MPLS Segment Routing
Driving a modern approach to MPLS transport

For many years, one of the biggest challenges in network design has been effectively managing traffic flow end-to-end across the network. Stated more specifically: How is traffic intelligently classified and path engineered throughout the network? Furthermore, how can this traffic be classified into differentiated levels of service, without adding unnecessary complication to management, control plane, or data plane state in this critically important part of the network?

Now, modern cloud and service provider networks in the SDN era require even more flexible control on steering of their traffic flows, at a much greater scale. With globally optimized SDN controllers utilizing parameters like distance, available bandwidth, latency, cost of long-haul links and business logic into their path computation algorithms, there is an increased need for flexible, intelligent source routing solutions. More and more operators have made a concerted effort to simplify their networks and eliminate the dependency on a wide set of protocols to achieve their goals. As native IPv6 deployments increase, there is a need for a native solution for IPv6 networks and IP backbones that provides the same capabilities as current MPLS networks.

This seems like an impossible set of opposing requirements to reconcile. Traditional MPLS implementations did address some of the original challenges, but they came with a complex array of protocols, suffered from scaling challenges and had rigid vendor specific path computation algorithms for Traffic Engineering (TE). They also did not provide native IPv6 support across the protocol stack.

Segment routing is the technology that elegantly addresses this varied set of requirements across many different networking domains: datacenter, content delivery networks and metro networks, to name a few. The basic philosophy of segment routing is to start with bare minimum distributed routing protocols and leverage the best out of them; the simplicity, scale, resiliency and flexibility of these protocols combined with the increasing demand for a centralized, global control approach make for end to end application-based traffic control.

This paper will discuss MPLS segment routing, Arista’s solutions and the use cases where they bring value compared to traditional MPLS protocols.
Segment Routing Operation
Segment routing divides the network into “segments” where each node and link could be assigned a segment identifier, or a SID, which gets advertised by each node using standard routing protocol extensions (ISIS/OSPF or BGP), eliminating the need to run additional label distribution protocols.

These SIDs are either globally unique SIDs or locally significant SIDs. Figure 1 illustrates these segments, where the globally significant node SIDs are dark green, and the locally significant adjacency SIDs are light green. The globally significant SIDs must be unique and can be assigned an absolute value from the Segment Routing Global Block (SRGB), that is advertised as a base value and label range. Alternatively, the SID value can be derived by each node using an index value and adding this to the SRGB base value.

There are three major variants of a global SID:
- **Prefix SID** - A segment identifier associated with a prefix.
- **Node SID** - A segment identifier allocated to a loopback that identifies a specific node. This a commonly used sub-case of a prefix SID
- **Anycast SID** - A shared segment ID assigned to a loopback shared by a set of routers. Again a commonly used sub-case of a prefix SID
- **Proxy SID** - A segment identifier associated with an IS-IS prefix originating from a router that does not support IS-IS SR

All nodes in the domain install the same value for the prefix SID. This notion of globally unique SIDs plays an important role in reducing the data plane state explosion across the nodes. This is illustrated below in the MPLS data-plane example, where the unique prefix SID index value for each node is flooded throughout the IGP domain. Using the same SRGB range, each node derives the same node SID label value for the remote nodes. The forwarding operations of push, swap and pop is also shown below.

![Figure 1: Segment Routing Node and Adjacency SIDs](image1)

![Figure 2: SR Prefix or Node SID Signaling & Forwarding Using IGP](image2)
Segment routing also preserves the notion of locally significant SIDs to allow for finer grained traffic engineering. Adjacency SIDs are only locally significant to each node and are advertised and installed only on directly connected neighbors, thus identifying a specific adjacency link. Adjacency SIDs are similar in nature to LDP assigned labels, which are also only locally significant and installed directly adjacent LDP peers. Adjacency SIDs can be used in segment routing traffic engineered paths as an entry in the order list, or label stack, specifying a specific link to traverse in the path.

