云计算之数据中心网络的发展

朱桂明 谢向辉 陆菲菲

（ 江南计算技术研究所 数学工程与先进计算国家重点实验室 无锡 214083 ）

摘 要 数据中心是云计算所依赖的核心基础设施，而数据中心网络则是数据中心的基础，关系到数据中心的性能、规模、可扩展性和管理性。本文分析了当前主要数据中心网络结构的不足，对当前数据中心网络的研究现状进行了分析和对比，并对未来数据中心网络的发展和应用作了展望。

关键词 数据中心网络；类树结构；扁平结构；服务器参与路由的结构

Data Center Networking Developing of Cloud Computing

ZHU Gui-ming XIE Xiang-hui LU Fei-fei

(*State Key Laboratory of Mathematical Engineering and Advanced Computing, Jiang Nan Institute of Computing Technology, Wuxi* 214083, *China*)

Abstract Data center is the kernel infrastructure of cloud computing, while data center is the base on data center networking, which relates to the performance, scalability and manageability of data center. This paper points out the disadvantage of the current main data center networking, analyzesand compares the state of art research work, and makes a forecast of the future data center networking.

Keywords data center networking; tree-similar structure; flattening structure; server-centric structure

1 引 言

云计算是对传统资源使用方式的一种变革，用户能够通过网络按需获取计算、存储和网络等资源。作为云计算所依赖的核心基础设施，数据中心成为各类资源的集中地、各种业务的服务中心、以及数据处理、存储和交换中心，发挥着越来越重要的作用。

随着云计算技术的发展，数据中心的规模日益增大，促使诞生了十万量级、甚至百万量级服务器的数据中心。现有数据中心网络互连结构主要依靠交换机、核心交换机、核心路由器将服务器连接起来构成树型结构，这对位于树型结构高层的核心路由器、核心交换机的性能要求极高，且其容易成为网络流量的瓶颈[1]。树型结构的固有弊端使得其远远不能达到数据中心网络所追求的高可扩展、容错性好、高聚集带宽等目标。针对传统数据中心网络结构的固有缺陷，研究人员纷纷开始探寻新的架构设计方案。目前，对于数据中心网络的研究和讨论主要集中在以下三大类结构中：类树结构、服务器参与路由的结构和扁平结构。

2 类树结构

图1给出了常见数据中心树型结构。Edge交换机通常采用48至288端口的千兆交换机，Aggregation和Core交换机通常采用32至128口的万兆交换机。这种结构对处于树型结构上层的交换机的性能要求极高，且代价也十分高昂。为此，加州大学的研究员Al-Fares等提出使用廉价的交换机构建胖树结构Fat-Tree[2]，如图2所示。该结构实现了系统的大规

作者简介：朱桂明，博士，工程师，研究方向为对等网络、数据中心网络；谢向辉，博士，研究员，研究方向为计算机体系结构；陆菲菲，博士生，助理工程师，研究方向为数据中心网络。

