云数据中心的 SDN 解决方案

钟翠，王蕾，罗兴
（华为技术有限公司，广东 深圳 518129）

摘　要：SDN 的设计理念是对网络资源进行池化管理，自动化实现网络资源的按需申请和调用，加速数据中心业务的上线周期。首先介绍了数据中心 SDN 的核心思想和 SDN 开放架构全景图，然后介绍了 SDN 网络模型和 SDN 业务部署方案，最后阐述了华为 CloudFabric 云数据中心网解决方案。

关键词：SDN；数据中心；自动化
中图分类号：TP399
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018213

SDN solutions for cloud data center

ZHONG Cui, WANG Lei, LUO Xing
Huawei Technologies Co., Ltd., Shenzhen 518129, China

Abstract: The design concept of SDN is to pool and manage network resources, automate the on-demand application and call of network resources, and accelerate the online cycle of data center services. Firstly, the core idea of data center SDN and the panoramic view of SDN open architecture was introduced. Then, the SDN network model and SDN service deployment plan was introduced. Finally, solution of Huawei CloudFabric cloud data center network was described.

Key words: software defined networking, data center, automation

1　引言

传统数据中心业务上线主要是通过人力操作繁杂的命令行模式，但随着云计算的发展，数据中心网络在过去几年有了翻天覆地的变化。网络带宽从吉比特演进到 10 万兆，接入端口数量从几百台增长至几万台，大型数据中心的服务器数量甚至超过了 8 万台，假定每台服务部署 20 台虚拟机，面对互联网新业务的快速上线和迭代诉求，数据中心网络管理员应对百万虚拟机的上线和变更需求，将会是一个巨大的挑战。除此以外，数据中心网络的管理人员还面临着各种各样的难题。

- 数据中心网络的基础设施技术越来越复杂，新技术不断推陈出新。
- 数据中心的业务需求经常变化，网络管理员如何应对？这么多配置如何进行调整？
- 数据中心中新服务越来越多，网络如何应对这些部署需求？

网络作为数据中心的基础设施，本质是上层业务的支撑系统，拥有“快速弹性的架构”和一

收稿日期：2018−05−28；修回日期：2018−07−05

个“资源池”，可以提供“按需自助服务”且这些服务是“可测量的服务”，而资源池里的服务最终是可以被“宽带接入”访问，成为新一代数据中心 IT 基础设施建设的一个普适性需求。

为了满足以上诉求，业界提出了软件定义数据中心的概念和方法，将 IT 基础设施资源变成服务对象，通过虚拟化方式进行抽象，通过自动化的流程和软件方式提供服务。SDN（software defined networking，软件定义网络）技术的出现适应了网络 IT 化、设备资源池化和架构标准化的趋势。

2 数据中心 SDN 核心思想

SDN 的最终目标是服务于多样化的业务应用创新，通过开放架构，北向为上层业务提供了丰富的北向 API，基于可视化应用模型，让租户以应用视角定义网络诉求。通过核心组件自我驱动将应用模型自动翻译为网络配置，按需进行网络资源申请和配置，灵活调用底层网络能力，屏蔽底层物理转发设备的差异，将底层的物理资源池化共享，实时、按需、动态地分配给不同租户的不同应用。

3 SDN 开放架构全景图

与传统网络不同，SDN 架构模型如图 1 所示，既要理解业务对网络的要求，通过业务意图的编排和理解将其映射为抽象的网络模型；同时将映射后的抽象网络模型翻译成各网元可以理解的转发策略，分发到数据中心的网络设备上。使得网络架构基于开放架构，分层解耦、各司其职，层次间通过标准化接口互联，满足 SDN 开放性、扩展性、生态融合的诉求。

如果将 SDN 架构比作人体系统，那么这个系统的不同层级和接口都有其独特的定位和作用。

3.1 SDN 的“五官”：业务呈现协同层

如同人体系统中负责与外界沟通交互的五官，业务呈现和协同层是 SDN 架构最上层的交互界面，是 SDN 对外显示信息的通道，主要负责理解用户的业务诉求，然后整合数据中心内计算、存储、网络资源，通过标准化接口实现资源协同配合，完

