# 分段路由(Segment Routing): 大规模SDN部署必备技术

思科大中华区电信运营商事业部首席技术官 兼 思科首席工程师

苏远超

2016年12月

# Cisco VNI:全球IP流量预测
## by 2020

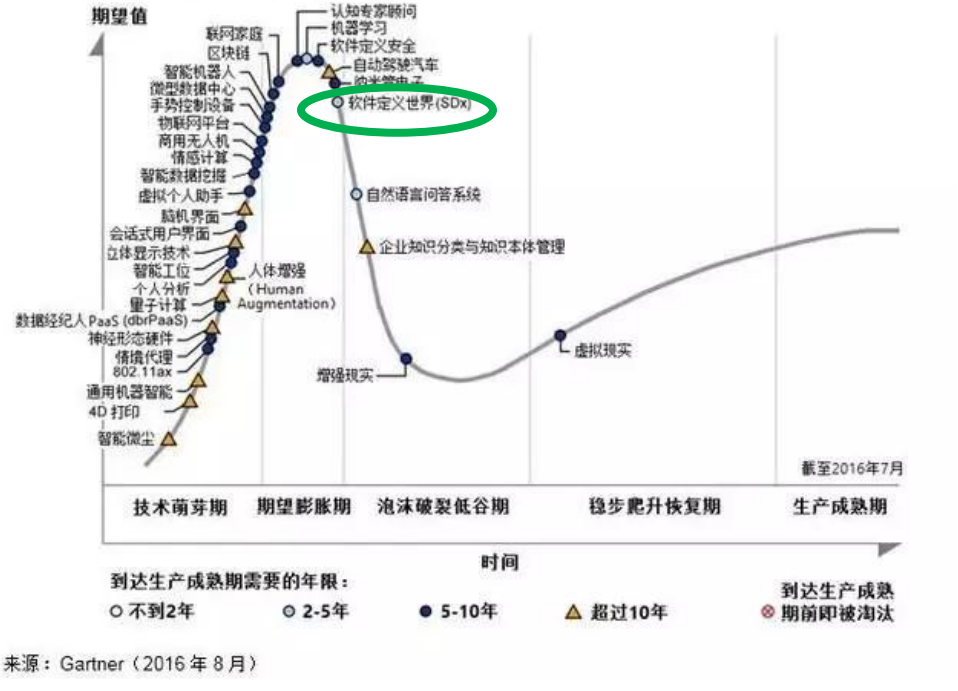
| | |
|---|---|
| Global Internet Users | **4.1 Billion** global internet users, representing **52%** of the global population |
| Global Devices/Connections | **3.4 devices/connections** per capita globally |
| Global IP Video Traffic | **82% of the world's IP traffic will be video**<br>**21% of IP video traffic to be 4K by 2019** |
| Global Mobile Traffic | **16%** of IP traffic will be carried over cell networks |
| Global Wi-Fi Traffic | Fixed Wi-Fi will generate **50%** of global IP traffic |
| Global IP Traffic | Global IP traffic will reach **194 EBs**/per month (**2.3 ZBs** annually) |

Source: Cisco Visual Networking Index Global IP Traffic Forecast, 2015–2020

# SDN未来5年将是”痛并快乐”阶段

## 性能及与应用的协同是关键

图一、2016 年新兴科技技术成熟度曲线




来源：Gartner（2016 年 8 月）

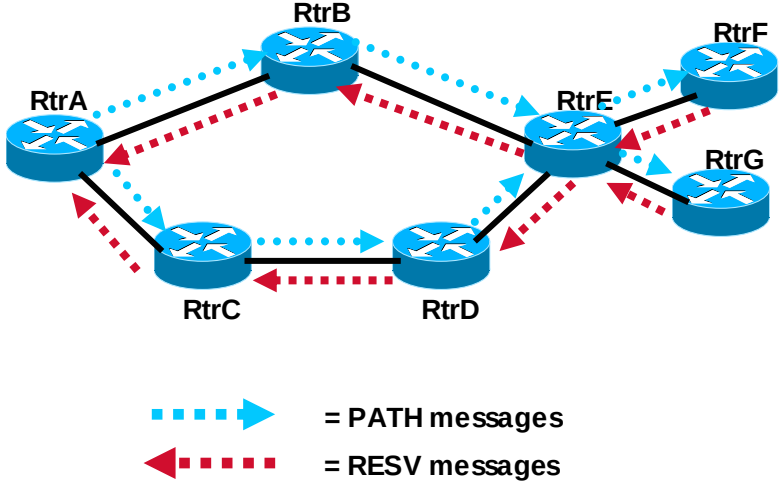
# Agenda

- Segment Routing(SR)解决了什么问题?
- SR原理
- SR典型应用场景
- 进阶话题
- 案例分享
- 思科SR解决方案

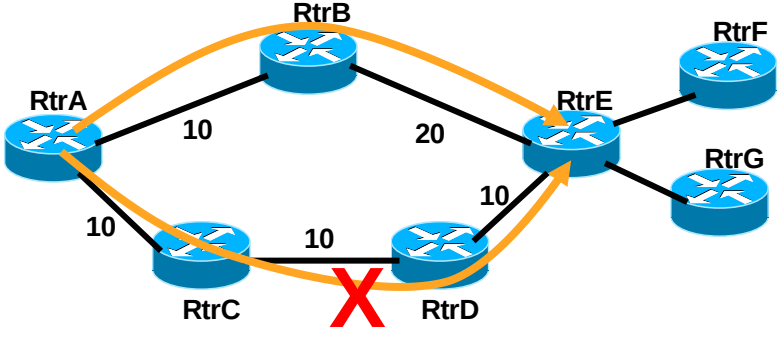
# Segment Routing(SR)解决了什么问题?

# RSVP-TE之痛

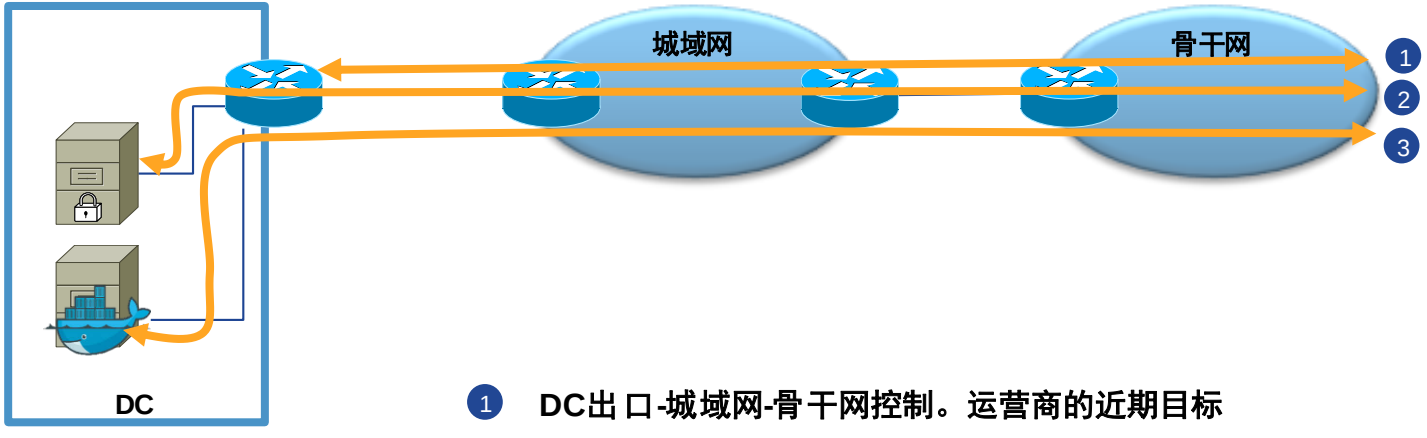
网络中每个节点均需维护大量的路径状态信息,可扩展性差




不支持ECMP,造成资源利用率低



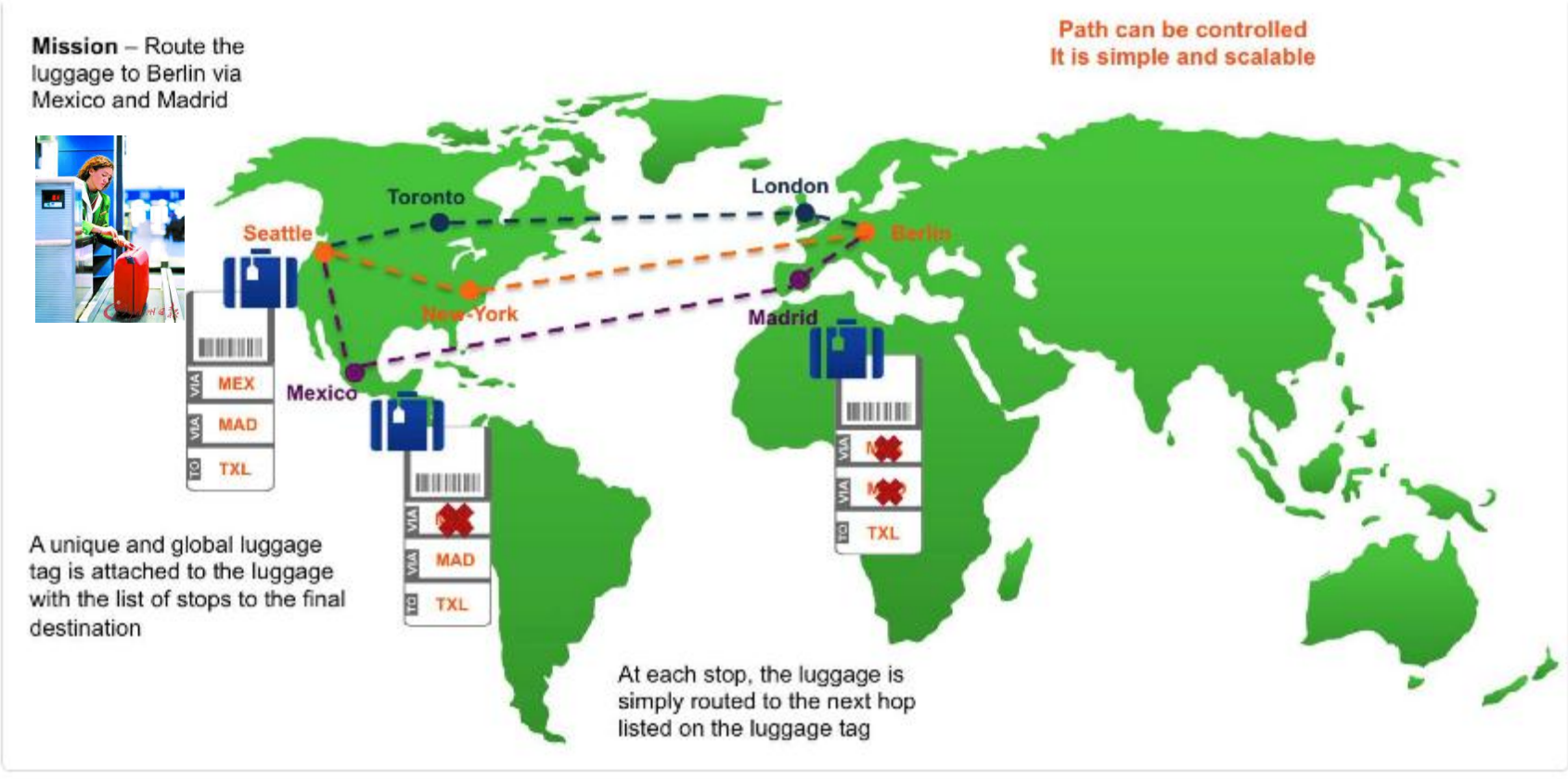
# 应用和网络还是离的很远



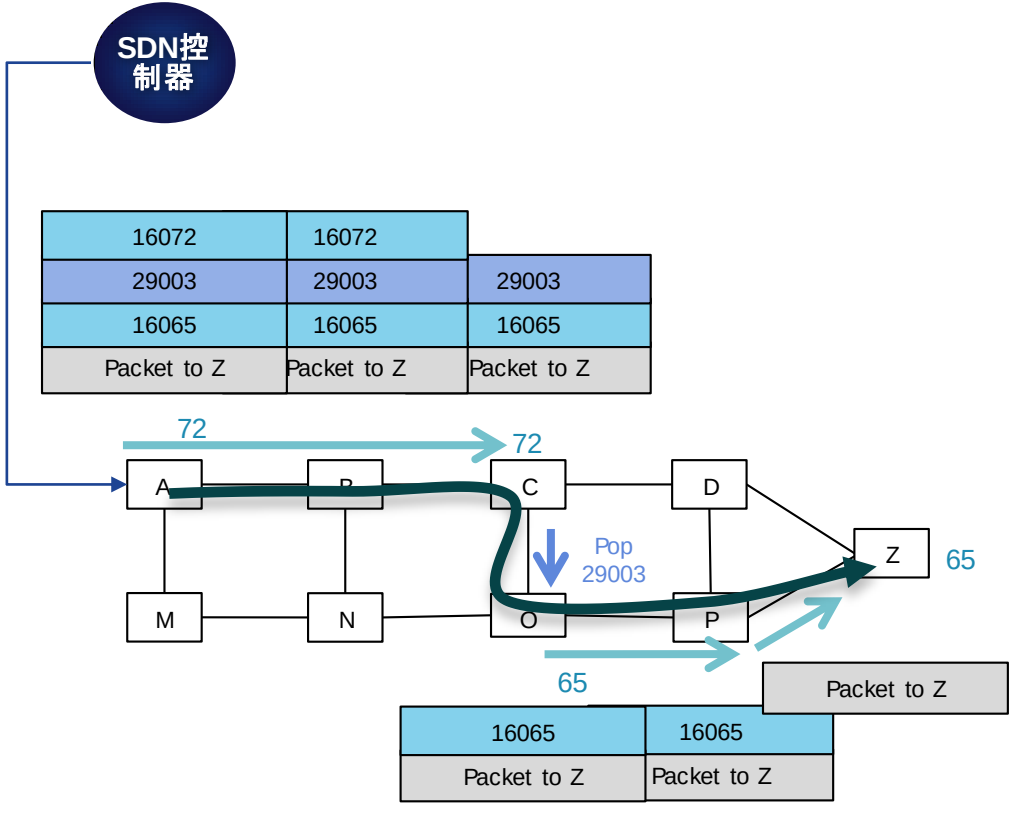

1. DC出口-城域网-骨干网控制。运营商的近期目标
2. 服务器-DC出口-城域网-骨干网控制。个别客户使用
3. Docker-服务器-DC出口-城域网-骨干网控制。概念验证阶段

# SR原理

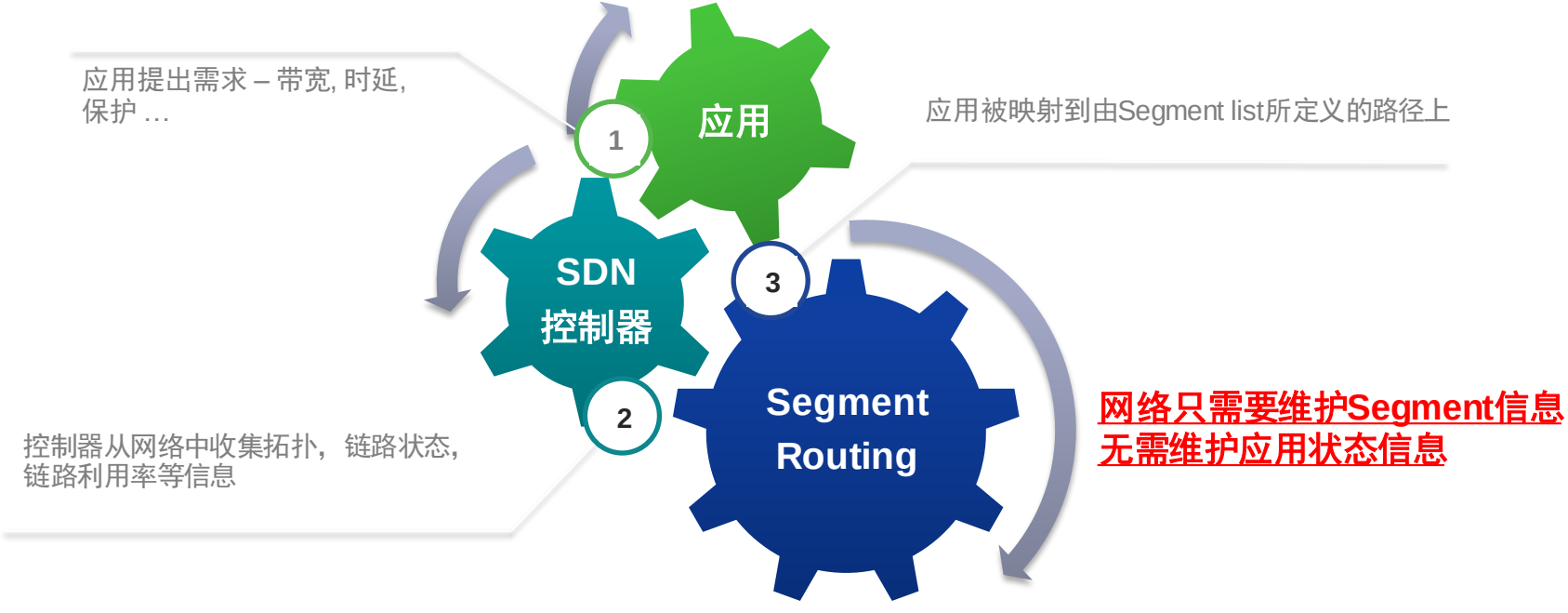
# 行李是如何被托运的...



