IDR WG

Segment Routing TE Policy

draft-previdi-idr-segment-routing-te-policy

Stefano Previdi – sprevidi@cisco.com
Clarence Filsfils – cfilsfil@cisco.com
Arjun Sreekantiah – asreekan@cisco.com
Siva Sivabalan – msiva@cisco.com
Paul Mattes – pamattes@microsoft.com

IETF95 Buenos Aires, April 2016

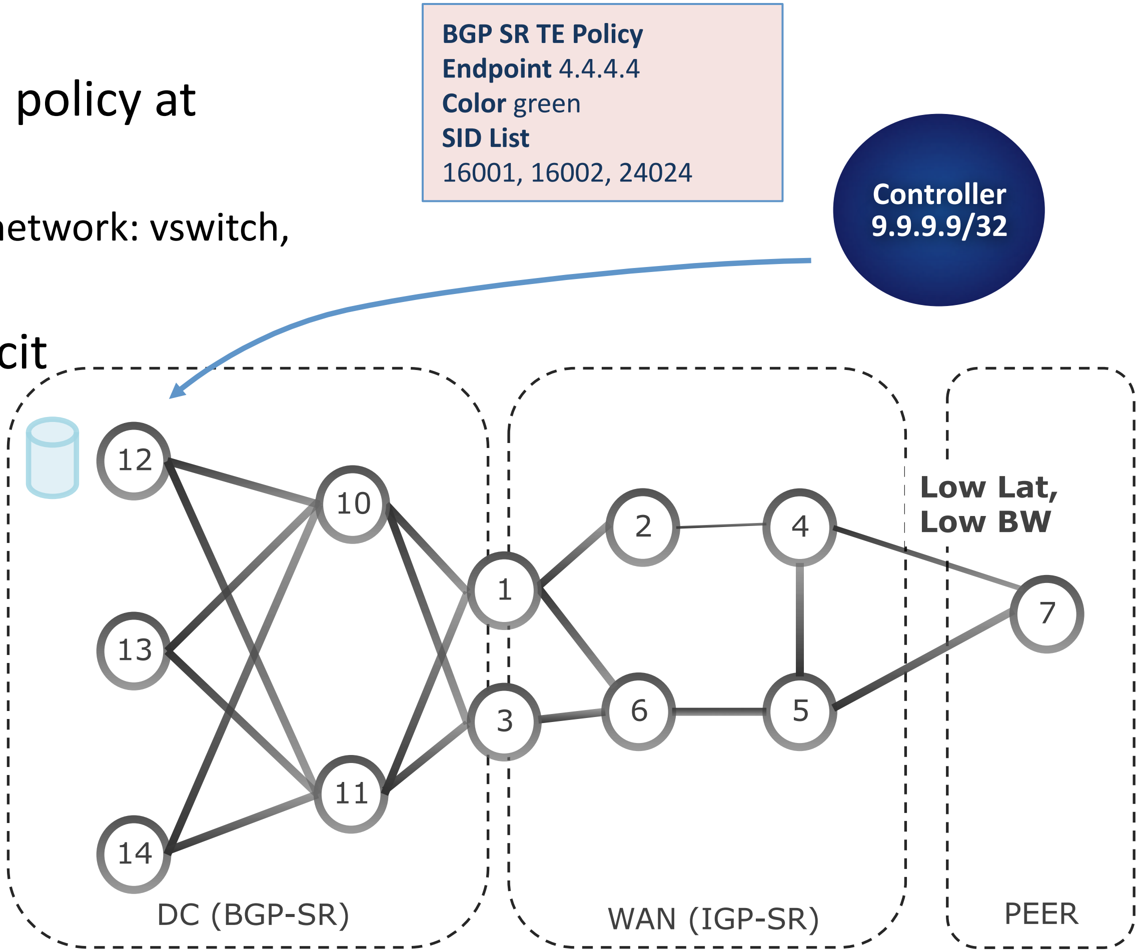
Introduction

- What is it ?
  - An ability to advertise in BGP a TE policy (e.g., low latency path, disjoint path, etc.) including a [u|e]cmp set of explicit paths
  - An ability to classify traffic into a TE policy
- What is the motivation ?
  - Ever growing interest in simplifying network operations
  - TE policy is advertised by a BGP speaker as a list of segments
  - No need to configure tunnels and the associated traffic steering mechanisms such as PBR
  - Existing mechanisms like BGP PIC FRR are preserved.
  - Policies are ingress related, i.e., two ingress routers may have different policies for reaching the same egress

