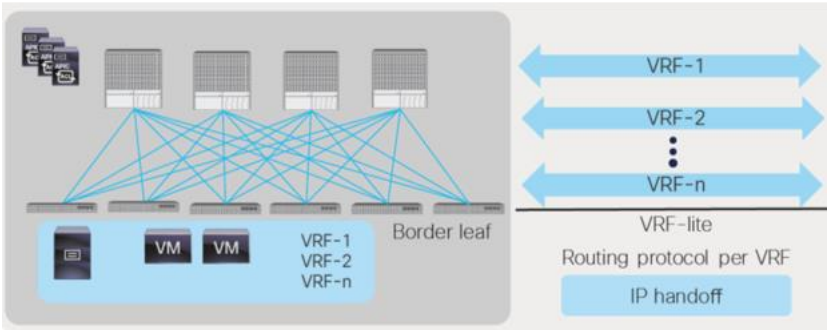
ACI SR-MPLS Handoff
(the L3Out for Segment Routing)

At the beginning there was the "L3Out", used to provide IP routing connectivity, between internal ACI Bridge Domains subnets and the external world (because Data Centers in the end must talk with someone outside there ☺), enabled at Tenant level associated to each single VRF.
You can find more information on the previous article https://bit.ly/3sTHo6Z.

As you can see on the figure below, this approach that uses normal *IP Handoff*, doesn't scale because per each single VPN, a physical or logical interface and a routing protocol session is required between the border leaf and the Provider Edge router using the *VRF-lite* paradigm (or, if you prefer, talking the same language of ISPs, a kind of Inter AS Option A must be implemented).



GOLF scenario has been introduced then to overcome the previous limits (look here https://bit.ly/2VcEo6y for more details), relying on a single BGP EVPN session between Spines and an external WAN router that performs the interworking function between EVPNoVXLAN and L3VPNoMPLS environments, able to transport all the prefixes and relative attributes necessary to identify the single VRFs.
GOLF brings advantages in terms of scaling, however, it requires still VXLAN to be enabled as overlay layer on the external WAN router (fabric side); some ISP platforms could not support VXLAN, so it was necessary to move a step forward and CISCO did it again… ☺

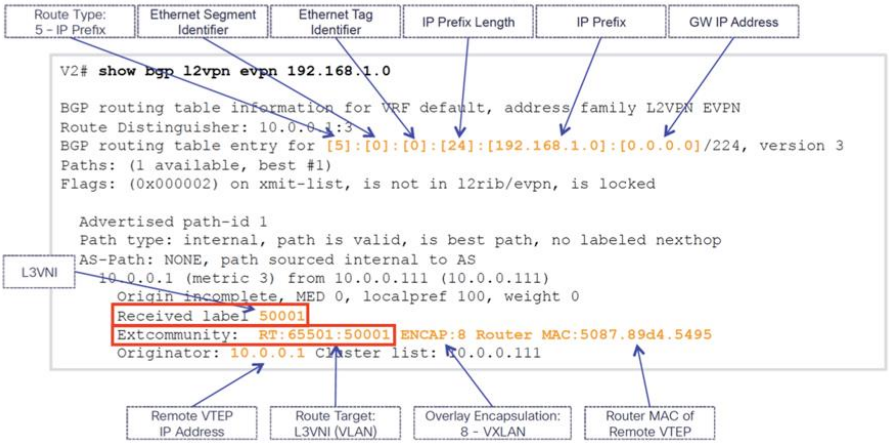
By **Release 5.0(1**), CISCO introduced the *SR-MPLS Handoff,* where VXLAN is no longer necessary because the only protocols involved (and this time between Border Leaf and not Spine anymore, and the external router DC-PE) are:

- **eBGP EVPN** (iBGP EVPN is not supported yet) that carries the VPNv4 prefixes and bottom MPLS label that identify the VRF
- **BGP-lu** (iBGP BGP-lu is not supported yet) that provides Loopback's reachability between the DC-PE and the BL and their own Segment ID (SID) indexes exchange to rise up the eBGP EVPN sessions

The DC-PE one more time, as it was for the GOLF, must re-originates the EVPN prefixes in L3VPN to provide end2end reachability in the SP core and vice versa.

…but, before talking of the different flavors of BGP involved in the *SR-MPLS Handoff* solution, I want to recap some concepts I think useful, to better understand the whole picture.

As you know, in a VXLAN environment, the identifier of a VRF, that classifies uniquely the VRF that the specific prefix belongs to, is the L3VNI ID (in the example below, 50001); that information is transported on EVPN updates type 5 (AFI/SAFI = 25/70) as Extended Community, the Route Target (RT) in the form AS:L3VNI (in the example below, 65501:50001).



Easy, isn't it? ☺

That could mean that the eBGP EVPN session between BL and DC-PE is all I need to propagate the L3VPN prefixes ingress/egress from/to the fabric; actually, something is still missing. eBGP EVPN represents the overlay but we need still the underlay protocol, represented by BGP-lu to transport loopbacks prefix and label information (SR label obtained as the minimum value of the SRGB range, 16000, plus the label index value, SID associated to each single BL and DC-PE) necessary to overlay eBGP EVPN to establish the sessions and defines the SR LSP path between BL and DC-PE. Basically, to make a parallelism, if you consider an ISP provider, it has MPLS with LDP acting as underlay layer to realize the LSP path between loopbacks of PEs, LDP is used to negotiate and propagate the labels and in the meanwhile, BGP VPNv4 is used to propagate as overlay, the L3VPN prefixes end2end …again, making possible the forwarding of traffic in VPN; we are just changing the dialects, but the language is still the same ☺

One moment; CISCO introduced a new solution for L3OUT, they named it as *SR-MPLS Handoff;* it is based on the SR concept as MPLS transport protocol because easier to use and easily interoperates with an ISP environment, however, the protocol used as underlay to propagate the Loopbacks and the SR labels is neither ISIS (with the new sub-TLVs) nor OSPF (with extended Prefix/Link Opaque LSAs), as it normally would be in an ISP environment.

…they adopted BGP-lu, why?

Honestly, I don't have an answer for this question.
ISIS was already not supported on a normal L3Out deployment, BGP was, instead; BGP-lu well accomplishes the Loopback prefixes propagation and SR labels

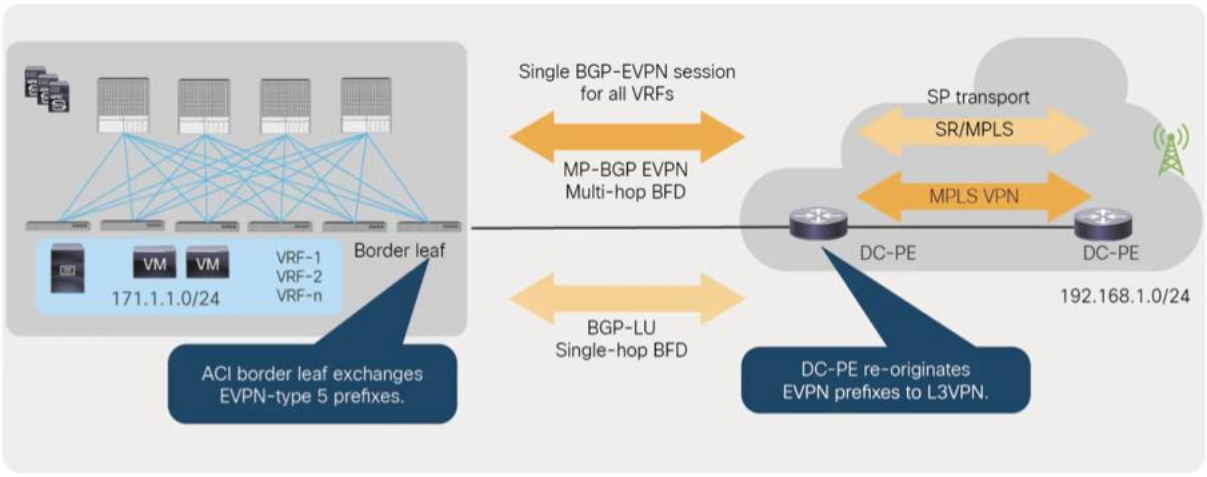
exchanges. You could say BGP is slower than ISIS in convergence, but… BFD is fully supported for BGP-lu since day 0 on SR-MPLS Handoff solution.

