专题：智算光互联

大规模智算中心光电交换网络架构演化综述

叶通，胡卫生
（上海交通大学，上海 200240）

摘　要：随着智算中心规模向百万卡级演进，以“数据中心光互联（data center optical interconnection，DCI）+电分组交换（electrical packet switching，EPS）”为特征的传统智算中心网络面临功耗高、时延高、可靠性不足的挑战。近几年工业界开始探索引入光子技术的方案，以降低智算中心网络的功耗并增强其扩展性、灵活性和可靠性。回顾了工业界研究的“DCI+EPS+光线路交换（optical circuit switching，OCS）”和“DCI+光分组交换（fast optical switching，FOS）”两类智算中心网络架构。结合工业界头部企业的实际案例及科研机构的相关探索，探讨了两种架构的技术路径、性能优势及待研究问题，为未来智算中心网络的设计提供参考。

关键词：智算中心；光电交换网络；算力集群
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025116

Overview of large-scale optical-electrical switching networks for artificial intelligent data center (AIDC)

YE Tong, HU Weisheng
Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: With the explosive growth in the scale of artificial intelligence data centers (AIDC), traditional AIDC networks characterized by “data center optical interconnection (DCI) + electrical packet switching (EPS)” are increasingly challenged in terms of power consumption, communication latency, and reliability. To address this issue, photonic technologies have been introduced in recent years to reduce the power consumption and enhance the scalability, flexibility, and reliability of AIDCs. Two types of network architectures—“DCI + EPS + optical circuit switching (OCS)” and “DCI + fast optical switching (FOS)”—that had been studied were reviewed. Combining the practices of leading enterprises and academic institutions, the technical pathways, performance advantages, and issues yet to be studied of these proposals were discussed. Insights were provided to guide the design of future large-scale AIDC networks.

Key words: artificial intelligence data center, optical-electrical switching network, computing power cluster

收稿日期：2025−03−18；**修回日期：**2025−04−09
通信作者：叶通，yetong@sjtu.edu.cn
基金项目：国家自然科学基金资助项目（No.62271306，No.62331017）
Foundation Items: The National Natural Science Foundation of China (No.62271306, No.62331017)

0　引言

自 2025 年年初 DeepSeek 轰动全球以来，人工智能（artificial intelligence，AI）大模型进入开源应用和迅猛发展的新阶段。缩放准则（scaling law）作为大模型发展的核心原理，揭示了其性能随参数量、数据集和算力规模幂率增长的规律[1]。大规模算力的支撑对大模型能力进化起到至关重要的作用。

为抢占 AI 技术发展先机，全球主要大国正积极推动大规模智算中心的建设。所谓智算中心是指以图形处理单元（graphics processing unit，GPU）为算力单元的数据中心。2024 年，美国政府设立智算中心基础设施工作组[2]，美国科技巨头则相继建成或规划了超大规模智算集群，如 xAI 的 10 万卡智算中心、Meta 的 35 万卡智算集群、微软联合 OpenAI 提出的“星际之门”百万卡集群计划。2025 年 1 月美国政府宣布投入 5 000 亿美元，将“星际之门”提升为美国 AI 基础设施建设的重要项目。我国则从 2021 年开始出台政策推动建设进程[3]，已建成十余个万卡以上的智算中心，如 2024 年 8 月投入使用的中国移动（哈尔滨）的 1.8 万卡智算中心、中国电信上海临港即将建成的 10 万卡（规划 30 万卡）智算中心等。2024 年 1 月工业和信息化部等七部门联合发文提出“建设超大规模智算中心，满足大模型迭代训练和应用推理需求”[4]，同年 12 月国内行业联盟开放数据中心委员会（ODCC）提出聚焦百万卡集群的 Mega Scale Out 项目。目前，智算中心规模正从十万卡迅速向百万卡演进。

区别于传统数据中心，智算中心流量具备新的特征。在大规模智算中心，大模型训练通过智算中心网络分配给大量 GPU 并行处理，GPU 则定期通过网络进行数据同步，其产生的 AI 流量呈现出“三超”特征。

（1）超高速率，每对 GPU 通信量大，通信速率将达 Tbit/s 以上。

（2）超低时延，参训 GPU 的通信同步发起且要求尽快同步完成，任一通信时延增大都会减慢整个训练进度。

（3）超高可靠，若任一参训 GPU 或网络设备发生故障且无法快速恢复，将导致训练中断重启，造成损失。

目前商用智算中心采用由数据中心光互联（data center optical interconnection，DCI）和电分组交换（electrical packet switching，EPS）构成的拓扑静态的“DCI+EPS”网络，多方面限制了智算中心的规模扩展。现有智算中心网络利用 EPS 细粒度交换能力和数据中心光互联技术将大量 GPU 整合成算力池提供训练服务。现有架构仍然是应用主力，但在大规模扩展时会遇到以下困难。

（1）功耗高，EPS 能力提升伴随功耗增长，且带来逐跳光电光转换，进一步增加功耗[5]。

（2）时延高，AI 流跨越 EPS 次数（电跳数）多，流量均衡和拥塞控制复杂，维持稳定时延十分困难[6]。

（3）可靠性不足，EPS 间以及 EPS 和服务器间的连接关系固定，故障恢复时间长[7]。例如，故障可使十万卡级集群可用率低于 60%[8]。

为满足智算中心规模快速增长的需求，对传统智算中心网络架构进行突破势在必行。与此同时，多种光子器件已经成熟，为网络架构的变革提供了新的选项。2024 年 9 月，华为公司推出了基于 3D 微电子机械系统（micro-electro mechanical system，MEMS）技术的 256 端口光线路交换（optical circuit switching，OCS）设备 OptiXtrans DC808。已经商用化的光子器件还包括半导体光放大器（semiconductor optical amplifier，SOA）、阵列波导光栅路由器（array waveguide grating router，AWGR）、快速可调激光器等。在此背景下，工业界已开始研究如何将光子技术应用于智算中心网络，增强网络扩展性。

目前，按照所使用的光子器件和对智算中心EPS的替代程度来分，已经提出的架构大体可以分为基于大端口OCS的“DCI+EPS+OCS”网络[9]和基于快速光交换（fast optical switching，FOS）的“DCI+FOS”网络。事实上，已提出的智算中心光网络架构与通用云数据中心和超算中心的网络架构有继承关系，由于篇幅所限，本文聚焦于回顾智算中心光网络架构的演进过程。

