IP/MPLS EVOLUTION: SEGMENT ROUTING

Wouter Reynders
Oct 8, 2014
AGENDA

1. Overview of Segment Routing
2. Segment Routing in Shortest Path Routing Applications
3. Segment Routing in Traffic Engineering Applications
4. WAN SDN Controller with PCE
5. Feature Planning
MPLS OVERVIEW - HISTORY

- Two main protocols: LDP or RSVP
  - LDP for scale and simplicity – extensions to LFA/LFA policies/RLFA
  - RSVP for TE and FRR for some time
- To scale MPLS we enabled
  - LD Po RSVP
  - Seamless MPLS: LBL-BGP with LDP or RSVP
- Traffic engineering: RSVP based
- Services through:
  - BGP/IGP shortcuts, PW (T-LDP/BGP), VPLS (LDP/BGP), VPRN (BGP), MVPN (BGP/mLDP/P2MP RSVP)
- Issues:
  - TE solutions don’t scale when we want more granularity/dynamicity, RLFA too complex (dynamic T-LDP signaling)

<table>
<thead>
<tr>
<th>LDP</th>
<th>RSVP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>Multipoint to point</td>
</tr>
<tr>
<td>Operation</td>
<td>Simple</td>
</tr>
<tr>
<td>Dependencies</td>
<td>Relies on IGP</td>
</tr>
<tr>
<td>LBL allocation</td>
<td>Local significant per node (interface)</td>
</tr>
<tr>
<td>Traffic Engineering</td>
<td>No</td>
</tr>
<tr>
<td>Scaling</td>
<td>1 LBL per node (interface)</td>
</tr>
<tr>
<td>Fast Reroute</td>
<td>LFA, LFA Policies, Remote LFA - &lt;100% coverage</td>
</tr>
<tr>
<td>Multicast</td>
<td>mLDP</td>
</tr>
<tr>
<td>IPv6</td>
<td>Extensions required</td>
</tr>
</tbody>
</table>
SEGMENT ROUTING OVERVIEW

- Network elements modeled as segments
  - Prefix segment identifies shortest path to a given prefix
  - Node segment is a type of prefix segment and corresponds to node-id (loopback)
  - Adjacency segment identifies a given interface/next-hop
  - Shortest Path Route = \{node segment\}
  - Source Route = Path = ordered list of node and adjacency segments
  - When MPLS data path is used: segment-id (SID)=label

- Control Plane:
  - SID distribution by IGP to support MPLS Data Path
    - Network-wide unique prefix/node SID indices
      - SID index is an offset from start of label range
      - Labels drawn from pre-assigned MPLS label range on node
    - Locally significant adjacency SID labels
      - Drawn from dynamic MPLS label range on each node
    - Supports both IPv4 and IPv6 SID using MPLS labels
  - Native IPv6 Support
    - SID support in new Routing Header extension (draft-previdi-6man-segment-routing-header-01)
Inherently supports IPv6 -> no extension required for LDP or RSVP

Inherently supports ECMP

Simplifies RLFA such that no additional T-LDP signaling is required for transport LSPs

Better than RLFA as it guarantees 100% FRR Coverage

Supports TE by leveraging Segment/LBL stacks on the packet (Node Segment ID(s) or Adjacency Segment ID(s))

No need for N^2 signaling of paths like RSVP – SR allows to keep state out of the network

Allows the use of a SDN controller to optimize the TE paths in the network

OAM: full path exercising possible
SEGMENT ROUTING
MAIN APPLICATIONS

• Shortest Path Routing Applications
  - LDP transport LSP infrastructure replacement or backup coverage improvement
  - Shorted path forwarding with ECMP: admin-group, SRLG support
  - Achieve full protection coverage with LFA and remote LFA
  - Can co-exist with LDP or RSVP in seamless MPLS deployments