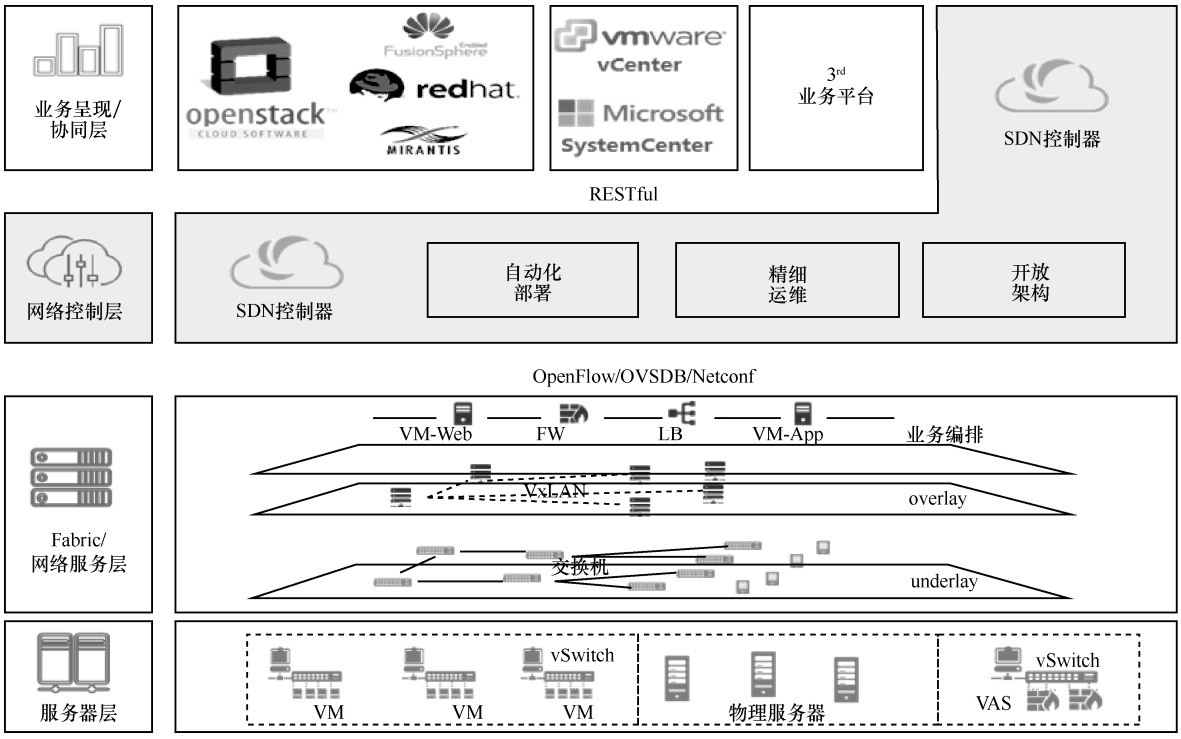

图 1 SDN 架构模型

成业务编排、自动发放、服务保障等功能。

3.2　SDN 的“大脑”：网络控制器

网络控制层是 SDN 整体架构的“大脑中枢”，起着贯通南北的作用，这一层的展现实体就是控制器。控制器北向支持 RESTful API 对接上述业务呈现和协同组件，南向使用 OpenFlow、OVSDB、Netconf 等接口实现虚拟网络设备和传统物理网络设备，如交换机、防火墙、负载均衡器等的统一纳管。

首先，协同层使用 RESTful API 按需申请网络资源，通过控制器创建逻辑网络模型；然后，控制器根据逻辑网元的属性和逻辑拓扑关系，将模型转换为物理和虚拟网元可以识别的配置或流表（这种转换过程称为映射），如图 2 所示，下发到网络设备来开通网络业务。

