文章编号:1007-5321(2017)01-0057-05 **DOI**:10.13190/j.jbupt.2017.01.010

软件定义数据中心网络研究

于 洋, 梁满贵, 王 哲
(北京交通大学 计算机与信息技术学院, 北京 100044)

摘要: 提出一种软件定义数据中心网络方案. 采用多控制器进行层次多域管控,引入向量地址作为数据交换标签,结合自主研发低造价交换设备和商业级服务器设计,并实现数据中心网络模型. 实验结果表明,该方案具有良好的实用性和可编程特性,有效地解决了数据中心网络扩展性差、运维成本高和管控方式复杂等问题,为构建未来绿色节能数据中心网络提供了重要思路.

关 键 词: 数据中心; 软件定义网络; 控制器; 可编程性; 向量地址; 流表
中图分类号: TP393 **文献标志码**: A

Software Defined Data Center Networks

YU Yang, LIANG Man-gui, WANG Zhe
(Department of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: With employing vector address as the switching label, a new type of data center network deployment scheme named software defined data center network was proposed. The restriction of the processing ability of controllers and the bandwidth for control channel was solved in the scheme by adopting hierarchical multiple controllers. A self-developed OpenFlow switch was applied combing with commercial servers in constructing the new data center network. It is shown that the scheme has good practicability and programmable features, and it effectively solves the problems such as poor scalability, high maintenance cost and complex management in data center network. The scheme also provides a valuable reference for the construction of green data center network.

Key words: data center; software defined network; controller; programmable; vector address; flow table

目前 OpenFlow 技术在流表的可扩展性和能耗等方面存在限制因素. 研究者提出一些改进策略,如内存能耗优化策略[1]、多级流表映射算法[2]和基于资源复用的流表存储优化方案[3]等. 但数据中心的业务复杂性和规模不断增大,上述方法不能从根本上解决流表存储容量受限问题. 为使数据中心向绿色节能和弹性演进,必须在组间级、设备级及网络架构级不断研究出新的方法与技术[4]. 为此,提出软件定义数据中心网络方案(SDDC,software defined data center),引入向量地址作为数据交换标签并采用多控制器进行层次多域管控,有效解决数据中心网络扩展性差、运维成本高和管控方式复杂等问题. 为未来数据中心网络建设提供必要的借鉴和参考.

1 向量地址和 SDDC 网络模型

1.1 向量地址

向量地址(VA,vector address)是一种新的网络地址编码技术[5],交换机为其所有端口按序依次分

收稿日期: 2016-05-04
基金项目: 国家重点基础研究发展计划(973 计划)项目(2011CB302203); 国家高技术研究发展计划(863 计划)项目(2007AA01Z203); 文化部民族民间文艺发展中心数字文化研究基地项目(136023522); 国家自然基金联合项目(U1636109)
作者简介: 于 洋(1987—), 女, 博士生, E-mail:12112075@bjtu.edu.cn; 梁满贵(1963—), 男, 教授, 博士生导师.

配一个本地序号(PI,port index). 将通信路径上的交换机的本地端口序号按序依次组合形成的序列形成向量地址. 以图1所示的网络为例详细说明VA. 网络包含端系统A、B、C、D和向量交换机(VS, vector switch)E、F、G、H、I、J. 设从$A\sim D$的通信路径为$A\rightarrow G\rightarrow I\rightarrow J\rightarrow D$,对应输出端口序号依次为$A$:1、$G$:2、$I$:4和$J$:2,将这些端口号依次组合形成的序列就是从$A\sim D$的向量地址Vad = 1242. 用二进制编码表示{1,10,100,10},整合为{11010010},该序列独立完整地标识了从$A\sim D$的一条通信路径. 其中,各交换机的地址位数可以不同,需根据端口数量提前配置到交换机中. 向量交换就是以向量地址作为数据交换地址的包交换过程.

