数据中心的网络互联结构和流量协同传输管理

郭得科 陈涛 罗来龙 李妍 著

清华大学出版社

数据中心的网络互联结构和流量协同传输管理

郭得科 陈涛 罗来龙 李妍 著

清华大学出版社
北京

内 容 简 介

本书对数据中心及其网络互联结构的现状和发展趋势进行了深度剖析，深入介绍了一些新型网络互联结构的设计与优化方法，力求满足数据中心网络的高带宽、高容错、高可扩展性等方面的需求，并通过引入数据中心内关联性流量的协同传输机制，实现对数据中心现有传输能力的高效利用。

本书可用作高等学校计算机专业、软件工程专业、信息系统工程专业以及其他相近专业的教材或教学参考书，也可供这些专业的研究人员和工程技术人员阅读。

图书在版编目(CIP)数据

数据中心的网络互联结构和流量协同传输管理/郭得科等著. —北京：清华大学出版社，2016

ISBN 978-7-302-44810-5

Ⅰ. ①数… Ⅱ. ①郭… Ⅲ. ①计算机网络－网络结构－研究 ②计算机网络－流量－研究 Ⅳ. ①TP393

中国版本图书馆 CIP 数据核字(2016)第 200801 号

责任编辑：薛　慧
封面设计：何凤霞
责任校对：王淑云
责任印制：王静怡

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　**邮　　编**：100084
社 总 机：010-62770175　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　**印　　张**：17.75　**插　　页**：1　**字　　数**：318 千字
版　　次：2016 年 11 月第 1 版　**印　　次**：2016 年 11 月第1次印刷
印　　数：1～1500
定　　价：59.00 元

产品编号：062164-01

前言

1. 背景

数据中心的出现源于人们对海量数据的高效存储和处理需求。互联网的蓬勃发展和社会的数字化变革，导致网络上的数据呈爆炸式增长，出现了越来越多需要进行大规模数据存储和处理的应用需求。数据中心的规模和应用领域不断扩展，已经渗透到经济、科技、军事以及人们日常生活等各个方面。总体而言，数据中心旨在依据特定网络结构互联大规模服务器和网络设施等硬件资源，形成计算、存储、网络等资源的规模效应和整体优势，进而面向各类上层应用提供网络化存储、网络化计算等弹性服务。

现代社会信息量的爆炸式增长、资源复用技术的成熟和宽带网络的普及，共同促进了云计算的诞生和发展。数据中心可为云计算提供大规模可扩展的基础物理资源，并在云平台的辅助下为用户提供多种类型的云服务。云计算模式的出现拓展了数据中心的使用方式，而数据中心的建设和发展也为云计算的推广和应用奠定了坚实的基础。另外，随着大数据时代的来临，如何从类型多样、规模巨大的大数据中快速提取有价值的信息成为关键。数据中心可为大数据应用提供基础平台，并在解决大数据的存储、大规模分析处理等难题方面具有天然的优势。

云计算和大数据等新技术和应用推动了现代数据中心的快速发展，并使其成为国家和 IT 企业的核心信

息基础设施。数据中心具有巨大的商业价值和社会效益，其应用领域非常广泛，涉及信息社会的诸多行业和领域。数据中心的网络化存储和网络化计算为很多大规模数据处理模式提供基础服务，而这类数据处理模式在大数据领域、物联网、科学应用等领域也得到了广泛的重视和应用。同时，数据中心也面临着不断的变革，新一代数据中心由数以万计的服务器组成，并通过特定的网络结构互联为一个整体，共同形成一个分布式的计算和存储网络。新一代数据中心因为内在的高可扩展性、高容错性等优势得到了业界的高度关注，并向着虚拟化、软件定义、模块化、绿色节能以及自动化运行维护等方向迈进。