1 “DCI+EPS+OCS”网络架构

基于3D MEMS技术的商用化OCS端口数可达320[10]。大端口OCS具备多个优势。（1）插入损耗（插损）低，在3 dB以内。（2）能以较低功耗提供大容量的交换能力。（3）对速率透明，一次部署即可支持端口速率平滑升级[11]。因此大端口OCS技术引起工业界的广泛关注。然而，OCS重构速度慢，交换状态切换时间在10 ms量级，频繁重构会带来较大的带宽浪费和系统负担。故而OCS可以提供带宽大但是数量少的光路，对细粒度但是连接性要求高的通信需求并不友好。因此，如何降低其灵活性不足带来的限制，是应用过程中需要考虑的问题。

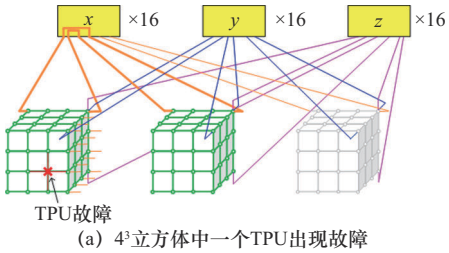
1.1 工业界动态

谷歌在2023年ACM ISCA会议上报道了TPUv4[7]。TPUv4由面向传统超算中心网络的3D-Torus架构演化而来。在这个架构中，64颗张量处理器（tensor processing unit，TPU）芯片通过一系列6×6 EPS和短电缆互联成4^3立方体，立方体表面的TPU则与128端口的OCS相连。通过重构OCS，可按需将多个4^3立方体拼接成各种形状的大规模长方体，不仅增加了组网灵活度，还可以绕过含有故障TPU的4^3立方体来提升系统可用性。TPUv4可以很好地支持包括All-Reduce和小规模All-to-All通信在内的流量。谷歌的测试结果显示，虽然单颗TPU芯片算力只有英伟达A100 GPU的88%，但是凭借“DCI+EPS+OCS”网络架构赋能的算力集群，其整体运算能力可以比基于传统电交换网络的英伟达算力集群快1.2~1.7倍，能耗效率则是1.3~1.9倍。目前谷歌已经将TPUv4部署于实际系统中。

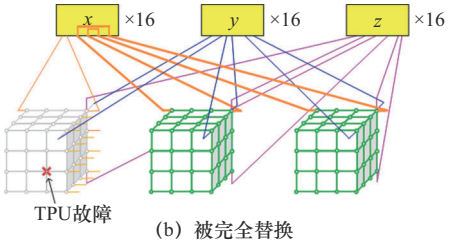
TPUv4资源调配示意图如图1所示，受OCS端口数限制，TPUv4资源调配粒度粗，图1（a）4^3立方体中一个TPU出现故障，图1（b）则被完全替换。基于TPUv4，还有如下两个问题值得思考。

（1）如图1所示，资源调配的最小粒度为4^3立方体，单个TPU失效即替换整个4^3立方体，进一步提升算效或许还有空间。

（2）该架构对连接性要求高的大规模All-to-All通信支持能力还有待研究和测试。谷歌采用Apollo架构[9]对TPUv4进行规模扩展，其中由一层OCS和一层EPS构成叶脊（Leaf-Spine）结构，每个EPS下面连接一个TPUv4。



（a）4^3立方体中一个TPU出现故障

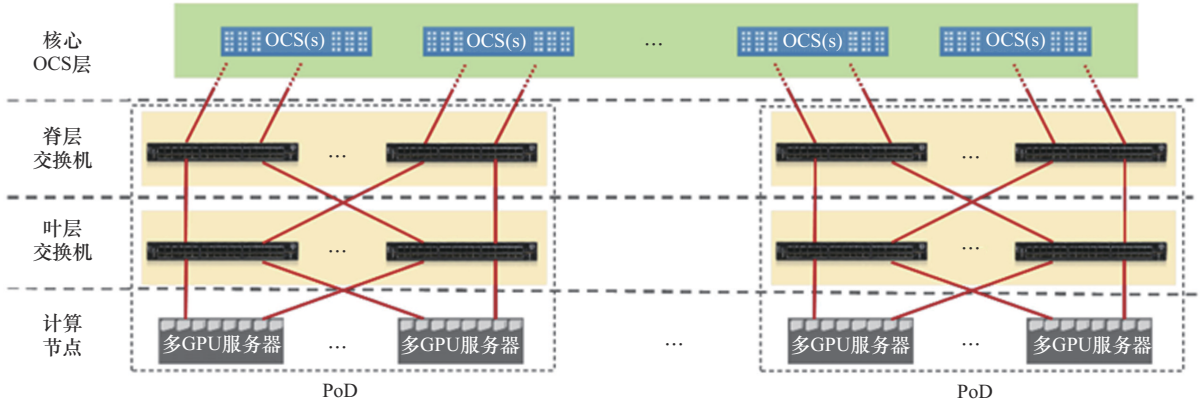


（b）被完全替换

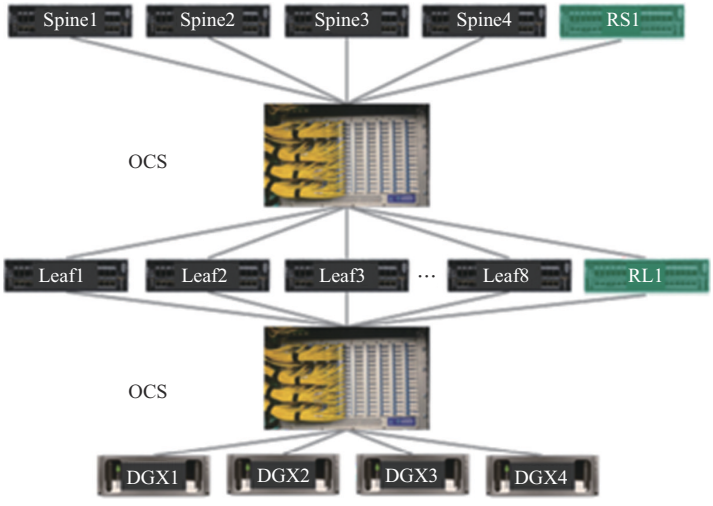
图1 TPUv4[7]资源调配示意图

谷歌的探索激发了工业界对“DCI+EPS+OCS”架构的兴趣，英伟达和百度等巨头开始投入研究。不同于谷歌的TPUv4，英伟达和百度从传统通用数据中心网络出发，或用OCS来替换叶脊结构的一层或多层EPS，或在各层EPS间加入一层OCS。