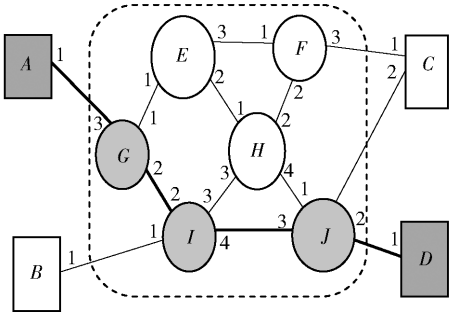

图1 VA转发原理

1.2 SDDC模型

1.2.1 SDDC网络架构

为研究和验证软件定义网络(SDN,software defined networks)和向量交换技术在数据中心网络建设和管理方面的作用,设计了软件定义数据中心网络方案. 如图2所示,系统由3部分组成,以下将逐一介绍.

1) SDDC管理器

用于实现整个数据中心业务逻辑功能,同时与SDN网络操作系统和虚拟机管理接口统一计算和存储网络资源,以保证计算资源与网络资源同步调度,其内部保存了每个租户虚拟机与网络资源信息的对应关系,为每个用户提供逻辑视图,且与其他网络租户相互隔离.

2) 虚拟机管理

用以实现资源管理、计算与存储和上报虚拟机创建、迁移等事件,其内部包含资源调度、高可用性和虚拟机迁移等模块.

3) SDN网络操作系统

用于实现全网络资源管理,通过OpenFlow协议控制数据面转发设备. 其软件部分包括NOS核心,由VS、OpenFlow交换机(OFS, OpenFlow switch)、OpenFlow控制器(OFC,OpenFlow controller)和向量控制器(VAC,vector address controller)组成. 其中OFC由OpenFlow全局控制器(OFGC, OpenFlow global controller)和OpenFlow局部控制器(OFLC, OpenFlow local controller)构成.

① VS

网络边缘使用OFS,进入网络核心后,采用VS负责数据的高速转发,省去使用三态内容寻址存储器(TCAM, ternary content addressable memory)带来的成本和能耗等问题. 当有新数据流出现时,仅需在两个边沿网关OFS上各增加一条流表项,位于核心网的VS无需增加任何表项存储和查找代价.

② OFS

作为边沿网关连接终端和VS,它的任务包括路由请求,收到新数据流后,向OFC发送路由请求,进行流表更新操作;流表查找,用以实现向量包与其他类型的包格式转化,如可提供IP、多协议标记交换技术(MPLS, multi-protocol label switching)等各种类型终端机的支持. 同OpenFlow标准相比,OFS仅在指令操作集中新添加数据包的封装与解封装操作.

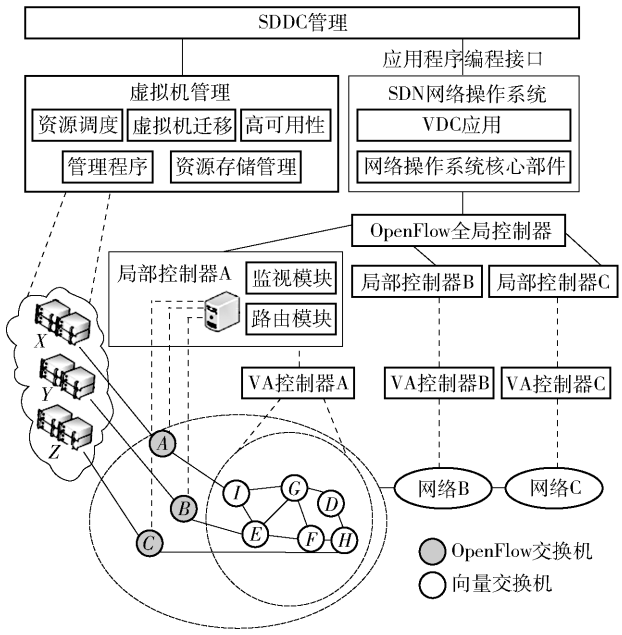

图2 SDDC网络架构

③ SDDC控制器

如图3所示, OFLC直接负责控制和响应OFS,通过向量通道与VAC通信,并在VAC的协助下进

行本域内网络拓扑收集和与路由计算. OFLC 通过 OpenFlow 通道添加新流表项至 OFS 中,流表项新添加对原始数据的封装与解封装功能. VAC 通过向量网控制面协议管理核心网 VS 及拓扑信息,并将搜集的信息上报至 OFLC 中,VS 结构简单,只需配合完成极少信令处理.

