软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(2):295−316 [doi: 10.3724/SP.J.1001.2013.04336]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

数据中心网络的体系结构*

魏祥麟[1,2], 陈 鸣[2], 范建华[1], 张国敏[2], 卢紫毅[1]

[1](南京电讯技术研究所,江苏 南京 210007)

[2](解放军理工大学 指挥信息系统学院,江苏 南京 210007)

通讯作者: 魏祥麟, E-mail: wei_xianglin@ieee.org

摘 要: 在新的应用模式下,传统层次结构数据中心网络在规模、带宽、扩展性和成本方面存在诸多不足.为了适应新型应用的需求,数据中心网络需要在低成本的前提下,满足高扩展性、低配置开销、健壮性和节能的要求.首先,概述了传统数据中心网络体系结构及其不足,并指出了新的需求;其次,将现有方案划分为两类,即以网络为中心和以服务器为中心的方案;然后,对两类方案中的代表性结构进行了详细的综述和对比分析;最后指出了数据中心网络未来的发展方向.

关键词: 数据中心;网络;体系结构;拓扑;路由

中图法分类号: TP393 **文献标识码**: A

中文引用格式: 魏祥麟,陈鸣,范建华,张国敏,卢紫毅.数据中心网络的体系结构.软件学报,2013,24(2):295−316. http://www.jos.org.cn/1000-9825/4336.htm

英文引用格式: Wei XL, Chen M, Fan JH, Zhang GM, Lu ZY. Architecture of the data center network. Ruanjian Xuebao/Journal of Software, 2013,24(2):295−316 (in Chinese). http://www.jos.org.cn/1000-9825/4336.htm

Architecture of the Data Center Network

WEI Xiang-Lin[1,2], CHEN Ming[2], FAN Jian-Hua[1], ZHANG Guo-Min[2], LU Zi-Yi[1]

[1](Nanjing Telecommunication Technology Research Institute, Nanjing 210007, China)

[2](College of Command Information Systems, PLA University of Science and Technology, Nanjing 210007, China)

Corresponding author: WEI Xiang-Lin, E-mail: wei_xianglin@ieee.org

Abstract: Under the new application mode, the traditional hierarchy data centers face several limitations in size, bandwidth, scalability, and cost. In order to meet the needs of new applications, data center network should fulfill the requirements with low-cost, such as high scalability, low configuration overhead, robustness and energy-saving. First, the shortcomings of the traditional data center network architecture are summarized, and new requirements are pointed out. Secondly, the existing proposals are divided into two categories, i.e. server-centric and network-centric. Then, several representative architectures of these two categories are overviewed and compared in detail. Finally, the future directions of data center network are discussed.

Key words: data center; network; architecture; topology; route

信息服务的集约化、社会化和专业化发展使得因特网上的应用、计算和存储资源向数据中心迁移.商业化的发展促使了承载上万甚至超过 10 万台服务器的大型数据中心的出现.截至 2006 年,Google 在其 30 个数据中心拥有超过 45 万台服务器,微软和雅虎在其数据中心的服务器数量也达到数万台[1].随着规模的扩大,数据中心不仅承载了传统的客户机/服务器应用,而且承载了包括 GFS 和 MapReduce 在内的新应用[2].这种趋势一方面突

* 基金项目: 国家自然科学基金(61070173, 61103225, 61201216); 国家重点基础研究发展计划(973)(2012CB315806); 中国博士后科学基金(201150M1512); 江苏省自然科学基金(BK2010133); 国防科技重点实验室基金(9140C020302110C0206)

收稿时间: 2012-08-09; 定稿时间: 2012-10-19; jos 在线出版时间: 2012-11-23

CNKI 网络优先出版: 2012-11-23 12:17, http://www.cnki.net/kcms/detail/11.2560.TP.20121123.1217.006.htm

出了数据中心作为信息服务基础设施的中心化地位,同时也凸显了传统分层数据中心在面临新应用和计算模式时的诸多不足.

在新计算模式面前,分层数据中心的主要不足包括:服务器到服务器连接和带宽受限、规模较小、资源分散、纵向扩展成本高、路由效率低、配置开销较大、不提供服务间的流量隔离和网络协议待改进等.这些问题使其难以满足日益发展应用的需求,并在近几年得到了广泛关注.为了适应新型应用的需求,新型数据中心网络需要满足的要求包括:大规模、高扩展性、高健壮性、低配置开销、服务器间的高带宽、高效的网络协议、灵活的拓扑和链路容量控制、绿色节能、服务间的流量隔离和低成本等.这些要求也是当前研究的目标和出发点.

目前,关于数据中心网络结构的研究可以分为两类:网络为中心的方案和服务器为中心的方案.在网络为中心的方案中,网络流量路由和转发全部是由交换机或路由器完成的.这些方案大多通过改变现有网络的互联方式和路由机制来满足新的设计目标.在服务器为中心的方案设计中,采用迭代方式构建网络拓扑,服务器不仅是计算单元,而且也充当路由节点,会主动参与分组转发和负载均衡.

本文首先指出了现有分层结构存在的不足并分析了针对新型数据中心网络的需求,然后系统分析了当前具有代表性的多种数据中心网络方案,进行了综述和综合对比,最后对今后的研究方向进行了展望和探讨.

本文第 1 节介绍数据中心网络的基本结构和新的需求.第 2 节和第 3 节分别系统地讨论网络为中心和服务器为中心的多种具有代表性的方案,并指出仍然存在的问题.第 4 节综合分析数据中心网络的各种方案,进行对比,并展望未来的研究方向.第 5 节总结全文.

1 数据中心网络:基本结构及新的需求

1.1 传统数据中心网络体系结构

数据中心网络(data center network,简称 DCN)是指数据中心内部通过高速链路和交换机连接大量服务器的网络[1].传统数据中心网络主要采用层次结构实现,且承载的主要是客户机/服务器模式应用.多种应用同时在同一个数据中心内运行,每种应用一般运行在其特定的服务器/虚拟服务器集合上.每个应用与一个或者多个因特网可路由的 IP 地址绑定,用于接收来自因特网的客户端访问.在数据中心内部,来自因特网的请求被负载均衡器分配到这个应用对应的服务器池中进行处理.根据传统负载均衡的术语,接受请求的 IP 地址称为虚拟 IP 地址(virtual IP address,简称 VIP),负责处理请求的服务器的集合称为直接 IP 地址(direct IP address,简称 DIP).一个典型的传统数据中心网络体系结构如图 1 所示[1,3].

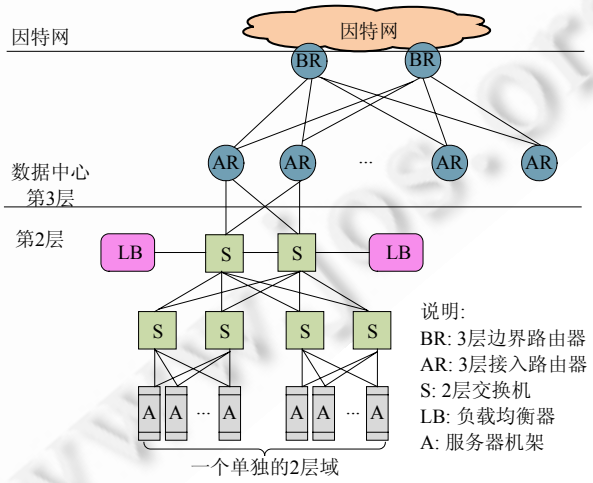

Fig.1 Traditional architecture of the DCN

图 1 传统的数据中心网络层次体系结构

在图 1 中,根据其 VIP 地址,来自因特网的请求通过 3 层的边界路由器(BR)和接入路由器(AR)被路由到 2 层域.应用对应的 VIP 地址被配置在图 1 中连接在上层交换机(S)的负载均衡器(LB)中.对于每个 VIP,负载均衡器为其配置了一个 DIP 列表,这个列表包含的通常是服务器(A)的内部私有地址.根据这个列表,负载均衡器将接收到的请求分配到 DIP 对应的服务器池中进行处理.

1.2 传统数据中心网络体系结构不足

面向云计算等新型计算模式的数据中心已经不同于企业数据中心[4],它呈现出许多新的特点:

(1) 规模不断扩大,需要支持的服务器数量达到数十万或更高的量级.

(2) 在流量特征方面,MapReduce 应用、虚拟机迁移以及其他带宽密集型应用等触发的数据中心内部流量显著增加,达到总流量的 80%左右,从而使得网络带宽经常成为稀缺资源[1].

(3) 出于成本考虑,数据中心规模的急剧扩大要求其采用成熟的普通商业化(commodity)网络设备达到横向扩展,而不是采用昂贵的高性能设备进行纵向扩展[1,3,5,6].

(4) 一些新型数据中心网络结构具有不同于传统网络的结构,比如立方体[7,8]、随机图[9]、无标度网络[10,11]、多根树[5]等,这些网络结构可以用来辅助设计高效的路由算法.

(5) 为了保证服务质量和安全性,需要为各个服务的流量提供一定的流量隔离[12,13].

(6) 虚拟化已经成为数据中心的重要理念,其需要数据中心网络支持任意一个虚拟机的任意迁移和部署,且不影响已经存在应用层状态.

(7) 在大量服务器和交换机存在的前提下,数据中心网络不应该引入过多的交换机配置开销,需要做到即插即用.

(8) 一方面,数据中心网络已经成为全球能耗不可忽视的部分;另一方面,数据中心的超过 80%的链路负载非常轻[14,15].

(9) 由于数据中心网络中广泛采用了低成本的低端设备,从而存在链路失效、服务器失效和交换机失效等多种故障和差错,而持续的可靠服务能力需要数据中心网络提供高效的失效恢复策略和容错机制[1,12,16].

根据这些特点以及图 1 展示的网络体系结构,可以总结出传统层次数据中心网络体系结构的不足主要包括如下的几个方面:

(1) 服务器到服务器的连接和带宽受限.层次体系结构意味着隶属不同 2 层域的服务器间的通信流量需要经过 3 层.但出于成本考虑,2 层~3 层的链路经常是超额认购(oversubscription)的,也就是说,接入路由器与边界路由器的链路容量显著低于连接到接入路由器的服务器的总输出容量.这就导致了隶属不同 2 层域的服务器间的可用带宽非常受限(取决于 3 层的超额认购比值以及流量分布情况).

(2) 规模较小.如图 1 所示,所有连接到 1 对接入路由器的服务器构成单个 2 层域.如果使用传统的网络体系结构和协议,受限于快速失效恢复的需要,单个 2 层域的规模约为 4 000 个服务器.由于广播流量(由 ARP 等引起)的开销的限制,2 层域大多被配置于 2 层交换机上的 VLAN 划分为子网.这个规模难以用来构建十万级甚至百万级规模的数据中心.

(3) 资源分散.目前流行的负载均衡技术,比如目的 NAT(半 NAT)和直接服务器返回(direct server return)等,要求所有 VIP 的 DIP 池位于同一个 2 层域.这个限制意味着应用不能使用其他 2 层域的服务器,这导致了资源的分散和较低的资源利用率.通过源 NAT 或者全 NAT 的负载均衡允许服务器分散在 2 层域.但这种情况下,服务器通常不能看到客户端的 IP,这使得服务器无法利用客户端的 IP 地址信息提供个性化服务或者进行数据挖掘类的工作,这对于服务器来说是无法接受的.

(4) 采用专用硬件纵向扩展,成本高.在传统的体系结构中,负载均衡器成对使用.当负载变得太大时,运营商使用新的拥有更大容量的均衡器代替现有的均衡器,这个纵向扩展的策略成本很高[3].另外,3 层的路由器在超额认购比例发生变化或者拓扑发生变化时,需要纵向升级到更昂贵的路由器,而不是横向升级.由于目前的高端设备与普通商业交换机/路由器的价格差别巨大,因此这种策略的升级成本

高昂.

