

一种基于 SDN 的多管理域路由机制

张 俊, 沈苏彬

(南京邮电大学 计算机学院, 江苏 南京 210003)

摘 要:随着软件定义联网 (software defined networking, SDN) 的不断发展与完善, 未来互联网将会由多个 SDN 网络互联, 各管理域通过其域内控制平面, 对其数据平面进行集中式管控。在单一管理域的应用场景下, SDN 能够方便地为应用提供转发路径, 但是当应用流需要跨越多个管理域时, 由于当前互联网域间路由协议 BGP 的分布式部署, 以及管理域间缺乏协商, 将难以提供跨域路径。为了解决该问题, 提出了一种基于 SDN 的多管理域路由机制 (SDN-based multi-domain routing mechanism, SDNMDR)。该机制考虑应用的需求, 并通过扩展标准控制器的功能模块, 以及相邻管理域控制器间进行交互与协商, 采用 OpenFlow 协议控制 SDN 交换机, 为跨多管理域的应用提供端到端路径, 并且在各管理域网络资源允许的情况下, 能够为应用提供服务质量保证。通过 Mininet 和 Ryu 搭建实验环境, 证明了 SDNMDR 的可行性和有效性。

关键词:软件定义联网; OpenFlow; 多管理域; 路由

中图分类号: TP393

文献标识码: A

文章编号: 1673-629X(2018)08-0086-05

doi:10.3969/j.issn.1673-629X.2018.08.018

A SDN-based Multi-domain Routing Mechanism

ZHANG Jun, SHEN Su-bin

(School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: With the continuous development and improvement of software defined networking (SDN), the Internet will be interconnected by multiple SDN network in the future, managing their data plane through their control plane respectively. Under a single administrative domain of application scenarios, SDN can easily provide the forwarding path for application. But when the application flow across multiple administrative domains, because of distributed deployment of BGP, the current Internet inter-domain routing protocol, as well as the lack of consultation between management domain, it will be difficult to provide the cross-domain path. For this, we propose a SDN-based multi-domain routing mechanism (SDNMDR). With consideration of the requirements for applications, through the extension of the function module of the controller and the interaction and negotiation between intra-domain controller, it uses the OpenFlow to control SDN switch, which can provide end-to-end paths for applications across multiple management domains, and service quality assurance for applications where the network resources of each management domain allow. A prototype and testbed established with the aid of Mininet and Ryu have been implemented to prove feasibility and effectiveness of the SDNMDR.

Key words: software defined networking; OpenFlow; multiple management domain; routing

1 概 述

互联网的蓬勃发展, 驱动着网络应用的不断创新与丰富, 从高清视频会议到健康监测, 以及一些远程控制系统, 如森林防火系统等, 对于网络延迟、带宽和可用性等方面都有着严格的要求。并且随着互联网规模的不断扩大, 在多数情况下, 应用流都需要进行跨域传输, 这将会给 IP 骨干网带来艰巨的挑战。然而, 目前针对多管理域场景, 域间路由信息交互的协议大多采

用 BGP^[1], 该协议的分布式部署, 虽然满足了可扩展性和可靠性, 但是由于缺乏集中控制, 使得策略表达及配置相对比较困难。同时, 由于在选路过程中缺乏域间协商, 各管理域域内的局部最优路由算法无法保证全局最优, 使得应用流的 QoS 需求在目前这种“尽力而为”的网络传输中难以得到保障, 从而造成用户的 QoE 下降。

为了改善网络性能, 以及保障应用的 QoS 约束与

收稿日期: 2017-09-16

修回日期: 2018-01-15

网络出版时间: 2018-04-28

基金项目: 国家自然科学基金 (61502246); 未来网络前瞻性研究项目 (BY2013095-1-08); 南京邮电大学科研启动基金项目 (NY215019)

作者简介: 张 俊 (1992-), 男, 硕士研究生, 研究方向为软件定义联网、未来网络; 沈苏彬, 博导, 教授, CCF 高级会员 (E200005482S), 研究方向为计算网络、下一代电信网及网络安全。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20180427.1640.054.html>

终端用户的 QoE,学术界和工业界提出了一些解决方法。对于单个管理域,有 IntServ^[2]、DiffServ^[3]、MPLS^[4]等;对于多个管理域,有 BandwidthBrokers^[5]等。互联网中对于服务质量保证的问题,根据现有网络状态进行约束路径计算是一个关键组成部分,然而在当今互联网分布式体系中却难以实现。并且,现有网络协议的多样化,以及每个路由器为了保证 E2E 流量工程和应用业务 QoS 所需承载的计算量,使得当前网络中的路由器工作过于繁忙,不利于未来网络的持续发展与创新。