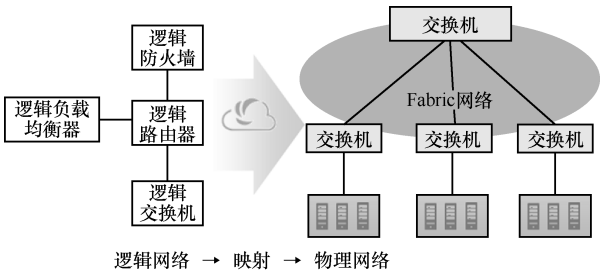

图 2　逻辑到物理网络映射关系

3.3　SDN 的“躯干”：Fabric 网络

如同人体的躯干，Fabric 网络（如图 3 所示）构成了 SDN 架构的主体，是网络流量和业务的主要承载组件，由骨干节点交换机、叶子节点交换机、虚拟交换机及防火墙和负载均衡器组成。

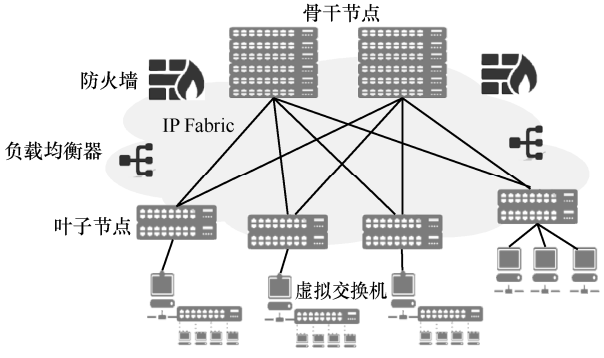

图 3　Fabric 网络

3.4　SDN 的“手足”：虚拟平台

虚拟平台层主要由虚拟化服务器构成，是 SDN 架构的边缘节点，如同人体系统的末端——手足。虚拟化服务器通常由计算虚拟化厂商提供。SDN 对虚拟化服务器主要关注两个方面。

- 将虚拟交换机 vSwitch 安装部署到虚拟化服务器的 Hypervisor 中，利用 vSwitch 软件灵活性与计算实例亲和性的优势，满足数据中心用户越来越多的定制化和安全诉求。
- 在无云平台的虚拟化场景下，控制器同 VMM 联动感知计算实例上下线和迁移细节，并为其自动化开通网络；下发安全策略、流量控制策略，实现面向应用的网络自动化发放。

3.5　SDN 的“神经”：组件接口

如同大脑要通过神经系统控制全身各处，SDN 控制器也需要通过标准接口与各组件进行交互。如图 4 所示，以控制器为中心，接口分为北向（连接协同层）、南向（连接 Fabric 层）和东西向（连接外部组件）接口。

（1）北向接口

RESTful API 表示层状态转换接口，互联网软件架构拥有一些明显的特征：使用 URI 标记资源；利用 HTTP 的 get/post/put 等标准接口实现交互；交互过程中服务器端不记录客户端状态。基于上述特点，RESTful API 非常适合用于海量并发请求的场景。

（2）南向接口

南向连接的网络设备种类繁杂，因此支持的接口也种类繁多。

Netconf 接口：基于 CLI 接口演化，是一种基于 XML 语言可批量执行的设备配置接口，主要解决 CLI 接口安全性差、命令执行效率低、无法描述复杂逻辑关系的问题，对物理网络设备范围内兼容性最佳。

