Interconnecting Millions Of Endpoints With Segment Routing
draft-filsfils-spring-large-scale-interconnect-03

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 1, 2016.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction .............................. 3
2. Reference Design .......................... 3
3. Control Plane ............................ 5
4. Illustration of the scale ............... 5
5. Optional Designs .......................... 6
6. Deployment Model .......................... 7
7. Benefits .................................. 7
8. IANA Considerations ...................... 8
9. Manageability Considerations ............ 8
10. Security Considerations ................. 8
11. Acknowledgements ....................... 9
12. References ............................. 9
Authors’ Addresses .......................... 9
1 Introduction

This document describes how SR can be used to interconnect millions of endpoints.

1.1. Terminology

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].

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agg</td>
<td>Aggregation</td>
</tr>
<tr>
<td>BGP</td>
<td>Border Gateway Protocol</td>
</tr>
<tr>
<td>DC</td>
<td>Data Center</td>
</tr>
<tr>
<td>DCI</td>
<td>Data Center Interconnect</td>
</tr>
<tr>
<td>ECMP</td>
<td>Equal Cost MultiPathing</td>
</tr>
<tr>
<td>FIB</td>
<td>Forwarding Information Base</td>
</tr>
<tr>
<td>LDP</td>
<td>Label Distribution Protocol</td>
</tr>
<tr>
<td>LFIB</td>
<td>Label Forwarding Information Base</td>
</tr>
<tr>
<td>MPLS</td>
<td>Multi-Protocol Label Switching</td>
</tr>
<tr>
<td>PCE</td>
<td>Path Computation Element</td>
</tr>
<tr>
<td>PCEP</td>
<td>Path Computation Element Protocol</td>
</tr>
<tr>
<td>PW</td>
<td>Pseudowire</td>
</tr>
<tr>
<td>SR</td>
<td>Segment Routing</td>
</tr>
<tr>
<td>TI-LFA</td>
<td>Topology Independent - Loop Free Alternative</td>
</tr>
</tbody>
</table>

2 Reference Design

For example, an operator could do the following:

- Independent ISIS-OSPF/SR instance in core (C)
- Independent ISIS-OSPF/SR instance in Metro1 (M1)
- Independent ISIS-OSPF/SR instance in Metro2 (M2)
- BGP/SR in DC1
- BGP/SR in DC2
- Agg routes are redistributed from C to M and from M to DC domains. Nothing else is distributed
- Same homogenous SRGB throughout the domains (e.g. 16000-23999)
- Allocate unique SRGB sub-ranges to each metro and core domains: 16000-16999 to the core, 17000-17999 to the metro, 18000-18999 to the metro2. Specifically, Agg3 is 16003 and the anycast SID for (Agg3, Agg4) is 16006. DCI3 is 17003 and the anycast SID for (DCI3, DCI4) is 17006
- Re-use the same SRGB sub-range for each DC: e.g. 20000-23999. Specifically A and Z are both 20001.

3. Control-plane

   It is out of the scope of this document to describe how the SRTE Policies are computed and programmed at the source nodes.

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

   The service orchestration programs A with a PW to a remote next-hop Z with a given SLA contract (low-latency path, be disjoint from a specific core plane, be disjoint from a different PW service, etc.).

   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 compute block. The multi-domain topology is continuously refreshed from BGP-LS feeds from each domain. The compute block implements TE algorithms designed specifically for SR path expression. Upon receiving the PCEP request, the SR PCE computes the solution (e.g. {16003, 16005, 18001}) and provides it to A.

   The SR PCE logs the request as a stateful query and hence recomputes another solution upon any multi-domain topology changes that invalidates the previous solution.

   A receives the PCEP reply with the solution. A installs the received SRTE policy in the dataplane. A automatically steers the PW on that SRTE policy.

4. Illustration of the scale

   1 core domain and 100 leaf domains

   Core domain has 200 core nodes. Assume two nodes per each leaf domain, with specific node segment and anycast segments, it’s 300
prefix segments in total. Assume a core node connects only one leaf domain.

Each leaf domain has 6,000 leaf node segments. Each leaf-node has 500 endpoints attached, thus 500 adjacency segments. In total, it is 3M endpoints per leaf domain.

