



域间 SDN 互联技术 WE-Bridge 及其实验床的研究进展^{*}

毕 军^{1,2}

(1.清华大学信息网络科学与网络空间研究院 北京 100084;

2.清华信息科学与技术国家实验室 北京 100084)

摘 要:在讨论 SDN 管控域间互联发展的 3 个阶段的基础上,介绍了清华大学提出的一种协作式的域间 SDN 互联技术 WE-Bridge,然后介绍了基于 WE-Bridge 技术构建的跨洲际的域间 SDN 实验床。

关键词:软件定义网络;域间 SDN;WE-Bridge

doi: 10.3969/j.issn.1000-0801.2014.08.005

Research Progress of Inter-Domain SDN Inter-Networking Technology WE-Bridge and Its Testbed

Bi Jun^{1,2}

(1.Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China;

2.Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, China)

Abstract: Firstly three phases of the development of inter-domain SDN internetworking were discussed, then a cooperation based inter-domain SDN technology WE-Bridge was introduced. Finally, an inter-continental inter-domain SDN testbed based on WE-Bridge technology was introduced.

Key words: software defined networking, inter-domain SDN, WE-Bridge

1 引言

SDN 将复杂的网络控制与管理从传统采用分布式协议通信的网络设备中剥离出来,使数据平面网络设备负责高速转发,而网络的控制与管理由逻辑集中的控制平面基于全局网络视图与状态进行统一决策。近年来通过大量的研究和实践,学术界和工业界已经对 SDN 技术及其未来应用前景较高的评价。从学术界看,以 SIGCOMM、NSDI 等为代表的顶级学术会议里,近年研究 SDN 的论文数量呈显著增长趋势;从工业界看,以开放式网络基金会(ONF)、开源项目 OpenDaylight 等为代表的组织以及 Google、华为等众多国内外企业正在将软件定义网络 (software defined

networking,SDN)技术向实际应用场景大力推进。

但这种数控分离和集中控制机制也导致在大规模网络下的 SDN 控制平面出现可扩展性问题,特别是尝试将 SDN 技术扩展到全球范围时将更加严峻。近年 SDN 在校园网、企业网、数据中心以及私有广域网中的研究集中在单个自治域(AS)内采用逻辑集中的控制平面,而全球互联网已经有超过 40 000 个自治域,且根据亚太网络信息中心(APNIC)首席科学家 Geoff Huston 对核心网络路由信息的统计,AS 表现出超线性趋势。因此 SDN 在如何应用于分布式管控的互联网域间场景,以提高域间路由等 SDN 域间应用的可编程性仍有很大挑战。同时,SDN 控制平面的灵活性也为域间网络资源的最优调度提供了可能。因此,

^{*} 国家高技术研究发展计划(“863”计划)基金资助项目(No. 2013AA013505)

研究域间 SDN 的挑战与机遇并存。

本文中的“SDN 域”是指“SDN 管控域”,即运行者所能独立管控的一个部署了 SDN 技术的网络域。SDN 域可以是一个自治系统(AS),也可以是由多个自治系统组成的管控域,未来甚至可以是没有 AS 号的 SDN 部署域。本文所研究的域间 SDN 技术是指如何解决全球互联网规模的 SDN 管控域间的控制协作问题。如果这一问题无法得到解决,SDN 技术只能被用在小规模的实验环境或小规模的单自治域网络中,无法成为改变大规模互联网的一次重要变革。

本文首先讨论 SDN 管控域间互联发展的 3 个阶段,然后介绍清华大学提出的一种协作式的域间 SDN 互联技术 WE-Bridge,最后介绍在中美学术网络未来互联网工作组推动下基于 WE-Bridge 技术所构建的跨洲际的域间 SDN 实验床。

2 SDN 管控域间互联发展的 3 个阶段

域间 SDN 是一个较新的问题。SDN 的当前研究主要集中在域内地理小范围内集中管控的单个控制器,或者多控制器控制一个管控域内大量交换机的情况。如何提高 SDN 的可扩展性,控制大规模互联网范围内不同 SDN 管控域的互联与控制协作,仍然处在起步阶段。

