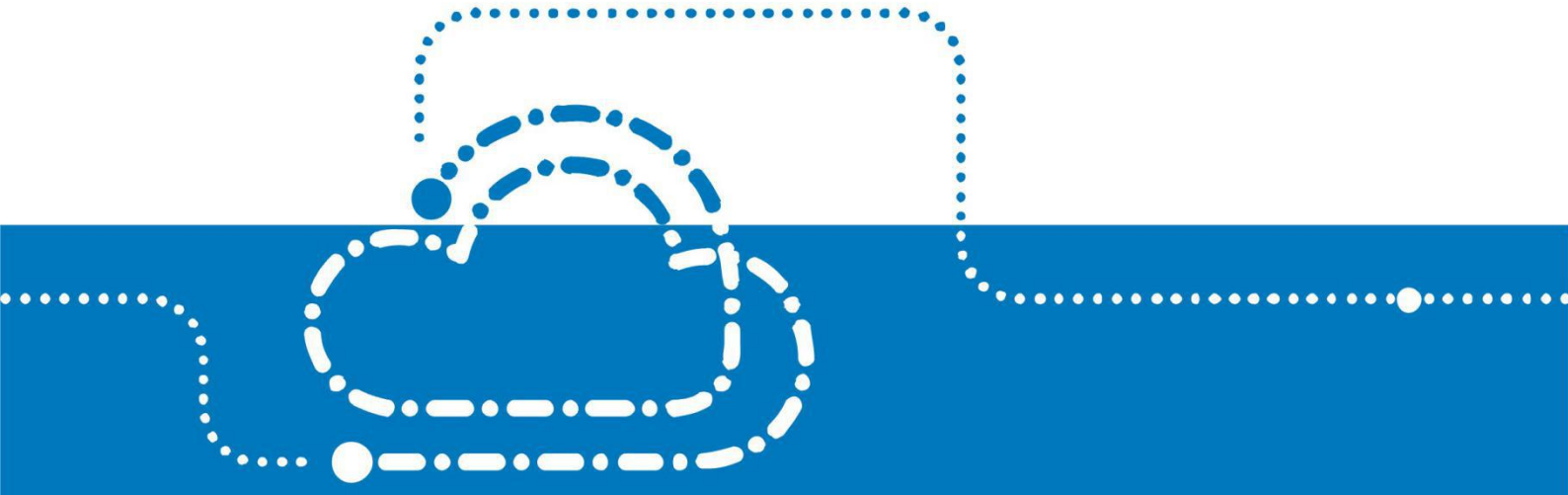
ZTE中兴

SRv6 技术白皮书



© ZTE Corporation. All rights reserved.
版权所有 中兴通讯股份有限公司 保留所有权利

SRv6 技术白皮书

版本	日期	作者	审核者	备注
V1.0	2020/07/27	ZTE	ZTE	新建

关键词：分段路由，SRv6

摘要：SRv6 是指将 Segment Routing 技术应用于 IPv6 数据平面，通过在 IPv6 报文中插入一个路由扩展头 SRH（Segment Routing Header），在 SRH 中包含了由 IPv6 地址列表表示的 segment list，报文的目的地址将逐段的被更新，完成逐段转发。本文介绍了 SRv6 的基本概念、实现方案和基本应用及现状。

© 2020 ZTE Corporation. All rights reserved.
2020 版权所有　中兴通讯股份有限公司　保留所有权利
版权声明：
本文档著作权由中兴通讯股份有限公司享有。文中涉及中兴通讯股份有限公司的专有信息，未经中兴通讯股份有限公司书面许可，任何单位和个人不得使用和泄漏该文档以及该文档包含的任何图片、表格、数据及其他信息。
本文档中的信息随着中兴通讯股份有限公司产品和技术的进步将不断更新，中兴通讯股份有限公司不再通知此类信息的更新。

目录

中兴通讯版权所有未经许可不得扩散

图目录

表目录

中兴通讯版权所有未经许可不得扩散

1 技术背景与分析

1.1 产生背景

移动互联网及云业务的快速发展，对承载网络的基础网络能力提出了更高的要求，包括智能管控、快速服务构建、可信的差分化服务保障、云网一体化等。

SDN+NFV 架构是符合当前及未来业务发展的主流网络架构：SDN 主要实现网络路径集中式规划，NFV 定义虚拟化网络功能云化部署，上层编排器负责业网融合，云网一体规划。

传统的 MPLS，SR-MPLS、SFC 等技术往往仅能支撑路径或者业务独立部署，而 SRv6 独特的 Locator+Function SID 设计同时提供路径和业务的规划能力，为两者的结合提供了完美的手段。在网络和业务编排器的支撑下，SRv6 能够实现云网路径拉通及业务定义能力，为云网融合、端到端业务定义提供了极好的技术选择。

SRv6 一经问世，即赢得了业界的高度关注，被认为是未来网络的基础性支撑技术。运营商网络正面临网络结构和运营模式的巨大变革点，创新力度空前，也面临各种技术挑战。SRv6 应运而生，正处于技术发展和试点应用的前导期，与上层业务模式、管控支撑系统互促互进，逐渐成熟。

1.2 技术特点

SR（Segment Routing）是源路由技术的一种，SRv6 是 SR 技术在 IPv6 网络平面的应用。SRv6 技术在 IPv6 报文中新增 SRH（Segment Routing Header）报头，用于存储 128bit IPv6 地址格式的 SRv6 SID（segment ID）列表。

128 位 SRv6 SID 主要由三部分组成，标识节点位置的 LOC 字段（IPv6 前缀格式，可路由）、标识服务和功能的 FUNC 字段（本地识别）以及存储相关参数的 ARG 字段（见图 1）。一个标准的 SRv6 SID 可以定义特定节点的路径信息及服务和功能信息。

图 1 128 位 SRv6 SID

由基本的 SRv6 定义，我们可以看到 SRv6 的基本特征：SID 可路由、通过 SID

可同时定义节点路径和功能服务信息。

■SID 可路由

SID 的 LOC 即为标准的 IPv6 地址前缀，在 IPv6 网络中可直接路由。

可路由特性对优化现有网络带来较多的便利：

- 在 IPv6 网络中实现离散型 SRv6 部署。非 segment list 的中间节点只要支持 IPv6 转发即可。作为对比，一般情况下，SR-MPLS 需要中间节点普遍支持 MPLS 转发。举例来说，离散型部署特性在仅升级 VPN PE 节点情况下即可实现基于 SRv6 的 VPN 业务，而不必网络中所有节点进行 SRv6 升级。
- SRv6 在跨域 LSP 路径时，不需要复杂的路由扩散，转发面只要 SID IPv6 路由可达，简化跨域路径建立。
- 简化 SR 技术的网络转发面需求，不再需要专门的 MPLS 转发面支撑，有 IPv6 转发面基础即可。

当前的 SRv6 试点和部署主要是利用了 SRv6 的可路由属性，对现有网络业务进行继承和优化。

■通过 SID 可同时定义节点路径和功能服务信息

SRv6 更为核心的特性是融合了路径和业务编排能力，能够预先规划特定的路径以及路径中每一个节点的 Function 动作。yitihua

2 SRv6 技术实现

2.1 概述

2.1.1 SRH 封装

SRv6 扩展了 RFC2460 中的 Routing Header 定义，新增一种 Segment Routing Header(SRH)，以包含 Segment List。如下图，IPv6 Header 中的 Next Header

中兴通讯版权所有未经许可不得扩散

取值为 43 表示下层头为 Routing Extension Header，Routing Extension Header 中的 Routing Type 为 4 表示该 Routing Extension Header 是一个 Segment Routing Header (SRH) 。