Creating an SRTE Policy

- Controller programs an SR TE policy at ingress
  - This could be anywhere in the network: vswitch, spine, DCI, PE, Agg …
- SR TE Policy defines the explicit path from ingress to policy endpoint
- An SR TE Policy is identified through:

  <Color, Endpoint>



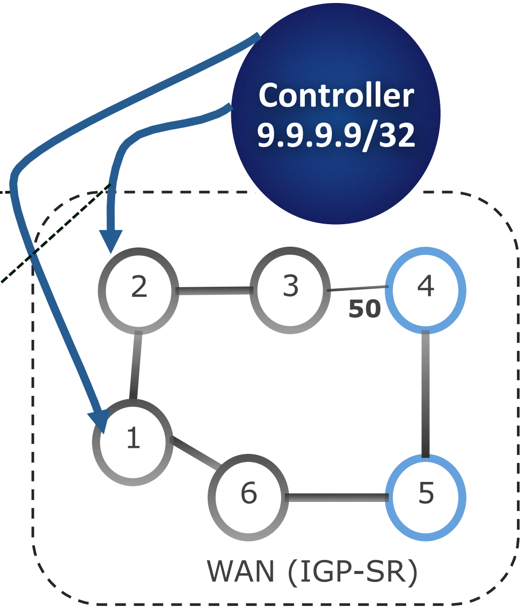
BGP SR TE Policy
Endpoint 4.4.4.4
Color green
SID List
16001, 16002, 24024

Creating an SRTE Policy

- Same SR TE Policy may be expressed with different content for different ingress nodes

```
SR TE Policy:
Endpoint: 4.4.4.4
Color: green
Segment List:
           16003
           24034
Binding SID: 4001
```

```
SR TE Policy:
Endpoint: 4.4.4.4
Color: green
Segment List:
           Weight: 100
           16003
           24034
Segment List:
           Weight: 200
           16006
           16004
Binding SID: 4001
```



SR TE Policy Advertisement in BGP

- A BGP speaker (router or controller) advertises SR TE policies in the form of SID list, Weight, etc.
- Multiple objects define a SR TE Policy
  - Segment List
  - Weight (unequal cost multipath)
  - Binding SID (request allocation of BSID)