• Traffic Engineering (TE) Applications
  - Traffic engineered tunnels with state at ingress LER only
    - stateless TE LSR nodes
  - Distributed or node-level TE for resilience: admin-group, SRLG support
  - External TE Controller with Resource Discovery for more advanced TE
AGENDA

1. Overview of Segment Routing
2. Segment Routing in Shortest Path Routing Applications
3. Segment Routing in Traffic Engineering Applications
4. WAN SDN Controller with PCE
5. Feature Planning
• Segment Routing can have the following deployment options
  • Remote LFA next-hop only uses segment routing in IP forwarding
    • Remote LFA next-hop needs a tunnel while primary next-hop and regular LFA next-hop do not
  • Both primary and LFA/Remote LFA next-hops use segment routing in IP forwarding
  • Hybrid LDP/Segment Routing
    • Primary next-hop and regular LFA next-hop use LDP
    • Remote LFA next-hop uses segment routing
REMOTE LFA NEXT-HOP IN SEGMENT ROUTING
SHORTEST PATH ROUTING EXAMPLE

- Node B computes SPF for prefix X
  - primary next-hop (NH) is via link B-A
  - LFA NH does not exist
  - Remote LFA NH requires tunnel to node N (or Z)
    - Puts packet back into shortest path while still avoiding link B-A
  - B programs LFA NH with NHLFE={node N SID label=60, prefix X node-SID=X (optional)}
- Node B forwards packet for prefix X
  - Unlabelled (with label X) over link B-A
  - Pushes label 60 when link B-A fails
  - Node N pops label 60 and looks up packet in FIB
  - Forwards packet unlabelled (with label X) to A via link N-Z
REMOTE LFA NEXT-HOP IN SEGMENT ROUTING SOURCE ROUTING EXAMPLE (Directed LFA)

- Link D-N cost is now 100
- Node B computes SPF for prefix X
  - primary next-hop (NH) is via link B-A
  - LFA NH does not exist
  - Remote LFA NH requires tunnel to node N
    - Shortest path to N does not avoid link B-A
    - Need source routing via D
  - B programs LFA NH with NHLFE={node D SID label=80, adjacency SID=9101, prefix X node-SID=X (optional)}
- Node B forwards packet for prefix X
  - Unlabelled (with label X) over link B-A
  - Pushes label stack 80/9101 when link B-A fails
- Node D pops label 80 and looks up label 9101
  - Forwards packet via link D-N
- Node N looks up packet in FIB and forwards it unlabelled (with label X) to A via link N-Z

Maximum Pushed Segment Routing Label Stack Size = 3
REMOTE LFA IN DEPLOYED LDP NETWORK

- Hybrid LDP/ Segment Routing
- Node B computes SPF for prefix X
  - primary next-hop (NH) is via link B-A using LDP
- Remote LFA NH requires tunnel to node N
- B programs LFA NH with NHLFE={node N SID label=60, prefix X node SID=X}
- Node N pops label 60 and swaps label X
- Packet travels using ISIS segment label all the way to destination/advertising router

Note LDP FEC is “stitched” to a SR Label Route when RLFA NH is activated
SEGMENT ROUTING IN SEAMLESS MPLS NETWORK

- LDP and Segment Routing domains stitched via BGP label routes
- VPRN prefix X advertized by PE2 in BGP and next-hop changed by ABR/ASBR node A to self
- Node B resolves prefix X as follows:
  - Pushes label stack \{50, BGP-LBL(PE2), VC-LBL(X)\} when forwarding to primary next-hop
  - Pushes label stack \{80, 9101, 50, BGP-LBL(PE2), VC-LBL(X)\} when forwarding to remote LFA next-hop
- ABR/ASBR pops segment routing labels, swaps BGP label, and pushes LDP label of PE2: \{LDP-LBL(PE2), BGP-LBL(PE2), VC-LBL(X)\}
INTRODUCTION OF SEGMENT ROUTING INTO PRODUCTION NETWORK

