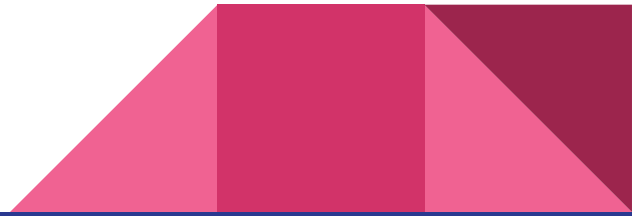
Auto-bandwidth con SR-TE

Dmytro Shypovalov
dmytro@vegvisir.ie

RSVP-TE Auto-bandwidth

- RSVP-TE Puede hacer reserva de ancho de banda por LSP
- También puede medir la utilización del ancho de banda de un LSP y ajustar la reserva
- Por ejemplo: ajustar cada pocas horas
- O ajustarlo tras cambios repentinos (overflow/underflow)



Problemas con RSVP-TE

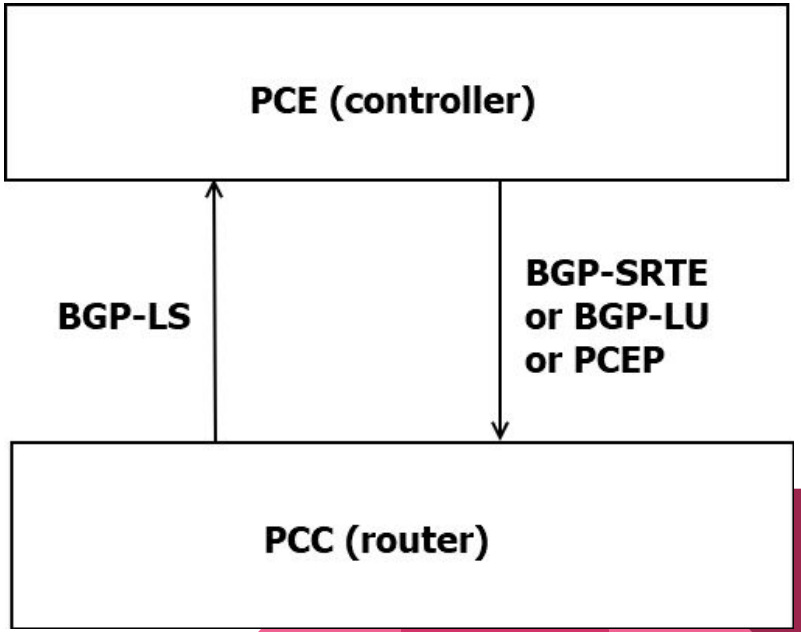
- Complejidad operativa y escalabilidad, poca interoperabilidad entre los distintos fabricantes de enrutadores.
- Sin ruteo determinístico
- Auto-bandwidth no puede reaccionar adecuadamente por los cambios en los destinos de BGP
- No existe un estándar para asignar diferentes servicios en el mismo PE a diferentes LSP
- Añadir nuevos enrutadores requiere configuración adicional en enrutadores existentes. (Completar topología de LSP al nuevo destino).

Ingeniería de tráfico con Segment Routing

- SR es más sencillo de implementar que RSVP y ofrece una excelente interoperabilidad.
- SR también es más escalable que RSVP
- Mapeo del tráfico de un LSP es más sencillo agregando comunidades extendidas (Color).
- Una salvedad: Las reservas de ancho de banda en SR requieren un controlador (PCE)
    - Esto se debe a que SR es stateless, por lo un enrutador P/transito no tiene conocimiento de los LSP que transitan a través de él
- Al utilizar un PCE con una base de datos de ancho de banda centralizada, es posible asignar restricciones de ancho de banda estáticas a los LSP de SR
- ¿Qué hay sobre el auto-bandwidth?

Conceptos básicos de SR-TE PCE

- El PCE recibe la topología de red por medio de BGP-LS
- PCE calcula las políticas con la información obtenida
- El PCE puede enviar las políticas a través de cualquiera de los siguientes protocolos:
  - **BGP-SRTE (RFC9830)**: Mejor opción
  - **BGP-LU (RFC8277):** Enrutadores Legacy
  - **PCEP** (alrededor de 50 RFC, aún sin estandarizar adecuadamente)
  - **GNMI/Netconf/cualquier API** específica del proveedor