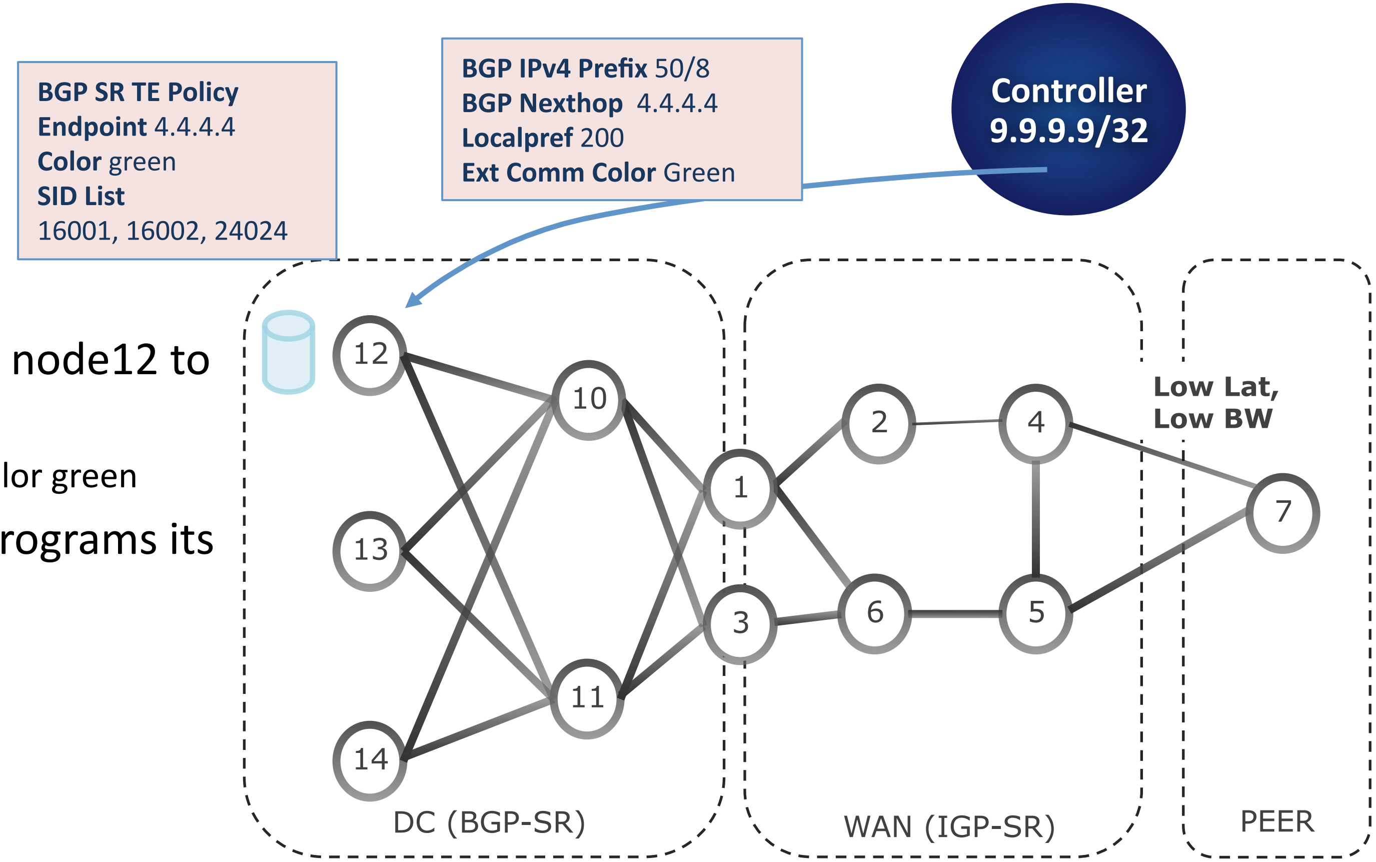
Role of the client

- Receive the policy
- Program dataplane with SR TE Policy instantiation
- The client does not need to do any TE optimization. The SID list is given explicitly by the controller

Classification and Traffic Steering

- A steering mechanism is also needed so to use a SR TE Policy for a given traffic flow
  - Steering onto an SR Policy involves the classification of packets into the specified SR policy: color extended community
- A destination prefix is steered into a policy if
  - the color of the destination prefix matches the color of the policy AND
  - the next-hop of the destination prefix matches the endpoint of the policy (if present)

Steering traffic on an SR TE Policy



BGP SR TE Policy
Endpoint 4.4.4.4
Color green
SID List
16001, 16002, 24024

BGP IPv4 Prefix 50/8
BGP Nexthop 4.4.4.4
Localpref 200
Ext Comm Color Green

Controller
9.9.9.9/32

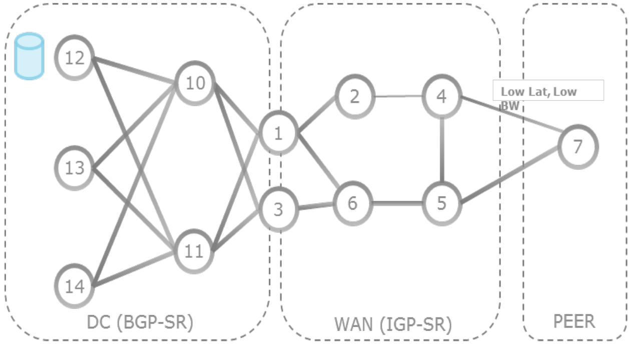
- Controller programs node12 to steer 50/8 via:
  - bgp nhop 4.4.4.4, color green
- As a result, TOR12 programs its FIB accordingly

WECMP within a (nhop, color) path

- When traffic is steered into a policy
  - Weighted ECMP ia used across SID lists, according to “weight” value

BGP SR TE Policy
Endpoint 4.4.4.4
Color green
SID List (set)
16001, 16002, 24024, weight 2
16003, 16002, 24024, weight 1

BGP IPv4 Prefix 50/8
BGP Nexthop 4.4.4.4
Localpref 200
Ext Comm Color Green



ECMP between Policies

BGP SR TE Policy
Endpoint 4.4.4.4
Color green
ERO SID List set
16001, 16002, 24024, weight 2
16003, 16002, 24024, weight 1

