



专题:SDN & NFV

基于 SDN 的大型 IP 网络 BGP 路由优化方案

唐宏,朱华虹,曹维华,邹洁

(中国电信股份有限公司广州研究院,广东 广州 510630)

摘要:针对 IP 骨干网路由规模大、路径多、重叠度高而易绕转的问题,在分析传统 BGP 路由选路机制缺陷的基础上,采用 SDN 控制技术,提出了一种支持传统路由设备和 OpenFlow 设备的路由反射优化方法,并给出了具体实现算法及部署方案。典型应用场景的测试结果表明,国际访问时延平均缩短了 30%,验证了方法的有效性。

关键词:软件定义网络;OpenFlow;边界网关协议;OpenDaylight

中图分类号:TN915.41

文献标识码:A

doi: 10.11959/j.issn.1000-0801.2016112

Route optimization method for BGP based on SDN in large-scale IP network

TANG Hong, ZHU Huahong, CAO Weihua, ZOU Jie

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

Abstract: Aiming at rotation problem of route because of large-scale, multi-path, overlap in IP backbone networks, an optimization method for route reflection supported by both traditional routing devices and OpenFlow devices based on SDN controller technology was proposed. At first, the defects of traditional BGP routing mechanism was analyzed in detail. Then, a specific algorithm implementation and deployment scenarios were given. Test results of typical application scenarios demonstrate the validity of the proposed method. The average delay of international access reduces by 30%.

Key words: software defined networking, OpenFlow, border gateway protocol, OpenDaylight

1 引言

近年来,随着产业变革和新技术的发展,互联网迅速成为影响社会经济发展、改善人民生活品质的重要基石。互联网应用不断丰富,宽带用户数高速增长,尤其是“宽带中国”战略进一步促进了宽带网络能力的跃升,网络流量每年以 60% 的速度高速增长,给 IP 骨干网运营带来巨大的挑战。互联网路由条目数不断增加,但受限于传统分布式的路由算法以及匮乏的整体网络拓扑,大量的重叠路由

导致流量绕转、用户感知下降等问题。因此,随着骨干网规模的增加及流量的增长,如何优化路由选路策略成为一个重要而有价值的研究课题。

软件定义网络 (software defined networking, SDN) 技术^[1]为骨干网路由优化提供了有效手段,但骨干网设备数量多、改造成本高,全网设备的升级替换较为困难,需要考虑兼容现有设备能力的解决方案。本文在分析现有路由反射器选路机制缺陷的基础上,提出了一种基于源和目的地址的路由反射优化方法,并给出了骨干网部署方案。最后,在

典型应用场景下进行测试,验证了方法的有效性。

2 传统路由反射器路由选路机制

边界网关协议(border gateway protocol,BGP)^[2]是一种自治系统间的动态路由发现协议,它的基本功能是在自治系统间自动交换无环路的路由信息,通过交换带有自治系统号(AS)序列属性的路径可达信息,构造自治区域的拓扑图,从而消除路由环路并实施用户配置的路由策略。在大规模网络中,通过部署路由反射器来减少对等体连接关系,如图1所示。路由反射器收到多个指向同一IP地址前缀但下一跳不同的路由信息,路由反射器按照BGP路由选择机制来确定最优路由,也就是选择下一跳,然后转发给客户机和非客户机。选路的规则如下^[3]:

- (1)如果 next-hop 无法到达,则不考虑;
- (2)首选具有最大 weight 的路由(Cisco 特有);
- (3)如果路由具有相同 weight,则使用本地优先级最高的路由;
- (4)如果具有相同本地优先级,则首选来自本身路由器的 BGP 路由;
- (5)如果没有来自本身路由器上的 BGP 路由,则选择 AS 长度最短的路由;
- (6)如果所有的路由具有相同的 AS 长度,则选择具有最低 origin code 的路由;
- (7)如果 origin code 相同,则选择 MED 值最小的路由;
- (8)如果 MED 相同,则首选外部路由,而不是内部路由;
- (9)如果仍然相同,选择最近的 IGP 邻居的路由;
- (10)如果仍然相同,选路由器 ID 最小的路由;
- (11)如果仍然相同,选 cluster_list 最短的路由。