针对传统网络所呈现的相关问题,SDN(software defined networking)^[6-7]作为一种新型网架构与技术应运而生。其突出特点包括:

(1)转控分离:解耦传统网络设备的控制平面与转发平面,转发平面只具备数据流量转发的能力,控制平面通过南向控制协议(如 OpenFlow、NetConf、OVS-DB 等)控制转发平面的流量行为。

(2)集中管控:控制器拥有全局网络信息,如拓扑和网络资源等。网络管理员能够根据这些信息完成资源的合理分配与调度,解决负载均衡、网络资源分配不均等问题,简化网络管理。

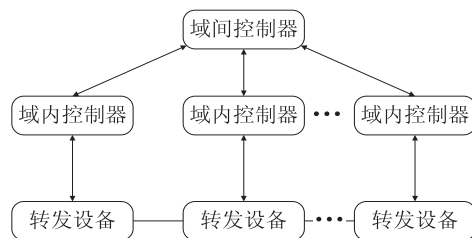
(3)网络可编程:通过控制器北向接口(如 REST API),应用平面告知控制平面如何进行网络资源管理才能更好地满足应用约束。并且,因特网服务提供商能够调用开放的北向接口,实现网络服务的定制,加快业务的动态部署。

OpenFlow^[8-9]作为 SDN 网络架构的一种实现方式,由斯坦福大学提出,其规范由 ONF(开放网络基金会)制定。其设计理念为:无需设计新的硬件,只对现有硬件更新其软件。这在很大程度上降低了部署网络应用所需的成本。在传统网络中,二层交换机主要采用 MAC 地址和 VLAN 标签对数据包进行处理与转发,而在 SDN 网络中,OpenFlow 作为构建网络的一种标准南向协议与规范,通过解析数据包或帧的包头域,将其 MAC 地址、VLAN 标签、IP 地址、TCP/UDP 端口号等特征作为“Flow”进行处理,通过在交换机中添加流表进行匹配,就能够灵活便捷地决定应用流的转发路径。

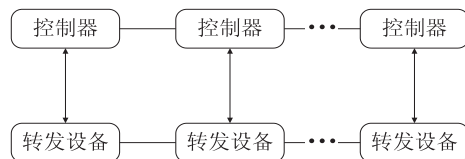
针对大规模多管理域(SDN 域)应用场景,业界提出了一些解决方案,主要是对控制平面进行了扩展,大致分为两类:分层分布式和完全分布式。

对于分层分布式控制平面(如图 1(a)所示),其域内控制器管控各个管理域域内网络,包括网络抽象、链路发现等,负责域内路由,并且向域间控制器上传特定的拓扑信息和主机信息。域间控制器通过获取来自域内控制器的信息,计算域间路由。例如, OXP^[10]针对

异构控制器间无法直接交互和目前接口协议效率较低的问题,设计了一种高效的、支持多种模式的东西向协议。 OXP 采用分层分布式的控制平面来解决 SDN 网络中单一控制平面遇到瓶颈的问题,域间控制器负责域间路由,域内控制器只负责域内路由。当发现应用流的目的 IP 地址不属于当前管理域时,域内控制器将数据包包头转发至域间控制器。域间控制器基于全局网络信息计算跨域路径,最后由各域内控制器完成流表的下发。



(a) 分层分布式



(b) 完全分布式

图 1 控制平面架构

对于完全分布式控制平面(如图 1(b)所示),每个管理域的控制器同时负责其域内路由,以及到下一管理域的域间路由,各管理域控制器之间进行信息的交互与协商,从而进行路由的决策。例如, HyperFlow^[11]采用 WheelFS 实现分布式的发布/订阅系统,该系统保证了发布事件的存储是永久的,并且保证事件的触发顺序与源控制器一致,在网络划分时具备良好的可伸缩性。但是由于所有事件的处理都需要全局信息,造成控制器间交互的信息量巨大,此外,还会造成网络状态不一致等问题。因此,这种模式只能处理一些发生率较低的事件,而不适用于大规模多管理域网络。 WE-Bridge^[12-13],虽然可实现异构控制器之间的协同工作,但是当网络视图发生改变时,由于控制器间采用全连接的方式,需要在所有同类节点中同步数据,导致消耗大量带宽,效率低下。 SDNi 虽然作为一种控制器间信息交互的协议,但是草案中并没有对网络如何存储这些信息以及如何在多个域之间共享这些信息给出具体的描述。 DISCO^[14]采用基于消息导向的通信总线,提供一个分布式控制平面。 DISCO 将整个网络分成多个域,用多个控制器进行管理。每个控制器管理一个域,并提供唯一轻量级的可管理的控制通道,每个控制器上的代理动态嵌入通道来获取其他域的信息。代理间互相共享全网信息,因此可以为应