OpenFlow 接口：随 SDN 理念一起推出，也

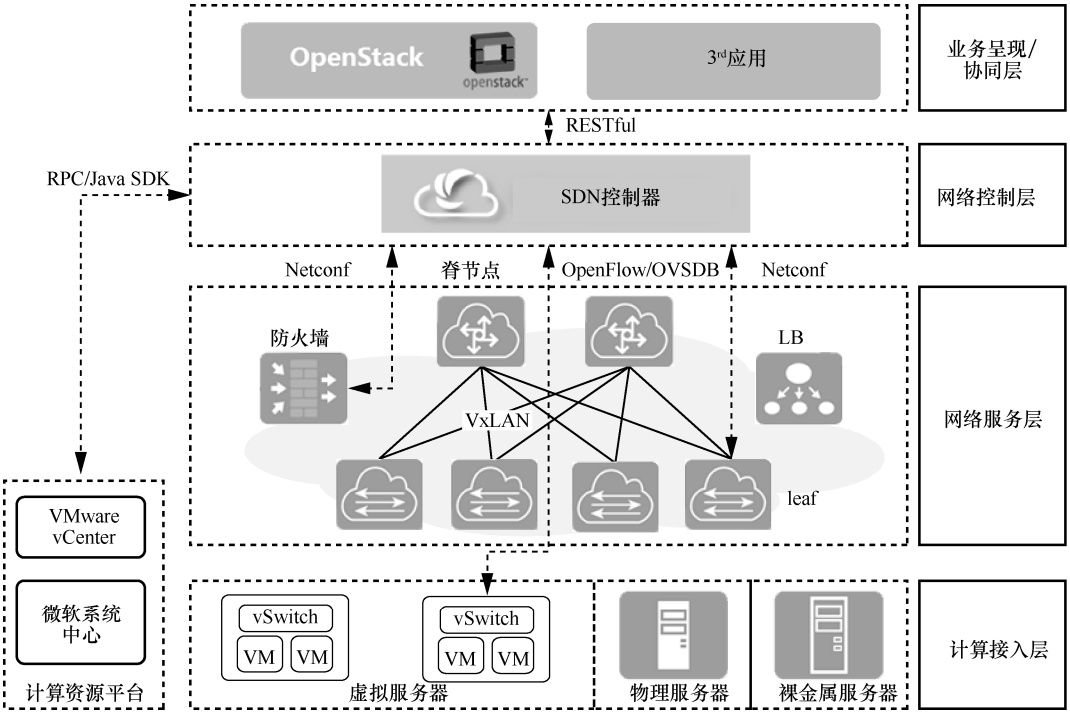

图 4　开放接口

是转发型 SDN 的标准协议，从 2009 年至今已经发展到 1.4x 版本，它通过多级流表（flow table）的方式支持流量的细化控制，在新的协议版本中通过支持组表（group table）实现了多播功能，同时支持多控制器并发连接。

OVSDB 接口：同开源 OVS（open virtual switch）相伴而生。将 OVS 组件拆开可以看到 OVSDB-server 模块，作用是作为中间数据库向控制器和 OVS 传递配置、状态信息。当控制器有配置需要下发到 OVS 时，如 mirror、LACP 等，利用 OVSDB 接口向 OVSDB-server 写入配置数据，再由 OVS 自动加载；当 OVS 发生状态、配置变化需要通知控制器时，也将这些变化的信息写入 OVSDB-server 数据库，控制器通过读取数据库数据感知 OVS 的变化，如感知虚拟机端口上线事件等。

除此以外，南向接口还包括 CLI、SNMP 等接口，不再赘述。

（3）东西接口

目前东西接口主要是指同 VMM 平台之间的连接关系，包括两类接口 Java SDK 和 RPC。

Java SDK 是 VMware vCenter 面向第三方厂商开放的可编程接口；通过此接口第三方厂商可以感知 ESXi 虚拟机上线、下线、迁移等事件；查询 vCenter 范围内 VM 的详细信息以及 VSS/VDS 交换机部署信息。

4　SDN 的网络模型

随着 IT 基础架构云计算时代的不断演进，虚拟机数量呈几何数增长，越来越多的虚拟机交互业务需要通过网络实现，东西向流量开始取代南北向流量，成为数据中心的主流。流量模型的转变为数据中心的网络架构提出了全新的升级诉求。传统 3-tier 网络架构，导致流量时延、拥塞，扩展性差，亟须扁平化、无阻塞的网络架构演进。起源于大型 OTT 的 spine-leaf（叶脊）网络架构呼之欲出。与传统 3-tier 架构相比，spine-leaf 通过大二层接入，实现网络中任意两个服务器通过 leaf-spine-leaf 三跳可达。spine-leaf 通过 full-mesh

