Segment Routing

Santanu Dasgupta, Consulting Engineer

APRICOT 2014
Agenda

• Segment Routing – Introduction
• Segment Routing – Use Cases
Operator Partnership & Standardization Work at IETF

- Fundamental to the velocity and success
- Significant commitment
  technical transparency, multi-vendor commitment
  beta and PoC

<table>
<thead>
<tr>
<th>Topic</th>
<th>IETF Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>draft-filsfils-rtgwg-segment-routing</td>
</tr>
<tr>
<td>MPLS</td>
<td>draft-filsfils-spring-segment-routing-mpls</td>
</tr>
<tr>
<td>I Pv6</td>
<td>New draft to be submitted</td>
</tr>
<tr>
<td>Use Cases</td>
<td>draft-filsfils-rtgwg-segment-routing-use-cases</td>
</tr>
<tr>
<td>SR/LDP</td>
<td>draft-filsfils-spring-segment-routing-ldp-interop</td>
</tr>
<tr>
<td>TE</td>
<td>draft-shakir-rtgwg-sr-performance-engineered-lsps</td>
</tr>
<tr>
<td>OAM</td>
<td>draft-geib-spring-oam-usecase</td>
</tr>
<tr>
<td>ISIS</td>
<td>draft-previdi-isis-segment-routing-extensions</td>
</tr>
<tr>
<td>OSPF</td>
<td>draft-psenak-ospf-segment-routing-extensions</td>
</tr>
<tr>
<td>FRR</td>
<td>draft-francois-segment-routing-li-lfa</td>
</tr>
<tr>
<td>PCEP</td>
<td>draft-sivabalan-pce-segment-routing</td>
</tr>
</tbody>
</table>
Operators’ Desire from the Network

• Simplicity
  Less numbers of protocols to operate & troubleshoot
  Less numbers of protocol interactions to deal with
  Deliver automated FRR for any topology

• Scale
  Avoid thousands of labels in LDP database
  Avoid thousands of MPLS Traffic Engineering LSP’s in the network
  Avoid thousands of tunnels to configure

• Leverage all services supported over MPLS today (L3/L2 VPN, TE, IPv6)
  Requires evolution and not revolution

• Bring the network closer to the applications

• IPv6 data plane a must, and should share parity with MPLS
  IPv6 SR routing extension header, includes the list of segments
Segment Routing

- **Source Routing**: the source chooses a path and encodes it in the packet header as an ordered list of segments

- **Segment**: an identifier for any type of instruction
  - Service
  - Context
  - Locator
  - IGP-based forwarding construct
  - BGP-based forwarding construct
  - Local value or Global Index
Segment Routing

• **MPLS**: an ordered list of segments is represented as a stack of labels
  a completed segment is popped

• **IPv6**: an ordered list of segments is represented as a routing extension header, see 4.4 of RFC2460
  Type 0 could be used. A new type is proposed to enhance functionality while improving forwarding performance and security
  upon completion of a segment, the pointer is incremented
Segment Routing – Technology Basics
Simple Extension to IGP

• Simple extension to IS-IS or OSPF, automatically builds and maintains Segments
  Nodal Segment – A Shortest path to the related node
  Adjacency Segment – One hop through the related adjacency

• Excellent Scale: a node installs N+A FIB entries
  N = nodal segments; A = adjacency segments
Node Z advertises its node segment (loopback 0)
  e.g. in ISIS its just a simple ISIS sub-TLV extension

All remote nodes install the node segment to Z in the MPLS dataplane
Adjacency Segment

- Node C allocates a local label for CO link segment
- C advertises the adjacency label in IGP
e.g. for ISIS, it’s a simple sub-TLV extension
- C is the only node to install the adjacency segment in MPLS dataplane (FIB)

A packet injected at node C with label 9003 is forced through datalink CO
Combining Nodal Segments to Engineer Path

- **ECMP**
  - Node segment
- **Per-flow state only at head-end**
  - not at midpoints
- **Source Routing**
  - the path state is in the packet header
Combining Nodal & Adjacency Segments

• Source Routing along with the explicit path, stack of nodal and adjacency segments

• Any explicit path can be expressed: e.g. ABCOPZ

• ECMP
  Node segment

• Per-flow state only at head-end
  not at midpoints

• Source Routing
  the path state is in the packet header
Central Optimization with Path Computation Element (PCE)

Centralized Optimization
- find a path meeting the SLA requirement
- encode it as a list of nodal and adjacent segments

Agility and Scalability

Hybrid Central/Distributed CP

ABCOPZ meets SLA. I account the BW. I encode the path as nodal segment to C, adj segment to O, nodal segment to Z

