



分段路由网络研究及其在流量工程中的应用

何晓明¹, 卢泉¹, 邢亮²

(1. 中国电信股份有限公司广州研究院, 广东 广州 510630;

2. 中国电信集团公司, 北京 100032)

摘要: 分段路由网络是对 IP 路由技术的继承和创新, 在实现网络虚拟化、快速重路由、网络可编程、负载均衡和流量工程等方面具有独特优势。深入研究了分段路由的实现机制, 全面介绍和剖析了 IGP 和 BGP 支持分段路由网络的协议扩展, 探讨了分段路由在运营商网络中典型应用场景的流量工程实现。随着技术的成熟和标准的完善, 分段路由网络将具有广阔的应用前景。

关键词: 分段路由; 流量工程; 路由协议; SDN 控制器

中图分类号: TN919.21

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016179

Study on segment routing network and its application in traffic engineering

HE Xiaoming¹, LU Quan¹, XING Liang²

1. Guangzhou Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China

2. China Telecommunications Corporation, Beijing 100032, China

Abstract: Segment routing network is the inheritance and innovation of IP routing technology, which has many unique advantages in the realization of network virtualization, fast re-routing, network programming, load balancing and traffic engineering, and so on. Segment routing mechanism was studied in depth, and the protocol extensions of interior gateway protocols (IGP) and border gateway protocol (BGP) for segment routing were introduced and analyzed in detail. Furthermore, the realization of traffic engineering for segment routing in typical application scenarios of the operator networks was explored. With the maturity of technology and standardization, segment routing network will promise to have broad application prospects.

Key words: segment routing, traffic engineering, routing protocol, SDN controller

1 引言

为 IP 报文指定转发路径、而不是按照传统 IP 网络的最短路径转发的网络能力, 称为基于源路由网络 (the source packet routing in networking, SPRING)^[1]。SPRING 的核心思想是由报文头部维护策略状态, 无需转发路径的中

间节点维护策略状态, 这样可大大简化网络为维护策略状态信息所需复杂信令协议的支持。SPRING 有助于实现网络虚拟化、快速重路由 (fast reroute, FRR)、网络可编程、OAM、负载均衡和流量工程等网络功能。MPLS (multi-protocol label switching, 多协议标签交换) 作为一种成熟可商用的标签交换技术, 已经在各大运营商网络得到广泛部署, 主

要用于提供二/三层 VPN 服务。然而, MPLS 支持流量工程及快速重路由需要更为复杂的信令协议, 如 CR-LDP^[2]、RSVP-TE^[3]等, 网络设备需要维护每条 LSP 的状态信息, 设备开销大, 存在严重的扩展性问题, 因此, 除了某些特殊应用 (如为骨干网少数关键节点和链路提供局部保护) 外, 一般极少使用。尽管采用独立的路径计算单元 (path computation element, PCE) 替代边缘入口设备 (label edge router, LER) 选择 TE LSP 可以大大减轻设备控制面的负担 (维护流量工程数据库需要足够多的存储, 路径计算需要消耗大量计算资源)^[4], 已部分体现了 SDN 思想, 但网络设备仍需运行 RSVP-TE 等信令来维护每条 LSP 的状态信息, 对控制面的功能要求高。

SDN 近年来受到广泛关注, 并成为运营商、设备制造商、学术界研究和应用的热点领域。根据开放网络基金会 (Open Networking Foundation, ONF) 对 SDN 的定义, 在 SDN 架构中, 控制平面和数据平面解耦, 网络智能和状态逻辑集中, 底层网络基础设施抽象于上层应用。SDN 终极目标是简化网络设备的复杂性, 实现网络虚拟化和网络应用的开放可编程, 快速提供业务部署。当前最具革命性的 SDN 技术当属由 ONF 主导的 OpenFlow 协议, 基于 OpenFlow 的 SDN 通过集中的 SDN 控制器发现和维护全网拓扑信息, 根据自定义策略为每一跳转发设备下发流转发表, 底层转发设备依据流表实现业务流的转发、过滤和丢弃等动作。然而, OpenFlow 网络当前仍面临流表的扩展性、硬件芯片技术制约网络转发性能、新旧设备共存及网络演进等问题^[5,6]。由 IETF 主导的 I2RS 工作组希望在继续保留分布式路由优势的同时, 仍可通过应用来改变路由决策^[7]。实践已经证明, 传统路由协议如 OSPF、ISIS、BGP 在大规模网络 (如 Internet) 取得了巨大成功。通过对传统路由协议进行扩展改良, 能够适应网络应用和业务创新发展, 从而实现现有网络基础设施向 SDN 的平滑演进。

SPRING 的目标与 IETF 主导的基于 I2RS 的 SDN 是一致的, 其目的在于通过对传统 IGP、BGP 进行扩展和改进, 进而简化网络设备的复杂性, 更好地实现网络弹性和可编程、网络负载均衡和流量工程、网络虚拟化等功能, 为业务提供最佳适配。

本文全面介绍和剖析 IGP 和 BGP 支持分段路由网络的协议扩展, 并分析了分段路由网络的 OAM 机制, 同时探讨分段路由在运营商网络中典型应用场景的流量工程实现。

