



SR 软件路由器的设计与实现

高晓敏¹, 刘雪晶², 李佳伟¹, 罗洪斌¹, 董平¹

(1. 北京交通大学电子信息工程学院, 北京 100044;

2. 中国电子科技集团公司电子科学研究院, 北京 100041)

摘要:分段路由(segment routing, SR)是近几年 IETF 提出的一种路由机制,其目的是使 IP/MPLS 网络面向服务并且更高效。SR 减少现有 MPLS 网络控制平面复杂的协议(LDP/RSVP-TE),并实现流量工程等机制。与基于 RSVP 的显式路径中隧道中间节点维护报文流的状态不同,基于 SR 的显式路径只在入口边界路由器中维护报文流的状态,从而提高了网络的可扩展性,因此 SR 对于下一代互联网具有重要的研究意义。对 SR 的实现机制进行了研究,利用软件路由平台 Click,对基于 PCE 架构的 SR 控制平面子系统和基于 MPLS 的 SR 数据平面子系统的各个模块进行实现,并在实际中进行部署,以验证 SR 软件路由器的可行性。

关键词:分段路由;软件路由器;路径计算单元;多协议标签交换;下一代互联网

中图分类号:TP393

文献标识码:A

doi: 10.11959/j.issn.1000-0801.2017032

Design and implementation of software router supporting segment routing

GAO Xiaomin¹, LIU Xuejing², LI Jiawei¹, LUO Hongbin¹, DONG Ping¹

1. School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China

2. Electronic Science Research Institute of China Electronic Technology Group Corporation, Beijing 100041, China

Abstract: Segment routing (SR) is a routing mechanism proposed by IETF in recent years, which aims to make the IP/MPLS network service-oriented and more efficient. SR eliminates the complex control plane protocol (LDP/RSVP-TE) of the MPLS network, while SR implements the traffic engineering mechanism. Different from the explicit path in the RSVP tunnel which maintains the state of each flow in the middle of the tunnel, the explicit path based on SR only maintains the state of flow at the entrance border router of the SR domain. For this reason, SR improves the scalability of the network. Therefore the research of the SR mechanism is very important to the next generation internet. The realization mechanism of SR was studied, using the click router. The SR control plane subsystem based on PCE architecture and each module of MPLS-based SR data plane subsystem were implemented and deployed in practice to verify the feasibility of the software router in actual deployment.

Key words: segment routing, software router, PCE, MPLS, next generation network

收稿日期:2016-12-11;修回日期:2017-01-23

基金项目:国家重点基础研究发展计划(“973”计划)基金资助项目(No.2013CB329100);国家自然科学基金资助项目(No.62171200, No.61422101, No.61232017);教育部高等学校博士学科点专项科研基金资助项目(No.20130009110014)

Foundation Items: The National Basic Research Program of China(973 Program)(No.2013CB329100), The National Natural Science Foundation of China (No.62171200, No.61422101, No.61232017), The Research Fund for the Doctoral Program of Higher Education of China (No.20130009110014)

1 引言

互联网的迅猛增长对当前基于 IP(internet protocol)的网络提出了挑战。新的应用层出不穷,而传统的 IP 网络没有数据属性的判别能力,也就谈不上按数据属性来区分服务;网络规模迅速增长,但现有的 IP 路由技术及组网方式不适应网络的扩展。MPLS (multiprotocol label switching,多协议标签交换)^[1] 技术是一种在开放的通信网上利用标签引导数据高速、高效传输的技术。其主要优点是兼容现有各种主流网络技术,在提供 IP 业务时能确保 QoS(quality of service)和安全性,具有流量工程(traffic engineering,TE)能力。然而 MPLS 控制平面复杂的网络协议 LDP(label distribution protocol)或 RSVP-TE(reservation protocol with TE extension)都大大增加了 MPLS 网络实际部署的难度。为此,IETF(Internet Engineering Task Force,国际互联网工程任务组)提出了分段路由(segment routing,SR)机制^[2],将 MPLS 的数据平面应用于 SR,在保留了 MPLS 优越性的同时,不使用包含复杂通信信令的控制协议,大大简化了网络部署难度。

软件定义网络(software defined networking,SDN)是一种新兴的网络技术,它将网络的控制平面与数据平面分离,使得原本紧密结合的独立网络设备,变成了可访问、可编程的计算设备。SR 属于一种 SDN 的实现方式,其将控制平面和数据平面分离,提高了网络的灵活性和可扩展性,受到了国内外学者的广泛关注。参考文献[3]论述了在线和离线两种情况下决定 SR 路径计算的最优参数,主要解决了 SR 中流分割值的最优参数的选取,这些值的选取依据等效多路径(equal cost multi path,ECMP),从而将离线情况下最坏情况的链路利用率最小化。本文除考虑链路利用率外,还考虑了时延/跳数,能够更好地满足用户服务层需求。参考文献[4]提出了在部署大规模分布式设备时的能量消耗问题,利用基于 SR 能量效率的流量工程来减少骨干网络中的能源消耗,利用 SDN 的方法,可以选择性地关闭一些链路,以节省能源。主要目标是决定网络设备的状态,看是需要关闭还是打开。参考文献[5]介绍了将 SDN 模型和基于 SR 的流量工程结合在一起的架构,主要解决如何将计算的路径映射到 SR 路径上,提出了一种 SR 路径分配算法,以此来计算 SR 路径。参考文献[6]提出了一种 SR 路径标识编码算法,尽量减少 SR 路径标识的长度。本文除了考虑 SR 路径标识的长度问题外,更注重考虑如