笔者认为,随着 SDN 域的日益部署,SDN 管控域间的互联发展存在 3 个阶段。第 1 个阶段和第 2 个阶段都是过渡阶段,其目标都是如何在新的 SDN 管控域与传统的 BGP 网络域之间实现 IP 分组的互通。但根据是否需要替换现有设备(即部署的难易程度),又具体分为两个阶段:第 1 个阶段不需要替换现有网络设备,第 2 个阶段采用新的 SDN 设备(如 OpenFlow 交换机)替换 AS 内的 BGP 路由器等现有网络设备。第 3 个阶段是当已经存在很多部署了 SDN 的区域后,需要研究 SDN 管控域如何互联和控制协作,以充分发挥 SDN 方式在域间路由或控制上的优势。

BTSDN 技术^[1]是第一阶段的一个例子。BTSDN 的原理是,在当前一个 AS 里部署新的 SDN Area(区域),但不替换已经存在的 BGP 路由器。控制平面上,SDN 区域的控制器通过运行 iBGP(内部 BGP)的应用与位于本 AS 边界的 BGP 路由器交换路由信息,然后通过 eBGP 与其他运行 eBGP 的 AS 互通;数据平面上,采用 ARP proxy 和 MAC rewrite 等机制保证 IP 分组在 SDN 区域与 BGP 路由器之间的互通。从中可以看出,该方法部署的代价很小,不需要替换现有的 BGP 路由器,仅需在 AS 内部署新的 SDN 区

域,在控制器上实现控制平面的功能,并对数据平面完成所需要的控制。

第二阶段有很多代表性技术,包括 RouteFlow^[2]、SDN-IP^[3]、SDX^[4]等,这些方法的共性是实现了 SDN 管控域与 BGP 网络域的互联,而且采用新型 SDN 设备实现这种互联。根据不同的目标,这些方案有各自所面向的场景。RouteFlow 着重提供虚拟路由器服务,通过把商用路由器替换成可被远程编程控制的 OpenFlow 交换机,路由器的逻辑控制功能被移动到虚拟机上,每台 OpenFlow 交换机设备对应一台虚拟机;SDN-IP 着重基于 BGP 实现 SDN 管控域和传统 BGP 路由域的无缝对等互联(peering),并实现 SDN 管控域作为传统 BGP 路由域的穿越网络,以促进逐步部署 SDN 域来替换传统的 AS;SDX 着重基于 SDN 改造传统的互联中心(IXP)和路由服务器(route server),通过软件定义的互联中心实现多个 AS 之间的互联。

第三阶段的目标是实现 SDN 管控域之间的互联和控制协作。在 RFC 草案^[5]中提出,在数据中心间的联邦 SDN 控制器之间传输 IP 路由可达性时,仍然采用 BGP。在参考文献[6]中提出了一种 XSP(extensible session protocol),XSP 是一个位于传输层之上的高级会话层协议,被作为网络应用和网络服务之间的交互接口。XSP 着重设计一个高级的会话层协议本身,仅针对某项具体应用而并不解决域间 SDN 的基础互联机制。此外还有一些项目试图为上层网络应用提供 NSI(network service interface),以实现 SDN 多域间的虚拟资源共享,例如欧盟 FP7 项目资助的 FELIX^[7]等,但目前这方面的具体技术还未见发表,而且其主要目的是高层的域间资源共享。上述工作虽然各自有其应用场景和价值,但都不是一种普适的域间 SDN 互联机制。

第三阶段的域间 SDN 研究旨在提出解决 SDN 在全球互联网规模的互联和控制协作方面的一般性机制和方法,为新型域间路由等大规模互联网下 SDN 域间应用的可编程性提供关键机制和方法。这类研究在国际上还处在起步阶段,清华大学项目团队是在国际上较早进行此方向研究的,提出了一种新型的协作式域间 SDN 机制 WE-Bridge^[8],下面在第 3 节中予以介绍。