2 分段路由网络研究

由 Cisco 发起的分段路由 (segment routing, SR) 网络是一种基于源路由网络, 目前正由 IETF 的 SPRING 工作组进行标准化。根据 IETF“分段路由结构”draft^[8]的定义, 分段路由是指主机或网络边缘入口节点为 IP 报文添加一个包含一串按序排列的、被称为“分段”的指令列表的报文头部, 中间节点根据报头列表中的顶层段标识进行 IP 报文的逐跳转发。分段路由的思想是把报文流经过的一条端到端的路径划分成多个分段, 每一分段以段标识 (SID) 来标识。对于 MPLS 网络, “分段路由结构”draft 定义了 5 种段标识: IGP 段标识、BGP peer 段标识、LDP LSP 段标识、RSVP-TE LSP 段标识和 BGP LSP 段标识。对 IGP SID 而言, 可以是节点标识 (node SID)、地址前缀标识 (prefix SID), 也可以是邻接标识 (Adj-SID)。对于纯 IPv6 网络, SID 编码成 IPv6 地址, 为支持分段路由功能, IETF 专门定义了一种新的 IPv6 路由扩展头部^[9]。

在每个分段里, 转发设备根据转发表 (forwarding information base, FIB) 查找段标识对应的下一跳, 最终转发到该段标识的发起者。由于 SID 是通过 IGP 向全网通告, 这样, IP 报文在每个分段里都按照 ECMP 感知的最短路径优先 (shortest path first, SPF) 算法进行转发。由一串段标识列表指定的端到端路径通常由 SDN 控制器根据网络资源的利用情况以及业务需求等因素来确定, 当然也可手工配置。网络设备只需支持 OSPF、ISIS 等基本路由协议, 无需通过信令协议来维护报文流的状态信息, 极大地简化了网络设备的复杂性。

2.1 分段路由的 IGP 扩展

在一个 IGP 拓扑域里, 分段路由通过网络通过一串按序排列的分段标识灵活指定了一条端到端的转发路径, 这些段标识通过链路状态路由协议 (如 OSPF 或 ISIS) 向全网通告。一个 prefix SID 通常代表多跳路径, 具有全局唯一性; 而一个 Adj-SID 在多数情况下是一个单跳路径, 一般只具有本地意义。为了通告各种 SID, ISIS 和 OSPFv3 在链路状态协议数据单元 (label switched path, LSP) 分别为它们定义了新的 TLV^[10,11]。由于 OSPFv2 协议采用固定格式的链路状态通告 (link-state advertisement, LSA) 报文, 为了支持新特性, OSPFv2 在 RFC7684 专门定义了 2 个新的 opaque LSA: OSPFv2 扩展 prefix opaque LSA 和 OSPFv2 扩展 link opaque LSA, 并且分别为它们各自增加了一个新的 TLV:



OSPFv2 扩展 prefix TLV 和 OSPFv2 扩展 link TLV,用于通告附加的前缀和链路属性。以下介绍 OSPFv2 协议扩展^[12]与 SID 通告相关的几个重要的 TLV。

(1)SID/标签 TLV

SID/标签 TLV 用于通告前缀和链路关联的 SID 或标签,SID/标签 TLV 格式如图 1 所示。

类型	长度
SID/label (可变)	

图 1 SID/标签 TLV 格式

SID/label 字段含义:若长度字段设置为 3,那么 SID/label 值最右边的 20 bit 代表一个标签;若长度字段设置为 4,则 SID/label 值代表 32 bit 的 SID。

(2)prefix SID TLV

prefix SID TLV 是 OSPFv2 扩展 prefix TLV 或 OSPFv2 扩展 prefix range TLV 的子 TLV,用于通告前缀绑定的 SID 属性。prefix SID TLV 格式如图 2 所示。

类型		长度	
flags	保留	MT-ID	算法
SID/index/label (可变)			

图 2 prefix SID TLV 格式

flags 字段:8 bit,定义如图 3 所示的标志位。

	NP	M	E	V	L		
--	----	---	---	---	---	--	--

图 3 prefix SID TLV flags 标志位

NP 标志位:no-PHP 标志。若设置为 1,则次末跳向下一跳转发报文时不应该弹出 prefix SID;若设置为 0,则次末跳必须弹出 prefix SID,这等同于 MPLS 网络中的次末跳弹出机制。

M 标志位:映射服务器标志。若设置为 1,该 SID 由映射服务器通告。

E 标志位:explicit-null 标志。若设置为 1,prefix SID 发起者的上游邻居在转发报文前必须用一个 explicit-null 值(对于 IPv4 地址为 0)代替 prefix SID。

V 标志位:值/索引标志。若设置为 1,则 prefix SID 携带一个绝对值(SID);若设置为 0,prefix SID 代表一个索引。

L 标志位:本地/全局标志。若设置为 1,则 prefix SID 携带的值或索引具有本地意义;若设置为 0,则具有全局意义。

其他标志位:保留,这些标志位发送时必须置 0,接收时忽略。

MT-ID:由 RFC4915 定义的多拓扑 ID。

algorithm:一个字节字段,标识 prefix SID 关联的路径选择算法。IGP 扩展目前定义了两种路径选择算法:当该字段置“0”时,表示缺省基于度量的最短路径优先(SPF)算法,允许任何节点根据本地路径选择策略来改变基于缺省 SPF 选路;当该字段置“1”时,表示沿路径的所有节点必须遵循严格 SPF 算法,不允许本地策略改变路径选择。