2025年英伟达在谷歌Jupiter架构[11]的基础上，提出在各电交换层之间加入若干OCS增强可靠性[12]。英伟达光电交换网络架构如图2所示，首先，用大端口OCS替换核心层EPS（如图2（a）），



（a）OCS替换核心层EPS



（b）EPS交换层引入OCS增强网络鲁棒性

图2　英伟达光电交换网络架构[12]

核心层OCS和PoD（每个PoD为一个千卡级GPU集群）之间采用叶脊结构，即二者之间构成全互联拓扑，从而得到类似于谷歌的Jupiter架构；然后，再在脊（Spine）层EPS和叶（Leaf）层EPS之间以及叶层EPS和GPU服务器之间加上一层OCS（如图2（b））。核心层OCS负责在PoD之间建立可重构的光拓扑，用来处理经过两层EPS汇聚且相对平稳的PoD间流量。EPS层间的OCS则用于故障恢复。当某个EPS出现故障时，通过重构层间OCS，将相邻的EPS或GPU连接到备用EPS，从而在物理层直接提供保护倒换的能力，可实现快速的故障恢复。相比传统的“DCI+EPS”架构，该架构可以缩小电跳数，将能耗降低41%[11]，并提升了可靠性。然而，架构仅能在PoD间提供可重构拓扑，PoD内仍然是“DCI+EPS”静态网络，存在资源调度不够精细的问题。EPS层间的OCS仅用于故障恢复，重构优势未充分利用。此外，文献[12]并未给出层间OCS的具体实施方式。

文献[12]同时还提出LEAN网络架构。英伟达LEAN架构如图3所示，该架构用一层大端口OCS替换传统胖树（fat-tree）核心交换层和汇聚交换层，OCS和柜顶交换机（top of rank，ToR）之间采用分轨的方式进行互联，即将ToR分为若干组并进行编号，每个OCS仅与不同组编号相同的ToR相连。同时，ToR和GPU也采用分轨连接方式。LEAN的优点是交换网络层次较少，可以

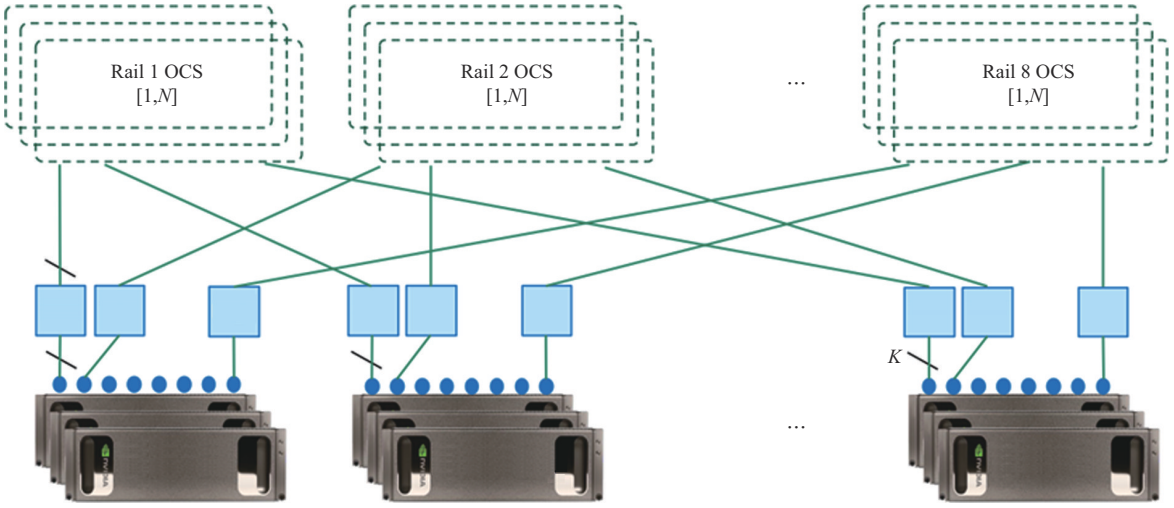


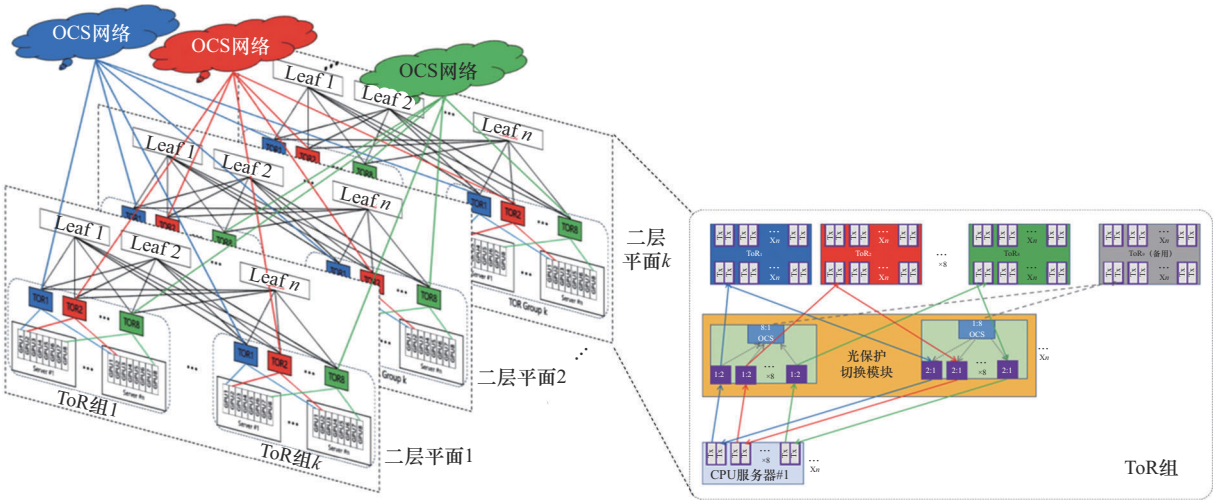
图3　英伟达LEAN架构[12]