*Note: There are other SIDs defined for segment routing such as Peer SIDs (BGP only for EPE) and binding SIDs for interworking with LSPs signaled via RSVP.*

The simplest way to leverage the distributed SR label state in the network is for traffic to follow IGP shortest path, while leveraging ECMP through the network. In this case, an ingress LER will simply push an SR label (derived from node SID) onto the incoming traffic corresponding to the next-hop towards that destination. Alternatively, if any traffic engineering is desired, the ingress LERs can then use the SIDs, encoding them in the packet header to define a path through the network, either using a single SID to the destination node or as an ordered list of SIDs defining an alternate path from the IGP shortest path.

**Comparison to Existing Label Signaling Protocols**

One of the fundamental premises of segment routing is to eliminate the need for additional signaling and label distribution protocols, such as LDP and RSVP-TE, and leverage the routing protocol itself (IGP, BGP) for label distribution. Let us take a look at how this compares with LDP and RSVP-TE.

**Comparing Segment Routing to LDP**

Segment routing label distribution in the IGP and LDP label distribution are similar in the sense that they are both “plug and play”. Hence, they are easy to configure, labels are automatically advertised amongst the routers when an adjacency peering is formed and there are no label switched paths (LSPs) to manually configure. Additionally, both LDP and SR form stateless MP2P LSPs derived automatically for each node. It is this simplicity in deployment that made LDP popular and is one reason segment routing is being adopted so readily.

This, however, is where the similarity ends. Fundamentally, LDP relies on IGP state, and features like the LDP-IGP sync were introduced so that they always remain in sync, reducing the possibility of traffic blackholing. With label distribution directly done by IGP, this issue is now moot. In segment routing, nodes and prefixes have globally unique labels assigned throughout the domain, whereas in LDP, these labels are locally significant, assigned a unique value at every hop. Global labels significantly reduce data plane state at every network hop. Depending on vendor implementation choices, an implementation with LDP independent LSP control mode can proliferate unnecessary control and data plane state, resulting in scaling challenges.

SR also enables TE capabilities and has support for V6 address family derived inherently from the routing protocol, but CR-LDP was never deployed for traffic engineering. V6 support in LDP is generally lacking from both the implementation and deployment fronts.

**Comparing Segment Routing to RSVP-TE**

Traffic engineering and Fast re-route (FRR) are the main reasons that RSVP-TE is widely deployed in networks today. It is the improved traffic engineering scalability and flexibility provided by segment routing while addressing some of the new SDN requirements that are driving its adoption. Traffic Engineering with RSVP-TE enables the following:

a. Computation of paths using constraints such as bandwidth, shared link risk groups and signaling of such explicit paths, which don’t necessarily have to follow the IGP shortest path

b. Bandwidth reservation on the computed path and tracking of available link bandwidth, allowing for improved utilization on the long-haul links

c. Fast link and node protection on failure using switchover to pre-computed backup paths using FRR capabilities
However, many deployments use statically configured explicit paths using offline path computation tools with no bandwidth reservation. In many deployments, FRR alone is the reason to use RSVP-TE. RSVP-TE networks are built using full mesh of point-to-point TE tunnels, which install control and data plane state for every tunnel that originates, transits or terminates at that network hop. There is also inherently no leveraging of ECMP in the network. Additionally, upon any failure in the core, the LSPs must be re-sigaled, causing churn and control-plane state change on every device running RSVP-TE. These scalability issues, along with limited coarse-grained flow control, have plagued RSVP-TE.

Segment routing addresses some of these challenges by taking an SDN-centric approach to traffic engineering. Instead of encapsulating the source routed explicit path in the control plane, it defines the traffic engineered path as a list of SIDs encapsulated in the data plane. Only the edge or ingress routers (LERs) impose the traffic engineering path, as a stack of MPLS labels, thereby leaving the core free of per flow control plane state.

Any number of constraints or attributes can be used by the centralized controller to define the TE path, not just the limited set of attributes offered in RSVP-TE. To that end, segment routing offers improved scalability and provides the granular control required for traffic engineering in modern networks.

Segment Routing Advantages

Here is a table that summarizes the key functional, operational and scale differences between LDP, RSVP-TE and SR.

