数据中心中光交换技术研究的现状与挑战

Optical Switching for Data Center: Current Status and Challenging

郭秉礼/GUO Bingli，黄善国/HUANG Shanguo

（北京邮电大学信息光子学与光通信国家重点实验室，北京 100876）

(State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Post and Telecommunication, Beijing 100876, China)

摘要：对构建下一代灵活数据中心互连网络的问题，提出了结合不同维度光交换技术的光电混合互连网络搭建思路，使未来数据中心网络具备动态拓扑重构、灵活带宽调整等特性，可以有效应对数据中心发展中由于硬件重构、业务多样等趋势导致的业务突发性强、通信模式差异大等问题。同时，在光电混合互连网络的构建与实用化过程中，仍需在智能控制体系架构、高速突发接收、低延时与低抖动控制等方面取得突破。

关键词：数据通信；数据中心；光交换；拓扑重构

Abstract: In this paper, a hybrid optical and electrical interconnection network with multi-dimension of optical switching technologies is proposed for the construction of next generation flexible data center interconnection network. The future data center network (DCN) has the characteristics of dynamic topology reconfiguration and flexible bandwidth adjustment which can effectively deal with the problems such as traffic burstness and various communication pattern due to hardware disaggregation and service diversification. At the same time, there are still lots of technique bottlenecks that need to be broken, including intelligent network control system, high speed optical burst receiver and low latency and low jitter network control system in the approach of building the optical/electrical hybrid interconnection network.

Key words: data communication; data center; optical switching; topology reconfiguration

DOI：10.12142/ZTETJ.201905004

网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20190929.1636.002.html

网络出版日期：2019-09-30

收稿日期：2019-08-16

视频业务、在线游戏等个人业务的快速兴起与以云计算/大数据为代表的企业互联网业务的强势推广，对目前数据服务过程中的计算、交互与存储能力提出了前所未有的挑战。同时，随着人工智能和机器学习等计算密集型服务的繁荣，也极大地提高了对数据计算、存储执行效率与资源利用率的要求。在传统技术手段中，往往通过增加数据中心（DC）空间来容纳更多的机架和服务器，进而达到增加数据处理能力的目的。然而，随着业务需求的增加，线性扩容系统的方式使数据中心正在逼近能耗极限，目前需要寻找新的技术手段来最大限度地提高计算能力和效率。在节点算力与使用效率的提升方面，信息技术（IT）领域的研究人员提出诸如硬件解耦、与高性能计算体系融合等多方面的解决方案，在本文中不再赘述，后续部分着重讨论数据中心互连网络（DCN）方面的进展与目前面临的挑战。

在DC带宽密度提升方面，在高速率、低功耗需求的驱动下，相同容积的光模块需要具备更大的数据传输量，多通道、光子集成与混合集成技术可以将光组件做得很紧凑，顺应光模块小型化趋势，方便使用成熟自动化集成电路（IC）封装工艺，有利于量产，是未来数据中心用光模块提升带宽密度的行之有效的

技术手段。

另一方面，连接无数计算节点的互连网络承担了海量数据的传输与交换功能，不再只是一个流量转发的、仅需追求稳定的平台，它逐渐成为一个重要的生产环节。上述业务的演进趋势对数据通信中的互连网络，在时延和吞吐量方面提出了更高的需求，使数据中心网络业务承载能力的提升逐渐成为一件亟待解决的事情。本文中，我们通过分析目前DCN在应对突发业务与带宽灵活调度等方面面临的一些挑战，进一步探讨光交换技术在帮助DCN应对上述挑战中可能起到的积极作用，最后总结了光电混合DCN可行性及其仍需解决的一些技术难题。

1 数据中心互连网络面临的主要挑战

目前DC通过电交换设备形成各种形态的互连拓扑，把大量通用服务器互连。随着DC规模的急速增加以及服务器性能的提升，近年来的接入路由器和核心路由器端口速率需求将会随之达到40 Gbit/s和400 Gbit/s。然而如图1所示，电交换机的能效随着交换容量的增大而无法继续提升[1]，这使得数据中心在能耗、带宽提供这两方面遇到瓶颈。DC亟须解决能耗问题带来的扩容瓶颈，才能以合理的功耗继续提升网络带宽。有研究表明DC中99%的链路利用率不足10%[2]，同时欧盟FP7框架下开展的面向光通信的数据速率和功率感知的自适应收发器（ADDAPT）项目[3]的研究显示，某些DC中链路无效数据传输时间高达90%，IBM研究人员通过实验测试指出光模块具有快速启动与突发接收功能时能耗可节省85%。因此，如何提高DC系统能效，需要突破现有框架，从DC业务特点出发，探索新的思路。