3 一种协作式的域间 SDN 互联技术 WE-Bridge

WE-Bridge 机制是清华大学在国家高技术研究发展计划(“863”计划)“未来网络体系结构和创新环境”中提出的



一种新型的 SDN 东西向互联机制。图 1 展示了该“863”项目所提出的 FINE[®]体系结构中,WE-Bridge 所处的地位。其中, FN 指转发节点 (forwarding node)。WE-Bridge 是 FINE 的网络操作系统层和虚拟平台层的扩展,基于 WE-Bridge 交互各管控域的虚拟视图信息,为各种域间应用(如新型的域间路由协议或路径计算)提供协作的接口,使 SDN 应用之间可以实现跨域协作。WE-Bridge 的 SDN 跨域协作的东西接口和协议,包括部署 SDN 的互联网自治域之间对等连接的机制,SDN 跨域机制所交互的信息(各管控域的路由策略、虚拟网络视图等)的优化分发机制、报文格式和协商过程。此基础上,进一步研究了新型细粒度域间路由等可编程的 SDN 域间应用,以验证所提出机制的有效性。

SDN 管控域间首先需要解决 SDN 跨域的视图共享和控制协作问题,从兼容不同管控域所采用的不同存储系统和不同数据结构出发,确定多管控域间进行协作需要规范协商的内容,制定通用的域间交互机制和相应的域间接口(即 SDN 跨域的东西向接口),主要包括如下几个方面。

(1)SDN 管控域网络视图的抽象

目前在 WE-Bridge 中已经考虑的 SDN 视图抽象包括:SDN 管控域的可达性 IP 地址信息;拓扑信息,包括节点(如交换机、服务器、主机、控制器、防火墙、负载均衡器等)、链接、链接属性、端口吞吐量、链路连接情况;网络服务功能,如 SLA (service level agreement, 服务等级协议)、GRE (generic routing encapsulation, 通用路由封装)和 SSL (secure sockets layer, 安全套接层);QoS (quality of service)

参数;网络状态,即流表 (flow table) 在每个交换机的条目信息、拓扑结构中的实时带宽利用率、在网络中所有数据流的路径等。

(2)SDN 管控域网络视图的自动学习

上述网络视图中的拓扑信息可以表示为实体(节点、虚拟节点、链路、虚拟链路)和实体属性的无向图。目前,SDN 控制器一般使用传统网络的 LLDP (链路层发现协议)来发现域内的拓扑结构。通常情况下,每个 SDN 域控制器指示每个连接的 OpenFlow 交换机向所有端口发送 LLDP 报文(LLDP 报文中携带源交换机标识、输出端口以及其他能力)。因此,一旦相邻交换机收到 LLDP 报文,将报文直接送至控制器,控制器将提取和分析 LLDP 报文的信息,建立设备的连接关系。为了获取更完善的网络视图信息,还增加了 SDN 数据平面版本信息(如 OpenFlow 版本号、每个节点所支持的流表数量、链路利用率和流表条目数量等)。此类信息可以通过计算某个时间片内 OpenFlow 交换机的所有流表上与某个端口相关的数据分组数量来获取,并通过扩展 LLDP 学习链路利用率。

(3)SDN 管控域网络视图的传输格式

SDN 跨域协商的网络视图传输格式描述语言需要满足以下需求:独立于供应商和应用程序,以便网络视图消息传输格式可以独立于不同 SDN 域的不同存储系统;允许自定义数据结构(选择元素易于扩展的语言),以便网络视图中的消息格式更灵活并容易扩展。目前满足此需求的语言主要有 JSON (JavaScript object notation)、XML (extensible

