数据中心网络的研究进展与趋势

李丹[1,2)]，陈贵海[3)]，任丰原[1,2)]，蒋长林[1,2)]，徐明伟[1,2)]

[1)](清华大学计算机科学与技术系，北京 中国 100084)

[2)](清华大学科学与技术国家实验室(筹)，北京 中国 100084)

[3)](上海交通大学计算机科学与工程系，上海 中国 200240)

摘 要 作为云计算的基础设施和下一代网络技术的创新平台，数据中心网络的研究成为了近年来学术界和工业界关注的热点。本文围绕数据中心网络研究的基本问题，介绍了国际国内的研究现状，包括数据中心网络拓扑设计、传输协议、无线通信、增强以太网、虚拟化、节能机制和 SDN（软件定义网络）等，并展望了数据中心网络的发展趋势。

关键词 数据中心网络；虚拟化；软件定义网络

中图法分类号 TP393 DOI 号：

Data Center Network Research Progress and Trends

LI Dan[1, 2)], CHEN Gui-Hai[3)], REN Feng-Yuan[1, 2)], JIANG Chang-Lin[1, 2)], XU Ming-Wei[1, 2)]

[1)](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[2)](Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, China)

[3)](Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract As the key infrastructure of cloud computing and innovation platform for next-generation networking, recently data center network has been a hot research topic in both academia and industry. Based on the fundamental research problems in data center network, we first describe the international and domestic research progress of this area in this paper, including data center network topology design, transport protocol, wireless communication, enhanced Ethernet, virtualization, energy saving, SDN (software defined networking), etc., and then prospect the research trend.

Key words data center network; virtualization; SDN

1 引言

作为云计算的核心基础设施，数据中心在近年来得到了学术界和工业界的极大关注。数据中心网络是连接数据中心大规模服务器进行大型分布式计算的桥梁，因此更成为了热点中的热点。究其原因，主要有以下几点。

第一，云计算的核心价值之一在于大数据的集中处理。随着数据中心流量从传统的“南北流量”为主演变为“东西流量”为主，对网络带宽和性能提出了很高的挑战。由于传统数据中心网络不能很好地满足云计算大数据处理的带宽要求，数据中心网络已经成为现代云计算的瓶颈所在[1-3]。设计新型的数

本课题得到国家自然科学基金(No. 61170291)和 973 项目(2014CB347800)资助. 李丹，男，1981 年生，博士，副教授，硕士生导师，主要研究领域为互联网体系结构和协议设计、数据中心网络、软件定义网络. E-mail: tolidan@tsinghua.edu.cn. 陈贵海，男, 1963 年生, 博士，教授，博士生导师，主要研究领域为分布式系统协议、数据中心网络. E-mail: gchen@cs.sjtu.edu.cn. 任丰原，男，1970 年生，博士，教授，博士生导师，主要研究领域为网络拥塞控制、数据中心网络. E-mail: renfy@tsinghua.edu.cn. 蒋长林，男 1980 年生，博士生，主要研究领域为互联网体系结构、数据中心网络和网络路由. E-mail:jiangchanglin@csnet1.cs.tsinghua.edu.cn. 徐明伟，男，1971 年生，博士，教授，博士生导师，主要研究领域为计算机网络体系结构、互联网路由和高性能路由器. E-mail: xmw@cernet.edu.cn.

据中心网络拓扑结构和传输协议，是提高云计算性能和用户体验、推动云计算发展的重要需求。

第二，云计算的另一重要特点是资源的统计复用，因此虚拟化技术在云计算中尤其重要。传统的计算虚拟化和存储虚拟化技术相对成熟，而网络虚拟化技术则发展缓慢（部分原因是缺乏云计算数据中心网络这样的需求平台）。由于网络资源的共享特性，为云计算租户的虚拟数据中心网络之间提供安全隔离、带宽保障和灵活调度，是实现云计算资源复用的必然要求。

第三，作为可控可管的大规模网络环境，云计算数据中心为网络技术发展提供了良好的创新平台。互联网技术创新的主要难点之一在于众多运营商之间的协调和博弈，缺乏部署实现的激励机制。而数据中心网络往往为云计算提供商所独有，云计算提供商为了提高服务性能和收益，有较强动力进行网络技术革新。今年早些时候Google公布已在数据中心网络全面部署OpenFlow①，就是一个典型的例子。

在标准化工作方面，国际互联网标准化组织IETF 成立了以数据中心网络为主要应用场景的工作组 Software Driven Networks，IEEE 也成立了针对数据中心网络的任务组 DCB（Data Center Bridge）。在工业界，Cisco、Juniper、华为等设备厂商先后推出了数据中心交换机产品；Amazon、Google、Microsoft、Facebook 等云计算提供商也在世界各地修建能容纳数万台甚至数十万台服务器的大型数据中心，并在网络架构、节能示范等方面进行了大胆革新。

本文主要介绍数据中心网络研究现状，并讨论数据中心网络的研究趋势。

2 研究现状

本节介绍数据中心网络的研究现状，包括拓扑设计、传输协议、无线通信、虚拟化、增强以太网、节能机制和软件定义网络（Software Defined Networking, SDN）等方面。

2.1 数据中心网络拓扑设计

传统数据中心网络普遍采用树型拓扑方案②。典型的拓扑由三层交换机互联构成，分别是接入层交换机、汇聚层交换机和核心层交换机。但实践证明这种拓扑方案已经不能很好地适应当前云计算数据中心的业务需求[1-3]。第一，树型拓扑对顶层网络设备的要求高，尤其当网络规模较大时；第二，网络存在单点失效问题，容错性差；第三，树形拓扑的网络带宽不足，无法较好地支持以“东西流量”为主的数据中心分布式计算。因此，近年来学术界对数据中心网络拓扑展开了广泛的研究。

当前提出的新型数据中心网络拓扑方案可以分为两类，分别是以交换机为核心的拓扑方案和以服务器为核心的拓扑方案。其中，在以交换机为核心的拓扑中，网络连接和路由功能主要由交换机完成。这类新型拓扑结构要么采用更多数量的交换机互联，要么融合光交换机进行网络互联，因此要求升级交换机软件或硬件，但不用升级服务器软硬件。代表方案包括 Fat-Tree[1,4]、VL2[2]、Helios[5]、c-Through[6]、OSA[7]等。在以服务器为核心的拓扑中，主要的互联和路由功能放在服务器上，交换机只提供简单的纵横式（crossbar）交换功能。此类方案中，服务器往往通过多个接口接入网络，为更好地支持各种流量模式提供了物理条件，因此需要对服务器进行硬件和软件升级，但不用升级交换机。具体方案包括 DCell[3]、BCube[8]、FiConn[9]、CamCube[10]、MDCube[11]、uFix[12]等。

2.1.1 以交换机为核心的拓扑方案

Fat-Tree[1,4]仍然采用三层拓扑结构进行交换机级联，如图 1 所示，但与传统树型结构不同的是，接入交换机和汇聚交换机被划分为不同的集群。在一个集群中，每台接入交换机与每台汇聚交换机都相连，构成一个完全二分图，每个汇聚交换机与某一部分核心交换机连接，使得每个集群与任何一个核心层交换机都相连。Fat-Tree 结构中提供足够多的核心交换机保证 1:1 的网络超额订购率（oversubscription ratio），提供服务器之间的无阻塞通信。典型 Fat-Tree 拓扑中所有交换机均为 1G 端口的普通商用交换机。

① OpenFlow – Enabling Innovation in Your Network [EB/OL], http://www.openflow.org/

② Cisco Data Center Infrastructure 2.5 Design Guide[EB/OL], http://www.cisco.com/application/pdf/en/us/guest/netsol/ns107/c649/ccmigration_09186a008073377d.pdf

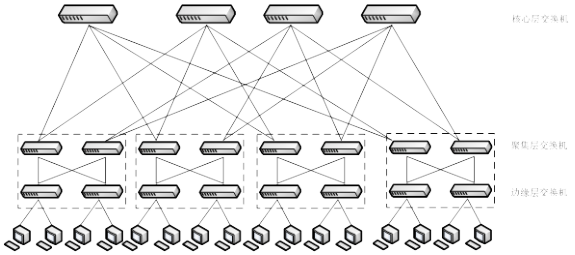

图 1　Fat-Tree 拓扑结构

与 Fat-Tree 一样，VL2[2]也通过三层级联的交换机拓扑结构为服务器之间的通信提供无阻塞网络交换。但不同的是，VL2 中的各级交换机之间都采用 10G 端口以减小布线开销。VL2 方案中，若干台（通常是 20 台）服务器连接到一个接入交换机，每台接入交换机与两台汇聚交换机相连。每台汇聚交换机与所有核心交换机相连，构成一个完全二分图，保证足够高的网络容量。

Helios[5]网络是一个两层的多根树结构，主要应用于集装箱规模的数据中心网络，其拓扑图如图 2 所示。Helios 将所有服务器划分为若干个集群，每个集群中的服务器连接到接入交换机。接入交换机同时还与顶层的分组交换机和光交换机同时相连。该拓扑保证了服务器之间的通信可使用分组链路，也可使用光纤链路。一个集中式的拓扑管理程序实时地对网络中各个服务器之间的流量进行监测，并对未来流量需求进行估算。拓扑管理程序会根据估算结果对网络资源进行动态配置，使流量大的数据流使用光纤链路进行传输，流量小的数据流仍然使用分组链路传输，从而实现网络资源的最佳利用。