用提供端到端的网络服务。其缺点在于全局事件的处理需要一致的全局信息,控制器之间交互的信息过多。

针对现有域间路由技术存在的相关问题及不足,文中提出一种基于 SDN 的多管理域路由机制(SDN-MDR),以及相关的实现和实验方案,并通过实验进行验证。

2 基于 SDN 的多管理域路由机制

虽然现有 SDN 域间路由技术在一定程度上解决了网络扩展性的问题,但是大多数方案都只是针对单一网络提供商实体而言的,并且域间交互的信息量过大,将会给网络负载带来巨大的压力。此外,当多个管理域由不同的网络提供商实体进行管控时,出于安全性考虑,大家都不愿公开其内部网络信息。因此,上文所列举的现有 SDN 域间路由技术将不再适用。

针对 BGP 分布式部署存在的诸多问题,以及现有 SDN 域间路由技术的不足,结合 SDN 网络可编程带来的优势,设计了 SDNMDR 机制。该机制采用完全分布式控制平面,主要通过控制器上的三个模块来实现,如图 2 所示。

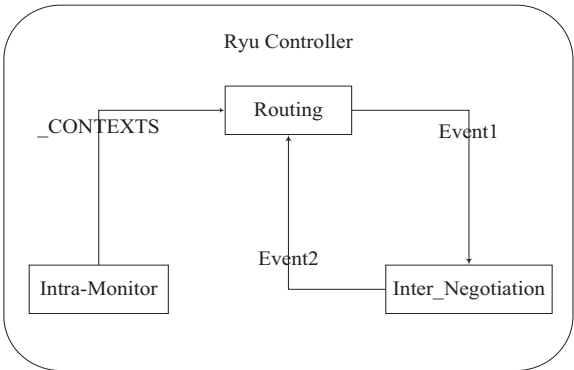


图 2 控制器功能模块

(1)域内监控模块(Intra_Monitor)。该模块根据 SDN 域内集中管控的思想,采用 LLDP 协议获取域内交换机信息,以及网络的实时情况,包括网络资源和网络流量信息等,从而分别为域间交互模块和路由模块提供协商和选路依据。

(2)域间交互模块(Inter_Negotiation)。对于网络应用流的传输,可达性是最基本的要求。首先,该模块参照 BGP^[15] 协议交互网络可达性信息,在目的网络可达的情况下,为了满足应用流的端到端需求约束,相邻管理域再通过该模块进行应用需求的请求与协商。

(3)路由模块(Routing)。主要完成两项工作:第一,对于跨域的应用流,结合网络可达性信息和域内网络资源信息,进行域内路径以及到相邻管理域的域间路径的计算;第二,根据计算结果,通知域间交互模块与相邻管理域进行协商,协商完成后再由该模块完成流表的下发。

目前业界大多数开源控制器,如 Ryu、OpenDaylight、Floodlight、ONOS 等,都已经实现域内监控和域内路由。由于 SDN 的初衷在于单个管理域内的集中控制,所以对于 SDN 域间如何交互与协商的问题,即东西向接口问题,仍是当前将 SDN 应用于大规模网络中亟需解决的一个热点问题。以下为文中提出的一种解决方案。

实现端到端服务质量约束路由的基础是多管理域间的互联互通,在 SDN 网络中,为了构建跨域的流表,相邻管理域控制器间需要进行通信,交互网络层可达性信息。文中控制器之间的交互参照了 BGP,在域间交互模块中实现,域间控制器的通信流程如图 3 中①~④所示,具体过程如下:

(1)使控制器具备 BGP speaker 的能力,每个 BGP speaker 包含状态机逻辑,并且当控制器启动时,BGP 功能体触发连接事件。每个管理域的 BGP 信息由网络管理员进行配置,控制器与邻居控制器建立 TCP 连接。