# Segment Routing是如何转发的

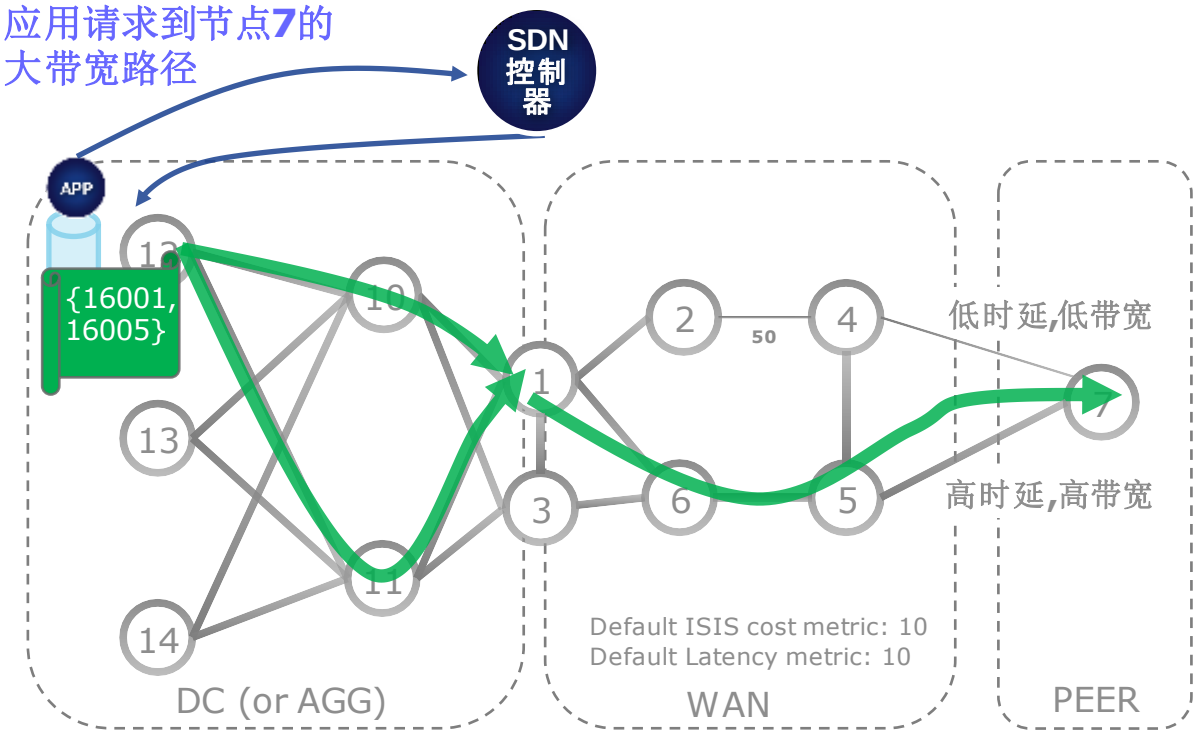


# 应用驱动网络:Segment Routing 简化 & 可扩展



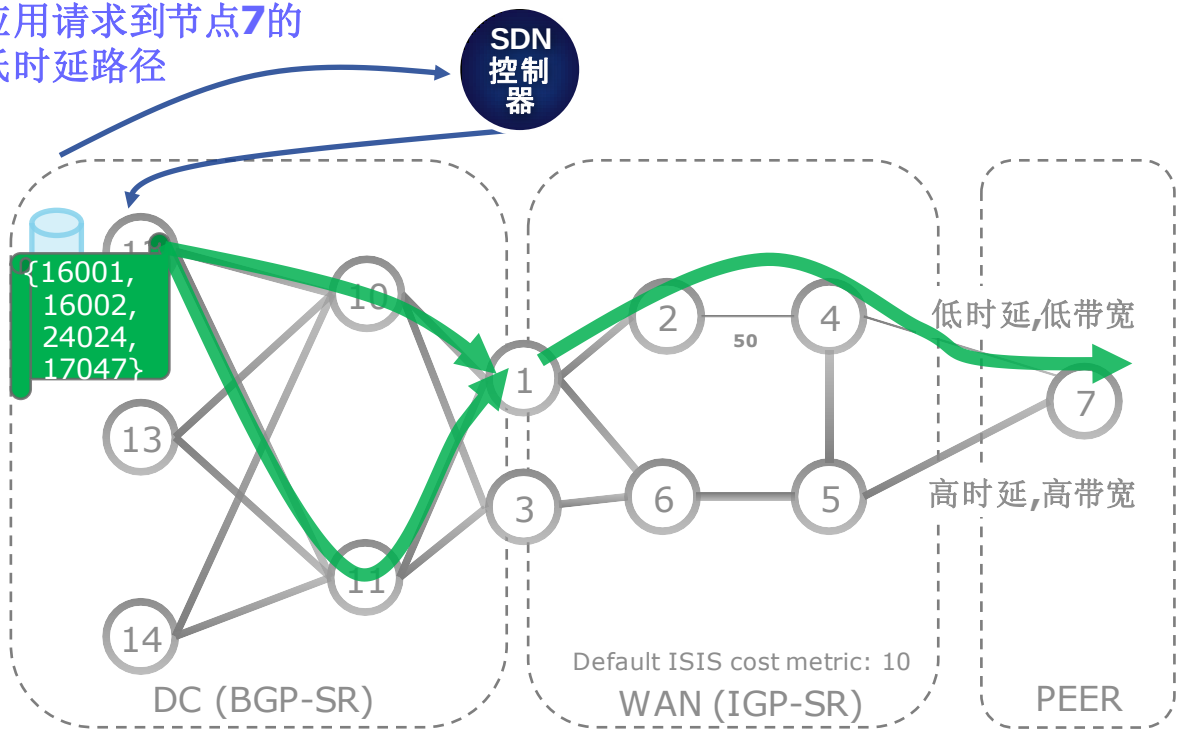
# 应用驱动网络示例1

- Applications program the network on a per-flow basis
- End-to-End policy
  - DC, WAN, AGG, PEER
- Millions of flows
  - No per-flow midpoint state
  - No reclassification at boundaries
- Simple
  - BGP and ISIS/OSPF



# 应用驱动网络示例2

- Controller computes that the green path can be encoded as
  - 16001
  - 16002
  - 24024
  - 17047
- Controller programs a single per-flow state to create an application-engineered end-to-end policy

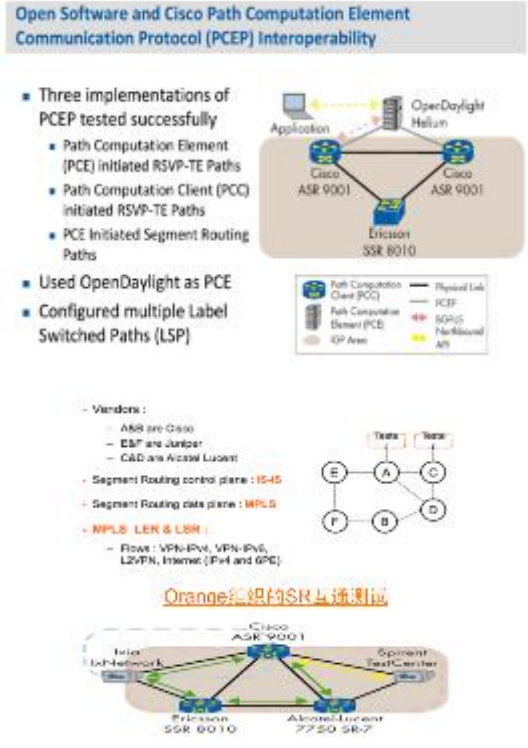


# Segment Routing已经被产业界广泛接受

## Segment Routing IETF draft

- draft-ietf-spring-segment-routing-03 - Segment Routing Architechtur
- draft-ietf-isis-segment-routing-extensions-04 - ISIS extension for segment-routing
- draft-ietf-ospf-segment-routing-extensions-02 - OSPF extension for segment-routing
- draft-ietf-ospf-prefix-link-attr-06 - OSPF extension for segent routing SID
- draft-ietf-idr-bgpls-segment-routing-epe
- draft-ietf-pce-segment-routing - PCEP extension for segment routing
- draft-ietf-pce-lsp-setup-type - PCEP selection ISP type of rsvp-te or SR
- draft-gredler-idr-bgp-ls-segment-routing-extension-02

## EANTC/Orange Inter-op Test




## ODL Supports SR

OPEN DAYLIGHT

BGP LS PCEP:PCEP

Contents [hide]

1 PCEP overall architecture
- 1.1 PCEP
  - 1.1.1 Session handling
  - 1.1.2 Parser
    - 1.1.2.1 Registration
    - 1.1.2.2 Parsing
    - 1.1.2.3 Serializing
- 1.2 PCEP IETF stateful
  - 1.2.1 Configuration
- 1.3 PCEP segment routing
  - 1.3.1 Configuration
- 1.4 PCEP topology
- 1.5 PCEP tunnel

2 Programming overall architecture
- 2.1 Programming
- 2.2 Programming topology
- 2.3 Programming tunnel

# Segment Routing概念

- Source Routing
  - the source chooses a path and encodes it in the packet header as an ordered list of segments
  - the rest of the network executes the encoded instructions
- Segment: an identifier for any type of instruction
  - forwarding or service
- MPLS: an ordered list of segments is represented as a stack of labels
  - SR re-uses MPLS data plane without any change
- IPv6: an ordered list of segments is represented as a routing extension header, see 4.4 of RFC2460
- IGP-based segments require minor extension to the existing link-state routing protocols (OSPF and IS-IS).
- BGP-based segments BGP Egress Peering Engineering(EPE) and BGP-LU

# 全局和本地Segment

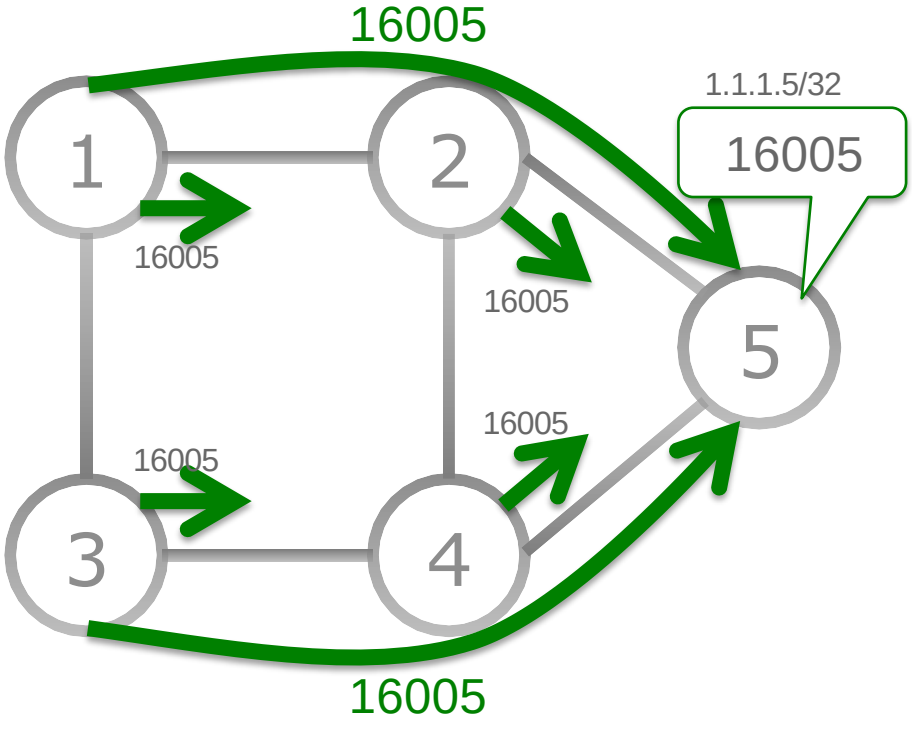
- Global Segment
  - Any node in SR domain understands associated instruction
  - Each node in SR domain installs the associated instruction in its forwarding table
  - MPLS: global label value in Segment Routing Global Block (SRGB)
- Local Segment
  - Only originating node understands associated instruction
  - MPLS: locally allocated label

# 全局Segment–全局标签索引

- Global Segments always distributed as a label range (SRGB) + **Index**
  - **Index** must be unique in Segment Routing Domain
- Best practice: same SRGB on all nodes
  - “Global model”, requested by all operators
  - Global Segments are global label values, simplifying network operations
  - Default SRGB: 16,000 – 23,999
    - Other vendors also use this label range

# IGP Segment之Prefix-SID

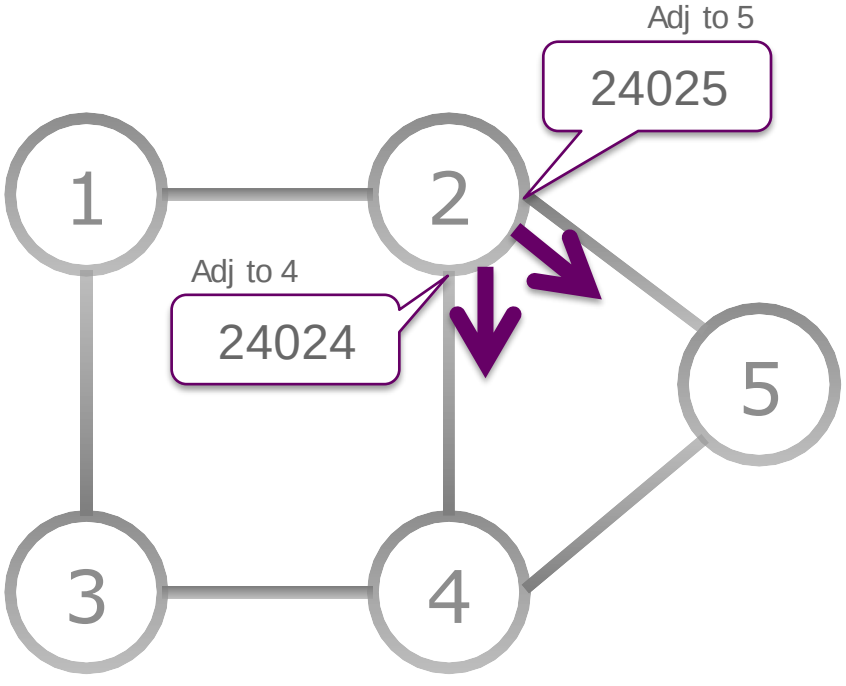
- Shortest-path to the IGP prefix
  - Equal Cost MultiPath (ECMP)-aware
- Global Segment
- Label = 16000 + Index
  - Advertised as index
- Distributed by ISIS/OSPF




All nodes use default SRGB
16,000 – 23,999

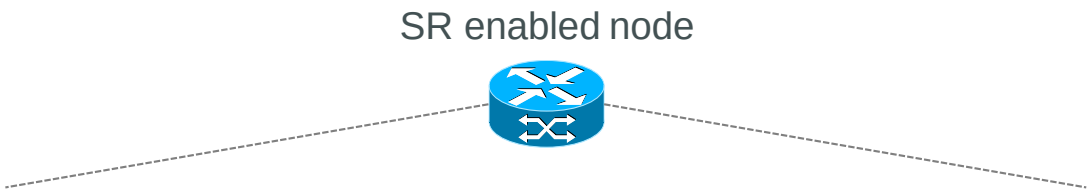
# IGP Segment之Adjacency-SID

- Forward on the IGP adjacency
- Local Segment
- Advertised as label value
- Distributed by ISIS/OSPF




All nodes use default SRGB
16,000 – 23,999

# SID编码



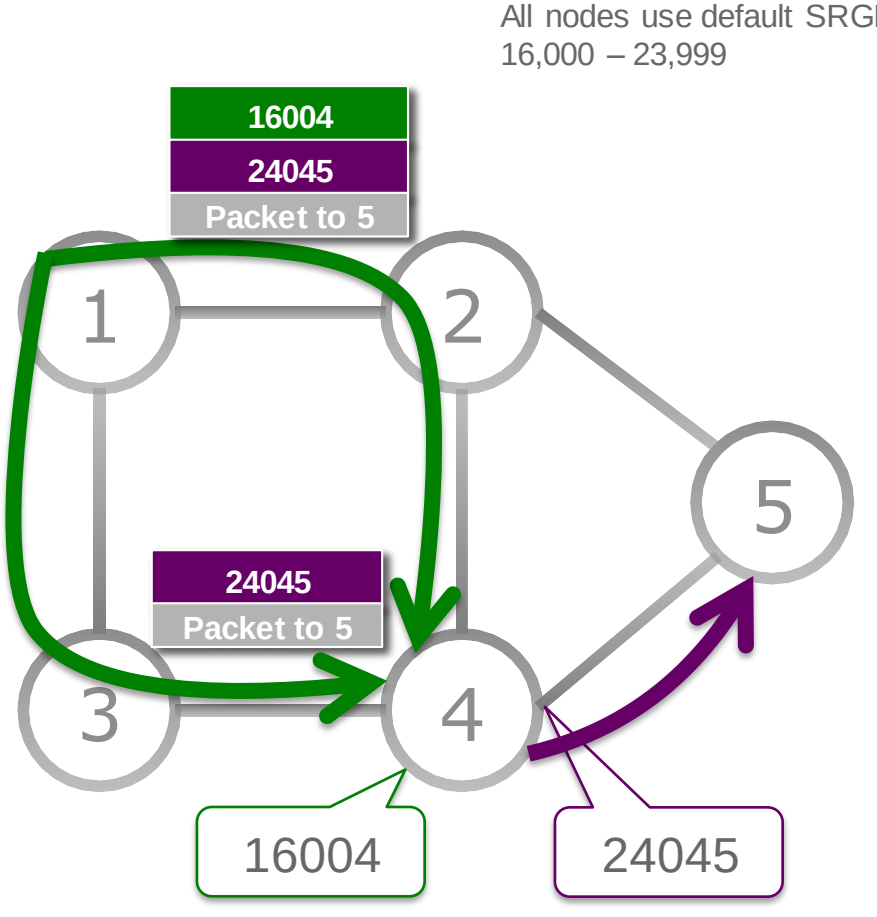

- Prefix SID
  - Uses SR Global Block (SRGB)
  - SRGB advertised with router capabilities TLV
  - In the configuration, Prefix-SID can be configured as an absolute value or an index
  - In the protocol advertisement, Prefix-SID is always encoded as a globally unique index

    Index represents an offset from SRGB base, zero-based numbering, i.e. 0 is 1st index

    E.g. index **1** → SID is 16,000 + **1** = 16,001
- Adjacency SID
  - Locally significant
  - Automatically allocated for each adjacency
  - Always encoded as an absolute (i.e. not indexed) value

# 组合使用IGP Segment

- Steer traffic on any path through the network
- Path is specified by list of segments in packet header, a stack of labels
- No path is signaled
- No per-flow state is created
- Single protocol: IS-IS or OSPF



# SR IS-IS控制平面

- IS-IS Segment Routing functionality
  - IPv4 and IPv6 control plane
  - Level 1, level 2 and multi-level routing
  - Prefix Segment ID (Prefix-SID) for host prefixes on loopback interfaces
  - Adjacency Segment IDs (Adj-SIDs) for adjacencies
    - Non-protected adj-SIDs and protected (since IOS XR 5.3.2) adj-SIDs
      - See SRTE presentation for more information
  - Prefix-to-SID mapping advertisements (mapping server)
  - MPLS penultimate hop popping (PHP) and explicit-null signaling

# IS-IS TLV扩展

- SR for IS-IS introduces support for the following (sub-)TLVs:

| | |
|---|---|
| SR Capability sub-TLV (2) | IS-IS Router Capability TLV (242) |
| Prefix-SID sub-TLV (3) | Extended IP reachability TLV (135) |
| Prefix-SID sub-TLV (3) | IPv6 IP reachability TLV (236) |
| Prefix-SID sub-TLV (3) | Multitopology IPv6 IP reachability TLV (237) |
| Prefix-SID sub-TLV (3) | SID/Label Binding TLV (149) |
| Adjacency-SID sub-TLV (31) | Extended IS Reachability TLV (22) |
| LAN-Adjacency-SID sub-TLV (32) | Extended IS Reachability TLV (22) |
| Adjacency-SID sub-TLV (31) | Multitopology IS Reachability TLV (222) |
| LAN-Adjacency-SID sub-TLV (32) | Multitopology IS Reachability TLV (222) |
| SID/Label Binding TLV (149) | |

- Implementation based on *draft-ietf-isis-segment-routing-extensions-02*

# SR OSPF控制平面

- OSPF Segment Routing functionality
  - OSPFv2 control plane
  - Multi-area
  - IPv4 Prefix Segment ID (Prefix-SID) for host prefixes on loopback interfaces
  - Adjacency Segment ID (Adj-SIDs) for adjacencies
    - Non-protected adj-SIDs and protected (since OSPF SRTE release) adj-SIDs
  - MPLS penultimate hop popping (PHP) and explicit-null signaling

