计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20150962
53(9):2085-2106,2016

云计算数据中心网络设计综述

王斌锋[1] 苏金树[1,2] 陈 琳[1]
[1](国防科学技术大学计算机学院 长沙 410073)
[2](国防科学技术大学并行与分布处理国防重点实验室 长沙 410073)
(binhai.feng@163.com)

Review of the Design of Data Center Network for Cloud Computing

Wang Binfeng[1], Su Jinshu[1,2], and Chen Lin[1]
[1](*College of Computer, National University of Defense Technology, Changsha* 410073)
[2](*Science and Technology on Parallel and Distributed Processing Laboratory, National University of Defense Technology, Changsha* 410073)

Abstract Under the influence of the development of cloud computing, the data center network is going through tremendous changes, which are reflected not only in the improvement of network size, bandwidth, abundant links, scalability, agility and the cutting down of the cost, but also in other aspects, such as the support for the VMotion dynamically, the network virtualization etc. On the basis of the main challenges of data center network for cloud computing, the paper firstly describes the network architecture of data center for cloud computing in two different aspects, which regards switches and servers as the forwarding center respectively. Next, from the perspective of the flexible deployment of network resources, the paper deeply analyzes the related proposals and key technologies concerning the VMotion dynamically in data center for cloud computing. Further, in order to summarize the application of the network virtualization in data center for cloud computing in detail, the paper mainly analyzes a variety of designs for the virtual network architecture, which includes two different types: the scalability type and the performance guaranteed type. Finally, taking the rapid development of advanced technologies into account, the paper puts forward several points of future prediction in terms of the development of data center network for cloud computing.

Key words data center network; cloud computing; network architecture; mobility; virtualization

摘 要 在云计算技术趋势的影响下,数据中心网络正发生着深刻的变革,不仅体现在规模、带宽、多链路、扩展性、灵活性的提升和成本的降低上,而且还体现在对虚拟机动态迁移的支持以及网络虚拟化等方面.围绕云计算数据中心网络所面临的主要挑战,首先,从交换机为中心和服务器为中心的2个角度分别展开论述了云计算数据中心的网络体系架构;其次,从网络资源灵活部署的角度,详细分析了云计算数据中心虚拟机动态迁移的相关协议与关键技术;然后,从虚拟网络体系架构的角度分析了虚拟化技术在云计算数据中心网络中的应用;最后预测了云计算数据中心网络未来的发展趋势.

关键词 数据中心网络;云计算;网络体系架构;移动性;虚拟化

中图法分类号 TP391

收稿日期:2015-11-09;**修回日期**:2016-03-22

基金项目:国家"九七三"重点基础研究发展计划基金项目(2012CB315906);国家自然科学基金项目(61303264)

This work was supported by the National Basic Research Program of China (973 Program) (2012CB315906) and the National Natural Science Foundation of China (61303264).

云计算作为一种服务提供模型,可以依托网络实现随时随地、按需地访问数据中心的各种资源,包括计算资源、网络资源、存储资源等,且在云计算数据中心内部可以实现不同用户间资源的动态分配和调整. 目前,主要有 3 种服务模型,分别是 IaaS, SaaS, PaaS. 与传统的 IT 服务模式相比,云计算自诞生以来就具有显著的特点和巨大的优势,体现在建设成本、扩展性、可靠性、服务质量以及远程访问等方面[1],因而吸引着学术界和工业界的极大关注. 数据中心作为云计算的核心基础设施,为了更好地承载云计算相关的应用服务,更有利地响应租户的资源访问请求,其正经历着前所未有的变革,带来发展机遇的同时也带来了前进的诸多挑战,主要体现在以下 6 个方面:

1) 网络规模日益庞大,存在多种网络形态

为了提高海量数据的集中处理能力,数据中心网络规模日益庞大,互连的计算节点数量能达到 10^5 以上的量级,而交换节点的数量也接近 10^4 量级,如 Google 在其 30 多个数据中心中就拥有超过 100 万台服务器,日渐扩大的网络规模对网络架构、传输协议以及网络性能管理等都提出了新的设计要求[2]. 另外,由于数据中心网络相对封闭、管理较为集中,在实际的部署过程中受不同性能需求的牵引,呈现出多种网络形态共存的情形,如 InfiniBand 高速互连存储网络、增强以太网及用于高性能计算的专用高速网络.

2) 流量工程难度大

数据中心网络流量复杂,难以实现有效地预测与调度,究其原因主要体现在 2 方面:①由于数据中心网络流量具有高动态及高突发的特性,如 many-to-one 通信模式所导致的 Incast 问题,且这些流量大部分耗时极短、不易检测[3];②由于计算密集型应用服务的发展(如 MapReduce, Hadoop 等应用),以及虚拟化技术的大范围引进(如虚拟机在线迁移 VMotion 等),致使数据中心网络"东西流量"占据了总流量的将近 80%,给网络不仅带来了严重的传输负载,而且也增加了网络中流量行为的复杂性.

3) 网络纵向扩展成本高

当前数据中心所采用的网络架构大多是单根或多根的树形层次结构,在网络规模不断扩大以及"东西流量"占据总流量比例较高的情况下,为了提高网络的对分带宽,保证应用服务的服务质量,数据中心通常需要配置更昂贵更具专业化的转发硬件或者负载均衡器,这种纵向扩展的模式无疑成本非常高. 况且,由于树形结构的核心链路存在收敛比的问题[4],随着节点数目的增加,纵向扩展的方式显得不具有可持续性.

4) 网络资源利用率低

当多个应用服务同时部署于数据中心时,通常需要对其使用的网络资源进行相互隔离(如 VLAN),以避免它们之间相互产生干扰. 然而,由于传统通信标识 IP 地址不但表示了应用服务的位置信息,而且也表示了应用服务的身份信息,这就使得隔离措施很大程度地限制了资源调度的灵活性,致使网络资源的使用率普遍较低. 无论是出于提高数据中心可用网络容量的考虑,还是为了降低投入产出比,对网络资源利用率的提高都是很有必要的.

5) 网络故障率呈上升趋势

长期的实践表明,网络故障率会随着系统节点数的增加而快速增长. 目前,数据中心的网络故障大体可划分为 4 类:软件故障、硬件故障、网络配置故障及不明原因的故障[5],各自所占的比率分别为 21%, 18%, 38%, 23%,可以看出,网络配置故障占据的比例最多,不明原因的故障次之. 随着网络规模的扩大,节点之间的互连关系变得越来越复杂,致使网络配置难度增加,局部的配置失误也可能引起网络大范围的失效. 另外,当前云计算数据中心网络趋向于大面积使用成本低廉的商业交换机,而商业交换机的处理能力却十分有限,在网络流量高度突发、行为异常复杂的情况下,容易导致原因不明的网络故障,如交换机突然停止转发流量.

6) 网络运维成本高

在云计算技术趋势的影响下,数据中心的高运维成本主要突显在 3 个方面:网络性能管理、网络配置和能耗管理. 部署于数据中心的应用大多都是在线或者需即时响应的服务,如网络游戏、电子商务、在线视频会议等,这些应用或对网络延迟敏感,或对带宽的分配敏感,或者对二者都很敏感,因而细粒度地实现对应用服务质量的保证是十分复杂的. 网络配置开销大的原因有 2 方面:①由于网络中基于不同使用需求而部署的传输、路由协议比较多样;②由于当前网络配置的自动化程度不高,而人工配置的出错率又较大的缘故. 关于数据中心的能耗,IT 设备系统约占 50%,空调系统约占 37%[6],这二者占据了绝大多数的能耗开支,为了节约能源,绿色数据中心的建设已经逐渐成为关注的焦点.

针对云计算数据中心发展变革所面临的诸多挑战,学术界和工业界都积极提出了多种不同的设计

1）东西向流量：数据中心内部服务器之间的流量，实现虚拟机之间的互通、迁移、存储数据的同步
2）跨数据中心流量：不同数据中心之间的流量；
3）南北向流量：数据中心之外的访问点或数据源与数据中心内部服务器之间的流量。

方案，并有效地推动了云计算应用服务的飞速发展. 通过对这些设计方案的分析与研究，我们认为目前云计算数据中心网络的研究主要有 3 方面的挑战：1）数据中心网络体系架构的设计；2）云计算数据中心支持虚拟机动态迁移的设计；3）云计算数据中心网络虚拟化的设计. 具体可以描述如下：

1）高性价比数据中心网络体系架构的广泛探究. 在云计算环境下，企业为了追求服务可靠、运行简单、成本合理，往往将自身的数据中心移往云端，致使云计算数据中心规模越来越大. 不仅如此，伴随着应用服务的不断繁荣发展，数据中心的东西流量比南北流量将近多出 3 倍，占据着数据中心的绝大多数，改变了传统的 20%～80%原则. 为了保证较好的网络性能，在数据中心网络体系架构的设计过程中，不仅仅要求有较强的可扩展性，集装箱式数据中心就是典型的案例[7]，而且要达到较大的网络带宽，点与点之间需存在多条通信链路. 除此之外，还需要考虑建设成本的问题. 在上述诸多因素的驱动下，当前云计算数据中心网络体系架构的研究成果颇丰，依据转发中心的不同可划分为两大类：以交换机为转发中心的设计和以服务器为转发中心的设计，各有优缺点，二者都取得了较大发展，且在不断优化与创新.

2）对数据中心网络虚拟机迁移(virtual machine motion，VMotion)支撑技术的研究. 快速便捷地实现网络资源(虚拟机)的灵活部署是云计算数据中心 IT 管理流程自动化的重要环节. 为了克服传统数据中心网络中资源(虚拟机)分配不能跨越 3 层设备的缺陷，云计算数据中心网络采用了 2 种解决思路：①设计“大二层”网络，以扩大虚拟机分配的范围；②借助隧道技术，使虚拟机的分配能够在“虚拟 2 层”中得以进行. 通过这些措施，期望实现虚拟机的灵活部署，从而提高网络资源的利用率.

3）对数据中心网络虚拟化技术的积极探索. 为了实现云应用服务的性能隔离、灵活管理以及简易部署等，云计算数据中心大面积多层次引入了虚拟化技术. 作为云计算数据中心重要的技术支撑，虚拟化的对象不仅仅包括服务器，还包括网络(路由器、交换机及链路等[8]). 网络虚拟化的发展使得数据中心的整体服务能力得以大幅提升，主要表现为：提高了网络资源的利用率；增强了应用服务管理的灵活性；改善了应用服务部署的简易性；可以为不同的应用服务提供流量隔离；由于可以限制网络故障的影响区域，推动了新网络协议的测试等.

本文围绕云计算数据中心网络所面临的主要挑战，从云计算数据中心网络体系架构、云计算数据中心 VMotion 支撑和云计算数据中心网络虚拟化 3 个方面，针对性地阐述了相关的研究现状，并讨论了其未来的发展趋势.

1 云计算数据中心网络体系架构

传统数据中心网络承载的主要是客户机/服务器模式的应用服务，一般采用 2 层或 3 层的树形结构：核心层、接入层，或者是核心层、汇聚层、接入层，如图 1 所示为 3 层的传统数据中心网络. 但是，传统的树形层次结构存在着多方面的局限性：1）所有的服务器都位于同一个 2 层广播域中；2）汇聚和核心交换机是服务器之间通信的带宽瓶颈；3）易于发生“单点故障”，如果一个核心交换机出现问题，可能会影响到上千台服务器的正常工作. 随着新型模式应用服务的发展，数据中心的网络体系结构也逐渐发生着演变，Clos 网络[9]在树形结构的基础上，增加了核心交换机的数量，拓展了服务器之间的带宽，如图 2 所示为 Clos 网络的 3 层结构，可以看出上一层

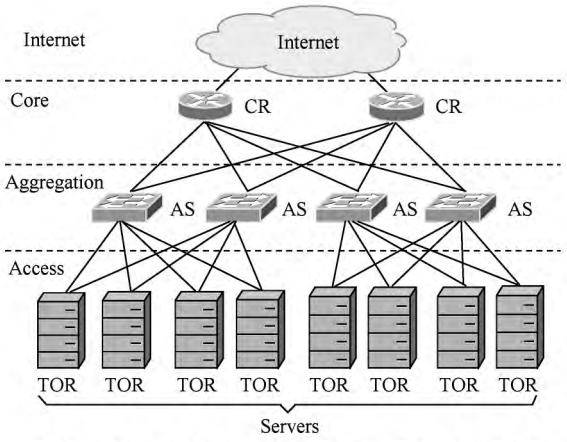


Fig. 1 Traditional data center network architecture.

图 1 传统数据中心网络架构

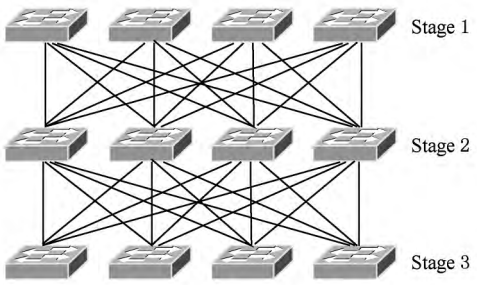


Fig. 2 Three-stage Clos topology.

图 2 Clos 网络 3 层结构

的交换机和下一层的所有交换机都相连. 而 Fat-Tree[10]是 Clos 网络的一个特例,其以 Pod 为单位对接入和汇聚交换机进行划分,体现了汇聚、接入 2 层逐渐融合的扁平化思想. 为了进一步追求高带宽、多链路、低成本及较强的可扩展性,数据中心网络体系结构或者引入无线技术,或者突破传统结构的局限,重新设计新型的网络体系结构.

目前针对数据中心网络体系架构的研究变化多样,大体上可以划分为两大类:以交换机为转发中心的体系架构和以服务器为转发中心的体系架构. 在数据中心网络中引入无线技术的设计,从根本上讲可以归类为以交换机为转发中心的设计类别,但是为了突出其特殊性,本文在 1.1.3 节专门进行详细描述.

