

可重构数据中心网络研究综述



张登科¹ 王兴伟¹ 何强² 曾荣飞³ 易波¹

1 东北大学计算机科学与工程学院 沈阳 110819

2 东北大学医学与生物信息工程学院 沈阳 110819

3 东北大学软件学院 沈阳 110819

(zhangdk@mail.neu.edu.cn)

摘要 超大规模数据中心成为数字社会的关键基础设施。用户端应用的激增使得数据中心网络(Data Center Networks, DCNs)的东西向流量呈指数级增长,同时端应用的多样化也导致了严重的流量倾斜问题。此外,后摩尔时代的到来和 Dennard 缩放的失效使得数据中心网络设备容量的增速趋缓。数据中心网络面临用户激增、流量倾斜和 CMOS 性能墙等多重压力。为解决上述问题,可重构数据中心网络(Reconfigurable Data Center Networks, RDCNs)应运而生。文中首先介绍 RDCNs 的 5 个研究驱动力,重点概述了两类物理层使能技术;其次,详细阐述 RDCNs 研究分类和链路重构、层重构以及拓扑重构这三大设计空间关键技术的研究现状;然后,简述 RDCNs 理论的研究进展;最后,展望未来研究方向并总结全文。

关键词: 可重构数据中心网络;使能技术;链路重构;层重构;拓扑重构

中图法分类号 TP393

State-of-the-art Survey on Reconfigurable Data Center Networks

ZHANG Deng-ke¹, WANG Xing-wei¹, HE Qiang², ZENG Rong-fei³ and YI bo¹

1 College of Computer Science and Engineering, Northeastern University, Shenyang 110819, China

2 College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110819, China

3 College of Software, Northeastern University, Shenyang 110819, China

Abstract Hyper-scale data centers have become the key infrastructure in digital society. The prosperity of user applications has caused exponential growth of east-west traffic in the Data Center Networks (DCNs), and simultaneously, the diversification of user applications has led to serious traffic skew problems. On the other hand, the growth of network equipment capacity becomes slow in the post-Moore era, with the breakdown of Dennard scaling. Reconfigurable Data Center Networks (RDCNs) emerge when data centers are facing the pressures from the surge of users and the skew traffic as well as the CMOS performance wall. Firstly, this paper presents five motivations of RDCNs. Then, two types of enabling physical technologies for RDCNs are summarized and the research status of RDCNs is elaborated in detail in terms of the three-design space, i. e., link-level reconfiguration, layer-level reconfiguration and topology-level reconfiguration. In addition, theoretical research results of RDCNs are introduced. Finally, the future work is presented and the whole paper is concluded.

Keywords Reconfigurable data center networks, Enabling technologies, Link-level reconfiguration, Layer-level reconfiguration, Topology-level reconfiguration

1 引言

众所周知,21 世纪被称为数据时代,数据已经成为重要的生产要素,可与劳动力、资本、土地等相提并论。数据中心(Data Centers, DCs)是我国七大新基建的组成之一,成为支撑社会运行的关键基础设施^[1]。DCs 灵活地提供计算资源,促进了互联网经济的可持续发展。为应对更多的用户和网络应用,国内外大型互联网企业均在持续推进超大型数据中心

的建设,建设规模已经上升至数百万台服务器^[2-3]。超大型数据中心具有较强的成本优势、规模效应和可扩展性,已经成为国家竞争力的新指标^[4]。研究报告预计,到 2021 年,全球范围的超大规模数据中心数量将再增加 50%^[4-5]。

随着 5G/B5G、边缘计算和物联网等新技术落地和规模化部署,用户端应用激增,将进一步促进超大规模数据中心的发展,应用的增长使得数据中心内部流量以 25% 的年增长率急剧增长^[4-5]。可预见的还有下一代通信密集型工作负载会

到稿日期:2020-11-04 返修日期:2021-01-27 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家自然科学基金(61872073,61572123)

This work was supported by the National Natural Science Foundation of China (61872073,61572123).

通信作者:王兴伟(wangxw@mail.neu.edu.cn)

将更多流量导向 DCNs,这类负载包括使用硬件加速器(如智能网卡和 TPU^[6])的分布式机器学习和人工智能等。DCs 交换机的端口速率已经高达 100Gbps。然而,最高性能的 GPU 处理器已经能够处理 2.4Tbps 的网络流量^[7]。随着后摩尔时代的到来,硅基 CMOS 网络交换机处理能力的增速逐年降低^[8-9]。数据中心流量增长和交换机处理能力滞滞之间的差距将进一步拉大,如图 1 所示,不断增加的间隙将导致高功耗、高成本和高延迟。

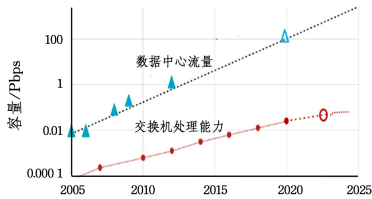


图 1 数据中心流量和交换机处理能力^[10]

Fig. 1 DCNs traffic and switch processing capacity^[10]

数据中心网络中的流量和传统因特网流量不同,其以东西向流量为主要组成部分,其中“大带宽”应用和“低延迟”应用并存,“大象流”和“老鼠流”并存。特定大规模机器学习应用产生的东西向流量占比甚至高达 70%^[11]。南北向和东西向流量混合等问题对网络运行和资源利用带来了巨大的挑战。用户端应用的多样性也为 DCNs 带来严重的流量倾斜问题^[12]。

目前,DCNs 大都使用固定线缆连接服务器,固定拓扑网络通常可以满足一种或几种通信模式的需求,但是在其他通信模式下表现不佳^[13]。网络拓扑对称性和用户流量倾斜性的天然矛盾使得资源利用率低下,DCNs 的使用率很少能超过 50%^[14]。

前述 CMOS 性能墙、流量倾斜问题和网络资源利用率低下激发了对一种新 DCNs 架构的需求。为解决固定拓扑灵活性不足造成的前述问题,可重构数据中心网络应运而生,其本质是通过重新配置数据中心拓扑结构进行网络优化。RDCNs 的研究由新兴物理层技术驱动,由软件定义网络/可编程网络提供的数据平面/控制平面分离思想支撑。RDCNs 的开创性探索工作是微软研究团队于 2009 年提出的“飞路”(Flyways)^[15]方案。“飞路”方案在运行时重新配置拓扑,根据流量需求在固定有线网络拓扑中增加无线 Ad hoc 链路。受这一思想启发,涌现出了大量研究成果^[16-19]。随着底层使能技术的发展,可以用作“飞路”的物理层使能技术从最初的毫米波无线技术(60GHz)发展到了 2D/3D MEMS(2D/3D Micro-Electro Mechanical Systems)等光交换技术、可见光以及最新硅光和自动扇出技术(本文第 2.3 节将详细介绍使能技术)。使能技术的发展使得 RDCNs 的研究内容逐渐扩大:从最初的按需提供基于链路的重构以实现热点消除,扩展到了按需搬运网络容量的层级重构以及关注自身重构力的全拓扑重构研究^[20-21]。RDCNs 的重构研究已经触及数据中心网络几乎所有网元,包括链路、网卡、虚拟交换机、柜顶交换机(Top of Rack, ToR)和聚合交换机等。最近有学者将人工智能引入这一领域并取得了进展^[18,22]。

RDCNs 目前还处于实证研究阶段,大量的研究工作由实际需求驱动,还未形成理论体系。因此,本文创造性地引入“设计空间”作为梳理主线,对本领域现有研究工作进行全景式调研和全面综述,以挖掘并扩大 RDCNs 研究范畴。据笔者所了解,这是首篇针对 RDCNs 的全面性综述文章。

本文第 2 节具体介绍 RDCNs 研究的驱动力;第 3 节和第 4 节分别详细阐述 RDCNs 研究分类和 3 个设计空间的研究进展;第 5 节简述 RDCNs 涉及的理论和算法;最后总结全文并展望未来研究方向。

2 RDCNs 研究驱动力

DCNs 架构对称性和用户流量倾斜性之间的天然矛盾是可重构网络研究成为热点的原动力。DCs 向超大规模发展引发社会对其成本和能耗控制的关切,本节称之为“规模税”,这形成了 RDCNs 的经济驱动力。新兴底层物理通信技术的出现和成熟使得可重构有了物质基础驱动力。软件定义网络(Software Defined Networking, SDN)/可编程网络的控制与数据解耦思想给了可重构网络的软件驱动力。大型数据中心拥有者一般是单一的,相对独立的管理环境允许网络研究创新,这构成了 RDCNs 的环境驱动力。

2.1 拓扑对称性和流量倾斜的天然矛盾

计算机互连网络的拓扑结构是决定网络性能的重要因素,而网络结构的高对称性和正则性是传统网络设计者所追求的目标之一。随着 DCs 规模的不断增大,遵循对称性的网络设计变得更加具有挑战性。为了满足用户严格的延迟需求,当前的 DCs 通常追求全对分带宽,并在超额订购情形下运行。主流 DCNs 采用折叠式 Clos 对称性拓扑结构(也可称为多根树、胖树或者 k 次 n 元树),具有良好的扩展性和经济性。其他常见的具有对称性结构的拓扑还包括 VL2^[23]和 BCUBE^[24]等。分组交换是现在互联网采用的标准数据交换技术,其灵活性造就了当前互联网的繁荣。然而,在 DCNs 场景中,链路固化导致网络无法应对流量倾斜,应用多样性会引入更多倾斜流量,容易导致部分链路拥塞,而其他链路则未得到充分利用。网络对称性和链路固化导致的流量倾斜成为根本矛盾。RDCNs 是向解决流量倾斜问题迈出的重要一步。