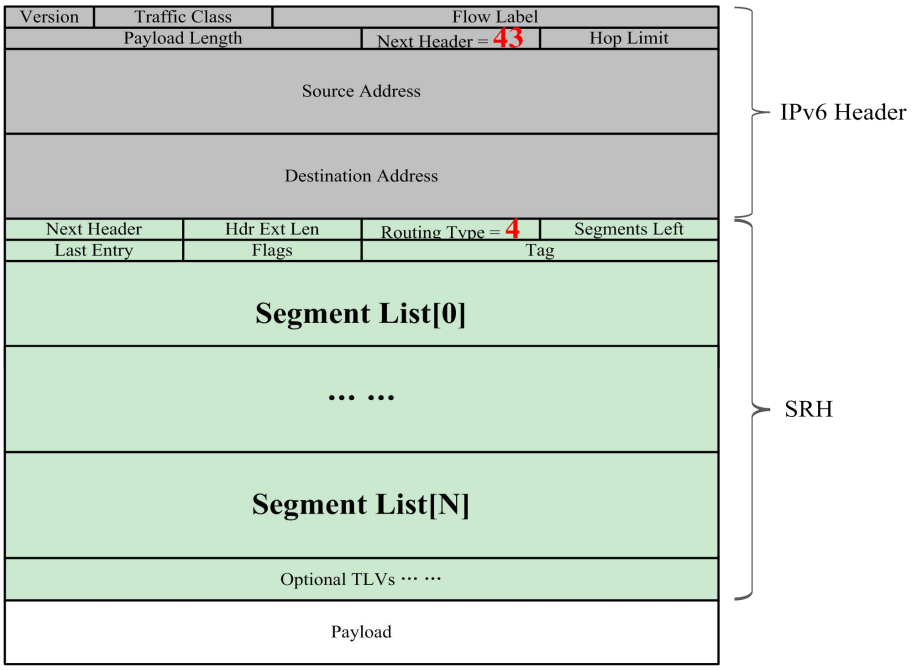


图 2 SRH 封装格式图

SRH 中包含的各字段解释如下：

字段名	解释
Next Header	标识该 SRH 所封装的下层头类型。
Hdr Ext Len	该 SRH 头的长度（以 8 字节为基本单位），不包括前 8 字节。
Routing Type	4
Segments Left	表示还剩下多少个 segment 待转发。头节点在发出报文时，Segment Left 设置为 n-1，n 为 SR policy 的 segment 的总数。
Last Entry	索引，是指 SRH 中的 segment list[]数组中最后一个数组元素对应的下标（由于 SRH 中是逆序存放，Last Entry 实际上是逻辑上第一个 SID 的下标），它的作用是给出 SRH 中实际包含的 segment 的总数，取值一般为 n-1，reduced-SRH 模式时为 n-2，假设 n 为 SR policy 的 segment list 中所包含的 segment 的总数。
Flags	定义了以下标志： O：OAM flag，OAM 报文时设置。

Tag	用于标识报文属于同一类或同一组，比如具有相同的属性集合的报文。
Segment List[ ]	Segment List[0]表示最后一个 segment，Segment List[n-1]表示第一个 segment。
Optional TLVs	目前仅定义了 HMAC TLV 和 PAD TLV，注意这些 TLVs 不用于路由，不用于指导转发。 PAD TLV 用于使得 SRH 整体为 8 字节的整数倍。 HMAC 全称为 keyed Hashed Message Authentication Code，该 TLV 可选，用于校验报文的源头是否允许在报文的 DA 中使用当前 segment，并确保报文在传输时没有被修改。

2.1.2 Segment 类型

SRH 中直接采用 IPv6 地址表示 Segment，可以灵活的支持非常多的类型，将不同类型的 Segment 结合在一起使用以完成特定的功能。大体上，Segment 可以分为两类：表示路径信息的 Segment；表示业务信息的 Segment。

典型的几种 Segment 类型如下：

Segment 类型		解释
路径信息	END	路径中包含的节点。相当于 SR-MPLS 中的 node-sid。
	END.X	路径中包含的链路。相当于 SR-MPLS 中的 adjacency-sid。
	END.B6	路径中包含的 SRv6 policy 子路径。报文进入该子路径时还可区分 Encaps/Encaps.RED 等行为。
	END.BM	路径中包含的 SR-MPLS policy 子路径。
L2VPN 业务	END.DX2	Egress PE 上标识 VPWS 业务，内层 L2 载荷向 END.DX2 指定的出接口转发。
	END.DX2V	Egress PE 上标识 EVPN LXC（灵活交叉）业务，内层 Ethernet 载荷向指定 CE 侧转发时继续根据内层 VLAN 查表区分不同的出向接口。
	END.DT2U	Egress PE 上标识 EVPN 单播业务，内层 Ethernet 载荷查询私网 MAC 表转发。
	END.DT2M	Egress PE 上标识 EVPN 组播业务，内层 Ethernet 载荷查询私网广播表转发。
L3VPN 业务	END.DX6	Egress PE 上标识 L3VPN 业务，内层 IPv6 载荷向 END.DX6 指定的出接口转发。
	END.DX4	Egress PE 上标识 L3VPN 业务，内层 IPv4 载荷向 END.DX4 指定的出接口转发。
	END.DT6	Egress PE 上标识 L3VPN 业务，内层 IPv6 载荷查询私

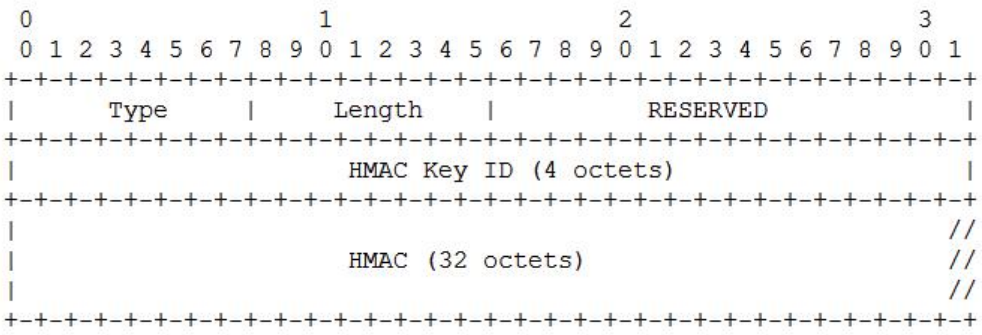
中兴通讯版权所有未经许可不得扩散

		网 IPv6 路由表转发。
	END.DT4	Egress PE 上标识 L3VPN 业务，内层 IPv4 载荷查询私网 IPv4 路由表转发。
	END.DT46	Egress PE 上标识 L3VPN 业务，内层 IPv6 或 IPv4 载荷查询私网 IPv6 或 IPv4 路由表转发。

2.1.3 SRH 中的安全字段

SRv6 在 SRH 中插入 HMAC 信息以校验报文的源头是否允许在报文的 DA 中使用当前 segment，并确保报文在传输时没有被修改。HMAC 信息是根据某个 KEY 和算法对样本 TXT 进行计算得到的结果。DA 对应的目的节点会使用本地保存的同样的 pre-shared key 和算法对同样的样本 TXT 进行加密，然后将加密结果与 HAMC 字段进行对比看是否一致。

SRH 中的 HMAC TLV 格式如下：



HMAC Key ID：是一个索引，唯一标识<pre-shared key, algorithm>，可基于对 TXT（IPv6 Header + SRH）文本内容运行加密算法得到 HMAC 字段。HMAC Key ID 如果为 0，则不包含 HMAC。

HMAC：采用 pre-shared key 和特定算法对 IPv6 Header + SRH 加密得到的内容。

HMAC 的计算方法可参考 RFC2104。

2.2 SR 域内部署模型

对于处于 SR domain 外的节点，它们是不可信的，不能直接使用 SR domain 内的 SID。因此可使用以下两级 ACL 来保证安全：

- SR domian 的边界节点检查任何从非 SR domain 进入 SR domain 的报文如果其目的 IP 是 SR domain 内的某个 SID，则将被丢弃。如果无法实施这种入向过滤策略，整个 SR domain 将暴露于 RFC5095 描述的源路由攻击中。
- SR domain 内的每个节点都检查报文的源 IP 如果处于非 SR domian，则丢弃报文。

2.3 SRv6-TE 路径的建立

如下图，建立一条从 R1 至 R9 的 SRv6-TE 路径，其路径信息为<node-R3, link-R3R7, node-R9>，则沿该路径转发的报文将先沿最短路径转发至节点 R3，然后沿链路 R3R7 转发至节点 R7，最后沿最短路径转发至节点 R9。该路径可以是在头结点 R1 上静态配置，或通过 CSPF 计算得到，或者从控制器下发（通过 netconf/pcep/bgp 等南向通道）。R1 上可以以 SRv6-TE tunnel 或者 SRv6 Policy 的形式维护该 SRv6-TE 路径对应的实例。