- Upgrade to SW release which supports Segment Routing
- Configure network global node SID label space
- Assign node SID index/label for prefixes of loopback interfaces to be advertised in a IGP instance
- Enable Segment Routing in ISIS or OSPF instance
  - Router advertises the Segment Routing Capability Sub-TLV to routers in all areas/levels
  - It advertizes the assigned index for each configured node SID in the new prefix SID sub-TLV with the N-flag (node-SID flag) set; programs ILM.
  - It assigns and advertizes automatically an adjacency SID label for each formed adjacency over an interface in the new Adjacency SID sub-TLV; programs ILM.
  - It resolves received prefixes and programs ILM and primary/LFA NHLFE
- Programmed node SID prefixes can be used:
  - as tunnels by services and as BGP shortcuts
  - for resolving both primary and LFA/RLFA next-hops (IGP prefixes) or remote LFA next-hops only (IGP and LDP prefixes)
Segment Routing vs T-LDP based RLFA

<table>
<thead>
<tr>
<th></th>
<th>Segment Routing RLFA/DLFA/TI LFA</th>
<th>T-LDP RLFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>100%</td>
<td>&lt;100%</td>
</tr>
<tr>
<td>Control Plane Overhead</td>
<td>Low</td>
<td>High (many extra T-LDP sessions)</td>
</tr>
<tr>
<td>Troubleshooting</td>
<td>Simple</td>
<td>Complex</td>
</tr>
<tr>
<td>LBL scaling</td>
<td>1 LBL per node and interface</td>
<td>High (extra labels advertised over T-LDP sessions)</td>
</tr>
</tbody>
</table>

• TI LFA = Transport Independent LFA, has the capability to push more node SID’s and adjacency SID’s for the backup path. Guarantees 100% FRR Coverage
### SEGMENT ROUTING

**Comparison**

- Agreement to be held on the protocol semantics between vendors
- Multicast to be sorted
- LBL stack depth concerns to be analyzed with respect to TE
- TE protocol functionality to be worked out with respect to:
  - Topology discovery
  - TE path placement
  - Service/Flow mapping
  - Path statistics collection

<table>
<thead>
<tr>
<th></th>
<th>LDP</th>
<th>RSVP</th>
<th>SR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overview</strong></td>
<td>Multipoint to point</td>
<td>Point to point</td>
<td>Multipoint to point</td>
</tr>
<tr>
<td><strong>Operation</strong></td>
<td>Simple</td>
<td>LSP per destination/TE-path</td>
<td><strong>Simple</strong></td>
</tr>
<tr>
<td><strong>Dependencies</strong></td>
<td>Relies on IGP</td>
<td>Relies on IGP TE</td>
<td>Relies on IGP + offline TE</td>
</tr>
<tr>
<td><strong>LBL allocation</strong></td>
<td>Local significant per node (interface)</td>
<td>Local significant per node (interface)</td>
<td><strong>Global</strong></td>
</tr>
<tr>
<td><strong>Traffic Engineering</strong></td>
<td>No</td>
<td>Yes</td>
<td><strong>yes</strong></td>
</tr>
<tr>
<td><strong>Scaling</strong></td>
<td>1 LBL per node (interface)</td>
<td>Nx(N-1)</td>
<td><strong>1 LBL per node/local interface</strong></td>
</tr>
<tr>
<td><strong>Fast Reroute</strong></td>
<td>LFA, LFA Policies, RLFA - &lt;100% coverage</td>
<td>Link/Node protection (detour/facility) – 100% coverage</td>
<td>LFA, LFA Policies, RLFA/DLFA - can get to <strong>100% coverage</strong> (better than LDP with RLFA)</td>
</tr>
<tr>
<td><strong>Multicast</strong></td>
<td>mLDP</td>
<td>P2MP RSVP</td>
<td>TBD</td>
</tr>
<tr>
<td><strong>IPv6</strong></td>
<td>Extensions required</td>
<td>Extensions required</td>
<td><strong>Native</strong></td>
</tr>
</tbody>
</table>
AGENDA