So, why not?

The topology used to interconnect BL with DC-PE is quite easy and clear (one hop away; even though more complex SR topologies could be realized to bind the fabric with the SR external world).

After this digression, we can have a look in deep, trying to join the two worlds, ACI and Segment Routing; I want to highlight some aspects of the solution that makes it attractive not only from a marketing standpoint, but also for the cross connections that it involves between the different protocols used.

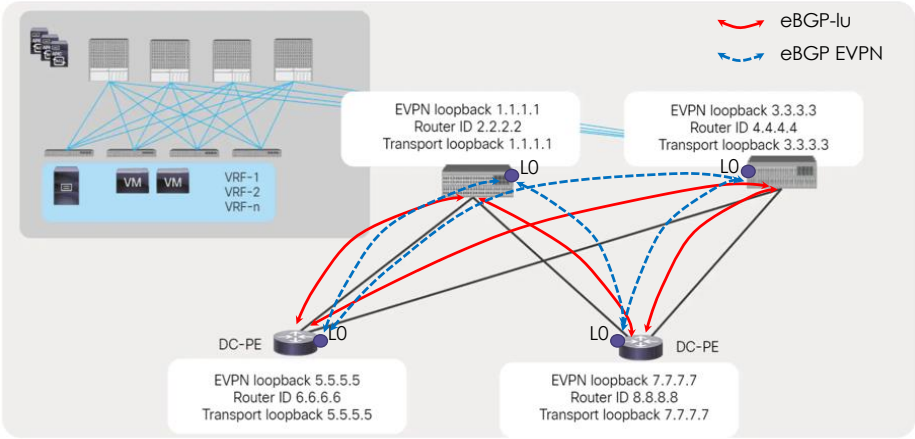
In one picture, all the actors involved are quoted here:



…more specifically, the BGP-lu sessions are configured between the direct connected interfaces of BL and DC-PE and the BGP EVPN are configured between the Loopback addresses, announced by BGP-lu. We are referring now, to one-hop scenario, that means *implicit-null label* announced for the transport Loopback0 by both BL and DC-PE.

When you must configure SR-MPLS Handoff on ACI, the first thing to do is to activate the *SR-MPLS infra L3Out* on border leaf. Three IP addresses must be provided (I don't see any reason why *BGP EVPN loopback* and the *MPLS transport loopback* should be different, so I prefer for them to have the same IP address):

- **Router ID**: it identifies the border leaf and requires to be unique per BL
- **BGP EVPN loopback**: it is configured on the border leaf to build BGP EVPN session with the DC-PE
- **MPLS transport loopback:** is used for data-plane purposes. The BGP prefixes exchanged between the BL and the DC-PE have the next hop as the MPLS transport loopback

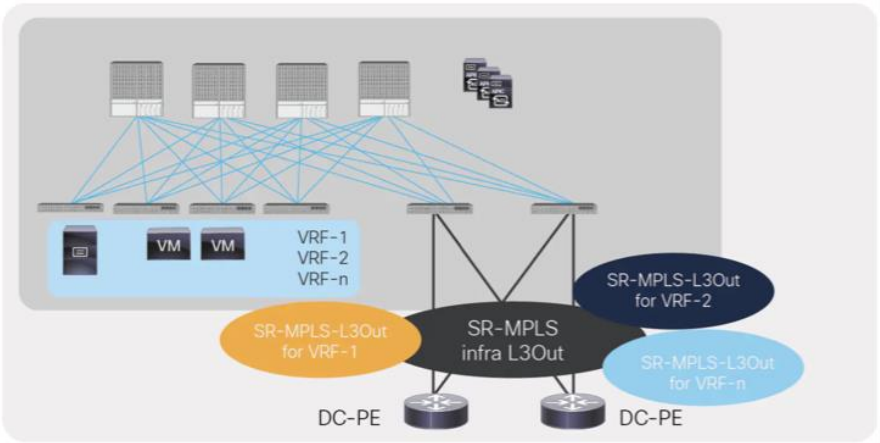


A first thing that give more flexibility to the solution respect GOLF, is the different L3 physical/logical interfaces scenarios you can adopt on BL:

- routed interface or sub-interface
- routed port channel or port channel sub-interface
- for sub-interfaces, any VLAN tag is supported (not only the VLAN ID 4 as for GOLF ☺)

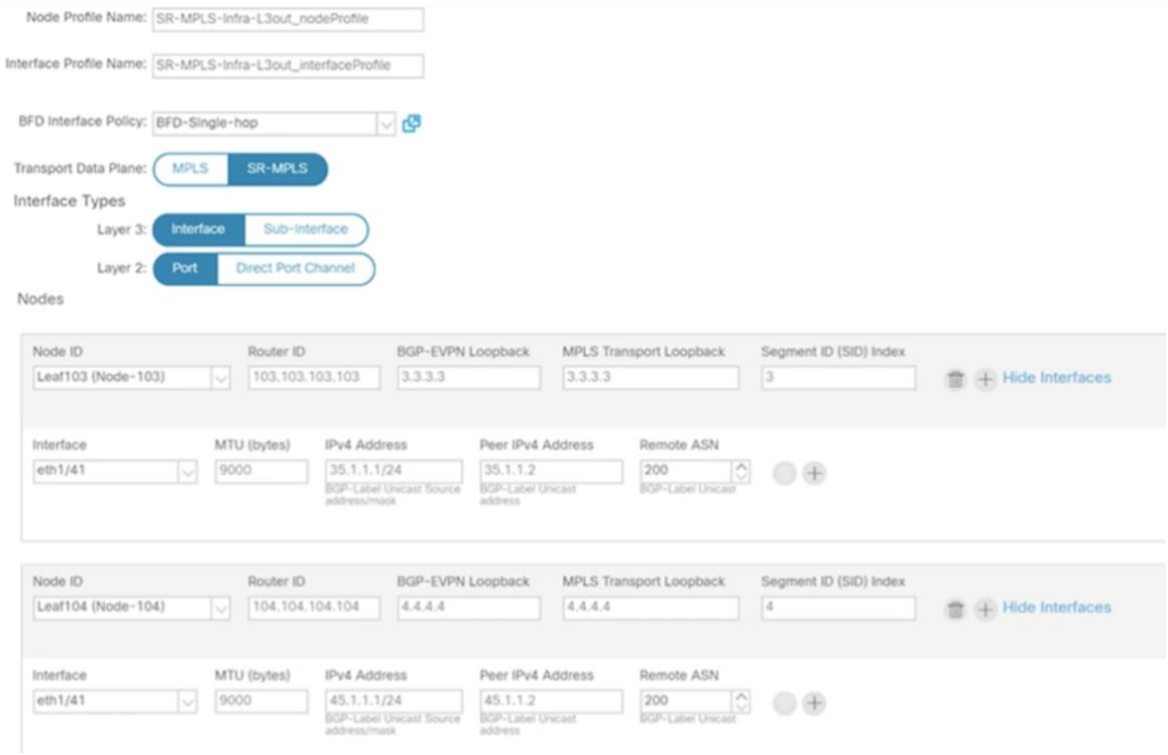
A second difference, this time, with the normal IP L3OUT deployment, is that no *associations* are necessary to be configured between internal BDs and the *SR-MPLS VRF L3Outs* to advertise externally the BD prefix; you need to make sure that an *outbound route-map* is configured correctly to advertise BD prefixes otherwise they are not announced by default (differently, a rout-map in ingress is optional, not strictly necessary).
The *contract* between the internal EPG and the external EPG defined on SR-MPLS VRF L3Out for traffic forwarding between the EPG and SR-MPLS L3Out must be still configured (as it should be, being the contract used to classify the external prefixes to which is allowed to talk with internal EPGs).

We have so, two different kind of L3OUT to be configured (also on GOLF solution we had the "fake" L3OUTs configured at Tenant level, one per each VRF, and the real one configured on the *infra* Tenant ☺), the *SR-MPLS infra L3out* and the many *SR-MPLS VRF L3OUT*.