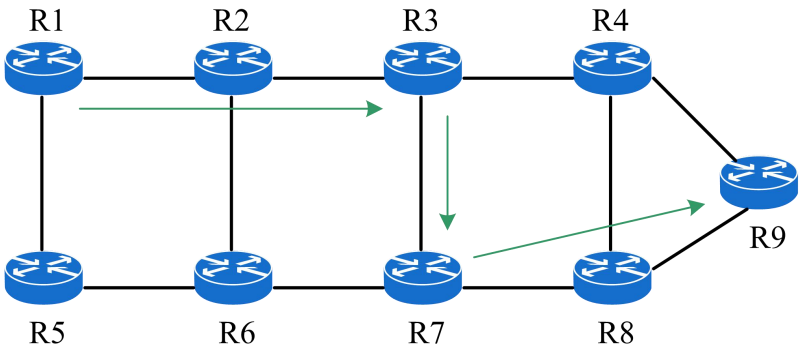


图 3 SRv6-TE 路径建立

需要将形如<node-R3, link-R3R7, node-R9>的路径信息转换成 Segment List，这个转换动作可以由控制器完成，也可以由 R1 节点自己完成。R1 节点可以通过 IGP 泛洪学习到 R3 节点的 END SID 与 END.X SID，比如：

节点 R3 上可以静态配置 END SID 为：block:a3::1

节点 R3 上为链路 R3R7 静态配置或动态产生的 END.X SID 为：block:a3::2

类似的，节点 R9 上可以静态配置 END SID 为：block:a9::1

这些表示路径信息的 SRv6 SID 均可通过 IGP 泛洪，R1 学习到后，则可将<node-R3, link-R3R7, node-R9>转换成 Segment List {block:a3::1, block:a3::2, block:a9::1}。当然，有些场景下 Segment List 可以直接静态配或者从控制器下发。

中兴通讯版权所有未经许可不得扩散

需要注意 SR domain 中所有节点的 SID 资源均处于同样 SID block 范围内。

每个支持 SRv6 功能的节点都需要为它自身的 SRv6 SID 准备相应的 Local SID 表项，一般就实现为路由表项，在相应的 Local SID 表项中给出转发动作以及特定的 Function，比如 R3 上可建立如下 Local SID 表项：

prefix/prefix-length	operation	function
block:a3::/64	本地终结	无
block:a3::1/128	本地终结	END
block:a3::2/128	向链路 R3R7 转发	END.X

R3 节点的 SRv6 SID 都从 block:a3::/64 范围中分配，这里 R3 将首先对外传统的泛洪前缀 block:a3::/64，然后在 SRv6 SID 分配好后，再对外泛洪 SRv6 SID。

每个支持 SRv6 功能的节点绝对不能为学习到的远端 SRv6 SID 建立相应的 remote SID 表项。

2.4 报文沿 SRv6-TE 路径的转发流程

如下图，R1 上将流量导入至上述创建好的 SRv6-TE 路径时，将直接在载荷前封装 IPv6 header + SRH，SRH 中包含的 Segment List 为{block:a3::1, block:a3::2, block:a9::1}，反序存放。在这个简单的例子中，只会封装一层 IPv6 header + SRH，不涉及到 SRH 嵌套。

具体转发流程如下：

① IPv6 header 中的 DA 首先拷贝为第一个 segment 即 block:a3::1，SRH 中的 Segment Left 字段减 1 变为 2，报文沿最短路径向目的 R3 节点转发。

② 报文到达 R2，R2 节点不处于 Segment List 中，它无需支持 SRv6 功能。R2 继续将报文按最短路径向 R3 转发。

③ 报文到达 R3，R3 上发现 DA(block:a3::1) 命中了 local SID 表项(block:a3::1/128)，其 function 为 END。于是 R3 继续根据 SRH 中的 Segment Left 获取下一个 segment(block:a3::2)，拷贝至 DA，SRH 中的 Segment Left 字段减 1 变为 1。将报文查路由表继续转发。

④ R3 上发现 DA(block:a3::2)命中了 local SID 表项(block:a3::2/128)，其 function 为 END.X。于是 R3 继续根据 SRH 中的 Segment Left 获取下一个 segment(block:a9::1)，拷贝至 DA，SRH 中的 Segment Left 字段减 1 变为 0。将报文向链路 R3R7 继续转发。

⑤ 报文到达 R7，R7 节点不处于 Segment List 中，它无需支持 SRv6 功能。R7

继续将报文按最短路径向 R8 转发。

⑥ 报文到达 R8，R8 节点不处于 Segment List 中，它无需支持 SRv6 功能。R8 继续将报文按最短路径向 R9 转发。

⑦ 报文到达 R9，R9 上发现 DA(block:a9::1)命中了 local SID 表项(block:a9::1/128)，其 function 为 END。R9 发现 SRH 中的 Segment Left 为 0，知道自身是最后一个 segment，则剥除 IPv6 header + SRH，载荷上送至控制平面处理。

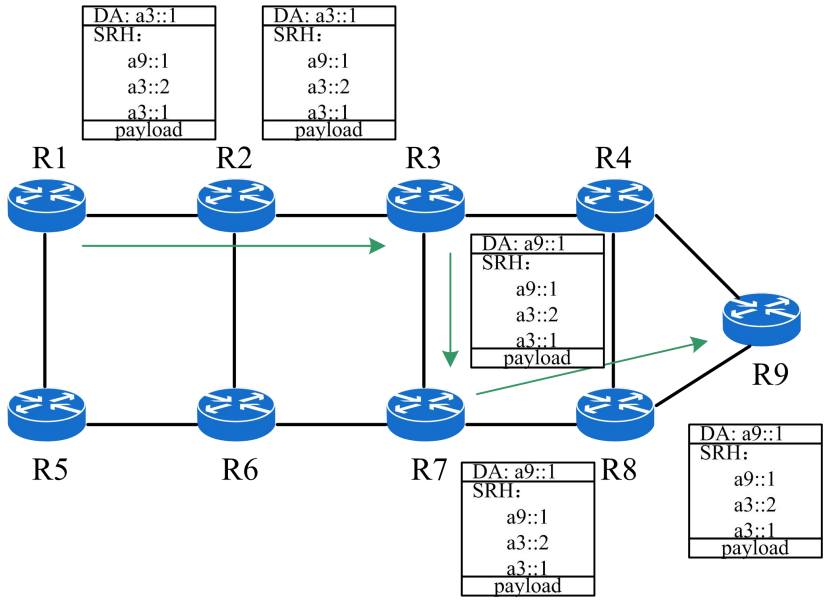


图 4 报文沿 SRv6-TE 路径的转发流程

上述第 4 步中，R3 可根据 block:a3::2/128 路由表项中含有 PSP 标志（SRH 提前弹出标志）并且 Segment Left 已更新为 0，则可直接将 SRH 提前剥除，保留 IPv6 header，DA 为 block:a9::1，将报文向链路 R3R7 转发。

2.5 L3VPN over SRv6

L3VPN 路由如想迭代 SRv6-plain（或称 SRv6-BE）外层隧道，egress PE 只需要在向 ingress PE 通告 VPN 路由的时候携带 SRv6 Service SID，后续报文转发时，将封装外层 IPv6 header 并且其目的地址就是 egress PE 通告的 SRv6 Service SID。此时 ingress PE 至 egress PE 沿途的转发均是 plain IPv6 转发，无需支持 SRH 解封装。

L3VPN 路由如想迭代 SRv6-TE 外层隧道，egress PE 需要在向 ingress PE 通告 VPN 路由的时候除了携带 SRv6 Service SID，还携带 color extended community

中兴通讯版权所有未经许可不得扩散

（表示相应 LSA），后续报文转发时，将封装外层 IPv6 header + SRH，SRH 中包含满足相应 SLA 的 SRv6 policy 的所有 SID，后跟 SRv6 Service SID。Ingress PE 至 egress PE 沿途的转发，那些处于 SRH 中的节点需要支持 SRv6 转发。

上述 SRv6 Service SID 在 egress PE 上有本地含义，由 egress PE 按每 CE 或每 VRF 分配，实际使用中它可以是 END.DT4/6 SID 或者 END.DX4/6 SID。

SRv6 Service SID 有两个作用：可用于 egress PE 所处 AS 内其它节点向本 egress PE 的报文的路由；可用于指代相应的 L3VPN ID。