<table>
<thead>
<tr>
<th></th>
<th>LDP</th>
<th>RSVP-TE</th>
<th>SR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>MP2P</td>
<td>P2P</td>
<td>MP2P</td>
</tr>
<tr>
<td>Operation</td>
<td>Simple</td>
<td>Difficult</td>
<td>Simple</td>
</tr>
<tr>
<td>Separate Label</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Distribution Protocol</td>
<td>Relies on IGP</td>
<td>Relies on IGP extensions</td>
<td>Relies on IGP</td>
</tr>
<tr>
<td>Dependencies</td>
<td>Label Allocation</td>
<td>Locally significant</td>
<td>Locally significant</td>
</tr>
<tr>
<td>MPLS ECMP</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Traffic Engineering</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>TE Scale</td>
<td>N/A</td>
<td>Medium/Low</td>
<td>N(N-1)</td>
</tr>
<tr>
<td>Fast Reroute</td>
<td>Partial LFA (&lt;100%)</td>
<td>Yes Node/Link Protection</td>
<td>Yes TI-LFA</td>
</tr>
<tr>
<td>Multicast</td>
<td>Yes mLDP</td>
<td>Yes</td>
<td>P2M LSP</td>
</tr>
<tr>
<td>IPv6</td>
<td>Limited Extensions Required</td>
<td>Limited Extensions Required</td>
<td>Native</td>
</tr>
</tbody>
</table>

Table 1: Comparison of SR Operation and Capabilities Compared to LDP & RSVP-TE
It is normally the simplest solution that elegantly addresses an issue, or set of issues, that is most successful. The key advantages of segment routing include:

**Single Protocol (Label advertisement in IGP)** - The main benefit with this approach is that an additional label distribution protocol is eliminated from the requirement, making the implementation simpler and reducing risk in day-to-day operations and management of the network.

**Simplicity** - Simplified configuration and more plug and play capabilities similar to LDP is one of the huge benefits of segment routing. The fact that the labels are distributed in the routing protocol itself (IGP or BGP) and that the core has no requirements to maintain per flow state makes segment routing simpler to operate the network at scale.

**ECMP, Macro TE and Anycast SID** - New requirements are broadening the applicability of MPLS data plane beyond just traditional WAN networks, and these networks inherently have more ECMP. The ability to perform MPLS ECMP routing is very limited with traditional MPLS traffic engineering approaches. There is a need for flexible traffic engineering, including macro-TE, where an operator can group a set of nodes under a common prefix SID, called anycast SID. Segment routing offers ECMP routing capabilities in non-traffic engineering domains similar to current LDP MPLS networks. With the concept of the global Prefix/Node SID, anycast SID, traffic engineering can be realized while still ECMP routing between nodes, or sets of nodes. The capability to load-balance across the label switched network between nodes, or to a set of nodes, affords great flexibility as macro traffic engineered paths can be defined that make use of all ECMP links between nodes, or set of nodes. These can remain valid even upon a node failure when using an Anycast SID in the path. This flexibility could simply not be realized if using traditional traffic engineering methodologies.

**Scalability** - Segment routing offers a solution that can scale far beyond current traffic engineering solutions. In segment routing, there is little to no resource impact on the core nodes in the network and no per tunnel control or data plane state in the core of the network. This affords huge scale for full-mesh end to end MPLS (overlay networks), where the traffic engineered paths are only programmed on the edge of the domain.

**MPLS Data-Plane Support** - One of the biggest enablers for the adoption of segment routing in present-day networks is the support for MPLS. MPLS data plane is not only well understood and deployed but offers some inherent attributes like source routing with an MPLS label stack and support for MPLS-oGRE to tunnel MPLS over IP fabric. Plus, MPLS forwarding capabilities are prevalent in commodity merchant silicon chips, making it very attractive to cloud and service providers.
IPv6 Support - Segment routing was conceived as a technology for IPv6 networks and as such, natively supports IPv6, derived from the routing protocol V6 address family support, which is consistent with IPv6 Specification (RFC2460).

Interworking With Existing Label Distribution Protocols - Segment routing can easily interwork with existing LDP and/or RSVP-TE deployments. With LDP, a simple mapping of SR to LDP labels is configured, as illustrated below. Likewise, with RSVP, a binding segment identifier is configured. This makes inter-working between SR ad non-SR domains simple and scalable, allowing a migration path for operators dealing with existing brownfield networks.