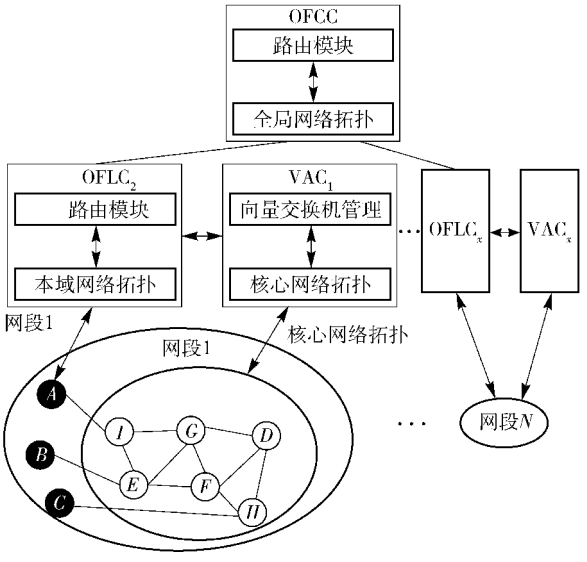

图 3　SDDC 控制面结构

1.2.2　SDDC 网络通信过程

在图 2 所示的 SDDC 网络中,以 X 和 Z 通信为例说明 SDDC 网络数据通信过程,假设通信路径为:$X\rightarrow A\rightarrow I\rightarrow E\rightarrow F\rightarrow H\rightarrow C\rightarrow Z$. 步骤说明如下:

① X 收到 Z 的数据包 Pkt;

② 收到数据包 Pkt 后,查找流表,若命中,则转至步骤⑥;否则,A 向 OFLC 发送路由请求;

③ OFLC 检查 A 和 C 是否在同一个网域,若不在同一个网域,则向 OFGC 发送路由请求,OFGC 启动路由模块功能,计算出通信路径,回应 OFLC 并将新的流表项发送至下层两个 OFLC,继而由 OFLC 下发流表至对应 OpenFlow 交换机中;否则,转至步骤④;

④ OFLC 发现 A 和 C 在同一网域,启动路由模块,并在 VAC 的协助下得到通信路径,回应 A 的路由请求并将新的流表项下发至 A 和 C 中;

⑤ A 接收 OFLC 发来的路由回应,并添加新的流表项至本地流表中,开始进行数据包 Pkt 转发操作;

⑥ A 根据匹配成功的流表项,添加向量包头和 VA 至 Pkt 中以实现对 Pkt 的封装过程,此时,变为向量包 Pktva(Pktva = Head + VA + pkt). 假设路由得到的路径序列是 $A\rightarrow I\rightarrow E\rightarrow F\rightarrow H\rightarrow C$,则将 Pktva 转发至向量交换机 I;

⑦ I 收到向量包 Pktva 后,按照向量交换方式转发至 E,依次类推,重复上述过程直至送达 C,VA 耗尽;

⑧ C 收到向量包 Pktva 后,查找流表,仅需去除向量包头即可完成解封装操作,转化为原始数据包 Pkt,并将其送达至 Z;

⑨ Z 接收数据,完成通信过程.

Z 到 X 的数据传输过程相同,仅方向相反,继而 X 与 Z 建立起相互信任的支持任意数据类型的通信.

2　实验比较

在 SDDC 网络模型中,采用多控制器进行层次多域管理,可有效地解决控制器处理能力和控制通道带宽的约束问题. 同时,采用向量交换技术对数据中心数据面进行改进,以下将分别从交换机路由和转发两方面分析 SDDC 网络优势.