2.5.1 L3VPN over SRv6-plain（SRv6 BE）

以 VPNv4 over SRv6-plain 为例，如下图，R9 节点上按每 VRF 分配 SRv6 VPN SID，即分配一个 END.DT4 类型的 SRv6 VPN SID (block:a9::10/128)，R9 通过 BGP 向 R1 通告相应的 VPNv4 NLRI 路由，在 BGP Prefix-SID Attribute 中携带 SRv6 Service SID (block:a9::10)。注意 R9 还需通过 IGP 向外泛洪 block:a9::/64 前缀。

R1 收到该 VPNv4 路由通告后，导入至本地的 VRF 路由表。后续报文根据相应的 VPN 路由表转发时，直接封装外层 IPv6 header，DA 直接填写为 block:a9::10，不封装 SRH。

报文将沿最短路径转发直至 R9 节点，R9 上根据 DA 命中 local SID 表项，且 function 为 END.DT4，则 R9 将外层 IPv6 header 剥除，将内层 IPv4 载荷继续查私网路由表转发。

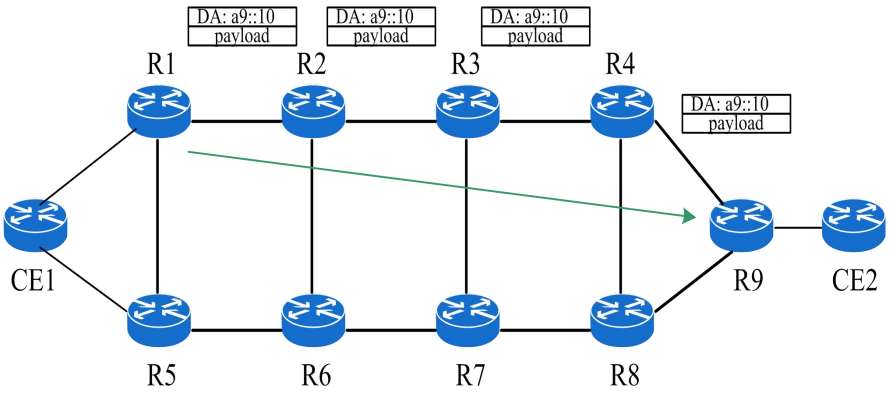


图 5 VPNv4 over SRv6-plain 的转发流程

2.5.2 L3VPN over SRv6-TE

R1 上可以为 VRF 配置隧道策略，使得 VPN 路由强行迭代至某个 SRv6-TE 路径

上，或者 R9 向 R1 通告 VPN 路由时同时携带 color community 属性，也将使得 R1 将 VPN 路由迭代至具有相应 color 的目的地址为 R9 的 SR-TE 路径。

以 VPNv4 over SRv6-TE 为例，如下图，R1 上将 VPN 路由迭代至 SR-TE 路径，该 SR-TE 路径对应的 Segment List 为{block:a3::1, block:a3::2, block:a9::1}，报文转发时，SRv6 VPN SID(block:a9::10)可以添加在尾部，即得到{block:a3::1, block:a3::2, block:a9::1, block:a9::10}。R1 可以将其中的 block:a9::1 优化掉，得到{block:a3::1, block:a3::2, block:a9::10}。

具体报文转发的流程不再赘述，报文将封装 IPv6 header + SRH，沿 SR-TE 路径转发直至 R9 节点，R9 上根据 DA 命中 local SID 表项，且 function 为 END.DT4，则 R9 将外层 IPv6 header + SRH 剥除，将内层 IPv4 载荷继续查私网路由表转发。

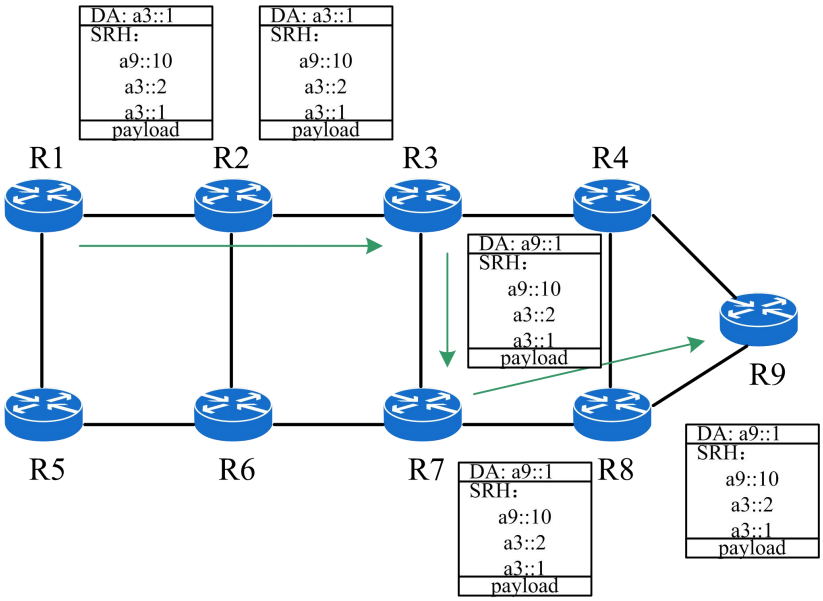


图 6 VPNv4 over SRv6-TE 的转发流程

2.6 EVPN over SRv6

Rfc7432 定义了 EVPN，主要讨论基于 MPLS 的 EVPN，也可扩展采用基于 IP 的 EVPN。

其中定义了 4 种路由类型，携带 prefix 以及 MPLS label 属性，每个标签对 MPLS 封装的 EVPN 流有特定的用途。draft-ietf-bess-evpn-prefix-advertisement 中定义了第 5 种路由类型，也携带 MPLS label 信息。

以下是 EVPN 相关的各种路由类型：

中兴通讯版权所有未经许可不得扩散

- Ethernet Auto-discovery Route (Route Type 1)
- MAC/IP Advertisement Route (Route Type 2)
- Inclusive Multicast Ethernet Tag Route (Route Type 3)
- Ethernet Segment route (Route Type 4)
- IP prefix route (Route Type 5)
- Selective Multicast Ethernet Tag route (Route Type 6)
- IGMP join sync route (Route Type 7)
- IGMP leave sync route (Route Type 8)

为了支持基于 SRv6 的 EVPN，SID 需要在以上的 1,2,3,5 路由类型中通告。与 L3VPN over SRv6 是类似的，egress PE 通告这些路由类型时也携带 BGP Prefix-SID Attribute，包含 SRv6 Service SID 信息。

具体内容不再赘述。

2.7 SRv6 Policy

Segment Routing 技术可以在报文入口节点上为报文流指定转发路径，而不需要在中间节点上维护报文流的路径状态。SR policy 是入口节点上转发路径的管理实体，通过 SR policy 进行报文转发，会把 SR policy 中的路径信息加入到报文头中，指导报文的转发，是 Segment Routing 技术的实现主体。

SR policy 以三元组<headend，color，endpoint>为键值。每个 SR policy 可以包含多个 Candidate Path，每个 Path 有一个 Preference；Candidate Path 可以包含多个 Segment List，每个 Segment List 拥有一个权重。SR policy 中会选择一个有效且优先级最高的 candidate path 作为 active path，来承载业务。active path 中的多个 segment list，按权重分担承载的业务。

SR Policy 通过 Segment List 来实现流量工程意图，Segment List 对数据包在网络中的任意转发路径进行编码。List 中的 Segment 支持多种类型，如 MPLS 标记、SRv6 SID 等。以 SRv6 SID 作为 Segment 的 SR policy 即可称为 SRv6 policy。

完整的 SRv6 policy 主要包含路径计算，配置下发，导流，保护等内容，典型应用案例可参见 3.1.1.2 部分。

中兴通讯版权所有未经许可不得扩散

2.8 基于 SR 转发机制的 TI-LFA

TI-LFA 提供了一种 local repair 机制，需要依赖 SR 转发机制，当故障发生时能够恢复端到端的连接。如下图，S 节点为 Point of Local Repair（PLR），需要找到修复节点 Q，以便至 D 的流量所经过的链路 L、或者包括 L 在内的 SRLG、或者节点 F 出现故障时能绕行至 Q 得到保护，并且绕行转发路径需尽量与收敛后的路径保持一致，避免流量在不同的路径上多次切换。