(5) 流量工程难度大.数据中心流量是高动态和突发的,约 80%的流量都是内部流量[14,17,18],且流量持续变化,难以预测,从而使得传统流量工程方法无法有效工作.

(6) 自动化程度不高[4].当服务需要在服务器间重分配时,传统数据中心网络的地址空间分片会导致巨大的人工配置成本,且人工操作出错的概率很高[6].在云服务数据中心网络中,提高自动化程度可以控制 IT 员工与服务器成本的比值,并且能够降低由于员工操作失误带来的风险,使得网络更加健壮.

(7) 配置开销大.3 层结构需要为每个交换机配置子网信息,并同步 DHCP 服务器以基于主机的子网分配 IP 地址[12];另外,VIP 与 DIP 列表的对应关系需要配置在负载均衡器上.当交换机或者网络设备故障或 VIP 与 DIP 对应关系发生变化时,引入的配置开销较大,同时增加了误操作的风险.

(8) 不提供服务间的流量隔离.在数据中心承载多个服务的同时,网络并没有阻止一个服务的流量影响其周围的其他服务.当一个服务经历了流量洪泛时,位于其同一个子树的其他服务就会承担洪泛引起的伤害[12].

(9) 网络协议待改进.在数据中心中大量沿用网络传输协议(TCP),而原有的 TCP 协议是面向互联网开发的,没有考虑数据中心网络的低时延和高带宽特点.研究发现,在多对一通信模式时,TCP 会出现链路利用率不足的吞吐量崩溃现象[19,20].

1.3 数据中心网络的新需求

为了满足新型计算模式和应用的需求,新型数据中心网络需要满足如下的要求:

(1) 服务器和虚拟机的便捷配置和迁移.允许部署在数据中心的任何地方的任何服务器作为 VIP 服务器池的一部分[3],使得服务器池可以动态地缩减或扩展;并且,任意虚拟机可以迁移到任何物理机.迁移虚拟机时无须更改其 IP 地址,从而不会打断已经存在的应用层状态.

(2) 服务器间的高传输带宽.多数数据中心应用的服务器间的流量总量远大于数据中心与外部客户端的流量.这意味着体系结构应该提供任意对服务器间尽可能大的带宽[3].

(3) 支持数十万甚至上百万台的服务器.大的数据中心拥有数十万级或者更多的服务器,并允许增量的部署和扩展[3].

(4) 低成本且高扩展[1]:第一,物理结构可扩展,需要以较小的成本物理连接服务器,不依赖于高端交换机来纵向扩展,而是采用普通商业化的组件实现横向扩展;第二,可以通过增加更多的服务器到现有结构上实现增量扩展,且当添加新的服务器时,现有的运行的服务器不受影响;第三,协议设计可扩展,比如路由协议可扩展.

(5) 健壮性.数据中心网络必须能够有效地处理多种失效,包括服务器失效、链路断线或者机架失效等.

(6) 低配置开销.网络构建不应引入过多的人工配置开销.理想情况下,管理员在部署前无须配置任何交换机[16].

(7) 高效的网络协议.根据数据中心结构和流量特点,设计高效的网络协议.

(8) 灵活的拓扑和链路容量控制.数据中心网络流量是高动态和突发的,从而使得网络中某些链路由于超额认购产生拥塞,成为瓶颈链路,而很多其他链路则负载很轻[14,18],因此,网络需要能够灵活地调配负载或者灵活地调整自身拓扑和链路容量,从而适应网络流量变化的需求[21].

(9) 绿色节能.新一代数据中心在当今能源紧缺与能源成本迅猛增长的情况下需要综合考虑能源效率问题,提高利用率,降低流量传输和制冷开销.

(10) 服务间的流量隔离.某服务的流量不应被其他服务的流量所影响,从而提供安全性和服务质量保证.

根据这些需求,研究者从多个角度出发,提出了多种设计方案.由于每个工作的出发点不同,为了能够清晰地描述,本文着重综述数据中心网络体系结构方面的研究,并分析其存在的问题.为了便于描述,将目前的研究划分为网络为中心的方案和服务器为中心的方案两类,并在第 2 节和第 3 节分别进行描述.

2 网络为中心的方案

在网络为中心的方案中,网络流量路由和转发全部是由交换机和路由器完成的.这些方案大多通过改变现有网络的互联方式和路由机制来满足新的设计目标.相关方案主要包括 FatTree[5],ElasticTree[22],Monsoon[3],VL2[12],PortLand[16],SecondNet[13]和 Jellyfish[9]等.一些研究者引入光交换技术到数据中心网络,提出了 Helios[23],DOS[24],c-Through[25],OSA[21]以及光电混合交换方案[26].另一些研究者则将无线技术引入数据中心,提出了 Flyway[18]和 WDCN[27,28]等技术.王聪等人也提出了一种低成本高连通性的数据中心网络体系结构[29].本节将选择具有代表性的方案进行综述.

2.1 FatTree

2.1.1 拓扑构建

Al-Fares 等人借鉴 Charles Clos 等人 50 年前在电话网络领域中的做法,提出使用 FatTree 来互联以太网交换机[5].FatTree 结构中交换机分为 3 层:核心交换机、聚合交换机和边缘交换机.一个简单的 FatTree 的结构如图 2 所示.图 2 展示的是一棵 k(=4)叉树:FatTree.图中有 4 个 Pod,每个 Pod 中包含 k 个交换机,其中 $k/2$ 个是边缘交换机,$k/2$ 个是聚合交换机.并且边缘交换机有 k 个端口,其中,$k/2$ 个连接到端主机,$k/2$ 个连接到聚合交换机.聚合交换机的 $k/2$ 个端口连接到边缘交换机,另外 $k/2$ 个连接到核心交换机.有$(k/2)^2$个核心交换机.核心交换机的 k 个端口分别连接到 k 个 Pod 的聚合交换机.

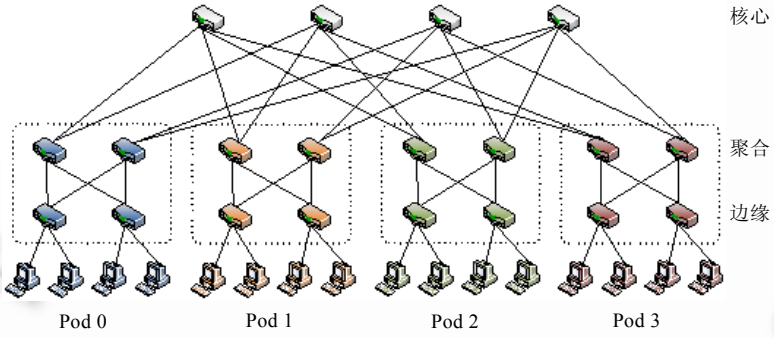

Fig.2 Topology of a simple FatTree network

图 2 一个简单的 FatTree 网络拓扑结构

2.1.2 地址配置

Al-Fares 等人在私有的 10.0.0.0/8 块内设置网络中的所有 IP 地址.Pod 交换机的地址形式是 10.*pod*.*switch*.1,其中,*pod* 代表 Pod 号(属于[0,k–1]),*switch* 代表交换机在 Pod 中的位置(属于[0,k–1],从左到右,自底向上).核心交换机地址形式为 10.k.j.i,其中,j 和 i 代表了交换机在$(k/2)^2$个核心交换机网格中的坐标(从左上角开始).主机的地址跟随其连接到的 Pod 交换机,主机的地址形式是 10.*pod*.*switch*.*ID*,其中,*ID* 是主机在子网中的位置(在$\lceil 2, k/2+1 \rceil$之间,从左到右).因此,每个底层交换机负载一个/24 子网的 $k/2$ 个主机.

2.1.3 两级路由表和路由

为了使 Pod 间的流量尽可能均匀地分布于核心交换机,FatTree 实现了两级路由表以允许两级前缀查询.一些路由表的表项会有个额外的指针到一个二级路由表(*suffix*,*port*)项.图 3 展示了一个两级路由表的例子,其中,左上角是一级路由表,右下角是二级路由表.一级路由前缀如果不包含任何二级前缀,这个表项就称为终结性表项,比如图 3 中一级路由表的前两个表项.一级路由表是左匹配的,二级路由表则从右开始匹配.图 3 中, 0.0.0.0/0 路由表项拥有二级路由表,包含两个表项,分别对应输出端口 2 和输出端口 3.

前缀	输出端口
10.2.0.0/24	0
10.2.1.0/24	1
0.0.0.0/0	

前缀	输出端口
0.0.0.2/8	2
0.0.0.3/8	3

Fig.3 Two-Level route table of FatTree

图 3 FatTree 两级路由表

FatTree 的任意两个不同 Pod 主机之间存在 k 条路径,从而提供了更多的路径选项,并且可以将流量在这些路径之间进行分散.任意给定 Pod 的低层和高层交换机对位于本 Pod 的任意子网都有终结性表项.因此,如果一个主机发送一个分组到同 Pod 的不同子网的另一个主机,那个 Pod 的所有的高层交换机会拥有指向目的子网交换机的终结性表项.

对于 Pod 间流量,Pod 交换机有一个拥有二级表的/0 前缀匹配主机的 ID.使用目的主机 ID 作为确定性熵的源,这会使流量均匀地分布于所有的核心交换机[5].在核心交换机,为所有网络的 ID 分配终结性的一级表项指向包含那个网络的合适的 Pod.一旦一个分组到达核心交换机,仅有 1 条链路到目的 Pod,那个交换机就会为这个分组的 Pod(10.*pod*.0.0/16,port)包含一个终结性的"/16"前缀.一旦一个分组到达了目的 Pod,接收的高层 Pod 交换机会包含一个(10.*pod*.*switch*.0/24,*port*)前缀,从而将分组指向其目的子网交换机.目的子网交换机最终将分组发送到目的主机.

2.1.4 ElasticTree

由于数据中心流量的突发性,Heller 等人认为,FatTree 为所有的端主机之间提供完全带宽(full bandwidth)是没有必要的,并从节能的角度出发提出了 ElasticTree[22].其主要出发点是:关掉不用的链路和交换机,仅在需要时才开启.其系统结构如图 4 所示,其中,优化器(optimizer)负责找到满足当前流量条件的最小能量网络子集;其输入是拓扑、流量矩阵、每个交换机的电量模型和期望的容错属性.优化器输出一个活动组件的集合给电量控制(power control)和路由(routing)模块.电量控制模块切换端口、板卡和整个交换机的能量状态.路由模块则为所有的流选择路径,然后将路由放入网络.在图 2 所示的 FatTree 拓扑中,当每个主机仅有 0.2Gbps(主机网卡容量为 1Gbps)流量跨越核心交换机时,ElasticTree 的拓扑如图 5 中的实线所示.

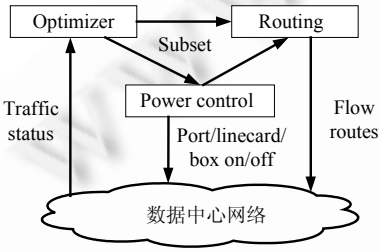

Fig.4 Architecture of ElasticTree

图 4 ElasticTree 体系结构

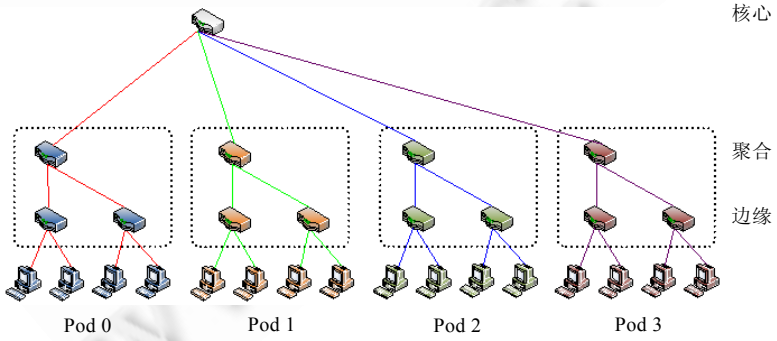

Fig.5 Topology of a simple ElasticTree network

图 5 一个简单的 ElasticTree 网络拓扑

2.1.5 FatTree 和 ElasticTree 分析

与传统层次结构相比,首先,FatTree 结构消除了树形结构上层链路对吞吐量的限制,并能为内部节点间通信提供多条并行链路;然后,其横向扩展的尝试降低了构建数据中心网络的成本;最后,FatTree 结构与现有数据中心网络使用的以太网结构和 IP 配置的服务器兼容.