2.1　SDDC 和 OpenFlow 路由比较

为比较 SDDC 和 OpenFlow 的路由,下面将统计在相同的路由信息和网络拓扑情况下两者在路由时所消耗的信令数,假设数据流都经过 n 个网络节点进行转发.

在 OpenFlow 网络中,接收数据包的 OFS 路由需要的信令数为 3(1 条 PACKET-IN 消息,1 条 Modify State Message 和 1 条 PACKET-OUT 消息),其他 $n-1$ 台 OFS 各需要信令数为 1(modify state message),因此,共消耗 $n+2$ 条信令数,与路径跳数 n 呈线性比例.

在 SDDC 网络模型中,分 2 种情况讨论:一种情况是边沿两台 OFS 在同一个网域时,则这两台 OFS 所需要的信令数是 4(1 条 PACKET-IN 消息,2 条 modify state message 和 1 条 PACKET-OUT 消息),其余 $n-2$ 台 VS 无需流表更新信令,共需信令数为 4;另一种情况是边沿两台 OFS 不在同一个网域内,这两台 OFS 所需要的信令数仍是 4,对应上层 OFLC 所需要的信令数共 3 条(1 条 PACKET-IN 消息和 2 条modify state message),此时所需要的信令总数是 7,与网络跳数无关. 可知 SDDC 与 OpenFlow 在路由更新所消耗的信令数之比是 4:(n +2)或 7:(n +2). 根据中国因特网网络跳数平均为 13[6],将

n 取值为 13，则 SDDC 所需要的信令数平均约占 OpenFlow 信令数的 36%. 这说明路由和流表更新过程得到大大简化.

2.2 OpenFlow 交换机与向量交换机的比较

网络现场可编程门阵列(NetFPGA, net field-programmable gate array)[7]是一个为研究者提供的低成本可重用的网络硬件平台，使得研究者可以在该硬件上搭建吉比特的高性能网络系统模型. 基于 NetFPGA 实验平台设计实现了 SDDC 交换机，可完成4个千兆接口的交换功能，其数据包转发处理模块如图4所示. 当接收到数据包后，首先进行包头解析，读取当前分量 PI，调度机制根据 PI 进行交换阵列调度，VA 处理结束后，数据包分别进入采用片上广播识别存取法(BRAM, broadcast recognition access method)实现的缓冲队列 0～3 中，这里不再使用片外动态随机存取存储器(DRAM, dynamic random access memory)和静态随机存储器(SRAM, static random access memory). 这样不仅解决了片外存取的速度限制问题，也降低了资源开销和成本.

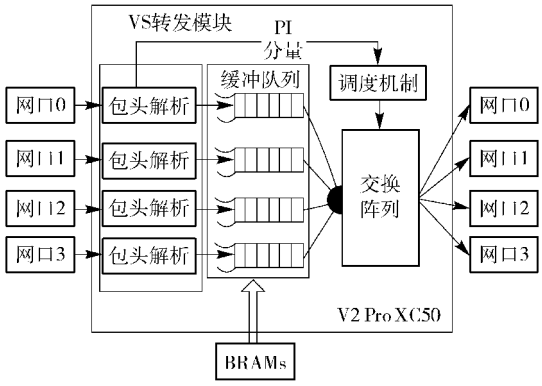

图4 基于 NetFPGA 的 VS 结构

该研究基于 NetFPGA 的交换机，并对硬件资源消耗做出了详细统计分析，其中包括 LUT、Slices 和片外存储等，同时，将 VS 与 OFS、IPv4 Router 和以太网交换机(Ethernet switch)等参考设计进行实验比较，对比结果如表1所示.

上述统计结果可以看出，将 VS 用于大规模数据中心网络数据转发中具有以下优势：VS 片内资源消耗非常少，DFF 和 LUT 等资源消耗平均仅为 OFS 的43%；VS 无需 SRAM 片外存储，而 OFS 需要使用 SRAM 作为流表存储器，且在流表更新时需外部设备互连(PCI, peripheral component interconnect)总线支持，降低了流表更新速率；VS 采用片内 BRAM 存储，而 OFS 采用片外 DRAM 缓冲数据包，对片外 DRAM 占用偏高.