(3)Adj-SID TLV

Adj-SID 是 OSPFv2 扩展 link TLV 中一个可选的子 TLV,用于标识一条或一组单向链路。Adj-SID 格式如图 4 所示。

类型		长度	
flags	保留	MT-ID	weight
SID/index/label (可变)			

图 4 Adj-SID 格式

flags 字段:8 bit,定义标志位如图 5 所示。

B	V	L	S	保留
---	---	---	---	----

图 5 Adj-SID TLV flags 标志位

B 标志位:备份标志,若置“1”,该 Adj-SID 标识的链路符合链路保护条件。

V 标志位:值/索引标志,若置“1”,则该 Adj-SID 携带一个绝对值;若置“0”,则该 Adj-SID 代表一个索引。

L 标志位:本地/全局标志。若设置为 1,则 prefix SID 携带的值或索引具有本地意义;若设置为 0,则具有全局意义。

S 标志位:链路组标识,若置“1”,则 prefix SID 表示一个链路组,流量经由一组链路负载分担转发到邻居节点。

MT-ID:由 RFC4915 定义的多拓扑 ID。

weight 字段:用于流量在一组并行链路间的负载均衡权重,这对于一组非等速链路是很有用的。

2.2 分段路由的 BGP 扩展

由于 IGP 扩展对 SID 的通告只局限于单个 IGP 域,只能获取 IGP 域内的链路状态数据库(link state database, LSDB)和流量工程数据库(traffic engineering database, TED),因此,单纯使用 IGP 不足以构建跨多个 IGP 域或跨 AS 域的分段路由网络。为了满足某些需要跨域拓扑可视的应用需求,IETF 正在标准化的 BGP-LS^[13]允许 BGP 携带

链路状态信息。BGP-LS 定义了一种新的 BGP 网络层可达信息(network layer reachability information,NLRI)编码格式及一种新的 BGP-LS 属性,节点、链路及前缀这 3 种关键的链路状态对象分别被编码在 NLRI 中,与这些链路状态对象关联的属性被编码在 BGP-LS 属性中。图 6 展示了一种典型的跨 IGP 域或 AS 域的部署场景。在每个 IGP 域中,一个或多个节点配置使能 BGP-LS,这些 BGP 发言者通过连接路由反射器(route reflector,RR)形成 IBCP 全互联。这样,所有的 BGP 发言者(特别是 RR)都获取了全部 IGP 域的链路状态信息,外部部件如 SDN 控制器可以从 RR 获取全网链路状态信息。

类似 SDN 控制器这种外部部件,通过收集跨 IGP 域的分段路由来构建端到端的路径,指示 IP 报文按照指定的分段路径转发。

为支持分段路由网络,BGP-LS 扩展了 6 个新的 BGP-LS 属性 TLV:prefix SID TLV、Adj-SID,SR 能力 TLV、SR 算法 TLV、IGP binding TLV 和映射服务器通告前缀与 SID 绑定 TLV^[14]。限于篇幅,不再详细介绍。

2.3 分段路由的域间多出口工程 BGP 扩展

为了解决域间多出口工程 (egress peer engineering, EPE) 问题,BGP-LS 扩展^[15]定义了 3 种 BGP peering SID: peer-node-SID;peer-Adj-SID;peer-set-SID。peer-node-SID 用于标识 BGP 会话邻居(peer),peer-Adj-SID 用于标识 BGP 会话的本地接口或互联链路,peer-set-SID 用于标识远端 AS 的所有 peer。由于域间多出口连接性描述基于链路和远端 peer/AS,因此,可以重用 BGP-LS 中的 link NLRI 通告 BGP peering SID 的可达性,并定义了一种新的 protocol-ID,用于区分 link NLRI 携带的 EPE 描述信息与 IGP 链路状态信息。

(1)BGP-LS link NLRI

BGP-LS 定义了用于携带 EPE 描述消息的 link NLRI 格式,如图 7 所示。

协议 ID			
identifier (64 bit)			
本地节点描述符			
远端节点描述符			
链路描述符			

图 7 链路 NLRI 格式

本地节点描述符(local node descriptor)TLV 必须包含 3 个消息:BGP router ID、AS、BGP-LS 标识。

远端节点描述符(remote node descriptor,RND)TLV 可以包含两个消息:BGP router ID、AS。

链路描述符(link descriptor)TLV 包含的消息跟通告的 peering SID 相关,若通告 peer-node-SID 或 peer-set-SID,链路描述符 TLV 必须包含 BGP 会话的本地接口地址和 peer 接口地址;若通告 peer-Adj-SID,链路描述符 TLV 必须包含 4 byte 的链路本地标识符和 4 byte 的 0 值,可以包含 BGP 会话的本地接口地址或 peer 接口地址。

(2)BGP-LS 扩展 link 属性

BGP-LS 扩展 link 属性定义的 peer-node-SID TLV、peer-Adj-SID TLV 和 peer-set-SID TLV,它们都具有如图 8 所示的格式。

类型		长度
flags	weight	保留
SID/index/label (可变)		