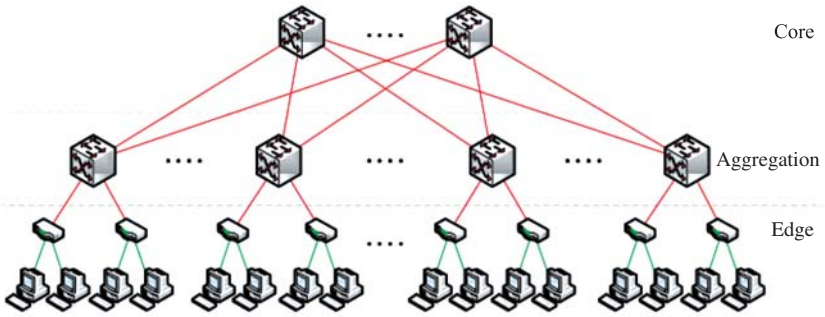


图1　常见数据中心树型互连结构

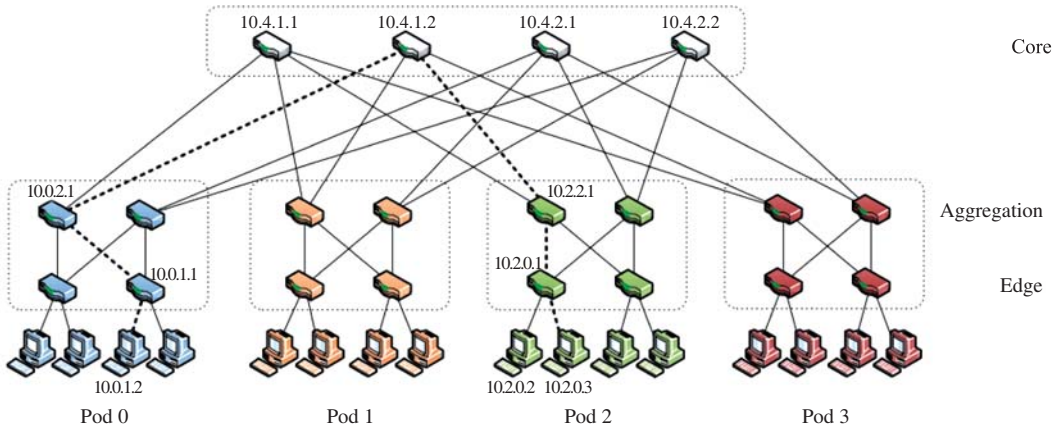
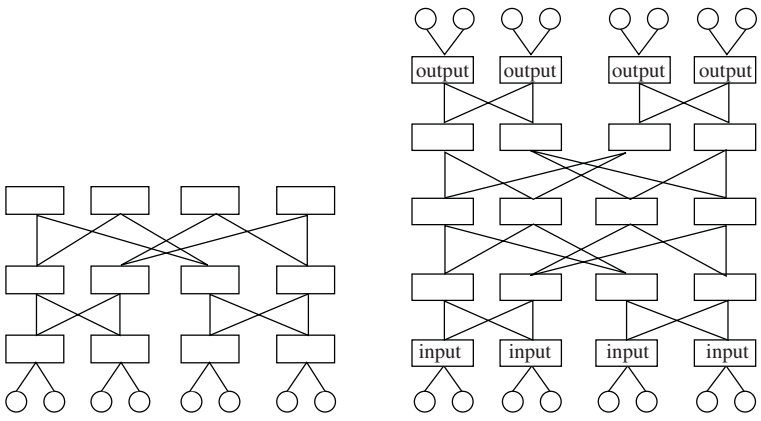


图2　基于廉价交换机的简单Fat-tree结构

模互连以及服务器之间的高通信带宽。Fat-tree共有核心层（Core level）、聚集层（Aggregation level）、边界层（Edge level）三层交换机；有n个Pod，每个Pod含有聚集层和边界层各$n/2$个交换机；边界层的每个交换机使用$n/2$个端口连接服务器，其余$n/2$端口连接聚集层的$n/2$个交换机；核心层有$(n/2)^2$个n口交换机，每个交换机有一个端口连接一个Pod。

这一发展趋势并不是偶然的，上世纪五十年代在电话交换领域也面临过类似的挑战。贝尔实验室的Charles Clos博士首次提出应用大量小型商业交换机构建Clos网络应对高带宽交付的问题[3]，后来被广泛应用于TDM网络。Fat-tree结构的应用动机与Clos网络相似，而且同构于5级Clos网络，如图3所示。

胖树结构相对于传统多根树结构的优点是聚合层的节点之间具有更高的数据传输性能和容错性，且各层的链路数相等，解决了普通的树型结构在核心层的带宽瓶颈问题，使得每一层都具有相同的聚合带宽，而且具有良好的容错特性。这种结构的缺点主要表现



(a) 简单的Fat-tree拓扑　　(b) 5级Clos结构（folded-Close）

图3　Fat-tree结构与Clos结构

在：它的设计思想是“scale out”模式，要求核心层交换机具有较大的连接度数，即其扩展性受限于交换机端口数量，存在扩展性不足的缺点；当网络规模增大时，会对交换机的转发表造成巨大的压力，同时会面临地址解析的压力（ARP广播）；处理交换机故障能力不足及路由协议容错性不强；而它的树型特征也决定了其不能很好地支持one-to-many（一对多）和several-to-several（多对多）的网络通信服务。

3 服务器参与路由的结构

为了能更好地达到数据中心网络所追求的高性能目标，越来越多的以服务器为中心的数据中心网络结构在近几年被提了出来。借助服务器的多网络NIC端口以及数据转发功能，可以利用大量服务器之间的连线以及低端交换机来实现大型数据中心网络的高效互连，其中基于复合图（Compound Graph）并采用层次网络的设计思想成为当前以服务器为中心的数据中心网络互连的主流方法。