何使网络尽可能达到负载均衡,并减少请求拒绝率,以减少 SR 路径标识的平均长度,而不只是单纯地缩短 SR 路径标识的长度。参考文献[7]描述了 SR 在多层网络中的应用,并使用 SDN 控制器作为控制平面,但并没有讨论 SDN 控制器使用何种算法计算路径。参考文献[8]将 SR 的理念应用于现有的 CE(carrier Ethernet,电信级以太网),提出了一种分层策略,使其满足服务提供商的需求。

现有的国内外针对 SR 相关机制的研究主要是在仿真平台进行实现,不能在实际中部署;另一方面,由于现有的硬件路由器集成度高,都是针对现网而设计的,无法对其内部结构进行修改,不适于下一代互联网的研究与实现。因此,本文的研究基于软件路由器进行实现,既避免了现有硬件的限制,又能够进行实际的部署。本文选用了 Click^[9] 软件路由平台提供的转发框架进行 SR 软件路由器的实现,包括 SR 控制平面子系统和 SR 数据平面子系统的设计与实现。

2 SR 协议简介

SR 利用源路由模型,节点通过一系列的 segment(段)来转发分组。segment 可以表示拓扑里面的任何实体或者服务。一个 segment 可以具有本地意义也可以具有全局意义。用户可以将任何路径编码为一系列的 segment,即 SL(segment list),将其封装在数据分组头,使数据分组沿着特定的路径进行转发,即 SR 允许强制一个流通过任意拓扑路径和数据服务链,同时只在 SR 域内的入口节点处维护每条数据报流的状态。

segment 通过 SID(segment ID)来标识。当 SR 数据平面应用 MPLS 时,一个 SID 编码成为一个 MPLS 标签^[10]。一系列 SID 被编码成一个标签堆栈。一旦 SID 完成其相应的动作,就会有相应的标签出栈。

SR 利用节点 SID(node SID)来表示数据分组沿着到达该节点的最短路径转发,即同一个 SR 域的每个节点都在本地转发表保存了到达其他节点的信息,作为默认路由,节点 SID 是全局的。另外,SR 还使用邻接 SID(adjacency SID)来控制流的转发,邻接 SID 表示数据分组沿着某条链路转发,即规定了在某个节点沿着特定的网口转发,邻接 SID 是局部的。不同于节点 SID 的是,邻接 SID 的每个节点在转发表中保存本地的邻接 SID 规则即可。SR 将一系列节点 SID 和邻接 SID 按照特定的方式组合成 SL,用来标识一条流量工程的路径(即源路



由)。SR 入口路由器将 SL 封装在数据分组的头部;SR 中间节点路由器只需根据 SL 进行分组转发即可。因此,SR 的中间节点不需要保留数据报流的任何状态,从而增强了可扩展性。

每个 SR 路由器在本地维护了一个 LFIB (label forwarding information base, 标签转发信息库),如图 1 所示。LFIB 定义了分组转发的下一跳和对 SR 标签堆栈执行的操作,操作包括 3 种:push(标签入栈)、continue(标签不变)、next(标签出栈)。LFIB 条目可以由本地策略产生,或者从 IGP 通告协议得知。当分组到达时,根据栈顶标签(入标签)以及入口,查找下一跳出口以及出标签和相应的 SR 操作。

入接口	入标签	出接口	出标签	操作
-----	-----	-----	-----	----

图 1 标签转发信息库示意

SR 的控制平面利用 PCE(path computation element)架构中^[1]的 PCE 作为控制器。与 OpenFlow 协议中的控制器最大的不同是,由于 SR 是一种源路由,PCE 只需要为入口路由器配置路由转发条目,而不需要像 OpenFlow 中的控制器需要为数据平面的每个交换机配置转发条目以及其他的转发规则,大大降低了协议的复杂度。SR 系统示意图如图 2 所示。

本文实现的 SR 系统由 1 个 SR 域和 2 个 IP 域构成。SR 域的主要功能是连通两个 IP 域。SR 系统主要分为控制平面子系统和数据平面子系统两部分。控制平面由 PCE 作为控制器,数据平面由域边界路由器(R1、R6)和域内路由器(R2~R5)组成。控制平面 PCE 控制器和

数据平面边界路由器通过 PCEP (path computation element communication protocol)^[12]进行通信。本文主要介绍控制平面子系统和数据平面子系统的具体实现,着重介绍控制平面子系统中启发式的具有带宽保证的流量工程算法。

3 SR 控制平面子系统实现

SR 的控制平面基于 PCE 架构实现。PCE 架构由 PCE 和 PCC(path computation client)两部分组成。当 PCE 应用于 SR 的控制平面时,PCC 即作为入口路由器,主要负责为进入 SR 域内的数据分组封装 SR 分组头(SL);PCE 即作为 SR 的控制器,主要功能是为到达 PCC 的数据分组计算一条满足服务层需求的 TE 路径,并用 SL 来标识此 TE 路径。PCE 和 PCC 之间通过 PCEP 进行通信,PCEP 主要通过两个消息完成核心功能:PCReq(path computation request, 路径计算请求)消息和 PCRep(path computation reply, 路径计算应答)消息。