# OSPF扩展

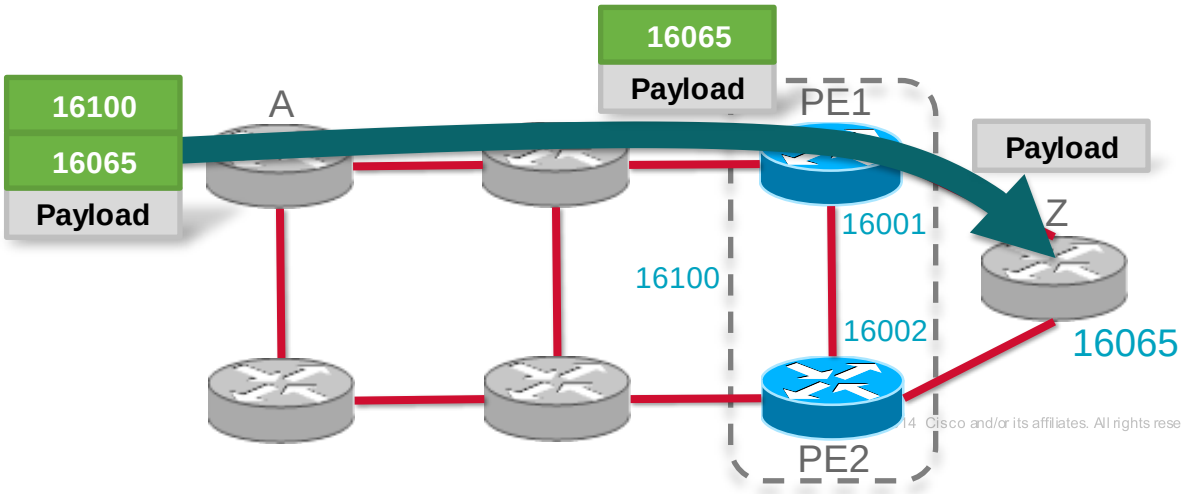
- OSPF adds to the Router Information Opaque LSA (type 4):
  - SR-Algorithm TLV (8)
  - SID/Label Range TLV (9)
- OSPF defines new Opaque LSAs to advertise the SIDs
  - OSPFv2 Extended Prefix Opaque LSA (type 7)
    - OSPFv2 Extended Prefix TLV (1)
      - Prefix SID Sub-TLV (2)
  - OSPFv2 Extended Link Opaque LSA (type 8)
    - OSPFv2 Extended Link TLV (1)
      - Adj-SID Sub-TLV (2)
      - LAN Adj-SID Sub-TLV (3)
- Implementation is based on
  - draft-ietf-ospf-prefix-link-attr-01 and draft-ietf-ospf-segment-routing-extensions-02

# 任播(Anycast)Prefix-SID

- Anycast prefixes: same prefix advertised by multiple nodes
- Anycast prefix-SID: prefix-SID associated with anycast prefix
  - Same prefix-SID for the same prefix!
- Traffic is forwarded to one of the Anycast prefix-SID originators based on best IGP path
- If primary node fails, traffic is auto re-routed to the other node
- Note: nodes advertising the same Anycast prefix-SID must have the same SRGB

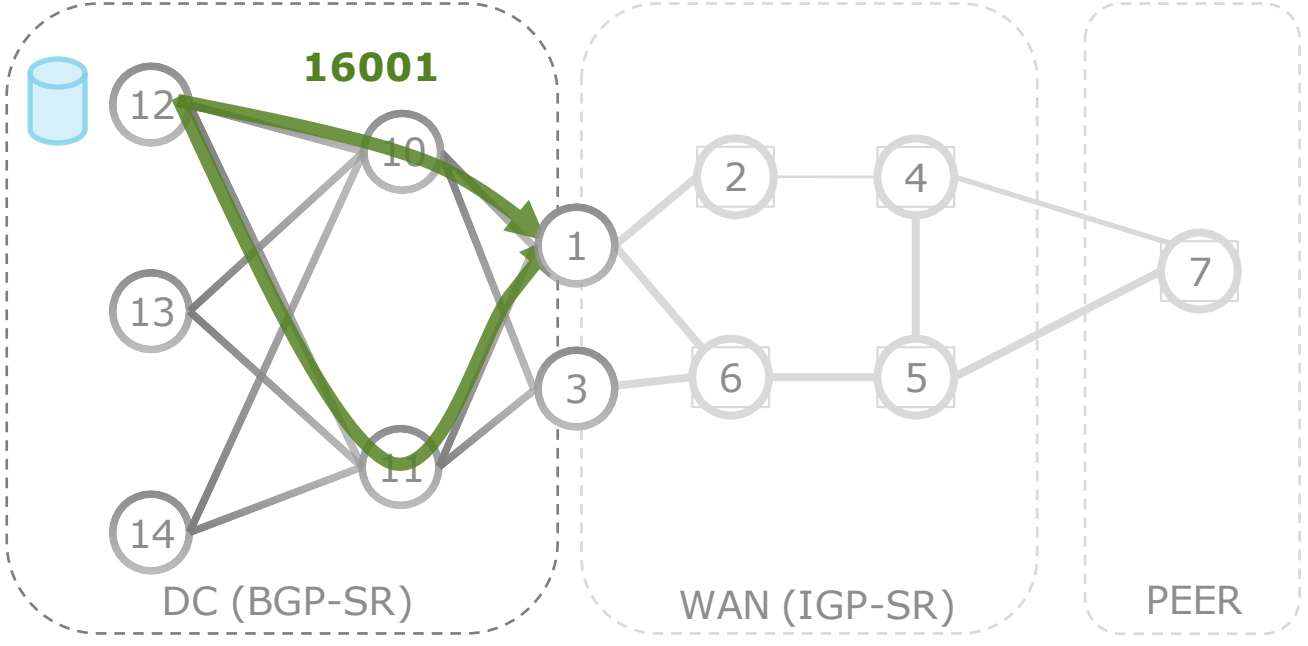
# 任播Prefix-SID的用处

- Coarse Grained Traffic Engineering, steering traffic via groups of routers (with common Anycast-SID)
- High-availability
  - if one of the Eastern routers fail, the policy survives
- Typical for service virtualization
  - nearest firewall/DPI etc.



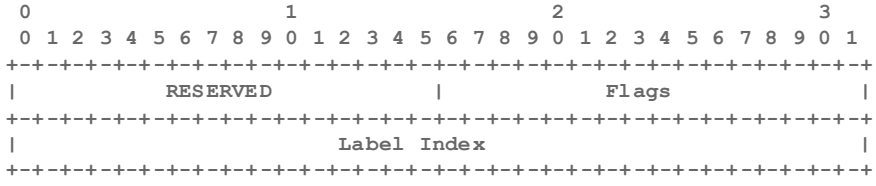
# BGP Segment之Prefix-SID

- Shortest-path to the BGP prefix
- Global
- 16000 + Index
- Signaled by BGP



# BGP Prefix-SID

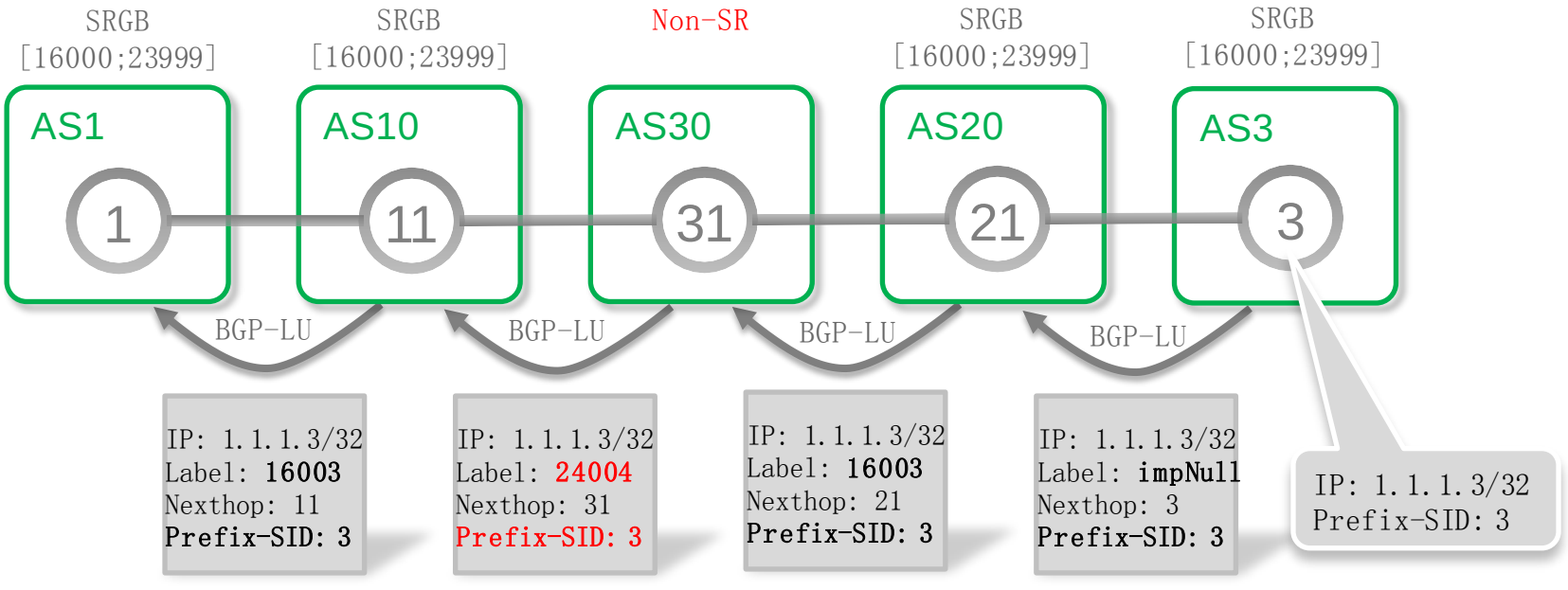
- New attribute (type 40) BGP-Prefix-SID
  - Reserved 2 bytes
  - Flags 2 bytes
  - Label Index 4 bytes
- Example:
  - SAFI: Labeled Unicast
  - NLRI: 1.1.1.3/32
  - Label: 16003
  - Prefix-SID: 3




```
 Update Message (2), length: 75
          Multi-Protocol Reach NLRI (14), length: 17, Flags
[OE]:
            AFI: IPv4 (1), SAFI: labeled Unicast (4)
            nexthop: 99.3.21.3, nh-length: 4, no SNPA
              1.1.1.3/32, label:3 (bottom)
            0x0000:  0001 0404 6303 1503 0038 0000 3101
0101
            0x0010:  03
          Origin (1), length: 1, Flags [T]: IGP
            0x0000:  00
          AS Path (2), length: 6, Flags [T]: 3
            0x0000:  0201 0000 0003
          Multi Exit Discriminator (4), length: 4, Flags
[O]: 0
            0x0000:  0000 0000
          BGP-Prefix-SID (40), leng
            0x0000:  0000 0000 0000
```

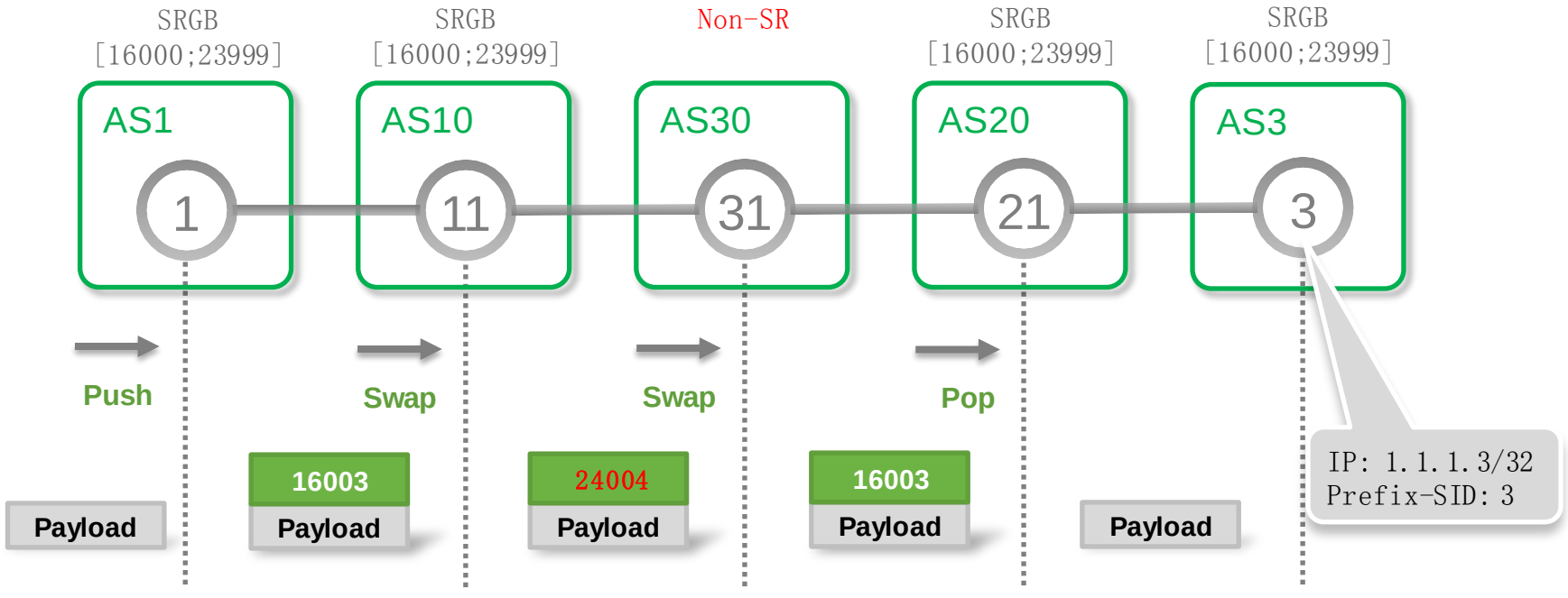
Optional Transitive

# BGP Prefix-SID – 支持与Non-SR设备进行互操作



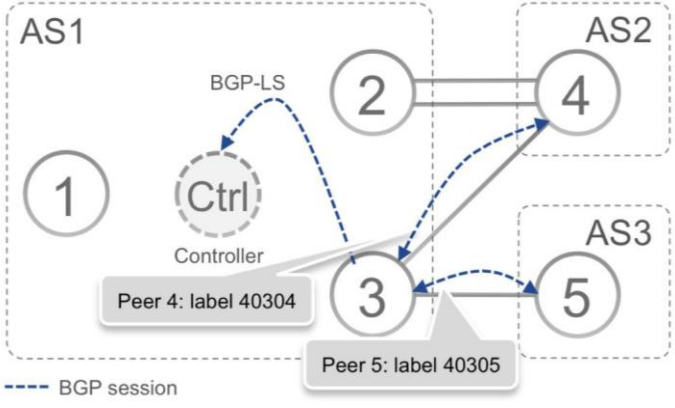

- Node31 is not SR enabled, it will allocate dynamic labels for the 3107 prefixes, while still propagating the BGP-Prefix-SID attribute (Transitive)

# BGP Prefix-SID – 支持与Non-SR设备进行互操作

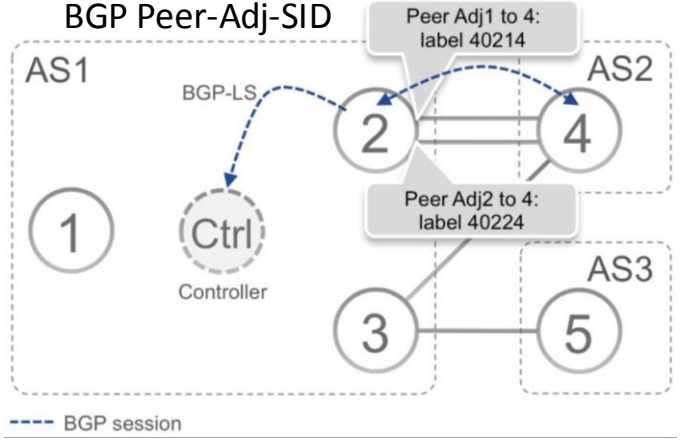


# BGP Segment之EPE(Egress Peer Engineering)

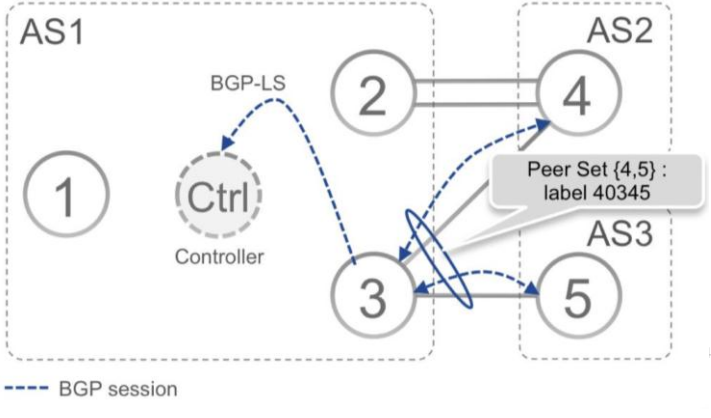
BGP Peer-Node-SID




BGP Peer-Adj-SID




BGP Peer-Set-SID



# SRTE

- No signaling protocol, unlike RSVP-TE
- Traffic steering by pushing a stack of labels (or SRv6 prefix-SIDs)
- Directly benefit from existing Ti-LFA and micro-loop avoidance
- SRTE label/SRv6-SID stack can be signaled from a PCE or configuration
- Constraint SPF
  - Afinity, SRLG-disjoint
  - Static and dynamic path options