1. Overview of Segment Routing
2. Segment Routing in Shortest Path Routing Applications
3. Segment Routing in Traffic Engineering Applications
4. WAN SDN Controller with PCE
5. Feature Planning
Distributed Traffic Engineering

- Objective:
  - Compute disjoint tunnels or disjoint primary and secondary paths for the same tunnel to a destination PE
  - Use tunnels for forwarding service/shortcut packets
- Node computes disjoint paths with following constraints:
  - admin-group and/or SRLG
  - Include/Exclude node-SID/Adjacency-SID
  - No bandwidth constraint
    - LSR nodes have no state of the TE tunnels and cannot update link BW in IGP TE link TLVs
- Each path locally protected using LFA/remote-LFA next-hop when possible
- Alternative uses shortest path with separate IGP instances
  - TE tunnels required if additional constraints other than path diversity exist
Distributed Traffic Engineering (continued)

- CSPF path of a segment routing tunnel can include node SID and/or adjacency SID
- CSPF must minimize label stack overhead
  - Adjacency SID included in path computation only if constraints require it
  - Otherwise, selection of next-hop to downstream node SID left to the local LSR
- LSR node is unaware of segment routing path constraints in PE node
  - Selection of primary next-hop (if no adj-sid in packet’s header) and LFA next-hop to downstream node SID based on SPF calculation
  - User must configure relevant constraints in main SPF and LFA SPF policies and apply them at LSR
- Detection of primary path failure and activation of secondary path requires operating BFD on tunnels
AGENDA

1. Overview of Segment Routing
2. Segment Routing in Shortest Path Routing Applications
3. Segment Routing in Traffic Engineering Applications
4. WAN SDN Controller with PCE
5. Feature Planning
WHY AN EXTERNAL TRAFFIC ENGINEERING CONTROLLER?

• Lack of Network Wide Path Scheduling with Distributed TE
  - PE routers resignal LSP paths asynchronously which leads to RSVP session collision and retry
    - Churn is more frequent for resignalizing in auto-bandwidth Make-Before-Break (MBB)
    - Steady state LSP bandwidth load-balancing over network links not deterministic

• Enforcing Constraints Across Paths Originating/Terminating on different Routers
  - Example: path diversity and bi-directionality

• Simplifying Protection Strategy
  - Router computes local FRR protection for both IP prefixes and TE tunnels
  - external TE controller computes e2e protection for TE tunnel

• Bandwidth Management for Segment Routing TE Tunnel
  - LSR has no state for the tunnel and cannot report link bandwidth usage
STATEFUL PATH COMPUTATION ELEMENT (PCE) SERVER

- Industry has been pushing for a standard based external TE controller
- Original PCE Standard (RFC 5440) intended for stateless operation
  - Router helper function for more complex path computations, e.g., inter-domain TE
- Evolution to stateful operation driven by:
  - Need for path scheduling capability
  - Support of application and demand driven path computation requests
- PCE Based External TE Controller Functions:
  - Discovers TE resources and topology by tapping into IGP Segment Routing updates (IGP, BGP-LS) and collecting link statistics
  - Accepts requests from node or management system for computing paths (PCEP)
  - Communicates with PCEs in other domains/areas for multi-domain/area path computation
  - Regularly re-optimizes and downloads to nodes optimal placement of paths given measured real-time traffic demand, LSP statistics, and network state (Stateful PCE)
  - Supports both RSVP and Segment Routing LSP types
PCE Server Architecture

Stateful PCE Server

- Statistics Database
- Path Computation and Re-optimization Module (CSPF)
- Traffic Engineering Database (TED)
- LSP Path Database
- State and Statistics Collection Manager
- PCEP Session Manager
- Network Topology Discovery

