wim Henderickx
Nokia
Copernicuslaan 50
Antwerp 2018
Belgium

Email: wim.henderickx@nokia.com

Dave Cooper
CenturyLink

Email: Dave.Cooper@centurylink.com