Enhanced Traffic Protection and Fast Reroute in Segment Routed Networks - Segment routing provides Topology Independent Loop Free Alternate (TI-LFA) as its main solution for FRR. Compared to LDP, there are no targeted LDP sessions to manage and maintain with segment routing. Unlike LDP, it provides failure coverage even in ring or complicated partial-mesh topologies. One can even argue the TI-LFA provided by segment routing is superior to the one-to-one or facility backup link protection provided by RSVP-TE. In RSVP-TE, if a link fails, the protecting LSP will re-route traffic to the protected next-hop, which often tends to be the sub-optimal path. This does not happen with TI-LFA, as the backup path for a link always just follows the optimal IGP best path, assuming the failed link is out of the topology.

Arista MPLS SR Solution

The key pillars of the Arista MPLS SR solutions are Arista's Extensible Operating System (EOS) and Arista's R3-series Universal Leaf and Universal Spine platforms with FlexRoute capabilities.

Arista EOS

EOS is perfectly suited to meet the demands of a segment routing network as it is built on the strong foundations of a multi-process state-sharing architecture with modularity, programmability, fault containment and resiliency as the core software building blocks. EOS combines these attributes with key infrastructure innovations including EOS SDK, Go programming language, NetDB for improved route scale and convergence, support for Docker containers and real-time state streaming and analytics for live monitoring and historic forensic troubleshooting. These strengths allow leveraging of Arista platforms in various network roles beyond the traditional cloud and service provider datacenter environment, while ensuring seamless customer experience and high software quality. In addition to a rich L3 routing stack (BGP, ISIS, OSPF, PIM, etc.), EOS also has support for an increasing set of MPLS control plane protocols including static MPLS, LDP, ISIS-SR, BGP-LU and BGP SR-TE.

Built with a modern MPLS infrastructure in mind, EOS has complete flexibility in label management, unlike traditional MPLS routing platforms.

```
router(config)#mpls label range ?
bgp-sr Specify labels reserved for BGP SR global segment identifiers (SIDs)
dynamic Specify labels reserved for dynamic assignment
isis-sr Specify labels reserved for IS-IS SR global segment identifiers (SIDs)
l2evpn Specify labels reserved for L2 EVPN routes
srlb Specify labels reserved for SR local segment identifiers (SIDs)
static Specify labels reserved for static MPLS routes
```
This allows an operator to carve out the label allocation ranges in EOS to suit the needs of their specific deployment. For example, an operator enabling SR in the Metro still looking to interoperate with LDP in the WAN may want to set aside equal chunk of global and dynamic labels. On the other hand, a cloud provider using an SDN controller for TE may want no labels in the dynamic range and may instead choose to allocate static and SR labels only. In case of segment routing deployments, this flexibility allows an operator to match SRGB on an EOS platform to match that of any other vendor OS in the network. This enables homogenous SRGB in the network, which is recommended and desirable.

```
router(config)#mpls label range isis-sr 16000 8000
```

**Arista R3-series platforms**

![Arista R-series Universal Leaf and Universal Spine with Arista FlexRoute](image)

Arista's 7280R3, 7500R3 and 7800R3 400G Universal 400G platforms with FlexRoute and AlgoMatch, powered by Arista EOS®, are key components of the MPLS segment routing solutions. Built using merchant silicon, these platforms continue to deliver improved scale and bandwidth while reducing price per port. At the same time, these platforms have differentiated hardware and platform software capabilities well aligned to address the needs of segment routing deployments in the years to come. One of the key requirements that segment routing drives that is different from traditional MPLS is the encapsulation of the source routed path information in the data plane with an MPLS label stack. This requires segment routing capable platforms to a) support a larger number of simultaneous label pushes to encapsulate the hop information and b) support good ECMP hashing capabilities at the transit hops. Moreover, the use of MPLSoGRE is very prevalent in cloud networks where the label stack is usually encapsulated at the host, but traffic needs to traverse a DC IP Fabric before getting to the MPLS core. Class Based Forwarding (CBF) is essential to segregate different classes of traffic into different SR tunnels. Some of the key Arista's R3-series platform attributes include:

- Support multi-label MPLS push support
- Advanced ECMP and LAG hashing
- IPoMPLS and EoMPLS
- MPLSoGRE
- Class Based Forwarding
- No restrictions on forwarding incoming packets with large MPLS label stacks
Segment Routing solution options

There are four main approaches that a customer can take for a segment routing solution on Arista R3-series platforms. Depending on which stage of segment routing adoption phase you are in, one or more of these approaches may be suitable.