表1 基于 NetFPGA 的几种交换机硬件资源统计

硬件资源	VS	OFS	Ethernet Switch	IPv4 Router
Slices	1 868	6 806	——	——
LUTs	2 690	10 524	3 786	14 398
DFFs	2 121	7 326	984	3 712
BRAMs	26	13	92	16
片外存储	无	DRAM、SRAM	——	DRAM、SRAM

3 结束语

针对现代数据中心网络高带宽、大流量和绿色节能等需求，将 SDN 技术引入其中，采用多控制器进行层次多域管理，不仅提高了控制器处理能力和路由效率，而且有效解决了控制通道带宽和控制面负载能力限制问题. 在 OpenFlow 研究成果上，将向量地址作为数据交换标签，解决了流表存储容量、能耗及成本等限制问题，提出软件定义数据中心网络模型，为未来数据中心网络建设提供必要的借鉴和参考，有利于推动未来数据中心网络新协议、新业务和新功能的快速实现与部署.

参考文献：

[1] 彭宏玉, 陈刚, 张英海, 等. SDN 架构下数据中心内存能耗优化策略[J]. 北京邮电大学学报, 2015, 38(2): 78-82.
Peng Hongyu, Chen Gang, Zhang Yinghai, et al. The memory energy optimization strategy in data center based on SDN technology[J]. Journal of Beijing University of Posts and Telecommunications, 2015, 38(2): 78-82.

[2] 刘中金, 李勇, 苏厉, 等. TCAM 存储高效的 OpenFlow 多级流表映射机制[J]. 清华大学学报(自然科学版), 2014, 54(4): 437-442.
Liu Zhongjin, Li Yong, Su Li, et al. The research of OpenFlow multistage flow table mapping mechanism based on TCAM[J]. Journal of Tsinghua University (Natural Science), 2014, 54(4): 437-442.

[3] Li Xiangwen, Ji Meng, Cao Min, et al. The optimization of OpenFlow flow table based on resource reuse[J]. Optical Communication Research, 2014, 40(2): 8-11.

[4] 马文婷. 基于 OpenFlow 的 SDN 控制器关键技术研究[D]. 北京: 北京邮电大学, 2015.

[5] Zhao Aqun, Liang Mangui. A new forwarding address for next generation networks[J]. Frontiers of Information Technology and Electronic Engineering, 2012, 13(1): 1-10.

[6] 赵阿群, 梁满贵, 廉松海, 等. 向量地址平均长度研究[J]. 高技术通讯, 2011, 21(12): 1246-1251.
Zhao Aqun, Liang Mangui, Lian Songhai, et al. The research in the average length of the vector address[J]. High Technology Letters, 2011, 21(12): 1246-1251.

[7] Naous J, Erickson D, Covington G A, et al. Implementing an OpenFlow switch on the NetFPGA platform[C]// ACM/IEEE Symposium on Architecture for Networking and Communications Systems. 2008: 1-9.

(上接第 35 页)

[5] Pan Xin, Xu Hanchen, Song Jie, et al. Capacity optimization of battery energy storage systems for frequency regulation[C]// IEEE International Conference on Automation Science and Engineering. Goteborg: IEEE, 2015: 1139-1144.

[6] Si Jennie, Wang Yutsung. On-line learning control by association and reinforcement[J]. IEEE Transactions on Neural Networks, 2001, 12(2): 264-276.

[7] Schaltz E, Khaligh A, Rasmussen P O. Influence of battery/ultracapacitor energy-storage sizing on battery lifetime in a fuel cell hybrid electric vehicle[J]. IEEE Transactions on Vehicular Technology, 2009, 58(8): 3882-3891.

[8] Houari A, Abbes D, Labrunie A, et al. Hybridization of electrical energy storage for intelligent integration of photovoltaics in electric networks[C]// European Conference on Power Electronics and Applications. Geneva: IEEE, 2015: 1-10.