1.1 以交换机为中心的设计方案

在以交换机为转发中心的网络架构设计方案中,网络的连接及流量的路由、转发等功能全部是由交换机和路由器来完成的,这类设计方式最为直接、简单,通过优化网络连接以及路由机制,或者升级交换机的硬件和软件就能加以实现,代表性的方案有 Fat-Tree[10]、Facebook 数据中心架构[11]、REWIRE[12]、SPAIN[13]等. 为了方便描述,本文依据设计方案的网络结构是否规则进行了再划分,并就 2 个方面典型的设计进行了简述.

1.1.1 规则网络结构

1) Fat-Tree

Fat-Tree 是对传统树形层次结构的一个优化,由 Al-Fares 等人[10]提出,具体如图 3 所示,其仍然可分为核心、汇聚和接入 3 个层次. 与传统树形层次结构相比,Fat-Tree 有 3 方面的独特之处:①在网络中大量采用了统一的廉价的商业交换机作为转发设备,避免了纵向扩展所带来的高成本;②由于各个层次都采用了经济型的商业交换机,为了保证网络的可靠性,所有的汇聚和接入交换机都被划分为不同的集群 Pod(其中有 $k/2$ 台汇聚交换机、$k/2$ 台接入交换机),在集群内汇聚和接入交换机之间采用了全相连的方式;③采用多路径路由技术,为服务器之间的通信提供了较大的对分带宽.

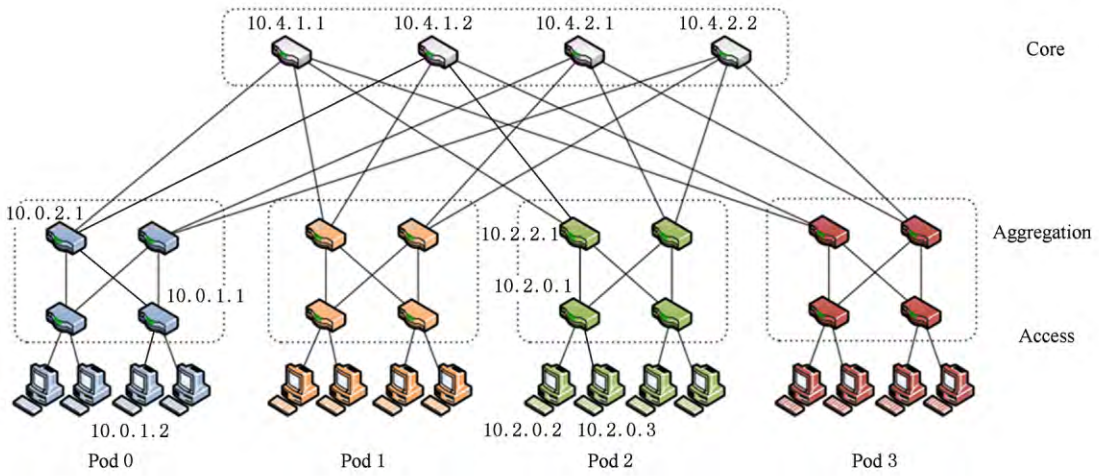


Fig. 3 The network architecture of Fat-Tree.

图 3 Fat-Tree 网络结构

为了在服务器之间实现多路径的通信,Al-Fares 等人主要采取了 2 方面的措施:①制定了地址分配的具体规则;②建立了 2 级路由表及相应的路由机制. 假设使用 10.0.0.0/8 地址块对网络中的所有设备进行地址分配,则集群 Pod 内交换机的地址形式为 10.*pod*.*switch*.1,其中 *pod* 表示集群号,*switch* 表示交换机在集群中的位置;核心交换机的地址形式为 10.*k*.*j*.*i*,其中 *j* 和 *i* 表示该交换机在 $(k/2)^2$ 个核心交换机中的位置;服务器的地址形式为 10.*pod*.*switch*.*ID*,其中 *ID* 从 2 开始编号. 基于这样的地址分配,作者建立了 2 级路由表及相应的路由机制,具体如图 4 所示,左端为第 1 级路由表,右端为第 2 级路由表. 对于不同 Pod 的终端,该结构提供有 k 条互不相同的传输路径,主要依赖第 2

Prefix	Output Port
10.2.0.0/24	0
10.2.1.0/24	1
0.0.0.0/0	

Suffix	Output Port
0.0.0.2/8	2
0.0.0.3/8	3

Fig. 4 Two level routing table.

图 4 2 级路由表

级路由表来完成,能够实现 Pod 间流量最大可能地均匀分布于核心交换机之间;对于同一 Pod 的终端,其路由依赖于第 1 级路由表中的终结性表项,会将相应的流量从特定的端口转发出去.

虽然 Fat-Tree 网络结构为服务器与服务器之间的通信提供了多条路径,但是如果网络流量在多条路径上分布不均,就会影响链路整体的使用效率,特别是对于多播流量. 针对这一问题,文献[4]研究了在保证链路收敛比 O 的情况下核心交换机最少应当部署的数目,并指出通过提高链路收敛比来降低数据中心建设成本的方法是存在局限性的,从而为如何低成本高效率地构造多播 Fat-Tree 结构提供了理论依据.

2) Facebook 数据中心架构

对于普通的数据中心而言,网络的"东西流量"占据着总流量的将近 80%,但是,作为社交网络巨头的 Facebook 有着其自身的特殊性,其数据中心网络的"东西流量"比"南北流量"将超出近 1 000 多倍[11]. 在这种情况下,Fat-Tree 结构明显存在较大的局限性:①纵向扩展的成本比较高;②即使部署了很多昂贵的转发设备,也不一定能有效地支撑这种规模的网络流量. 为了解决网络较大的对分带宽需求,Facebook 最新建立的数据中心采用了一种新型的网络结构,具体如图 5 所示:

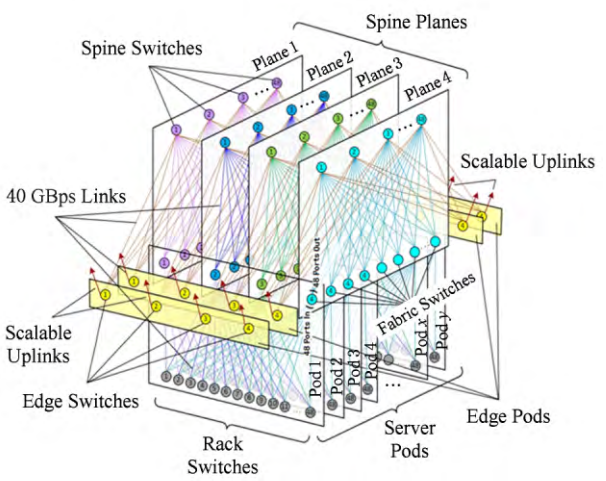


Fig. 5 The data center network of Facebook.

图 5 Facebook 数据中心网络

从图 5 可以看出,该 Facebook 网络结构也是在网络中全面部署了廉价的架顶交换机,但与 Fat-Tree 结构不同的是其优化了服务器之间通信的路由机制. 在 Facebook 网络结构中,集群 Pod 依然是基本的划分单位,包括 48 台 10 GBps 带宽的架顶交换机以及 4 台光纤交换机,其中每台架顶交换机向上以 40 GBps 的上行链路与光纤交换机相连,向下与底层服务器相连. 另外,该结构还设计了 4 个相互独立的 Spine 交换机平面(每个平面可以扩展至 48 台设备).

为了实现所有服务器的互连互通,集群 Pod 中的每一台光纤交换机都会与所在 Spine 平面的所有设备进行互连. 这样的互连结构一方面满足了可扩展性强的实际需求,集群 Pod 和骨干平面 Spine Plane 组成了一个模块化的网络拓扑;另一方面,无论是 Pod 内的通信还是 Pod 间的通信,该结构都提供了较多的传输路径,可以实现比较大的网络对分带宽,最高可扩展至几个 PB.

该结构实际是传统树形层次结构的进一步演化,有 2 点特征:①注重增大服务器与服务器之间的对分带宽,通过巧妙的网络连接减轻了核心交换机的瓶颈限制;②强调模块化构造数据中心的思想,在云计算环境下数据中心的网络规模是不断扩张的,增加 Server Pod 就可实现扩张数据中心的目的,而增加 Spine Plane 就可实现增加服务器与服务器通信对分带宽的目的.

1.1.2 非规则网络结构

除了类似于 Fat-Tree,Facebook 网络结构等这样有组织的、规则的网络体系架构设计之外,研究者也对数据中心无组织的、非规则的网络体系架构展开了大量研究,主要为了解决如何在现有数据中心网络体系架构的基础上进行扩展以满足云计算应用服务的部署需求,这种设计方式既避免了对已有资源的浪费又降低了云计算数据中心的建设成本,REWIRE 就是其中典型的一例.

REWIRE 是针对数据中心网络架构的设计而提出的一种框架,通过使用相应的优化算法,该框架能够设计出一种无组织的、非规则的数据中心网络结构:在满足用户自定义约束的同时实现了最大对分带宽和最小端到端延迟的性能目标. 传统数据中心网络架构的设计通常采用规则的方式,主要原因在于无规则、任意网络结构的管理、运行尚存在很多未解决的问题,如路由、负载均衡等. 但随着相关路由、负载均衡等技术的发展,针对无规则、任意网络结构的高效管理和运行也逐渐变得可行,文献[13-14]从不同的方面探索了无规则、任意网络结构部署应用的可行性.

REWIRE 框架在进行网络架构设计时,根据原有网络是否需要扩展,其优化算法可以有 2 种不同的输入:一种是原有的网络结构;另一种是原有的

网络结构和一些孤立的新的交换机. 如果输入是第1种情况,则 REWIRE 优化算法具体可以分为4步:①暂不考虑网络性能,确定满足用户自定义约束的网络结构空间;②随机选取一候选网络结构作为初始值,并评估其网络性能;③执行 local search 操作,通过修改前一候选网络结构的本地属性值,转向另一更符合需求的候选网络结构;④循环步骤③,直至找到近似最优的网络结构. 在整个过程中,该算法只是优化网络结构的互连关系,并不向网络中增加新的交换机. 如果输入是第2种情况,则首先通过任意连线使原有网络结构和所有孤立的新的交换机之间变得连通,然后按照第1种输入情况的处理方式进行网络结构的设计.

通过比较验证,基于 REWIRE 框架设计出的无规则网络结构具有较优异的网络性能,但是也存在着缺陷与不足,比如在原有网络基础上进行扩展设计时该方法需要一一考虑所有新增交换机的组合,导致设计效率随着新增交换机数量的增长而变低.

1.1.3 采用无线技术的网络结构

随着网络规模的不断扩大、应用服务带宽需求的逐步提高,传统数据中心网络所采用的有线静态链路连接方式面临着巨大挑战:1)网络规模越大,每个服务器平均对应的数据链路就会越多,直接导致管理运营成本的增加;2)由于每个 rack 的实际带宽需求难以预先估计,链路的分配通常只能按照等量的原则来进行,但这样无法有效应对网络中的突发流量;3)有线网络链路的改动成本高,而无线通信技术的快速发展为这些问题的有效解决提供了新思路. 将无线通信技术引入到数据中心的构建中,不仅可以减小实际布线的复杂度、降低建设与运营管理的成本,而且可大幅提高数据中心的网络性能. 目前,针对该问题的研究主要集中在2方面:一是发展无线链路技术,如 radio-frequency[15], 3D-beamforming[16], steered-beamforming[17], Free-Space Optics[18]等;二是探索无线数据中心构建,如 Flyway[19]、FireFly[20]、全无线数据中心等[21]. 以下对这2方面分别进行了描述:

1) 无线链路技术

文献[22]首次提出将60 GHz无线通信技术应用到数据中心网络中,旨在提高网络的对分带宽、减小链路发生拥塞的概率. 选择使用60 GHz射频,主要原因有2点:①60 GHz频段有着接近7 GHz宽的可用频谱,使其能够提供 Gbps 量级速度的多条链路;②60 GHz频段拥有抗干扰、防监听的双重技术优势. 然而,60 GHz无线通信技术也存在先天性的不足[23]:无线信号强度会随着距离的增加而迅速减小,覆盖半径极其有限;在距离较远处,多路效应会对信号的变动产生较大影响. 为了提高无线信号的传输强度,在实际部署过程中研究人员对60 GHz射频的传播进行了改进,采用了有向天线技术,该技术可以隔离不同的无线链接使信号浮动大幅减小.

为了优化数据中心中的无线链路,除了采用有向天线技术,波束成形和波束转向等技术也得到了普遍应用,3D-beamforming 就是其中典型的一例,具体如图6所示. 3D-beamforming 技术比较于普通的波束成形技术,主要有2方面的改进:一是 rack 之间的通信路径采用了反射的点到点形式而没有使用直接的视线链路,避免了链路易受小障碍物阻塞的问题;二是减小了波束之间潜在的相互干扰,拓宽了每条链路的有效范围.

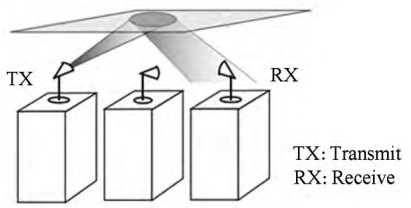


Fig. 6 The schematic diagram of 3D-beamforming.

图6 3D-beamforming 示意图

FSO(free-space optical communication)是研究者优化 rack 之间通信的另一项重要技术,相比于3D-beamforming,由于该技术使用可调制的可见或者红外光束在自由空间传输,因而能够在更大范围内实现数据的高速传输,同时避免信号之间的相互干扰[24].

2) 无线数据中心构建

依据无线通信技术在数据中心中的普及程度,目前可将无线数据中心的构建划分为两大类:①在局部范围内应用无线技术;②在全网范围内应用无线技术.

