MPLS SR协议介绍

引入

- Segment Routing（直译为分段路由，缩写为SR）技术，脱胎于MPLS，但是又做了革命性的颠覆和创新，它代表的是一种新的网络理念——应用驱动网络。自从诞生那一刻起，SR技术便被誉为网络领域最大的黑科技，因其与SDN天然结合的特性，也逐渐成为SDN的主流网络架构标准。本文为大家梳理了SR技术的起源，引出SR技术的基本概念和优势，并展望SR下阶段的演进方向。

课程目标

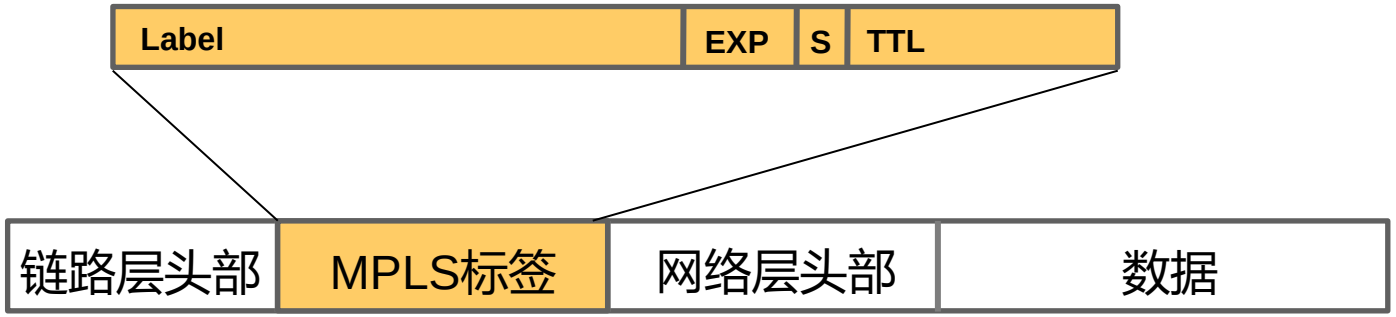
学习完本课程，您应该能够：

- 了解SR产生的背景
- 掌握SR的基本概念和原理
- 了解SR的应用案例

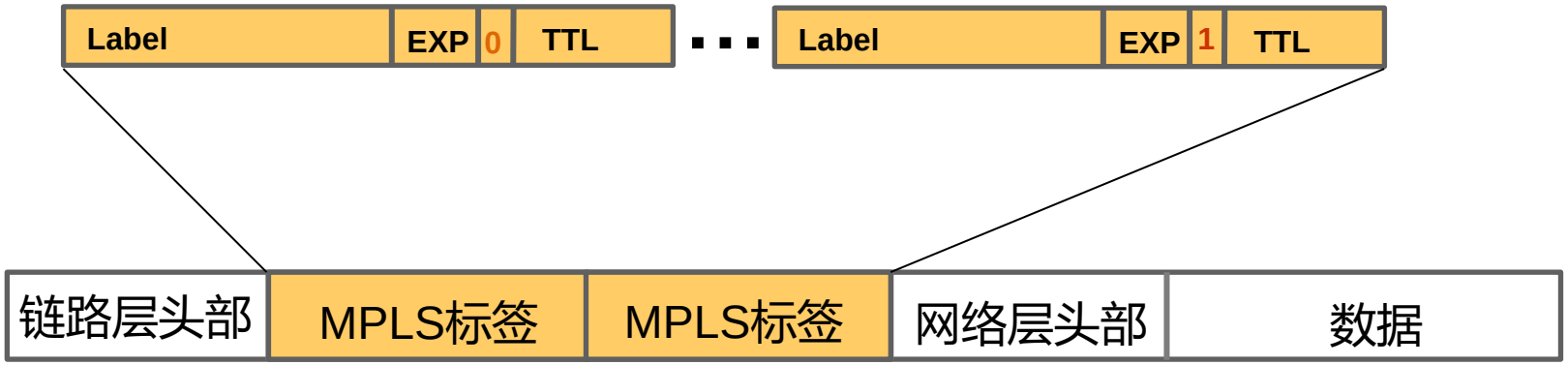
目录

MPLS介绍

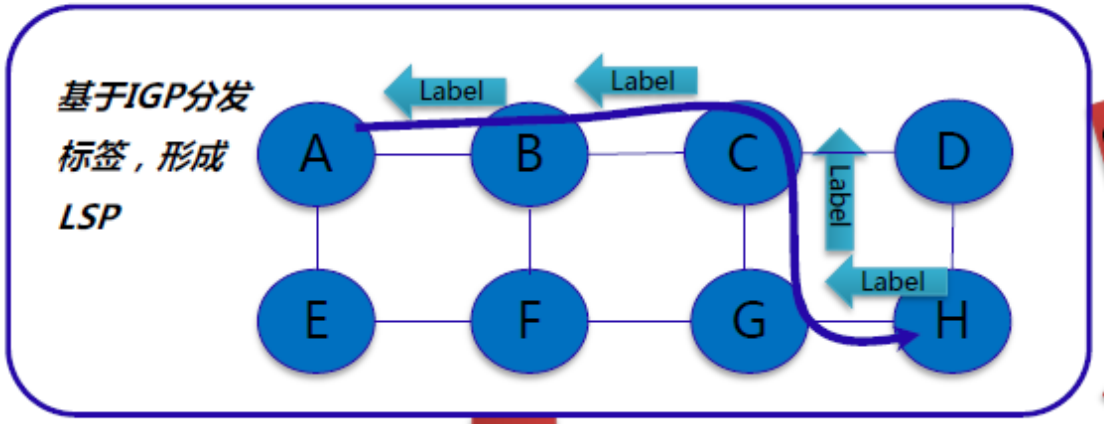
- MPLS标签位于链路层和网络层之间



- 通过MPLS标签的S值可以实现多层MPLS标签嵌套

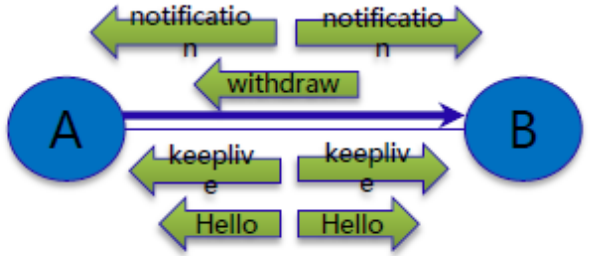


LDP困境



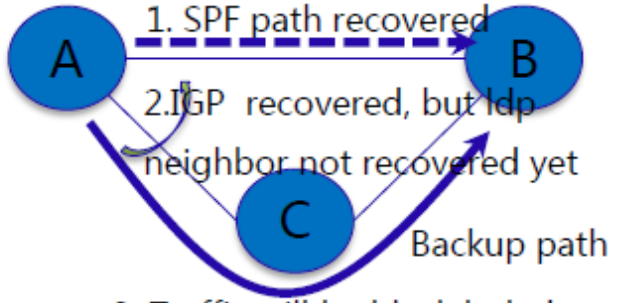
defects

11 types of LDP messages added, occupy the CPU and bandwidth



defects

➢Based on IGP, traffic will be black holed when LDP&IGP asynchronous.

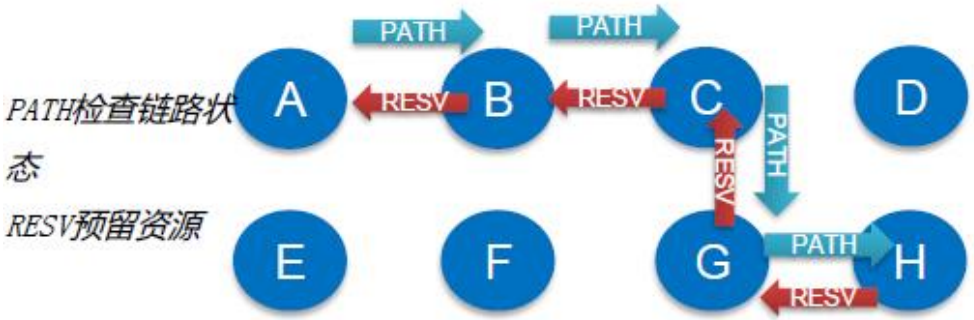


Benefits

- LDP（Label Distribute Protocol），标签分发协议
- 根据不同的目的地址来分配标签，使用标签转发代替了IP转发，隔离公网私网路由
- 跟随路由的最佳路径选择转发路径，支持ECMP
- 配置简单

LDP基于IGP做最短路径转发，无法实现流量工程

RSVP-TE困境

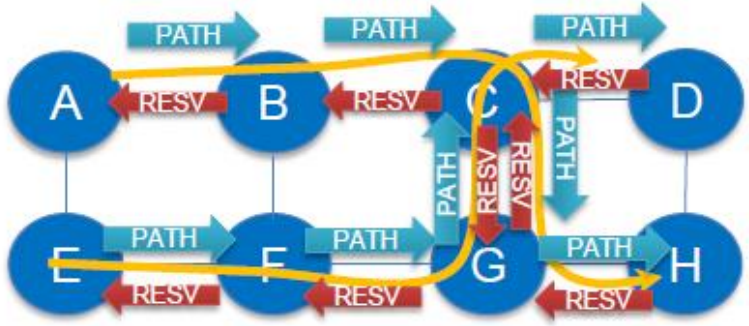


Benefits

- 为了解决传统IP网络只能最优路径转发，无法规划路径的问题，引入了RSVP-TE。
- RSVP-TE引入带来了很多好处，可以做路径的显式规划、带宽资源预留以及多种的保护。