但是,FatTree 的扩展性受限于核心交换机端口数量,目前比较常用的是 48 端口 10GB 核心交换机,在 3 层树结构中能够支持 27 648 台主机,因此,FatTree 存在扩展性不足的缺点[6].FatTree 的另一个缺点是容错性不够,具体表现为处理交换机故障能力不足以及路由协议容错性不强.研究结果表明[30],FatTree 对边缘交换机故障非常敏感,严重影响系统性能.因为 FatTree 仍然是树结构,本质上具有树结构的缺陷.另外,构建 FatTree 需要的交换机数量为 $5N/n$,其中,N 是服务器的数量,n 是交换机的端口数量.当 n 较小时,连接 FatTree 需要的交换机数量庞大,从而增加了布线和配置的复杂性.最后,FatTree 的流量调度依赖于中心服务器进行,这限制了其扩展性,面临单点失效的问题.

ElasticTree 基于 FatTree 构建,但是考虑了节能的问题,在轻载时降低了数据中心能耗,从而降低了成本.但它没有根本上调整网络结构,从而面临与 FatTree 相似的问题.

2.2 Monsoon

Monsoon 体系结构如图 6 所示[3],其中,所有的服务器连接到一个共同的 2 层网络,并且网络中没有超额认购链路,这意味着任何服务器都可以与其他服务器的网络接口以 1Gbps(假定主机网卡速度为 1Gbps)的速度进行通信.Monsoon 网络的 3 层部分将数据中心连接到因特网,使用等耗费多路径(equal cost MultiPath,简称 ECMP)将从因特网收到的请求均匀地分配到所有的路由器上.请求到达 2 层域后,接入路由器使用一致哈希将针对每个 VIP 的请求均匀分布到负载均衡器上.最后,负载均衡器使用特定的负载均衡函数将请求均匀分布到服务器池上.

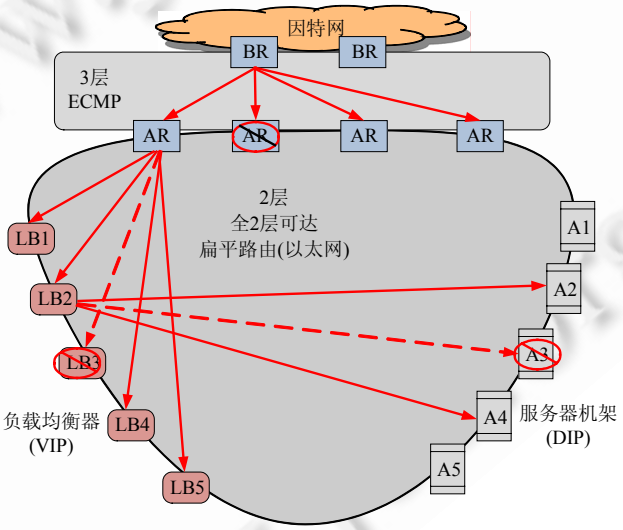

Fig.6 Architecture of Monsoon

图 6 Monsoon 体系结构

Monsoon 在 2 层商业以太网交换机构成的网络上使用 VLB(valiant load balancing)实现了无热点的支持任意流量模式的核心基础设施.Monsoon 利用扁平地址空间构建了单个可以支持 10 万台服务器的 2 层域,从而解决了资源分散的问题.另外,Monsoon 使用造价低廉的硬件负载均衡硬件将服务的请求分散到服务器池中,从而进一步消除了潜在的热点主机.

为了实现 Monsoon 的设计目标,Monsoon 使用了 MAC-in-MAC 技术,实现了 MAC 层隧道.另外,禁用了传统的 ARP 功能,将其替换为一个用户态进程.增加一个新的称为封装器的 MAC 接口来封装输出的以太网帧.这使得 Monsoon 与传统的使用以太网的网络不能直接兼容.

Monsoon 还采用控制平面来维护交换机的转发表,并维护目录服务,用于跟踪服务器连接到的交换机及其 IP 和 MAC 地址.在收到关于 IP 的查询时,目录服务返回这个 IP 地址对应的多个 MAC 地址及其使用的 VLB 中继交换机的集合.这使得控制平面的正常工作对 Monsoon 极为重要,面临单点故障.

2.3 VL2

图 7 是一个开关网络(clos network)的例子,也是 VL2 网络结构的示意图.与传统拓扑相似,机架(top of rack,简称 ToR)交换机连接到两个聚合交换机.但是每两个聚合交换机都可以通过中继交换机互联,从而形成了大量的可能路径.这意味着如果有 n 个中继交换机,任何一个的失效仅会减小 $1/n$ 的双向带宽,这增加了路径数量和健壮性.构建一个没有超额认购的开关网络很容易且成本较低,比如在图 7 中,如果使用拥有 D_A 个端口的聚合交换机和拥有 D_I 个端口的中继交换机,那么每层间的容量是 $D_ID_A/2$ 乘以链路容量.

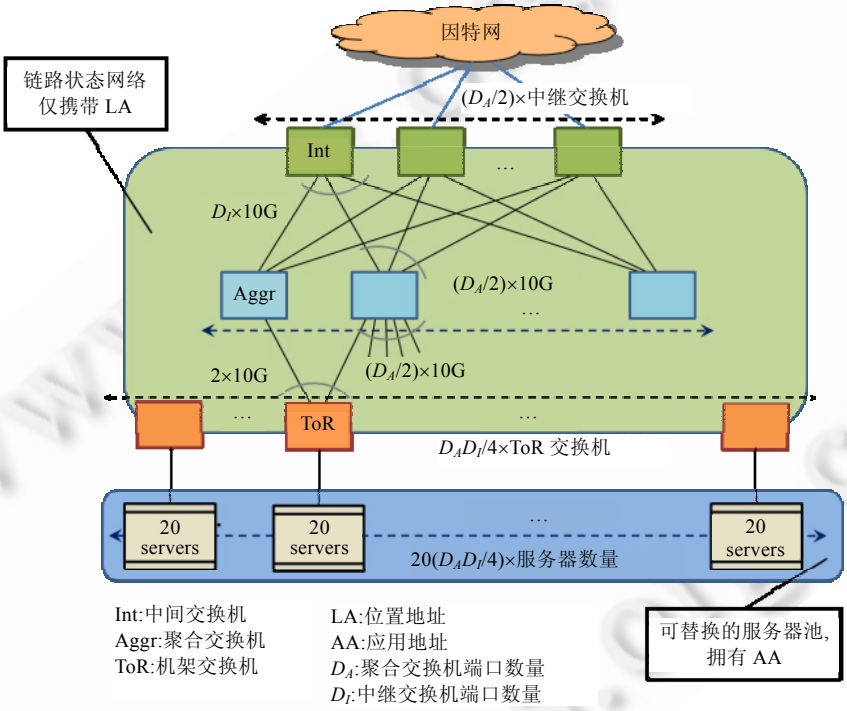

Fig.7 Structure of VL2 network

图 7 VL2 网络结构

这个开关网络拓扑特别适合 VLB,因为通过在网络顶层的一个中继交换机间接转发流量,网络可以为任何服从软管模型(hose model)的流量矩阵提供带宽保证.同时,路由很简单且富有弹性,采用一个随机路径到达一个随机中继交换机,然后沿一个随机路径到达目的 ToR 交换机.

在 VL2 中,IP 地址仅仅作为名字使用,没有拓扑含义.VL2 的寻址机制将服务器的名字与其位置分开.VL2 使用可扩展的、可靠的目录系统来维持名字和位置间的映射.当服务器发送分组时,服务器上的 VL2 代理开启目录系统以得到实际的目的位置,然后将分组发送到目的地.VL2 代理也辅助消除由 2 层网络 ARP 产生的扩展性问题.因此,VL2 同样依赖于中心化的基础设施来实现 2 层语义和资源整合,面临单点失效和扩展性问题.

2.4 PortLand

以 FatTree 网络结构为基础,Mysore 等人设计了一种可扩展和容错的 2 层路由和转发协议,称为 PortLand[16].PortLand 引入了基本结构管理者(fabric manager),并使用层次伪 MAC 地址(pseudo MAC address,简称 PMAC)进行分组转发:

- 基本结构管理者.PortLand 设置了一个逻辑上中心化的基本结构管理者,其维护诸如拓扑之类的网络配置信息的软状态.管理者是一个运行于特定主机上的用户态进程,负责辅助 ARP 解析、容错和多播.管理者可以是冗余连接的主机,也可以运行于一个单独的控制网络上.
- 层次伪 MAC 地址.PortLand 为每个端主机分配一个唯一的 PMAC 地址.PMAC 将端主机在拓扑中的位置进行编码.比如,所有的位于同一个 Pod 的端节点的 PMAC 拥有相同的前缀.通过维持 PMAC 到 AMAC 的映射,端主机无须修改,仍然认为自身维持其真实的 MAC(actual MAC,简称 AMCA).为了得到目的主机的 MAC 地址,主机会发起 ARP 请求以得到目的主机的 PMAC 地址.得到目的主机 PMAC 后,所有的分组转发都基于 PMAC 进行,从而使得转发表很小.出口交换机进行 PMAC 和 AMAC 头的重写,以使目的主机维持未修改的 MAC 地址.

PortLand 的边缘交换机学习唯一的 pod 号以及 pod 里的一个唯一的位置号.交换机使用位置发现协议来分配这些值.对于所有直接相连的主机,边缘交换机为其分配一个 48 比特的 PMAC,形式是 *pod.position.port.vmid*,其中,*pod*(16bit)反映了这个边缘交换机的 pod 号,*position*(8bit)是交换机在 pod 中的位置,*port*(8bit)是这个主机连接到的端口号,*vmid*(16bit)用来将位于同一个主机的多个虚拟机进行多路表示.边缘交换机为每个在给定端口观察到的新 MAC 地址的 vmid 递增编号.PortLand 会设定 vmid 超时,如果在一定时间内没有任何来自于某 vmid 的流量,会重用这个 vmid.

图 8 展示了建立 PMAC 与 AMAC 的映射关系的一个例子.