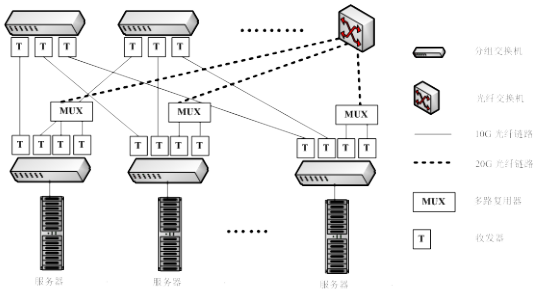

图 2　Helios 拓扑结构

c-Through[6]拓扑方案在传统的三层树型拓扑基础上，将所有接入交换机通过光交换机连接起来，构成一个融合了分组交换和电路交换的混合型网络拓扑，如图 3 所示。与 Helios 方案类似，c-Through 方案同样使用一个集中式的控制器对网络流量进行统计。与 Helios 不同的是，c-Through 方案对机架之间的流量（而不是各个服务器之间的流量）进行统计，并根据统计结果指示光交换机进行通信链路的动态配置。

OSA[7]在网络内部采用了全光信号传输，仅在服务器与接入交换机之间使用电信号传输。OSA 的应用场景是集装箱规模的数据中心网络。OSA 通过光交换机将所有接入交换机连接起来。由于服务器发出的都是电信号，因此 OSA 在接入交换机中放置光收发器（Optical Transceiver），用于光电转换；然后利用波长选择开关 (Wavelength Selective Switch)将接收到的不同波长映射到不同的出端口；再通过光开关矩阵(Optical Switching Matrix)在不同端口之间按需实现光交换。为了更有效地利用光交换机的端口，通过使用光环流器 (Optical Circulator)实现在同一条光纤上双向传输数据。Helios 和 c-Through 都只提供了一跳光信号传输，而 OSA 则实现了多跳光信号传输。不过在中间每一跳，都需要进行“光-电-光”的转换。OSA 的最大特点是利用光网络配置灵活的特点，能够根据实际需求动态调整拓扑，大大提高了应用的灵活性。

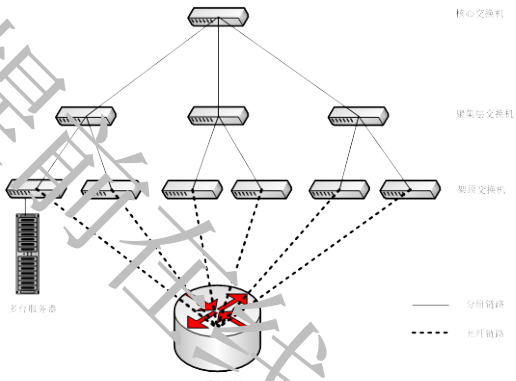

图 3　c-Through 拓扑结构

总的来说，引入光交换的拓扑方案的优势在于依靠光交换网络提供的高带宽、网络资源动态配置等优势为上层应用提供灵活的服务，而且有效地降低了组网的复杂度。但由于光交换网络是面向连接的网络，将不可避免引入时延，这将对搜索等对时延要求较高的应用带来影响。另外，与分组交换相比，光交换对于突发流量的支持不好。

2.1.2　以服务器为核心的拓扑方案

在以服务器为核心的拓扑方案中，大部分方案都具有层次性。这种分层拓扑结构的共同点是第 0

层网络都是由一台交换机连接若干台服务器构成，高层网络通过连接若干个低层网络构成。连接低层网络的方法有两种。一种是不通过交换机、直接通过服务器端口进行连接（即拓扑的层数受限于服务器的端口数），因此交换机只出现在第 0 层网络中。代表拓扑方案是 DCell[3]和 FiConn[9]。另一种是通过交换机进行连接，代表拓扑方案是 BCube[8]和 MDCube[11]。

在构建完整的 DCell[3]网络过程中，在由较低层次的网络互联构成较高层次的网络时，需要的低层次网络个数等于每个低层次网络中的服务器个数。互联的标准是每个低层次网络中的每台服务器分别与其它每个低层次网络中的某台服务器相连。DCell 拓扑结构如图 4 所示。如果将每个低层次网络看作一个虚拟结点，则高层次 DCell 网络是由若干个低层次 DCell 网络构成的完全图。DCell 拓扑的优势是网络可扩展性好。例如使用 6 口的小型交换机，构建三层 DCell 网络最多可以互联 3,263,442 台服务器。

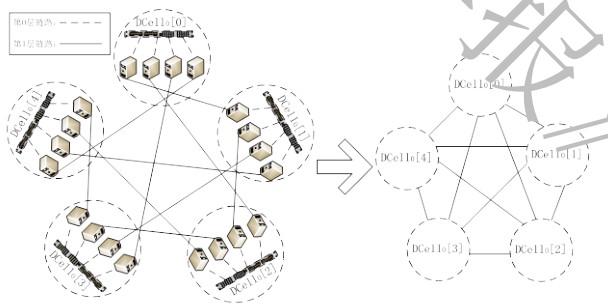

图 4 DCell 拓扑结构

FiConn[9]网络的构建方式与 DCell 网络相似，其拓扑结构如图 5 所示。但与 DCell 不同的是，FiConn 只使用具有两个端口(一个主用端口，一个备用端口)的服务器。其中主用端口用于连接第 0 层网络，备用端口用于连接高层网络。在进行层次化网络互联的过程中，每个低层 FiConn 网络中备用端口空闲的一半服务器会与其它相同层次的 FiConn 网络中备用端口空闲的服务器连接，构建高层次的 FiConn 网络。与 DCell 类似，高层次的 FiConn 网络是由若干个低层次的 FiConn 网络构成的一个完全图。该拓扑方案的优点是不需要对服务器和交换机的硬件做任何修改。

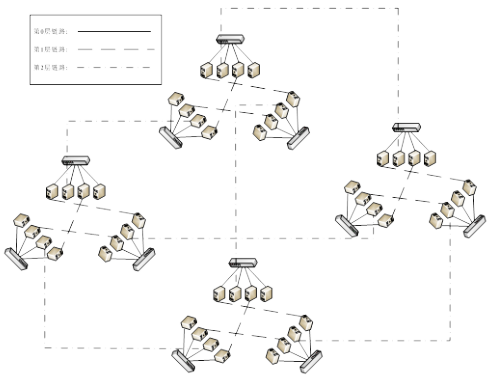

图 5 FiConn 拓扑结构

BCube[8]与 DCell 和 FiConn 的主要区别在于 BCube 使用交换机进行层次化网络的构建，拓扑结构如图 6 所示。BCube 通过若干个交换机将多个低层 BCube 网络互联起来，其中每个高层交换机与每个低层 BCube 网络都相连。在互联时，每层需要的交换机个数由 0 层网络中服务器的个数、以及层数决定。BCube 主要为集装箱规模的数据中心设计，其最大优势是链路资源非常丰富，同时采用集装箱构造有利于解决布线问题。

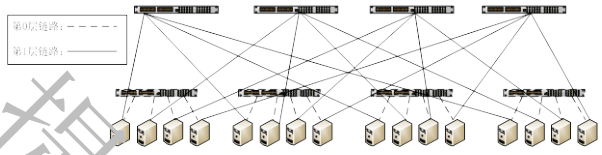

图 6 BCube 拓扑结构

MDCube[11]用于连接采用 BCube 构建的数据中心集装箱。MDCube 是一个多维的拓扑结构，它可以互联的数据中心集装箱的个数是所有维度上可容纳的数据中心个数的乘积。MDCube 使用光纤连接数据中心集装箱。如果我们将每个采用 BCube 拓扑构建的数据中心集装箱看作是一个虚节点的话，二维 MDCube 实质就是一个 HyperCube 网络。

此外，CamCube[10]同样是应用于集装箱数据中心的网络拓扑，其实质是一个 k-ary d-cube 网络。即该网络是一个 d 维网络，且每维有 k 台服务器。该拓扑方案的显著特点是整个网络中不使用任何交换机，是一个完全由服务器直接相连构成的网络。在每一维度中，所有服务器首尾相连，构成一个环。

uFix[12]也是为了对数据中心集装箱进行互联而提出的拓扑结构，但对集装箱内部所采用的拓扑结构没有限制，即可以连接异构的数据中心集装

箱。uFix 要求每个数据中心集装箱至少有一定数量的服务器空闲端口进行集装箱之间的互联网，互联时需要使用的端口数量由互联规模决定。在构建超大型数据中心网络时，使用 uFix 不需要对原有的数据中心网络做任何更新，只是利用服务器空闲端口进行互联，不需要增加额外的网络设备。

2.2 数据中心网络传输协议

与互联网一样，当前数据中心网络仍然广泛采用 TCP 传输数据。但一方面由于数据中心网络应用的特殊性，造成在特定情况下 TCP 会遭遇性能急剧下降，典型的现象是 TCP Incast[13]。另一方面，数据中心网络链路资源比互联网丰富很多，而传统的单路径传输层协议不能很好地利用拓扑中丰富的链路资源，导致网络资源的浪费。此外，数据中心不同的应用可能需要不同的传输协议，为数据中心特定应用定制相应的传输协议，也是近来的研究热点。

与此同时，数据中心网络的很多应用都具有典型的组通信模式（如分布式文件存储系统、软件升级等），而组播技术在互联网部署时遭遇的缺乏合理计价机制、安全性等问题在数据中心的集中可控环境中不复存在，因此数据中心也为组播提供了一个很好的应用平台。但数据中心组播的部署也面临新的挑战。目前数据中心网络的设计趋势是使用低端交换机进行网络互联，而用户需求可能导致大量的组播组，超过低端交换机的硬件限制。此外，云计算服务的服务质量需求对组播数据传输的可靠性也提出了较高的要求。

