MASTER IN INTERNETWORKING

MINT 709 - CAPSTONE PROJECT

Segment Routing Protocol Analysis

By

Byron Arévalo Torres

Mentor

Juned Noonari

Edmonton, Alberta

Abstract:

The objective of this project is to make a deep dive into the principles, uses cases, advantages, possible drawbacks, and some practical implementation aspects of Segment Routing (SR) technology.

This project is going to be divided into two main parts, theoretical and practical. For the first part, IETF standards, workshops, and books about Segment Routing will be analyzed in order to create a strong theoretical basis about concepts and applications using SR.

The starting point will be a brief and inevitable comparison with current MPLS signaling protocols because one of the main reasons to create SR was simplified MPLS' control plane. Then SR's control plane and data plane will be studied in detail, so concepts and principles about IGP (OSPF and ISIS) segments and BGP segments will be analyzed. Concepts about using the current MPLS data plane to run SR will also be explained in this section.

The next section will be about more complex topologies such as Multi-area or multi-level Segment Routing operations, SR in Seamless MPLS architecture, and applied to the Data Center. Mapping Server functions as a component of SR and LDP internetworking will also be analyzed.

Redundancy in SR will be studied through Topology Independent Loop-Free Alternate (TI-LFA) technique, all computation principles, protection options and micro-loop avoidance mechanisms are going to be detailed in this section.

On the other hand, in the practical part, several scenarios will be created using simulators, or real hardware if that support Segment Routing. Topologies, configurations, and captures of results will be included in this part to show how this protocol works in some use cases.

Table of Contents:

Table of Figures:

Table of Tables:

Table of Equations:

1 Introduction

Since MPLS started to be deployed in different networks, Label Distribution Protocol (LDP) and Resource Reservation Protocol (RSVP) has been used as control plane protocols for label distribution and Traffic Engineering (TE). LDP is simple to operate because it builds Label Switched Paths (LSP) automatically and using LFA techniques could achieve a Fast Reroute (FRR) protection closer to 100%, but that does not have TE tools. On the other hand, RSVP needs all LSPs being explicitly created, but its FRR technique reaches almost 100% of protection, and its TE tools are robust. Most of the networks have to use both protocols at the same time to try to accomplish all the redundancy and flow steering requirements, but RSVP-TE solutions do not scale well when a granularity is required, and LDP with RLFA techniques for FRR are too complex to be implemented in real cases [1].

On the other hand, Segment Routing is a source-routing architecture where a node can specify a forwarding path different to the main shortest path that a packet will traverse [2] [1]. SR was designed to address some of the issues mentioned in the previous paragraph without the addition of any other signaling protocol, so MPLS tunnels can be created using only the current IGP protocol, and a backup path can be created in any topology without additional signaling. Note that Segment Routing can be used as an alternative solution to LDP or RSVP protocols in an MPLS transport network, but at the same time SR can co-exist with both protocols in the same MPLS domain or multiple domains, environment using seamless MPLS.

Control plane in MPLS has proven through the years been not scalable and very complex, so first implementations of Segment Routing were focused on superseding control plane in existent MPLS networks and applying SR to mature and well-deployed MPLS data plane [3]. This application is very direct because a segment becomes a label and a list of segments becomes a stack of labels added to a packet header.

Next some notions about Control Plane and Data Plane and different Data Bases used by these planes in MPLS and Segment Routing will be detailed, then a brief explanation of MPLS control plane protocols and their evolution is going to be included.

1.1 MPLS Basics Concepts

In order to improve efficiency, a router splits its functions into two planes, a Control Plane and a Data Plane. These functions are usually assigned to different hardware components within the router.

1.1.1 Control Plane

All communications with other routers via protocols takes place here, and one kind of Control Processor is the hardware component in charge of these functions. Some databases reside here, so when a link-state protocol as OSPF or IS-IS is configured in two neighbor routers, they start to form adjacencies and exchange routing information updates in order to synchronize their topology databases and build their Routing Information Base (RIB). In Figure 1-1 RIB is build using routing updates between two routers [4].

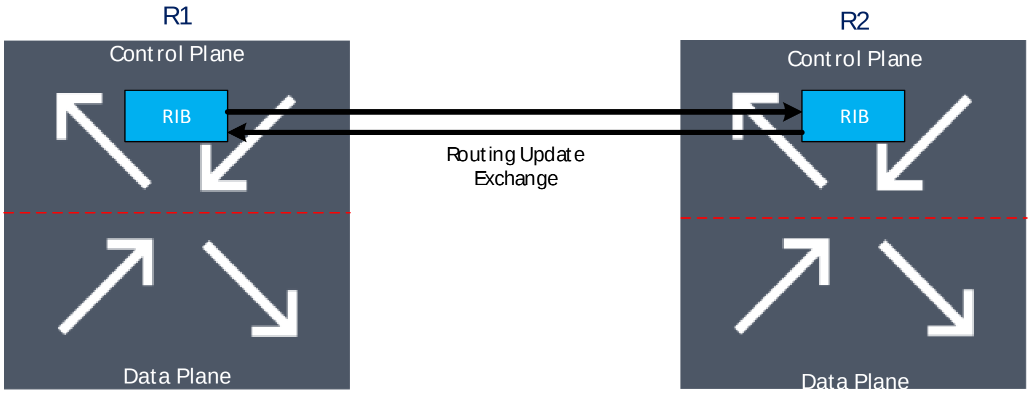


Figure 1-1: RIB creation

Inside RIB might exist all routes from different routing protocols enable in the network, so the best route has to be calculated to install it inside of the Route Table or Routing Table. If several routes towards the same node and coming from the same protocol exist, a lower metric value will be used to choose the best route. A lower metric is preferred because IGP link-state protocols use the Shortest Path First (SPF) algorithm which is a function of the physical bandwidth, so a higher bandwidth means a lower metric. In case several routes to the same destination were learned from different protocols, a preference value is used to choose the best route to be installed in the Routing Table. The preference is a predefined value for each routing protocol, lower is better as in metric. To sum up, the best route between the same protocols is the one with the lower metric, and the best route between different protocols is the one with a lower routing protocol preference. Figure 1-2 shows the relation between RIB and Route Table [4].

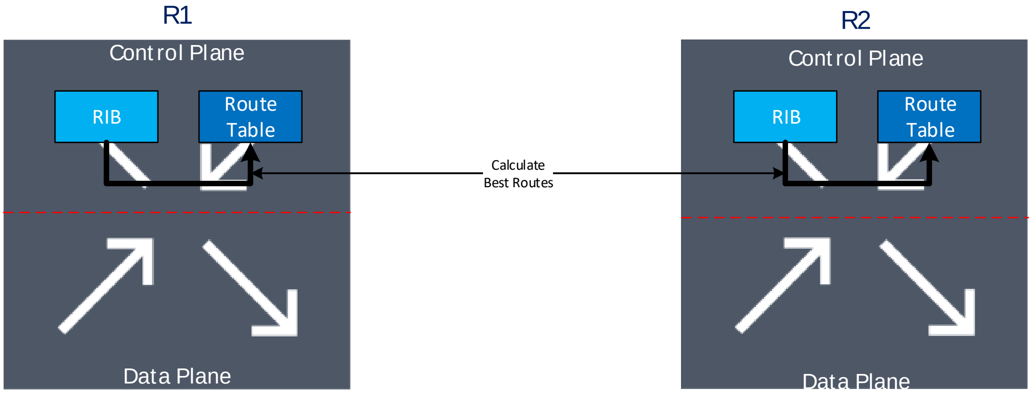


Figure 1-2: Route Table process

After an IGP protocol has converged that means the routing table is created and FIB is installed in Data Plane (process explain in the next point), a label exchange protocol such as LDP or RSVP could establish sessions with its neighbors and start to exchange label bindings for creating the Label Information Base or LIB. This process will be detailed in point 1.2.1, but Figure 1-3 shows how the label bindings exchange creates LIB [4].

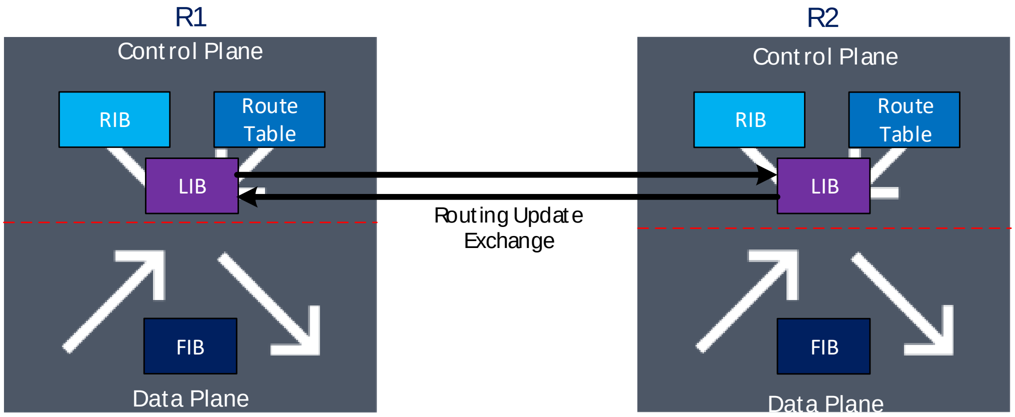


Figure 1-3: LIB creation process

1.1.2 Data Plane

Data packet processing and forwarding take place here; these functions are performed by the line cards in a routing distribute architecture. In order to execute the data packet forwarding function the Data Plane uses the Forwarding Information Base (FIB), which is a lightened copy of the Route Table without extra info related to the Control Plane. Every line card needs an identical copy of the FIB and that copy needs to be synchronized and up to date. Figure 1-4 shows the relationship between the routing table and the FIB [4].

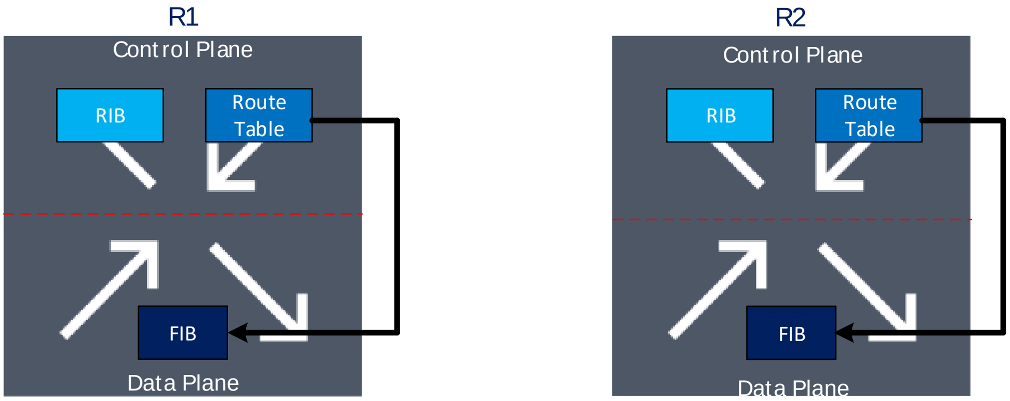


Figure 1-4: FIB and Route Table replication

The same process required for IP traffic to create FIB from the routing table is used for an MPLS protocol to create a Label Forwarding Information Base (LFIB) from LIB. In this process all redundant label entries which reside in LIB are not going to be installed inside LFIB, only the best entry will be installed there. Further explanation about this process will be developed in point 1.2.1. Figure 1-5 shows how LIB and LFIB are related [4].

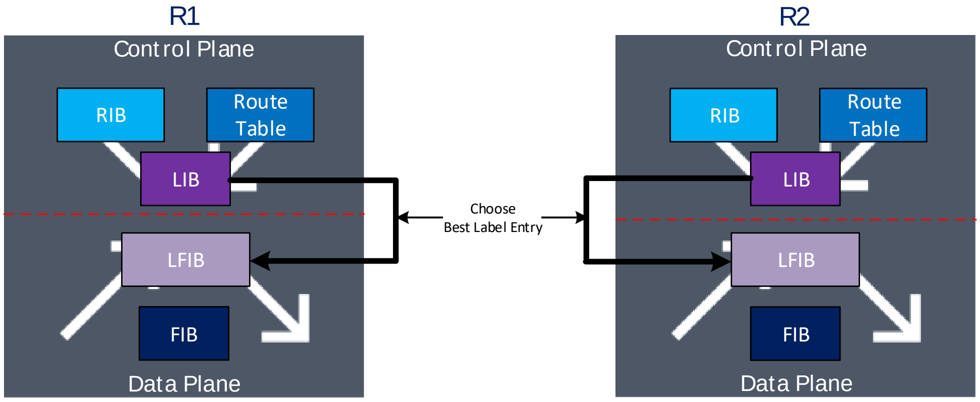


Figure 1-5: LFIB creation from LIB

1.2 MPLS Control Plane Evolution

The first step to enable MPLS in a network is to have a converged IGP protocol like OSPF or IS-IS, that is needed because routing information to Router-IDs loopbacks is used by Label Distribution Protocols, LDP and RSVP. Both protocols need to know how to reach those Router-IDs to build its Label Information Base (LIB) which is created with labels assigned to each router loopback.

1.2.1 Only LDP

LDP was the first standard protocol to be defined [5] and implemented in the first MPLS networks to distribute label information, that is very simple to configure and operate because router interfaces only have to be added to create new adjacencies with all neighbors, and LDP use Interior Gateway Protocol (IGP) short path information to create LSPs. However, LDP never gets Traffic Engineering mechanism and in initial stages that did not have traffic restoration techniques, so LDP relied completely on IGP to restore traffic after a path failure. Figure 1-6 shows as LDP adjacencies are enabled over the same interfaces were IGP adjacencies were created, that permit labels distribution to all routers in the network once LDP sessions are established.

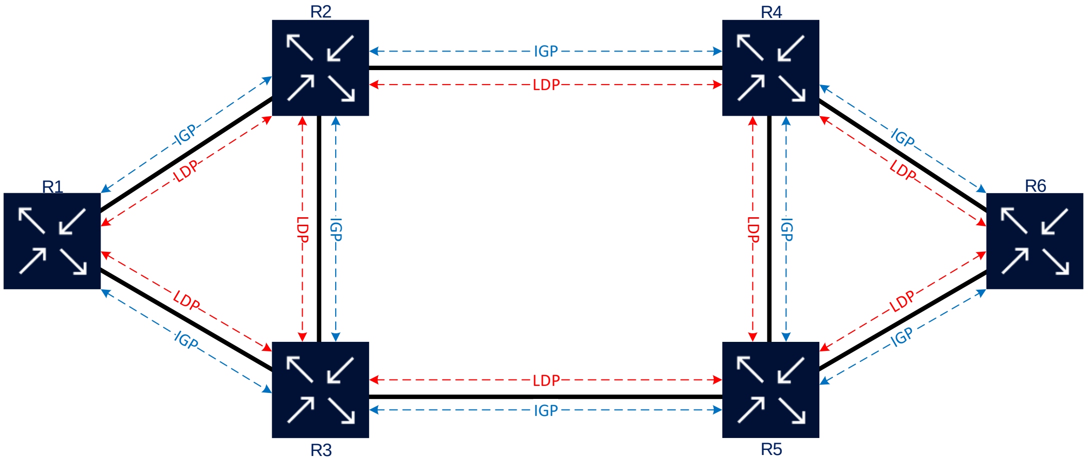


Figure 1-6: LDP Adjacencies

In order to form an LSP or LDP tunnel a label is bind with its Router-ID IP address, and that is distributed by each router in the network to all routers with LDP enable using downstream unsolicited mode. This mode implies a flooding behavior similar to IGP, which means a router distributes a label binding to its LDPs neighbors without being asked [4].

Figure 1-7 shown an example of label distribution from the router R6 perspective. First, R6 chooses the label that it wishes its peers to use when forwarding data to R6 Router-ID IP address (**10.0.0.6/32**), in this example the label is *131071*, then this label is distributed to neighbors R4 and R5. Second, routers R4 and R5 receive the label binding for prefix **10.0.0.6/32** and write it into their LIB, as an egress label towards router R6, then R4 and R5 generate their label binding for R4 Router-ID and distribute it to their neighbors. R4 distributes to R2 and R5 the label *131070*, and R5 send to R3 and R4 the label *131069*. Third, the same procedure from the second step is repeated in routers R3 and R2, so R4 and R1 receive a label binding (*131068*) for prefix **10.0.0.6/32** from R2 and router R5 and R1 receives the label *131067* from router R3. In each router all those labels received from its neighbors could potentially use the peer that advertised the label as a next-hop to send labeled data packets. Finally, router R1 must choose from among the two alternatives what label and egress interface to install them in the LFIB, in order to do that R1 checks the FIB to identify the IP next-hop for prefix **10.0.0.6/32**. In this case router R2 is the next-hop, therefore, router R1 installs the egress label *131068* in its LFIB, so when any unlabeled data packets coming in on R1 will be imposed that egress label and forward this traffic to router R2. If the main path goes down, router R3's egress label is going to be used after IGP has converged, so traffic will be forwarded to R3 in order to reach router R6 [4].

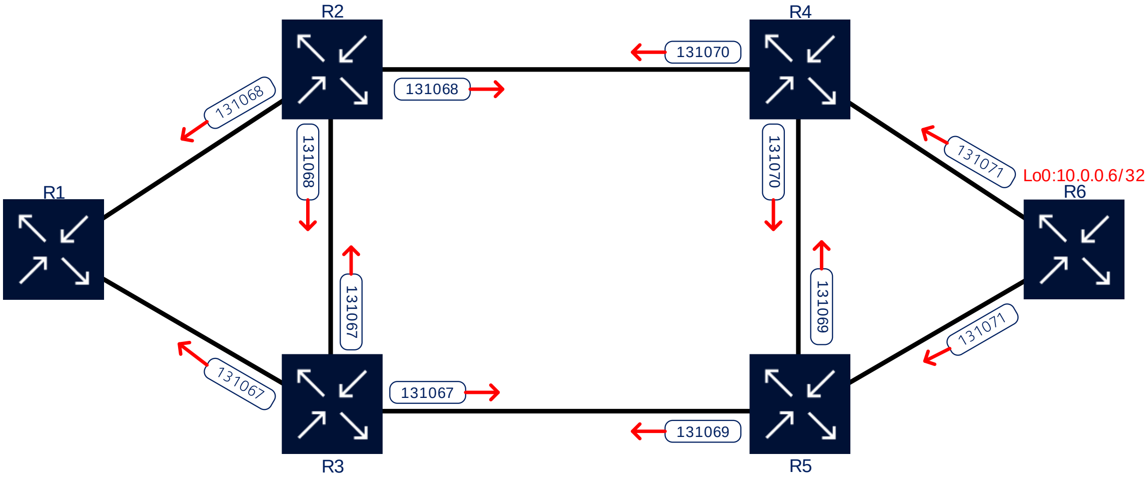


Figure 1-7: Label Distribution using LDP

After every single router in a network generates and distributes its label binding to its Router-ID prefix, all routers will have several entries for each prefix in their LIB and only the best option for each prefix will be installed in their LFIB. That entries detailed the ingress label, egress label, egress interface and operation. The three data plane MPLS operations, Push, Swap and Pop, will be explained in detail in chapter 2. Figure 1-8 shows an example of data traffic using label switched after the label distribution is completed in the control plane. When an incoming data packet with router R6 as the final destination arrives at router R1, that packet is labeled with an MPLS label value of *131068* and forwarded to R2. This implies a “Push” operation on router R1. Then in router R2 ingress label is swapped and transmitted with an egress label of *131070* towards R4. This labeled data arriving at router R4 is swapped again with *131071* as egress label and forwarded to router R6. Finally, the data traffic arriving from R4 with an ingress label of *131071* is popped on router R6. Example in Figure 1-8 is a simplification of this process because in normal conditions the label-switching process is only in the outer label or transport label and one or more service labels will be present after the service label is popped. All the remaining labels must be processed by router R6 in this case [4].

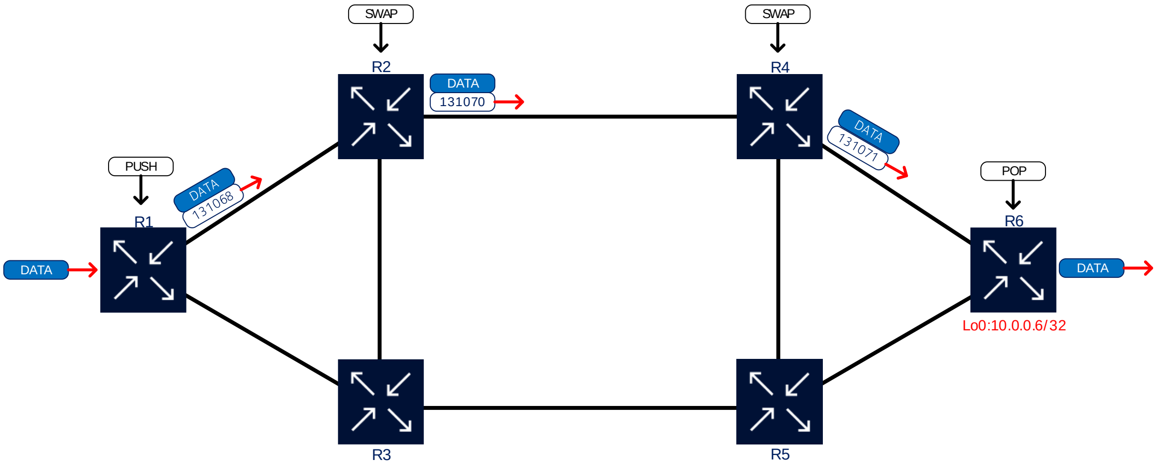


Figure 1-8: Data Plane using LDP

First MPLS networks permitted Service Providers to offer services isolated one to another reusing their IP networks, using pseudowires, L2 and L3 VPN services. These service providers found in LDP a simple protocol to signaling MPLS transport tunnels where service tunnels run inside. Yet their initial lack of traffic protection mechanism and no traffic engineering support, become a big problem when network convergence started. At that time, most IP/MPLS networks were used for the Core section and many different technologies such as SDH, ATM, etc., were used as access networks. These networks have a recovery time after failures below 50 ms, but that time could not be reached by LDP because of its complete dependency on IGP convergence time which was several seconds in the best case. Network convergence tries to consolidate all network sections in one only technology that means IP/MPLS when this migration started a different protocol had to be added in order to grant recovery times comparable with SDH and a rich set of Traffic Engineering tools.

1.2.2 LDP and RSVP-TE

RSVP-TE brings to the MPLS’ control plane many Traffic Engineering tools such as explicit strict defined paths, resource allocation along a path, link colors, etc. Besides, some protection mechanisms as a secondary path and Fast Reroute (FRR) technique in which restoration time is much shorter than convergence times offered by IGP [6]. Moreover, to implement RSVP-TE only a software upgrade in the node was needed to install OSPF TE [7] or IS-IS TE [8] extensions. Although all these benefits, RSVP-TE also carries on a high grade of complexity especially in the administrative task because this protocol used Downstream on Demand mode for label distribution in an MPLS network. That means LSPs have to be explicitly configured in order to signaling start, and two LSPs are required in each direction because an RSVP-TE LSP is unidirectional [9]. If a full mesh LSP is required, configuration and maintenance tasks could be long, complex and prone to errors. For instance, add a new to the network means configure LSPs from every single existent node to the new node, and from the new node to all existent nodes.

Figure 1-9 shows how RSVP-TE is added to the network, the very first step is adding all routers interfaces to this protocol in order to form sessions with their networks. This schema is used mainly in Core sections where services could ingress and egress in any Provider Edge (PE) router. In that case, LSPs are configured only when a service needs a TE mechanism or FRR protection while most services use LDP transport tunnels to avoid configuration complexity.

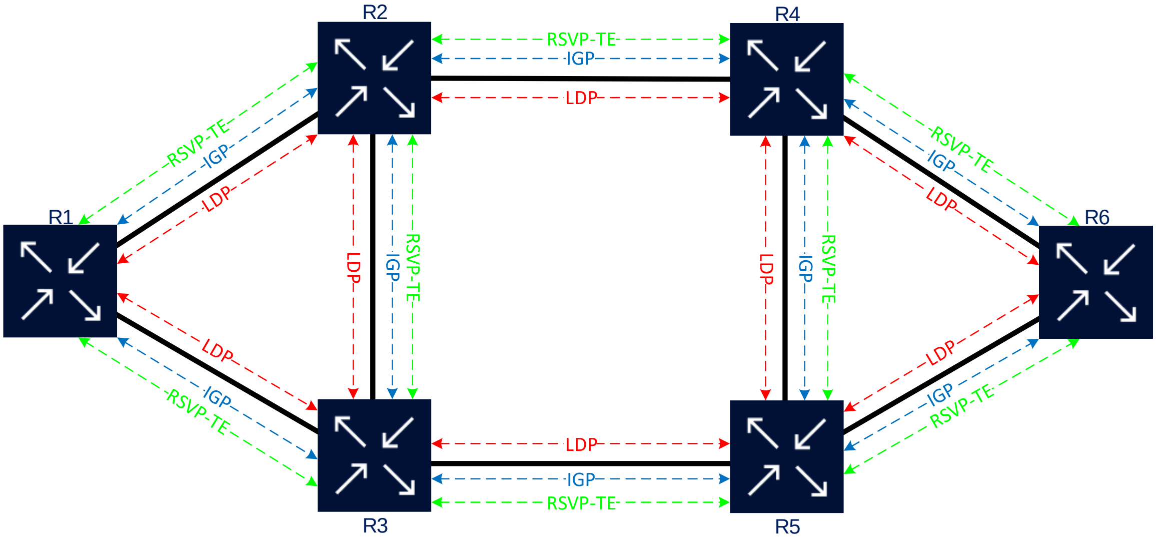


Figure 1-9: RSVP and LDP Adjacencies

Some access networks only require tens of routers as access nodes, but other networks are formed by hundreds or even thousands of access nodes. These access nodes were smaller (pizza-box size), with a few resources such as memory or processing power, but requires an under 50 ms recovery time and TE tools available. In this kind of network LDP is not a feasible option because label distribution is made without asking for it and all labels are stored in LIB no matters that are used or not, so scalability limits could be reached in small routers and that could drive to a major outage. Therefore, RSVP-TE was the label distribution protocol of choice because an LSP is only signaling when they are explicitly created, and labels are cleared if not needed. Those characteristics made of RSVP-TE a lightweight label distribution protocol ideal for small platforms and that also have traffic engineering and protection mechanisms.

Mobile Backhaul (MBH) or IPRAN is an example of this kind of network because access nodes are collocated in most of the cases with indoor radio units, so in a Mobile

Operator network region those access nodes could be hundreds. MBH uses a three tiers model to subdivide the network, smaller nodes or cell site routers form LowRAN, medium routers aggregating cell site rings are part of MidRAN, and the biggest platforms directly connected to Controllers constituted HighRAN. These tiers could have also been mapped to the traditional Access, Aggregation and Core model [10]. RSVP-TE fitted perfect for 2G and 3G transport services because in these technologies communication is always between controller and base station, for instance, BSC (Base Station Controller) with BTS (Base Transceiver Station) in 2G or RNC (Radio Network Controller) with NodeB in 3G. Communication between base stations was not required, so a couple of LSPs from access nodes to concentrator routers and vice versa are enough for 2G and 3G. Figure 1-10 shows an example of an MBH network with RSVP-TE as sole label distribution protocol, and only two LSPs from the LR1 access node to HR1 and HR2 core routers, LSPs from HR1 and HR2 to LR1 are also shown. With these LSPs, any Layer 2 or Layer 3 VPN service could be deployed in the network.

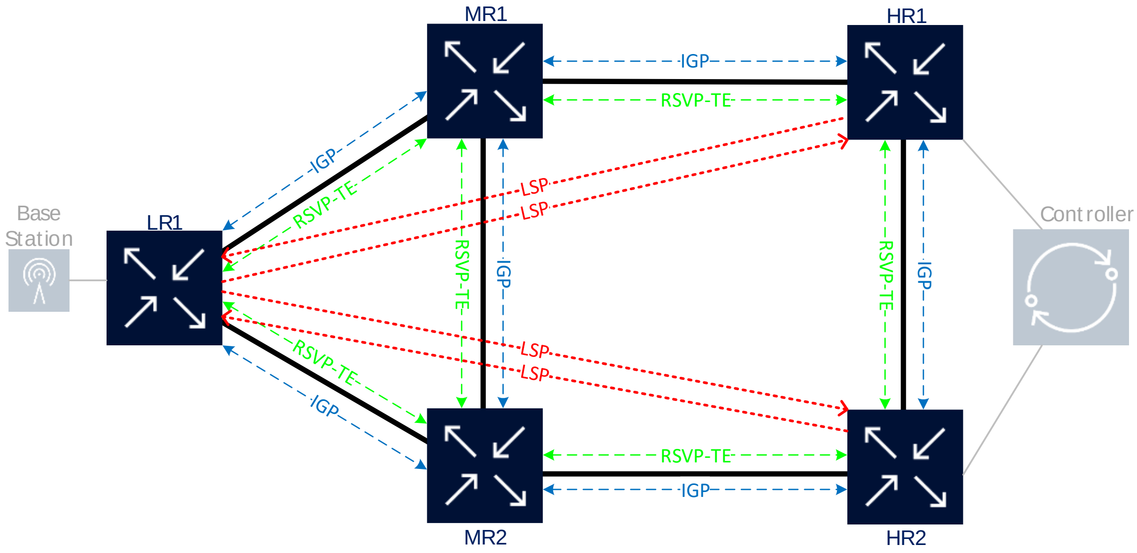


Figure 1-10: MPLS network with RSVP-TE only

This model started to become outdated when LTE or 4G was implemented in the MBH network because communication between base stations (X2 signaling) was mandatory. A full mesh of LSPs in a network with hundreds of routers is not practical because N*(N-1) LSPs are required (N is the number of routers), so several temporal but complex solutions at service level or even at hardware level were implemented to overcome this problem. For instance, upgrade many routers to support LDP and used that to transport X2 communications could be costly and long, as changing architecture from VPN L3 in the Cell Site Router to a hybrid L2 VPN in the LowRAN and L3 VPN in the HighRAN.

Another drawback of this model is this works well in a single area, but a single area could become a big scalability problem because any fault in the IGP domain could affect all routers in the area. An IGP hierarchical model that split the routing domain into multiple areas has many advantages over a flat IGP model, such as scalability increased, and better fault isolation but in multi-area topology LSPs cannot be signaling end-to-end by RSVP.

The reason for that this occurs is due to Traffic Engineering have a local area scope and cannot be performed across multiples areas. There is a solution for that, LDP over RSVP (LDPoRSVP) might be implemented to extend TE capabilities beyond area boundaries. Yet, that solution increase configuration and management complexity because RSVP LSP had to be split into several LSPs from source to Area Border Router (ABR) in each area, and a Targeted LDP session must be configured in the same LSP's segments in order to stitch RSVP LSPs in each ABR [4].

Targeted LDP or T-LDP is also used for signaling Virtual Leased Lines (VLL) and Virtual Private LAN Service (VPLS) service tunnels because the main characteristic of T-LDP is form sessions between two nodes not adjacent each other and exchange service labels. Multi-Protocol Border Gateway Protocol (MP-BGP) has a similar function for L3 VPN, and in new iteration could also be used for signaling VLL and VPLS [4]. MP-BGP is going to be further studied in the next chapters because is also used for Segment Routing.

Figure 1-11 shows all signaling protocols needed to create L2 and L3 VPN services from LR1 to HR1, assuming flat IGP area, RSVP-TE as a label distribution protocol, and unsupported MP-BGP as the service signaling protocol for VLL and VPLS in LowRAN nodes. It is easy to observe that even in a simple topology were too many protocols involved in order to create a service so that miss a configuration part is highly probable and troubleshooting tasks could become complex.

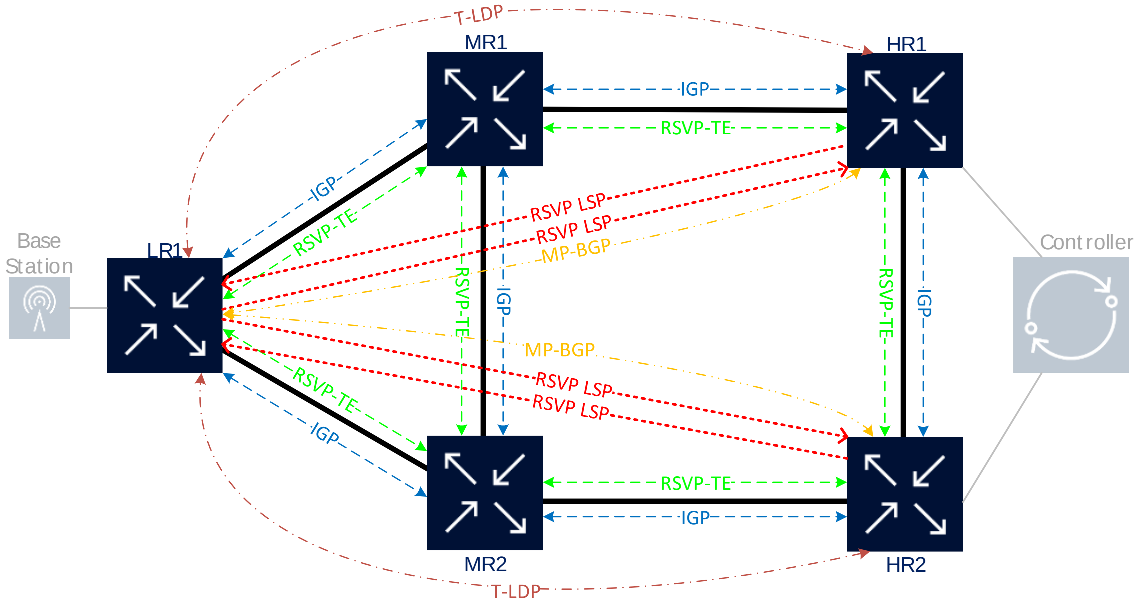


Figure 1-11: Transport and Service signaling protocols

1.2.3 LFA and Seamless MPLS

Loop-Free Alternate (LFA) is a technique that permits implement IP and MPLS/LDP Fast Reroute in order to node and link protection post-failure and pre-convergence [11]. This goal is achieved through the use of IGP pre-calculated backup next-hops that are loop-free and safe to use until the distributed network convergence process completes [12]. This technology did not achieve 100% protection especially in ring topologies because it is topology dependent, but that was a useful tool to avoid use RSVP-TE FRR as a protection mechanism in an LDP only network. New versions of LFA improve coverage protection, first with Remote LFA (RLFA) that provides additional backup connectivity for point-to-point link failures when none can be provided by the basic mechanism [13], so coverage goes to 95-99% [14]. Then Topology Independent LFA (TI-LFA) makes the coverage protection comparable to RSVP FRR that means 100%, and sub 50 ms recovery time. TI-LFA will be explored in later sections because it is the protection mechanism of choice for Segment Routing.

On the other hand, Seamless MPLS is a technique that uses existent protocols such as BGP, LDP, RSVP-TE, OSPF, and IS-IS to extends the core domain and integrates aggregation and access domains into a single MPLS domain, considering the limited feature set and scale of access nodes [15]. BGP based on RFC 3107 to route distribution through different areas, instances or even different IGPs. When BGP is used to distribute a route, it can also distribute an MPLS label for that route. The label is piggybacked in the same BGP update message used to distribute the route itself [16]. Therefore, BGP Label (RFC 3107) signaling transport tunnels End-to-End through different areas or instances.

Transport tunnels intra-region (an area or instance) still uses LDP or RSVP-TE, so inside each region LDP FRR LFA or RSVP-TE FRR protects against link or local ABR failures. On the other hand, BGP Edge Prefix Independent Convergence (PIC), also known as BGP FRR, protects against remote ABR failures for inter-region traffic protection. This technique uses pre-computation of BGP backup paths in the control plane to support fast reroute of BGP traffic around unreachable or failed next hops. That convergence is based on IGP convergence at the ABR, hence a bit slower compared to link-failure detection [15].

The general architecture of seamless MPLS is shown in figure 1-12, only one Autonomous System (AS) is used through this network subdivided into three IS-IS instances. Cell Site Routers (CSR) in instance 5 and instance 6, and Concentrators (CN) only known IGP routes inside their instance. However, ABRs knew routes from both instances, and for this reason ABRs become BGP Route-Reflector for their instance. That means ABRs first learn Router-ID IP addresses with a label associate from instance 0 and the other instance, then each ABR sends these routes to all CSRs in its instance, finally, the ABR sends BGP label routes from its instance to instance 0 and the other instance. All routes reflected from the ABR change next-hop info to ABR’s Router-ID IP address because CSRs and CN only know how to reach each other using ABR. With this architecture, each IGP instance is isolated from each other, so to create a service from a router from an instance to other transport tunnels is going to build based on BGP label end-to-end.

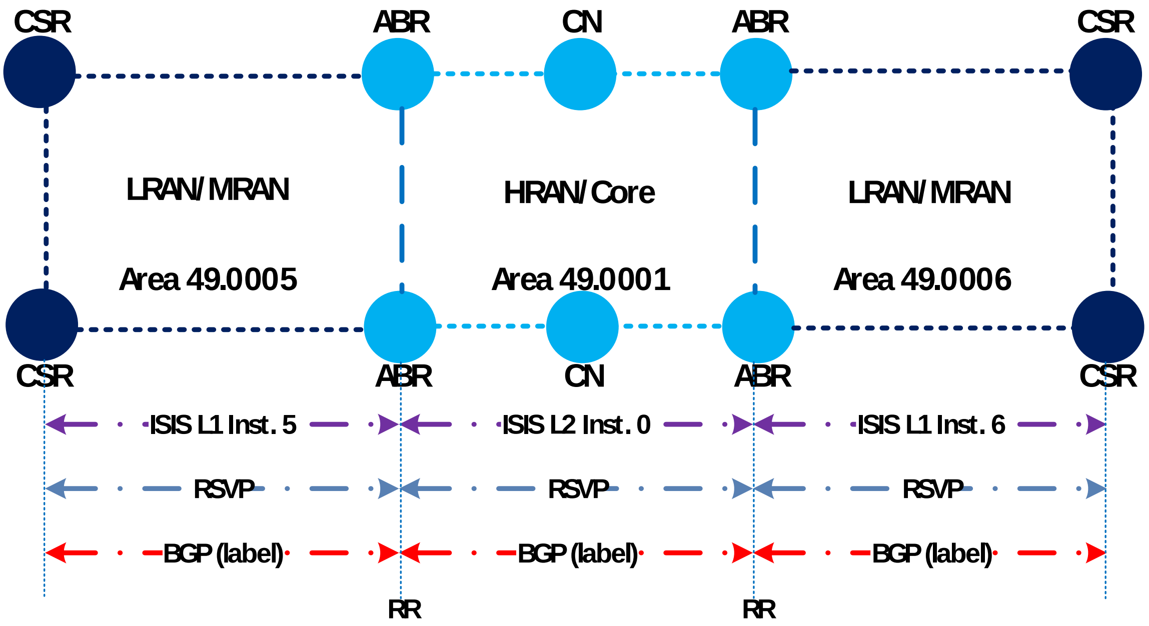


Figure 1-12: Seamless MPLS signaling

Finally, table 1-1 shown a comparison of the main characteristics of LDP and RSVP as label distribution protocols for MPLS' control plane.

	LDP	**RSVP**
Overview	Multipoint to point	Point to point
Operation	Simple	LSP per destination/TE-path
Dependencies	Relies on IGP	Relies on IGP TE
LBL allocation	**Local significant** per node (interface)	**Local significant** per node (interface)
Traffic Engineering	**No**	**Yes**
Scaling	1 LBL per node (interface)	**Nx(N-1)**
Fast Reroute	LFA, LFA Policies, **Remote LFA - <100% coverage**	Link/Node protection (detour/facility) – 100% coverage
Multicast	mLDP	P2MP RSVP
IPv6	Extensions required	Extensions required

Table 1-1: LDP and RSVP summary [17].

1.3 SR vs. MPLS Control Plane

As shown in Table 1-1 both label distribution protocols in MPLS cannot work alone if a network needs simplicity, scalability, redundancy and Traffic Engineering. LDP could be good for scale and operational simplicity, also fast-reroute is now possible using some flavor of LFA reaching similar coverage protection as RSVP-TE. Yet Traffic Engineering needs RSVP-TE to enable throughout the network to work. However, TE solutions using RSVP-TE do not scale properly when more granularity is required. RSVP-TE also has scalability problems because its full mesh of LSPs requirement and routers must maintain states all along LSP's path [17].

Both protocols have scalability problems when a flat area schema is used, so in order to create a hierarchical topology with several areas or instances LDPoRSVP or Seamless MPLS has to be implemented. LDP over RSVP is complex to configure and manage, so Seamless MPLS is the better option to grown to thousands of nodes using the current control plane protocols. Seamless MPLS also uses BGP, which is already deployed in most of MPLS networks, to distribute prefix information associated with a label. Seamless MPLS could serve as an intermediate step to migrate a network to Segment Routing. That approach permits integrates new nodes using SR and still has communication with the existent sections which are using LDP or RSVP [15].

RSVP-TE is not optimized for Software Defined Networks (SDN) where a controller is used to compute dynamic paths on-demand based in-service requirements as latency or bandwidth. Also, it is really difficult to apply dynamic traffic rerouting based on network conditions such as congestion or optical transport paths availability. To sum up, current MPLS control plane protocols, LDP and RSVP, are unsuitable for the continuous network growing, network automation, and management.

On the other hand, Segment Routing is a very scalable protocol because it uses current IGP, OSPF or IS-IS, for segment label distribution, so no other protocol has to be implemented as a control plane to tunnel services from ingress to egress using or not an explicit path. In addition, P or transit routers do not track any packet state because labels in the packet contain all instructions to get a destination [17]. That simplifies signaling a lot as it is shown in figure 1-13.

SR uses TI-LFA to get a 100% coverage protection with under 50 ms link and node traffic recovery time, this flavor of LFA uses a source route backup path to expand coverage to any topology. Also, TE capabilities of SR could be distributed in each router or centralized in an external controller, that is why protocol's creators consider SR has become the de-facto network architecture for SDN [18]. It is recommended we use a centralized traffic engineering controller, better known as Path Computation Element (PCE) server, to flow steering and service chaining management.

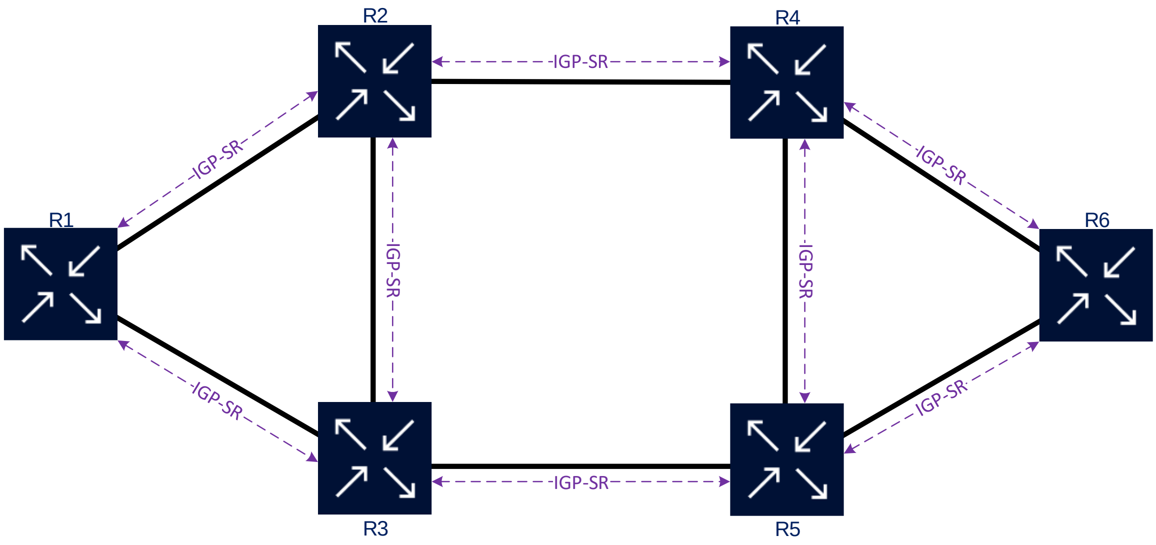


Figure 1-13: IGP with Segment Routing Extensions

2 SR Basic Concepts

2.1 Segment Routing Concept

Segment Routing (SR) is a mechanism to implement source routing in a network, which permits to a Label Egress Router (LER) to specify a unicast forwarding path, with traffic engineering capabilities, that a packet will traverse through an ordered list of segments. Current IGP will be reuse only adding SR extensions with a software upgrade, that simplifies control-plane because no other protocol is needed [1].

Unlike RSVP-TE based explicit-routed LSPs where midpoints hold the state of each LSP, Segment Routing requires that only the ingress PE holds state. So, no state is held in the network, which means the intermediate nodes only apply a set of instructions as forward, terminate or prepend new segments, without the need for any signaling protocol. In the end, this gives to Segment Routing a major scaling advantage [1].

Segment Routing starts in one simple idea based in real life, that is when someone wants to make a trip from point A to point B and the path is unknown, a GPS could be used with a turn by turn approach. Yet, when all the route is known a few steps are needed to get the destination, some path options are also available in most cases. The GPS option is similar to the RSVP-TE functions in MPLS because a state update is shown in each step along the route. On the other hand, the few intermediate notable points in the path are very similar to the Segment Routing operation. For instance, someone wants to go from Edmonton to Banff, so the shortest path is to get Calgary using Queen Elizabeth II Highway, and then take the route to Banff. But sometimes a peaceful and slow way is preferable, in that case, the same Highway could be used until taking Pigeon Lake’s route and then use Rocky Mountain House and Cochrane as intermediate points to get Banff avoiding Calgary and its busy roads. Figure 2-1 shows the highlighted steps along both routes.

The same way a trip could be planned with only a few reference points, SR could use a small number of shortest paths hops through intermediate devices of the network. That means it is possible to use one or two segments that describe the entire path [19]. Also, when multipath with equal bandwidth exists in a network, SR will build paths using Equal Cost Multipath (ECMP) to load balance traffic along different routes.

2.2 Segments

A segment can represent an instruction topological or service-based which a node executes on the incoming packet [20]. A segment list can represent either a specific topological path to a node or link or service. The segments can be thought of as a set of instructions from the ingress PE or LER such as “go-to router R5 using the shortest path”, “follow the shortest path for prefix P,” or “use link L” [17]. Figure 2-2 shown an example of topological paths as a node or link.

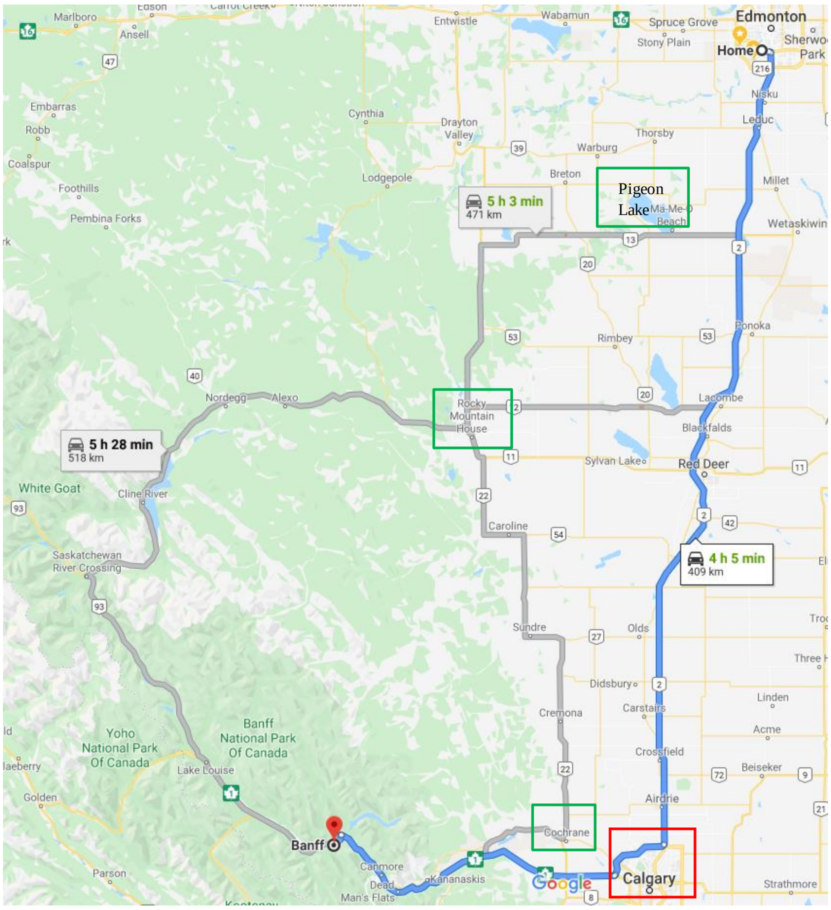


Figure 2-1: Reference points in Edmonton-Banff routes

Segments could be local or global, so if a segment is *local* that is relevant only for the local router and that may indicate use a determinate output interface or a specific link between several links in a LAG. A segment *global*, by contrast, is an instruction with significance within all SR domains, for example a specific path to a destination different to the shortest path calculated.