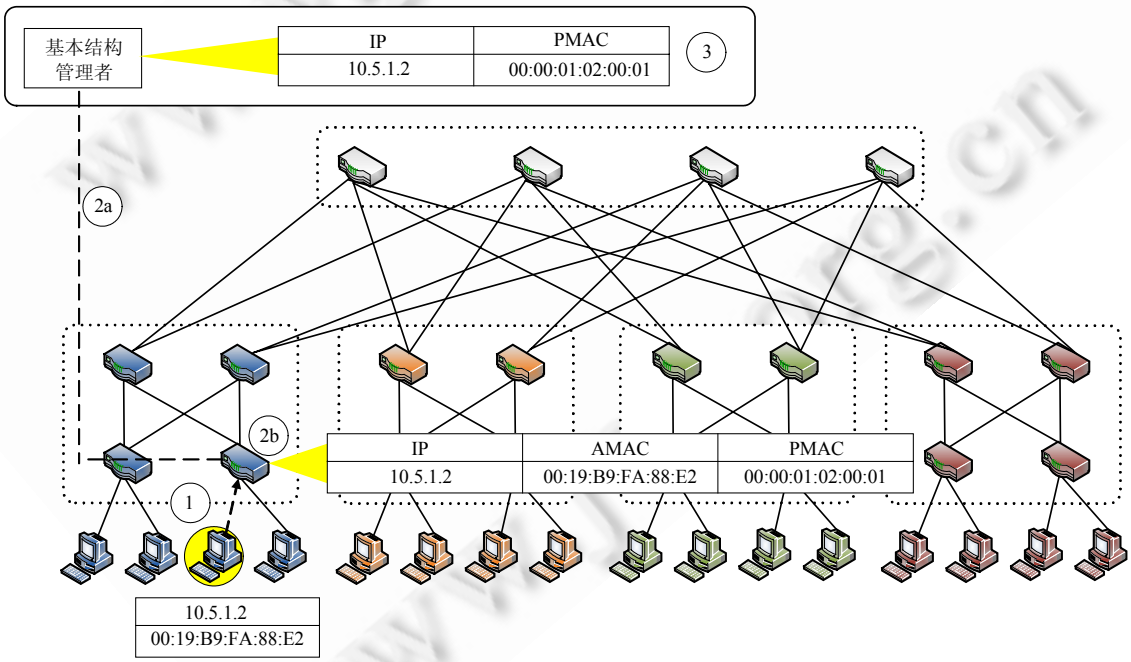

Fig.8 AMAC to PMAC mapping

图 8 AMAC 与 PMAC 的映射

在图 8 中,第 1 步,位于最左边 Pod 的一台主机发送分组到其直接连接的入口交换机(ingress switch);第 2 步,

入口交换机为其分配 PMAC 地址,并将其 IP、AMAC 和 PMAC 的映射关系放入转发表(如图 8 中的 2a 步所示),并发送这个映射到基本结构管理者(如图 8 中的 2b 步所示);第 3 步,基本结构管理者将收到的 IP 与 PMAC 地址映射关系放入映射表格,这个表格可以用来响应 ARP 请求.

当一台主机需要与另一台主机通信时,其通过基本结构管理者进行 ARP 查询,获得目的主机的 PMAC 地址.在分组将到达目的主机时,目的主机直接连接的交换机将 PMAC 地址重写为目的主机的 AMAC 地址,从而实现了转发对主机的透明.当发生虚拟机迁移时,由基本结构管理者存储迁移后的新映射,并向虚拟机原始连接的交换机发送更新报文.

PortLand 在 FatTree 的基础上构建了 2 层路由和转发协议,可以较好地进行容错路由和转发,支持虚拟机的迁移和系统的扩展.但 PortLand 对交换机的修改,使得需要升级原始交换机才能符合其要求;而且,它依赖于中心化的基本结构管理者的体系结构,使得其面临单点故障和失效的威胁.

2.5 SecondNet

虚拟数据中心(virtual data center,简称 VDC)被定义为一组虚拟机的集合、客户提供的 IP 地址范围和一个服务等级约定(service level agreement,简称 SLA).SLA 不仅定义了计算和存储需求,而且定义了虚拟机的带宽需求.VDC 允许 SLA 根据客户的动态需求进行调整.为了支持虚拟数据中心的抽象,Guo 等人从高扩展性、低成本和易伸缩的角度出发,设计了数据中心网络虚拟化体系结构,称为 SecondNet[13].

为了提供带宽保证,需要在网络中维护带宽分配状态.出于扩展性考虑,Guo 等人并没有将这个状态维持在交换机上,而是将带宽分配状态放在服务器的系统管理程序(hypervisor)上,且管理程序仅需维持其本身运行的虚拟机的状态.

为了进行带宽分配,Guo 等人提出了一种中心化的启发式算法.在这个算法中,将邻居服务器分为规模不同的组.当需要分配一个 VDC 时,仅需搜索合适的组而不是整个物理网络,从而极大地减小了分配的时间.分配的服务器距离较近,也使得 VDC 内的带宽较高;然后,使用高效的最小耗费流算法来将虚拟机映射到物理服务器上,利用物理服务器网络的丰富连接进行路径分配.

完成带宽分配之后,Guo 等人又引入了基于端口交换的源路由(port-switching based source routing,简称 PSSR).因为网络拓扑已知,PSSR 将一个路由路径表示为一系列交换机输出端口的序列.PSSR 可以使用现有商业交换机的 MPLS 功能进行部署.因此,SecondNet 可以在多数当前提出的数据中心网络结构上部署,比如 FatTree,VL2 和 BCube 等.

SecondNet 的体系结构如图 9 所示.

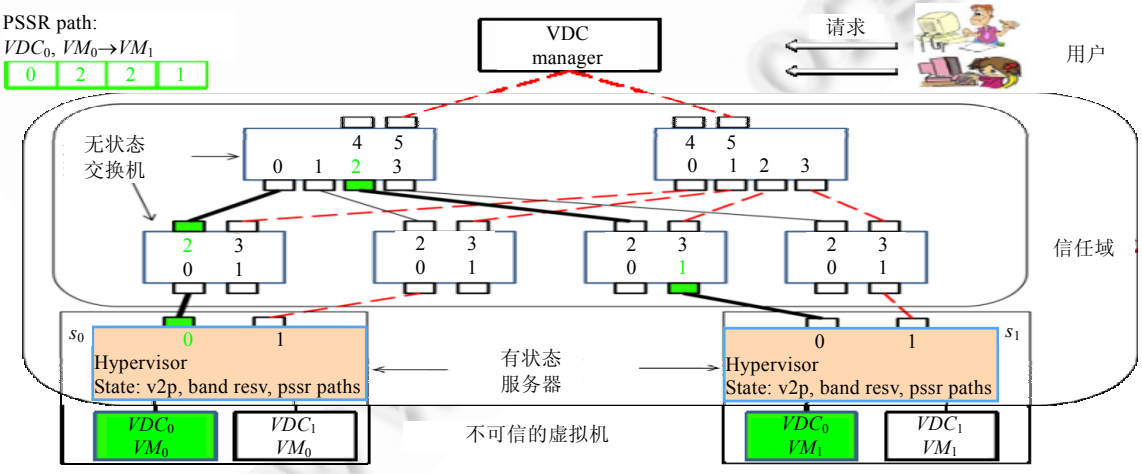

Fig.9 Architecture of SecondNet

图 9 SecondNet 体系结构

从图中可以看出,SecondNet 引入了 VDC 管理者(VDC manager)进行 VDC 创建、调整和删除.VDC 管理者、服务器管理软件和交换机构成了可信计算基础.VDC 管理者通过生成树来管理服务器和交换机,如图 9 中的虚线所示.图 9 中还用粗实线显示了一条 PSSR 路径.

SecondNet 依赖于中心化的 VDC 管理者进行带宽分配和虚拟数据中心管理,使得它在 VDC 管理者处形成瓶颈,面临单点失效的风险.

2.6 Jellyfish

Singla 等人认为,严格的网络结构会限制网络的扩展性,因此提出通过随机互联网络中的交换机构成一张随机图,从而获得更短的平均路径长度,并潜在地减小网络耗费[9].根据这种思想,提出了 Jellyfish.

Jellyfish 的方法是在 ToR 交换机层上构建一个随机图.每个 ToR 交换机 i 有 k_i 个端口,其中,r_i 个连接到其他 ToR 交换机,使用剩下的 k_i-r_i 个端口连接到服务器.在最简单的情况下,每个交换机有相同数量的端口并承载相同数量的服务器,即任取 $i,k=k_i,r=r_i$,有 N 个机架,网络可以支持 $N(k-r)$个服务器.在这种情况下,网络是一个随机规则图.图 10 是一个 Jellyfish 网络拓扑的示意图.

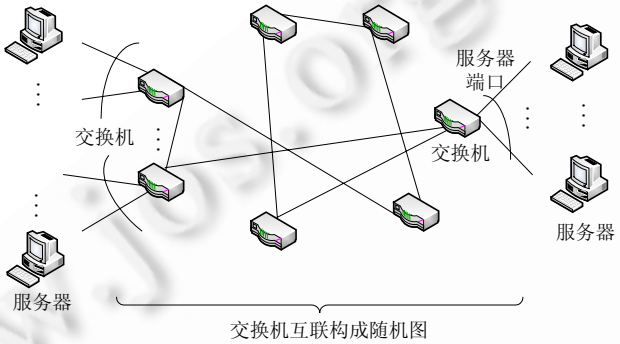

Fig.10 Topology of Jellyfish

图 10 Jellyfish 拓扑

为了构建 Jellyfish 网络,可以采用如下方法:随机选取一对拥有空闲端口的非邻居交换机,使用一段链路将其连接,并重复这个过程,直到网络中不能再添加新的链路为止.当网络中新添加交换机或者仍然有交换机的空闲端口大于 2 时,假定空闲端口为(p_1,p_2),那么随机去掉一条已经存在的链路(x,y),并添加两条新的链路(p_1,x)和(p_2,y).

通过这种构建方法,根据现有的关于随机图的理论结果,Singla 等人指出:在网络中存在的服务器数量小于 900 时,Jellyfish 可以比 FatTree 多承载 27%的服务器,并且在网络规模扩大时优势更加明显.此外,Jellyfish 网络的平均路径长度小于 FatTree 的平均路径长度,并且在能耗方面也拥有优势.

Jellyfish 主要从网络结构出发,取得了比 FatTree 更短的平均路径长度,并提供了比 FatTree 更宽的路径容量.但随机的网络结构也使得布线成为一个挑战,并使得机架之间的位置摆放受限.特别是需要互联由 Jellyfish 构成的集装箱数据中心时,如何处理互联数据中心的光纤成为一个挑战.另外,如何实现 Jellyfish 中假定的最优路由,是随机拓扑面临的最大挑战.

2.7 OSA

2.7.1 动机和技术基础

通过分析数据中心流量的特征,Chen 等人认为,在数据中心中提供均匀的高性能网络场景没有必要,并指出,如果网络能够根据流量变化调整自身的拓扑和链路带宽,将能提供更大的灵活性和传输带宽.他们还引入了光网络技术[21].

光网络支持按需提供网络连接和链路容量,可以为服务器池构建灵活的连接.光链路可以使用比铜线更少

的电量在长距离上支持更高的比特率.另外,光交换机比电交换机运行时产生的热量更少,使得其热耗散和制冷成本更低.鉴于此,Chen 等人提出了基于光交换的数据中心网络体系结构 OSA(optical switching architecture),其体系结构如图 11 所示.OSA 中主要引入了光交换矩阵和波分选择交换机作为技术基础.

- 光交换矩阵(optical switching matrix,简称 OSM).大部分光交换模块是双向 $N\times N$ 矩阵,其中,任意输入端口可以连接到任意的输出端口.目前流行的 OSM 技术使用 MEMS(micro-electro-mechanical switch)实现,它可以在 10ms 以内通过机械地调整镜子的微排列来更改输入和输出端口的连接.
- 波分选择交换机(wavelength selective switch,简称 WSS).一个 WSS 是一个 $1\times N$ 交换机,由一个通用和 N 个波长端口组成.它将从通用端口进入的波长的集合在 N 个波长端口分开,这个过程可以在运行时以毫秒级进行配置.例如,如果通用端口收到了 80 个波长,它可以在端口 1 路由第 1 个~第 20 个波长,将第 30 个~第 40 个波长和第 77 个波长通过端口 2 路由等.