2.2.1 数据中心网络TCP协议

（1）TCP Incast

数据中心网络有一种通信模式是一台客户端向多台服务器同时发出数据请求（对应数据中心的典型分布式文件系统场景），这些服务器在收到请求后，会向发出请求的客户端同时返回数据。然而由于数据中心普遍采用低端交换机，这种通信模式会造成交换机的缓冲区溢出，发生丢包，从而导致网络吞吐量急剧下降，这种现象被称作 TCP InCast。发生 TCP InCast 的前提条件包括[14]：1)网络具有高带宽、低延时特性，而且交换机的缓存小；2)网络中存在同步的多对一流量；3)每条 TCP 连接上的数据流量较小。

针对 TCP Incast，学术界首先对其导致网络吞吐量下降的原因进行了分析。当交换机缓冲区溢出导致丢包时，TCP 通过两种方式进行数据重传，一是通过定时器机制触发超时重传，二是通过接收到重复的 ACK 报文触发快速重传。传统 TCP 的重传定时器（RTO）一般不低于 200ms，而数据中心网络环境的往返延迟（RTT）一般在微秒数量级，因此一旦发生超时重传，将会导致网络链路长时间处于空闲，从而造成网络吞吐量严重下降。文献[14-15]均认为丢包导致超时重传、且 TCP 重传 RTO 的超时时间与 RTT 值严重失配是导致 TCP InCast 的根本原因。文献[16]在此基础上，将通信过程中发生超时重传的类型划分为 BHTO (Block Head TimeOut)和 BTTO (Block Tail TimeOut)两类。其中，BHTO 一般发生于通信过程的开始阶段，由于整个窗口的数据被丢弃，从而发生超时重传。BTTO 一般发生于通信过程的结束阶段，由于最后的三个报文中至少有一个被丢弃，从而造成无法触发快速重传，导致超时重传。

近年来学术界提出了若干种方案解决 TCP Incast 问题。这些方案大致可以分为三类：1）减少 RTO 值，使之与 RTT 匹配[14-15]；2）设计新的拥塞控制算法[17-18]；3）采用基于编码的传输方案[19]。

文献[14-15]均通过减小发送方 RTO 定时器时间至微秒级别来减轻 TCP InCast 对系统性能的影响。由于 RTO 的值与 RTT 匹配，即使发生超时重传，发送端也能及时地重传丢失的数据，进行数据恢复，不会造成链路长时间处于空闲状态，从而保证网络吞吐量不会大幅下降。但将 RTO 值修改为微秒级别往往需要升级操作系统甚至硬件。另外，在光交换网络中，即使微秒级别的 RTO 对性能的改善也不会太明显。

ICTCP[17]和 DCTCP[18]通过设计新的拥塞控制算法来解决 TCP Incast 问题。ICTCP 通过实时监测接收方的流量速率来动态调整接收方的接收窗口，从而有效控制发送方的发送速率。ICTCP 对接收窗口的调节机制同样采用与拥塞窗口相同的机制：慢启动和拥塞避免。DCTCP 的核心思想是在不影响网络吞吐量的前提下，尽量保持交换机中队列长度较短。DCTCP 的实现利用了显式拥塞通知（Explicit Congestion Notification, ECN）功能，在交换机队列长度超过一定阈值时，向源端通告并让源端降低发送速度，避免丢包。同时，发送方收到拥塞通告后，不再单一地将拥塞窗口减半，而是根据当前网络的拥塞轻重情况相应地减小拥塞窗口大小。

文献[19]提出的解决方案摒弃了 TCP，使用 UDP

传输数据。由于 UDP 没有类似 TCP 的拥塞控制机制，因此即使发生丢包，也不会造成发送方降低发送速度或停止发送。但采用 UDP 同时也带来了挑战：一是 UDP 无法保证数据的可靠传输，二是 UDP 会不公平地抢占网络中其它 TCP 流的带宽。该方案使用了数据喷泉码保证数据的可靠传输，并通过部署 TFRC（TCP Friendly Rate Control）[20]保证 TCP 友好。

（2）多径 TCP

传统 TCP 协议在一对源端和目的端之间只建立一条连接传输数据。但如 2.1 节所述，为了解决数据中心网络的诸多问题，学术界在近几年提出了若干种新的“富连接”拓扑方案，如 Fat-Tree、VL2 和 BCube 等。这些拓扑中任意一对源端和目的端之间同时存在多条路径，为了充分利用链路资源、提高网络吞吐率，多径 TCP（MPTCP）[21-23]被引入到数据中心网络中。MPTCP 在同一对源端和目的端之间建立多个连接，源端将数据拆分成若干部分，使用不同的连接同时进行数据传输。MPTCP 相较于标准 TCP 的不同的地方有：第一，MPTCP 的连接建立操作比标准 TCP 复杂。MPTCP 在连接建立阶段，要求服务器端向客户端返回服务器端所有的地址信息，用于客户端建立连接使用。不同子流的源/目的地址信息可以相同，也可以不同。第二，各个子流维护自己的序列号和拥塞窗口。由于数据同时使用多个子流同时传输，因此接收端需要增加额外的操作，用于将从不同子流接收到的数据组装为原来的顺序。第三，MPTCP 同样采用 AIMD 机制维护拥塞窗口，但各个子流的拥塞窗口增加与所有子流拥塞窗口的总和有关[22]，从而能够保证将拥塞链路的流量往拥塞程度更轻的链路上转移。

（3）为应用定制的传输协议

与互联网为所有应用提供公共的传输协议不同，不同的云计算数据中心往往运行不同的典型应用，因此为特定的数据中心应用定制不同的传输协议，是数据中心网络传输协议研究的一个重要方向。比较有代表性的协议是 D^3[24]。D^3 针对数据中心的实时应用，通过分析数据流的传输数据大小和完成时间需求，为每个流显示分配传输速率。当网络资源紧张时，主动断开某些无法按时完成传输的数据流，从而保证更多的数据流能按时完成传输。实验表明，与传统 TCP 协议相比，D^3 可以大大增加数据中心的吞吐率。

2.2.2 数据中心网络组播协议

（1）数据中心可扩展组播

目前学术界在数据中心网络环境下提出的可扩展组播的解决方案大致可以分为两类：一类是将组播应用映射为网络层组播和单播，代表方案是 MCMD[25]。一类是采用 Bloom Filter 解决组播可扩展性问题，代表方案包括 ESM[26]和 MBF[27]。

MCMD 的核心思想是利用基于 Gossip 的控制协议[28]公告组成员信息，实现全网范围内所有节点中组信息的松散一致。为了优化组播地址的使用，MCMD 截获组播报文，并基于 Gossip 控制平面选出一个网络节点，根据组播组的规模和设置的策略将应用层的组播会话翻译为网络层组播地址或一组单播地址。MCMD 方案具有以下特点：1)可扩展性强，可以进行策略设置（例如决定每个组播组的成员数量或是否禁止使用 IP 组播）；2)对应用层透明，无需升级硬件，易于部署；3)使用基于 Gossip 的控制平面传播组成员信息，容错性好。

ESM[26]使用 Bloom Filter 保证组播方案的可扩展性。在分组内增加使用 Bloom Filter 编码的组播组信息虽然能够消除硬件限制对组播方案可扩展性的限制，但同时也增加了网络带宽的利用。因此该方案对成员较多的组播组使用传统的往交换机中注入路由转发表的方式实现组播，对成员较少的组播组(这类组播组在数据中心网络中占多数)采用在分组内增加 Bloom Filter 编码信息、利用编码信息实现路由转发。

MBF[27]通过理论分析和实验发现，在 Bloom Filter 中，如果每个元素的出现概率不同，为每个元素分配不同数量的哈希函数，结果产生的假真率比标准 Bloom Filter 低。基于该想法，MBF 按照每个组播组的出现概率相应地决定哈希函数的个数，出现概率大的组分配较少数量的哈希函数，而为出现概率小的组分配较多数量的哈希函数，从而降低了采用 Bloom Filter 进行组播数据转发带来的流量开销。

（2）数据中心可靠组播

由于数据中心普遍采用低端交换机并且链路资源丰富，传统互联网的可靠组播方案并不适用于数据中心网络。RDCM 是一种针对数据中心网络设计的可靠组播方案[29]。RDCM 在组播树上结合数据中心网络拓扑特征建立显示的覆盖网络，并在组成员中以 P2P 的方式恢复丢失的报文。RDCM 的优势包括：1)由终端恢复丢失的报文，不需要网络设备

的支持；2)利用数据中心网络中丰富的链路资源，采用组成员协作的方式实现故障隔离，有效地提高网络吞吐量。

2.3 数据中心无线通讯技术

由于传统数据中心普遍采用以太网静态链路和有线网络接口，大量的高突发流量和高负载服务器会降低数据中心网络的性能，而无线网络的广播机制可以顺利克服这些限制。同时由于极高频技术（特别是 60GHz 无线通信技术）的产生，使得无线通信亦可高速传输数据（吞吐量可达 4Gbps），因此近年来研究者开始尝试在网络瓶颈部分使用无线链接来分流有线链路的数据流。这种新技术可以降低布线复杂度，减小数据中心成本，并大幅提高数据中心网络的性能。此外，把 60GHz 无线通信技术应用在数据中心网络还有其它几大优势：1）7GHz 的可用频谱（57-64GHz）使 60GHz 无线通信技术能够提供 Gbps 量级速度的多条链接；2）60GHz 频段在减少无线信号干扰的同时也减少了被监听的机会；3）无线网络更益于数据中心网络的扩容和提升；4）无线网络可以按需建立，它能动态改变数据中心网络的拓扑结构，使其更适合当前网络环境。

2008 年出现了首篇讨论在数据中心网络中应用 60GHz 无线通信技术的论文[30]，随后另有多篇学术论文[31-39]从设计构建和性能优化两个方面讨论无线通信技术在数据中心网络的应用。