数据中心研究的一个重要理论基础是数据中心网络，其作为基本要素在新一代数据中心中具有至关重要的地位。数据中心网络所考虑的不仅是设备之间的通信协议，更主要的是把交换机和服务器作为一个整体进行拓扑互联、性能优化、资源管理和能耗控制，形成数据中心基础设施在网络化计算能力、网络化存储能力和网络通信能力等方面的综合优势。换句话说，数据中心网络不仅是连接大规模服务器的桥梁，而且是承载网络化存储和网络化计算的基础。数据中心支持的业务往往伴随着服务器之间密集的数据交互，网络资源已成为影响数据中心服务质量的瓶颈，且直接关系到各类用户对数据中心的使用体验。因此，迫切需要开展数据中心网络方面的基础理论研究，从而推动数据中心及相关应用领域的发展。另外，数据中心网络是互联网的重要组成部分，该领域的研究进展也会对下一代互联网的发展产生一定的推动作用。与互联网相比，数据中心具有集中管控等鲜明特点，这有利于开展网络创新技术的探索。数据中心运营商为了提高服务性能和收益，会根据应用需求定制网络架构和协议，并进行创新网络技术的部署。

当前，数据中心的可持续发展面临着一系列关键基础性问题，而数据中心网络是其中一个重要的研究方向。数据中心网络的研究面临很多基础理论和关键技术方面的挑战，例如网络功能的灵活定制、横向可扩展的互联结构、网络资源的高效复用、网络虚拟化、关联性流量的协同传输以及网络能耗的协同控制等。

（1）网络功能的灵活定制。传统网络设备的控制平面和数据平面紧密耦合，不具有动态性和灵活性，致使传统的数据中心网络只能提供有限且已知的网络功能和服务。如今，数据中心的网络应用需求变得日益丰富和灵活多样，为了提高数据中心的服务质量，需要针对不同的应用需求对网络功能进行灵活配置。而数据中心网络流量难以预测、网络设备可靠性低等环

境特征，也对网络的可动态灵活配置功能提出了新的需求，致使数据中心的传统网络架构面临着严峻的挑战。

（2）横向可扩展的网络互联结构。数据中心必须具备计算、存储、网络资源按需扩展的能力。有效互联上万台甚至更大规模的服务器，是数据中心提供网络化计算和网络化存储的前提。依靠扩充交换机端口数量或提升端口速率的纵向扩展方法已经远远不能满足数据中心的规模扩展需求。因此，迫切需要对数据中心的规模扩展方式进行改进，探索各种横向扩展模式，进而连接更多的交换机和服务器以实现计算性能和存储容量的按需扩展。

（3）数据中心网络资源的高效复用。数据中心所采用的网络协议基本源自广域网环境，致使数据中心网络资源的利用率很低。在软件定义的可定制网络框架下，如何设计新型路由和传输协议，以提高数据中心网络的资源利用率并进而提升上层应用的性能，是非常具有挑战性的问题。此外，软件定义的数据中心网络架构，为网络资源、计算资源和存储资源的联合优化提供了新的发展机会。

（4）数据中心的网络虚拟化。在大量用户竞争使用网络资源的数据中心环境下，需要实现网络资源的有效共享和安全隔离。虚拟化是保障数据中心安全和实现资源复用的重要技术，每个用户租用的多个虚拟机之间形成了虚拟数据中心网络。不同用户的虚拟网络之间竞争使用实际的物理网络，而且从安全考虑需要被有效隔离。当前，数据中心普遍采用“尽力而为”的方式共享网络资源，不能很好地支持虚拟数据中心网络的流量隔离和带宽保障需求。

（5）关联性流量的协同传输问题。数据中心中密集的数据交互行为产生了庞大的“东西向流量”。Multicast、Incast 以及 shuffle 传输是“东西向流量”的主要组成部分。此外，在组成一个 multicast、shuffle、incast 的众多数据流之间存在很大的数据关联性，进而存在非常大的数据流聚合增益。通过和上层应用的联合设计优化，可在不影响应用效果的前提下大幅降低关联性流量造成的网络传输开销，进而降低对数据中心稀缺网络带宽的消耗。

（6）数据中心网络能耗的协同控制。对数据中心进行高效的能耗管理具有重要的经济效益和社会影响。为此，需要从底层硬件节点、上层协议运行、外围供能系统等环节进行能耗控制的联合优化。其中，数据中心网络层面的能耗控制必不可少，且其能耗控制策略也直接影响到计算设备层面的