PCE 控制器的组成模块如图 3 所示。PCE 控制器主要包括以下模块:SR-TE 路径计算模块、SL 优化模块、感知消息处理模块、PCEP 会话管理模块。当接收到来自数据平面的数据分组时,首先对数据分组进行分类;若为 PCEP 消息,则发送至 PCEP 会话管理模块;若为感知消息,则发送给感知消息处理模块。PCEP 会话管理模块将 PCReq 消息发送至 SR-TE 路径计算模块进行路径计算,将计算好的 SR-TE 路径发送至 SL 优化模块对 SL 进行优化,之后将优化的 SL 发送至 PCEP 会话管理模块,封装成 PCRep 消息发送至数据平面。感知消息处理模块负责接收来自数据平面的实时链路状态信息,并对流量工程数据库进行更新。

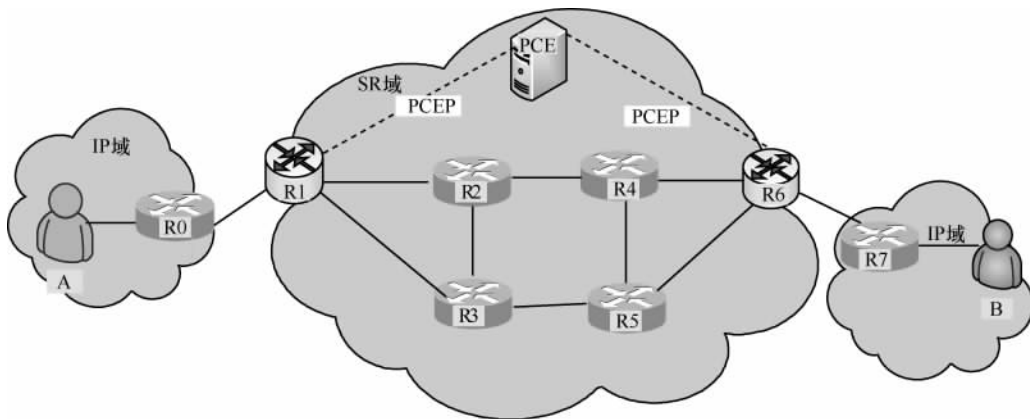
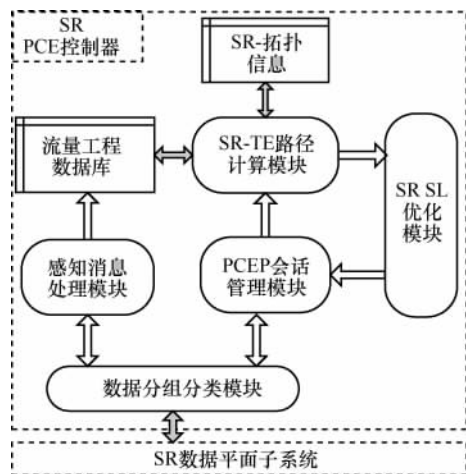


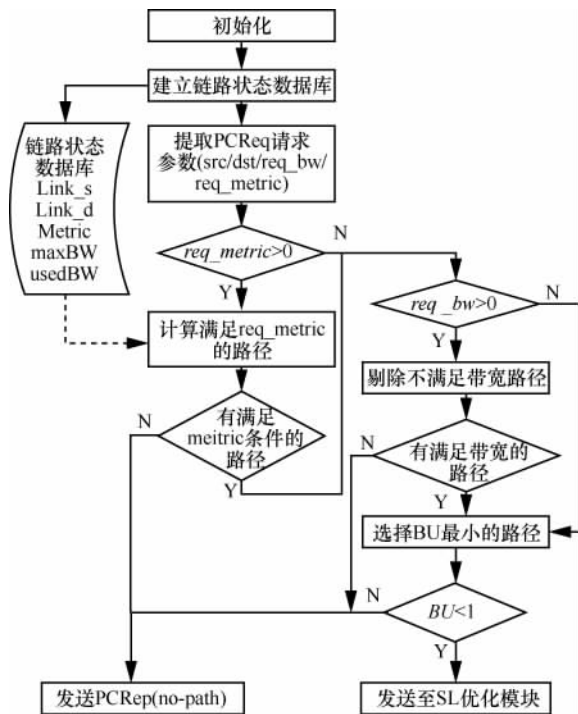
图 2 SR 系统示意



3.1 SR-TE 路径计算模块

SR-TE 路径计算模块主要负责为进入域内的数据分组计算一条符合服务层需求的流量工程路径,使用的算法是一种启发式的具有带宽保证的路由算法,此算法不仅实现了使用流量工程来达到负载均衡的目的,同时也减少了网络的分组头开销。本算法的限制条件为跳数/时延、链路带宽利用率。目标是尽可能使网络达到负载均衡,并尽可能减少路径请求拒绝率,从而提升网络的性能。