SR-TE auto-bandwidth: enfoque de la IETF

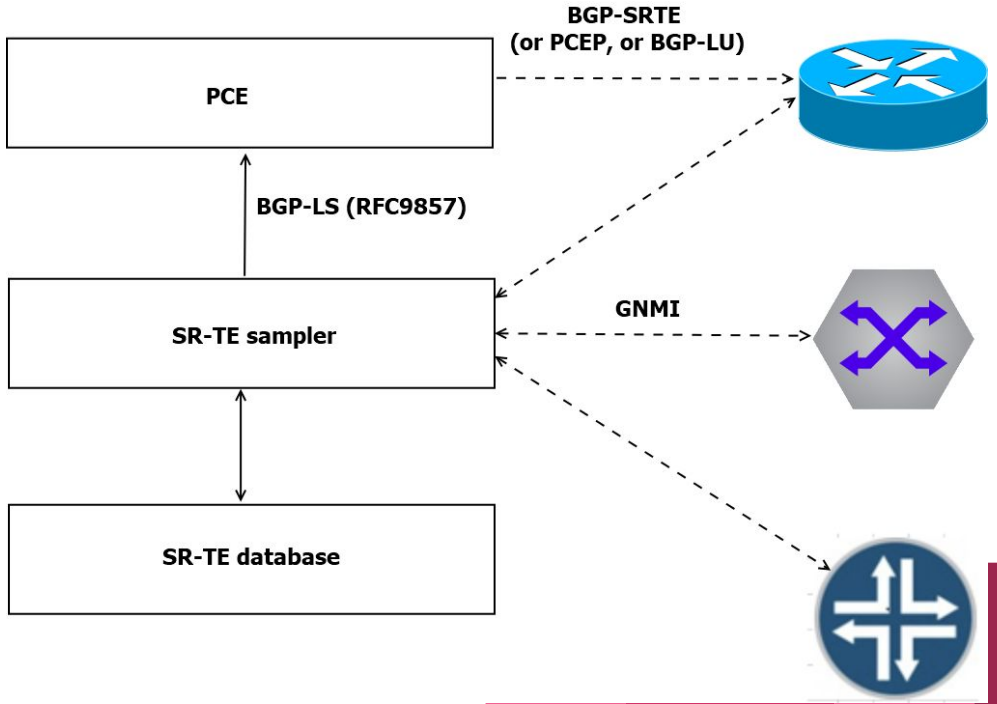
- **RFC 8773:** PCEP extensions for MPLS-TE LSP Auto-Bandwidth adjustment
- Similar al RSVP-TE auto-bandwidth, pero con un PCE
  - El PCC mide las tasas de tráfico en el LSP y comunica la utilización al PCE
- Se requiere una sesión PCEP entre el PCE y cada PCC
- El estado de la implementación/estándares no está claro
- En general, PCEP no interopera bien debido a la complejidad en la escalabilidad

SR-TE auto-bandwidth: enfoque practico

- **RFC 9857:** Advertising Segment Routing policies using BGP-LS
- SR-TE utiliza BGP-LS para obtener las base de datos de los enrutadores, lo cual mejora la visibilidad de los LSP en SR
- Muchas implementaciones ya son compatibles con este RFC
- El PCC or PCE puede anunciar los LSP
    - SR bandwidth constraint TLV (sección 5.6.3) se puede utilizar para anunciar el ancho de banda actual en cada LSP
- La mayoría de las implementaciones de SR tienen contadores por cada LSP
- En la práctica todo lo que necesitamos es: Contadores + RFC9857

Diseño del SR-TE sampler

- El muestreo se obtiene a través de GNMI
- Con los datos se actualiza el ancho de banda para cada LSP y lo comunica al PCE a través de BGP-LS
- El PCE ajusta la reserva de ancho de banda
- Actualmente funciona con Cisco, Arista y Juniper
- Si un enrutador puede anunciar el ancho de banda de LSP según RFC9857, no se necesitan muestras, pero sigue siendo útil disponer de una base de datos de ancho de banda centralizada



Configuración y salidas del muestreo

```
router bgp 65001
   router-id 100.2.2.2
   !
   neighbor 10.10.10.202
      remote-as 65001
   !
   neighbor 192.168.102.102
      remote-as 65002
      ebgp-multihop 10
!
sampling options
   sampling interval 10
   adjust interval 60
   adjust threshold 10
!
telemetry profiles
   !
   profile EOS_PROFILE
      os eos
      port 6030
      auth password
      username admin
      password admin
!
telemetry clients
   !
   group EOS_CLIENTS
      profile EOS_PROFILE
      client 192.168.102.107
      client 192.168.102.108
```

```
lmk-vm103-dev-bw-sampler#show sampling policies
Sampling policies information
Number of policies: 12, active 12, stale 0
Status codes: ~ stale

    Policy                                          Rate                 Last updated
    [2.2.2.2][1.1.1.1][101]                    40.564 Gbps                      0:00:08
    [2.2.2.2][7.7.7.7][101]                    40.408 Gbps                      0:00:08
    [2.2.2.2][8.8.8.8][101]                    40.345 Gbps                      0:00:08
    [1.1.1.1][2.2.2.2][101]                    39.924 Gbps                      0:00:08
    [1.1.1.1][7.7.7.7][101]                    39.297 Gbps                      0:00:08
    [1.1.1.1][8.8.8.8][101]                    39.191 Gbps                      0:00:08
    [7.7.7.7][1.1.1.1][101]                    40.391 Gbps                      0:00:08
    [7.7.7.7][2.2.2.2][101]                    39.368 Gbps                      0:00:08
    [7.7.7.7][8.8.8.8][101]                    40.281 Gbps                      0:00:08
    [8.8.8.8][1.1.1.1][101]                    39.950 Gbps                      0:00:08
    [8.8.8.8][2.2.2.2][101]                    39.947 Gbps                      0:00:08
    [8.8.8.8][7.7.7.7][101]                    39.841 Gbps                      0:00:08

lmk-vm103-dev-bw-sampler#show sampling policies [1.1.1.1][2.2.2.2][101]
Detailed sampling policies information
Number of policies: 1, active 1, stale 0

Sampled policy entry for [1.1.1.1][2.2.2.2][101]
  Router-id: 1.1.1.1
  Endpoint: 2.2.2.2
  Color: 101
  Rate 39.924 Gbps, calculated from 6 samples within 50.4 seconds
  Last updated: 0:00:49 ago
```