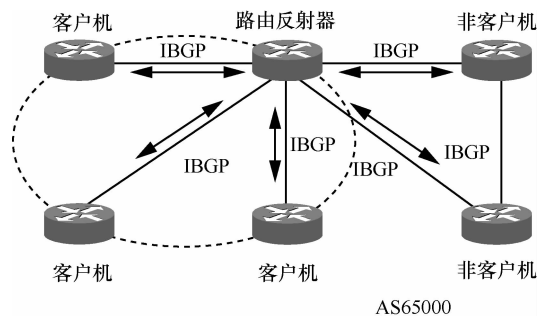


图1 路由反射示意

因此,从客户机角度看,经过路由反射器选择后的下一跳可能不是最佳选择——只是距离路由反射器最近的路由,而不是源和目的地址间距离最近的路由,导致次优路由的产生,如图2所示。

图2中,上海客户机和广州客户机1都有ICP的路由prefix 1,并将该路由向路由反射器进行通告。路由反射器根据BGP选路规则进行选路,当(1)~(8)的属性都相同,无法判断时,根据规则(9)选择距离自身IGP最近的上海客户机作为下一跳反射给所有客户机,导致广州客户机2接入的用户经上海访问prefix 1,造成路由绕转。

3 基于源和目的地址的路由反射方法

3.1 基于源和目的地址的路由反射方法

传统路由反射器的选路规则在(9)中是从路由反射器自身角度计算到下一跳的IGP最短距离,因此所有的客户机都将收到同样的路由,对于某些客户机来说,该路由并非最优路径。在很多情况下,可能导致流量的绕转,造成时延增大,用户感知下降。针对该问题,IETF也有相关草案,路由器支持add-path功能^[4],反射器反射多条路由,由客户

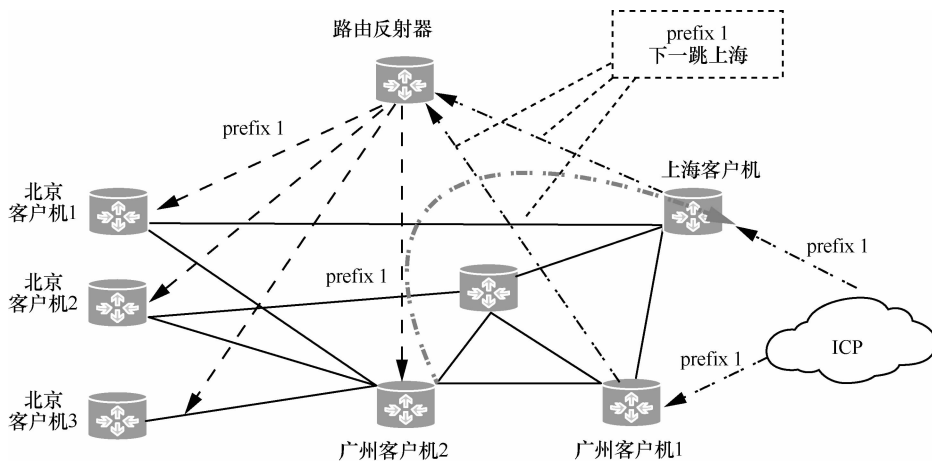


图2 路由绕转示意



机自行计算最佳路由。考虑目前互联网路由数量超过 50 万条^[5],且波动较大,因此,对反射器的性能要求较高,客户机需要接收的路由条目也较多,实际应用中实施困难。为此,对该条选路规则进行修改,路由反射器反射路由时,对不同的客户机计算客户机到下一跳的 IGP 最短距离,从而选择源和目的地址间的路径最短路由。基于 OpenFlow 技术^[6],在网络中部署 OpenFlow 控制器,对不同的设备下发不同的流表实现最优路径的选择。图 3 为 OpenFlow 1.3^[7]的流表结构,对于相同的路由前缀,针对不同的客户机计算其与各下一跳之间的 IGP 距离,选择距离最小的下一跳作为最优路由下发流表。

