



面向应用的云数据中心网络技术研究

陈楠,樊勇兵,何晓武,刘艺,黄志兰

(中国电信股份有限公司广东研究院 广州 510630)

摘要:提出了具有应用感知、网络可编程、网络虚拟化、弹性与高可靠性等特点的面向应用的云数据中心网络架构,并基于该架构对各层所涉及的关键技术及应用情况进行了分析,该网络较好地满足了云计算环境下对数据中心网络的要求。

关键词:软件定义网络;数据中心;网络功能虚拟化;业务链;面向应用;云计算

doi: 10.3969/j.issn.1000-0801.2014.09.018

Research on the Technology of Application-Oriented Cloud Data Center Network

Chen Nan, Fan Yongbing, He Xiaowu, Liu Yi, Huang Zhilan

(Guangdong Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China)

Abstract: The architecture of application-oriented cloud data center network, which is application-aware, network programmable, network virtualization, flexible and high available was proposed. The key technologies and their applications in each layer were analyzed. The proposed network architecture meets the requirements of data center network for cloud computing.

Key words: software defined networking, IDC, network function virtualization, service chain, application-oriented, cloud computing

1 引言

各类云业务的高速发展与应用,对云数据中心网络提出了更高的要求。数据中心网络除实现“高性能、大带宽进行数据传输”这一基本目标外,还应具备对应用的感知能力。在传统模式下,用户需要配置底层各类网络参数才能确保应用的正常运行,浪费了大量时间和精力,同时还要求具有较高的网络知识水平与能力。是否可以像服务器虚拟化技术一样,用户只关心应用所需的资源与功能要求,而由系统自动实现逻辑资源与物理资源的转化与自动化部署配置?本文基于网络策略模型和软件定义网络(SDN)技术,提出了一种面向应用的云数据中心网络模型,实现网络对应用的感知与自动化部署管理。

2 面向应用的云数据中心网络

2.1 云数据中心网络的挑战

云数据中心是指利用云计算技术,自动化地按需提供各类云计算服务的新一代数据中心^[1]。与传统数据中心网络相比,云数据中心网络面临下述几个方面的挑战。

(1)网络规模

随着服务器虚拟化技术的广泛应用,单台服务器上承载的业务数量较传统数据中心有10倍以上的增长,同时数据中心所容纳服务器的整体数量也在不断增加。以一个具有5万台物理服务器的数据中心为例,其MAC地址数量高达150万个^[2]。不难推断,对应的业务数量和租户规模相比传统数据中心也提升了超过一个量级。

(2) 网络性能

云数据中心基础设施实现了资源池化,用户访问应用的效率不再取决于单台网络设备的性能,而是整个网络所能提供的端到端性能。同时,在多租户场景下,具有各种不同服务等级的网络流量相互叠加,对网络性能保证提出了更高的要求。

(3) 网络架构

传统数据中心流量以南北向为主,而云数据中心包含了新型计算模式(Hadoop 等)、容灾备份、迁移等应用类型,存在大量东西向流量,同时流量整体呈网状,具有不确定性。因此对传统数据中心承载以南北向流量为主,具备高收敛比的接入—汇聚—核心的三级架构提出了挑战。

(4) 网络部署

传统数据中心承载的应用数量少,对于网络的要求较单一。为了配合应用的开通,需要对网络中所涉及的网元逐个进行配置,甚至需调整整个网络拓扑,不但耗时长还很很灵活。随着应用数量的增多,这种方式难以维系,因此对网络部署效率与灵活性都提出了更高的要求。

(5) 网络运维

传统数据中心一般通过网管系统对各种网络设备进行管理。实际上,由于各厂商设备间存在迥异的操作界面、版本及设备型号,自动化程度很低,无论是配置变更还是升级打补丁都面临着耗时长、配置复杂的问题。故障发生时还难以定位,对承载的应用造成较大的影响。

2.2 面向应用的云数据中心网络特点

为了应对云数据中心网络面临的挑战,本文提出了面向应用的云数据中心网络,通过引入网络抽象层实现应用需求与网络配置之间的适配,基于 SDN 架构实现控制层与转发层的分离,在控制层实现全网统一管理,在转发层提供虚拟化的多租户网络,利用网络功能虚拟化(NFV)技术实现灵活的网络服务提供,基于无阻塞网络拓扑及网卡优化技术实现高速、高效、高可靠的网络连接与传输。具体而言,该网络具备下述特点。