The *SR-MPLS infra L3Out* that is configured on *infra* Tenant and on the *overlay-1* VRF, where all the BGP EVPN and BGP-lu sessions are configured (including interfaces, BFD one-hop for BGP-lu and multi-hop for BGP EVPN), then we have the different *SR-MPLS VRF L3Out* related to each single VRF of which we want to announce the prefixes outside to put in communication with the external world still in VPNv4 (including the definition of route-maps, External EPG, contracts…).

SR-MPLS infra L3Out



…generic ***SR-MPLS VRF L3Out***

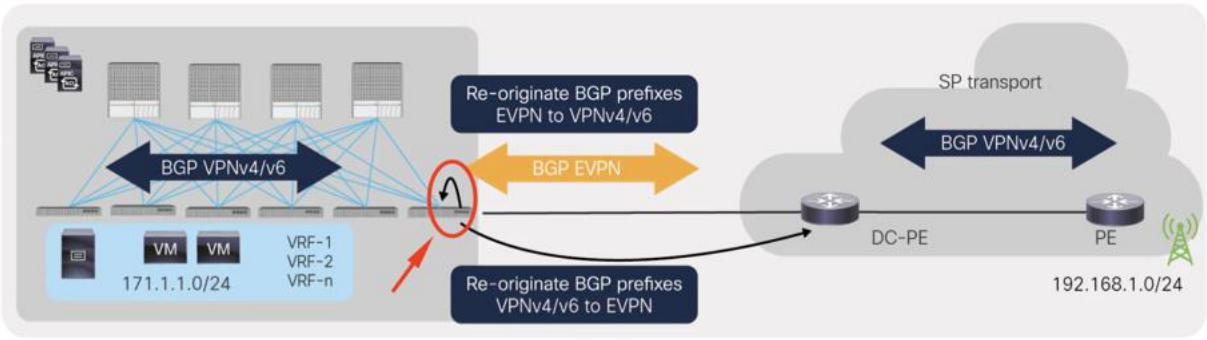
The joining of the two kind of L3Out is done on the single *SR-MPLS VRF L3Out* the referencing the *SR-MPLS infra L3Out* by name.

Before we have talked about the BFD one-hop (ACI supports a minimum timer of 50msec and a minimum-detect multiplier of 3) for BGP-lu and multi-hop (ACI supports a minimum timer of 250msec and a minimum-detect multiplier of 3) for BGP EVPN; why there are two kind of BFD sessions?

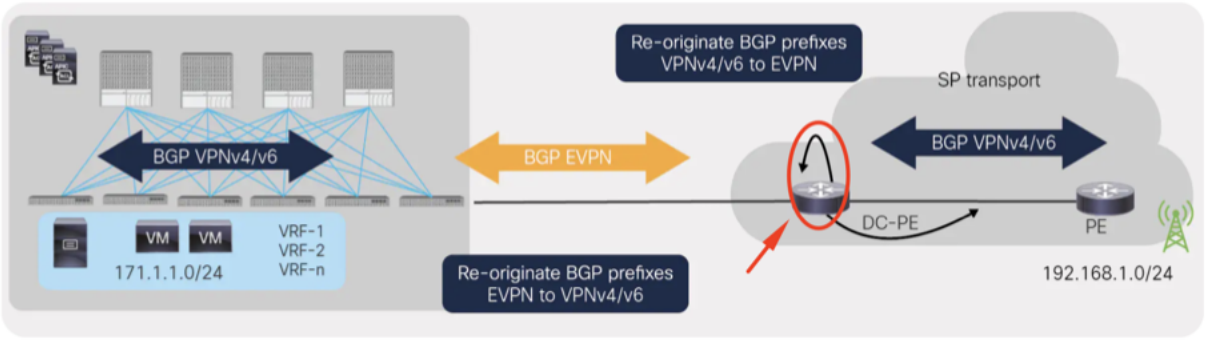
Basically, BGP-lu will be always one-hop sessions, even though the BL vs DC-PE network infrastructure had to include a real SR topology with P routers in the middle; BGP EVPN instead will always be between BL and DC-PE Loopbacks, so for their nature, multi-hop.

Before, going through an example concerning one-hop SR-MPLS Handoff deployment, I'd like to cover another quite important topic, that gives evidence of how the real "SR + BGP VPNv4" world is melting into the "SR (BGP-lu) + BGP EVPN" one, and which are the main actors of this magic.

As you know, BGP VPNv4/v6 address-family are used within the fabric to learn external IPv4/IPv6 prefixes; the solution we are talking about uses BGP EVPN to exchanges prefixes between the BL and the DC-PE. It requires so that BL to re-originate as BGP VPNv4/v6 and advertise them within the fabric, the EVPN prefixes learned by the DC-PE, through the eBGP EVPN session. Similarly, it needs to re-originate the VPNv4/v6 prefixes learned from the fabric into EVPN prefixes and then advertise them to the DC-PE again through the eBGP EVPN session. This is done automatically on BL.



What we have seen before for BL, is not automatically done on DC-PE; DC-PE normally uses the BGP VPNv4/v6 address-family to exchange prefixes (backbone side) whereas, in our scenario, it uses BGP EVPN to exchange prefixes with BL. Therefore, the DC-PE also needs to re-originate the EVPN prefixes to VPNv4/v6 prefixes and re-originate VPNv4/v6 prefixes to EVPN prefixes toward BL in the opposite direction.
This operation is done by configuration on DC-PE.



The process of re-origination and advertising of prefixes from/to EVPN and VPNv4/VPNv6 is based on the concept of *stitching* and *non-stitching RTs*. An EVPN prefix that is received by DC-PE from BL brings with itself an RT configured on ACI, defined as *stitching RT*; the EVPN prefix with stitching RTs is then imported in the local VRF on DC-PE and re-originated into VPNv4/VPNv6 using a non-stitching RT value (stitching RT and non-stitching RT can have the same value or not). Reverse processing happens when a VPNv4/VPNv6 prefix is received by DC-PE with a non-stitching RT, is then imported in the VRF and re-originated into EVPN using a stitching RT.

The extract of configuration on DC-PE that does what just described is quoted here:

…in *BGP EVPN address-family* peering with BL:

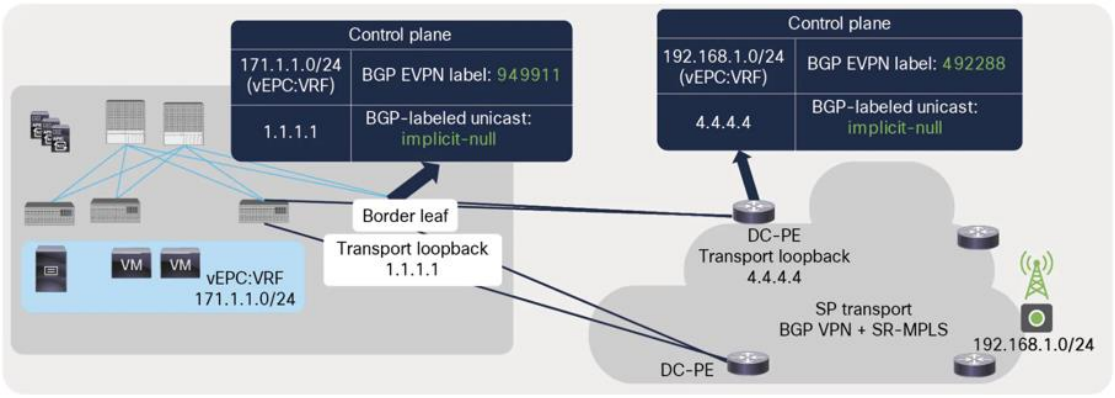
```
router bgp 1
  neighbor-group ACI-site-EVPN
  address-family l2vpn evpn
   import stitching-rt re-originate
   advertise vpnv4 unicast re-originated stitching-rt
   advertise vpnv6 unicast re-originated stitching-rt
   !
  !
```

…in *BGP VPNv4/VPNv6 address-family* peering with the Router Reflector on SR backbone:

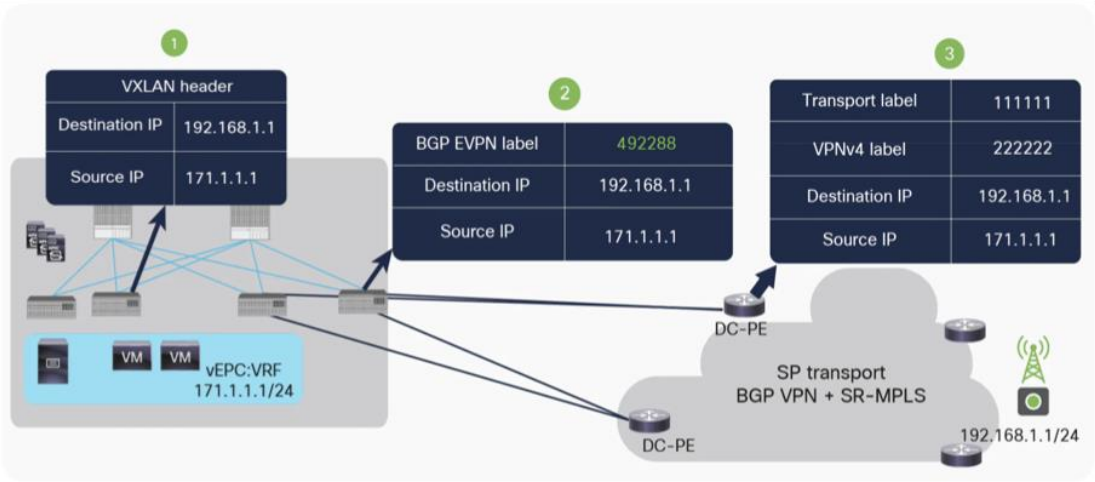
```
router bgp 1
 neighbor-group RR-VPNunicast
  address-family vpnv4 unicast
   import re-originate stitching-rt
   advertise vpnv4 unicast re-originated
   !
  address-family vpnv6 unicast
   import re-originate stitching-rt
   advertise vpnv6 unicast re-originated
```

Let's see now, with an example, how control plane works in terms of labels assigned (consider that the example is concerning the one-hop topology, so an implicit null - label 3 – is announced by both BL and DC-PE each other via BGP-lu for the transport layer).

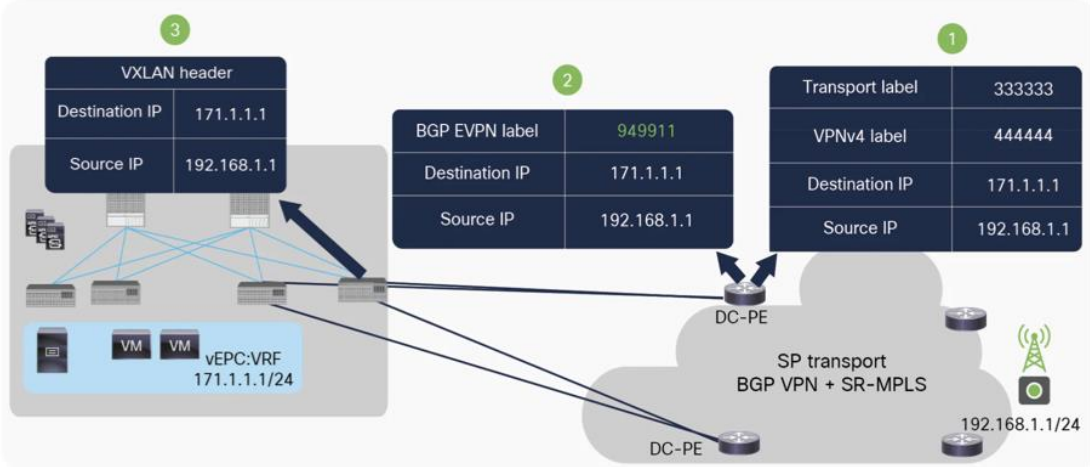
The bottom labels announced via BGP EVPN by BL (949911) ad DC-PE (492288) each other as aggregate labels for VRF, are the *service labels;* they are used locally to recognize which VRF the packet received must be associated to. Once the packet is received respectively by BL and DC-PE, it is de-encapsulated and route lookup in the VRF routing table executed; based on that route lookup, the new VPN and transport label are applied and send finally to the destination (once entered on the ACI fabric, actually, the packet is VXLAN encapsulated, so no Transport label will be applied).



Here it is illustrated the direction ACI to SR transport data-path:



Here it is illustrated the direction SR transport to ACI data-path:



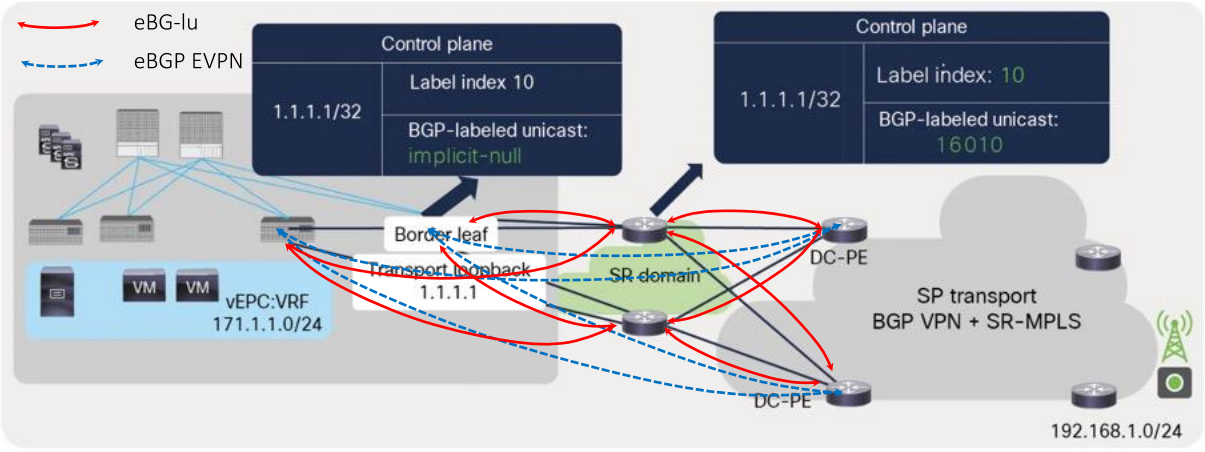
The *Transport label* on DC-PE in the direction SR to ACI, would be absent, because of penultimate hop popping task (it depends on the configuration adopted on DC-PE, SR-MPLS backbone side).

The scenario where the transport *Top label* comes into the scene is when between BL and DC-PE there is a real L3 SR-MPLS network that requires the SR label to correctly forward the packet.
In that case, the *SID to Loopback* assignment is done uniquely by each SR router and label announced using the BGP-lu along the SR LSP path.
On SR domain who decides the LSP path is the head-end of the SR LSP, in fact normally for the SR packet forwarding, the *source-routed path* is expressed as an ordered list of segments imposed by the ingress router; the FEC classification is done by the ingress router and remains the same along the whole LSP path.

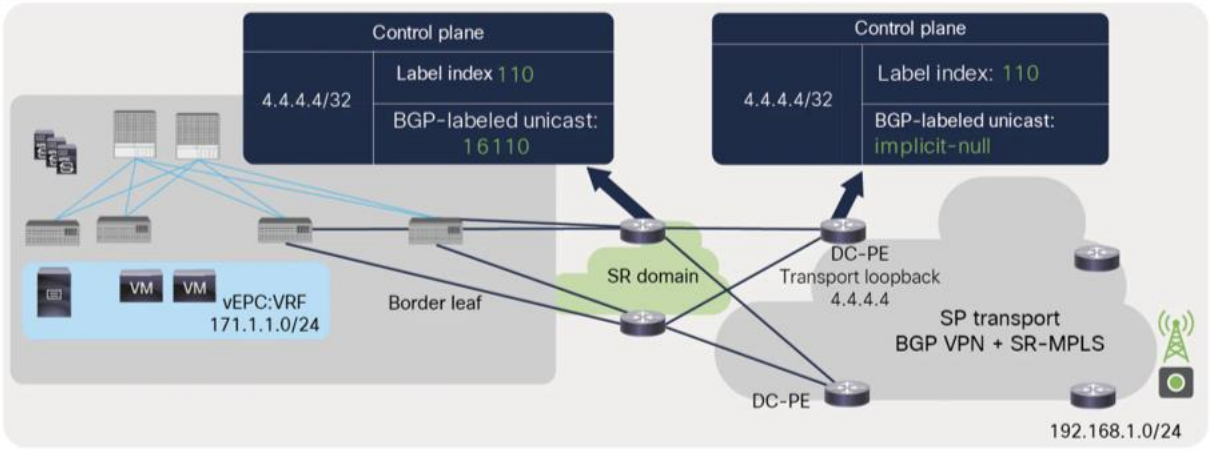
The extract on configuration in XR necessary to assign the SID to the loopback is shown here:

```
router bgp "local AS#"
 bgp router-id "L0"
 address-family ipv4 unicast
  network "L0"/32 route-policy SID("prefix-SID index") → set prefix-SID index
  allocate-label all
 !
…
route-policy SID($SID)
  set label-index $SID
end-policy
!
```