2.2 规模税

工业界 DCNs 使用折叠式 Clos 的分层拓扑结构,该拓扑结构大都使用低基数(也可称为低端口数)电气交换机。这样它们只需增加更多的层级就可以任意扩展,并且在每一层级均可进行超额订购以降低成本。

然而,具有良好可扩展性的层次结构却为 DCs 的建设增加了规模税,即随着网络规模扩大,网络功率和延迟随之增加。根据公开的网络设备功率数据^[25],仅收发器和光纤直接连接两个节点就消耗约 50 W/Tbps,交换机和收发器总电量消耗约为 480 W/Tbps。根据文献^[25-26]提供的数据中心历史流量数据推测,DCs 实现 1 Tbps 的等分带宽需要的总功率约为 480 kW;若实现 100 Pbps 的等分带宽,其所需总功率将是令人望而却步的 480 MW。单位对分带宽网络功率可以通过增减层次结构来按比例缩放,如表 1 所列。

表 1 DCNs 功率
Table 1 DCNs power

层次 (Clos)	基数	等分速率 /bps	网络 功率
3-level	32	1 T	450 kW
4-level	64	100 P	450 MW

随着网络层级的增加,网络成本开销和延迟也会受到类似的影响。能否用更低成本实现数据中心同等规模的增长,成为工业界和学术界共同关注的一个问题。规模税的存在促进了 RDCNs 的研究进展。

2.3 新兴使能技术

传统上,DCNs 基于电子元件,针对可扩展性、构建成本、容量、布线复杂性、延迟或健壮性等静态特性进行优化^[27]。RDCNs 研究则不同,需要使能技术提供低重构时间的敏捷性和高扇出能力,以支持大量不同的节点在任何给定时间或者随时建立连接。目前,允许动态改变两对 TOR 之间容量的最常用的底层驱动技术是无线通信技术和光通信技术,还有少数其他底层技术。

2.3.1 无线通信技术

解决流量热点问题的方法是实现一种“灵活”的网络,在易受影响的节点之间按需建立链路。无线通信被认为是一种灵活的方案,可以帮助 RDCNs 应对非确定性不均衡流量,缓解拥塞热点。无线通信技术可以通过增加机架间的无线链路来扩展现有固定拓扑的有线基础设施,或者通过纯无线网络完全取代有线基础设施来实现更灵活的 RDCNs 设计。在后者中,无线通信链路同时用于执行机架内和机架间通信。

推动 RDCNs 研究的 3 种无线通信技术分别是 60GHz 无线、自由空间光通信(Free-Space Optics,FSO)和可见光通信(Visible Light Communication,VLC)。

(1)60GHz 无线技术,又称为毫米波无线技术。该频段属于 ISM 频段,可以在全球范围内使用,并且无需得到许可。该频段包含超过 802.11b/g(2.4 GHz)80 倍的带宽,并支持数 Gbps 的数据传输速率。60 GHz 可以支持短距离(1~10 m)、高带宽无线链路在一个未经许可的频段运行。其通过在机架上安装无线天线,可在一定程度上消除网络瓶颈处的链路拥塞问题;使用天线阵列克服机械转向延迟,以较低的扇出和较大的干扰换取 $O(10)ns$ 范围内的较小延迟,有关讨论见文献[28]。也有学者探索纯无线数据中心网络,有望解决热点和布线复杂性问题^[29-30]。

(2)自由空间光学。FSO 是另一种新兴技术,通常依赖于激光器,通过“自由空间”实现节点(如机架顶部交换机)之间的大容量通信。一般来说,FSO 能够在远距离上提供比 60GHz 通信更低的干扰和更高的带宽。使用 FSO 通信链路,本质上是使用激光束在空中无线传输数据。尽管室外无线光通信链路在恶劣的天气条件下运行具有挑战性,但在温度控制的 DCs 环境中使用时,在信息处理能力方面具有巨大的尚未开发的潜力。应用 FSO 的网络为每个服务器机架配备大量的光通信设备,每个设备都可以被实时地“引导”到与不同机架上的另一个光通信设备进行无线通信。FSO 设备的转向能力使动态网络能够适应当前的网络流量。与无线链路类

似,FSO 链路的机械转向(垂直和旋转运动)或可切换反射镜的使用,导致了相对较高的切换时间(例如 FireFly^[31]中的切换时间大约为 20 ms)。Projector^[15]中使用改进的数字微镜器件(Digital Micromirror Device,DMD)支持 FSO,延迟降低至 7~12 s。这种基于 DMD 的解决方案还能够增加扇出,比基于 MEMS 的解决方案更易扩展。

(3)可见光通信。在无线频谱资源逐渐匮乏的困境中,一种新型的通信方式——可见光通信出现在人们的视野中。可见光光谱带宽约为 400 THz,远远大于现有的无线通信频谱。对于 VLC,数据传输通过可见光 LED 或激光二极管的强度调制来实现。它是最简单的数字通信形式,使用开关键控调制方案,其中“1”和“0”分别表示光的存在和不存在。在微芯片的控制下,LED 可以在 1 s 内完成数百万次的开关。2013 年,研究人员已经基于 LED 的 VLC 链路实现了在 10 m 距离达到 10 Gbps 的数据传输。虽然当前很少有数据中心采用 VLC 链路,但已经有研究证明了其应用潜力^[32]。

上述 3 类无线技术对于 RDCNs 仅起到有限改善网络性能的作用,即增加了数据中心热点之间的网络带宽。从数据中心的机架层看,这类架构并没有从本质上改变数据中心物理网络拓扑的固定结构。RDCNs 研究人员继而引入了光交换技术。

2.3.2 光交换技术

光域中的高带宽交换成为当前 RDCNs 的热门解决方案,其将光路交换机(Optical Circuit Switches,OCSs)集成到网络结构中,实现重新配置网络拓扑以改变和动态地适应预测的流量,进而实现网络拓扑结构的可重构性。OCSs 可以被理解为“可直接对光束操作的 0 层交换机,并且不对任何分组进行解码”^[33]。OCSs 的核心光交叉部件是基于 MEMS 实现的。MEMS 通常可定义为使用微制造技术(即沉积材料层通过光刻和蚀刻形成图案以生产所需器件)制造的微型电子机械器件(特别是具有运动结构的器件)。MEMS 得益于其在单个芯片上集成电气、机械和光学元件的独特能力。这类系统工作在微观尺度,与宏尺度系统相比具有不同的特性(开关速度、可靠性和可扩展性)和不同的使用方法。MEMS 光开关已用于长距离传输大带宽数据的通信网络中。根据光信号在开关内部的传播,基于 MEMS 的开关可分为自由空间开关和波导开关两大类。在自由空间光开关中,光信号在自由空间中传播,并沿着其路径被引导至所需的输出光纤。这是通过使用微小的可移动微镜实现的,微镜直径约为数百微米。另一方面,在波道光开关中,光信号在波导中传播而不逃逸到自由空间。波导切换是通过将波导或耦合器放置在特定位置来实现的。MEMS 的工作原理一直是众所周知的。然而通过广泛的研究,直到最近几年才取得了很大的进展。

基于 3D MEMS 的交叉器能够提供可扩展性(几百个输入/输出端口)、可接受的切换时间(几毫秒或几十毫秒)、低插入损耗和低串扰。典型的 3D MEMS 架构采用两个微镜阵列,每个微镜阵列对准准直输入或输出光纤阵列,只需要对 N 个端口使用 $2N$ 个镜像,远远少于传统 2D MEMS 架构。该开关由 1 个输入光纤阵列、1 个输入透镜阵列、三维空间中

2个平行的 MEMS 反射镜阵列、1个输出透镜阵列和1个输出光纤阵列组成,如图 2 所示。每个输入光纤将光束引导至输入阵列上的反射镜,而输入反射镜将光束引导至输出反射镜,输出反射镜又将光束引导至输出光纤。由于采用对称设计,输入镜和输出镜都要求具有相同的偏转能力。3D MEMS 产品已经上市,其切换时间为 1~25 ms^[34-35]。比如,美国 Calient 公司已经量产 3D-MEMS 全光交换机,其最高密度型号拥有 640×640 阵列端口,能在 1260~1630 nm 的波长范围内实现小于 3.5 dB(典型)的插入损耗,链路总延迟增加少于 30 ns,功耗为 90 W。

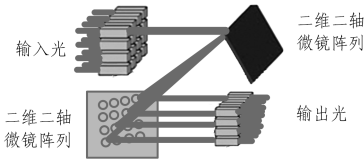


图 2 3D MEMS 转换器件结构
Fig. 2 Structure of 3D MEMS transformer

理想的 RDCNs 解决方案是在大范围内以低成本部署快速交换机。然而,这样的 MEMS 交换机并不存在。实现光电路交换的 3D MEMS 交换机硬件配置时延在 10~100 ms,控制层优化调度的时延在 100 ms 到 1 s 之间,因此光电路交换仅能部署在网络的核心层^[36]。

因此,文献[37-38]提出了结合慢速和快速光开关的混合解决方案,部分波长通过慢速 OXC 路由,而其他波长通过快速 OXC 路由。通常会考虑将 3D MEMS 和 SOA 的开关组合使用。这允许集成不同开关技术的优势,即慢速 3D MEMS 开关的高可扩展性和基于 SOA 的小型开关的高开关速度。电路和突发流量通过慢速 MEMS 开关路由,而短突发和分组通过基于 SOA 的快速光开关路由。基于 SOA 的交换机^[39-40]可以扩展到 32×32 或 64×64 的端口数量。