在不同的ToR之间建立可重构拓扑。现有ToR可以连接32个GPU，因此资源调度的粒度比图3所示架构更加精细。按照标准服务器包含8个GPU，LEAN可连接65 536个GPU。由于OCS慢重构的限制，该架构的不足在于分轨拓扑对All-to-All通信的支持能力不足，这个情况在多租户的情况下更显著[13]。虽然文献[12]展示了针对128个GPU的All-to-All通信的配置例子，但是对于更大规模的All-to-All通信的支持能力还有待深入研究。

为了支持All-to-All通信以及多租户场景，2024年百度提出的架构增加了不同轨ToR之间的互联[13]。百度智算中心光电交换网络架构如图4所示。如图4（a）所示，百度所提架构与LEAN的不同之处在于，8个ToR编为一组，若干组ToR合在一起构成一个平面，然后在一个平面的ToR上加一层与其全连接的Leaf EPS。目前，关于这种架构是否能很好地支持All-to-All通信业务还有待进一步研究。百度所提架构可支持的算卡数量与LEAN一致。文献[13]还给出了一种在ToR和算卡之间采用小规模OCS实施保护倒换的方案，如图4（b）所示。

1.2　学术界动态

与此同时，学术界也做出了积极的探索，其



（a）百度基于OCS的智算中心网络架构　　（b）ToR组内连接情况

图4　百度智算中心光电交换网络架构[13]

中包括来自麻省理工学院（MIT）、西安电子科技大学和香港科技大学的研究团队。区别于工业界所提架构与通用数据中心和超算中心网络架构有较强的继承和延续性，学术界则侧重探索更加新颖的架构。

2023 年 MIT 联合 Meta 提出了 TopoOpt 架构[14]。在此架构中，GPU 服务器和大端口 OCS 互连为两层 Leaf-Spine 结构。该架构可以通过把多台服务器放置在同一个 ToR 下面，然后让 OCS 和 ToR 互连为两层 Leaf-Spine 结构进行规模扩展。按照标准服务器包含 8 个 GPU 计，TopoOpt 可连接 65 536 个 GPU。TopoOpt 采用 One-Shot 的工作方式，即在一个新训练任务到达时，根据任务需求配置 OCS，为任务分割出一个计算资源子网。由于存在 OCS 粒度大但连接性差的问题，文献[14]提出利用 GPU 进行流量中转的思路，其仿真和实验结果表明，该架构可以很好地支持 All-Reduce 等流量，能显著降低通信开销、缩短迭代周期。然而，当存在大规模 All-to-All 流量时，流量通常需要经历多次 GPU 中继，不仅容易造成网络拥堵，还会增加逐跳转发的时延。

同年，西安电子科技大学研究团队在光纤通信会议和展览（OFC）提出了 X-Next+网络架构[15]。该架构把所有 ToR 划分成若干个 PoD。每个 PoD 内部的 ToR 分成两个集合，并通过低速电缆连接成二分电网络，同时通过高速光链路和一系列 OCS 连接成 Leaf-Spine 结构。在 PoD 间，同编号的 ToR 集合和跨 PoD 的 OCS 连接成 Leaf-Spine 结构，其中不同集合内的同编号 ToR 均与同一组 OCS 连接。PoD 内和 PoD 间的 OCS 根据整网的流量矩阵预设若干个交换状态，不同交换状态有不同的时长权重，这些 OCS 按权重周期重构依次执行预设的交换状态。此外，一旦探测网络流量矩阵有显著变化，网络控制器则更新这些预设的交换状态。X-Next+的优点在于可以适配全网流量矩阵的变化，不足之处在于 OCS 交换状态重构会影响训练任务的数据同步。

香港科技大学团队则针对混合专家（mixture of experts，MoE）并行带来的 All-to-All 流量提出 mFabric 架构[16]。mFabric 架构如图 5 所示。该项工作主要基于 MoE 模型 All-to-All 通信的两个特征：（1）参与同一个 All-to-All 通信操作的 GPU 数量通常不是很多；（2）在每个训练迭代中，只有少量 GPU 之间有较大的流量，其余 GPU 之间的流量较小。文献[18]在英伟达分轨优化网络的基础上，将若干台服务器划分为一个高带宽域，然后引入小规模 OCS 为每个域内的服务器提供灵活可变的光互联。在每个训练迭代的过程中，利用模型训练的时间，将 OCS 按照 All-to-All 流量分布重构成不同的交换状态，从而在数据同步阶段为流量大的 GPU 对提供带宽。结果表明，mFabric 在获得和 fat-tree、分轨优化网络相似的时延性能的情况下，可以大幅降低网络成本。

总体来说，基于慢速大端口 OCS 的“DCI+

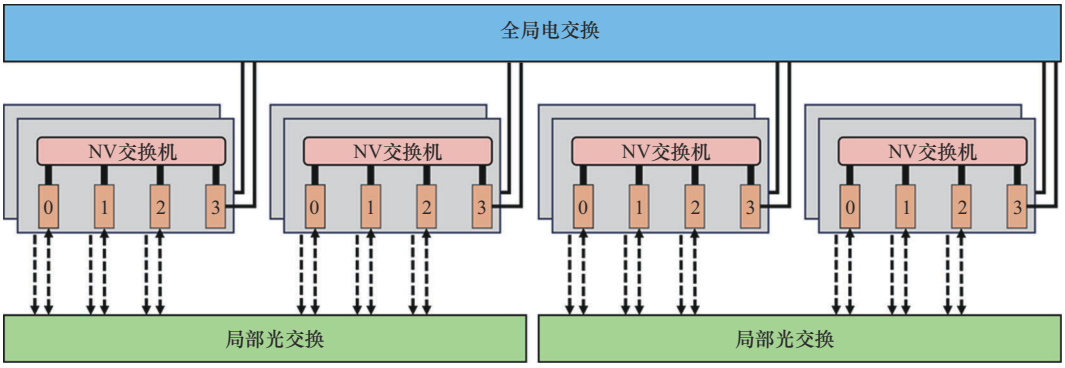


图 5　mFabric 架构[16]