然而,在骨干网中,仍然存在大量传统路由设备,对 OpenFlow 的支持有限。为了在现网中实现该方法,依然需要考虑基于 BGP 对网络设备进行控制。控制器主要包含状态信息采集、数据中心、策略管理、网络建模、统一计算以及指令适配模块,具体如图 4 所示。

状态信息采集模块采集 IP 骨干网拓扑及网络基础设施和互联网业务路由、业务流量流向和业务质量等信息数据,同时将这一系列的大数据入库到数据中心;统一计算模块从 BGP 路由表中选出上述选路规则中(1)~(8)全相同的路由条目,同时计算设备间 IGP metric 矩阵,并对不同的客户机计算下一跳 IGP 最短的最优路径;经路由仿真模块验证策略后,最后进行统一下发。主要算法实现如下所示。

步骤 1 从当前 BGP 路由表查找选路规则(1)~(8)中

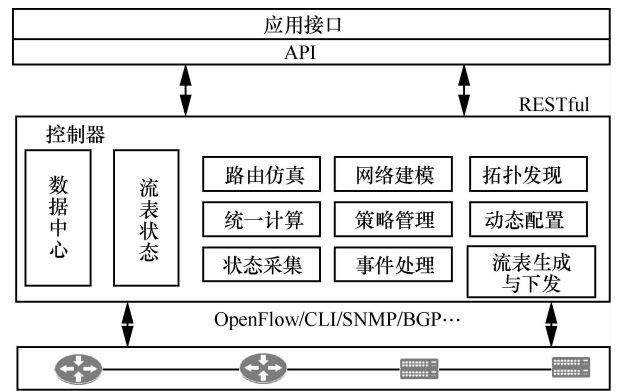


图 4 控制器主要功能

路由属性相同(如 local-preference、MED、AS path length)的路由,即经过路由反射器可能产生非优选的路由。

步骤 2 将步骤 1 中查到的路由数据复制到数据表 BGP prefix 中(先清空 BGP prefix 中已有数据,再写入)。

步骤 3 由于 BGP 路由更新频繁,为了便于比较更新的路由,数据表 prefixSnap 用于存放以前采用步骤 1 获取的路由。将 BGP prefix 表中的 prefix 与数据表 prefixSnap 中的 prefix 进行比较,如果相同,说明路由没有更新,不做处理;如果不同,则将 BGP prefix 表中的 prefix 增量更新到数据表 prefixSnap 中。

步骤 4 采集当前网络中 IGP 拓扑信息,生成设备间 IGP metric 矩阵。

步骤 5 获取当前控制器的客户机列表 igpDevMetric,为了便于比较更新的拓扑,peerIpMetricSnap 用于存放以前采集的客户机列表。将 igpDevMetric 与 peerIpMetricSnap 进