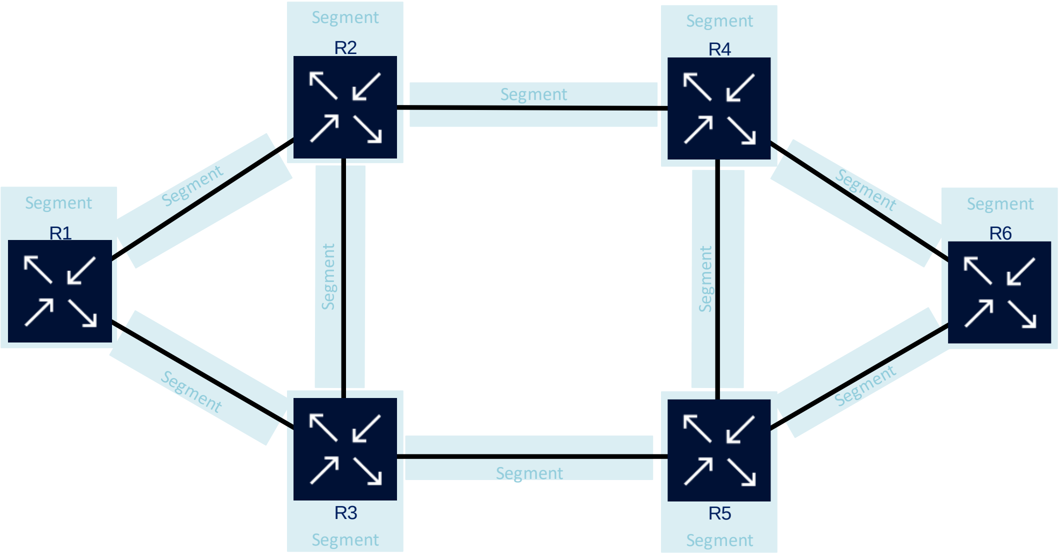


Figure 2-2: Segments Topological

An entity called Segment ID (SID) is associated with each segment, and this is advertising by IGP. SID format is based in the data plane used for SR, for instance, a SID in an MPLS Data plane is defined as a 32 bits field, where the 20 rightmost bits are the MPLS label [21]. Although RFC 8402 stated that used SID in place of the term SID is technically imprecise [20], all vendors used the terms SID (Segment ID) and Segment interchangeably because a SID is tied to only one segment.

2.2.1 Segment List Operations

A SID list is carried in the header of the packets, where the segment used by the receiving router to process the packet and follow the instructions coded is called **Active Segment**. This segment is located at the top of the stack [19]. If an active segment contains an order like "go-to router R5 using the shortest path", that means each transit node will forward the packet using its short path until getting the destination.

Each node executes one of three operations in the segment list in order to process a packet:

- **Push:** Could insert one or several SIDs at the head of the segment list, the top most SID becomes the Active Segment.
- **Continue:** The Active Segment instruction is incomplete, so it must remain active to the next hop**.**
- **Next:** The Active Segment instruction is done, so the next segment in the list become active [3]**.**

When MPLS Data plane is used by SR these operations are equivalent to **Push**, **Swap** and **Pop** respectively. IPv6 also is defined to be used as a data plane for Segment Routing. A segment list operations example is shown in Figure 2-3, the instruction *Push* in R1 is "go-to router R5 using the shortest path", in R3 a *Continue* operation is executed because the destination is not got yet, then in R5 a *Next* is executed because Active Segment instruction is complete. No more SID is in the segment list, so data is delivered in R5.

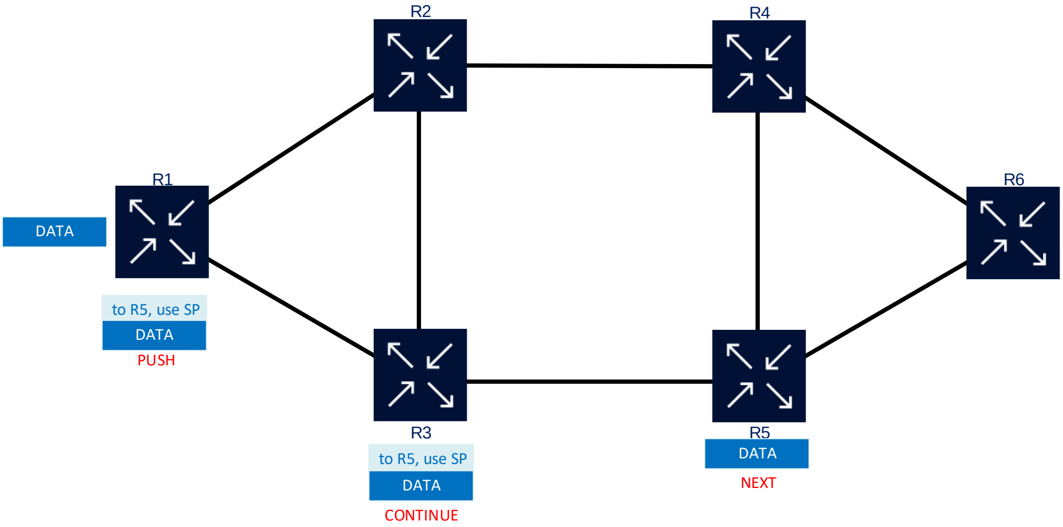


Figure 2-3: Segment List Operations

2.3 IGP Segments

In the early stages of networking, there were a bunch of IGP protocols, but currently, only OSPF and IS-IS are suitable to been used in a modern network. So, when Segment Routing was conceived using these two protocols to flood segment information was the logical step to follow. New extensions were defined to permit OSPF and IS-IS to distribute IGP segments information within an SR domain, so no other protocol as LDP or RSVP-TE is needed in the SR control plane. There are two types of IGP segments, IGP Prefix Segment associated with router ID and IGP Adjacency segment tie to network links.

OSPF and IS-IS are link-state protocols where any node inside the IGP domain shares the same topological database, which means each router has routes to all links and Router IDs inside an area. On the other hand, for inter-area communication, a link-state protocol does not share links information but prefix information or Router ID between routers in different areas. These statements are also valid when a router becomes SR enabled because inside an area both types of IGP segments are propagated but in a multi-area topology only IGP Prefix-SID are distributed. That means SR could build transport tunnels end-to-end in a multi-area environment [3].

Figure 2-4 outlines IGP segments types and their relationship with Router ID and link network. All *P-number* segments are Prefix-SID associated with each router loopback RID, and all *A-number* segments are Adjacency SID ties to each interface active in a router.

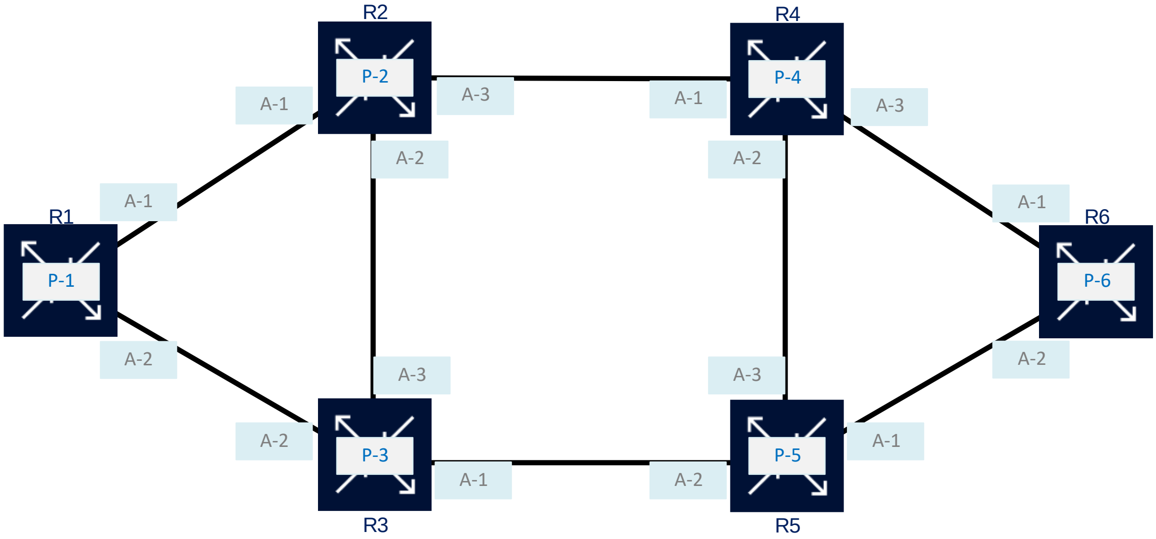


Figure 2-4: IGP Segments Types

2.3.1 Prefix-SID

Also known as Node-SID because identifies each router inside a Segment Routing domain and is associate with the router ID or loopback that means a /32 or host prefix. This segment is always global because it must be a unique value for all networks. Although RFC 8402 stated that an IGP prefix segment could be associated with any kind of prefix and could also be a local segment, all vendors limit that Prefix-SID are associated exclusively to host loopbacks or RID and prefix segment are always global [20].

Even though a node SID is a special case of prefix ID where a flag must indicate when a segment is a node segment [22], both terms are used interchangeably by most of the vendors and papers [19] [21]. That is because all practical implementations of SR using only /32 prefixes for prefix ID, and this host address most of the time is the router ID (RID). It is a best practice in design using only one IP address as RID for all protocols such as OSPF adjacencies, BGP session and so on.

A prefix SID in a packet header represents the shortest path route to the node which advertises this nodal segment associated with its global identifier. This prefix segment is ECMP aware which means when two or more paths have the same cost, traffic will send for all of them using load balancing. Only one prefix SID is enough to send traffic from anywhere in the network to any specific destination using the shortest path calculated by IGP along with multiple hops. Figure 2-5 shown all these prefix SID features; in this example packets have *P-6* prefix SID as an active segment, so they are going to travel from R1 to R6 using both shortest paths with the same cost available in the network.

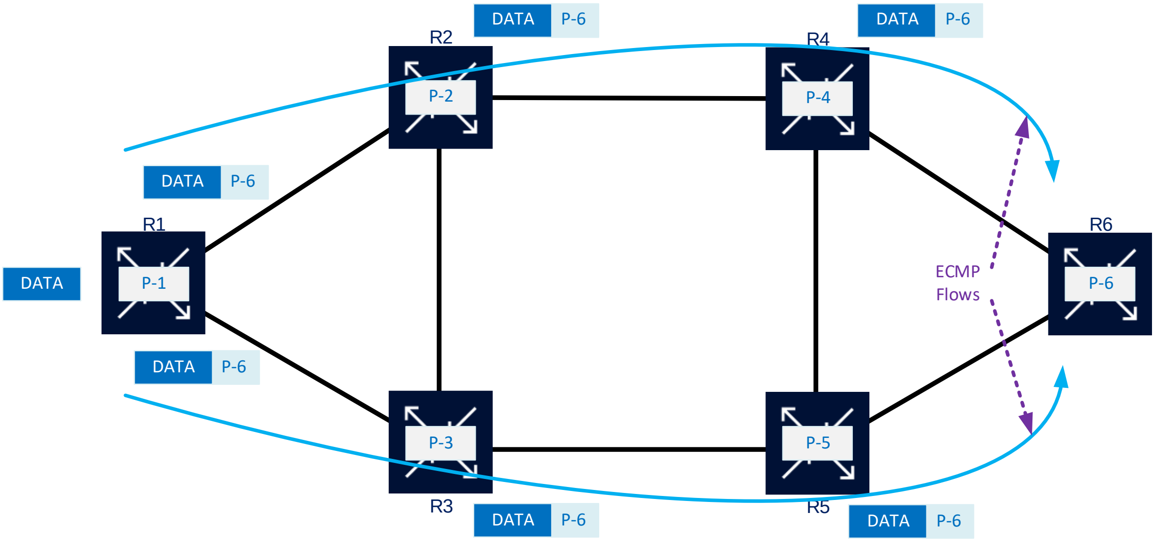


Figure 2-5: Prefix-SID and ECMP

Prefix-SID allocation must be assigned in the network planning phase, along with Router ID IP address because both will remain unchanged since implementation onwards. In fact, node SID could derivate from the RID IP address, for instance node segment's three last digits could be the same as the loopback last octet. This assignment must be registered because each Node SID must be unique along with all the SR domain, like loopback IP address cannot be repeated to identify a node. Automatic tools could be used to assign, log and avoid node segment repetition. Figure 2-6 shows how a node SID could be a tie to the router loopback, their routers loopbacks use 10.0.0.0/24 range, and prefix segments use 16000 to 16999 range. The RID IP address last octet and node segment three last digits are the same.

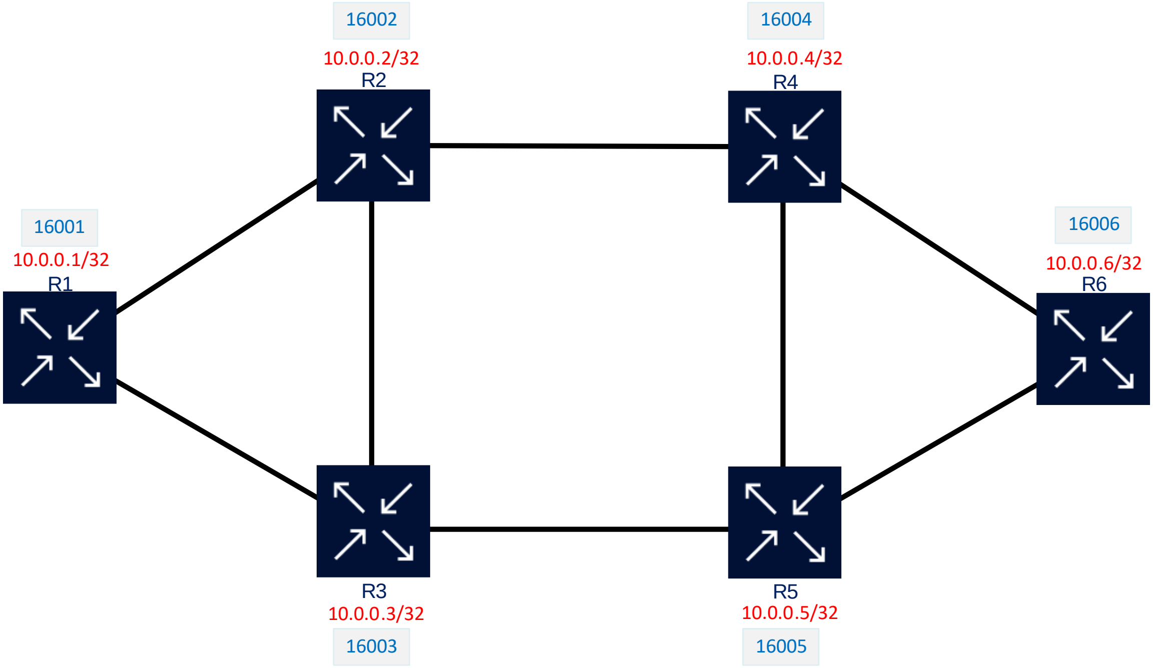


Figure 2-6: Router ID and Node SID

2.3.1.1 *Prefix-SID Algorithm*

By default, the prefix segment uses the traditional Shortest Path First (SPF) algorithm used in OSPF and IS-IS. As already was mention in point 1.1, the shortest path to any destination is which has the lowest cost value to that destination. RFC 8402 also defines a second algorithm which is a slight variant of SPF called "Strict Shortest Path First". The main difference between both algorithms is that any routing policy in a router could change the path calculate by SPF but using Strict SPF any local policy will be ignored and the packet will follow the SPF path no matter what. If this algorithm is going to be used in an SR domain, all nodes must support that because a path calculated could be changed by a router that does not support Strict SPF. Algorithm 0 was assigned to SPF and algorithm 1 to Strict SPF. [20]. Currently, besides Cisco, there is no evidence another vendor is using this new algorithm [23].

The fact that more than one algorithm exists in SR resulting in different prefix SID have to be used for each algorithm, so two different prefix segments are needed for the same prefix if both algorithms are going to be used. Most of the time only one prefix SID is used for each node inside an SR domain because Strict SPF is still not deployed in many networks.

When a router receives a prefix SID, that write a forwarding entry with the Active segment, next hop's outgoing interface, next hop's IP address and the pertinent segment list operation (these operations were explained in point 2.1.1). Each prefix SID is calculated and distributed by the current IGP deployed in the network, IS-IS or OSPF, so an SDN controller could only select and use the specific prefix and hold the only state at the source. Also, each prefix SID could use TI-LFA to protect traffic against failures with a restoration time minor to 50 ms [22].

2.3.1.2 IGP Anycast Segment

Anycast IP address is a networking technique that has been using a long time ago for several kinds of server pools to load balance and high availability. This technique uses only one IP facing the client and several servers that have the same function and are represented by this unicast IP address. The server topologically closest to the client will be who answer to the customer request. A traditional example of this is DNS servers which receive thousands of queries to the same IP address, but each query could be attended by a different server [24]. Figure 2-7 shown a client making a query to its configured DNS server, all queries coming from this client will be resolved by Server A because only router R1 is between customer and server. If this server or path fails, a query will be automatically switched to Server B which is the second closer to the client and so on. Although Anycast IP is widely deployed for servers, there are other applications for anycast, for example, a Rendezvous Point (RP) IP address for the distribution of multicast channels where two or more routers sharing IGMP join messages from customers.

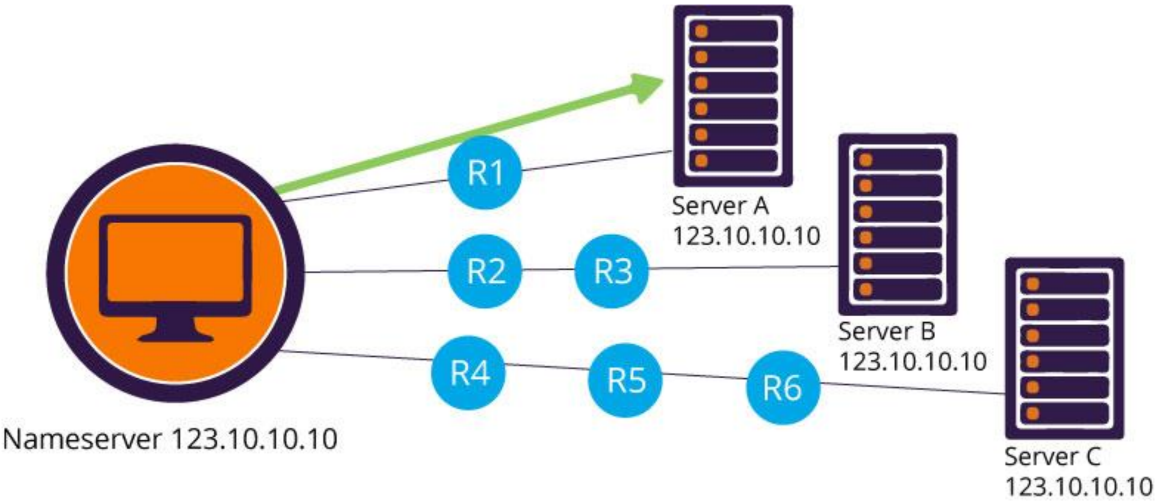


Figure 2-7: Anycast IP address for DNS server [24]

From an SR point of view, an Anycast SID is a special type of Prefix segment that identifies a group of nodes instead of a specific router. In general, an anycast segment is tied to a loopback prefix which represents several nodes not a specific one. The instructions of an anycast segment still are the same as a Prefix-SID, which means send traffic for all ECMP shortest paths available to the anycast prefix tie to the anycast segment. If two nodes from the anycast set are at the same topological distance from the source, traffic will be load balancing. On the other hand, if one node from the anycast group is closest to the traffic source, only this anycast node will receive traffic. Figure 2-8 shown these two cases: First, router R1 want to send traffic to Anycast group represented by IP address 10.0.0.11/32 which is attached to the anycast SID 16011, so traffic will be send using ECMP to routers R2 and R3 because both are part of this anycast set and have the same cost from router R1. Second, router R4 is going to send packets to the same Anycast SID 16011, so when this router calculates the shortest path to 10.0.0.11/32 found router R2 is the closest one to it and will send all the traffic to R2 exclusively.

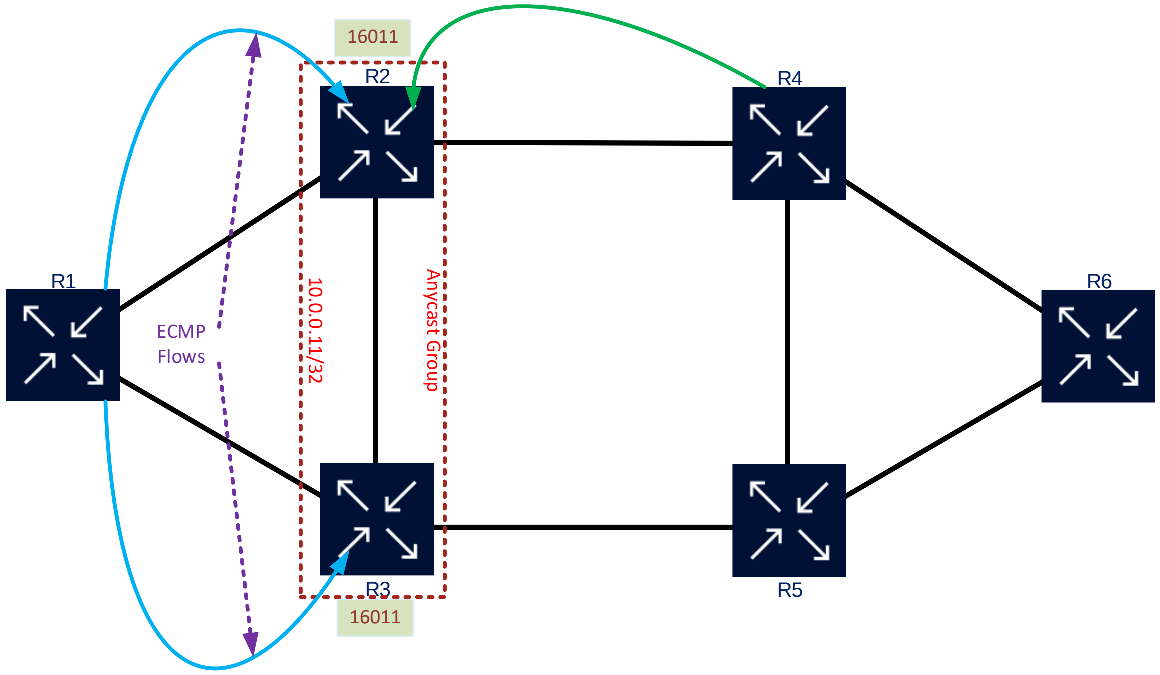


Figure 2-8: Anycast SID traffic behavior

However, Anycast SID is used mainly for load sharing that also is used for high availability. When two or more ECMP flows are present and one of the routes which are receiving fails, this flow will be redirected to another member of the anycast group. If only one shortest path is used then after this path fails, traffic is rerouted to the second-best short path to the anycast SID. This process is made automatically without node changes. Figure 2-9 shows this behavior for a fail in router R2 and redundancy in figure 2-8 scenario.

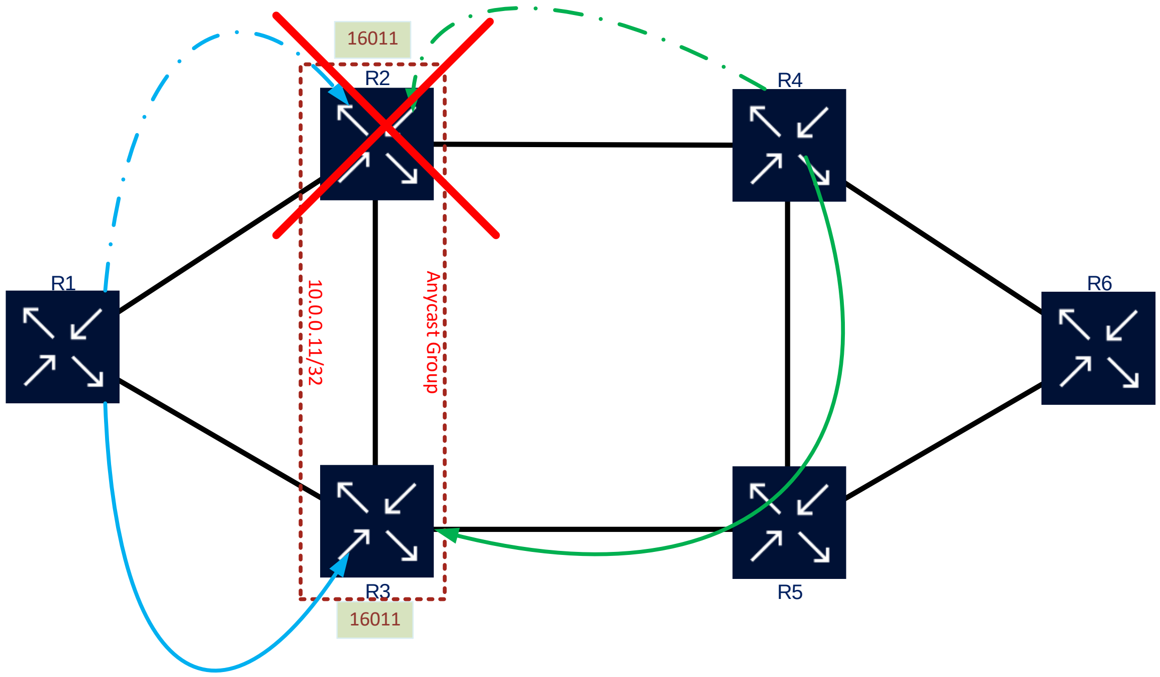


Figure 2-9: Anycast SID redundancy

2.3.2 Adjacency SID

The Adjacency segment is a SID attached to a single or a set of unidirectional adjacencies. It is formed by the local and the remote (normally the adjacent) node and cannot be the IGP node itself. This IGP segment identifies a given interface or next-hop where there will be one local identifier for each adjacency [21]. Similar to Prefix-SID, RFC 8402 stated that an Adj-SID could be a global segment, which is not applicable for any vendor because there are not benefits to make Adjacency Segment globally unique. In fact, the approach most used in practical implementations is made Adj-SID locally unique where each node in the SR domain could use the same space [1].

A packet with an Adj-SID as an active segment will be forwarded using the interface associated with this segment in this particular node, so the same Adjacency segment in another node could use a completely different interface to forward traffic. An Adj-SID instruction is locally defined by each router inside an SR domain, and that could be used to create paths different from the shortest path because the cost is ignored. This characteristic permit creates explicit paths expressed by a list of segments [22]. Figure 2-10 shown the way Adjacency SID could have the same values along with different routers, in this example if the active segment is *A-1* in router R1, the packet will be forwarded using the interface to router R2. The same Adj-SID value *A-1* in router R3 will forward the packet using the interface to router R5.

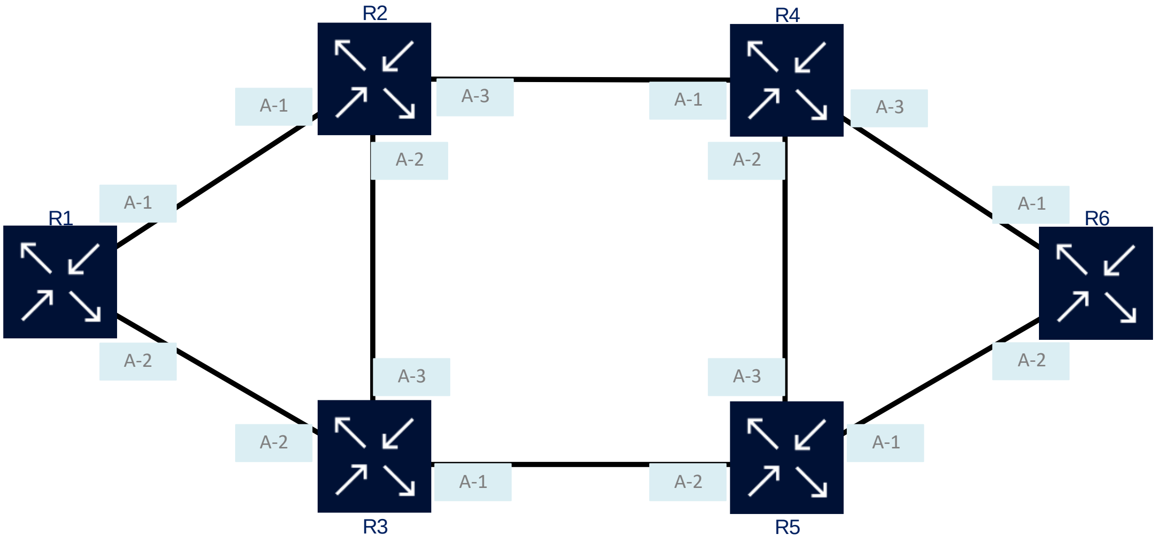


Figure 2-10: Adjacency SID local meaning

The local nature of the Adjacency segment force that a node SID is always needed before an Adjacency SID because a specific node where the Adj-SID will be used is mandatory to forward a packet for a particular interface. Figure 2-11 shown the way Adjacency segments are used: Router R1 wants to send a packet to router R6, but the SPF path is not going to be used. R1 is going to use a path that uses R5 to R6 interface in router R5, previous and next hops could use the shortest path. As a result router R1 add to header Node SID *P-6*, then Adj-SID *A-3* and finally Node segment *P-5* as the active segment. This packet is forwarded to R5 using the shortest path, which means by router R3. In router R3 a CONTINUE operation is made and the packet is sent to router R5 using a directly connected interface. In router R5, Adj-SID *A-3* becomes an active segment and its instruction is executed, so data is sent only with *P-6* Node SID using interface towards router R4. Finally, router R4 sent the packet to router R6 using its shortest path, which means interface R4 to R6.

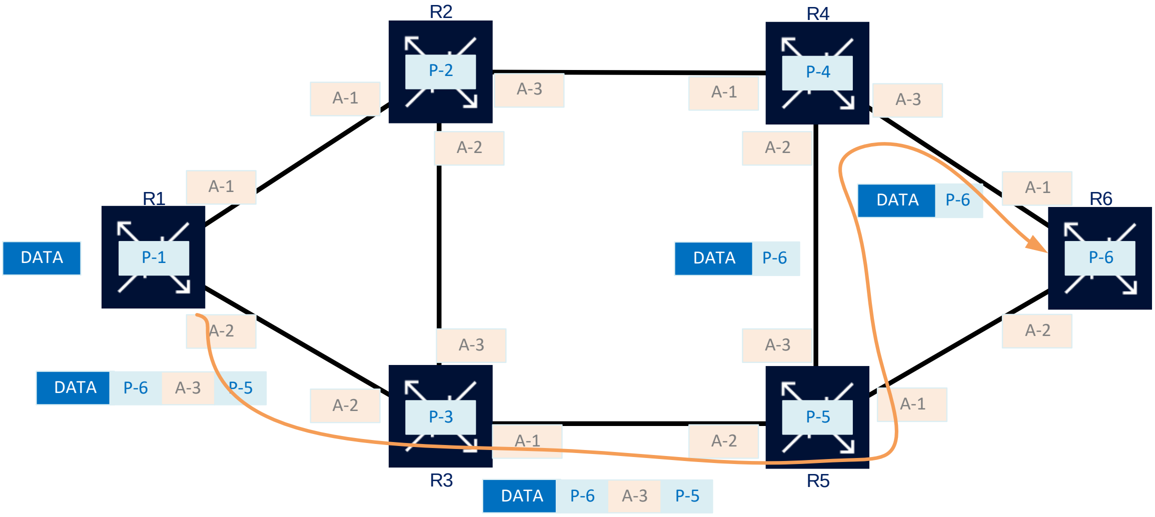


Figure 2-11: Adjacency SID traffic forwarding

2.3.2.1 Layer 2 Adjacency SID

Link Aggregation Groups (LAGs) can be configured to increase the bandwidth available between two network devices, depending on the number of links installed. LAG also provides redundancy if one or more links participating in the LAG fail. All physical links in a given LAG link combine to form one logical interface [25]. Only one Adj-SID will be associated with the LAG L3 interface, so it is not possible to tie individual links inside a LAG to and Adjacency SID. To overcome this problem Layer 2 Adjacency SID was created, which is a special kind of Adjacency Segment that is associate with a specific LAG member link. That means L2 Adj-SID permit choose between different Layer 2 links inside a LAG to forward traffic.

Figure 2-12 shown Adj-SID *40103* tied to interface R3 to R5 in router R3, this interface uses all four links in LAG-1 to load balancing traffic. This Adjacency Segment is related to the Layer 3 interface that uses LAG-1 as Layer 2 transport and an IP subnet in L3. Router R3 also generates a Layer 2 Adjacency segment for each link which is part of LAG-1, so in this example link 4 receives the L2 Adjacency SID *40143*. To steer traffic using this specific link, segment *40143* had to be added to the segments list after router R3 Node-SID. The same way an Adj-SID has local meaning, also L2 Adjacency Segment has a local span.

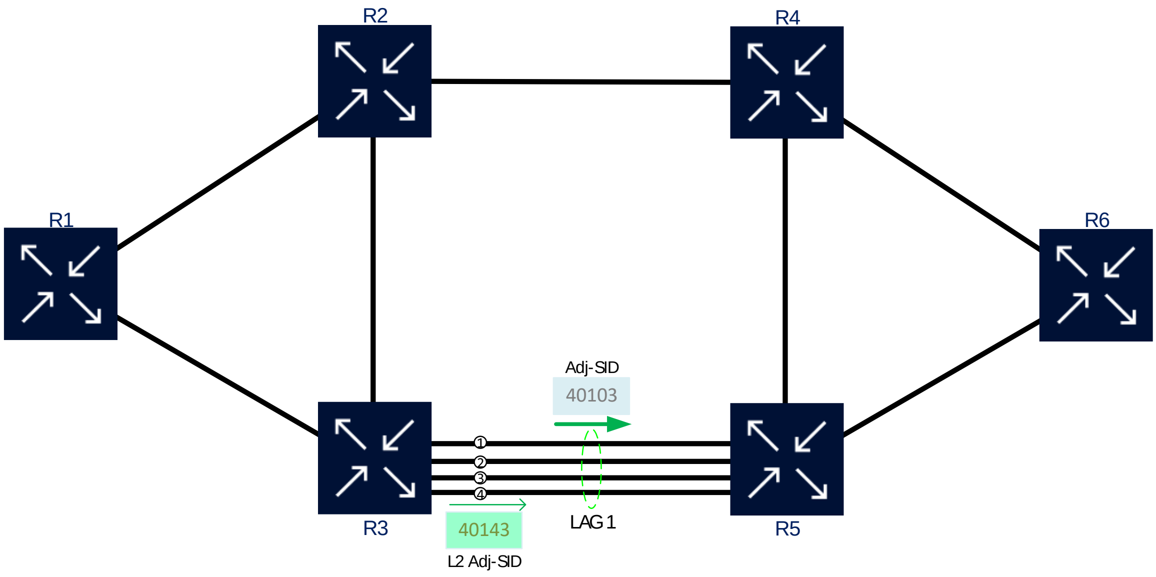


Figure 2-12: L2 Adjacency SID and Adj-SID

2.4 BGP Segments

BGP could also be used as a control plane protocol in Segment Routing, so IGP segments have their equivalence in BGP. Assigned more signaling functions to BGP has become a tendency in last years because that is a simpler protocol for design, is well supported by all vendors, is really scalable and modular, and is already deployed in almost all service providers networks. All these reasons that made some web-scale Data Centers (DC) from big players used BGP as IGP, some DC use spine-leaf topologies where only eBGP sessions are established between Spines, Leaf and Top of the Rack (ToR) switches [26].

Initially, BGP was created for peering between different Autonomous System (AS), but soon that start to be used as a signaling protocol for VPN L3. At this point BGP had become Multi-Protocol BGP (MP-BGP) which permits add new address-families to support different protocols like IPv6, L2TP, MVPN, BGP-LU (RFC3107), etc. For instance, VPLS or VPN L2 could be signaling using BGP to create service tunnels in order to avoid use T-LDP. Currently, most of the service provider networks have at least a couple of Route-Reflector (RR) up and running, and these are reflecting hundreds of routes from multiples address families to their clients. That is a big advantage because any protocol that is going to be deployed only need to add the respective address-family to the session between a node and the RR.

2.4.1 BGP Prefix-SID

A BGP prefix segment is attached to a BGP prefix, the same way an IGP prefix SID is attached to an IGP prefix. In an IGP free environment BGP prefix, SID must be used to tie prefixes learn by BGP to their respective SID [20]. Similar to the IGP prefix segment a BGP prefix segment is always global, so all nodes inside the Segment Routing domain could execute instructions inside this SID. A BGP segment command to forward and load balancing traffic using all BGP paths available in the network.

Figure 2-13 shown a Spine-Leaf DC topology where Spine switch is in AS 1 and Leaf switches are in AS 2, a full mesh eBGP session is assumed between both layers. All nodes forming this network are going to use BGP to distribute its BGP prefix SID attached to its node's loopback. In this example, a packet in switch Leaf L1 has a destination prefix 10.0.0.14/32, so the BGP Prefix-SID *16014* is added as segment active. L1 is going to send this packet using all shortest paths available, in this case, there are two paths using switches S1 and S2. Packets will be load balanced using both paths, and when they arrive at S1 and S2 they are going to forward packets to switch L4.

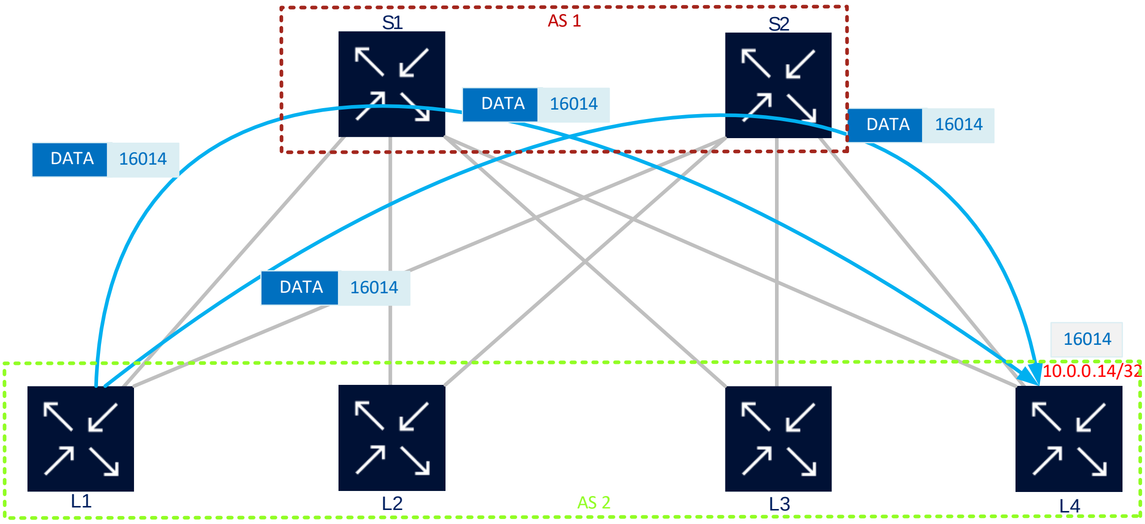


Figure 2-13: BGP Prefix-SID forwarding

2.4.1.1 BGP Anycast Segment

Anycast prefix can also be used in BGP, so a group of nodes could be identified by only one unicast IP address for redundancy and load balancing. Anycast is linked with the BGP protocol which ensures that all of a router's neighbors are aware of the networks that can be reached through that router and the topological distance to those networks [24]. A BGP Anycast SID is a special type of BGP Prefix segment that is a tie to an anycast prefix. The instructions of a BGP anycast segment still is the same as a BGP Prefix-SID, which means to send traffic for all ECMP shortest paths available to the anycast prefix tie to the anycast segment. Figure 2-14 illustrates how a BGP anycast segment is forwarded in a Spine-Leaf network. Switch L1 is going to send traffic to Anycast group of spine switches S1 and S2 represented by IP address 10.10.0.1/32 which is attached to the anycast SID *16101*, so traffic will be send using ECMP for a different path to switches S1 and S2 because both are part of this anycast group and have the same cost from switch L1. If one of the paths to S1 or S2 fails all traffic is going forward to the only link available.

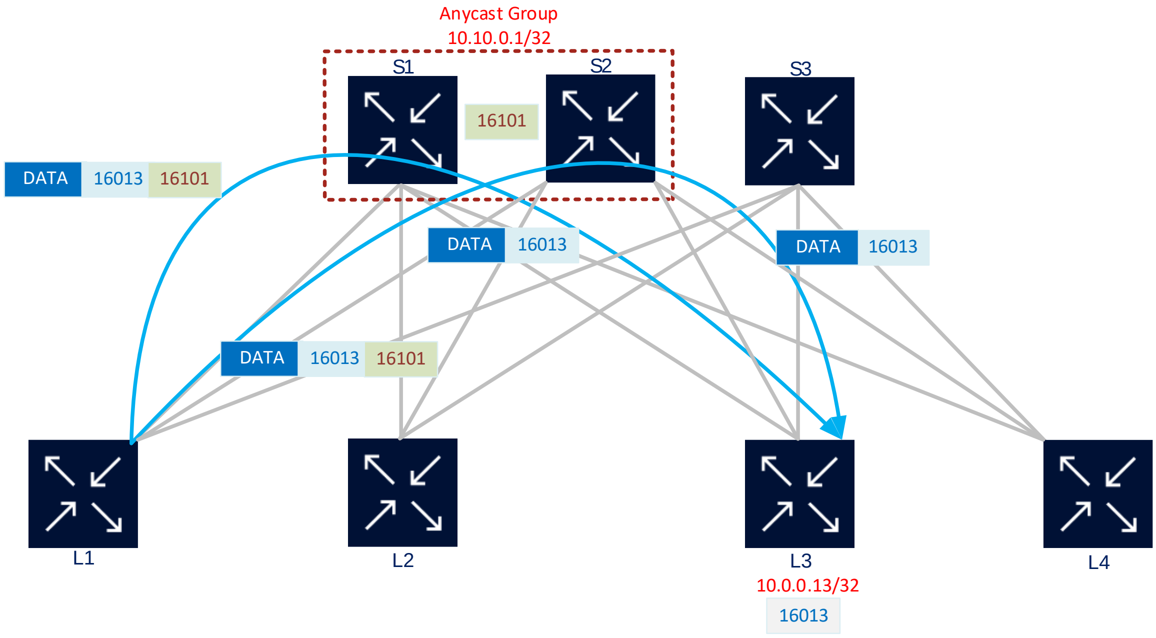


Figure 2-14: BGP Anycast SID forwarding

2.4.2 BGP Peer SID

A BGP peer segment is attached to a BGP internal or external neighbor sessions. These SIDs are local and are assigned to all peering sessions inside a node using the new address family BGP Link-State (BGP-LS), protocol defined in RFC 7752 [27]. Similar to the IGP Adjacency segment, RFC 8402 stated that a BGP Peer SID could be a global segment, which is not applicable for any vendor because these segments have to be distributed to all routers in the SR domain even to routers without BGP.

There are three types of BGP peer segments, all are used for SR Traffic Engineering and will be explained later in the TE chapter.

2.4.2.1 BGP Peer Node SID

This BGP peer segment is attached to a specific peering session, that is intended to override the normal BGP decision process and forward traffic using all ECMP paths available to that BGP neighbor.

Figure 2-15 shows a typical multi-peering schema in a service provider network, here in AS 1 router R1 have two eBGP sessions to router R2 in AS 2, and router R3 in AS 3. Router R1 generates local BGP Peer-Node-SID segments *40102* for router R2 and *40103* for router R3, these segments are distributed to the Controller or Path Computation Element (PCE) using BGP-LS. PCE is an external entity in charge of path computation because that has a full view of the network and the TE database. When a BGP peer-node segment as *40103* arrives to router R1, it will send to router R3 no matter what next-hop BGP calculates in normal conditions. For segment *40102* the process is the same, any packet with this peer node SID will be forwarded to router R2.

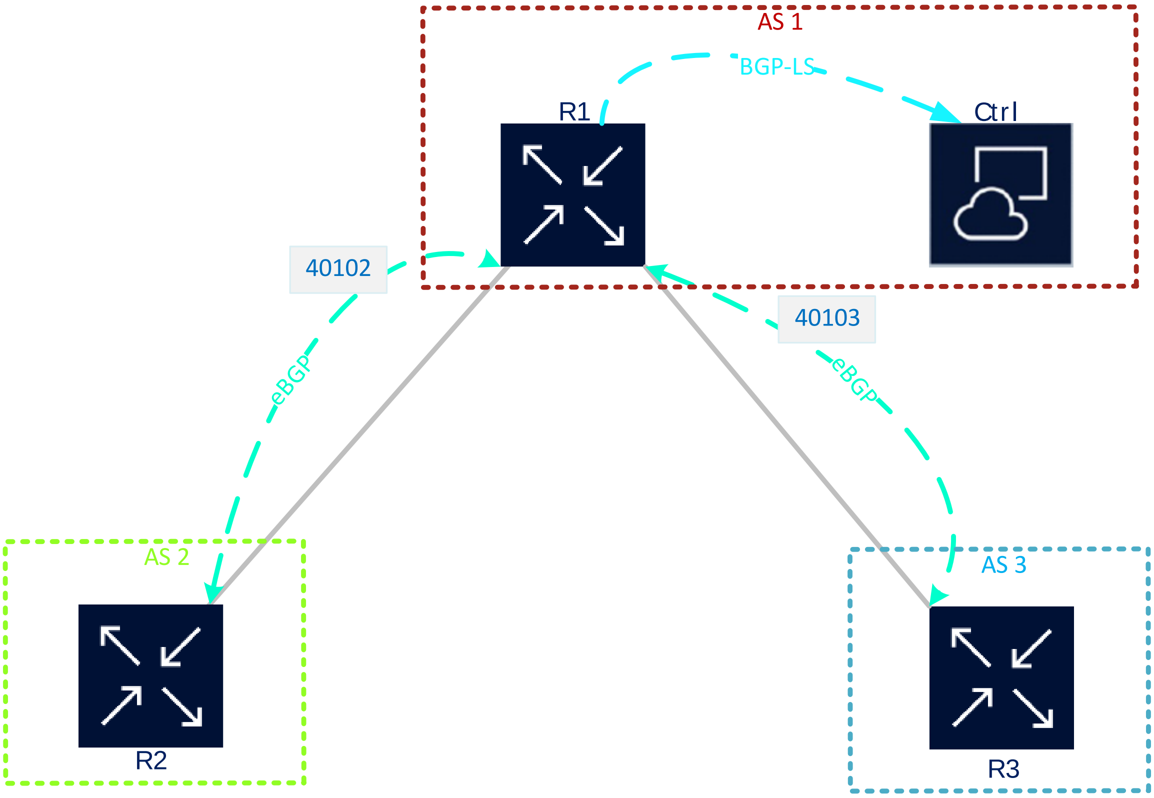


Figure 2-15: BGP Peer Node SID

2.4.2.2 BGP Peer Adjacency SID

An eBGP session established to a neighbor based on loopback interfaces may pass through multiple physical links [28], so a BGP Peer-Adj SID is attached to a specific link used by this session. A BGP Peer Adjacency Segment commands to forward traffic using a specific port to a neighbor. This segment also overrides the normal BGP decision procedure, so when a packet came with a BGP Peer-Adj SID that will be steering to the appropriate link attached to the segment even if there are better paths or ECMP paths.

In figure 2-16 there are two Autonomous Systems, in AS 1 resides router R1 which has an eBGP session using its loopback to router R2 in AS 2. Two links between both routers are used for redundancy and traffic load balancing. Router R1 generates a local BGP Peer-Adj-SID segment for each link, so segment *40212* is used for link 1 to router R2 and *40222* for link 2 to the same router. These segments are distributed to the Controller or Path Computation Element (PCE) using BGP-LS for path computation because that has a full view of the network and the TE database. When a BGP peer-node segment as *40212* arrives, in a header's packet, to router R1, it will be headed to router R2 using only link 1 no matter what next-hop BGP calculate in normal conditions. Traffic will not be load balanced using ECMP, though that two links have the same cost because the segment's instruction permits send a packet using only the first link. For segment *40222* the process is the same, any packet with this peer adjacency SID will always be forwarded to router R2 using only link 2.