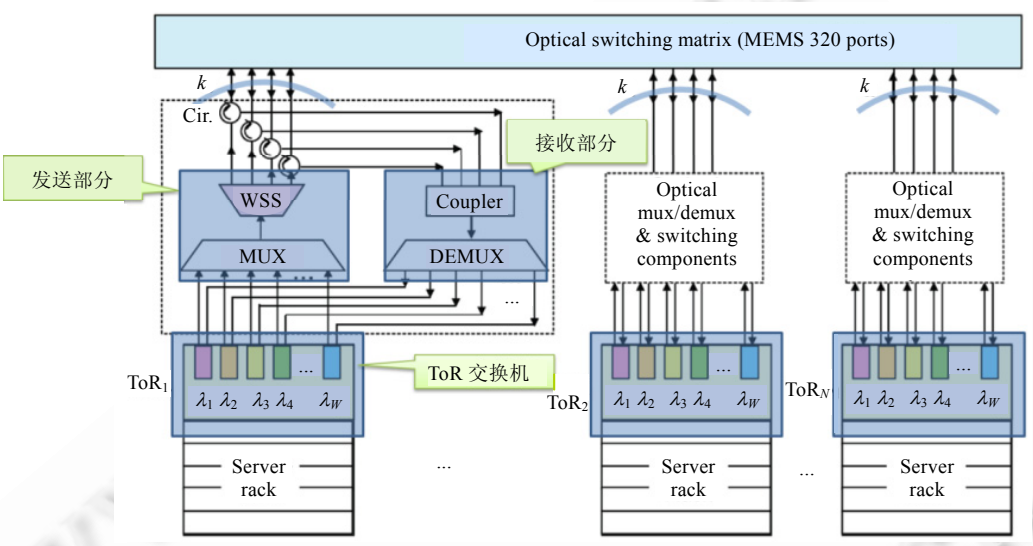

Fig.11 Architecture of OSA

图 11 OSA 体系结构

2.7.2 OSA 网络体系结构和路由

OSA 通过利用 MEMS/OSM 的重配置能力达到灵活的拓扑.开始时,将 N 个 ToR 中的每一个连接到一个 N 端口 MEMS 中的一个端口.假定 MEMS 双向端口匹配,这意味着每个 ToR 在给定的时刻可以与一个其他的 ToR 进行通信.如果将 N/k 个 ToRs 连接到 MEMS 的 k 个端口,则每个 ToR 可以同时与 k 个其他 ToR 通信,其中,k>1 是一个 ToR 的度.

为了实现多跳连接,OSA 使用逐跳交换来达到网络范围的连通性.多跳路径上的每一跳都将分组从光转换为电信号,然后转回光信号,并在 ToR 进行交换.OSA 通过管理拓扑来调整每个连接的容量和路由路径.OSA 采用管理平面实现拓扑、链路容量和路由的计算与配置.

2.7.3 OSA 和光数据中心网络分析

光交换比点交换方式具有潜在的更高的传输速度、更灵活的拓扑结构,并且其制冷成本更低,因而是数据中心网络很重要的研究方向.

但现有的 OSA 结构仍面临一些问题.(1) 在当前的 OSA 体系结构中,一些时延敏感的短流会受到光设备重配置的约 10ms 的影响.通常,数据中心的时延小于 1ms,并且数据中心网络的控制流很多是短流,这会影响数据中心在控制平面对整个系统的操作时效性.(2) OSA 当前的设计针对的是集装箱规模的数据中心,其规模有限.如何以其为基础,从体系结构和管理的角度设计和构建大规模数据中心网络很有挑战性.(3) OSA 中采用的 ToR

交换机使用的交换机是同时支持光传输和电信号传输的交换机,这使得其与传统数据中心的纯电信号交换机难以兼容.

Helios 和 c-Through 是两个早期提出的混合电/光结构,在它们的模型中,每个 ToR 与一个电交换网络和一个光网络有连接.电网络是一个 2 层或 3 层的具有特定超额认购比例的树.在光部分,每个 ToR 仅有 1 个连接到其他 ToR 的光链路,但是这个链路容量很大,其成本较高并且没有充分利用光交换的灵活性.

2.8 无线数据中心网络

2.8.1 典型体系结构

无线技术可以在不必进行重新布线的情况下灵活调整拓扑,因此,Ramachandran 等人在 2008 年将无线技术引入了数据中心网络.随后,Kandula 等人设计了 Flyways[18,31],通过在 ToR 交换机间增加无线链路来缓解机架的拥塞问题,从而最小化最大传输时间.但是无线网络很难单独满足所有的针对数据中心网络的需求,包括扩展性、高容量和容错等.比如,由于干扰和高传输负载,无线链路的容量经常是受限的.因此,Cui 等人引入无线传输来缓解热点服务器的拥塞,并将无线通信作为有线传输的补充,提出了一个异构的以太网/无线体系结构,其体系结构如图 12 所示,这里称其为 WDCN[27].

为了不引入过多的天线和彼此干扰,Cui 等人将每个机架作为一个无线传输单元(wireless transmission unit,简称 WTU),如图 12 底部所示.这样设计使得机架不会阻塞视线传输.

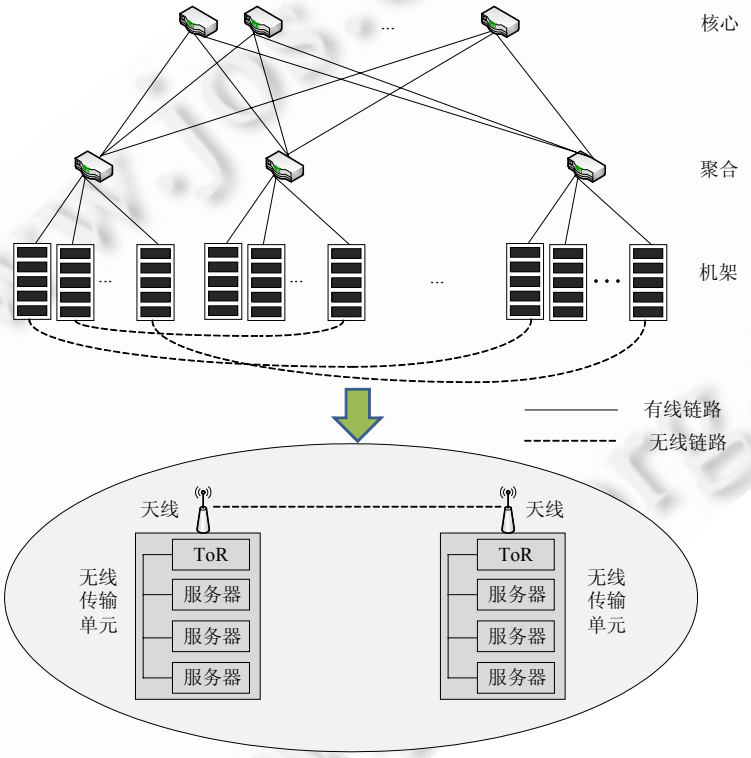

Fig.12 Architecture of WDCN

图 12 WDCN 的体系结构

2.8.2 无线链路调度机制

Cui 等人提出的无线链路调度机制,包括收集流量需求、分配链路、链路调度这 3 个部分:

- 收集流量需求:一个 MTU 内的特定服务器被指定为 MTU 的单元头.单元头负责收集本地流量信息并执行调度算法.每个单元头安装了控制天线,所有的单元头通过共同的 2.4/5GHz 信道以推的方式广播

其流量负载.因此,所有的单元都可以得到全局流量负载分布,并可以独立进行无线链路调度.

- 链路调度:在收集了流量需求信息之后,头服务器需要为无线传输分配信道.Cui 等人提出了一种启发式的分配方法[27].

2.8.3 全无线架构数据中心网络

基于 60GHz 无线通信技术,Shin 等人提出了一种全无线架构的数据中心网络[31,32].Shin 等人将交换结构聚合到服务器节点,期望将服务器节点布置得紧密相连、低伸展且支持失效恢复.为了达到这个要求,将服务器的网卡替换为 Y-交换机[32].另外,还将服务器布置在圆柱型机架里,从而可以方便地建立机架间和机架内的通信信道,并使得这些连接一起构成了一个紧密链接的网状结构.由于其网络连接属于 Cayley 图的一种,因此称其为 Cayley 数据中心(cayley data center,简称 CayleyDC).根据这种结构,Shin 等人设计了全新的路由协议,可以让服务器用少量内存短时间内计算路由,并保证路由的有效性和较短的跳数.CayleyDC 本质上属于服务器为中心的结构,但为了讨论方便,本文在此集中讨论.

正如 Shin 等人指出的那样,Cayley 数据中心面临着几个问题:一是 MAC 层竞争极大地影响了系统的性能;二是无线网络的性能受其网络跳数影响较大;三是多跳的性能问题限制了 Cayley 数据中心的扩展性.

2.8.4 无线数据中心网络分析

- 无线/有线混合结构

无线技术的应用使得网络拓扑不再固定不变,并且省去了复杂的布线工作,使其在数据中心网络环境具有一定的应用前景.Flyways 以及 WDCN 引入无线技术缓解热点主机的带宽问题,取得了一定的效果,并使得流量需求与无线链路调度成为研究的焦点.但无线技术在提供足够带宽的前提下,其传输距离是有限的,因而限制了其在大规模数据中心中的部署.另外,WDCN 采用广播方式收集流量需求,使得其面临时钟同步以及通信开销较大的问题.而且,测量结果显示,数据中心流量是持续改变的,这使得热点主机的位置是不确定的,从而对拓扑的调整提出了更大的挑战.

- 全无线结构

像第 2.8.3 节所指出的那样,全无线数据中心网络在规模、扩展性和性能方面还面临着许多问题,但是,无线技术的优势和不断的发展,使得使用无线结构构建中小型、面向特定应用的数据中心成为可能.

3 服务器为中心的方案

在服务器为中心的方案设计中,采用迭代方式构建网络拓扑,服务器不仅是计算单元,而且也充当路由节点,会主动参与分组转发和负载均衡.这类方案通过迭代设计避免了位于核心层交换机的瓶颈,服务器之间拥有多条可用的不相交路径.这一类的典型设计方案包括 DCell[1],BCube[7],MDCube[8],PCube[15],FiConn[33],雪花结构[3]和 HFN[2]等.

3.1 BCube和MDCube

3.1.1 拓扑结构与构建方法

BCube 网络中主要包括服务器和交换机两种组件[7].BCube 采用了递归的构建方法,$BCube_0$ 的结构如图 13 所示.从图 13 可以看出,$BCube_0$ 就是将 n 个服务器连接到一个 n 端口的交换机.在图 13 中,n=4.$BCube_1$ 结构如图 14 所示,它由 4 个 $BCube_0$ 和 4 个 4 端口交换机构成.更一般的情况是,$BCube_k$ 由 n 个 $BCube_{k-1}$ 和 n^k 个 n 端口交换机组成.每个 $BCube_k$ 中的服务器有 k+1 个端口,标记为 level 0 到 level k.因此,一个 $BCube_k$ 有 $N=n^{k+1}$ 个服务器和 k+1 层交换机,每一层有 n^k 个 n 端口交换机.

$BCube_k$ 的构建规则如下:将 n 个 $BCube_{k-1}$ 标记为 0~n−1.每个 $BCube_{k-1}$ 中的服务器计数为 0~n^k−1.然后将第 j 个 $BCube_{k-1}$ 中的服务器 i 的第 k 层端口连接到第 k 层交换机的第 i 个交换机的第 j 个端口.