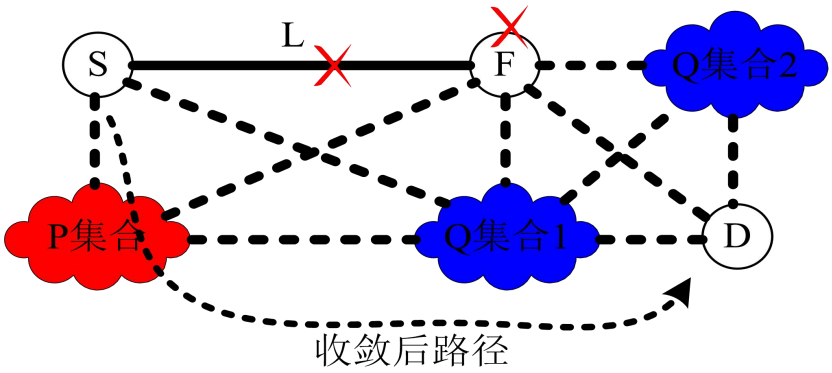


图 7 TI-LFA 网络拓扑图

S 节点上计算至 D 的 TI-LFA 路径的步骤如下：

- 计算 P 集合：

首先计算 P-Space P(S,X)：表示一个节点集合，即故障 X（X 是指故障链路 L 或者故障节点 F，下同）发生前，节点 S 至这些节点沿最短路径不需经过 X。P 集合中默认是包含 S 自身的。

如果 P-Space P(S,X)计算结果为空，则尝试计算 P-Space P(S->N,X)：表示一个节点集合，即故障 X 发生前，节点 S 的邻居 N 至这些节点沿最短路径不需经过 X。邻居 N 建议仅是故障 X 发生后的新的最优 SPF 下一跳（如前所述绕行转发路径需尽量与收敛后的路径保持一致，避免流量在不同的路径上多次切换）。

- 计算 Q 集合：

Space Q(D,X)：指一个节点集合，即这些节点至 D 节点的可达路径不需经过 X。

中兴通讯版权所有未经许可不得扩散

Q 集合中默认是包含 D 节点自身的。

- 确定 P 节点与 Q 节点：

按故障 X 发生后的 S->D 的最短路径，与 P 集合取交集确定候选 P 节点子集，与 Q 集合取交集确定候选 Q 节点子集。

从候选 P 节点子集中获取 P 节点，以及从候选 Q 节点子集中获取 Q 节点。如果候选 P 节点子集与候选 Q 节点子集存在交集，则应获取一个节点即作为 P 节点也作为 Q 节点；

否则获取的 P 节点与 Q 节点在沿 S->D 的最短路径上应尽量靠近。

- 确定 TI-LFA 路径：

计算 TI-LFA FRR 路径时，可以优先按节点 F 失效计算，如果计算结果为空，则再按照链路 L 失效计算，一般情况下，只要物理拓扑上存在冗余路径，则可以 100%的计算出 TI-LFA FRR 路径。

如下图，Segment List 为{N1,N2,N3}，当 N2 发生故障时，N2 的直连上游邻居 PLR 应该放弃尝试将流量发送给 N2，而是直接发给 N3（可能会沿至 N3 的 TI-LFA 路径发送）。

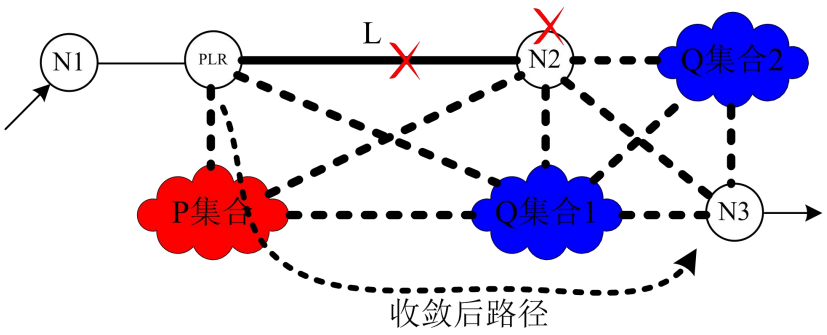


图 8 Segment List 重导向转发示意图

2.9 基于 SR 转发机制的微环避免

如下图，R4 与 R3 间链路的 metric 为 100，其它链路的 metric 为 1，当 R2 与 R3 之间的链路出现故障时，则 S 发往 D 的流量可能出现环路。比如 R0 比 R5 先收敛，比如 R5 比 R4 先收敛。

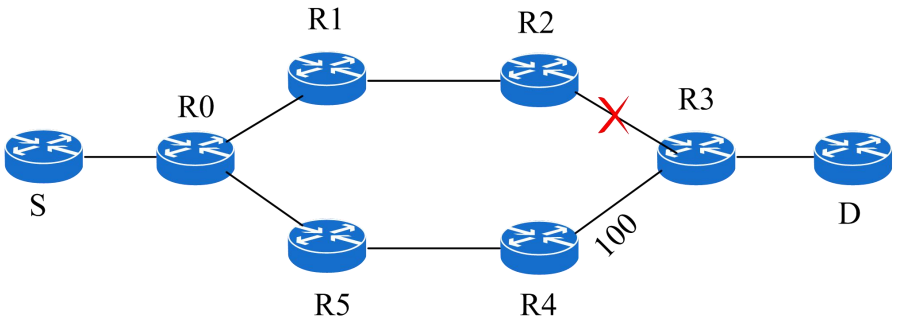


图 9 基于 SR 转发机制的微环避免

R0 收到链路 R2-R3 的故障泛洪后，可以先采用 SR policy 将流量无环的发往 D，比如采用 Segment List [NodeSID(R4), AdjSID(R4->R3), D]。稍后，再采用普通的路由转发。即转发分为两个阶段：最坏收敛时间之内走 SR policy 转发；之后走普通路由转发。防微环的 Segment List 的计算与 TI-LFA 备份路径的计算是类似的，只不过前者是感知链路故障时触发计算，此例中，P=R4，Q=R3。

同理，链路 R2-R3 的故障恢复后，S 发往 D 的流量也可能出现环路。比如 R0 比 R1 先收敛，比如 R1 比 R2 先收敛。类似的，R0 收到链路 R2-R3 的故障恢复泛洪后，可以先采用 SR policy 将流量无环的发往 D，比如采用 Segment List [NodeSID(R2), AdjSID(R2->R3), D]。稍后，再采用普通的路由转发。此时 P=R2，Q=R3。

2.10 SRv6 Ping/Trace-route

RFC4443 描述了在 IPv6 网络中使用 ICMPv6 进行网络诊断与错误报告，由于 SRv6 只不过是在 IPv6 的路由扩展头基础上新增了 SRH 类型，所以现有的 ICMPv6 ping/traceroute 机制是可以用于 SRv6 网络的。

本段以如下图来描述 Ping/Trace-route 流程。

中兴通讯版权所有未经许可不得扩散

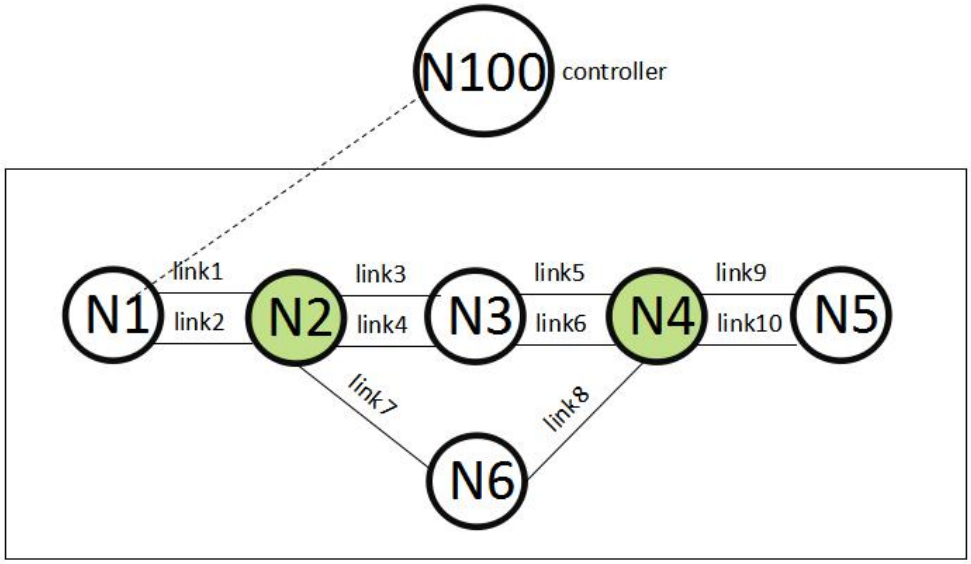


图 10 SRv6 Ping/Trace-route 网络拓扑

其中，N1/2/4 是支持 SRv6 的节点，N3/5/6 是传统的纯 IPv6 节点，N100 是控制器。

记节点 K 的全局 loopback 地址记为 A:k::/128，处于 IP block A 内；节点 X 与 Y 之间的第 n 条链路 X 侧全局 IP 地址记为 2001:DB8:X:Y:Xn::，节点 X 与 Y 之间的第 n 条链路 Y 侧全局 IP 地址记为 2001:DB8:X:Y:Yn::；节点 K 的具有 function 为 F 的 SID 记为 B:k:F::，处于 SID block B 内；特别的，节点 K 通过第 j 条链路至邻居 i 对应的 END.X 记为 B:k:Cij::。