图 8 BGP peering SID 格式

flags 字段:8 bit,定义标志位如图 9 所示。

V	L	保留
---	---	----

图 9 flags 标志位

V 标志位:若置“1”,Adj-SID 携带一个绝对值,缺省置位。

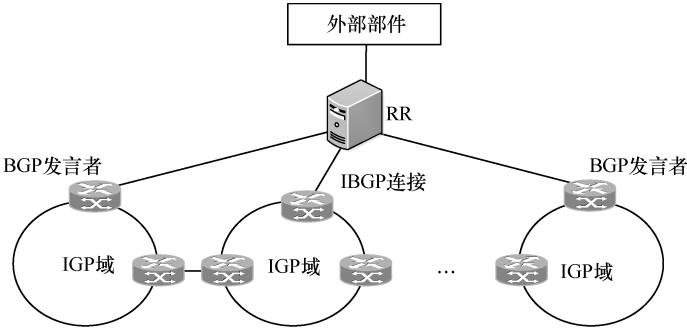


图 6 跨域链路状态信息收集



L 标志位:若置“1”,Adj-SID 携带值具有本地意义,缺省置位。

其余 bit,必须置“0”。

weight 字段意义同图 5,用于链路负载均衡。

为了便于理解,以图 10 为例,对分段路由的 BGP 扩展操作做进一步说明。

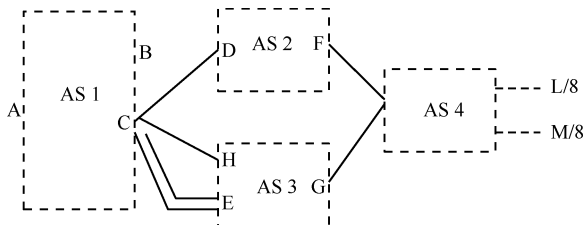


图 10 EPE 拓扑示意

图 10 中,假定为出口节点 C 及 peer D、H、E 及其接口分配如下地址。

- 链路 CD 中连接节点 C 的接口地址:1.0.1.1/24,连接节点 D 的接口地址:1.0.1.2/24。
- 链路 CH 中连接节点 C 的接口地址:1.0.2.1/24,连接节点 H 的接口地址:1.0.2.2/24。
- 链路组 CE 的两条并行链路中上一条连接节点 C 的接口地址:1.0.3.1/24,连接节点 E 的接口地址:1.0.3.2/24。
- 链路组 CE 的两条并行链路中下一条连接节点 C 的接口地址:1.0.4.1/24,连接节点 E 的接口地址:1.0.4.2/24。
- 链路组 CE 的两条并行链路中上一条连接节点 C 的链路本地标识符:0.0.0.1.0.0.0.0。
- 链路组 CE 的两条并行链路中下一条连接节点 C 的链路本地标识符:0.0.0.2.0.0.0.0。
- 节点 E 用于 eBGP 多跳连接的 loopback (环回)地址:1.0.5.2/32。
- 节点 C 的 loopback 地址:3.3.3.3/32, 并为其分配 SID:64。
- 节点 C、D、E、H 的 BGP router-ID 分别为:3.3.3.3,4.4.4.4,5.5.5.5,6.6.6.6。

同时,出口节点 C 配置 3 种 BGP peering SID 如下所示。

- 为 peer D、H、E 分配的 peer-node-SID 分别为:1012、1022、1052。
- 为连接到 peer E 的上、下链路分配的 peer-Adj-SID 为 1032 和 1042。

- 为连接到 AS3 的 peer H 和 E 分配的 peer-set-SID 为 1060。

- 为节点 C 配置的本地 BGP-LS 标识为 10000。

BGP-LS 更新消息由节点 C 产生,并通过 iBGP 或 eBGP 向邻居通告。

下面描述出口节点 C 通告上述 BGP peering SID 的 BGP-LS 扩展。

①为 peer D 通告的 peer-node-SID(1012)

BGP-LS link NLRI 包含:

- 本地节点描述符 (BGP router ID、AS、BGP-LS 标识):3.3.3.3, AS1, 10000;
- 远端节点描述符 (BGP router ID、AS):4.4.4.4, AS2;
- 链路描述符 (本地接口地址和 peer 接口地址):1.0.1.1, 1.0.1.2。

BGP-LS link 属性包含 peer-node-SID(1012)

②为 peer H 通告的 peer-node-SID(1022)和 peer-set-SID (1060)

- 本地节点描述符 (BGP router ID、AS、BGP-LS 标识):3.3.3.3, AS1, 10000;
- 远端节点描述符 (BGP router ID、AS):6.6.6.6, AS3;
- 链路描述符 (本地接口地址和 peer 接口地址):1.0.2.1, 1.0.2.2。

BGP-LS link 属性包含 peer-node-SID (1022) 和 peer-set-SID(1060)。

③为 peer E 通告的 peer-node-SID(1052)和 peer-set-SID (1060)

- 本地节点描述符 (BGP router ID、AS、BGP-LS 标识):3.3.3.3, AS1, 10000;
- 远端节点描述符 (BGP router ID、AS):5.5.5.5, AS3;
- 链路描述符 (本地接口地址和 peer 接口地址):3.3.3.3, 1.0.5.2。

BGP-LS link 属性包含 peer-node-SID (1052) 和 peer-set-SID(1060)。

④为连接到节点 E 的上一条链路通告的 peer-Adj-SID (1032)