Cálculo de auto-bandwidth LSP en PCE

- El PCE con soporte para RFC9857 asigna las rutas BGP-LS recibidas a LSP de SR-TE, utilizando los valores ***(router-id, endpoint, color)***
- De esta forma, conseguimos retroalimentación entre la utilización del ancho de banda y la reserva
- Para esta demostración, he implementado la compatibilidad con RFC9857 en Traffic Dictator
- En la implementación actual, el intervalo de ajuste y el umbral están controlados por el muestreo (sampler); sin embargo, es posible añadir restricciones adicionales en el PCE
- De esta forma, auto-bandwidth funciona con SR-TE
    - Pero no es suficiente
    - Simplemente hemos replicado el comportamiento de RSVP-TE de hace 25 años (con mejor escalabilidad y compatibilidad con diferentes fabricantes de enrutadores)
    - Sin embargo, esto es solo un paso hacia la excelencia

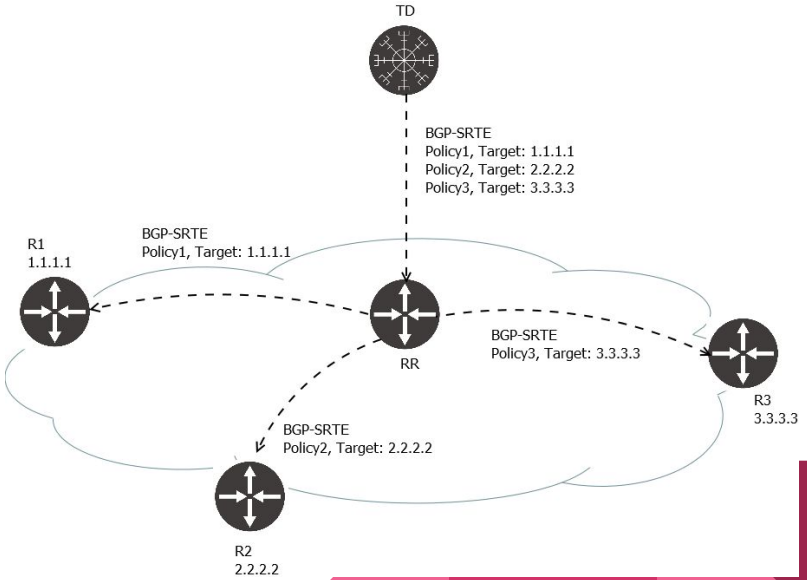
SR-TE mesh-templates: la base para un TE escalable

- Aunque el **control plane** de SR-TE se escala mejor que el de RSVP-TE, el **management plane** es el mismo
  - Aún hay que configurar muchos LSP
  - Al añadir un nuevo enrutador, hay que configurar un nuevo LSP en cada enrutador existente.
  - Si se utiliza un controlador (PCE) para aprovisionar LSP con PCEP, se requiere una sesión PCEP con cada router
- **Mesh-templates:** Genera muchos LSP similares con poca configuración
- Los Mesh-templates son totalmente dinámicos, añaden o eliminan LSP automáticamente cuando se añaden o eliminan enrutadores
- Se utiliza BGP-SRTE para anunciar los LSP generados de mesh-templates, solo se requiere una sesión con el RR (O dos en el caso de tener redundancia)

