Network Working Group
Internet-Draft
Intended status: Informational
Expires: June 24, 2018

C. Filsfils, Ed.
S. Previdi
G. Dawra, Ed.
Cisco Systems, Inc.
D. Cai
Individual
W. Henderickx
Nokia
D. Cooper
Level 3
T. Laberge
S. Lin
Individual
B. Decraene
Orange
L. Jalil
Verizon
J. Tantsura
Individual
R. Shakir
Google

December 21, 2017

Interconnecting Millions Of Endpoints With Segment Routing
draft-filsfils-spring-large-scale-interconnect-08

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 DC's and/or large aggregation networks. Forwarding tables of midpoint and leaf nodes only require a few tens of thousands of entries.

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].

Status of This Memo

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

This Internet-Draft will expire on June 24, 2018.

Table of Contents

1. Introduction

This document describes how SR can be used to interconnect millions of endpoints.The following terminology is used in this document:

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 MultiPathing
FIB	Forwarding Information Base
LDP	Label Distribution Protocol
LFIB	Label Forwarding Information Base
MPLS	Multi-Protocol Label Switching
PCE	Path Computation Element
PCEP	Path Computation Element 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 Alternative

3. Reference Design

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

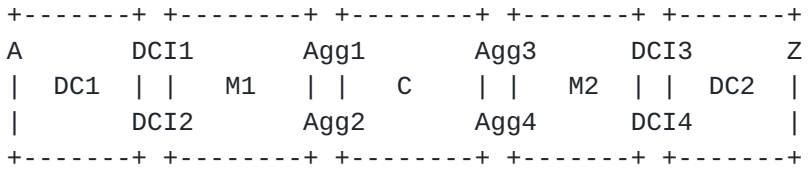

Figure 1: Reference Topology

The following applies 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 SRGB is used throughout the domains (e.g. 16000-23999).

Unique SRGB sub-ranges are allocated to each metro (M) and core (C) domains:

16000-16999 range is allocated to the core (C) domain/region.

17000-17999 range is allocated to the M1 domain/region.

18000-18999 range is allocated to the M2 domain/region.

Specifically, Agg3 router has SID 16003 allocated and the anycast SID for Agg3 and Agg4 is 16006.

Specifically, DCI3 router has SID 18003 allocated and the anycast SID for DCI3 and DCI4 is 18006.

The same SRGB sub-range is re-used within each DC (DC1 and DC2) region. for each DC: e.g. 20000-23999. Specifically, range 20000-23999 range is used in both DC1 and DC2 regions and nodes A and Z have both SID 20001 allocated to them.

4. Control Plane

This section provides a high-level description of an implemented control-plane.

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

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

Node A automatically detects that it does not have reachability to Z. It then automatically sends a PCEP request to an SR PCE for an SRTE policy that provides reachability to Z with the requested SLA.

The SR PCE is made of two components. A multi-domain topology and a computation engine. The multi-domain topology is continuously refreshed through BGP-LS feeds from each domain. The computing engine implements Traffic Engineering (TE) algorithms designed specifically for SR path expression. Upon receiving the PCEP request, the SR PCE computes the requested path. The path is expressed through a list of segments (e.g. {16003, 16005, 18001} and provided to node A.

The SR PCE logs the request as a stateful query and hence is capable 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 dataplane. Node A then automatically steers the PW into that SRTE policy.

5. Illustration of the scale

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

- There's 1 core domain and 100 of 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 brings up to 300 prefix segments in total.
- A core node connects only one leaf domain.
- Each leaf domain has 6000 leaf node segments. Each leaf-node has 500 endpoints attached, thus 500 adjacency segments. In total, it is 3 millions 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 millions of 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 6 smaller domains so that each leaf node only need to learn 1000 leaf node segments + 300 core node segments + 500 adjacency segments which gives a total of 1,800 segments.

6. Design Options

This section describes multiple design options to the illustration of previous section.

6.1. SRGB Size

In the simplified illustrations of 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.

6.2. Redistribution of Agg nodes routes

The operator might choose to not redistribute the Agg nodes routes 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 2-tier design (Core/Metro) or a 3-tier (Core/Metro/DC).

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 VM's). 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, 17006, 20001, 40001}. Such local segment could represent the NID (Network Interface Device) in the context of the SP access network, or VM in the context of the DC network.

6.5. Compressed SRTE policies

As an example and according to Section 3, we assume node A can reach node Z (e.g., with a low-latency SLA contract) via the SRTE policy consisting 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 SID's allocated to "intermediate" SRTE Policies allow to compress end-to-end SRTE Policies.

The segment list {16001, 4001, 20001} expresses the same path as {16001, 16002, 16003, 18006, 20001} but with 2 less segments.

The Binding SID also provide for an 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 be deployed as a green field but as well in interworking (brown field) with seamless-mpls design as described in [I-D.ietf-mpls-seamless-mpls].

8. Benefits

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

8.1. Simplified operations

Two protocols have been removed from the network: LDP and RSVP-TE. No new protocol has been introduced. The design leverage the core IP protocols: ISIS, OSPF, BGP, PCEP with straightforward SR extensions.

8.2. Inter-domain SLA

Fast reroute and resiliency is provided by TI-LFA with sub-50msec FRR upon Link/Node/SRLG failure. TI-LFA is described in [I-D.francois-rtgwg-segment-routing-ti-lfa].

The use of anycast SID's also provide an improvement in availability and resiliency.

Inter-domain SLA's can be delivered, e.g., latency vs. cost optimized path, disjointness from backbone planes, disjointness from other services, disjointness between primary and backup paths.

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

8.3. Scale

In addition to having eliminated two control plane protocols, per-service midpoint states have also been removed from the network.

8.4. ECMP

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

9. IANA Considerations

TBD

10. Manageability Considerations

TBD

11. Security Considerations

TBD

12. Acknowledgements

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

13. References

13.1. 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>.

13.2. Informative References

[I-D.francois-rtgwg-segment-routing-ti-lfa] Francois, P., Bashandy, A., Filsfils, C., Decraene, B., and S. Litkowski, "Abstract", draft-francois-rtgwg-segment-routing-ti-lfa-04 (work in progress), December 2016.

[I-D.ietf-mpls-seamless-mpls] Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07 (work in progress), June 2014.

Authors' Addresses

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

Email: cfilsfil@cisco.com

Stefano Previdi
Cisco Systems, Inc.
Via Del Serafico, 200
Rome 00142
Italy

Email: stefano@previdi.net

Gaurav Dawra (editor)
Cisco Systems, Inc.
USA

Email: gdawra.ietf@gmail.com

Dennis Cai
Individual

Wim Henderickx
Nokia
Copernicuslaan 50
Antwerp 2018
Belgium

Email: wim.henderickx@nokia.com

Dave Cooper
Level 3

Email: Dave.Cooper@Level3.com

Tim Laberge
Individual

Steven Lin
Individual

Email: slin100@yahoo.com

Bruno Decraene
Orange
FR

Email: bruno.decraene@orange.com

Luay Jalil
Verizon
400 International Pkwy
Richardson, TX 75081
United States

Email: luay.jalil@verizon.com

Jeff Tantsura
Individual

Email: jefftant@gmail.com

Rob Shakir
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043

Email: robjs@google.com