Encoded path for traffic to Z {66, 9001, 65}

Collect network status information

Need 2Gbps from A to Z with SLA

Collect network status information

Congested

API

Visualization/Analytics

Bandwidth Orchestrator

Collector

Programming

Apps

Need 2Gbps from A to Z with SLA

Collect network status information

Encoded path for traffic to Z {66, 9001, 65}

ABCOPZ meets SLA. I account the BW. I encode the path as nodal segment to C, adj segment to O, nodal segment to Z

• Centralized Optimization
  - find a path meeting the SLA requirement
  - encode it as a list of nodal and adjacent segments

• Agility and Scalability

• Hybrid Central/Distributed CP
Use Case: Simple & Efficient Transport of MPLS services: L3/L2VPN

- Efficient packet networks leverage ecmp-aware shortest-path!
  node segment!

- Simplicity
  no complex LDP/ISIS synchronization to troubleshoot
  one less protocol to operate

All VPN services ride on the node segment to PE2

IPv4 over MPLS/IGP
VPN over MPLS/IGP
Internet over MPLS/IGP
PW over MPLS/IGP
IPv6 over MPLS/IGP
Use Case: Simple and Scalable Traffic Engineering

- SR router scales much more than with RSVP-TE
  - The state is not in the router but in the packet
    - Node + Adj vs. Node^2
- No requirement of RSVP-TE protocol
  - And knobs such as LDPoRSVP etc.
Topography Independent LFA (TI-LFA)

draft-francois-segment-routing-ti-lfa-00

- Guaranteed Link/Node FRR in any topology
  even with asymmetric metrics
- No Directed LDP session
- Simplicity
  entirely automated (no need for customization)
- Incremental deployment
  Applicable to LDP and IP primary traffic
  Only the repair tunnel is SR-based
- For networks with symmetric metric & link protection
  No extra computation
  Simple repair stack
  Node segment to P node
  Adjacency segment from P to Q
- Demo available
Benefits

• 100%-coverage 50-msec link and node protection
• Simple to operate and understand
  automatically computed by the IGP
• Prevents transient congestion and suboptimal routing
  leverages the post-convergence path, planned to carry the traffic
• Incremental deployment
  applicable to primary IP and LDP traffic
  only the repair tunnel needs to be SR-enabled
• Demo available
Explicit Post-Convergence Path

- What is the more optimal and natural path upon a failure?
  the post-convergence path

- Why have we never used it before SR?
  the post-convergence path may not be an LFA and hence may loop

- Thanks to SR, we can always use the post-convergence path
  Explicit Post-Convergence (EPC): the non-LFA portion of the path is encoded as an explicit list of segments
Explicit Post-Convergence Path

- Computation leverages proven and existing LFA technology
  intersection of post-convergence SPT with P and Q spaces

- Number of Segments to form the Repair Tunnel
  Symmetric network, link protection: Proven: <= 2 segments to get into Q space
  Asymmetric network or node protection:
    No theoretical bound
    In reality, as we already saw for RLFA, things are much simpler!

- Orange use-case
  100% link protection
  100% use <= 2 segments
  100% node protection (<=4 segments)
  99.72% use <= 2 segments
  0.24% use 3 segments
  0.04% use 4 segments
Applicability on a Large SP Network
TI-LFA for Path Optimality
Applicability on a Large SP Network
TI-LFA for Path Optimality

Comparison of 4 different FRR paths.

TI-LFA FRR path is the optimum because it reflects the post-convergence path.
Dual Plane Core

• Each pop has two core routers
  a blue one and a red one
  typically in different building/locations

• The blue routers are interconnected and form the blue plane
  the red routers are interconnected and form the red plane

• The grey links between blue and red routers have bad metric
  once a packet is within a plane, it reaches its destination without leaving the plane (except if the plane is partitioned)
Use Case: Simple Disjointness in Dual Plane Core
TE Without Bandwidth Admission Control – Anycast Node Segment

- A sends traffic with [65]
  Classic ECMP “a la IP”

- A sends traffic with [111, 65]
  - All the blue routers advertise the same anycast loopback (1.1.1.1/32) with the same anycast nodal segment 11
  - Packets get attracted in blue plane and then use classic ECMP

ECMP-awareness!
Segment Routing – In Summary

• Wide Applicability
• Simple to deploy and operate
• More scalable and functional IP and MPLS
• Agile Wan Orchestration with hybrid centralized/distributed
• Massive operator interest and support
• ISIS/SR demonstrated in Feb 2013
• TI-LFA demonstrated in Oct 2013
• Much more happening! Join the community.
Thank you.