连接，spine 和 leaf 均可支持横向扩容，从而构建出了一个无阻塞、可扩容的数据中心优选组网模型。

4.1　underlay 组网

传统 Fabric 为单层网络，经历了 xSTP、CSS/iStack、M-LAG（multi-chassis link aggregation group，跨设备链路聚合）等多种组网技术。

xSTP 是最为传统的 Fabric 组网技术。通过设备控制面运行 xSTP，在二层环路中阻塞部分链路，将环状网络拓扑映射为树状无环网络拓扑。当转发链路出现故障时，xSTP 将阻塞链路打开，恢复流量转发。xSTP 协议族历经 STP、RSTP、MSTP 多次优化，目前已经发展得非常稳定，在传统数据中心有非常多部署案例。但是由于协议阻塞了部分链路，因此链路利用率较低，因此 xSTP 组网近些年越来越多地被无环组网技术替代。

CSS/iStack 是典型的无环组网技术。框式交换机利用 CSS 特性或者盒式交换机利用 iStack 特性可以将同系列的两台或者多台物理网络设备虚拟为一台逻辑网络设备，逻辑网络设备由逻辑系统主控节点统一管控。逻辑设备间互联，设备之间的多条链路建议配置为聚合链路，Fabric 天然无环，充分发挥硬件设备转发性能优势。同时设备虚拟化后网元数量减少，简化运维界面，可有效降低数据中心运营成本。无环组网是目前推荐的组网方案。

伴随着数据中心网络规模急速膨胀，网络设备控制面（一般是指 ARP 处理）处理能力逐渐无法满足组网要求，上述 CSS/iStack 逻辑系统控制面由主控节点统一管控，承担着巨大升级压力，如果骨干节点系统崩溃，将对数据中心业务造成严重影响。因此一种新的虚拟化技术 M-LAG 应运而生。

构建 M-LAG 虚拟逻辑系统的网络设备各自拥有独立的控制面，分别处理收到的 ARP 请求报文，分担组网压力。M-LAG 作为网关设备时部署 VRRP，使用阻断 VRRP 心跳的方式构建 VRRP 双主系统，实现本地优先转发，既分担了控制面压力，提升设备级可靠性又充分利用了设备带宽资源。

4.2　overlay 组网

为了满足数据中心海量虚拟机带着 IP/MAC（不变）地址迁徙的刚性需求，横贯整个数据中心网络或者跨数据中心的网络大二层需求应运而生。应对挑战的理想方案是在传统单层网络基础上叠加一层逻辑网络。overlay 网络包括物理基础层（underlay）和逻辑叠加层（overlay）两个层次，物理基础层可以使用 underlay 的传统组网技术部署，只要数据中心网络上任意两点路由可达即可；逻辑叠加层则需满足大二层及动态化要求。

overlay 网络的技术多种多样，目前业界主推的 VxLAN 是一种基于 IP-IP 的隧道封装技术，如图 5 所示，在原始报文的基础上添加 L2/L3/L4（UDP）头部信息，外层 L2 信息在转发过程中逐跳修改封装，外层 L3 信息作为隧道标识，只在隧道端点封装/解封装，外层 L4-UDP 信息包括一组扩展为 24 bit 达 16 Mbit/s 的子网 ID，用于隧道端点设备识别报文转发到哪一个网络。

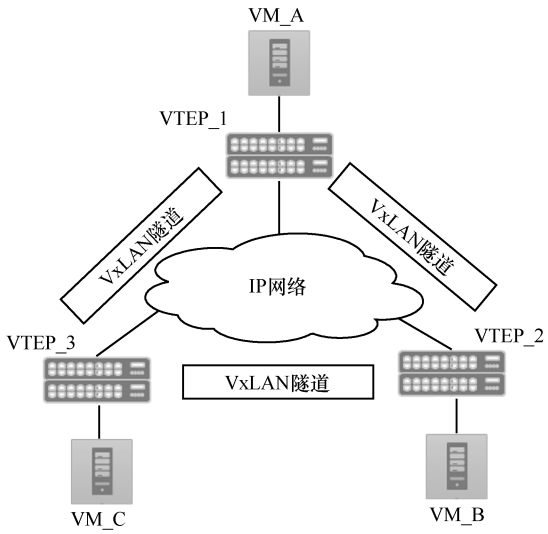

图 5　VxLAN 组网

报文添加 VxLAN 封装后，VxLAN 头部是标准的 TCP/IP 封装，因此传统 IP 技术的操作、管

理与维护工具、运维理念以及设备兼容性等方面均有非常好的表现，如果将 VxLAN 网络边缘延伸到服务器内部的 vSwitch 虚拟交换机，那么数据中心已有的大部分网络设备都可以重复利用，降低数据中心的部署成本。