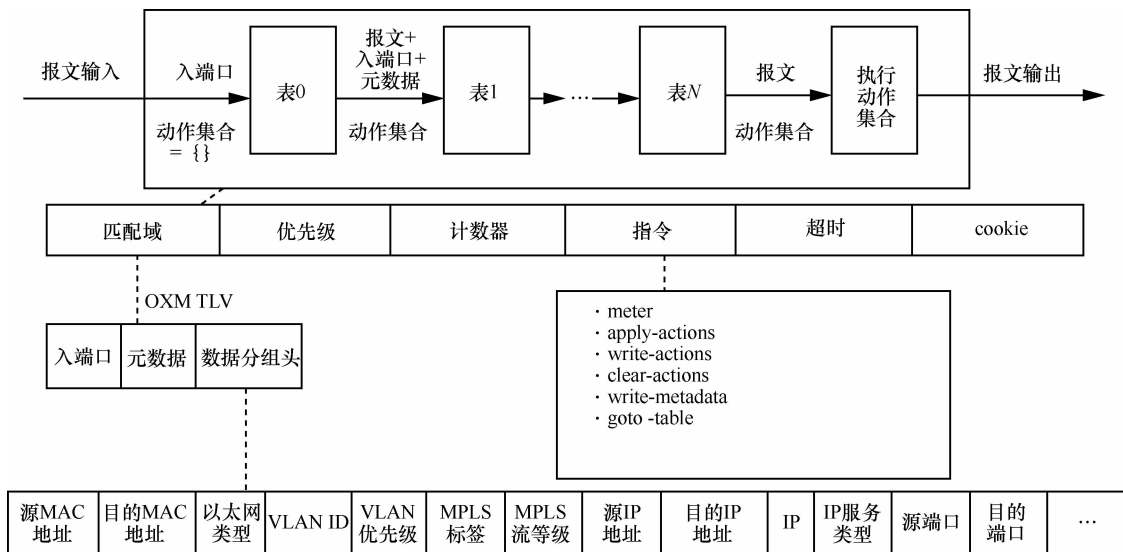


图 3 OpenFlow 1.3 流表结构

行比较,如果相同,说明拓扑没有更新,不做处理;如果不同,则将 igpDevMetric 增量更新到 peerIpMetricSnap 中。

步骤 6 获取 peerIpMetricSnap 中的客户机列表,针对每个客户机,分别以该客户机为根节点,基于 IGP metric 矩阵,采用 SPF 算法,对数据表 prefixSnap 中的相同 prefix 计算根节点到各下一跳的 metric,将 metric 最小的下一跳作为该 prefix 的优选路由。

步骤 7 无论是否需要配置下发,都将上述最优路由进行统计,并将其放入数据表 WorkStatus 中。

3.2 网络部署方案

软件定义网络技术为传统 IP 的优化提供了重要手段,然而,IP 网络全面实现软件自定义还有很长的过程。首先是技术的成熟度还不适合大规模现网运营的要求,如高可靠性、高安全性以及电信级 SLA 要求;其次,现网设备的技术支持能力也成为应用推广的关键。考虑到骨干网仍然以传统网络设备为主及新技术引入的可能风险,网络中的部署方案以增量叠加为主:在网络中部署 SDN 控制器,支持 OpenFlow 和 BGP,对传统设备采用 BGP 的更

新方式,对 OpenFlow 设备下发流表进行控制。其部署方案如图 5 所示。

控制器与路由反射器、相关的客户机建立 IBGP 邻居关系,并仅接收路由反射器反射的路由,同时获取 IGP metric 矩阵信息。结合 BGP 路由数据库及链路状态数据库,提取多路径路由,计算源到路由接收段之间的 SPF 计算,针对不同的客户机提取最优路径,经校验路由可达后分别对不同的客户机反射相关路由,或者下发流表,具体算法见第 3.1 节中的描述。客户机同时收到传统路由反射器及控制器的路由,根据 BGP 选路信息可以进一步得到最佳路径,放入路由表。该部署方案的好处在于,如果控制器发生故障或者计算错误,可以直接退出服务,原有 IP 地址仍然起效,不会对网络运营造成巨大影响。为了更好地对全网路由进行维护和监控,系统提供了相关的展示功能,如图 6 所示。