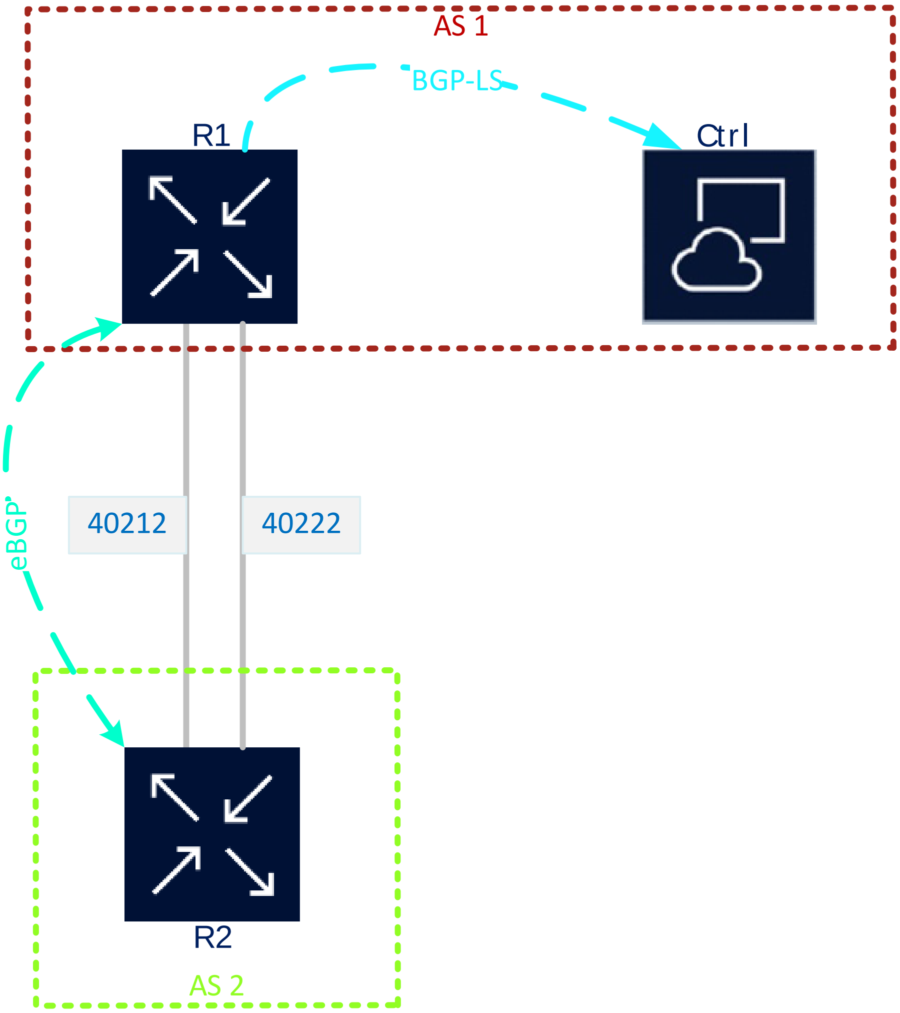


Figure 2-16: BGP Peer Adjacency SID

2.4.2.3 BGP Peer Set SID

This BGP peer Set segment is attached to a group of peers which form a set of neighbors with some common characteristic, so two or more nodes could be part of this set. A BGP Peer Set Segment instruction permits send traffic to nodes that form a group using all ECMP paths available. That segment is intended to override the normal BGP calculation process and forward traffic using all ECMP paths available to the BGP neighbors' group [22].

In figure 2-17 there are three Autonomous Systems, in AS 1 resides router R1 which has an eBGP session to router R2 in AS 2 and router R3 in AS 3. Routers R2 and R3 form a group of BGP peers. Router R1 generates a local BGP Peer-Set-SID segment for that group of routers, so the peer set segment is *40423*. These segments are distributed to the Controller using BGP-LS for path computation because that has a full view of the network and the TE database. When a BGP peer-node segment as *40423* arrives, in a header's packet, to router R1, it will be headed to router R2 and router R3 load balancing the traffic using ECMP. Packets with this Set segment behave the same way no matter what next-hop BGP calculates in normal conditions.

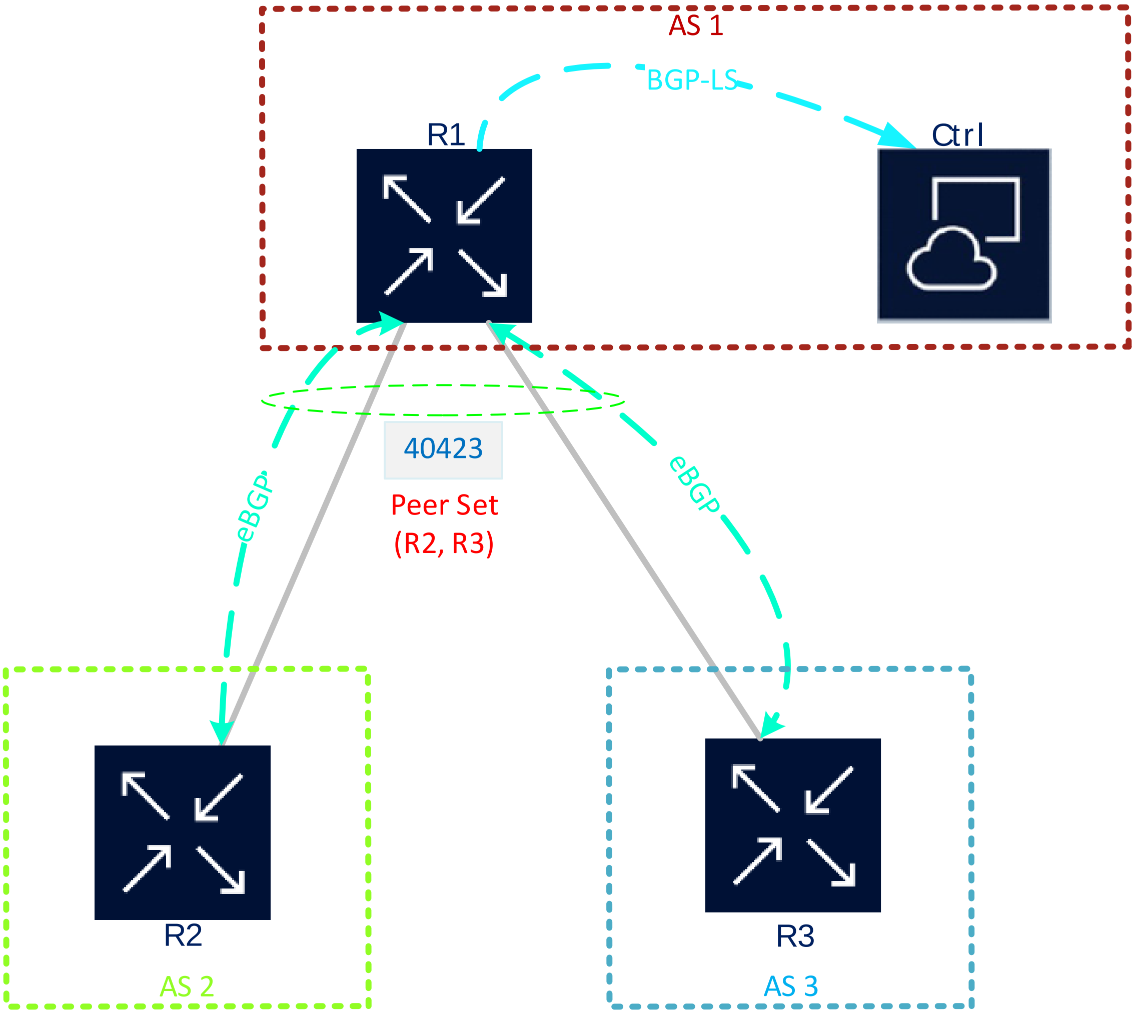


Figure 2-17: BGP Peer Set SID

2.5 IGP and BGP Combined

Different types of segments could be combined to forward traffic end to end using specific nodes and based on latency or bandwidth parameters. Traffic flow could cross through different domains, for instance, a packet could start in an IGP domain and then cross to an only BGP domain. This combination of different segments is used mainly in TE applications and TI-LFA FRR scenarios [22].

According to Ref. [22] a SID list to forward traffic end to end only needs a maximum of 3 or 4 SIDs to express TE or TI-LFA policies in real networks. That means a very detailed explicit path could be defined using a few segments because each segment in the SID list is attached to a significant node where a specific action is made to change an SPF path.

An example of creating an end-to-end path using a few segments could be seen in figure 2-18. The main goal in this example is traffic beyond node L1 need to reach equipment beyond router R6 using router R4's interface to R6 because that is low bandwidth and latency link. In DC network only BGP-SR is used, and in the service provider network, IS-IS with SR extensions enabled where the default metric is used (10) in most of the links except for R2-R4 (20) and R4-R6 (200). To forward traffic end-to-end only three SIDs are used, the way is detailed next: First, BGP Prefix-SID *16002* which instruct traffic to be forwarded using all ECMP paths to border router R2, those possible paths are used L1-S1-R2 and L1-S2-R2. Second, from router R2 an IGP Node-SID *16004* is used to reach router R4 using ECMP paths R2-R4 and R2-R1-R4, those paths have the same cost though the second one has an additional hop. Finally, IGP Adj-SID *30406* is used to get router R6, this segment instructs router R4 to use its R4-R6 interface even though it is not the shortest path to get R6.

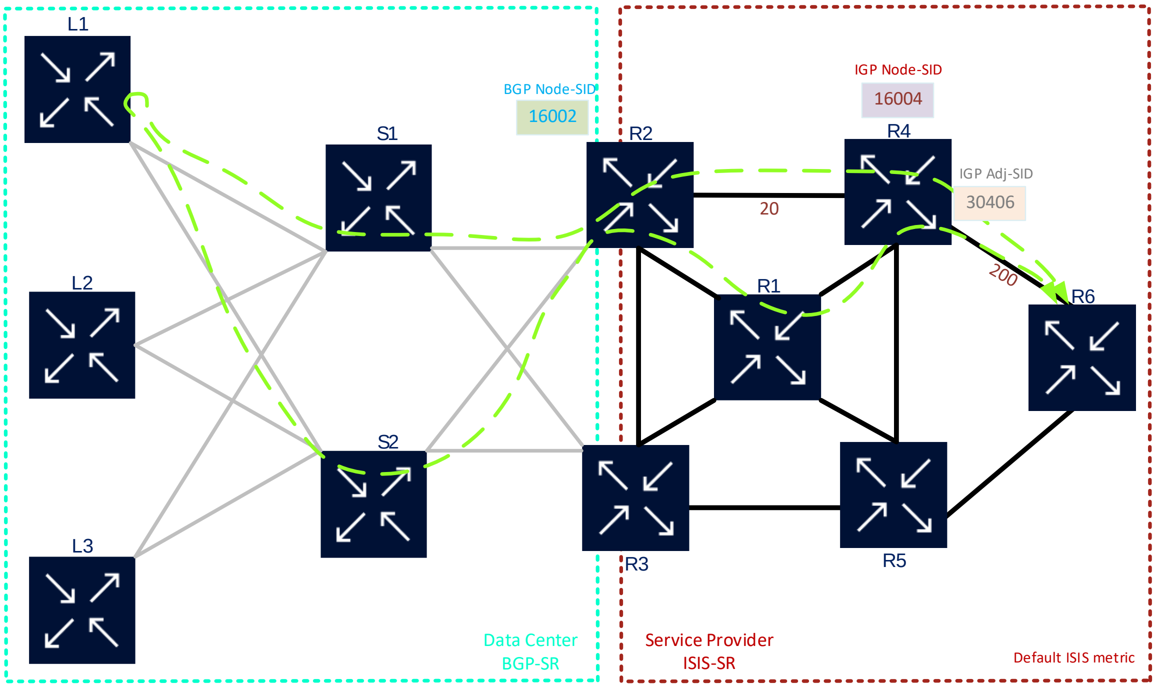


Figure 2-18: Combining different types of Segments

3 Segment Routing Data Plane

Segment routing can use MPLS or IPv6 as Data Plane, however in this document IPv6 data plane is not going to be analyzed in detail. SR IPv6 data plane or simply SRv6 is a very promising technology, but only for greenfield or newer implementations because in part that needs hardware support on each node handling the new SR header [29]. SRv6 also cannot be widely deployable because only a couple of vendors are investing in this technology [30] and the standardization process is still immature [31], so in a multi-vendor environment, SR IPv6 data plane is not an option.

3.1 MPLS Data Plane

As it already been said SR could run directly over MPLS data plain with a simple software upgrade to the nodes where Segment Routing is going to be deployed. In this case, a SID is a 32-bit header where the MPLS label is represented by the 20 rightmost bits of the segment, which is shown in figure 3-1. A segment list is a label stack where the Active Segment is the top label, so when the action represented by this segment is executed this label is popped from the stack [1].

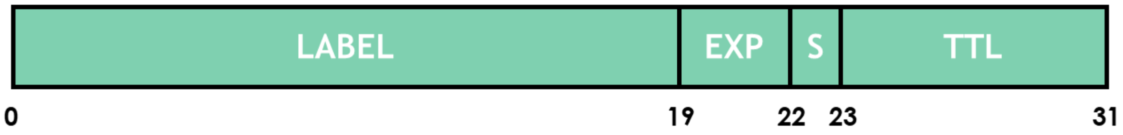


Figure 3-1: Segment ID in MPLS Data Plane [1]

RFC 3031, which defines MPLS architecture, also introduces the concept of Forwarding Equivalence Class (FEC). An FEC is a group of IP packets forwarded in the same manner, over the same path, and with the same forwarding treatment. In MPLS, FECs can be defined based on destination IP prefixes, and other administrative criteria. FEC lookup determines next-hop and labels the source router pushes onto the packet, this process is made only at the ingress Label Edge Router (LER) on incoming data packets. The label imposition process is equivalent to select a specific Label Switched Path (LSP) previously signaling for a label distribution protocol as LDP or RSVP-TE. Then LSP is selected, the label is only swapped on the Label Switching Router (LSR) along the way until its destination [4] [32].

In SR, the mapping of a prefix to an FEC remains the same and the only difference is that the label corresponding to this FEC is derived from its Prefix Segment, which is signaled by any IGP protocol or BGP, instead of being signaling via traditional hop-by-hop label distribution protocol like LDP or RSVP-TE.

3.2 Segment Routing Global Block

In an MPLS architecture, SRGB is the set of local labels reserved for global segments or SIDs in a particular node, where a SID is a label value or an index in a label block. SRGB is used for all global segments, that means Node-SID and Anycast-SID allocates labels using this block. However, Adjacency-SID is a local segment that uses a different block sometimes called SRLB or Segment Routing Local Block.

Prefix SIDs need to be globally unique within an SR domain, but SRGB is a range with local significance and network could be multi-vendor. In that case is hard to agree on one common MPLS label range to allocate the SRGB, to overcome these two techniques could be used to guarantee a global segment is unique [1] [3] [33].

3.2.1 SRGB Indexing

Indexing allows for the use of different SRGB's or SR label range, but all routers in the SR domain are expected to configure and advertise the same Prefix-SID index range for a given IGP instance. The label value used by each router to represent a prefix 'n' can be local to that router by the use of an offset label, referred to as a start label. To calculate the local label value, the next formula is applied [1]:

$$\textit{Local Label (for Prefix-SID)} = \textit{(local) start-label} + \{\textit{Prefix SID index}\}$$

Equation 3-1: Local Label Value

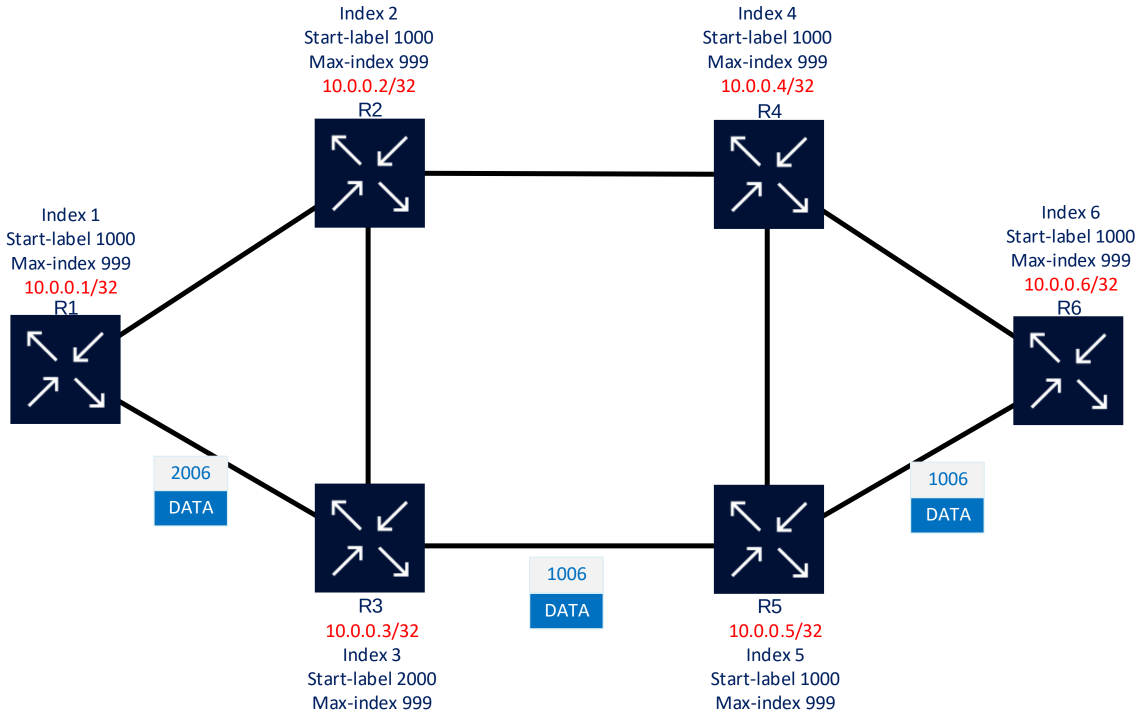


Figure 3-2: Indexing Prefix-SID

For example, all routers in the figure 3-2 have an SRGB of *{1000-1999}* except for R3, which has an SRGB of *{2000-2999}*. All routers have a Prefix-SID index range of *{1-*

999} and each SR router in the domain defines a start point in the SRGB, better known as *start-label*, and advertises an offset label for the *Prefix-SID Index*. To obtain a *Local Label* for and prefix the formula above is used. In this case, path R1-R3-R5-R6 is preferred and router R6 advertises its loopback *10.0.0.6/32* with a prefix index of *6*, its local label for that prefix is *{1000+6} = 1006*; for router R5's local label for the prefix is also *{1000+6} = 1006*; and router R3's local label for the prefix is *{2000+6} = 2006*. To reach router R6 with an SR tunnel, router R1 pushes on label *2006* and forwards to router R3; this router swaps *2006* for *1006* and forwards to router R5 where label *1006* swaps for the same *1006* and finally forwards to router R6 [33].

3.2.2 SRGB Absolute SID

The use of absolute SID values requires a single consistent SRGB and start-label on all SR routers throughout the IGP domain, and that is not always possible, particularly in multi-vendor environments. This approach is preferred for many operators due to the operational simplicity of using the same SRGB and label value along with each node inside the SR domain. This type of implementation may not be possible if consistent SRGB and start-label cannot be identified with the installed base. If this is the case, the indexing method must be used [33].

In figure 3-3 is shown an SR domain where the absolute domain approach is used. First, router R6 advertises its loopback address 10.0.0.6/32 with Node-SID 600. Then, to forward traffic from router R1 to router R6, router R1 pushes on label 600, which is the Node-SID for router R6 on top of the label stack. Again, it is assuming a preferred path non-ECMP using R1-R3-R5-R6. That means Label (SID) does not change hop-by-hop until it reaches router R6 [33].

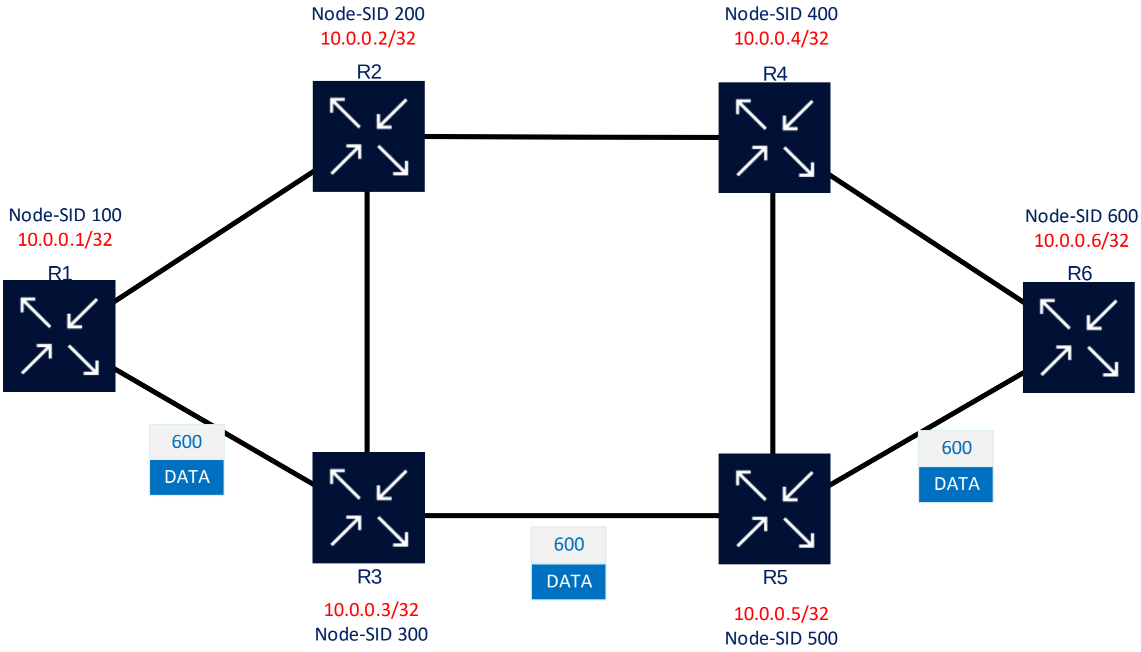


Figure 3-3: Absolute Prefix-SID

In the end, absolute or homogenous SRGB is only a special case of indexing where each node calculates the same label for Prefix-SID. SR was designed to operate with a consistent SRGB and to ensure that a default SRGB block of *{16000-23999}* was proposed, but it still is possible to configure a custom range in each node [3].

3.2.3 SRGB Size

The SRGB size is one of the first parameters to be configured in a router when SR is being implemented, which will be explained in detail in the practical part. This parameter indicates the maximum number of Prefix-SID that a router could handle or in other words the number of labels in the block, so that must be defined in the design process with loopbacks assignment and Node-SID as discussed before in point *2.3.1*. Even though the size of SRGB could increase after the implementation phase, that is not recommended because it is a local parameter that needs to be changed in all nodes running SR to allocate the new size [34].

In order to avoid make size changes in the future, adequate planning needs to be executed previous to SR deployment. To calculate the size of SRGB all global SID has to be considered, which includes a Node-SID for each router in the network, any other Prefix-SID tie to different loopbacks and any Anycast-SID. Enough margin of free segments for future growth also needs to be provided. A best practice is maintaining the same size of SRGB along with all nodes because the smallest SRGB size limits the maximum SID index. For instance, if a node receives a SID index of 2500 but its SRGB size is only 2000 then the incoming SID will be invalid, and that leads to packets drops or data wrongly routed because this router will not have a valid SID index for its neighbour [34].

In Nokia routers, there is not a default range for SRGB, so this value has to be explicitly configured to reserve from the dynamic label-range in each router from *18432* to *524287* [33]. On the other hand, Cisco IOS XR routers have a default label range from *16000* to *23999* or 8000 SIDs, this range could be increased to near a million limit [34]. Although RFC 8402 stated a multi-range SRGB with discontinuous blocks is possible [20], both vendors only support a continuous label range.

It is easy to see that in a multivendor environment use the same label range is not always possible, but the same SRGB size and continuous label range must be used in every single router inside the SR domain to make Prefix-SID indexing possible.

3.2.4 SRGB Label Range

In most cases, in a multivendor network there will be different SRGBs and indexing Prefix-SID is used, but the index range must be the same to ease SR administrative tasks. For instance, if a network has Cisco and Nokia routers, the Cisco portion could use the default label range and start in *16000*; in Nokia, routers range might be set from *32000* to *39999,* so each index can be mapped efficiently. Figure 3-4 shown how this mapping works, there all routers used SRGB range of *16000-23999* except for router R3 and R4 which uses SRGB range of *32000-39999*. When router R6 propagates its SID index *6* attached to prefix *10.0.0.6/32,* each node uses SID index *6* to map a correspondent local label value. For routers R1, R2 and R5 this prefix segment is *16006*, and for routers, R3 and R4 is *32006*. In this example, when a packet from router R1 has router R6 destination, label push to router R3 is *32006* and to router R2 is *16006*. Then, in router R3 this label is swapped to *16006* and is send to router R5, and the label is swapped in *32006* from router R2 to router R4. Finally, from routers, R5 and R4 labels are swapped to *16006* before sent to router R6 [34].

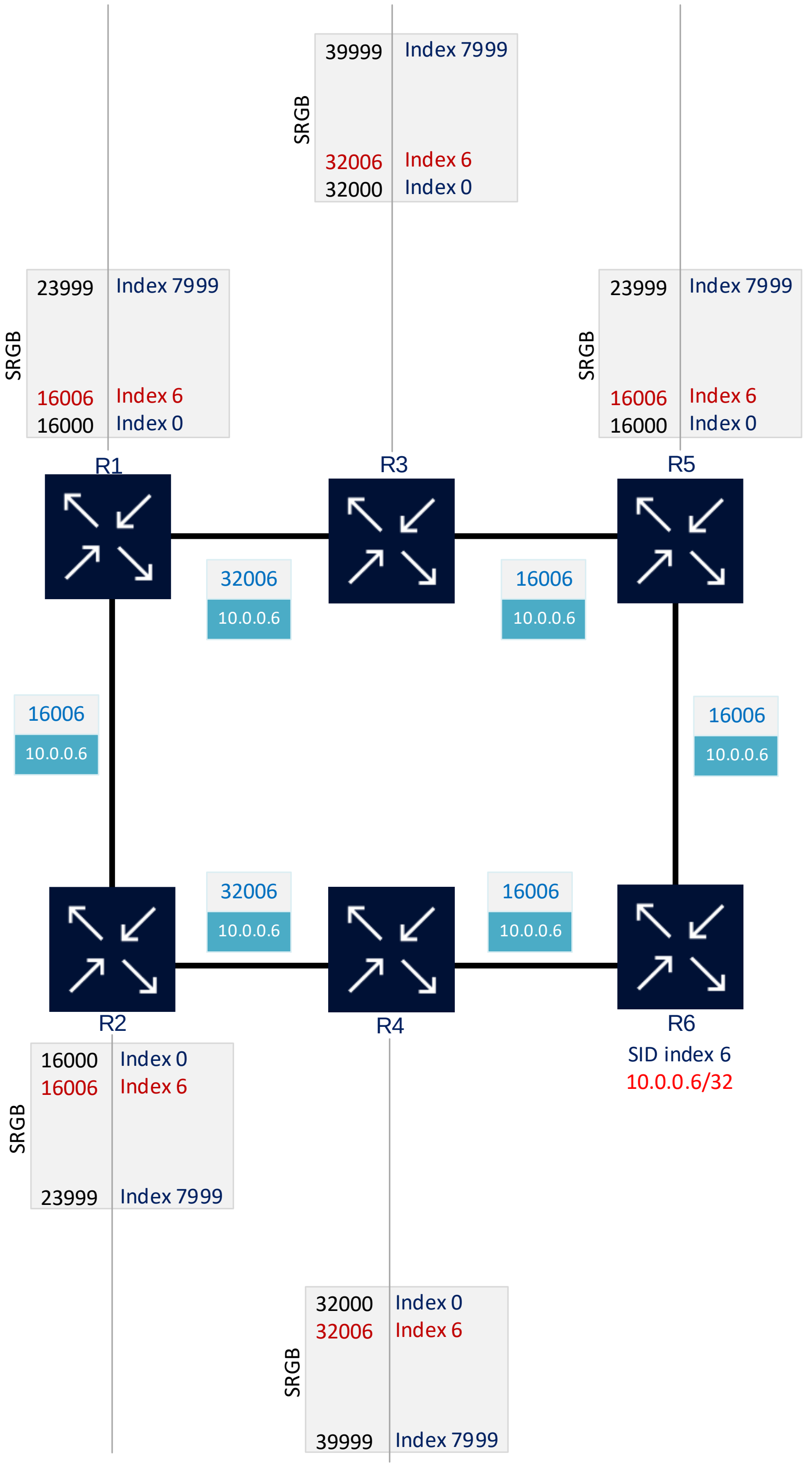


Figure 3-4: SRGB Allocation with different ranges

Use the same SRGB for all routers in the SR domain is a strong suggestion from RFC 8402's authors because these simplify operations and the troubleshooting end-to-end [34]. Under this premise the best approach is changing Cisco default SRGB to a common value between vendors because the default value is not feasible for all of them, par example Nokia SRGB could be allocated from *18432* so *16000* value is out of this range. Any particular case could be individually analysed but a good value for SRGB might be *32000* to *39999*. Figure 3-5 shows the same case of figure 3-4, the only difference is that all routers use the same SRGB, in this case, {*32000-39999*}, so a packet always uses the same label along with all routers until it reaches its destination [34].

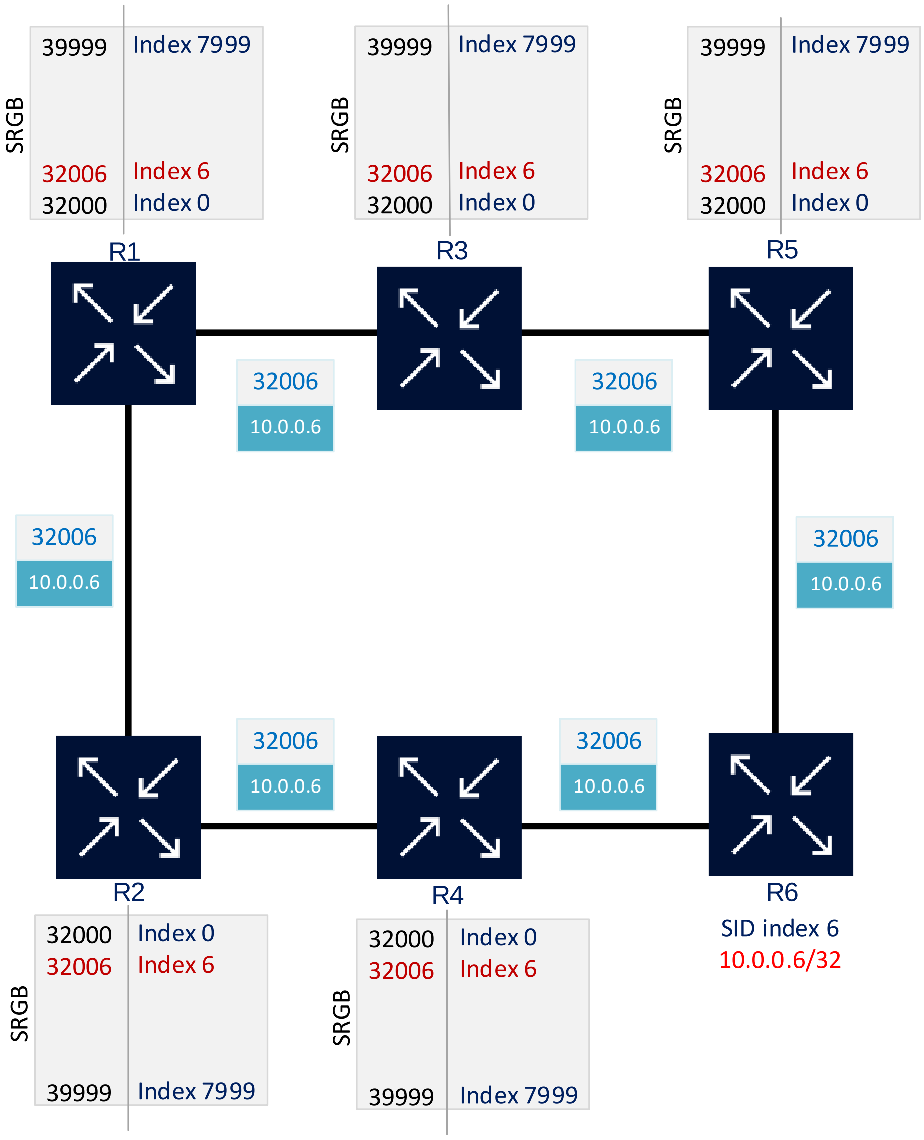


Figure 3-5: SRGB allocation with the same range

3.2.5 SRGB for Anycast-SID

Even though it is a good practice to use the same SRGB for all routers in an SR domain that is not mandatory because detailed planning can avoid any problem indexing Node-SID in each node. Yet, in the Anycast-SID case, all routers that are part of the anycast group must have the same SRGB. Figure 3-6 shown what problems arise when anycast nodes use different SRGB, in this example all routers used SRGB {*32000-39999*} except router R3 that uses {*64000-71999*}. Router R1 wants to send a packet to router R6 using anycast group formed by routers R2 and R3, so two labels have to be pushed in the packet header: Anycast-SID with SID index *11* and Node-SID with SID index *6*. As routers R2 and R3 have different SRGB, router R1 pushes label *32011* to router R2 and label *64011* to router R3 as Anycast-SID. Also, Node-SID for router R6 has to be added, but there are two possible values *32006* and *64006* and that makes it impossible to have different SRGB in routers which are part of an anycast group [34].

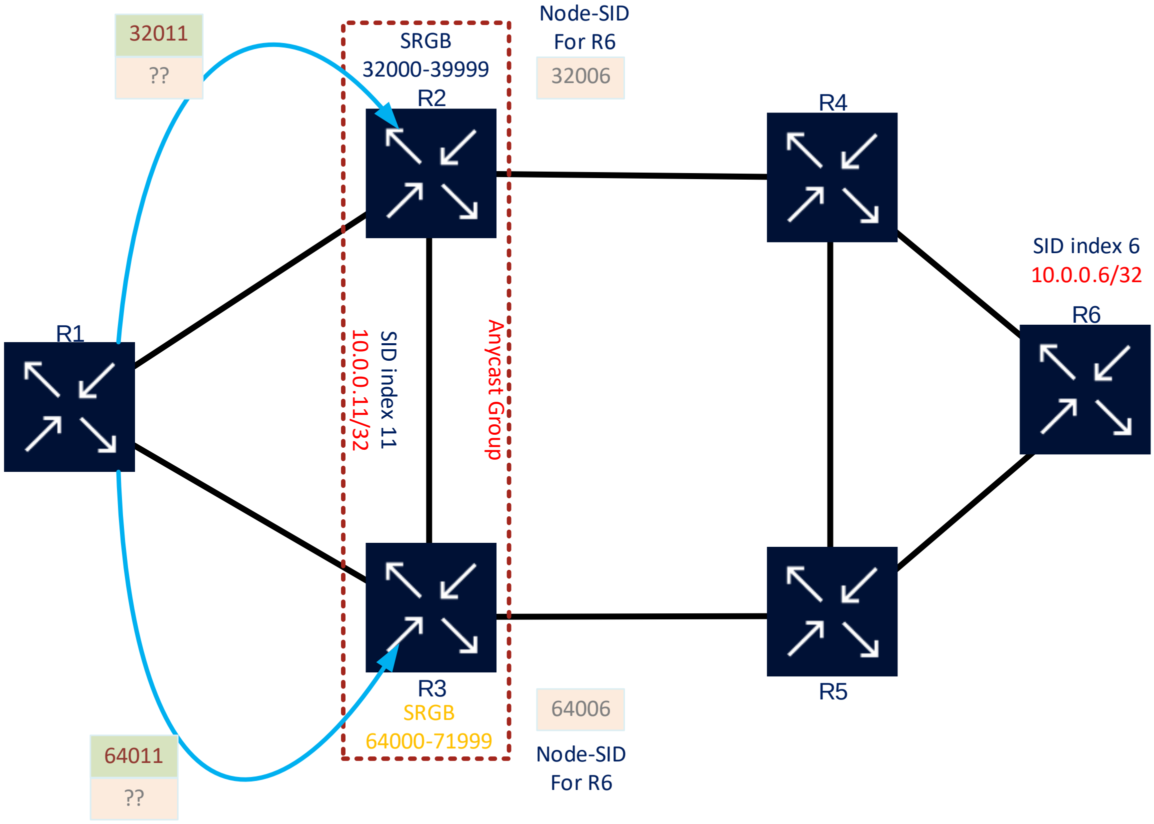


Figure 3-6: Anycast-SID with different SRGB

On the other hand, there is no problem when all nodes inside an anycast group using the same SRGB, but this range is different from other routers in the network. Figure 3-7 proved this statement, in this example, all routers used SRGB {*32000-39999*} except routers R2 and R3 that use {*64000-71999*}. Again, router R1 wants to send a packet to router R6 using anycast group formed by routers R2 and R3, so two labels have to be pushed in the packet header: Anycast-SID with SID index *11* and Node-SID with SID index *6*. As routers R2 and R3 have the same SRGB, router R1 pushes label 64011 to both routers as Anycast-SID. In addition, Node-SID *64006* for router R6 has to be added, then in routers R2 and R3 traffic use both ECMP paths to reach router R6, popping first the anycast label. Routers R4 and R5 only receive label 64006 to reach router R6.

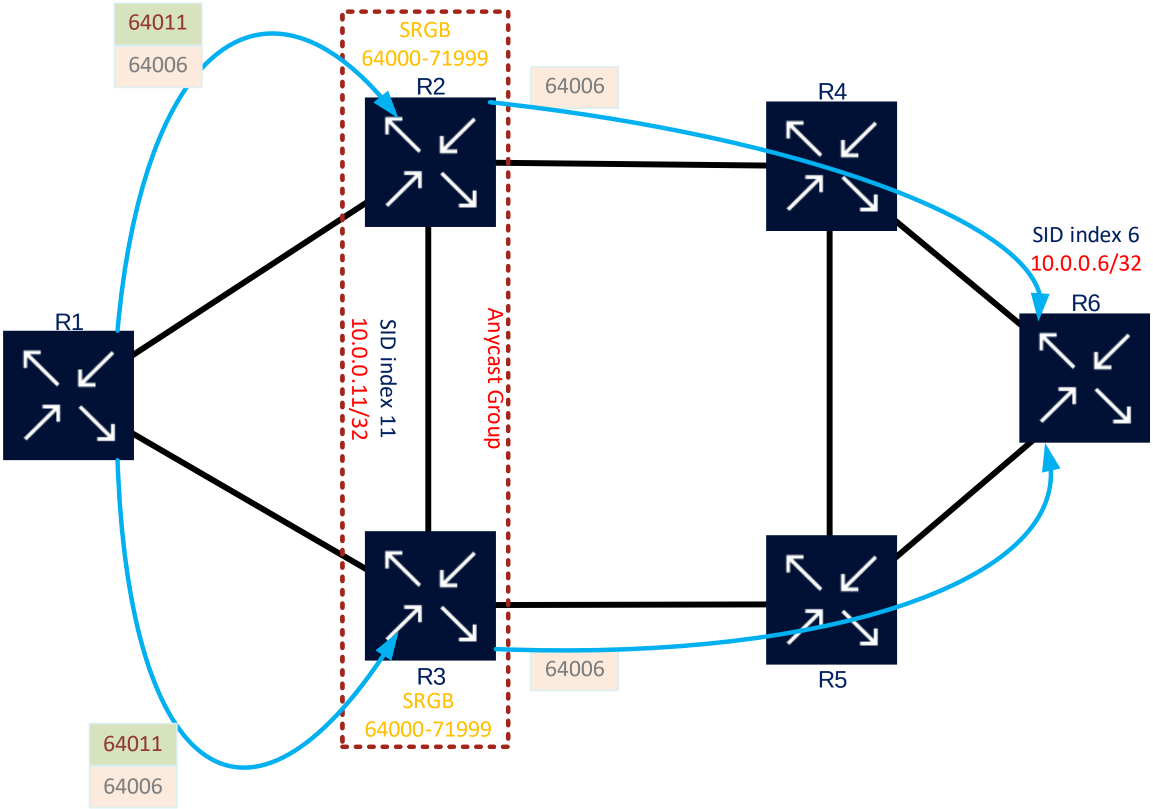


Figure 3-7: Anycast-SID with the same SRGB

3.3 SR MPLS Label Operations

As already described in point 2.2.1, segment list operations are directly applied over traditional MPLS label stack operations. Table 3-1 describes each MPLS forwarding operation and its mapping with the SID list operation.

Segment List Operation	MPLS label Operation	Description
PUSH	PUSH	Insert a label in the packet header as the outermost label.
CONTINUE	SWAP	The topmost label is replaced with a different label.
POP	POP	Eliminate the outermost label form the label stack.

Table 3-1: Segment List vs. MPLS label operation [3]

3.3.1 PUSH

A Prefix-SID label is pushed in a pure IP packet header if destination prefix is installed in the FIB and also has a Prefix-SID associated with this prefix. Another requisite to impose a label is that the next-hop router needs to be SR enabled to handle the prefix segment label [3].

Figure 3-8 shown how a label insertion is made in an SR network that uses an MPLS forwarding plane. In this example, all routers are SR capable and LDP is not enable in the network, routers also use the same SRGB {*32000-39999*}. Only path R1-R3-R5-R6 will be used because all links use metric 10 except link R1-R2 which metric is 200. First, a pure IP packet arrives to router R1 with destination address *10.0.0.6*, so router R1 matching this address with prefix *10.0.0.6/32* installed in FIB and there is a Prefix-SID *32006* tie to this prefix. Then, label *32006* is imposed on the packet and forward to the shortest path to router R6.

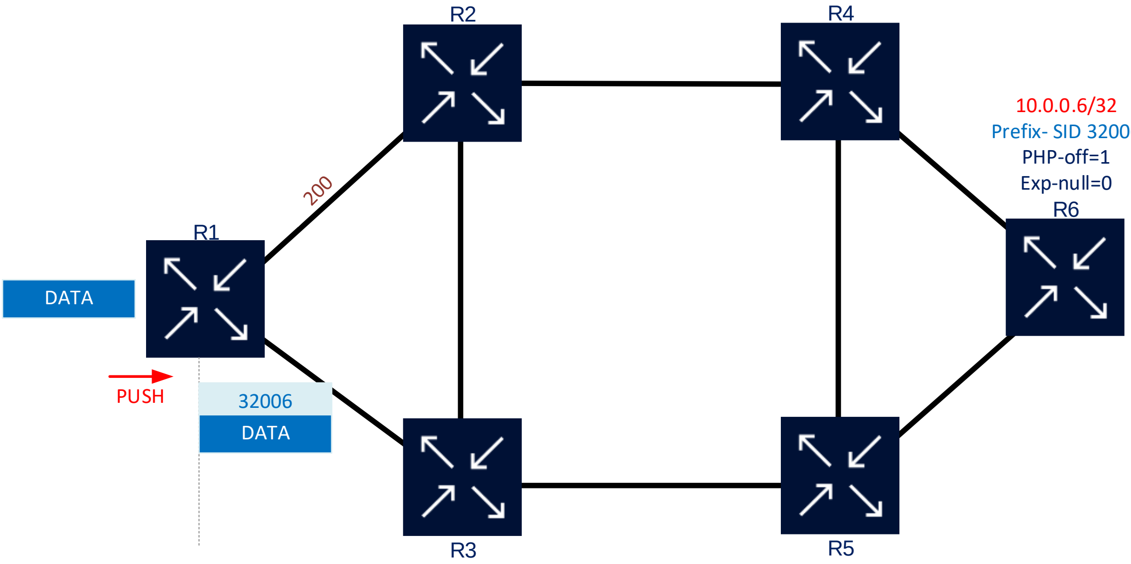


Figure 3-8: MPLS label Push

3.3.2 SWAP

This operation is executed in all LSR along the path until the final destination, here incoming label is swapped for an outgoing label and forwarded to the respective out interface. Therefore, LFIB contains an entry with the arriving label mapped to an outgoing label and outgoing interface. If different SRGB range is used the outgoing label will be different from the incoming label, but if an identical SRGB is used in all nodes both labels will be the same.

Figure 3-9 shown this operation continuing with an example from figure 3-8. Here router R3 also uses SRGB {*32000-39999*} as router R1, so its Prefix-SID for prefix 10.0.0.6/32 is the same as router R1 *32006*. In this router there are no ECMP paths, that means inside its LFIB there is an entry with incoming label *32006*, outgoing label *32006* and out interface R3-R5. As a homogenous SRGB model is used in this network, router R3 will swap label *32006* with label *32006* and the packet will be forwarded using link R3-R5.

If Penultimate Hop Popping (PHP) is not used in the network, the swapping process in router R5 will be the same as in router R3. However, two other possibilities will explain in the next point.

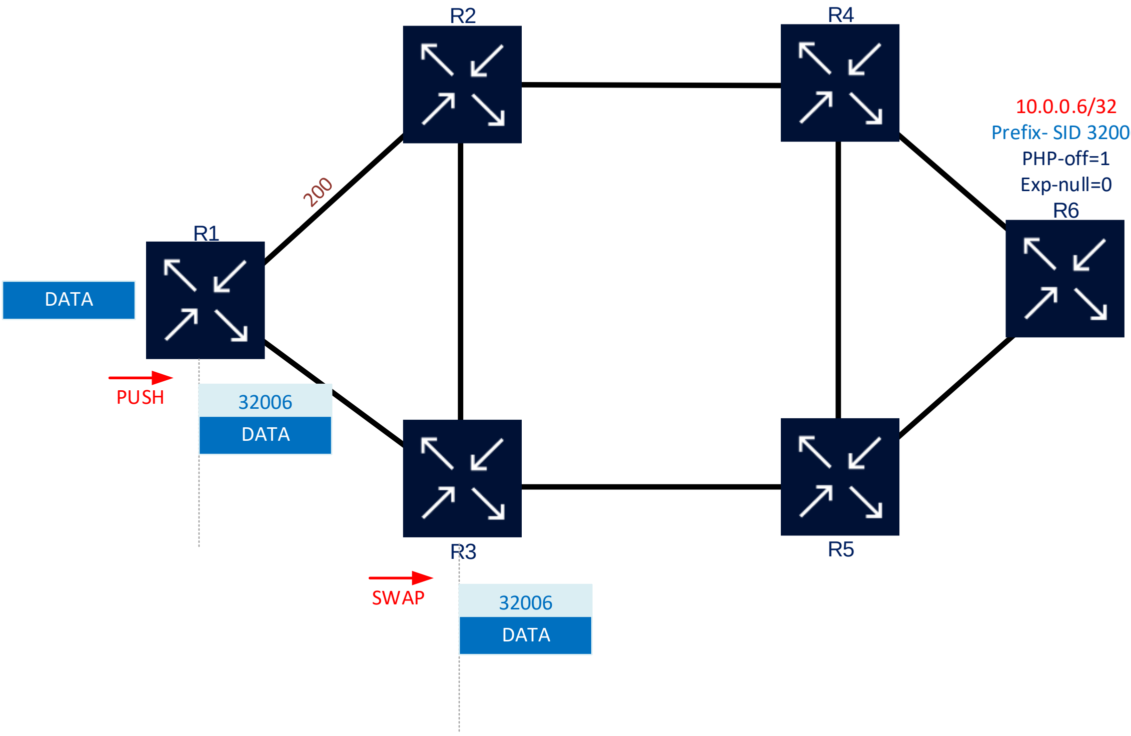


Figure 3-9: MPLS Label Swap

3.3.3 Penultimate Hop Router

The destination router controls what MPLS operation the penultimate router must execute to incoming label, three options are available: Swap with Prefix-SID label, Pop or Swap with Explicit-null label. Each option has some advantages and drawbacks which will be presented below.

3.3.3.1 Swap with Prefix-SID Label

In this model, the destination router which is the Prefix-SID originator demands receive its Prefix-SID label as an incoming label. In general, this is applicable when experimental bits (EXP) in the MPLS header are used to applied Quality of Service (QoS) end-to-end. MPLS label is shown in figure 3-1 where bits 19 to 22 are used for experimental bits. If the MPLS label would be popped in the penultimate router QoS information encoded in these bits will be lost and a complex mapping to DSCP might be needed to overcome this problem [35].

One drawback of this model is the final router needs to make a double lookup to each packet, first in LFIB and then in FIB. This behavior might have an impact on the forwarding performance; however, this statement was true when routers were very limited in memory and Forwarding Processors were not as specialized and powerful as now. Therefore, sometimes it is better to guarantee a correct treatment of different traffic flows using QoS than make some savings in tables lookups [4] [35].

Cisco traditionally avoids this model also call Ultimate Hop Popping (UHP), so when a packet arrives with the local Prefix-SID it will be dropped [3]. On the other hand, Nokia uses this model by default, so an interoperability problem could arise if this parameter is left unchanged in both vendors in the implementation.

Figure 3-10 shown how this model works, when router R6 distributes its Prefix-SID ask to router R5 to send its Prefix-SID label that means flag for PHP-off is enabled. Then, router R5 will swap incoming label *32006* for the same label, and the packet will be forwarded to router R6 using the interface R5-R6. EXP bits will be copied from incoming labels to outgoing labels preserving this way the same QoS treatment along with all network links.