所有链路的 IGP metric 为 10，除了 N2 与 N3 之间的两条链路为 100。

2.10.1 SRv6 Ping

一、传统的 ping（PING 一个属于 IP block 的远端地址）

SRv6 网络中，传统的纯 IPv6 节点对于 PING 的处理在软硬件上不需要做任何修改。

假设 N1 ping N5，沿 segment list <B:2:C31, B:4:C52>，则过程如下：

- Node N1 发出的报文：

(A:1::, B:2:C31)(A:5::, B:4:C52, B:2:C31, SL=2, NH=ICMPv6)(ICMPv6 Echo Request)

- Node N2 是 SRv6 节点，执行标准的 SRH 处理，此处执行 END.X (B:2:C31) 将报文沿 link3 发给 N3；
- Node N3 是一个纯 IPv6 节点，执行标准的 IPv6 处理，它根据报文的 DA=B:4:C52 转发报文；
- Node N4, 是 SRv6 节点，执行标准的 SRH 处理，此处执行 END.X (B:4:C52)带 PSP 标志，将 SRH 移除，沿 link10 发给 N5。
- Node N5 将收到不含 SRH 的 IPv6 报文，Node N5 是一个传统的纯 IPv6 节点，执行标准的 IPv6/ ICMPv6 处理，并回复 ICMP。

二、PING 远端 SID（PING 一个属于 SID block 的远端地址）

上述传统的 PING 流程无法满足 PING 一个 remote SID，比如 N1 ping 一个属于节点 N4 的 SID B:4:C52，沿 segment list <B:2:C31, B:4:C52>，则 N1 发出的报文为：(A:1::, B:2:C31)(B:4:C52yew, B:2:C31, SL=1;NH=ICMPv6)(ICMPv6 Echo Request)，这个报文在 N2 上就可以 PSP 了，最终 N4 收到报文后，发现 DA 为 local END.X SID，但下层载荷却是 ICMPv6，不是 SRH（因为 END.X 不能作为最后一个 segment，需要继续从 SRH 中获取下一个 segment）。显然，如果任由 END.X 将报文转发至下一跳，下一跳还会将报文环回来。

为了解决这个问题，需要显式的在 SRH 里识别出下层载荷是 OAM 报文，以做特殊处理。有两种方法：

- 在 segment list 中最后一个 SID 前插入 END.OP SID 或 END.OTP SID。

最后的 SRv6 节点，执行标准的 SRH 处理，执行 END.OP 流程：获得时间戳并将报文发给 OAM 进程处理，检查 SRH 中下一个 SID 是否是 local SID，如果不是，则产生 ICMPv6 错误消息并回复，否则回复成功。

注：第 5 版 SRv6 OAM 草案（draft-ietf-6man-spring-srv6-oam-05）已经删除了使用 END.OP/OTP SID 的方案，目的节点收到 PING 报文时，判断目的 IP 为 local SID，根据报文内层 ICMPv6 类型回复 ICMPv6 应答。

- 设置 SRH.Flags.O-flag 为真；

注意基于 O-flag 的流程中，segment list 中每个 segment 节点都会回复 ICMP（有

中兴通讯版权所有未经许可不得扩散

点 traceroute 的意思），这样能够提供证明报文确实经过了这些 segment 节点。此时最后一个 SID 必须是 USP。

中间的 SRv6 节点，执行标准的 SRH 处理，由于 O-flag 为真，则拷贝报文并获得时间戳，拷贝的报文将扔给 OAM 进程处理，做相应回复。注意如果不支持 O-flag，则简单忽略 O-flag 即可。

目的 SID 所在的 SRv6 节点，执行标准的 SRH 处理，由于 O-flag 为真，则拷贝报文并获得时间戳，拷贝的报文将扔给 OAM 进程处理，并做相应回复。注意如果 Node N4 不支持 O-flag，则简单忽略 O-flag 即可，但此时 PING 会失败。

2.10.2 SRv6 Trace-route

一、传统的 traceroute（Traceroute 一个属于 IP block 的远端地址）

SRv6 网络中，传统的纯 IPv6 节点对于 Traceroute 的处理在软硬件上不需要做任何修改。

假设 N1 traceroute N5，沿 segment list <B:2:C31, B:4:C52>，则输出结果如下：

> traceroute A:5:: via segment-list B:2:C31, B:4:C52

Tracing the route to B5::
1 2001:DB8:1:2:21:: 0.512 msec 0.425 msec 0.374 msec
SRH: (A:5::, B:4:C52, B:2:C31, SL=2)
2 2001:DB8:2:3:31:: 0.721 msec 0.810 msec 0.795 msec
SRH: (A:5::, B:4:C52, B:2:C31, SL=1)
3 2001:DB8:3:4::41:: 0.921 msec 0.816 msec 0.759 msec
SRH: (A:5::, B:4:C52, B:2:C31, SL=1)
4 2001:DB8:4:5::52:: 0.879 msec 0.916 msec 1.024 msec

可见输出结果是沿途每一跳根据 Hop-limit 超时，向源节点发送 ICMP 消息其中 SA 为原始报文接收时的入接口 IP 地址。Segment list 虽然控制了报文的路径，但是 SR 节点与非 SR 节点在向源节点发送 ICMP 消息时没有任何差异。

二、traceroute 一个 remote SID（traceroute 一个属于 SID block 的远端地址）

上述传统的 Traceroute 流程无法满足 Traceroute 一个 remote SID，比如 N1 Traceroute 一个属于节点 N4 的 SID B:4:C52，沿 segment list <B:2:C31, B:4:C52>。

其原因与前述 PING 章节是一样的，即最终 N4 收到 traceroute 报文后，发现 DA 为 local END.X SID，但下层载荷却是 UDP，不是 SRH。

为了解决这个问题，需要显式的在 SRH 里识别出下层载荷是 OAM 报文，以做特殊处理。如上述 PING，通过 O-flag 或通过 local SID 报文内层 ICMPv6 类型回复 ICMPv6 应答

该文档仅举例说明 SRv6 SID 的 ICMP ping 和基于 UDP 的 traceroute，这些过程同样适用于对 SRv6 SID 进行探测的其他 IPv6 OAM（例如 BFD、SBFD、TWAMP Light 和 STAMP）。具体来说，只要本地配置允许上层报头处理特定 SRv6 SID 的有效 OAM 负载，现有的 IPv6 OAM 技术就可以用于将探针定位到（远程）SID。

2.11 SRv6 SRH 优化

网络节点内 ASIC/NPU 收到数据包后，会把数据包存在外置的内存中。ASIC/NPU 读取固定长度的报头内容（一般是 96~128 字节），然后查找芯片本地/外部内存中的转发表，进行转发。如果报文头太长，无法在一个处理周期完成读取，则需要使用两个处理周期进行读取（Recycle），这将导致吞吐量下降一半。

在 SR-MPLS 下，协议引入的开销较小，因此现有的大多数网络设备硬件均可以在一个处理周期内读取完 SR 报头信息，完成转发，意味着现有的硬件无须替换，只需升级软件即可支持 SR-MPLS。这是 SR-MPLS 能迅速得到大量部署的技术基础之一。

但 SRv6 引入的协议开销远大于 SR-MPLS，Segment 所对应的操作也比 SR-MPLS 复杂，因此 SRv6 对网络设备提出了非常高的要求。如果按照目前的 SRv6 协议实现，要么需要替换掉绝大多数的网络设备，要么网络吞吐降低一半（Recycle），这对于很多用户而言是难以接受的。所以 SRv6 需要迫切解决在协议开销、承载效率、MTU 和对硬件要求方面的问题。这几个问题，其实本质上都是同一个问题，即如何提高 SRv6 Segment 效率。