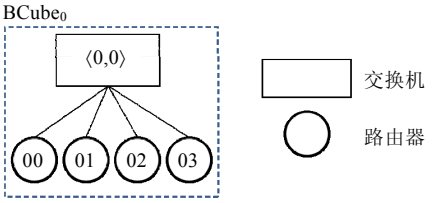

Fig.13 $BCube_0$ network

图 13 $BCube_0$ 网络

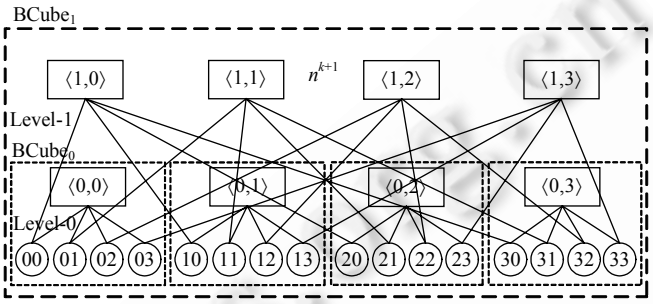

Fig.14 $BCube_1$ network

图 14 $BCube_1$ 网络

3.1.2 单播路由和多路径计算

给定两个服务器 $A=a_ka_{k-1}\ldots a_0$ 和 $B=b_kb_{k-1}\ldots b_0$,其单路径路由算法如图 15 所示.从图 15 可以看出,从第 k 位到第 0 位,BCubeRouting 每次遇到 A 与 B 不同的位时,就添加一个节点到 A 与 B 的路径中,最终成功到达节点 B.在 $BCube_k$ 中,任意两个服务器间存在 k+1 条并行路径.另外,为了进行负载均衡和容错,Guo 等人为 BCube 设计了源路由协议 BSR[7].

```
/*A=a_k a_{k-1}…a_0 和 B=b_k b_{k-1}…b_0;A[i]=a_i;B[i]=b_i;
  IT=[π_k,π_{k-1},…,π_0]是[k,k-1,…,1,0]的一个组合*/
BCubeRouting(A,B,IT):
    Path(A,B)={A,};
    I_Node=A;
    for (i=k; i≥0; i--)
        if (A[π_i]≠B[π_i])
            I_Node[π_i]=B[π_i]
            将 I_Node 加入到 path(A,B)
    return path(A,B);
```

Fig.15 Single-Path routing algorithm in BCube

图 15 BCube 中的单路径路由算法

3.1.3 MDCube

BCube 设计的目标是集装箱数据中心,而如何互联集装箱数据中心、构建更大规模的数据中心,是 MDCube 的主要目标[8].互联集装箱数据中心面临的 3 个挑战是:集装箱间的高带宽需求、互联结构的成本和布线的复杂性.

MDCube 使用 BCube 中交换机的高速接口来互联多个 BCube 集装箱.为了支持数百个集装箱,它使用光纤作为高速链路.每个交换机将其高速接口作为其 BCube 集装箱的虚拟接口.因此,如果将每个 BCube 集装箱都当作一个虚拟节点,它将拥有多个虚拟接口.出于扩展性考虑,MDCube 引入了维度.每个 BCube 集装箱的交换机被划分为组,作为连接到不同维度的接口.一个集装箱由其映射到一个多维数组的 ID 来标识.假定维度是 D,那么

一个集装箱的标识是 D 元组 $cID=c_Dc_{D-1}\ldots c_0(c_d[0,m_d-1],d[0,D])$.在维度 d 上,两个仅在维度 d 上 ID 元组不同的集装箱之间存在一条链路.通过这种方法,BCube 集装箱互相连接形成一个超立方结构网络[34].

3.1.4 BCube 类方案分析

BCube 类方案采用的服务器为中心的体系结构充分利用了服务器和普通交换机的转发功能,在支持大量服务器的同时降低了构建成本,成为数据中心网络的重要研究方向.BCube 类方案提供多路径,并且提供了负载均衡,不会出现明显的瓶颈链路,并增加了可靠性.当发生服务器或者交换机失效时,BCube 可以做到性能的优雅下降,从而维持了服务的可用性.

BCube 也存在一些不足:

(1) BCube 中使用普通商业交换机连接大量的服务器,从而使得其需要很多交换机和链路,这增加了其布线的难度和出错的概率.

(2) BSR 执行时,BCube 会探测服务器间存在的 k+1 条路径,从而确定最佳路径.在“所有到所有(all-to-all)”通信模式中,这个探测过程会造成较大的通信和计算开销.

(3) BCube 要求每个服务器都要有 k+1 个端口,这使得目前的很多现有服务器难以符合其要求,需要进行升级改造.

3.2 FiConn

普通商用服务器通常有 2 块网卡,但目前往往仅使用 1 块,另一块作为备用.这启发了 Li 等人使用备用端口进行服务器互联的想法,从而设计了一个互联结构,称为 FiConn[33].每个 FiConn 中的服务器使用两个网卡端口,一个连接到交换机,另一个连接到另外的 FiConn 服务器.

FiConn 是一个递归定义的结构.高层 FiConn 由一些低层 FiConn 构建.Li 等人将 k 层 FiConn 标识为 FiConn_k. FiConn_0 是基本的构建单元,由 n 个服务器和一个 n 端口交换机连接.每个 FiConn 中的服务器有 1 个端口连接到 FiConn_0,称其为 0 层端口.连接到 0 层端口和交换机的链路称为 0 层链路.如果服务器的备用端口没有连接到其他服务器,则称其为可用备用端口.比如在 FiConn_0 中,初始有 n 个服务器有可用备用端口.

如果一个 FiConn_{k-1} 中共有 b 个服务器拥有可用备份端口,那么,FiConn_k 中 FiConn_{k-1} 的数量 $g_k=b/2+1$.在每个 FiConn_{k-1} 中,b 个服务器中的 $b/2$ 个拥有备用端口的服务器使用其备用端口连接到其他 FiConn_{k-1}.这 $b/2$ 个被选择的服务器称为 k 层服务器,k 层服务器上被选择的端口称为 k 层端口,连接 k 层端口的链路称为 k 层链路.如果将 FiConn_{k-1} 看作一个虚拟服务器,那么 FiConn_k 事实上是一个由 k 层链路连接的 FiConn_{k-1} 的网络.

使用数字 u_k 来标识中 FiConn_k 的一个服务器 s.假定在 FiConn_k 中服务器数量是 N_k,那么 $0\leqslant u_k\leqslant N_k$.另外,服务器 s 可以使用 k+1 元组标识,也就是$[a_k,\ldots,a_1,a_0]$,其中,a_0 在 FiConn_0 中标识 s.那么,为了标识方便,s 也可以被标识为$[a_k,u_{k-1}]$,$[a_k,a_{k-1},u_{k-2}]$等.

图 16 的算法展示了在 g_k 个 FiConn_{k-1} 基础上,FiConn_k 的构建.在每个 FiConn_{k-1},满足的服务器被选择作为 k 层服务器,它们在图 16 的第 4 行~第 6 行中被互联.

```
FiConnConstruct(k){
    for (i1=0; i1<gk; i1++)
        for (j1=i1×2^k+2^(k-1)−1; j1≤Nk−1; j1=j1+2^k)
        i2=(j1−2^(k−1)+1)/2^k+1
        j2=i1×2^k+2^(k−1)−1
        将[i1,j1]与[i2,j2]相连
    return
    }
```

Fig.16 Constructing method of FiConn_k

图 16 FiConn_k 的构建方法

我们以图 17 为例.其中,n=4,k=2.FiConn_0 由 4 个服务器和 1 个 4 端口交换机组成.用于组成 FiConn_1 的 FiConn_0 的数量是 4/2+1=3.服务器[0,0],[0,2],[1,0],[1,2],[2,0]和[2,2]被选择作为 1 层服务器,将[0,0]连接到[1,0],

将[0,2]连接到[2,0],将[1,2]连接到[2,2].

在每个 $FiConn_1$,有 6 个拥有可用备用端口的服务器,因此,用于构成 $FiConn_2$ 的 $FiConn_1$ 的数量是 6/2+1=4.连接选择的 2 层服务器如下:[0,0,1]连接到[1,0,1],[0,1,1]连接到[2,0,1],[0,2,1]连接到[3,0,1],[1,1,1]连接到[2,1,1],[1,2,1]连接到[3,1,1],[2,2,1]连接到[3,2,1].这样就构成了如图 17 所示的拓扑结构.

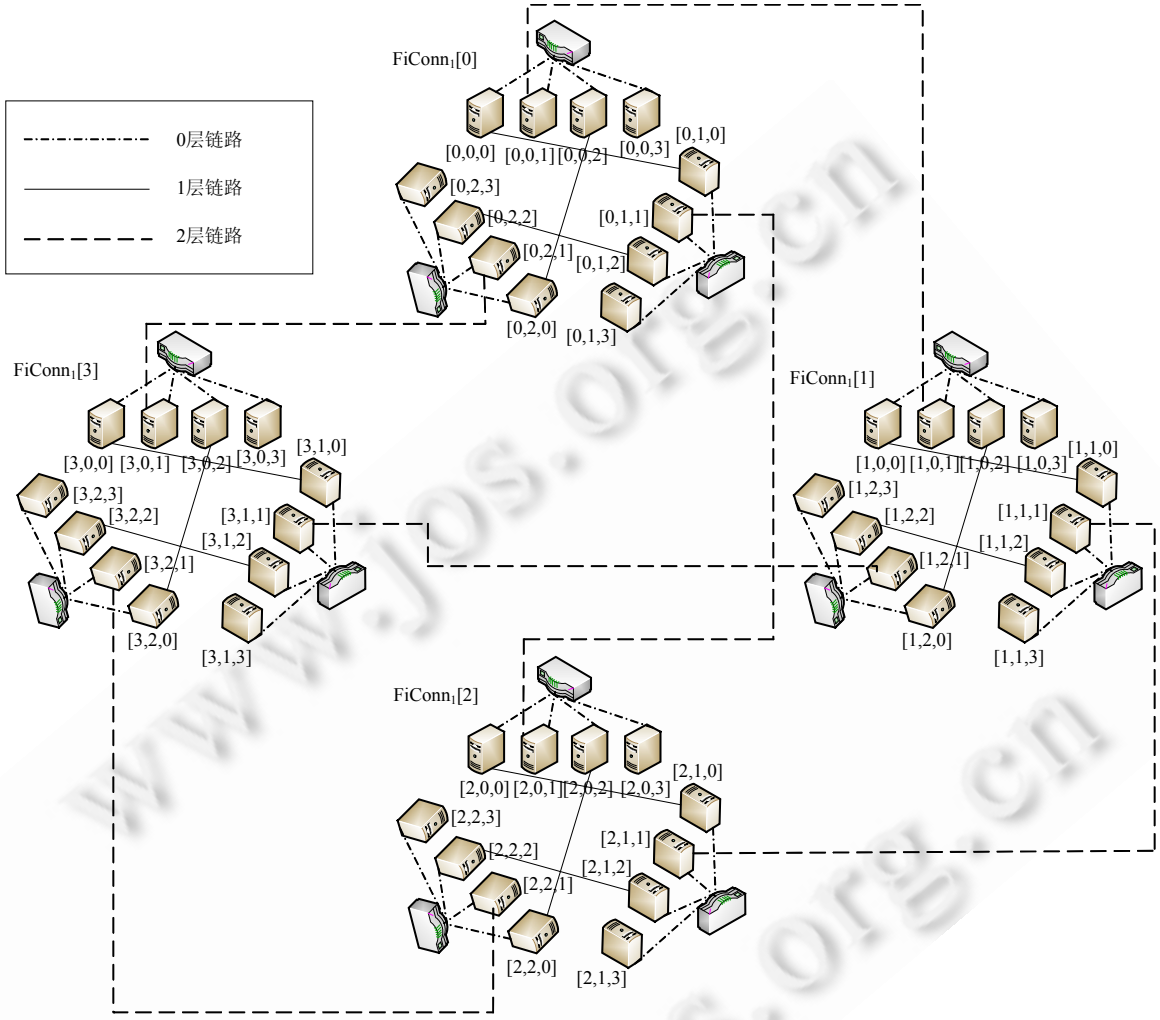

Fig.17 An example of $FiConn_2$ with n=4

图 17 当 n=4 时,一个 $FiConn_2$ 的例子

FiConn 在不大量增加服务器端口的前提下,讨论了服务器为中心的结构,在一定程度上提高了服务器间的联通度,增加了服务器间通信的带宽.但从图 17 也可以看出,每个 $FiConn_0$ 对外连接的链路仍然有限,这使得 FiConn 的容错性较弱,且其路径长度较大,路由效率不高.

4 数据中心网络综合分析及未来的研究方向

4.1 数据中心网络综合分析

第 2 节和第 3 节分别综述了网络为中心的方案和服务器为中心的方案,并简要分析了各个方案的优缺点.

这里将从总体上从以下几个测度对有代表性的方案进行集中对比.