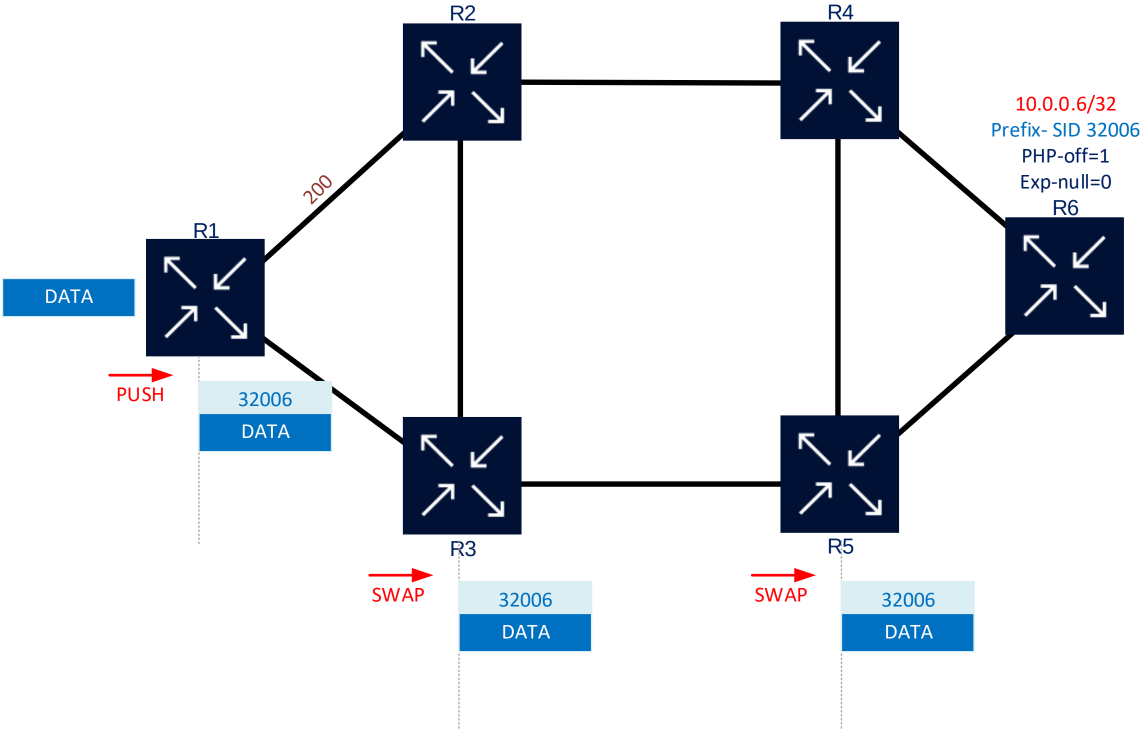


Figure 3-10: Swap in Penultimate router

3.3.3.2 *Pop Label*

In Penultimate Hop Popping or PHP model, the Prefix-SID originator router sends a request to its neighbors to pop the outermost label before sending it to the packet. This model is the complete opposite of the previous model because PHP advantages are drawbacks for UHP, and vice versa. That means the goal for this model is to avoid a double lookup in each packet because LFIB has not to be used and the only lookup will be in FIB before delivering the packet to its final destination. However, if EXP bits in MPLS label are used for QoS this info will be lost and all incoming traffic flows would be treated as Best Effort [3].

This model was created by Cisco and always have been the default behavior for its equipment. Other vendors like Nokia use UHP as default behavior, so in a multivendor environment this parameter is the paramount importance.

Figure 3-11 shown that router R6 ask to R5 for popping Prefix-SID label using parameter PHP-Off disable. In this example, when an incoming packet gets router R5 with Prefix-SID label *32006*, this router lookup it's LFIB and found a POP operation has to be applied to this packet. Router R5 will send this packet without top label using interface R5-R6 towards router R6.

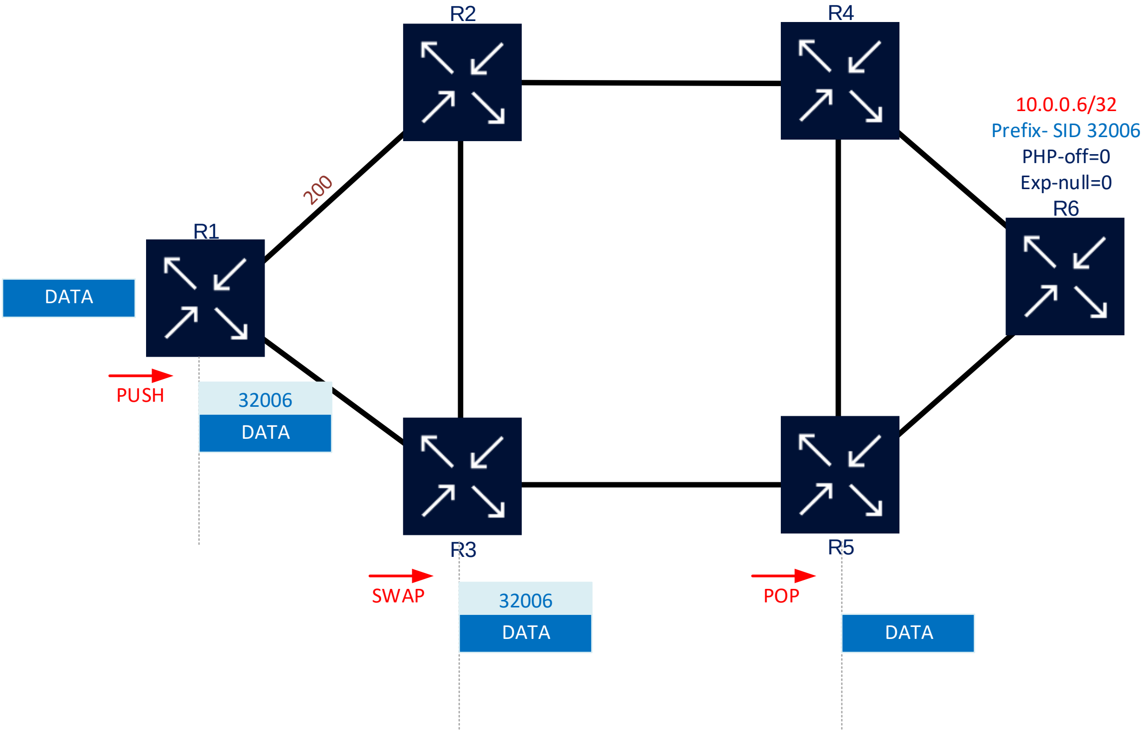


Figure 3-11: PHP behavior

3.3.3.3 Swap with Explicit-Null Label

This option avoids penultimate router send to destination router its Prefix-SID label, but the QoS information coded inside EXP bits is preserved until the last node. Here the originator Prefix-SID router asks for an explicit-null label to the penultimate node, so when this node receives a packet with the last Prefix-SID label that swaps this label to Explicit-Null label. This is only a reserved label whose value is *0* for IPv4 and *2* for IPv6 [3].

This model has the same drawback of UHP, which means the final router needs to make a double lookup to each packet, first in LFIB and then in FIB. Yet, as the label is swapped EXP bits are copied from the incoming label to Explicit-Null label, and no QoS info is lost in this process. Another advantage of this model is avoiding the packet would be dropped because vendors like Cisco by default discard any packet coming with the local Prefix-SID label, but a packet with the Explicit-Null label will be preserved [3].

In figure 3-12 originator Prefix-SID router R6 turned on the Exp-null flag in order to ask router R5 send and Explicit-Null Label when that receives router R6's Prefix-SID. Then, router R5 will swap incoming label *32006* for the Explicit-null label in this case *0* because prefix *10.0.0.6/32* is IPv4. The packet will be forwarded to router R6 using the interface R5-R6. EXP bits will be copied from incoming labels to outgoing labels preserving this way the same QoS treatment along with all network links, and no matter what vendor is router R6 this packet will not be discarded.

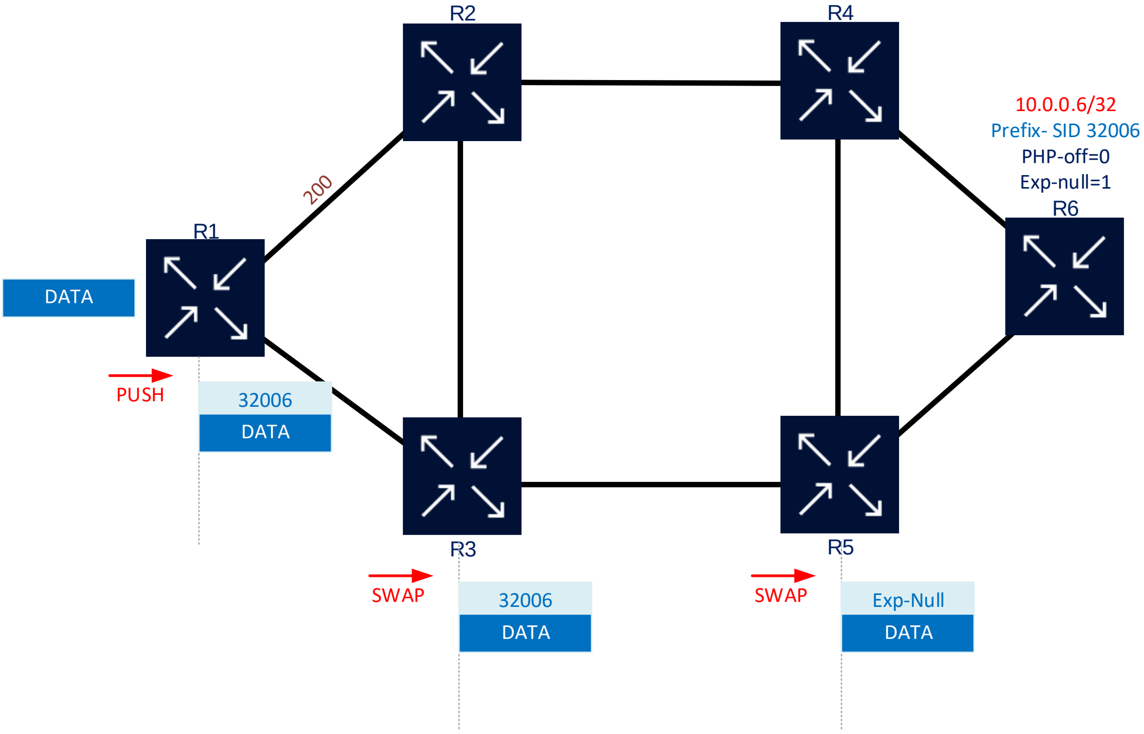


Figure 3-12: Swap Explicit-null

3.3.4 Last Hop Router

In figures, 3-8 to 3-12 router R6 is the last router, that will make lookups to LFIB and FIB or only FIB depending on what model is used in the penultimate router before sending a packet to its final destination. In most of the cases, there will be more than a label in the label stack after Prefix-SID had been popped, and these labels generally correspond to services. That means the packet will be forwarded to the respective service VPN L2 or L3 [3].

3.4 Unlabeled Packets

In some cases, a router inside a Segment Routing domain could be Non-SR capable or LDP enabled, but an LSP might still be built using packet unlabelled. The basic premise of unlabelled operation is forwarding a packet as unlabelled if there are not label for a prefix, also a packet could be forwarded if the incoming packet header has only one label or is pure IP otherwise packet will be dropped. In some cases, the “unlabelled” tag may represent an erroneous condition in the LSP path [3].

Figure 3-13 shown unlabeled packets behavior in two different cases, in this network router R3 is non-SR or LDP enable, so router R1 will forward unlabeled packets by R3 path. In the first case, router R1 is going to forward a packet with destination router R5. If this packet is unlabeled, router R5 will forward this packet as pure IP towards router R3 and then router R5. If the incoming packet has label *32005* and no more label resides in the label stack, again the packet will be forwarded as pure IP after stripping off the incoming label. Finally, if the incoming packet has label *32005* but there are more labels in the label stack, this packet will be dropped, and an error will be generated [3].

In the second case, router R1 is going to forward a packet with destination router R6 having two ECMP paths available. If this packet is unlabeled, router R5 will forward this packet as pure IP towards router R3 also that will impose label *32006* and then forward this packet to router R2. If the incoming packet has label *32006* and no more label resides in the label stack, again the packet will be load balancing and forwarded as pure IP to router R3 after dropping the incoming label. In router R2 direction label will be swapped to *32006* and headed to router R2. Finally, if the incoming packet has label *32006* but there are more labels in the label stack, only path through router R2 will be used because path over router R3 is considered unavailable in this case [3].

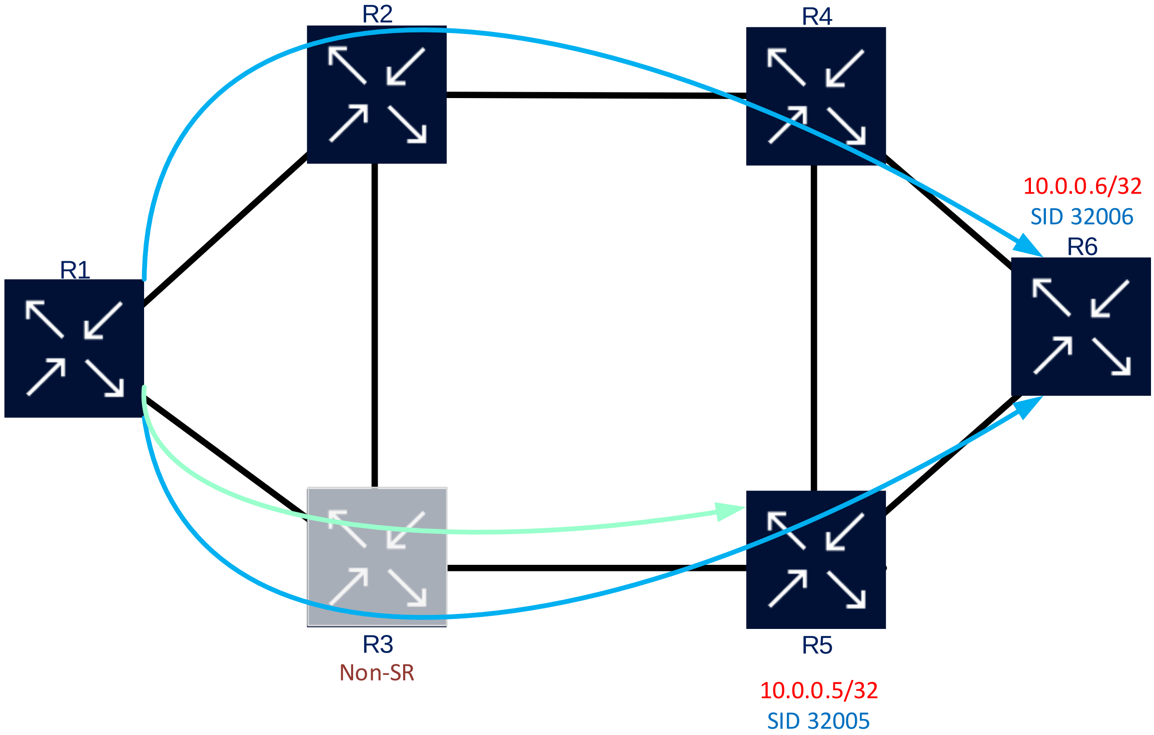


Figure 3-13: Unlabeled packets

3.5 MTU Considerations

In most cases, the Maximum Transmission Unit (MTU) in an Ethernet link is 1514 bytes, so an 1500 bytes packet payload could traverse this link without fragmentation. The additional 14 bytes coming from the Ethernet header, 6 bytes for source physical address or MAC, 6 for destination MAC and the last 2 bytes for type/length field. Inside this 1500 bytes limit a packet IP or MPLS must fit including its header, so the real capacity is decremented by 4 bytes for each label in MPLS case. That means default MTU must be incremented to make room to MPLS header and leave payload maximum in 1500 bytes [3].

How many must be increased depends on the maximum number of labels to be used in a network, for instance, if 4 labels are going to be used 16 bytes need to be added to 1514 bytes MTU, which means 1530 bytes. That means 1530 bytes is the minimum MTU to be configured in an interface Ethernet to have a maximum 4 label stack. Some vendors like Cisco use the IP MTU concept in addition to L2 MTU where both values have to be configured to work properly, but other vendors as Nokia only use L2 MTU value to derivate maximum label stack depth and payload in a link.

In current networks where Fibre Optics is used as transport links, it is a good practice to configure the value of a Jumbo frame of 9212 to prevent any dropping packets problem related to the MPLS header. This also eases the implementation of new techniques as FRR using TI-LFA or several TE tools because packet header size is not a problem anymore [3].

4 SR IGP Control Plane

In a traditional MPLS control plane an additional protocol is needed to distribute label information, LDP or RSVP-TE, however, SR uses current IGP running in the network to advertise IGP segments. That means current topological and forwarding database distribution will continue the same adding IGP segments information. Therefore, the forwarding database also will contain Prefix-SID and Adj-SID labels, so any traffic might be forwarded using these SIDs information [36]. Only link-state protocols have extensions to distribute IGP SIDs between each router inside an SR domain since these are the only option in modern networks. OSPF [36], OSPFv3 [37] and IS-IS [38] have their extensions defined in RFCs recently approved, so interoperability inter-vendor is guaranteed.

SR support for OSPFv3 defined in RFC 8666 has not gained traction along with the vendors because this protocol is almost identical to traditional OSPFv2, only the IPv6 support is the biggest advantage from OSPFv3 over OSPFv2. Not many live networks had deployed version 3 of OSPF because IS-IS has IPv6 native support. To sum up, SR extensions for OSPFv3 are not going to be studied here because they are similar to OSPF SR extensions and are not supported for most of the vendors.

4.1 IS-IS Extensions for SR

IS-IS has native support for IPv4 and IPv6, so its extensions for Segment Routing also have support for both address families and level-1 and level-2 routing [36]. Triplets Type Length Value (TLV) is used to identify the type of information encoded in the IS-IS update frame, that permits to IS-IS carried different kinds of data making this protocol extendable. IS-IS uses layer 2 frames to send these updates where those frames always have the same header, but the TLV varies according to the type of update being sent. Figure 4-1 shows an ISIS packet and its different sections [40].

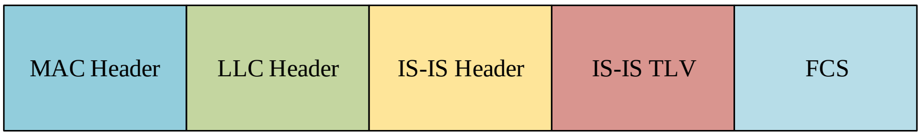


Figure 4-1: IS-IS packet

There are some TLVs consider essential for IS-IS such as Area Address, IS neighbors, IP interface address, Authentication and Protocols Supported. These TLVs need to be present in all IS-IS routers inside a router domain [40]. Yet, there are other TLVs that are not always present in all implementations, in this case, a router which not support a specific TLV must only ignore this TLV and continue with remain TLVs. If a sub-TLV is not supported by a router also that must ignore the sub-TLV [41]. This behavior prevents any disruption in a routing domain where routers with different capabilities are present.

4.1.1 IS-IS Router Capability TLV

In the SR case, IS-IS uses the Capability TLV, with type 242, to indicate support for Segment Routing. This TLV is formed of multiple sub-TLVs to inform capabilities such as TE, Graceful Restart, or IS-IS mesh group. TLV 242 was defined originally in RFC 4971 and then updated by RFC 7981 [41], also new sub-TLVs for SR were defined in RFC 8667 to be carried by this TLV [39]. New sub-TLVs to indicate support for SR are detailed in table 4-1.

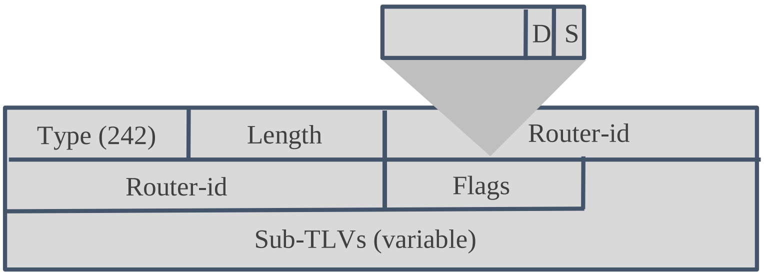


Figure 4-2: IS-IS Router Capability TLV

Figure 4-2 shown Capability TLV structure and fields where *router-id* must be a loopback specifically configured to be the Router-ID of a router. There is an order of preference in case router-Id was not explicitly configured but is a good practice assigned an IP address /32 and Prefix-SID index or label to identify each router inside a network along with all protocols. Flags are considered of flooding scope because if the *S* flag is enabled the TLV will be domain-wide, and if *D* flag is enabled the TLV will be leaked between levels.

SR sub-TLV	Carrier TLV	Description
SR-Capability (2)	Router Capability (242)	Used to indicate SR data-plane capability, label range and start number (SRGB).
SR-Algorithm (19)		Used to indicate algorithm in use for the router to calculate reachability (SPF, SSPF, etc.).
SR Local Block (22)		Used to indicate the range of labels the node has reserved for local SIDs (for a controller).
SRMS Preference (24)		Used to associate a preference with Segment Routing Mapping Server (SRMS) for a source.

Table 4-1: IS-IS Sub-TLVs to advertise router capabilities [21] [33].

Even though there are four sub-TLV that could be carried by Router Capability TLV, only SR-Capability is generally applied to most of the SR implementations. This sub-TLV will be studied in more detail below.

4.1.1.1 IS-IS SR Capability sub-TLV

This sub-TLV is carried by Router Capability TLV and is used to announce to all nodes inside an IS-IS instance that router is Segment Routing data-plane capable. In addition, SR Capability sub-TLV advertises the label range and label start number uses for SRGB [33].

Figure 4-3 shows the SR Capabilities sub-TLV format where flag *I* is used to indicates if a router support MPLS data plane with IPv4 and flag *V* to indicate if a router support MPLS data plane with IPv6. The Range field is the 24-bit length and contains the size of the SRGB label range. Finally, the SID/Label Sub-TLV advertises the label start number inside an SRGB where the 20 rightmost bits correspond to an MPLS label [36].

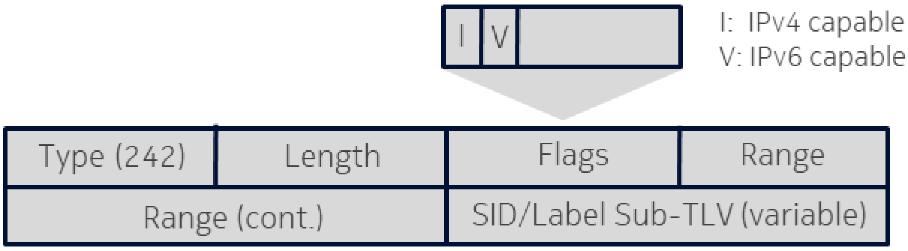


Figure 4-3: IS-IS SR Capability sub-TLV [33]

In figure 4-4 a small network of three routers is set, router *sr21* advertise its SR capabilities to the other routers in the IS-IS Level-1 area. Meanwhile in figure 4-5 a capture from Nokia router *sr01* is shown. Here an output from SR Capability sub-TLV from router *sr21* is highlighted in blue where that indicates *sr21* has support for MPLS IPv4 and IPv6. Also, SRGB size is *10000* labels and the start label is *20000*, so SRGB labels go from *20000* to *29999*. Other details in this output are that router ID of *sr21* is *1.0.0.21*, and flags *D* and *S* are disabled so this TLV must not be advertised between IS-IS levels [36].

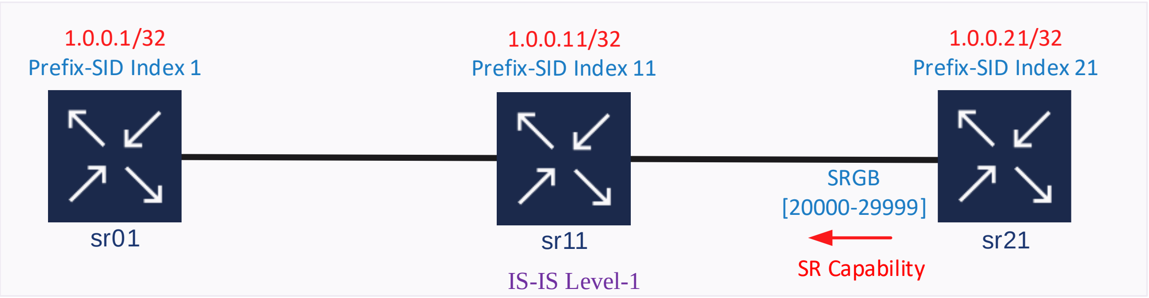
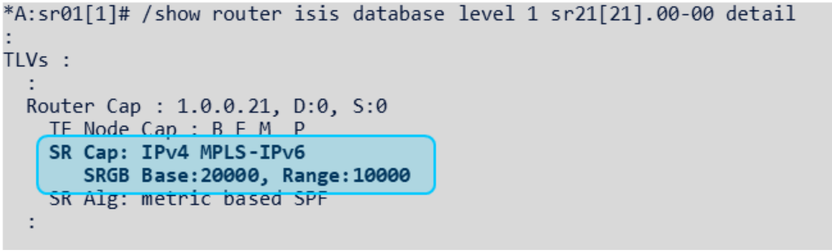


Figure 4-4: IS-IS network for SR Capability advertisement



```
*A:sr01[1]# /show router isis database level 1 sr21[21].00-00 detail
:
TLVs :
  :
  Router Cap : 1.0.0.21, D:0, S:0
    TE Node Cap : B E M  P
    SR Cap: IPv4 MPLS-IPv6
       SRGB Base:20000, Range:10000
    SR Alg: metric based SPF
  :
```

Figure 4-5: IS-IS SR Capability advertisement [21]

4.1.2 IS-IS Prefix-SID Sub-TLV

This sub-TLV type 3 is carried for any of the IP reachability advertisements and is used to advertise a Prefix-SID. Four TLVs could carry Prefix-SID sub-TLV, these are TLV 135 Extended IP Reachability, TLV 235 Multi-Topology IPv4 IP Reachability, TLV 236 IPv6 Reachability, and TLV 237 Multi-Topology IPv6 IP Reachability [33].

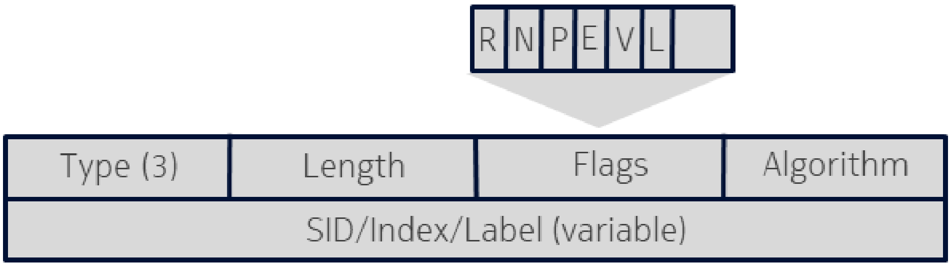


Figure 4-6: IS- IS Prefix-SID sub-TLV format [33]

Figure 4-6 shown the Prefix-SID sub-TLV format, where length field is the size of the SID which could be *5* or *6*. The algorithm field is the code of algorithm used to calculate reachability, it is normally *0* or the traditional SPF based in metric. Flags meaning is explained in table 4-2 and no further details are needed except for flags *V* and *L* because these flags change what is the content of the SID/Index/Label field [21].

When flags *V* and *L* are unset the SID/Index/Label field contains a 32-bit index defining the offset in the SID/Label space advertised by this router, this is the only possible combination in Cisco routers. If flags *V* and *L* are set, the SID/Index/Label field contains a 24-bit label where the 20 rightmost bits are used for encoding the label value. Finally, the SID/Index/Label field could be a variable-length SID as an IPv6 address SID [33] [36].

Flag	**Description**
R	Re-advertise Flag, the prefix has been propagated from other level or protocol redistribution.
N	Node-SID Flag used to identify Prefix-SID attached to Router-ID or loopback prefix (/32 for IPv4 and /128 for IPv6) as a Node-SID. It is enabled by default.
P	no-PHP Flag, if enabled penultimate hop router must not pop top label before forwarding. By default, set in Cisco, unset in Nokia.
E	Explicit-Null Flag, if an enabled upstream neighbor of the Prefix-SID originator replaces the Prefix-SID label with Explicit-Null.
V	Value Flag. If set Prefix-SID carries a value instead of an index. By default, unset.
L	Local Flag. If set Prefix-SID has local significance. By default, unset.

Table 4-2: IS-IS Prefix-SID sub-TLV Flags [33]

In figure 4-7 a small network of three routers is set, router *sr21* advertise its Node-SID to the other routers in the IS-IS Level-1 area. Meanwhile in figure 4-8 a capture from Nokia router *sr01* is shown. Here an output from Prefix-SID sub-TLV from router *sr21* where that indicates *sr21* is using Prefix-SID index *21* that means Prefix-SID label will be *20021* assuming the same SRGB of figure 4-4 is used. In addition, the algorithm used is well-known SPF based on metrics because the *Algo* field is unset. Flags set are *N* and *P*, so prefix *1.0.0.21* is the router ID or loopback of *sr21* because this Prefix-SID is a Node-SID. Also, PHP is off, so the Prefix-SID label will not be popped before forward the packet to router *sr21* [36].

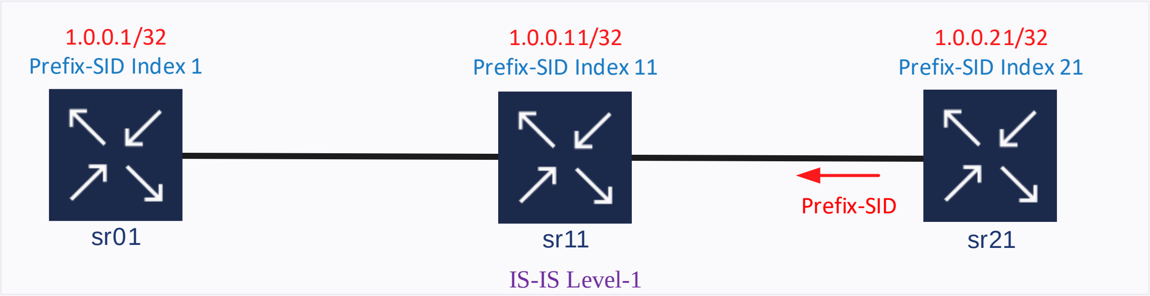
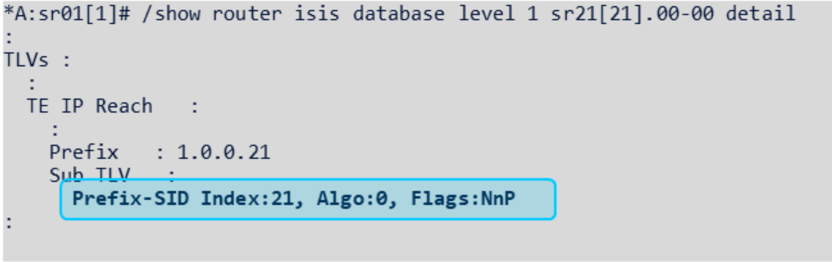


Figure 4-7: IS-IS network for Prefix-SID advertisement



```
*A:sr01[1]# /show router isis database level 1 sr21[21].00-00 detail
:
TLVs :
  :
  TE IP Reach    :
    :
    Prefix    : 1.0.0.21
    Sub TLV     :
     Prefix-SID Index:21, Algo:0, Flags:NnP
:
```

Figure 4-8: IS-IS Prefix-SID advertisement [21]

4.1.3 IS-IS Adjacency-SID Sub-TLV

Each adjacency in a router that is up and running will be assigned an Adj-SIDs (from a different range of SRGB) and advertised to its neighbors. In the same way, IS-IS generates different Adjacency segments for different levels or different versions of IP, so a neighbor adjacency could have one Adj-SID for level-1 and another for level-2. Also, two different Adj-SID for IPv4 and IPv6 [36].

Although adjacencies on a broadcast network as an Ethernet LAN are possible and Adj-SID tie to these adjacencies, modern networks avoid this approach using links point-to-point in all cases. This simplify setup a network because there is no waste of resources in IP address allocation and speed up the adjacency process. Configuring a link as point-to-point even in a broadcast technology as Ethernet means /30 or /31 [42] (for IPv4) subnet might be used, making big savings in IP subnetting. To sum up, Adjacencies SIDs for broadcast networks are defined in the RFC but its inherent complexity and rare use do not justify making a detailed study of this topic.

The Adjacency-SID sub-TLV type 31 is carried for one of the IS neighbor advertisements and is used to advertise an Adj-SID. Four TLVs could carry Adj-SID sub-TLV, these are: TLV 22 Extended IS Reachability, TLV 222 Multi-Topology IS, TLV 23 IS Neighbor Attribute and TLV 223 Multi-Topology IS Neighbor Attribute [33].

Figure 4-9 shown the Adj-SID sub-TLV format, where length field is the size of the SID which could be *5* or *6*. Weight field is used for load-balancing across a number of adjacencies giving a specific weight to the Adj-SID. Flags meaning is explained in table 4-3 and no further details are needed except for flags *V* and *L* because these flags change what is the content of the SID/Index/Label field [21].

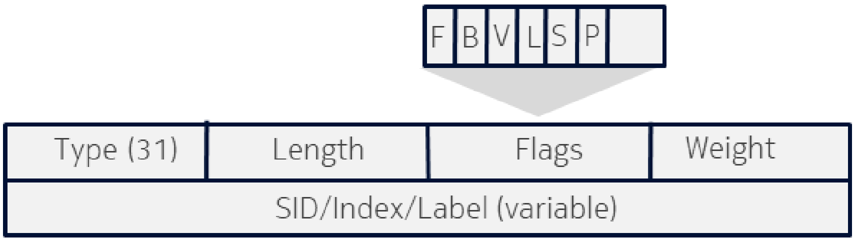


Figure 4-9: IS-IS Adj-SID sub-TLV format [33]

When flags *V* and *L* are unset the SID/Index/Label field contains a 32-bit index defining the offset in the SID space advertised by this router. If flags *V* and *L* are set, the SID/Index/Label field contains a 24-bit label where the 20 rightmost bits are used for encoding the label value, this is the only possible combination in Cisco routers. Finally, the SID/Index/Label field could be a variable-length SID as an IPv6 address SID [33] [36].

Flag	**Description**
F	Address Family Flag, if unset Adj-SID is IPv4. If set Adj-SID is IPv6.
B	Backup Flag, if set Adj-SID is eligible for protection using FRR techniques. It is enabled by default.
V	Value Flag, if set Adj-SID carries a value instead of an index. By default, set.
L	Local Flag, if set Prefix-SID has local significance. By default, set.
S	Set Flag, when set Adj-SID refers to a set of adjacencies.
P	Persistency Flag, Adj-SID value is kept if router reboots.

Table 4-3: IS-IS Adj-SID sub-TLV Flags [33]

In figure 4-10 a small network of three routers is set, router *sr21* advertise its only Adj-SID to the other routers in the IS-IS Level-1 area, that is attached to its *sr11* adjacency with metric *10* and subnet mask */31*. Meanwhile in figure 4-11 a capture from Nokia router *sr01* is shown. Here an output from Adj-SID sub-TLV from router *sr21* where neighbor *sr11* IS-IS ID is exposed (*sr11[11].00*), metric (*10*), IP address local (*2.11.21.1*) and remote (*2.11.21.0*). The weight parameter is set to 0, so Adj-SID is unique and there is no load balancing possible. In addition, the label value is presented (*524283*), not index, this label is outside of SRGB range and that have local significance because flags *V* and *L* are set. Other flags set to indicate that IPv4 address-family is being used (*v4*) and there is a backup route (B) to this Adj-SID [36].

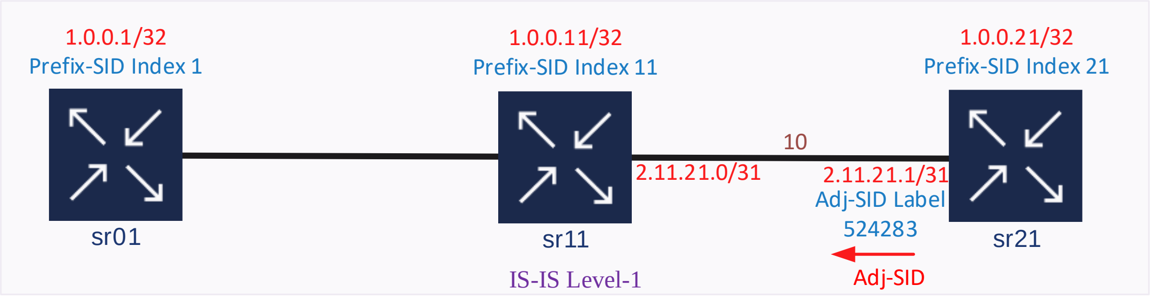
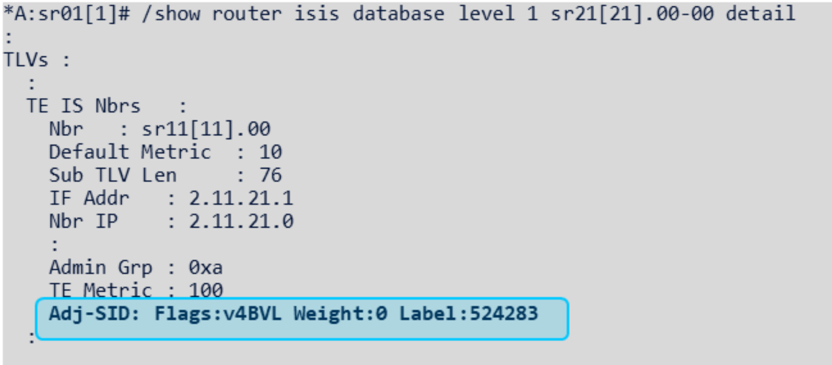


Figure 4-10: IS-IS network for Adj-SID advertisement



```
*A:sr01[1]# /show router isis database level 1 sr21[21].00-00 detail
:
TLVs :
  :
  TE IS Nbrs    :
    Nbr    : sr11[11].00
    Default Metric  : 10
    Sub TLV Len     : 76
    IF Addr    : 2.11.21.1
    Nbr IP     : 2.11.21.0
    :
    Admin Grp : 0xa
    TE Metric  : 100
    Adj-SID:  Flags:v4BVL Weight:0 Label:524283
  :
```

Figure 4-11: IS-IS Adj-SID advertisement [21]

4.2 OSPF Extensions for SR

OSPF has native support for IPv4 only, so its extensions for Segment Routing also have support for this address family and multi-area routing. Similar to IS-IS, in OSPF a Node-SID must be configured, and Adjacencies segments will be automatically generated [36].

4.2.1 OSPF Extended Opaque LSA

Traditional OSPF is based in length fixed LSAs, but this change with the introduction of Opaque LSAs based in Type Length Value (TLV) Triplets which permit attach new attributes to links or prefix like TE. RFC 7684 introduces two Extended Opaque LSA, Extended Prefix Opaque and Extended Link Opaque. In the same way IS-IS TLV, OSPF Opaque LSAs only need to add new sub-TLV for adding new functionalities, so new sub-TLV were aggregated to support SR segments and capabilities [43].

These LSAs are opaque because they could be sent to OSPF routers which not support this kind of LSA, but still will be forwarded until a router could use the information coded in this LSA. These opaque LSAs have the format of TLVs within the fields of the LSA, which is the same format as used by Traffic Engineering extensions to OSPF [33].

4.2.1.1 OSPF Extended Prefix TLV

The OSPF Extended Prefix TLV resides inside the OSPF Extended Prefix LSA and is used for advertising additional attributes associated with the prefix [33]. Figure 4-12 shown how the Extended Prefix TLV is nested inside the Extended Prefix Opaque LSA.

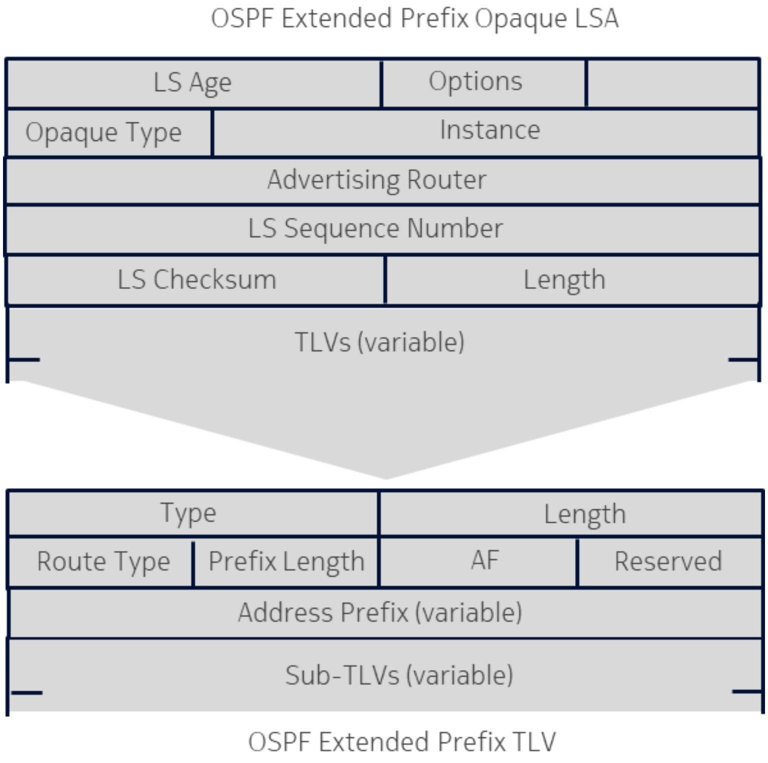


Figure 4-12: Extended Prefix Opaque LSA & Extended Prefix TLV [33]

A brief description of all fields inside an Extended Prefix TLV could be found in table 4-4. In the Sub-TLVs field is going to be the new sub-TLV defined for SR in RFC 8665, this is mainly Prefix-SID sub-TLV.

Field	Description
Route Type	0=unspecified, 1=intra-area, 2=inter-area, 5=external, 7=NSSA external
Prefix Length	Length of the prefix.
AF	0=IPv4 unicast
Address Prefix	Prefix encoded as an even multiple of 32-bit words.
Route Type	0=unspecified, 1=intra-area, 2=inter-area, 5=external, 7=NSSA external

Table 4-4: OSPF Extended Prefix TLV fields [33]

4.2.1.2 OSPF Extended Link TLV

The OSPF Extended link TLV is carried by the OSPF Extended Link LSA and is used to advertise additional attributes associated with the interfaces in an adjacency link [36]. Figure 4-13 shown how the Extended Link TLV is nested inside the Extended Link Opaque LSA.

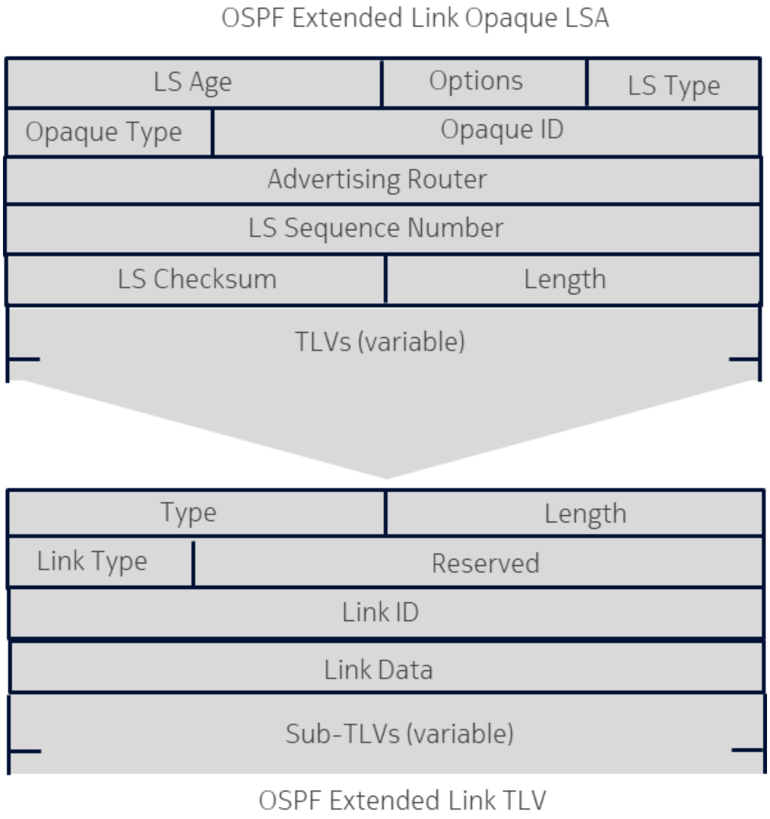


Figure 4-13: Extended Link Opaque LSA & Extended Link TLV

A brief description of all fields inside an Extended Link TLV could be found in table 4-5, here Link ID and Link Data fields are dependable of a Link type value. In the

Sub-TLVs field is going to be the new sub-TLV defined for SR in RFC 8665, this is mainly Adj-SID sub-TLV.

Link Type	Link ID	Link Data	Description
1 (PtoP)	Neighbor Router ID	Interface IP address	PtoP link
		Interface Index (ifIndex)	PtoP link unnumbered
2 (transit network)	DR IP address	Interface IP address	Connection to a transit network
3 (stub network)	Network IP address	Network IP mask	Connection to a transit network
4 (virtual link)	Neighbor Router ID	Interface IP address	Virtual Link

Table 4-5: OSPF Extended Link TLV fields

4.2.2 OSPF Opaque Router Information LSA

This Router Information (RI) LSA was defined first in RFC 4970 and then updated by RFC 7770, which goal was interchange information about optional capabilities between an OSPF router and its neighbors. The Opaque RI LSA could be advertised in a link-scoped (type 9), area-scoped (type 10), or AS-scoped (type 11). [44].

Segment Routing capabilities are distributed using the RI LSA area-scoped, here the SID/Label Range TLV will be nested in the Opaque Router Information to advertise the SRGB. Figure 4-14 shows the SID/Label Range TLV and sub-TLV where Range Size contains the size of the SRGB label range and that is a 24-bit length field. This TLV carried a SID/Label sub-TLV to represent the first SID/Label from the advertised range [33].

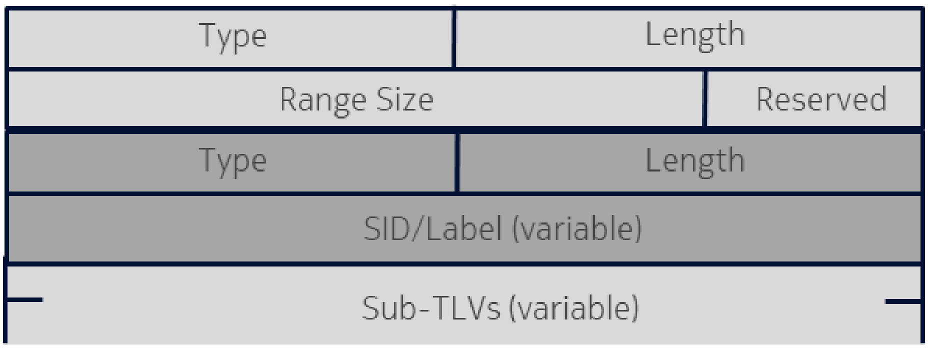


Figure 4-14: OSPF SID/Label Range TLV [36]

In figure 4-15 a small network of three routers is set, router *sr21* advertise its SR capabilities to the other routers in the OSPF area 0. Meanwhile in figure 4-16 a capture from Nokia router *sr01* is shown. Here an output from Opaque Router Information LSA from router *sr21* is highlighted in blue where that indicates *sr21*'s SRGB size is *10000* labels and the start label is *20000*, so SRGB labels go from *20000* to *29999*. Other details in this output are that the router ID of *sr21* is *1.0.0.21*, the scope of the advertisement is Area and the algorithm used is IGP based metric.

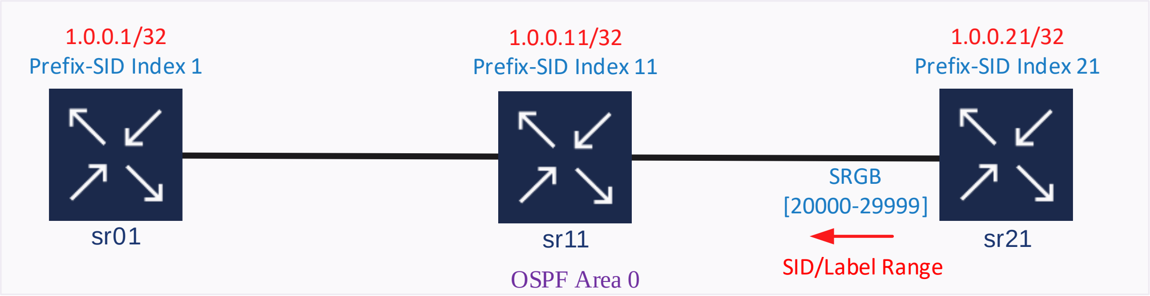
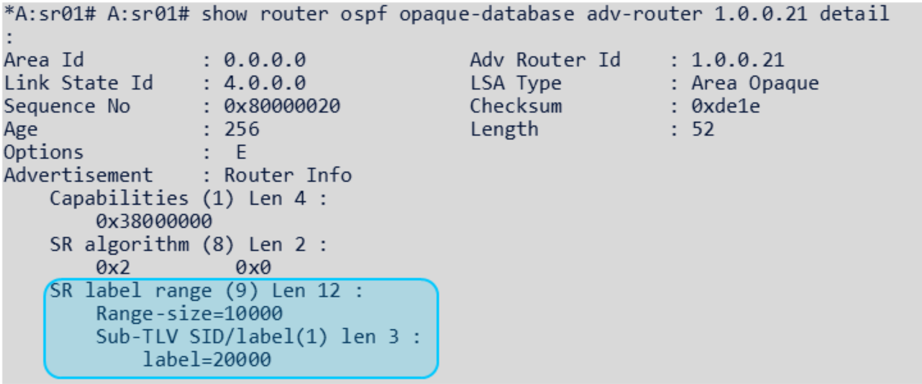


Figure 4-15: OSPF network for SR Capability advertisement



```
*A:sr01# A:sr01# show router ospf opaque-database adv-router 1.0.0.21 detail
:
Area Id             : 0.0.0.0                   Adv Router Id     : 1.0.0.21
Link State Id       : 4.0.0.0                   LSA Type          : Area Opaque
Sequence No         : 0x80000020                Checksum          : 0xde1e
Age                 : 256                       Length            : 52
Options             :  E
Advertisement       : Router Info
    Capabilities (1) Len 4 :
        0x38000000
    SR algorithm (8) Len 2 :
        0x2         0x0
    SR label range (9) Len 12 :
        Range-size=10000
        Sub-TLV SID/label(1) len 3 :
            label=20000
```

Figure 4-16: OSPF SR Capability advertisement

4.2.3 OSPF Prefix-SID Sub-TLV

This sub-TLV, as defined in the RFC 8665, that is carried for the OSPF Extended Prefix TLV and is used for advertising a Prefix-SID. The OSPF Extended Prefix TLV is nested to the Extended Prefix Opaque LSA, and it could have an area or AS scope [37].