(1) 应用感知

基于网络组策略模型^[3,4],实现应用对网络的高层次需求与物理设备配置参数间的转化,从而使应用不必关心物理设备的具体实现与配置命令,只需根据业务特点定义网络,网络则通过相关技术,利用组策略模型,在网络内实现业务定义的逻辑网络与物理设备间的映射,从而根据应用的特点,在统一的物理网络之上,基于不同应用/流量提供

个性化的网络连接与业务。

(2) 网络可编程

SDN 的本质就是提供网络可编程能力。引入 SDN 架构,实现了控制平面与转发平面的分离、控制平面的逻辑统一以及网络能力的软件化。网络作为一个整体将各种能力通过编程接口提供给用户/应用,大大简化了网络部署配置的方式与时间,加快了应用创新速度与部署效率,避免了以往单厂商,逐个设备配置、管理、运维的难题。

(3) 虚拟化

云数据中心网络面临的一大挑战就是海量的应用与租户规模。应用网络虚拟化技术,通过定义新的租户标识方法,为每个租户提供了一张逻辑上相互独立的业务网络,基于 NFV 和业务链相关技术^[9],将各类功能网元基于虚拟机承载,基于租户实现灵活的组合与编排,在网络的灵活性与隔离性之间实现了平衡。

(4) 弹性与高可靠性

网络可以根据应用负载情况进行动态调整,通过统一的控制平面及时下发相关策略,更新或新建网络连接,还可基于 NFV 技术灵活、快速地实时增减功能网元,实现网络能力的水平扩展(scale out)。在故障发生时,统一的控制平面可以快速定位故障,并基于全局拓扑指引流量通过其他等价路径/备选路径,提高网络的可靠性。

3 面向应用的云数据中心网络功能架构

如图 1 所示,面向应用的云数据中心网络可分为网络抽象、网络控制、业务网络、物理网络 4 层。

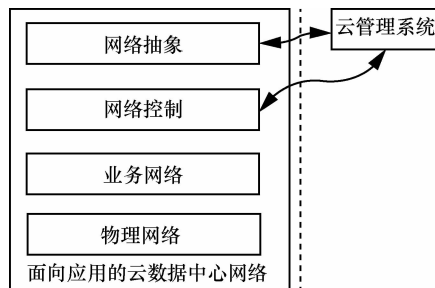


图 1 面向应用的云数据中心网络功能架构

网络抽象层面面向应用,向上提供 API,基于统一的网络策略模型,将面向应用的高级网络需求转换为面向设备的低级网络需求,即将逻辑网络需求映射为物理网络设备配置策略,同时配合云管理系统提供计算、存储、网络资源的统一编排。举例而言,应用 A 需要定义一个横



跨多台物理服务器的虚拟交换机,实现多服务器承载的多虚拟机端口间的网络连接,网络抽象层通过 API 获取该网络需求后,基于网络策略模型将需求转换为底层所涉及的交换机,服务器端口的转发、配置策略,并发送到下层。

网络控制层实现全局拓扑的统一管理,全网网元的配置管理、变更管理、系统管理以及路径(由)计算等功能。其中拓扑管理通过收集网元及链路信息,生成并维护多个层面的网络拓扑信息,为各类应用提供网络拓扑抽象模型并用于路径计算;配置管理基于网络抽象层所获取的应用需求,实现对纳管的各类网元(物理设备和虚拟设备)的流表或转发策略、端口策略等配置的生成、下发与回收;变更管理基于对纳管网元状态的监控,根据应用需求的动态变化,自动化实现设备变更操作,并支持变更控制及回滚;系统管理支持与网管系统对接,收集全网维护管理信息、告警故障信息等;路径(由)计算利用拓扑管理提供的信息并结合业务需求进行路径计算,生成满足不同业务需求和管理需求的可选路径。

业务网络层提供灵活、多租户隔离的网络连接与服务。每个租户独享一张虚拟网络,通过虚拟交换机、虚拟路由器实现 L2/L3 的网络连接,同时基于业务链(service chain)对虚拟防火墙(vFirewall)、虚拟负载均衡器(vLoadbalance)、虚拟入侵检测器(vIDS)等 L4~L7 设备进行灵活编排,提供基于租户的灵活的网络服务。

物理网络层提供高速、高效、高可靠的物理网络连接与数据传输。通过构建适合流量流向特点的低时延、高性能物理网络以满足高速转发需求,利用网卡加速技术优化服务器网卡转发性能。