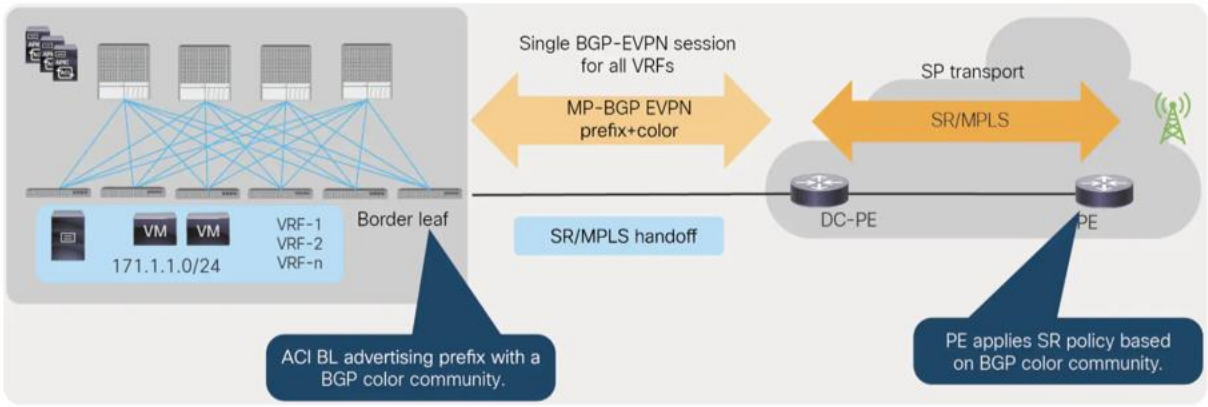
In the direction DC-PE to BL for instance, the packet coming from DC-PE would have at first hop the top label 16010 (SID = 10 is assigned by BL), then it will be removed (because of *penultimate hop popping*) to reach finally the BL.



In the direction BL to DC-PE, the packet coming from BL would have at first hop the top label 16110 (SID = 110 is assigned by DC-PE), then it it will be removed to reach finally the DC-PE.



We have seen previously that an outbound route-map in SR-MPLS VRF L3Out is required; it must be explicitly configured to select the ACI fabric prefixes that need to be leaked/announced outside to the SR network.
One of the coolest applications that the ACI into SR-MPLS Handoff integration involves is the setting of route policies such as the *BGP color community* for the prefixes advertised to the SR network.
The BGP color community can be used in fact in the SR network to select SR policies and path selection based on different requirements providing a better automation and policy consistency across data center and transport domains.



To understand the power of color community, let's remark some aspects of this attribute:

- Color is a BGP Extended Community expressed as 32-bit values of a segment routing policy that determines the sets of traffic flows that are steered by the specific SR policy it is referencing to
- SR Policy colors are used to identify an intent or SLA (an intent is a set of constraints). Such color is used to match a service route that requires a given SLA to an SR Policy that provides the path that satisfies this SLA (an SLA can be based on minimum delay path, shortest IGP path, jitter…)
- In general, a BGP route has (among the other ones) two attributes:
    - the *next-hop* that indicates where to go
    - the *color* that indicates *how to go* the *next-hop* (the intent)

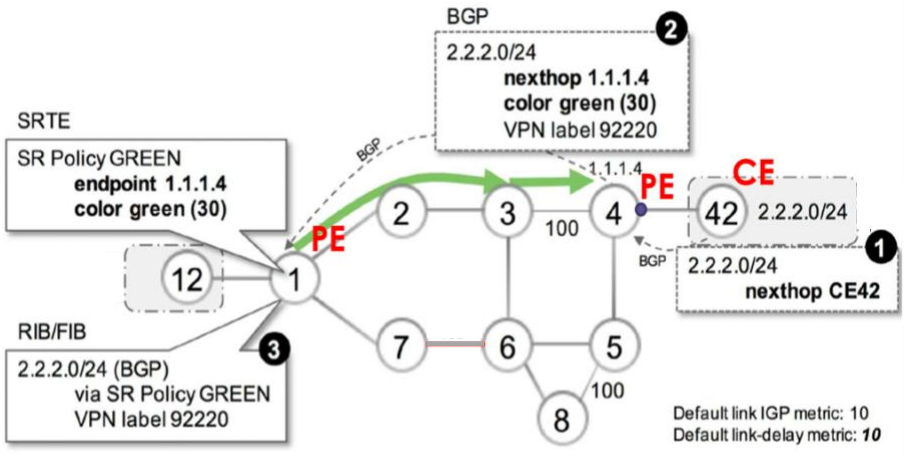
As example, it is illustrated a generic scenario where we show the use of color community (the goal is to setup a TE LSP path with the minimum end2end delay, for a pool of destination prefixes).

The delay is measured based on telemetry that relies on the *Performance Measurement (PM) functionality;* it provides a generic framework that dynamically measures various characteristics (delay, loss…) of a network.
The measurement method relies on timestamping in Query and Response packets, as defined in RFC 6374 "Packet Loss and Delay Measurement for MPLS Networks", RFC 4656 "One-way Active Measurement Protocol (OWAMP)", and RFC 5357 "Two-Way Active Measurement Protocol (TWAMP)", exchanged between the p2p links.

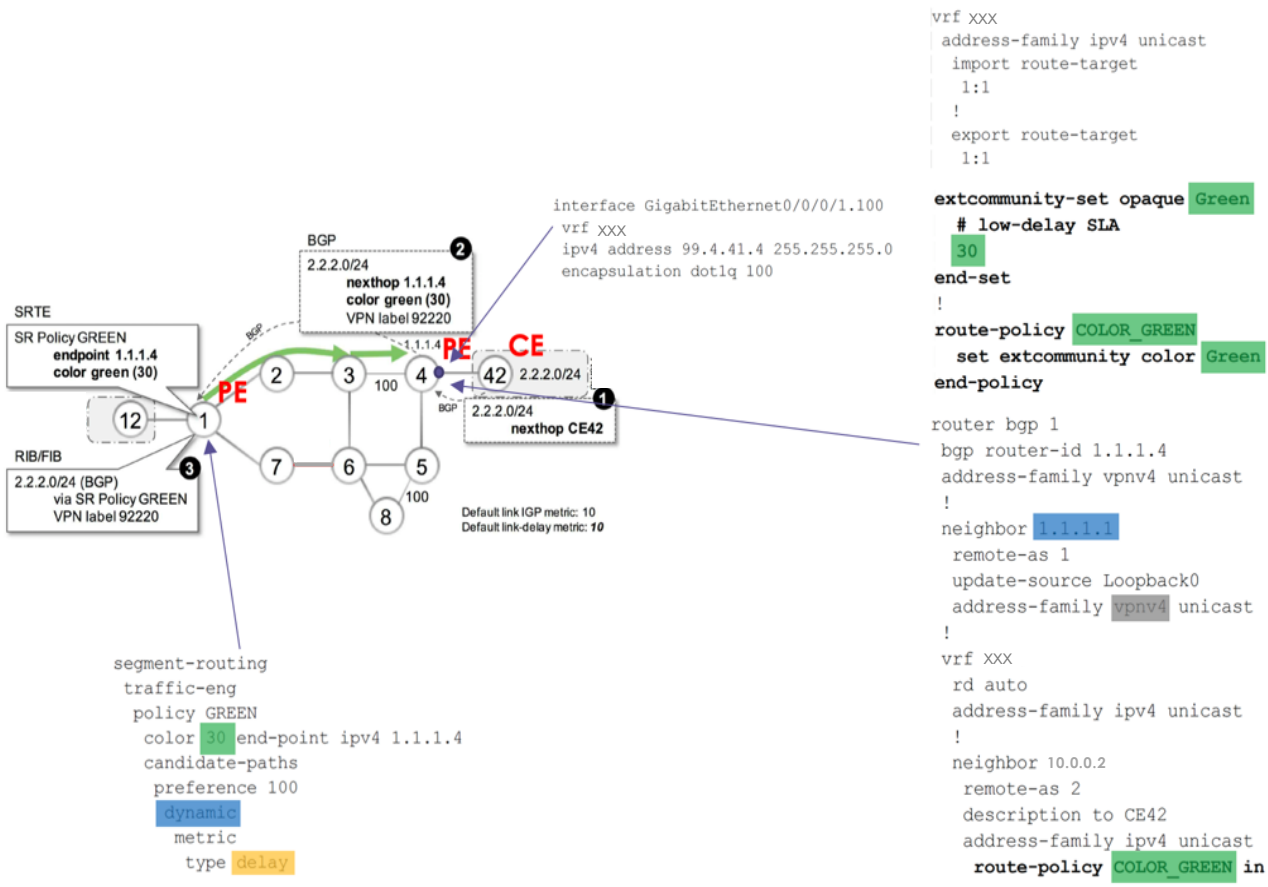
Here is highlighted an example that shows how to enable delay measurement on an interface:

```
performance-measurement
 interface GigabitEthernet0/0/0/0
  delay-measurement
 !
 interface GigabitEthernet0/0/0/1
  delay-measurement
```