SR-TE mesh templates: ejemplo

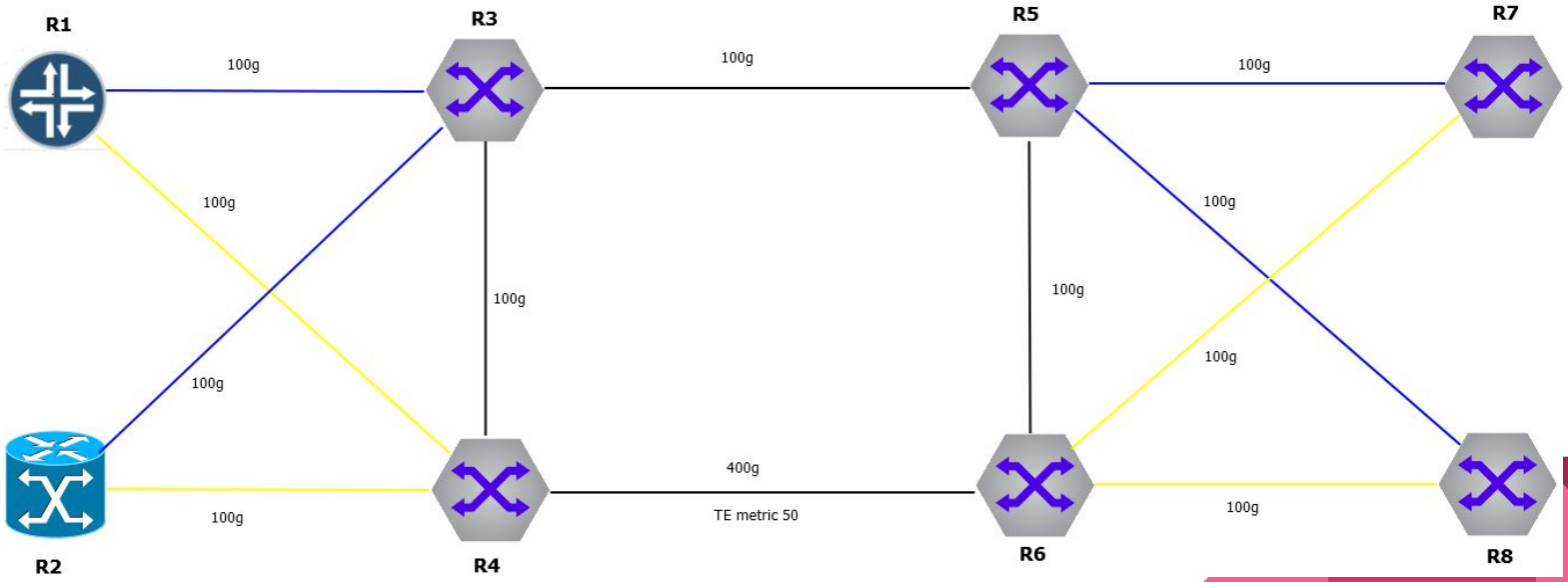
- Esta configuración genera un full mesh de LSP SR-TE en una topología con el ID de BGP-LS 101
- Cada LSP tendrá el color 11101
- La restricción es affinity-set **CORE_LINKS**
- Todos los LSP se anuncian a RR con BGP-SRTE
- RR anuncia el LSP a los routers pertinentes

```
traffic-eng mesh-templates
   !
   template TOPO_101_BLUE
      topology-id 101
      color 11101
      install indirect srte peer-group TOPO_101_RR
      !
      candidate-path preference 100
         affinity-set CORE_LINKS
```