Flyway 是由美国微软研究院提出的将无线通信技术引入数据中心的著名设计方案,旨在解决网络中部分节点过热的问题,主要思路是在过热的交换机之间添加新的链路、分摊网络流量、实现缓解网络拥塞、提高数据中心整体性能的目的. 如何在保证网络性能的同时实现建设成本的降低,一直以来是数据中心设计的难点,FireFly 设计方案通过引入无线技术在这方面做出了积极探索,其数据中心的构建主要有以下3个突出特征:①rack 间的所有链路

都是无线链路;②rack 间的所有链路都可配置,灵活性高;③网络中只存在架顶交换机,不存在汇聚、核心交换机.另外,通过采用自由空间光通信 FSO 技术以及天花板反射的相关原理[25],FireFly 单跳就可实现任意 2 台 rack 之间的通信,极大地缩小了端到端的通信延迟,从而提高了网络的整体性能.

除在数据中心中部分应用无线技术之外,完全采用无线连接方式在理论和逻辑上也是可行的.基于这种思想,来自美国康奈尔大学和微软的研究团队提出了一种全无线数据中心结构.考虑到无线链路有限的传播半径,以及为端到端提供多条冗余链接的需要,该结构采用了圆柱形的机架组织方式,具体如图 7 所示.可以看出,环形排列的所有服务器形成了一紧密的连接网,相比于传统有线数据中心,其端到端的平均延迟将会显著减少,总带宽将会更大.另外,由于无线链接自身的灵活性,全无线数据中心将会有更强的容错能力.

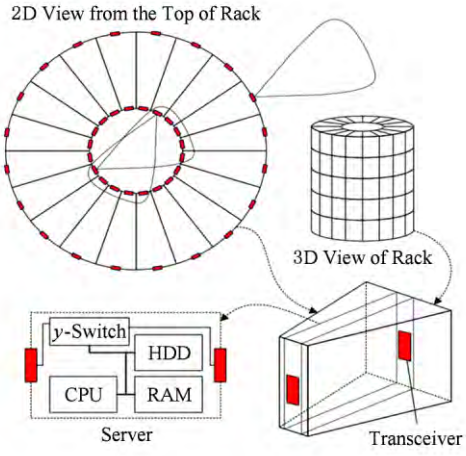


Fig. 7 The wholly wireless data center architecture.

图 7 全无线数据中心结构

1.2 以服务器为中心的设计方案

除了上述以交换机为转发中心的网络架构设计,以服务器为转发中心的网络架构设计也是一个研究热点.根据设计过程中所使用服务器的端口数目,该研究又可以被分为两大类:以多端口的服务器为转发中心和以双端口的服务器为转发中心.

1.2.1 以多端口为转发中心的网络结构

DCell[26]、BCube[27]、雪花结构[28]、文献[29]都是以多端口的服务器为设计出发点,分别提出了具有扩展性强、对分带宽大等特性的不同新型网络体系架构,满足了云计算数据中心网络越来越多应用服务的部署需求.

1) DCell 网络结构

考虑到树形层次结构的缺陷与不足,DCell 以一种完全不同的设计方式(以服务器为转发中心)进行网络结构的构造,如图 8 所示为 DCell_1 的网络结构.

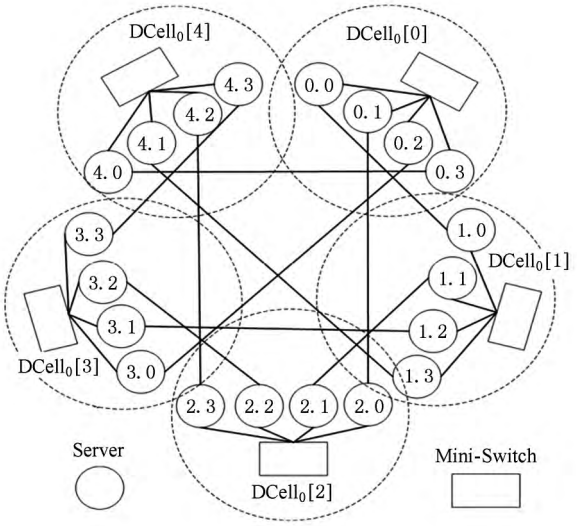


Fig. 8 The data center architecture of DCell_1.

图 8 DCell_1 网络结构

DCell 网络结构采用了递归的方式来定义,即高等级的 DCell 网络结构由低等级的 DCell 网络结构按照一定的连接方式构成,假设 DCell_{k-1} 中有 t_{k-1} 个 servers,则 DCell_k 就由 $t_{k-1}+1$ 个 DCell_{k-1} 采取全连接的方式构成,如果在 DCell_0 中 $t_0=4$,那么 DCell_1 就由 5 个 DCell_0 构成.依据 DCell 的定义方式,DCell_k 服务器数量的计算公式为 $g_k=t_{k-1}\times(t_{k-1}+1)$,当 $k=3$,$n=6$ 时,服务器数量可以达到 3.26×10^6,表明了 DCell 网络具有极好的可扩展性,能够适应日益增大的数据中心网络规模.另外,通过理论分析,DCell_k 网络的对分带宽可以达到 $\frac{t_k}{4t_k\log_n t_k}$,为服务器相互之间的通信提供了强有力的支撑.

2) BCube 网络结构

同 DCell 类似,BCube 网络结构也采用递归的定义方式,如图 9 所示为 Level_1 BCube 网络结构.

Level_k BCube 网络结构由 nk 个小交换机和 n 个 Level_{k-1} BCube 网络结构组成,其中每个交换机与每个 Level_{k-1} BCube 都有连接,并且第 i 个交换机只连接每个 Level_{k-1} BCube 的第 i 个服务器.通过这种构造方式,BCube 网络也能够提供较好的可扩展性以及较高的网络对分带宽.

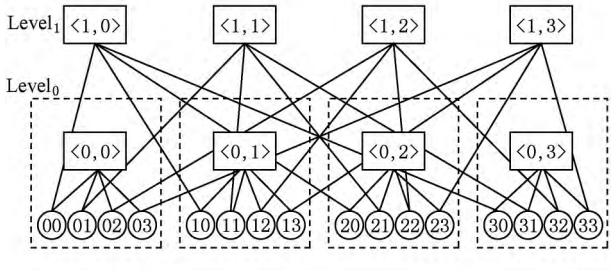


Fig. 9 The data center architecture of level$_1$ BCube.

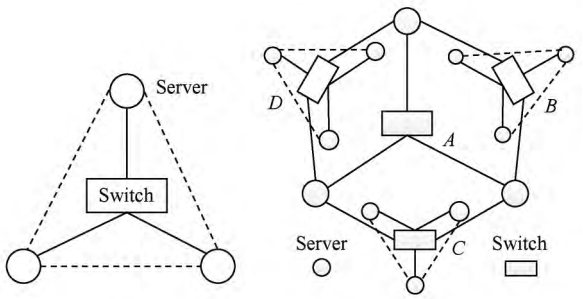
图 9 Level$_1$ BCube 网络结构

在 BCube 网络中，任意 2 台服务器之间存在 $k+1$ 条并行的通信路径. BCube 网络使用 $k+1$ 维行向量对所属的服务器进行标识，如服务器 A 可被表示为 $a_k a_{k-1}\cdots a_0$，其中 a_i 表示 A 在 Level$_i$ BCube 网络结构中的坐标. 基于服务器的标识方法，这 $k+1$ 条并行通信链路可理解为由 2 部分组成：①只考虑通信两端服务器标识中对应维相同的部分，计算通信路径；②只考虑通信两端服务器标识中对应维不相同的部分，计算通信路径. 以 Level$_1$ BCube 为例，假设 $A=00$，$B=30$，如果只考虑 $a_i=b_i$，则 A 到 B 的路径为 00→01→31→30；如果只考虑 $a_i\neq b_i$，则 A 到 B 的路径为 00→30，即 Level$_1$ BCube 为服务器 A 与 B 提供了 2 条通信路径. BCube 网络中服务器之间的多路径为网络流量的负载均衡提供了有效的支撑机制，缓解了网络拥塞的发生，从而保证了应用服务的性能质量.

3) 雪花结构

雪花结构同样采用递归定义的方式，n 级雪花结构可通过在 $n-1$ 级雪花结构上添加若干个 0 级雪花结构来实现. 0 级雪花结构由一个微型交换机和 k 个服务器组成，k 通常取为 3～8，如图 10(a)所示为 0 级雪花结构.

在雪花结构的具体构造过程中，有 2 个比较重要的概念：虚连接和实连接. 在图 10(a)的 0 级雪花结构中有 3 条服务器之间连接的虚线，这 3 条虚线就是所谓的虚连接，虚连接实际并不存在. 相对于虚连接，实连接是真实存在的连线，当雪花结构进行扩展时，断开虚连接并添加 0 级雪花结构，所形成的 2 条虚线都连往新添 0 级雪花结构的交换机，由这 2 条虚线所形成的网络链接称为实连接. 如图 10(b)所示为由 0 级雪花结构扩展成的 1 级雪花结构，可以发现图 10(a)中有 3 条虚连接和 0 条实连接，图 10(b)中有 6 条虚连接和 6 条实连接. 当雪花结构向更高等级扩展时，需断开每一处虚连接和实连接，然后按照规则再添加 0 级雪花结构.



(a) Level$_0$ snowflake architecture (b) Level$_1$ snowflake architecture

Fig. 10 The snowflake architecture.

图 10 雪花结构

可以验证，n 级雪花结构包含的服务器数目为 $k\times(k+1)^n$，即该结构服务器的数量会随着 n 值的增加而呈指数次方不断增长，满足了云计算数据中心对大规模网络的需求. 另外，当 k 取 3～8 时，n 级雪花结构任意 2 台服务器之间的最短路径将不超过 $2n+1$ 跳，而且它们之间存在至少 2 条、至多 2^{2n} 条并行路径，既保证了网络具有较小的网络直径，也保证了服务器之间较大的网络对分带宽.

1.2.2 以双端口为转发中心的网络结构

虽然以多端口为转发中心的网络架构设计方案具有很好的网络属性，如可扩展性强、对分带宽大等，但是大量多端口服务器的构造需要额外添加不少硬件，相对于双端口服务器而言，不具有普遍性而且成本比较高. 针对这些问题，研究者对基于双端口服务器网络架构的设计展开了积极探索，目前具有代表性的方案有 DPillar[30]，HCN[31]，BCN[31]，SWCube[32]，SWKautz[32]，FiConn[33]等.

1) DPillar 网络结构

相比于传统数据中心网络架构的设计，DPillar 网络结构的构造有 2 方面的优势：①大量采用了商用的 2 层交换机，容易获取且价格相对低廉；②使用双端口的服务器作为转发中心，而现有的商业服务器本身就拥有 2 个万兆以太网端口，因而 DPillar 结构以很小的部署代价就可以实现网络的大规模扩展.

尽管使用的服务器仅仅只有 2 个端口，DPillar 结构却为服务器之间的通信提供了富余的链路，有效地支撑了带宽密集型应用服务的部署. 在 DPillar 网络的构造过程中，服务器和交换机会被分别划分为 k 列，并以特定的连接规则进行互连，因而使用 (n,k) 就可以对 DPillar 网络进行唯一地标识，其中 n 表示交换机的端口数.

假设 k 列服务器分别记为 $H_0 \sim H_{k-1}$,k 列交换机分别记为 $S_0 \sim S_{k-1}$,DPillar(n,k)网络将 k 列服务器和 k 列交换机交替排列形成环状结构,具体如图 11 所示,其中服务器每列有 $(n/2)^k$ 台,交换机每列有 $(n/2)^{k-1}$ 台. 对于 H_i 中的每一台服务器,其一个端口连接 S_i 列的交换机,一个端口连接 $S_{(i+k-1)\%k}$ 列的交换机;对于 S_i 列的每一台交换机,其 $n/2$ 的端口连接 H_i 列的 $n/2$ 台服务器,另外 $n/2$ 端口连接 $H_{(i+1)\%k}$ 列的 $n/2$ 台服务器.

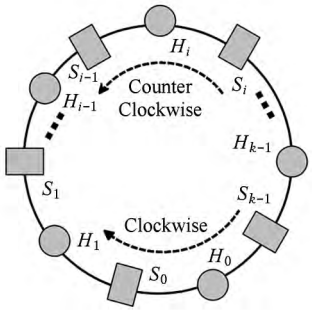


Fig. 11 The architecture of DPillar(n,k).

图 11 DPillar(n,k)网络结构

在 DPillar(n,k)网络中,每一台服务器都可以使用(C,$V^{k-1}\cdots V^0$)序列进行唯一地标识,其中 C 表示该服务器位于 H_c 列,$V^{k-1}\cdots V^0$ 表示该服务器在 H_c 列的坐标位置. 基于服务器的标识(C,$V^{k-1}\cdots V^0$),DPillar 网络的具体互连关系可以进行如下描述:将 H_c 列和 $H_{(c+1)\%k}$ 列的 $2(n/2)^k$ 台服务器划分成 $(n/2)^{k-1}$ 组,设服务器的标识为(C,$V^{k-1}\cdots V^c\cdots V^0$),在不考虑 V^c 位的情况下,每组 n 台服务器的其他位都相同;对于划分的每一组,其 n 台服务器都连接于 S_c 列的同一交换机. 如图 12 所示为 DPillar(8,2)网络结构的 2 维视图.

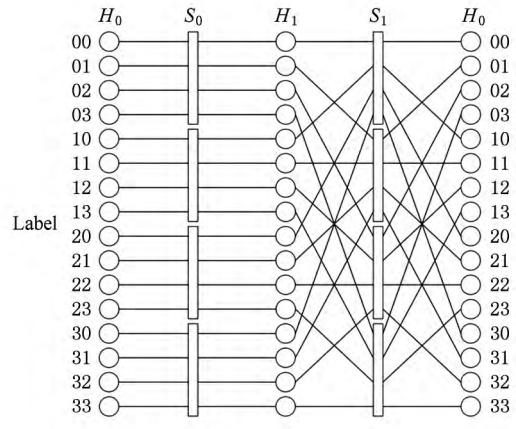


Fig. 12 The 2D-View of the DPillar(8,2) architecture.