MEMS 技术仍然是现在和将来的全光通信网络的关键交换技术之一,可以预期研究将集中在光开关的优化,可能研发出质量有显著改进的新技术,比如基于绝热和定向耦合器的波导开关可能成为下一代光开关。与 3D MEMS 相比,基于定向耦合器的波导开关的开关时间快一千倍以上。定向耦合器可以被调谐到特定波长,因此它可以用于波长选择开关(Wavelength Selective Switch,WSS)。使用 WSS 的重构架构如图 3 所示^[36]。最近的研究表明,基于绝热耦合器的光开关既能提供快速的切换时延(小于 1 s),又能提供较大带宽^[20]。

另一个相关的使能技术是波导光栅路由器(Arrayed Wave guide Grating Router,AWGR)。AWGR 利用波长路由特性,实现优于严格无阻塞的多路并行无竞争交换,不仅可以互连任意空闲的输入输出,还允许任意输出端口同时接收来自不同输入端口的信号。AWGR 一般和环回式共享缓存共同使用。在波分复用(Wavelength Division Multiplexing,WDM)通信系统中,AWGR 被用作分离单个波长的多路复用器,或作为多路复用器将它们组合起来。所用耦合器是一种无源光器件,用于光网络中信号的聚合和分配。AWGR 可以支持多个输入输出。

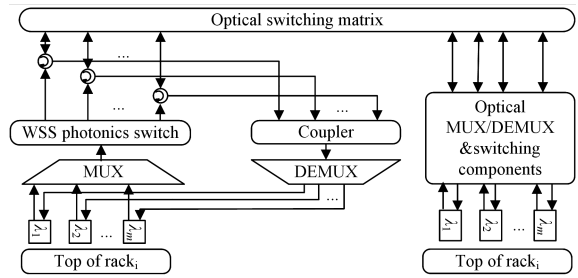


图 3 WSS 重构架构
Fig. 3 Reconfigurable WSS

基于 MEMS、波长选择和 AWGR 的光波长交换技术的对比如表 2 所列。

表 2 光交换使能技术的对比

Table 2 Comparison of OWS enablers

	MEMS	WSS	AWGR
延迟	毫秒级	微秒级	微秒级
灵活性	低	中	高
成本	中	低	高

2.3.3 其他使能技术

其他使能技术主要有光学转子、智能网卡和 WDM 等。光学转子可以更高速地完成光路切换^[21],基于 ASIC 或 FPGA 的智能网卡给主机端网卡带来了接入可重构的可能^[41]。WDM 技术通常在单个光纤上可以使用 40 个或更多波长。例如,Mordia 原型^[42]在 24 个端口下可实现 O(10)s 的重新配置时间。这样的交换机也可以组合和扩展,达到 O(100k)端口。具有纳秒量级重配置时间的技术也开始出现^[11],可参考文献[43]获得进一步的细节。

2.4 控制和数据解耦

RDCNs 研究的第 4 个驱动力来自软件定义网络/可编程网络的控制和数据解耦的思想。控制平面与数据平面解耦使得可重构网络具备了 SDN 平台,获得了面向网络遥测(In Network Telemetry,INT)的可编程交换架构的支撑,重构所必需的实时修正控制和反馈成为可能。

OpenFlow^[44]是 SDN 中采用最多的协议标准,它提供一个南向 API。SDN 将网络的控制平面(配置、路由、监控和其他网络控制功能)与数据平面(分组转发、分组字段修改和其他分组处理功能)分离,OpenFlow 使用远程 SDN 控制器来添加、修改和删除流规则,并从交换机检索统计信息,例如每个流携带的数据包数目、字节数目等。因此,启用 OpenFlow 的交换机不需要实现复杂的路由和网络管理协议;控制器实现管理功能,并用实现所需功能的流规则配置交换机中的数据平面硬件。

数据平面硬件的一个趋势是白盒交换机,也称为裸交换机,由不采用高端交换机特性的商品交换机组成。白盒交换机进一步降低了成本,同时提供开放和可扩展的接口,允许配置和控制交换机的数据平面硬件。白盒交换机通常使用开源操作系统,如 Linux 白盒交换机支持 OpenFlow 协议。与传统交换机中的数据平面硬件一样,白盒交换机中的数据平面硬件采用三态内容寻址存储器(Ternary Content Addressable

Memory, TCAM), 交换机允许控制器使用 OpenFlow 配置 TCAM, 以便以线路速率处理数据包。交换机在 DRAM 中存储新的 OpenFlow 规则。一旦传入的数据包与流规则匹配, 交换机将该规则缓存在 TCAM 中, 这意味着来自同一流的后续数据包将以线速率转发。因此, 白盒交换机提供了在 ToR 交换机中存储整个 RDCNs 绑定表所必需的内存层次结构。

可编程语言 P4^[45] 提供了一个框架, 能够更灵活地对交换机进行编程以处理数据包, 其匹配/动作功能比 OpenFlow 更丰富。在 INT 环境中, P4 能够利用网络设备(如交换机和主机端网卡)的计算能力的最新进展来增强监控能力。最新的带内 INT 支持在数据平面上收集和报告网络状态, 提供了对网络元件性能和负载的粒度监控。INT 使流量源能够在数据包中嵌入遥测指令, 避免单独探测或基于管理的监控。INT 接收器节点通过检索由不同 INT 信息源追加的 INT 元数据指令来跟踪和收集度量。通过相对较少的基于 SNMP 的轮询或通过 RMON^[46] 的事件检测, 或通过使用探测数据包进行性能测量, INT 能够对用户数据包进行及时、精确地测量。INT 的细粒度网管信息将激励更灵活和实时反馈更好的 RDCNs 研究取得进展。

2.5 创新环境驱动力

当代公有云大都为单一实体拥有, 比如国内的阿里云^[47] 和华为云^[48], 国外的微软^[49]、Facebook^[50] 和谷歌^[51] 等, 这便于运营商在自己的数据中心内使用定制性技术, 以提升网络性能。相对简单和相对独立的管理环境也为网络技术变革提供了有利条件。工业界认为网络架构已经成为整体系统不可缺少的一部分, 承担着将单一资源互连成资源池的重任, 网络对数据中心性能有决定性影响。

综上, RDCNs 研究的五大驱动力如图 4 所示。

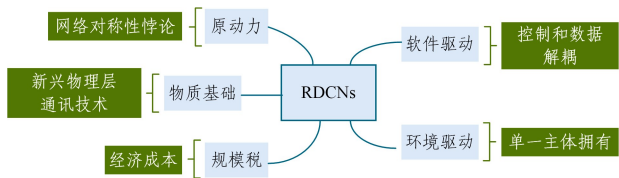


图 4 RDCNs 研究的驱动力
Fig. 4 Driving force of RDCNs

对称性拓扑和倾斜性流量的天然矛盾促使了流量感知的早期 RDCNs 研究, 底层通信技术、经济成本和控制数据解耦的思想则为 RDCNs 研究提供了多方面的基础, 超大规模数

据中心的单一拥有主体使得 RDCNs 的实实践更便捷。在上述研究动机的共同作用下, 近年来无论在学术界还是在工业界, 无论在物理层使能技术还是在网络设计优化方面, 面向 RDCNs 的技术研究都取得了快速发展。

3 RDCNs 分类

随着应用数量的增加和数据中心规模的快速增长, 数据中心服务器之间的互联模式迫切需要与之对应的快速响应式的动态重构。为了解决这个问题, RDCNs 利用新型底层通信技术开发调度算法来管理何时通过哪些交换机连接服务器。解决这一挑战将有助于数据中心资源更有效地用于数据分析和其他计算密集型任务。RDCNs 的出现已经影响到现代数据中心的实际网络设计, 促进了具有高连接性和低交换成本的下一代光交换机的开发。RDCNs 领域的实证研究成果表明, 可重构技术具有巨大潜力。两个常用的 RDCNs 模型是分层模型和抽象模型: 分层模型如图 5(a) 所示, 模型中可重构链路实现 RDCNs 的可重构设计, 后续基于链路和基于层的 RDCNs 设计方案(详见第 4.1 节和第 4.2 节)一般基于该模型。抽象模型是一种理想模型(如图 5(b) 所示), 模型中网络抽象成一个“大交换机”, 重构由其内部设备或链路实现, 后续基于拓扑的 RDCNs 设计多采用抽象模型(详见第 4.3 节)。

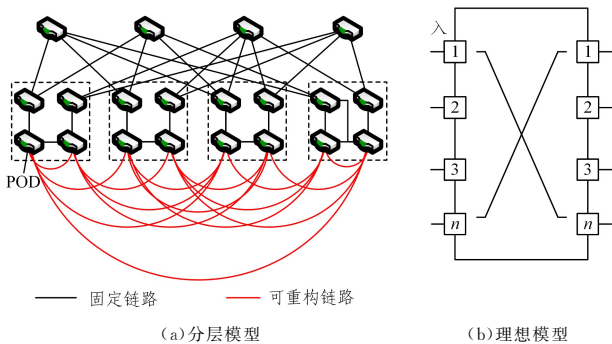


图 5 RDCNs 模型
Fig. 5 Model of RDCNs

本节使用 3 个维度对 RDCNs 研究进行分类。一个维度是依据流量矩阵(Traffic Matrixes, TMs), 如图 6 所示。流量无关 RDCNs 研究, 根据固定的调度策略周期性地改变拓扑结构, 该策略与流量需求无关。流量感知 RDCNs 研究依据流量模式进行拓扑结构的按需调整。