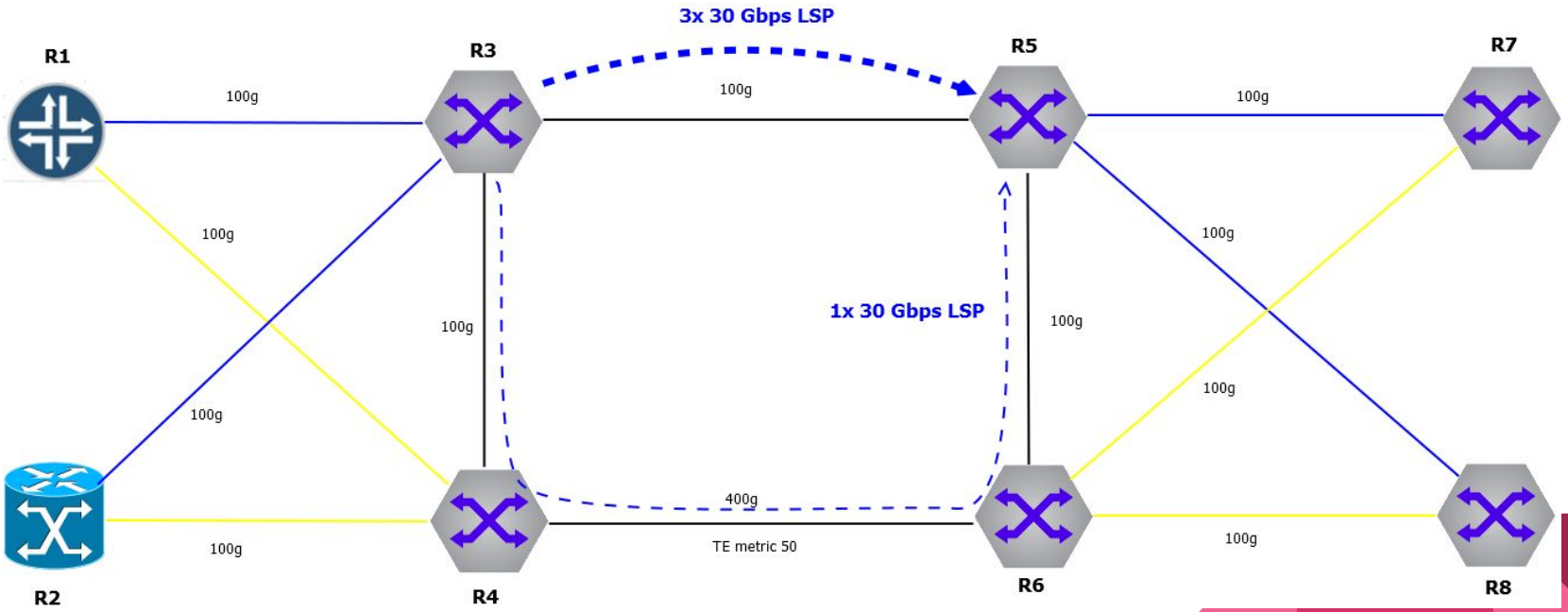


SR-TE mesh templates + auto-bandwidth

- Necesitamos un mesh de LSP entre PE (R1, R2, R7, R8)
- El enlace central R3-R5 tiene menor latencia que R4-R6, pero tiene capacidad limitada