4 面向应用的云数据中心网络关键技术与应用

4.1 网络抽象层

一般地,应用的部署需应用管理员与网络管理员协商确定网络的拓扑与配置细节,应用管理员通常对应用本身有较深刻的认识,但对于网络并没有太多了解,而网络管理员则恰恰相反,因此相对于计算和存储资源的部署而言,网络的部署往往要花上更多的时间和精力。通过网络抽象层,利用一种基于声明式(declarative)的策略模型^[3,4]实现对网络能力的抽象,将应用对网络的需求从底层网络基础设施的配置细节中分离出来,使得应用管理员通过抽象的策略来表述对于网络的需求,对于具体的实现则由系统负责。

开源组织 OpenDaylight 的 Group Policy Plugin^[4]项目正在进行策略模型制定的工作。如图 2 所示,整体网络行为可抽象为一系列具有相同策略属性的端点(endpoint)组成的端点组(endpoint group),该端点组根据合同(contract)规定的内容(由分类器(classifier)、指令(directive)等构成)与其他端点组进行通信。基于该模型,可以使应用管理员和基础设施(网络管理员)分工合作,各有侧重:应用管理员处理高层次的网络抽象而并不关心底层网络如何实现;基础设施(网络)管理员处理基础设施(设备)具体的实现而不需要了解应用具体的考虑(如某两个组件之间的网络连接策略)。

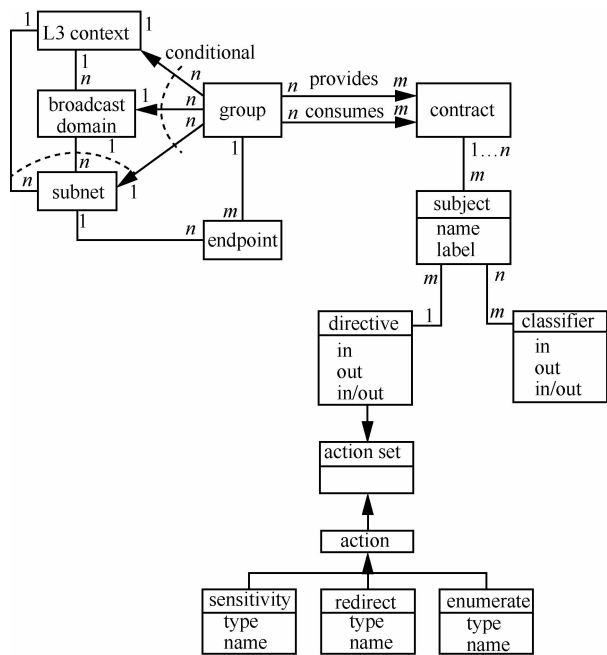


图 2 OpenDaylight 组策略信息模型

实际上,另一个重要的开源组织 OpenStack 也在从事相关工作。网络组件 Neutron 的 Group Policy 项目也计划基于组策略的抽象(group-based policy abstractions)利用声明式策略驱动连接模型扩展 Neutron API,向用户提供简单的、面向应用的接口^[3]。在目前已商用的产品中,阿尔卡特朗讯公司的 Nuage 产品的 Virtualized Services Directory 组件也部分实现了该功能。

4.2 网络控制层

网络控制层要实现全网网元的管理及拓扑维护,需要通过协议与各网元进行信息交互。具体而言,网络控制层与各类网元间的信息可分为两类:网元的状态/配置信息和路由/转发信息。各网元通过状态/配置信息的交互将

设备状态上报给网络控制层,网络控制层利用该类信息对网元状态进行监控,并通过该类信息下发对各网元的配置,从而实现对网元的配置管理与变更管理。另外,还可以基于此类信息向网管系统提供相应的维护管理信息和告警故障信息。网络控制层通过与全网网元交互路由/转发信息获取并维护全局拓扑,同时基于上层应用的需求及全网拓扑情况,为特定的流计算端到端路径,并将所需的路由/转发信息下发至相关网元,完成路径建立。这两类信息可分别由不同的协议承载,各自的特点与应用场景不尽相同,具体信息见表 1。

4.3 业务网络层

4.3.1 多租户网络虚拟化

业务网络层主要解决将网络分成多个虚拟网络,面向不同的租户相互隔离,并基于租户提供可灵活组合的网络业务。网络划分即在数据平面实现网络虚拟化。其实现技术无非两类,采用点到点的租户标识技术或叠加网(overlay)技术。