5 SDN 业务部署方案

5.1 云网一体化方案

为了解决传统数据中心业务部署效率低、资源利用率低、运维管理复杂的问题，数据中心需要联手云计算架构场景演进。OpenStack 是一个开源的 IaaS（基础设施即服务）云计算平台，类似一个数据中心的“操作系统”，管理着数据中心的计算（Nova）、对象存储（Swift）、网络（Neutron）等资源。Neutron 是提供网络即服务（networking as a service）能力的工程，其提供了比较完整的 L2~L7 层网络模型和对应的 API；Neutron 南向提供了灵活的 plugin 和 driver 机制，方便厂商对接控制器和设备。

云平台提供计算和网络统一管理界面，控制器与云平台开放对接如图 6 所示。SDN 控制器通过 Neutron 的 API 实现网络即服务能力。业务管理员通过云平台界面统一创建计算资源和网络资源。

业务管理员通过云平台将网络资源分配给指定的业务或应用。云平台将业务下发指令传递给网络控制器，再由网络控制器将配置明细自动下发至设备，无需人工配置。

业务管理员通过云平台进行计算和存储资源的创建、删除和迁移等操作。云平台、网络控制器、网络设备和服务器之间自行进行协调交互，无需人工干预。

5.2 计算虚拟化方案

在计算业务管理系统庞杂，或计算管理和网络管理融合度不高又无法构建统一云平台的情况下，适用计算联动场景。

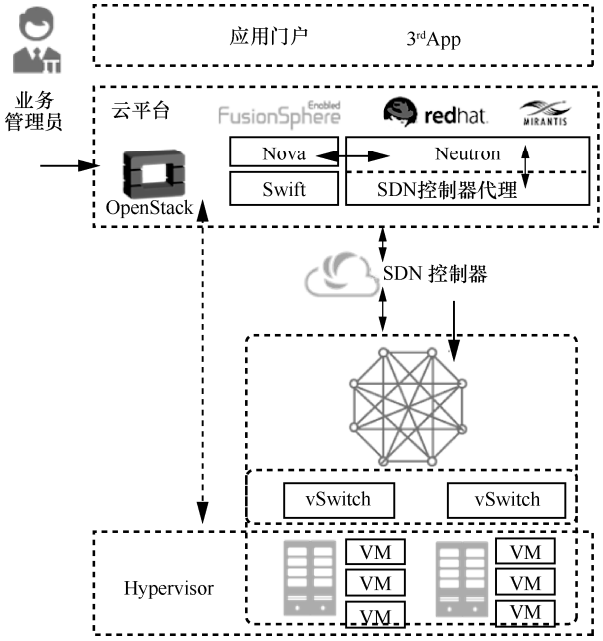

图 6 云网一体化方案

通过 SDN 控制器对接计算虚拟化平台，由控制器和计算虚拟平台一同承担业务下发职责，实现计算与网络协同发放。计算虚拟化方案如图 7 所示。

业务发放包括以下两个部分。

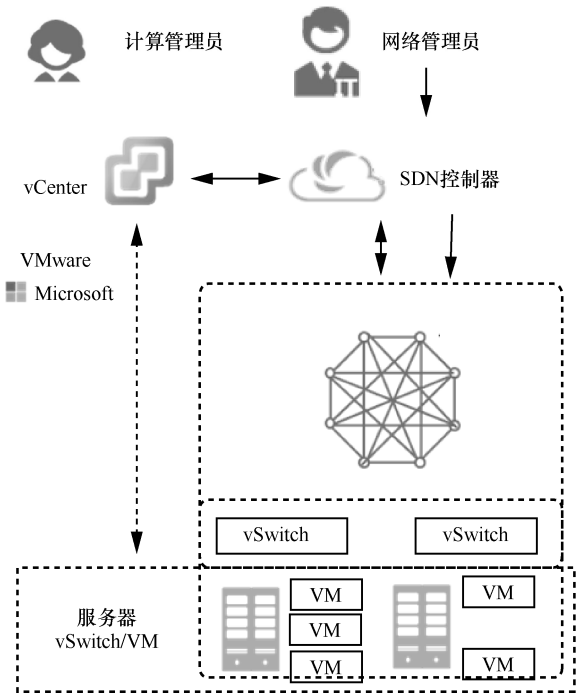

图 7 计算虚拟化方案

- 网络业务发放：网络管理员通过 SDN 控制器将网络资源分配给指定的业务或应用。其中 overlay 网络的业务配置及 VM/物理机相关接入配置均由 SDN 控制器自动下发完成。
- 计算业务发放：计算管理员通过 VMM 进行计算和存储资源的创建、删除和迁移等操作。SDN 控制器可自动感知 VMM 对计算资源的操作，最终实现资源上线后的网络打通。

5.3　机架出租

机架出租场景是指出租方（通常为运营商）向承租方提供机房、机柜空间、网络接入和增值服务的租用服务。机架出租场景直接由网络管理员通过 SDN 控制器进行网络资源业务发放，无需对接云平台/计算资源/其他第三方系统。机架出租场景根据租户是否自带网关设备可进一步区分为如下两种应用场景。