Going back to our example, let's suppose to have a CE, Node42 advertising prefix 2.2.2.0/24 via BGP to PE Node4.
PE Node4 allocates a VPN label 92220 for the prefix 2.2.2.0/24 in VRF and advertises this prefix via BGP VPNv4 to PE Node1 with the BGP next-hop set to its loopback address 1.1.1.4.
PE Node4 adds also a color "green" (tag 30) to the BGP update for prefix 2.2.2.0/24 identifying with that, the low-delay SLA.
PE Node1 receives the BGP VPNv4 route that finds the matching SR Policy "green", with color 30 and NH 1.1.1.4; automatically installs the prefix 2.2.2.0/24 in the forwarding table pointing to the SR Policy "green", instead of pointing to the BGP next-hop directly.
All traffic destined for route 2.2.2.0/24 is then forwarded following what established by the SR Policy "green".



… the generic configuration is shown here below:



This is an example of *Automated Steering (AS)* functionality. It automatically steers service prefixes into the SR Policy that provides the desired intent or SLA; it automates the steering of the service traffic on the SR Policies that deliver the required SLA.

Where is the beauty of this solution?

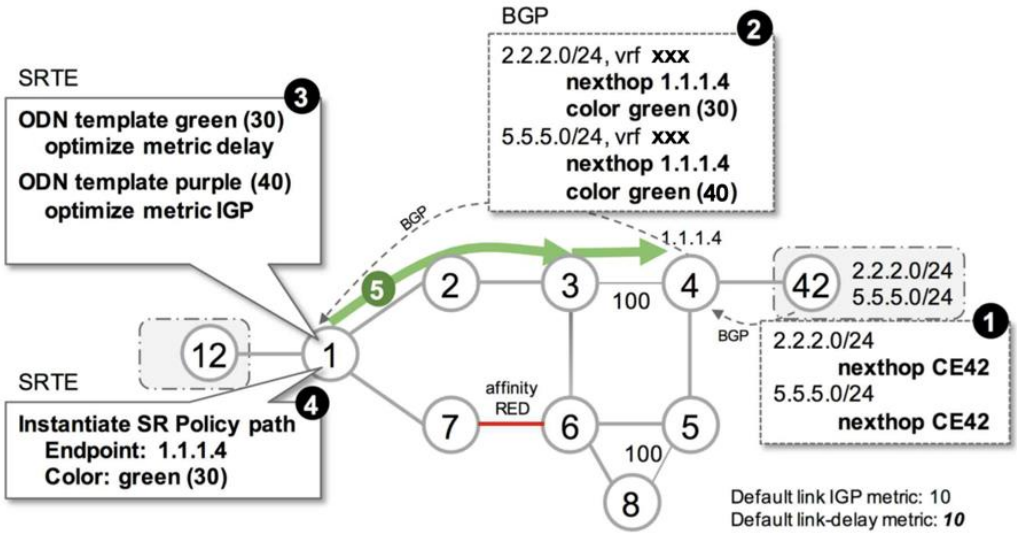
First, the *PM functionality* that makes available to the network the measurements hop by hop of the delay constraint; secondly, the ease of implementation: once associated the specific color to the prefix, the head end of the SR LSP path (PE1 in our example) will decide dynamically the right path that satisfies the requirements specified on the SR policy, the LSP at minimum delay in our case, and creates the SR labels stack to be pushed on the packet.

The SR policy, however, not necessarily has to be instantiated statically (like in the configuration example), it can be done dynamically based on a color template and endpoint's Next Hop; we are referring to the *On-Demand Next-Hop (ODN)* feature. This is a fundamental simplification as one no longer needs to preconfigure any SR Policy.

Moreover, considering that the focus of a network operator is to support services, ODN can be seen as the key achievement for SR-TE implementation.
The traditional RSVP-TE solution would involve a complex mesh of pre-configured tunnels and a lot of operation activities to steering traffic into the RSVP tunnels.

With ODN, the solution is much easier to implement; the service prefixes are tagged, and the related SR Policies are automatically instantiated and used for steering.
Let's analyze still another example, this time with two services, one (2.2.2.0/24) that must be supplied following the SR LSP at minimum delay, the second one (5.5.5.0/24) that must be supplied following the SR LSP at minimum IGP cost (best metric).

Associating two different colors 30 and 40 to the two services, the SR policies can be instantized dynamically using the ODN paradigm, as shown in the extract of configuration adopted on PE Node1.

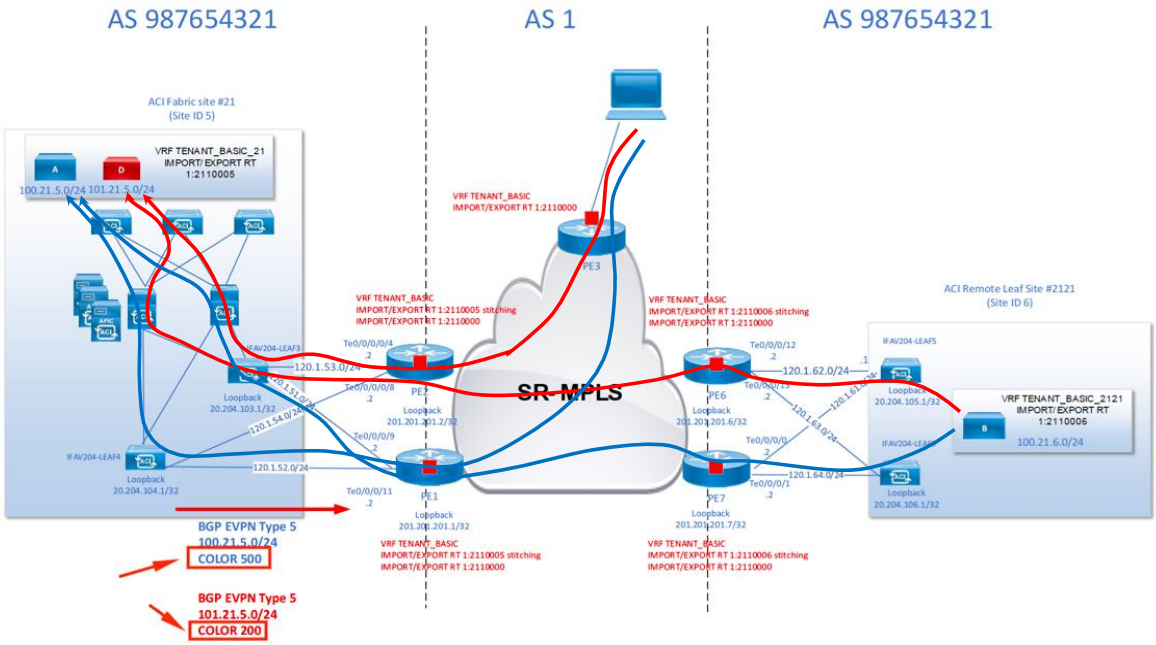



```
segment-routing
 traffic-eng
  on-demand color 30
     dynamic
      metric
        type delay
  !
  on-demand color 40
   dynamic
    metric
     type igp
```

Going back to our main topic, the Data Center ☺, it's clear that the "coloring" process based on different requirements per each service, makes this tool very attractive, above all for the automation that it potentially involves; this functionality allows seamless integration (like Unified MPLS does with BGP-lu for interdomain MPLS segmented LSP) between the DC and the SR MPLS transport network.