(2)BGP 使用 TCP 作为其传输层协议,当 TCP 连接建立后,BGP speaker 将会互相发送 OPEN 消息,并且状态将会变成 OPEN。

(3)BGP speaker 在 OPEN 状态下协商会话能力,控制器通过 OpenFlow 南向协议获取其管理域网络能力信息。

(4)成为 BGP Peers 后,控制器进入 ESTABLISHED 状态,双方在这个状态下互换 BGP UPDATE 消息,如 NLRI 和带宽等。

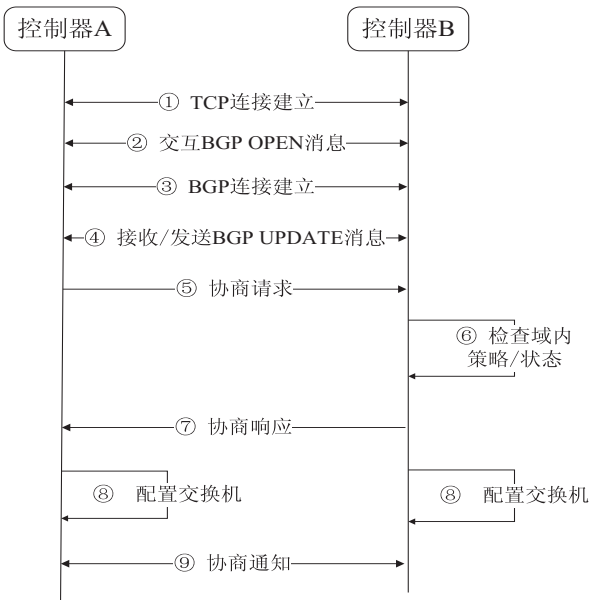


图 3 控制器建立连接及协商过程

为了给用户带来良好的网络体验,应用流在跨域传输的过程中,需要加以服务质量约束。由于各管理域只能决定其域内路由,彼此间缺乏协商机制,使得

业务的 QoS 约束难以得到保障。针对该问题,文中设计了一种域间协商协议,通过协商来实现应用流的端到端 QoS 约束路由。协商协议交互过程如图 3 中⑤~⑦所示。域间协商协议的消息如表 1 所示。

表 1 域间协商消息

消息类型	语义
协商请求	appId, reqId, srcIp, srcPort, dstIp, dstPort, protocol, start, duration, constraint (bandwidth, packetLoss, packetDelay, jitter)
协商响应	appId, reqId, ACCEPT(COST)/REJECT
协商通知	appId, reqId, CONFIRM/CANCEL

域间协商协议具体过程如下:

(1)控制器 A 发现应用流的目的地址是在域外,基于 BGP 信息,选择最佳下一域 B,从本地数据存储中检索边界信息,并且发送一个包含端到端应用约束的协商请求给控制器 B。

(2)控制器 B 收到这样的消息,与域内策略作比较,如果符合域内策略且有能够满足应用需求的配置路径,则更新这些路径以满足额外需求,否则将通过路由模块计算新的路径。此外,还会检查域间链路的统计,如果有可用的资源,然后会回复控制器 A 一个包含相应的代价 ACCEPT 协商响应,否则回复 REJECT。

(3)如果控制器 A 收到一个 ACCEPT,则表示协商成功。

(4)如果目的 IP 不在 B,则重复以上步骤,采用递归的方式。

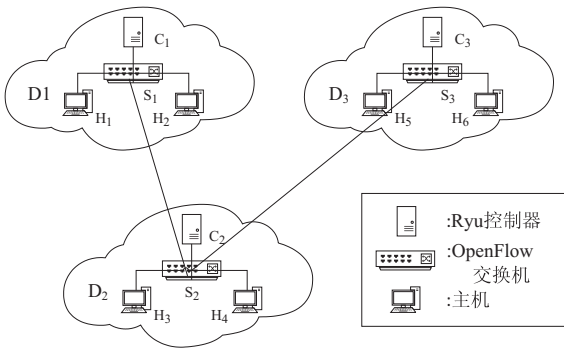
根据协商结果,各管理域中的控制器根据路由模块中计算好的路径,采用 OpenFlow 协议对沿路的交换机通过下发流表进行配置。