DC承载业务类型多样，流量分布不均且具有很强的突发性[4]，现有DCN流量工程机制复杂，无法快速应对流量的波动。一些热点机架承载着数据中心中绝大部分的流量[4]，造成热点机架间的路径出现拥塞，端到端数据延时加大，而其他位置的网络资源却处于闲置状态。互联网协议（IP）层带宽调度技术又过于复杂，无法满足DC业务时效性和网络运维灵活性的需求。上述情况造成了互连带宽的浪费，限制了整个数据中心的吞吐量与业务承载能力。

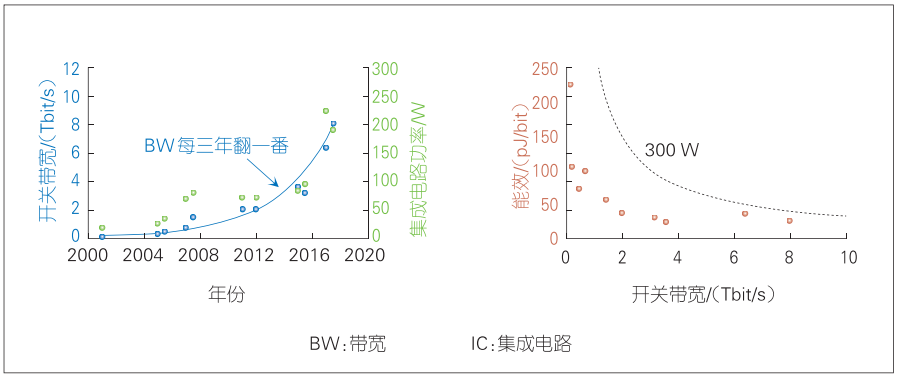
近年来以Facebook等互联网巨头为代表的数据中心用户希望通过硬件解耦[5]，即在硬件层将同类资源聚合为资源池，如中央处理器（CPU）池、内存（Memory）池、存储（Storage）池，然后根据应用的需求分配具备相应特性的资源组合，来提高DC中资源的利用率和灵活性，同时降低资本性支出（CAPEX）和运营成本（OPEX）。资源池间互连网络的带宽提供能力（大瞬时带宽、低延时、高动态）是决定资源解耦范围的关键因素之一。因此，全局资源调度与高效网络重构能力也是硬件解耦等数据中心技术演进趋势对DCN提出的新要求。

为了应对上述挑战，DC互连网络亟须在能效提升和带宽资源灵活调度方面寻找新的解决思路。由于光交换技术具有速率透明、低功耗、可重配置等优势，基于光交换的互连网络被认为是一种解决DC面临的问题的有效方法，以满足DC日益增长的高带宽、低延迟和高能效等方面的需求。

2 光交换相关技术的发展现状与挑战

2.1 光交换矩阵技术发展现状

光交换矩阵可以实现光束在时间、空间、波长等维度上的切换，是光通信、光计算机、光信息处理等光信息系统的关键器件。通常来说，光交换矩阵的性能由开关单元、切



▲图1 电分组交换专用集成电路能效随带宽的变化[2]

换机制和互连结构等多方面因素决定。表1中，我们对目前具备商用可能性的光交换矩阵的相关性能进行了分析比较。其中，基于微机电系统（MEMS）和波束控制（Beam-Steering）的开关矩阵已经达到商用成熟阶段，规模已经可以达到数百端口；纳秒级切换时间的光交换矩阵规模仍较小，依赖于模块多级级联，因此插损偏高。在近年来的光交换矩阵的研究中，研究人员在驱动集成、片上放大、偏振不敏感方面做了大量的研究，结果表明光交换矩阵整体向着高可靠、低损耗、小功耗、小体积以及大规模方向发展。同时，近几年光交换矩阵在电信领域（光传送网（OTN）建设中需要构建大量的可重构型光分插复用器（ROADM）、光交叉连接（OXC）节点，光交换矩阵是搭建这些节点的基础模块）、网络测试领域有大规模应用的趋势。