# SR典型应用场景

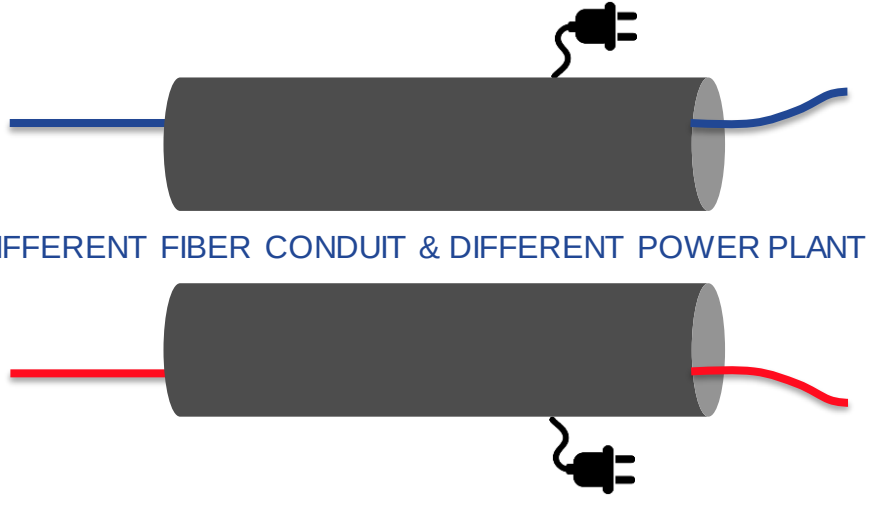
# 应用场景1: 分离路径服务

## Without Segment Routing

SAME FIBER CONDUIT & SAME POWER PLANT

**NO GUARANTEE OF SERVICE**

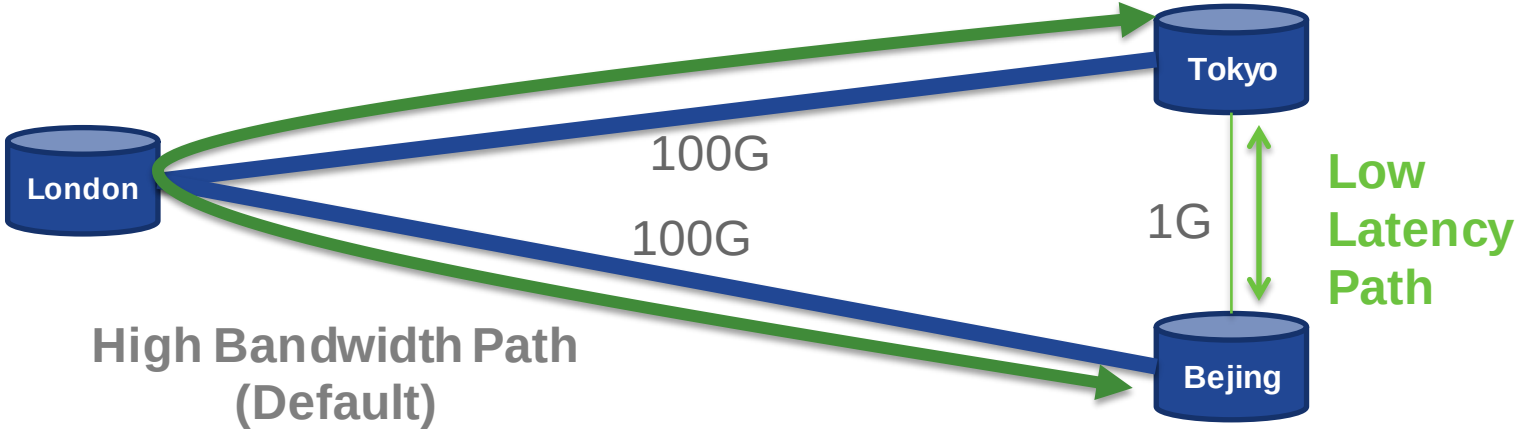
## With Segment Routing



DIFFERENT FIBER CONDUIT & DIFFERENT POWER PLANT

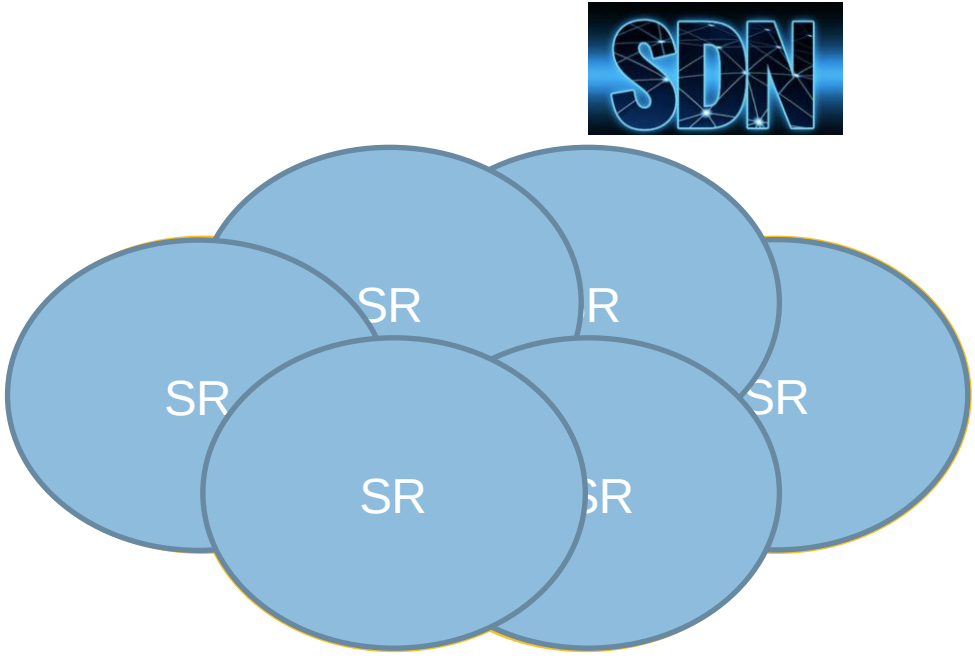
**GUARANTEED SERVICE**

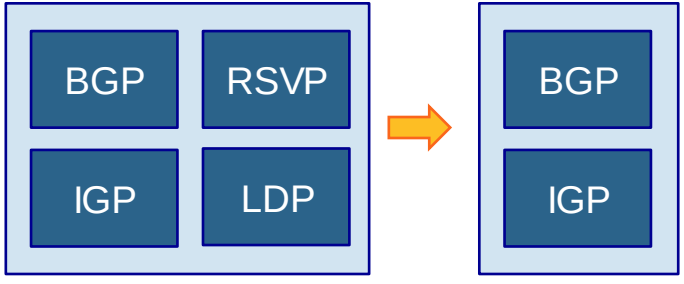
# 应用场景2: 低时延路径服务



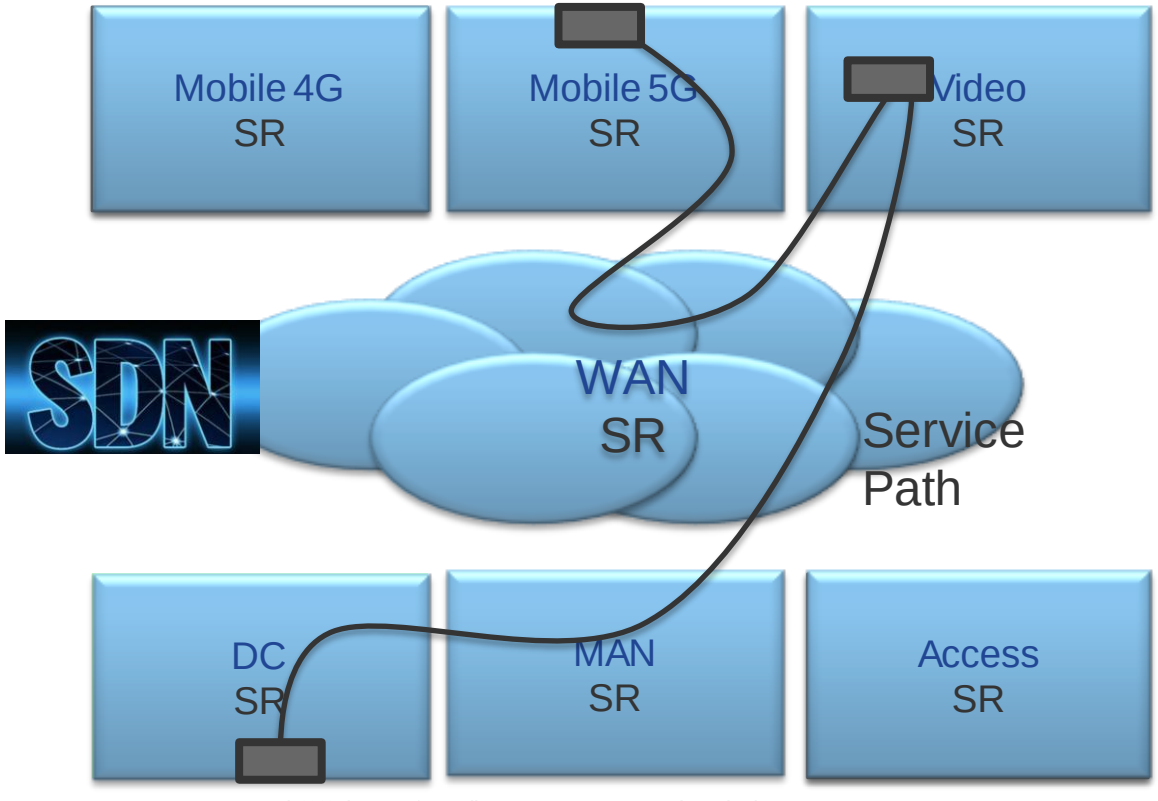
# 应用场景3: 自动50ms保护

# 应用场景4: 传统MPLS网络升级



# 应用场景5: 端到端统一传送平面



# 进阶话题

# SR与LDP的共存/互操作模型

SR+LDP
(Ship in the Night)

SR-Prefer

LDP to SR

不需要配置

SR to LDP

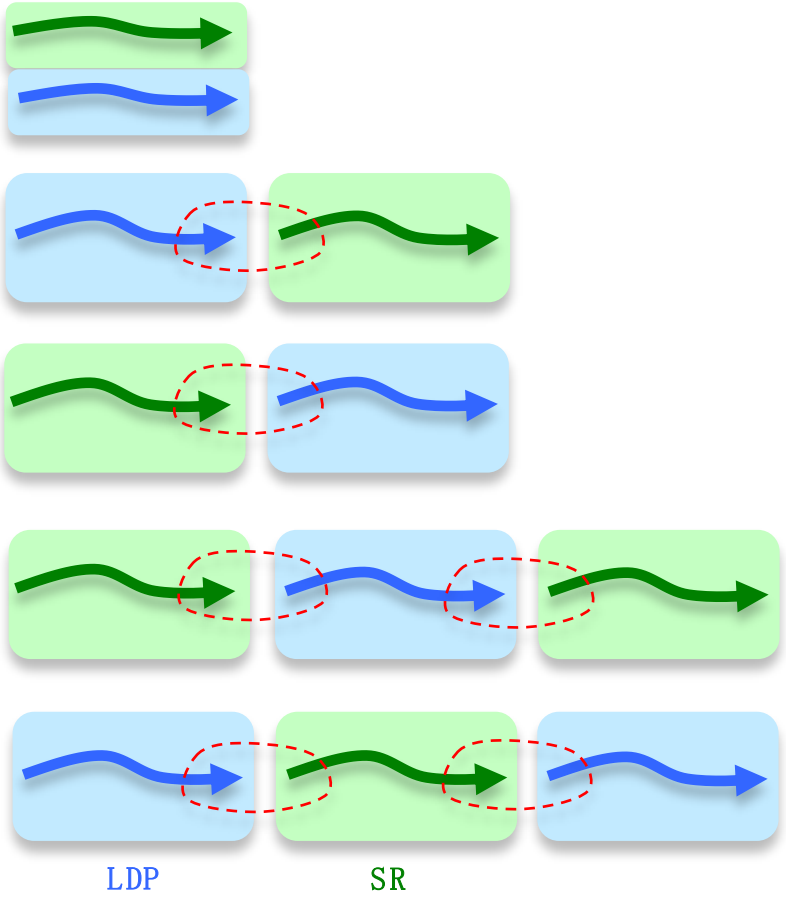
需要映射服务器

SR over LDP

=(SR to LDP)+(LDP to SR)

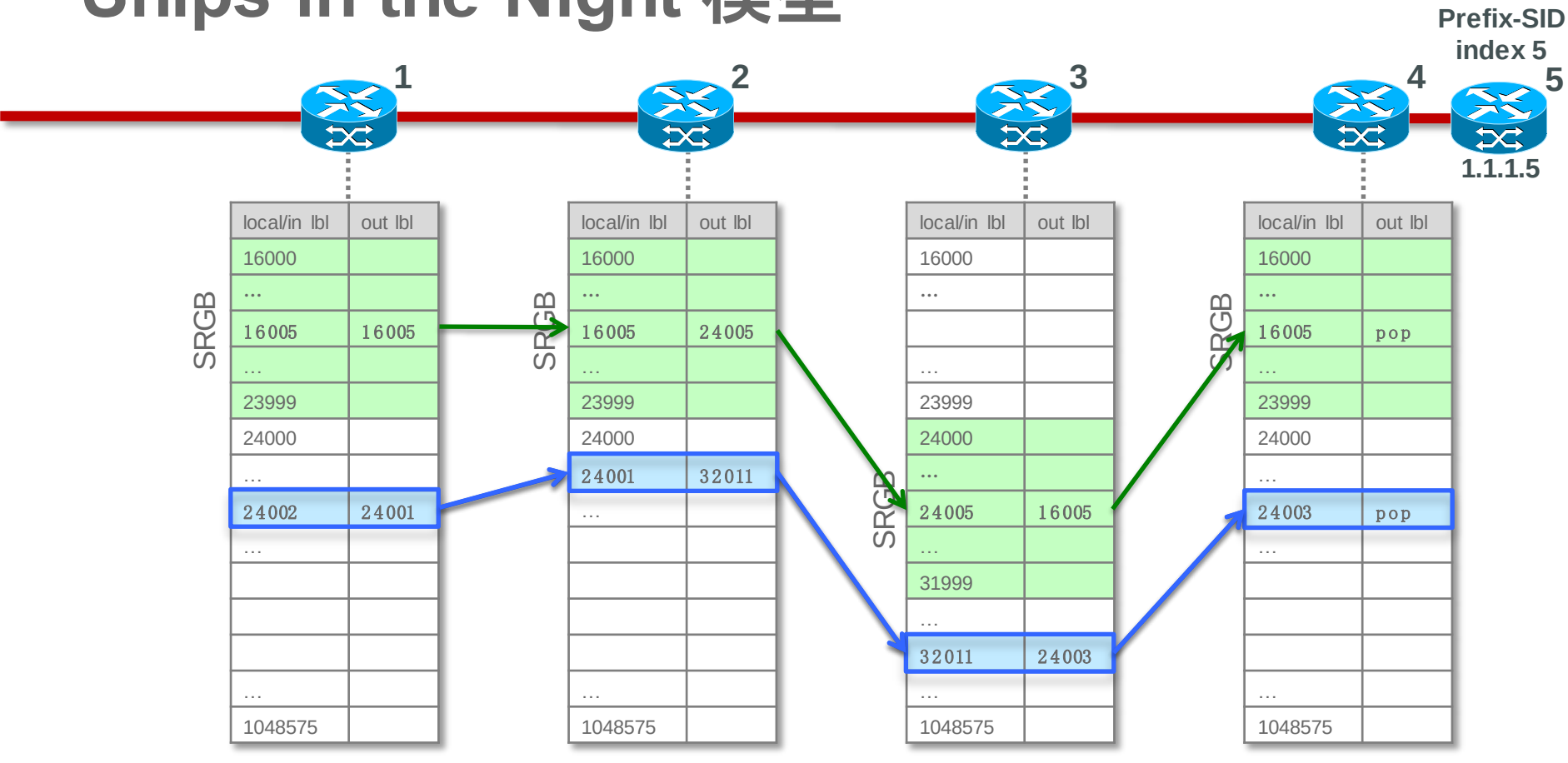
LDP over SR

=(LDP to SR)+(SR to LDP)



All nodes: SR + LDP SR LDP

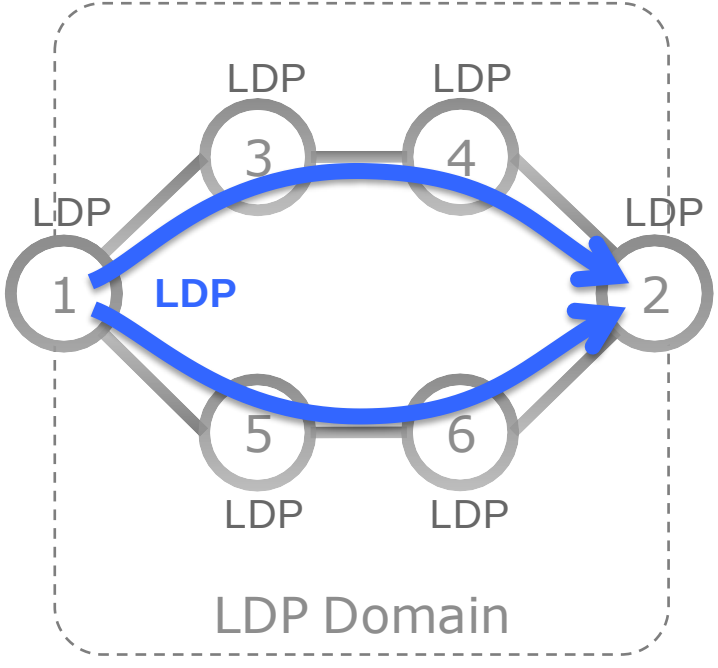
# ‘Ships in the Night’模型



# LDP向SR迁移示例

Assumptions:
- all the nodes can be upgraded to SR
- all the services can be upgraded to SR

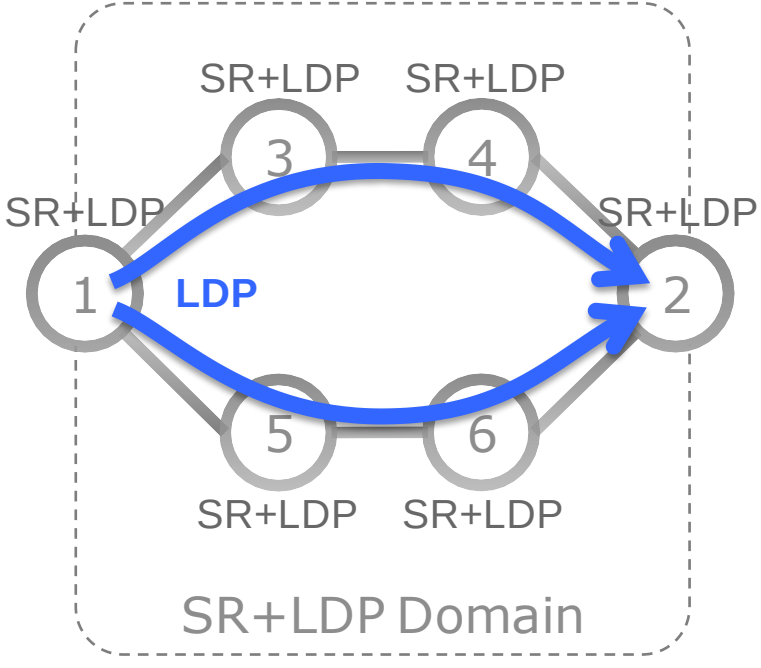
- Initial state: All nodes run LDP, not SR



# LDP向SR迁移示例

Assumptions:
- all the nodes can be upgraded to SR
- all the services can be upgraded to SR