启发式的具有带宽保证的路由算法的具体实现流程如图4所示。工作流程介绍如下。



步骤 1 初始化, 根据网络拓扑, 建立链路状态数据库, 包括每条链路的两端点、记录链路的 metric 值、链路的最大容量以及已用带宽和剩余带宽。

步骤 2 当接收到来自 PCC 的 PCReq 消息时,从消息中提取源地址、目的地址、metric 限制值 *req_metric*、需求带宽值 *req_bw*。

步骤 3 若 $req_metric > 0$, 则表明路径请求中包含了 metric 限制, 利用改进的 K 最短路径, 计算出满足 req_metric 的路径, 并存入 TE 路径库; 若 $req_metric = 0$, 则表明路径请求不包含 metric 限制, 则利用 K 最短路径计算出前 K 条路径, 存入 TE 路径库。

步骤 4 判断 TE 路径库中是否包含计算出的路径：
若没有，则向 PCEP 会话模块发送 PCRep no-path 消息，即拒绝请求；若有路径，则转到步骤 5。

步骤 5 若 $req_bw > 0$, 剔除 TE 路径库中不满足带宽限制的路径, 若此时 TE 路径库中没有路径, 则表明不能满足条件, 向 PCEP 会话模块发送 PCRep no-path; 若 TE 路径库中有路径, 再从剩余的路径中选出链路带宽利用率最小的路径, 以达到负载均衡的目的; 若 $req_bw = 0$, 则表明路径请求中不包含 bw 限制, 直接依据链路状态数据库, 在 LSP 路径库中选出链路带宽利用率最小的路径; 若最后计算出 TE 路径, 则表明该限制条件下 TE 路径存在, 则将此 SR-TE 路径转发给 SL 优化模块。

如上所述,该算法是一种具有带宽保证的路由算法,能够满足数据流的所需带宽、metric 等服务层需求,并且能够使网络尽可能达到负载均衡。

3.2 SL 优化模块

当 SR-TE 模块计算出满足条件的 TE 路径时, 使用路径上节点的 SID 组成 SL 来标识此 TE 路径。SID 可以表示多跳, 为了增加有效带宽利用率, 要使 SL 尽可能的短, 最理想的情况是 SL 只有 1 个标签, 即目的节点 SID, 但这只能标识最短的路径, 且无法区分等效多路径 (ECMP)。为了标识 SR-TE 模块计算出的满足服务层需求的 TE 路径, 首先使用每个节点 SID 组成初始 SL, 之后根据具有 ECMP 感知的最短路径压缩算法, 对初始 SL 进行优化, 计算出尽可能短的 SL。SL 优化算法流程如图 5 所示。

由于 SR 中的节点 SID 可以表示多跳, 即可以表示最短路径, 因此在减少 SL 中的 SID 时, 可以利用这一点作为依据。即当路径中的某一个节点 SID 不能表示该路径时,

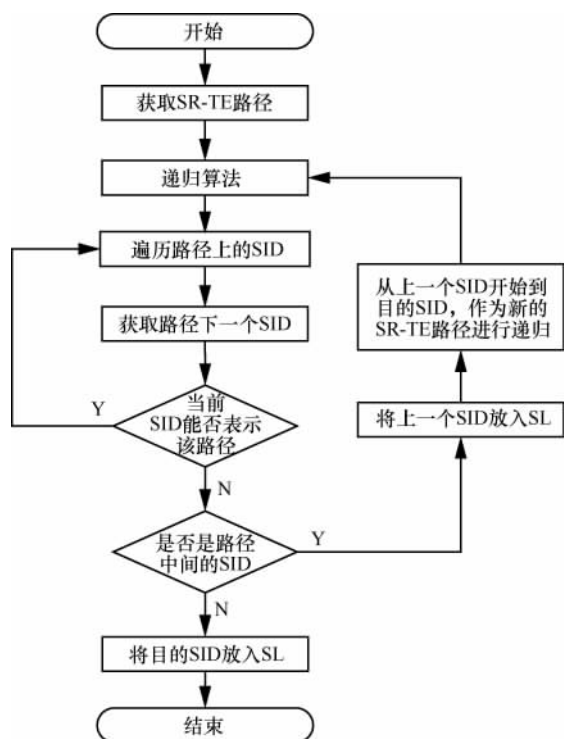


图5 SL优化算法流程

就需要将前一个节点SID加入SL中。不能表示路径有两种情况：不是最短路径和具有等价多路径。本算法使用Floyd算法计算SR域内任意两个节点之间的最短路径，同时也可以计算任意两点间是否有等价多路径。

3.3 其他辅助模块

(1) 感知消息模块

感知消息模块的主要功能是收集数据平面每条链路的实时带宽，是SR-TE路径计算模块的基础，以保证链路剩余带宽的实时性、增加TE路径计算的灵敏度和可靠性，是提高QoS的关键。

感知消息是通过数据平面路由器进行主动更新的，PCE控制器被动更新。当PCE收到来自数据平面路由器的感知更新提醒消息时，立即返回感知更新请求消息，以请求更新；当收到感知更新消息时，根据收到的实时流量更新链路状态库中对应链路的已用带宽，为SR-TE路径计算模块中的路径计算提供实时的带宽数据，以提高路径计算的准确性和实时性。