2.2 光收发节点技术

在传统的点对点光纤通信或光电路交换（OCS）系统中，光接收机一般接收另外一个固定节点发送的连续模式光信号，并从中检测出电信号。使用快速光交换的收发系统，是因为光信号的非连续性：除了满足传统光接收机所要求的高灵敏度外，还要有较大的动态范围和快速的响应能力，即突发模式接收技术。光突发模式接收机主要由信号整形、突发同步和数据恢复3大部分组成。表2中，我们对传统接收机与突发模式接收机相关性能要求进行了比较。其中，对于突发模式信号，两相邻突发分组信号间有相位突变。在这种情况下，要避免使用传统的交流耦合方式。因光接收机在交流耦合之后，要对信号进一步放大，再进行整形和判别输出；而突发信号的不均衡，其直流成分（均值）发生漂移，要影响到后面放大器的直流工作点，使其不能稳定工作。此外，判决电路对幅度不均衡信号进行判决时，要么会出现小信号的丢失，要么会出现大信号的脉宽失真。上述原因都是研制高速突发接收模块的需要解决的技术难题。

目前，商用的突发模式光接收机主要应用在各种各样的无源光网络（PON）中，支持1.25 Gbit/s以及10 Gbit/s的速率。电子设备工程（EEE）下一代以太网无源光网络（NG-EPON）正致力于25 Gbit/s单波长和50 Gbit/s双波长的解决方案。这与最近数据中心传输速率从10 Gbit/s迅速转变为25 Gbit/s的趋势是一致的[6]。近来，针对高速光突发模式接收机的研究也取得了一些进展。IBM在国际晶体管电路讨论会（ISSCC）2015上报道了突发模式时钟和数据恢复（BM-CDR）以25 Gbit/s的速率在18.5 ns锁定时间下的成功演示实验[7]。IBM和瑞士洛桑联邦理工学院（EPFL）在国际固态电路（ISSCC）2018上报道了使突发模式光接收机（BM-Optical RX）从10 Gbit/s提高到56 Gbit/s的实验，该实验演示56 Gbit/s BM-Optical RX通过链接协议完成384UI（6.8 ns）中的唤醒和CDR锁定[8]。在2018年光纤通信展览会及研讨会（OFC）上，IBM报道了一种由850 nm光电二极管（PD）阵列组成的、以低成本垂直腔面发射激光器（VCSEL）为基础的、14 nm互补金属氧化物（CMOS）的4×40 Gbit/s

▼表1 采用不同技术的光开关性能比较

	MEMS开关	波束控制开关	LCoS开关	铌酸锂光电开关	半导体放大器开关	热光开关
开关时间	~20 ms	~ 20 ms	>100 ms	~ ns	~ ns	~10 ms
插损	中等	低	高	高	中等	低
能耗	中等	低	高	高	低	低
稳定度和可扩展性	高	高	中等	中等	中等	低
端口数	320 x 320	384 x 384	1 x 20	32 x 32	16 x 16	8 x 8
可靠性	低	低	高	高	中等	高
成本	中等	中等	低	高	高	低

LCoS：硅基液晶　　MEMS：微机电系统

▼表2 传统接收机与突发模式接收机比较

接收机类型	传统光接收机	突发模式光接收机
耦合方式	交流	直流
判决门限建立方式	固定	动态
幅度和时钟恢复时间	微秒	纳秒

2 pJ/bit 光接收器(RX)。该 RX 可以实现低至 8 ns 的 Power-on 和 CDR-Lock 时间[9]。综上所述,高速光突发模式接收机的相关技术研究也得到了显著进步,有望支撑未来 100 G 以内光突发接收模块的相关研制。

2.3 全光交换网络相关技术

依赖于不同维度的光开关器件的研究进展,基于光交换的 DCN 近年来得到广泛关注,包括 IBM、Google 在内的大量企业与研究机构在数据中心内也进行了大量的尝试与实验。表 3 和表 4 为目前业界主要的主要光互连方案在技术特性、成熟度等方面的比较。其中,开放式可插拔规范(OPS)、光突发交换技术(OBS)需要复杂的冲突避免机制,需要在光缓存器件、光逻辑器件等方面进行技术突破;光电路交换(OCS)相关技术的成熟度较高,光时隙交换次之,光时隙交换系统依赖于快速光交换器件;收发模块方面,除 OCS 外,其他交换机制的实现均依赖于突发模式收发技术。综上所述,光时分复用(OTDM)系统在数据延时、控制时效性等方面有一定的优势,混合波分复用(WDM)的 OTDM 系统可以作为一种实现数据中心内动态光互连拓扑重构的可行方案;而 OCS 机制适合于可以提前预知流量变化的场景。