Figure 4-17 shows the OSPF Prefix-SID sub-TLV format, where type is always *2* and length field is the size of the SID which could be *7* or *8* bytes depending on *V* flag. This sub-TLV also has the support of Multi-Topology OSPF using the MT-ID field. The algorithm field is the code of algorithm used to calculate reachability, it currently could be only *0* or the traditional SPF based in metric. Flags meaning is explained in table 4-6 and no further details are needed except for flags *V* and *L* because these flags change what is the content of the SID/Index/Label field [21].

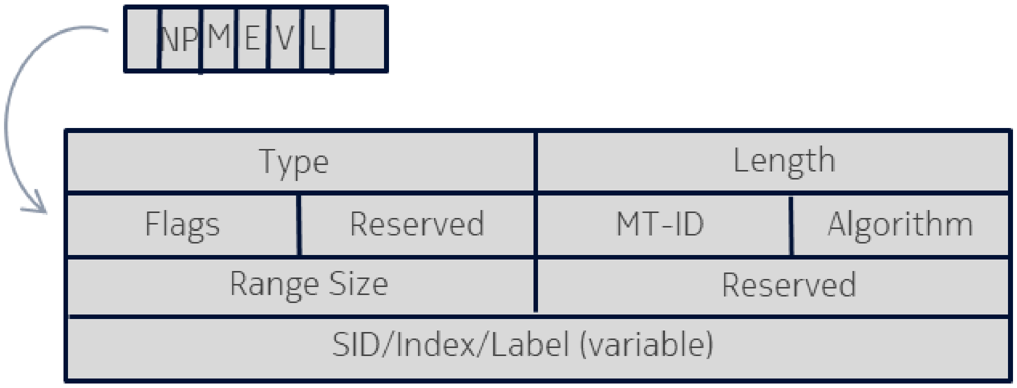


Figure 4-17: OSPF Prefix-SID sub-TLV format [33]

When flags *V* and *L* are unset the SID/Index/Label field contains a 32-bit index defining the offset in the SID/Label space advertised by this router, this is the only possible combination in Cisco routers and default in Nokia routers. If flags *V* and *L* are set, the SID/Index/Label field contains a 24-bit label where the 20 rightmost bits are used for encoding the label value. These two are the only possible combination for *V* and *L* flags, any other combination will be ignored [33] [37].

Flag	**Description**
NP	no-PHP Flag, if enabled penultimate hop router must not pop top label before forwarding. By default, set in Cisco, unset in Nokia.
M	Mapping Server Flag. If set, the SID is advertised from the Segment Routing Mapping Server.
E	Explicit-Null Flag, if an enabled upstream neighbor of the Prefix-SID originator replaces the Prefix-SID label with Explicit-Null.
V	Value Flag. If set Prefix-SID carries a value instead of an index. By default, unset.
L	Local Flag. If set Prefix-SID has local significance. By default, unset.

Table 4-6: OSPF Prefix-SID sub-TLV Flags [33]

In figure 4-18 a small network of three routers is set, router *sr21* advertise its Node-SID to the other routers in the OSPF area 0. Meanwhile in figure 4-19 a capture from Nokia router *sr01* is shown. Here an output from Prefix-SID sub-TLV from router *sr21* where that indicates *sr21* is using Prefix-SID index *21*, which means Prefix-SID label will be *20021* because the same SRGB of figure 4-16 is used. In addition, the algorithm used is well-known SPF based on metrics because the *Algorithm* field is unset. Only flag set is *NP*, so PHP is off and the Prefix-SID label will not be popped before forward the packet to router *sr21*. Also, prefix *1.0.0.21* is the router ID or loopback of *sr21* because this Prefix-SID is a Node-SID [36].

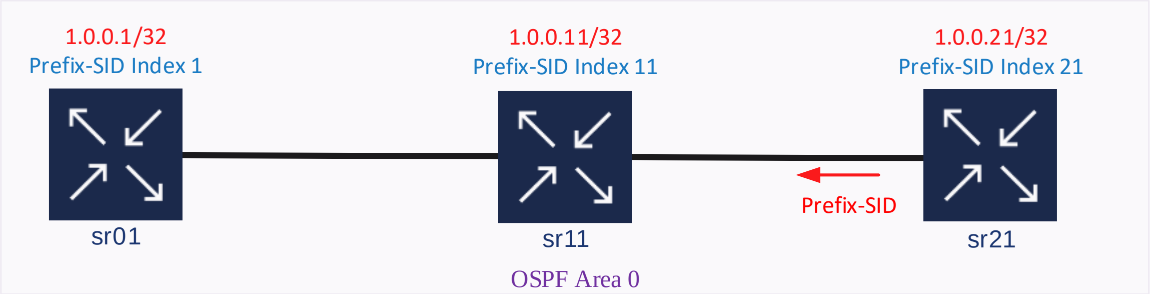
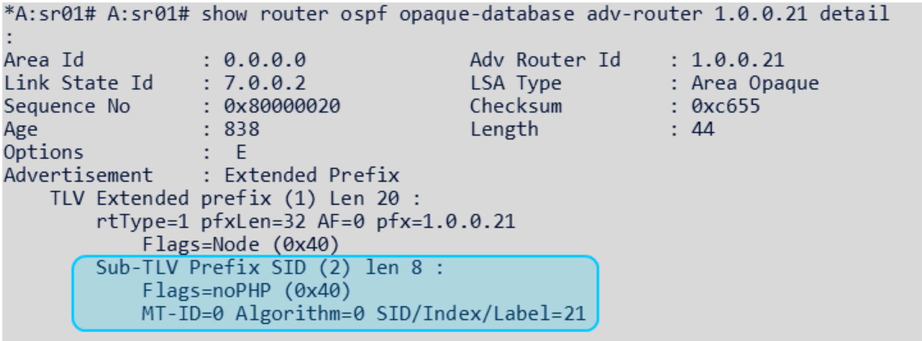


Figure 4-18: OSPF network for Prefix-SID advertisement



```
*A:sr01# A:sr01# show router ospf opaque-database adv-router 1.0.0.21 detail
:
Area Id          : 0.0.0.0                Adv Router Id    : 1.0.0.21
Link State Id    : 7.0.0.2                LSA Type         : Area Opaque
Sequence No      : 0x80000020             Checksum         : 0xc655
Age              : 838                    Length           : 44
Options          :  E
Advertisement    : Extended Prefix
    TLV Extended prefix (1) Len 20 :
        rtType=1 pfxLen=32 AF=0 pfx=1.0.0.21
            Flags=Node (0x40)
        Sub-TLV Prefix SID (2) len 8 :
            Flags=noPHP (0x40)
            MT-ID=0 Algorithm=0 SID/Index/Label=21
```

Figure 4-19: OSPF Prefix-SID advertisement

4.2.4 OSPF Adjacency-SID Sub-TLV

In the same way as IS-IS, in OSPF each adjacency in a router that is up and running will be assigned an Adj-SIDs with local significance and from a different range of SRGB, then these segments will be advertised to its neighbors. If any backup path is available, the Adjacency-SID will be protected, and this will be advertised as type *1* [36].

In OSPF Adjacencies SIDs or LAN Adjacency-SID for broadcast networks are defined in the RFC, but modern networks avoid this approach using links point-to-point in all cases even in a broadcast technology like Ethernet. That means a subnet /30 or /31 for IPv4 might be used, making big savings in IP subnetting.

The Adjacency-SID sub-TLV is an optional sub-TLV of the Extended Link TLV and may appear multiples times in this TLV. The Extended Link TLV is nested in an Extended Link Opaque LSA with area-scope to avoid that will be advertised outside of base OSPF and is used for advertising additional attributes associates with the link. See point 4.2.1.2 for further details [36].

Figure 4-20 shown the Adj-SID sub-TLV format, where length field is the size of the SID which could be *7* or *8* depending on the *V* flag. Weight field is used for load-balancing across several adjacencies giving a specific weight to the Adj-SID. This sub-TLV also has the support of Multi-Topology OSPF using the MT-ID field. Multiple Adj-SIDs can be allocated to a single adjacency, or the same Adj-SID can be allocated to multiple adjacencies. Flags meaning is explained in table 4-7 and no further details are needed except for flags *V* and *L* because these flags change what is the content of SID/Index/label and length fields [37].

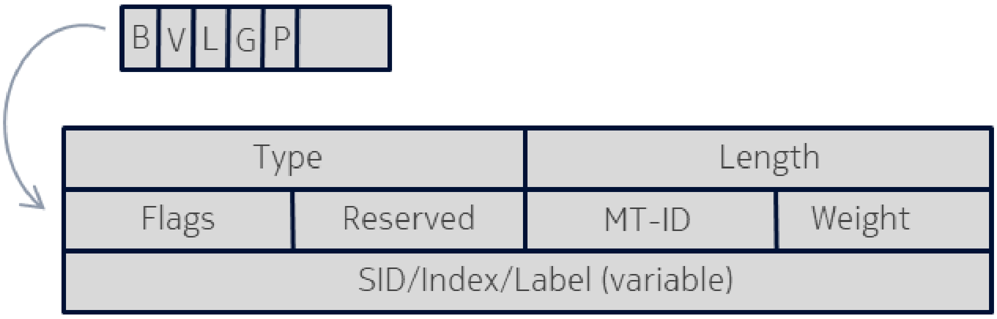


Figure 4-20: OSPF Adj-SID sub-TLV format [33]

When flags *V* and *L* are unset the SID/Index/Label field contains a 32-bit index defining the offset in the SID space advertised by this router. If flags *V* and *L* are set, the SID/Index/Label field contains a 24-bit label where the 20 rightmost bits are used for encoding the label value, this is the only possible combination in Cisco routers, and by default combination in Nokia routers [33] [36].

Flag	**Description**
B	Backup Flag, if set Adj-SID is eligible for protection using FRR techniques. It is enabled by default.
V	Value Flag, if set Adj-SID carries a value instead of an index. By default, set.
L	Local Flag, if set Prefix-SID has local significance. By default, set.
G	Group Flag, when set Adj-SID refers to a group of adjacencies.
P	Persistency Flag, Adj-SID value is kept if router reboots.

Table 4-7: OSPF Adj-SID sub-TLV Flags [33]

In figure 4-21 a small network of three routers is set, router *sr21* advertise its only Adj-SID to the other routers in the OSPF area 0, that is attached to its *sr11* adjacency with metric *10* and subnet mask */31*. Meanwhile in figure 4-22 a capture from Nokia router *sr01* is shown. Here an output from OSPF Adj-SID sub-TLV from router *sr21* where neighbor *sr11* router ID is exposed (*1.0.0.11*) and IP address local (*2.11.21.1*). MT-ID and Weight parameter are set to 0, so there are not Multi-topology and Adj-SID is unique therefore is no load balancing possible. In addition, the label value is presented (*524287*), not index, this label is outside of SRGB range and that have local significance because flags *V* and *L* are set (*Flags=Value Local*). Link-type used is a point to point (*P2P*) and Adj-SID sub-TLV is *7* bytes length [36].

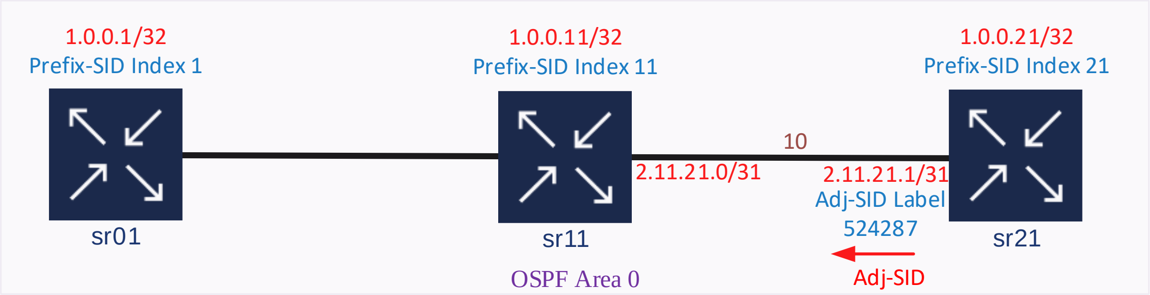
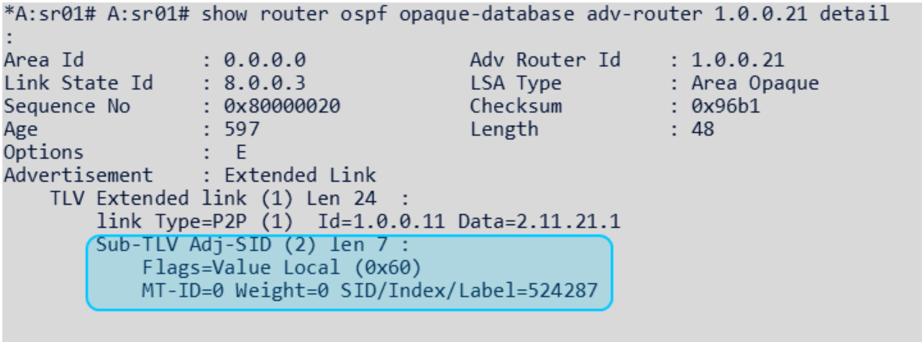


Figure 4-21: OSPF network for Adj-SID advertisement



```
*A:sr01# A:sr01# show router ospf opaque-database adv-router 1.0.0.21 detail
:
Area Id          : 0.0.0.0              Adv Router Id    : 1.0.0.21
Link State Id    : 8.0.0.3              LSA Type         : Area Opaque
Sequence No      : 0x80000020           Checksum         : 0x96b1
Age              : 597                  Length           : 48
Options          :  E
Advertisement    : Extended Link
    TLV Extended link (1) Len 24  :
        link Type=P2P (1)  Id=1.0.0.11 Data=2.11.21.1
        Sub-TLV Adj-SID (2) len 7 :
            Flags=Value Local (0x60)
            MT-ID=0 Weight=0 SID/Index/Label=524287
```

Figure 4-22: OSPF Adj-SID advertisement

5 SR BGP Control Plane

At the start, in RFC 4271 BGP was defined for interchange Internet IPv4 prefixes only between different Service Provider networks [45], later MP-BGP appeared using the Capabilities parameter in the OPEN message to add new address-families to support different protocols like IPv6, L2VPN, MVPN, VPN-IPv4, VPN-IPv6, etc. This Capabilities parameter allows routers which are establishing a session using a given capability advertised by both BGP speakers in the OPEN exchange process [46].

Address Family Identifier (AFI) and Subsequent Address Family Identifier (SAFI) parameters are used to identifying a network layer protocol and associate that with Next-Hop information and Network Layer Reachability Information (NLRI) field. That means what address families are supported in a given BGP session. AFI and SAFI values are assigned by IANA, for instance, a VPN-IPv4 prefix is represented as *AFI 1* (*IPv4*) and *SAFI 128* (*MPLS-Labeled VPN address*) [47]. Table 5-1 shows the most used values of AFI and SAFI for different address families.

AFI	**Description**	**SAFI**	**Description**
1	IPv4	1	Unicast
2	IPv6	2	Multicast
		4	MPLS Label (BGP-LU)
		5	MCAST-VPN
		65	Virtual Private LAN Service (VPLS)
		70	BGP EVPNs
		71	BGP-LS
		73	SR-TE Policy
		74	SD-WAN Capabilities
		128	MPLS-labeled VPN address
		132	Route Target constrains

Table 5-1: AFI and SAFI values for MP-BGP [48] [49].

Since MP-BGP is widely deployed around the world, most of the service provider networks only need to add the respective address-family to BGP sessions to have support for a new network layer protocol. In some cases, a router's software upgrade is required but still, a new address family could be added without a hardware upgrade.

The final goal of BGP SR is direct traffic towards a destination creating a BGP SR-TE LSP using an MPLS data plane encapsulation.

5.1 BGP-LU

As it was briefly mentioned in point 1.2.3, BGP Label Unicast (BGP-LU) was defined in RFC 3107 and is the main component in a seamless MPLS architecture. With BGP-LU a label is mapped to a specific prefix (loopback or router ID), and then distributed both in the same update message. In this address-family *AFI 1* (*IPv4*) and *SAFI 4* (*MPLS Label*) is used to indicate the NRLI contains a label, so this NRLI is encoded as a triplet of the form <*length, label, prefix*> [15] [16].

BGP-LU permits deploy services end-to-end between two or more Autonomous Systems (AS) that is why that is used to implement Inter-AS model C to extend VPN L3 services from one AS to another. This Inter-AS model uses eBGP Label Unicast sessions and sessions inside a Seamless MPLS network to use iBGP-LU because in most cases the same AS is used in specific network sections. Architecture and uses cases for Seamless MPLS or Unified MPLS (Cisco terminology) are described in *draft-ietf-mpls-seamless-mpls-07* [50].

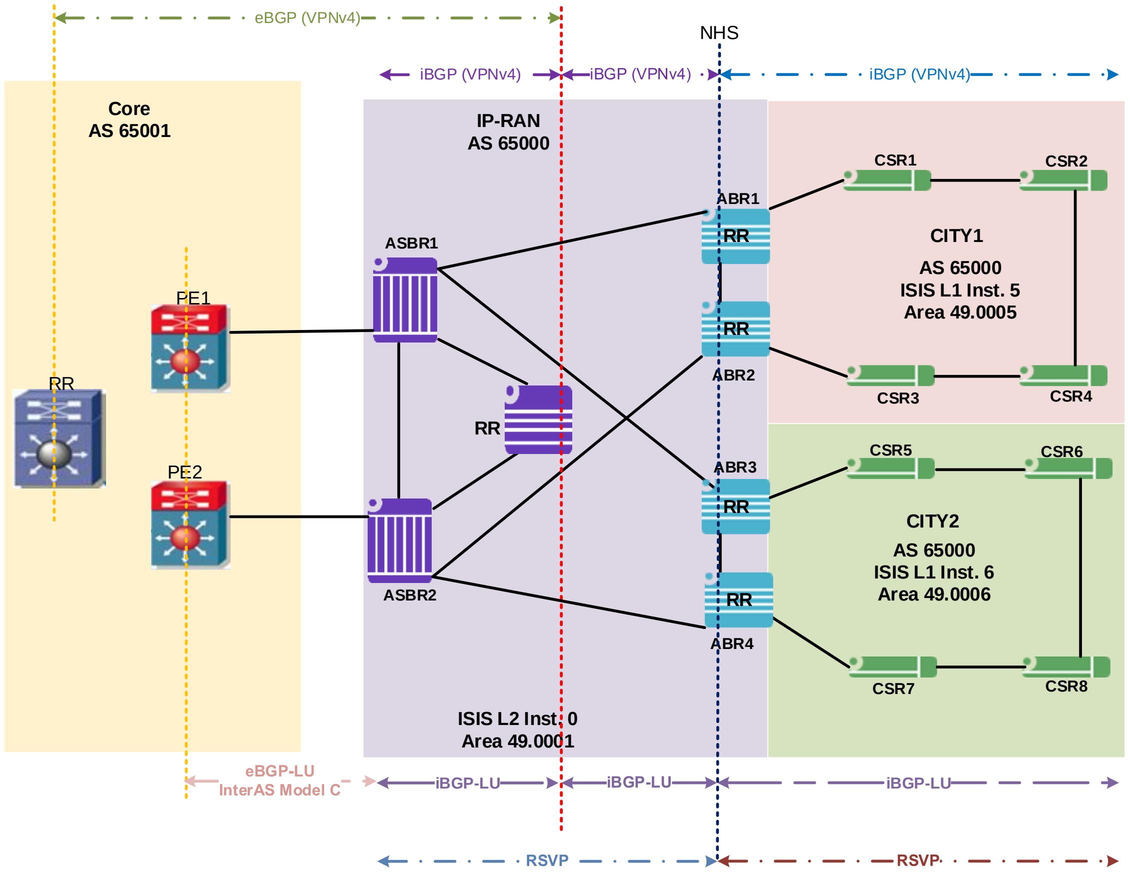


Figure 5-1: Inter-AS and Seamless MPLS network

An example of both Inter-AS model C and Seamless MPLS is displayed in figure 5-1, here a Mobile Backhaul (MBH) or IP-RAN network is using a Seamless MPLS topology to interconnect different portions of this network, and also the MBH uses an Inter-AS model C interconnection to the Core to extend a VPN L2 or L3 from the IPRAN to the Core and vice versa. Both techniques used BGP 3107 or BGP Label Unicast to provide end-to-end reachability.

In the Seamless MPLS part, the same AS (*65000)* is used across the different sections of the IPRAN network, but diverse IS-IS instances are used to prevent the routing information from one section reach other sections. Instance *0* is used by ABRs, ASBRs and main RRs, instance *5* is used for *City 1* region with CSRs *1* to 4 and ABRs *1* and *2* inside. Instance 6 is used for *City 2* region with CSRs *5* to *8* and ABRs *3* and *4* inside. This topology improves scalability and fault tolerance because thousands of routers could be added in different instances and also fail like a routing loop in a section won't affect the others. In this architecture, two aggregation routers become the ABRs which function is become RRs regionals and translate a route from one instance to another. These routers activate the *next-hop-self* (NHS) function to redistribute routes towards both instances because LDP or RSVP-TE tunnels will work only inside an instance. The way a CSR reaches an ASBR is by using a BGP-LU tunnel using labels distributed by the RRs. Any service L2 or L3 could be created using these labels to build service tunnels end-to-end because each node carries all loopbacks of all PE routers inside the network. In an MBH network, not only the CSRs are PEs because Base Station units are connected to the geographically closer router, and sometimes this is an ABR or even an ASBR. Inside a network region traffic is protected by RSVP-TE FRR or LFA techniques, but to protect an end-to-end communication BGP Prefix Independent Convergence (BGP PIC) better known as BGP FRR is used to. BGP FRR brings together indirection techniques in the forwarding plane and pre-computation of BGP backup paths in the control plane to support the fast reroute of BGP traffic around failed next-hops [51].

In the Inter-AS model C section of figure 5-1, the ASBR routers in the IPRAN have a direct connection to PE routers in the Core. This connection is used to establish an eBGP session with IPv4 label unicast enabled between IPRAN's ASBRs with AS *65000* and Core's PEs with AS *65001*. All router's loopbacks from each AS are interchange with a label associated to each one using the eBGP session, then RRs in both AS establish an eBGP multihop session with VPN-IPv4 address-family in order to interchange VPN L3 prefixes between both AS. Those prefixes permit extend a VPN L3 (also known as VRF in Cisco and VPRN in Nokia) from one AS to another in an automatic way because this VRF is configured in each PE to reach any point inside the VPRN. No configuration has to be made in intermediate nodes as ASBRs or ABRs because each PE will send its VPN-IPv4 prefixes to the regional RRs. Pseudowire and VPLS services also can be configured through both AS using existent BGP-LU tunnels for the end-to-end communication and LDP or RSVP-TE inside each AS. Again, as in Seamless MPLS each AS could have different IGP protocol and label advertisement protocol, so in the example AS *65000* use ISIS as IGP protocol and AS *65001* uses OSPF. Also, the Core section uses LDP to distribute labels and the IPRAN network uses RSVP-TE. The same traffic protection mechanism as in Seamless MPLS applies to Inter-AS model C because inside each AS a local method could be used and for the end-to-end protection BGP PIC must be used.

Inter-AS model C is a good option to consolidates two different AS from the same company as in a merge or acquisition because integration is very straightforward with few changes in the installed base. There are no changes in ASs, IGP protocols or MPLS protocols, and in most of the cases, only the BGP-LU unicast has to be added to each session from a PE to the RRs in order to propagate its loopback associate with a label and receive all loopbacks from other routers in both ASs. Once configuration in ASBRs and RRs is done, any kind of service could be created or extended from an AS to another only configuring services in the respective PEs.

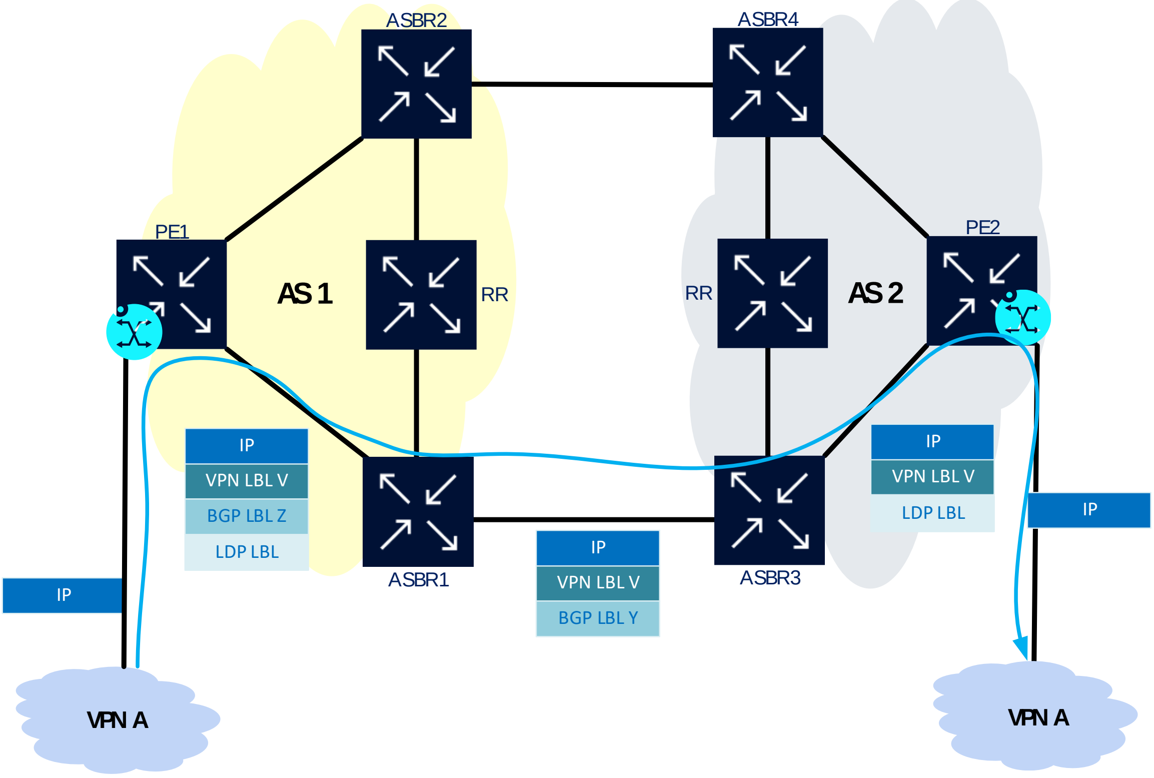


Figure 5-2: Inter-AS Model C Data Plane

In figure 5-2, Inter-AS model C is used to interconnect AS *1* and AS *2*, assuming all loopbacks from each AS have already been interchanged and eBGP session between RRs is up and running. LDP is the label advertisement protocol in both ASs. A VPN L3 is configured in PE1 and PE2 to use BGP-LU tunnels in order to establish an end-to-end transport tunnel. In this example when an IP packet arrives at PE1 in the interface where VPN A is configured, this packet will be encapsulated with three labels at the outbound interface. The first label is the service label *V*, then BGP-LU label *Z* and finally the LDP label or transport label to reach ASBR1. At this point, ASBR1 pops the LDP label and swaps the BGP label to *Y* to send the packet to ASBR3. Here ASBR2 pops BGP label and pushes an LDP transport label to the packet, then that packet is forwarded to PE2. In PE2, the transport label is popped, VPN label *V* is used to get the correct FIB and then is popped. Finally, a pure IP packet is delivery to the destination node.

5.2 BGP Prefix-SID

Prefix-SID is a new BGP path attribute that associates a SID with a prefix IPv4 or IPv6 advertised by BGP. As in SR IGP, Prefix Segments are global unique segment indexes used to identify specific Prefix-SID from the SRGB range. Also, when a Prefix-SID tied to a prefix installed in a packet header arrives at an SR capable router that forwards this packet on the best BGP path towards the prefix originator, using all available multipath computed by BGP. The Prefix-SID is useful to extend segment routing across admin boundaries that do not have a common IGP and also in Data Centres where an IGP protocol is not used. [52] [53].

To distribute Prefix-SID to other SR capable routers BGP-LU is used, each Prefix-SID is attached to a BGP-LU prefix advertisement. This Prefix-SID is originated by the node where the prefix resides or where that was redistributed from a different protocol like IGP. As already mentioned a Prefix-SID is propagated as a path attribute in the BGP update message.

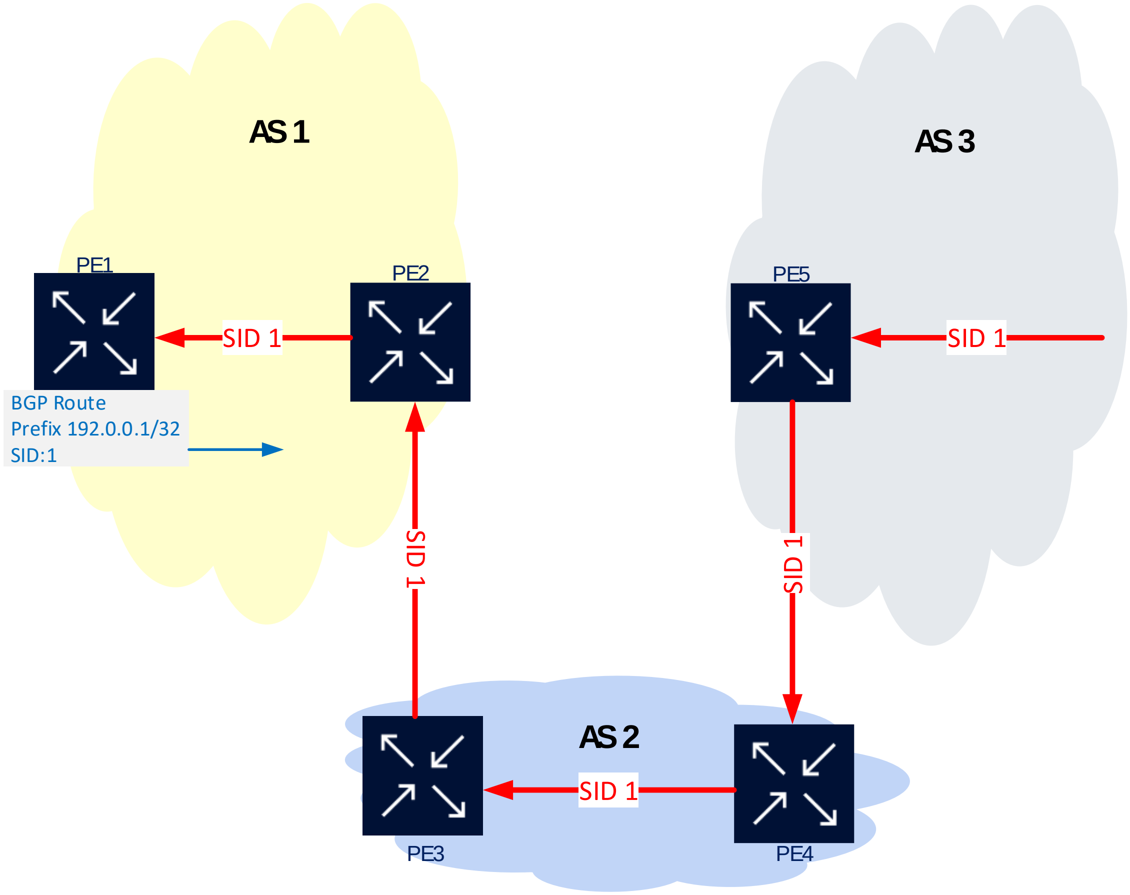


Figure 5-3: BGP Prefix-SID through ASs

In order to make a BGP Prefix-SID global across ASes interconnected with each other (as in figure 5-3), all these ASes need to be part of the same SR domain. Each BGP speaker needs to be configured with a Segment Routing Global Block (SRGB) label range. As was already mentioned is a best practice using the same SRGB across all the nodes within the SR domain, but the SRGB of a node has local meaning and could be different on different routers [54].

In figure 5-3 a prefix 192.0.0.1/32 is originated by PE1 and distribute to SR routers in AS *1*, AS *2* and AS *3* using BGP-LU. This prefix is associated with Prefix-SID *1* and advertise together with the prefix. Then when a packet arrives at PE5 in AS *3* with SID *1*, that router is going to forward the packet towards router PE4 in AS *2*. Inside this AS, PE4 will send the packet to the router closest to AS *1* which is PE3. Then router PE3 forward this packet to AS *1* using router PE2. Finally, router PE2 is going to send the packet to the router who generates the Prefix-SID advertisement, which is PE1.

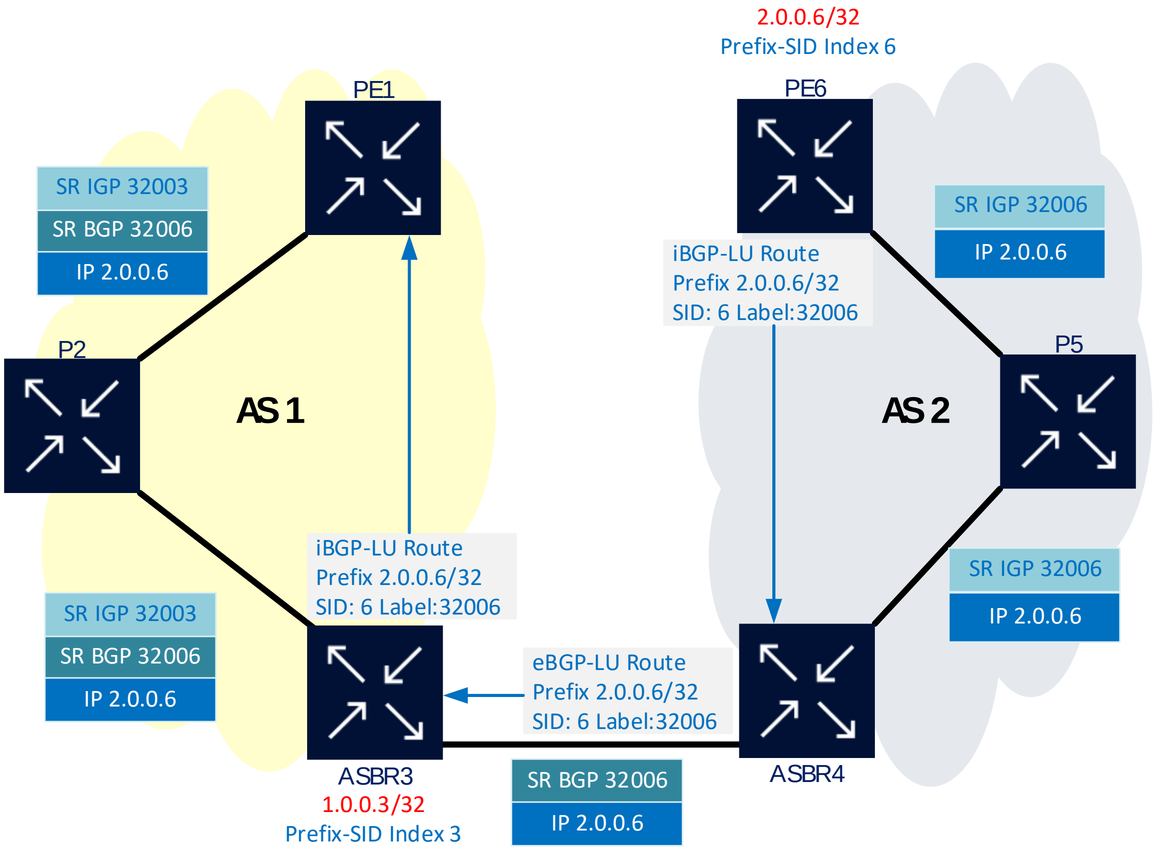


Figure 5-4: Inter-AS with BGP SR enabled

In figure 5-4, there is an Inter-AS interconnection between AS *1* and AS *2* through an eBGP-LU session between ASBR3 and ASBR4, all routers have IGP and BGP SR enabled with [*32000-39999*] as SRGB. As SR is enabled BGP-LU will use labels from the SRGB range using the label-index information propagated in the update message. The IGP SR in both ASs will also provide the same labels no matter what protocol is that, and there is no LDP or RSVP-TE enable in the network. Inside each AS iBGP-LU sessions are established in a full-mesh way in order to propagate loopback prefixes tied to Prefix-ID, but in this example only relevant sessions are shown. Router PE6 advertises its loopback prefix *2.0.0.6/32* with label-index *6* to ASBR4, this router learns that prefix and using SID index assign label *32006* as Prefix-SID for PE6. Then ASBR4 will follow a similar process to distribute this prefix and Prefix-ID using the eBGP-LU session to ASBR3. Finally, router ASBR3 send prefix and Prefix-SID information from PE6 to router PE1.

Also, in figure 5-4, when an IP packet arrives at PE1 with destination address *2.0.0.6*, this packet will be encapsulated with two labels, *32006* for SR BGP and *32003* for SR IGP. This IGP label corresponds to ASBR3 because this is the next hop for SID *32006*. In ASBR3 the SR IGP label will be removed from the packet and forwarded to ASBR4, where BGP Prefix-SID label will be popped and IGP Prefix-SID *32006* is imposed and forward to router P5. In that router, a label swap with the same value will be performed and then the packet is forwarded to its final destination router PE6.

5.2.1 BGP Prefix-SID Advertisement

The mechanism to advertise BGP Prefix-SID is described in RFC 8669 where the BGP prefix attribute is defined as an optional transitive BGP path attribute type 40. That means if a router is not SR capable still this attribute will be passed to other nodes. The Prefix-SID attribute could be attached to prefixes advertises using BGP Label Unicast (RFC 3107 [16]) when MPLS data plane is used in SR, and also BGP unlabelled IPv6 unicast (RFC 4760 [46]) could be used when IPv6 is used as data plane (SRv6) [54].

The BGP Prefix-SID attribute uses two TLVs for the MPLS data plane to advertise a prefix segment. These are Label Index TLV and Originator SRGB TLV [54].

5.2.1.1 Label Index TLV

This TLV must be present in the BGP Prefix-SID attribute attached to an IPv4 or IPv6 Labelled Unicast prefix. If Label Index TLV is present in a different kind of prefix it must be ignored [54].

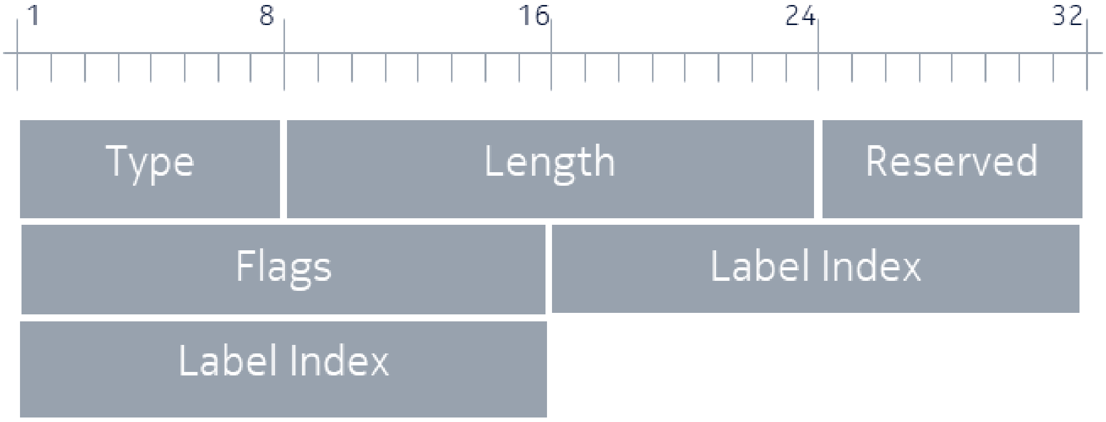


Figure 5-5: BGP - Label-Index TLV format

Figure 5-5 shown the Label Index TLV format, where Type is equal to *1*, Length is always *7* bytes and no Flags have been defined in the RFC. The index value pointing to a label in the SRGB space is coded in the 32-bit field called Label Index [54].

5.2.1.2 Originator SRGB TLV

This TLV is optional to the BGP Prefix-SID attribute attached to an IPv4 or IPv6 Labelled Unicast prefix. The Originator SRGB TLV carried the SRGB used by the router which originates the prefix and this value must not be changed along with the advertisement to different routers when different SRGB ranges are used. If Originator SRGB TLV is present in a different kind of prefix it must be ignored [54] [53].

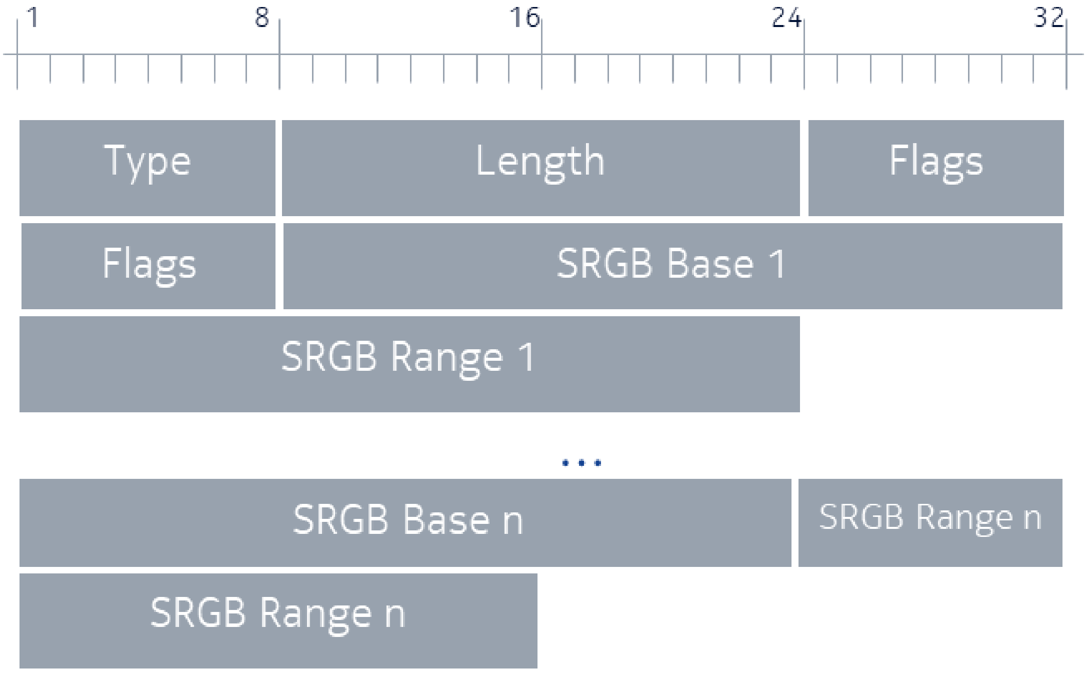


Figure 5-6: BGP – Originator SRGB TLV

Figure 5-6 shows the Label Index TLV format, where Type is equal to *3*, Length is variable depending on how many SRGB ranges are defined in the node, for instance when only one SRGB is used this value is *8* bytes. As in Label-Index no Flags have been defined in the RFC. The first label in the SRGB space is coded in the 3 bytes field called SRGB Base, and the other 3 bytes field called SRGB Range carried the number of labels in the range. For instance, if a router's SRGB range is [*32000-39999*], SRGB Base will be *32000* and SRGB Range will be *10000*. The SRGB field could be repeated several times to indicates there is more than one continuous SRGB range, which is possible but not recommended [54].

The operation of BGP-LU is the same with Prefix-SID, in this case an originator node allocates its local label (from the SRGB) associated with its loopback prefix and included the Prefix-SID index inside the Label Index TLV, in the advertisement packet to its peers. Each receiving node built its MPLS FIB with all Prefix-SID indexes received from its neighbours mapping them to its local labels inside the SRGB. Then, this local label allocated is sent along with the BGP Prefix-SID attribute to its peers in the prefix update [53].

5.2.2 SR BGP and non-SR BGP Interoperation

As long as the BGP Prefix-SID attribute is optional transitive, all BGP-LU non-SR capable routers are going to ignore this attribute in an update packet. That means the SID label carried in the Label Index TLV, is not going to be installed in the MPLS FIB associated with a prefix by a non-SR router. This router instead will allocate a random label from the dynamic label range and forward that to its peers. If one of these peers is SR capable that will receive the original Label Index TLV because the BGP Prefix-SID attribute is forwarded unchanged, and this will use the SID index to allocate the local label [53].

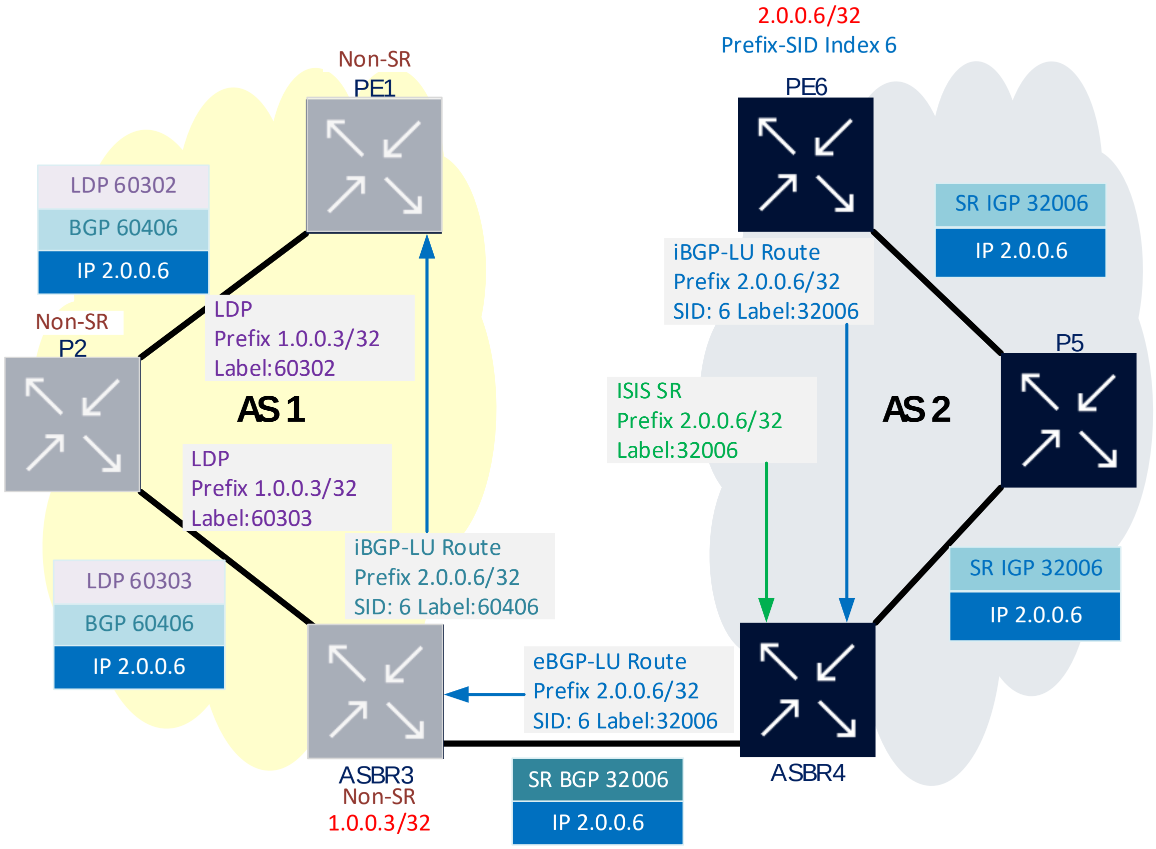


Figure 5-7: Inter-AS between SR BGP and non-SR BGP routers

Figure 5-7 shown an Inter-AS connection between AS *1* and AS *2* through an eBGP-LU session between ASBR3 and ASBR4, all routers inside AS *2* have IGP and BGP SR enabled with [*32000-39999*] as SRGB. All routers inside AS *1* are non-SR capable and have LDP enable them to distribute labels inside the AS. As SR is enabled in AS *2* BGP-LU will use labels from the SRGB range using the label-index information propagated in the update message inside the Label Index TLV. Also, the IGP SR will provide the same labels no matter what protocol is that, and there is no LDP or RSVP-TE enable in the AS *2*. On the other hand, iBGP-LU inside AS *1* will allocate labels from the dynamic label range and LDP also uses this range to assign labels in each node. Inside each AS iBGP-LU sessions are established in a full-mesh to propagate loopback prefixes tied to a label, in the AS *2* this label will get from the Prefix-ID, but in this example, only relevant sessions are shown. Router PE6 advertises its loopback prefix *2.0.0.6/32* with label-index *6* to ASBR4, this router learns that prefix and using SID index assign label *32006* as Prefix-SID for PE6. Then ASBR4 will follow a similar process to distribute this prefix and Prefix-ID using the eBGP-LU session to ASBR3. When router ASBR3 receives the update packet with the Label Index TLV inside, that is not to be understood by this router and label *60406* from the dynamic range is allocated and the next hop is setting to ASBR3. Inside AS *1*, LDP allocates label *60303* attached to prefix *1.0.0.3/32* which is the loopback of ASBR3 and the label *60302* from P2 for the same prefix.