(2) PCEP会话管理模块

PCEP会话管理模块负责对数据平面发来的PCEP相关消息进行处理，当PCE收到数据平面的入口边界路由器发来的PCReq消息时，将消息传送给SR-TE路径计算

模块；收到PCRep消息时，将消息传送给数据平面的入口边界路由器。

4 SR数据平面子系统实现

SR的数据平面是基于MPLS的数据平面进行实现的。数据平面由域内路由器（center router, CR）、边界路由器（border router, BR）组成。

边界路由器的主要功能是将到达SR域的数据分组转化成能够被SR域内路由器识别的SR数据分组，即为其封装相应的SL标签堆栈，并在其离开SR域时为其剥除SR标签堆栈。同时，边界路由器作为SR域内的一部分，需要具备SR域内路由器的全部功能。因此，下面只详细介绍边界路由器。

SR边界路由器组成模块如图6所示。主要包括IP数据分组处理模块、SR数据分组处理模块、PCEP会话处理模块、检查更新模块、感知消息处理模块以及相关信息表（主要包括本地流量工程路径库（traffic engineer data base, TE-DB）、转发等价类表（forwarding equal class, FEC）、标签转发信息表（LFIB））。IP数据分组处理模块为进入SR域的IP数据分组封装SL，若在本地的TE-DB中没有查找到相应的SL，则将数据分组缓存在本地，并向PCEP会话处理模块发送路径请求，若查找到SL，则将SL封装在IP分组的分组头转换成SR数据分组，再转发给SR数据分组处理模块。PCEP会话处理模块在接收到IP数据分组处理模块的路径请求时，通过PCEP向控制平面请求路径，并将控制平面返回的路径存入TE-DB，同时检查更新模块定期对TE-DB中的路径条目更新。感知消息处理模块定期将各网口的流量信息汇报给控制平面，以便于控制平面计算流量工程路径。SR数据分组处理模块从域内SR路由器或者IP数据分组处理模块接收SR数据分组，根据分组头的SL对数据分组进行转发。

4.1 IP数据分组处理模块

IP数据分组处理模块负责处理IP数据分组。IP数据分组若是从域外转发至域内，称为OI分组，若是从域内转发至域外，称为IO分组。

(1) OI分组处理流程

步骤1 当收到数据分组，解析数据分组头，根据版本字段判断是否为IP分组，若是，则转到步骤2，否则，丢弃。

步骤2 解析IP数据分组，提取目的IP地址、需求带

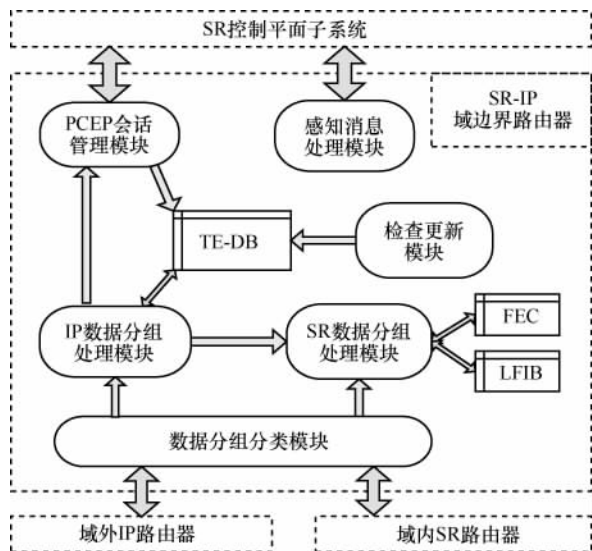


图6 SR边界路由器模块

宽、需求 metric, 依据目的 IP 地址、需求带宽、需求 metric 查找转发等价类表(FEC), 找到目的 SID, 查找本地路径库(TE-DB); 若查找到相对应的 SL, 则为 IP 分组封装 SR 分组头, 转发给 SR 数据分组处理模块; 若查找不到相对应的 SL, 则将数据分组进行缓存, 并将路径请求转发给 PCEP 会话处理模块进行路径请求。

(2) IO 分组处理流程

步骤 1 当收到数据分组, 解析数据分组头, 根据版本字段判断是否为 IP 分组, 若是, 则转到步骤 2, 否则, 丢弃。

步骤 2 解析 IP 数据分组, 提取目的 IP 地址, 若目的 IP 地址是自己, 则接收此数据分组; 否则, 为 IP 分组封装 MAC 层, 转发至相应的网口。

4.2 SR 数据分组处理模块

SR 数据分组处理模块负责处理 SR 数据分组。工作过程如下。

步骤 1 当收到数据分组, 判断目的 MAC 地址是否是自己。若是, 则转到步骤 2; 否则, 丢弃。

步骤 2 判断数据分组是否是 SR 数据分组, 判断依据是 MAC 层 type 字段是否为 0x8847 (这里使用 MPLS 的 MAC 层 type 字段的值)。若是, 则剥掉 MAC 层, 转到步骤 3; 否则, 丢弃。

步骤 3 解析 SR 数据分组头, 提取 SR 分组头 SL 中的第一个 SID, 查找本地标签转发信息表(LFIB), 依据 SL 中第一个 SID 即入口 SID 查找出口 SID、标签操作和转发网口。

步骤 4 进行标签操作 (continue、next): 当出口 SID 标识下一跳时, 执行 next 操作, 即将此 SID 从 SR 分组头中剥掉; 当出口 SID 不是标识下一跳时, 执行 continue 操作, 即 SR 分组头不变。