Pain Spot of RSVP-TE

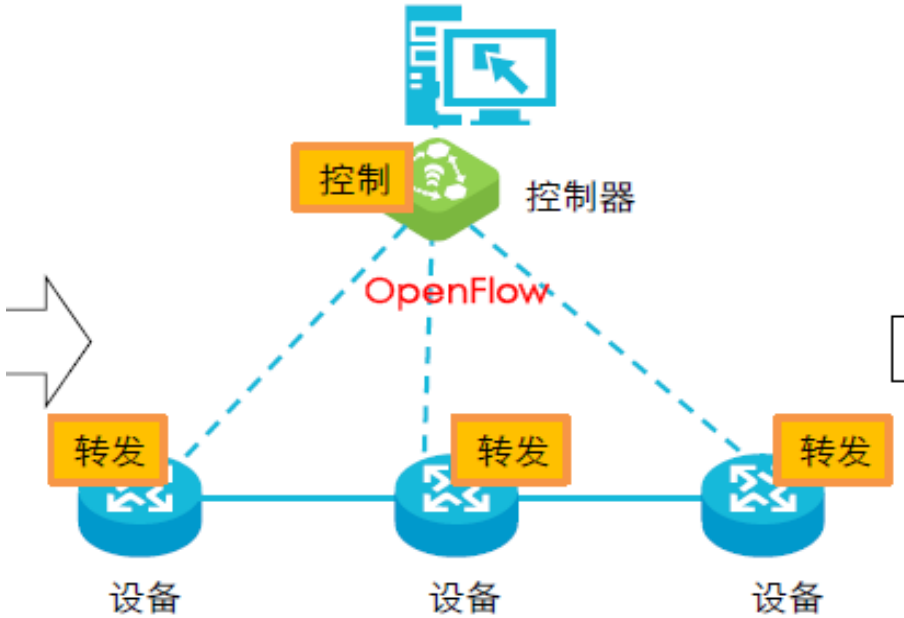
Question



- RSVP依赖IGP建立后，会维护自己的邻居、链路状态，控制面复杂化，增加网络维护、问题定位复杂度。
- RSVP-TE需要建立多条隧道来完成ECMP功能。
- LSP状态维护需要PATH、RESV不断刷新，业务增多后，带来了中间点的性能问题（承受业务叠加），如图中的C/G。
- 而且RSVP-TE隧道的配置量，一直被运营商所诟病。平均一个隧道要8条配置。

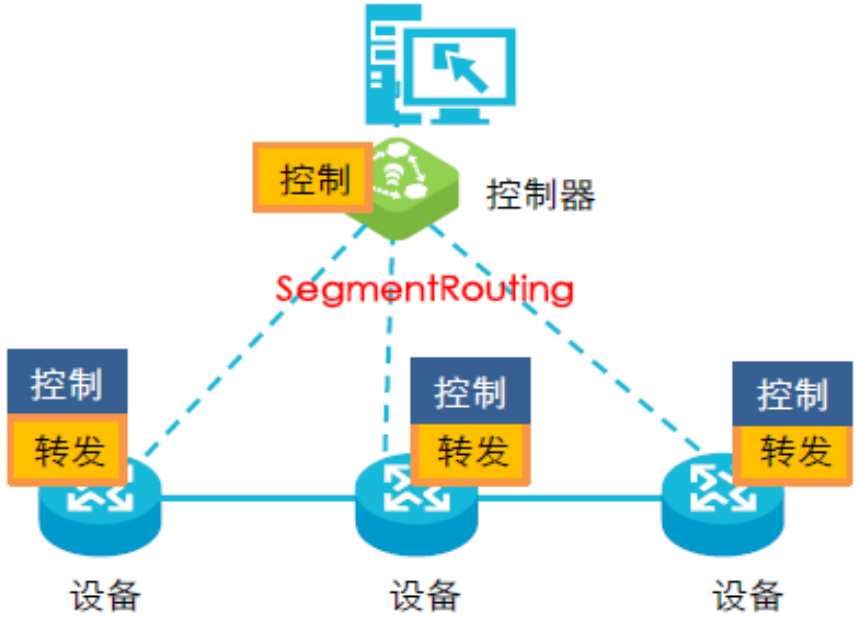
具备路径规划，资源预留的能力，但配置和状态维护太过复杂，不具备可扩展性

SDN的影响



革命型SDN网络

- 网络故障依赖控制器才能恢复，可靠性差
- 大规模网络流表数量大，流表下发速率有性能瓶颈
- 对传统网络是一个颠覆，运营商现网改造及规模应用难度极大



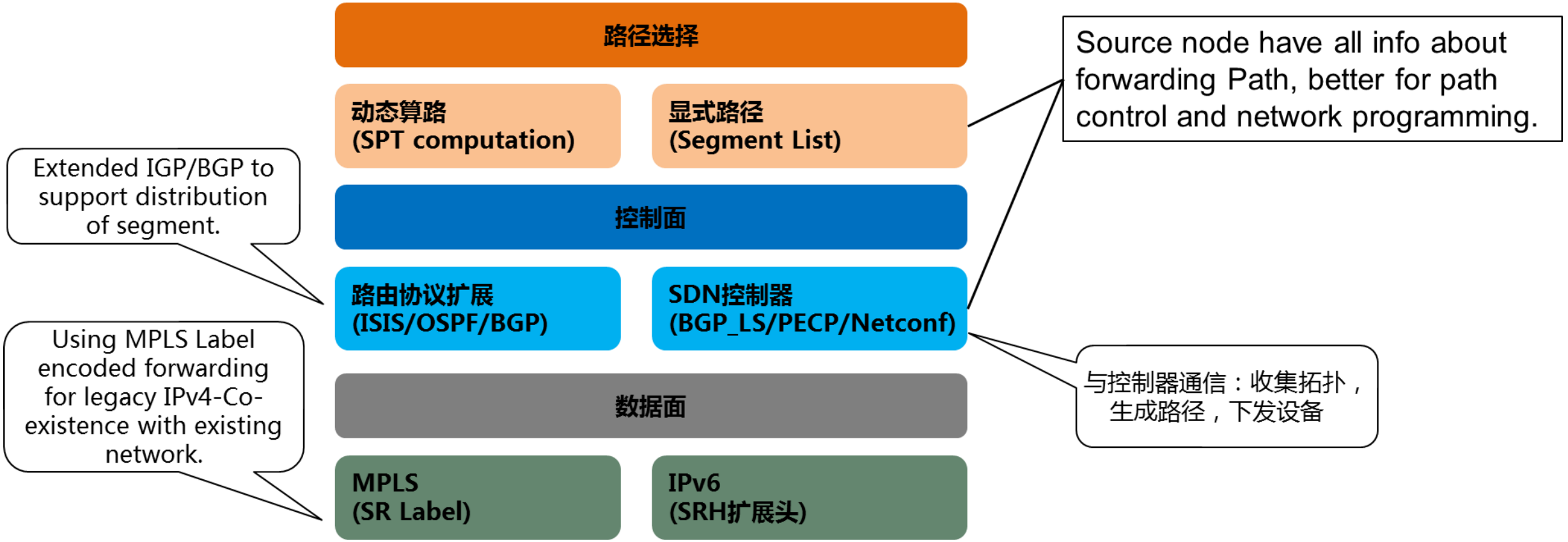
增量型SDN网络

- 通过对现有协议进行扩展，能更好的平滑演进
- 提供集中控制和分布式之间的平衡
- 采用源路由技术，提供网络和上层应用快速交互的能力

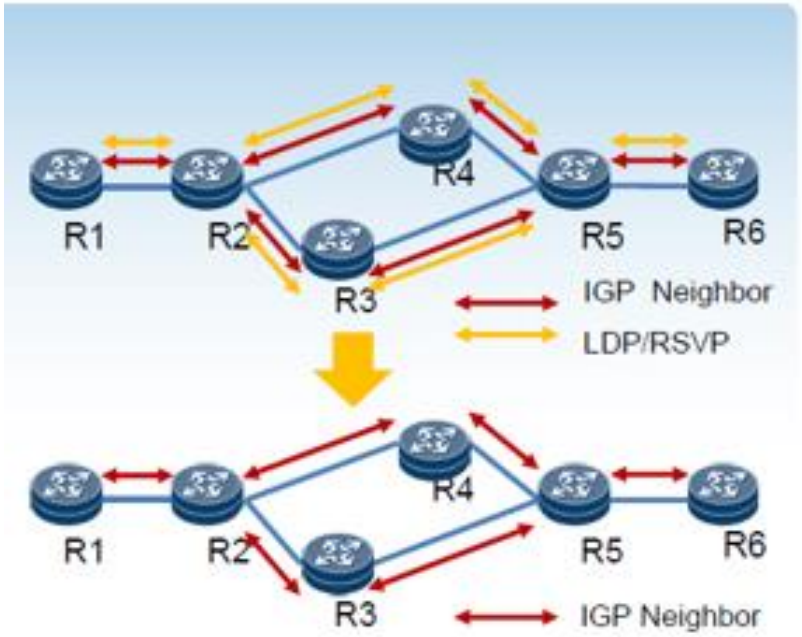
SR技术框架

Segment Routing定义

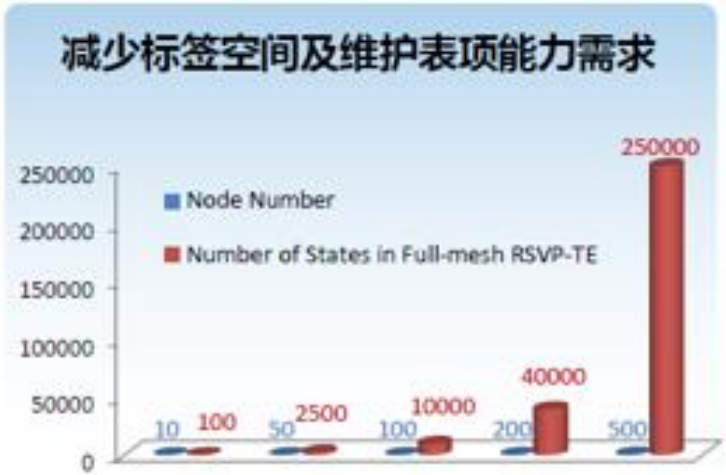
- 一种只需在源（显式路径加载的节点）节点给报文增加一系列的段标识，便可知道报文转发的技术方案。
- Cisco & Orange首次与2013年3月在IETF提出，同年7月组建SPRING（Source Packet Routing in Networking）工作组。