OpenFlow 流表的匹配域能够匹配数据包包头的二层至四层地址信息。首先,分别对源主机和目的主机所连接的交换机配置,通过控制器中存储的 MAC-PORT 映射表,采用匹配源目的 MAC 地址的方式构建流表;对非源目主机所在管理域的边界交换机配置,采用匹配目的 IP 地址的方式构建流表,使得经过该交换机的应用流能够转发至协商好的下一管理域的入口交换机,如图 3 中⑧~⑨所示。此外,各管理域还需周期性检查域间链路状态以及相关 QoS 参数,确保能满足应用流的服务质量约束和管理域间所约定的 SLA 参数,当不满足时,重新进行协商与配置。

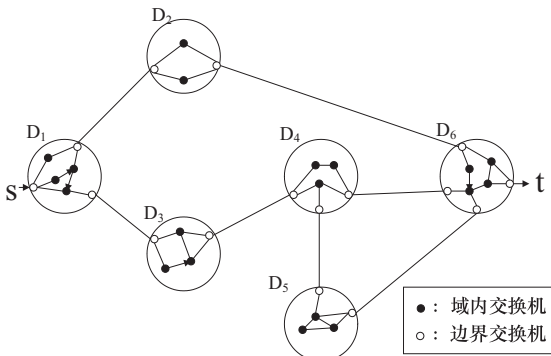
3 多域路由机制的实验和测试

为了满足应用流的跨域传输,首先需要实现的是多管理域的互联互通。实验选取 Ryu 开源控制器,通过 Quagga 使其具备 BGP speaker 的能力,采用 Mininet 仿真实验环境。实验拓扑如图 4(a)所示,各管理域控

制器采用 BGP 消息进行交互,获取 NLRI 后,通过 OpenFlow 配置交换机,执行 pingall 命令,H₁~H₆主机间能够互相 ping 通。通过测试,验证了该多 SDN 域实验平台的可行性。



(a) 3管理域仿真拓扑



(b) 6管理域抽象拓扑

图 4 实验拓扑

当源管理域控制器收到多条关于目的地址前缀的路由通告时(即在源节点与目的节点之间存在多条物理链路),模拟抽象网络拓扑如图 4(b)所示。在该场景下,应用流需要从 D₁ 中的主机 s(172. 16. 1. 1)传输至 D₆ 中的主机 t(172. 16. 6. 1),此时存在多条域间链路(①D₁-D₂-D₆;②D₁-D₃-D₄-D₆;③D₁-D₃-D₄-D₅-D₆)。假设该应用流的总时延约束为 100 ms,为了保障端到端约束路由,控制器间需要进行协商,并且根据协商的结果下发流表,配置各自域内交换机和边界交换机。

假设域间链路时延均为 10 ms,代价均为 10。各管理域域内聚合链路(边界路由器之间的链路)的时延及代价如图 5 所示。

首先根据 BGP 交互的网络层可达性信息以及从域内监控模块获取的域内拓扑和资源信息,D₁中控制器结合应用流服务质量约束与路由模块中的路由算法(文中采用最小代价算法),选择可以到达目的 IP 地址的下一管理域,通过交互模块发送协商请求,相邻管理域根据请求消息返回请求结果和所需代价。测试结果如下:

(1)D₁选择时延 20 ms,D₁-D₂时延为 10 ms,D₂选

择时延 20 ms, D_2-D_6 时延 10 ms, D_6 选择时延 20 ms。共 80 ms, 满足约束。递归返回代价 70。

D_1	10 ms(10) , 20 ms(5)
	10 ms(10) , 15 ms(5)
D_2	10 ms(35) , 20 ms(30)
D_3	15 ms(15) , 20 ms(10)
D_4	10 ms(15) , 15 ms(8)
	10 ms(20) , 30 ms(5)
D_5	20 ms(20) , 30 ms(10)
	10 ms(20) , 20 ms(15)
	5 ms(20) , 10 ms(15)
D_6	10 ms(20) , 20 ms(15)

图 5 域内链路(时延(代价))

(2) D_1 选择时延 15 ms, D_1-D_3 时延为 10 ms, D_3 选择时延 20 ms, D_3-D_4 时延 10 ms, D_4 选择时延 30 ms, D_4-D_5 时延 10 ms, D_5 选择时延 30 ms 和 20 ms 均不满足约束。

(3) D_1 选择时延 15 ms, D_1-D_3 时延为 10 ms, D_3 选择时延 20 ms, D_3-D_4 时延 10 ms, D_4 选择时延 15 ms, D_4-D_6 时延 10 ms, D_6 时延 10 ms, 共 90 ms, 满足约束。递归返回代价 68。