EPS+OCS”架构可以降低网络的功耗和成本，对All-Reduce等流量的支持度也很好，可以降低数据同步的开销，加快模型的训练速度。然而，由于OCS慢重构的限制，该类架构对大规模All-to-All的支持还存在不足。

2 “DCI+FOS”网络架构

FOS技术的研究始于20世纪90年代，通过超快光学实现纳秒或亚纳秒级重构能力，执行细粒度的光交换。如果将重构速度推向亚纳秒级，FOS技术则可执行光分组交换（optical packet switching，OPS）。大体来说，FOS可以分为基于半导体光放大器（semiconductor optical amplifier，SOA）和光耦合器（optical coupler，OC）组合的广播选择结构[17]，基于可调波长变换器（或快速可调激光器）和AWGR组合的交换结构[10, 18]以及基于马赫-曾德尔光开关阵列的Benes网络[19-20]等。FOS虽然可以执行分组级细粒度交换，但其插损、功耗、光同步等方面的问题还值得研究[21]。

近两年也有研究组提出“DCI+FOS”网络架构，即用FOS完全替代所有的EPS。“DCI+FOS”架构的优势在于网络扁平化，通信时延低，但在插损、功耗、成本和实现技术难度方面还有提升空间。按动态可调的程度，此类架构可进一步分为全动态、半动态和全静态3类。

2.1 全动态网络

全动态网络是指，网络中几乎每个光子元件都进行动态重构，以决定源目的GPU之间的交换状态。

2024年英国伦敦大学学院基于大端口光耦合器、SOA和AWGR，提出一种纳秒级快速可重构的光交换架构RAMP [22]。基于光耦合器、SOA和AWGR组合的RAMP架构如图6所示，包含x个通信组，每个通信组包含J个机架，每个机架包含Λ个节点，每个节点配备x个可调光发射模块组，每组包含b个光发射机。每个可调光发射模块组连接一个$1\times x$光耦合器，耦合器的每个输出端口连接一个SOA，每个SOA连接到不同的中间级子网（subnet）。$1\times x$光耦合器上的x个SOA的开关状态决定光发射模块组发出的光信号连接到哪个中间级子网。一个中间子网两侧分别与一个特

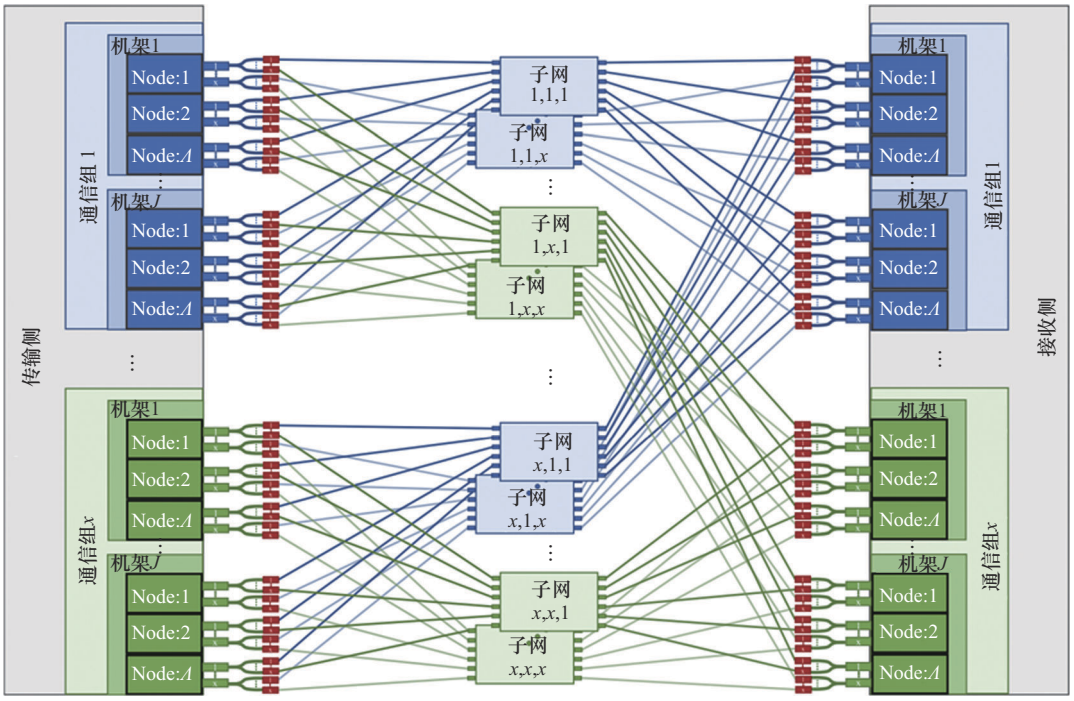


图6 基于光耦合器、SOA和AWGR组合的RAMP架构[22]

定的通信组相连。中间级子网可以由耦合器和SOA构成的空分广播选择网络构成，也可以由AWGR和SOA构成，目的是决定将输入的信号送往特定通信组的哪一个节点。因此，源节点通过本地SOA选择中间级子网来选择去往哪个目的通信组，相应的中间级子网则最终确定目的通信组的目的节点。该架构可支持超过4 096个GPU，且可以根据流量分布的改变快速重构GPU间的连接关系以支持All-to-All通信。该架构不足之处在于，多个地方使用光耦合器导致端到端的插损高，且需要大量使用能耗高的SOA器件。

同年，日本KDDI研究所采用多级超快2×2光开关网络和阵列波导光栅（array waveguide grating，AWG），提出一种纳秒级快速可重构的光交换网络架构Modoru[23]。MD-DD-WSS结构和Modoru网络架构如图7所示，Modoru的核心部件是KDDI提出的一种基于AWG和纳秒级2×2光开关网络复合而成的多维度双向波长选择开关（MD-DD-WSS）。MD-DD-WSS的两端分别是用于波长复用和解复用的AWG，中间级由若干2×2电控法拉第旋转器晶体光开关构成的可重构无阻塞Clos交换网络构成。在Modoru中，每个GPU通过一个16×64 MD-DD-WSS与核心层交换矩阵相连，每个交换矩阵均为一个2×2光开关构成的可重构无阻塞Clos交换网络。该架构可支持65 536张GPU，但插损也相对较高，可达到19 dB。此外，大规模光开关矩阵的协同切换控制以及多级交换网络的重构算法无疑增加了网络运行的负担。如何降低网络运行开销值得深入探讨。