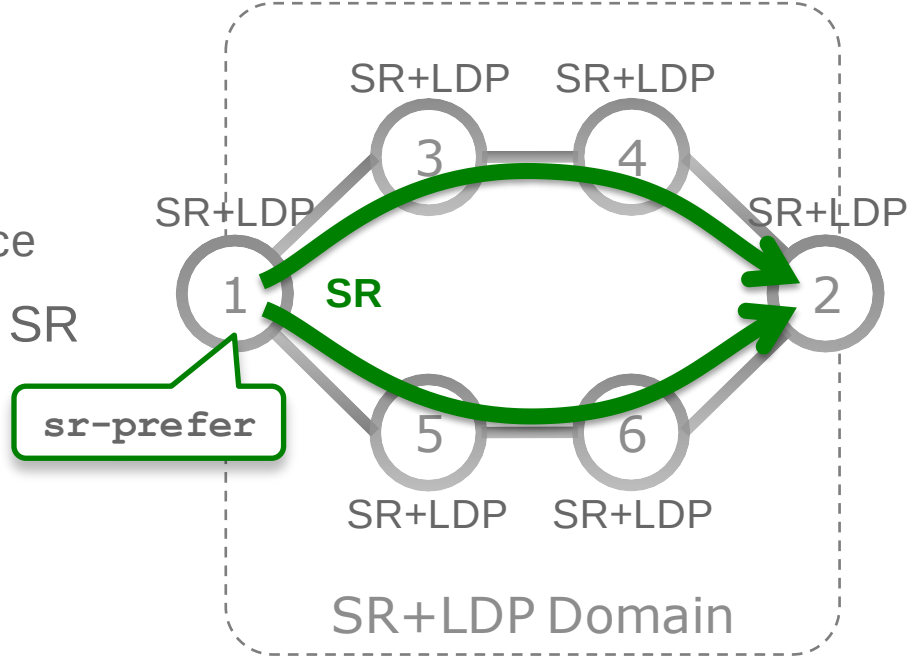
- Initial state: All nodes run LDP, not SR
- Step1: All nodes are upgraded to SR
  - In no particular order
  - leave default LDP label imposition preference



# LDP向SR迁移示例

Assumptions:
- all the nodes can be upgraded to SR
- all the services can be upgraded to SR

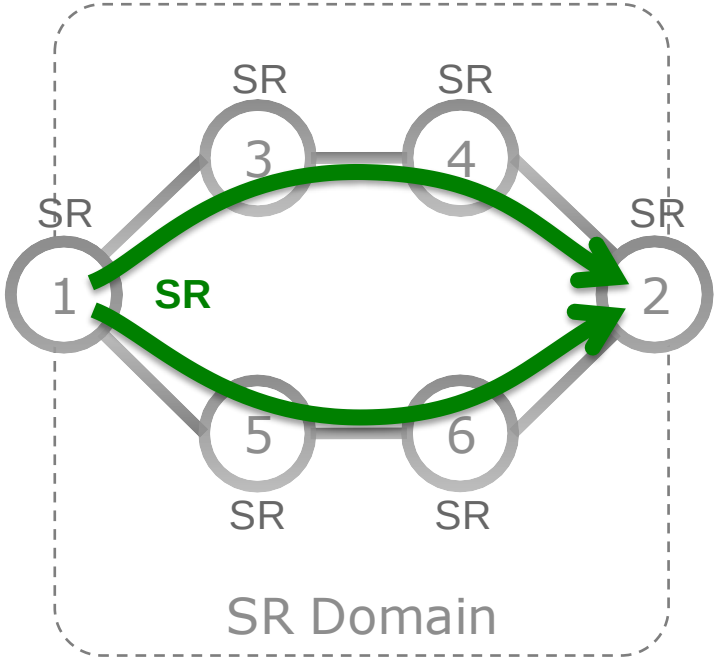
- Initial state: All nodes run LDP, not SR
- Step1: All nodes are upgraded to SR
  - In no particular order
  - leave default LDP label imposition preference
- Step2: All PEs are configured to prefer SR label imposition
  - In no particular order



# LDP向SR迁移示例

Assumptions:
- all the nodes can be upgraded to SR
- all the services can be upgraded to SR

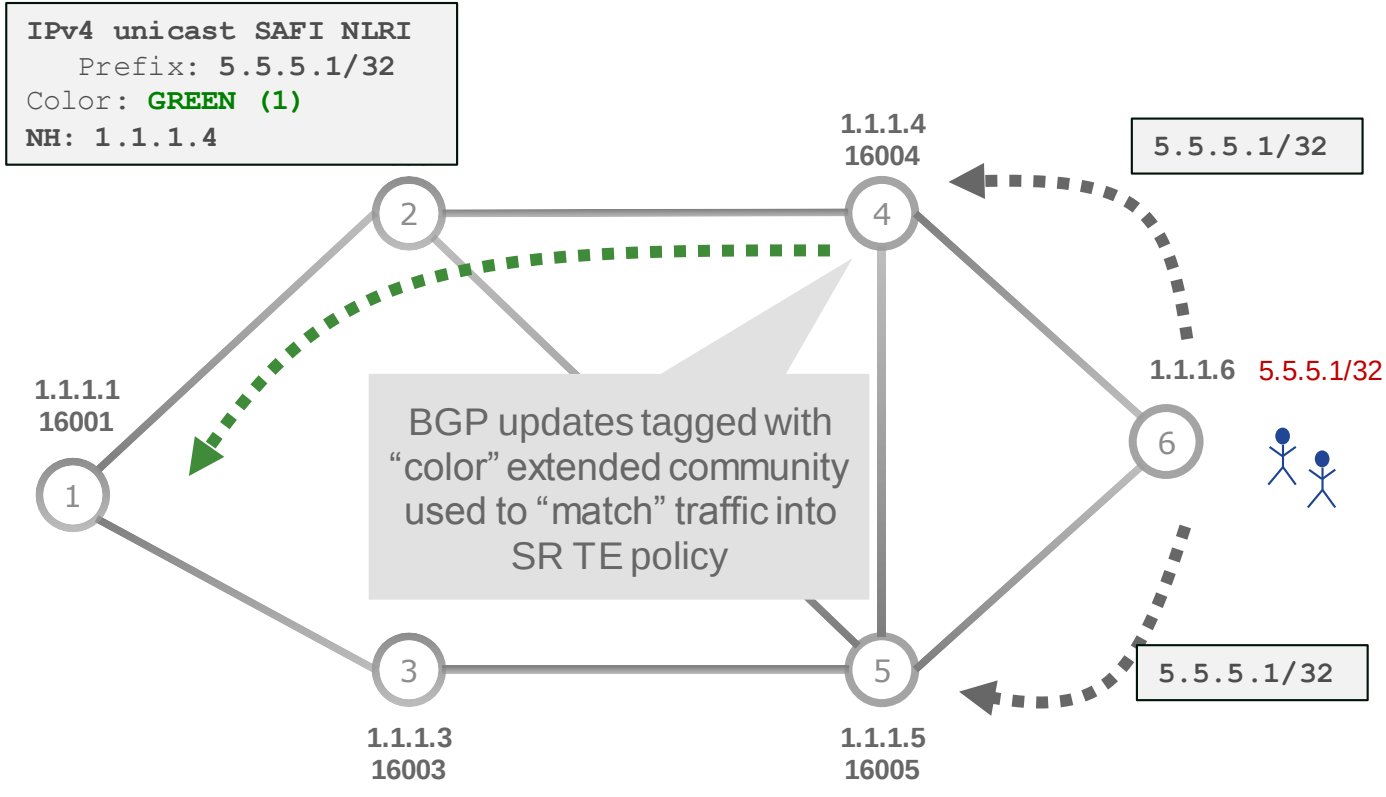
- Initial state: All nodes run LDP, not SR
- Step1: All nodes are upgraded to SR
  - In no particular order
  - leave default LDP label imposition preference
- Step2: All PEs are configured to prefer SR label imposition
  - In no particular order
- Step3: LDP is removed from the nodes in the network
  - In no particular order
- Final state: All nodes run SR, not LDP



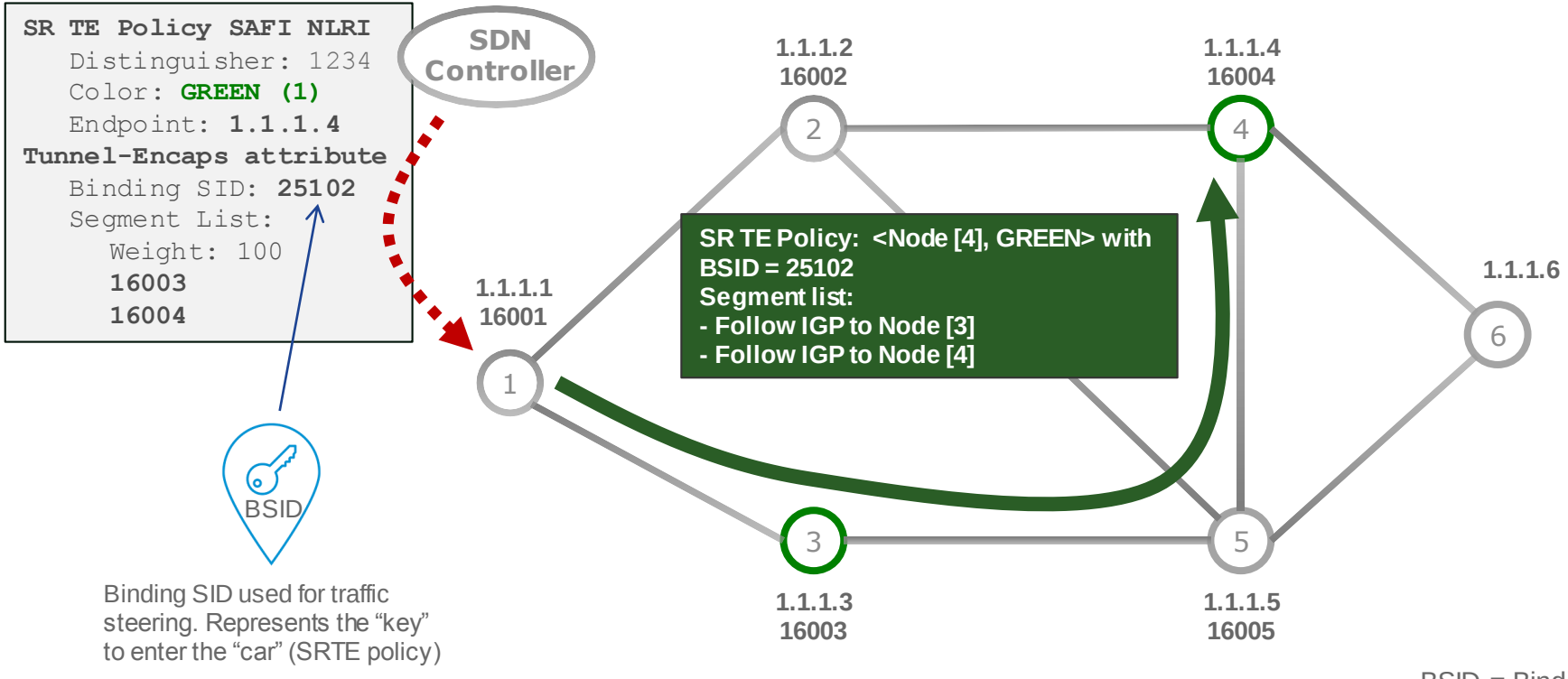
# BGP SRTE

- Advertisement of SR TE policies via BGP
- Automatic instantiation of SR TE policies
- Automatic traffic steering into SR TE policies, eliminate the need for PBR