- 租户自带网关：租户托管的设备包括服务器、L2 交换机、网关设备和防火墙。租户设备通过 L3 方式接入运营商网络。
- 运营商提供网关：租户托管的设备只有服务器和 L2 交换机，不包括网关设备和防火墙。租户设备通过 L2 方式接入运营商网络，网关和增值服务都由运营商提供。

6　华为 CloudFabric 云数据中心网解决方案

为了帮助客户快速适应互联网下各种云业务的变化，华为推出了面向新一代云数据中心的 CloudFabric 解决方案，旨在为客户构筑敏捷、智能、超宽、开放的云数据中心网络，支撑数据中心云业务高效发展。基于大二层无阻塞网络，将 SDN 控制器 Agile Controller 作为华为云数据中心网络的核心组件，可实现对网络资源的统一控制和动态调度，快速部署云业务。支持与 AI 和大数据分析系统联动，不断调整和优化网络能力，简化网络运维形成网络闭环管理，最终实现可以自治自愈的自循环智简网络系统。

6.1　基于 spine-leaf 架构的 AI 应用网络

spine 节点由华为 CE 系列的高性能交换机担当，可实现线速的 VxLAN 转发；通过业界独家的 iPCA 技术对实际业务逐跳进行精准的分组丢失、时延、抖动检测，实现故障快速定位；同时它采用全可编程架构，可灵活快速满足客户定制需求；技术自主可控，关键器件 100%国产化；leaf 节点可提供 25GE/40GE 接口，支持内置 FPGA 专用芯片，可以实现纳秒级低时延转发。

提供基于以太网的低时延、零丢失分组 AI 应用网络。通过精准反压和动态水线等技术，不仅能够在初期对网络进行有效的拥塞调节，实现网络零丢失分组；同时可大幅提升吞吐率，有效助力数据中心网络提升业务执行效率，降低运维复杂度和建网成本。

6.2　端到端的云化数据中心网络

6.2.1　图形化网络规划与设计

提供网络规划和设计阶段的图形化工具、网络业务变更和应急故障处理指导、网络冗余与容灾设计辅助，更高效、准确地规划设计网络。

6.2.2　SDN 控制器 Agile Controller

基于开放架构提供南北向标准接口、北向抽象网络资源和服务，支持与标准云平台或第三方应用无缝集成，同时支持与 vCenter/SystemCenter 主流计算平台联动，以业务为中心，自动化实现网络按需部署。南向适配不同设备和网络实现，实行资源池化管理，实现跨数据中心物理与虚拟网络资源池化纳管。华为 SDN 控制器 Agile Controller 自身提供高可靠集群能力，系统采用负载分担方式，对南北向业务进行处理，满足数据中心业务的高可靠性要求。