1. Static MPLS Push and Next-hop Groups Solution

An MPLS network is running a routing protocol with SR extensions such as ISIS-SR for label distribution. If TE is desired for traffic traversing this network, then at the ingress LER, one easy option is to configure a route pointing to a label stack via the CLI or use eAPI in EOS to push a label stack. Multiple tunnels can also be configured to load balance traffic. Class-based forwarding service policy can be applied to segregate different types of traffic on to different tunnels. This approach is fairly simple to develop and manage and a good first step in the direction of using segment routing. Figure 9 illustrates pushing a label stack for a service prefix 11.1.0.0/16.

![Figure 9 ISIS-SR Solution With Static Label Stack Push](image1)

Figure 10 illustrates the configuration for a class-based service policy. This can be extended or combined with a prefix and port matching policy for more granularity.

![Figure 10 ISIS-SR Solution Using MPLS NHG For Class Based Forwarding](image2)

2. Controller Based Solution Using EOS SDK

Arista EOS is built on the foundations of programmability and with the philosophy of allowing customers to develop their own software on this platform. Using EOS SDK, customers can develop their own customized EOS applications in C++ or Python. This EOS development model allows third party applications to be first-class citizens of EOS along with other EOS agents. One of the highly leveraged use cases of EOS SDK is for MPLS Segment Routing deployments, where the custom agent using EOS SDK libraries programs SR multi-label stack tunnel mappings using MPLS nexthop-groups at the LER. However, with this approach, customers can develop their own control plane in the agent, interact with their controller to accept multiple ECMP tunnels or CBF attributes, build health checks and liveness engine for tunnels and notify the controller of failures. The EOS SDK integration provides more functionality, resiliency and a deeper integration with EOS than using eAPI, because the EOS SDK client is running as a process in EOS, interacting directly with other EOS agents and reacting to state changes. The controller can learn the ISIS topology updates as well as all the SIDs (node, adjacency, prefix) either by running a passive ISIS listener over ISISoGRE peering with any node in the ISIS topology. EOS also offers rich real-time state streaming telemetry solutions to export various states from the platform to an external analytics engine. This allows customers to enhance their TE decisions and enables troubleshooting to changing conditions in the network.

![Figure 10 ISIS-SR Solution Using MPLS NHG For Class Based Forwarding](image2)
3. BGP-LU with ISIS-SR solution

The main enhancement available with this option compared to other options is that an external controller can signal to the LERs a tunnel end point or a service route with an SR traffic engineering label stack using BGP labeled unicast (RFC 3107). This solution allows dynamic signaling of the SR tunnels using standards based BGP-LU protocol instead of either a static tunnel config or a custom agent. EOS allows the operator to separate the signaling of tunnels (BGP-LU) from the signaling of service route prefixes (BGP IPv4 or BGP IPv6) pointing to the tunnels. This enables the solution to scale better due to the resulting indirection. This way, an underlying SR tunnel route or nexthop change can be executed in a single pass, without having to re-signal and re-program all the service routes pointing to it. The remaining aspects of this solution for label distribution, topology and SID discovery are identical to option 2.

4. BGP-LU with ISIS-SR solution

Segment Routing Traffic Engineering Policy (SR-TE) aka SR Policy makes use of Segment Routing (SR) to allow a headend to steer traffic along any path without maintaining per flow state in every node. A headend steers traffic into an “SR Policy”. EOS adds support for SR Policy for the MPLS dataplane (SR-MPLS) for Type-1 SR Policy segments with BGP and locally configured policies as sources of SR Policies.
Common MPLS Segment Routing Use Cases

In the following section, four common compelling use-cases for segment routing will be examined, each detailing the specific benefits segment routing brings.