Also, in figure 5-7, when an IP packet arrives at PE1 with destination address *2.0.0.6*, this packet will be encapsulated with two labels, *60406* for BGP-LU and *60302* for LDP. This LDP label corresponds to ASBR3 because this is the next hop for label *60406,* then in P2 the LDP label will be swapped for label *60303* leaving the BGP label untouched. In ASBR3 the LDP label will be removed from the packet, and the BGP label will be swapped to *32006* and forwarded to ASBR4, where BGP Prefix-SID label will be popped and IGP Prefix-SID *32006* is imposed and forward to router P5. In that router, a label swap with the same value will be performed and then the packet is forwarded to its final destination router PE6.

5.3 SR BGP in Data Centre

First Data Centres (DC) implementations used the classical three-layer architecture composed by Access, Aggregation, and Core, this topology is used for most of the telecommunication networks until now because most of the traffic needs to get the Core section. For instance, in an MBH network most of the traffic going from Access nodes to Controllers in the Core or the Internet beyond the Core. Here links of low capacity between Access and Aggregation are acceptable, and high capacity links are reserved to Aggregation-Core or Core-Core connections where traffic from the Access section is consolidated. First DC networks also have *north-south* traffic pattern behaviour because almost all the traffic was client-server, so traffic from the client came from different access networks or the Internet have to pass by the Core network to reach the DC network where Servers were installed [53].

Nowadays most of the traffic inside a Data Centre is between servers, also known as *east-west* traffic patterns, so the three-layer model is not suitable anymore for a modern DC network. Moreover, most large cloud DC networks are built with a *spine/leaf* architecture which deals better with server-to-server traffic than the classical three-layer approach. The *spine/leaf* networks sometimes also referred to as Clos networks and Fat

Tree networks because links become bigger closer to the root, for instance, 1G links towards the servers, 10G links towards tier 1, 40G links towards tier 2, etc. Also, a very distinctive characteristic of these networks is that every Leaf switch is connected to every Spine switch, as is shown in figure 5-8. That means any communication between two servers in different Leaf switches always must going across the Spine switches [55].

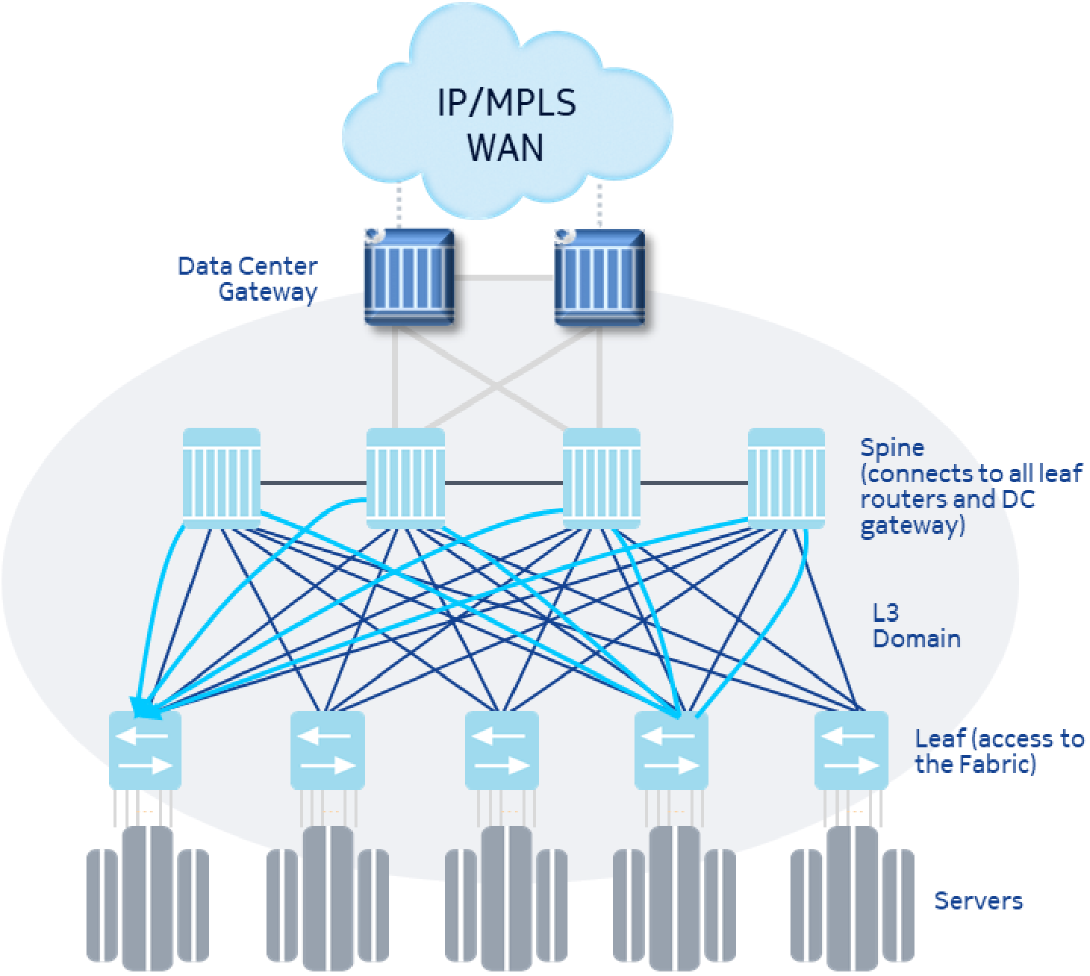

IP/MPLS WAN
Data Center Gateway
Spine (connects to all leaf routers and DC gateway)
L3 Domain
Leaf (access to the Fabric)
Servers

Figure 5-8: Spine and Leaf network [55]

Tiers in a Clos are increasing from Leaf to Spine where the higher Tier is always the Spine layer, and the lower tiers could be several layers of Leaf layers. As the Leaf switches always are connected to all the Spine switches, there is not aggregation layer because the bandwidth is the same between each layer. The Leaf switches are usually installed in the same rack of a group of servers, so they also known as Top-Of-Rack (ToR) switches and might be installed in pairs to give redundant connection to the servers. When over the ToR layer another Leaf layer is added usually this last one becomes the Leaf layer, so the network will be formed for three tiers, Tier 1 or ToR, Tier 2 or Leaf and Tier 3 or Spine; as is shown in figure 5-9 [53].

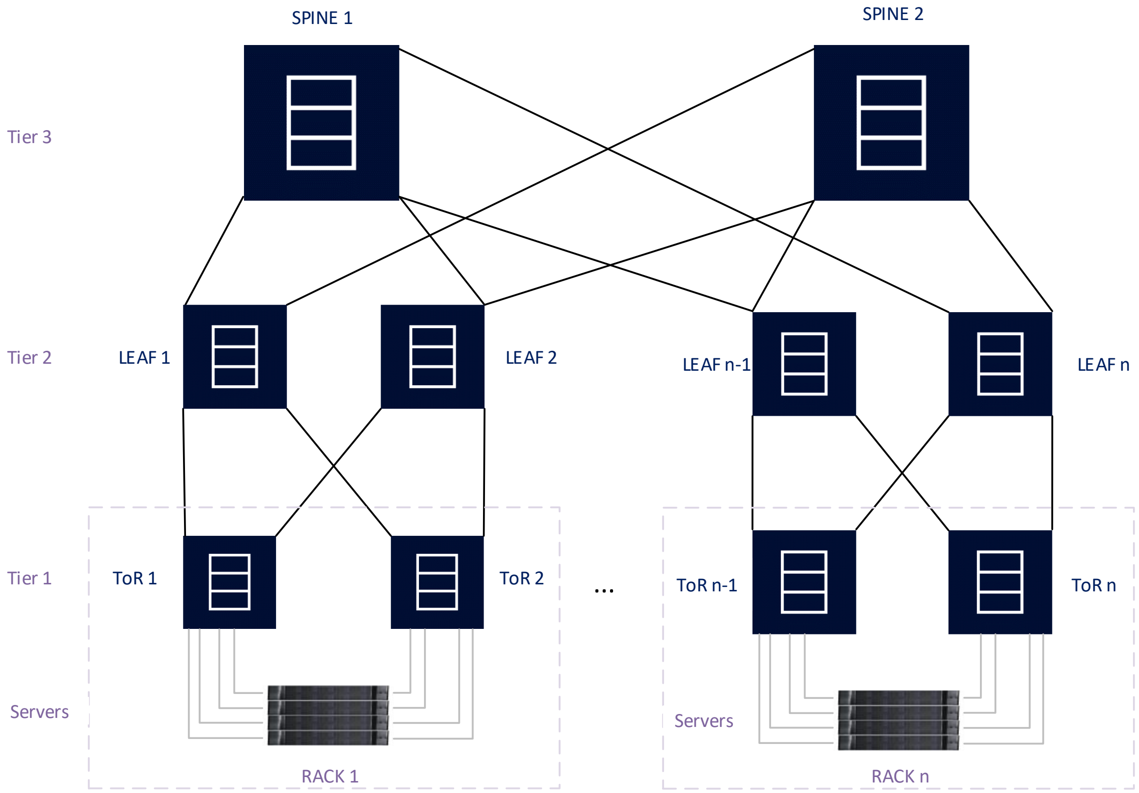


Figure 5-9: Three Tier Leaf-Spine Network

5.3.1 BGP in L3 DC

Most of the web-scale Data Centres running Layer 3 on it because only Layer 2 solution is not scalable for a hundred thousand of servers, but the Internet is the proof that Layer 3 protocols can scale without problems. In a Layer 3 Leaf and Spine network, each link is a routed link with its subnet. Equal Cost Multipath (ECMP) is used to load balancing traffic across all links between Leaf and Spine layers. In many cases a hybrid approach is used with a Layer 2 topology from the Servers to the Leaf switches and Multi-Chassis Link Aggregation (MC-LAG) to aggregate bandwidth in Layer 2. This L2 network section is required mainly for Virtual Machine (VM) migration between servers [26] [55].

There are several options to use as Layer 3 protocol as IGP, BGP or a combination of both, but IGP protocols are considered too complex and send to much routing information that is not used in most of the cases. In the BGP case, eBGP has some advantages over iBGP such as a full mesh session requirement is not needed, AS-path as routing loop prevention is easy to manage, and all routes learn from one peer are automatically send to other peers. For those reasons External BGP (eBGP) is used as sole Layer 3 protocol in very large Data Centres because it is simpler than any IGP and even iBGP, so eBGP sessions are established between Leaf and Spine switches. Different private ASes need to be used to separate the Spine from the Leaf, and because leaf switches are always many more than Spine switches that are going to be split in several private ASes to ease the management [26] [53].

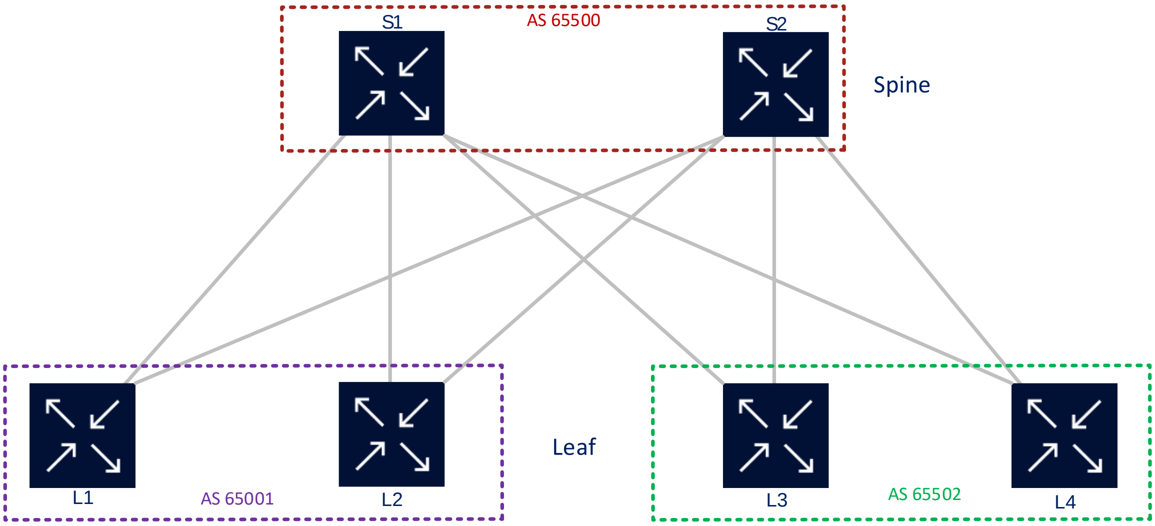


Figure 5-10: eBGP routing in the DC network

Figure 5-10 shown three different private ASes used to divide the Spine and Leaf layers, for Spine AS *65500* is used and the Leaf layer is split into two Ass *65001* and *65002*. Each link will be a network link and correspond to an eBGP session where prefix will be exchanged. Switches from AS *65502* will be propagated their internal prefixes to the Spine AS and then their switches will re-advertise those prefixes to Spine switches L1 and L2 inside AS *65001*.

The eBGP session between two switches is usually BGP-LU enable because a label will be associated with the node's loopback, and services could be deployed end-to-end using BGP-LU labels to build transport tunnels. This approach could be extended to support SR inside the Data Center network.

5.3.2 BGP Prefix-SID in DC

The same principles already studied applies when Segment Routing is active in a Data Center. Figure 5-11 shown a single path inside a DC Clos network with three tiers and using different AS for each layer, so each link between two switches has an eBGP-LU session where each switch loopback prefix and Prefix-SID index will be interchanged. This eBGP session will be established between two private ASes, for instance, AS *65022* is used for switch TOR2 and AS *65012* for switch L2. All switches have BGP SR enabled with [*32000-39999*] as SRGB and no IGP or LDP/RSVP-TE protocol is configured. As SR is enabled BGP-LU will use labels from the SRGB range using the label-index information propagated in the update message to allocate a local label.

ToR Switch TOR2 advertise its loopback prefix *1.0.0.12/32* with label-index *12* to leaf switch L2, this router learns that prefix and using SID index assign label *32012* as Prefix-SID for switch TOR2. Then, leaf switch L2 will follow a similar process to allocate a local label (*32012*) and distribute this prefix and Prefix-ID using the eBGP-LU session to the spine switch S1. Next, spine switch S1, after allocating a local label *32012*, advertise prefix and Prefix-ID from switch TOR2 using the eBGP-LU session to leaf switch L2. Finally, leaf switch L2 first allocates local label *32012* and then sends the prefix and Prefix-

SID information from switch TOR2 to switch TOR1. This last switch will be using the Prefix-SID index value (*12*) received from switch L1 to allocate a local label *32012* mapped to the loopback of switch TOR2.

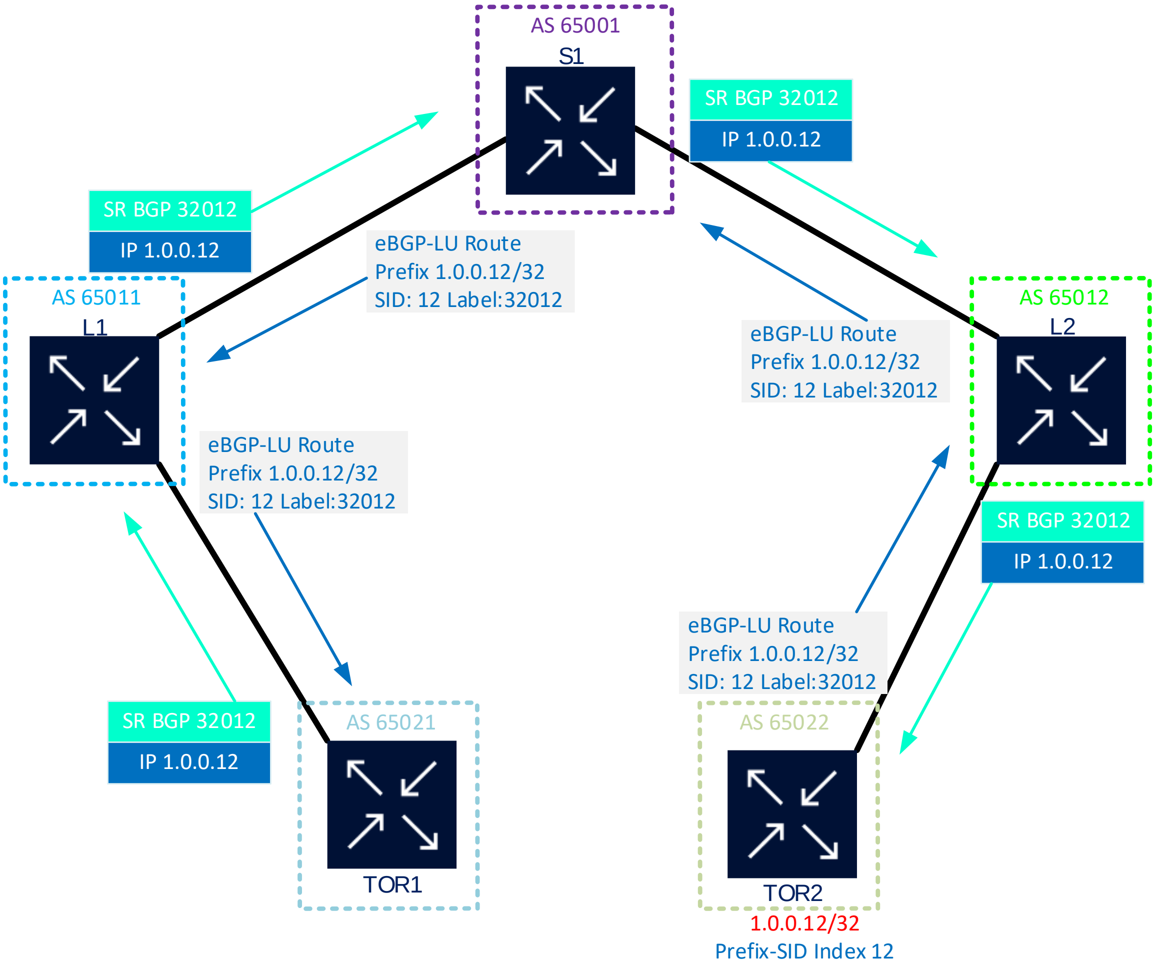


Figure 5-11: SR BGP in a Spine&Leaf three-tier network

Also, in figure 5-11, when an IP packet arrives to switch TOR1 with destination address *1.0.0.12*, this packet will be encapsulated with label *32012* for SR BGP and will be forwarded to switch L1 using TOR1-L1 interface as next-hop. In switch L1 the SR BGP label will be swapped for the same value and forwarded to switch S1. Then, in switch S1 SR BGP label *32012* will be swapped again with label *32012* and forwarded to switch L2. In this last hop, if switch TOR2 would send an explicit null label to switch L2, a pop action will be executed. Yet in this example switch TOR2 not used PHP behaviour, so SR BGP label will be swapped again and then the packet is forwarded to its final destination switch TOR2.

As long as BGP Anycast-SID is only a special case of Prefix-SID, the same process is followed for this kind of segment. The only consideration for Anycast-SID is that the label must be different from any Node-SID in the network. Another special case is interoperability between SR BGP capable routers and non-SR BGP capable nodes, as already seen in section 5.2.2 this is possible and could work without many considerations.

6 Topology Independent LFA

Fast Reroute is a resiliency mechanism that defines ways of automatically establishing protection paths for traffic before a failure arises. The router where the protection path originates is called Point of Local Repair (PLR), when a failure occurs the PLR installs the new path in the data plane and the traffic start flowing again after a brief interruption. The PLR also starts the IGP convergence to update the new forwarding entries along with the IGP domain [4].

The time interval between a fail happens and the traffic flows are re-established known as restoration time and must be equal or less than 50 ms, that time is an old standard time inherited from SDH/SONET transport networks. This time is dependable of the detection time that is usually 30 ms when Bidirectional Forwarding Detection (BFD) is used. BFD establishes an IPv4 or IPv6 session between a point-to-point link where an echo packet is sent in a periodic time basis (usually 10 ms), then if a specific number the packets are missing (usually 3) failure is detected and Fast Reroute process starts [56] [14].

The first technique in MPLS to restore traffic was RSVP-FRR where a primary LSP-path of an LSP signalled by RSVP-TE will have an alternative path calculated and signalled automatically. Router closest to the point of failure will recover the traffic using the protection tunnel calculated previously. This protection tunnel could be one-to-one where each LSP-path has a path signalled and dedicated to protecting only that LSP-Path. On the other hand, if an LSP is configured with facility FRR protection, a common bypass tunnel can be used to protect multiple primary LSP-Paths. Both types of FRR use by default Node Protection that protects against the failure of the next router, but if protection is needed in the link level Link protection could be used [4] [6].

During several years the only way to have FRR protection was using RSVP-TE along LDP until Loop-Free Alternate (LFA) was defined in RFC 5286. In that RFC IP-FRR appears and that could be used for LDP FRR also. From the configuration perspective using LFA instead of RSVP-FRR was easy, but as LFA is a topology dependant mechanism its coverage was lower than 100%. To improve this coverage several LFA iterations were defined until TI-LFA reach 100% network coverage [12].

Details of each LFA iteration is studied below.

6.1 Loop-Free Alternate

IP-FRR is a prefix-based FRR solution where the PLR pre-computes a path for each prefix in a similar way to one-to-one protection. IP-FRR is based on using a pre-computed alternate path so that when a failure is detected with the primary path, the alternate path can be rapidly used until an SPF is run by the IGP protocol and a new primary next hop is installed in the FIB. This approach has some advantages over facility mechanism because permits ECMP could be used to load balancing traffic for all paths available after a failure.

Also, an optimal backup path could be used as repair paths compared with RSVP-FRR where any path will be used as a repair path even the sub-optimal ones [14] [33].

IP-FRR provides local protection, so the first node to detect a failure active the protection path and start the IGP convergence. Failure condition is not propagated to all routers in the network. In the end each router calculates its backup paths for any destination without any signalling between nodes, that permits have a Node and link protection post-failure and pre-convergence [14].

To calculate a backup path a PLR must calculate a route whose shortest path follows different hops than the primary path, and this route cannot traverse the first hop of the primary path. If a node found a possible repair path for a specific destination produces a Micro-Loop no backup path will be created, so in this case coverage cannot reach 100%. The existence of a suitable loop-free alternate (LFA) and percentage of FRR coverage is dependent on topology [33].

6.1.1 Suitable Loop-Free Alternate.

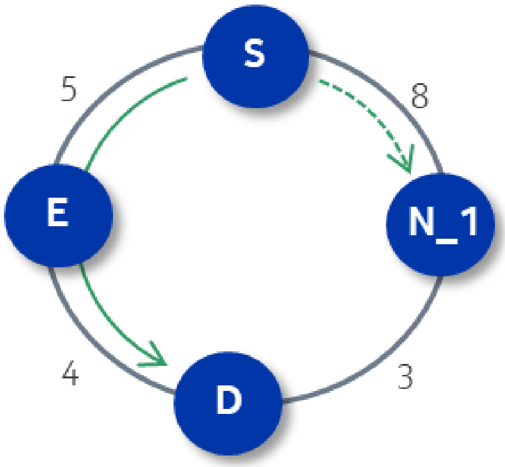


Figure 6-1: LFA path possible [33]

Figure 6-1 shown a ring topology used very often in the MBH network. In this topology Router, S computes the shortest path to router D and uses router E as its primary next-hop. Router S also computes that router N_1 is a feasible alternate next-hop for destination router D and installs it as a repair path [33].

Router N_1 is a loop-free alternate path because it’s metric to router D using its direct connection is *3* while the path that uses router S and D hops has a metric of *17*. Resulting router N_1 will follow its direct link path because that has a lower metric value than the other one.

If the link between routers S and E fails, router S forwards traffic to a pre-computed backup path; that means router N_1 becomes the first next hop of the repair path. Then, router N_1 forward traffic to its directly connected link to router D. Note router N_1 is not aware of S-E link fail, so from its point of view both paths are still available and the only reason to send traffic to N_1-D link is the lower metric [33].

6.1.2 No suitable loop-free alternate.

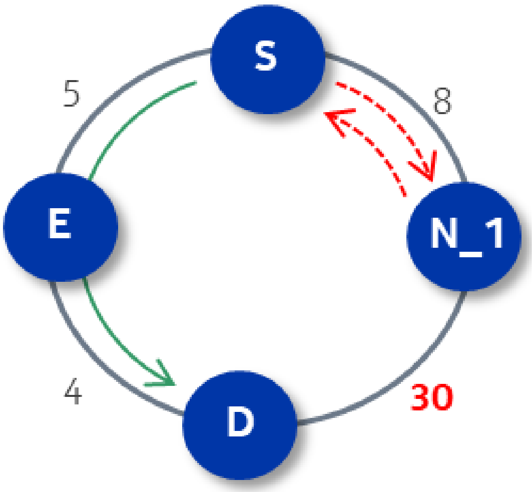


Figure 6-2: LFA path no possible [33]

Figure 6-2 shows the same topology as in figure 6-1, but this time link N_1-D metric is *30*. Again, router S computes the shortest path to router D and uses router E as its primary next-hop. Yet, in this case, router N_1 is not a suitable loop-free alternate, because the cost of the path from router N_1 to router D via router S would be *17*, whilst the cost from router N_1 directly to router D would be *30*. In short, if link S-E failed and router S forwarded traffic to router N_1, we would have a transient loop or micro-loop until the next SPF completes [33].

Again, router N_1 is not aware of S-E link fail, so from its point of view both paths are still available, and this time, the path through routers S and E has a lower metric than N_1-D link. Resulting router N_1 is going to send traffic back to router S and this one will send back the traffic to router N_1, producing a loop until IGP converges and route through link S-E is deleted from router N_1 FIB.

6.1.3 LFA loop-free condition.

A neighbour N can provide a loop-free alternate path only if it meets the inequality criterion:

Link Protect:

$$\textit{Distance}\ (N, D) < \textit{Distance}\ (N, S) + \textit{Distance}\ (S, D)$$

Equation 6-1: LFA condition for Link Protect [12]

Node Protect (Node E):

$$Distance\ (N, D) < Distance\ (N, E) + Distance\ (E, D)$$

Equation 6-2: LFA condition for Node Protect [12]

N in equations 6-1 and 6-2 is a neighbour of the PLR different from the main path resides, *E* is the PLR's neighbour where the main path traverse, and *D* is the destination node.

These equations could be applied validated mathematically examples analysed in 6.1.1 and 6.1.2. In the first case, distance (N_1, D) is equal to *3*, distance (N_1, S) is equal to *8* and distance (S, D) is equal to *9* (S-E=*5* + E-D=*4*). That means *3 < 8 + 9*, so LFA in point 6.1.1 is possible because *3* is minor than *17*.

In the second case, distance (N_1, D) is equal to *30*, distance (N_1, S) is equal to *8* and distance (S, D) is equal to *9* (S-E=*5* + E-D=*4*). That means *30 > 8 + 9*, so LFA in point 6.1.2 is not possible because *30* is greater than *17*.

Although the LFA FRR mechanism is really simple, it is not suitable for some topologies because its coverture will be minor than 100%, especially in ring topologies where this coverage could be as low as 60%. These drawbacks compelled to design a better mechanism with improving coverage.

6.2 Remote Loop-Free Alternate

Remote LFA (RLFA) was defined in RFC 7490 to extends the classic LFA repair mechanism to increase FRR coverage in the event of a failure of a router or link along the primary path between a given source and destination. If a link cannot be protected with any local LFA neighbours, RLFA tries to create a virtual LFA by using a repair tunnel to carry packets to a point in the network where they will not be looped back [13] [57].

When the primary downstream neighbour fails, the remote LFA mechanism is described using the reference point of a computing router (*S router*) that calculates an alternative path comprising of a repair tunnel to a point in the network where traffic will not be looped back to use the broken link (*P space*). At the point where traffic emerges from the repair tunnel router must be able to forward towards a destination router (*E router)* without using a failed link. If *P space* and *Q space* intersect in a given router (*PQ router*), then a repair tunnel could be used from *S* router to *PQ* router. That means a full LFA coverage could be reached in some cases [57].

When remote LFA is enabled, that classic LFA is also implicitly enabled, so any prefix protected using classic LFA may not be protected by a remote LFA repair tunnel. Classic LFA is always calculated first, then Remote LFA is calculated only for prefixes unprotected after classic LFA calculations [57].

It is no coincidence that figures used to define the RLFA mechanism in RFC are a ring topology because in a ring is where classic LFA coverage is really low and need to improve. Figure 6-3 shown a simple ring topology where link S-E cannot be fully protected by router S. To get router C from router S there are ECMP paths, but router D and router E cannot be protected by any LFAs. That is easy to validate applying equation 6-1 to this topology. First, if the destination is C, then distance (E, C) is equal to *200*, distance (E, S) is equal to *100* and distance (S, C) is equal to *300*. That means *200 < 100 + 300*, so LFA to C is possible because *200* is minor than *400*. On the other hand, if the destination is D main path is going through E, then distance (A, D) is equal to *300*, distance (A, S) is equal to *100* and distance (S, D) is equal to *200*. That means *300 = 100 + 200*, so LFA to router D is not possible because *300* is equal not minor than *300*.

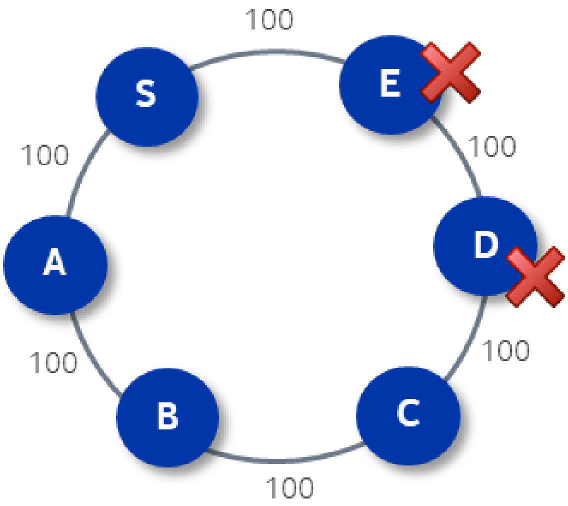


Figure 6-3: LFA partial coverage [33]

To calculate an RLFA repair path for link S-E, first, a group of routers which can be reached from router S without traversing the link S-E must determine, and then these routers must match with the set of routers that can reach router E without traversing the link S-E. These are the already mentioned router S's *Extended P space*, and router E's *Q space* respectively [57].

6.2.1 S's Extended P space:

A Short Path First calculation is made, using router S as root to calculate, the routers that can be reached without going through link S-E, so S's *P space* covers only routers A and B. Yet the only neighbour available, if a failure in link S-E occurs, is router A, so router A's *P space* must be calculated. This is referred to as the *Extended P space* of router S. By extending router S's SPF to include its neighbours, and including ECMP paths, the router S's *Extended P space* includes routers A, B, and C [33].

Table 6-1 shown how inequality equation 6-1 is applied to routers B, C, and D as destination router (*D*), router S as source router (*S*) and router A as a neighbour router (*N*), in order to find the *Extended P space* in figure 6-3. Note, distances are the optimum one that means the lower cost between two routers. In the end, only routers B and C met the conditions to be included in the *Extended P space* [57].

Router (D)	Distance (N, D)	Distance (N, S)	Distance (S, D)	Inequality condition met?
B	100	100	200	Yes (100<100+200)
C	200	100	300	Yes (200<100+300)
D	300	100	200	No (300<100+200)

Table 6-1: Extended P space calculation [57]

6.2.2 E's Q space:

The idea behind the calculation of E's *Q space* concerning the S-E link is found in which routers can reach router E without going through link S-E. A reverse SPF computation must be done to find this space that means calculate cost towards the root router rather than from it and yields the best paths towards the root router from other nodes in the network. In figure 6-3 example, this compute will be rooted at router E, and the result equates to routers D and C [33].

Table 6-2 shown how inequality equation 6-1 is applied to routers B, C, and D as destination router (*D*), router S as source router (*S*) and router E as a neighbour router (*N*), in order to find the *Q space* in figure 6-3. Note, the reference point is interchanged because the reverse SPF, for instance now is the distance from router Neighbour to router Destination instead of router Destination to router Neighbour. In the end, only routers C and D met the conditions to be included in the *Q space* [57].

Router (D)	Distance (D, N)	Distance (S, N)	Distance (D, S)	Inequality condition met?
B	300	100	200	No (300<100+200)
C	200	100	300	Yes (200<100+300)
D	100	100	200	No (100<100+200)

Table 6-2: E's Q space calculation [57]

Figure 6-4 shown routers A, B and C conform to the router S's *Extended P space*, and router C and D are part of the router E's *Q space*.

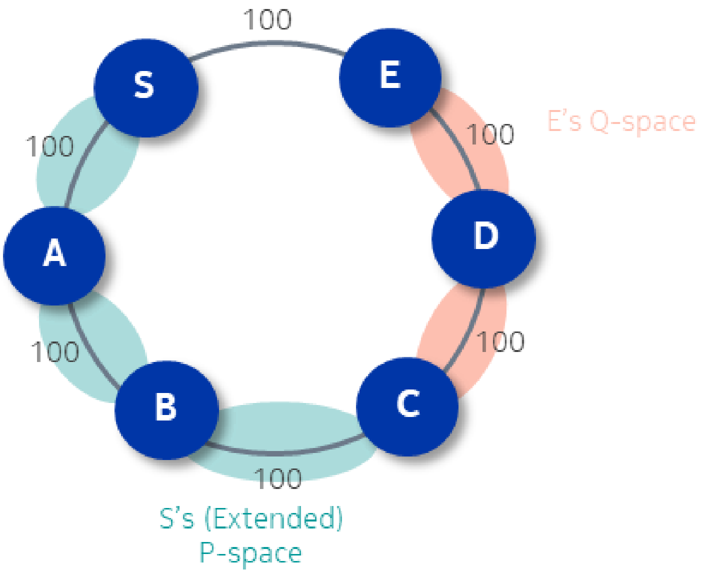


Figure 6-4: Remote LFA P space and Q space [33]

6.2.3 PQ Router:

The node where the *P space* and *Q space* intersect is referred to as the *PQ router* and is the point to which a repair tunnel can be built. In figure 6-4 case, correspond to router C. So, a repair tunnel is built from router S to C, and node C becomes a neighbour of router S through that tunnel, that would become a Remote LFA for routers D and E. Note that this repair tunnel can only be used for repair traffic and not for service transport. Figure 6-5 shows how the repair tunnel is built from router S to router C and *PQ router* location in router C [33].

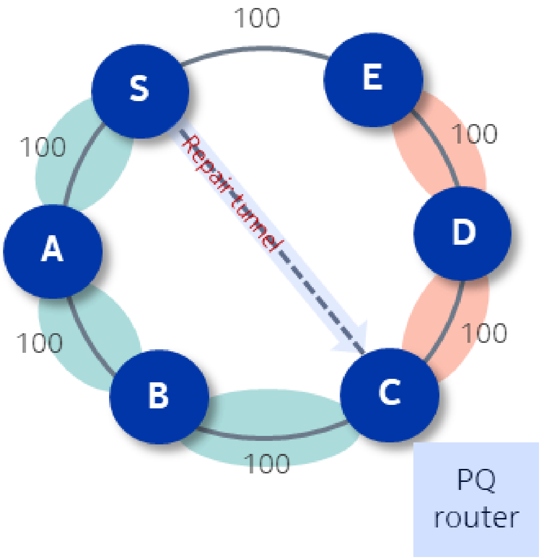


Figure 6-5: R-LFA Repair Tunnel and PQ router [33]

6.2.4 Repair Tunnel:

RFC 7490 makes no assertions about what MPLS protocol or SR is used to create the repair tunnel, but if this is an MPLS network then a label stack may be used to provide the tunnel that will protect link S-E. As long as IP FRR is intended for LDP in an MPLS network, because RSVP-TE has its FRR mechanisms, so prior to the introduction of Segment Routing, the repair tunnel must be LDP-over-RSVP, or LDP-over-LDP [33].

If LDP-in-RSVP is used, the RSVP LSP must be manually established from the source to the repair point and configured for LFA SPF only. Also, targeted LDP runs inside the RSVP LSP so a peer relationship between source and *PQ router* must be established [57].

On the other hand, LDP-in-LDP requires dynamic creation of targeted LDP sessions from the source router to potential *PQ routers* in order to signal inner labels. After a link failure, these sessions need to be dynamically torn down, cleaned up, and re-established to the new post-convergence repair points. This dynamic behaviour is not favoured by many operators and is not currently supported by most of the vendors [57].

RLFA could use SR to build the repair tunnel. In this case, if link S-E fails, S router pushes node SID of router C to the top of the stack and forwards to router A. Then, router A looks up node SID for router C and forwards the packet to router B. At this point, router B looks up for node SID for router C and forwards the packet to router C. Finally, at router C, Node SID is popped, and the packet is forwarded to router D because this is the shortest path to router E [57].

Although RLFA could reach a coverage of 95-99% according to real data sets analysed in RFC 7490, T-LDP sessions and RSVP-TE tunnels have to be created if SR is not enabled in the network. In addition, these issues create management and scalability problems, and for that reason in some cases is preferable to maintain active the old, reliable RSVP-TE FRR than migrate to RLFA. To sum up, RLFA extends coverage of classic LFA but is complex to understand and implement because it requires additional signalling protocols and manual configurations to work [14].

6.3 TI - Loop-Free Alternate

As already mentioned, classic LFA is a really simple protocol to configure and implement but is topology dependent and its performance is especially low in ring topologies. Remote LFA tries to fix the coverage problem reaching in the best case to 99% of coverage, but that increases administrative complexity because T-LDP and even RSVP-TE tunnels must be configured. Furthermore, both mechanisms not always select the most optimal repair path and manual tuning need to be applied to overcome this problem [14].

In order to resolve all these limitations while it is keeping simple as classic LFA but with 100% coverage and optimal repair paths, Topology Independent LFA (TI-LFA) was defined in *draft-ietf-rtgwg-segment-routing-ti-lfa-02*. TI-LFA relies on SR to build a repair tunnel, so Segment Routing must be already deployed in the network before TI-LFA begins to work; however in a hybrid network where LDP is still running, TI-LFA could protect LDP or IP traffic [14].

Classic LFA and RLFA calculate a pre-computed backup path, at the PLR so that they can switch from primary next-hop to backup next hop in less than 50 ms. During a failure, however, there are frequently two transitions in the repair path; a pre-convergence path to FRR, and then FRR to post-convergence path. When both paths are different, the repair path before IGP convergence is suboptimal [33].

Figure 6-6 shown an example where paths pre and post-convergence are different. In this topology, router S attempts to find an LFA for router R2 to protect link S-D. No local LFA next-hop exists via the shortest path (metric 40) through routers N2-C-D because inequality equation is not met (30<10+10), so router S must use the longer path (metric 60) via routers N1-A-B-D for the backup path; here inequality equation is met (30<30+10). At the event of a failure of link S-D, the first transition moves traffic to the backup path through routers N1-A-B-D. Once the IGP has re-converged, the second post-convergence transition moves traffic to the shortest path via routers N2-C-D. This applies to classic LFA and RLFA because the last one is unable to find a PQ router for the shortest path, so both algorithms will have a backup path by a suboptimal path until the IGP convergence [11] [33].

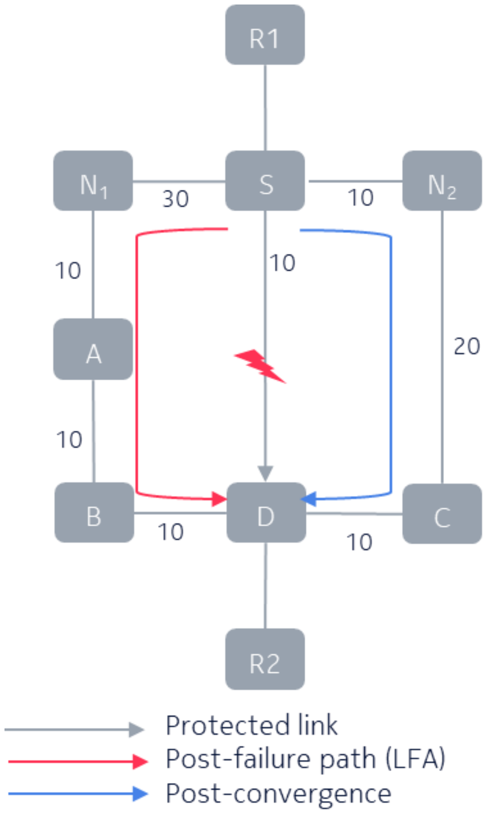


Figure 6-6: LFA and RLFA path transitions [11]

TI-LFA intends to reduce this double transition to a single one, which means pre-convergence to post-convergence transition only. This goal is achieved using IGP to automatically calculate the repair path, so no other protocols are required and the post-FRR path will be the same as IGP post-convergence. Aside from reducing the number of service disruptions, it also allows backup paths to be better capacity planned because that will be always the optimal backup path. By using Segment Routing with RLFA, 100% coverage also is guaranteed against link and node failure [33].

For each prefix or destination router this mechanism runs first a classic LFA calculation, then computes an Explicit Post-Convergence (EPC) repair path from the PLR for destination router D by computing the intersection of the candidate P-Q nodes with the post-convergence SPF for each destination. Finally, RLFA is run for prefixes unprotected after LFA and TI-LFA computation [11] [33].

6.3.1 TI-LFA Calculation

In figure 6-7, the protected path is link S-D and the shortest path pre-failure between R1 and R2 is through routers R1-S-D-R2. The Post-convergence path is through routers R1-S-N2-C-D-R2 because it has a cost of 40 which is lower than via N1 whose cost is 60.

6.3.1.1 Compute LFA for N2

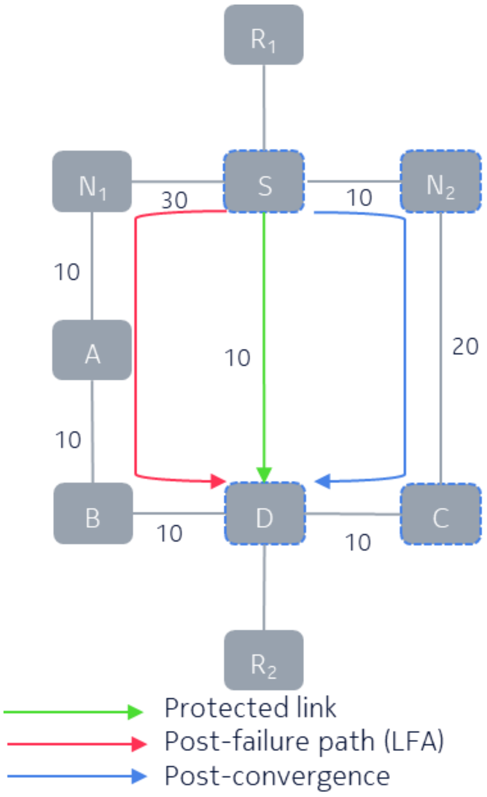


Figure 6-7: LFA for router N2 [11]

Path R1-S-N2-C-D-R2 does not satisfy the inequality equation (equation 6-1) because distance from router N2 to D is *30*, but the addition of distance N2 to S (*10*) plus distance S to D (*10*) is only *20*. So, *30<10+10* is false. That means there is no LFA available in this backup path.

6.3.1.2 Compute LFA for N1

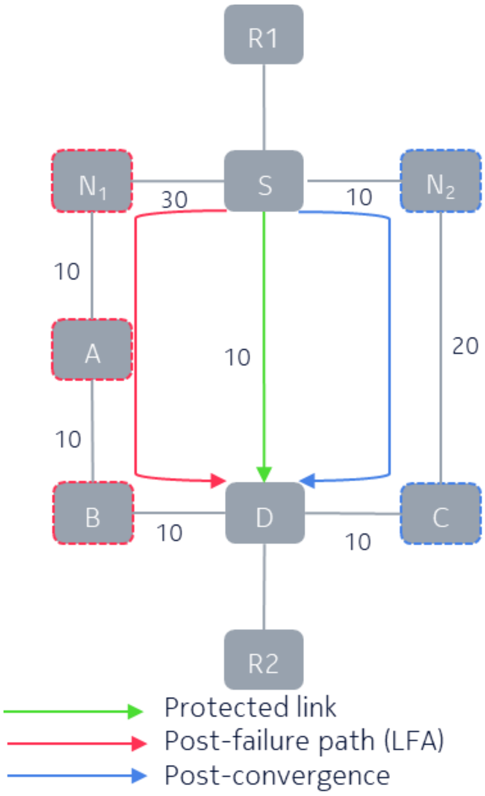


Figure 6-8: LFA for router N2 [11]

Figure 6-8 shown path R1-S-N1-A-B-D-R2 which satisfy the inequality equation (equation 6-1) because distance from router N1 to D is *30*, and the addition of distance N1 to S (*30*) plus distance S to D (*10*) is *40*. So, *30<30+10* is true. That means router N1 is a valid LFA next hop but is not the post-convergence path.

6.3.1.3 Compute TI-LFA on post-convergence path

The post-convergence and post-failure path must be the same, however, after classic LFA calculation post-failure path is going via router N1 and the post-convergence path is going via router N2.

Figure 6-9 shown the post-convergence path has no Remote LFA coverage because the *Extended P space* and *Q space* are disjointed. The *Extended P space* of router S is only formed by nodes N2 and S itself because router S cannot reach C without using protected link S-D. *Q space* of router D is only formed by nodes C and D itself because router C is the only node that router D can reach using reverse SPF without using protected link S-D [11].

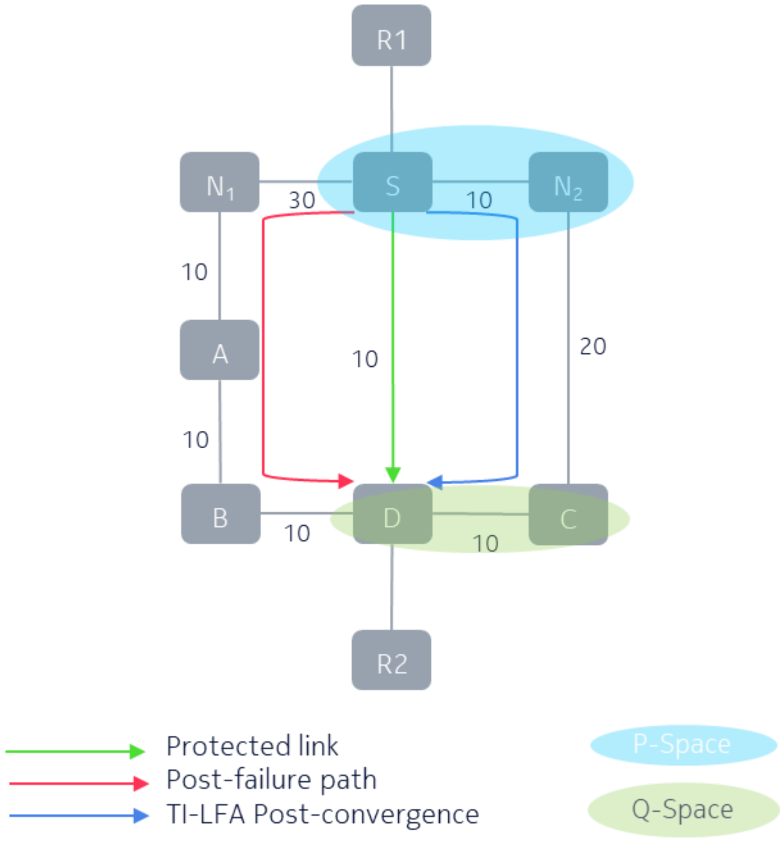


Figure 6-9: P space and Q space after RLFA

To solve this problem, router S could install an SR repair tunnel to router C which is the closest *Q node* in the post-convergence path.

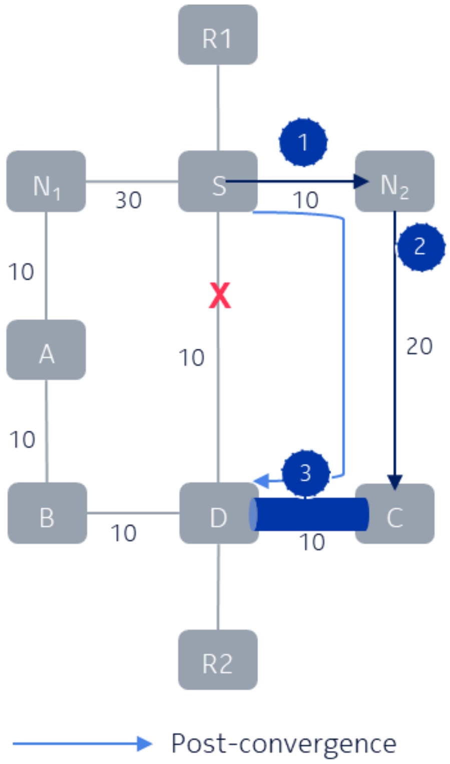


Figure 6-10: Repair tunnel to router C with 1 FRR label

Figure 6-10 shown steps to build the repair tunnel to router C: First, traffic is forwarded to router N2, router N2's node-SID is not required, as N2 is an adjacent neighbour of the repairing router S. Second, an Adj-SID is imposed to direct traffic emerging from the tunnel at router N2 towards router C. Third, node-SID of router D must be imposed, so that when Adj-SID is popped at router N2, router C can read the top label and forward on to D. The Repair tunnel for link S-D use one FRR label imposed at router S, that is the Adj-SID of N2-C at router N2. However, router D node-SID is not an FRR label. [11].