能耗控制策略。为了实现多维度的协同能耗控制，数据中心网络需要研究如何实时感知网络能耗、如何在不影响网络性能和可靠性的前提下实现节能流量工程，并尽可能地使用清洁能源。

我们在数据中心的横向可扩展的网络互联结构和关联性流量的协同传输领域进行了一系列深入而系统的研究工作。本书以数据中心的可扩展网络互联结构为基础，深入探讨了一些新型网络互联结构的设计与优化方法，力求满足数据中心网络的高带宽、高容错、高可扩展性等方面的需求，并通过引入数据中心内关联性流量的协同传输机制，实现对数据中心现有传输能力的高效利用。本书绝大部分内容取材于我们近期在国内外重要学术期刊和会议上发表的论文，全面系统地展示了相关领域的很多新的研究成果和进展。

2. 内容安排

本书共 10 章，从结构上可分为 3 个部分。

第 1 部分是对数据中心及数据中心网络互联结构发展现状的介绍，包括第 1 章和第 2 章。

第 1 章首先介绍了数据中心的起源和发展、云计算和大数据对数据中心的内在需求以及新一代数据中心的发展趋势。在此基础上，从应用角度详述了数据中心在网络化存储、网络化计算以及数据分析处理等领域的应用现状。最后，从网络功能的灵活定制、横向可扩展的互联结构、网络资源的高效复用、网络虚拟化、关联性流量的协同传输、网络能耗的协同控制等角度探讨了数据中心网络发展所面临的重要挑战。

第 2 章对当前数据中心网络的最新互联结构进行了综述，从构建规则、路由算法、网络性能等方面进行了对比分析。同时，将当前数据中心的网络互联结构按照 5 种类型进行归类，以揭示数据中心网络互联结构设计理念的发展和变化，分别是交换机为核心的互联结构、服务器为核心的互联结构、模块化数据中心的互联结构、随机型数据中心的互联结构以及无线数据中心的互联结构。最后，对未来数据中心网络互联结构的发展趋势提出了一些观点。

第 2 部分介绍数据中心的新型网络互联结构，包括第 3～6 章。

第 3 章介绍以服务器为核心的数据中心互联结构 HCN 和 BCN。在服务器网络端口数目固定不变的情况下，设计了由这类同构服务器互联而成的数据中心网络互联结构。首先采用复合图(compound graph)理论设计

以服务器为核心的网络互联结构 HCN，为只具有两个网络端口的普通服务器和多端口的低成本网络交换机提供高效互联，使其兼具无损可扩展和持续可扩展能力。DCell 和 BCube 等同类数据中心的网络互联结构中每台服务器的网卡和连线会依据网络规模的扩大而增长，其最大规模被每台服务器的网卡数目所限定，不具备持续可扩展能力。在此基础上，在相同的服务器网卡配置和网络直径前提下，构造出规模尽可能大的数据中心网络互联结构 BCN。随后介绍了这两种网络互联结构的高效及容错路由机制，并通过数学分析和综合仿真模拟结果验证了 HCN 和 BCN 的良好拓扑特征。

第 4 章介绍模块化数据中心互联结构 DCube。构建大规模数据中心有两种截然不同的趋势：第 1 种趋势是通过类似 DCell、BCube、HCN 等扩展性网络互联结构，构造出单体的大规模数据中心；第 2 种趋势是在大量单体数据中心的基础上，通过模块化的互联结构构造大规模数据中心。本章介绍为模块化数据中心设计的一组模块内网络互联结构 DCube，包括 H-DCube 和 M-DCube，每个 DCube 互联大量配备双网卡的服务器和低成本交换机。大量 DCube 互联结构的数据中心模块进一步互联可形成全新的模块化数据中心。DCube(n,k)由k个互联的子网络构成，每个子网络都由许多基本构建模块和标准超级立方体结构(或其变种结构 1-möbius)按照复合图理论构成。在此基础上，分析了 H-DCube 和 M-DCube 的路由机制，并与 BCube 和 Fat-Tree 进行了性能分析比较。如果采用诸如 Twisted cube、Flip MCube 和 Fastcube 等标准超级立方体结构的其他变种结构，本章提出的设计方法仍然适用。