Cloud WAN

To handle the ever increasing inter-DC traffic due to distributed clusters and changing application traffic patterns, most cloud providers have built a Cloud WAN to interconnect their DCs with Petabytes of traffic traversing it today and growing. WAN links also tend to be expensive and there is a need to drive a high degree of utilization on those links. Cloud networks need fine-grained control for traffic engineering based on a holistic global view of their end-to-end network. They need the ability to compute an optimal traffic engineered path based on the global topology, distance, bandwidth availability, congestion conditions, traffic type, latency sensitivity and business logic. This is where segment routing using an external SDN controller to drive Traffic Engineering decisions around path computation and re-optimization is a very attractive solution and provides the perfect paradigm for intelligent software-driven source routing.

With the complexity of per flow TE state and additional protocols eliminated, the physical network can now be designed for high performance routing and switching, running distributed IP routing protocols and providing rich network telemetry in order to support this new software driven traffic engineering approach. Different providers have taken different approaches to rolling out segment routing deployment in their WAN. Some have built islands around their current LDP/Rsvp-TE networks requiring SR interworking with these traditional protcols, while others have built SR networks parallel to their traditional TE networks. Another approach is enabling both SR and RSVP-TE simultaneously on the same network and segregating the traffic that is carried on these tunnels.

Content Distribution Network (CDN)

CDN providers require software driven traffic engineering to select the optimal egress path from the CDN POP (DC) towards their content consumers. Similar to the Cloud WAN use case, this decision is driven by various constraints like cost of interconnect links, geography, traffic type and the location of the end consumer.
While some CDN providers have stuck to an all IP solution by absorbing all BGP paths in the controller, some others have adopted the MPLS label encapsulation to identify and enforce the selection of egress point on the Edge. Since the network nodes below the Edge are IP today, several customers will choose an MPLSoGRE encapsulation to steer traffic to a specific interface on an Edge routing device. This mapping is often statically configured using CLI or APIs. With segment routing enhancements to BGP or using BGP-LU, the MPLS label mapping to the egress can be dynamically signaled to the controller to scale this operation more effectively. Furthermore, large CDNs with numerous exit edge routers tend to need scale-out leaf-spine design tiers to connect these Edges to the cache nodes. To completely eliminate the need for any IP based forwarding in the intermediate nodes and IP route scale while keeping the overhead of running MPLS low, MPLS segment routing with either IGP or BGP offers an elegant design within such CDNs.

**Next Gen Telco NFV Cloud**

Telco providers want to build cost-effective, scale-out, software driven cloud data centers to offer NFV services via their central offices. The Telco NFV DCs interconnected via their MPLS backbone look very similar to cloud provider distributed DCs interconnected by an inter-DC WAN. Telco NFV DCs are essentially cloud-like leaf-spine designs which need to be easy to operate, leverage ECMP, offer rich telemetry and allow for traffic engineering by leveraging a globally optimized SDN controller.
Due to the inherent nature of services like Ethernet access, L2 and L3 VPN and subscriber management, the Telco NFV cloud may have a need for technologies like EVPN for network overlay. Using segment routing in the underlay offers a simple, easy to manage and elegant solution for the Telco NFV cloud.

**Next Gen Metro Transport Solution**

Metro is another domain that can benefit significantly from using a segment routing transport solution. In this case, the demand is not driven as much for traffic engineering, but a) for segment routing to solve the “Ring” topology protection problem and b) to get a better MPLS redundancy and ECMP solution, as more metro aggregation networks resemble L-S networks. Segment routing as a transport solution is very compelling as it can natively provide the fast reroute topology independent loop free alternate (TI-LFA) path for any failure. Additionally, Anycast SIDs can be easily used to provide a simple elegant solution for a resilient gateway out of the rings, as shown below.

Now, metro providers have a simple solution that addresses the main issues that exist in this domain in the ring access. Segment routing also provides a simple way to integrate the rings into existing LDP or RSVP-TE core environments using the SR-LDP mapping functionality or a binding SID.
Summary
New requirements emerging from changing traffic engineering needs driven by new business problems, as well as an overarching desire to simplify network design is driving the adoption of MPLS segment routing. It is a technology that enables an architecture representing the perfect balance between a distributed and centralized control plane. Arista’s R3-series platforms powered by EOS offer differentiated capabilities and support for various solutions with segment routing to address various use cases in the cloud, content and service provider networks.