- 本地节点描述符 (BGP router ID、AS、BGP-LS 标识):3.3.3.3, AS1, 10000;
- 远端节点描述符 (BGP router ID、AS):5.5.5.5, AS3;
- 链路描述符 (本地接口标识和 peer 接口地址):0.0.0.1.0.0.0.0, 1.0.3.2。

BGP-LS link 属性包含 peer-Adj-SID(1032)。

⑤为连接到节点 E 的下一条链路通告的 peer-Adj-SID (1042)

- 本地节点描述符 (BGP router ID、AS、BGP-LS 标识): 3.3.3.3, AS1, 10000;
- 远端节点描述符 (BGP router ID、AS): 5.5.5.5, AS3;
- 链路描述符 (本地接口标识和 peer 接口地址): 0.0.0.2.0.0.0.0, 1.0.4.2。

BGP-LS link 属性包含 peer-Adj-SID(1042)。

2.4 分段路由的 OAM 机制

根据“分段路由结构”draft 的定义,分段路由能够直接应用于 MPLS 数据平面:段标识(SID)等同于 20 bit MPLS 标签,分段路由头部等同于 MPLS 标签栈。由于现有的 MPLS 架构需要 LDP 或 RSVP 等特定控制平面协议的支持,而分段路由结构中标签的分发基于 IGP 的泛洪机制,分段路由这种技术特点引发了不同于传统 MPLS 网络 LSP 故障检测和隔离的考虑。

“分段路由网络的 OAM 需求”draft^[16]对 OAM 需求提出了全面的描述,包括:

- SR 的 OAM 必须支持按需(on-demand)和连续性故障诊断功能;
- SR 的 OAM 报文必须准确跟随数据平面的流量路径;
- SR 的 OAM 报文必须有能力发现并执行等价多径路径(equal cost multipath, ECMP)和非等价多径路径(unequal cost multipath, UCMP);
- SR 的 OAM 报文必须有能力发现并执行可用路径,而不仅仅只执行最佳路径;
- SR 的 OAM 报文必须有能力检测到任何基于 node-SID 或 Adj-SID 路径转发失败;
- SR 的 OAM 必须有能力由 SR 域中任意节点发起到其他节点的连通性检测和连续性检查;
- 当出现连续性检查失败时,SR 的 OAM 可以支持快速故障定位以隔离故障节点;
- SR 的 OAM 应具有由集中控制器发起故障诊断的能力;
- 当集中控制器发起 OAM 检测时,必须有能力向 SR 域中任何边缘节点警示对应的路径或业务失败,收到警示的节点可以采取本地保护行动或弹出提示性消息;

- SR 的 OAM 必须有能力使用主动或被动方式测量报文丢失、报文时延或时延抖动等性能指标。

为了检验 SR 网络的路径(包括节点和链路)有效性, IETF 制定了 SR 网络中使用 MPLS 数据平面的 LSP ping/trace 测试规范^[17],该草案对 SR 网络存在问题进行阐明,同时在 target FEC stack TLV 定义了 3 个新的 SID 子 TLV: IPv4 IGP-prefix segment ID; IPv6 IGP-prefix segment ID; IGP-adjacency segment ID,用于解决 SR 网络路径有效性可能存在的问题。另外, draft-ietf-spring-oam-usecase^[18]还提出了一种可扩展的 MPLS 数据平面监视系统,并给出了应用场景,这进一步丰富了 SR 网络中的 OAM 机制。

3 分段路由在流量工程中的应用

由于分段路由是在网络边缘入口节点为 IP 报文添加一个包含一串按序排列 SID 列表的报文头部,中间节点只需根据报头列表中的顶层 SID 进行 IP 报文的逐跳转发。所有路径选择的策略状态信息都由入口节点维护,中间节点无需运行复杂信令协议来维护每流的状态信息,大大简化了网络设备的复杂度,同传统基于 MPLS 流量工程相比,具有良好的扩展性和路径选择的灵活性。因此,SR 实现网络的负载均衡和流量工程具有独特优势。下面从 3 个应用场景分别描述 SR 在负载均衡和流量工程中的应用。

3.1 IGP 域内负载均衡

传统 IP 网络基于最短路径转发报文,导致网络流量分布极度不均衡,一部分链路严重拥塞,而另一部分链路长期处于轻载状态。无法通过 IGP metric 调整来达到全网负载均衡的目的。图 11 中,用户终端连接 R1,访问的内容源连接 R2,假定拓扑中所有链路的 metric 值都相等。用户访问站点 1.1.0.0/16 的全部流量最终经过最短路径 R1-R2 到达目的站点,其他链路无法承载流量,处于空闲状态。这种情形在现实网络中普遍存在。这个问题在 SR 网络中可以得到轻松解决。

在 SR 网络中,假设为 R2、R3、R4、R5 的环回地址分配的 node-SID 分别为 102、103、104、105,并通过 IGP 扩展向全网通告。这样,网络中所有节点的 FIB 都维护着这些 SID 的标签转发信息。网络中的入口节点 R1 通过规划流量路径(一般由 SDN 控制器负责),可以基于多种流分类策略(如接入端口、源地址、报文五元组等)为 IP 报文添加带有 SID 列表的头部。例如,为 VIP 流选择 R1-R2 的最短路径, VIP 报文采用缺省方式转发,不携带包含 SID 列表