提供高于业界 10 倍管理能力的核心组件 SDN 控制器 Agile Controller，可通过图形化拖拽定义业务模型，自动化完成业务到命令行的翻译和下发，实现网络业务快速发放；可一键式实现网络端到端打通，简化业务上线复杂度，提高准确率，业务部署周期从周缩短到分钟级。

6.2.3 网络智能分析

构筑的数据中心领域大数据分析器，提供无处不在的网络应用分析与可视化呈现，打通应用和网络的边界，提供技术创新到商业创新的连接。与传统聚焦资源状态的监控方式相比，通过 Telemetry 方式提供全网真实流秒级采集，并根据大数据智能算法对网络数据进行分析、呈现，实时感知 Fabric 的状态、应用的行为状态；由传统设备监控变为实时业务监控，打破网络和应用的边界，从应用视角看清网络，帮助客户及时发现网络与应用的问题，保障应用的持续稳定运行。

通过 FabricInsight 组件实现 AI 的智能运维，实时收集真实业务流数据，支持百亿级数据的秒级检索和智能算法分析，实现异常流数据的快速发现和定位。系统可自动实现应用和网络的关联，呈现丢失分组严重的主机和故障网络路径，快速定位到故障节点，减少端到端故障处理时间，故障定位从天缩短到分钟级。

6.3 跨数据中心容灾和双活管理

为了满足多个远距离数据中心之间的容灾和业务双活，华为 CloudFabric 解决方案提供了多 DC 互联互通方案，可以构建同城双活、异地容灾（如两地三中心模式）的数据中心业务架构，同时跨 DC 的业务可以实现灵活按需互通；支持将同一套业务系统分别部署在两个独立 DC 中，可实现系统间 L2/L3 互通，并对外提供跨 DC 的双活服务，进一步提升服务质量。提供跨 DC 的微分段和业务链，使业务部署更加灵活。

截止到 2017 年底，华为 CloudFabric 云数据中心网解决方案已完成 180 多个 SDN 商用方案交付，广泛部署在中国、西欧、日本、韩国、南太平洋、俄罗斯、中东等全球各地，市场覆盖涉及运营商、互联网、金融、制造、政府等多个行业。

7 结束语

SDN 作为网络领域的一次技术革新，有效地支撑了商业化数据中心上层业务的服务需要，提升了网络上线和运维效率。随着网络技术的不断演进，网络将逐步走向以用户为中心的意图驱动解决方案，通过在商业意图和物理网络之间构建数字孪生，借助 ABC（人工智能、大数据、云计算）技术，以智慧、极简、超宽、安全、开放的理念，打造意图驱动的智简网络，从以设备为中心的网络向以用户为中心的网络转型，实现自动化、智能化，并最终走向自治。

参考文献：

[1] 韦乐平. SDN 的战略性思考[J]. 电信科学,2015, 31(1): 7-12.
WEI L P. Strategic thinking on SDN [J]. Telecommunications Science, 2015, 31(1): 7-12.

[2] 何晓明, 冀晖, 毛东峰, 等. 电信 IP 网向 SDN 演进的探讨[J]. 电信科学, 2014, 30(6): 131-137.
HE X M, JI H, MAO D F, et al. Discussion of evolution of carrier IP network to SDN [J]. Telecommunications Science, 2014, 30(6): 131-137.

[3] 赵慧玲, 史凡. SDN/NFV 的发展与挑战[J]. 电信科学, 2014, 30(8): 13-18.
ZHAO H L, SHI F. Development and challenge of SDN/NFV[J]. Telecommunications Science, 2014, 30(8): 13-18.

[4] 李晨, 段晓东, 陈炜, 等. SDN 和 NFV 的思考与实践[J]. 电信科学, 2014, 30(8): 23-27.
LI C, DUAN X D, CHEN W, et al. Thoughts and practices about SDN and NFV [J]. Telecommunications Science, 2014, 30(8): 23-27.

[作者简介]

钟翠（1981−），女，华为技术有限公司高级工程师兼高级营销经理，主要研究方向为新网络技术及解决方案。

王蕾（1985−），女，华为技术有限公司高级工程师，主要研究方向为数据中心网络技术和解决方案。

罗兴（1986−），男，华为技术有限公司工程师，主要研究方向为数据中心技术及解决方案。