More use cases: See www.segment-routing.net
Segment Routing – The last 12 months

- 2012/Oct: NAG: first presentation
- 2012/Nov: Lead Operator group formed
  - use-cases identified
- 2013/Feb: 5.2.0 beta available
- 2013/Mar: IETF draft released and first Public Presentation
  - MPLS World Congress and IPv6 Conference – Paris
  - ALU and Ericsson
- 2013/Jul: 8 IETF drafts released
  - Huawei
  - JNPR (ISIS/OSPF protocol extension)
- 2013/Oct:
  - TI-LFA FRR beta available
  - SR-TE Central Optimization and Orchestration beta available
  - 12 SR drafts and working-group formed (SpRing)
  - [www.segment-routing.net](http://www.segment-routing.net)
SR can express deterministic non-ECMP path as a list of adjacency segments

A specific non-ECMP path i.e. ABCDEFGHZ can be expressed by a label stack {9001, 9002, 9003, 9004, 9001, 9002, 9003, 9004}

The label stack can be compressed by following –

The use of nodal segment of E as 101 and Z as 109, the same path can be expressed as {101, 109}

Use of Forwarding Adjacency between node B and H with explicit path BCDEFGH and Adjacency Segment ID of 9007, the same path can be expressed as {9001, 9007, 9004}
• A SR head-end router can map the result of its distributed CSPF computation into an SR segment list
• The operator configures a policy on A → Z destined traffic must avoid SRLG1. SRLG1 is link BC
• The SRLG get flooded in the link state IGP. A may implement the policy like the following way –
  Prunes the links affected by the SRLG1, computes an SPF on the rest topology and picks one SPF paths, say ABDCZ
  Translates the path as a list of segments – so ABDCZ can be expressed as two nodal segments {104, 109}
  It monitors the status of the LSDB and upon any change impacting the policy, it either re-computes a path meeting the policy or update its translation as a list of segments
• The network is simple, can respond to rapid changes and is programmable
  
  perfect support for centralized optimization efficiency, if required
The network is simple, can respond to rapid changes and is programmable.

The Central Path Computation and Optimization system (PCE) may have Northbound API’s through which applications can make requests (such as BW 2G from A to Z with max latency of “X” milliseconds).

The router nodes in the network needs to have Programmatic interfaces such as PCEP or I2RS to facilitate southbound programming of the network by the PCE system to reflect changes.
IPv6 Segment Routing
IPv6 Segment Routing

• A segment is represented by a 128-bit IPv6 address
  Prefix segments, Node segments, Adjacency segments, as defined in the segment routing architecture, are identified through IPv6 addresses.

• A segment identifies a forwarding instruction such as: Service, Context, Locator, IGP-based or BGP-based forwarding construct, others…

• Terminology
  Segment List: list of segment forming the path
  Active segment: segment currently used by the packet
  Next Segment: segment following the active segment in the segment list
IPv6 Segment Routing

- Segment Routing defines a new routing header type: Segment Routing Header (SRH)
  Described in draft-previdi-6man-segment-routing-header
  A new Routing Header type is proposed in order to enhance functionality, improve forwarding performance and address security
  The SRH contains
    - The segment list representing the path
    - A pointer identifying the next segment
    - Policy information (ingress, egress SR nodes)
    - Flags
    - HMAC authentication

- SR-IPv6 use cases:
  - Network Resources optimization (TE)
  - Service Chaining, in conjunction with draft-quinn-sfc-nsh
    - SR for path forwarding
    - NSH for service metadata

- Running code under network operators evaluation
SR-IPv6 Example: Network Resources Optimization

- Desired behavior: Optimize resources usage:
  - Route packets from A to Z across explicit path: B-C-D-H-I-J-K
- Scale infrastructure by allowing any number of explicit paths
  - Do not maintain state across the path

- Packet classification at ingress:
  - SRH is inserted with segment list representing the path
- Packet is forwarded according to active segment
SR-IPv6 Example: Service Chaining

- **Step-1: packet classification**
  - Insertion of SR header with network path information including Service Segments

- **Step-2: Segment Routing to first segment**
  - Plain IPv6 in the core
  - First segment is a service segment

- **Step-3: Segment Routing forwarding through service chains**
  - SR header update by each segment endpoint

- **Step-4: Exit the SR domain**
  - After last segment is inspected, and according to flags, SR header is removed
  - Packets is forwarded towards destination

**Packet Classification and service chain construction at ingress (e.g.: through a SDN/WAN Ctrl)**

**Insertion of SR header**

**Packet goes to the last segment and SR header is removed.**

**After processing, SR header is updated and packet continues its path to next service either through core or directly.**