- 规模:方案的现有设计能够支持的服务器的数量大小.分为 3 个层次:小、中、大.
- 带宽:在使用主流配置和数据中心的流量模式下,服务器间的带宽相对于服务器网卡带宽的比例.网络的超额认购比例越高,服务器间路径越少,其带宽就越小;反之则越大.分为 4 个层次:小、中、较大、大.
- 容错性:主要从该方案能否能够有效处理服务器故障、交换机故障和链路故障来衡量.分为 4 个层次:差、中、较好、好.
- 扩展性:主要从该方案是否存在中心化设施、单点失效以及是否容易增量部署几个方面进行衡量.分为 3 个层次:差、中、较好.
- 布线复杂性:主要衡量方案引入的布线的复杂程度.一般地,结构越复杂、引入的端口和链路越多,其复杂性就越高.分为 5 个层次:低、较低、中、较高、高.
- 成本:主要从方案使用的交换机类别/价格、交换机数量和链路数量进行衡量.分为 4 个层次:低、中、较高、高.
- 兼容性:主要衡量各个方案是否与使用 IP 和以太网的传统数据中心兼容,包括服务器的端口数量、网络协议和路由方式等.分为 4 个层次:低、中、较高、高.
- 配置开销:主要反映方案中交换机配置、中心化基础设施的配置以及服务器配置的工作量.分为 3 个层次:中、较高、高.
- 流量隔离:反映方案是否考虑到服务的部署以及在服务之间提供流量隔离.可以看到,在诸多方案中,SecondNet 关注了这一点,其他方案多数没有明确指出这一点.
- 灵活性:是指方案构建的网络规模、网络拓扑和链路容量能否灵活变化.分为 4 个层次:低、中、较高、高.

4.2 未来的研究方向

从前文的综述可以看出,数据中心网络已经得到了极大的关注,研究者提出了诸多方案.但从表 1 的分析与第 1.3 节的需求可看出,目前提出的方案很难完全满足当前诸如新型计算和应用模式对数据中心提出的需求.

Table 1 Comparison of researches on the architecture of data center network

表 1 数据中心网络体系结构研究的对比

		规模	带宽	容错性	扩展性	布线复杂性	成本	兼容性	配置开销	流量隔离	灵活性
网络为中心的方案	分层结构	小	小	差	差	中	高	高	高	无	低
	FatTree	中	中	中	中	较高	较高	高	较高	无	低
	ElasticTree	中	中	中	中	较高	较高	高	高	无	中
	PortLand	中	大	较好	中	较高	较高	较高	较高	无	中
	Monsoon	大	大	中	中	较高	较高	中	较高	无	中
	VL2	大	大	中	中	较高	较高	中	较高	无	中
	Jellyfish	大	大	较好	较好	高	中	中	较高	无	较高
	OSA	小	大	差	中	较低	较高	低	中	无	高
	WDCN	小	大	较好	中	较低	中	中	中	无	高
	SecondNet	大	可变	较好	中	高	中	中	中	有	高
服务器为中心的方案	DCell	大	较大	较好	较好	高	较高	中	较高	无	较高
	BCube	小	大	好	较好	高	较高	中	较高	无	较高
	FiConn	大	较大	较好	较好	较高	中	中	较高	无	较高
	雪花结构	大	较大	中	较好	高	较高	中	较高	无	较高
	MDCube	大	大	较好	较好	高	高	中	较高	无	较高
	CayleyDC	小	中	中	较好	低	低	低	中	无	高

在未来的研究中,重点关注的方向包括如下几个方面:

(1) 新颖网络结构的应用和研究.网络结构在分布式系统中已经得到了广泛的研究,研究者也已经提出了多种网络结构[3,5,9,33,35].在数据中心网络环境下,现有成熟网络结构的应用需要进行验证和研究.尤其是在服务器为中心的结构中,新兴网络结构研究的空间较大.

(2) 数据中心网络协议的研究与改进.数据中心网络协议包括从 MAC 层到运输层的协议[36,37].数据中心网络在管理和结构上都显著区别于现有的因特网体系结构.管理方面,数据中心网络往往由单个实体进行管理,因而其全局拓扑、流量信息、失效信息和各种日志信息都是可以得到的.利用这些信息辅助协议设计和网络结构设计具有很大的研究价值[13].数据中心网络在结构上比因特网的结构更加严格.利用这种结构信息提出适合于特定结构的协议可以增加运行的效率.

(3) 流量以及失效规律的测量与建模.运行不同类型应用的数据中心的流量模式是不同的,从而使得其流量特征也不相同.采用不同架构的数据中心网络,其失效规律也不一样.目前,已经有了一些关于数据中心流量特征和失效规律的测量和研究[14,17,38,39].全面理解数据中心流量和失效特征有待进一步的测量与建模分析.测量分析结果可以帮助设计者理解数据中心的特性和探索新的网络结构及失效管理机制.

(4) 地址自动配置.诸如 PortLand、BCube、雪花结构等体系结构中将位置和拓扑信息编码到服务器或交换机的地址中,从而提高路由的性能.因此,传统的诸如 DHCP 类的协议无法在这种场景下应用,而人工配置如此大数量的交换机或服务器又是不可能完成的任务.因此,这些结构对自动地址配置提出了新的要求.另外,自动配置可以降低人力成本并减小配置出错的风险.因此,针对拓扑已知甚至未知的数据中心网络,提出低开销、高可靠和易管理的自动地址配置方法成为目前重要的研究方向,并且已经得到许多关注[40,41].

(5) 数据中心的流量工程.数据中心网络中的路由机制期望考虑和满足时延、可靠性、吞吐量和节能方面的要求,这也是流量工程问题.包括数据中心内和数据中心间的流量工程问题,目前采用的方案包括 ECMP 和 VLB 等.但是 ECMP 不检查路径分配的负载,从而导致可能的路径拥塞.VLB 的随机分配方式也可能导致拥塞.注意到,数据中心网络环境下的流量工程的前提不同于因特网环境,比如其虚拟机位置可变、拓扑可知并且可以使用中心化的方法进行流量工程.因此,设计可靠、负载均衡和节能的流量工程方法是重要的研究方向[42−44].

(6) 光交换技术和无线传输技术的进一步应用.如果根据布线以及设计的复杂性和能量消耗来衡量,光/电混合结构是优于传统电交换体系结构的.但是光设备仍然比较昂贵且尚未在数据中心网络中得到应用,因此,除了设计体系结构之外,如何降低其应用成本也是研究的重要内容[45].全无线结构的布线复杂性最低,但如何在多跳环境下设计可靠且高性能的结构仍然是个难题.在无线/有线混合结构中[46],无线技术可以有效缓解热点主机的负载,如何感知流量需求并进行高效的无线链路调度是这个方向研究的重点.

(7) 节能机制[22].低碳、节能减排的社会发展大趋势,对数据中心结构以及路由的能耗提出了新的挑战.数据中心网络的节能可以通过以下几个方面共同实现:设备节能、路由节能、虚拟机和任务调度的节能.ElasticTree 方案和 Shang 等人提出的方案都通过调节设备的状态来实现节能[47].关于虚拟机和任务调度的节能,可参见文献[48].另外,如何根据服务产生流量的特点和所采用的路由算法提出综合的节能方案是有效的节能思路,这方面的研究目前尚且不多.我们认为,服务的部署不能够仅从数据中心网络的结构出发,还应该考虑应用的特点和流量特性,为服务定制合适的网络结构.

(8) OpenFlow 等未来网络技术在数据中心网络中的应用.数据中心网络管理实体单一,使得其可以比因特网更适合进行中心化的控制与管理.OpenFlow 是斯坦福大学在 2008 年提出的[49],其基本思想是,将路由器的控制平面和数据平面相分离,并采用流表来扁平化网络处理层次.目前,Jellyfish 使用 OpenFlow 寻找 k 最短路,MicroTE 使用 OpenFlow 实现流量工程[50],ElasticTree 在 OpenFlow 平台上实现了其原型系统,PortLand 使用 OpenFlow 构建实验床等.我们认为,OpenFlow 在未来的数据中心将会有更多、更广泛的应用.另外,数据中心网络与因特网相对独立,为了效率和成本,它完全有可能采用具有革命性的新架构,而未来网络领域的诸多研究值得其借鉴,比如其数据的标识、内容的定位与传输方法和路由的选择等.

5 结束语

首先,本文总结了在新型计算模式和应用面前传统分层数据中心呈现的不足以及面临的诸多需求.其次,将现有研究方案分为两类进行了综述和对比分析,即以网络为中心的方案和以服务器为中心的方案.分析发现,两类方案各有优缺点,且其典型方案尚不能够满足目前对新型数据中心的需求.最后,对各种典型方案进行了综合

分析对比,并指出了未来的研究方向.

References:

[1] Guo C, Wu H, Tan K, Shi L, Zhang Y, Lu S. Dcell: A scalable and fault-tolerant network structure for data centers. ACM SIGCOMM Computer Communication Review, 2008,38(4):75−86. [doi: 10.1145/1402958.1402968]

[2] Ding ZL, Guo DK, Shen JW, Luo AM, Luo XS. Researching data center networking topology for cloud computing. Journal of National University of Defense Technology, 2011,33(6):1−6 (in Chinese with English abstract).

[3] Greenberg A, Lahiri P, Maltz DA, Patel P, Sengupta S. Towards a next generation data center architecture: Scalability and commoditization. In: Proc. of the ACM Workshop on Programmable Routers for Extensible Services of Tomorrow. 2008. 57−62. [doi: 10.1145/1397718.1397732]

[4] Greenberg A, Hamilton J, Maltz DA, Patel P. The cost of a cloud: Research problems in data center networks. ACM SIGCOMM Computer Communication Review, 2009,39(1):68−73. [doi: 10.1145/1496091.1496103]

[5] Al-Fares M, Loukissas A, Vahdat A. A scalable commodity data center network architecture. ACM SIGCOMM Computer Communication Review, 2008,38(4):63−74. [doi: 10.1145/1402958.1402967]

[6] Liu XQ, Yang SB, Guo LM, Wang SL, Song H. Snowflake: A new type network structure of data center. Chinese Journal of Computers, 2011,34(1):76−86 (in Chinese with English abstract).

[7] Guo C, Lu G, Li D, Wu H, Zhang X, Shi Y, Tian C, Zhang Y, Lu S. BCube: A high performance, server-centric network architecture for modular data centers. ACM SIGCOMM Computer Communication Review, 2009,39(4):63−74. [doi: 10.1145/1592568.1592577]

[8] Wu H, Lu G, Li D, Guo C, Zhang Y. MDCube: A high performance network structure for modular data center interconnection. In: Proc. of the 5th Int'l Conf. on Emerging Networking Experiments and Technologies. 2009. 25−36. [doi: 10.1145/1658939.1658943]

[9] Singla A, Hong CY, Popa L, Godfrey PB. Jellyfish: Networking data centers randomly. In: Proc. of the 9th USENIX Symp. on Networked Systems Design and Implementation 2012 (NSDI 2012). San Jose, 2012. 17−31. [doi: 10.1.1.224.9720]

[10] Gyarmati L, Trinh TA. Scafida: A scale-free network inspired data center architecture. ACM SIGCOMM Computer Communication Review, 2010,40(5):4−12. [doi: 10.1145/1880153.1880155]

[11] Shin JY, Wong B, Sirer EG. Small-World datacenters. In: Proc. of the 2nd ACM Symp. on Cloud Computing (SOCC). 2011. 15−27. [doi: 10.1145/2038916.2038918]

[12] Greenberg A, Hamilton JR, Jain N, Kandula S, Kim C, Lahiri P, Maltz DA, Patel P, Sengupta S. VL2: A scalable and flexible data center network. ACM SIGCOMM Computer Communication Review, 2009,39(4):51−62. [doi: 10.1145/1592568.1592576]

[13] Guo C, Lu G, Wang HJ, Yang S, Kong C, Sun P, Wu W, Zhang Y. Secondnet: A data center network virtualization architecture with bandwidth guarantees. In: Proc. of the 6th Int'l Conf. 2010. 15−27. [doi: 10.1145/1921168.1921188]

[14] Benson T, Anand A, Akella A, Zhang M. Understanding data center traffic characteristics. In: Proc. of the 1st ACM Workshop on Research on Enterprise Networking (WREN). 2009. 65−72. [doi: 10.1145/1672308.1672325]

[15] Huang L, Jia Q, Wang X, Yang S, Li B. PCube: Improving power efficiency in data center networks. In: Proc. of the 2011 IEEE Int'l Conf. on Cloud Computing (CLOUD). 2011. 65−72. [doi: 10.1109/CLOUD.2011.74]

[16] Mysore RN, Pamboris A, Farrington N, Huang N, Miri P, Radhakrishnan S, Subramanya V, Vahdat A. PortLand: A scalable fault-tolerant layer 2 data center network fabric. ACM SIGCOMM Computer Communication Review, 2009,39(4):39−50. [doi: 10.1145/1592568.1592575]

[17] Kandula S, Sengupta S, Greenberg A, Patel P, Chaiken R. The nature of data center traffic: Measurements & analysis. In: Proc. of the 9th ACM SIGCOMM Conf. on Internet Measurement (IMC 2009). 2009. 202−208. [doi: 10.1145/1644893.1644918]

[18] Kandula JPS, Bahl P. Flyways to de-congest data center networks. In: Proc. of the 8th ACM Workshop. Hot Topics in Networks. 2009. 1−6.