6.3.1.4 SR Label stack for TI-LFA

The segment list for the TI-LFA explicit repair tunnel will depend on the location of the repair node:

- If the repair node is a direct neighbour, the repair list is empty as in a local LFA repair.
- If the repair node is a PQ node, the repair list is made of a single node segment to that node as in an RLFA repair tunnel.
- If the repair node is a Q node, a neighbour of the last P node, the repair list is made of two segments; a node-SID to the adjacent P node, and an Adj-SID from that node to the repair node as is shown in figure 6-11.
- If there is a topology where P and Q nodes are not adjacent to the post-convergence path, the PLR can impose additional segments to compute a loop-free path between P and Q [33].

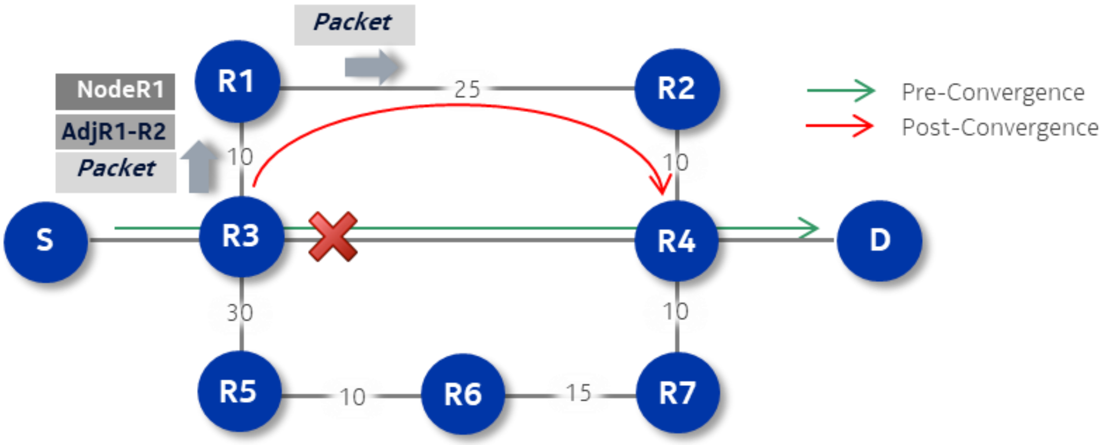


Figure 6-11: SR FRR repair tunnel

Most of the cases analysed in real network scenarios (99.72%) shown a maximum depth of the FRR stack is only 2 segments for link protection in networks with symmetric IGP metrics. More may be required for some node protection cases, 99.96% of destinations are protected with up to 3 segments, and no more of 4 segments are needed for a 100% coverage [14] [58].

6.3.1.5 P router not adjacent to router S

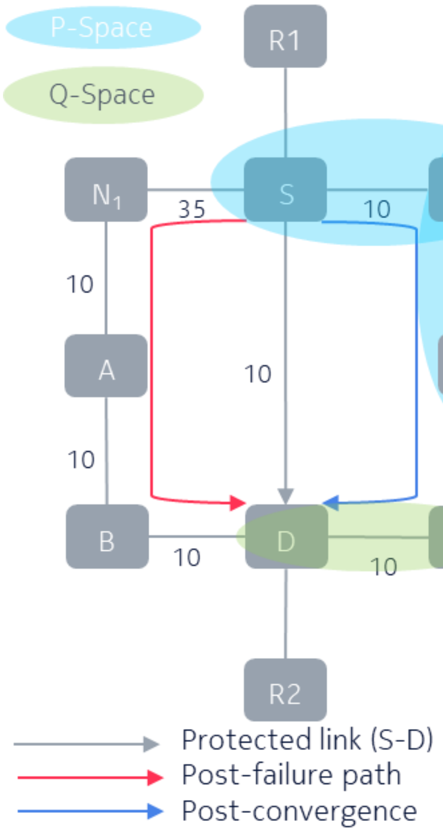


Figure 6-12: P router not adjacent to router S [11]

Figure 6-12 shown a case where P router is not adjacent to computing router S in contrast with figure 6-10 where there is a P router adjacent to router S. In this topology *P Space* is extended to include router E, and router S installs a repair tunnel to closest Q node, router C. In order to build that tunnel router S will push two additional labels, one is router E's Node-SID which correspond to the Shortest Path tunnel to E, and Adj-SID of link E-C at router E, to direct traffic emerging from the tunnel at router E towards router C.

By definition, TI-LFA must provide a post-failure and post-convergence on the same path, so in figure 6-13 shown steps to build the repair tunnel to router C via N2: First, traffic is forwarded to router E using node-SID of router E. Second, an Adj-SID is imposed to direct traffic emerging from the tunnel at router E towards router C, using link E-C. Third, node-SID of router D must be imposed, so that when Adj-SID is popped at router E, router C can read the top label and forward on to D. The Repair tunnel for link S-D use two FRR labels imposed at router S, that is node-SID of router E and Adj-SID of E-C link at router E. However, router D node-SID imposed at the end is not an FRR label. [11].

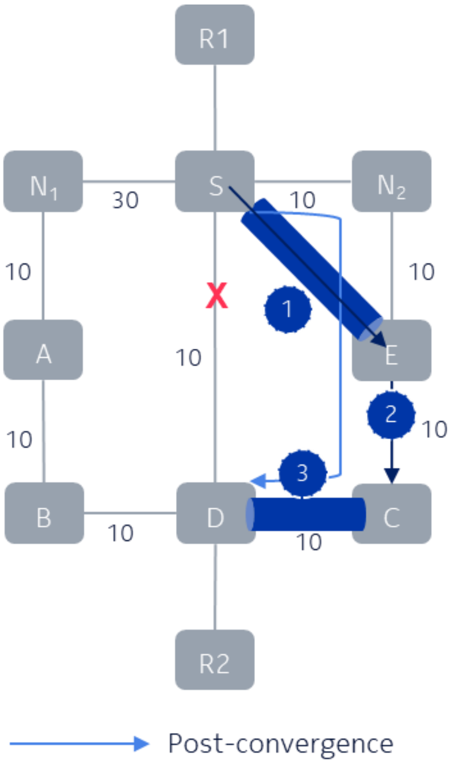


Figure 6-13: Repair tunnel with 2 FRR labels

6.3.2 TI-LFA for LDP

6.3.2.1 SR to LDP Interworking

In order to protect a network MPLS with LDP using TI-LFA, SR routing must be enabled and must interoperate with LDP. To interoperate, both forwarding entries have to be installed into the MPLS data-plane, so those entries need to be unique to be installed [33].

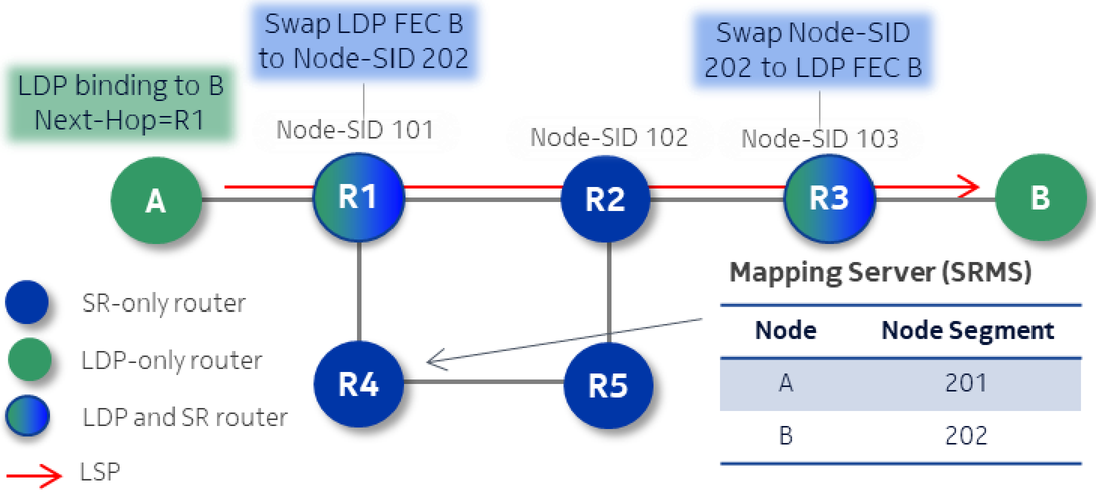
Also, these control planes could co-exist in two scenarios. First, LDP and SR are present on all routers in the network, and preference for LDP or SR is a local decision at the head-end router. In this scenario, SR can also be used to enhance FRR coverage [33].

Second, when SR is only present in parts of the network, an SR Mapping Server (SRMS) is needed to LDP and SR can be interworked to provide an end-to-end tunnel or an FRR tunnel. Also, key nodes on the repair path need to be SR enabled in order to the SR/LDP internetworking functionality built an end-to-end labelled path [14].

Figure 6-14 shown an example where one or more SRMS are used to advertise Node-SIDs on behalf of non-SR routers. For example, router R4 advertises Node-SIDs 201 and 202 respectively for the LDP-only routers A and B. Router A forward packets to router R1 using LDP. Then, router R1 does not have an LDP label binding for its next-hop router R2 but does have an SR Node-SID to this router, so it swaps its local LDP label for FEC

B to Node-SID 202 and forwards to router R2. This router is SR only, so it swaps the inbound Node-SID label 202 for the same value before sending the packet to router R3.

When the packet arrives to router R3 knows that router B is not SR-capable because router B did not advertise SR capability in the IGP protocol, so router R3 swaps Node-SID for LDP FEC B.



Mapping Server (SRMS)

Node	Node Segment
A	201
B	202

Figure 6-14: SR and LDP internetworking using SRMS [33]

6.3.2.2 SR to LDP Fast Reroute

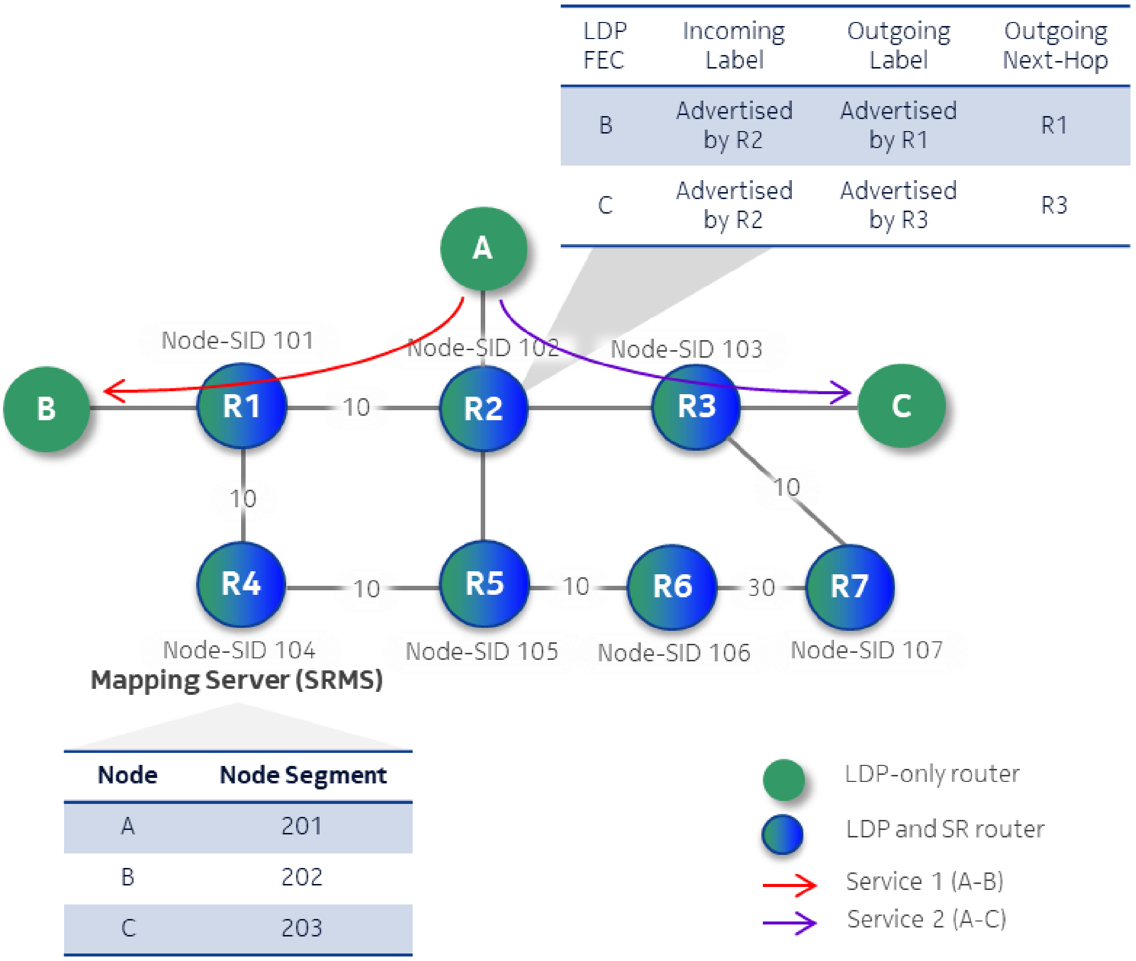
An analogous procedure with LDP-SR interworking can be used to provide FRR coverage improvement, so as already mention two cases are possible:

First, SR is present only in parts of the network, so applying TI-LFA to this network has the potential for increased coverage. Also, an SRMS is needed for mapping labels between LDP and SR. Note full coverage is not possible here.

Second, if SR is present on all routers a full TI-LFA coverage will be reaching for the LDP traffic and a Mapping Server is no mandatory [33].

In figure 6-15 LDP is used by all routers in the network, and SR is enabled only in routers R1 to R7. Router R4 make functions of SRMS, so that mapped LDP only routers router-ID prefixes with a Node-SID. Then Mapping Server advertises Node-SID 201, 202, 203 respectively for the LDP only routers A, B, and C to all SR capable routers. Also, routers R1 to R7 have LFIB entries with labels associated with LDP FECs of routers B and C, so an LSP using LDP could be automatically created and used as tunnel transport [33].

In router A two services had been created, service 1 in direction to router B and service 2 towards router C. Router A is the head-end node, so that choose LDP as the preferred transport protocol because this is the only signalling protocol locally enabled. In this example the goal is to protect link R2-R1 for service 1, and link R2-R3 for service 2 using TI-LFA [33].



LDP FEC	Incoming Label	Outgoing Label	Outgoing Next-Hop
B	Advertised by R2	Advertised by R1	R1
C	Advertised by R2	Advertised by R3	R3

Node	Node Segment
A	201
B	202
C	203

Figure 6-15: TI-LFA in an LDP/SR network

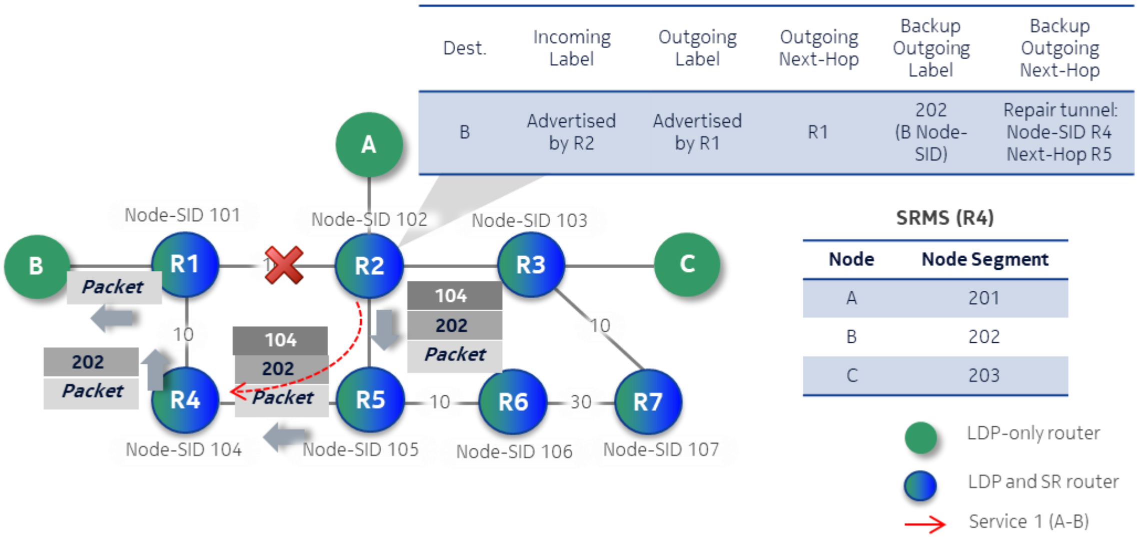
Figure 6-16 shown how TI-LFA is calculated to protect link R2-R1 and thus Service 1 with an LFA for router B. First, routers R1 to R7 advertise Node-SID and Adj-SIDs for their IGP adjacencies as usual in a Segment Routing domain. Router R4 is acting as Mapping Server for LDP only routers A, B and C [33].

In normal conditions LDP is used as the preferred transport tunnel for Service 1 because the source router A is LDP only and routers R2 and R1 support LDP as well. This service is going through routers A, R2, R1 and B.

When a failure arises in link R2-R1, router R2 swaps the incoming top label which is an LDP label with the Node-SID 202 for router B. Then, router R2 impose router R4 Node-SID 104 and sends the packet into a repair tunnel headed to router R4.

After the packet arrives to router R5 that swaps the top label 104 for the same label which is the Node-SID of router R4 and then forwards the packet to router R4 because PHP is not enabled in this case.

Finally, router R4 pops top label 104 and swaps label 202 for the same label which is the Node-SID of router R1 and forwards to router R1. Router R1's next-hop to router B is not SR-capable, so router R1 swaps label 202 for the LDP label announced by router B, in this case is implicit-null [33].



Dest.	Incoming Label	Outgoing Label	Outgoing Next-Hop	Backup Outgoing Label	Backup Outgoing Next-Hop
B	Advertised by R2	Advertised by R1	R1	202 (B Node-SID)	Repair tunnel: Node-SID R4 Next-Hop R5

SRMS (R4)

Node	Node Segment
A	201
B	202
C	203

Figure 6-16: Tunnel repair for Service 1

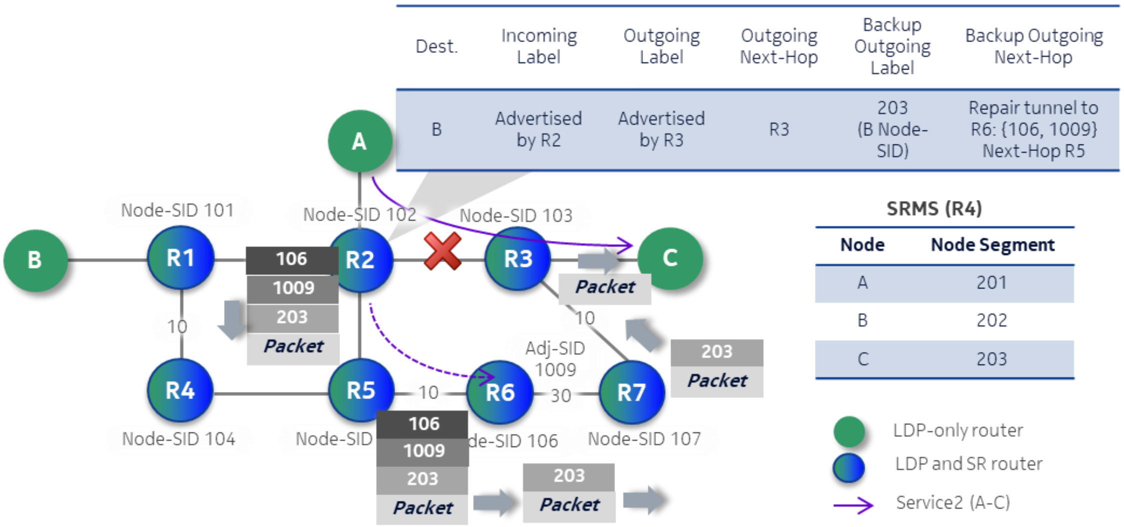
Figure 6-17 shown how a tunnel repair is calculated to protect link R2-R3 and thus Service 2 with an LFA for router C. Routers R1 to R7 advertise Node-SID and Adj-SIDs for their IGP adjacencies as usual in a Segment Routing domain. Router R4 is acting as Mapping Server for LDP only routers A, B and C [33].

In normal conditions LDP is used as the preferred transport tunnel for Service 2 because the source router A is LDP only and routers R2 and R3 support LDP as well. This service normally goes through routers A, R2, R3 and C.

In a failure case in link R2-R3, router R2 swaps the incoming top label which is an LDP label with the Node-SID 203 for router C. Then, router R2 imposes router R6 Node-SID 106 followed by Adj-SID 1009 which represents the R6-R7 link in router R6.

Then, Router R2 forwards the label stack comprised by {106, 1009, 203} to router R5. When this packet arrives to router R5 that swaps top label 106 for the same label which is the Node-SID of router R6 and forwards the packet to router R6 with PHP disabled.

Router R6 pops Node-SID label 106 and Adj-SID label 1009, then swaps label 203 for 203 and forwards the packet to router R7. When the packet gets router R7 that swaps label 203 for the same label which is the Node-SID of router R3 and forwards to router R3. Router R3's next-hop to router C is not SR-capable, so router R3 swaps label 203 for the LDP label announced by router B which is implicit-null [33].



Dest.	Incoming Label	Outgoing Label	Outgoing Next-Hop	Backup Outgoing Label	Backup Outgoing Next-Hop
B	Advertised by R2	Advertised by R3	R3	203 (B Node-SID)	Repair tunnel to R6: {106, 1009} Next-Hop R5

SRMS (R4)

Node	Node Segment
A	201
B	202
C	203

Figure 6-17: Tunnel repair for Service 2

7 Segment Routing Applications:

At this point all basic principles of Segment Routing and technologies associated have been analysed carefully, therefore SR practical applications are going to be proved since the basic steps to enable SR in a network using OSPF and ISIS and without LDP or RSVP-TE. TI-LFA is going to be deployed as an FRR mechanism in an SR enabled network. Distributed Traffic Engineering will be used for the sake of simplicity.

In this section topologies, configurations and shows are going to be displayed to test Segment Routing in several scenarios and applications.

7.1 Basic Premises

A server located in Ottawa is going to be used to build all scenarios, virtual Nokia Service Router 7750 SRc12 using Service Routing Operation System (SROS) images version 16.0.6 will be deployed. This virtual platform has a Controller card active and an interface card with 5 ports GE.

Most of the tests are going to use topology shown in figure 7-1 if a scenario needs a different topology that will be displayed in the pertinent section. This topology is formed by six Nokia virtual routers interconnected by interfaces shown in this figure and detailed in table 8-1 along with the IP planning for network links and router loopbacks.

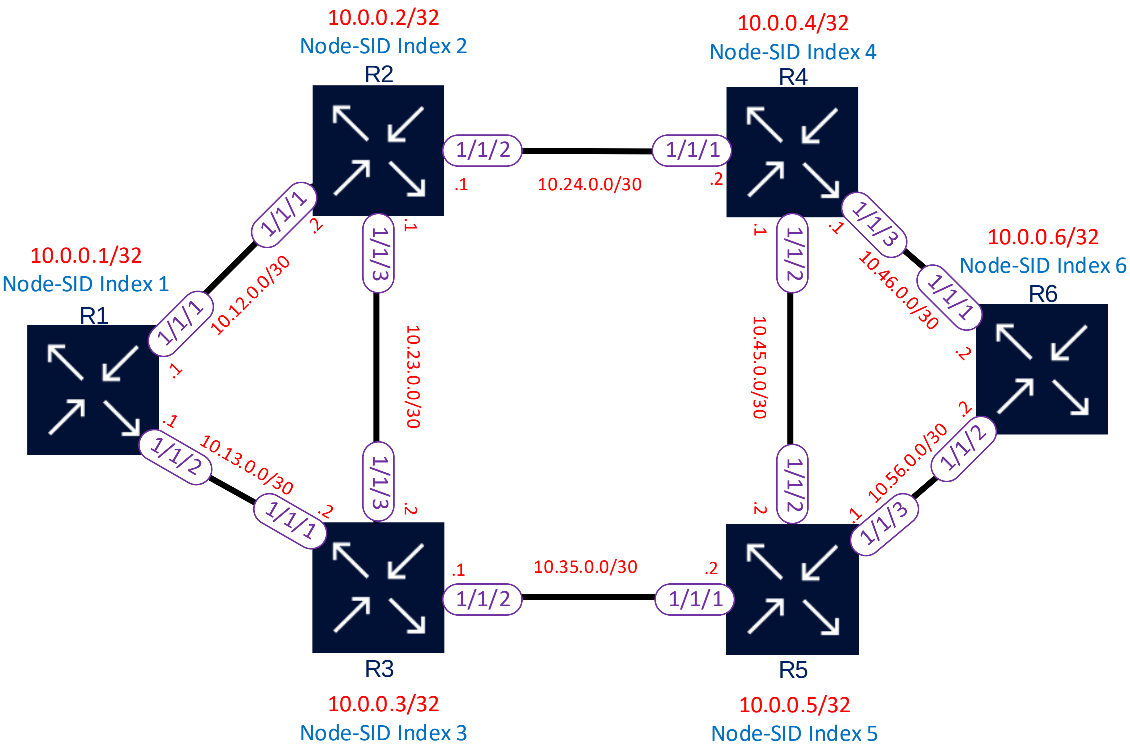


Figure 7-1: Basic Topology for Test

Router	Router ID /32	Node-SID	Node-SID Index	Port	Network IP /30
R1	10.0.0.1	32001	1	1/1/1	10.12.0.1
				1/1/2	10.13.0.1
R2	10.0.0.2	32002	2	1/1/1	10.12.0.2
				1/1/2	10.24.0.1
				1/1/3	10.23.0.1
R3	10.0.0.3	32003	3	1/1/1	10.13.0.2
				1/1/2	10.35.0.1
				1/1/3	10.23.0.2
R4	10.0.0.4	32004	4	1/1/1	10.24.0.2
				1/1/2	10.46.0.1
				1/1/3	10.45.0.1
R5	10.0.0.5	32005	5	1/1/1	10.35.0.2
				1/1/2	10.45.0.2
				1/1/3	10.56.0.1
R6	10.0.0.6	32006	6	1/1/1	10.46.0.2
				1/1/2	10.56.0.2

Table 7-1: Planning IP Test SR

7.1.1 Basic Configurations

Previous to enable SR in a router, ports (layer 2) and interfaces (layer 3) configurations need to be deployed using the information from figure 7-1 and table 7-1. Nokia differentiates layer 2 configurations and layer 3 configurations using ports and interfaces respectively.

Note in Nokia routers a special loopback called “system” need to be configured as router-ID, this loopback must use /32 masks and be configured inside all routing protocols enabled in a router.

7.1.1.1 R1 Ports and Interfaces

```
configure
#--------------------------------------------------
echo "Port Configuration"
#--------------------------------------------------
    port 1/1/1
        description "to_R2"
        ethernet
        exit
        no shutdown
    exit
    port 1/1/2
        description "to_R3"
        ethernet
        exit
        no shutdown
    exit
#--------------------------------------------------
echo "Router (Network Side) Configuration"
#--------------------------------------------------
    router Base
        interface "system"
            address 10.0.0.1/32
            no shutdown
        exit
        interface "toR2"
            address 10.12.0.1/30
            port 1/1/1
            no shutdown
        exit
        interface "toR3"
            address 10.13.0.1/30
            port 1/1/2
            no shutdown
        exit
```

7.1.1.2 R2 Ports and Interfaces

```
configure
#--------------------------------------------------
echo "Port Configuration"
#--------------------------------------------------
    port 1/1/1
        description "toR1"
        ethernet
        exit
        no shutdown
    exit
    port 1/1/2
        description "toR4"
        ethernet
        exit
        no shutdown
    exit
    port 1/1/3
        description "toR3"
        ethernet
        exit
        no shutdown
    exit
#--------------------------------------------------
echo "Router (Network Side) Configuration"
#--------------------------------------------------
    router Base
        interface "system"
            address 10.0.0.2/32
            no shutdown
        exit
        interface "toR1"
            address 10.12.0.2/30
            port 1/1/1
            no shutdown
        exit
        interface "toR3"
            address 10.23.0.1/30
            port 1/1/3
            no shutdown
        exit
        interface "toR4"
            address 10.24.0.1/30
            port 1/1/2
            no shutdown
        exit
```

7.1.1.3 R3 Ports and Interfaces

```
configure
#--------------------------------------------------
echo "Port Configuration"
#--------------------------------------------------
    port 1/1/1
        description "toR1"
        ethernet
        exit
        no shutdown
    exit
    port 1/1/2
        description "toR5"
        ethernet
        exit
        no shutdown
    exit
    port 1/1/3
        description "toR2"
        ethernet
        exit
        no shutdown
    exit
#--------------------------------------------------
echo "Router (Network Side) Configuration"
#--------------------------------------------------
     router Base
        interface "system"
            address 10.0.0.3/32
            no shutdown
        exit
        interface "toR1"
            address 10.13.0.2/30
            port 1/1/1
            no shutdown
        exit
        interface "toR2"
            address 10.23.0.2/30
            port 1/1/3
            no shutdown
        exit
        interface "toR5"
            address 10.35.0.1/30
            port 1/1/2
            no shutdown
        exit
```

7.1.1.4 R4 Ports and Interfaces

```
configure
#--------------------------------------------------
echo "Port Configuration"
#--------------------------------------------------
    port 1/1/1
        description "toR2"
        ethernet
        exit
        no shutdown
    exit
    port 1/1/2
        description "toR5"
        ethernet
        exit
        no shutdown
    exit
    port 1/1/3
        description "toR6"
        ethernet
        exit
        no shutdown
    exit
#--------------------------------------------------
echo "Router (Network Side) Configuration"
#--------------------------------------------------
     router Base
        interface "system"
            address 10.0.0.4/32
            no shutdown
        exit
        interface "toR2"
            address 10.24.0.2/30
            port 1/1/1
            no shutdown
        exit
        interface "toR5"
            address 10.45.0.1/30
            port 1/1/2
            no shutdown
        exit
        interface "toR6"
            address 10.46.0.1/30
            port 1/1/3
            no shutdown
        exit
```

7.1.1.5 R5 Ports and Interfaces

```
configure
#--------------------------------------------------
echo "Port Configuration"
#--------------------------------------------------
    port 1/1/1
        description "toR3"
        ethernet
        exit
        no shutdown
    exit
    port 1/1/2
        description "toR4"
        ethernet
        exit
        no shutdown
    exit
    port 1/1/3
        description "toR6"
        ethernet
        exit
        no shutdown
    exit
#--------------------------------------------------
echo "Router (Network Side) Configuration"
#--------------------------------------------------
     router Base
         interface "system"
            address 10.0.0.5/32
            no shutdown
        exit
        interface "toR3"
            address 10.35.0.2/30
            port 1/1/1
            no shutdown
        exit
        interface "toR4"
            address 10.45.0.2/30
            port 1/1/2
            no shutdown
        exit
        interface "toR6"
            address 10.56.0.1/30
            port 1/1/3
            no shutdown
        exit
```

7.1.1.6 R6 Ports and Interfaces

```
configure
#--------------------------------------------------
echo "Port Configuration"
#--------------------------------------------------
    port 1/1/1
        description "toR4"
        ethernet
        exit
        no shutdown
    exit
    port 1/1/2
        description "toR5"
        ethernet
        exit
        no shutdown
    exit
#--------------------------------------------------
echo "Router (Network Side) Configuration"
#--------------------------------------------------
     router Base
        interface "system"
            address 10.0.0.6/32
            no shutdown
        exit
        interface "toR4"
            address 10.46.0.2/30
            port 1/1/1
            no shutdown
        exit
        interface "toR5"
            address 10.56.0.2/30
            port 1/1/2
            no shutdown
        exit
```

7.2 SR-OSPF

7.2.1 OSPF and SR Configuration

All routers are going to be inside the OSPF Area 0, and all network interfaces are going to be point-to-point. The interface “system” must be included inside Area 0 to function as router-ID to form adjacencies.

In the Segment Routing part, the same SRGB range (32000 to 39999) must be defined in all routers, then the Prefix-SID range is declared Global and SR is enabled inside OSPF. Finally, each router sends its SR capabilities in an area scope, and a Node-SID index is assigned inside the “system” interface according to table 7-1.

7.2.1.1 R1 OSPF and SR Configuration

```
configure
    router Base
#--------------------------------------------------
echo "MPLS Label Range Configuration"
#--------------------------------------------------
```

```
        mpls-labels
            sr-labels start 32000 end 39999
        exit
#--------------------------------------------------
echo "OSPFv2 Configuration"
#--------------------------------------------------
        ospf 0
            traffic-engineering
            advertise-router-capability area
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            area 0.0.0.0
                interface "system"
                    node-sid index 1
                    no shutdown
                exit
                interface "toR2"
                    interface-type point-to-point
                    no shutdown
                exit
                interface "toR3"
                    interface-type point-to-point
                    no shutdown
                exit
            exit
            no shutdown
        exit
```

7.2.1.2 R2 OSPF and SR Configuration

```
configure
    router Base
#--------------------------------------------------
echo "MPLS Label Range Configuration"
#--------------------------------------------------
        mpls-labels
            sr-labels start 32000 end 39999
        exit
#--------------------------------------------------
echo "OSPFv2 Configuration"
#--------------------------------------------------
        ospf 0
            traffic-engineering
            advertise-router-capability area
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            area 0.0.0.0
                interface "system"
                    node-sid index 2
                    no shutdown
                exit
                interface "toR1"
                    interface-type point-to-point
                    no shutdown
                exit
                interface "toR4"
                    interface-type point-to-point
                    no shutdown
                exit
                interface "toR3"
                    interface-type point-to-point
                    no shutdown
                exit
            exit
            no shutdown
        exit
    exit
```

7.2.1.3 R3 OSPF and SR Configuration

```
configure
     router Base
#--------------------------------------------------
echo "MPLS Label Range Configuration"
#--------------------------------------------------
        mpls-labels
            sr-labels start 32000 end 39999
        exit
#--------------------------------------------------
echo "OSPFv2 Configuration"
#--------------------------------------------------
        ospf 0
            traffic-engineering
            advertise-router-capability area
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            area 0.0.0.0
                interface "system"
                    node-sid index 3
                    no shutdown
                exit
                interface "toR1"
                    interface-type point-to-point
                    no shutdown
                exit
                interface "toR5"
                    interface-type point-to-point
                    no shutdown
                exit
                interface "toR2"
                    interface-type point-to-point
                    no shutdown
                exit
            exit
            no shutdown
        exit
```

7.2.1.4 R4 OSPF and SR Configuration

```
configure
     router Base
#--------------------------------------------------
echo "MPLS Label Range Configuration"
#--------------------------------------------------
        mpls-labels
            sr-labels start 32000 end 39999
        exit
#--------------------------------------------------
echo "OSPFv2 Configuration"
#--------------------------------------------------
        ospf 0
            traffic-engineering
            advertise-router-capability area
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            area 0.0.0.0
                interface "system"
                    node-sid index 4
                    no shutdown
                exit
                interface "toR2"
                    interface-type point-to-point
                    no shutdown
                exit
                interface "toR5"
                    interface-type point-to-point
                    no shutdown
                exit
                interface "toR6"
                    interface-type point-to-point
                    no shutdown
                exit
            exit
            no shutdown
        exit
```

7.2.1.5 R5 OSPF and SR Configuration

```
configure
    router Base
    #--------------------------------------------------
    echo "MPLS Label Range Configuration"
    #--------------------------------------------------
            mpls-labels
                sr-labels start 32000 end 39999
            exit
    #--------------------------------------------------
    echo "OSPFv2 Configuration"
    #--------------------------------------------------
            ospf 0
                traffic-engineering
                advertise-router-capability area
                segment-routing
                    prefix-sid-range global
                    no shutdown
                exit
                area 0.0.0.0
                    interface "system"
                        node-sid index 5
                        no shutdown
                    exit
                    interface "toR3"
                        interface-type point-to-point
                        no shutdown
                    exit
                    interface "toR4"
                        interface-type point-to-point
                        no shutdown
                    exit
                    interface "toR6"
                        interface-type point-to-point
                        no shutdown
                    exit
                exit
                no shutdown
            exit
```

7.2.1.6 R6 OSPF and SR Configuration

```
configure
     router Base
#--------------------------------------------------
echo "MPLS Label Range Configuration"
#--------------------------------------------------
        mpls-labels
            sr-labels start 32000 end 39999
        exit
#--------------------------------------------------
echo "OSPFv2 Configuration"
#--------------------------------------------------
        ospf 0
            traffic-engineering
            advertise-router-capability area
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            area 0.0.0.0
                interface "system"
                    node-sid index 6
                    no shutdown
                exit
                interface "toR4"
                    interface-type point-to-point
                    no shutdown
                exit
                interface "toR5"
                    interface-type point-to-point
                    no shutdown
                exit
            exit
            no shutdown
        exit
```

7.2.2 SR-OSPF Signaling

Interchange Router Capabilities is a needed step in the neighbour adjacency establishment between two OSPF routers, here each router announces its SR capabilities and other like TE. In the end, each router knows all router capabilities of the rest in the network such as SRGB range and algorithm used.

```
A:sr06# show router ospf capabilities

===============================================================================
Rtr Base OSPFv2 Instance 0 Capabilities
===============================================================================
scope     Router Id           Capabilities
-------------------------------------------------------------------------------
Area      10.0.0.1            OSPF Informational Capabilities (0x38000000):
                                  Stub Router support
                                  TE support
                                  P2P-VLAN
                              SR Algorithm:
                                  Backup-constrained-SPF
                                  IGP-metric-based-SPF
                              SR Label Range: start label 32000 range 8000
Area      10.0.0.2            OSPF Informational Capabilities (0x38000000):
                                  Stub Router support
                                  TE support
                                  P2P-VLAN
                              SR Algorithm:
                                  Backup-constrained-SPF
                                  IGP-metric-based-SPF
                              SR Label Range: start label 32000 range 8000
Area      10.0.0.3            OSPF Informational Capabilities (0x38000000):
                                  Stub Router support
                                  TE support
                                  P2P-VLAN
                              SR Algorithm:
                                  Backup-constrained-SPF
                                  IGP-metric-based-SPF
                              SR Label Range: start label 32000 range 8000
Area      10.0.0.4            OSPF Informational Capabilities (0x38000000):
                                  Stub Router support
                                  TE support
                                  P2P-VLAN
                              SR Algorithm:
                                  Backup-constrained-SPF
                                  IGP-metric-based-SPF
                              SR Label Range: start label 32000 range 8000
Area      10.0.0.5            OSPF Informational Capabilities (0x38000000):
                                  Stub Router support
                                  TE support
                                  P2P-VLAN
                              SR Algorithm:
                                  Backup-constrained-SPF
                                  IGP-metric-based-SPF
                              SR Label Range: start label 32000 range 8000
Area      10.0.0.6            OSPF Informational Capabilities (0x38000000):
                                  Stub Router support
                                  TE support
                                  P2P-VLAN
                              SR Algorithm:
                                  Backup-constrained-SPF
                                  IGP-metric-based-SPF
                              SR Label Range: start label 32000 range 8000
-------------------------------------------------------------------------------
No. of LSAs: 7
===============================================================================
```

Once OSPF has converged, each router assigns a specific label from the SRGB range to each Node-SID using its index, also and Adj-SID is assigned locally to each interface. This is shown in the next capture where a tunnel is created for each Node-SID received.

```
A:sr01# show router tunnel-table

===============================================================================
Flags: B = BGP backup route available
       E = inactive best-external BGP route
       k = RIB-API or Forwarding Policy backup route

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
10.0.0.2/32       ospf (0)  MPLS  524293    10       10.12.0.2      100
10.0.0.4/32       ospf (0)  MPLS  524296    10       10.12.0.2      200
10.0.0.5/32       ospf (0)  MPLS  524297    10       10.12.0.2      300
10.0.0.6/32       ospf (0)  MPLS  524298    10       10.12.0.2      300
10.12.0.2/32      ospf (0)  MPLS  524294    10       10.12.0.2      0
10.13.0.2/32      ospf (0)  MPLS  524299    10       10.13.0.2      0
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
       k = RIB-API or Forwarding Policy backup route
===============================================================================
```

Analyzing in detail router R6’s tunnel entry (*10.0.0.6/32*), the tunnel label *32006* is the expected one because Node-SID Index for router R6 is *6*. OSPF SR propagated this information and that is going to use MPLS as a data plane.

```
A:sr01# show router tunnel-table 10.0.0.6/32 detail

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination      : 10.0.0.6/32
NextHop          : 10.12.0.2
Tunnel Flags     : entropy-label-capable
Age              : 00h03m45s
CBF Classes      : (Not Specified)
Owner            : ospf (0)              Encap            : MPLS
Tunnel ID        : 524298                Preference       : 10
Tunnel Label     : 32006                 Tunnel Metric    : 300
Tunnel MTU       : 8686                  Max Label Stack  : 1
-------------------------------------------------------------------------------
Number of tunnel-table entries          : 1
Number of tunnel-table entries with LFA : 0
===============================================================================
```

A very detailed view of each TLV and sub-TLV used by OSPF to propagate information of SR is shown next. Here in the Router Info the start label (*32000*) and range size (*8000*) of SRGB is advertised, in Extended Prefix TLV the IP prefix (*10.0.0.6*), SID Index (*6*) and flags (*noPHP*) are shown. Finally, two Adj-SID is advertised by router R6 with its respective IP address and local label.

```
A:sr01# show router ospf opaque-database adv-router 10.0.0.6 detail
..
-------------------------------------------------------------------------------
Opaque LSA
-------------------------------------------------------------------------------
Area Id          : 0.0.0.0               Adv Router Id    : 10.0.0.6
Link State Id    : 4.0.0.0               LSA Type         : Area Opaque
Sequence No      : 0x8000001c            Checksum         : 0xa52a
Age              : 1335                  Length           : 52
Options          :  E
Advertisement    : Router Info
    Capabilities (1) Len 4 :
        0x38000000
    SR algorithm (8) Len 2 :
        0x2         0x0
    SR label range (9) Len 12 :
        Range-size=8000
        Sub-TLV SID/label(1) len 3 :
            label=32000
-------------------------------------------------------------------------------
Opaque LSA
-------------------------------------------------------------------------------
Area Id          : 0.0.0.0               Adv Router Id    : 10.0.0.6
Link State Id    : 7.0.0.2               LSA Type         : Area Opaque
Sequence No      : 0x8000001d            Checksum         : 0xde5b
Age              : 1148                  Length           : 44
Options          :  E
Advertisement    : Extended Prefix
    TLV Extended prefix (1) Len 20 :
        rtType=1 pfxLen=32 AF=0 pfx=10.0.0.6
            Flags=Node (0x40)
        Sub-TLV Prefix SID (2) len 8 :
            Flags=noPHP (0x40)
            MT-ID=0 Algorithm=0 SID/Index/Label=6
-------------------------------------------------------------------------------
Opaque LSA
-------------------------------------------------------------------------------
Area Id          : 0.0.0.0               Adv Router Id    : 10.0.0.6
Link State Id    : 8.0.0.3               LSA Type         : Area Opaque
Sequence No      : 0x8000001d            Checksum         : 0xc277
Age              : 1456                  Length           : 48
Options          :  E
Advertisement    : Extended Link
    TLV Extended link (1) Len 24  :
        link Type=P2P (1)  Id=10.0.0.4 Data=10.46.0.2
        Sub-TLV Adj-SID (2) len 7 :
            Flags=Value Local (0x60)
            MT-ID=0 Weight=0 SID/Index/Label=524285
-------------------------------------------------------------------------------
Opaque LSA
-------------------------------------------------------------------------------
Area Id          : 0.0.0.0               Adv Router Id    : 10.0.0.6
Link State Id    : 8.0.0.4               LSA Type         : Area Opaque
Sequence No      : 0x8000001e            Checksum         : 0x48e5
Age              : 1771                  Length           : 48
Options          :  E
Advertisement    : Extended Link
    TLV Extended link (1) Len 24  :
        link Type=P2P (1)  Id=10.0.0.5 Data=10.56.0.2
        Sub-TLV Adj-SID (2) len 7 :
            Flags=Value Local (0x60)
            MT-ID=0 Weight=0 SID/Index/Label=524284
===============================================================================
```

7.3 SR-ISIS

7.3.1 IS-IS and SR Configuration

In the IS-IS case, all routers are inside Area 49.0001, and all network interfaces are going to be point-to-point. The interface “system” must be included inside IS-IS to function as router-ID to form adjacencies. In this example topology is shown in figure 7-2 is used, here routers R1 and R2 are in IS-IS Level 1only and routers R2 to R5 are Level 1/Level 2.

In normal conditions router R1 only receives Prefix-SID from routers R2 and R3, because L2 prefixes are not propagated to L1. In this example prefix R6 and R1 are going to be redistributed to each other, so not only the prefix is advertised but the SID index.

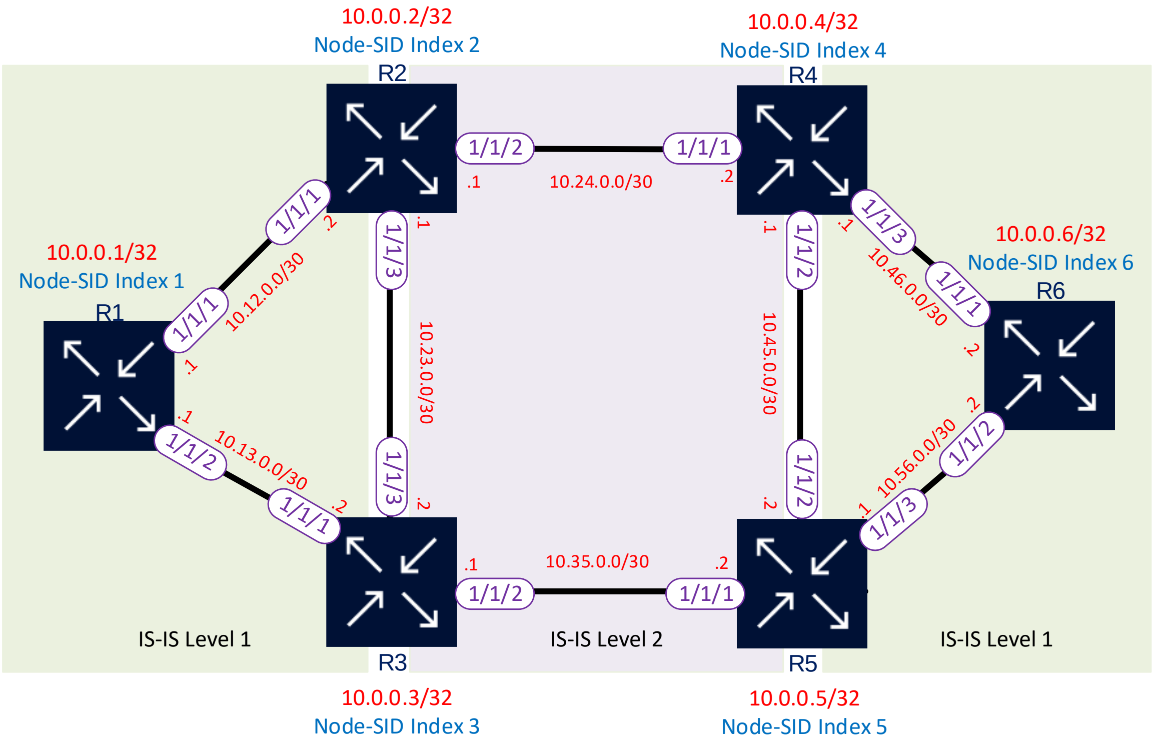


Figure 7-2: SR IS-IS with different levels.

In the Segment Routing part, the same SRGB range (32000 to 39999) must be defined in all routers, then enabled SR using SRGB. Finally, each router sends its SR capabilities in an AS scope, and a Node-SID index is assigned inside the “system” interface according to table 7-1.