(4) D_1 中控制器的交互模块比较 1 和 2 中的代价, 选择 $D_1-D_3-D_4-D_6$ 这条链路, 并且发送确认消息。

(5) 各管理域根据协商结果, 通知路由模块下发流表配置交换机。

配置完成后, 通过 `ovs-ofctl dump-flows` 命令查看 D_1 与 D_3 相连的边界交换机流表, 可以看到包含“ip, nw_dst=172.16.6.1, actions=output:4”字段信息的流表, 查看 D_6 中与主机 t 相连的交换机流表, 可以看到包含“in_port=3, dl_dst=00:00:00:00:00:02, actions=output:3”字段信息的流表。

实验结果表明, SDNMDR 能够实现多管理域间的互联互通, 并且当源目的主机间存在多条可选域间路径时, 以及各管理域网络资源允许的情况下, 服务提供商能够根据管理域间的协商, 综合考虑应用需求与传输过程中所需代价, 选择满足约束的端到端路由。

4 结束语

通过扩展标准 Ryu 开源控制器的功能模块, 结合 Mininet 网络仿真工具, 实现了一种基于 SDN 的多管理域路由机制。通过模拟与测试, 验证了该机制能够在满足应用需求, 为用户提供良好体验的前提下, 降低服务提供商的所需成本。并且该方案采用完全分布式控制平面, 可缩放性较强, 能够适应当前互联网快速发展的迫切需求。

参考文献:

[1] GRIFFIN T G, WILFONG G. An analysis of BGP convergence properties[J]. ACM SIGCOMM Computer Communication Review, 1999, 29(4): 277-288.

[2] BARZILAI T P, KANDLUR D D, MEHRA A, et al. Design and implementation of an RSVP-based quality of service architecture for an integrated services Internet[J]. IEEE Journal on Selected Areas in Communications, 2006, 16(3): 397-413.

[3] MYKONIATI E, CHARALAMPOUS C, GEORGATSOS P, et al. Admission control for providing QoS in DiffServ IP networks; the TEQUILA approach[J]. IEEE Communications Magazine, 2003, 41(1): 38-44.

[4] XIAO Xipeng, HANNAN A, BAILEY B, et al. Traffic engineering with MPLS in the Internet[J]. IEEE Network, 2002, 14(2): 28-33.

[5] BOURAS C, STAMOS K. An efficient architecture for bandwidth brokers in diffserv networks[J]. International Journal of Network Management, 2010, 18(1): 27-46.

[6] 沈苏彬. 软件定义联网的建模与分析[J]. 南京邮电大学学报: 自然科学版, 2014, 34(3): 1-9.

[7] 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.

[8] MCKEOWN N, ANDERSON T, BALAKRISHNAN H, et al. OpenFlow: enabling innovation in campus networks[J]. ACM SIGCOMM Computer Communication Review, 2008, 38(2): 69-74.

[9] 左青云, 陈 鸣, 赵广松. 基于 OpenFlow 的 SDN 技术研究[J]. 软件学报, 2013, 24(5): 1078-1079.

[10] 杨 帆, 李 呈, 黄 韬. OXP: 一种面向 SDN 移动自组网的东西向协议[J]. 电信工程技术与标准化, 2016, 29(9): 32-37.

[11] TOOTOONCHIAN A, GANJALI Y. HyperFlow: a distributed control plane for OpenFlow[C]//Proceedings of the 2010 internet network management conference on research on enterprise networking. [s. l.]: [s. n.], 2010.

[12] 林萍萍. 软件定义网的东西向对等互联机制研究[D]. 北京: 清华大学, 2014.

[13] LIN Pingping, BI Jun, CHEN Ze, et al. WE-bridge: west-east bridge for SDN inter-domain network peering[C]//IEEE conference on computer communications workshops. Toronto, ON, Canada: IEEE, 2014: 111-112.

[14] PHEMIUS K, BOUET M, LEGUAY J. DISCO: distributed multi-domain SDN controllers[C]//Proceedings of the IEEE network operations and management symposium. Kraków, Poland: IEEE, 2014: 1-4.

[15] QUOITIN B, PELSSER C, SWINNEN L, et al. Interdomain traffic engineering with BGP[J]. IEEE Communications Magazine, 2003, 41(5): 122-128.