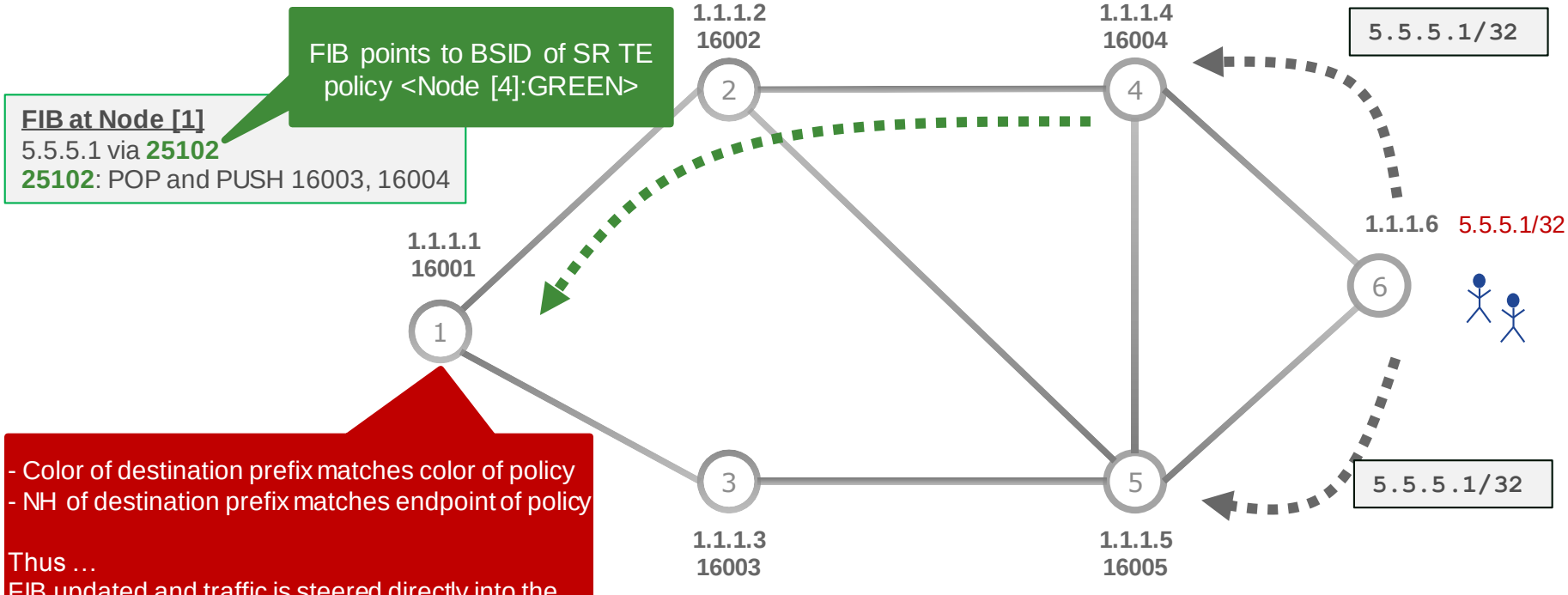
# 通过BGP Community标记需要获得的SLA



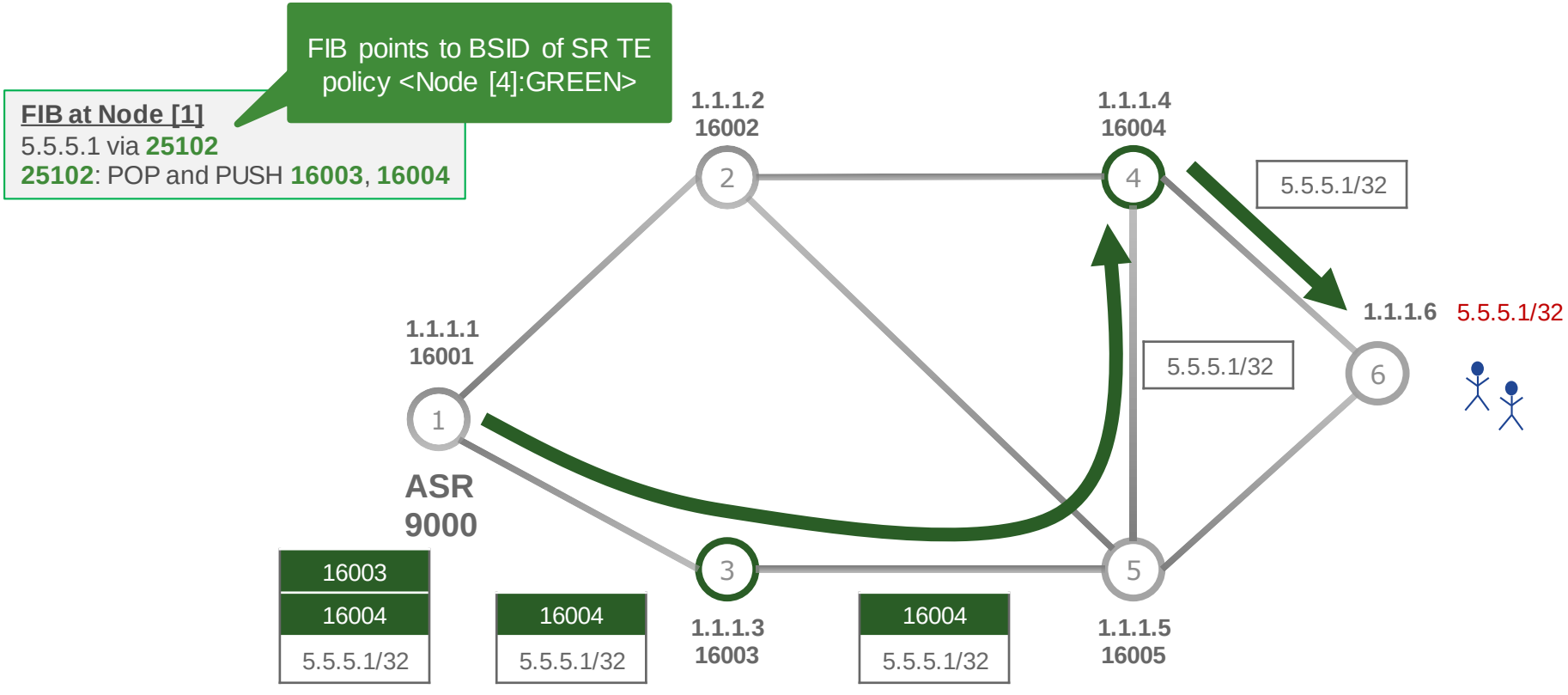
# BGP SRTE策略: 通过BGP更新或在设备静态配置



# BGP SRTE自动生成隧道并引流,BSID实现关联



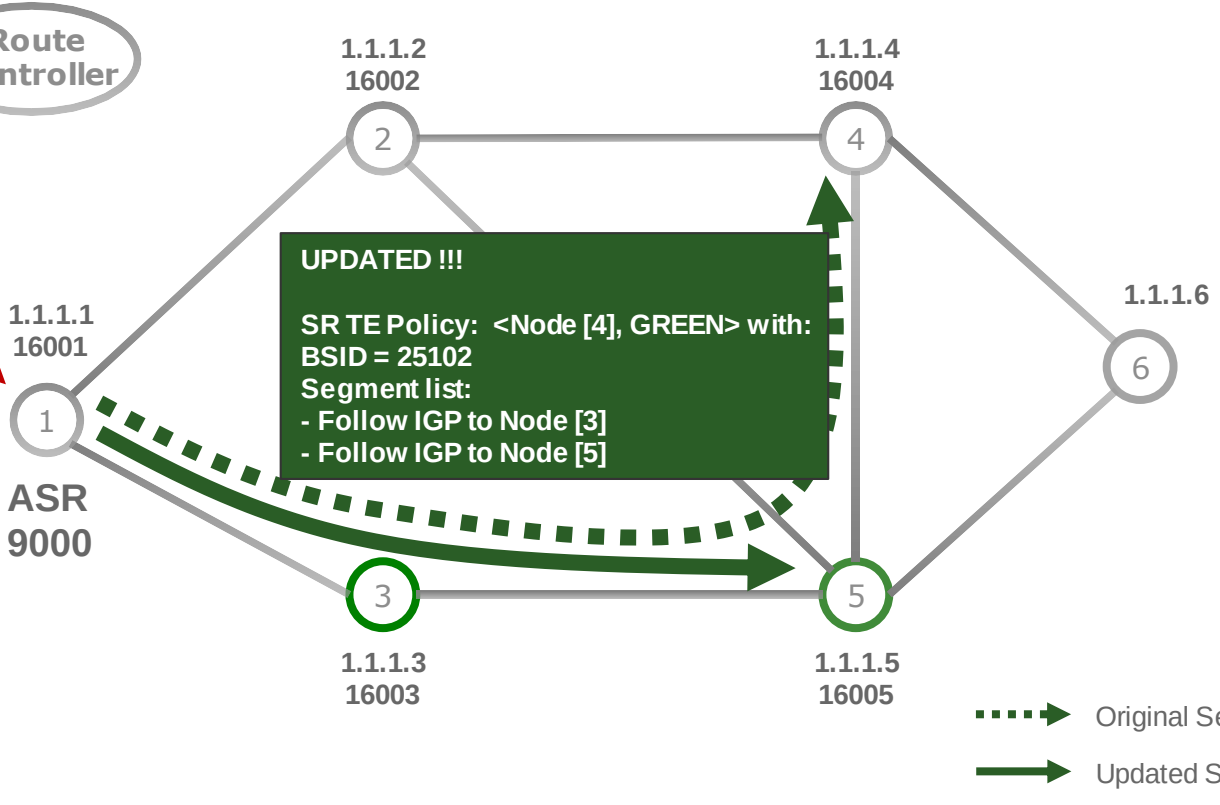
# 转发过程



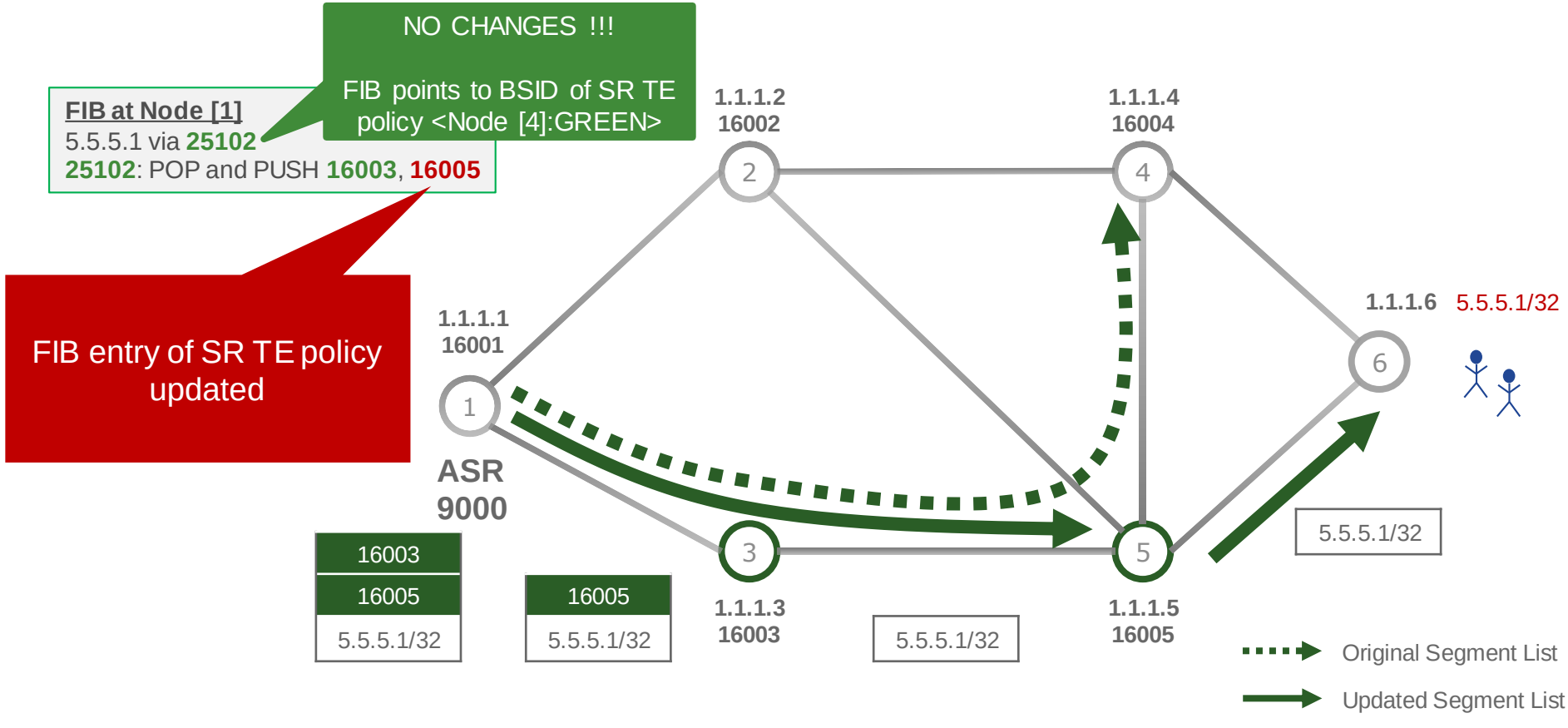
# 只要SRTE Policy不改变,则BSID不改变,但Segment List可以改变

```
SR TE Policy SAFI NLRI
   Distinguisher: 1234
   Color: GREEN (1)
   Endpoint: 1.1.1.4
Tunnel-Encaps attribute
   Binding SID: 25102
   Segment List:
      Weight: 100
      16003
      ~~16004~~ 16005
```

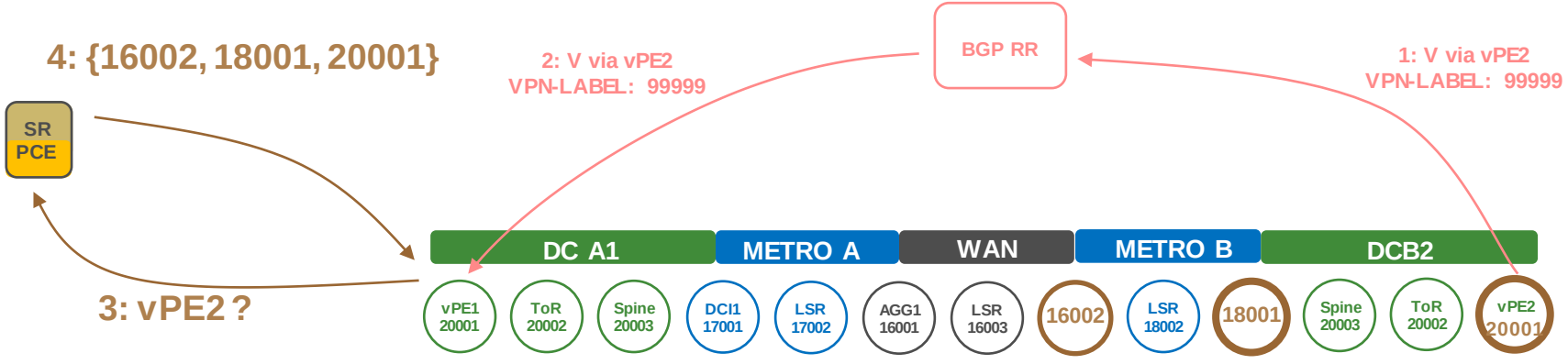
SR Policy NLRI is updated e.g. Segment list is modified



# 转发过程(更新)

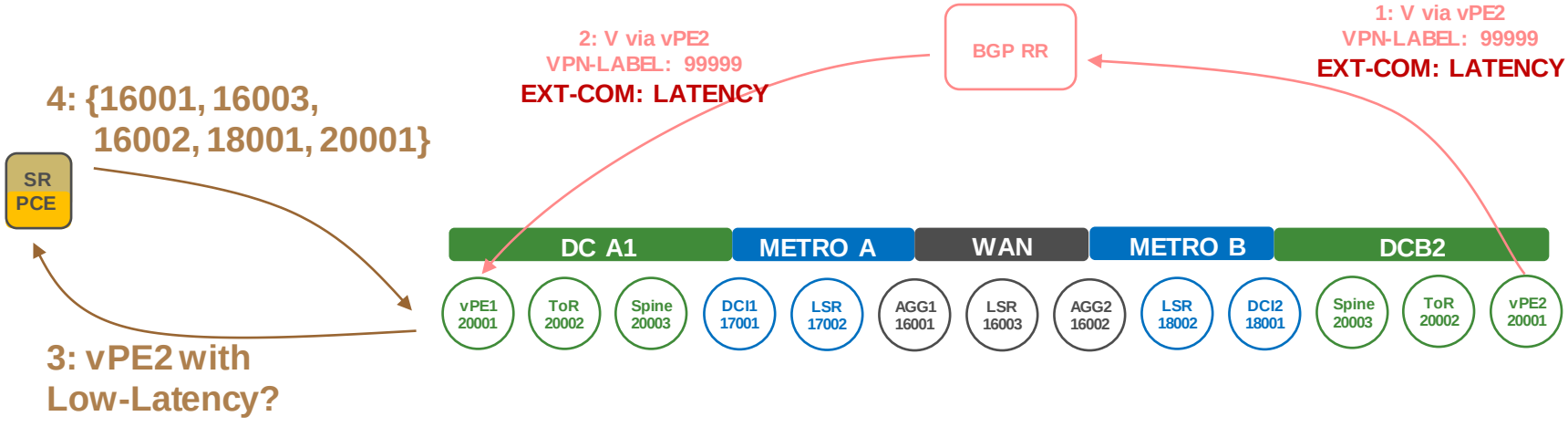


# 按需下一跳ODN
# 降低边缘设备FIB要求



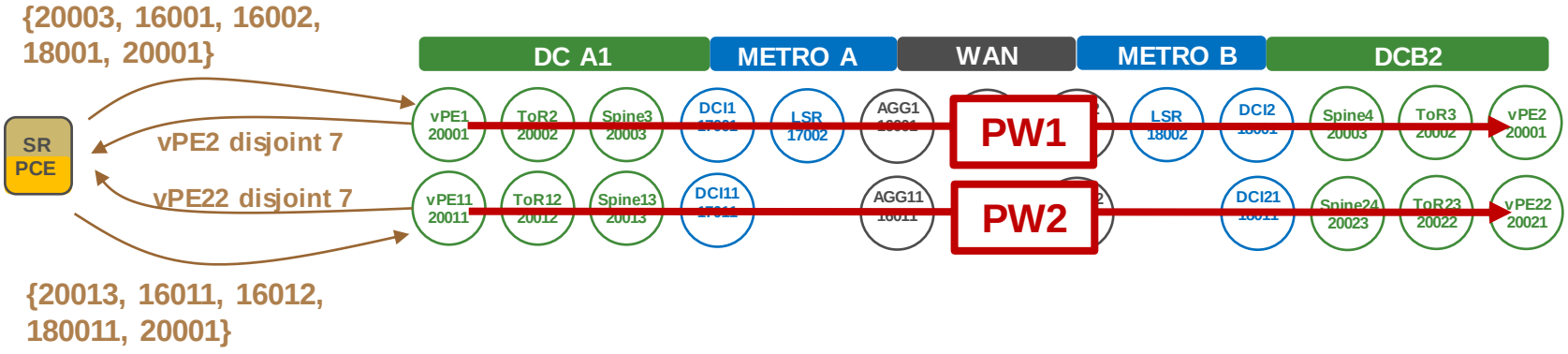

- vPE1's ODN functionality automatically request a solution from SR-PCE
- Scalable: vPE1 only gets the inter-domain paths that it needs
- Simple: no BGP3107 pushing all routes everywhere

# ODN结合SLA需求



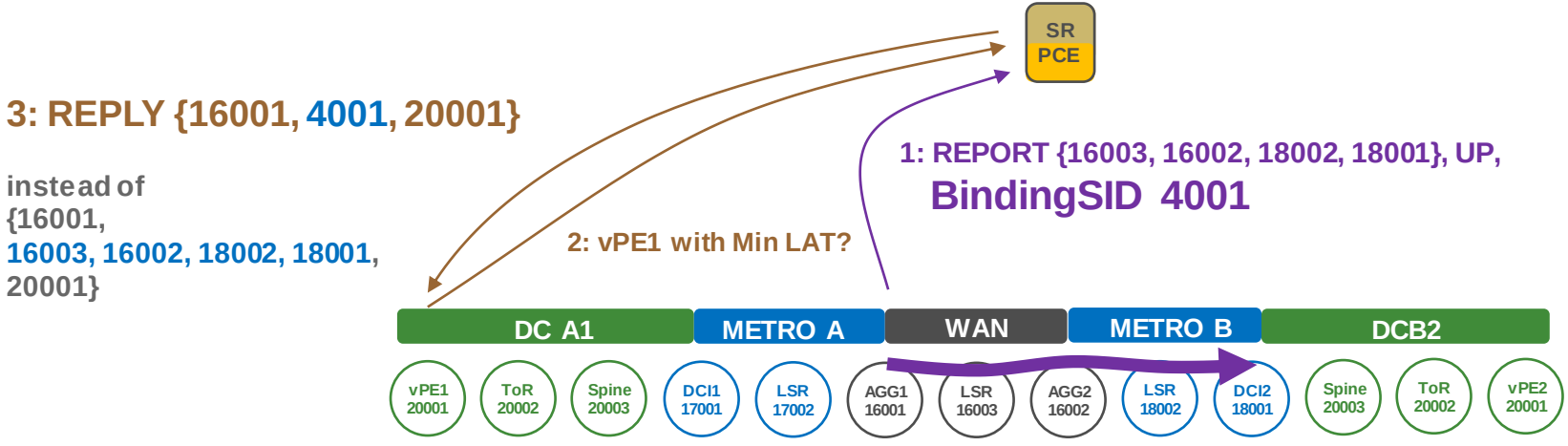

- Inter-domain SLA with scale and simplicity
  - No RSVP, no midpoint state, no tunnel to configure !!

# ODN结合分离路径需求(Disjoint Path)




- ODN/SR-PCE automated compute disjoint paths for PW1 and PW2
- PW1 and PW2 do not share the same headend, neither the same tailend
- Inter-domain SLA with scale and simplicity
  - No RSVP, no midpoint state, no tunnel to configure !!

# ODN结合BSID

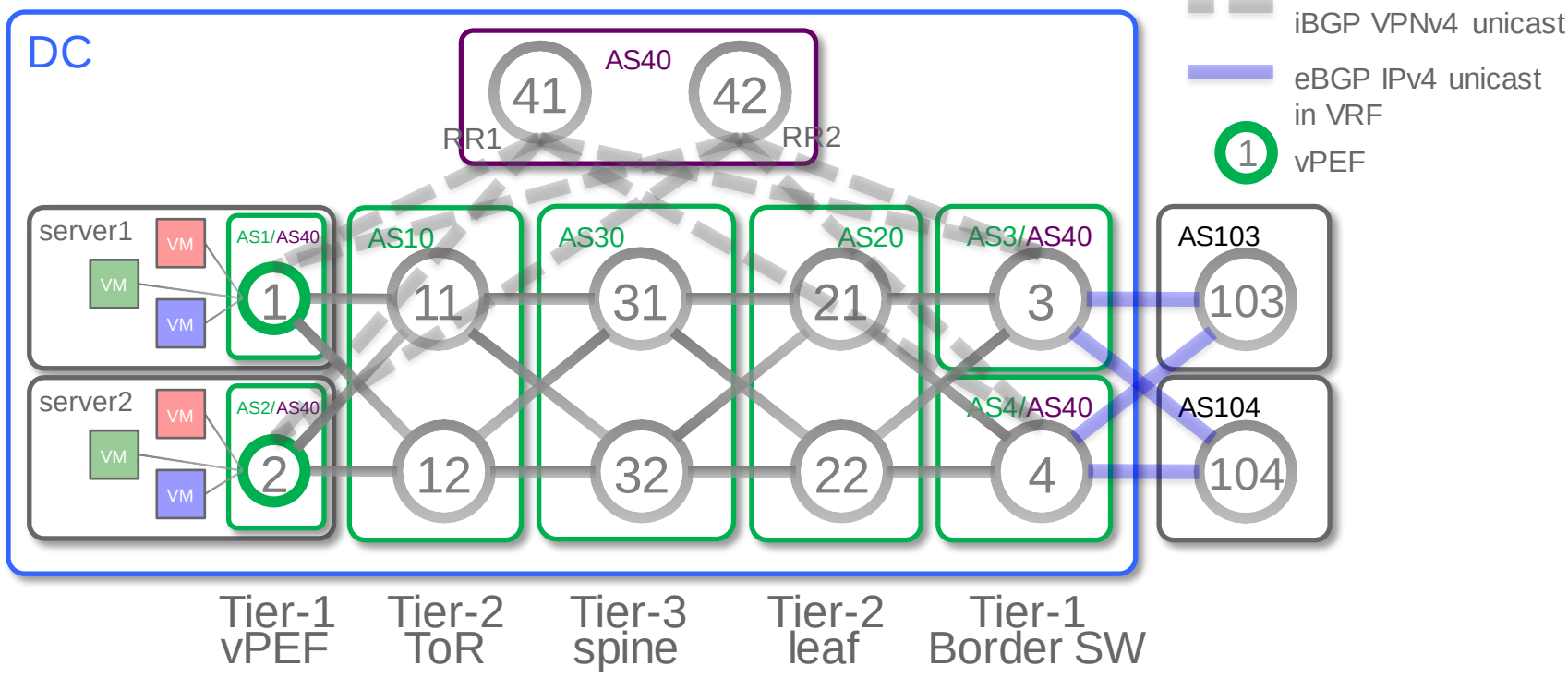



- End-to-end policies can be composed from more basic ones
  - An SRTE policy is bound by default to a Binding SID
  - RSVP-TE tunnels can also be bound to a Binding SID and hence RSVP-TE tunnels can be used within an end-to-end SR policy
- **Shorter SID list and churn isolation between domains**
  - Even if the WAN-MetroA sub-path changes, the related Binding SID 4001 is constant