当前业界提出了以下四种方案：

- CISCO 提出的 uSID carrier 方案，即在一个 128 bits 的 IPv6 地址中存放多个 uSID，每个 uSID 一般只占 16 bits。需新增 SRv6 FUNCTION，并在控制器平面通告 uSID 的路由可达性。此方案需要 SRv6 domain 内的所有 Segment 节点的 SID 部署相同在 uSID block 内。
- Juniper 提出的 Compress-SRH 方案，直接在 SRH 中增加标志支持以更短小的索引表示 Segment。需要建立索引至 IPv6 地址的映射表。

中兴通讯版权所有未经许可不得扩散

- ZTE 提出的 Unified-SRH 方案，直接在 SRH 中增加标志支持 MPLS label 或多种字长的 SRv6 SIDt。支持以上几种映射及合并压缩方式的融合
- Huawei 提出的 Common-SRH 方案，将 SID List 中所有 SID 的公共部分存放在 DA 中，SRH 中只存放差异部分。

目前头压缩方案正处于标准整合阶段，IETF 成立了专门的 design team 讨论头压缩标准化问题，争取推出统一的头压缩方案。

3 SRv6 的典型应用及应用现状

3.1.1 SRv6 的典型应用

3.1.1.1 L2/3 VPN over SRv6 BE

如 2.5/2.6 节所述，SRv6 能够支持快速部署 L2/L3 VPN 业务。

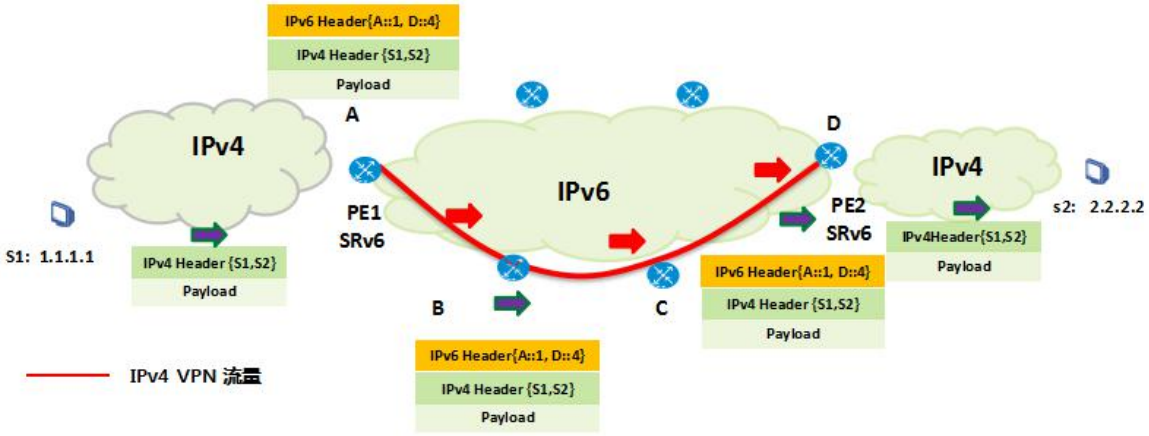


图 11 L2/L3VPN over SRv6 BE 网络拓扑

如图 11，在 IPv6 网络边缘 PE 节点支持 SRv6，开启 L3/L2 VPN over SRv6 BE，中间节点支持普通 IPv6，即可完成 L2/L3 VPN 业务部署。如在城域或者骨干网络部署 L2/L3 VPN，不再需要复杂的 MPLS 支撑。

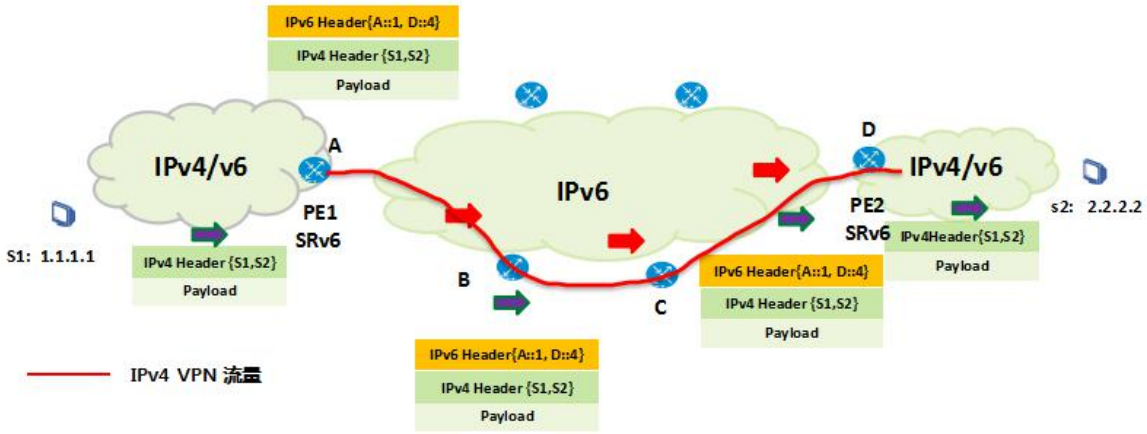


图 12 L2/L3VPN over SRv6 BE 网络拓扑

也可以如上图在两侧边缘网络增加 SRv6 网关设备，穿透中间 IPv6 网络，快速构建边缘网络的 L2/L3 VPN 业务。如政企分支 VPN 互联，运营商跨骨干同省份城域 VPN 业务部署等。

L2/L3 VPN over SRv6 BE 能够仅通过部署支持 SRv6 的 PE 设备或者其他类设备，快速开通 L2/L3 VPN 业务，提供类似 MPLS VPN，云专线，云接入等服务。在这种方式下，中间网络仅需支持 IPv6，无需开启 SRv6；但 PE 间仅支持尽力而为，依照路由优先转发。对于有质量要求的 VPN 业务，需要结合切片、QOS 设置等方式提供一定程度的保障。

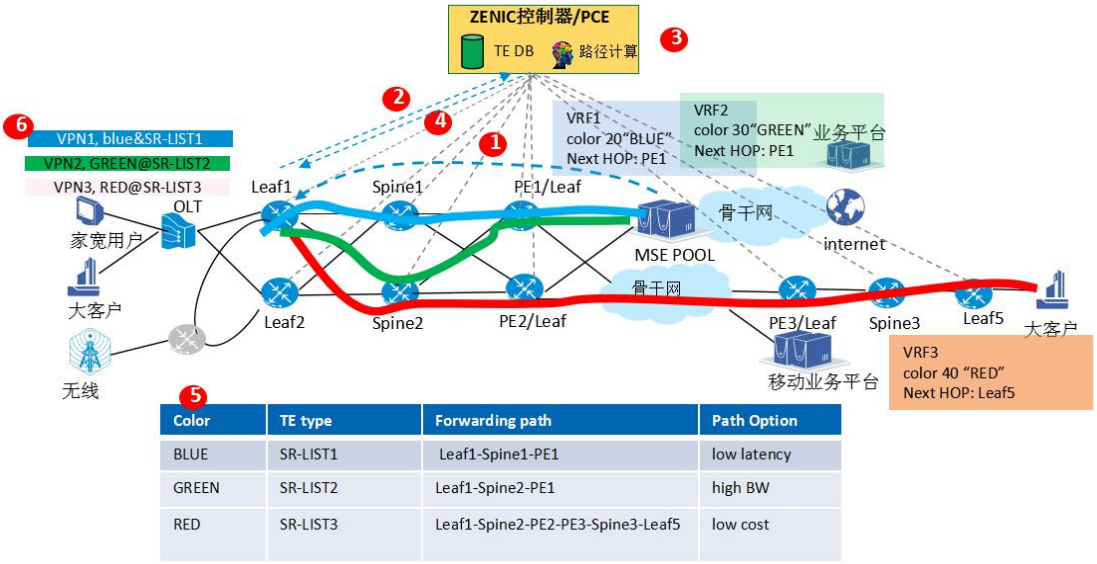
3.1.1.2 L2/3 VPN over SRv6 TE（SRv6 Policy）

SRv6 Policy 完全抛弃了隧道接口的概念，是重新设计的一套 TE 体系，通过 Segment List 来实现流量工程意图。Segment List 对数据包在网络中的任意转发路径进行编码。