图 12 DPillar(8,2)网络结构的 2 维视图

DPillar(n,k)网络包含的服务器数量为 $k(n/2)^k$,且对分带宽可以达到 $(n/2)^k$,以 $n=48$,$k=5$ 为例,其支持的服务器就多达 4×10^6 多台,表明了 DPillar 网络能够实现大规模扩展且可以提供较高的对分带宽. 除此之外,由于网络结构的对称性,DPillar 可以为服务器之间的通信提供简单高效的路由机制. 对于任意一对服务器,假设源为(C_s,L_s),目的为(C_d,L_d),二者之间的通信可以被分为 2 个阶段:helix 阶段和 ring 阶段. 在 helix 阶段,数据包从源(C_s,L_s)转发至中间节点(C_i,L_d),之间最多有 k 跳;在 ring 阶段,数据包从中间节点(C_i,L_d)转发至目的(C_d,L_d),之间最多有 $k/2$ 跳.

2) HCN 和 BCN 网络结构

HCN 和 BCN 网络结构以双端口的服务器为转发中心,二者都实现了较小的网络直径和较高的对分带宽. 其中,HCN 网络规则化程度高、对称性好且可扩展性较强,适合于模块化数据中心的设计;BCN 网络可以为节点之间的通信提供多条互不相交的路径,具有较好的容错能力,而且在节点度为 2、网络直径为 7 的情况下,BCN 是目前规模最大的网络结构.

HCN 网络结构采用了递归的定义方式,即 HCN(n,h)结构由多个下一级 HCN(n,$h-1$)结构组成,其中 n 表示交换机的端口数目,h 和 $h-1$ 表示 HCN 结构的等级. HCN(n,0)是 HCN 网络结构的基本构造单元,由 1 台交换机和 n 台双端口服务器组成. 对于 HCN(n,h)结构的具体构建过程,可以进行如下简单描述:由 n 个 HCN(n,$h-1$)结构组成,且每个 HCN(n,$h-1$)结构通过全相连的方式与其他 $n-1$ 个同等级的结构相连,由于每个 HCN(n,$h-1$)结构有 n 个可用的服务器端口,则不同 HCN(n,$h-1$)结构互连之后所构建的 HCN(n,h)网络仍然有 n 个可用的服务器端口用于进一步扩展,如图 13 为 HCN(4,2)网络结构.

HCN 网络结构虽然具有很好的网络属性,但就对分带宽而言,其仍然存在局限性. 通过对 HCN 网络结构的优化与改进,研究者提出了一种 2 维分层的 BCN 网络结构,该结构同样采用递归的定义方法. BCN(α,β,0)是 BCN 网络结构的基本构造单元,由一台 n 端口的交换机、α 台主服务器和 β 台从服务器组成,其中 $\alpha+\beta=n$. BCN 网络结构分为 2 个维度进行扩展:第 1 维度使用主服务器进行扩展;第 2 维度使用从服务器进行扩展. 第 1 维度的扩展与 HCN

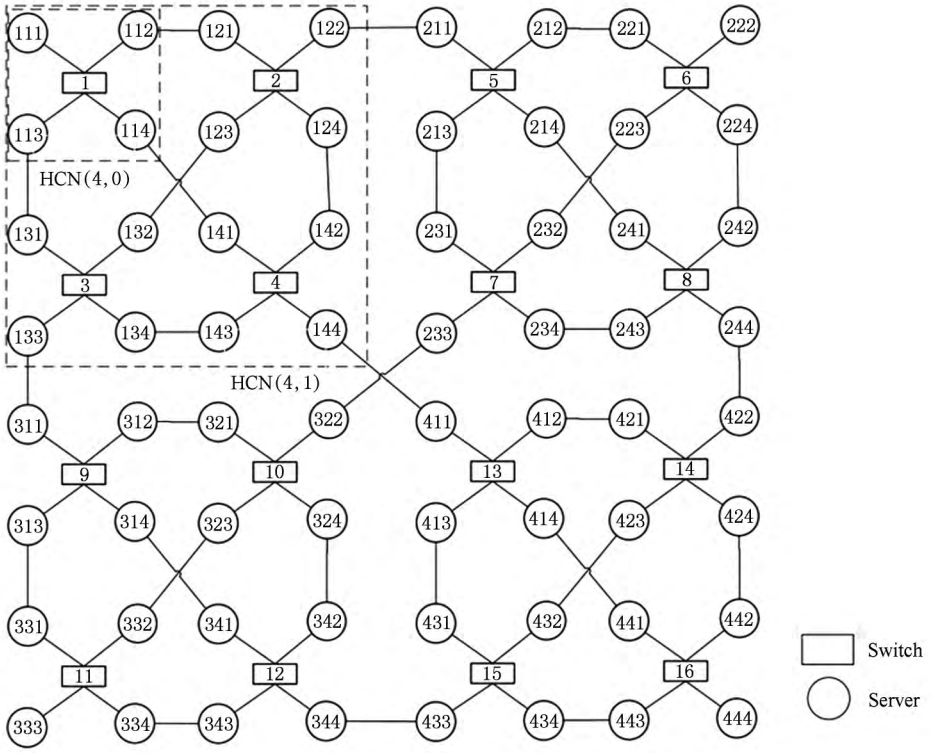


Fig. 13 The architecture of HCN(4,2).

图 13 HCN(4,2)网络结构

结构的扩展类似,而且每一级的 BCN(α,β,i)结构都存在 α 个可用端口用于下一级的扩展. 假设第 2 维度扩展的基本构造单元为 BCN(α,β,γ)(该结构含 $\alpha^{\gamma}\beta$ 台从服务器),使用 $\alpha^{\gamma}\beta+1$ 个 BCN(α,β,γ)结构进行全相连. 根据上述构造过程,可以看出 BCN 网络在第 1 维度尚可扩展,在第 2 维度已经不能扩展. 记 2 维分层的 BCN 网络为 BCN(α,β,h,γ),其中 h 表示 BCN 网络在第 1 维度扩展到了 h 级,γ 表示 BCN 网络在第 2 维度的基本构造单元为 BCN(α,β,γ). 如果 $h<\gamma$,则需先将第 1 维度扩展到 γ 级才可进行第 2 维度的扩展;如果 $h>\gamma$,由于 BCN(α,β,h)网络中会有 α^{h-r} 个 BCN(α,β,γ)结构,为了避免第 2 维度互连所使用的 BCN(α,β,γ)成为性能瓶颈,对 α^{h-r} 个 BCN(α,β,γ)结构进行编号,同编号的 BCN(α,β,γ)在第 2 维度都进行全相连. 如图 14 为 BCN(4,4,1,0)网络结构,其中黑色圆圈代表主服务器,白色圆圈代表从服务器.

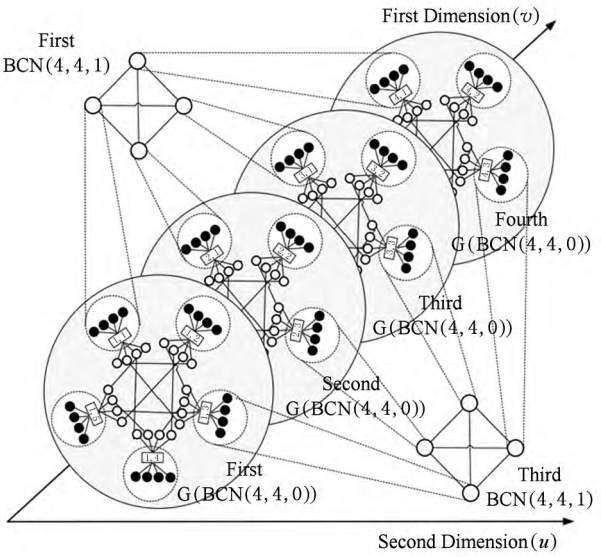


Fig. 14 The architecture of BCN(4,4,1,0).

图 14 BCN(4,4,1,0)网络结构

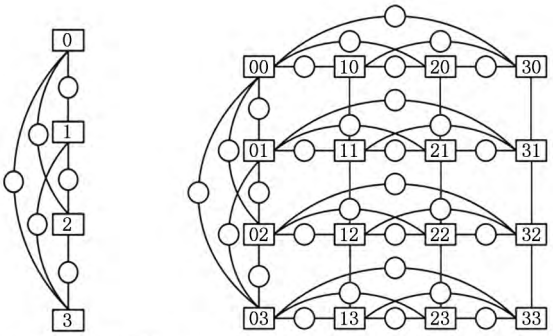
相比于 HCN 网络,BCN(α,β,h,γ)网络在任意服务器之间都提供有 $\alpha-1$ 条并行通信链路,无论 h 大于还是小于 γ,都有效地提高了网络的对分带宽. 此外,BCN(α,β,h,γ)网络具有较小的网络直径,在 $h<\gamma$ 时最多为 $2^{\gamma}+2^{h}-1$,其中 h 和 γ 都是很小的整数.

3) SWCube 和 SWKautz 网络结构

BCN 和 DPillar 网络结构都具有较强的可扩展能力,但是在给定条件下,如网络直径或者交换机的端口数目取恒定值,其所容纳的服务器数量远远没有达到理论的上限值[34]. 在给定网络直径长度 d 和交换机端口数目 n 的情况下,如何进行网络架构的

设计才能覆盖更多的双端口服务器,针对这一问题,研究者提出了2种新颖的网络架构SWCube和SWKautz.

SWCube网络结构的设计以Hypercube为基础,具体的构造可描述如下:①将Hypercube结构中的节点替换成交换机;②在交换机与交换机之间插入服务器.假定记SWCube网络结构为SWCube(r,k),其中k表示Hypercube的维数,r表示Hyper-cube每一维的基数,如图15所示分别为SWCube(4,1)和SWCube(4,2)网络结构.可以证明,在网络直径d和交换机端口数目n确定的情况下,SWCube(r,k)网络结构所包含的服务器数量可达到$n(n/(d-1)+1)^{d-1}/2$.当取$n=16$,$k=8$时,服务器的数量就可达52 488台之多.

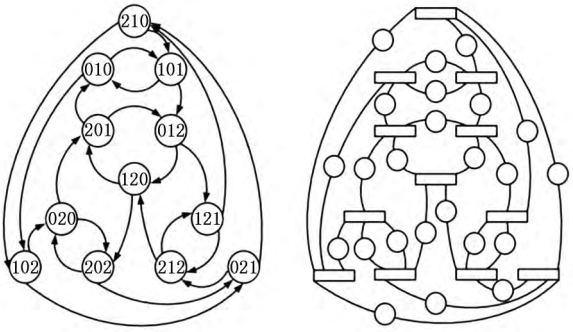


(a) The SWCube(4,1) architecture (b) The SWCube(4,2) architecture

Fig. 15 The SWCube architecture.

图15 SWCube结构

SWKautz网络结构的设计以Kautz graph为基础,具体的构造可描述如下:①将KA(n/2,k)图中的每一个节点都替换成n端口的交换机;②去除交换机与交换机之间链路的方向;③在2交换机之间插入一双端口的服务器.假定记SWKautz网络结构为SWKautz (r,k),其中k表示Kautz graph的维数,r表示Kautz graph节点标识的每一位取值空间的大小,如图16为KA(2,3)和SWKautz (2,3)结构.



(a) KA(2,3) (b) SWKautz(2,3)

Fig. 16 KA(2,3) and SWKautz(2,3).

图16 KA(2,3)和SWKautz(2,3)网络结构

可以证明,SWKautz(n/2,k)网络的网络直径为$k+1$,并且在网络直径为d、交换机端口数目为n的情况下,其可支持的服务器数量为$(n/2)^d+(n/2)^{d-1}=(n/2)^k+(n/2)^{k+1}$,如$n=16$,$k=5$时,服务器的数量多达30万台左右,体现了该结构具有较强的可扩展性.

综合以上关于云计算数据中心网络结构设计的概述,表1就多个不同指标对其进行了对比分析.可以看出,以服务器为中心的设计方案相比于以交换机为中心的设计方案,在网络规模、带宽、扩展性、灵活性等方面可以实现较好的性能,特别是以双端口服务器为中心的设计方案;在布线复杂性和成本方面,无线的设计方案显得更优,而以多端口服务器为中心的设计方案虽然布线复杂性优于以双端口为中心的设计,但是后者在成本方面却略胜一筹;由于以服务器为中心的设计方案不同于传统的设计模式,地址、路由等都不同,因而配置开销比较高.