7.3.1.1 R1 IS-IS and SR Configuration

```
configure
    router Base
#--------------------------------------------------
echo "MPLS Label Range Configuration"
#--------------------------------------------------
        mpls-labels
            sr-labels start 32000 end 39999
        exit
#--------------------------------------------------
```

```
echo "ISIS Configuration"
#--------------------------------------------------
        isis 0
            level-capability level-1
            area-id 49.0001
            traffic-engineering
            advertise-router-capability as
            level 1
                wide-metrics-only
            exit
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            interface "system"
                ipv4-node-sid index 1
                no shutdown
            exit
            interface "toR2"
                level-capability level-1
                interface-type point-to-point
                no shutdown
            exit
            interface "toR3"
                level-capability level-1
                interface-type point-to-point
                no shutdown
            exit
            no shutdown
        exit
```

7.3.1.2 R2 IS-IS and SR Configuration

```
configure
    router Base
#--------------------------------------------------
echo "MPLS Label Range Configuration"
#--------------------------------------------------
        mpls-labels
            sr-labels start 32000 end 39999
        exit
#--------------------------------------------------
echo "ISIS Configuration"
#--------------------------------------------------
        isis 0
            area-id 49.0001
            export "L2L1"
            traffic-engineering
            advertise-router-capability as
            level 1
                wide-metrics-only
            exit
            level 2
                wide-metrics-only
            exit
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            interface "system"
                ipv4-node-sid index 2
                no shutdown
            exit
            interface "toR1"
                interface-type point-to-point
                no shutdown
            exit
            interface "toR4"
                level-capability level-2
                interface-type point-to-point
                no shutdown
            exit
            interface "toR3"
```

```
                level-capability level-2
                interface-type point-to-point
                no shutdown
            exit
            no shutdown
        exit
#--------------------------------------------------
echo "Policy Configuration"
#--------------------------------------------------
        policy-options
            begin
            prefix-list "R6"
                prefix 10.0.0.6/32 exact
            exit
            policy-statement "L2L1"
                entry 10
                    from
                        prefix-list "R6"
                        level 2
                    exit
                    to
                        level 1
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

7.3.1.3 R3 IS-IS and SR Configuration

```
configure
    router Base
#--------------------------------------------------
echo "MPLS Label Range Configuration"
#--------------------------------------------------
        mpls-labels
            sr-labels start 32000 end 39999
        exit
#--------------------------------------------------
echo "ISIS Configuration"
#--------------------------------------------------
        isis 0
            area-id 49.0001
            export "L2L1"
            traffic-engineering
            advertise-router-capability as
            level 1
                wide-metrics-only
            exit
            level 2
                wide-metrics-only
            exit
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            interface "system"
                ipv4-node-sid index 3
                no shutdown
            exit
            interface "toR1"
                interface-type point-to-point
                no shutdown
            exit
            interface "toR5"
                level-capability level-2
                interface-type point-to-point
                no shutdown
            exit
            interface "toR2"
                level-capability level-2
                interface-type point-to-point
                no shutdown
            exit
            no shutdown
        exit
#--------------------------------------------------
echo "Policy Configuration"
#--------------------------------------------------
        policy-options
            begin
            prefix-list "R6"
                prefix 10.0.0.6/32 exact
            exit
            policy-statement "L2L1"
                entry 10
                    from
                        prefix-list "R6"
                        level 2
                    exit
                    to
                        level 1
                    exit
                    action accept
                    exit
                exit
            exit
            commit
     exit
```

7.3.1.4 R4 IS-IS and SR Configuration

```
configure
    router Base
#--------------------------------------------------
echo "MPLS Label Range Configuration"
#--------------------------------------------------
        mpls-labels
            sr-labels start 32000 end 39999
        exit
#--------------------------------------------------
echo "ISIS Configuration"
#--------------------------------------------------
        isis 0
            area-id 49.0001
            export "L2L1"
            traffic-engineering
            advertise-router-capability as
            level 1
                wide-metrics-only
            exit
            level 2
                wide-metrics-only
            exit
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            interface "system"
                ipv4-node-sid index 4
                no shutdown
            exit
            interface "toR2"
                level-capability level-2
                interface-type point-to-point
                no shutdown
            exit
            interface "toR5"
                level-capability level-2
                interface-type point-to-point
                no shutdown
            exit
            interface "toR6"
                interface-type point-to-point
                no shutdown
            exit
            no shutdown
        exit
#--------------------------------------------------
echo "Policy Configuration"
#--------------------------------------------------
        policy-options
            begin
            prefix-list "R1"
                prefix 10.0.0.1/32 exact
            exit
            policy-statement "L2L1"
                entry 10
                    from
                        prefix-list "R1"
                        level 2
                    exit
                    to
                        level 1
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

7.3.1.5 R5 IS-IS and SR Configuration

```
configure
    router Base
#--------------------------------------------------
echo "MPLS Label Range Configuration"
#--------------------------------------------------
        mpls-labels
            sr-labels start 32000 end 39999
        exit
#--------------------------------------------------
echo "ISIS Configuration"
#--------------------------------------------------
        isis 0
            area-id 49.0001
            export "L2L1"
            traffic-engineering
            advertise-router-capability as
            level 1
                wide-metrics-only
            exit
            level 2
                wide-metrics-only
            exit
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            interface "system"
                ipv4-node-sid index 5
                no shutdown
            exit
            interface "toR3"
                level-capability level-2
                interface-type point-to-point
                no shutdown
            exit
            interface "toR4"
                level-capability level-2
                interface-type point-to-point
                no shutdown
            exit
            interface "toR6"
                interface-type point-to-point
                no shutdown
            exit
            no shutdown
        exit
#--------------------------------------------------
```

```
echo "Policy Configuration"
#--------------------------------------------------
        policy-options
            begin
            prefix-list "R1"
                prefix 10.0.0.1/32 exact
            exit
            policy-statement "L2L1"
                entry 10
                    from
                        prefix-list "R1"
                        level 2
                    exit
                    to
                        level 1
                    exit
                    action accept
                    exit
                exit
            exit
            commit
         exit
```

7.3.1.6 *R6 IS-IS and SR Configuration*

```
configure
    router Base
#--------------------------------------------------
echo "MPLS Label Range Configuration"
#--------------------------------------------------
        mpls-labels
            sr-labels start 32000 end 39999
        exit
#--------------------------------------------------
echo "ISIS Configuration"
#--------------------------------------------------
        isis 0
            level-capability level-1
            area-id 49.0001
            traffic-engineering
            advertise-router-capability as
            level 1
                wide-metrics-only
            exit
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            interface "system"
                ipv4-node-sid index 6
                no shutdown
            exit
            interface "toR4"
                level-capability level-1
                interface-type point-to-point
                no shutdown
            exit
            interface "toR5"
                level-capability level-1
                interface-type point-to-point
                no shutdown
            exit
            no shutdown
         exit
```

7.3.2 SR-ISIS Signaling

Routers R2 and R3 send prefix and Node-SID from router R6 after redistribution to router R1. On the other way, it is the same. Redistributing routers must set the R flag in the Prefix-SID to avoid redistributing an L1 LSP back into L2, and Adj-SID information only has area scope, so it is not visible across levels.

```
A:sr01# show router isis database sr03.00-00 detail

===============================================================================
Rtr Base ISIS Instance 0 Database (detail)
===============================================================================

Displaying Level 1 database
-------------------------------------------------------------------------------
LSP ID    : sr03.00-00                                Level     : L1
[..]

TLVs :
[..]
    Default Metric  : 20
    Control Info: D S, prefLen 32
    Prefix   : 10.0.0.6
    Sub TLV   :
      Prefix-SID Index:6, Algo:0, Flags:RNnP
```

Next capture is showing what is the label assigned to Node-SID for router R6, outbound interface and IP next-hop.

```
A:sr01# show router fp-tunnel-table 1 10.0.0.6/32

===============================================================================
IPv4 Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                  Protocol            Tunnel-ID
  Lbl
    NextHop                                                      Intf/Tunnel
-------------------------------------------------------------------------------
10.0.0.6/32                                  SR-ISIS-0             -
  32006
  10.12.0.2                                                    1/1/1
-------------------------------------------------------------------------------
Total Entries : 1
-------------------------------------------------------------------------------
===============================================================================
```

A very detailed view of each sub-TLV used by IS-IS to propagate information of SR is shown next. Here in the Router Info the start label (*32000*) and range size (*8000*) of SRGB is advertised, router-ID (*10.0.0.4*), SID Index (*4*) and flags (*NnP*) are shown. Finally, Adj-SID is advertised by router R4 with its respective IP address and local label.

```
A:sr06# show router isis database level 1 sr04.00-00 detail

===============================================================================
Rtr Base ISIS Instance 0 Database (detail)
===============================================================================

Displaying Level 1 database
-------------------------------------------------------------------------------
LSP ID    : sr04.00-00                                Level     : L1
Sequence  : 0x13                   Checksum  : 0xff14   Lifetime  : 589
Version   : 1                      Pkt Type  : 18       Pkt Ver   : 1
Attributes: L1L2                   Max Area  : 3        Alloc Len : 161
SYS ID    : 0100.0000.0004         SysID Len : 6        Used Len  : 161

TLVs :
  Area Addresses:
```

```
  Area Address : (3) 49.0001
Supp Protocols:
  Protocols     : IPv4
IS-Hostname     : sr04
Router ID   :
  Router ID   : 10.0.0.4
Router Cap : 10.0.0.4, D:0, S:0
  TE Node Cap : B E M  P
  SR Cap: IPv4 MPLS-IPv6
     SRGB Base:32000, Range:8000
  SR Alg: metric based SPF
I/F Addresses :
  I/F Address   : 10.46.0.1
  I/F Address   : 10.0.0.4
TE IS Nbrs   :
  Nbr   : sr06.00
  Default Metric  : 10
  Sub TLV Len     : 19
  IF Addr   : 10.46.0.1
  Nbr IP    : 10.46.0.2
  Adj-SID: Flags:v4VL Weight:0 Label:524286
TE IP Reach   :
  Default Metric  : 10
  Control Info:    , prefLen 30
  Prefix   : 10.46.0.0
  Default Metric  : 0
  Control Info:   S, prefLen 32
  Prefix   : 10.0.0.4
  Sub TLV   :
    Prefix-SID Index:4, Algo:0, Flags:NnP
  Default Metric  : 20
  Control Info: D S, prefLen 32
  Prefix   : 10.0.0.1
  Sub TLV   :
    Prefix-SID Index:1, Algo:0, Flags:RNnP
```

7.4 Services using SR

7.4.1 L3VPN with SR

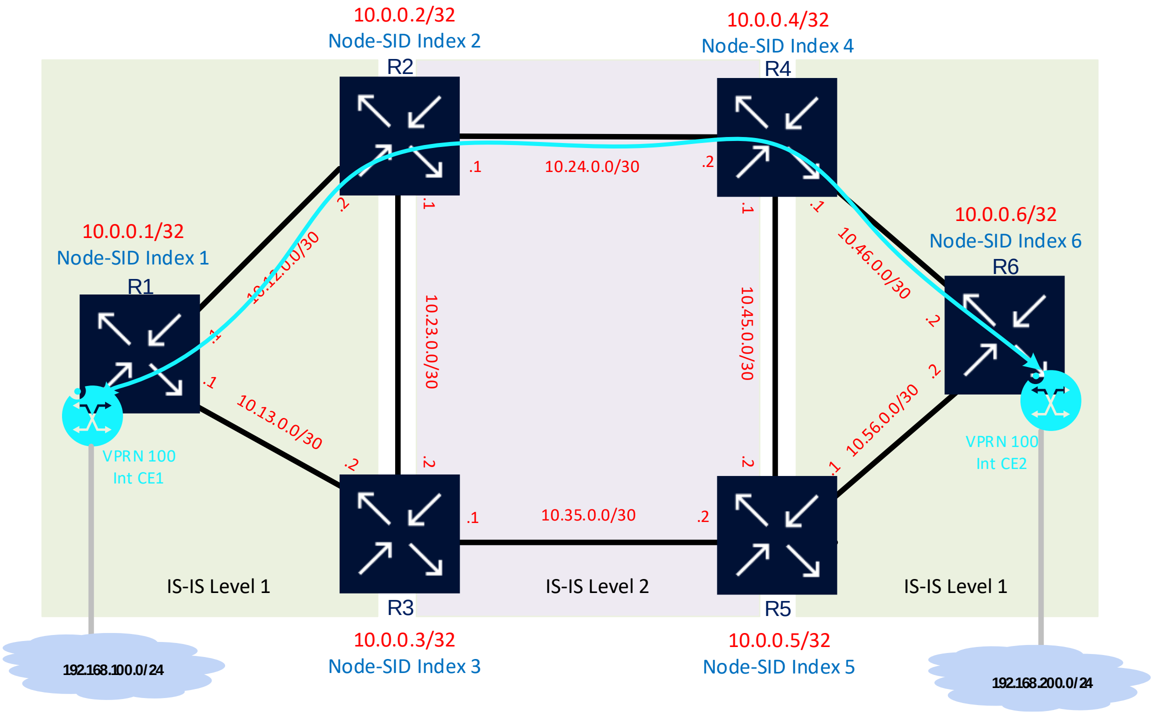


Figure 7-3: L3VPN (VPRN) using SR

A Virtual Private Routing Network (VPRN) will be created in router R1 and R6, using an SR tunnel between both routers as transport. Any L3VPN need also MP-BGP to automatically propagate any VPN-IPv4 prefix to all sites where this service is also created to build the service tunnel.

In this example an iBGP session will be established between routers R1 and R6, both routers using the same Autonomous System (AS), itself router-ID and adding only VPNIPv4 address-family to the session. Then the specific service will be created using all parameters detailed in table 7-2, to use the SR tunnel between R1 and R6 a filter is configured inside the *auto-bind-tunnel* parameter.

Router	Router ID /32	AS	VPRN ID	Route Target	Route Distinguisher	CE IPAddress /24
R1	10.0.0.1	65000	100	65000:100	65000:100	192.168.100.1
R6	10.0.0.6	65000	100	65000:100	65000:100	192.168.200.1

Table 7-2: VPRN (L3VPN) Parameters

7.4.1.1 R1 L3VPN Configuration

```
configure
    router
       autonomous-system 65000
        router-id 10.0.0.1
#--------------------------------------------------
echo "BGP Configuration"
#--------------------------------------------------
        bgp
            min-route-advertisement 5
            enable-peer-tracking
            rapid-withdrawal
            group "INTERNAL"
                family vpn-ipv4
                type internal
                neighbor 10.0.0.6
                exit
            exit
            no shutdown
        exit
    exit
    service
           vprn 100 name "L3VPN-SR" customer 1 create
                route-distinguisher 65000:100
                auto-bind-tunnel
                    resolution-filter
                        sr-isis
                    exit
                    resolution filter
                exit
                vrf-target target:65000:100
                interface "toCE1" create
                    address 192.168.100.1/24
                    loopback
                exit
                no shutdown
            exit
    exit
```

7.4.1.2 R1 L3VPN Configuration

```
configure
    router
```

```
         autonomous-system 65000
        router-id 10.0.0.6
#--------------------------------------------------
echo "BGP Configuration"
#--------------------------------------------------
        bgp
            min-route-advertisement 5
            enable-peer-tracking
            rapid-withdrawal
            group "INTERNAL"
                family vpn-ipv4
                type internal
                neighbor 10.0.0.6
                exit
            exit
            no shutdown
        exit
     exit
    service
         vprn 100 name "L3VPN-SR" customer 1 create
            route-distinguisher 65000:100
            auto-bind-tunnel
                resolution-filter
                    sr-isis
                exit
                resolution filter
            exit
            vrf-target target:65000:100
            interface "toCE2" create
                address 192.168.200.1/24
                loopback
            exit
            no shutdown
        exit
     exit
```

7.4.1.3 L3VPN Results

The first verification step consists of view if the iBGP session is established and how many prefixes are interchanged between both routers. In this example only one prefix will be sent, received and installed in the respective VRF.

```
A:sr01# show router bgp summary
===============================================================================
 BGP Router ID:10.0.0.1          AS:65000       Local AS:65000
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 1           Total Peers                 : 1
Total VPN Peer Groups   : 0           Total VPN Peers             : 0
Total BGP Paths         : 13          Total Path Memory           : 4392

[..]
Total VPN-IPv4 Rem. Rts : 1           Total VPN-IPv4 Rem. Act. Rts: 1
Total VPN-IPv6 Rem. Rts : 0           Total VPN-IPv6 Rem. Act. Rts: 0
Total VPN-IPv4 Bkup Rts : 0           Total VPN-IPv6 Bkup Rts     : 0
Total VPN Local Rts     : 2           Total VPN Supp. Rts         : 0
Total VPN Hist. Rts     : 0           Total VPN Decay Rts         : 0

[..]

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
10.0.0.6
               65000     1878    0 15h45m39s 1/1/1 (VpnIPv4)
                         1876    0
```

```
-------------------------------------------------------------------------------
```

Next capture show routing table inside VPRN 100 with a local route and one route coming from router R6, there is specified SR-ISIS as the protocol to build the specific transport tunnel.

```
A:sr01# show router 100 route-table

===============================================================================
Route Table (Service: 100)
===============================================================================
Dest Prefix[Flags]                              Type    Proto     Age         Pref
      Next Hop[Interface Name]                                      Metric
-------------------------------------------------------------------------------
192.168.100.0/24                                Local   Local     15h48m26s   0
       toCE1                                                        0
192.168.200.0/24                                Remote  BGP VPN   15h48m22s   170
       10.0.0.6 (tunneled:SR-ISIS:0)                                 30
-------------------------------------------------------------------------------
No. of Routes: 2
```

More detail about the route learned from router R6, could be seen in the next capture. Route Distinguisher and Route Target are highlighted in red, also the VPN Label used to build the service tunnel.

```
A:sr01# show router bgp routes 192.168.200.0/24 vpn-ipv4 hunt
===============================================================================
 BGP Router ID:10.0.0.1         AS:65000       Local AS:65000
===============================================================================
 [..]
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network         : 192.168.200.0/24
Nexthop         : 10.0.0.6
Route Dist.     : 65000:100              VPN Label      : 524285
Path Id         : None
From            : 10.0.0.6
Res. Nexthop    : n/a
Local Pref.     : 100                    Interface Name : toR2
Aggregator AS   : None                   Aggregator     : None
Atomic Aggr.    : Not Atomic             MED            : None
AIGP Metric     : None
Connector       : None
Community       : target:65000:100
Cluster         : No Cluster Members
Originator Id   : None                   Peer Router Id : 10.0.0.6
Fwd Class       : None                   Priority       : None
Flags           : Used  Valid  Best  IGP
Route Source    : Internal
AS-Path         : No As-Path
Route Tag       : 0
Neighbor-AS     : n/a
Orig Validation: N/A
Source Class    : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified   : 15h47m20s
VPRN Imported   :  100

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
```

A traceroute from router R1 and using the SR-ISIS tunnel to router R6 is shown in the next capture. Here is easy to see that the path followed by the transport tunnel going through routers R1-R3-R4 and R6.

```
A:sr01# oam lsp-trace sr-isis prefix 10.0.0.6/32
lsp-trace to 10.0.0.6/32: 0 hops min, 0 hops max, 104 byte packets
1  10.0.0.2  rtt=2.32ms rc=8(DSRtrMatchLabel) rsc=1
2  10.0.0.4  rtt=6.26ms rc=8(DSRtrMatchLabel) rsc=1
3  10.0.0.6  rtt=19.1ms rc=3(EgressRtr) rsc=1
```

In order to observe how is the packet walkthrough along the path described above, next captures follow inbound labels, outbound labels and outbound interface at each hop. As expected, label *32006* is swapped with the same label along the network (CONTINUE operation) until get router R6 where label is popped (NEXT operation).

```
##################################R1#####################################################

A:sr01# tools dump router segment-routing tunnel in-label 32006
===================================================================================================
Legend: (B) - Backup Next-hop for Fast Re-Route
        (D) - Duplicate
label stack is ordered from top-most to bottom-most
===================================================================================================
--------------------------------------------------------------------------------------------------+
 Prefix                                                                                           |
 Sid-Type        Fwd-Type       In-Label  Prot-Inst                                               |
                 Next Hop(s)                                     Out-Label(s) Interface/Tunnel-ID |
--------------------------------------------------------------------------------------------------+
 10.0.0.6
 Node            Orig/Transit   32006     ISIS-0
                 10.12.0.2                                       32006       toR2
--------------------------------------------------------------------------------------------------+
No. of Entries: 1
--------------------------------------------------------------------------------------------------+

##################################R2#####################################################

A:sr02# tools dump router segment-routing tunnel in-label 32006
===================================================================================================
Legend: (B) - Backup Next-hop for Fast Re-Route
        (D) - Duplicate
label stack is ordered from top-most to bottom-most
===================================================================================================
--------------------------------------------------------------------------------------------------+
 Prefix                                                                                           |
 Sid-Type        Fwd-Type       In-Label  Prot-Inst                                               |
                 Next Hop(s)                                     Out-Label(s) Interface/Tunnel-ID |
--------------------------------------------------------------------------------------------------+
 10.0.0.6
 Node            Orig/Transit   32006     ISIS-0
                 10.24.0.2                                       32006       toR4
--------------------------------------------------------------------------------------------------+
No. of Entries: 1
--------------------------------------------------------------------------------------------------+

##################################R4#####################################################

A:sr04# tools dump router segment-routing tunnel in-label 32006
===================================================================================================
Legend: (B) - Backup Next-hop for Fast Re-Route
        (D) - Duplicate
label stack is ordered from top-most to bottom-most
===================================================================================================
--------------------------------------------------------------------------------------------------+
 Prefix                                                                                           |
 Sid-Type        Fwd-Type       In-Label  Prot-Inst                                               |
                 Next Hop(s)                                     Out-Label(s) Interface/Tunnel-ID |
--------------------------------------------------------------------------------------------------+
 10.0.0.6
 Node            Orig/Transit   32006     ISIS-0
                 10.46.0.2                                       32006       toR6
--------------------------------------------------------------------------------------------------+
No. of Entries: 1
--------------------------------------------------------------------------------------------------+
#############################R6###################################################

A:sr06# tools dump router segment-routing tunnel in-label 32006
===============================================================================================
Legend: (B) - Backup Next-hop for Fast Re-Route
        (D) - Duplicate
label stack is ordered from top-most to bottom-most
===============================================================================================
----------------------------------------------------------------------------------------------+
 Prefix                                                                                       |
 Sid-Type           Fwd-Type        In-Label   Prot-Inst                                      |
                    Next Hop(s)                                      Out-Label(s) Interface/Tunnel-ID |
----------------------------------------------------------------------------------------------+
 10.0.0.6
 Node               Terminating     32006      ISIS-0
----------------------------------------------------------------------------------------------+
No. of Entries: 1
----------------------------------------------------------------------------------------------+
```

As both interfaces inside VPRN *100* are loopbacks, they will be always on and the reachability is guarantee. Next ping from both end point to the other.

```
#############################R1###################################################
A:sr01# ping router 100 192.168.200.1
PING 192.168.200.1 56 data bytes
64 bytes from 192.168.200.1: icmp_seq=1 ttl=64 time=5.05ms.
64 bytes from 192.168.200.1: icmp_seq=2 ttl=64 time=31.9ms.
64 bytes from 192.168.200.1: icmp_seq=3 ttl=64 time=96.9ms.
64 bytes from 192.168.200.1: icmp_seq=4 ttl=64 time=4.49ms.
64 bytes from 192.168.200.1: icmp_seq=5 ttl=64 time=12.9ms.

---- 192.168.200.1 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 4.49ms, avg = 30.2ms, max = 96.9ms, stddev = 34.8ms

#############################R6###################################################
A:sr06# ping router 100 192.168.100.1
PING 192.168.100.1 56 data bytes
64 bytes from 192.168.100.1: icmp_seq=1 ttl=64 time=12.4ms.
64 bytes from 192.168.100.1: icmp_seq=2 ttl=64 time=50.0ms.
64 bytes from 192.168.100.1: icmp_seq=3 ttl=64 time=4.94ms.
64 bytes from 192.168.100.1: icmp_seq=4 ttl=64 time=25.5ms.
64 bytes from 192.168.100.1: icmp_seq=5 ttl=64 time=7.10ms.

---- 192.168.100.1 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 4.94ms, avg = 20.0ms, max = 50.0ms, stddev = 16.6ms
```

7.5 TI-LFA

In order to protect traffic when a link becomes unavailable TI-LFA is enabled in all routers within the SR domain, as already said TI-LFA needs SR was already enabled to work properly. For the sake of simplicity, the topology used is all nodes inside IS-IS Level 2 because without the information of Adj-SID is not possible by default to have a backup path using TI-LFA.

7.5.1 TI-LFA to protect Services

In the next example an additional site is added to VPRN 100 in router R4, then a failure will be simulated in the R2-R4 link to force to TI-LFA imposes an FRR label to reach router R4 from router R1. BGP and VPRN configuration needed to add the new site are not shown because they are the same preciously shown at point 7.4.

Figure 7-4 shows the expected path through router R1-R2-R3-R5-R4 when failure appears in the R2-R4 link.

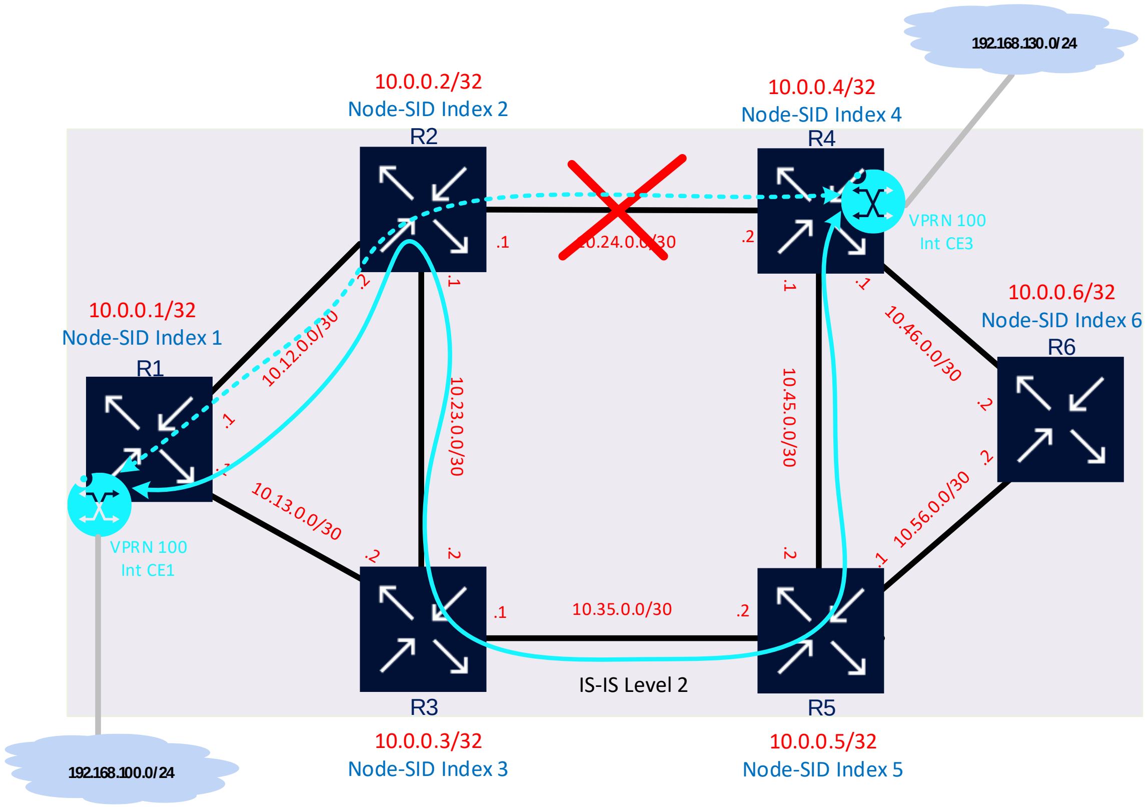


Figure 7-4: Protecting path using TI-LFA

7.5.1.1 R1 to R6 TI-LFA Configuration

After SR is operating in a network enabling TI-LFA in a router is made with only one command, which is shown next.

```
configure
     router
        isis
            loopfree-alternates
                ti-lfa
                exit
            exit
        exit
     exit
```

7.5.1.2 Traffic in normal conditions

Next capture show router R1 has a main path headed to router R2 and a backup path through router R3, both paths using the same label to the destination router. Before TI-LFA was enabled only one path and label reside in this table (see point 7.3.2), here captures from router R6 are shown along router R4 to compare the results after and before activated TI-LFA.

```
A:sr01# show router fp-tunnel-table 1 10.0.0.6/32

================================================================================
IPv4 Tunnel Table Display
```

```
Legend:
B - FRR Backup
===============================================================================
Destination                                  Protocol            Tunnel-ID
  Lbl
    NextHop                                                      Intf/Tunnel
-------------------------------------------------------------------------------
10.0.0.6/32                                  SR-ISIS-0             -
  32006
  10.12.0.2                                                    1/1/1
  32006
  10.13.0.2(B)                                                 1/1/2
-------------------------------------------------------------------------------
Total Entries : 1
-------------------------------------------------------------------------------
===============================================================================
A:sr01# show router fp-tunnel-table 1 10.0.0.4/32

===============================================================================
IPv4 Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                  Protocol            Tunnel-ID
  Lbl
    NextHop                                                      Intf/Tunnel
-------------------------------------------------------------------------------
10.0.0.4/32                                  SR-ISIS-0             -
  32004
  10.12.0.2                                                    1/1/1
  32004
  10.13.0.2(B)                                                 1/1/2
-------------------------------------------------------------------------------
Total Entries : 1
-------------------------------------------------------------------------------
===============================================================================
```

Because the failure is going to be in the R2-R4 link, the PLR is router R2, so here TI-LFA adds an FRR label to router R5 in the backup path. So, when a failure happens in R2-R4 traffic will be headed by router R2 to router R5 using its Node-SID *32005* and then router R5 will forward traffic to router R4 using it's Node-SID *32004*.

```
A:sr02>config>router>isis>if# show router fp-tunnel-table 1 10.0.0.4/32

===============================================================================
IPv4 Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                  Protocol            Tunnel-ID
  Lbl
    NextHop                                                      Intf/Tunnel
-------------------------------------------------------------------------------
10.0.0.4/32                                  SR-ISIS-0             -
  32004
  10.24.0.2                                                    1/1/2
  32004/32005
  10.23.0.2(B)                                                 1/1/3
-------------------------------------------------------------------------------
Total Entries : 1
-------------------------------------------------------------------------------
```

LFA coverture expected is 100% could be verified in next screen.

```
*A:sr01>config>router>bgp>group# show router isis sr-lfa-coverage

===============================================================================
Rtr Base ISIS Instance 0 SR LFA Coverage
===============================================================================
MT-ID   SidType      Level Proto LFA       RLFA      TILFA     Coverage
-------------------------------------------------------------------------------
0       node-sid     L2    ipv4  0(0%)     0(0%)     5(100%)   5/5(100%)
0       adj-sid      L2    ipv4  0(0%)     0(0%)     2(100%)   2/2(100%)
===============================================================================
*A:sr01>config>router>bgp>group# show router isis lfa-coverage

===============================================================================
Rtr Base ISIS Instance 0 LFA Coverage
===============================================================================
Topology          Level   Node           IPv4                IPv6
-------------------------------------------------------------------------------
IPV4 Unicast      L1      0/0(0%)        11/11(100%)         0/0(0%)
IPV6 Unicast      L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast    L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast    L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Unicast      L2      5/5(100%)      11/11(100%)         0/0(0%)
IPV6 Unicast      L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast    L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast    L2      0/0(0%)        0/0(0%)             0/0(0%)
===============================================================================
```

VPRN site in router R4 is reachable from router R1 using ping, and this use path through R1-R2-R4 in normal conditions.

```
*A:sr01>config>router>bgp>group# ping router 100 192.168.230.1
PING 192.168.230.1 56 data bytes
64 bytes from 192.168.230.1: icmp_seq=1 ttl=64 time=101ms.
64 bytes from 192.168.230.1: icmp_seq=2 ttl=64 time=3.09ms.
64 bytes from 192.168.230.1: icmp_seq=3 ttl=64 time=87.3ms.
64 bytes from 192.168.230.1: icmp_seq=4 ttl=64 time=2.85ms.
64 bytes from 192.168.230.1: icmp_seq=5 ttl=64 time=35.4ms.

---- 192.168.230.1 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 2.85ms, avg = 46.0ms, max = 101ms, stddev = 41.4ms

*A:sr01>config>router>bgp>group# oam lsp-trace sr-isis prefix 10.0.0.4/32
lsp-trace to 10.0.0.4/32: 0 hops min, 0 hops max, 104 byte packets
1  10.0.0.2  rtt=19.7ms rc=8(DSRtrMatchLabel) rsc=1
2  10.0.0.4  rtt=8.75ms rc=3(EgressRtr) rsc=1
```

7.5.1.3 *Traffic after failure*

A failure is simulated in router R4 port to R2, so this router switches the traffic from router R1 to R4 to its alternate path through R3-R5-R4. This new path follows hop by hop the behaviour described previously, so in this case TI-LFA only needed one FRR label to protect traffic in router R2.

```
A:sr01>config>router>bgp>group# oam lsp-trace sr-isis prefix 10.0.0.4/32
lsp-trace to 10.0.0.4/32: 0 hops min, 0 hops max, 104 byte packets
1  10.0.0.3  rtt=2.15ms rc=8(DSRtrMatchLabel) rsc=1
2  10.0.0.5  rtt=3.38ms rc=8(DSRtrMatchLabel) rsc=1
3  10.0.0.4  rtt=16.4ms rc=3(EgressRtr) rsc=1
```

The routing table inside VPRN 100 continues been the same from the router R1 point of view.

```
A:sr01>config>router>bgp>group# show router 100 route-table

===============================================================================
Route Table (Service: 100)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                   Metric
-------------------------------------------------------------------------------
192.168.100.0/24                              Local   Local     19h54m12s  0
       toCE1                                                      0
192.168.200.0/24                              Remote  BGP VPN   00h00m12s  170
       10.0.0.6 (tunneled:SR-ISIS:0)                               30
192.168.230.0/24                              Remote  BGP VPN   00h00m12s  170
       10.0.0.4 (tunneled:SR-ISIS:0)                               20
-------------------------------------------------------------------------------
No. of Routes: 3
```

Also, only one packet was lost when the LFA protection was active. This is not a very accurate way to measure transition time but gives a good idea of how rapid could be a transition from the main path to backup.

```
A:sr01>config>router>bgp>group# ping router 100 192.168.230.1 count 1000
PING 192.168.230.1 56 data bytes
64 bytes from 192.168.230.1: icmp_seq=1 ttl=64 time=59.5ms.
64 bytes from 192.168.230.1: icmp_seq=2 ttl=64 time=3.67ms.
64 bytes from 192.168.230.1: icmp_seq=3 ttl=64 time=4.80ms.
64 bytes from 192.168.230.1: icmp_seq=4 ttl=64 time=76.9ms.
64 bytes from 192.168.230.1: icmp_seq=5 ttl=64 time=22.2ms.
64 bytes from 192.168.230.1: icmp_seq=6 ttl=64 time=11.6ms.
64 bytes from 192.168.230.1: icmp_seq=7 ttl=64 time=8.40ms.
64 bytes from 192.168.230.1: icmp_seq=8 ttl=64 time=5.35ms.
64 bytes from 192.168.230.1: icmp_seq=9 ttl=64 time=3.62ms.
64 bytes from 192.168.230.1: icmp_seq=10 ttl=64 time=3.41ms.
64 bytes from 192.168.230.1: icmp_seq=11 ttl=64 time=4.04ms.
64 bytes from 192.168.230.1: icmp_seq=12 ttl=64 time=29.5ms.
64 bytes from 192.168.230.1: icmp_seq=14 ttl=64 time=146ms.
64 bytes from 192.168.230.1: icmp_seq=15 ttl=64 time=51.3ms.
64 bytes from 192.168.230.1: icmp_seq=16 ttl=64 time=4.51ms.
64 bytes from 192.168.230.1: icmp_seq=17 ttl=64 time=7.52ms.
64 bytes from 192.168.230.1: icmp_seq=18 ttl=64 time=4.38ms.
Request timed out. icmp_seq=13.
64 bytes from 192.168.230.1: icmp_seq=19 ttl=64 time=95.9ms.
64 bytes from 192.168.230.1: icmp_seq=20 ttl=64 time=12.6ms.
64 bytes from 192.168.230.1: icmp_seq=21 ttl=64 time=68.4ms.
64 bytes from 192.168.230.1: icmp_seq=22 ttl=64 time=4.96ms.
64 bytes from 192.168.230.1: icmp_seq=23 ttl=64 time=4.94ms.
[..]
64 bytes from 192.168.230.1: icmp_seq=49 ttl=64 time=25.9ms.
^C
ping aborted by user

---- 192.168.230.1 PING Statistics ----
49 packets transmitted, 48 packets received, 2.04% packet loss
round-trip min = 3.41ms, avg = 28.3ms, max = 146ms, stddev = 0.000ms
```

8 Conclusions:

As has been demonstrated in previous sections, Segment Routing simplifies enormously the MPLS control plane using IS-IS or OSPF as the sole signaling protocol, it also gives a powerful and simple FRR mechanism using TI-LFA. SR also permits using an external server to compute the short paths giving a specific constrain, that headed network automation from a central point.

MP-BGP completes the SR signaling in multi-area or multi-instance environments using BGP-LU to interconnect different network sections even with traditional MPLS signaling protocols. Also, in Data Centre environments, a BGP only signaling protocol is used in Web-Scale implementations.

Almost all SR deployments use MPLS Data-Plane because it is a proven and well mature technology, so SR implementation is simpler because an IGP protocol is always implemented in a Service Provider network and MPLS is also used to transport services.

On the other hand, using IPv6 as Data-Plane is a future bet because currently, this technology is being developed by only a couple of vendors and brownfield networks would be difficult to migrate to an IPv6 only schema.

To choose what IGP protocol use as Control-Plane in SR, most Service Providers (SPs) will use the current IGP used in their networks, so if an SP uses OSPF as IGP protocol the logical step is to upgrade this protocol to support SR. However, if this SP is going to implement IPv6 in his Global Routing schema OSPF is not a valid choice because this protocol does not support IPv6. OSPFv3 is not a good option as well since it is not supported by any vendor to run SR. The only valid option in an SP with IPv4 and IPv6 addressing is IS-IS, SR over IS-IS is a priority for all vendors because its inherent capabilities to add new extensions support using TLVs and sub-TLVs.

To sum up, ISIS-SR with BGP-SR are the protocols best suited as Control-Plane in most network scenarios and MPLS Data-Plane will continue been used to forward packets in SR networks.

9 References:

[1] Alcatel-Lucent, "Segment Routing - Deep Dive," Alcatel-Lucent, 2015.

[2] I. e. Cisco Systems, "Segment Routing," Cisco Systems, Inc. employees., 2019. [Online]. Available: http://www.segment-routing.net/.

[3] C. Filsfils, K. Michielsen and K. Talaulikar, "SR MPLS Data Plane," in *Segment Routing, Part I*, Cysco Systems, Inc, 2016, p. 22.

[4] Nokia, "Nokia Multiprotocol Label Switching Student Guide v.2.3.1," Nokia Service Routing Certification Program, 2019.

[5] A. A. L. Andersson, "RFC 5036 - LDP Specification," IETF, October 2006. [Online]. Available: https://tools.ietf.org/html/rfc5036.

[6] P. Pan, G. Swallow and A. Atlas, "RFC 4090 - Fast Reroute Extensions to RSVP-TE for LSP Tunnels," IETF, May 2005. [Online]. Available: https://tools.ietf.org/html/rfc4090.

[7] D. Katz, K. Kompella and D. Yeung, "RFC 3630 - Traffic Engineering (TE) Extensions to OSPF Version 2," IETF, September 2003. [Online]. Available: https://tools.ietf.org/html/rfc3630.

[8] T. Li and H. Smit, "RFC 5305 - IS-IS Extensions for Traffic Engineering," IETF, October 2008. [Online]. Available: https://tools.ietf.org/html/rfc5305.

[9] D. Awduche, L. Berger and D. Gan, "RFC 3209 - RSVP-TE: Extensions to RSVP for LSP Tunnels," IETF, December 2001. [Online]. Available: https://tools.ietf.org/html/rfc3209.

[10] J. S. Esa Metsala, Mobile Backhaul, Wiley, 2012.

[11] Nokia, "Topology Independent-Loopfre Alternate R15 v0.3," Nokia, 2017.

[12] A. Atlas and A. Zinin, "RFC 5286 - Basic Specification for IP Fast Reroute: Loop-Free Alternates," IETF, September 2008. [Online]. Available: https://tools.ietf.org/html/rfc5286.

[13] S. Bryant, C. Filsfils and S. Previdi, "RFC 7490 - Remote Loop-Free Alternate (LFA) Fast Reroute (FRR)," IETF, April 2015. [Online]. Available: https://tools.ietf.org/html/rfc7490.

[14] C. Filsfils, K. Michielsen and K. Talaulikar, "TI-LFA," in *Segment Routing, Part I*, Cisco Press, 2016, p. 383.

[15] Alcatel-Lucent, "Seamless MPLS," ALU IPD, Prague, 2013.

[16] Y. Rekhter and E. Rosen, "RFC 3107 - Carrying Label Information in BGP-4," IETF, May 2001. [Online]. Available: https://tools.ietf.org/html/rfc3107.

[17] Alcatel-Lucent, "Seamless MPLS & Segment Routing," SReXperts Alcatel-Lucent, Tokyo, 2015.

[18] C. Filsfils, K. Michielsen and K. Talaulikar, "Introduction," in *Segment Routing, Part I*, Cisco Press, 2016, pp. 9-15.

[19] G. Maila, I. Marius and C. Victor, "Segment Routing," in *THE 10th INTERNATIONAL SYMPOSIUM ON ADVANCED TOPICS IN ELECTRICAL ENGINEERING*, Bucharest, Romania, 2017.

[20] S. P. L. G. C. Filsfils, "RFC 8402: Segment Routing Architecture," *IETF,* 2018.

[21] Nokia, "Segment Routing PCC and PCE initiated LSPs," in *SReXperts*, Anaheim, 2019.

[22] C. Filsfils, K. Michielsen and K. Talaulikar, "Fundamentals of Segment Routing," in *Segment Routing, Part I*, Cisco Press, 2016, pp. 27-50.

[23] Cisco, "Using SSPF with Segment Routing," Cisco, 16 January 2019. [Online]. Available: https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/seg_routing/configuration/xe-16-9/segrt-xe-16-9-book/sr-sspf.html.

[24] Z. Leyes, "How Anycast Works to Bring Content Closer to Your Visitors," Imperva, 21 February 2017. [Online]. Available: https://www.imperva.com/blog/how-anycast-works/.

[25] Nokia, "INTERFACE CONFIGURATION GUIDE - RELEASE 19.10.R1," Nokia, 2019.

[26] Microsoft, "BGP is a better IGP," Microsoft, 2012. [Online]. Available: https://archive.nanog.org/meetings/nanog55/presentations/Monday/Lapukhov.pdf.

[27] H. Gredler, J. Medved and S. Previdi, "RFC 7752 - North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP," IETF, March 2016. [Online]. Available: https://tools.ietf.org/html/rfc7752.

[28] Huawei, "BGP for SR - Segment Routing 01," Huawei, Jul 2019. [Online]. Available: https://support.huawei.com/enterprise/en/doc/EDOC1100093196/5b56c2ff/bgp-for-sr.

[29] EANTC, "Multi-Vendor Interoperability Test," in *MPLS+SDN+NFV World Congress*, Paris, 2019.

[30] EANTC, "Interoperability Showcase," in *MPLS+SDN+NFV World Congress*, Paris, 2018.

[31] Nokia, "GSS update Segment Routing & NFIX," in *SReXperts*, Athens, 2019.

[32] E. Rosen, A. Viswanathan and R. Callon, "RFC 3031 - Multiprotocol Label Switching Architecture," IETF, January 2001. [Online]. Available: https://tools.ietf.org/html/rfc3031.

[33] Nokia, "Segment Routing Overview and Applicability v2.1," in *PLM Summit*, Murray Hill, 2019.

[34] C. Filsfils, K. Michielsen and K. Talaulikar, "Management of the SRGB," in *Segment Routing, Part I*, Cisco Press, 2016, pp. 85-93.

[35] Nokia, "Nokia Quality of Service StudentGuide v4.1," Nokia Service Routing Certification Program, Plano-Texas, 2018.

[36] C. Filsfils, K. Michielsen and K. Talaulikar, "Segment Routing IGP Control Plane," in *Segment Routing, Part I*, Cisco Press, 2016, pp. 105-226.

[37] P. Psenak, S. Previdi, C. Filsfils, H. Gredler and W. Henderickx, "RFC 8685 - OSPF Extensions for Segment Routing," IETF, December 2019. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8665.txt.

[38] P. Psenak and S. Previdi, "RFC 8666 - OSPFv3 Extensions for Segment Routing," IETF, December 2019. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8666.txt.

[39] S. Previdi, L. Ginsberg, C. Filsfils, A. Bashandy and H. Gredler, "RFC 8667 - IS-IS Extensions for Segment Routing," IETF, December 2019. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8667.txt.

[40] Nokia, "Nokia Interior Routing Protocols," Nokia Service Routing Certification Program, 2019.

[41] L. Ginsberg, S. Previdi and M. Chen, "RFC 7981- IS-IS Extensions for Advertising Router Information," IETF, October 2016. [Online]. Available: https://tools.ietf.org/html/rfc7981.

[42] A. Retana, R. White, V. Fuller and D. McPherson, "RFC 3021 - Using 31-Bit Prefixes on IPv4 Point-to-Point Links," IETF, December 2000. [Online]. Available: https://tools.ietf.org/html/rfc3021.

[43] P. Psenak, H. Gredler, R. Shakir, W. Henderickx, J. Tantsura and A. Lindem, "RFC 7684 - OSPFv2 Prefix/Link Attribute Advertisement," IETF, November 2015. [Online]. Available: https://tools.ietf.org/html/rfc7684.

[44] A. Lindem, N. Shen, J. Vasseur, R. Aggarwal and S. Shaffer, "RFC 7770 - Extensions to OSPF for Advertising Optional Router Capabilities," IETF, February 2016. [Online]. Available: https://tools.ietf.org/html/rfc7770.

[45] Y. Rekhter, T. Li and S. Hares, "RFC 4271 - A Border Gateway Protocol 4 (BGP-4)," IETF, January 2006. [Online]. Available: https://tools.ietf.org/html/rfc4271.

[46] T. Bates, R. Chandra, D. Katz and Y. Rekhter, "RFC 4760 - Multiprotocol Extensions for BGP-4," IETF, January 2007. [Online]. Available: https://tools.ietf.org/html/rfc4760.

[47] C. Bookham, "Chapter1: Getting Started," in *Versatile Routing and Services with BGP*, Indianapolis, John Wiley & Sons, 2014.

[48] IANA, "Address Family Numbers," IANA, 04 11 2019. [Online]. Available: https://www.iana.org/assignments/address-family-numbers/address-family-numbers.xhtml.

[49] IANA, "Subsequent Address Family Identifiers (SAFI) Parameters," IANA, 05 March 2019. [Online]. Available: https://www.iana.org/assignments/safi-namespace/safi-namespace.xhtml.

[50] N. Leyman, B. Decraene, C. Filsfils, M. Konstantynowicz and D. Steinberg, "Seamless MPLS Architecture - draft-ietf-mpls-seamless-mpls-07," IETF, 28 June 2014. [Online]. Available: https://tools.ietf.org/html/draft-ietf-mpls-seamless-mpls-07.

[51] A. Bashandy, C. Filsfils and P. Mohapatra, "BGP Prefix Independent Convergence - draft-ietf-rtgwg-bgp-pic-11," IETF, 10 February 2020. [Online]. Available: https://www.ietf.org/id/draft-ietf-rtgwg-bgp-pic-11.txt.

[52] Nokia, "Cloud centric evolution of Unicast Routing," in *SReXperts*, San Diego, 2019.

[53] C. Filsfils, K. Michielsen and K. Talaulikar, "Segment Routing BGP Control Plane," in *Segment Routing, Part I*, Cisco Press, 2016, pp. 229-276.

[54] S. Previdi, C. Filsfils, A. Lindem, A. Sreekantiah and H. Gredler, "RFC 8669 - Segment Routing Prefix Segment Identifier Extensions for BGP," IETF, December 2019. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8669.txt.

[55] Nuage Networks, "WBX running Nuage SROS," in *SReXperts*, Anaheim, 2017.

[56] D. Katz and D. Ward, "RFC 5881 Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)," IETF, June 2010. [Online]. Available: https://tools.ietf.org/html/rfc5881.

[57] Nokia, "Remote LFA Node Protection," in *PLM Summit*, Anaheim, 2019.

[58] S. Litkowski and B. Decraene, "Topology independant LFA - Orange use case & applicability," in *MPLS SDN congress*, Paris, 2014.

[59] D. Awduche, J. Malcolm and J. Agogbua, "RFC 2702 - Requirements for Traffic Engineering Over MPLS," IETF, Sep 1999. [Online]. Available: https://tools.ietf.org/html/rfc2702.

[60] C. Filsfils, K. Michielsen and K. Talaulikar, "Segment Routing, Part I," in *Segment Routing BGP Control Plane*, Cisco Press, 2016, pp. 229-290.

[61] S. Litkowski, A. Bashandy, C. Filsfils, B. Decraene, P. Francois and D. Voyer, " Topology Independent Fast Reroute using Segment Routing - draft-ietf-rtgwg-segment-routing-ti-lfa-02," IETF, 18 January 2020. [Online]. Available: https://www.ietf.org/id/draft-ietf-rtgwg-segment-routing-ti-lfa-02.txt.