在设计构建方面，当前的研究主要借助特殊物理环境[31]和波束成形 (Beamforming) [32]、有向天线 (Directional Antenna)[33]技术使 60GHz 无线链路能够有效部署在数据中心网络中。该方向的研究目标是证实无线通信技术应用在数据中心网络中的可行性和优越性。

在性能优化方面，目前研究方向主要集中于无线通信在数据中心网络中的调度问题，即无线网络中的信道分配问题。自 2009 年文献[39]指出可以增加新的“飞路”（Flyways）来缓解部分“热节点”（Hot Node）的拥塞状况以来，其后的研究[34-36]多尝试应用启发式算法通过分流解决部分节点过热问题，使得数据中心网络整体吞吐量最大化，或整体利用率最大化。

下面分别针对无线网络设计构建与无线通信优化问题两个方面进行论述。

2.3.1 无线数据中心网络的设计与构建

为了在有线数据中心网络中使用无线通信，首先应从物理层面考虑在数据中心网络中使用 60GHz 无线通信技术的可行性及实用性，并通过合理的机柜摆放及无线节点空间排布，形成有效的整体系统结构，使数据中心网络性能得到大幅提高。现有工作对此无线网络可行性方面的研究主要包括：1）针对天线技术的讨论；2）机柜摆放对无线网络的影响；3)全无线数据中心网络的可行性。

（1）天线技术

在对“飞路”系统[33]的研究中指出，使用有向天线可以让信号变得清晰，还可以避免干扰，保证无线链接的稳定性。不仅是“飞路”系统，几乎所有的数据中心无线通信设计方案都使用了有向天线技术。另外还使用了波束成形（Beamforming）技术[32]和波束转向（Beamsteering）技术[37]。这些技术可以提高链接传输速率并增大数据传输带宽。

（2）机柜摆放

由于 60GHz 无线通信技术的天线覆盖范围极其有限，广播半径只有 10 米左右[33]，广播半径成为空间排布问题中首要因素。通过合理的空间排布，能够使得 60GHz 无线通信技术的覆盖范围内，服务器对无线链路的使用率最大化，并且如果采用有线与无线相结合的方式，便可以很好的权衡布线的数量与布线的长度，从而减少机柜之间的布线成本。

（3）全无线数据中心网络

除了在有线数据中心网络中添加无线链接，学术界还提出了更大胆的设计。康奈尔大学和微软雷德蒙德研究院应用 60GHz 射频技术设计了一个全无线数据中心[38]，并设计了配套的拓扑结构和路由协议，降低了数据中心费用，增大了容错能力。尽管目前该技术尚未投入实用，但这给后来的研究工作展示了全新的发展方向。

总体来说，应用 60GHz 无线通信技术在数据中心网络中建立新的链路是可行且高效的，通过合理的空间利用，数据中心网络的整体性能可大幅提高，甚至理论上有可能搭建出全无线数据中心。

2.3.2 无线数据中心网络的通信优化问题

除了设备与物理环境，还需要从系统性能优化的角度来设置无线链路，使得无线通信技术能够准确、高效的使用在数据中心网络中。下面讨论两种无线数据中心网络的设计方案以及相关的优化问

题，即“飞路”方案和无线信道分配问题。

（1）“飞路”方案

“飞路”是利用无线通信技术解决数据中网络中部分过热点的著名设计方案[39]。此方案的主要思想是通过在原有数据中心网络拓扑结构中添加一些新的链接（即“飞路”）分流过热的交换机之间的数据流，从而突破传输瓶颈，提高数据中心的整体性能。主要思路是运用贪心算法将网络中流量最大的链路分摊至其他可行路径，由此得到效用最高的无线链接方式。

之后，研究者在“飞路”方案基础上提出了一个更为细致准确的系统[33]，不仅从硬件上提出可行方案，同时通过模拟仿真验证了“飞路”方案在很大程度提高了数据中心网络流量，缓解了部分节点过热的问题。总的来说，“飞路”系统多数情况下可使数据中心网络流量提速 45%。但是，由于数据中心网络中某些流量的不可预计性或不可跟踪性，“飞路”算法有可能失效。

（2）无线信道分配问题

关于“飞路”系统的研究普遍忽视了无线信道间的干扰问题，因此研究者在建立优化模型时加入了对此问题的讨论[35]。他们设定了带干扰限制的最大化问题，目标函数是所有无线链接效用之和，同时基于匈牙利算法设计了一种启发式算法来解决该问题。仿真结果显示，运用此算法后热节点的负载大幅减小。此外，还应考虑自适应传输速率，着眼于全局工作完成时间[34]。

2.4 数据中心增强以太网

目前数据中心包含独立的三种交换网络，分别服务于存储业务、计算集群和互联网业务。其中，存储网络中数据是以块为单位传输的，分组丢失将导致整个数据块的重传，从而严重影响存储网络的性能；高性能计算需要网络具有较低的传输延迟，同时要求 I/O 设备处理数据的延迟低；服务于互联网业务的则是常用的以太网。

2.4.1 扁平一体化数据中心交换网络

寻求一体化的网络交换结构，不仅可以减少数据中心中冗余的设备和接口，降低成本，而且可以降低网络的复杂度。将三种交换网络整合起来，有利于协调数据中心各种资源的管理，建立统一的资源优化管理策略使服务器、存储和网络等子系统协调一致的工作，提高资源的统计复用率。数据中心一体化交换网络应该具有以下特性：1）必须具有低延时无丢失的特性，以满足高性能计算和存储的需要；2）必须同时支持现有的典型的互联网、存储和高性能计算的协议，如 TCP/IP、FC、InfiniBand 等；3）必须为网络设备开发统一的网络接口，以实现减少冗余设备，实现以太网、存储网络和高性能计算网络的互连。同时，该接口应该具有一定网络流量的管理支持；4）应该在一定程度上实现互联网、存储和高性能计算相关流量共享带宽的公平性和高利用率。

因为成熟的以太网技术得到了普遍应用，且有潜力提供超高带宽，10G以太网已经投入商业应用，40/100G以太网络标准正在由IEEE P802.3ba工作组负责研究制定 ①，增强以太网被选择为数据中心中的一体化交换网络技术。目前，IEEE 802.1 数据中心桥接（Data Center Bridging, DCB）工作组正致力于标准化以太网的一系列增强机制 ②。众多主流网络设备供应商，如Cisco、Juniper、华为、IBM、NetAPP、HP、Brocade、Fulcrum、Qlogic、Nuova等都参与了其中的标准化工作。这一增强的以太网被称为数据中心以太网（Data Center Ethernet, DCE）、融合增强的以太网（Converged Enhanced Ethernet，CEE）或者数据中心光纤（Data Center Fabric, DCF）。同时，INCITS T11 工作组开发了FCoE技术，使得DCF能够承载使用FC协议的存储流量 ③；OFED开发了RoCEE，使得DCF能够承载使用RDMA协议的高性能计算流量 ④，Myricom亦公布了MXoE，使得DCF能够承载使用MyriNet协议的高性能计算流量 ⑤。另一方面，设备厂商如思科（Cisco）、博科（Broadcom）、Juniper等纷纷推出在数据中心使用DCF的商业解决方案 ⑥⑦⑧。

① IEEE P802.3ba: 40Gb/s and 100Gb/s Ethernet Task Force[EB/OL], http://www.ieee802.org/3/ba/

② Data Center Bridging Task Group[EB/OL], http://www.ieee802.org/1/pages/dcbridges.html

③ Fibre Channel over Ethernet[EB/OL], http://www.t11.org/fcoe

④ Remote direct memory access over the converged enhanced ethernet fabric: evaluating the options[EB/OL], http://www.hoti.org/hoti17/program/slides/Panel/Talpey_HotI_RoCEE.pdf

⑤ Open-MX: Myrinet Express over Generic Ethernet Hardware[EB/OL], http://open-mx.gforge.inria.fr/

⑥ http://www.cisco.com/en/US/netsol/ns340/ns394/ns224/products.html

⑦ http://www.juniper.net/us/en/solutions/enterprise/data-center/

⑧ http://www.brocade.com/solutions-technology/industry/data-center/index.page

2.4.2 数据中心增强以太网流量控制

因为数据中心链路复用率高，流量突发性强，而且一体化交换网络必须满足存储流量对无丢失的需求和高性能计算流量对低延迟的需求，要将以太网从“尽力而为（Best-effort）”的传输模式改造成更适合数据中心环境的使用一体化交换网络技术，流量控制必不可少。目前，IEEE 802.1 DCB工作组中，主要有两个子工作组在负责标准化流量控制相关的增强机制。1）工作组 IEEE 802.1 Qbb负责标准化优先级流控（Priority-based Flow Control, PFC）机制[①]，该机制主要使用基于优先级的Pause的方法，避免因流量突发导致丢包，满足存储流量无丢失的需求。2）工作组 IEEE 802.1 Qau负责标准化端到端的拥塞控制机制[②]。通过PFC和端到端的拥塞控制这两个增强机制的配合，增强的以太网应该能够适应成为数据中心中的一体化交换网络技术。

PFC 机制首先是根据存储、高性能计算和交互流量的不同，为流量划分了优先级，从而使得要求低延迟的高性能计算流量得到延时保障。另外，PFC 主要是将 802.3x 定义的 Pause 机制，应用于不同的优先级上面。在同一优先级的流量过大导致交换机缓存将要溢出时，Pause 机制被触发，该优先级的流量被禁止注入交换机缓存，从而避免交换机缓存溢出。文献[40]探讨了 Pause 的时间，给出了一种动态设置 Pause 的时间的方案。文献[41]建议为不同的优先级流量设置不同的拥塞控制参数。