步骤 5 为 SR 数据分组封装 MAC 层分组头, 根据查找到的转发网口将 SR 数据分组转发至相应的网口。

4.3 其他辅助模块

(1) 相关数据信息表

TE-DB: 用于存放向 PCE 控制器请求的 TE 路径, 并由检查更新模块定时更新。

FEC: 用于区分不同的服务, 对到达的数据流进行分类, 并将同种类型的服务请求按照同一种转发方式进行转发。

LFIB: 本地标签转发信息表。

(2) 检查更新处理模块

负责维护本地 TE-DB, 维护一个定时器, 定期检查本地 TE 路径条目, 根据条目中维护的时间戳, 判断该条目是否失效 (即是否超时)。若失效, 则将此条目从 TE-DB 中删除; 否则, 继续检查其他 TE 路径条目。

(3) PCEP 会话处理模块

负责处理 PCEP 的相关消息。作为边界路由器, 当收到从域外发送至域内的数据分组时, 首先查找本地 TE-DB, 若本地没有相应的路径, 边界路由器则需要作为 PCE 架构中的 PCC, 使用 PCEP 向 PCE 控制器请求路径。当收到 IP 数据分组处理模块发送的路径请求消息时, 将请求信息封装成 PCReq 消息发送给控制器; 当收到控制器回送的 PCRep 消息时, 将应答消息中的 SL 发送给 IP 数据分组处理模块。

(4) 感知消息处理模块

负责监测并记录各网口的实时流量, 并维护定时器, 定期向控制器汇报各网口的实时流量, 以便于控制器更合理地进行流量工程路径计算。

5 测试结果及分析

为验证 SR 软件路由器的功能并测试其性能, 在实验室实体机柜上部署了测试拓扑。SR 域中标识 R1~R6 的节点 SID 分别为 16001、16002、16003、16004、16005、16006; 链路带宽除 R3-R5 和 R4-R6 两条链路带宽为 50 Mbit/s 外, 其余链路带宽均为 100 Mbit/s。实验中只使用了节点 SID, 没有涉及邻接 SID。



为了测试 SR 软件路由器的功能,IP1 域中的 A 用户向 IP2 域中的 B 用户发送数据,数据请求带宽 req_bw 为 80 Mbit/s,请求 metric 限制值为 10。

数据分组在 SR 域内的转发过程如下:

步骤 1 当 IP 数据分组到达入口边界路由器 R1 时,在本地路径库中查找不到对应的 SL,则向 PCE 发送 PCReq 消息,如图 7 所示,该消息包含 end-point object(记录了源 IP 地址、目的 IP 地址)、bandwidth object(记录了请求带宽)、metric object(记录了请求 metric 限制值)。

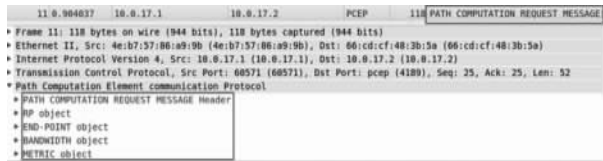


Figure 7 shows a Wireshark capture of a PCReq message. The packet list shows a Path Computation Request Message (Type: MPLS label switched packet) with a length of 118 bytes. The packet details show the Path Computation Request Message Header, End-point object, Bandwidth object, and Metric object.

图 7 PCReq 消息

步骤 2 PCE 收到路径请求消息,为其计算的 SR-TE 路径为 R1→R2→R4→R5→R6,对应的 SL 为{16004,16005,16006},PCE 将此标签堆栈封装在 PCRep 消息中的 ERO (explicit route object)显式路径字段中,如图 8 所示,并将此消息返回给入口边界路由器 R1。

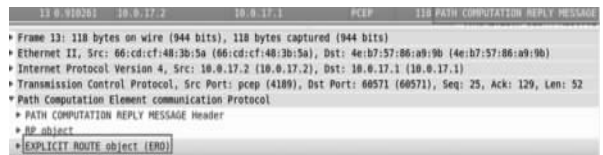


Figure 8 shows a Wireshark capture of a PCRep message. The packet list shows a Path Computation Reply Message (Type: MPLS label switched packet) with a length of 118 bytes. The packet details show the Path Computation Reply Message Header, End-point object, Bandwidth object, and Metric object.

图 8 PCRep 消息

步骤 3 入口边界路由器 R1 将 SR 标签堆栈封装在请求分组的头部进行转发,如图 9 所示(由于使用 Wireshark 抓取分组,其将 SR 分组识别为 MPLS 数据分组)。

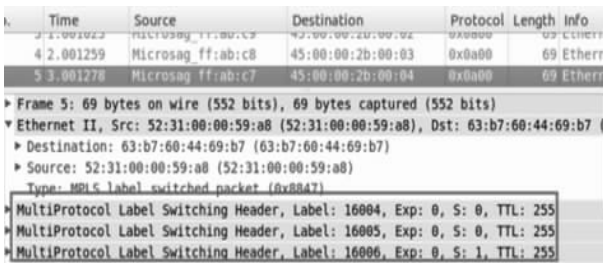


Figure 9 shows a Wireshark capture of a data packet from R1. The packet list shows a packet of 69 bytes. The packet details show the Ethernet II header, Internet Protocol Version 4 header, and MultiProtocol Label Switching Header (Label: 16004, Exp: 0, S: 0, TTL: 255).