SR技术优势



- 运维简化，网络更可靠，有效减少网络OPEX；
- 可扩展性好，天然和IGP同步，ECMP天然支持



- SR标签表： N+A vs N*A
- SR TE状态维护对比RSVP TE维护： N+A vs N^2
  - N： 代表网络中的节点个数；
  - A： 代表各节点上的SR链路数

- 更灵活高效，业务路径调整只需头节点上感知
- 更有效，控制器集中算路
- 链路资源充分利用，SR TE与SDN的完美兼容
- 支持对IPv6网络的SRH编排

简化、可扩展、高性能、高可用、全连接，更好的实现SDN

SR和传统MPLS对比

	SR	LDP	RSVP-TE
使用的协议	IGP扩展	IGP+LDP	IGP+RSVP
LSP信息是否与IGP同步	是	否，可能存在流量黑洞	否，可能存在流量黑洞
配置复杂度	简单	简单	复杂
是否支持ECMP	支持	支持	需额外配置
隧道维护	仅头节点	所有节点	所有节点
标签属性	全局或本地	本地	本地
是否支持SDN路径调优	是	否	部分（PCEP）

SR → SRv6

SR-MPLS（成熟应用）

简单

- 简化网络协议，不需要运行LDP/RSVP/T-LDP等协议，通过IGP/BGP传递标签
- 状态维护少，设备的增减对网络的影响小

SDN

- 自动化配置与管理
- 路径可编排，流量灵活与智能调度

应用场景广

- 能实现TI-LFA FRR，在任何拓扑下支持FRR
- 平滑升级支持MPLS VPN专线业务

SRv6（目标方案）

极简

- 控制面精简到只剩下IPv6 IGP(ISISv6/OSPFv3)/BGP4+
- 转发面全部IPv6

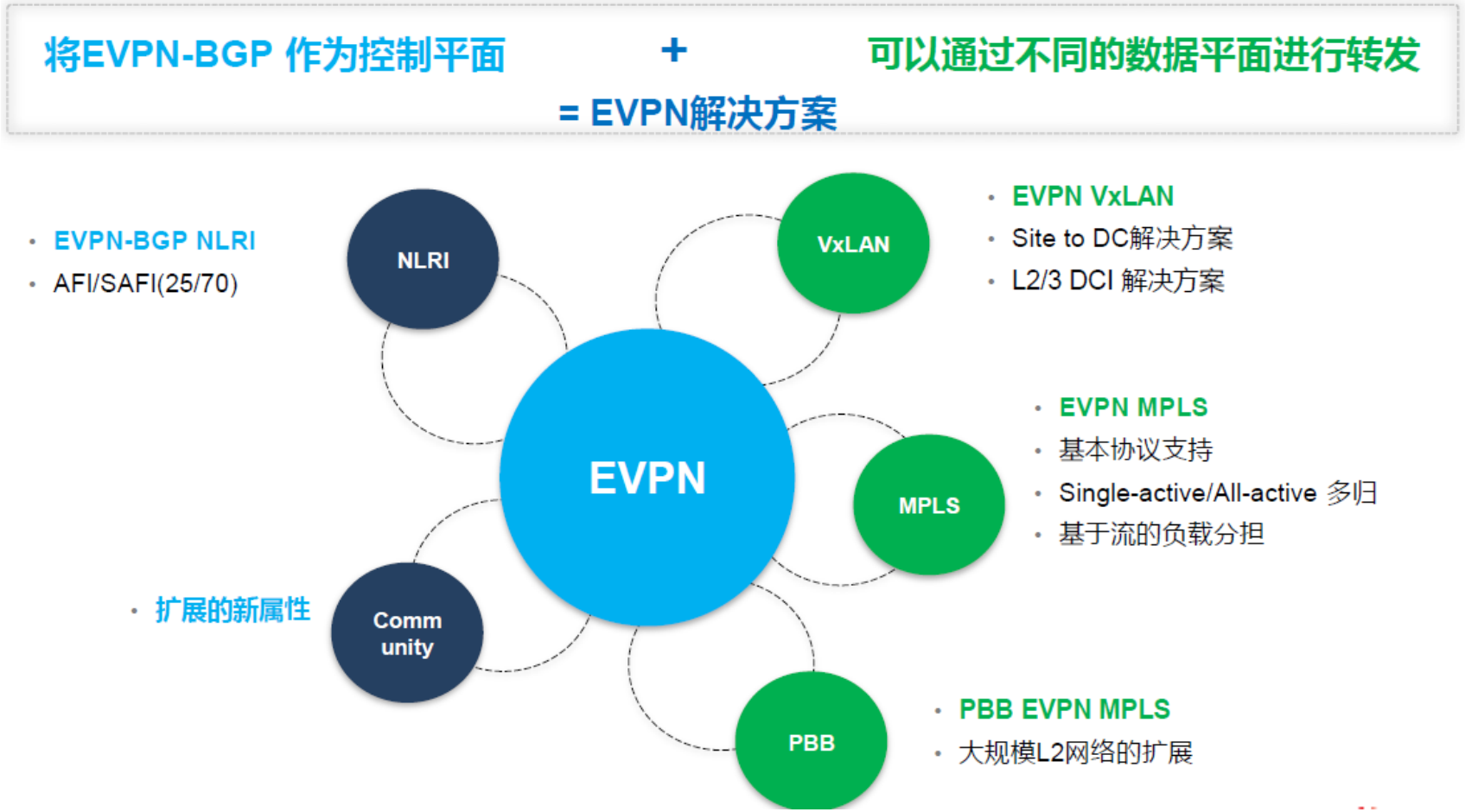
SDN

- 自动化配置与管理
- 路径可编排，流量灵活与智能调度
- 网络可编程，无缝支持服务链与跨域互连能力

端到端业务应用部署

- 能实现TI-LFA FRR，在任何拓扑下支持FRR
- 主机侧Linux-网络侧全程部署SRv6

EVPN



EVPN的优势：

- EVPN通过扩展BGP协议使二层网络间的MAC地址学习和发布过程从数据平面转移到控制平面，从而管理MAC地址实现负载分担（相同MAC地址不同下一跳）
- 通过使用EVPN技术，利用BGP的RR特性，运营商骨干网上的PE设备之间不再需要建立全连接，只需要部署反射器反射EVPN路由即可，从而减少了网络部署成本
- 采用BGP发布MAC，避免流量泛洪方式的MAC扩散

目录

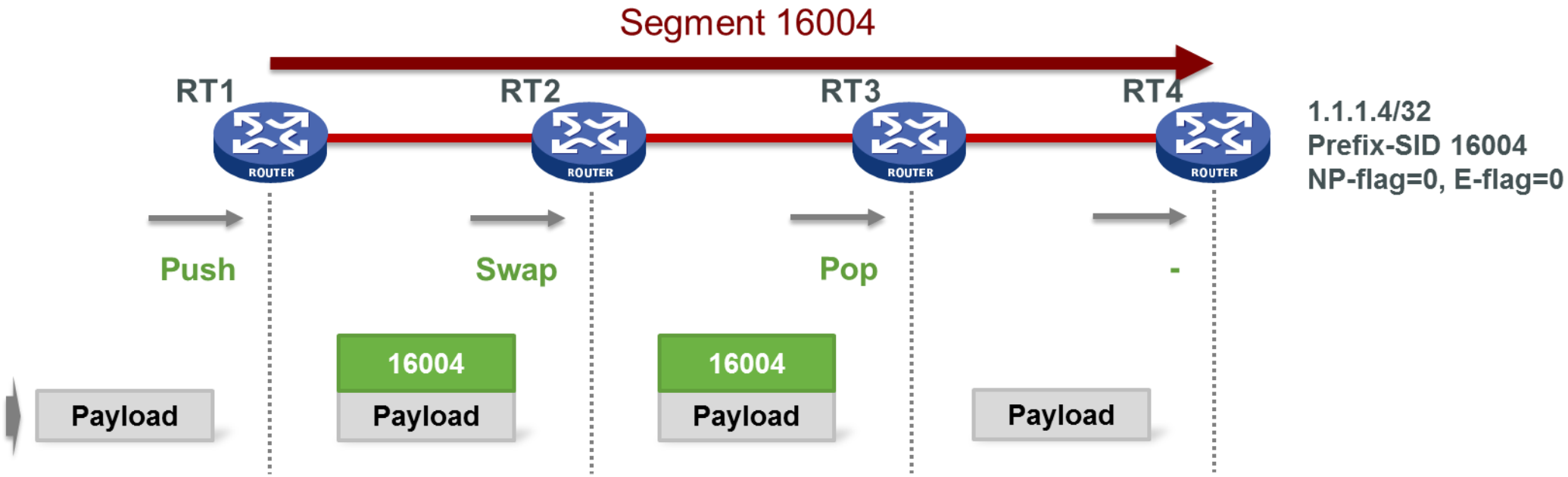
MPLS SR概念

- SRGB（段路由全局标签段）
  - 用户指定的为Segment Routing预留的本地标签集合
- Prefix Segment（前缀标签）
  - 用于标识网络中的某个目的地址前缀
  - 通过IGP协议扩散到其他网元，全局可见，全局有效
  - 接收端会根据自己的SRGB计算实际标签值用于生成MPLS转发表项
- Node Segment（节点标签）
  - 特殊的Prefix Segment，用于标识特定的节点
  - 节点的环回口下配置IP地址作为前缀，这个节点的Prefix SID实际就是Node SID
- Adjacency Segment（邻接标签）
  - 用于标识网络中的某个邻接
  - 通过IGP协议扩散到其他网元，全局可见，本地有效
  - 通过Adjacency Segment ID（SID）标识。 Adjacency SID为SRGB范围外的本地SID