3 光电混合 DCN 发展趋势

大量研究显示,全光交换技术在特定场景下比电交换技术在能效等方面有一定的优势,但无法全面替代电交换技术细粒度的业务调度能力,所以如何设计光电混合的 DCN 成为目前业界所研究的重点。需要充分发挥各自的一些优势,使其能够适应 DC 内多样、突发的业务流量。

3.1 数据中心内业务特征

数据中心网络业务的第 1 个特征是南北向流量与东西向流量的“二八定律”。在数据中心发展的早期,出于用户对服务器上大容量存储数据的访问需求,大量流量流向机架外部,然而随着互联网和云产业的迅猛发展,现阶段的数据中心中,这种南北向流量已降低至 20% 左右[10]。预计到 2021 年,94% 的工作负载和计算实例将由云数据中心处理;传统数据中心处理的比例仅为 6%[11]。在新兴的云数据中心中,应用和其所依赖的组件大多部署在同一个机架内,网络流量具有明显的特征:75% 以上的流量停留在机架内部,核心链路利用率低于 25%[12]。

数据中心网络的第 2 个特征是大象流与老鼠流的混合。一个数据中心通常需要承载各种各样的业务,为用户提供包括网页搜索、直播视频、基于 IP 的语音传输(VoIP)、数据存储、资源下载、即时通信等丰富多彩的云服务。这些应用程序产生的流量具有不同的特征,可以根据其传输数据量的多少分为大象流和老鼠流。大象流通常产生自带宽敏感型业务,例如数据库同步、存储

▼表 3 不同光交换技术各方面特性比较

交换技术	光开关性能要求	收发模块性能要求	其他	交换粒度	控制时效性
OCS	~10 ms	连续模式	支持 WDM	波长	~200 ms, 链路端到端预约配置
OPS	~10 ns	突发模式	单波、可混合 WDM	光包	标签配合路由表转发,不依赖实时控制
OBS	~1 μs-~10 ns	突发模式	单波、可混合 WDM	光突发包	信令预约时隙,实时协议交互
OTDM	~100 ns	突发模式	单波、可混合 WDM	光时隙	支持实时按需调整

OCS:光电路交换技术
OPS:开放式可插拔规范
OBS:光突发交换技术
OTDM:光时分复用
WDM:波分复用

▼表 4 不同光交换技术成熟度对比

交换技术	光开关	收发模块	交换粒度	控制时效性
OCS	成熟度高,端口可达 512	成熟度高	成熟	控制平台与协议成熟
OPS	成熟度低,实验室端口可达 8	10 Gbit/s 突发接收技术成熟、高速突发模式亟待研究	依赖光标签处理技术	要求低
OBS			依赖光开关性能	不成熟
OTDM	成熟度中等,实验室端口可达 16		依赖光开关性能	相对成熟

OCS:光电路交换技术
OPS:开放式可插拔规范
OBS:光突发交换技术
OTDM:光时分复用

备份、数据分析等需要占用大带宽的业务；老鼠流通常产生自时延敏感型业务，例如社交网络、搜索引擎等实时性业务。相关研究表明，传输数据量不足1 MB的突发性老鼠流占数据中心网络流数量的90%以上，而传输数据量不超过100 MB的老鼠流占到数据中心网络流数量的98%左右[13]。传输数据量大于100 MB的大象流的数量虽然比较少，却承载了网络中90%以上的数据量，即90%以上的流量被认为是老鼠流，而90%以上的数据量在大象流中。

数据中心网络的第3个特征是流量分布的突发性和不均匀性。局部的hot spots承载了大量的流量，其他地方闲置的链路造成了网络资源的浪费。有相关研究指出，数据中心网络内86%的链路会因为突发的大象流而产生超过10 s的网络拥塞[14]。