Escenario A: 1 (un) mesh; cada LSP tiene unos 30 Gbps de tráfico



Escenario A: configuración y resultados

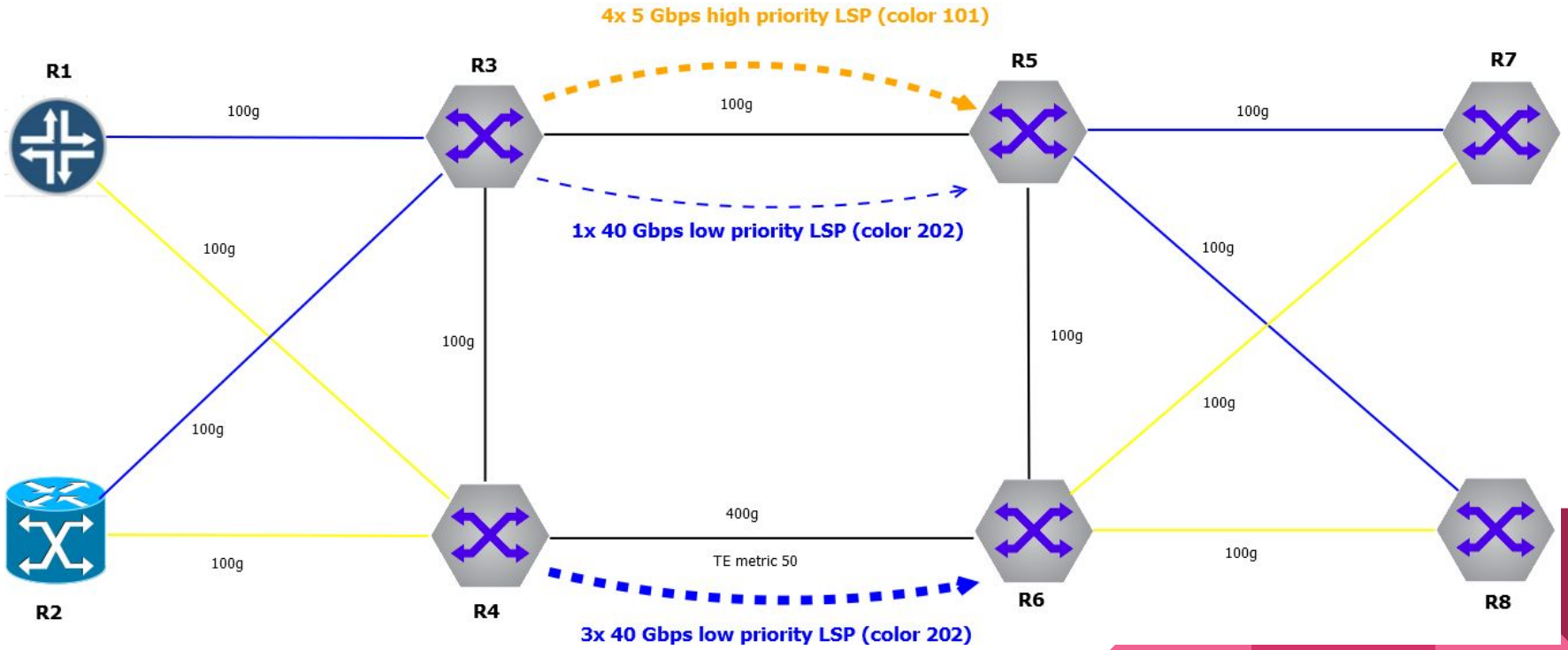
```
traffic-eng capacity-profiles
   !
   profile AUTO_BW
      bandwidth auto
!
traffic-eng mesh-templates
   !
   template TOPO101_BLUE_AUTOBW
      topology-id 101
      color 101
      access-list ipv4 ALL_PE_IPV4
      install indirect srte peer-group RR
      !
      candidate-path preference 100
         metric te
         affinity-set BLUE_ONLY
         capacity-profile AUTO_BW
```

```
lmk-vm102-dev-td1#show traffic-eng mesh-template policies mesh_h2.2.2.2_e8.8.8.8_c101 detail
Detailed traffic-eng policy information:

Traffic engineering policy "mesh_h2.2.2.2_e8.8.8.8_c101"

    Generated from mesh-template TOPO101_BLUE_AUTOBW
--------------snip------------
    Bandwidth type: auto
    Reserved bandwidth: 29.783 Gbps
--------------snip------------
    Candidate paths:
        Candidate-path preference 100
--------------snip------------
            This path is currently active

    Calculation results:
        Aggregate metric: 90
        Topologies: ['101']
        Segment lists:
            [900003, 900004, 900006, 900005, 900008]
```

```
BGP routing table entry for
[SP][SR][I0][N[c65001][b0][q2.2.2.2][2.2.2.2]][C[po2][f0][e8.8.8.8][c101][as650
01][oa2.2.2.2][di100]]
NLRI Type: sr_policy
Protocol: None
Identifier: 0
Local Node Descriptor:
      AS Number: 65001
      BGP Identifier: 0.0.0.0
      BGP Router Identifier: 2.2.2.2
      TE Router Identifier: 2.2.2.2
SRTE Policy CP Descriptor:
      Protocol origin: SR Policy
      Flags: 0
      Endpoint: 8.8.8.8
      Color: 101
      AS Number: 65001
      Originator Address: 2.2.2.2
      Discriminator: 100
Paths: 2 available, best #1
  Last modified: January 30, 2026 18:25:02
  Local
    10.10.10.203 from 10.10.10.203 (100.2.2.2)
      Origin igp, metric 0, localpref 100, weight 0, valid, internal, best
      Link-state: SRTE Bandwidth rate bps 29783246848
```

```
lmk-vm102-dev-td1#show traffic-eng mesh-template policies
Traffic-eng policy information per mesh-template
-----
Mesh-template: TOPO101_BLUE_AUTOBW
Status codes: * valid, > active, s - admin down
       Policy name                          Headend          Endpoint         Color            Protocol          Reserved bandwidth      Priority   Status/Reason
    *>  mesh_h8.8.8.8_e7.7.7.7_c101         8.8.8.8          7.7.7.7          101              SR-TE/indirect    30.167 Gbps(auto)       7/7        Active/Installed
    *>  mesh_h7.7.7.7_e8.8.8.8_c101         7.7.7.7          8.8.8.8          101              SR-TE/indirect    30.161 Gbps(auto)       7/7        Active/Installed
    *>  mesh_h1.1.1.1_e8.8.8.8_c101         1.1.1.1          8.8.8.8          101              SR-TE/indirect    31.015 Gbps(auto)       7/7        Active/Installed
    *>  mesh_h8.8.8.8_e2.2.2.2_c101         8.8.8.8          2.2.2.2          101              SR-TE/indirect    30.311 Gbps(auto)       7/7        Active/Installed
    *>  mesh_h1.1.1.1_e7.7.7.7_c101         1.1.1.1          7.7.7.7          101              SR-TE/indirect    29.849 Gbps(auto)       7/7        Active/Installed
    *>  mesh_h7.7.7.7_e2.2.2.2_c101         7.7.7.7          2.2.2.2          101              SR-TE/indirect    30.860 Gbps(auto)       7/7        Active/Installed
    *>  mesh_h2.2.2.2_e8.8.8.8_c101         2.2.2.2          8.8.8.8          101              SR-TE/indirect    29.783 Gbps(auto)       7/7        Active/Installed
    *>  mesh_h8.8.8.8_e1.1.1.1_c101         8.8.8.8          1.1.1.1          101              SR-TE/indirect    29.886 Gbps(auto)       7/7        Active/Installed
    *>  mesh_h1.1.1.1_e2.2.2.2_c101         1.1.1.1          2.2.2.2          101              SR-TE/indirect    30.326 Gbps(auto)       7/7        Active/Installed
    *>  mesh_h7.7.7.7_e1.1.1.1_c101         7.7.7.7          1.1.1.1          101              SR-TE/indirect    29.838 Gbps(auto)       7/7        Active/Installed
    *>  mesh_h2.2.2.2_e7.7.7.7_c101         2.2.2.2          7.7.7.7          101              SR-TE/indirect    30.085 Gbps(auto)       7/7        Active/Installed
    *>  mesh_h2.2.2.2_e1.1.1.1_c101         2.2.2.2          1.1.1.1          101              SR-TE/indirect    30.009 Gbps(auto)       7/7        Active/Installed
```