MPLS SR概念

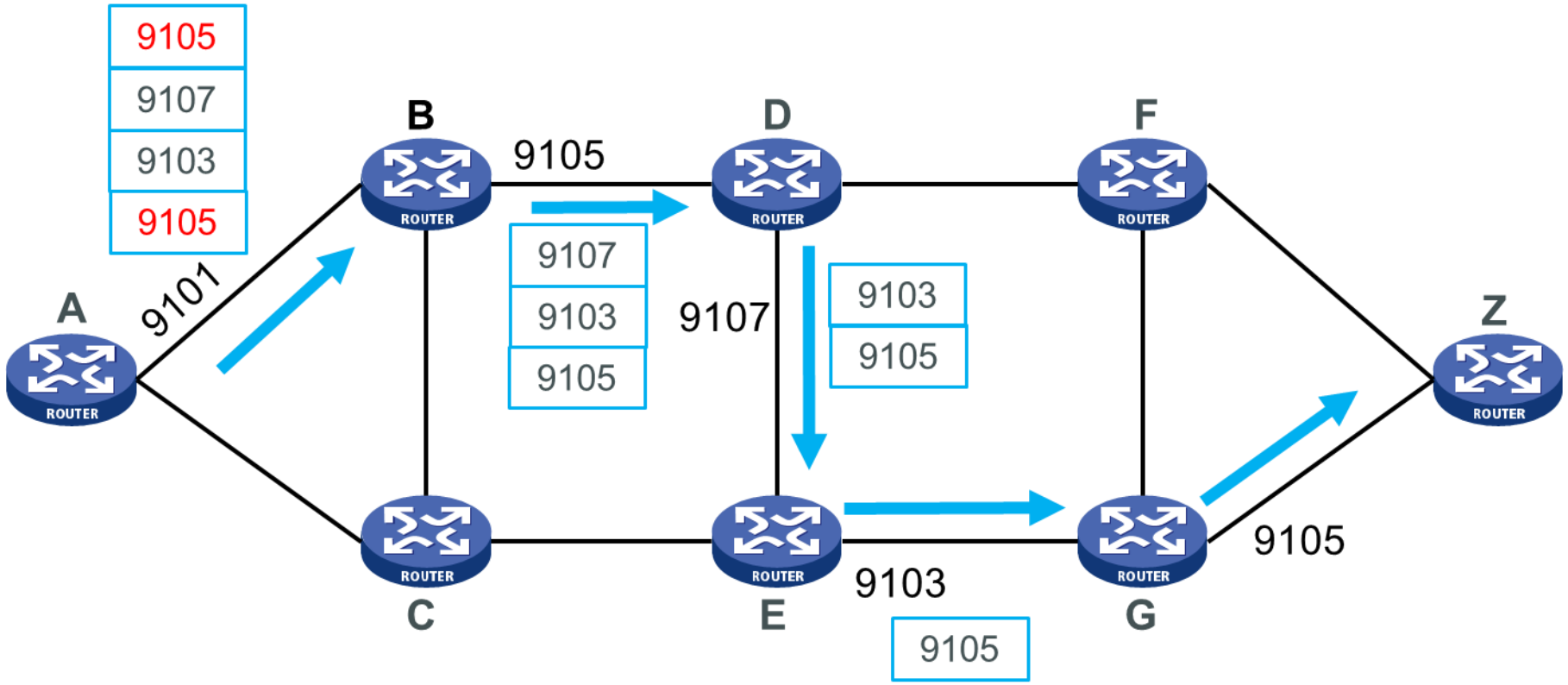
- Segment List（段路由列表）
  - 用来表示报文转发路径的一个有序的Segment列表，在MPLS中为标签栈，在IPv6中为封装在一个Segment Routing Header（SRH）中的IPv6 Address列表
  - 采用Global segment与Local Segment联合组成路径
- SR LSP（基于段路由的LSP）
  - 使用SR技术建立的标签转发路径，报文的路径称为SRLSP；由一个Prefix或Node Segment指导数据包转发
  - SR LSP的创建过程和数据转发与LDP LSP类似，这种LSP不存在Tunnel接口
- SR Tunnel
  - 在网络头节点上，将Segment List封装到报文头中的隧道，可以由管理员手工创建，也可以是控制器通过NETCONF或PCEP等接口协议自动创建。一个SR隧道既可用于TE流量工程应用，也可用于FRR等目的

MPLS SR转发



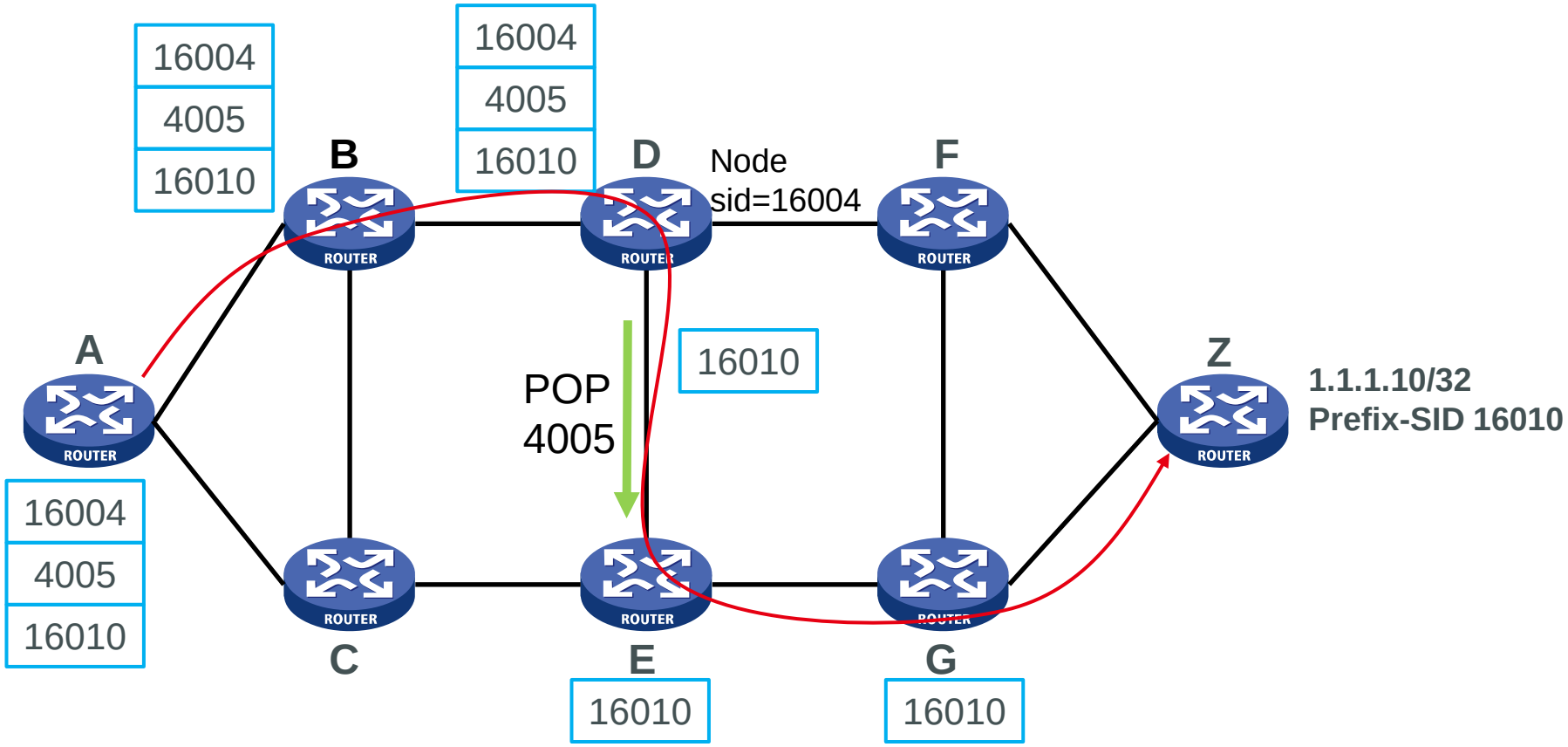
- 节点4由IGP通告其Loopback ipv4 prefix1.1.1.4/32及其prefix-SID 16004
- 节点4请求默认的PHP功能(noPHP-flag=0, ExpNull-flag=0)

MPLS SR转发



节点为每一条链路分配邻接标签，比如为D-E链路分配了9107，D节点将所有的邻接标签通过IGP协议的扩展TLV通告到全网，MPLS数据层面，整个网络中所有节点只安装自己分配的邻居标签转发表，不会安装其它设备通告的邻接标签，比如D分发出去的9107邻接标签只在D节点上安装。

MPLS SR转发



邻接标签加节点标签转发的方式。

SRGB

- 强烈推荐在所有节点上使用相同的SRGB
  - 这是所有的运营商期望的模式
  - 简单、直接
  - 虽然SR支持使用不同的SRGB，但是对用户来说操作复杂
- 所有节点上的SRGB大小必须相同
  - 不同设备缺省SRGB不一样，我司RA5300和12500R设备缺省为16000-55999，大小为40000
- 非默认的SRGB可以在16000到1010151间分配
  - 不同设备SRGB可配范围不一样，还是得根据产品要求来
- 同一SR域内，每个prefix-sid的index值不一样。