上述DC内的流量特征决定了DCN流量调度问题的复杂性以及传统互联网解决方案在应对上述一些问题时也必将会面临着大量的不适应性。

3.2 基于通信模式的拓扑重构

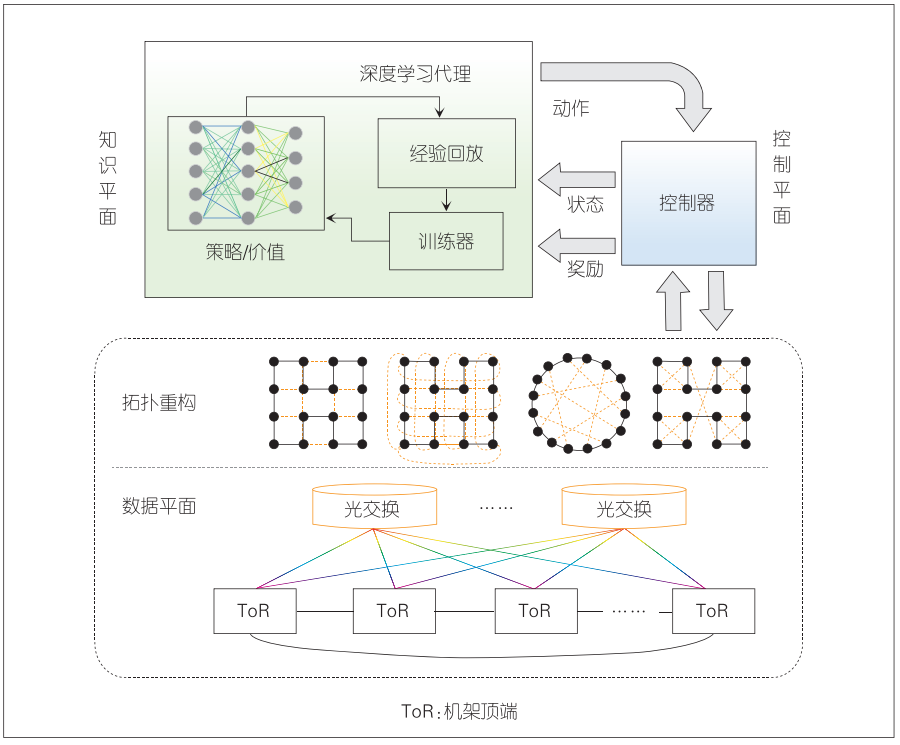
针对上述流量特征，电交换网络适合针对老鼠流进行灵活分发，而光交换网络提供了可重配置的快速光通道，为突发的大象流业务按需提供实时的高速连接。为了达到上述目的，需要在数据中心中部署知识平面、智能控制平面来实现流量的高效感知和光电混合网络的实时控制。

图2是基于AI流量分析的光电混合DCN重构体系。在该体系中，首先通过知识平面对数据中心业务进行感知与分析，可以充分利用sFlow、NetFlow等基于报文随机采样的网络流量监测控制技术。这些技术可以实时完整地提供全网范围的数据链路层、网络层和传输层的网络流量信息，进而对网络流量进行实时的分析与分类，从而与网络控制平面形成联动关系，然后再根据业务需求实时改变网络拓扑，在数据面实现相应流量的高效汇聚以及转发。

同时，在数据平面拓扑构造方面，通过电交换设备和点到点光链路构成DCN基础拓扑，使其具备基本的连通性，再通过光交换矩阵连接必要的节点，如接入层机架顶端（ToR）或汇聚层ToR构成可重构的高速互连拓扑。

3.3 低延时或确定性延时控制技术

为了满足光电混合网络对动态业务实时调度的要求，需要极大提升现有网络控制平面的时效性，包括有效降低控制软件的响应时间及其抖动，降低控制消息传递时延及其抖动。传统网络控制系统（如软件定义网络控制器）响应时间随网络负载差异较大，业务响应时延基本保持在百毫秒到秒级；控制消息传递的时延与抖动也无法有效控制。如果实时网络控制系统的时延抖动过大，会引起网络协议振荡，最终导致网络稳定性变差。为了提升控制效率，软件加速技术、国际互联网工程任务组（IETF）（DetNet）[15]和IEEE 802.1时间敏感网络（TSN）等



▲图2 基于人工智能的光电混合网络重构架构

确定性网络低延时传输技术、控制系统与收发节点的高精度时间同步技术都将是提升控制系统时效性的关键手段。

4 结束语

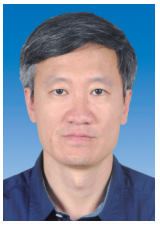
随着移动互联网业务的迅猛发展与普遍接入，用户使用各种互联网服务的行为产生了大量的数据。以5G为代表的通信网络的快速推广使得更高速的数据传输成为可能，而数据中心作为存储、处理和分析这些数据的重要基础设施，其节点算力逐渐增强，规模逐渐增大，要求数据中心互连网络具备提供高带宽、低能效、可应对突发数据的承载能力。结合不同维度的光交换技术的光电混合数据中心互连网络将成为提升目前数据中心带宽调度灵活性的关键技术手段，该技术亟待在流量分析、智能与高效控制多方面取得突破。