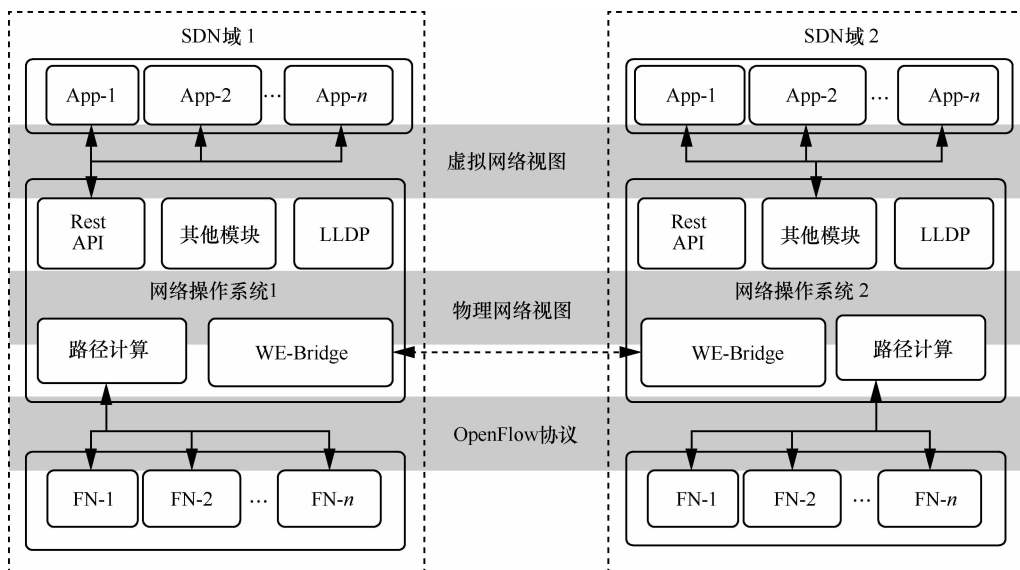


图 1 域间 SDN WE-Bridge 机制

markup language)、YANG (RFC6020)等,WE-Bridge 当前版本的实现采用 JSON 作为传输格式。

(4)采用 SDN 管控域的虚拟网络视图以支持管控域的机密性和可扩展性

网络视图通常指整个网络,出于对隐私的考虑,每个 SDN 管控域往往只愿意向对方透露其域的一部分信息,而非全局网络视图信息。在这种情况下,协议需要支持将物理网络抽象成虚拟网络。此处设计两种虚拟网络视图:将物理网络抽象成一个只包含网络边缘交换机的虚拟网络,虚拟网络的出口交换机与入口交换机的路由路径段可以有 SLA 等级路径的属性,如时延、可靠性、带宽、分组丢失率;将物理网络抽象成一个虚拟节点。网络抽象后,每个 SDN 管控域的控制平面应该存储物理网络与虚拟网络之间的映射表。

(5)高性能网络交换机制

所有 SDN 对等体彼此是等价的,并且构成非结构化网络连接。抽象所有 SDN 对等体拓扑到一个无向图 G 。每个对等体都可表示为一个顶点 V ,每两个 SDN 对等体之间的连接表示为一条边 E 。考虑到每个节点只有有限的硬件资源(如带宽、计算能力),只能处理有限的连接。在这里,用度数 D 表示最大连接数,实时连接数目记为度数 d 。在一个有对等方的 SDN 中,如果每个节点都连接了剩余的其他节点,那么虚拟拓扑将是全网(full-mesh)状拓扑结构,在拓扑中的连接数越多,数据传输就越稳定、越快速。基于“最大连接度数”原则,每个节点应该与对等方建立尽可能多的连接。WE-Bridge 设计了一个基于最大度数的可扩展网络虚拟视图分发机制。

4 基于 WE-Bridge 技术的跨洲际的域间 SDN 实验床

2012 年,中美学术网络成立未来互联网工作组,由美国 Internet2 CTO Stephen Wolff 和笔者等担任共同主席,其工作组 Charter 的主要内容就是开展域间 SDN 的实验床,进行域间 SDN 的创新应用试验。工作组研究认为,域间 SDN 实验床不能采用集中式的结构,采纳了清华大学所提出的 WE-Bridge 方案。2013 年 1 月,在美国夏威夷举行的 Internet2 和 APAN 联合大会上,工作组讨论确定了清华大学的技术方案。因为会议在夏威夷大学的 East West Center 召开,受此启发,将此机制命名为 WE-Bridge (west-east Bridge),希望通过此技术架起域间