路由表中的数据可以按“路由前缀”(prefix 字段)、归属 AS(destAS 字段)、next-hop 进行查询及显示,方便人员进行操作。

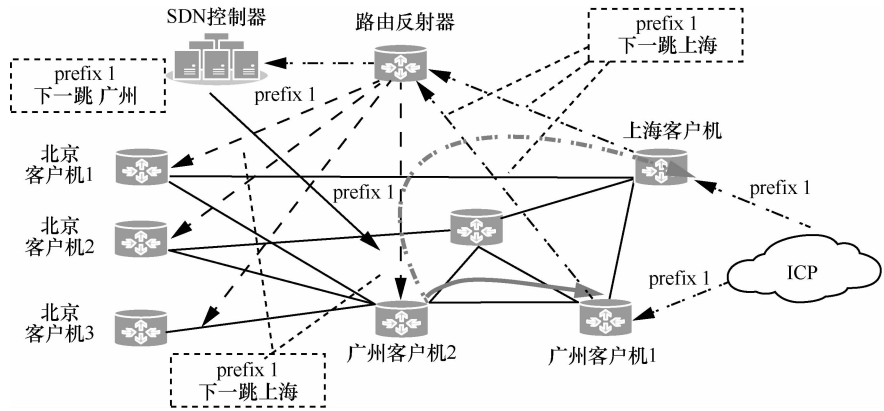


图 5 网络部署方案

多路径路由发布监控

当前状态 | 运行中

停止

即时更新

状态起始时间 | 2015-01-21 13:38:34

修改

路由发布范围 | 2222,2225

修改

路由接收统计 |

总路径数	60 021
接收路由叫数	40 011
多路径路由数	0

多路径路由由详细

路由前缀 |

归属 AS |

next-hop |

查询

图 6 多路径路由发布界面



4 测试结果分析

SDN 控制器主要是一个软件实体,目前主流的开源控制器主要有 NOX、POX、Ryu 等。本文提出的控制器主要基于 OpenDaylight 开源平台^[8]实现,采用 OSGI 框架和 Java 开发,南向支持 SNMP、BGP、OpenFlow 等协议,北向提供 RESTful 接口^[9],便于实现开放性。

测试的典型场景如图 5 所示,大量 ICP 会在多地接入骨干网,例如从上海、广州两地的 ASBR 均拥有 ICP 的路由,经骨干网路由反射器后只优选一条下一跳为上海节点的路由进行反射,导致广州接入段的用户需要绕转到上海节点访问 ICP,造成时延增加,用户体验下降。尤其在国际网络环境下,绕转的距离将大幅度增加,从而裂化访问质量。在骨干网(100 多台路由器,50 万多条互联网路由)中部署 SDN 控制器,采用本文所提方法采集全网拓扑信息和路由数据,针对广州客户机 2,计算出到达 ICP 的最优路径为广州客户机 1,于是对广州客户机 2 下发下一跳为广州客户机 1 的 ICP 路由,实现路由最优化,降低单向访问时延约 15 ms,如图 7 所示。进一步地,国际访问时延平均可降低 30%。测试结果表明,本文方法可实现路由端到端优化,提高互联网访问质量,验证了方法的可行性。

5 结束语

SDN 作为一种优化和简化网络操作的体系结构方式,

具有更大的灵活性和敏捷性,为基础互联网设施提供了智能化选择,成为当前网络领域最热门和最具发展前途的技术之一。本文针对 BGP 的缺陷导致 IP 骨干网路由不佳、流量绕转问题,提出了一种基于 SDN 的路由反射方法,并给出了网络规模部署方案。测试结果表明,新方法可减少流量绕转情况,国际互联网访问质量可大幅度提升,证明了方法的有效性。然而,SDN 作为一项系统工程,仍有大量的技术研发及实例化工作,后续将进一步完善控制器功能,实现网络的高质量运营。

参考文献:

- [1] 张朝昆,崔勇,唐嵩祯,等. 软件定义网络(SDN)研究进展[J]. 软件学报, 2015, 26(1): 62-81.
ZHANG C K, CUI Y, TANG H Y, et al. State-of-the-art survey on software-defined networking (SDN)[J]. Journal of Software, 2015, 26(1): 62-81.
- [2] 李道丰,王高才,王志伟,等. 标准模型下可证明安全的 BGP 路由属性保护机制[J]. 计算机学报, 2015, 38(4): 859-871.
LI D F, WANG G C, WANG Z W, et al. Provable secure mechanism for BGP path protection in the standard model[J]. Chinese Journal of Computers, 2015, 38(4): 859-871.
- [3] 徐建锋,朱华虹. 改进 BGP 实现大型复杂 IP 网络的负载均衡[J]. 电信科学, 2004, 20(10): 15-19.
XU J F, ZHU H H. Load sharing implementation in large complicated IP networks by improving BGP [J]. Telecommunications Science, 2004, 20(10): 15-19.