Table1 Comparison of Data Center Network Architecture for Cloud Computing

表1 云计算数据中心网络体系架构的对比

Category			Property						
			Size	Bandwidth	Scalability	Flexibility	Wiring Complexity	Cost	Configuration Overhead
Switch as the Forwarding Center	Wired	Regular Layered	Small	Little	Bad	Low	Medium	High	High
		Regular FatTree	Medium	Medium	Medium	Medium	Higher	Higher	Higher
		Regular Facebook Architecture	Medium	Large	Medium	High	Higher	Medium	High
		Irregular REWIRE	Medium	Medium	Medium	Low	Medium	Medium	Higher
	Wireless	FlyWay	Medium	Medium	Medium	High	Low	Medium	High
		FireFly	Medium	Large	Good	High	Low	Low	Higher
		Completely Wireless	Medium	Large	Bad	High	Low	Low	Higher

Continued (Table1)

Category			Property						
			Size	Bandwidth	Scalability	Flexibility	Wiring Complexity	Cost	Configuration Overhead
Server as the Forwarding Center	Multi-Port Server	DCell	Large	Larger	Better	Higher	High	Higher	Higher
		BCube	Small	Large	Better	Higher	High	Higher	Higher
		Snowflake	Large	Larger	Better	Higher	High	Higher	Higher
	Dual-Port Server	DPillar	Medium	Large	Better	Higher	Higher	High	High
		HCN	Medium	Large	Better	Higher	High	High	High
		BCN	Large	Large	Better	Higher	High	High	High
		SWCube	Large	Larger	Better	Higher	Higher	High	Higher
		SWKautz	Large	Larger	Better	Higher	Higher	High	Higher
		FiConn	Large	Larger	Better	Higher	High	High	Higher

2 云计算数据中心 VMotion 的支撑

在云计算环境下,数据中心基础设施提供商 InP 响应来自服务提供者 SP 的应用资源请求. 对于 InP 而言,一方面需要满足 SP 所请求的应用资源,另一方面需要保证自身计算资源(CPU、存储、网络等等)的高利用率,这就要求云计算数据中心对计算资源的调度要有足够的灵活性,而这种灵活性很大程度上体现在虚拟机(virtual machine, VM)的动态迁移. 但是,由于传统通信实体标识 IP 地址同时绑定了应用服务的位置信息和身份信息,如果一台 VM 发生漂移,那么对 IP 地址的处理将会变得很复杂. 如果修改 IP 地址,其所代表的应用服务身份信息将会改变,需要重新对其进行识别;如果不修改 IP 地址,该 VM 在新的网段内将无法正常通信. 针对如何实现 VM 的灵活迁移,目前的研究思路主要有 2 类:1)改进传统的 2 层网络,研究可以支持更大规模服务器通信的云 2 层网络;2)在不改变现有 2 层网络通信机制的情况下,研究如何跨越 3 层交换机或路由器来实现 VM 的迁移.

2.1 "大二层"网络体系架构

考虑到 VLAN 对网络资源的隔离以及传统通信实体标识 IP 地址对身份信息与位置信息的同时绑定,解决"VM 迁移"最直接的思路就是扩大 2 层网络. 但是,无论从 2 层生成树协议 STP 的缺陷来讲,还是从 2 层交换机转发表自学习不利于扩展的角度来讨论,对传统 2 层网络不能仅仅是简单地扩大即可,而需要对原有的通信体制进行优化改进,使其满足云计算环境下的资源迁移需求. 依据其所依赖的具体实现技术,"大二层"网络可划分为 3 类:1)基于位置/身份相分离技术的"大二层"网络;2)基于改进 2 层控制平面方法的"大二层"网络;3)基于 VPN 隧道的跨数据中心"大二层"网络.

2.1.1 位置/身份分离技术

为了实现灵活、可扩展性强的数据中心网络,VL2[35]破除了传统 IP 地址对身份和位置信息的绑定,设计了 2 套地址族,分别为 AA(application address)和 LA(locator address),其中 AA 不再有拓扑含义,仅仅作为名字使用,而 LA 仅用于数据包的路由. 在源站点处,数据包添加 AA 地址后被发往源架顶交换机;由源架顶交换机查询 LA,并依据 LA 发往目标架顶交换机;目标架顶交换机去除数据包外层的 LA,基于 AA 最终发往目标站点. 很显然,AA-LA 映射关系在数据包的传输过程中起着重要的枢纽作用,因而,学习、维护有效的 AA-LA 映射关系十分关键. 在 VL2 网络中,AA-LA 间的映射关系由一可扩展的、可靠的目录服务系统来维持,具体如图 17 所示. 该目录服务系统使用了 2 级层次结构,第 1 级处理读优先的查询操作,包括 2 步:请求查询和回复;第 2 级处理写优先的更新操作,包括 6 步:请求更新、转发更新请求到 RSM(replicated state machine)、RSM 间信息同步、RSM 返回确认、DS(directory server)返回确认和 DS 间分发更新后的 AA-LA 映射关系.

不同于 VL2,PortLand[36]虽然采用的依然是位置信息和身份信息相分离的设计思路,但并没有使用 2 套地址族,而是将表征位置信息的地址族使用伪 MAC 地址 PMAC 进行取代. 对于网络中的每一台主机,PortLand 都会为其分配唯一的 PMAC,该 PMAC 包含了主机在网络中的位置信息. 为了实现基于 PMAC 的网络通信,主机在获取 PMAC 的

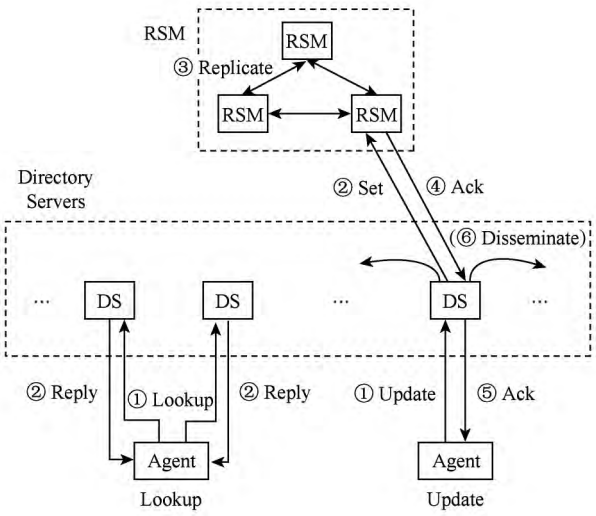


Fig. 17 The VL2 directory service system.

图 17 VL2 目录服务系统

同时会生成 MAC-IP-PMAC 三者的映射关系,并将该映射关系发往网络的集中管理器 Fabric Manager(负责管理和维护全网的 MAC-IP-PMAC 映射).当数据包发向网络时,源交换机通过查询 Fabric Manager,将数据包的源 MAC 和目的 MAC 改写为源 PMAC 和目的 PMAC,并基于 PMAC 进行数据包的转发,到达目标交换机后再进行 PMAC 向 MAC 的转换.

VL2 和 PortLand 都是依赖于集中式方法来管理身份信息和位置信息之间的映射关系,在网络规模逐渐扩大的情况下,这种方式使用十分受限.LISP[37]同样是基于位置/身份分离技术来研究"大二层"网络的,其 2 类地址信息可描述为:表明位置的路由标识符 RLOCs(routing locators)和表明身份的节点标识符 EIDs(endpoint identifiers).为了在新地址机制下完成通信,LISP 为网络中的每一个站点都分配有独立的 EID,并引入了 2 种新的网元:ITR(ingress tunnel router)和 ETR(egress tunnel router),二者部署在 LISP 网络的边界.站点数据包的转发采取的是 Map-and-Encap 机制,并由 ITR/ETR 完成 EID 与 RLOCs 映射关系的转换.为了在大规模网络环境下高效地实现 EID 与 RLOCs 之间的映射,LISP 引入了 LISP-ALT(LISP alternative topology)概念,即架构于基础网络之上的 LISP 覆盖网.该覆盖网包含网络中所有的 ITR 和 ETR,并通过在这些节点之间运行 BGP 协议,实现 EID 可达信息的交换,最终建立全网的 EID-RLOCs 映射关系表,该方式较 VL2 和 PortLand 的集中管理方式更具可扩展性.

2.1.2 改进 2 层控制平面

针对 2 层网络的广播风暴、网络环路以及交换机缓存小等传统问题,研究者将 3 层路由技术引入 2 层网络,期望通过改进控制平面实现 2 层网络的多链路负载均衡、快速收敛及高可扩展性等特点,并最终达到扩张的目的.

Monsoon[38]网络架构为了避免内部服务器资源被分离以及链路存在收敛比等情况,将网络中的所有服务器连接于同一个 2 层域中.Monsoon 强化了控制平面的作用,在禁用 ARP、自学习算法的基础上,建立了维护 IP-MAC 映射关系的目录服务系统.初始情况下,Monsoon 由每台架顶交换机维护着其自身所连主机的 IP-MAC 映射关系;随后,通过运行链路状态协议 LSA 实现所有架顶交换机 IP-MAC 映射信息相互之间的交换,从而建立起全局的 IP-MAC 映射目录服务系统.Victor[39]体系架构同样采用此种设计思路实现了 VM 在不同网络之间的灵活迁移.Monsoon 架构通过优化 IP-MAC 的学习方式,可以实现更大规模的 2 层网络,但是由于网络设备的 MAC 地址无规律可循,导致运行链路状态协议所学到的路径并非最优,而且将其调整成最优的代价很大.

FabricPath[40]是 Cisco 针对"大二层"环境所设计的一个技术方案,与 Monsoon 相比,其主要的改进表现在新定义了全新的地址空间:switch ID.对于 FabricPath 网络中的每一台设备,其在初始接入时都会被分配唯一的 switch ID,用于网络设备身份的标识,也作为数据包路由寻址的依据.当数据包发送到 FabricPath 网关时,原数据帧会被添加新的帧头,该新帧头除包括源、目的 switch ID 外,还包括 TTL 值.TTL 字段类似于 IP 包头的 TTL,其作用是为了防止数据帧在网络中被无限次的转发.如图 18 所示为 FabricPath 的组网图.类似于 Monsoon 不依赖于 MAC 地址进行寻址,FabricPath 网络也通过运行链路状态协议(IS-IS 协议),为网络设备建立全网拓扑图,以计算最优路径来进行数据包的转发.此处,IS-IS 协议使用的路由地址为 switch ID,其比 MAC 地址更有规律性,从而降低了节点之间最优路径选择的复杂度.新的地址空间加上 IS-IS 协议,FabricPath 为 2 层网络建立起了拥有更多功能特征的控制平面,可以有效地实现多链路负载均衡和较强的可扩展性.虽然维护简单且性能优越,但由于 FabricPath 定义了新的地址空间和帧头,而

传统设备并不支持这些功能,因此为了部署 FabricPath 就需要额外购置新的硬件设备,成本较高.

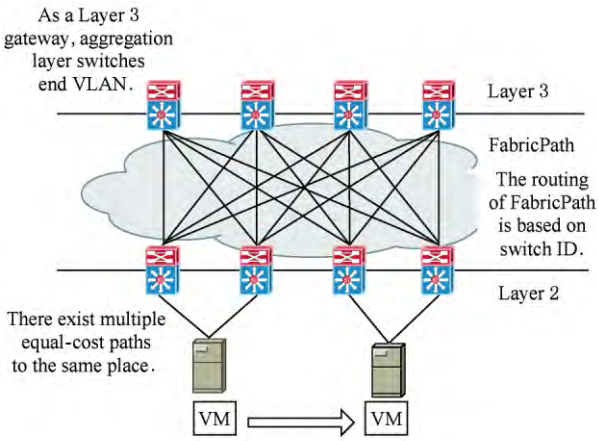


Fig. 18 The FabricPath network.

图 18 FabricPath 组网

同 Monsoon, FabricPath 一样, SPB[41] (shortest path bridging)也是瞄准未来 2 层网络发展而研发的新技术. 在控制平面, SPB 依然采用 IS-IS 协议来保证节点之间数据传输的最优路径. 但是,在数据平面 SPB 并没有设计新的帧头结构,而是采用了复用现有以太网技术的 2 种转发策略:基于 VID 的最短桥接路径(shortest path bridging VID, SPBV)和基于 MAC 的最短桥接路径(shortest path bridging MAC, SPBM),这样就无需破坏现有以太网帧结构,对于网络设备而言,理论上只需升级软件即可,部署成本较低.

2.1.3 跨数据中心的"大二层"互连

在云计算环境下,资源的灵活分配不仅仅局限于数据中心内部,在数据中心之间也存在这种需求,这就要求 2 层网络能够跨广域网进行延伸.

虚拟私有网络服务(virtual private LAN service, VPLS)[42]搭建于 MPLS 核心网之上,并基于隧道技术实现了 2 层网络的跨域延伸. 在 VPLS 网络中,主要包含有 3 类设备:CE(customer edge)设备、PE(provider edge)设备和 P(provider)设备,其中 PE 设备为核心,一方面要负责 MPLS 隧道的建立,另一方面又要负责 MAC 地址学习. VPLS 网络的基础为 MPLS 的 Full Mesh 连接,因为数据包的转发依赖于 MPLS 所提供的交换能力. 在 Full Mesh 之上,VPLS 利用伪连接 PW(pseudo wire)建立起了 PE 之间的邻接关系,从而形成了覆盖所有 PE 的 VPN 网络. 为了实现 2 层数据的转发,每个 PE 都保存着一个 VPLS 实例,该实例记录了同属于一个局域网的所有 PE 节点,并维护了相对应的所有 MAC 地址信息. 正是 PE 设备对 PW 邻居关系的维护以及整个局域网 MAC 地址的记录,使 VPLS 得以为不同地理位置的节点提供跨广域网的 2 层互连. 但是,VPLS 只考虑了以太网的扩张,并没有对广播风暴、网络环路等传统以太网问题进行优化,这将会造成广域网链路资源的极大浪费.

上层传输虚拟化(overlay transport virtualization, OTV)[43]针对 VPLS 缺乏对以太网进行优化的不足,通过进一步强化 2 层网络的"控制平面",同时减弱对"数据平面"的要求,在优化以太网、增加灵活性的基础上实现了 2 层网络的扩张. 在数据平面,OTV 简化了数据帧对传输链路的要求,即由本地发往广域网的数据帧只进行 UDP 封装即可,从而使其应用范围更加广泛. 在控制平面,OTV 基于 IS-IS 协议对 MAC 进行寻址,避免了使用 ARP 所造成的广播风暴等,另外,依据广域网是否支持组播,OTV 将控制平面的报文交互机制设置为 2 种:基于组播的信息同步和无组播环境下基于邻接服务器的信息同步,以提高 MAC 地址表同步的效率.