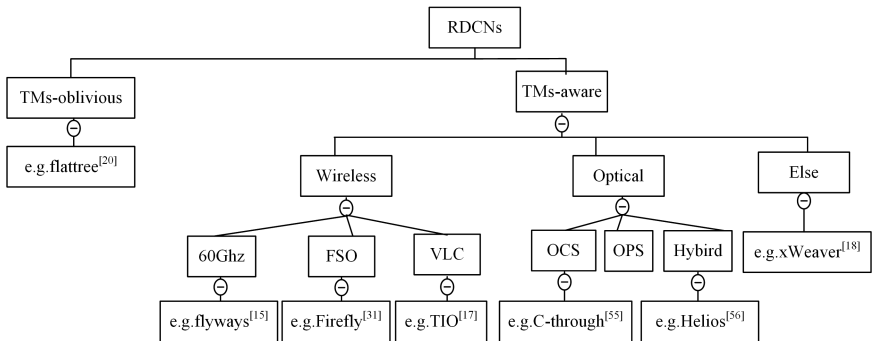
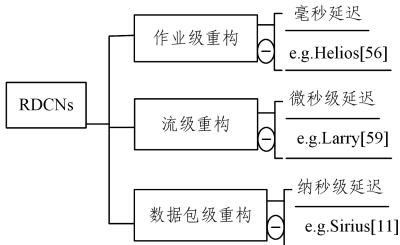
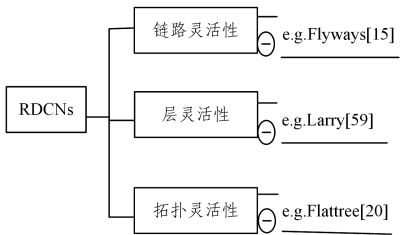


图 6 基于 TMs 的 RDCNs 分类
Fig. 6 Classification of RDCNs based on TMs

另一种是依据重构延迟指标。秒级延迟的 RDCNs 设计适用于作业级应用,该类应用的软件堆栈延迟将掩盖 RDCNs 的拓扑切换延迟。毫秒级延迟的 RDCNs 设计用于流级重构的场景,而纳秒级延迟的 RDCNs 设计则可用于用户数据包级的重构场景,如图 7(a)所示。



(a) 基于延迟的 RDCNs 分类



(b) 基于网元灵活性的 RDCNs 分类

图 7 流量矩阵无关的 RDCNs 分类

Fig. 7 TMs-independent classification of RDCNs

可重构需要灵活性的网络元件,基于这个考虑的研究包括链路灵活性、层灵活性和拓扑灵活性等类别,如图 7(b)所示。不同类别的 RDCNs 研究工作共同组成了 RDCNs 相对完整的设计空间,本文第 4 节将从设计空间这一新颖角度对 RDCNs 的研究进展进行梳理总结。

4 RDCNs 设计空间

随着新使能技术和网络控制范式的应用,RDCNs 的设计空间越来越丰富,本节根据图 7(b)所示的 RDCNs 分类方案梳理研究进展,以帮助研究人员识别并创造新的研究机遇。

4.1 基于链路的重构

新兴底层通信技术为 RDCNs 体系提供了连接灵活性,可以按需使用任意一对转发器形成链路。这一领域的研究挑战需要高度复杂的硬件设计和软件控制器,以确保正确及时地进行收发器之间的链路重构。这类 RDCNs 的代表工作有使用 60GHz 无线的“飞路”^[15]、使用 FSO 的 Projector^[16]、使用 VLC 的 TIO^[17]以及引入人工智能算法的 xWeaver^[18]。

“飞路”技术是美国华盛顿大学的 Daniel 等于 2011 年在 ACM Sigcomm 会议上发布的研究成果,它开创式地通过 60Ghz 有向无线技术实现了数据中心网络无线通信设计方案。这一设计就像在 TOR 交换机之间实现了多条的“飞行链路”,实现部分连接形态的可重构拓扑,如图 8 所示。关于重构时间,文献[52]提到天线可以在 $O(100)\mu s$ 内操纵,但其仿真中忽略了此类开销。

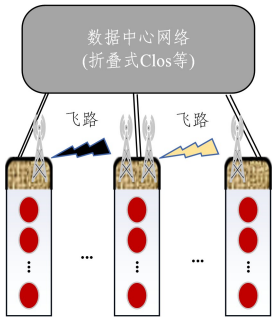


图 8 Flyways 架构^[15]

Fig. 8 Flyways architecture^[15]

受“飞路”思想的启发,大量研究者致力于研究 60GHz 毫米波无线技术在 RDCNs 的应用^[15-19]。这些工作大都通过改进 60GHz 的布局实现。还有研究人员采用了镜子,例如挂在天花板上球形镜面^[19],有效地延长无线链路的“视线”,并减少了干扰。“飞路”可以提供更多重构模式,同时提高无线资源的使用效率。由于存在干扰,60GHz 的模式会对邻近无线模块产生影响。这增加了 RDCNs 中路由和网络管理的复杂性,降低了网络吞吐量。

Projector^[16]和 Firefly^[31]使用 FSO 技术来解决衰减和传播损耗问题,FSO 自身的高载波频率和相对较大的检测器面积提供了避免多径衰减的基础能力。相对毫米波无线技术,FSO 链路的设计更简单。上述两种设计的架构如图 9 和图 10 所示。FSO 在 RDCNs 中的研究工作还包括文献[52-54],这些研究利用 FSO 的特性实现了不同的重构方式,甚至能够支持组播等复杂的传输模式。

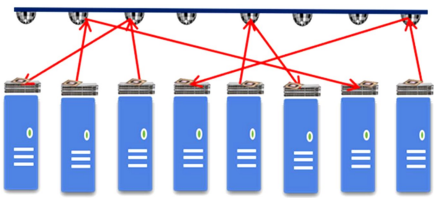


图 9 Projector 架构^[16]

Fig. 9 Projector architecture^[16]

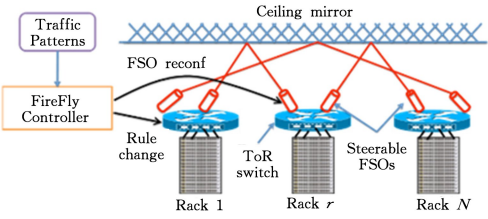


图 10 Firefly 架构^[31]

Fig. 10 Firefly architecture^[31]

文献[18]提出了 xWeaver,首先使用人工智能方法实现了可重构链路的选择,其架构如图 11 所示。它采用神经网络来提供流量驱动的拓扑自适应能力,通过深度学习找到最佳的拓扑配置以支持动态流量需求。xWeaver 能对不同的性能指标和网络体系结构进行拓扑优化,通过设计适当结构的神经网络,自动从数据跟踪中导出关键流量模式,并学习流量模

式和特定于目标数据中心的拓扑配置之间的基础映射。离线训练后，xWeaver 可以在线生成优化(或接近最佳)的拓扑配置。类似的研究工作还包括 Deepconf^[22]。

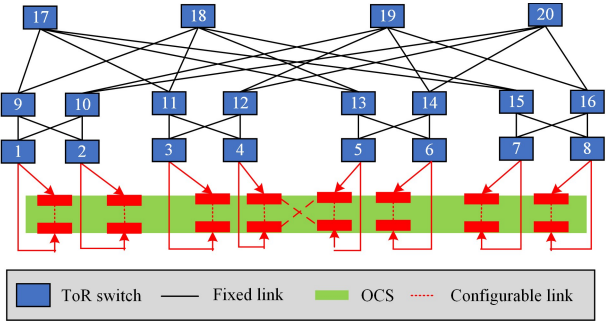


图 11 xWeaver 架构^[18]
Fig. 11 xWeaver architecture^[18]

文献[17]采用 VLC 技术实现了一种名为 TIO 的 RDCNs 架构，它是一种将无线 Jellyfish 和有线折叠式 Clos 拓扑无缝集成的混合 RDCNs 设计，采用 VLC 链路来构建一个无线的稳定网络。与上述其他混合设计不同的是，它放弃了光学设备上复杂的控制机制，支持即插即用，且容易部署。TIO 在机架上方空间分配 VLC 链路，如图 12 所示。因此，它不需要反射镜或其他装置来中继光信号，且易于部署和维护。

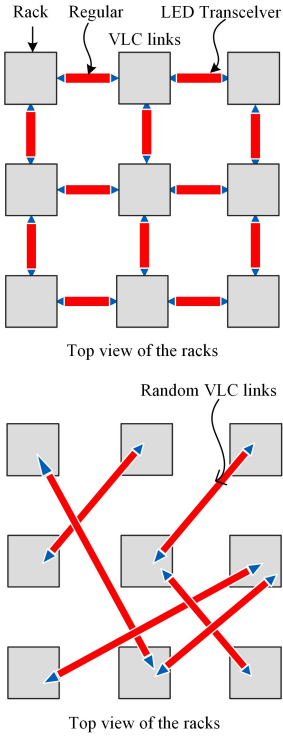


图 12 TIO 架构^[17]
Fig. 12 TIO architecture^[17]

4.2 基于层的重构

静态网络拓扑结构无法适应应用空间局部性导致的倾斜流量问题。为了应对流量倾斜，基于层的 RDCNs 重构研究着重在层次化固定拓扑中添加灵活配置层来实现拓扑的可重构。灵活配置层的研究有两条技术路线。第一条使用 OCSs 引入电路交换层，其原理如图 13 所示。此类研究的典型代表

是 C-through^[55] 和 Helios^[56]，最新的典型研究是 Sirius^[11]。

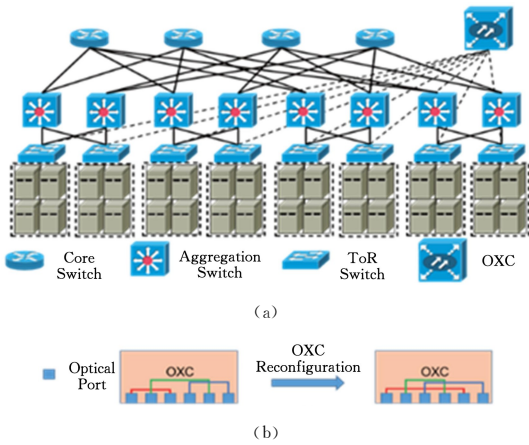


图 13 基于 OCSs 的 RDCNs 架构^[55]
Fig. 13 RDCNs architecture based on OCSs^[55]