DCell[4]提供了可扩展的结构来构建大型数据中心网络，其利用低端交换机以及具有多网络NIC端口的服务器迭代地构建服务器之间的互连结构，利用分布式容错性路由协议DFR（DCell Fault-Tolerant Routing）实现近似的最短路路由，并将流量均匀分散到各条链路上，消除瓶颈问题。DCell使用迭代的方式构建，每个高层的DCell通过连接一定数量的低层DCell来构建，多个同层DCell之间彼此全连通。$DCell_0$是最基本的构建模块，它由n个服务器与一个n口微型交换机连接构成；$DCell_1$由n+1个$DCell_0$构成，图4给出了n=4的$DCell_1$拓扑结构。FiConn[5]采用与DCell类似的层次化方式构建，然而，对FiConn而言，每台服务器只需要配备两个NIC端口就能构建任意层次的FiConn结构，如图5所示。

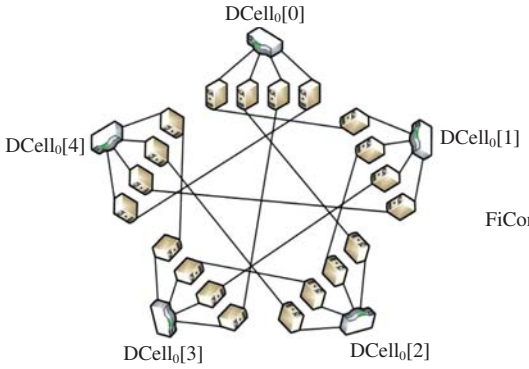


图4 $DCell_1$互连结构

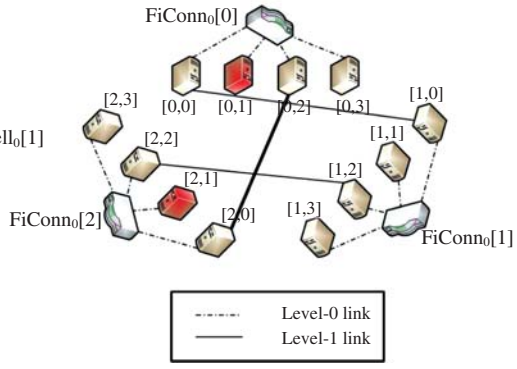


图5 $FiConn_1$互连结构

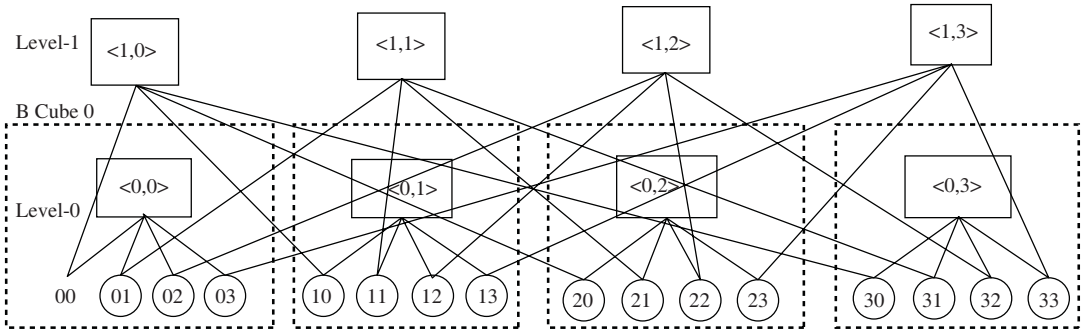


图6 $BCube_1$互连结构

BCube[6]是一种依层次化构建的、具有高吞吐量的新型数据中心网络结构。$BCube_0$由n个服务器连接一个n口交换机构成；$BCube_k$由n个$BCube_{k-1}$和n^k个n口交换机构成。图6给出了$BCube_1$互连结构，与DCell中服务器直接与其他cell中的服务器相连不同的是，BCube中不同cube之间通过服务器-交换机-服务器的方式连接，当出现部分交换机或服务器失效时，BCube的性能降低的速率比较缓慢。

HCN[7]和BCN[7]与FiConn有着相似的设计理念，即网络的规模不受限于服务器的网络NIC端口个数，同

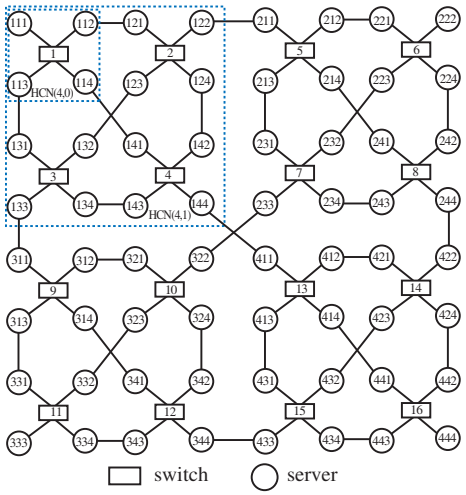


图7　HCN(4,2)