被广泛使用的 VLAN 就是一种点到点的租户标识技术,通过划分不同的 VLAN,实现租户间的 L2 隔离。但受报文格式限制,理论上仅能支持 4 096 个标识,同时还存在收敛速度、网络利用率不高等问题。为了突破该限制,必须引入规模更大、更灵活的租户标识与网络隔离技术。近年来兴起的 OpenFlow 技术^[6]为此提供了可能。基于 OpenFlow 典型架构,网络中所有交换机通过对控制器下发的 OpenFlow 流表中不同流表项的匹配实现对多同租户的隔离,当新的分组到达网络边缘时,网络边缘处的交换机会将首分组上送至 OpenFlow 控制器,控制器根据源目的地址等因素计算出路径,然后对该路径上所有网元下发流表。这种被动式逐跳的 OpenFlow 解决方案主要面临下述挑战:随着规模的扩大,支持 OpenFlow 协议的设备需存储更多的流表、转发表;由于采用被动式,首分组必须上送到控制器,控制器对该路径中所有节点进行通信,这势必带

来新的性能瓶颈和时延,因此该类方式的应用规模受到一定限制,并未被广泛部署应用。

另一种思路就是采用叠加网技术将网络划分为边缘和核心两层。访问控制、隔离等网络功能在网络边缘实现,让网络核心专注于流量转发。这样,既降低了物理设备中对于流表数量的压力(仅保存物理服务器地址而不需要保存虚拟机地址),同时实现租户网络与物理网络的相对隔离,保持底层物理网络的稳定性。一般将虚拟机监控器(VMM)中的虚拟交换机/路由器作为隧道的起始点(也可以将物理交换机/路由器作为隧道终结点),即上述的网络边缘节点,在底层物理网络之上基于隧道封装协议的标识实现端到端的网络隔离。面向应用的云数据中心在业务网络层基于叠加网技术方案,该方案的典型封装协议如 VxLAN、NVGRE、MPLS over GRE、GENEVE 等获得了业界软硬件厂商不同程度的支持与应用。值得注意的是,该类方案由于与底层物理网络完全解耦,因此无论在底层网络的排障还是对基于租户的 QoS 保证上均存在不足。

4.3.2 网络功能虚拟化与业务链

业务网络层另外一个关键问题就是如何实现基于多租户的可灵活组合的网络业务。如前所述,云数据中心无论是租户数量还是应用部署数量和速度都有较大幅度的提高,因此传统的基于单个应用分别部署相关功能网元的方式已不能适应多租户环境下各类云应用对网络的需求。所幸的是,随着 NFV 技术的发展,越来越多的基于传统网络设备承载的网元逐步向基于标准 x86 服务器承载转变。功能网元的 x86 化、虚拟化已成为云数据中心网络的一个重要发展趋势,这也为灵活提供各类网络业务提供了可能。而基于业务链(service chain)技术可引导网络业务报文次序通过多个业务处理节点,实现各类功能网元间灵活的串接,为不同的业务和用户提供个性化的端到端服务,或针对不同的流量类型,实现差异化的路径。业务链的实现有两种方式:一种是利用虚拟交换机/路由器与各类功能

表 1 网络控制层相关协议

协议名称	种类	标准化	简介与应用
Netconf	状态/配置信息	IETF 标准	基于 XML 的网络配置管理协议,一般应用于对物理网元的配置管理
OVS-DB	状态/配置信息	IETF 标准	OVS 交换机的配置协议,VMware NSX 采用该协议
OF-Config	状态/配置信息	ONF 标准	远程配置和控制 OpenFlow 交换机的协议,ALU Nuage 采用该协议
XMPP	状态/配置信息+路由/转发信息	IETF 标准	基于 XML 的可扩展通信和表示协议,Juniper Contrail 采用该协议
OpenFlow	路由/转发信息	ONF 标准	OpenFlow 控制器与交换机间流表交互协议,VMware NSX、ALU Nuage 采用该协议
IGP	路由/转发信息	IETF 标准	自治网络系统内路由信息交互协议,用于网络控制层与物理网络层设备交互
BGP	路由/转发信息	IETF 标准	路由域间的多路连接协议,用于网络控制层多控制器间信息的同步



网元连接,通过网络控制层对虚拟交换机/路由器相关端口进行动态配置与策略下发,从而实现上述功能,该方案现阶段已存在商用产品,但由于流量需绕行虚拟交换机/路由器,因此存在性能瓶颈;另一种思路是基于重叠网技术,通过隧道封装协议携带应用信息,相关网元通过识别该封装协议中的应用元数据,实现基于业务/租户的网络业务的灵活组合。IETF 的 GENEVE 标准草案正是该类技术的代表之一,目前还没有成熟的商用方案。