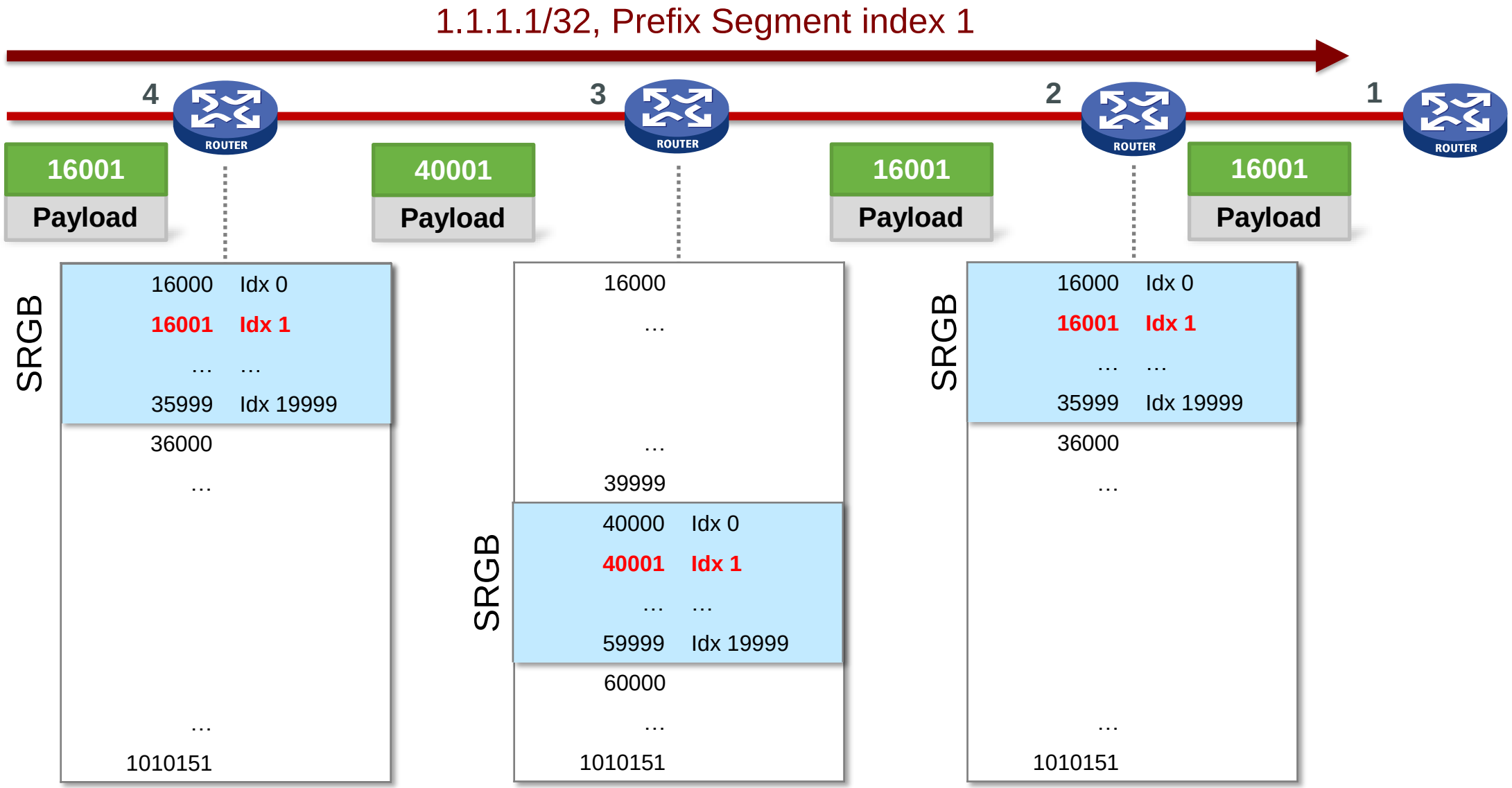
SRGB

```
<MCR002>display isis lsdb verbose     ----->查看同一SR域内网络节点的SRGB情况
LSPID                 Seq Num      Checksum      Holdtime      Length  ATT/P/OL
-------------------------------------------------------------------------------
MER001.00-00   0x00002d39  0x766f          1116          1039      0/0/0
 Source       0100.5322.1001.00
 HOST NAME    MER001
 ... ...
 Router ID    10.53.221.1
 Router capability
    Router ID: 10.53.221.1      Flags (D/S): 0/0
    Segment routing (I/V/H): 1/0/0
    SRGB base: 24000          SRGB range : 20000
```

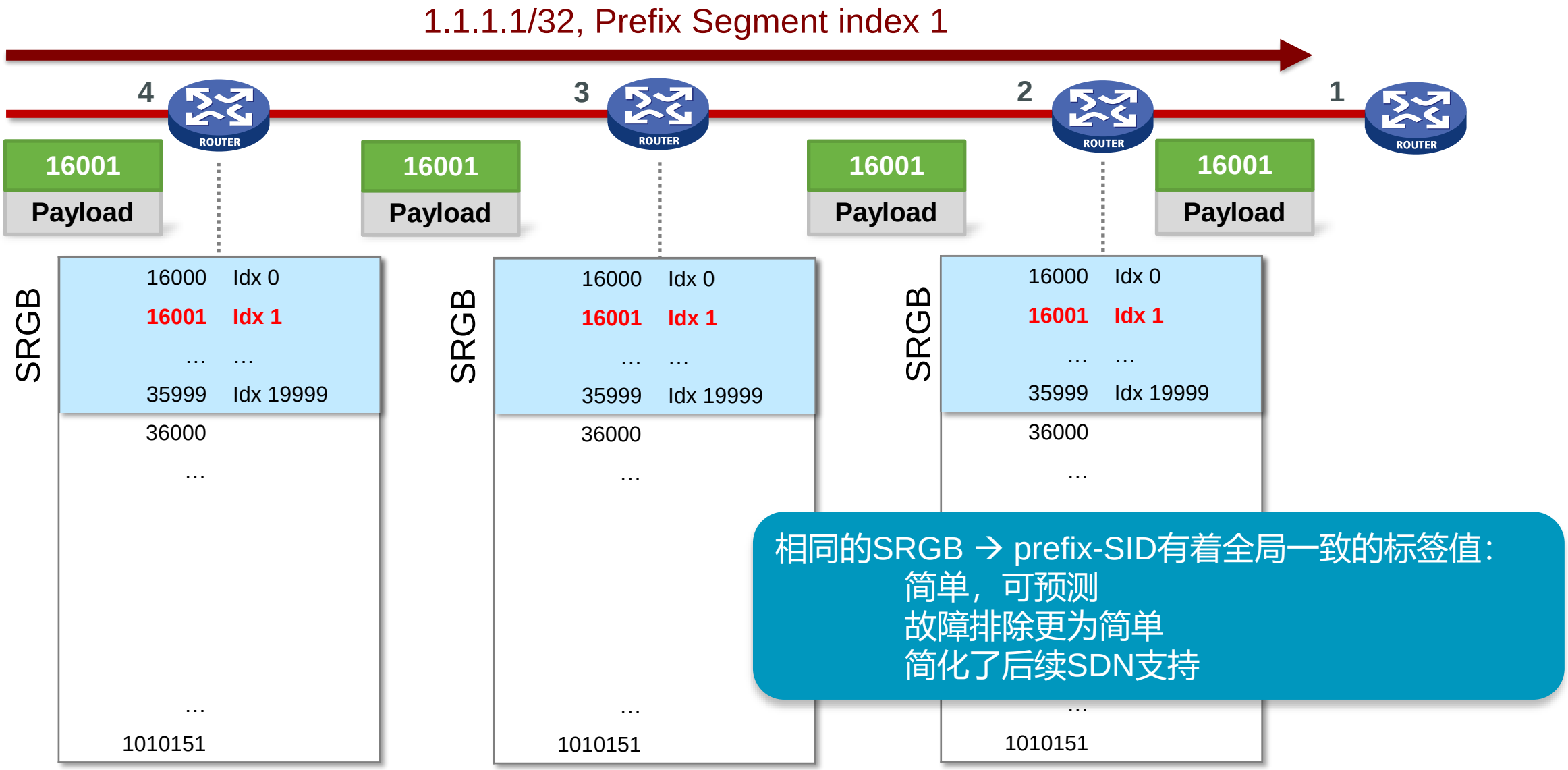
开始标签为24000

数量 = 20000

SRGB


1.1.1.1/32, Prefix Segment index 1
4
3
2
1
ROUTER
ROUTER
ROUTER
ROUTER
16001
Payload
40001
Payload
16001
Payload
16001
Payload
SRGB
16000 Idx 0
16001 Idx 1
… …
35999 Idx 19999
36000
…
…
1010151
SRGB
16000
…
…
39999
40000 Idx 0
40001 Idx 1
… …
59999 Idx 19999
60000
…
1010151
SRGB
16000 Idx 0
16001 Idx 1
… …
35999 Idx 19999
36000
…
…
1010151

SRGB


1.1.1.1/32, Prefix Segment index 1
4
3
2
1
ROUTER
ROUTER
ROUTER
ROUTER
16001
Payload
16001
Payload
16001
Payload
16001
Payload
SRGB
16000 Idx 0
16001 Idx 1
… …
35999 Idx 19999
36000
…
…
1010151
SRGB
16000 Idx 0
16001 Idx 1
… …
35999 Idx 19999
36000
…
…
1010151
SRGB
16000 Idx 0
16001 Idx 1
… …
35999 Idx 19999
36000
…
…
1010151
相同的SRGB → prefix-SID有着全局一致的标签值：
简单，可预测
故障排除更为简单
简化了后续SDN支持

节点标签静态配置

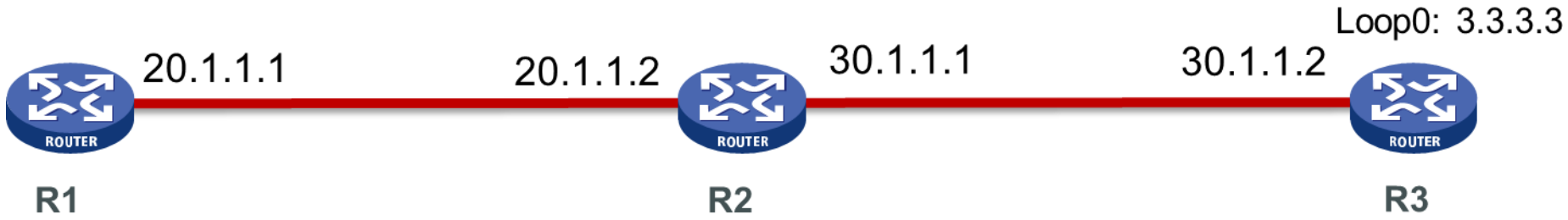
节点标签：在每台SR节点上为目的IP地址前缀静态配置入标签、出标签和下一跳。

[R1]static-sr-mpls prefix prefix-1 destination 3.3.3.3 32 in-label 16003 nexthop 20.1.1.2 out-label 17003