SDN 东西向互通的桥梁。

在 WE-Bridge 设计和实现的基础上,在中美学术网络未来互联网工作组的推动下,中国教育和科研计算机网(CERNET)、美国 Internet2、中国科技网(CSTNET)、荷兰学术网络 SURFnet 等合作,建立了首个基于协作式的跨洲际域间 SDN 实验床。该实验床从 2013 年 7 月开始运行,目前使用的设备包括神州数码公司的 DCRS-7604 和 DCRS-5980、华三公司的 S5820、戴尔公司的 S4820、Pica8 公司的 Pronoto3920、HP 公司的硬件 OpenFlow 交换机以及 OVS 的软件交换机。在 CERNET、Internet2、CSTNET 和 SURFnet 的多个 SDN 实验域内安装了清华大学开发的跨控制器平台的 WE-Bridge 软件。该域间 SDN 实验床先后在 2013 年 9 月中国杭州举行的 CANS 中美网络技术会议、2013 年 11 月美国 Denver 举行的 SuperComputing 国际超级计算会议、2014 年 1 月印度尼西亚万隆举行的 APAN 亚太先进网络会议、2014 年 3 月美国硅谷举办的 ONS 全球开放式网络峰会等国际会议上成功演示。该实验床也被 IEEE INFOCOM2014 学术会议录取为演示,在 2014 年 4 月加拿大多伦多举行的 INFOCOM 会议上发表^[9]。

目前该实验床上运行的一项域间创新应用是通过基于 SDN 域间可编程的新型细粒度域间路由,对 WE-Bridge 机制进行有效性验证。目前 BGP 路由策略是基于目的 IP 地址前缀的。通过 SDN 跨域协作机制提供的东西向接口及其信息协商格式,使相邻的 SDN 自治域之间可以协商基于源和目的地址前缀、端口范围等细粒度字段的路由策略。这样相邻自治域之间可以建立更细粒度的策略表达,协商基于特定源和目的 IP 地址前缀的路由策略,实现比传统网络更灵活的多路径路由。如图 2 所示,其中图 2(a)中粗实线路径是目的驱动的单路径路由,图 2(b)中粗虚线和粗实线路由是源地址驱动的到相同目的前缀的多路径路由。

在该实验床上进行试验的其他域间创新应用包括:为中美基因组研究所需的大数据传输进行动态建立和拆除保证服务质量的 VIP 通道、一种域间真实源地址验证技术 ICS 等。

Internet2 CTO Stephen Wolf 在 Internet2 官网上发表了对此技术的评价^[10],认为目前 SDN 技术仅能用于单域环境,而此项开拓性工作展示了 SDN 技术可以扩展到全球科学合作的关键多域环境,对于美国部署 SDN 的校园网与 Internet2 国家基础设施互通也有重要意义。