迄今为止，借鉴互联网拥塞控制机制的经验，802.1Qau 工作组共提出了 4 个提案，分别为后向的拥塞通告（Backward Congestion Notification, BCN），显式拥塞通告（Explicit Ethernet Congestion Notification, E2CN），前向显式拥塞通告（Forward Explicit Congestion Notification, FECN）和量化的拥塞通告（Quantized Congestion Notification, QCN）。目前，QCN 协议已经于 2010 年被批准为相应的标准。QCN 的目标是把瓶颈链路的队列长度控制在目标点，使得既不会出现缓存溢出、也不会出现缓存排空的情形。这样做能为应对突发流量预留缓存，降低排队延迟，抑制长期拥塞，保证链路利用率。然而，交换机的缓存一般较小，远小于路由器的缓存大小，这对链路层拥塞控制算法提出了很高的稳定性要求。文献[42]使用 Averaging Principle 的方法，分析了 QCN 的稳定性，证明 QCN 在延时不超过给定界限的情形下是稳定的，但是该文使用了经典的频域分析方法，该方法无法抓住 QCN 系统在加速和减速之间的切换过程的特征。因此文献[43]采用相平面分析方法分析 QCN，发现 QCN 可能会持续在加速和减速之间的切换，亦即进行滑模运动，直至进入平衡状态。但是 QCN 是启发式的设计，无意识的利用了滑模运动与系统参数和网络环境无关的特性。某些情况下，QCN 可能会无法进入滑模运动状态，变得不稳定。因而文献[44]设计了 SMCC 协议。SMCC 在任意情况下，都能够进入滑模状态，从而到达平衡点。

另外，QCN 的设计中，公平性没有被考虑。文献[45]发现 QCN 的公平性很差, 认为 QCN 反馈方式导致不公平，提出了 AF-QCN 协议，将 AFD 算法引入 QCN，以改善 QCN 的公平。文献[46]认为 QCN 的加速周期的设计使得它自己不公平。通过改变 QCN 的加速周期设计，改进了 QCN 的加速算法，使得 QCN 变得公平。文献[41,47]仿真测试了实际使用 DCF 的网络并得出结论: PFC 能提升 TCP 的性能，而 QCN 对 TCP 性能的影响强烈依赖于环境和参数。

2.5 数据中心网络虚拟化

数据中心有海量的计算和存储资源。随着技术的进步，计算资源和存储资源已经分别完成虚拟化。例如 Citrix 公司的 XenServer，VMware 公司的 vSphere 和 Microsoft 的 Hyper-V 为服务器虚拟化提供了良好的基础平台；EMC 公司的 Rainfinity 全局文件虚拟化网络存储管理系统，HP 公司的 StorageWorks 虚拟阵列以及 IBM 公司的 SAN Volume Controller（SVC）可将存储区域网中的各种存储设备整合成一个虚拟存储池。通过虚拟机（Virtual Machine, VM）技术，云计算租户（Tenant）可以使用数据中心的计算资源而不用担心物理计算机的管理、维护和升级。通过网络存储技术，用户可以在数据中心存储几乎无限量的数据，而不用担心数据的备份和安全。通过这种计算和存储的虚拟化，按需使用、按需付费的理念正在变成现实，并成为未来人们使用数据中心的主要方式之一。

当用户所需的计算资源多于一个虚拟机时，则需要租用多个虚拟机。通常多个虚拟机之间需要进行信息交互，而虚拟机之间需要网络互联和通信，

① 802.1Qbb - Priority-based Flow Control[EB/OL],http://www.ieee-802.org/ 1/ pages/802.1bb.html

② 802.1Qau - Congestion Notification, http://www.ieee802.org/1/ pages/ 802.1au.html

因此这些虚拟机组成了虚拟网络。这些虚拟机所组成的网络之所以被称为虚拟网络，是因为这些网络实际上共存于同一物理网络之上。相对于服务器和存储系统的虚拟化，网络虚拟化技术的发展相对滞后，主要原因在于之前没有类似于数据中心这种特定应用需求推动它的发展。

数据中心网络虚拟化面临的技术挑战主要包括如下几个方面。1）虚拟网络隔离。出于安全考虑，不同租户的虚拟机所形成的虚拟网络需要进行隔离，属于不同虚拟网络的虚拟机在缺省配置下应不能互相通信。不同租户可能使用相同的 IP 地址或 MAC 地址。传统的虚拟局域网（VLAN）虽然可以隔离不同的广播域，但 VLAN 子网分割的 4096 主机数限制影响网络规模的扩展性。2）虚拟机迁移。为适应数据中心资源共享和服务器整合的需要，虚拟机应该具备实时迁移的能力。基于二层网络的虚拟机迁移方案受二层网络的规模限制，难以扩展；基于三层网络的传统移动 IP 机制实现开销过大，难以适应大规模数据中心较为频繁的虚拟机实时迁移。3）带宽共享和保障。由于不同的虚拟网络共享同一个物理网络，设计公平而高效的带宽共享机制非常重要。传统网络由服务器通过基于流的 TCP 机制来竞争网络带宽，但租户可以通过创建众多的 TCP 流来获得更高的带宽。下面分别介绍相关研究成果。

2.5.1 数据中心网络隔离和虚拟机迁移

VXLAN[48]和 NVGRE[49]是当前 IETF 制定的网络虚拟化的报文格式标准。VXLAN 把虚拟机的以太网数据分组封装在 UDP/IP 数据分组内，而 NVGRE 是利用标准 GRE 格式来封装来自于虚拟机的以太网分组。在这两种格式中，用户的虚拟网络 ID 均是 24 比特，从而大大提高了数据中心可以支持的虚拟网络数量，并且可以跨三层实现虚拟机迁移。这两种格式目前均得到了工业界的支持。

NetLord[50]提出了把租户虚拟机的以太网分组封装在第三层的 IP 分组上的做法来实现网络的虚拟化及多租户支持。通过封装，租户虚拟机使用的 MAC 地址并不会出现在转发表中，不但解决了不同租户的 MAC 地址空间重叠的问题，而且大大减小了交换机的转发表空间。这种方案并不需要对网络核心设备进行升级。

VL2[2]使用两层 IP 地址空间，即 AA（Application Address）和 LA（Location Address）来解决租户 IP 地址冲突的问题，AA 由应用程序使用，LA 用于路由。PortLand[4]采用了类似的方法，引入了两层 MAC 地址空间，即 PMAC（Psuedo MAC）和 AMAC（Actual MAC），AMAC 由虚拟机实际使用，而 PMAC 用于路由交换。

2.5.2 数据中心网络带宽共享机制

目前数据中心网络的带宽共享机制主要基于两种思路，一是基于竞争，二是基于分配。基于竞争的方案的基本思路是在虚拟机或租户级别实现带宽竞争。和传统的基于 TCP 流的竞争方式不同，这种竞争方式可以防止应用程序通过增加流数目的方式来骗取网络资源，确保一定的公平性。典型的基于竞争的带宽共享机制包括 Seawall[51]，Netshare[52]和 Faircloud[53]等工作。Seawall [51]在虚拟机的层面上设计了一个拥塞控制系统来避免不公平的带宽分配，Netshare[52]和 Faircloud[53]则通过一个集中控制的管理器，根据每个服务的不同需求为其分配对应的权值，最后根据权重对全网的带宽资源以 max-min fairness 的方式进行竞争。这样可以在提高利用率、不浪费带宽的同时，确保需求更大、优先级更高的应用程序可以获得更多的带宽。其中 Faircloud 深入研究了各种基于权值进行带宽分配机制的解决方案满足的特性。这种基于租户需求指定权值进行分配的模式相对于传统的模式而言具备更高灵活性。

基于分配的方案确切定义每个虚拟机或者每个租户对网络带宽的需求，直接给虚拟机分配足量的带宽，并且通过限速机制来确保每个虚拟机或者租户对带宽的利用不会超过分配的限额。多种模型被提出来定义租户的带宽需求。比较常见的模型包括流量矩阵模型和“软管”模型。流量矩阵模型确定每一对虚拟机之间的带宽需求，而软管模型只限制每个虚拟机的进出总带宽。在这两种分配模型上，比较典型的分配机制包括 SecondNet[54]和 Oktopus[55]的工作。相对于基于竞争的机制而言，基于分配的方案可以提供真正的带宽“保障”，但缺点是租户可能无法用足所申请的网络带宽，从而造成网络资源浪费。在基于带宽显式分配的模型下，文献[56-57]研究了云计算多租户的带宽计价策略。文献[58]则首次运用博弈论从理论角度深入分析了云计算多租户环境下带宽共享机制。

2.6 数据中心网络节能机制

数据中心日益增长的规模虽然满足了用户的需求，但同时大幅增加了数据中心网络的能耗。据

Cisco公司提供的数据，2009 年美国境内所有的数据中心网络支付的电费高达 33 亿美元[①]。

随着新能源（如太阳能、风能）的日益普及，一些研究致力于将新能源引入数据中心网络中、降低能耗开销。文献[59]是国内第一篇关于新能源在数据中心网络中应用的综述，并给出了其发展趋势。而文献[60]则通过严谨的理论分析给出了调度多种能源的算法 SmartDPSS，以保证运营商以最低的代价为用户提供可靠的电力。SmartDPSS 的基本思想是为实时应用采用即时服务，而将对延迟不敏感的服务推迟至新能源充足或电价便宜的时候再进行服务。