VPLS 和 OTV 都实现了 2 层网络的跨域延伸,且 OTV 对延伸后的 2 层网络进行了优化,但考虑到广域网的流量负载均衡、技术的开放性和成熟程度以及运维管理的成本等,跨数据中心的 2 层互连还值得进一步研究.

2.2 跨越 3 层的网络体系架构

在云计算环境下,为了实现如"VM 迁移"等网络资源的灵活分配,除了针对扁平化的"大二层"网络进行研究外,探索跨越 3 层的资源迁移方式也是另一种思路,典型的代表有 VXLAN[44], NVGRE[45]等.

VXLAN 有 2 个设计目标:一是破除虚拟化部署对广泛 2 层网络的要求;二是克服 VLAN 缺陷,并对 2 层标签技术进行彻底改革,为此,其主要有 3 方面的改进工作:

1) VXLAN 的数据平面

为了跨越 3 层网络, VXLAN 采用了隧道机制,在现有网络之上搭建了一叠加网络. 在这种机制下,数据包的转发过程可以描述为:首先, VXLAN 在网络中定义了 VTEP(VXLAN tunnel end point)实体;其次,由该实体对虚拟机产生的数据包添加 VXLAN 包头(UDP 包头);在叠加网中,数据包依据 VXLAN 包头进行转发,不再依赖虚拟机本身的 MAC 地址和 VLAN 信息. 基于隧道机制有 2 点好处:一是减小了对现网的改动,只需在需要通信的链

路两端部署一对封装设备即可,不需要改动中间设备;二是能够支持网络的快速变更,隧道在建立和拆除时都不会对其他网络链路造成影响.

2) 定义了新的包头

VXLAN 包头如图 19 所示,从外到内依次包括了两端通信实体 VTEP 的 MAC 地址、IP 地址以及 VXLAN 标签. VXLAN 标签的功能类似于 VLAN 标签,只不过 VXLAN 可以表示更多网段,其使用 24 b 的 VNI(VXLAN network identifier)来标识网段,只有具有相同 VNI 的虚拟机才可以互相通信.

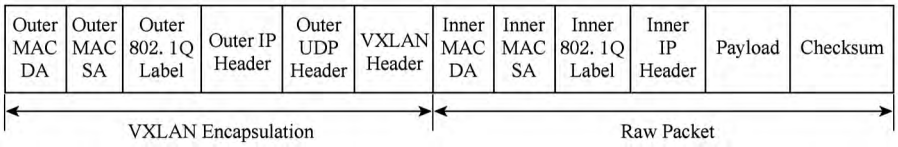


Fig. 19 The structure of VXLAN frame.

图 19 VXLAN 帧结构

3) VXLAN 的控制平面

由于采用了隧道机制,VXLAN 控制平面的一项主要工作是记录虚拟机、VNI 以及 VTEP 的对应关系. 通过这些对应关系,到达 VTEP 的数据包就能够准确地找到虚拟机的位置. 对于不认识的 MAC 地址,VXLAN 控制平面依靠类似广播的行为来获取路径信息,所谓类似广播的行为是 VXLAN 对传统广播行为的一个改进,具体描述为:VXLAN 选择使用 IP 组播来承载 2 层的广播流量,所有的 VTEP 都会被加入到一个特定的 IGMP 组播组,这样 ARP 请求等广播报文就会被局限在组播范围内,并不会影响到其他不相关的网络设备,从而提高了链路的使用效率,避免了网络拥塞的发生.

NVGRE 是 Microsoft 提出的研究方案,与 VXLAN 实现了类似的功能,与其不同之处在于 NVGRE 采用了 GRE 技术来搭建隧道,由于许多交换机芯片都支持 GRE 隧道功能,这就很便于 NVGRE 的普及. 但 NVGRE 未在添加的 NVGRE 包头中包含原始 2 层帧头的 Hash 结果,就没有能够同 VXLAN 一样提供多条路径上的负载均衡.

对云计算数据中心 VMotion 支撑技术的研究,极大地促进了网络资源的灵活部署,使得对底层物理网络的利用变得更加充分.

表 2 对 VMotion 相关支撑技术的优缺点进行了比较分析,表明"大二层"网络并非虚拟化软件部署的必要条件,同时也显示了 VMotion 支撑技术的发展有强化控制平面功能、保留或弱化数据平面功能的趋势.

Table 2 Comparison of the Supporting Technologies Concerning VMotion of Data Center for Cloud Computing

表 2 云计算数据中心 VMotion 支撑技术优缺点比较

Category		Advantages	Disadvantages	Instances
The Network Architecture of Big Two Layer	Locator/Identifier Separation Technology	A single application isn't tied to one server no longer and maybe resides on multiple servers, so mobility becomes simple and flexible.	It needs to maintain the mapping between identity information and location information at aggregation switches in the whole network, so its scalability is not strong.	VL2, PortLand, LISP etc.
	Improved Control Plane	Through improving the control plane of two-layer network, it avoids the spanning tree, broadcast storms etc and realizes the multi-path forwarding technology.	As running the routing protocol on the control plane, it leads to more complicated configuration management.	Monsoon, SPB, TRILL, FabricPath etc.
	the Big Two Layer Interconnection across Data Centers	It realizes the expansion of two layer network across different geographical positions of data centers.	The gateways in different regions will consume precious WAN resources when learning MAC addresses each other.	VPLS, OTV etc.
The Network Architecture across Three Layer		It simplifies the specific requirements for the deployment of network virtualization software.	The terminal entities of tunnels will consume precious WAN resources when learning MAC addresses each other.	VXLAN, NVGRE etc.

3 云计算数据中心网络虚拟化

“虚拟化”是云计算数据中心的一大重要特征，通过引入虚拟化，可以极大提高数据中心的资源利用率，并确保更多用户获取质量更优的应用服务. 在云计算环境下，虚拟化的对象不仅仅局限于服务器，交换机、路由器、网络链路等网络元素都属于可虚拟化的范畴. 从不同的角度讲，云计算数据中心网络虚拟化往往实现着不同的目标[46-54]，云租户期望实现对其所部署应用服务质量的保证；而云基础设施提供商期望能够提高物理数据中心的承载比，即同一底层物理网络承载更多的虚拟数据中心，从而增加经济收益，具体如图 20 所示. 为了在有限的物理基础设施之上承载更多的 VDC，并高可靠地满足应用服务对网络性能的多种需求，对云计算数据中心网络虚拟化的研究具有十分重要的意义.

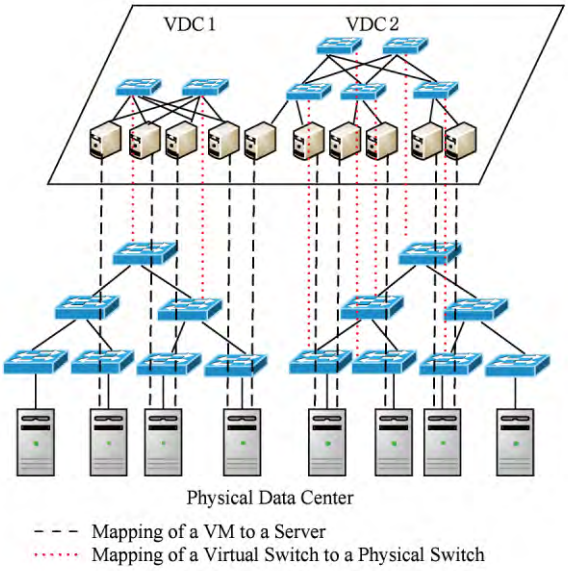


Fig. 20 The virtual data center.

图 20 虚拟数据中心

围绕云计算数据中心网络虚拟化，目前研究者已经提出了多种虚拟网络体系架构的解决方案，根据所解决问题侧重的不同，可以将其分为两大类：可扩展型虚拟网络体系架构，主要解决虚拟数据中心的可扩展性问题；性能保证型虚拟网络体系架构，主要解决应用服务性能质量的保证问题，如带宽、容错等.

3.1 可扩展型虚拟网络体系架构

可扩展型虚拟数据中心网络体系架构旨在突破现有物理网络设备的性能瓶颈，通过采用恰当的设计方式，实现对更大规模虚拟数据中心的支持.

文献[55]提出了一种虚拟网络体系架构 SecondNet，其能够部署于任意的物理网络拓扑结构之上，具有很强的适用性. SecondNet 虚拟网络体系架构如图 21 所示，其主要有 3 方面的特征：

1) hypervisor 取代交换机进行路径转发信息的维护，通常情况下，为了满足云租户的性能需求，需要对 VM-Pairs 路径的相关状态信息(如预留的带宽等)进行保存和维护，SecondNet 为了避免信息维护给中间交换机带来巨大的处理负担，选择使用 hypervisor 来进行状态信息的维护，由于服务器上驻守的 VM 数量有限，因而该方法切实高效可行；

2) 使用基于端口的源路由，数据中心网络存在单一管理实体，因而全网拓扑是可以获取的，基于这样的事实，为了实现部署的任意性和灵活性，SecondNet 采用了类似源路由的数据包转发方式，但不同于源路由，数据报文中并没有携带所有下一跳的 MAC 地址或者 IP 地址，而是携带所有下一跳的输出端口；

3) 建立健壮可靠的信息传输树，为了保证网络管理控制信息的畅通传输，SecondNet 建立了一个以 VDC Manager 为根的带内通信生成树，保证了 VDC Manager 和节点之间控制信息的可靠传输.

正是由于这 3 点特征，使得 SecondNet 具有较强的可扩展性. 但是，源路由的应用需要交换机支持端口交换和基于流的优先级的抢先调度.

NetLord[56]虚拟网络体系架构通过虚拟云租户的 2 层和 3 层地址空间，即允许云租户根据实际需要设计部署自己的地址空间，从而实现了对更多云租户的支持. 如图 22 所示为 NetLord 虚拟网络体系架构，其数据包的发送过程涉及 6 个实体：源虚拟机 VM-S、源 NetLord 代理 NLA-S、入口交换机 ES-S、出口交换机 ES-D、目的 NetLord 代理 NLA-D、目的虚拟机 VM-D. NetLord 数据包的发送过程可描述如下：VM-S 为数据包指定 2 层地址(源虚拟机的 MAC 和目的虚拟机的 MAC)，并发往 NLA-S；部署于物理服务器上的 NLA-S 为数据包增加 2 层包头和 3 层包头，2 层包头的地址分别为 ES-S 和 ES-D 的 MAC 地址，3 层包头的源 IP 为云租户 MAC 地址空间的 ID，3 层包头的目的 IP 为 ES-D 出端口和云租户 ID 的组合；通过 2 层网络，该数据包被传送到了 ES-D，去掉外层的 2 层包头，并依据目的 IP 数据包到达 NLA-D；数据包到达 NLA-D 后，依据云租户 ID 和 VM-D 的 MAC 就能到达目的虚拟机.

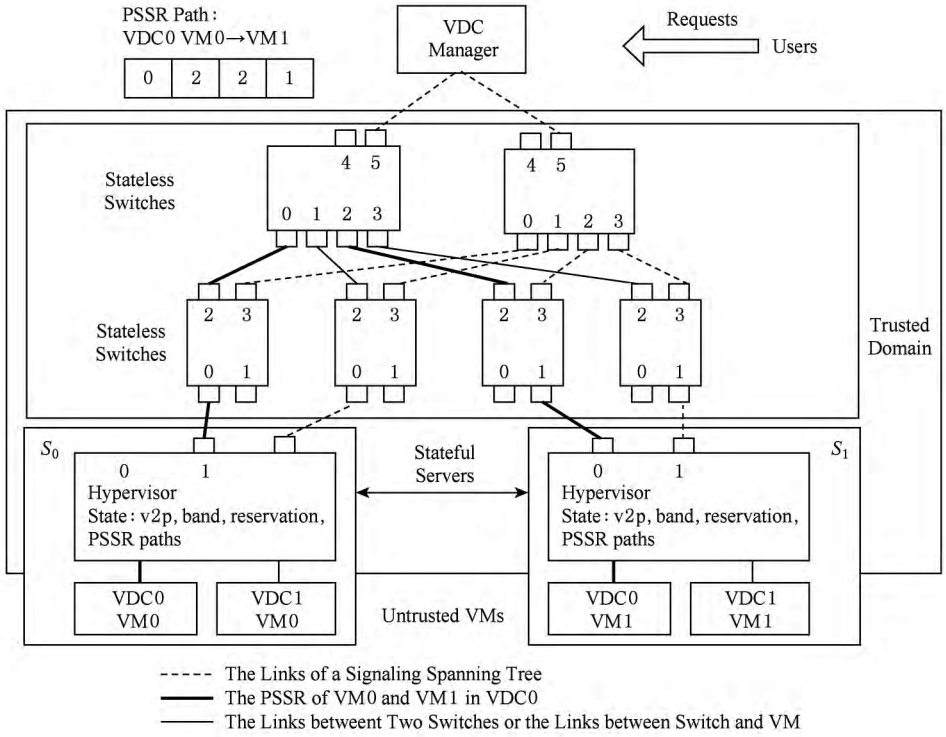


Fig. 21 The virtual network architecture of SecondNet.