BGP SR TE Policy
Endpoint 5.5.5.5
Color yellow
ERO SID List
16001, 16005, weight 1
16003, 16005, weight 1

- Traffic may be steered to different policies
  - E.g.: a destination prefix is advertised (add-paths) with different next-hops and different colors

BGP IPv4 Prefix 50/8
BGP Nexthop 4.4.4.4
Localpref 200
Ext Comm Color Green

BGP IPv4 Prefix 50/8
Add-Path
BGP Nexthop 5.5.5.5
Localpref 200
Ext Comm Color yellow

- Traffic is steered into the two policies
  - WECMP between Segment Lists according to weights

IMPORTANT Aspects of SR TE Policy

- Advertising a TE Policy is new in BGP
  - SR TE Policy is NOT a prefix advertisement and it is not related to any prefix
  - SR TE Policy is NOT a tunnel advertisement and it is not related to any tunnel
  - SR TE Policy is NOT an attribute of a prefix and it is not related to any specific prefix
  - IOW: a SR TE Policy is a new and self-contained BGP advertisement

IMPORTANT Aspects of SR TE Policy

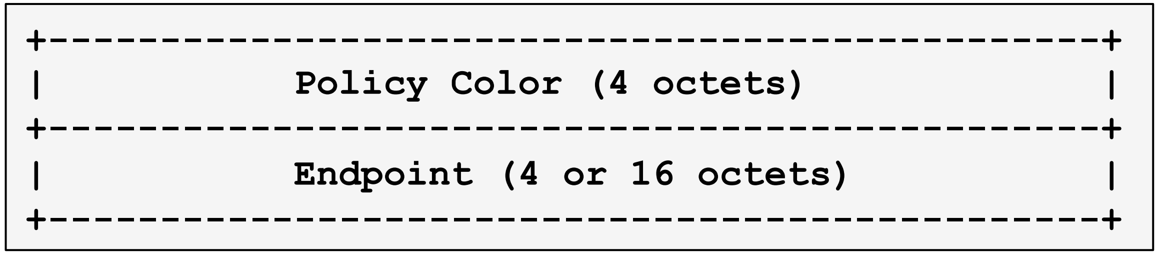
- Granularity is the policy, not the endpoint
  - Policy is identified by [<color><endpoint>] tuple
  - NOTE WELL: <endpoint> may be a generic/wildcard one
    - IOW: a Policy may not have an endpoint. It's valid.
- Scalability/Flexibility:
  - If a given policy changes (e.g., the Segment List) only that policy needs to be re-advertised
  - If a new policy is defined, only that new policy needs to be advertised
- Not bound to the BGP next-hop
  - Any destination can be steered to any policy. No need to honor BGP next-hop attribute
  - E.g.: a SR TE Policy may even not have any endpoint (service/application based)
- No message size (BGP MTU) issue

SR TE Policy Requirements

- Thousands of SR TE Policies may be advertised by a single node (controller)
  - The BGP speaker originating the SR TE Policies (typically a controller) will originate hundreds of policies for each ingress PE. In total the controller will originate several thousands of policies
- It MUST be possible to advertise, update, replace or withdrawn a single policy without requiring to re-advertise all of them.
  - While, in some cases, grouping policies within the same NLRI advertisement may be helpful, the implementation MUST be capable of originating and receiving a single policy per NLRI advertisement

Encoding Structure

- New SAFI: SR TE Policy
- New SR TE Policy SAFI NLRI



```
+------------------------------------------------+
|            Policy Color (4 octets)             |
+------------------------------------------------+
|            Endpoint (4 or 16 octets)           |
+------------------------------------------------+
```

- Characteristics of the Explicit Path described in Tunnel-Encaps attribute
  - draft-ietf-idr-tunnel-encap

Encoding Structure

- Example of SR TE Policy encoding

```
SR TE Policy SAFI NLRI: <Policy-Color, Endpoint>
  Attributes:
    Tunnel Encaps Attribute (23)
      Tunnel Type: SR TE Policy
        Binding SID TLV
        Segment List TLV
          Weight TLV
          Segment TLV
          Segment TLV
        Segment List TLV
          Weight TLV
          Segment TLV
          Segment TLV
```