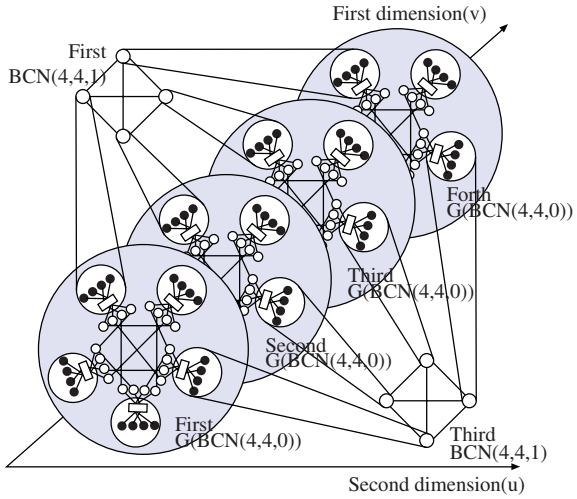


图8 BCN(4,4,1,0)

时，HCN结构具有高度的规则性、灵活性和系统性。但是，HCN和BCN专注于解决数据中心网络的无损可扩展性、持续可扩展性和渐进可扩展性问题，却忽略了对数据中心网络对网络性能的要求。图7给出了HCN(4,2)结构，图8给出了BCN(4,4,1,0)的结构。

微软2011年研发了集成有交换机芯片的PCI网卡ServerSwitch[8]，不但具备传统交换机的能力，而且能通过高速PCI-E接口实现CPU和ServerSwitch的高速互联。这利于借助服务器强大的计算能力甚至存储能力实现对网络流量的深入分析处理。ServerSwith利用了可编程交换芯片和网络接口芯片，做成了一块PCI-E接口卡，此卡与一台商用服务器共同构成了ServerSwith。ServerSwith可通过C/C++语言对交换芯片进行编程，确定转发机制，实现了BCube等各种以服务器为中心的网络协议。ServerSwitch大大降低了服务器参与路由转发而引起的处理器负载，和转发数据包的延迟，为服务器参与路由的方式数据中心网络结构大规模应用奠定了坚实的基础。

4　扁平结构

扁平结构（以交换机为中心），是ISCA、HPCA、SC等超级计算机体系结构领域的研究人员给出的思路，包括FBFLY[9]和HyperX[10]。基本思想是，用多端口交换机（例如64口）互联而成一个一般意义上的超级立方体，每个交换机的剩余端口连接一些服务器。其实BCube是一种变通的generalized hypercube，而FBFLY和HyperX在交换机层面则是切切实实的generalized hyupercube。

图9(a)是4-ary 2-fly butterfly结构，图(c)