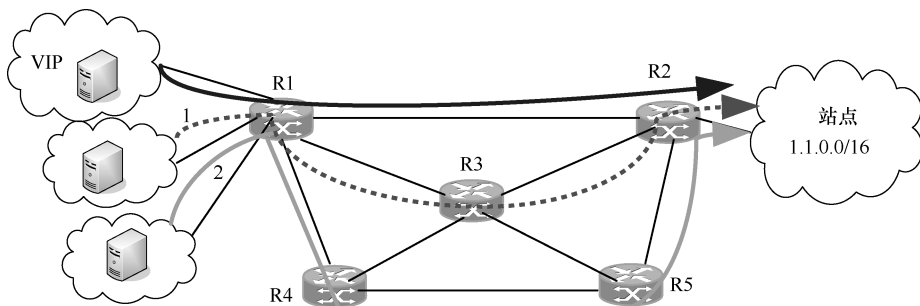


图 11 SR 实现 IGP 域内负载均衡

的头部；为来自端口 1 的报文流选择路径 R1-R3-R2，添加带有 SID 列表为{103}的头部；为源自端口 2 的报文流选择路径 R1-R4-R5-R2，添加带有 SID 列表为{104,105}的头部。入口节点 R1 通过合理规划流量路径，可实现流量在所有链路负载分担，达到全网流量均衡的效果。

3.2 骨干网双平面流量工程

随着互联网应用的飞速发展，过去几年建成的骨干网难以支撑互联网流量的爆炸式增长，各大运营商纷纷建设骨干网新平面，以承接汹涌而来的数字洪流。由于新老平面通常属于独立的自治域，无法实现两个平面的网络拓扑及流量数据的共享。采用 SDN 控制器可以收集跨域链路状态数据及流量工程数据，并根据新、老平面资源使用效率来动态调度城域网流量的合理分配，实现骨干网双平面流量工程。SDN 控制器与各个 AS 域中的边缘节点建立 BGP 会话，通过 BGP-LS 收集各个 AS 域的链路状态数据及流量工程数据，并根据需要为边缘入口节点的不同报文流选择流路径。入口节点还可以支持路径计算单元协议（path computation element protocol, PCEP），为特定业务流向 SDN 控制器请求路径选择。

图 12 中，城域网 A 与城域网 B 之间的互访流量通过新、老平面进行疏导。各个自治域都支持 SR，图 12 中各自治域的边缘节点 RA、RB、R1、R2、R3、R4 以 BGP-LS 方式向

SDN 控制器通告链路状态及流量工程数据，RA 为连接新平面的入口节点 R1 的链路分配的 Adj-SID 为 9001；为连接老平面的入口节点 R2 的链路分配的 Adj-SID 为 9002；新平面的出口节点 R3 为自身的环回地址分配的 node-SID 为 103；旧平面的出口节点 R4 为自身的环回地址分配的 node-SID 为 104。假设城域网 A 的用户和业务地址包含 2 个地址段：1.1.0.0/16 和 1.2.0.0/16，RA 为这 2 个地址段的业务流向 SDN 控制器请求路径选择。SDN 控制器根据这 2 个地址段承载的流量大小，并结合新、旧平面的网络资源利用情况，为地址段 1.1.0.0/16 的流量指定新平面路由，下发的路径标签为{9001,103}；为地址段 1.2.0.0/16 的流量指定旧平面路由，下发的路径标签为{9002,104}。RA 为源地址属于 1.1.0.0/16 地址段的报文添加路径标签为 {9001,103} 的头部，为源地址属于 1.2.0.0/16 地址段的报文添加路径标签为 {9002,104} 的头部，报文流在新、老平面遵循 ECMP 感知的最短路径转发到出口节点。这样，城域网 A 去往城域网 B 的流量便实现了新、老平面的负载分担。

3.3 域间多出口流量工程

在运营商骨干网与城域网及数据中心跨域互联中，通常存在多出口流量均衡问题。传统采用 BGP 路由策略来解决，存在配置复杂、流量调度不够灵活等局限，并且难以实现基于源地址的流量优化调度。第 2.3 节详细描述了分

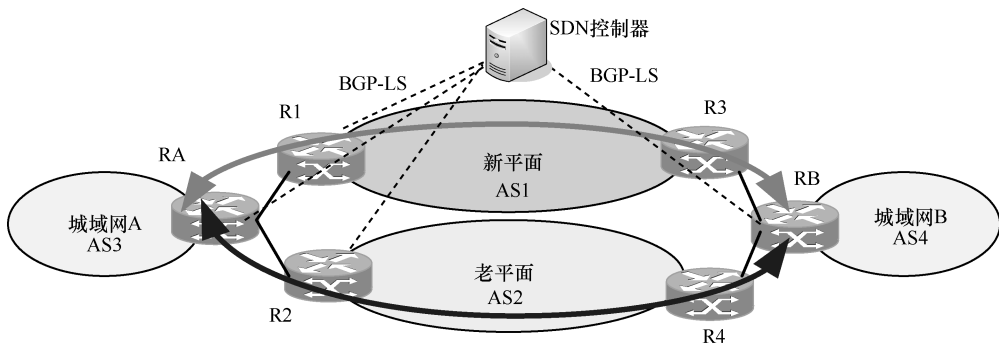


图 12 SR 实现骨干网双平面流量工程

段路由专门针对 EPE 的 BGP 扩展。下面以图 10 所示的 EPE 拓扑示意图为例,分析 SR 在多出口流量工程中的应用。

图 12 中,出口节点 C 通过 BGP-LS 方式向 SDN 控制器通告 BGP peering SID 关联的域间拓扑连接关系。同时,节点 C 的 FIB 维护 BGP peering SID 转发信息见表 1。