2.2　半动态网络

半动态是光网络的物理拓扑不频繁变化或者固定，仅通过动态调节多波长器件，例如快速可调谐光源或者快速可调滤波器（如硅基微环），改变网络的逻辑拓扑，从而降低网络控制运行的复杂性。

2021年苏州大学团队提出一种基于快速可调激光器和波长选择开关（wavelength selective switch，WSS）的Mesh-Torus网络[24]。基于WSS的2D Torus网络和WSS在网络中的连接关系如图8所示，每个WSS均为(4+p)×(4+p) WSS，其中p个输入/输出端口和本地服务器配备的可调光发

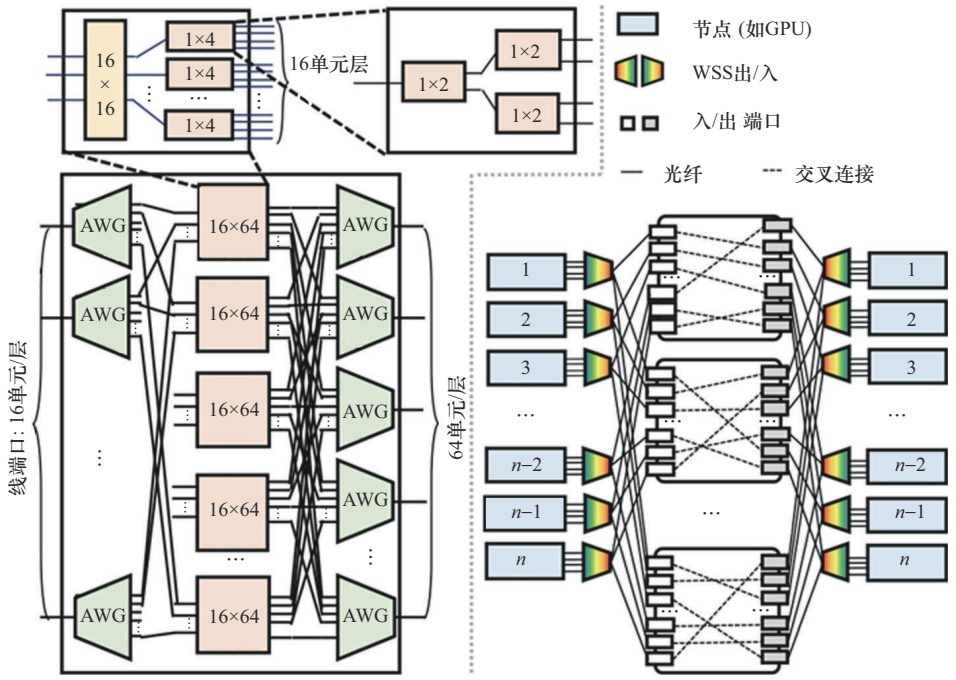
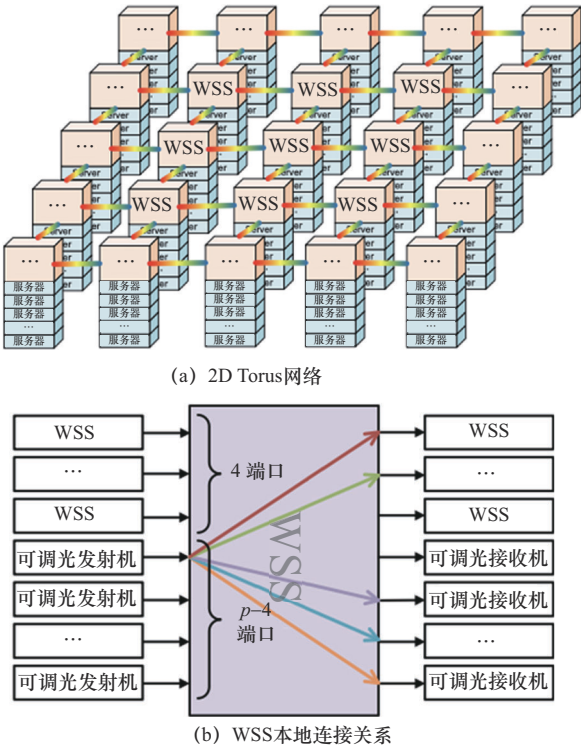


图7　MD-DD-WSS结构和Modoru网络架构[23]

射机/接收机相连，而剩余的4个端口和其他WSS相连构成2D Torus。所有WSS的交换状态提前配置好并固定不动。当有新的训练任务时，指派给该任务的服务器通过调谐可调光发射机来决定与其他相关服务器的连接关系，从而定义适合的逻辑拓扑。逻辑拓扑的优化目标是最小化训练完成时间，期间考虑每条光路经过的WSS个数，以防止光路损耗过大。然而，当网络规模扩大时，光路跨越的WSS数量将不可避免地增加，损耗将成为制约该结构扩张的因素。

基于类似原理，中国科学院计算所提出了ODDL架构[25]。中国科学院计算所ODDL架构如图9所示，ODDL将GPU节点与WSS互联成与BCube类似的网络拓扑。该架构在新的训练任务到达的时候配置好WSS交换网络，然后在训练迭代周期内，将可调激光器按照各个数据同步步骤内的通信需求快速切换到不同波长即可。即在整个训练期间，在不重构WSS交换状态的前提下完成网络逻辑拓扑的重构。该架构可以支持4 096个GPU。类似前面的方案，该架构存在插损大的问题，规模进一步扩展的能力也依赖于商用化WSS模块的规模。



(a) 2D Torus网络

(b) WSS本地连接关系

图8 基于WSS的2D Torus网络和WSS在网络中的连接关系[24]