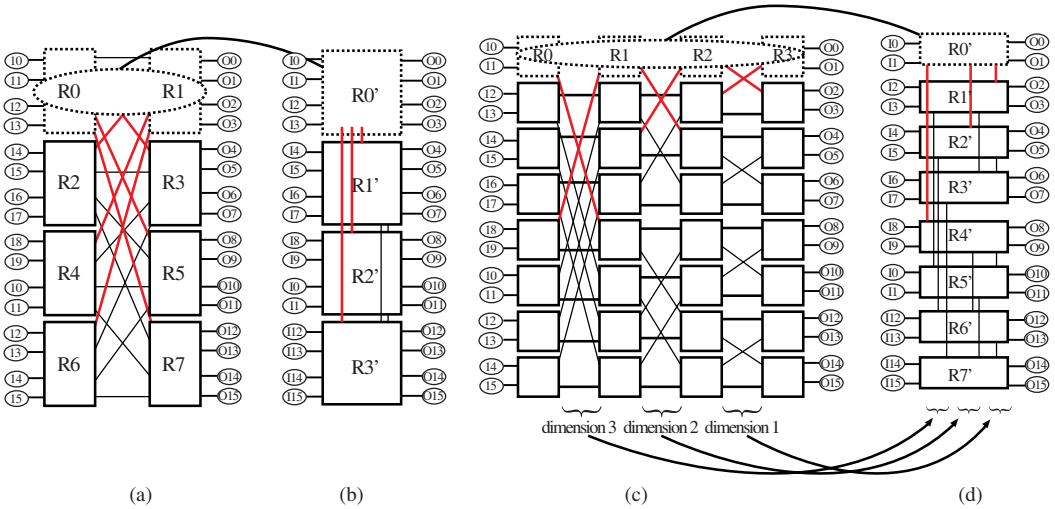


图9　Flattened Butterfly及其对应的Butterfly结构

是2-ary 4-fly butterfly结构，图(b)、图(d)分别是图(a)、图(c)对应的Flattened Butterfly结构。FBFLY（the k-ary n-flat Flattened Butterfly）结构利用了多端口交换机互联而成一个一般意义上的超级立方体，每个维度中的交换机需要与同维度的所有其它交换机互联，而服务器则只能使用交换机剩余的端口。

从目前市场的应用层面看，扁平化的网络架构主要体现在基于TRILL[11]标准等技术的发展上。一些大型的网络供应商们（包括Cisco、Brocade、H3C、Juniper等）认为数据中心的网络必须经历新一轮的扁平化、融合式的结构调整，以适应整体IT架构，因为整合了尽可能多的基础设施而弹性化的需求。不仅如此，目前这些相关的产品和技术也日渐涌现市场。它们主要是基于TRILL标准：Transparent Interconnection of Lots of Links简称TRILL，属于2层标准，其主要作用是整合了网桥和路由器的优点，将链路状态路由技术应用在二层，以克服生成树协议（STP）在规模上和拓扑重聚方面存在的不足，另外提高对单播和多播在多路方面（Multi-Pathing）的支持，并降低延迟。

5 数据中心网络发展展望

随着云计算技术应用热潮的到来，传统数据中心网络需要进行架构性变革已成为业内共识。但数据中心网络的发展方向和途径却众说纷纭。目前大多数数据中心网络是以clos网络、fat-tree结构为主。主要是这种结构对网络的改动比较小。但近年提出的BCube等结构在支持数据中心典型流量模式方面具有更大的优势，且容错性更好。预计未来各种各样的新型网络拓扑结构仍会不断涌现。由于一些非技术因素，以服务器为中心的结构也许在市场推广方面会有一定的难度，但我们认为其优异的性能、适合于各种通信模式（特别是“一对多”通信模式）的应用，会使其占据一定的市场。尽管目前扁平化网络架构似乎更受网络设备供应商的青睐，但是需要在新的网络设备方面有大量投入，因此，我们认为最终网络到底要不要扁平化还将归结为对性价比的评估。

参 考 文 献

[1] 陈贵海,吴盼,杨盘隆.数据中心网络 [J].中国计算机学会通讯, 2011, 7(7): 21-26.

[2] Fares M A, Loukissas A, Vahdat A. A scalable, commodity data center network architecture [C] // Proceedings of ACM International Conference on the applications, technologies, architectures, and protocols for computer communication, Seattle, USA, 2008.

[3] Clos C. A study of nonblocking switching network [J]. Bell System Technology, 1953, 32(5): 404-424.

[4] Guo C, Wu H, Tan K, et al. DCell: A scalable and fault-tolerant network structure for data centers [C] // Proceedings of Special Interest Group on Data Communication, Seattle, USA, 2008.

[5] Li D, Guo C, Wu H, et al. Ficonn: Using backup port for server interconnection in data centers [C] // Proceedings of IEEE International Conference on Computer Communications, Brazil, 2009.

[6] Guo C, Lu G, Li D, et al. BCube: a high performance, server-centric network architecture for modular data centers [C] // Proceedings of Special Interest Group on Data Communication, Barcelona, Spain, 2009.

[7] Guo D, Chen T, D. Li, et al. Expansible and cost-effective network structures for data centers using dual-port servers [C] // Proceedings of IEEE International Conference on Computer Communications, 2010.

[8] Lu G, Guo C, Li Y, et al. ServerSwitch: a programmable and high performance platform for data center networks [R]. Microsoft Techinical Report, 2011.

[9] Kim J, Dally J, Abts D. Flattened Butterfly: a cost-efficient topology for high-radix networks [C] // Proceedings of International Symposium on Computer Architecture, 2007.

[10] Binkert J A N, Davis A, et al. HyperX: topology, routing, and packaging of efficient large-scale networks [C] // Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2009.

[11] Perlman R, Dutt D, Gai S, et al. Routing bridges: base protocol specification [S]. Internet Engineering Task Force, 2011.