表 1 节点 C 的 FIB 维护 BGP peering SID 转发信息

入标签	操作	出接口
1012	POP	link to D
1022	POP	link to H
1032	POP	upper link to E
1042	POP	lower link to E
1052	POP	load balance on any link to E
1060	POP	load balance on any link to H or to E

为了确定目的网络的可达性,SDN 控制器需要收集域间 peer 通告所有可用路径,一种方式是利用 BGP 扩展支持的 ADD-path 能力^[9]。支持 ADD-path 能力的节点 C 将向 SDN 控制器通告相同网络前缀的所有 BGP 路由,如图 10 中节点 C 接收到的包含前缀 L/8 的 BGP 路由有:

- L/8 next-hop 1.0.1.2 AS path {AS 2, 4};
- L/8 next-hop 1.0.2.2 AS path {AS 3, 4};
- L/8 next-hop 1.0.5.2 AS path {AS 3, 4}。

SDN 控制器还需收集 IGP 域内链路状态拓扑信息,这可以通过 BGP-LS 方式来实现。

当 SDN 控制器获取了域内和域间完整的网络拓扑信息后,就可以根据策略规划流量出口工程。

EPE 策略可表述为携带 2 个 SID 的列表:第一个 IGP prefix SID 用于标识选择的出口节点;第二个 BGP peering SID 用于标识选择的域间 peer。SDN 控制器可以为目的前缀 L/8 选择如下路径:

- 出口节点 C 及域间 peer D:{64,1012};
- 出口节点 C 及域间 peer H:{64,1022};
- 出口节点 C 及域间 peer E:{64,1052};
- 出口节点 C 及连接域间 peer E 的上一条链路: {64,1032};
- 出口节点 C 及连接域间 peer E 的下一条链路: {64,1042};
- 出口节点 C 到 AS3 的任一条链路:{64,1060}。

SDN 控制器还可以指定 IGP 域内的显式路径,如到达目的地的 IP 报文必须经过节点 B 转发。假定节点 B 为自身的环回地址分配 SID 为 62,即可为目的前缀 L/8 选择一

条 SID 列表为{62,64,1 012}路径。

3.4 试验验证

目前,Cisco 应用于运营商网络从边缘、汇聚到核心的 ASR、NCS 及 CRS 全系列路由器都已支持分段路由技术。为验证其设备支持分段路由的功能和性能,选用了 ASR9000 (IOS-XR 5.2)、NCS6000 (IOS-XR 6.1.1)、CRS-3 (IOS-XR 6.1.1) 3 种型号设备混合组网,在实验室搭建了如图 3 所示拓扑测试环境。图 3 中所有互联链路带宽都为 10 Gbit/s。并根据第 3.1 节的描述配置网络参数,全网运行 OSPF 协议,在 R1 查看 LFIB 见表 2。

表 2 在 R1 查看 LFIB

入标签	出标签	下一跳
102	POP	R2
103	POP	R3
104	POP	R4
105	105	R4

测试仪表的 3 个 10 Gbit/s 端口分别同入口 R1 的端口 1、2、3 相连,用于发送测试流,其中,R1 的端口 3 模拟接收 VIP 流。测试仪表的 1 个 40 Gbit/s 端口同出口 R2 相连,配置为目标网络,用于接收测试流。测试仪表同时向每个端口发送 9 Gbit/s 流量,并通过 40 Gbit/s 端口进行流量回收,经过 24 h 的长期测试,测试结果无分组丢失现象。这完全验证了流量在所有链路负载分担,实现全网流量均衡。

同时,试验也验证了分段路由技术实现流量工程仅需路由设备支持基本的 IGP 扩展,设备配置较简单。而传统基于 MPLS 流量工程需要网络支持更多的信令协议,如 LDP、RSVP-TE 等,因此大大简化了网络运维的复杂度。由于分段路由无需中间设备维护流状态,仅需边缘入口设备对报文头部携带标签列表指示的流量转发路径进行编程,对流数目的支持具有良好的扩展性,非常适合运营商大规模网络的流量工程。

4 结束语

分段路由通过对 IGP 和 BGP 进行协议扩展实现段标识/标签的分发,网络节点维护段标识/标签的转发信息表。边缘入口节点为报文流维护策略状态信息,无需中间节点维护每报文流的策略状态信息,即可实现 IP 报文按指定路径在一个 IPG 域内甚至跨域的端到端转发,极大地简化了网络设备的复杂性。分段路由是对现有网络技术的继承和发展,并能实现同现有网络基础设施的互操作和平滑演进,



在网络虚拟化如二/三层 VPN 实现、网络节点及链路保护、负载均衡和流量工程等方面显示出了独特的优越性。分段路由由思想一经提出,便引起了网络界的极大兴趣,并得到设备制造商大力支持,目前正成为 IETF 的一项甚为活跃的标准化工作。由于分段路由技术的标准化工作起步较晚,目前还未达到成熟和完善阶段,特别是在 OAM、网络安全、与现有网络(包括 MPLS、IPv6)互操作等诸多方面^[20]还有许多工作要做。可以大胆预言,分段路由与 SDN 的结合,能够很好地实现网络的可编程性,进一步增强网络适配业务能力,为实现运营商大规模网络流量工程奠定基础。