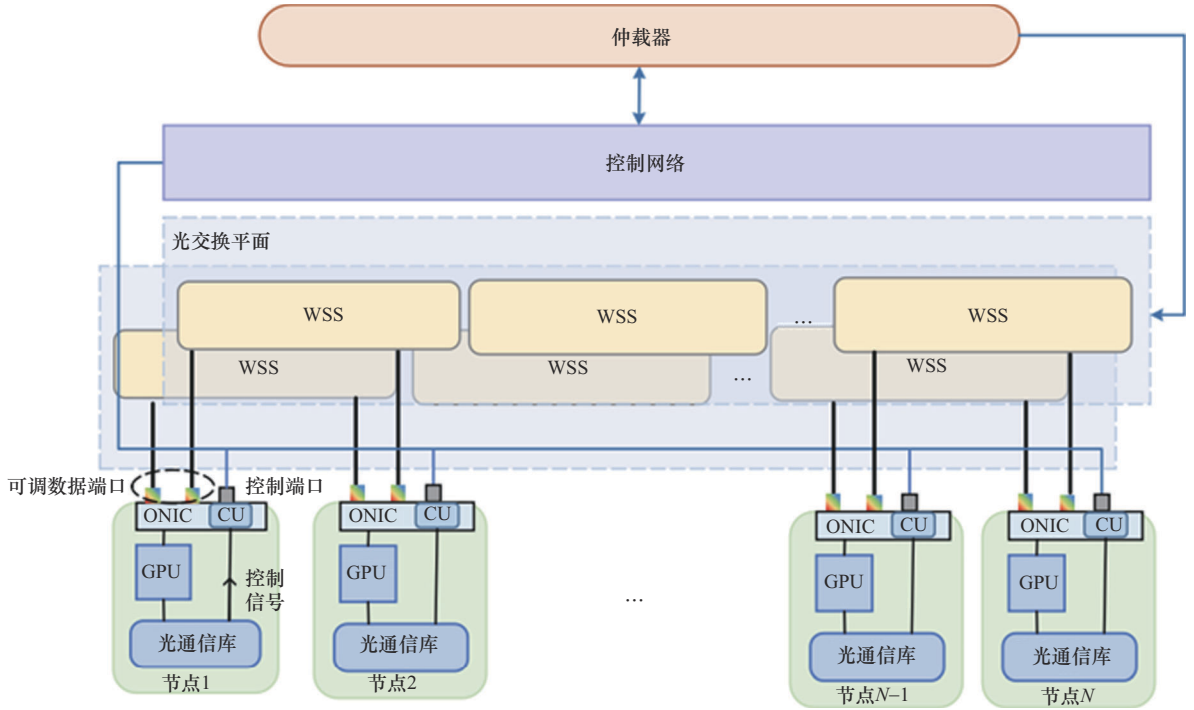


图9 中国科学院计算所ODDL架构[25]

另一项属于此类的工作是哥伦比亚大学团队提出的基于硅基微环的 SiP-Ring 架构[26]。SiP-Ring 架构如图 10 所示，若干个 GPU 和一组微环阵列构成一个节点，多个节点互联成环状拓扑。在一个节点中，通过调谐微环阵列将其他节点发往本地的波长下路，并将本地波长上路到环网中发往其他节点。显然，单个环网所能携带的 GPU 数量受限于可用波长数。因此，本文提及采用二维环进行网络规模的拓展。同样，当光路跨越的节点数增多时，损耗将变大。因此，SiP-Ring 在每个节点上都增加了一个光放大器补偿微环阵列以及光交织器带来的光损耗，这增加了系统成本。

2.3　全静态网络

为避免使用成本高昂的快速可调激光器，2025 年北京邮电大学团队进一步提出将 AWGR 互联网络[27-28]和固定波长激光器阵列相搭配的网络架构[29]。基于 AWGR 和固定波长激光器阵列的网络架构如图 11 所示，N^2 个 $k\times k$ AWGR 构成一个单层多波长互联网络，为 Nk 个 ToR 之间提供扁平化的、固定的一跳全连接。该架构需要的光模块数为 $(Nk)^2$，随着 ToR 数量 Nk 平方增长，连纤数量为 $2N^2k$。此外，端到端光路需要经过 2 个波长复用器和 1 个 AWGR，插损也比较高。严格来讲，该网络只在所有节点间提供固定的全连接，

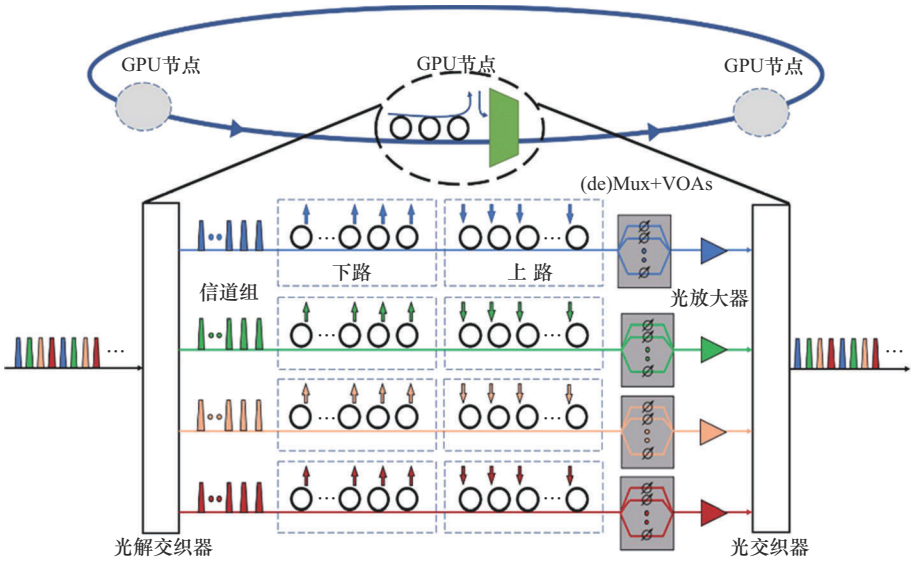


图 10　SiP-Ring 架构[26]