4.4 物理网络层

4.4.1 网络拓扑

物理网络层技术焦点集中在实现网络高速、高效、高可靠的连接与传输的问题。随着 10GE、40GE 技术的逐步成熟并在数据中心的应用,相比而言,组网,即网络拓扑就显得更为重要。如前所述,云数据中心网络流量呈网状特点,传统适合南北向流量为主的三层网络架构不能很好地适应新的流量特点。如何构建一个高扩展性、无阻塞的网络成为物理网络层的关键。CLOS 网络^[7]提供了一种解决思路,最初 CLOS 网络是一种经典的严格无阻塞的多级互连网络,由 Charles C 于 1953 年提出,并在 Benes 网络的基础上演变而来。该网络由于在通信网和多处理器计算机系统中被广泛采用,因而受到广泛重视。近年来,CLOS 网络也开始被用于数据中心的网络架构中,充分利用其无阻塞和高扩展性等特点,以满足当前云计算环境下对流量网状分布和对数据中心网络的需求。一般认为 CLOS 网络适应于数据中心内部流量(东西向流量)较大的场景,在实际部署上采用叶—脊(leaf-spine)两级架构。

4.4.2 网卡性能优化

由于业务网络层采用叠加网的方案,利用物理服务器内 VMM 生成的虚拟交换机/路由器实现隧道的终结,因此对于服务器网卡的性能提出了更高的要求,另外,如前所述,随着 NFV 技术的发展,越来越多的功能网元将基于 x86 服务器承载,其中一个关键问题就是解决 x86 服务器的网络转发性能。换句话说,物理网络层的范围从传统网络设备扩展到了服务器的网卡上,必须解决服务器网卡的性能问题。实际上,可应用 VMDq 和 SR-IOV 技术实现对特定应用网卡 I/O 性能的提高。VMDq 技术将数据分类功能从 VMM 转移到网卡中,因此提高了 I/O 性能;SR-IOV 技术实现了将 PCI 功能分配到多个虚拟接口以在虚拟化环境中共享一个 PCI 设备的资源,能够让网络传输绕过软件模

拟层,直接分配到虚拟机,降低了 VMM 中 I/O 的开销。这两类技术目前尚未规模应用,还面临着与 VMM 的兼容与适配及网卡成本的问题。DPDK 技术通过优化 I/O 处理机制、大页表、CPU 核绑定等方式可实现网卡对小分组转发性能的优化与提升,从而提高基于 x86 服务器承载的功能网元的转发能力,具有较好的应用前景,但目前对于网络特性的支持程度尚需要提高。

5 结束语

面向应用的云数据中心网络实现了网络对应用的感知,通过可编程接口实现网络部署、管理的自动化,利用虚拟化技术实现多租户网络的提供,同时具备较好的弹性和可靠性,较好地满足了云计算环境下对数据中心网络的要求。虽然所涉及的各类技术尚不完善,还在不断发展中,可以预期,随着相关技术的发展与成熟,面向应用的云数据中心网络将会获得更大程度的应用,实现云网合一,将云计算能力提升到一个新的高度。

参考文献

- 1 唐宏,陈楠. 云数据中心演进思路剖析. 电信科学, 2011, 10(6): 30~35
- 2 樊勇兵,丁圣勇,陈楠. 基于业务交换机的大规模云数据中心通用网络架构设计. 电信科学, 2013, 29(10):1~4
- 3 Neutron/GroupPolicy, Openstack. <https://wiki.openstack.org/wiki/Neutron/GroupPolicy>
- 4 Group Policy. OpenDayLight. https://wiki.opendaylight.org/view/Group_Policy:Main
- 5 ETSI. Network Functions Virtualisation—Introductory White Paper, 2012
- 6 ONF. OpenFlow Switch Specification 1.3.1, 2012
- 7 Charles C. A study of non-blocking switching networks. Bell System Technical Journal, 1953, 32(2): 406~424

[作者简介]



陈楠,男,博士,现就职于中国电信股份有限公司广东研究院,长期从事云计算、数据中心方面的技术跟踪研究与标准化制定工作,ITU-T Q14/SG11 联合报告人, Q18/SG13 云计算基础设施即服务功能需求及架构(Y.ccIaaS)编辑人,主要研究方向为 SDN、NFV、服务器虚拟化、数据中心网络、管理平台技术等。