C-through 和 Helios 均基于统计的业务需求矩阵进行拓扑重构，并使用延迟为几毫秒的 OCSs。如果没有可用的应用级视图，则获得的流量需求不精确且不及时，C-through 和 Helios 的电路利用率和应用性能可能会很差^[57]，因此其不适用于延迟敏感的 DCNs 应用。Sirius 则采用了一个单一高基数的交换机的抽象，可以直接连接数据中心网络中的数千个节点(包括机架或服务器)，同时实现纳秒级粒度的重新配置。它采用了新型可调谐激光器，实现了切换延迟小于 10ns 的 AWGR。Sirius 是引入电路交换最新技术的代表性工作。

第二条路线是添加可重构层，可重构层使用智能配线板(Rewire Patch Panel 或者 FPGA 卡)或者专有转子装置等，在 TOR、接入交换层、聚合层和 POD 之间进行灵活的连接配置，其原理如图 14 所示。图中数字 1—4 分别代表 RDCNs 可重构层的 4 个层间可选位置。此类研究的典型代表包括 Flexfly^[58]，Larry^[59]，MR^[60] 和 RDC^[61] 等，更多的工作研究了重构层布线评估和其他重构装置^[62]。大多数光电路交换 RDCNs 研究混合利用光电路交换与传统分组交换。光电路交换用于动态互连机架顶交换机(一个子集)，专为大象流所保留。传统分组交换用于老鼠流。不同的可重构网络因所使用的可重构设备的类型的不同而不同。

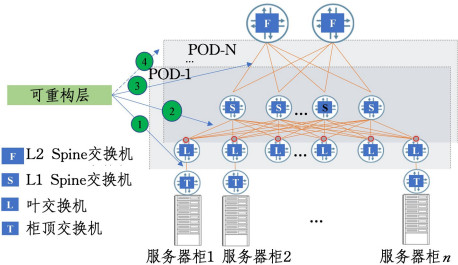


图 14 RDCNs 可重构层
Fig. 14 Reconfigurable layers of RDCNs

文献[56]提出了 Helios 架构。在 Helios 中，部分电分组交换机被光交换机所取代，实现了 POD(一个包含了数百个服务器的集群)之间的互连。文献[63]提出了 SSX(Single-

stage Shuffle Exchange)拓扑的光电混合网络架构,该网络架构解决了 MEMS 交换机较长配置时延带来的光电路利用率低的问题。但是,该网络对于通过光网络的流量存在限制条件,仅有部分业务在通过该网络时能获得更好的效益。

同时,研究人员还提出了多种全光互连的数据中心架构方案,典型的有 Proteus^[64] 和 OSA^[65] 网络架构。与传统基于电分组交换的 DCNs 相比,研究人员对引入光互连技术的光电混合数据中心网络的探索仍处于起步阶段,还未形成成熟的方案。OCSs 的引入本质上提高了分组交换数据中心网络的电路交换能力。基于传统分组交换的树形分层结构的 DCNs,正随着 DCNs 规模的不断增大而面临着越来越严重的效能和成本问题。

近年来,光电混合网络通过引入光互连技术,充分发挥了光网络的高带宽、低成本等优点,不仅提高了网络容量,也减少了传统数据中心网络中所需的电分组交换机的数量,从而降低了网络能耗和设备成本。在光电混合网络架构下,两种网络并行工作,不仅可以有效降低电交换网络的转发压力,也避免了单点失效等问题。但是,MEMS 光交换机较长的开关时延(约 25ms)会成为上述架构的性能瓶颈。此外,TOR 之间 OCSs 的带宽需由预先提供的收发信机决定,一旦网络建立,这些收发信机不能按需重新分配以服务于动态 DC 业务,从而激发研究人员设计更具灵活性的重构层。

微软学者提出的 Larry^[59] 实现了层级 RDCNs 的实用部署。Larry 允许机架根据流量需求动态地重构链路以适应聚合交换机,以满足需求较高的机架,使之能够使用来自其他未得到充分利用的邻接机架的上行链路,其体系结构如图 15 所示。Larry 实现了物理层的重构操作,适用于对延迟敏感的应用程序。

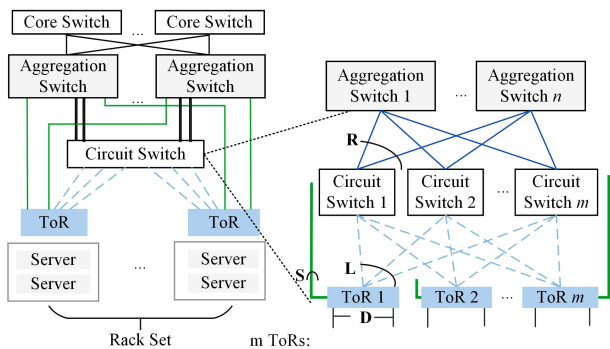


图 15 Larry 体系结构^[59]
Fig. 15 Larry architecture^[59]

层级 RDCNs 利用了各个网元间链路的灵活性,是 RDCNs 最热门的研究领域,从设计空间的角度可以发现其仍然有待探索的区域,比如如何利用端主机和机架间的灵活性等。随着 RDCNs 研究的深入,学者们开始探索利用拓扑灵活性进行全局式网络重构。

4.3 基于拓扑的重构

工业界数据中心网络的事实标准是层次化规则的折叠 Clos 架构,部署简单且易管理,但资源利用率低。学术研究则

追求高效率的随机拓扑,支持网络优化配置和资源的高效利用,但难以部署和管理。RDCNs 拓扑级重构研究旨在两者间取得折衷。这个方向的开创性工作 Flattree^[20],典型工作包括 RotorNet^[21],Flexspander^[66] 和 2-Tier^[12]。

2017 年,文献[20]提出 Flattree,这是首个基于拓扑灵活性的 RDCNs 研究工作,Flattree 基于不同拓扑支持不同的通信模式的事实,认为网络拓扑设计不应该“一刀切”。Flattree 支持网络在规则拓扑和随机拓扑之间切换以提供全对分带宽,降低超额订购率,提高网络吞吐量。

RotorNet 采用集中式重构与分布式重构的正交方法,完全不考虑当前需求,并按照预先计算的周期进行调度,其架构如图 16 所示。一个完美想法是循环所有可能的匹配集,然而这将带来巨大的时间开销。相反,RotorNet 只使用一小组匹配,即源和目的端点之间的连通性在一个匹配周期内得到保证。因此,该设置有效减少了交换机的重新配置时间。在均匀(延迟容忍)业务的情况下,这样的单跳转会使网络的对分带宽饱和。然而,对于倾斜流量需求,依然有闲置连接,网络容量仍未得到充分利用。为此,RotorNet 使用 Valiant 负载均衡^[67] 和多个可重构交换机来执行分布式多跳路由,同时配置对间接流量的缓冲,研究结果证明了该机制的有效性。

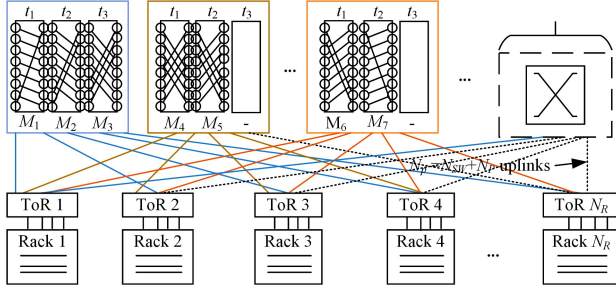


图 16 RotorNet 架构^[21]
Fig. 16 RotorNet architecture^[21]

Flexspander^[66] 是一种建立在扩展器基础上的 RDCNs 网络架构,它能够动态地适应倾斜流量。Flexspander 拓扑结构是通过 OCSs 层将一组吊舱(Pod)相互连接而构建的,如图 17 所示。通过对 OCSs 状态的重新配置,可以实现吊舱之间不同的吊舱级拓扑。每个吊舱包括一组“静态”互连的电分组交换机,使得吊舱内拓扑形成良好的扩展器网络。

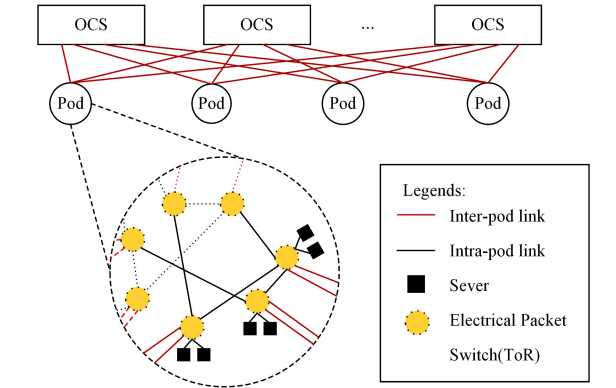


图 17 Flexspander 架构^[66]
Fig. 17 Flexspander architecture^[66]

笔者提出了 2-Tier 架构,首次引入补丁式拓扑子层的概念,为 RDCNs 研究设计提供了新的视角,如图 18 所示。2-Tier 由固定拓扑子层(底层)和移动式拓扑子层(上层)两层组成,可重构性由上层实现。固定拓扑子层可采用规则易部署折叠式 Clos 结构,移动拓扑子层则可灵活组网按需“搬运”网络容量。同时,两层结构中的路由、资源管理和网络同步性等仍然是开放问题,有望结合图谱理论和网络科学理论取得进展。