与此同时，如 2.1 节所述，为了解决传统树型拓扑对数据中心网络的性能限制，近年来提出了多种“富连接”的网络拓扑，保证在峰值网络流量条件下网络依然能够保持较好的性能。一方面，这些新型拓扑较传统拓扑使用了更多的网络设备，加剧了数据中心网络的能耗。另一方面，数据中心网络中的流量多数情况下远低于峰值，从而造成数据中心能耗使用效率低下。因此，学术界和工业界越来越关注对数据中心网络节能机制的研究。

目前对数据中心网络节能机制的研究成果大致可以分为两类：一类是通过降低硬件设备的能耗达到节能目的，另一类是通过设计新型的路由机制降低能耗。

2.6.1 硬件设备节能机制

对网络设备能耗模型的研究表明，网络设备的能耗不仅与启用的线卡和接口数量有关系，还与接口速率大小有关系[61]。基于这样的研究成果，目前硬件设备节能机制的研究成果主要可以划分为两类，分别是设备休眠和速率调整。

(1) 设备休眠技术

设备休眠技术的原理比较简单，即通过动态切换网络设备的休眠/工作模式达到节能的目的：当网络设备没有流量经过的时候，将处于空闲状态的网络设备或网络设备的线卡或接口置于休眠模式；当处于休眠模式的网络设备需要处理流量时，唤醒处于休眠模式的网络设备，恢复到工作模式。休眠与工作模式的切换机制是设备休眠技术研究的重点。

文献[62]首先提出了基于定时器的转换机制。该机制的原理是当监测到网络处于空闲状态时，则根据当前网络的负载情况设定定时器的值，然后启动定时器，并将网络设备切换到休眠模式。直到定时器超时，再将网络设备切换到工作模式。在网络设备处于休眠模式期间到达的所有报文都将被丢弃。不难看出，这种机制虽然实现简单，但会造成报文的丢失，这不仅会加剧网络负载，还会影响用户体验。

后续的研究成果都沿用了定时器机制。其中考虑到网络设备实施休眠模式与工作模式的切换与到达的流量息息相关，文献[63]在定时器的基础上引入了流量整形技术，即通过对流量进行整形、控制流量到达设备的时间间隔，减少网络设备实施模式切换的频率，增加设备休眠时间，达到更好的节能效果。这种机制虽然能够更有效地减小能耗，但会增加报文时延。同时流量整形的效果将会直接影响节能的效果。另外，为了减少单独使用定时器造成报文丢失的影响，文献[64]和[65]提出结合使用定时器和缓存的设备休眠机制。与文献[62]的区别在于，当设备处于休眠模式时，到达设备的报文不再被丢弃，而是被缓存起来，当设备唤醒后，再将这些缓存的报文进行发送。虽然这种机制能够减少报文的丢失，却大大增加了报文时延。

(2) 速率调整技术

网络设备接口速率的配置与其能耗有直接的关系[61]。因此，速率调整技术的核心思想是通过实时调整网络设备的接口速率达到节能的目的。

文献[66]提出了一种链路速率调整机制。该机制综合考虑当前的链路速率、缓存队列长度和链路利用率，从而确定是否增加或降低链路速率：当链路速率较低、且缓存队列长度高于某个阈值时，该机制会调高链路速率；当链路速率较高、且缓存队列长度和链路利用率均低于相应阈值时，该机制会降低链路速率。当决定调整速率时，链路一方会向对端发送一个速率调整请求，并在该请求中标明期望达到的链路速率。对端在接收到速率请求后，确定是否接受速率调整。该方案能够使链路速率动态适配于网络负载，因此能够有效地减少链路的能耗，但频繁的速率切换会带来严重的额外开销。

为了避免频繁地实施速率切换、增加方案的可行性，文献[67]提出了三种速率切换策略。第一种是基于缓存队列长度双阈值的速率切换策略：该机制为缓存队列长度设置了高、低两个阈值，只有当缓存队列长度落入该阈值区间外再实施速率切换。但这种策略可能会因为突发流量等因素引起抖动，

① http://www.cisco.com/web/partners/downloads/765/other/Energy_Logic_ Reducing_Data_Center_Energy_Consumption.pdf

因此提出基于链路利用率阈值和超时阈值的两种策略。这两种策略在使用第一种策略的基础上，分别通过设定链路利用率阈值和超时阈值，降低速率调整的频率，尽量减少速率的抖动。

2.6.2 节能路由机制

近年来，硬件功耗的降低使得网络设备在能耗优化方面已经取得了很大的进展，但研究表明，数据中心流量的增长对网络设备性能和能耗需求的增长远远快于硬件工艺的发展。而且基于单设备的节能技术依赖于网络流量分布，所带来的节能效果非常有限。另外，随着“富连接”网络拓扑的应用，每对服务器之间存在多条路径，等价多路径等路由策略被广泛应用于数据中心网络中，以提高网络链路利用率。但这些路由策略在实现网络负载均衡的同时并没有考虑网络能耗，不可避免地浪费了大量能耗。因此，节能路由机制的研究应运而生。

节能路由机制的核心思想是通过合理的选择路由，只使用一部分网络设备承载流量，并对没有流量经过的网络设备进行关闭或置于休眠模式，达到节能的目的。但关闭部分网络设备势必会对网络可达性造成一定的影响，从而影响网络性能。因此在路由机制设计的时候需要在节能效果和网络性能之间进行权衡。

弹性树 (Elastic Tree)[68]提出了一种全网范围内的能耗优化机制。该机制的核心思想是持续不断地监测数据中心网络的流量状况，在保证预期的网络性能和网络可靠性的前提下，实时调整链路和网络设备的状态，只选择全网范围内尽可能小的一部分网络设备和链路处理网络流量，提高能耗利用率。该机制探讨了多种实现算法，最优化算法、基于 bin-packer 的贪心算法、以及拓扑感知的启发式算法。其中最优化算法的节能效果最好，但其计算复杂度也最高，实时性较差。拓扑感知的启发式算法也仅限应用于 Fat-Tree 拓扑。

Shang 等人[69]提出一种基于吞吐率约束的数据中心节能路由算法。他们首先分析了节能路由问题的求解复杂度，证明其与经典的 NPC 问题（背包问题）的求解难度相同，之后提出一种交换机迭代删除算法。该算法的基本思想是：首先计算数据中心网络的初始路由和初始吞吐率，之后按照一定的策略依次删除网络拓扑中的交换机，直到网络吞吐率下降到一个预定义的网络性能阈值。最终获得节能路由，并将空闲的交换机休眠以实现数据中心网络能耗的节省。

2.7 软件定义网络与数据中心

软件定义网络（Software Defined Networking，SDN）[70] 是最近学术界关注的热点。在 SDN 中，数据分组的转发与控制被分开。网络智能被抽取到一个集中式的控制器（Controller）中。数据流的接入、路由等都由 Controller 来控制。而交换机只是按 Controller 所设定的规则进行数据分组的转发。由于数据中心是由一个单位（公司、政府、研究机构等）所建设和拥有，天然符合 SDN 所需要的集中控制要求。因此当前 SDN 最主要的应用场景都集中在数据中心。把 SDN 就用到数据中心主要有如下优点：

（1）可管理性。通过集中式控制，网络运营者能及时掌握网络设备的状况，如网络设备是否工作正常，是否有网络拥塞发生，网络服务质量是否在正常范围等。这些都是以往的分布式网络协议很难提供的。

（2）网络性能优化。在以往的分布式网络协议中，由于网络运营者对网络缺少细粒度的控制，为了保障网络服务质量在一定范围之类，网络运营者被迫降低网络利用率。一般而言，核心网络的带宽利用率只有 30%到 40%。在数据中心中，由于运营者知道网络的详细拓扑结构及核心应用的需求，可以大幅提高网络带宽利用率。最近 Google 公布了他们利用 SDN 技术把数据中心间的核心网络带宽利用率提高到了 100%。高带宽利用率意味着可以利用 SDN 来降低传输每比特的花费（cost per bit）。

（3）更快引入新的网络功能。在 SDN 中，由于控制功能被抽象到集中式控制器中，而各个网络设备只是受控制器的控制并执行相对简单的分组转发功能，因此当需要引入新的网络功能，如新的网络带宽分配算法，或者实施新的网络接入控制策略，只需要变更控制器中的软件。这样可以在很短时间内引入新的网络功能。由于新功能的引入只涉及软件更新，在进行更新之前还可以对软件进行充分的测试。这样也解决了分布式网络协议的更新、调试及测试所面临的困难。

SDN当前的主要研究者和推动者是Stanford大学 的 Nick McKeown 和 UC Berkeley 的 Scott Shenker。非盈利性组织开放网络基金会（Open

Network Foundation, ONF[①])负责大力推进SDN的标准化工作。ONF的参与者包括多家工业界的核心企业，如思科，Juniper等网络设备厂商，Google，Facebook等互联网公司，以及Microsoft等软件公司，还有学术界的顶尖机构。

当前，SDN已不止停留在概念层面，而是已经有了非常坚实的具体实现：OpenFlow[②]。OpenFlow是第一个针对SDN实现的标准接口，包括数据层与控制层之间的传输协议，控制器上的API等。OpenFlow起源于Stanford的“Clean slate”计划(斯坦福大学的“Clean slate ”计划是一个致力于研究重新设计互联网的项目)，在 2008 年开始发布并进行推广。其研发成员组成由最开始的Stanford大学高性能网络研究组(The High Performance Networking Group)，逐渐扩展为众多学术界顶尖机构如MIT、UC Berkeley等，还有工业界的领头企业，如Cisco、Juniper等。