图 21 SecondNet 虚拟网络体系架构

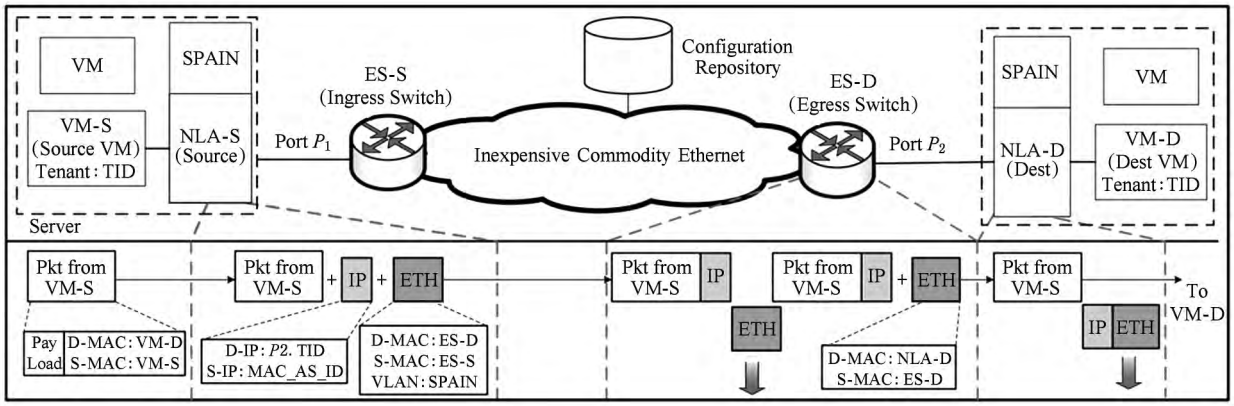


Fig. 22 The virtual network architecture of NetLord.

图 22 NetLord 虚拟网络体系架构

通过在服务器部署 NetLord 代理,将租户虚拟机的 MAC 地址封装起来,使其并不暴露于网络中,这样不但节约了交换机的转发表空间,而且允许不同云租户 MAC 地址空间的重叠,具有较好的扩展性. 但是 NetLord 额外的封装会增加数据包的长度,造成分片或者丢包. 另外,NetLord 使用 SPAIN 在 2 层网络上进行多路径转发,但由于 SPAIN 的转发操作是基于数据流的,这一定程度上限制了 NetLord 的可扩展性.

不同 SecondNet 使用源路由、NetLord 使用代理来增强虚拟数据中心网络体系结构的可扩展性,文献[57]提出了一虚拟网络向底层物理网络映射的再优化机制. 因为各个物理节点(边)在底层网络中所起的重要程度互不相同,所以应该有所区别地进行对待,如果不加以区别,就可能将虚拟节点(边)过多地映射到关键节点(边)上,从而影响更多虚拟网络的进一步映射. 该文通过检测、定位映射过程中的关键节点和瓶颈链路并对其进行重新映射,从而提高了底层物理网络对虚拟网络的承载比率,增强了虚拟网络体系结构的可扩展性.

3.2 性能保证型虚拟网络体系架构

性能保证型虚拟网络体系架构旨在更大程度地保证云租户应用服务的服务质量,如带宽、容错等,从而使云租户有更好的用户体验,减小其经济损失.

Oktopus[58]就是性能保证型虚拟网络体系架构的一例,其为了更清晰地描述云租户应用服务的资源需求,提出了2种网络抽象的实现,分别是virtual cluster和virtual oversubscribed cluster. 图23(a)所示为virtual cluster,即由不具有链路收敛属性的虚拟交换机连接所有VM,主要针对于数据密集型的应用,如MapReduce;图23(b)所示为virtual oversubscribed cluster,即通过多个virtual cluster形成一个具有链路收敛比属性的双层交换集群,主要针对于本地通信集中的应用. 基于这2个网络抽象的定义,云租户可以根据自身应用服务的通信特点,选择恰当的网络抽象方法以及链路收敛比,如$\langle N,B\rangle$或者$\langle N,S,B,O\rangle$,从而提高ISP实际所提供资源与云租户资源申请的匹配程度. 虚拟网络体系架构Oktopus通过采取一系列措施为云应用服务的正常运行提供了带宽保证,但与此同时却牺牲了网络链路的高利用率,Gatekeeper[59]针对这一问题,为每一对VM-Pair都定义了最低保证速率和最高允许速率,使得VM-Pair带宽在保证最低需求的基础上有着灵活的变动空间,从而一定程度上提高了链路的利用率. 而虚拟网络体系架构NetShare[60],SeaWall[61]则采用了不同的设计思路,提出基于权值的带宽分配策略,在不同应用服务间实现了相对公平的带宽共享,但并不提供固定的带宽保证,从而更进一步提高了链路的利用率.

但是TIVC[62]发现云应用服务在整个运行过程中,只有30%~60%的执行时间才会产生大量数据流量,如果在云应用服务初始运行时就分配固定的带宽资源,势必会造成数据中心资源的过度浪费. 为了合理地利用数据中心的带宽资源,TIVC提出根据云应用服务的流量模型[63],随时间动态地为其分配带宽资源.

TIVC模型具体如图24所示,为了清晰地描述如何随时间动态调整带宽资源,文献[62]以云计算数据中心典型的应用服务为例,归纳分析了常见的流量模型,共分为4类:Type 1(单一峰值);Type 2(重复固定宽度峰值);Type 3(变宽度峰值);Type 4(变宽度变高度峰值). 基于这4类,带宽有规律地调整,从而实现了数据中心网络资源高效利用.

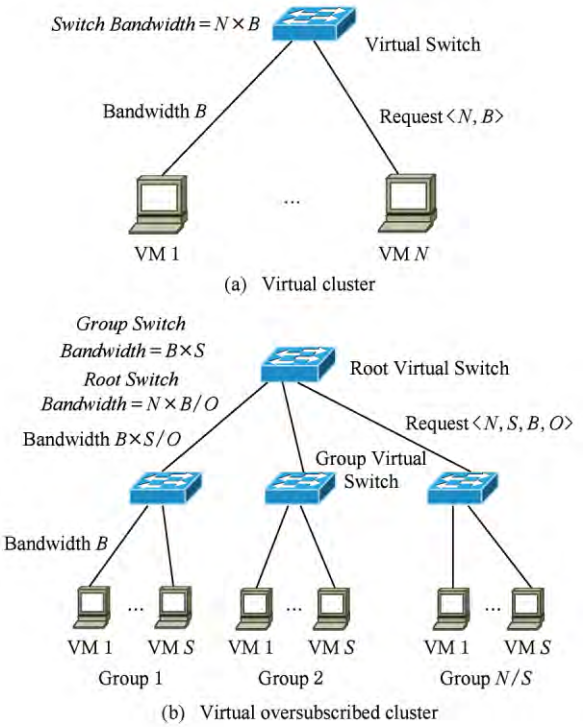


Fig. 23 The network abstractions in Oktopus.

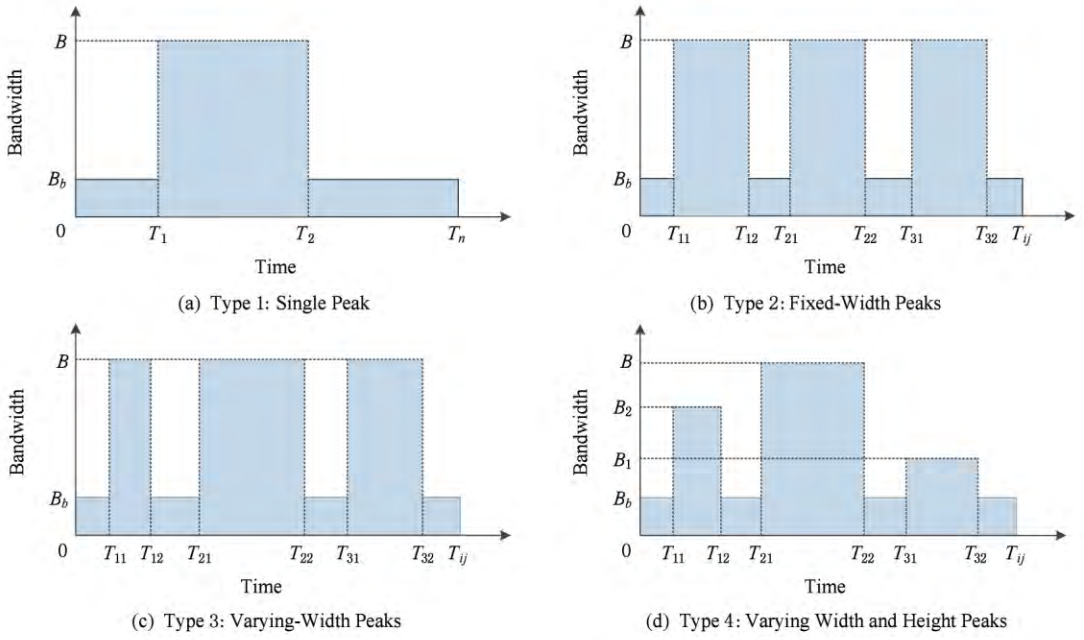
图23 Oktopus的网络抽象

然而,为支撑云应用服务的高效部署和管理,虚拟网络体系架构所面临的问题不仅仅限于如何进行VM分组、预留带宽等,还包括应用服务地址空间的定义、如何进行网络广播等问题,而目前的虚拟网络体系架构往往只能解决其中的部分问题. 旨在全面地解决虚拟网络体系架构所面临的问题,CloudNaaS[64]提供了一组原语,用于定义云应用服务的部署需求,这些原语包括应用服务地址空间的定义、网络广播、VM分组、预留带宽等. 为了实现上述目标,CloudNaaS利用OpenFlow技术,将部署云应用服务的过程分为以下3步:①云租户使用原语描述网络的资源需求,并发送给云控制器;②云控制器将网络资源需求的描述转化为通信矩阵;③云控制器确定VM到物理机的映射机制,并生成对应的交换机规则. 在第③步中,CloudNaaS针对如何实现VM到物理机的映射和如何降低交换机转发条目的数量等问题,提出了具体可行的方法. 此外,CloudNaaS可以通过重新规划虚拟数据中心VDC来在线支持网络策略的失效或者改变,但是,由于其对网络流传输路径的数目进行了限制,CloudNaaS可能存在易导致网络拥塞或资源利用率偏低等问题.

文献[65]就虚拟网络体系架构中的容错问题进行了研究,由于一条物理链路可能承载着多个虚拟

网络的多条虚拟链路,一旦底层物理网络发生故障,就可能影响到多个虚拟网络的正常运行,作者提出了一种 Hybrid 启发式策略,并结合节点迁移策略,可有效处置底层网络的多链路失效和单节点失效问题.

虚拟网络体系架构设计的合理性直接影响着云应用服务性能质量的保证,虽然有许多方案相继提出,但是其仍面临不少的问题,特别是随着云应用服务的繁荣快速发展,目前该领域仍是学术界和工业界的关注焦点.



(a) Type 1: Single Peak

(b) Type 2: Fixed-Width Peaks

(c) Type 3: Varying-Width Peaks

(d) Type 4: Varying Width and Height Peaks

Fig. 24 TIVC models.

图 24 TIVC 模型

4 未来发展趋势

在云计算技术趋势的影响下,目前,数据中心领域正发生着前所未有的深刻变革,本文围绕云计算数据中心网络所面临的主要挑战,将当前相关的研究工作从云计算数据中心网络体系架构、云计算数据中心 VMotion 支撑和云计算数据中心网络虚拟化 3 个方面进行了深入分析,可以看出,尽管数据中心已经获得了很大发展,但是仍尚未足够进化以满足新技术条件下具体应用服务的部署和使用需求,以下 3 个方面在未来仍值得继续探索与研究.

1) 新颖云计算数据中心网络架构的研究. 在目前数据中心的网络架构研究方案中,有线形式的研究已渐渐受到较少的关注,相反,为了追求更低廉的建设成本、降低网络布线的复杂性,无线形式的数据中心网络架构逐渐受到热捧,随着无线通信技术的不断发展,可以预期在这方面将产生不少的新成果. 另外,关于数据中心之间的互连拓扑也有较大的研究空间,通过数据中心之间的高效协同,可进一步推动云计算实际应用的大力发展.

2) "大二层"网络技术的深化与创新. 为了实现数据中心网络资源的灵活调度与部署,深化与创新"大二层"网络技术迫在眉睫,因为跨越 3 层的网络资源调度方式不够灵活且本身存在巨大的开销. "大二层"网络技术具有较大的研究空间,体现在以下 2 方面:①为了避免对现有网络设备的巨大改动,"大二层"网络技术需要更多地依托于现有网络设备(交换机)所支持的多种功能,而并非仅限于新设计的网络设备才可以支持;②发展跨数据中心的"大二层"网络技术,随着云计算的发展,数据中心之间的交互逐渐变得频繁,如何实现跨数据中心的资源调度又尽可能地节约宝贵的广域网资源、降低运维管理的复杂度,是值得探索的一个方向.

3) 云计算数据中心虚拟化的进一步发展. 虚拟化虽然已经在云计算数据中心环境中得到大量应用,但就其发展而言,仍有较大的研究空间,体现在 3 方面:①虚拟化的对象不仅仅局限于服务器、链路,还包括路由器、交换机、存储设备及安全系统等,扩大虚拟化对象的范围将有助于网络资源的进一步高效利用. ②虚拟网络在向物理网络映射时,除了保证带宽外,对于别的因素的考虑也很重要,比如如何

动态地映射以实现节能的目的、如何将虚拟网络映射到跨自治域的底层物理网络[66]等. ③在虚拟化的过程中也面临着不少新的安全问题,需要去研究解决,如不同虚拟网络对安全有着不同程度的要求,需要设计高效且可扩展的监控机制;虚拟网络之间的安全策略可能会互相影响,如何避免这种复杂性等.