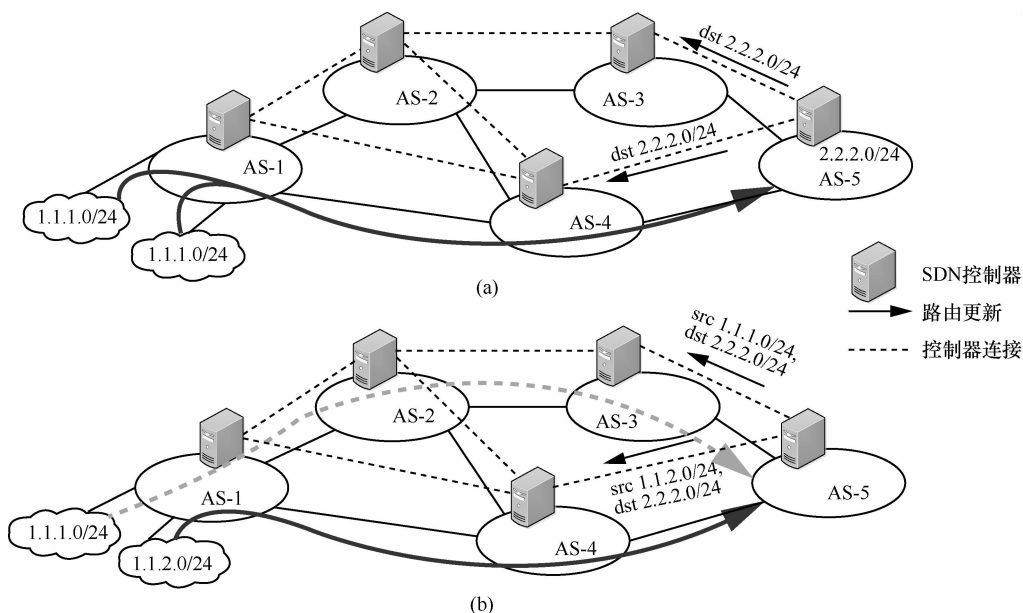


图 2 基于 SDN 的细粒度域间路由机制

5 结束语

近年来 SDN 已经获得了学术界和工业界的高度重视,并在集中管控的域内 SDN 技术和应用上取得了一定的进展。但如何将 SDN 技术扩展到域间,以支持新型的域间 SDN 应用,是一个机遇与挑战并存的研究课题。

本文提出了域间 SDN 发展的 3 个阶段,并介绍了各阶段的典型技术,指出在第三阶段即在 SDN 域间互联和控制协作以支持新型 SDN 应用上,国外尚没有可以借鉴的技术。本文介绍了清华大学依托国家“863”项目“未来网络体系结构和创新环境”提出了一种协作式的新型域间 SDN 互联技术 WE-Bridge,并基于 WE-Bridge 建立了首个跨洲际的域间 SDN 实验床,对新型细粒度域间路由等开展了试验。目前正在进一步完善 WE-Bridge 机制,并正在进行新的域间创新应用的研究与试验。

随着 SDN 域的逐渐部署,域间 SDN 互联将逐步成为研究的热点,希望本文对此领域的研究有一定参考价值。

致谢: 先后参与 WE-Bridge 设计和开发工作的科研人员包括清华大学信息网络科学与网络空间研究院网络体系结构和 IPv6 研究室博士生林萍萍、博士后王旻旻、博士生徐安民、硕士生陈泽等,胡虹雨博士也参与了域间 SDN 实验网的研究工作。感谢合作方美国 Internet2 CTO Stephen Wolff,中国科技网网络中心吴玉磊博士、鄂跃鹏博士,美国印第安纳大学的 John Hicks,

荷兰 SURFnet Ronald van der Pol 等在建设域间 SDN 实验网上的工作。

参考文献

- 1 Lin P P, Bi J, Hu H Y. BTSDN: BGP-based transition for the existing networks to SDN. Proceedings of the 6th IEEE International Conference on Ubiquitous and Future Networks (ICUFN14), Shanghai, China, 2014
- 2 Nascimento M R. Virtual routers as a service: the router flow approach leveraging software-defined networks. Proceedings of the 6th ACM International Conference on Future Internet Technologies (CFIT11), Seoul, Korea, 2011
- 3 Lin P P, Hart J, Kobayashi M, et al. Seamless interworking of SDN and IP. ACM Computer Communication Review, 2013, 43(4): 475~476
- 4 Gupta A, Vanbever L, Shahbaz M, et al. SDX: a software defined internet exchange. Proceedings of ACM SIGCOMM 2014, Chicago, USA, 2014
- 5 Balus F, Stiliadis D, Bitar N, et al. Federated SDN-Based Controllers for NVO3 (Network Virtualization Overlays). draft-sb-nvo3-sdn-federation-02, 2014
- 6 Kissel E, Swamy M. The eXtensible Session Protocol: A Protocol for Future Internet Architectures. Technical Report, 2012
- 7 FELIX Project. <http://www.ict-felix.eu/>
- 8 毕军. SDN 体系结构与未来网络体系结构创新环境. 电信科学, 2013, 29(8):7~15
- 9 Lin P P, Bi J, Chen Z, et al. WE-Bridge: west-east Bridge for

(下转第 46 页)