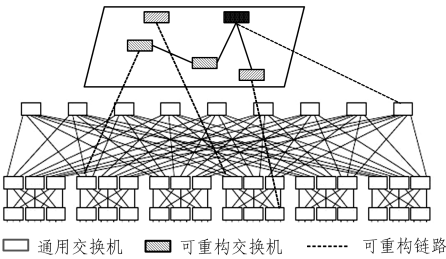


图 18 2-Tier 架构^[12]
Fig. 18 2-Tier architecture^[12]

5 RDCNs 算法理论

从本质上讲,链路级 RDCNs 涉及基本图论问题:如何用额外的边来重构拓扑结构,从而优化期望的目标函数。然而,与经典图增长^[68]不同的是,这些边不能随意地添加。相反,边缘添加受到诸多物理约束,例如,节点只有一个物理接收器或连接距离受限或存在阻塞信号等。在一些共享介质中,不同波长的着色算法也会发挥作用。网络节点连接到可重新配置的交换机(如 OCSs),该交换机在节点之间创建直接的物理连接,这些连接每次只能在两个节点之间形成,需要根据流量传输需求进行匹配。相应地,匹配的链路被添加到拓扑中,从而扩大网络以更好地满足需求。

层级和拓扑级 RDCNs 方案则提供了比链路级方案更大的灵活性;可以在不同拓扑间进行切换,例如提供高吞吐量的随机图拓扑和易部署管理的折叠 Clos 拓扑间切换。随机图理论、扩展图理论、自适应网络理论以及图谱理论等,被引入到 RDCNs 领域中以优化重构层和拓扑。

5.1 链路级 RDCNs 中的匹配算法

链路级 RDCNs 研究的基本思路是依据流量矩阵按需增加链路资源,其理想模型是图增长问题。但与经典图增长不同的是,这些扩充链路受到真实世界的资源约束,不能随意地添加,约束包括节点物理接收器数量、节点间距离或障碍物等。RDCNs 研究者提出了多种匹配算法。

Helios 和 C-through 采用了支持快速匹配的最大匹配 Edmond 算法^[69-70],为一个流量需求快照重新配置一次拓扑,并且唯一的目标是重构链路的单跳吞吐量最大化。Edmond 算法可以给出最优吞吐量的链路。然而受固定拓扑结构约束时,该算法性能会变差。这类方案大多没有从理论角度进行研究,单快照多重配置问题仍有待进一步探索。

Proteus^[71] 及其扩展 OSA^[65] 通过采用加权 B-匹配算

法^[72],为每个节点配置多个可重构链路。为了创建多跳路由连接图,OSA 还利用边缘交换操作。这两项研究工作未能在理论上提供多跳路由性能保证。文献[73-74]为了最小化平均加权路径长度,直接利用“最大 B-匹配算法”产生最佳结果。在动态可重构拓扑和固定拓扑进行多跳联合路由的非隔离路由问题是 NP 难问题^[73-74],仅有少量启发式求解算法^[75]。

为了降低硬件重新配置延迟,Mordia^[41]为单个流量需求快照多次重新配置网络。为此,其将业务需求矩阵转化为带宽分配矩阵,该矩阵表示理想调度中每个可能匹配的边应分配的带宽分数。然后采用计算效率高的 Birkhoff-von-Neumann 算法执行分解^[76],其复杂度为 $O(n^2)$ 。然而,该方案没有考虑最小化重构链路的数量。Solstice^[77]考虑了混合体系结构和时间约束(这使得问题变成 NP 难^[78]问题),但仅提供了贪婪调度启发式求解方式。Eclipse^[79]可以在具有重构延迟的混合交换机架构中提供吞吐量近似算法,但仅适用于沿单跳可重构连接的直接路由。间接多跳路由的情况仍然是开放性问题。

前述匹配算法的一个显著缺点是它们依赖于集中计算和协调。稳定匹配算法支持分布式方式实现,也可提供增量解决方案,即记忆当前拓扑状态。Projector^[16]采用 Gale-Shapley^[80]稳定的匹配算法,其目标是最小化延迟,通过老化过期请求的方式解决饥饿问题。对于吞吐量,该方案未从理论上证明其是否是 NP 难问题。

5.2 层级和拓扑级 RDCNs 中的图论

在纯固定拓扑的情境中,构建重构网络拓扑的设计被理解作为一种重新配置——只是一次拓扑重构而已。由于普通网络交换机的端口数受限,并且通常在整个数据中心网络设备都是相同的,使用随机图模型的 DCNs 设计的 Jellyfish^[81]可获得良好的吞吐量,随机拓扑降低了平均路径长度。Jellyfish 还设计了新型路由算法,将部分流量路由到未充分利用的网络节点,有效地应对了倾斜流量的冲击。这种思想依赖于 D-正则扩展图^[82]。文献[83]提出了如何实现固定数据中心拓扑扩展图的方案。这些随机图和扩展图的见解被应用于 RDCNs。文献[20]提出的 flattree 首次将图论应用到了 RDCNs 的拓扑设计中。扁平树通过部署其研制的专用重构设备,支持网络在折叠式 Clos 拓扑和近似随机图拓扑之间按需进行转换。类似的工作还有 Opera^[84]通过在其周期性重构中保持扩展图来扩展 RotorNet^[21]。尽管 Opera 的重构调度是确定的且不受影响的,但拓扑布局的预先计算仍然是随机的。

需求感知理论也被用于 RDCNs 的研究^[85-88]。文献[86]提出了自调整网络的理论,并将其应用于需求感知网络,优化目标是最小化网络拥塞。文献[73]刻画了 RDCNs 的算法复杂性。其认为,如果网络运营商将可重构的 DCNs 分成可重构的和静态的两部分,并通过将流标记为小流或大流来“独自”使用部分路由,将导致非最优路由。文献[75]提供了几种在 RDCNs 中联合优化拓扑和路由的算法。

RDCNs 的度量也引起了学者们的关注,笔者在 2-Tier 中

引入了网络科学理论中的有效介数作为度量,文献[89]在提出的 CLOVE 中设计了重新布线的度量指标。

RDCNs 领域已有的研究大都是实证型,还处于强经验驱

动的研究阶段。因此,目前还缺少统一系统的理论框架支持。针对 RDCNs 自身的理论研究仍然是个新领域。表 3 综合对比了近年来研究人员所提出的典型 RDCNs 架构。

表 3 RDCNs 典型架构综合比较
Table 3 Comprehensive comparison of RDCNs architectures

重构级别	名称,时间	主要特点	使能技术	重构目标	重构延迟	流量响应	应用支持	技术进展
链路级	Flyways ^[15] ,2009	应对流量热点,按需增加链路	60 GHz	高吞吐量	毫秒级	是	作业级	原型实现
	Firefly ^[31] ,2014	灵活重构,支持成组通信	FSO	吞吐量最大化	微秒级	是	作业级	原型实现
	Projector ^[16] ,2016	使用天花板镜面形成重构链路	FSO	无饥饿低延迟	微秒级	是	作业级	原型实现
	xWeaver ^[18] ,2018	引入人工智能优化重构链路	OCSs	吞吐量最大化	毫秒级	是	流级	仿真实现
	TIO ^[17] ,2019	可见光随机生成重构链路	VLC	吞吐量最大化	毫秒级	是	流级	仿真分析
层级	cThough ^[55] ,2010	增加 OCSs 层,隔离调度	OCSs	吞吐量最大化	毫秒级	是	作业级	原型实现
	Larry ^[59] ,2018	聚合层间增加 OCSs 层	OCSs	低延迟	微秒级	是	流级	原型实现
	RDC ^[61] ,2019	TOR 和聚合层间增加 OCSs 层	OCSs	低延迟	微秒级	是	流级	原型实现
	MR ^[60] ,2019	利用插线板最小化布线开销	插线板	低成本	微秒级	是	流级	原型实现
	DUA ^[41] ,2019	增加 FPGA 层,低延迟调度	智能网卡	低延迟	微秒级	是	流级	实际部署
	Sirius ^[11] ,2020	低延迟 OCSs	OCSs	低延迟	纳秒级	否	数据包级	原型实现
拓扑级	Flattree ^[20] ,2017	在随机和规则拓扑间切换	OCSs	吞吐量最大化	微秒级	否	流级	原型实现
	RotorNet ^[21] ,2017	周期性主动重构	光学转子	吞吐量最大化	微秒级	否	流级	原型实现
	Flexspander ^[66] ,2019	引入扩展器网络降低成本	OCSs	低延迟	纳秒级	否	数据包级	仿真实现
	2-Tier ^[12] ,2019	移动拓扑子层按需搬运容量	ALL	ALL	ALL	ALL	ALL	建模分析

结束语 为了推动 RDCNs 领域的未来发展,我们对 RDCNs 架构设计中的大量研究工作进行了分类分析,并首次从重构设计空间这一新颖角度对 RDCNs 研究的进展进行梳理,而非依据传统分层模型进行分类归纳。这也符合 RDCNs 的初衷,旨在解决传统对称性网络设计的问题。如前所述,多重驱动下的 RDCNs 研究工作正变得更加丰富,将产生更大的影响力。

随着新型大数据和人工智能应用以及松耦合数据中心的发展,未来 RDCNs 的设计空间将向更细粒度(如存储元器件^[90])和更大尺度(如广域网、大规模分布式机器学习环境^[91])延伸,总体上来说,RDCNs 的研究尚处于探索起步阶段,RDCNs 和软硬件界面分割关系密切,其研究将有利于促进数据中心体系结构的演进。已有的实践为理论探索和物理层硬件研发提出了许多需要解决的问题。现有 RDCNs 研究和预期规模化部署落地有望倒逼更具弹性(Resilience)的新型物理层通信技术的发展,例如 MEMS 的持续创新^[92]以及光学领域正在进行的创新^[93]有望在 RDCNs 的促进下取得更大的进展,也有望促进网络科学和图谱理论的发展。

参 考 文 献

[1] China launches ‘New Infrastructure’ an innovation-driven program [EB/OL]. <https://news.cgtn.com/news/2020-03-17/China-launches-New-Infrastructure-a-innovation-driven-program-OVo9mgDvGg/index.Html>.