[R2]static-sr-mpls prefix prefix-1 destination 3.3.3.3 32 in-label 17003 nexthop 30.1.1.2 out-label 18003

[R3]static-sr-mpls prefix prefix-1 destination 3.3.3.3 32 in-label 18003

节点标签转发表：设备根据静态配置的入标签、出标签以及下一跳的对应关系形成本地的标签转发表项。



R1节点标签转发表

Inlable	Nexthop	Outlable
16003	R2	17003

R2节点标签转发表

Inlable	Nexthop	Outlable
17003	R3	18003

R3节点标签转发表

Inlable	Nexthop	Outlable
18003	NULL	NULL

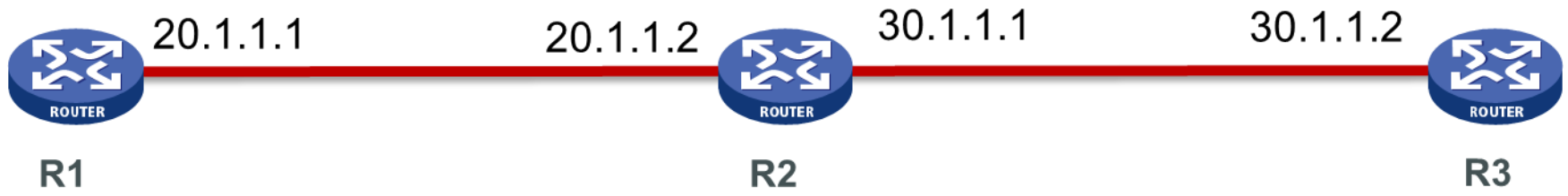
邻接标签静态配置

邻接标签：在每台SR节点上为与邻接设备相连的链路静态配置入标签和下一跳对应关系。

[R1]static-sr-mpls adjacency adjacency-1 in-label 102 nexthop 20.1.1.2

[R2]static-sr-mpls adjacency adjacency-1 in-label 203 nexthop 30.1.1.2

邻接标签转发表：设备根据静态配置的入标签和下一跳的对应关系形成本地的标签转发表项。



R1节点标签转发表

Inlable	Nexthop
102	R2

R2节点标签转发表

Inlable	Nexthop
203	R3

ISIS FOR SR

SR使用IGP协议进行拓扑信息、前缀信息、 SRGB和标签信息的通告。 IGP协议为了完成上述功能，对于协议报文的TLV进行了一些扩展。 IS-IS协议主要定义了针对SID和网元SR能力的子TLV（Sub-TLV）

名称	作用
Prefix-SID SubTLV	用于通告SR的Prefix SID
Adj-SID Sub-TLV	用于在P2P网络中通告SR的Adjacency SID
LAN-Adj-SID Sub-TLV	用于在LAN网络中通告SR的Adjacency SID
SID/Label SubTLV	用于通告SR的SID或MPLS Label
SID/Label Binding TLV	用于通告前缀到SID的映射
SR-Capabilities Sub-TLV	用于对外通告自己的SR能力
SR-Algorithm SubTLV	用于对外通告自己使用的算法

ISIS FOR SR

节点标签：每个SR节点手工为自己的Loopback地址指定全局唯一的index值，通过ISIS通告各个SR节点的SRGB标签段和Index值。

节点标签转发表：入标签为本地SRGB标签段基值 + Index，出标签为下一跳的SRGB基值 + Index。

R1配置示例

```
isis 1
segment-routing global-block 16000 16999
 address-family ipv4 unicast
  segment-routing mpls
interface xxx
 isis enable 1
```

R2配置示例

```
isis 1
segment-routing global-block 16000 16999
 address-family ipv4 unicast
  segment-routing mpls
interface xxx
 isis enable 1
```

R3配置示例

```
isis 1
segment-routing global-block 16000 16999
 address-family ipv4 unicast
  segment-routing mpls
interface xxx
 isis enable 1
interface loopback 0
 isis prefix-sid index 3
```

R1节点标签转发表

Inlable	Nexthop	Outlable
16003	R2	16003

R2节点标签转发表

Inlable	Nexthop	Outlable
16003	R3	16003

R3节点标签转发表

Inlable	Nexthop	Outlable
16003	NULL	NULL

ISIS FOR SR

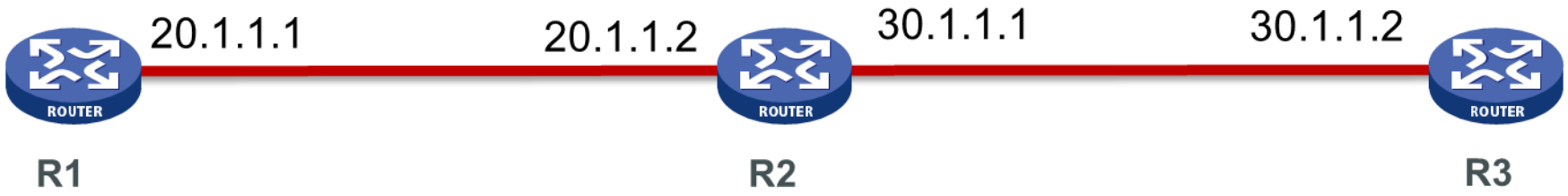
邻接标签：SR节点开启邻接标签分配功能，给直连链路分配邻接标签。

邻接标签转发表：设备根据动态分配的入标签和下一跳的对应关系形成本地的标签转发表项。

R1配置示例
```
isis 1
 address-family ipv4 unicast
  segment-routing mpls
  segment-routing adjacency enable
interface xxx
 isis enable 1
```

R2配置示例
```
isis 1
 address-family ipv4 unicast
  segment-routing mpls
  segment-routing adjacency enable
interface xxx
 isis enable 1
```

R3配置示例
```
isis 1
 address-family ipv4 unicast
  segment-routing mpls
  segment-routing adjacency enable
interface xxx
 isis enable 1
```



R1节点标签转发表

Inlable	Nexthop
1150	R2

R2节点标签转发表

Inlable	Nexthop
1151	R3

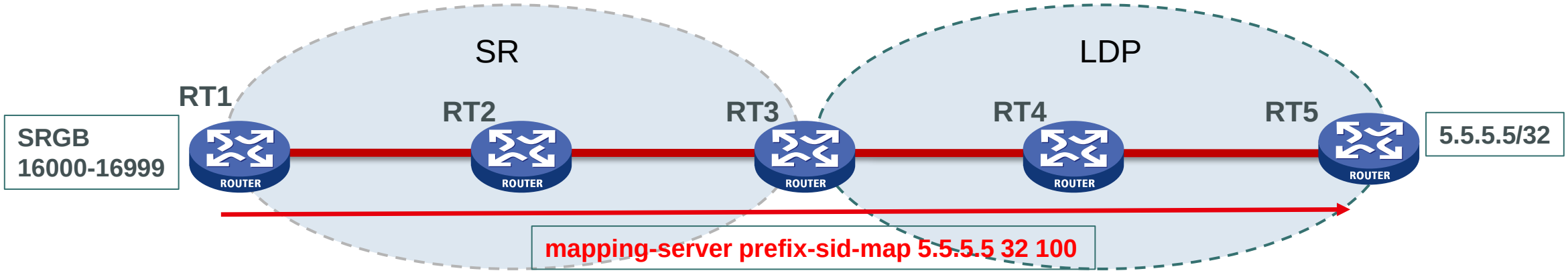
OSPF FOR SR

在OSPF协议中，通过定义一些新的Opaque LSAs，用来扩展包括SID、SR能力等的通告，这些新的Opaque LSAs是对现有LSA的补充，而不是取代

名称	作用
SR-Algorithm TLV	用于对外通告自己使用的算法
SID/Label Range TLV	用于通告SR的SID或MPLS Label范围
SRMS Preference TLV	用于通告网元做为SRMapping Server的优先级
SID/Label Sub-TLV	用于通告SR的SID或MPLS Label
Prefix SID Sub-TLV	用于通告SR的PrefixSID
Adj-SID SubTLV	用于在P2P网络中通告SR的Adjacency SID
LAN Adj-SID Sub-TLV	用于在LAN网络中通告SR的Adjacency SID

SR和LDP互通

- 在MPLS SR和LDP共存的网路环境中，需要解决SR网络和LDP网络之间互通的问题。为了实现SR与LDP网络互通，有如下几种方式
  - SR to LDP：通过将LDP网络的前缀映射为SR网络的SID，实现数据流量从SR网络转发到LDP网络
  - LDP to SR：通过IGP通告SID，将SID和LDP标签关联，实现数据流量从LDP网络转发到SR网络
  - SR over LDP：SR网络跨越LDP网络交互数据流量