入、增加资源的调度弹性、缩短业务的部署流程及周期、增强核心网业务能力等。但制约核心网虚拟化进程的因素也较多,如标准化的推进、设备厂商研发的进程、运营商部署的意向等。目前只考虑设备成熟的因素,虚拟化的核心网预计在 2015 年将具备外场测试条件,2017 年左右具备商用部署的能力。

虚拟化核心网的部署是对核心网架构的革命性演进,核心网云化已经是可预期的发展趋势。

6 结束语

核心网虚拟化是目前业界关注的焦点问题之一。本文主要介绍了虚拟化技术,分析了核心网引入虚拟化技术的应用场景。同时,梳理了核心网虚拟化涉及的相关问题,并探讨了需要解决的关键技术问题。最后对核心网虚拟化的进程及时间点进行分析。

核心网虚拟化对传统电信运营商来说是变革性的网络演进,首先需要在技术及运行维护方面解决相关关键问题,确保虚拟化核心网的功能及性能,其演进需要一个较长的过程。但核心网虚拟化是业内核心网发展的趋势,核心网的虚拟化势不可挡,在提升核心网能力及满足市场业务需求方面具有强劲的优势。

参考文献

- 1 ETSI GS NFV 001 V1.1.1. Network Functions Virtualization, Use Cases, 2012
- 2 Infonetics Research. SDN and NFV Strategies: Global Service Provider Survey, 2013

(上接第 32 页)

SDN inter-domain network peering. Proceedings of the 33rd IEEE International Conference on Computer Communications (INFOCOM14), Demo, Toronto, Canada, 2014:111~112

10 <https://www.internet2.edu/blogs/detail/5020/>

[作者简介]



毕军,男,于清华大学计算机系获得博士学位。曾赴美国留学,任贝尔实验室博士后、研究员。2003 年由清华大学引进回国工作,现任清华大学信息网络科学与网络空间研究院网络体系结构和 IPv6 研究室主任、教授、博士生导师,国家高技术研究

- 3 ETSI GS NFV 002 V1.1.1. Network Functions Virtualization, Architecture Framework, 2012
- 4 ETSI GS NFV 003 V1.1.1. Network Functions Virtualization, Terminology of Main Concepts in NFV, 2012
- 5 ETSI GS NFV 004 V1.1.1. Network Functions Virtualization, Virtualization Requirement, 2012
- 6 ETSI GS NFV-PER 002 V1.1.1. Network Functions Virtualization, Poof of Concepts, Framework, 2012

[作者简介]



胡晓娟,女,博士,中国电信股份有限公司北京研究院高级工程师,主要研究方向包括 LTE 网络、IMS 网络、软交换网络及 SDN/NFV 技术等,主要致力于网络解决方案的研究。



赵慧玲,女,中国电信股份有限公司北京研究院总工程师、教授级高级工程师,中国电信股份有限公司云计算研究中心主任,国家科技部 SDN 项目责任专家,清华大学中国电信下一代互联网技术联合实验室主任,主要从事宽带网络和下一代网络的技术研究以及通信网络发展战略研究工作,主持了我国网络标准的研究和制订工作。享受国家政府津贴,发表文章近百篇,出版技术专著 12 部;曾获得国家技术发明奖二等奖 1 项,国家科学技术进步奖二等奖 2 项,部级科学技术进步奖一等奖 3 项、二等奖 9 项、三等奖 8 项。

(收稿日期:2014-06-25)

发展计划(“863”计划)“未来网络体系结构和创新环境”项目首席专家。目前主要从事新型互联网协议和新型网络体系结构领域的研究工作。入选教育部“新世纪优秀人才”,获得国家科学技术进步奖二等奖、多项部委一等奖。在未来网络领域担任多个国际学术会议主席、亚洲未来互联网学会主席、中美学术网未来互联网工作组联合主席、中国 SDN 和开放式网络专业委员会常务副主任等学术兼职,应 ONF 邀请任研究顾问。近 5 年发表学术论文百余篇(SCI 收录源 20 余篇);以第一发明人获得授权国家发明专利 10 余项;在 IETF 制定国际标准草案 20 余项,颁布或已获批准 4 项。

(收稿日期:2014-07-12)