[2] CHOI S,BURKOV B,ECKERTA,et al. FBOSS:building switch software at scale[C]//Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). Budapest, Hungary, 2018:342-356.

[3] SINGH A,ONG J,AGARWAL A,et al. Jupiter rising:A decade of clos topologies and centralized control in Google’s datacenter network[J]. ACM Computer Communication Review, 2015, 45(4):183-197.

[4] HUAWEI,Global Connectivity Index[EB/OL]. [https://www.](https://www.huawei.com/minisite/gci/en)

[huawei.com/minisite/gci/en](https://www.huawei.com/minisite/gci/en).

[5] Hyperscale Data Centers Market Insights 2019,Global and Chinese Analysis and Forecast to 2024 [EB/OL]. <https://www.grandresearchstore.com/services/global-hyperscale-data-centers-2019-2024-542>.

[6] Google,Cloud TPU [EB/OL]. <https://cloud.google.com/tpu>.

[7] NVIDIA A100 Tensor Core GPU Architecture [EB/OL]. <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf>.

[8] HENNESSY,JOHN L,PATTERSON D A. A New Golden Age for Computer Architecture[J]. Communications of the ACM, 2019,62(2):48-60.

[9] MOORE S K. Another step toward the end of Moore’s law[J]. IEEE Spectrum, 2019,56(6):9-10.

[10] CHOWDHURY M,ZHONG Y,STOICA I. Efficient coflow scheduling with Varys[C]//Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). Chicago, Illinois, USA, 2014.

[11] BALLANI H,COSTA P,BEHRENDT,et al. Sirius:A Flat Datacenter Network with Nanosecond Optical Switching[C]//Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). New York, USA, 2020:782-797.

[12] ZHANG D K,WANG X W,HUANG M. 2-Tier:Data center network architecture with symbiotic traffic-skew mitigated capability and evolutionary deployment[J]. Journal of Southeast University (Natural Science Edition), 2020,50(2):402-408.

[13] XIA Y,SCHLANSKER M,NGT S E,et al. Enabling Topological Flexibility for Data Centers Using Omniswitch[C]//Proceedings of the 7th USENIX Conference on Hot Topics in Cloud Computing. Santa Clara, CA, 2015.

[14] VANINI E,PAN R,ALIZADEH M,et al. Let it flow:Resilient asymmetric load balancing with flowlet switching[C]//14th USENIX Symposium on Networked Systems Design and Implementation. Boston, USA, 2017:407-420.

- [15] KANDULA S, PADHYE J, BAHL V. Flyways to de-congest data center networks[C]//8th ACM Workshop on SIGCOMM logo Hot Topics in Networks. New York, USA, 2009; 1-6.
- [16] GHOBADI M, MAHAJAN R, PHANISHAYEE A, et al. Projector: Agile reconfigurable data center interconnect[C]//Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). Florianopolis, Brazil, 2016; 216-299.
- [17] QIN Y, GUO D, TANG G, et al. TIO: A VLC-Enabled Hybrid Data Center Network Architecture[J]. Tsinghua Science & Technology, 2019, 24(4): 484-96.
- [18] WANG M, CUI Y, XIAO S, et al. Neural Network Meets DCN: Traffic-driven Topology Adaptation with Deep Learning[C]//Proceedings of the ACM on Measurement and Analysis of Computing Systems. Irvine, USA, 2018.
- [19] ZHOU X, ZHANG Z, ZHU Y, et al. Mirror mirror on the ceiling: exible wireless links for data centers[C]//Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). Helsinki, Finland, 2012.
- [20] XIA Y, SUN X S, DZINAMARIRA S, et al. A Tale of Two Topologies: Exploring Convertible Data Center Network Architectures with Flat-Tree[C]//Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). Los Angeles, USA, 2017; 295-308.
- [21] MELLETTE W M, MCGUINNESS R, ROY A, et al. RotorNet: A Scalable, Low-Complexity, Optical Datacenter Network[C]//Proceedings of the Conference of the ACM Special Interest Group on Data Communication. 2017; 267-80.
- [22] SALMAN S, STREIFFER C, CHEN H, et al. DeepConf: Automating Data Center Network Topologies Management with Machine Learning[C]//Proceedings of the 2018 Workshop on Network Meets AI & ML. Budapest, Hungary, 2018; 8-14.
- [23] GREENBERG A, HAMILTON J R, JAIN N, et al. VL2: A Scalable and Flexible Data Center Network[J]. Communications of the ACM, 2011, 54(3): 95-104.
- [24] GUO C, LU G, LI D, et al. BCube: A High Performance, Server-Centric Network Architecture for Modular Data Centers[C]//Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). Barcelona, Spain, 2009; 63-74.
- [25] Broadcom. 25.6 Tb/s StrataXGS Tomahawk 4 Ethernet Switch Series [EB/OL]. <https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm56990-series>. 2020.
- [26] CALABRETTA N, MIAO W, MEKONNEN K, et al. Monolithically Integrated WDM Cross-Connect Switch for High-Performance Optical Data Center Networks[C]//2017 Optical Fiber Communications Conference and Exhibition (OFC). Los Angeles, USA, 2017; 1-3.
- [27] XIA W, ZHAO P, WEN Y, et al. A survey on data center networking (DCN): infrastructure and operations[J]. IEEE Communications Surveys and Tutorials, 2017, 19(1): 640-656.
- [28] CELIK A, SHIHADA B, ALOUINI M S. Optical Wireless Data Center Networks: Potentials, Limitations, and Prospects[J]. Broadband Access Communication Technologies XIII. 2019, 10945: 1094501.
- [29] ZHUO D, GHOBADI M, MAHAJAN R, et al. Understanding and mitigating packet corruption in data center networks[C]//Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). Los Angeles, USA, 2017.
- [30] ZOU S, WEN X, CHEN K, et al. Virtualknotter: Online virtual machine shuffling for congestion resolving in virtualized data-center[J]. Computer Networks, 2014, 67: 141-153.
- [31] HAMEDAZIMI N, QAZI Z, GUPTA H, et al. FireFly: A Reconfigurable Wireless Data Center Fabric Using Free-Space Optics[C]//Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). London, UK, 2015.
- [32] LUO L, GUO D, WU J, et al. Vlccube: A vlc enabled hybrid network structure for data centers[J]. IEEE Transactions on Parallel and Distributed Systems, 2016, 28(7): 2088-2102.
- [33] NIKOLOVA D, RUMLEY S, CALHOUND, et al. Scaling silicon photonic switch fabrics for data center interconnection networks[J]. Opt. Express, 2015, 23(2): 1159-1175.
- [34] S320 optical circuit switch [EB/OL]. <http://www.calient.net/members-area/?redirect-to=/download/s320-optical-circuit-switch-datasheet>, 2019.
- [35] Fujitsu's high speed MEMS mirror switch[J]. III-Vs Review, 2003, 16(8): 30.
- [36] KONG Q, HUANG S, ZHOU Y, et al. On the performance of a scalable optical switching architecture for flat intercluster data center network with centralized control[J]. Optical Engineering, 2014, 53(7): 075101.
- [37] ZERVAS G S, DE LEENHEER M, SADEGHIOON L, et al. Multi-granular optical cross-connect: Design, analysis, and demonstration[J]. IEEE/OSA Journal of Optical Communications and Networking, 2009, 1(1): 69-84.
- [38] FIORANI M, CASONI M, ALEKSIC S. Performance and power consumption analysis of a hybrid optical core node[J]. IEEE/OSA Journal of Optical Communications and Networking, 2011, 3(6): 502-513.
- [39] AW E T, WONFOR A, GLICK M, et al. Large Dynamic Range 32x32 Optimized non-blocking SOA based Switch for 2.56 Tb/s Interconnect Applications[C]//Optical Communication, VDE, 2007.
- [40] CHENG Q, DING M, WONFOR A, et al. The feasibility of building a 64x64 port count SOA-based optical switch[C]//International Conference on Photonics in Switching (PS). Florence, Italy, 2015; 199-201.
- [41] SHU R, CHENG P, CHEN G, et al. Direct universal access: Making data center resources available to FPGA[C]//16th USENIX Symposium on Networked Systems Design and Implementation. Boston, USA, 2019; 127-140.
- [42] AGUINALDO R, FORENCICH A, DEROCSEC, et al. Energy-efficient, digitally-driven "fat pipe" silicon photonic circuit switch in the UCSD MORDIA data-center network[C]//Lasers & Electro-optics. IEEE, 2014.