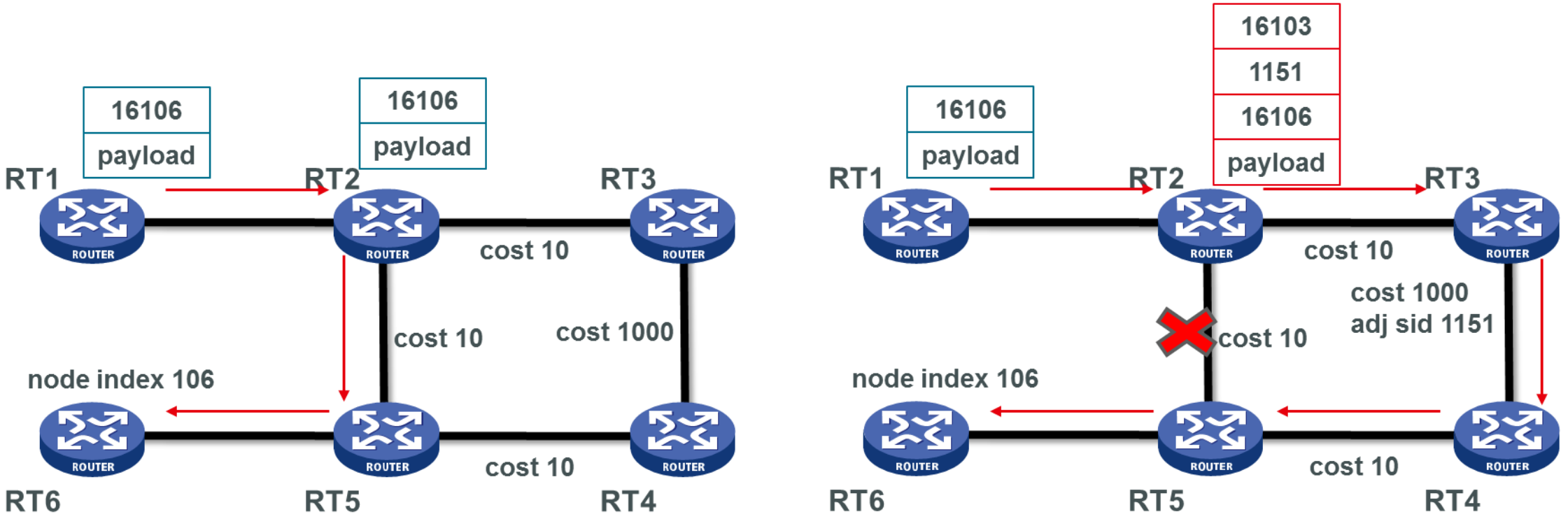
Type	In label	Out label
SR	16100	16100

Type	In label	Out label
SR	16100	-
LDP	-	20

Type	In label	Out label
LDP	20	-

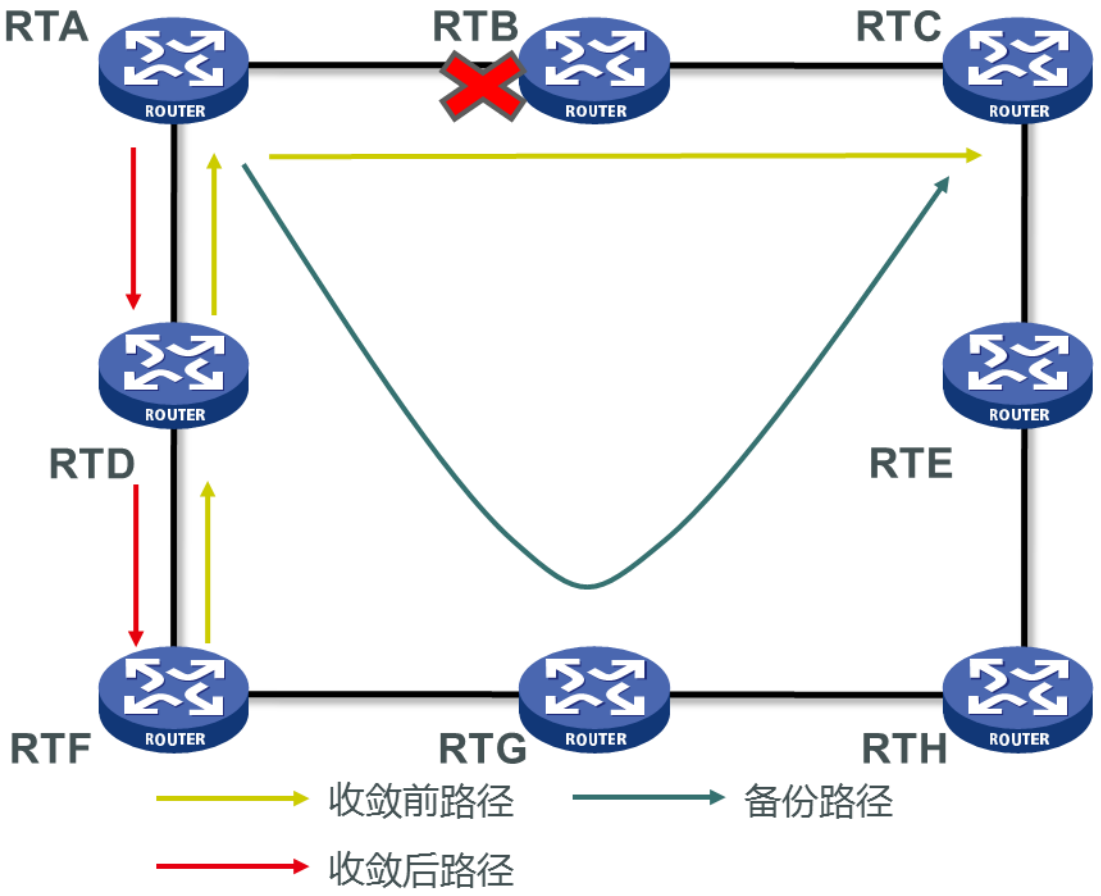
TI-LFA FRR

TI-LFA FRR (Topology-Independent Loop-free Alternate FRR)能为Segment Routing隧道提供链路及节点的保护。当某处链路或节点故障时，流量会快速切换到备份路径，继续转发。从而最大程度上避免流量的丢失



SR防微环

IGP协议的链路状态数据库是分布式的，在无序收敛时可能会产生环路。但这种环路会在转发路径的设备都完成收敛之后消失，这种暂态的环路被称为微环（MicroLoop）。微环可能导致网络丢包、时延抖动和报文乱序等一系列问题，所以必须予以重视，分为正切防微环和回切防微环



左图正切防微环举例：

- 正常流量，A → B → C；
- B故障后，A最先完成收敛，D、F完成收敛慢，A使用备份路径转发到C，报文经过D、F时，在A和F之间形成环路；
- 防微环功能，各个设备在感知故障完成收敛之前，仍然使用备份路径转发，并且等待一段时间后，再使用收敛后路径转发。

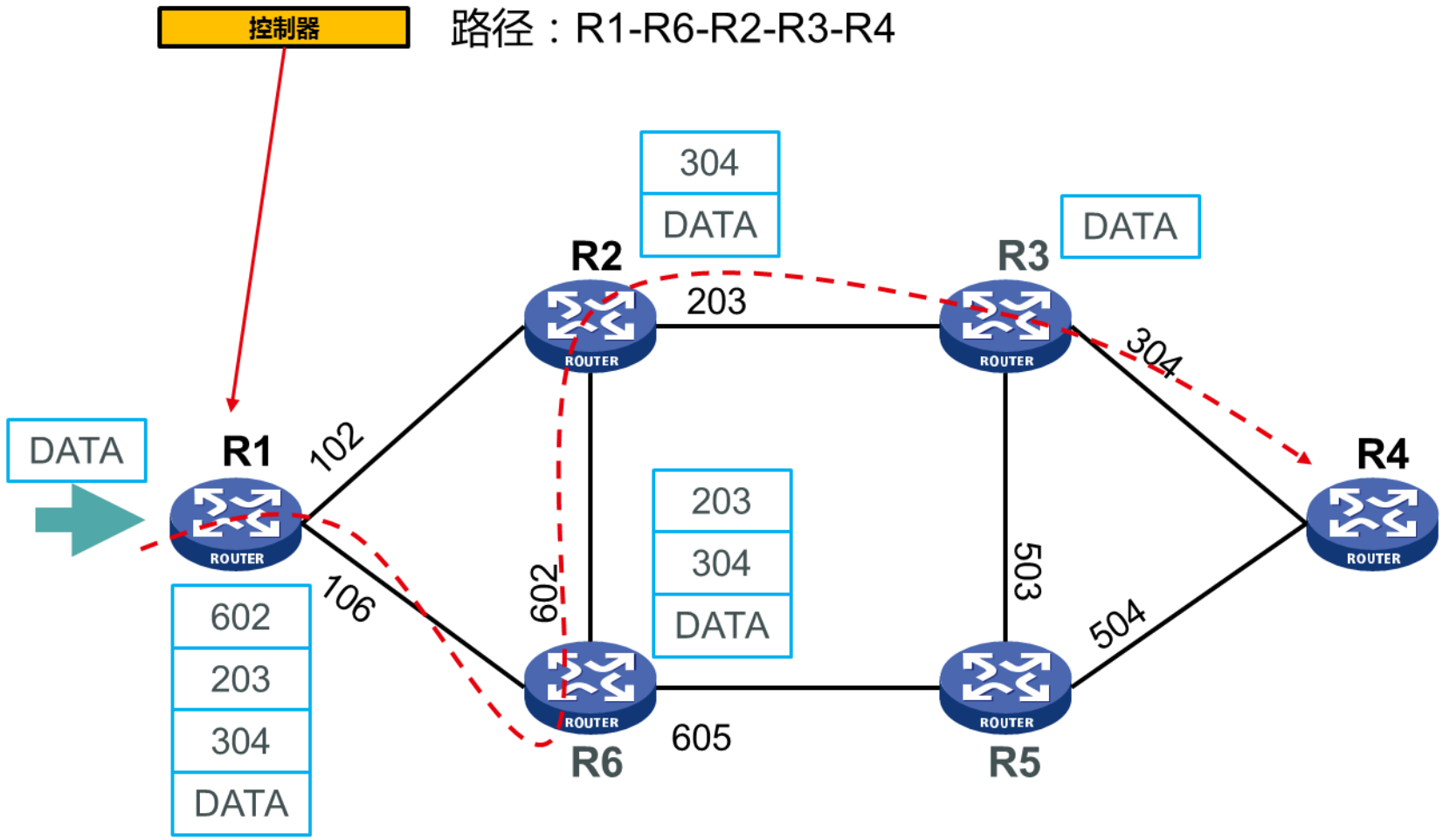
SR-MPLS TE与RSVP-TE比较

比较项	SR-MPLS TE	RSVP-TE
标签分配	通过IGP协议扩展进行标签分配和扩散。每条链路仅分配一个标签，所有的LSP共用这一个标签，减少了标签资源的占用，减轻了标签转发表的维护工作量。	通过MPLS协议扩展进行标签分配和扩散。每条LSP分配一个标签，当有多条LSP时，同一条链路上需要分配多个标签，标签资源占用的多，标签转发表的维护工作量大。
控制平面	信令控制协议为IGP协议扩展，无需专门的MPLS的控制协议，减少协议数量。	需要RSVP-TE作为MPLS的控制协议，控制平面较复杂。
可扩展性	网络中间设备不感知隧道，隧道信息携带在每个报文中，无需维护隧道状态信息，只需维护转发表项，可扩展性强。	需要维护隧道状态信息，也需要维护转发表项，可扩展性差。
路径调整和控制	网络中间设备不感知隧道，仅通过对入节点的报文进行标签操作即可任意控制业务路径，无需逐节点下发配置。当路径中的某个节点发生故障，首节点自动调整到到备份路径，或者有控制器给首节点下发新路径来转发。	无论是正常业务调整还是故障场景的被动路径调整，都需逐节点下发配置。