参 考 文 献

[1] Xu Libing. Tengyun: the Exploration of Network Technologies in the Era of Cloud Computing and Big Data [M]. Beijing: Posts & Telecom Press, 2013 (in Chinese)
(徐立冰. 腾云:云计算和大数据时代网络技术揭秘[M]. 北京:人民邮电出版社, 2013)

[2] Vamanan B, Hasan J, Vijaykumar T N. Deadline-aware datacenter TCP (D^{2}TCP) [J]. ACM SIGCOMM Computer Communication Review, 2012, 42(4): 115-126

[3] Rasley J, Stephens B, Dixon C, et al. Planck: Millisecond-scale monitoring and control for commodity networks [C] // Proc of the 2014 ACM Conf on SIGCOMM. New York: ACM, 2014: 407-418

[4] Guo Z, Yang Y. Multicast fat-tree data center networks with bounded link oversubscription [C] // Proc of IEEE INFOCOM'13. Piscataway, NJ: IEEE, 2013: 350-354

[5] Wu X, Turner D, Chen C C, et al. Netpilot: Automating datacenter network failure mitigation [J]. ACM SIGCOMM Computer Communication Review, 2012, 42(4): 419-430

[6] China Electronics Standardization Institute. Status and prospects of green data center [R]. San Francisco, CA: Sino-America Energy Efficiency Forum, 2011

[7] Farrington N, Porter G, Radhakrishnan S, et al. Helios: A hybrid electrical/optical switch architecture for modular data centers [J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 339-350

[8] Bari M F, Boutaba R, Esteves R, et al. Data center network virtualization: A survey [J]. Communications Surveys & Tutorials, 2013, 15(2): 909-928

[9] Dally W J, Towles B P. Principles and Practices of Interconnection Networks [M]. Amsterdam: Elsevier, 2004

[10] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture [J]. ACM SIGCOMM Computer Communication Review, 2008, 38(4): 63-74

[11] Gary B. Facebook fabric networking deconstructed [EB/OL]. 2014 [2015-02-15]. http://firstclassfunc.com/facebook-fabric-networking

[12] Curtis A R, Carpenter T, Elsheikh M, et al. Rewire: An optimization-based framework for unstructured data center network design [C] // Proc of IEEE INFOCOM'12. Piscataway, NJ: IEEE, 2012: 1116-1124

[13] Mudigonda J, Yalagandula P, Al-Fares M, et al. SPAIN: COTS data-center Ethernet for multipathing over arbitrary topologies [C] // Proc of NSDI'10. Berkeley, CA: USENIX Association, 2010: 265-280

[14] Chen K, Guo C, Wu H, et al. Generic and automatic address configuration for data center networks [J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 39-50

[15] Halperin D, Kandula S, Padhye J, et al. Augmenting data center networks with multi-gigabit wireless links [C] // Proc of the 2011 ACM Conf on SIGCOMM. New York: ACM, 2011: 38-49

[16] Zhang W, Zhou X, Yang L, et al. 3D beaming for wireless data centers [C] // Proc of the 10th ACM Workshop on Hot Topics in Networks. New York: ACM, 2011: 1-6

[17] Katayama Y, Takano K, Kohda Y, et al. Wireless data center networking with steered-beam mmwave links [C] // Proc of the IEEE Wireless Communications and Networking Conf. Piscataway, NJ: IEEE, 2011: 2179-2184

[18] Hamedazimi N, Gupta H, Sekar V, et al. Patch panels in the sky: A case for free-space optics in data centers [C] // Proc of the 12th ACM Workshop on Hot Topics in Networks. New York: ACM, 2013: 1-7

[19] Kandula S, Padhye J, Bahl P. Flyways to de-congest data center networks [C] // Proc of the ACM Workshop. New York: ACM, 2009: 32-41

[20] Hamedazimi N, Qazi Z, Gupta H, et al. FireFly: A reconfigurable wireless data center fabric using free-space optics [C] // Proc of the 2014 ACM Conf on SIGCOMM. New York: ACM, 2014: 319-330

[21] Shin J Y, Sirer E G, Weatherspoon H, et al. On the feasibility of completely wireless data centers [J]. IEEE/ACM Trans on Networking, 2013, 21(5): 1666-1679

[22] Ramachandran K, Kokku R, Mahindra R, et al. 60 GHz data-center networking: Wireless => worry less? [R]. Princeton, NJ: NEC Laboratories America, 2008

[23] Wei Wei, Wei Xuanzhong, Chen Guihai. Wireless technology for data center networks [J]. ZTE Technology Journal, 2012, 18(4): 1-6 (in Chinese)
(魏炜, 魏晅中, 陈贵海. 数据中心网络中的无线通信技术[J]. 中兴通讯技术, 2012, 18(4): 1-6)

[24] Kedar D, Arnon S. Urban optical wireless communication networks: The main challenges and possible solutions [J]. IEEE Communications Magazine, 2004, 42(5): 2-7

[25] Zhou X, Zhang Z, Zhu Y, et al. Mirror mirror on the ceiling: Flexible wireless links for data centers [J]. ACM SIGCOMM Computer Communication Review, 2012, 42(4): 443-454

[26] Guo C, Wu H, Tan K, et al. DCell: A scalable and fault-tolerant network structure for data centers [J]. ACM SIGCOMM Computer Communication Review, 2008, 38(4): 75-86

[27] Guo C, Lu G, Li D, et al. BCube: A high performance, server-centric network architecture for modular data centers [J]. ACM SIGCOMM Computer Communication Review, 2009, 39(4): 63-74

[28] Liu Xiaoqian, Yang Shoubao, Guo Liangmin, et al. Snowflake: A new-type network structure of data center [J]. Chinese Journal of Computers, 2011, 34(1): 76-86 (in Chinese)
(刘晓茜, 杨寿保, 郭良敏, 等. 雪花结构: 一种新型数据中心网络结构[J]. 计算机学报, 2011, 34(1): 76-86)

[29] Zhu Guiming, Xie Xianghui, Guo Deke, et al. High performance expandable data center networking structure [J]. Journal of Software, 2014, 25(6): 1339-1351 (in Chinese)
(朱桂明, 谢向辉, 郭得科, 等. 一种高吞吐量、高可扩展数据中心网络结构[J]. 软件学报, 2014, 25(6): 1339-1351)

[30] Liao Y, Yin D, Gao L. DPillar: Scalable dual-port server interconnection for data center networks [C] //Proc of the 19th Int Conf on Computer Communications and Networks. Piscataway, NJ: IEEE, 2010: 1-6

[31] Guo D, Chen T, Li D, et al. Expandable and cost-effective network structures for data centers using dual-port servers [J]. IEEE Trans on Computers, 2013, 62(7): 1303-1317

[32] Li D, Wu J. On the design and analysis of data center network architectures for interconnecting dual-port servers [C] // Proc of IEEE INFOCOM'14. Piscataway, NJ: IEEE, 2014: 1851-1859

[33] Li D, Guo C, Wu H, et al. FiConn: Using backup port for server interconnection in data centers [C] // Proc of IEEE INFOCOM'09. Piscataway, NJ: IEEE, 2009: 2276-2285

[34] Biggs N. Algebraic Graph Theory [M]. Cambridge, UK: Cambridge University Press, 1993

[35] Greenberg A, Hamilton J R, Jain N, et al. VL2: A scalable and flexible data center network [J]. ACM SIGCOMM Computer Communication Review, 2009, 39(4): 51-62

[36] Niranjan M R, Pamboris A, Farrington N, et al. PortLand: A scalable fault-tolerant layer 2 data center network fabric [J]. ACM SIGCOMM Computer Communication Review, 2009, 39(4): 39-50

[37] Farinacci D, Fuller V, Meyer D, et al. The Locator/ID Separation Protocol (LISP) [S/OL]. IETF, 2013 [2016-03-10]. https://tools.ietf.org/html/rfc6830

[38] Greenberg A, Lahiri P, Maltz D A, et al. Towards a next generation data center architecture: Scalability and commoditization [C] // Proc of ACM Workshop on Programmable Routers for Extensible Services of Tomorrow. New York: ACM, 2008: 57-62

[39] Hao F, Lakshman T V, Mukherjee S, et al. Enhancing dynamic cloud-based services using network virtualization [C] // Proc of the 1st ACM Workshop on Virtualized Infrastructure Systems and Architectures. New York: ACM, 2009: 37-44

[40] Hooda S K, Kapadia S, Krishnan P. Using TRILL, FabricPath, and VXLAN: Designing Massively Scalable Data Centers (MSDC) with Overlays [M]. Indianapolis, IN: Cisco Press, 2014

[41] Peter A. Shortest path bridging IEEE 802.1aq tutorial and demo [R]. Huawei: NANOG 50, 2010

[42] Wang P C, Chan C T, Lin P Y. MAC address translation for enabling scalable virtual private LAN services [C] // Proc of IEEE AINAW'07. Piscataway, NJ: IEEE, 2007: 870-875

[43] Cisco. Layer 2 everywhere: Overcoming overlay transport virtualization (OTV) site limitations within and between data centers [EB/OL]. 2012 [2015-05-15]. http://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/overlay-transport-virtualization-otv/white_paper_c11-702185.html

[44] Sridhar T, Kreeger L, Dutt D, et al. VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks Over Layer 3 Networks [S/OL]. IETF, 2014 [2016-03-10]. https://tools.ietf.org/html/rfc7348

[45] Mahalingam M, Duda K, Ganga I, et al. NVGRE: Network Virtualization Using Generic Routing Encapsulation [S/OL]. IETF, 2011 [2016-03-10]. https://tools.ietf.org/html/draft-sridharan-virtualization-nvgre-00

[46] Jing L, Jonathan T. Efficient mapping of virtual networks onto a shared substrate [R]. Saint Louis, Missouri: Washington University, School of Engineering & Applied Science, 2006

[47] Li X, Wang H, Ding B, et al. SPGM: An efficient algorithm for mapping MapReduce-like data-intensive applications in data centre network [J]. Int Journal of Web and Grid Services, 2013, 9(2): 172-192

[48] Li X, Wang H, Ding B, et al. Resource allocation with dynamic substrate network in data centre networks [J]. Communications, 2013, 10(9): 130-142

[49] Cheng X, Su S, Zhang Z, et al. Virtual network embedding through topology awareness and optimization [J]. Computer Networks, 2012, 56(6): 1797-1813

[50] Yu M, Yi Y, Rexford J, et al. Rethinking virtual network embedding: Substrate support for path splitting and migration [J]. ACM SIGCOMM Computer Communication Review, 2008, 38(2): 17-29

[51] Bienkowski M, Feldmann A, Jurca D, et al. Competitive analysis for service migration in VNets [C] // Proc of the 2nd ACM SIGCOMM Workshop on Virtualized Infrastructure Systems and Architectures. New York: ACM, 2010: 17-24

[52] Chowdhury N M, Rahman M R, Boutaba R. Virtual network embedding with coordinated node and link mapping [C] // Proc of IEEE INFOCOM'09. Piscataway, NJ: IEEE, 2009: 783-791

[53] Cheng X, Su S, Zhang Z, et al. Virtual network embedding through topology awareness and optimization [J]. Computer Networks, 2012, 56(6): 1797-1813

[54] Sun G, Yu H, Anand V, et al. A cost efficient framework and algorithm for embedding dynamic virtual network requests [J]. Future Generation Computer Systems, 2013, 29(5): 1265-1277

[55] Guo C, Lu G, Wang H J, et al. SecondNet: A data center network virtualization architecture with bandwidth guarantees [C] // Proc of the 6th Int CoNEXT. New York: ACM, 2010: 195-207

[56] Mudigonda J, Yalagandula P, Mogul J, et al. NetLord: A scalable multi-tenant network architecture for virtualized datacenters [J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 62-73

[57] Butt N F, Chowdhury M, Boutaba R. Topology-awareness and Reoptimization Mechanism for Virtual Network Embedding [M]. Berlin: Springer, 2010

[58] Ballani H, Costa P, Karagiannis T, et al. Towards predictable datacenter networks [J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 242-253

[59] Rodrigues H, Santos J R, Turner Y, et al. Gatekeeper: Supporting bandwidth guarantees for multi-tenant datacenter networks [C] // Proc of the 3rd Conf on I/O Virtualization. Berkeley, CA: USENIX Association, 2011: 73-85

[60] Lam T, Radhakrishnan S, Vahdat A, et al. NetShare: Virtualizing Data Center Networks across Services [M]. San Diego, CA: University of California, 2010

[61] Shieh A, Kandula S, Greenberg A G, et al. Sharing the data center network [C] // Proc of the 8th USENIX Conf on Networked Systems Design and Implementation. Berkeley, CA: USENIX Association, 2011: 309-322

[62] Xie D, Ding N, Hu Y C, et al. The only constant is change: Incorporating time-varying network reservations in data centers [J]. ACM SIGCOMM Computer Communication Review, 2012, 42(4): 199-210

[63] Kandula S, Sengupta S, Greenberg A, et al. The nature of data center traffic: Measurements & analysis [C] // Proc of the 2009 ACM Conf on SIGCOMM. New York: ACM, 2009: 202-208

[64] Benson T, Akella A, Shaikh A, et al. CloudNaaS: A cloud networking platform for enterprise applications [C] // Proc of the 2nd ACM Symp on Cloud Computing. New York: ACM, 2011: 53-64

[65] Rahman M R, Aib I, Boutaba R. Survivable Virtual Network Embedding [M]. Berlin: Springer, 2010

[66] Chowdhury M, Samuel F, Boutaba R. PolyViNE: Policy-based virtual network embedding across multiple domains [C] // Proc of the 2nd ACM SIGCOMM Workshop. New York: ACM, 2010: 49-56

Wang Binfeng, born in 1989. Received his BSc degree in computer science from China University of Petroleum in 2011 and his MSc degree in computer science from National University of Defense Technology in 2013. PhD candidate at the College of Computer, National University of Defense Technology. His main research interests include network management, data center network and cloud computing.

Su Jinshu, born in 1962. Received his MSc and PhD degrees in computer science from National University of Defense Technology in 1989 and 1999 respectively. Professor and PhD supervisor at the College of Computer, National University of Defense Technology. Member of the ACM and IEEE. His main research interests include high-performance routers, Internet routing, high-performance computing, cloud computing, wireless networks, and information security (sjs@nudt.edu.cn).

Chen Lin, born in 1976. Received her PhD degree in computer science from National University of Defense Technology in 2005. Associate professor at the College of Computer, National University of Defense Technology. Her main research interests include network management and data center network resource management (chenlin@nudt.edu.cn).