North-Bound Interfaces - TBD

- Streaming/IP FIX/SNMP
- PCEP
- BGP-LS/IGP
BGP-LS FOR TOPOLOGY DISTRIBUTION
draft-ietf-idr-ls-distribution

BGP speaker in each area/AS encodes link, node and prefix info from IS-IS/OSPF LSDB in BGP NLRI
Standard BGP rules apply - best path selection, route reflection, AS PATH, etc.
PCE (or ALTO) server does not need a layer 3 adjacency to each area in each AS to build an overall topology view
CO-ROUTED SERVICE NODE PROVISIONING

PCC INITIATED LSP – USE CASE 1

Step 1
- OSS provisions service on PE nodes:
  - Type of service: VPRN, VPLS, epipe, etc.
  - Local access interface (SAP, spoke-sdp)
  - Tunnel endpoints: remote and local
  - Tunnel type = segment routing, RSVP
  - Tunnel path control = PCE server
  - Path constraints = BW, Co-routed service diversity, bi-directionality

Step 1
- PE makes path computation request to the PCE server (PCC initiated LSP)
- Note: PCE path profile is used to convey the diversity constraint

Step 2
- PCE server computes and downloads the forward and reverse paths
- PE binds service to paths
- PE sends PCE report and delegates path control to PCE

Step 3
- PCE server monitors LSP stats and re-optimizes tunnels and downloads new paths to PE routers (same LSP-ID)
- PE node performs Make-Before-Break (MBB) and move flows to new path.
PCC INITIATED LSP - USE CASE 2
GLOBAL BANDWIDTH OPTIMIZATION

OSS provisions parallel infrastructure tunnels between a pair of PE nodes:
- Tunnel endpoints: remote and local
- Tunnel type = segment routing, RSVP
- Tunnel path control = PCE server
- Path constraints = min/max BW, diversity, admin-group

Step 1
- PE makes path computation request to the PCE server (PCC initiated LSP) server with path diversity constraints among the parallel set of tunnels
  - Note: this re-uses the SVEC object of PCEP as per RFC 5440

Step 2
- PCE server computes and downloads the paths for the tunnel set
  - PE instantiates the LSP and replies with a PCE report with delegation control flag set
  - flow mapper learns the set of LSP-ID values created between endpoints

Step 3
- External flow mapper pushes down the mapping of flow/prefix/destination to the set of parallel tunnels using OpenFlow or XMPP
  - PE instantiates the ACLs to map each flow to the designated LSP-ID.

Step 4
- PCE server monitors LSP stats and re-optimizes tunnels and downloads new paths to PE routers (same LSP-ID)
  - PE node performs Make-Before-Break (MBB) and move flows to new path.
AGENDA

1. Overview of Segment Routing
2. Segment Routing in Shortest Path Routing Applications
3. Segment Routing in Traffic Engineering Applications
4. WAN SDN Controller with PCE
5. Feature Planning
SEGMENT ROUTING AND PCE PLANNING

**13.0R1 - Committed**
- ISIS and OSPF support
- IPv4 shortest path tunnel for use by services and BGP shortcuts
- LFA and Remote LFA for IPv4 SR tunnel FRR protection

**13.0R4 - Not Committed**
- SR IPv4 TE tunnel with delegation of control to PCE server
- LFA and remote LFA for SR IPv4 TE tunnel FRR local protection
- WAN controller with PCE server support
- PCC and PCE initiated SR TE tunnel
- PCC/PCE IGP for topology discovery

**R14 (pre-DR0)**
- Segment Routing Shortest Path: IPv6 with MPLS data plane support: OSPFv3, ISIS
- Segment routing remote LFA for IP FRR
- Segment Routing remote LFA for LDP FRR (LDPv4/v6)
- Segment Routing: Topology Independent LFA (TI-LFA)
- IGP and static route forwarding over SR tunnel
- PCC/PCE BGP-LS for topology discovery
www.alcatel-lucent.com