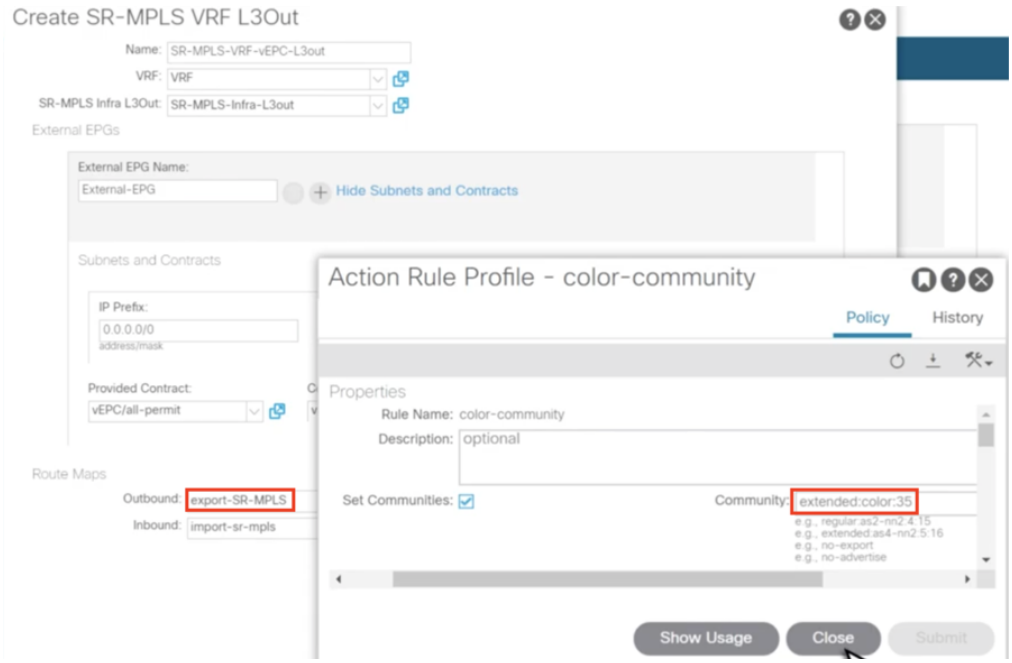
Let's see another example, this time concerning services A and D, hosted on a Data Center which require respectively a *best effort* transport and a *low latency* transport LSP paths as constraints (SR policies).
Let's suppose that the ACI fabric signals the color 500 associated to the *best effort* SR policy and the color 200 associated to the *low latency* SR policy; it advertises the prefix 100.21.5.0/24 of the service A with the BGP color 500 and the prefix 101.21.5.0/24 of the service D with the BGP color 200.

The BGP color, being a *transitive optional attribute*, is carried transparently by the connected DC-PE into the SP core and when a remote PE, like PE6 or PE3, receives the prefixes with their own colors, they leverage the *Automated Steering (AS)* or *On-Demand Next-Hop (ODN)* feature to automatically steer the traffic to the destination into the appropriate transport path that respect what defined on SR policies.



Where are set the colors on ACI?

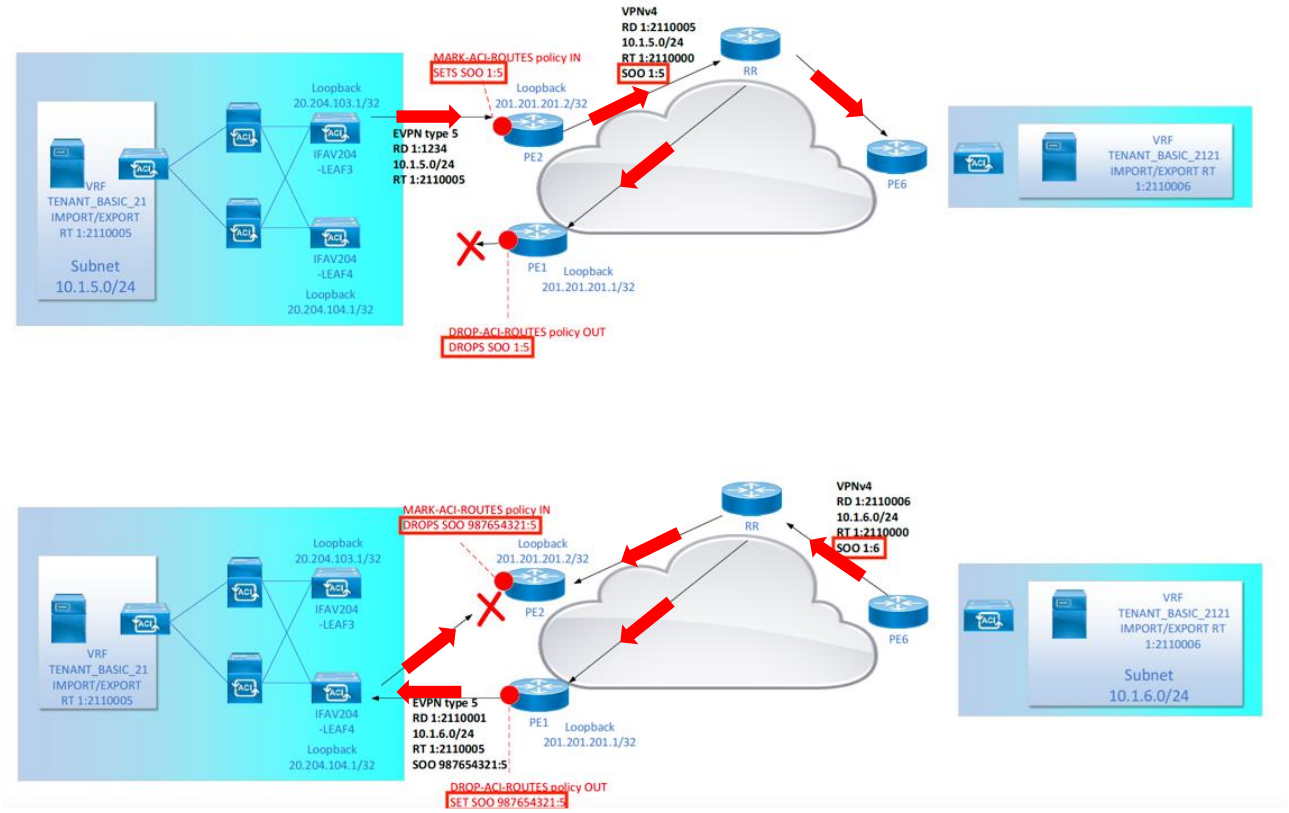
The outbound route-map is also used to set route policies such as the BGP color community for the prefixes advertised to the SR network (besides the RT import/export that has to match with the ones configured on DC-PE, the ones that previously we named *RT stitching*).



We are almost at the end of this paper; before, I'd like, however, to touch two topics still concerning the SR-MPLS Handoff implementation that I believe are worth considering; Loop prevention and the use of SR-MPLS Handoff instead of IPN network for the multi-POD scenario.

Let's start with the *loop prevention.*

It's usual to implement multihoming for redundancy between BL and DC-PE devices; routing loop is the first side effect because prefixes learned from a DC-PE would be reinjected back into the DC (the same is in the opposite direction); for instance, in the figure below, when PE2 learns a prefix from the fabric, it would propagate the prefix to the ISP core network and PE1 would receive it, propagating the prefix back again to the fabric (the same happens in the opposite direction).