参考文献:

- [1] PREVIDI S, FILSFILS C, DECRAENE B, et al. SPRING problem statement and requirements [J]. Der Orthop de, 2007, 36(4): 360-4.
- [2] FAMOUSSI B, ANDERSON L, CALLON R, et al. Constraint-based LSP setup using LDP[J]. Work in Progress Internet Draft, 2002, 38(2): 295-305.
- [3] AWDUCHE D, BERGER L, SWALLOW G, et al. RSVP-TE: extensions to RSVP for LSP tunnels[S]. RFC 3209, 2001.
- [4] FARREL A, VASSEUR J P, ASH J. A path computation element (PCE) based architecture [C]// IETF RFC, April 3-9, 2006, London, UK. New Jersey: IEEE Press, 2006.
- [5] 何晓明, 冀晖, 毛东峰, 等. 电信 IP 网向 SDN 演进的探讨[J]. 电信科学, 2014, 30(6): 131-137.
HE X M, JI H, MAO D F, et al. Telecom IP network to the exploration of the evolution of SDN [J]. Telecommunication Science, 2014, 30(6): 131-137.
- [6] 王熙. 撬动 2500 亿元市场的 SDN 技术为何依然难落地[EB/OL]. (2016-04-13) [2016-04-20]. <http://www.cww.net.cn/tech/html/2016/4/13/20164121826402633.htm>.
WANG X. Leveraged \$250 billion market SDN why technology is still difficult to be born[EB/OL]. (2016-04-13)[2016-04-20]. <http://www.cww.net.cn/tech/html/2016/4/13/20164121826402633.htm>.
- [7] HARES S, CHEN M. Summary of I2RS use case requirements[EB/OL]. (2016-02-16)[2016-04-05]. <http://www.ietf.org/>.
- [8] FILSFILS C, PREVIDI S, DECRAENE B, et al. Segment routing architecture [EB/OL]. (2016-03-06)[2016-04-05]. <http://www.segment-routing.net/home/ietf>.
- [9] PREVIDI S, FILSFILS C, FIELD B, et al. IPv6 segment routing header (SRH)[EB/OL]. (2016-02-11)[2016-04-05]. <http://www.ietf.org/>.
- [10] PREVIDI S, FILSFILS C, BASHANDY A, et al. IS-IS extensions for segment routing [EB/OL]. (2016-01-20)[2016-04-05]. <http://www.ietf.org/>.
- [11] PSENAK P, PREVIDI S, FILSFILS C, et al. OSPFv3 extensions for segment routing [EB/OL]. (2016-02-08)[2016-04-05]. <http://www.ietf.org/>.
- [12] PSENAK P, PREVIDI S, FILSFILS C, et al. OSPF extensions for segment routing[EB/OL]. (2016-03-17)[2016-04-05]. <http://www.ietf.org/>.
- [13] GREDLER H, MEDVED J, PREVIDI S, et al. North-bound distribution of link-state and TE information using BGP[EB/OL]. (2016-02-11)[2016-04-05]. <http://www.ietf.org/>.
- [14] PREVIDI S, PSENAK P, FILSFILS C, et al. BGP link-state extensions for segment routing [EB/OL]. (2016-03-09)[2016-04-05]. <http://www.ietf.org/>.
- [15] PREVIDI S, FILSFILS C, RAY S, et al. Segment routing egress peer engineering BGP-LS extensions [EB/OL]. (2016-02-15)[2016-04-05]. <http://www.ietf.org/>.
- [16] KUMAR N, PIGNATARO C, AKIYA N, et al. OAM requirements for segment routing network [EB/OL]. (2016-02-01)[2016-04-05]. <http://www.ietf.org/>.
- [17] KUMAR N, SWALLOW G, PIGNATARO C, et al. Label switched path (LSP) ping/trace for segment routing networks using MPLS dataplane[EB/OL]. (2016-02-07)[2016-04-05]. <http://www.ietf.org/>.
- [18] GEIB R, FILSFILS C, PIGNATARO C, et al. A scalable and topology-aware MPLS dataplane monitoring system [EB/OL]. (2016-02-09)[2016-04-05]. <http://www.ietf.org/>.
- [19] WALTON D, RETANA A, CHEN E, et al. Advertisement of multiple paths in BGP[EB/OL]. (2016-03-11)[2016-04-05]. <http://www.ietf.org/>.
- [20] FILSFILS C, PREVIDI S, BASHANDY A, et al. Segment routing interoperability with LDP [EB/OL]. (2016-03-21)[2016-04-05]. <http://www.ietf.org/>.

[作者简介]



何晓明(1968-),男,博士,中国电信股份有限公司广州研究院高级工程师,主要从事数据网络研究和支撑工作,研究兴趣包括网络架构、协议、流量工程、SDN/NFV等,已发表论文 30 余篇,申请专利 26 项,获授权专利 10 余项。

卢泉(1969-),女,中国电信股份有限公司广州研究院高级工程师,主要从事数据网络研究和支撑工作。

邢亮(1978-),男,中国电信集团公司网络运行维护事业部数据处工程师,主要从事数据网运行维护管理工作。