Escenario B: 2 (dos) meshes con prioridades diferentes



Escenario B: configuración y resultados

```
traffic-eng mesh-templates
   !
   template TOPO101_AUTO_BW_HIGH_PRIORITY
      topology-id 101
      color 101
      priority 5 5
      access-list ipv4 ALL_PE_IPV4
      access-list ipv6 DENY_IPV6
      install indirect srte peer-group RR
      !
      candidate-path preference 100
         metric te
         affinity-set BLUE_ONLY
         capacity-profile AUTO_BW
   !
   template TOPO101_AUTO_BW_LOW_PRIORITY
      topology-id 101
      color 202
      access-list ipv4 ALL_PE_IPV4
      access-list ipv6 DENY_IPV6
      install indirect srte peer-group RR
      !
      candidate-path preference 100
         metric te
         capacity-profile AUTO_BW
```

- Un mesh-template con prioridad alta para tráfico sensible a la latencia
- El otro mesh-template tiene una prioridad baja
- El color SR-TE se utiliza para asignar el tráfico de datos a las LSP pertinentes,
- Diferentes servicios en el mismo PE puedan recibir un tratamiento diferente

Escenario B: configuración y resultados (cont.)

```
lmk-vm102-dev-td1#show traffic-eng mesh-template policies
Traffic-eng policy information per mesh-template
-----
Mesh-template: TOPO101_AUTO_BW_HIGH_PRIORITY
Status codes: * valid, > active, s - admin down
       Policy name                          Headend          Endpoint         Color            Protocol         Reserved bandwidth       Priority   Status/Reason
    *>  mesh_h7.7.7.7_e8.8.8.8_c101         7.7.7.7          8.8.8.8          101              SR-TE/indirect   5.086 Gbps(auto)         5/5        Active/Installed
    *>  mesh_h2.2.2.2_e8.8.8.8_c101         2.2.2.2          8.8.8.8          101              SR-TE/indirect   4.989 Gbps(auto)         5/5        Active/Installed
    *>  mesh_h2.2.2.2_e7.7.7.7_c101         2.2.2.2          7.7.7.7          101              SR-TE/indirect   5.038 Gbps(auto)         5/5        Active/Installed
    *>  mesh_h7.7.7.7_e1.1.1.1_c101         7.7.7.7          1.1.1.1          101              SR-TE/indirect   4.999 Gbps(auto)         5/5        Active/Installed
    *>  mesh_h1.1.1.1_e2.2.2.2_c101         1.1.1.1          2.2.2.2          101              SR-TE/indirect   5.002 Gbps(auto)         5/5        Active/Installed
    *>  mesh_h7.7.7.7_e2.2.2.2_c101         7.7.7.7          2.2.2.2          101              SR-TE/indirect   4.894 Gbps(auto)         5/5        Active/Installed
    *>  mesh_h8.8.8.8_e1.1.1.1_c101         8.8.8.8          1.1.1.1          101              SR-TE/indirect   5.081 Gbps(auto)         5/5        Active/Installed
    *>  mesh_h8.8.8.8_e7.7.7.7_c101         8.8.8.8          7.7.7.7          101              SR-TE/indirect   5.083 Gbps(auto)         5/5        Active/Installed
    *>  mesh_h8.8.8.8_e2.2.2.2_c101         8.8.8.8          2.2.2.2          101              SR-TE/indirect   4.949 Gbps(auto)         5/5        Active/Installed
    *>  mesh_h1.1.1.1_e8.8.8.8_c101         1.1.1.1          8.8.8.8          101              SR-TE/indirect   5.056 Gbps(auto)         5/5        Active/Installed
    *>  mesh_h2.2.2.2_e1.1.1.1_c101         2.2.2.2          1.1.1.1          101              SR-TE/indirect   5.101 Gbps(auto)         5/5        Active/Installed
    *>  mesh_h1.1.1.1_e7.7.7.7_c101         1.1.1.1          7.7.7.7          101              SR-TE/indirect   5.003 Gbps(auto)         5/5        Active/Installed
Mesh-template: TOPO101_AUTO_BW_LOW_PRIORITY
Status codes: * valid, > active, s - admin down
       Policy name                          Headend          Endpoint         Color            Protocol         Reserved bandwidth       Priority   Status/Reason
    *>  mesh_h1.1.1.1_e8.8.8.8_c202         1.1.1.1          8.8.8.8          202              SR-TE/indirect   40.954 Gbps(auto)        7/7        Active/Installed
    *>  mesh_h2.2.2.2_e8.8.8.8_c202         2.2.2.2          8.8.8.8          202              SR-TE/indirect   40.049 Gbps(auto)        7/7        Active/Installed
    *>  mesh_h8.8.8.8_e2.2.2.2_c202         8.8.8.8          2.2.2.2          202              SR-TE/indirect   39.991 Gbps(auto)        7/7        Active/Installed
    *>  mesh_h1.1.1.1_e7.7.7.7_c202         1.1.1.1          7.7.7.7          202              SR-TE/indirect   40.873 Gbps(auto)        7/7        Active/Installed
    *>  mesh_h1.1.1.1_e2.2.2.2_c202         1.1.1.1          2.2.2.2          202              SR-TE/indirect   40.403 Gbps(auto)        7/7        Active/Installed
    *>  mesh_h8.8.8.8_e7.7.7.7_c202         8.8.8.8          7.7.7.7          202              SR-TE/indirect   39.728 Gbps(auto)        7/7        Active/Installed
    *>  mesh_h7.7.7.7_e8.8.8.8_c202         7.7.7.7          8.8.8.8          202              SR-TE/indirect   40.380 Gbps(auto)        7/7        Active/Installed
    *>  mesh_h7.7.7.7_e1.1.1.1_c202         7.7.7.7          1.1.1.1          202              SR-TE/indirect   40.239 Gbps(auto)        7/7        Active/Installed
    *>  mesh_h2.2.2.2_e1.1.1.1_c202         2.2.2.2          1.1.1.1          202              SR-TE/indirect   39.433 Gbps(auto)        7/7        Active/Installed
    *>  mesh_h2.2.2.2_e7.7.7.7_c202         2.2.2.2          7.7.7.7          202              SR-TE/indirect   39.860 Gbps(auto)        7/7        Active/Installed
    *>  mesh_h7.7.7.7_e2.2.2.2_c202         7.7.7.7          2.2.2.2          202              SR-TE/indirect   39.433 Gbps(auto)        7/7        Active/Installed
    *>  mesh_h8.8.8.8_e1.1.1.1_c202         8.8.8.8          1.1.1.1          202              SR-TE/indirect   38.796 Gbps(auto)        7/7        Active/Installed
```