[19] Zhang Y, Ansari N. On mitigating TCP incast in data center networks. In: Proc. of the IEEE INFOCOM 2011. 2011. 51−55. [doi: 10.1109/INFCOM.2011.5935217]

[20] Zhang J, Ren F, Lin C. Modeling and understanding TCP incast in data center networks. In: Proc. of the IEEE INFOCOM 2011. 2011. 1377−1385. [doi: 10.1109/INFCOM.2011.5934923]

[21] Chen K, Singla A, Singh A, Ramachandran K, Xu L, Zhang Y, Wen X, Chen Y, OSA: An optical switching architecture for data center networks with unprecedented flexibility. In: Proc. of the USENIX/ACM Symp. on Networked Systems Design and Implementation (NSDI). San Jose, 2012. 1−14.

[22] Heller B, Seetharaman S, Mahadevan P, Yiakoumis Y, Sharma P, Banerjee S, McKeown N. ElasticTree: Saving energy in data center networks. In: Proc. of the 7th USENIX Conf. on Networked Systems Design and Implementation (NSDI). 2010. 2−17.

[23] Farrington N, Porter G, Radhakrishnan S, Bazzaz H, Subramanya V, Fainman Y, Papen G, Vahdat A. Helios: A hybrid electrical/optical switch architecture for modular data centers. ACM SIGCOMM Computer Communication Review, 2010,40(4):339−350. [doi: 10.1145/1851275.1851223]

[24] Ye X, Mejia P, Yin Y, Proietti R, Yoo SJB, Akella V. DOS—A scalable optical switch for datacenters. In: Proc. of the ACM/IEEE Symp. on Architectures for Networking and Communications Systems (ANCS). 2010. 24−36. [doi: 10.1145/1872007.1872037]

[25] Wang G, Andersen DG, Kaminsky M, Papagiannaki K, Ng TS, Kozuch M, Ryan M. c-Through: Part-time optics in data centers. ACM SIGCOMM Computer Communication Review, 2010,40(4):327−338. [doi: 10.1145/1851275.1851222]

[26] Bazzaz HH, Tewari M, Wang G, Porter G, Ng TS, Andersen DG, Kaminsky M, Kozuch MA, Vahdat A. Switching the optical divide: Fundamental challenges for hybrid electrical/optical datacenter networks. In: Proc. of the 2nd ACM Symp. on Cloud Computing. 2011. 30−38. [doi: 10.1145/2038916.2038946]

[27] Cui Y, Wang H, Cheng X, Chen B. Wireless data center networking. Wireless Communications of IEEE, 2011,18(6):46−53. [doi: 10.1109/MWC.2011.6108333]

[28] Cui Y, Wang H, Cheng X. Wireless link scheduling for data center networks. In: Proc. of the 5th Int'l Conf. on Ubiquitous Information Management and Communication. 2011. 44−53. [doi: 10.1145/1968613.1968667]

[29] Wang C, Wang CR, Wang XW, Jiang DD. Network architecture design for data centers towards cloud computing. Journal of Computer Research and Development, 2012,49(2):286−293 (in Chinese with English abstract).

[30] Greenberg A, Hamilton J, Maltz DA, Patel P. The cost of a cloud: Research problems in data center networks. ACM SIGCOMM Computer Communication Review, 2009,39(1):68−73. [doi: 10.1145/1496091.1496103]

[31] Halperin D, Kandula S, Padhye J, Bahl P, Wetherall D. Augmenting data center networks with multi-gigabit wireless links. In: Proc. of the ACM SIGCOMM 2011 Conf. 2011. 38−49. [doi: 10.1145/2043164.2018442]

[32] Shin JY, Sirer EG, Weatherspoon H, Kirovski D. On the feasibility of completely wireless data centers. Technical Report, Cornell University, 2011. http://hdl.handle.net/1813/22846

[33] Li D, Guo C, Wu H, Tan K, Zhang Y, Lu S. FiConn: Using backup port for server interconnection in data centers. In: Proc. of the INFOCOM 2009. IEEE, 2009. 2276−2285. [doi: 10.1109/INFCOM.2009.5062153]

[34] Wang YJ, Sun WD, Zhou S, Pei XQ, Li XY. Key technologies of distributed storage for cloud computing. Ruanjian Xuebao/Journal of Software, 2012,23(4):962−986 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4175.htm [doi: 10.3724/SP.J.1001.2012.04175]

[35] Liao Y, Yin J, Yin D, Gao L. DPillar: Dual-port server interconnection network for large scale data centers. Computer Networks, 2012,56(8):2132−2147. [doi: 10.1016/j.comnet.2012.02.016]

[36] Alizadeh M, Greenberg A, Maltz DA, Padhye J, Patel P, Prabhakar B, Sengupta S, Sridharan M. Data center tcp (dctcp). ACM SIGCOMM Computer Communication Review, 2010,40(4):63−74. [doi: 10.1145/1851182.1851192]

[37] Raiciu C, Barre S, Pluntke C, Greenhalgh A, Wischik D, Handley M. Improving datacenter performance and robustness with multipath TCP. ACM SIGCOMM Computer Communication Review, 2011,41(4):266−277. [doi: 10.1145/2018436.2018467]

[38] Chen Y, Jain S, Adhikari VK, Zhang ZL, Xu K. A first look at inter-data center traffic characteristics via yahoo! datasets. In: Proc. of the INFOCOM 2011. IEEE, 2011. 1620−1628. [doi: 10.1109/INFCOM.2011.5934955]

[39] Gill P, Jain N, Nagappan N. Understanding network failures in data centers: measurement, analysis, and implications. ACM SIGCOMM Computer Communication Review, 2011,41(4):350−361. [doi: 10.1145/2043164.2018477]

[40] Chen K, Guo C, Wu H, Yuan J, Feng Z, Chen Y, Lu S, Wu W. DAC: Generic and automatic address configuration for data center networks. IEEE/ACM Trans. on Networking, 2012,20(1):84−99. [doi: 10.1109/TNET.2011.2157520]

[41] May X, Hu C, Chen K, Zhang C, Zhang H, Zheng K, Chen Y, Sun X. Error tolerant address configuration for data center networks with malfunctioning devices. In: Proc. of the ICDCS 2012. 2012. 708−717. [doi: 10.1109/ICDCS.2012.27]

[42] Viet HOT, Deville Y, Bonaventure O, Francois P. Traffic engineering for multiple spanning tree protocol in large data centers. In: Proc. of the 2011 23rd Int'l Conf. on Teletraffic Congress (ITC). 2011. 23−30.

[43] Dixit A, Prakash P, Kompella R. On the efficacy of fine-grained traffic splitting protocols in data center networks. ACM SIGCOMM Computer Communication Review, 2012,40(1):411−412. [doi: 10.1145/2254756.2254818]

[44] Chen K, Hu C, Zhang X, Zheng K, Chen Y, Vasilakos AV. Survey on routing in data centers: Insights and future directions. IEEE Network—The Magazine of Global Internetworking, 2011,25(4):6−10. [doi: 10.1109/MNET.2011.5958002]

[45] Xu L, Zhang S, Yaman F, Wang T, Liao G, Chen K, Singla A, Singh A, Ramachandran K, Zhang Y. All-Optical switching data center network supporting 100Gbps upgrade and mixed-line-rate interoperability. In: Proc. of the OSA/OFC/NFOEC 2011. 2011. 1−3.

[46] Zhou X, Zhang Z, Zhu Y, Li Y, Kumar S, Vahdat A, Zhao BY, Zheng H. Mirror mirror on the ceiling: Flexible wireless links for data centers. ACM SIGCOMM Computer Communication Review, 2012,42(4):443−454. [doi: 10.1145/2377677.2377761]

[47] Shang Y, Li D, Xu M. Energy-Aware routing in data center network. In: Proc. of the 1st ACM SIGCOMM Workshop on Green Networking. 2010. 1−8. [doi: 10.1145/1851290.1851292]

[48] Deng W, Liao XF, Jin H. Energy management mechanisms in virtualized data centers. ZTE Technology Journal, 2012,18(4):15−18 (in Chinese with English abstract).

[49] Mckeown N, Anderson T, Balakrishnan H, Parulkar G, Peterson L, Rexford J, Shenker S, Turner J. OpenFlow: Enabling innovation in campus networks. ACM SIGCOMM Computer Communication Review, 2008,38(2):69−74. [doi: 10.1145/1355734.1355746]

[50] Benson T, Anand A, Akella A, Zhang M. MicroTE: Fine grained traffic engineering for data centers. In: Proc. of the 7th Conf. on Emerging Networking EXperiments and Technologies. 2011. 1−12. [doi: 10.1145/2079296.2079304]

附中文参考文献:

[2] 丁泽柳,郭得科,申建伟,罗爱民,罗雪山.面向云计算的数据中心网络拓扑研究.国防科技大学学报,2011,33(6):1−6.

[6] 刘晓茜,杨寿保,郭良敏,王淑玲,宋滢.雪花结构:一种新型数据中心网络结构.计算机学报,2011,34(1):76−86.

[29] 王聪,王翠荣,王兴伟,蒋定德.面向云计算的数据中心网络体系结构设计.计算机研究与发展,2012,49(2):286−293.

[34] 王意洁,孙伟东,周松,裴晓强,李小勇.云计算环境下的分布存储关键技术.软件学报,2012,23(4):962−986. http://www.jos.org.cn/1000-9825/4175.htm [doi: 10.3724/SP.J.1001.2012.04175]

[48] 邓维,廖小飞,金海.基于虚拟机的数据中心能耗管理机制.中兴通讯技术,2012,18(4):15−18.

魏祥麟(1985－),男,安徽砀山人,博士,工程师,主要研究领域为对等计算,网络测量,分布式系统,无线自组织网络,数据中心网络.
E-mail: wei_xianglin@ieee.org

陈鸣(1956－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络测量,网络性能分析与建模,分布式系统,未来网络.
E-mail: mingchennj@163.com

范建华(1971－),男,博士,研究员,主要研究领域为云计算,无线网络.
E-mail: fjh0119@gmail.com

张国敏(1979－),男,博士,讲师,CCF 学生会员,主要研究领域为网络管理,分布式计算.
E-mail: zhang_gmwn@163.com

卢紫毅(1977－),男,工程师,主要研究领域为软件无线电,无线通信网络.
E-mail: systemlu@126.com