# 数据中心网络面临的挑战

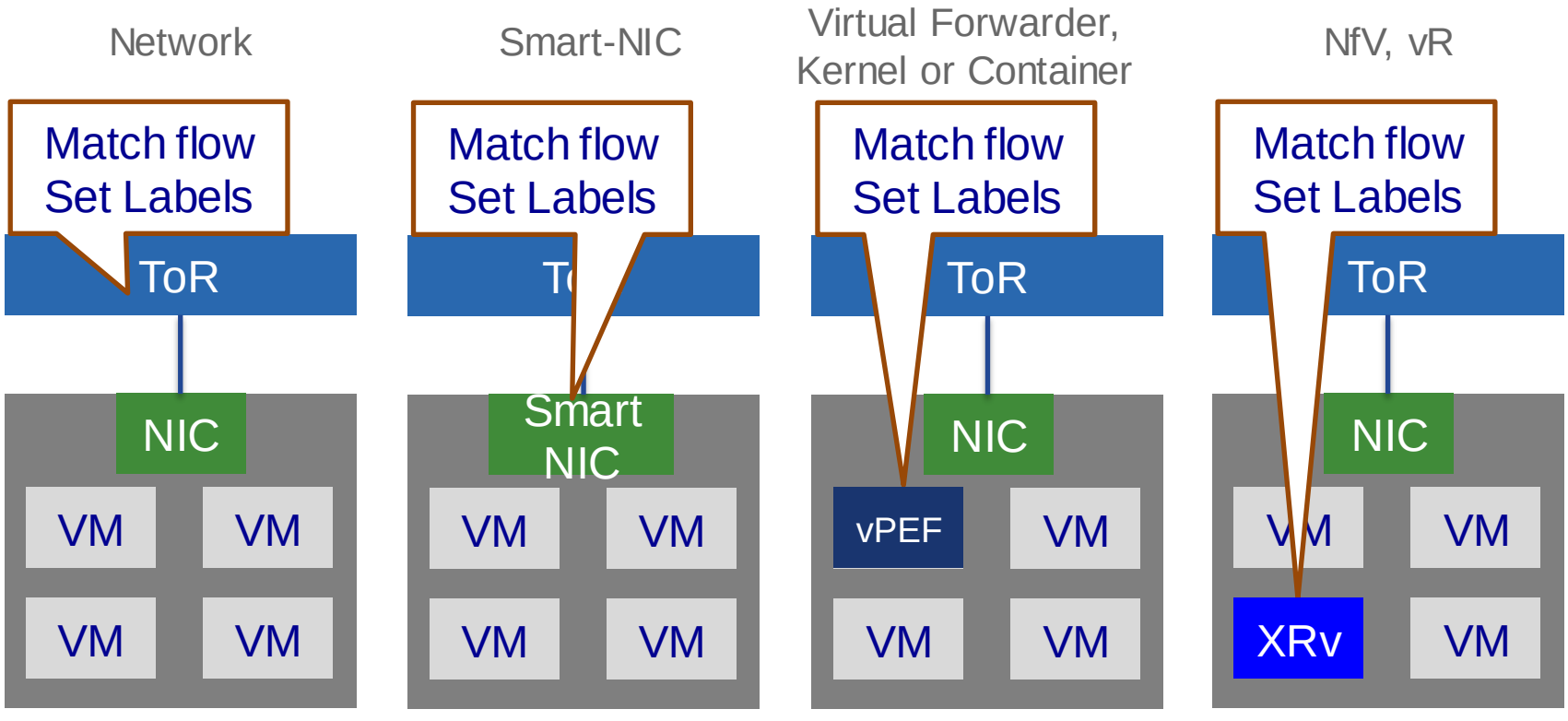
- Elephant Flows
  - Hashing over ECMP is flow based
  - Hence a long lived heavy flow overwhelm short-lived small flows
- Fault-isolation is hard
  - Non-determinism of exact path over ECMP due to many short-lived flows
- End-points oblivious to ECMP-based path selection
  - TCP treats the network as a blackbox
  - Difficult to re-route around congested points
- TE inside a DC
  - Different label values on different boxes*
  - Requires lots of signaling even with the presence of PCE/SDN-controller**

# SR数据中心

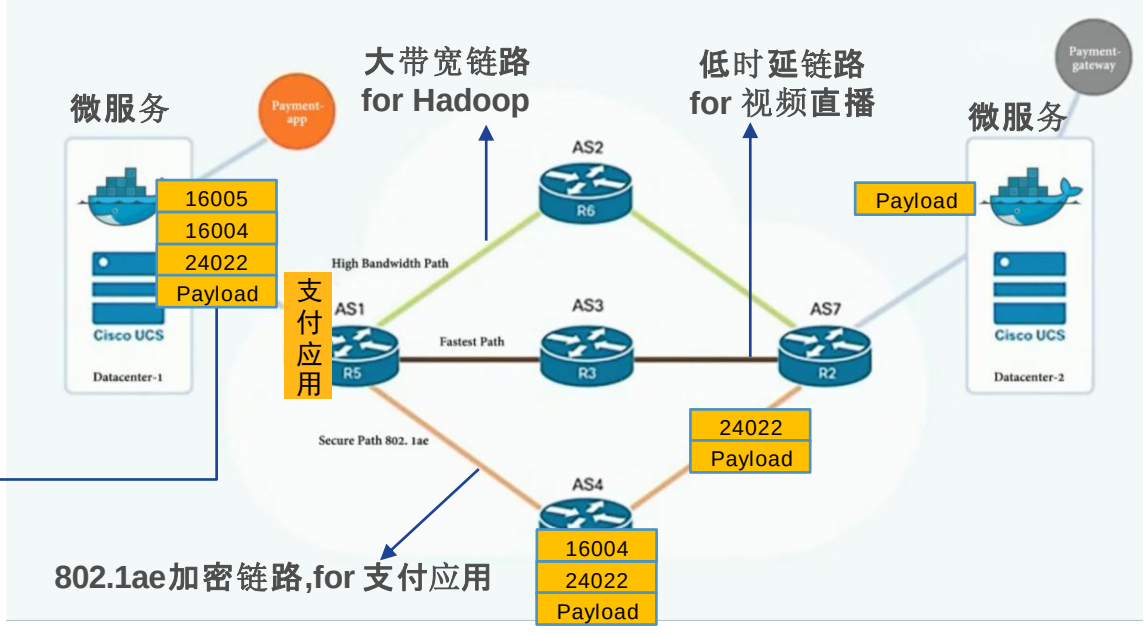



- Underlay uses SR using BGP-LU over IPv4
- Overlay uses SR using BGP-LU over VPNv4
- Hence on tier 1 routers (ToR or vPEF or Border SW), we have two BGP Instances
  - one for DC Fabric prefixes,
  - one for Overlay prefixes

# 在服务器上实现SR



# 在容器上实现SR
# Demo@Ciscolive

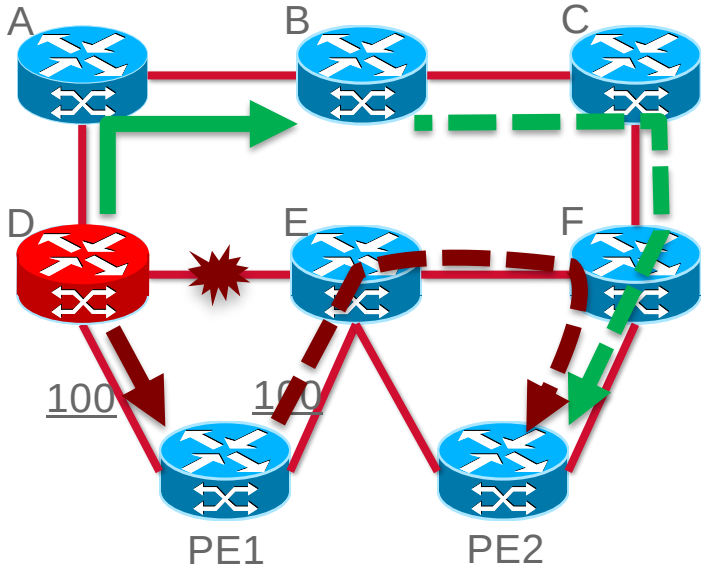



Contiv

Open Source

dev@host-A:$sudo tcpdump -I eth2 –c 1
tcpdump: verbose output suppressed use -v or -w for full protocol decode
listening on eth2.link-type EN10MB(Ethernet),capture size 262144 bytes
17:32:03.278955 MPLS (label **16005**,exp 0,ttl 64) (label **16004**,exp 0,ttl 64) (label **24022**,exp0,[S], ttl 64) IP10.200.1.3>10.200.1.4:ICMP echo request, id 27, seq 155 length 64

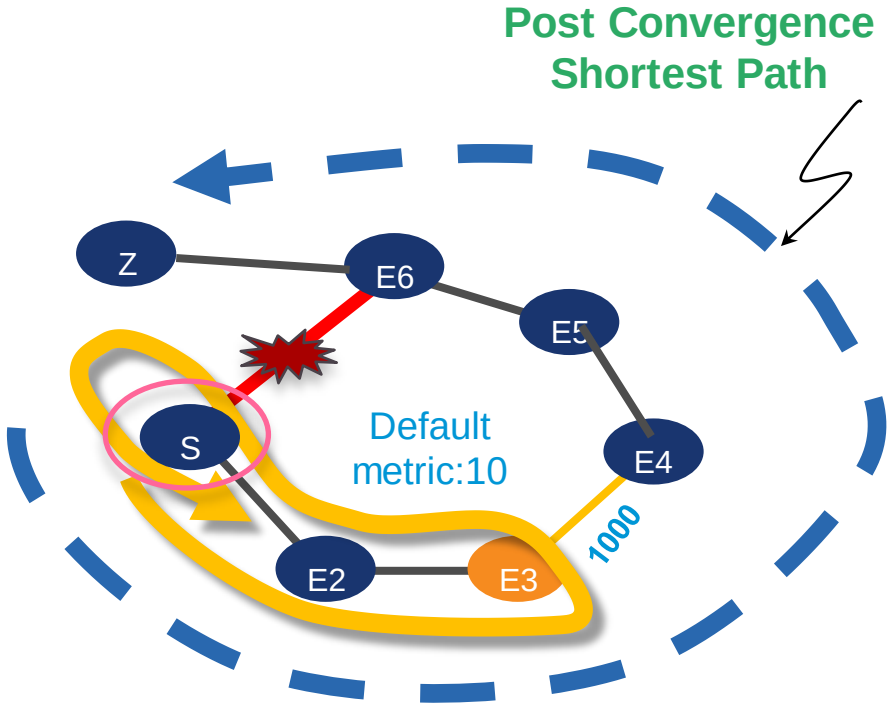
# LFA/RLFA存在的问题




Default metric: 10

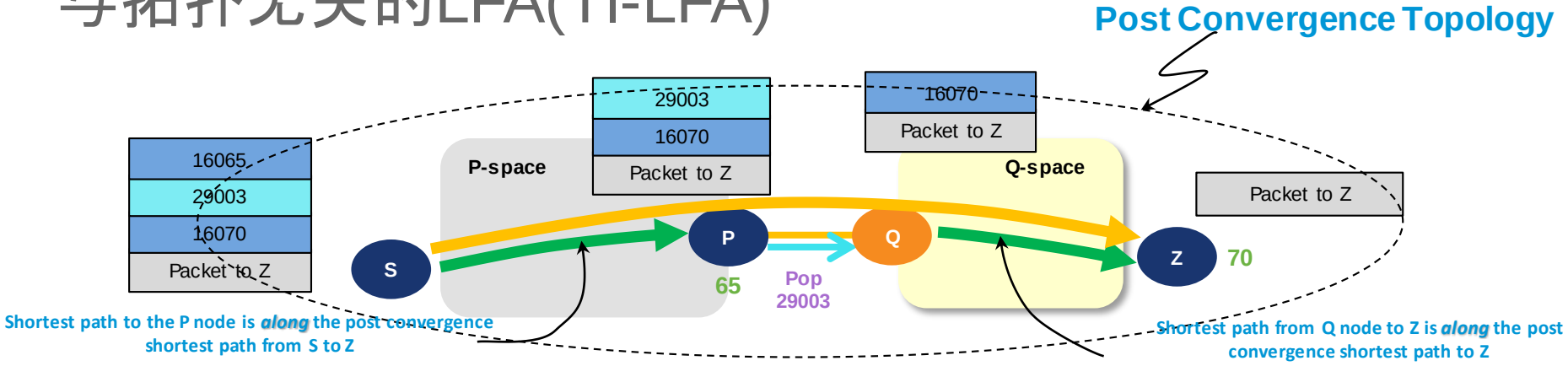
- Protecting link D-E on node D
- LFA: D switches all traffic destined to PE2 towards the edge node PE1

→ an edge node and edge links are used to protect the failure of a core link



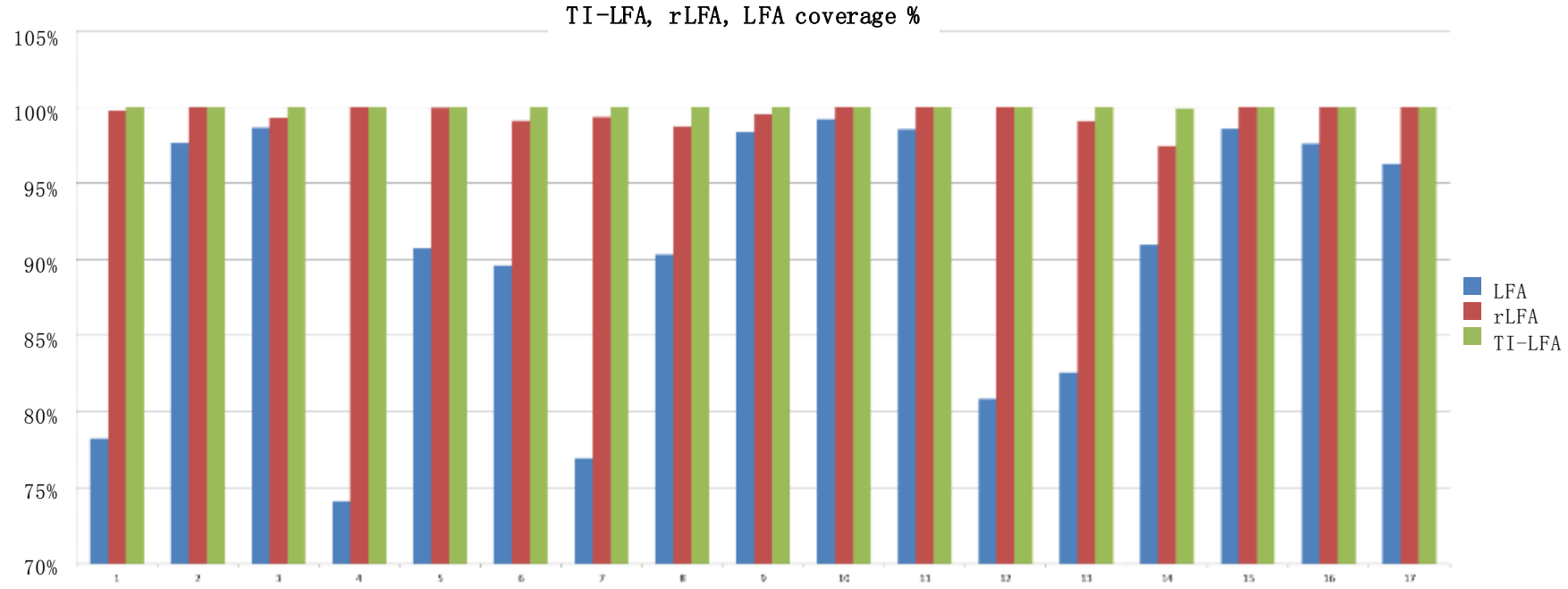

 

# 与拓扑无关的LFA(TI-LFA)




- P-Space:
  - Nodes that are reachable without the protected link
- Q-Space
  - Nodes that can reach the far end of the protected link without the link itself
- Calculate the post-convergence shortest path
- Find a PQ or P with adjacency Q along the ***post convergence*** shortest path
- Send the packet to a P node
- Force the packet to the adjacent Q node
- Let the packet flow freely to the destination

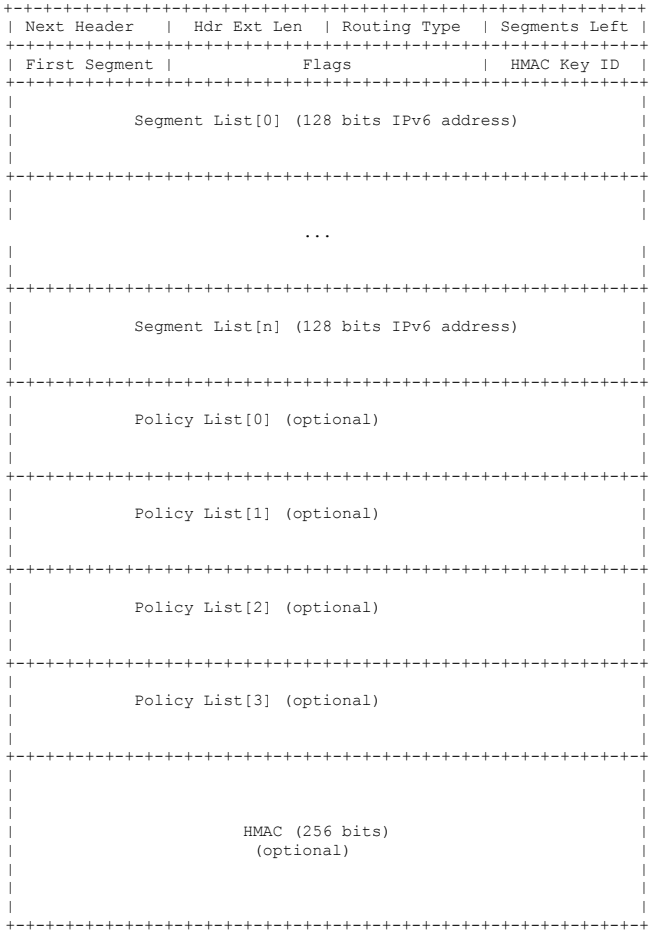
# TI-LFA能实现任意网络下的快速无环收敛



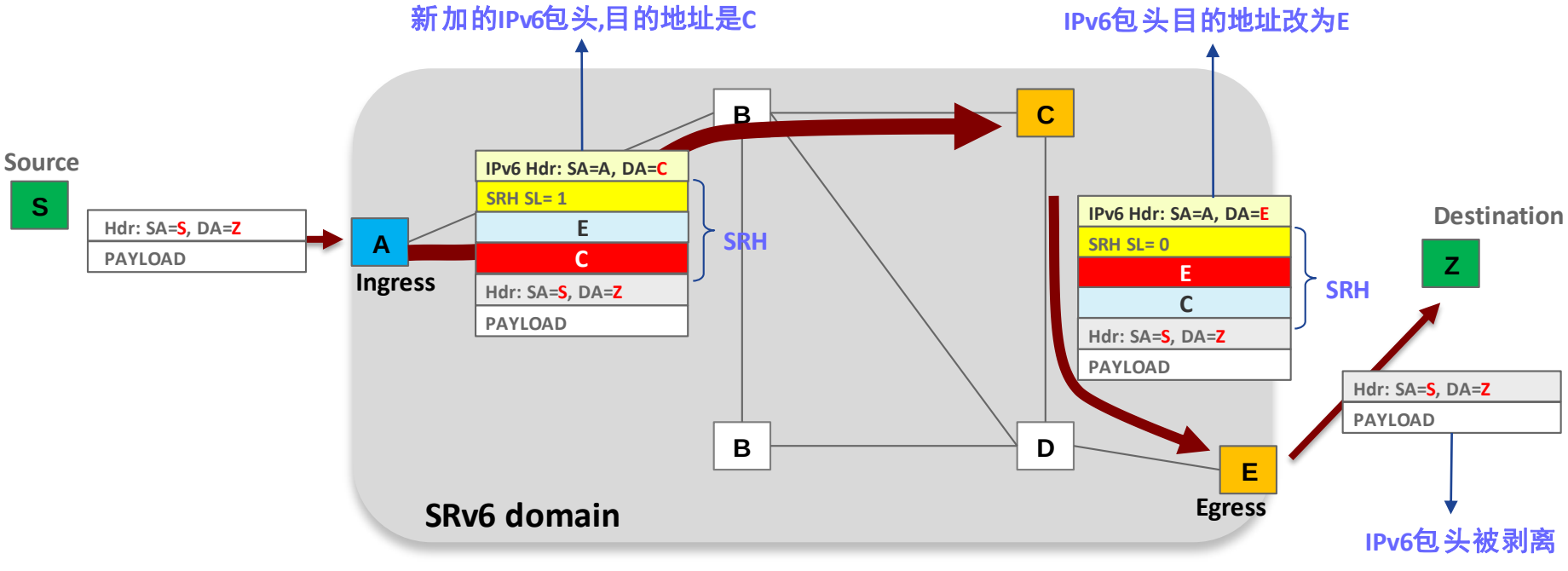

- Average over 17 SP WAN’s: 100.0%

# SRv6

- An SRv6 SID is simply a “forus” IPv6 address but…
- Treat that ***forus*** IPv6 address as an ***instruction***, rather than a destination
- Leverage existing Routing Header
  - Segment routing Extension header (***SRH***)
  - A secure superset of RH0
  - The SRH ***steers*** the packet into the desired path