SRv6 Policy 中引入 color 属性，通常代表意图（例如低延迟，排除 SRLG 等），这个新的基本概念用于实现 SR-TE 的自动化。基于 SRv6 Policy 的 SR-TE 将 BGP 路由置于解决方案的核心，通过对业务路由进行着色（用 color 标识）来实现按业务需要动态创建 SRv6 Policy 以及自动引流至 SRv6 Policy。从而大大简化配置。

本节以下图 L2/L3 VPN 业务为例，描述通过 color 着色创建 SRv6 Policy 并引流的流程。

中兴通讯版权所有未经许可不得扩散



Color	TE type	Forwarding path	Path Option
BLUE	SR-LIST1	Leaf1-Spine1-PE1	low latency
GREEN	SR-LIST2	Leaf1-Spine2-PE1	high BW
RED	SR-LIST3	Leaf1-Spine2-PE2-PE3-Spine3-Leaf5	low cost

图 13 L2/L3VPN over SRv6 TE 网络拓扑

① 远端 PE 通告带 color 的 VPN 路由
② 头端收到 VPN 路由，并向 PCE 请求路径计算
③ PCE 根据 BGP-LS 收集的拓扑拓扑信息进行路径计算
④ 路径结果下发
⑤ 头节点创建相应的 SR-LIST
⑥ 不同颜色的 VPN 路由引流到不同的 SR-LIST

此外，除了 Color 着色，进行 L2/L3 VPN 流量自动导流外，还可以通过策略路由等方式指定业务选择特定的 SRv6 Policy

3.1.1.3 SRv6 与 SR-MPLS 的互联

随着 SRv6 的发展，存量的 SR-MPLS 网络与新建 SRv6 网络有有如下互通性场景

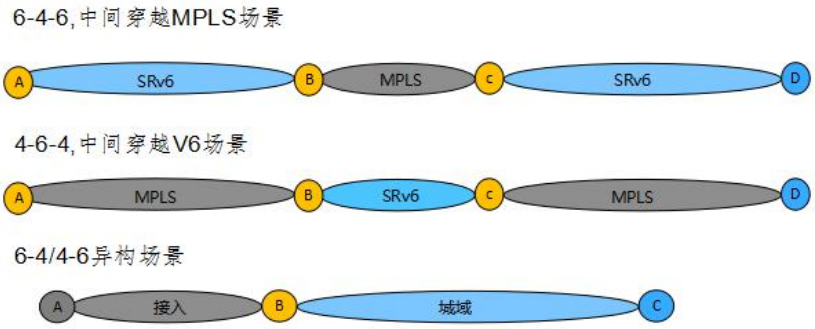
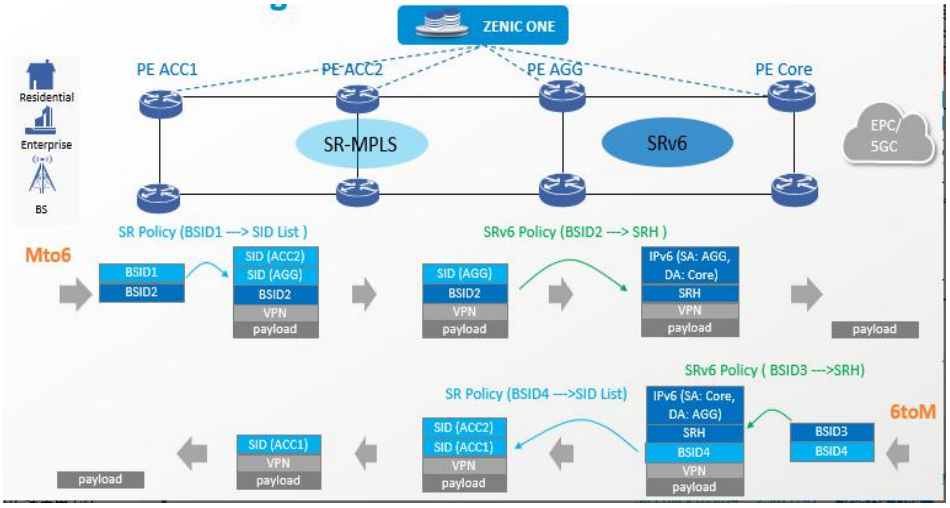


图 14 SRv6 与 SR-MPLS 互联场景

SRv6 与 SR-MPLS 互通时，传统上可以通过 Option A 的方式在域间进行 Overlay 层面的互导，下层 SRv6 和 SR-MPLS 隧道不需要直接互通

也可以直接创建 Overlay 端到端邻居关系，Underlay 创建端到端 SR-policy 隧道。如通过 BSID 的转换与映射实现，端到端的 SR policy={BSID1 for SR-MPLS policy，BSID2 for SRv6 policy}

此种 M to 6 场景中，当头节点接收到控制器下发的 SR policy {BSID1, BSID2}时，需要查找 BSID1 对应的 SR-MPLS LIST，并将展开的 SR-MPLS list 封装在数据包中转发{SR1，SR2，SR3，BSID2}，通过逐跳的 SR 路径送达到边界网关节点后，边界网关节点根据 BSID2，映射并展开相应的 SRH，将数据包继续传送到远端 PE。反向亦然。因而要求支持 MPLS 格式的 BSID 在 MPLS 域/SRv6 域中与 SR LIST/SRH 之间的映射。如下图所示



中兴通讯版权所有未经许可不得扩散

图 15 SRv6 与 SR-MPLS 端到端互联场景

SRv6 与 SR-MPLS 的互联场景较多，可根据不同场景采用不同的解决方案

3.1.2 SRv6 的应用现状

IPv6 网络是 SRv6 试点应用的基本条件，2017 年底国内开始的 IPv6 规模部署为 SRv6 应用提供了良好基础。过去一两年间，现网 SRv6 应用主要还是集中在承载网络侧。基于 SRv6 对 VPN 以及 TE 通道的支持和优化能力，结合当下新型城域网、云网融合、云专线、切片等新型网络和业务试点，进行一些应用场景的拓展。更多的是现有业务的继承性支持和优化。

同时，我们也要看到，受上下游产业链、设备支持能力以及相应编排和控制系统的限制，SRv6 的大规模网络拉通及云网拉通应用较少。SRv6 核心的路径和业务联合编排特性，受到现有业务模式和相应支撑系统的限制，尚没有规模应用试点。此外，经典 SRv6 128bit SID 字长引起的相关问题，也成为 SRv6 规模部署的主要障碍。

综上所述，SRv6 依旧处在应用的尝试阶段，大规模部署尚要相应条件的成熟。

4 总结和客户价值

SRv6 是未来网络的基础性技术，正处在高速发展的阶段。

SRv6 在 SRv6-BE 应用场景明确，域内及跨域 VPN 解决方案有一定优势；为政企和运营商客户提供新的 VPN 业务快速部署手段。结合 SRv6 Policy 与切片，以及相关 OAM、保护等技术，SRv6 正在逐渐尝试构建差分服务体系。

SRv6 在 SFC 领域标准化工作也在向前推进，为路径和业务编排提供了初步的实现手段，这也是未来运营商及政企客户业务开发的重要基础。

中兴通讯在 IP 领域具备深厚的技术积淀，相关 IP 产品已支持较为完善的 SRv6 功能集，并进行较大规模的现网试点工作。同时，在 SRv6 关键的头压缩等领域，中兴通讯处于标准及实现的前沿，推动 SRv6 向前发展。

中兴通讯版权所有未经许可不得扩散

5 术语及缩略语

表 5-1 术语及缩略语说明表

英文缩写	英文全称	中文全称
SRv6	Segment Routing IPv6	分段路由应用于 IPv6 转发平面
SRH	Segment Routing Header	分段路由扩展头
SID	Segment Identifier	段 ID
IP	Internet Protocol	互联网协议
IPv6	Internet protocol version 6	因特网协议版本 6
SR	segment routing	分段路由
SL	Segment Left	剩余 segment 数量
BGP	Border Gateway Protocol	边界网关协议
BGP-LS	BGP Link-State	BGP 链路状态
SDN	Soft defined network	软件定义网络
TE	Traffic Engineering	流量工程
VPN	Virtual Private Network	虚拟专用网
OAM	Operation Administration and Maintenance	运行、管理和维护
SBFD	Seamless bidirectional forwarding detection	无缝双向转发检测
TI-LFA	Topology Independent Loop free Alternate	拓扑无关的无环路备份

中兴通讯版权所有未经许可不得扩散