图 9 R1 数据分组 SR 标签堆栈

步骤 4 数据分组到达 R2,由于下一跳为 R4,故将 R4 的 SID 剥去,SL 变为{16005,16006},如图 10 所示。

步骤 5 数据分组到达 R4,由于下一跳为 R5,故将 R5 的 SID 剥去,SL 变为{16006},如图 11 所示。

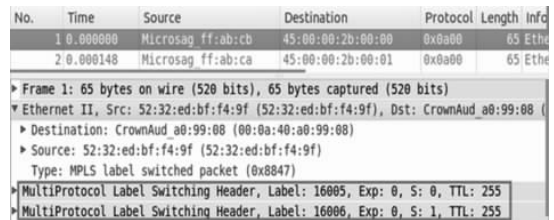


Figure 10 shows a Wireshark capture of a data packet from R2. The packet list shows a packet of 65 bytes. The packet details show the Ethernet II header, Internet Protocol Version 4 header, and MultiProtocol Label Switching Header (Label: 16006, Exp: 0, S: 1, TTL: 255).

图 10 R2 数据分组 SR 标签堆栈



Figure 11 shows a Wireshark capture of a data packet from R4. The packet list shows a packet of 61 bytes. The packet details show the Ethernet II header, Internet Protocol Version 4 header, and MultiProtocol Label Switching Header (Label: 16006, Exp: 0, S: 1, TTL: 255).

图 11 R4 数据分组 SR 标签堆栈

步骤 6 数据分组到达 R5,由于下一跳为 R6,故将 R5 的 SID 剥去,数据分组变为 IP 数据分组,如图 12 所示。

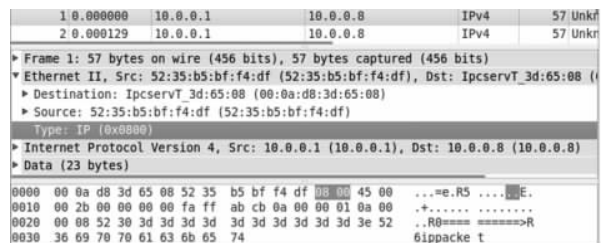


Figure 12 shows a Wireshark capture of a data packet from R5. The packet list shows a packet of 57 bytes. The packet details show the Ethernet II header, Internet Protocol Version 4 header, and Data (23 bytes).

图 12 R5 数据分组 IP 报文

步骤 7 数据分组到达 R6,域内网口接收到 IP 数据分组,则转发至域外网口,最终数据分组经过 R7 送达用户 B。

上述结果表明,SR 软件路由器能够按照原始设计正常工作。

为进一步验证 SR-PCE 控制器的性能,用户 A 无间隔地随机产生不同的请求数据分组,测得请求输入速率和 PCE 的 CPU 占用率,如图 13 所示。

根据实验数据可知,当 CPU 占用率接近 100%时,请求速率达到 160 000 分组/s。由于 SR 软件路由器是通过 Click 平台实现的,并且代码运行在用户层,且 Click 只能运行在单核上,不能够有效利用多核 CPU,因此未来需在 DPDK 等多核高速转发的平台部署 SR-PCE 控制器,以突破由 Click 平台产生的瓶颈。

6 结束语

本文详细阐述了 SR 系统中控制平面子系统和数据平

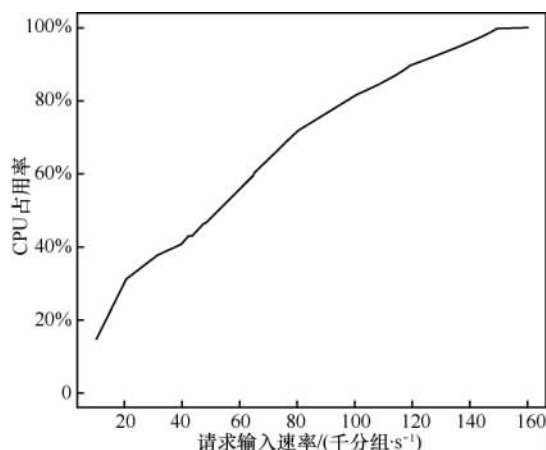


图 13 SR-PCE CPU 占用率

面子系统各个模块的具体软件实现,着重研究了基于 PCE 架构的控制平面子系统中一种具有带宽保证的流量工程路径算法,使网络尽可能地达到负载均衡,同时使用尽可能短的标签对 SR-TE 路径进行标识,提高有效带宽利用率,使转发更高效。

SR 作为一种 SDN 的实现方式,将控制平面和数据平面相分离,通过控制器和入口边界路由器利用源路由的思想实现分组的转发,提高了转发效率,并且本身就能够支持流量工程机制,使网络更高效。相比于 MPLS,SR 的网络更加简单,在实际网络中更便于部署,在下一代互联网中具有很好的应用前景。本文只描述了 SR 软件路由器系统各个模块的具体实现,没有将 SR 进行大规模部署,有待进一步的实现。

参考文献:

- [1] ROSEN E, VISWANATHAN A, CALLON R. Multi protocol label switching architecture[EB/OL]. (2001-01-20) [2016-12-11]. http://xueshu.baidu.com/s?wd=paperuri%3A%287a99238572d7ae9cedda5c41d7d3edb8%29&filter=sc_long_sign&sc_ks_para=q%3DRFC%203031%20Multiprotocol%20Label%20Switching%20Architecture&sc_us=14424292772978069001&tn=SE_aiduxueshu_c1gjeupa&ie=utf-8.
- [2] FILS C, NAINAR N K, PIGNATARO C. The segment routing architecture[EB/OL]. (2015-12-10) [2016-12-11]. http://xueshu.baidu.com/s?wd=paperuri%3A%28f7bfc9a91907a39c0b44e9526a386f0%29&filter=sc_long_sign&sc_ks_para=q%3DThe%20Segment%20Routing%20Architecture&sc_us=7997732307306009577&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8.
- [3] BHATIA R, FANG H, KODIALAM M, et al. Optimized network traffic engineering using segment routing [C]//IEEE INFOCOM,

April 26-May 1, 2015, Hong Kong, China. New Jersey: IEEE Press, 2015: 657-665.

- [4] CARPA R, GLUCK O, LEFEVRE L. Segment routing based traffic engineering forenergy efficient backbone networks [C]// IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Dec 14-17, 2014, New Delhi, India. New Jersey: IEEE Press, 2014: 1-6.
- [5] DAVOLI L, VELTRI L, VENTRE P L, et al. Traffic engineering with segment routing: SDN-based architectural design and open source implementation[C]//European Workshop Software Defined Networks(EWSDN), September 30-October 2, Bilbao, Spain. New Jersey: IEEE Press, 2015: 111-112.
- [6] LAZZERI F, BRUNO G, NIJHOF J, et al. Efficient label encoding in segment-routing enabled optical networks [C]// International Conference of Optical Network Design and Modeling (ONDM), May 11-14, 2015, Pisa, Italy. New Jersey: IEEE Press, 2015: 34-38.
- [7] SGAMBELLURI A, GIORGETTI A, CUGINI F, et al. First demonstration of SDN-based segment routing in multi-layer networks [C]//2015 Optical Fiber Communications Conference and Exhibition (OFC), March 22-26, 2015, Los Angeles, CA, USA. New Jersey: IEEE Press, 2015: 1-3.
- [8] BIDKAR S, GUMASTE A, SOMANI A. A scalable framework for segment routing in service provider networks: the omnipresent Ethernet approach[C]//2014 IEEE 15th International Conference on High Performance Switching and Routing (HPSR), July 1-4, 2014, Vancouver, BC, Canada. New Jersey: IEEE Press, 2014: 76-83.
- [9] KOHLER E. The click modular router[M]. Boston: MIT, 2001.
- [10] TANTSURA J, CRABBE E, SH R, et al. Segment routing with MPLS data plane[EB/OL]. [2016-12-11]. http://xueshu.baidu.com/s?wd=draft%3A+Segment+Routing+with+MPLS+data+plane&tn=SE_baiduxueshu_c1gjeupa&cl=3&ie=utf-8&bs=The+Click+Modular+Router.&f=8&rsv_bp=1&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_spt=3&rsv_n=
- [11] FARREL A, VASSEUR J P, ASH J. A path computation element(PCE)-basedarchitecture[EB/OL]. (2006-12-20)[2016-12-11]. http://xueshu.baidu.com/s?wd=paperuri%3A%283d7de298762b1ae4e164b1d1334d95af%29&filter=sc_long_sign&sc_ks_para=q%3DA%20Path%20Computation%20Element%20%28PCE%29-based%20Architecture&sc_us=&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8.
- [12] MEDVED J, TANTSURA J, CRABBE E, et al. PCEP extensions for segment routing[EB/OL]. (2006-01-20) [2016-12-11]. https://www.researchgate.net/publication/238698550_A_Path_Computation_Element_PCE-based_Architecture.



[作者简介]



高晓敏(1990-),女,北京交通大学电子信息工程学院下一代互联网互连设备国家工程实验室硕士生,主要研究方向为未来网络体系架构与现网的兼容性。



刘雪晶(1976-),女,中国电子科技集团公司电子科学研究院高级工程师,主要从事雷达信号处理、通信、系统总体设计等方面的研究工作。



李佳伟(1992-),男,北京交通大学电子信息工程学院下一代互联网互联设备国家工程实验室博士生,主要研究方向为软件定义网络、流量工程和信息中心网络。



罗洪斌(1977-),男,博士,北京交通大学教授、博士生导师,北京交通大学下一代互联网互联设备国家工程实验室副主任,中国通信学会第四届青年工作委员会委员,北京通信学会第八届理事会理事。入选教育部新世纪优秀人才支持计划,并获得国家自然科学基金委优秀青年基金资助项目资助。主要从事未来互联网体系架构、理论与关键技术的研究工作,正主持和参与多项国家“973”计划、“863”计划、国家自然科学基金项目。近年来在 IEEE/ACM Transactions on Networking (TON)、IEEE Journal on Selected Areas in Communications (JSAC)、IEEE Network 等本领域顶级期刊与国际会议发表论文 50 多篇。



董平(1979-),男,博士,北京交通大学副教授、博士生导师,主要研究方向为移动互联网、下一代互联网、网络安全。