参考文献

[1] LEE G B, DUPUIS N, PEPELIJUGOSKI P, et al. Photonic Switch Fabrics in Computer Communications Systems[J]. Journal of Lightwave Technology, 2015, 33(4): 768–777
[2] ROY A, ZENG H, BAGGA J, et al. Inside the Social Network's (Datacenter) Network[C]// SIGCOMM 2015. USA: ACM, 2015
[3] ADDAPT, ADDAPT Project[EB/OL].[2019–07–22]. http://www.addapt–fp7.eu/
[4] BENSON T, AKELLA A, MALTZ D. Network Traffic Characteristics of Data Centers in the Wild[C]//Internet Measurement Conference 2010. USA: ACM, 2010
[5] Disaggregated Data Centers: Great Idea, But Not Just Yet[EB/OL].(2015–07–21)[2019–07–22].http://www.datacenterdynamics.com/servers–storage/disaggregated–data–centers–great–idea–but–not–just–yet/94473.fullarticle
[6] YIN X, KERREBROUCK V J, GOUDYZER G, et al. Multi–Level High Speed Burst–Mode Receivers[C]//Optoelectronics & Communications Conference 2016. Japan: OECC, 2016
[7] RYLYAKOV A, PROESEL J, RYLOV S, et al. A 25 Gb/s Burst Mode Receiver for Rapidly Reconfigurable Optical Networks[C]//2015 IEEE International Solid–State Circuits Conference– (ISSCC) Digest of Technical Papers. USA: IEEE, 2015:400–401. DOI: 10.1109/ISSCC.2015.7063095
[8] OZKAYA I, CEVRERO A, FRANCESE A P, et al. A 56 Gb/s Burst–Mode NRZ Optical Receiver with 6.8ns Power–On and CDR–Lock Time for Adaptive Optical Links in 14nm FinFET CMOS [C]//IEEE International Solid–State Circuits Conference. USA: IEEE, 2018:16. DOI: 10.1109/ISSCC.2018.8310286
[9] CEVRERO A, OZKAYA I, MORF T, et al. 4x40 Gb/s 2 pJ/bit Optical RX with 8ns Power–on and CDR–Lock Time in 14nm CMOS[EB/OL]. [2019–07–22]. https://www.osapublishing.org/abstract.cfm?uri=OFC–2018–M2D.3
[10] Cisco. Forecast and Methodology, 2016–2021 White Paper, 2018[EB/OL].(2018–11–19)[2019–07–22].https://www.cisco.com/c/en/us/solutions/collateral/service–provider/global–cloud–index–gci/white–paper–c11–738085.html
[11] BENSON T, AKELLA A, MALTZ D A. Network Traffic Characteristics of Data Centers in the Wild[C]//Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. USA:ACM, 2010
[12] KREUTZ D, RAMOS F M, VERISSIMO P, et al. Software–Defined Networking: A Comprehensive Survey[J]. Proceedings of the IEEE, 2015, 103(1):14–76. DOI: 10.1109/JPROC.2014.2371999
[13] GREENBERG A G, HAMILTON J R, JAIN N, et al. VL2: A Scalable and Flexible Data Center Network[J]. ACM SIGCOMM Computer Communication Review, 2011,39 (4): 51. DOI:10.1145/1594977.1592576
[14] KANDULA S, SENGUPTA S, GREENBERG A, et al. The Nature of Data Center Traffic: Measurements & Analysis[C]//Proceedings of the 9th ACM SIGCOMM Conference on Internet measurement, 2009: 202. DOI: 10.1145/1644893.1644918
[15] Deterministic Networking Architecture[EB/OL].[2019–07–22]. https://tools.ietf.org/id/draft–ietf–detnet–architecture–04.html#rfc.section.4.5

作 者 简 介

郭秉礼，北京邮电大学光电信息学院副教授、硕士生导师；研究方向为数据中心与高性能计算中的光互连网络技术、天地一体化网络控制技术；先后主持和参加国家自然基金、国家重点研发等项目20余项；发表SCI/EI论文100余篇。



黄善国，北京邮电大学教授、博士生导师、理学院执行院长；研究方向为智能光网络与多维光交换技术、光网络规划与优化技术，以及高频光控波束形成；先后主持并参与10余项国家和省部级科研课题，包括国家自然科学基金、国家“863”计划、国家“973”计划合作课题等；发表SCI/EI检索论文150余篇，国际特邀报告7次，获得授权国家发明专利20项、IETF国际标准建议6篇，出版专著2部。