Network wide scale:
6,000x100=600,000 nodes
6,000x100x500=300M endpoints

Per-node segment scale:
Leaf node segment scale: 6,000 (leaf node segments) + 300 (core node segments) + 500 (adj segments) = 6,800
Core node segment scale: 6,000 (leaf domain segments) + 300 (core domain segments) = 6,300

In the above calculation, it didn’t count the link adjacency segments, which is local to the node. Typically it should be <100.

Note, depends on the leaf node FIB capability, we could split the leaf domain into multiple smaller domains. For the above example, we can split the leaf domain to 6 smaller leaf domains. So each leaf node only need to learn 1000 (leaf node segments) + 300 (core node segments) + 500 (adj segments) = 1,800 segments.

5 Optional Designs

5.1 SRGB size
In the simplified illustrations of this document, we picked a small homogenous SRGB range of 16000-23999. In practice, a large-scale design would use a bigger range such as 16000-80000, or even larger.

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

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

5.3 Sizing of the domains and number of Tiers
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 domains.

The operator is free to use a 2-tier design (Core/Metro) or a 3-tier (Core/Metro/DC).

5.4 Local Segments to Hosts/Servers
Local segments can be programmed at any leaf node (e.g. Z) in order to identify locally-attached hosts (or VM’s). For example, if Z has bound a local segment 40001 to a local host ZH1, then A uses the following SRTE Policy to reach that host: (16006, 17006, 20001, 40001). Such local segment could represent the NID (Network Interface Device) device in the context of the SP access network, or VM in the context of the DC network.

5.5 Compressed SRTE policies
We earlier saw that A could reach Z with a low-latency SLA contract via the SRTE policy (16001, 16002, 16003, 17006, 20001).

It is clear that the control-plane solution can install an SRTE policy (16002, 16003, 17006) at Agg1, collect the Binding SID allocated by Agg1 to that policy (e.g. 4001) and hence program A with the compressed SRTE policy (16001, 4001, 20001).

From A, 16001 leads to Agg1. Once at Agg1, 4001 leads to the DCI pair (DCI3, DCI4) via a specific low-latency path (16002, 16003, 17006). 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.

(16001, 4001, 20001) expresses the same path as (16001, 16002, 16003, 17006, 20001) but with 2 less segments.

Binding SID’s 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.

6 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 (draft-ietf-mpls-seamless-mpls).

7 Benefits
7.1 Inter-domain interconnection of millions of endpoints
We have illustrated how millions of endpoints across different domains can be interconnected.

7.2 Simplified operation
We have eliminated two protocols (LDP, RSVP-TE) and have not added any. The design leverage the core IP protocols: ISIS, OSPF, BGP, PCEP with straightforward SR extensions.

7.3 Inter-domain SLA
We leverage TILFA sub-50msec FRR upon Link/Node/SRLG failure.
We leverage the optional use of Anycast SID’s for further availability improvement.
We have shown how 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
We note that the existing inter-domain solutions (Seamless MPLS) do not provide any support for SLA contracts. They just provide a best-effort reachability across domains.

7.4 Scale
We have eliminated two protocols and not added any. We have eliminated midpoint states on a per-service basis.

7.5 ECMP
Each policy (intra or inter-domain, with or without TE) is expressed as a list of segments. As each segment is optimized for ECMP, therefore 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 L1 leaf domain to Core and 16002 load-shares across any gateway from Core to L2 leaf domain.

8. IANA Considerations
None

9. Manageability Considerations
TBD

10. Security Considerations
11. Acknowledgements

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

12. References

12.1. Normative References


12.2. Informative References


Authors’ Addresses

Clarence Filsfils (editor)
Cisco Systems, Inc.
Brussels
BE
Email: cfilsfil@cisco.com

Dennis Cai (editor)
Cisco Systems, Inc.
170, West Tasman Drive
San Jose, CA  95134
US
Email: dcai@cisco.com

Stefano Previdi
Cisco Systems, Inc.
Via Del Serafico, 200
Rome  00142
Italy
Email: sprevidi@cisco.com