SRTE概念

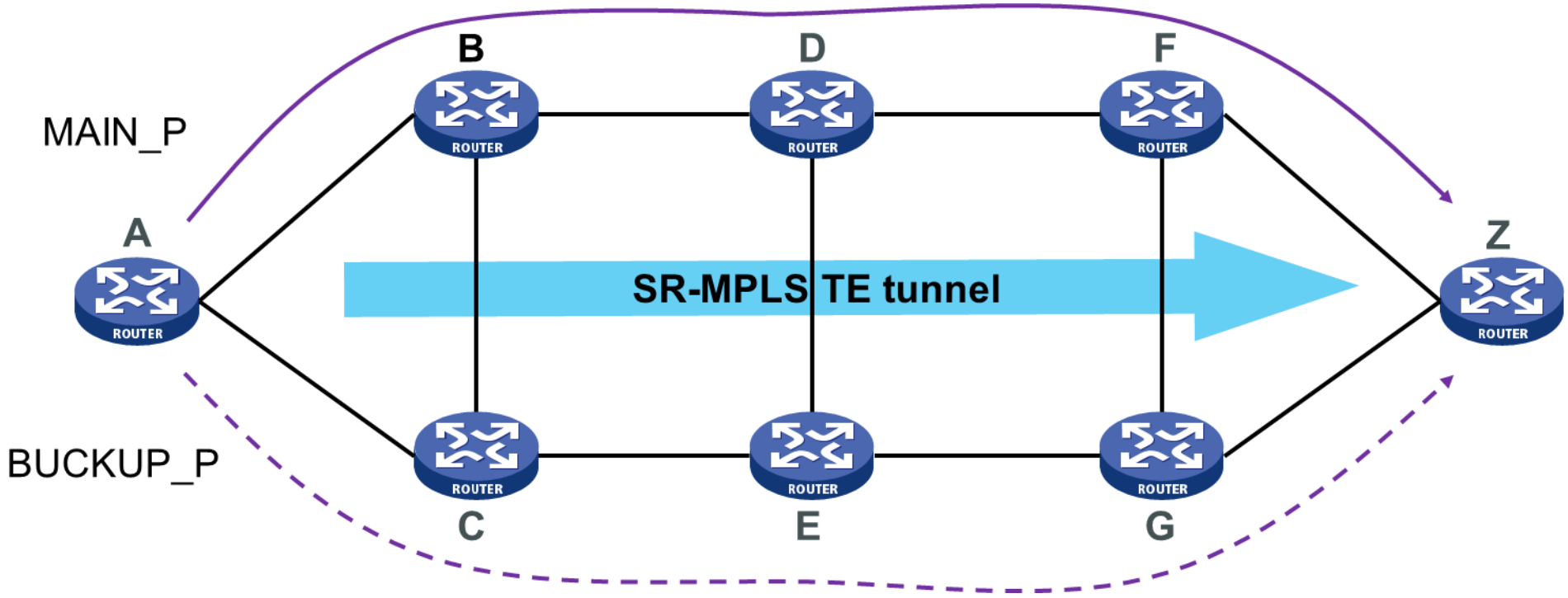
- SR-MPLS BE（Segment Routing-MPLS Best Effort）是指IGP使用最短路径算法计算得到的最优SR LSP，这种LSP不存在Tunnel接口。

- SR-MPLS TE（Segment Routing-MPLS Traffic Engineering）是指基于TE的约束属性，利用SR协议创建的隧道技术。在SR-MPLS TE隧道的入节点上，转发器根据标签栈，即可控制报文在网络中的传输路径。


控制器
路径：R1-R6-R2-R3-R4
304
DATA
DATA
R2
R3
203
304
DATA
R1
102
R4
203
304
DATA
602
203
304
DATA
106
602
503
504
605
R6
R5

BFD FOR SRTE

```
interface Tunnel 1000 mode mpls-te
 ip address unnumbered interface LoopBack0
 mpls te signaling segment-routing
 mpls te path preference 1 explicit-path MAIN_P no-cspf
 mpls te backup-path preference 1 explicit-path BACKUP_P no-cspf
 mpls bfd discriminator local 22 remote 12 template 10
 mpls tunnel-bfd template 50
 destination Z.Z.Z.Z
```



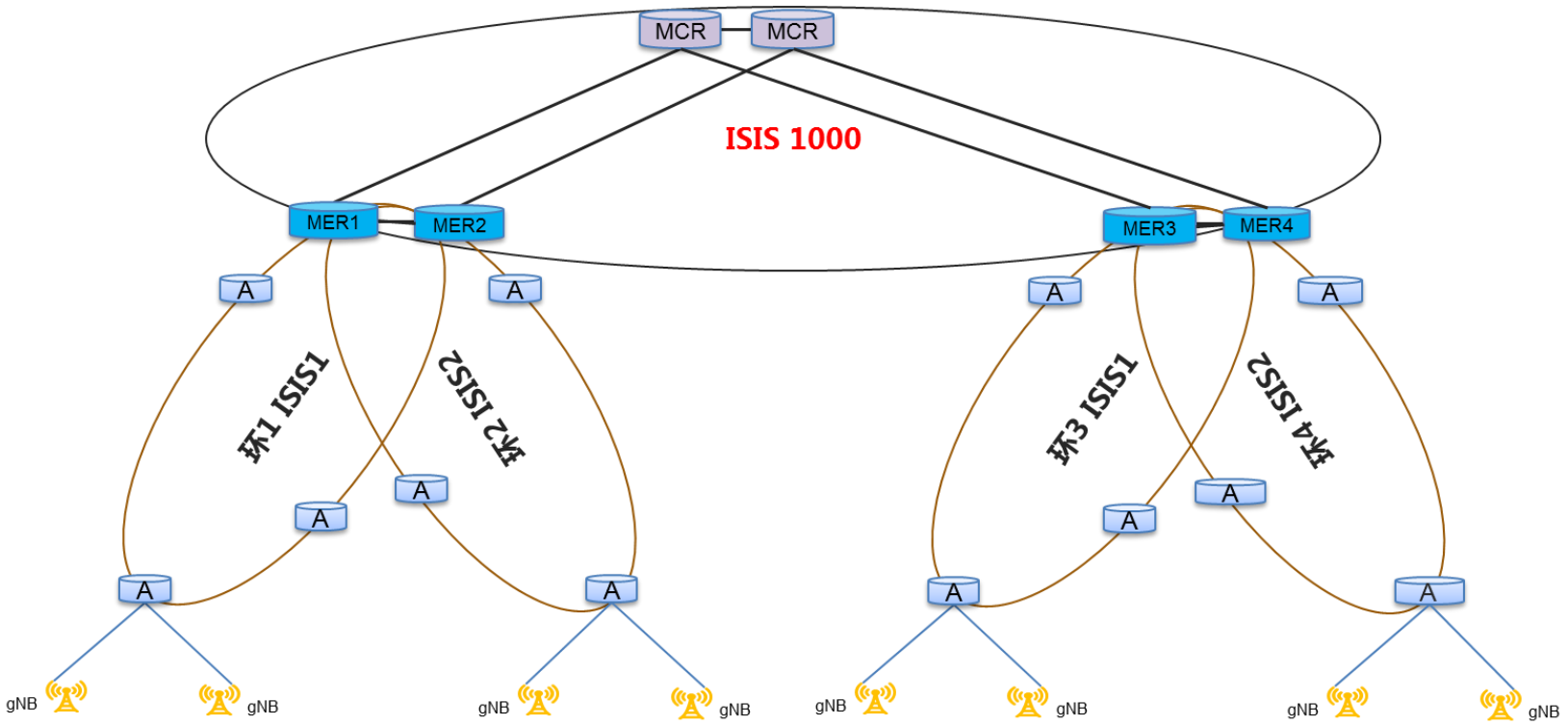
SR流量导入

- 在SR LSP（SR-MPLS BE）或者SR-MPLS TE隧道建立成功以后，还需要将业务流量引入SR LSP或者SR-MPLS TE隧道。常用方法有静态路由、隧道策略、自动路由等。

引入方式\隧道类型	SR LSP	SR Tunnel
静态路由	无Tunnel接口。可以配置静态路由指定下一跳，根据下一跳迭代SR LSP	有Tunnel接口，可以使用
隧道策略	可以使用隧道选择器，不能使用隧道绑定策略	可以使用隧道选择器和隧道绑定策略
自动路由	无Tunnel接口，所以不能使用	有Tunnel接口，可以使用
策略路由	无Tunnel接口，所以不能使用	有Tunnel接口，可以使用

目录

联通智能城域网



- 联通智网按地市为单位分为三级架构，核心、汇聚和接入，为一个SR区域
- 承载电信无线业务，专线业务，后期还会有固网业务
- 我司路由器产品涉及RA5300/5100，交换机涉及S12500R和S6890

课程总结

- 了解SR产生的背景
- 掌握SR的基本概念和原理
- 了解SR的应用案例

THANKS

www.h3c.com

www.h3c.com