OpenFlow由控制器和OpenFlow交换机组成。这里不叫做OpenFlow“路由器”是因为路由计算的任务都是在控制器完成，其自身只完成数据转发功能，因此叫做OpenFlow“交换机”更合适。控制器和OpenFlow交换机之间是加密的OpenFlow协议，用以传输控制器制定的转发表项给交换机和传输交换机收到的“陌生”数据包给控制器。OpenFlow以数据流作为生成转发表项的依据。传统路由器对数据流的标识通常是一个五元组：源IP地址，目的IP地址，源端口号，目的端口号，协议号。这种数据流标识无法全面的标识数据中心的数据流特征，因此OpenFlow对数据流的标识除了上述五元组外，还可以有源、目的以太网地址，VLAN号，VLAN优先级，数据流如端口等信息，从而使用户可以从更加细粒度的层面对数据流进行控制。控制器上的API为用户提供了编程接口，OpenFlow的API并不对编程语言进行限制，如C++、Python等都可以，另外由于OpenFlow的开放特性，使得API提供的都是最基本的功能模块，用户可以根据自己需求生成更加复杂的API。已经有一些开源的OpenFlow控制器供开发者直接使用，如NOX[③]。

OpenFlow 的简单工作流程是：用户通过 API 在控制器编写自己的路由策略，分为静态和动态两种。静态路由策略可以简单的看作是用户通过控制器直接向 OpenFlow 交换机写入转发表，数据流进入交换机后，若匹配相应表项则直接进行转发，不再需要控制器参与。动态路由策略则是在控制器上维护一个守护进程，交换机收到的数据流若在当前转发表中没有匹配，则将该数据流的第一个数据包转发给控制器，由控制器的守护进程进行计算，动态实时地进行转发规则生成，并写入交换机。该数据流后面的数据包将按照此时生成的转发规则被匹配。

3 发展趋势与展望

随着云计算和大数据应用的飞速发展，以及网络在数据中心中的核心地位，数据中心网络已经成为了近年来引人瞩目的研究热点。国际学术界、国际标准组织、网络设备厂商、云计算提供商等都对数据中心网络研究给予了非常大的关注。由于数据中心网络领域的研究与工业界结合紧密，技术创新易于部署，可以预计在未来数年内数据中心网络的研究还将持续成为焦点。学术界关于数据中心网络的研究将成为云计算发展的有力助推器，并推动计算机网络体系架构和协议本身的创新。

当前国内关于数据中心网络的研究基本与国际学术界保持同步，在部分技术方向甚至处于领先地位。因此，加强数据中心网络的研究，对于推动我国云计算和下一代互联网产业发展，并在国际新一轮 IT 技术革新浪潮中取得话语权，有非常重要的影响。

在本文介绍的数据中心网络相关研究方向中，关于数据中心内部拓扑方案的研究渐渐受到较少关注，而数据中心之间的互联拓扑还有很大的研究空间。数据中心网络传输协议、虚拟化、节能机制、SDN 等方向的研究方兴未艾，可以预期将产生不少创新成果，尤其是应用定制的传输协议、虚拟网络带宽保障机制、服务器与网络的联合节能优化、适用于数据中心网路的 SDN 可编程网络节点等。近年来，关于下一代互联网体系结构的研究项目非常多，如美国NSF资助的NDN、Nebula、Mobility First、XIA 等项目，以及国内 973 资助的 IPv6、SOFIA、一体化标识网络等项目。如何将这些创新体系结构与数据中心网络结合，通过数据中心网络的可控可管易部署的环境为这些新型网络架构提供创新平

① Open networking foundation, https://www.opennetworking.org/index .php

② OpenFlow – Enabling Innovation in Your Network [EB/OL], http://www.openflow.org/

③ http://www.noxrepo.org/

台，也是值得关注的研究方向。

参 考 文 献

[1] M. Al-Fares, A. Loukissas and A. Vahdat. A scalable, commodity data center network architecture//Proceedings of the Special Interest Group on Data Communication (SIGCOMM). Seattle, Washington, 2008:63-74

[2] A. Greenberg, J. Hamilton, N. Jain, etc. VL2: A scalable and flexible data center network//Proceedings of the Special Interest Group on Data Communication (SIGCOMM). Barcelona, Spain, 2009: 95-104

[3] C. Guo, H. Wu, K. Tan, etc. DCell: A scalable and fault-tolerant network structure for data centers//Proceedings of the Special Interest Group on Data Communication (SIGCOMM). Seattle, Washington, 2008: 75-86

[4] R. Mysore, A. Pamboris, N. Farrington, etc. PortLand: A scalable fault-Tolerant layer 2 data center network fabric//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Barcelona, Spain, 2009: 39-50

[5] N. Farrington, G. Porter, S. Radhakrishnan, etc. Helios: A hybrid electrical/optical switch architecture for modular data centers// Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). New Delhi, India, 2010: 339-350

[6] G. Wang, D. Andersen, M. Kaminsky, etc. c-Through: Part-time optics in data centers//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). New Delhi, India, 2010: 327-338

[7] K. Chen, A. Singla, A. Singh, etc. OSA: An optical switching architecture for data center networks with unprecedented flexibility// Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI). SAN JOSE, CA, 2012: 18-18

[8] C. Guo, G. Lu, D. Li, etc. BCube: A high performance, server-centric network architecture for modular data centers//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Barcelona, Spain, 2009: 63-74

[9] D. Li, C. Guo, H. Wu, etc. FiConn: Using backup port for server interconnection in data centers//Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Rio de Janeiro, Brazil, 2009: 2276-2285

[10] H. Abu-Libdeh, P. Costa, A. Rowstron, etc. Symbiotic routing in future data centers//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). New Delhi, India, 2010: 51-62

[11] H. Wu, G. Lu, D. Li, etc. MDCube: A high performance network structure for modular data center interconnection//Proceedings of the ACM International Conference on emerging Networking EXperiments and Technologies (CoNEXT). Rome, Italy, 2009: 25-36

[12] D. Li, M. Xu, H. Zhao, etc. Building mega data center from heterogeneous containers//Proceedings of the IEEE International Conference on Network Protocols (ICNP). Vancouver, BC Canada, 2011: 256-265

[13] D. Nagle, D. Serenyi and A. Matthews. The panasas activescale storage cluster: delivering scalable high bandwidth storage//Proceedings of the ACM/IEEE conference on Supercomputing. Pittsburgh, PA, 2004: 53-62

[14] V. Vasudevan, A. Phanishayee, H. Shah, etc. Safe and effective fine-grained TCP retransmissions for datacenter communication //Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Barcelona, Spain, 2009: 303-314

[15] Y. Chen, R. Griffith, J. Liu, etc. Understanding TCP incast throughput collapse in datacenter networks//Proceedings of the ACM workshop on Research on enterprise networking (WREN). Barcelona, Spain, 2009: 73-82

[16] J. Zhang, F. Ren, and C. Lin. Modeling and understanding TCP incast in data center networks//Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Shanghai, China, 2011: 1377-1385

[17] H. Wu, Z. Feng, C. Guo, etc. ICTCP: incast congestion control for TCP in data center networks//Proceedings of the ACM International Conference on emerging Networking EXperiments and Technologies (CoNEXT), Philadelphia, USA, 2010

[18] M.Alizadeh, A. Greenberg, D. Maltz, etc. Data center TCP (DCTCP) //Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). New Delhi, India, 2010: 63-74

[19] C. Jiang, D. Li and M. Xu. A coding-based approach to mitigating tcp incast in data center network//Proceedings of the International Conference on Distributed Computing Systems (ICDCS) Workshop on Data Center Performance. Macao, China, 2012: 29-34

[20] S. Floyd, M. Handley, J. Padhye, and J. Widmer. Equation-based congestion control for unicast applications//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). New York, NY, USA, 2000: 43 – 56

[21] A. Ford, C. Raiciu, M. Handley, etc. TCP extensions for multipath operation with multiple addresses. June 2012, IETF draft (work in progress).

[22] D. Wischik, C. Raiciu, A. Greenhalgh, etc. Design, implementation and evaluation of congestion control for multipath TCP//Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI). Berkeley, CA, 2011:8-8

[23] C. Raiciu, S. Barre, C. Pluntke, etc. Improving datacenter performance and robustness with multipath TCP//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Toronto, Ontario, Canada, 2011: 266-277

[24] C. Wilson, H. Ballani, T. Karagiannis, etc. Better never than late: meeting deadlines in datacenter networks//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Toronto, Ontario, Canada, 2011: 50-61

[25] Y. Vigfusson, H. Abu-Libdeh, M. Balakrishnan, etc. Dr. Multicast: Rx for data center communication scalability//Proceedings of the ACM Workshop on Hot Topics in Networks (HotNets). Calgary, Alberta, Canada, 2008

[26] D. Li, J. Yu, J. Yu, etc. Exploring efficient and scalable multicast routing in future data center networks//Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Shanghai, China, 2011: 1368-1376

[27] D. Li, H. Cui, Y. Hu, etc. Scalable data center multicast using multi-class bloom filter//Proceedings of the IEEE International Conference on Network Protocols (ICNP). Vancouver, BC Canada, 2011: 266-275

[28] R. Renesse, Y. Minsky, and M. Hayden. A gossip-based failure detection service. Proceedings of the IFIP International Conference on Distributed Systems Platforms and Open Distributed Processing. The Lake District, England, 1998: 55-70

[29] D. Li, M. Xu, M. Zhao, etc. RDCM: reliable data center multicast //Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Shanghai, China, 2011: 56-60

[30] K. Ramachandran, R. Kokku, R. Mahindra, etc. 60 GHz data-center networking: wireless=>worry less?. Princeton, USA: NEC Laboratories, Technical Report, 2008.