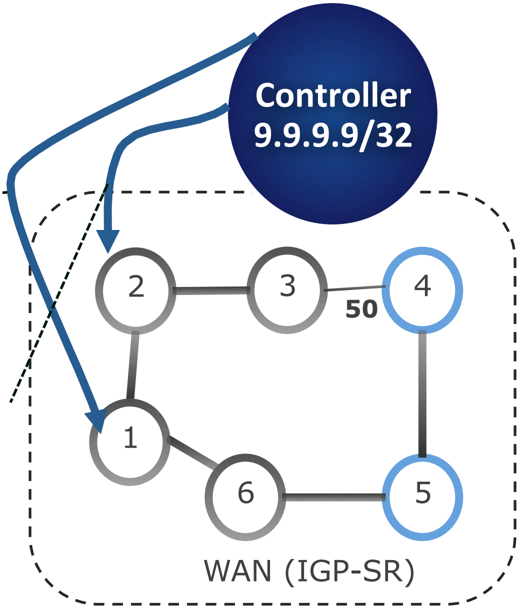
Encoding Structure

- In most of the cases, the SR TE Policy is intended for the receiver only
  - Use of NO_ADVERTISE community
- Therefore, a policy in the form of
  - <color, endpoint>

  May have different content (i.e.: different segment lists)

```
SR TE Policy:
Endpoint: 4.4.4.4
Color: green
Segment List:
          16003
          24034
Binding SID: 4001
```

```
SR TE Policy:
Endpoint: 4.4.4.4
Color: green
Segment List:
          Weight: 100
          16003
          24034
Segment List:
          Weight: 200
          16006
          16004
Binding SID: 4001
```

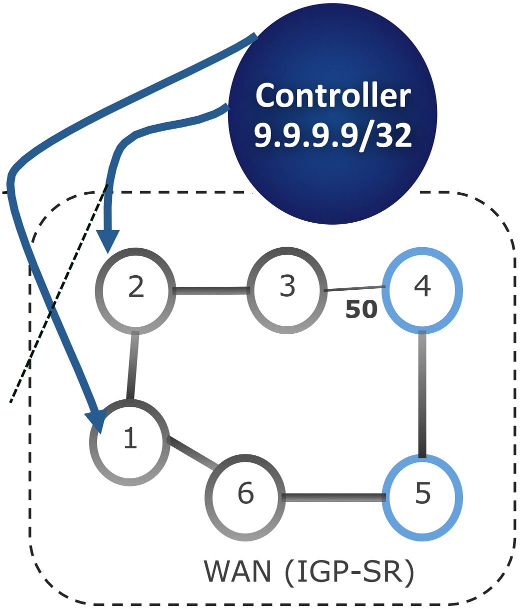


Encoding Structure

- In most of the cases, the advertisement is originated and sent by a controller directly to the receiver
  - No RR in the middle

```
SR TE Policy:
Endpoint: 4.4.4.4
Color: green
Segment List:
           16003
           24034
Binding SID: 4001
```

```
SR TE Policy:
Endpoint: 4.4.4.4
Color: green
Segment List:
           Weight: 100
           16003
           24034
Segment List:
           Weight: 200
           16006
           16004
Binding SID: 4001
```



Encoding Structure

- However, any BGP extension SHOULD work in presence of standard BGP propagation mechanisms (RR, confed, iBGP/eBGP)
- Therefore, the SR TE Policy MUST make use of either:
  - Add-paths
  - A form of "distinguisher"

  in order to distinguish multiple instances of the same policy
- Work in progress…
  - Add a "distinguisher" to the NLRI
  - Add a route-target community based mechanism for advertisement control
  - Report allocated Binding SID to controller (BGP-LS)

SR TE Policy Sub-TLVs

- Weight TLV
  - Encoded before the ERO TLV(s) so to assign a weight to it

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Type                  |      Length                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                              Weight                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- SID TLV
  - Multiple occurrences of the SID TLV are used for expressing a segment list

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Type                      |      Length                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   ST Type      |         Flags               |I|L|N|F|S|C|M|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
//                        SID (32 bits or 128 bits)            //
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
//                        NAI (variable)                       //
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- Binding SID TLV
  - Requires the receiver to bind a SID to the policy

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Type                      |      Length                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Binding SID(optional)                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```