- [43] TESTA F, PAVESI L. Optical Switching in Next Generation Data Centers[M]. Springer, 2018.
- [44] OpenFlow protocol API Reference [EB/OL]. <https://github.com/open-rpa/openflow>, 2020.
- [45] BOSSHART P, DALY D, GIBB G, et al. P4: Programming Protocol-independent Packet Processors [C] // ACM Computer Communication Review, 2014, 44(3): 87-95.
- [46] RMON Configuration Guide [EB/OL]. https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/rmon/configuration/xr-16/rmon-xr-16-book/Configuring_RMON_Support.html.
- [47] JIANG C, HAN G, LIN J, et al. Characteristics of Co-Allocated Online Services and Batch Jobs in Internet Data Centers: A Case Study from Alibaba Cloud [J]. IEEE Access, 2019, 7: 22495-22508.
- [48] SUN X, ANSARI N, WANG R. Optimizing Resource Utilization of a Data Center [J]. IEEE Communications Surveys and Tutorials, 2016, 18(4): 2822-2846.
- [49] AI at Scale - Microsoft Research [EB/OL]. <https://www.microsoft.com/en-us/research/project/ai-at-scale/>.
- [50] Facebook. Reinventing Facebook's data center network [EB/OL]. <https://code.fb.com/data-center-engineering/fl6-minipack>.
- [51] WHITTED W H, AIGNER G. Modular data center; U. S. Patent 7,278,273 [P]. 2007-10-9. <https://patents.google.com/patent/US7278273B1/en>.
- [52] BAO J, DONG D, ZHAO B, et al. Flycast: Free-space optics accelerating multicast communications in physical layer [J]. ACM Computer Communication Review, 2015, 45(5): 97-98.
- [53] HAMZA A, DEOGUN J, ALEXANDER D. Evolution of data centers: A critical analysis of standards and challenges for FSO links [C] // 2015 IEEE Conference on Standards for Communications & Networking. Tokyo, Japan, 2015: 100-105.
- [54] ALGHADHBAN A. F4Tele: FSO for Data Center Network Management and Packet Telemetry [J]. arXiv: 2006. 07419, 2020.
- [55] WANG G, ANDERSEN D G, KAMINSKY M, et al. C-Thorough: Part-Time Optics in Data Centers [C] // Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). Chicago, USA, 2010: 327-338.
- [56] FARRINGTON N, PORTER G, RADHAKRISHNAN S, et al. Helios: a hybrid electrical/optical switch architecture for modular data centers [J]. ACM Computer Communication Review, 2011, 41(4): 339-350.
- [57] BAZZAZ H H, TEWARI M, WANG G, et al. Switching the Optical Divide: Fundamental Challenges for Hybrid Electrical/Optical Datacenter Networks [C] // Proceedings of the 2nd ACM Symposium on Cloud Computing (SOCC'11). ACM, 2011.
- [58] WEN K, SAMADI P, RUMLEY S, et al. Flexfly: Enabling a Reconfigurable Dragonfly through Silicon Photonics [C] // Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Salt lake city, USA, 2016.
- [59] CHATZIELEFTheriou A, LEGTCHENKO S, WILLIAMS H, et al. Larry: Practical network reconfigurability in the data center [C] // 15th USENIX Symposium on Networked Systems Design and Implementation. Renton, USA, 2018: 141-156.
- [60] ZHAO S, WANG R, ZHOU J, et al. Minimal Rewiring: Efficient Live Expansion for Clos Data Center Networks [C] // 16th USENIX Symposium on Networked Systems Design and Implementation, 2019: 221-234.
- [61] WU D, WANG W, CHEN A, et al. Say No to Rack Boundaries: Towards A Reconfigurable Pod-Centric DCN Architecture [C] // Proceedings of the 2019 ACM Symposium on SDN Research, 2019: 112-118.
- [62] FIETZ J, WHITLOCK S, IOANNIDIS G, et al. VNTor: Network Virtualization at the Top-of-Rack Switch [C] // Proceedings of the Seventh ACM Symposium on Cloud Computing. Hawaii, USA, 2016: 428-441.
- [63] LUGONES D, KATRINIS K, THEODOROPOULOS G, et al. A Reconfigurable, Regular-Topology Cluster/Datacenter Network Using Commodity Optical Switches [J]. Future Generation Computer Systems, 2014, 30: 78-89.
- [64] SINGLA A, SINGH A, RAMACHANDRAN K, et al. Proteus: A Topology Malleable Data Center Network [C] // Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks. New Delhi, India, 2010: 8.
- [65] CHEN K, SINGLA A, SINGH A, et al. OSA: An optical switching architecture for data center networks with unprecedented flexibility [J]. IEEE/ACM Transactions on Networking, 2013, 22(2): 498-511.
- [66] TEH M Y, WU Z, BERGMAN K. Flexspander: Augmenting Expander Networks in High-Performance Systems with Optical Bandwidth Steering [J]. IEEE OSA Journal of Optical Communications and Networking, 2020, 12(4): B44-B54.
- [67] VALIANT L G. A Scheme for Fast Parallel Communication [J]. SIAM J, 1982, 11(2): 350-361.
- [68] MEYERSON A, TAGIKU B. Minimizing average shortest path distances via shortcut edge addition [C] // Proceeding of APPROX/RANDOM. Berlin, Germany, 2009: 272-285.
- [69] EDMONDS J. Paths, trees and flowers [J]. Canadian Journal of Mathematics, 1965, 17: 449-467.
- [70] SINGLA A, SINGH A, RAMACHANDRAN K, et al. Proteus: a topology malleable data center network [C] // Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). New Delhi, India, 2010.
- [71] ZHANG M, MYSORE R N, SUPITTAYAPORN PONG S, et al. Understanding Lifecycle Management Complexity of Data-center Topologies [C] // 16th USENIX Symposium on Networked Systems Design and Implementation. Boston, USA, 2019: 235-254.
- [72] PAPAGELIS M, BONCHI F, GIONIS A. Suggesting ghost edges for a smaller world [C] // Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM, 2011.
- [73] FOERSTER K T, GHOBADI M, SCHMID S. Characterizing the algorithmic complexity of reconfigurable data center architectures [C] // ACM/IEEE Symposium on Architectures for Networking and Communications Systems. New York, USA, 2018.

[74] FOERSTER K T, PACUT M, SCHMID S. On the complexity of non-segregated routing in reconfigurable data center architectures[J]. ACM Computer Communication Review. 2019, 49(2): 2-8.

[75] FENZ T, FOERSTER K, SCHMID S, et al. Efficient non-segregated routing for reconfigurable demand-aware networks[C]// 18th IFIP Networking Conference. Warsaw, Poland, 2019.

[76] GOEL A, KAPRALOV M, KHANNAS. Perfect matchings in $o(n \log n)$ time in regular bipartite graphs[J]. SIAM J. , 2013, 42(3): 1392-1404.

[77] LIU H, MUKERJEE M K, LI C, et al. Scheduling Techniques for Hybrid Circuit/ Packet Networks[C]// Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies. Heidelberg, Germany, 2015.

[78] LI X, HAMDI M. On scheduling optical packet switches with reconfiguration delay[J]. IEEE Journal on Selected Areas in Communications, 2003, 21(7): 1156-1164.

[79] BOJJA V S, ALIZADEH M, VISWANATH P. Costly circuits, submodular schedules and approximate caratheodory theorems [J]. Queueing System, 2018, 88(3/4): 311-347.

[80] GALE D, SHAPLEY L S. College admissions and the stability of marriage[J]. The American Mathematical Monthly, 1962, 69(1): 9-15.

[81] SINGLA A, HONG C Y, POPAL, et al. Jellyfish: Networking data centers randomly[C]// 9th USENIX Symposium on Networked Systems Design and Implementation. Lombard, USA, 2012: 225-238.

[82] VALADARSKY A, SHAHAF G, DINITZ M, et al. Xpander: Towards optimal-performance datacenters[C]// 12th International Conference on Emerging Networking Experiments and Technologies Irvine. USA, 2016: 205-219.

[83] KASSING S, VALADARSKY A, SHAHAFG, et al. Beyond fat-trees without antennae, mirrors, and disco-balls[C]// Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM). Los Angeles, USA, 2017: 281-294.

[84] MELLETTE W M, DAS R, GUO Y, et al. Expanding across time to deliver bandwidth efficiency and low latency[J]. CoRR, abs/1903.12307, 2019.

[85] AVIN C, HERCULES A, LOUKAS A, et al. RDAN: Toward Robust Demand-Aware Network Designs[J]. Information Processing Letters, 2018, 133: 5-9.

[86] AVIN C, MONDAL K, SCHMID S. Demand-aware network designs of bounded degree[J]. arXiv:1705.06024, 2017.

[87] AVIN C, SCHMID S. Toward demand-aware networking: A theory for self-adjusting networks[J]. ACM Computer Communication Review, 2018, 48(5): 31-40.

[88] CHEN K, WEN X, MA X, et al. Toward A scalable, fault-tolerant, high-performance optical data center architecture [J]. IEEE/ACM Transaction on Networking, 2017, 25 (4): 2281-2294.

[89] KATTA N, GHAG A, HIRA M, et al. Clove: Congestion-Aware Load Balancing at the Virtual Edge[C]// Proceedings of the 13th International Conference on Emerging Networking Experiments and Technologies. 2017: 323-335.

[90] ALACHIOTIS N, ANDRONIKAKIS A, PAPADAKIS O, et al. DReDBox: Materializing a Full-Stack Rack-Scale System Prototype of a next-Generation Disaggregated Datacenter[C]// 2018 Design, Automation & Test in Europe Conference & Exhibition. 2018: 1093-1098.

[91] AI at Scale-Microsoft Research [EB/OL]. <https://www.microsoft.com/en-us/research/project/ai-at-scale>.

[92] LUGONES D, KATRINIS K, COLLIER M. Are configurable optical/electrical interconnect architecture for large-scale clusters and datacenters[C]// Proceedings of the 9th Conference on Computing Frontiers. Cagliari, Italy, 2012.

[93] CLARK K A, CLETHEROE D, GERARD T, et al. Synchronous subnanosecond clock and data recovery for optically switched data centres using clock phase caching[J]. Nature Electronics, 2020, 3(7): 426-433.

A portrait photograph of ZHANG Deng-ke, a man with short dark hair, wearing a dark jacket over a light-colored shirt.

ZHANG Deng-ke, born in 1981, Ph. D candidate. His main research interests include network design and optimization of hyperscale data center and spectral graph theory.

A portrait photograph of WANG Xing-wei, a man with short grey hair and glasses, wearing a dark suit, white shirt, and a patterned tie.

WANG Xing-wei, born in 1968, Ph. D, professor, Ph. D supervisor. His main research interests include NGI, mobile wireless Internet, IP/DWDM optical Internet, etc.