第 5 章介绍一种数据中心网络的混合互联结构设计方法，并讨论了一种具体的混合互联结构 R3。如第 2 章所分析，当前的数据中心网络互联结构普遍采用两种设计思路，分别是完全规则的互联结构设计和完全随机的互联结构设计。虽然这两种类型的网络互联结构具有特定的优势，但是也存在内在的缺点和不足。本章论述了一种基于复合图理论的混合互联结构设计方法，可兼容当前的规则互联结构和随机互联结构。进一步地，还提出了一种混合互联结构 R3。该结构采用随机正则图作为基本单元，并采用通用超级立方体这种规则结构将这些基本单元进行互联。该混合结构兼具随机正则图和通用超级立方体的拓扑优势，并可有效避免二者的拓扑缺陷。

第 6 章介绍在数据中心中额外引入可见光通信(visible light communication，VLC)链路后，设计无线链路和有线链路混合的网络互联结构 VLCcube，从而提升数据中心的网络性能。具体而言，在 Fat-Tree 这一具有代表性的数据中心网络结构基础之上，在每个机架顶部安装 4 个

VLC 收发装置，即可提供 4 条 10Gbps 左右的无线链路，全体机架上的无线链路组网成为无线 Torus 结构。本章重点介绍了混合拓扑的构建规则、路由策略、批处理流量和在线流量的拥塞感知调度策略，并开展了相关实验验证工作。因为 Fat-Tree 中很多原本 4 跳的数据流可切换到无线 Torus 结构进行短距离传播，VLCcube 取得了比 Fat-Tree 更好的网络性能，而且设计的拥塞感知调度策略可使 VLCcube 的性能进一步得到提升。VLCcube 仅是数据中心中利用 VLC 链路的一种可选方案，未来供应商可基于不同的有线网络互联结构设计完全不同的混合网络互联结构。VLC 链路的引入不仅能和已有的数据中心网络良好地兼容，而且可有效提升数据中心的网络性能和网络设计的灵活性。

第 3 部分是数据中心内关联性流量的协同传输管理。虽然新型互联结构的研究不断提高数据中心的网络传输能力，但是对数据中心现有传输能力的高效利用同样重要。数据中心支持多种分布式计算框架，这些计算框架普遍采用流式计算模型，相邻处理阶段间的大量数据交互产生了严重的东西向流量，其中 multicast、incast、shuffle 等关联性流量占相当大的比重，进而严重影响到上层应用的性能。优化和管理这些关联性流量对高效使用数据中心的网络资源和提升处理作业的性能至关重要。这部分内容包括第 7～10 章。

第 7 章介绍关联性流量 incast 的协同传输管理问题，包括网内聚合和协同传输两个环节。当前很多上层应用决定了 incast 的全体数据流之间存在数据相关性，并在相同的接收端被执行聚合操作。这就促使我们考虑在这些关联性流量的网内传输阶段应尽可能早地而不是仅在流量的接收端进行数据聚合。本章首先以新型数据中心网络结构为背景讨论关联性流量之间数据聚合的可行性和增益，随后探讨实现该网内聚合所必须的基于 incast 树的协同传输方法。为最大化网内聚合的增益，我们为 incast 传输建立最小代价树模型，并设计了两种近似的 incast 树构造方法，其能够仅基于 incast 成员的位置和数据中心拓扑结构生成一棵有效的 incast 树。本章进一步介绍了 incast 树面临的多种动态和容错问题，最后通过实验发现我们所提出的网内聚合方法能大幅度降低 incast 流量造成的传输开销，从而节约了数据中心的网络资源。本章虽然选用 BCube 这种以服务器为核心的数据中心网络互联结构为研究背景，但是提出的关联性流量网内聚合理念也适用于其他类型的数据中心网络互联结构。本章工作是第 8 章和第 9 章的前提。