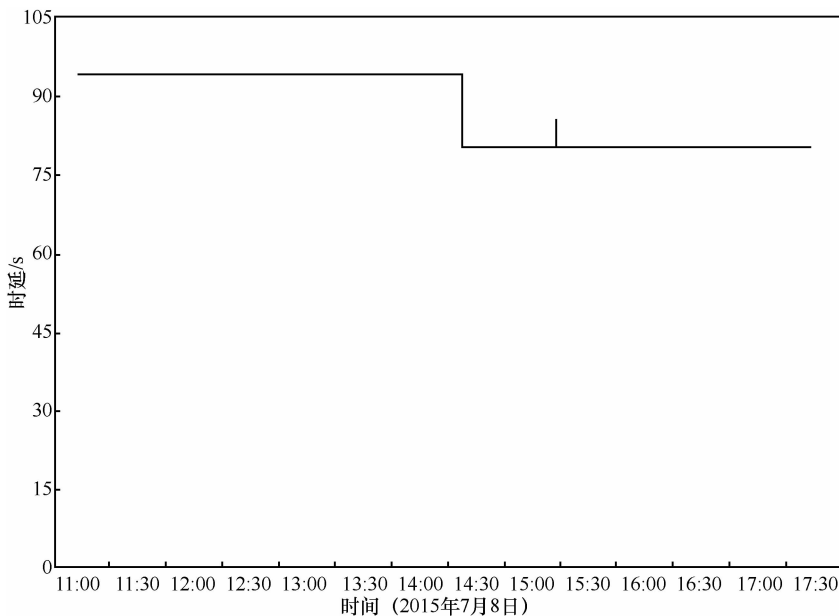


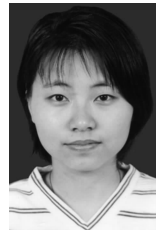
图7 优化前后访问时延结果对比

- [4] SCUDDER J, RETANA A, WALTON D, et al. Advertisement of multiple paths in BGP [EB/OL]. [2012-12-30]. http://xueshu.baidu.com/s?wd=Advertisement+of+Multiple+Paths+in+BGP&rsv_bp=0&tn=SE_baiduxueshu_c1gjeupa&rsv_spt=3&ie=utf-8&f=8&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_n=2.
- [5] 辛喆. 一种基于 SDN 的 IP 骨干网流量调度方案的研究与实现[D]. 北京: 北京邮电大学, 2015.
- XIN Z. Research and realization of IP backbone network traffic scheduling program based on OpenFlow [D]. Beijing: Beijing University of Posts and Telecommunications, 2015.
- [6] 左青云, 陈鸣, 赵广松, 等. 基于 OpenFlow 的 SDN 技术研究[J]. 软件学报, 2013, 24(5): 1078-1097.
- ZUO Q Y, CHEN M, ZHAO G S, et al. SDN technology research based on OpenFlow[J]. Journal of Software, 2013, 24(5): 1078-1097.
- [7] 徐秋伊. 基于 SDN 的路由映射算法的设计与实现 [D]. 北京: 北京邮电大学, 2015.
- XU Q Y. Design and realization of route mapping algorithm based on SDN [D]. Beijing: Beijing University of Posts and Telecommunications, 2015.
- [8] AHMED S, MARTINI B, GHARBAOUI M, et al. Orchestration algorithms for network-assisted virtual machine migrations using OpenDaylight controller[C]// 2015 2nd International Conference on Electrical Information and Communication Technology (EICT), December 10-12, 2015, Khulna, Bangladesh. New Jersey: IEEE Press, 2015.
- [9] WEI Z, LI L, MIN L, et al. REST API design patterns for SDN northbound API [C]// 2014 28th International Conference on Advanced Information Networking and Applications Workshops (WAINA), May 13-16, 2014, Victoria, BC, USA. New Jersey: IEEE Press, 2014: 358-365.

[作者简介]



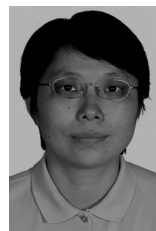
唐宏(1974-),男,中国电信股份有限公司广州研究院高级工程师,主要研究方向为 IP 网络技术、SDN 技、移动互联网技术。



朱华虹(1978-),女,中国电信股份有限公司广州研究院高级工程师,主要研究方向为 IP 路由技术、SDN 技术。



曹维华(1976-),女,中国电信股份有限公司广州研究院高级工程师,主要研究方向为 IP 技术、移动互联网技术。



邹洁(1975-),女,中国电信股份有限公司广州研究院高级工程师,主要研究方向为 IP 技术、SDN 技术。