The previous scenario shows the use of two route-policies configured on DC-PEs for Loop prevention on the EVPN BGP sessions with the BLs:

- inbound route-policy MARK-ACI-ROUTES
- outbound route-policy DROP-ACI-ROUTES

As multiple ACI sites may use the same AS number (in our case the different fabrics are using the same AS 987654321 and DC-PEs are in AS 1), then the embedded *ASPATH loop prevention* procedure based on AS# is normally disabled (*allow-as* feature is adopted).

Site of Origin must be used on DC-PEs in both directions (one SOO value per direction) to prevent loops.

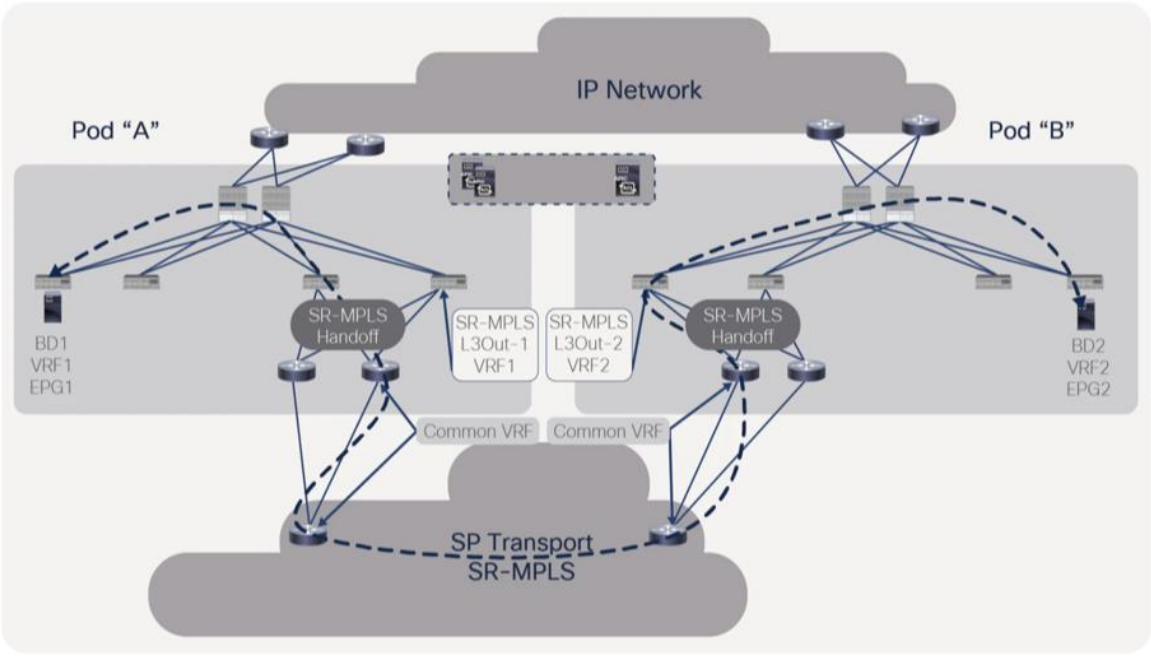
MARK-ACI-ROUTES route-policy marks the prefixes with an SOO associated to the pair of BLs (SOO 1:5) the DC-PE are facing with and drops prefixes with the SOO value associated to the pair of DC-PEs connected to the ACI site (SOO 987654321:5).
DROP-ACI-ROUTES drops the prefixes with an SOO associated to the pair of BLs (SOO 1:5) the DC-PE are facing with and marks the prefixes with an SOO associated to the pair of DC-PEs connected to the ACI site (SOO 987654321:5).

Let's examine now the last topic.

SR-MPLS Handoff can be configured on each POD. Normally the inter-POD traffic is forwarded through the IPN network following the VXLAN path, however it could be decided to use the SR-MPLS path.
A separate SR-MPLS infra L3OUT must be configured on each pod.

To use the SR-MPLS path for inter-POD traffic the following rules must be satisfied:

- a separate VRF needs to be configured on each ACI POD
- no contracts must be configured between EPGs in the different VRFs across PODs
- a common VRF must be configured on DC-PEs of each POD realizing the mapping of the VRFs of each POD with itself



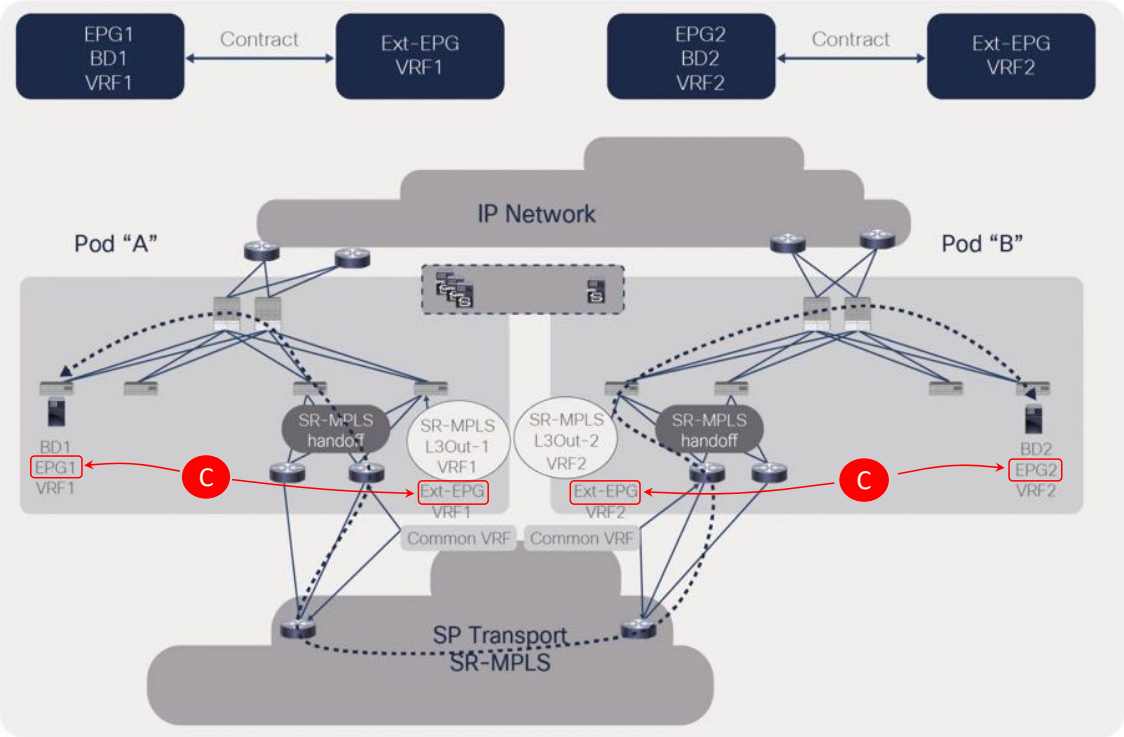
If the same VRF had to be configured on all PODs (same RT), the leaves of each ones would learn each other's prefixes directly over the IPN, preferring the internal path over the IPN to forward VXLAN traffic to each other.
If instead each POD is configured with different VRFs (different RT) and on the DC-PE of each POD, these VRFs are mapped to a common VRF, the BLs would have the reachability to the remote POD's prefixes via the SR-MPLS path only.

Example:

- Backbone (AS X):

    - VRF common
        - SOO X:1
        - RT X:111 import/export (on DC-PE)
- POD1 (AS 1):
    - VRF1
        - SOO 1:1
        - RT 1:1 stitched import/export (on SR-MPLS VRF L3OUT)
- POD2 (AS 1):
    - VRF2
        - SOO 1:2
        - RT 1:2 stitched import/export (on SR-MPLS VRF L3OUT)

Even though different VRFs had to be configured on each POD, the presence of contracts between EPGs referencing different VRFs, would involve the routes leaking between the VRFs, and one more time, the reachability between POD's prefixes would be provided via the IPN path.
To avoid this situation, the contract should be configured with a local external EPG in its own VRF; this would ensure that connectivity is always via the SR-MPLS path and that there is no route leaking across VRFs.



What seen for multi-POD scenario is still valid for multi-SITE scenario.

This time, we are really ☺ at the end of this paper. As usual, I hope you enjoyed going through the different aspects that characterize the quite new SR-MPLS Handoff feature and that I've been able to transmit you the potentialities it has above all considering the emerging Segment Routing framework as new paradigm for MPLS transport layer on ISP environments.