第 8 章介绍关联性流量 shuffle 的协同传输管理问题，包括网内聚合

和协同传输两个环节。受关联性流量 incast 网内聚合的启发，本章介绍了如何将关联性流量 shuffle 原本在诸多接收端执行的流量聚合操作推送到网络传输环节中执行，通过降低网络内的流量传输从而高效地利用网络资源。首先，针对新型数据中心结构 BCube 中的 shuffle 流量进行网内聚合问题的建模，并提出两种近似方法来高效地构建 shuffle 聚合子图，依据该结构进行流量的协同传输可有效实现预期的网内聚合。本章还介绍了基于布鲁姆滤波器(Bloom filters)的可扩展流量转发模式，从而为大量并存的 shuffle 传输实现各自预期的网内聚合效果。尽管本章选用以 BCube 为依托的网络互联结构，但是提出的关联性流量 shuffle 的网内聚合理念也适用于其他类型的数据中心网络互联结构。

第 9 章介绍不确定关联性流量 incast 的协同传输管理。当上层应用在数据中心内产生关联性 incast 流量时，其内部诸多数据流的发送端和接收端已经确定。第 7 章针对这类关联性流量介绍了如何实施网内聚合和协同传输。但是，很多数据中心应用面临计算节点和存储节点选择的多样性，不同的选择方案会导致对应的 incast 表现出不同的发送端和接收端。这类关联性流量被定义为不确定性 incast，与确定性 incast 流量相比，其具有更多机会来获取更大的数据流网内聚合增益。本章首先深入剖析了不确定 incast 流量的网内聚合问题，并设计了对应的协同传输方法以获取尽可能大的数据流网内聚合增益，包括为 incast 的各个数据流初始化发送端和构造 incast 聚合树两个环节。数据流发送端初始化的目标是令初始化后的全体发送端形成最少数目的群组，每个群组输出的数据流在传输的下一跳网络设备上即可被全部聚合为一个新数据流。为了充分利用初始化环节产生的这种优势，本章提出了两种 incast 聚合树的构建算法。实验结果表明，从减少网络流量和节省网络资源的角度来看，不确定性 incast 传输要优于确定性 incast 传输。

第 10 章介绍关联性流量 multicast 的协同传输管理。多播协议的出发点是从一个发送端将相同的内容传输给一组接收端，进而有效节约网络带宽并降低发送端的负载。数据中心的分布式文件系统为每个数据块提供多个副本，此时传统的 multicast 面临发送端的多样性问题，不再依赖于某个唯一选定的发送端，每个接收端只需从其中一个发送端获得发送内容。本章关注如何使这种发送端不确定的多播造成的网络传输代价尽可能地小，提出了对应的链路代价最小多播森林(minumum cost forest，MCF)模型。针对确定性 multicast 的方法不适用于 MCF 这个 NP 难问题，为此本章提出了两种高效的 MCF 近似算法，即 P-MCF 和 E-MCF。本章在 3 种类

型的数据中心互联结构(随机网络、随机正则网络以及无标度网络)下对MCF问题进行了评估。结果显示：3种网络互联结构下不确定性multicast的MCF都比确定性multicast的最小斯坦纳树占用更少的网络链路资源。

我们的研究工作得到了国家自然科学基金优秀青年科学基金项目(No.61422214)、国家重点基础研究发展计划("973"计划)青年科学家项目(No.2014CB347800)、湖南省自然科学杰出青年基金项目(No.2016JJ1002)、教育部新世纪优秀人才计划项目以及国防科学技术大学杰出青年基金项目(No.JQ14-05-02)的资助。国防科学技术大学的研究生赵亚威参与了本书1.1节和1.2节内容的撰写工作，研究生胡智尧、任棒棒、史良等同学参与了本书的排版、图表绘制、整理等工作，在此一并表示感谢。

由于作者水平所限，加之数据中心网络的互联结构和内部流量管理的研究仍处于快速发展和变化之中，书中错误和不足之处在所难免，恳请专家、读者予以指正。

郭得科

2016年2月于长沙

目录

第 1 部分　基础知识