Redundancia y conmutación por error

- La solución se basa en **estándares abiertos** (BGP-LS, BGP-SRTE), por lo que es robusta, fácil de entender y presenta fallos de forma predecible
- SR-TE sampler no necesita ser redundante
- PCE: despliega tantos como quieras, no se necesita sincronización (piensa en un RR)
- Cada PCE se configura con un **SRTE distinguisher** diferente, por lo que cada router recibe múltiples copias de la misma NLRI de SR-TE desde diferentes PCE
    - Si se retira uno, el router simplemente utiliza otro
- Fallo catastrófico de todos los PCE: el tráfico sigue la ruta más corta o la política local

SR-TE auto-bandwidth: conclusión

- Solución sencilla basada en estándares sin dependencia de un proveedor
- Configuración minimalista y con actualización automática: actívela una vez y olvídese
  - **Intent-based networking hecho correctamente**
- Se garantiza que el estado de enrutamiento en la red sea determinista (a diferencia de RSVP-TE)
- Podemos definir diferentes SLA de TE para diferentes servicios (por ejemplo, algunos servicios tienen prioridad)
- BGP-SRTE se escala extremadamente bien; implementa tantos LSP como quieras
- SR-TE sampler es open source https://github.com/Vegvisir-Systems/srte-sampler/
- Traffic Dictator no es open source, pero hay una versión gratuita disponible para su evaluación y estudio

SR-TE auto-bandwidth: futuro

- Por ahora se trata de un prototipo; SR-TE sampler v0.1 es experimental
- Es posible implementar LSP splitting/merging (TE++/container LSP)
- Gracias a la vista centralizada, podemos reaccionar a los cambios en los destinos de BGP
- Es posible parar o revertir reservaciones de ancho de banda para hacer troubleshooting
- Es posible ejecutar diversos análisis
    - Detectar patrones a largo plazo, mostrar la utilización general de la red y recomendar la actualización de determinados enlaces
    - Ofrecer recomendaciones para añadir un nuevo enlace, por ejemplo, si hay mucho tráfico entre los routers A y B pero no están conectados directamente

¿Alguna pregunta?