[31] H. Vardhan, N. Thomas, S. Ryu, etc. Wireless data center with millimeter wave network//Proceedings of the Global Telecommunications Conference (GLOBECOM). Miami, Florida, USA, 2010: 1-6

[32] W. Zhang, X. Zhou, L. Yang, etc. 3D beamforming for wireless data centers//Proceedings of the ACM Workshop on Hot Topics in Networks (HotNets). Cambridge, MA, 2011

[33] D. Halperin, S. Kandula, J. Padhye, etc. Augmenting data center networks with multi-gigabit wireless links//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Toronto, Ontario, Canada, 2011: 38-49

[34] Y. Cui, H. Wang, and X. Cheng. Channel allocation in wireless data center networks//Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Shanghai, China, 2011: 1395-1403

[35] Y. Cui, H. Wang, X. Cheng, etc. Wireless data center networking //Proceedings of the IEEE Wireless Communications, 2011, 18(6): 46-53

[36] Y. Cui, H. Wang, and X. Cheng. Wireless link scheduling for data center networks//Proceedings of the International Conference on Ubiquitous Information Management and Communication (ICUIMC). Seoul, Korea, 2011

[37] Y. Katayama, K. Takano, Y. Kohda, etc. Wireless data center networking with steered-beam mmwave links//Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC). Quintana Roo, Mexico, 2011: 2179-2184

[38] S. Yong, S. Gün, W. Hakim, etc. On the feasibility of completely wireless data centers// Proceedings of the eighth ACM/IEEE symposium on Architectures for networking and communications systems (ANCS). Austin, TX, USA, 2012: 3-14

[39] S. Kandula, J. Padhye, and P. Bahl. Flyways to de-congest data center networks//Proceedings of the ACM Workshop on Hot Topics in Networks (HotNets). New York, USA, 2009

[40] M. Hayasaka, T. Sekiyama, S. Oshima, etc. Dynamic pause time calculation method in MAC layer flow control//Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). Shanghai, China, 2010: 1-5

[41] S. Fang, C. H. Foh and K. Aung. Differentiated Ethernet congestion management for prioritized traffic//Proceedings of the IEEE International Conference on Communications (ICC). Cape Town, South Africa, 2010: 1-5

[42] M. Alizadeh, A. Kabbani, B. Atikoglu, etc. Stability analysis of QCN: the averaging principle//Proceedings of the ACM Special Interest Group on Performance Evaluation (SIGMETRICS). San Jose, CA, 2011: 49-60

[43] F. Ren and W. Jiang. Phase plane analysis of congestion control in data center Ethernet networks//Proceedings of the International Conference on Distributed Computing Systems (ICDCS). Genoa, Italy, 2010: 20-29

[44] W. Jiang, F. Ren, R. Shu, etc. Sliding mode congestion control for data center Ethernet networks//Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Orlando, FL, 2012: 1404 - 1412

[45] A. Kabbani, M. Alizadeh, M. Yasuda, etc. AF-QCN: approximate fairness with quantized congestion notification for multi-tenanted data centers//Proceedings of the IEEE High Performance Interconnects (HOTI). Mountain View, CA, 2010: 58-65

[46] Y. Hayashi, H. Itsumi and M. Yamamoto. Improving fairness of quantized congestion notification for data center Ethernet networks //Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS) Workshops. Minneapolis, MN, 2011: 20-25

[47] A. Anghel, R. Birke, D. Crisan, etc. Cross-layer flow and congestion control for datacenter networks//Proceedings of the Workshop on Data Center - Converged and Virtual Ethernet Switching (DC-CAVES). San Francisco, CA, 2011: 44-62

[48] M. Mahalingam, D. Dutt, K. Duda, etc. VXLAN: a framework for overlaying virtualized layer 2 networks over layer 3 networks. IETF draft, 2011

[49] M. Sridharan, K. Duda, I. Ganga, etc. NVGRE: network virtualization using generic routing encapsulation. IETF draft, 2011

[50] J. Mudigonda, B. Stiekes, P. Yalagandula, etc. NetLord: a scalable multi-tenant network architecture for virtualized datacenters //Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Toronto, Ontario, Canada, 2011: 62-73

[51] A. Shieh, S. Kandula, A. Greenberg, etc. Sharing the data center

network//Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI). Boston, MA, 2011: 23-23

[52] T. Lam, S. Radhakrishnan and A. Vahdat, NetShare: virtualizing data center networks across services. California, USA: UCSD, Technical Report: CS2010-0957, 2010

[53] L. Popa, A. Krishnamurthy, S. Ratnasamy, etc. FairCloud: sharing the network in cloud computing//Proceedings of the ACM Workshop on Hot Topics in Networks (HotNets). Cambridge, MA, 2011

[54] C. Guo, G. Lu, H. Wang, etc. SecondNet: a data center network virtualization architecture with bandwidth guarantees//Proceedings of the ACM International Conference on emerging Networking EXperiments and Technologies (CoNEXT), Philadelphia, PA, 2010

[55] H. Ballani, P. Costa, T. Karagiannis, etc. Towards predictable datacenter networks//Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). Toronto, Ontario, Canada, 2011: 242-253

[56] D. Niu, C. Feng, and B. Li. On pricing cloud bandwidth reservations under demand uncertainty//Proceedings of the ACM Special Interest Group on Performance Evaluation (SIGMETRICS). London, UK, 2012: 151-162

[57] D. Niu, C. Feng, and B. Li. A theory of cloud bandwidth pricing for video-on-demand providers//Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Orlando, FL, 2012: 711-719

[58] Jian Guo, Fangming Liu, Dan Zeng, John C.S. Lui, Hai Jin. A Cooperative Game Based Allocation for Sharing Data Center Networks//Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Italy, April, 2013

[59] Wei Deng, Fangming Liu, Hai Jin, Dan Li. Leveraging Renewable Energy in Cloud Computing Datacenters: State of the Art and Future Research. Chinese Journal of Computers, 2010, 36(3): 582-588(in Chinese)
(邓维, 刘方明, 金海, 李丹, "云计算数据中心的新能源应用：研究现状与趋势," 计算机学报, 2013, 36(3): 582-588)

[60] Wei Deng, Fangming Liu, Hai Jin, Chuan Wu. SmartDPSS: Cost-Minimizing Multi-source Power Supply for Datacenters with Arbitrary Demand//Proceedings of the International Conference on Distributed Computing Systems (ICDCS). Philadelphia, USA, July, 2013

[61] Mahadevan P, Sharma P, Banerjee S, Ranganathan P. A power benchmarking framework for network devices//Proceedings of the International Federation for Information Processing (IFIP) Networking'09. Aachen, Germany, 2009: 795-808

[62] M. Gupta, S. Grover, and S. Singh. A feasibility study for power management in LAN switches//Proceedings of the IEEE International Conference on Network Protocols (ICNP). Berlin, Germany, 2004: 361-371

[63] S. Nedevschi, L. Popa, G. Iannaccone, etc. Reducing network energy consumption via sleeping and rate-adaptation//Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI), San Francisco, CA, 2003: 323-336

[64] G. Ananthanarayanan and R. Katz. Greening the switch//Proceedings of the conference on Power aware computing and systems (HotPower). Berkeley, CA, 2008: 7-7

[65] M. Gupta and S. Singh. Using low-power modes for energy conservation in Ethernet LANs//Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). Anchorage, AK, 2007: 2451-2455

[66] C. Gunaratne1, K. Christensen1, B. Nordman. Managing energy consumption costs in desktop PCs and LAN switches with proxying, split TCP connections, and scaling of link speed. International Journal of Network Management, 2005, 15(5): 297-310

[67] C. Gunaratne, K. Christensen, B. Nordman, etc. Reducing the energy consumption of ethernet with adaptive link rate (ALR). IEEE Transactions on Computers, 2008, 57: 448–461.

[68] B. Heller, S. Seetharaman, P. Mahadevan, etc. ElasticTree: saving energy in data center networks//Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI). San Jose, CA, 2010

[69] Y. Shang, D. Li, M. Xu. Energy-aware routing in data center network //Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM) Workshop on Green Networking. New Delhi, India, 2010: 1-8

[70] K. Greene, Tr10: software-defined networking, Cambridge, USA: MIT. Technology Review, 2009

Li Dan, born in 1981, Ph.D., Associate professor. Master supervisor. His research interests include Internet architecture and protocol design, data center network, software defined networking. E-mail: tolidan@tsinghua.edu.cn.

Chen Gui-Hai, born in 1963, Ph.D., professor, Ph.D. supervisor. His research interests include distributed system protocol, data center network. E-mail:gchen@cs.sjtu.edu.cn.

Ren Feng-Yuan, born in 1970, Ph.D., professor, Ph.D. supervisor. His research interests include congestion control, data center network. E-mail: renfy@tsinghua.edu.cn.

Jiang Chang-Lin, born in 1980, Ph.D. candidate. His current research interests include Internet architecture, data center network, network routing. E-mail: jiangchanglin@csnet1.cs.tsinghua.edu.cn.

XU Ming-Wei, born in 1971, Ph.D., professor, Ph.D. supervisor. His research interests include computer network architecture, Internet protocol and routing, high-speed router architecture and green networking. E-mail: xmw@cernet.edu.cn

Background

Data center networks are employed to not only host on-line services but also execute back-end computations. As the key infrastructure of cloud computing and innovation platform for next-generation networking, data center network has unique characteristics compared to the Internet, and has been a hot research topic in both academia and industry in recent years. This paper conducts a survey study on the research progress of data center network, including data center network topology design, transport protocol, wireless communication, enhanced Ethernet, virtualization, software defined networking, etc. Then the future research trends in this area are discussed.

The work is supported by the National Natural Science Foundation of China under Grant No.61170291 and the National Basic Research Program of China (973 program) under Grant 2014CB347800. This project aims to make advances to the architecture of the data center networks. This paper summarizes the research progress of the data center networks in recent years.