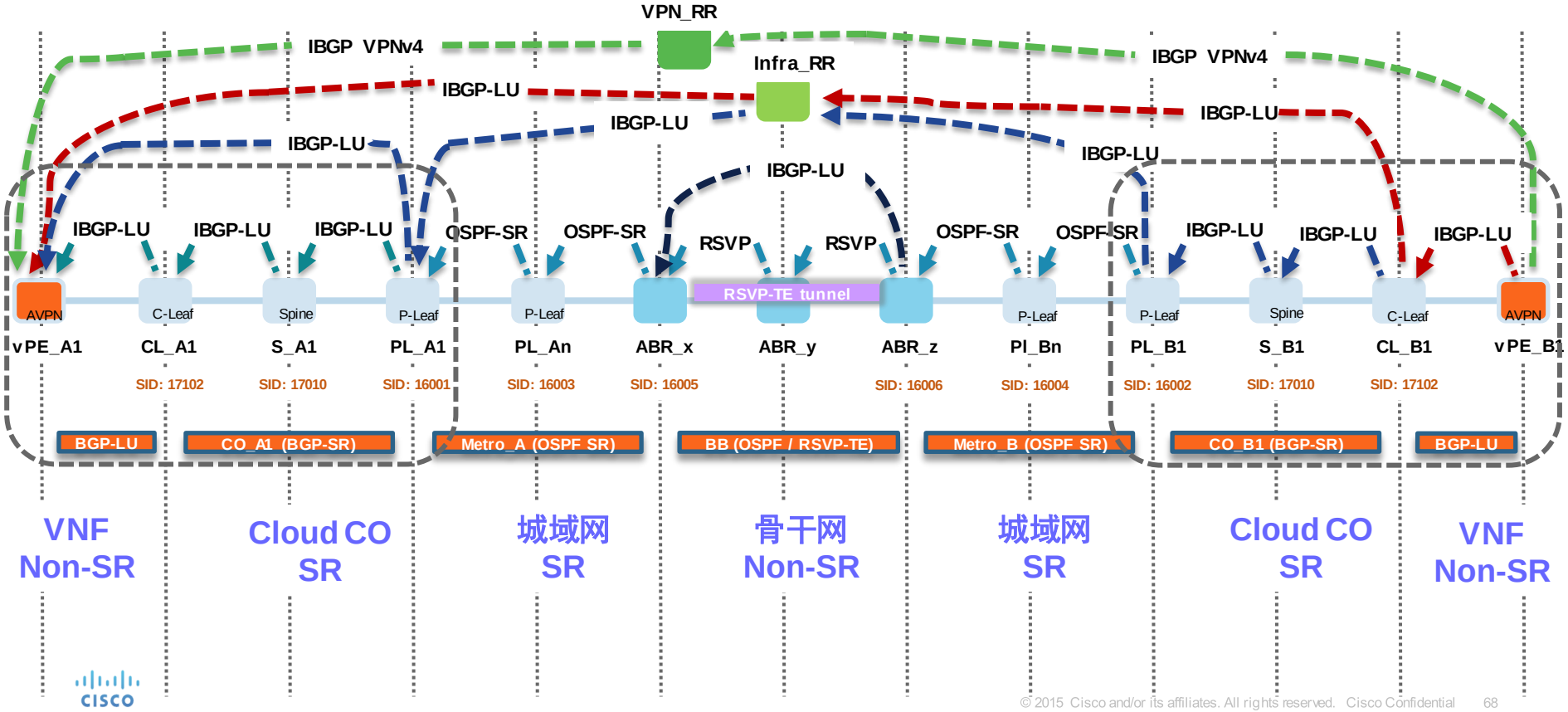
# SRv6转发过程




*如果原始报文本身是IPv6,也可选择直接在原IPv6包头基础上增加SRH的方式实现SRv6

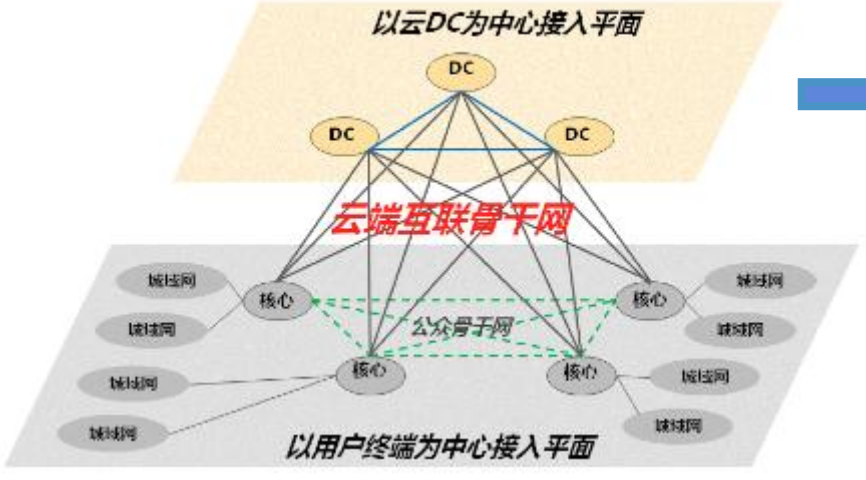
# 案例分享

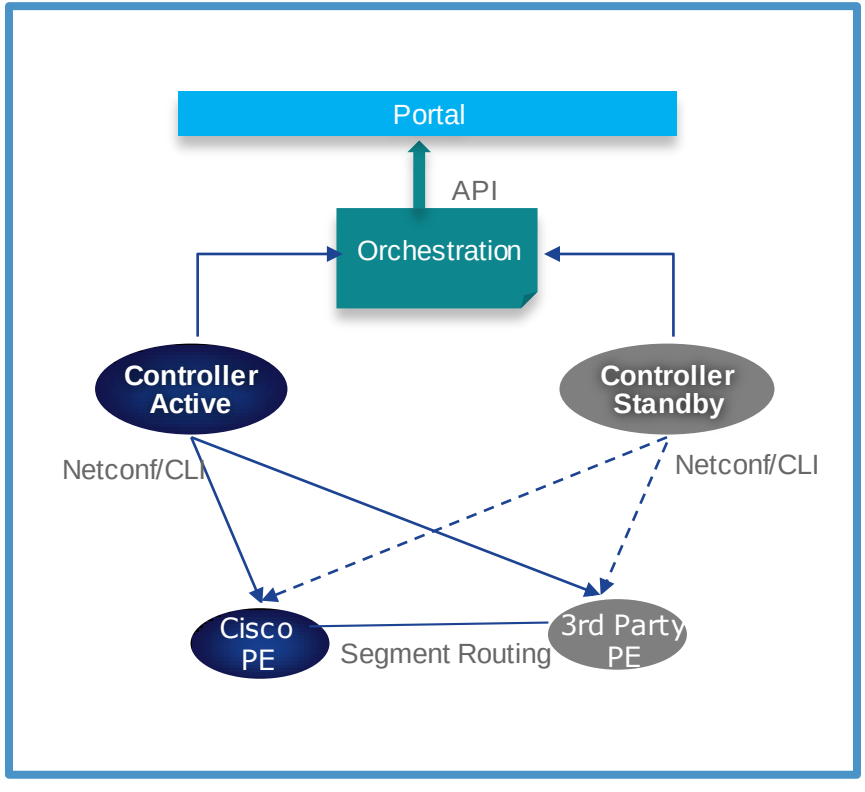
# 北美运营商NFV业务实现:SR+EVPN
## 端到端不再是问题,规模不再是问题!



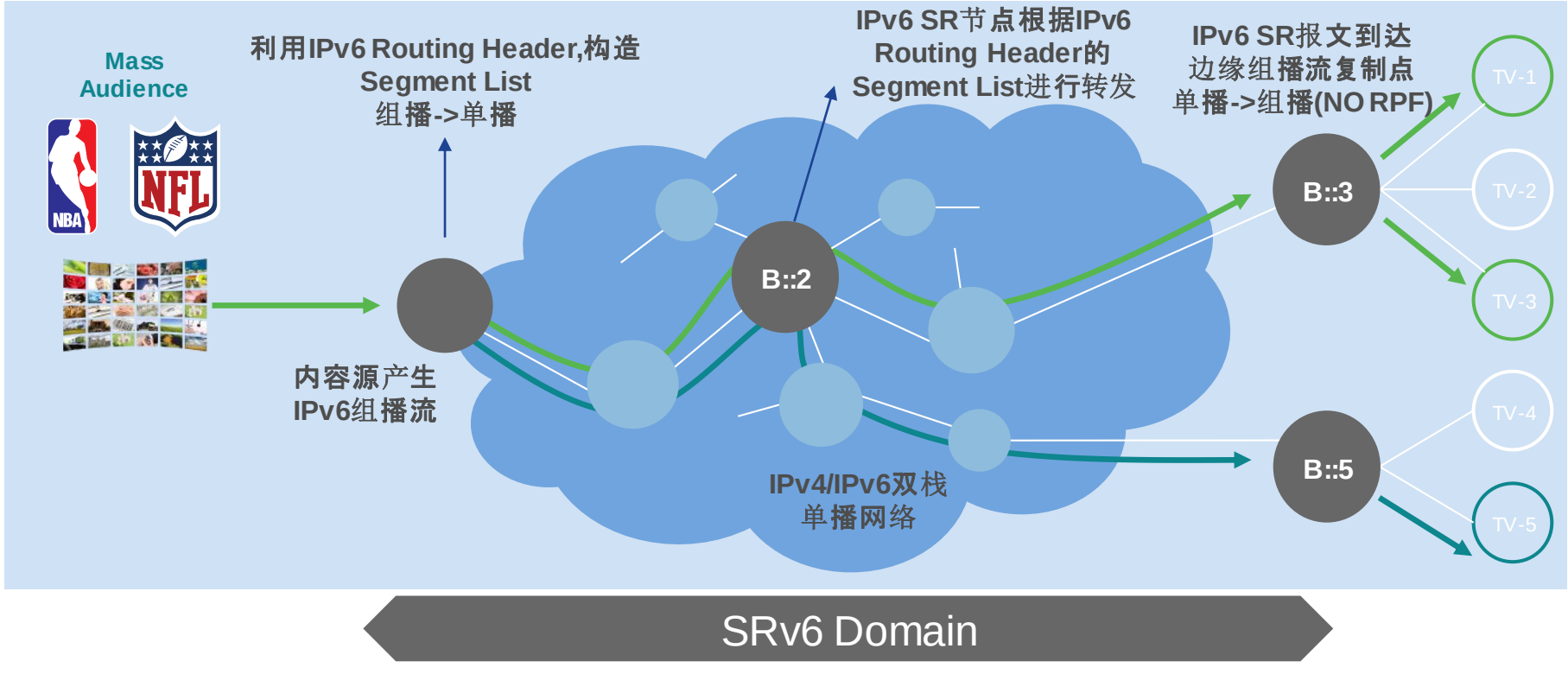

 

# 国内运营商承载网



# 北美运营商基于SRv6在双栈网络上分发直播内容



# 思科SR解决方案

# SR@Cisco: 全系列产品支持

IOS XR　　NX-OS

IOS classic　　Linux

NCS6000

CRS-3 / CRS-X

ASR9000

NCS5000
NCS5500

(NCS4000)

CSR1000v

*(XRV-9000)*

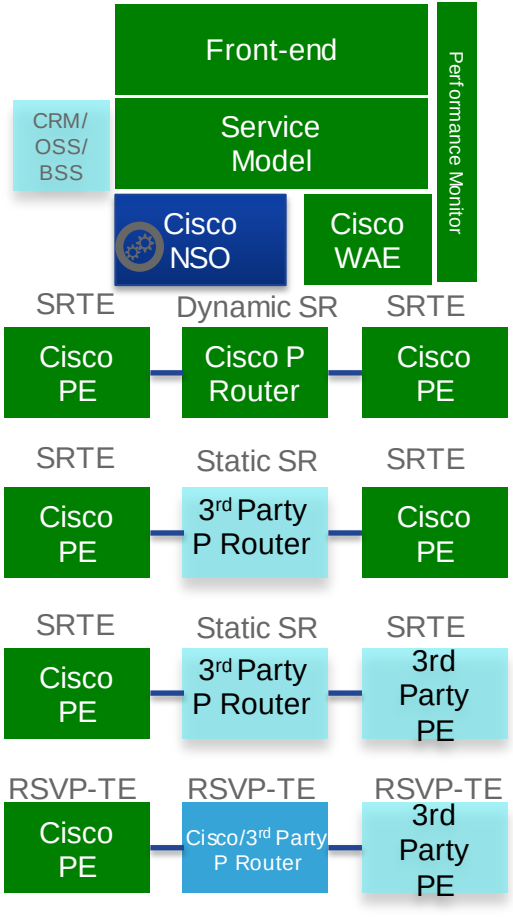
ASR900

ASR1000 / ISR400 / (*cBR8)*

(NEXUS 7000)
(NEXUS 8000)
NEXUS 9000


FD.io
WAE
ODL
(Docker)
(Linux
Kernel)

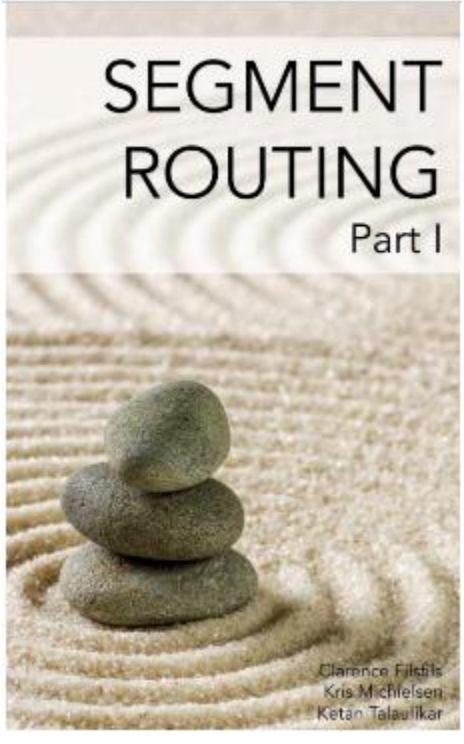
# SR@Cisco: SCN一体化解决方案




- WAE for network modeling, topology collection & path calculation
  - 10+ yrs path calculation engine
  - Deployed at Telstra Global/Comcast/PCCW/Facebook/Tencent
  - Only WAN controller supports both RSVP-TE & Segment Routing
  - ODL based, supports multi-vendor

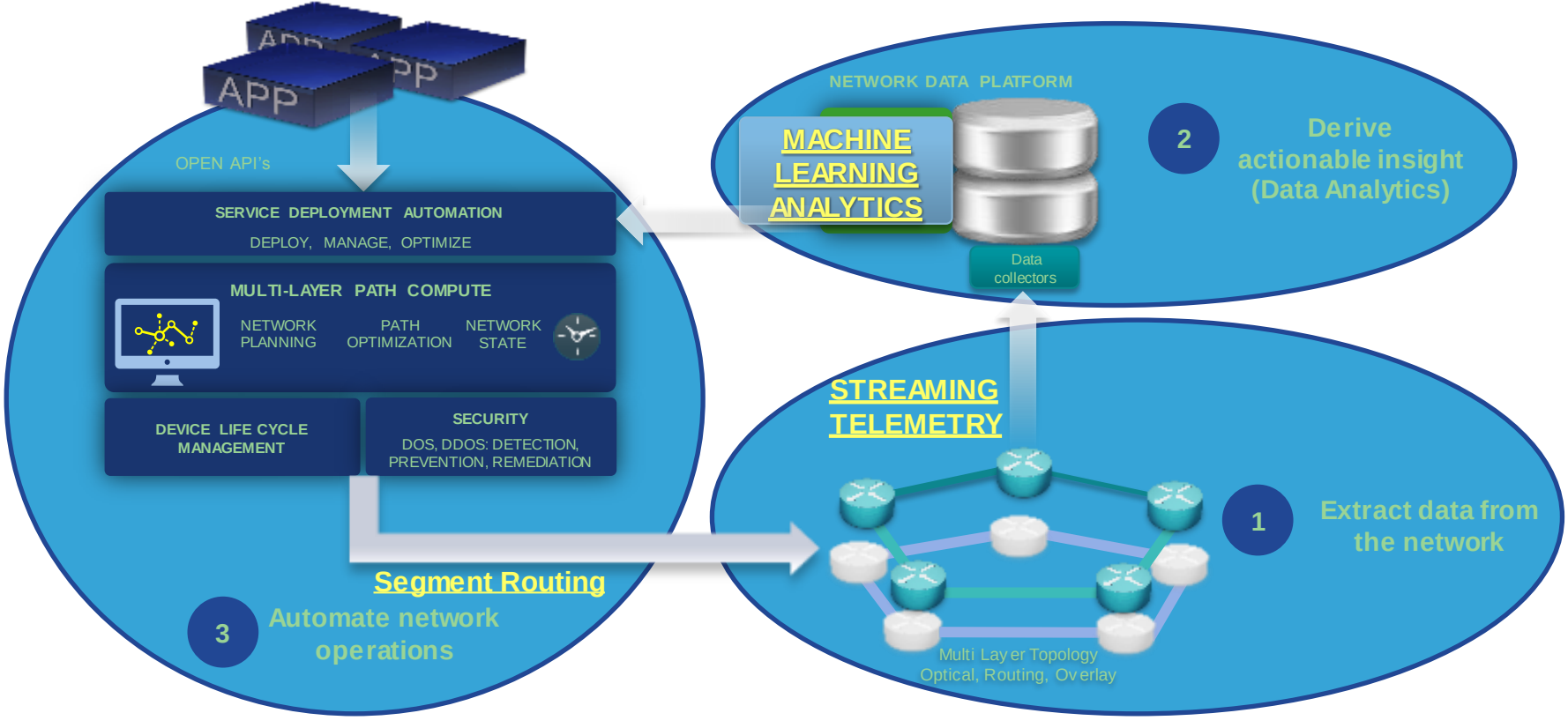
- NSO for traffic steering & TE/QoS provisioning, with extensive Multi-vendor support
  - YANG standard inventor. ConfD/NCS, industry de facto Netconf engine
  - No.1 Orchestration Software by Infonetics
  - Deployed at almost all Tier-1 SP, including ATT D2.0
  - Single NSO Cluster supports 10,000s devices

# SR@Cisco: Segment Routing首部专著




- 由Segment Routing之父Clarence及多位思科资深专家撰写，Kindle已经有售
- 中文版将于2017年出版!

# 总结: SR=应用驱动网络




CISCO
TOMORROW starts here.