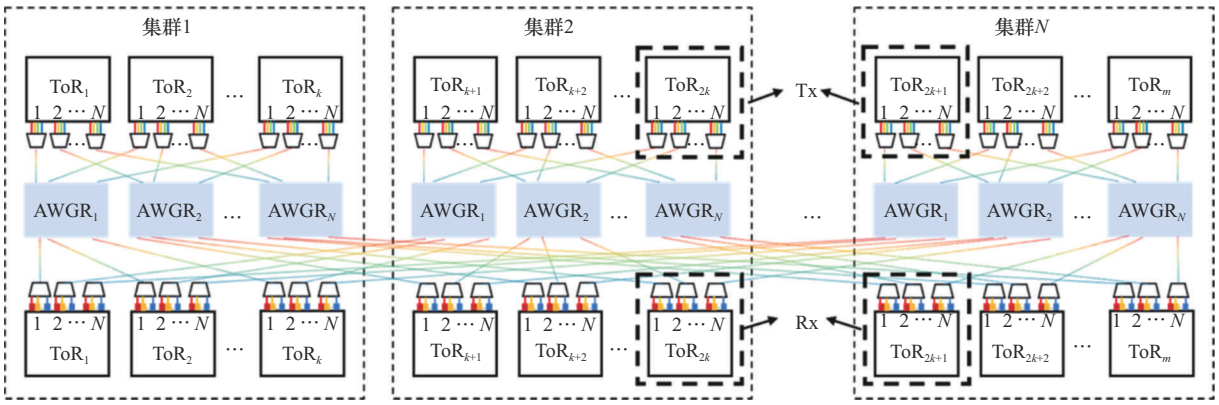


图 11　基于 AWGR 和固定波长激光器阵列的网络架构[29]

已经不属于FOS范畴。

总体来说，相比“DCI+EPS+OCS”架构，“DCI+FOS”架构的优点在于重构速度快，可以提供细粒度的交换能力，响应高突发业务。可以继续探讨的问题包括如何降低光插损和功耗以及与之相关的系统扩展性问题。

3 结束语

针对智算中心光电交换网络的研究主要集中在近几年，处于起步阶段，设计适配AI流量特征的大规模智算中心光电交换网络架构是一个值得深入探讨的问题。通过对提出的典型架构进行回顾和分析，初步得出以下结论。“DCI+EPS+OCS”架构通过OCS与EPS的协同，可以显著降低电跳数与功耗，并提升系统可靠性，但存在资源调度粒度较粗的问题，在高效支持大规模All-to-All通信流量方面仍面临挑战。相比之下，“DCI+FOS”架构可以提供细粒度的交换能力，对突发业务的响应速度快，但在插损、功耗以及扩展性方面还需优化。

参考文献：

[1] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[EB]. 2020.

[2] 中国算力大会. 中国智算中心服务发展报告[R]. 2024.
China Computational Power Conference. China artificial intelligence data center service development report[R]. 2024.

[3] 中国电信. 智算产业发展研究报告(2024)[R]. 2024.
China Telecom. Artificial intelligence data industry development research report (2024)[R]. 2024.

[4] 工业和信息化部, 等. 工业和信息化部等七部门关于推动未来产业创新发展的实施意见[R]. 2024.
MIIT, et al. Implementation opinions of seven departments including MIIT on promoting the innovation and development of future industries[R]. 2024.

[5] CHOPRA R. Looking beyond 400 G: a syetem vendor perspective[R]. 2020.

[6] 百度. 智算中心网络架构白皮书[R]. 2023.
Baidu. Artificial intelligence data center network architecture white paper[R]. 2023.

[7] JOUPPI N, KURIAN G, LI S, et al. TPU v4: an optically reconfigurable supercomputer for machine learning with hardware support for embeddings[C]//Proceedings of the 50th Annual International Symposium on Computer Architecture. New York: ACM Press, 2023: 1-14.

[8] 华为. 迈向智能世界白皮书2024—数据通信[R]. 2024.
HUAWEI. Stepping into the smart world white paper 2024-data communication[R]. 2024.

[9] LIU H, URATA R, YASUMURA K, et al. Lightwave fabrics: at-scale optical circuit switching for datacenter and machine learning systems[C]//Proceedings of the 2024 IEEE 37th International Conference on Micro Electro Mechanical Systems (MEMS). Piscataway: IEEE Press, 2024: 156-161.

[10] SATO K I, HASEGAWA H, NIWA T, et al. A large-scale wavelength routing optical switch for data center networks[J]. IEEE Communications Magazine, 2013, 51(9): 46-52.

[11] POUTIEVSKI L, MASHAYEKHI O, ONG J, et al. Jupiter evolving: transforming Google’s datacenter network via optical circuit switches and software-defined networking[C]//Proceedings of the ACM SIGCOMM 2022 Conference. New York: ACM Press, 2022: 66-85.

[12] PATRONAS G, TERZENIDIS N, KASHINKUNTI P, et al. Optical switching for data centers and advanced computing systems[J]. Journal of Optical Communications and Networking, 2025, 17(1): A87-A95.

[13] 朱宸, 周谞, 王佩龙. 可重构OCS技术在大模型预训练中的应用(特邀)[J]. 光通信研究, 2024(5): 29-38.
ZHU C, ZHOU X, WANG P L. Application of reconfigurable OCS technology for pre-training large language models[J]. Study on Optical Communications, 2024(5): 29-38.

[14] WANG W, et al. TopoOpt: co-optimizing network topology and parallelization strategy for distributed training jobs[C]//Proceedings of the 21st USENIX Symposium on Networked Systems Design and Implementation. New York: ACM Press, 2023: 739-767.

[15] GU H X, YU X S, LU Y F, et al. X-NEST+: a high bandwidth and reconfigurable optical interconnects for distributed machine learning and high-performance computing[C]//Proceedings of the 2023 Optical Fiber Communications Conference and Exhibition (OFC). Piscataway: IEEE Press, 2023: 1-3.

[16] LIAO X D, SUN Y J, TIAN H, et al. mFabric: an efficient and scalable fabric for mixture-of-experts training[EB]. 2025.

[17] YANG Y, WANG J. Designing WDM optical interconnects with full connectivity by using limited wavelength conversion[C]//Proceedings of the 18th International Parallel and Distributed Processing Symposium, Piscataway: IEEE Press, 2004: 35.

[18] YE T, LEE T T, HU W S. AWG-based non-blocking Clos networks[J]. IEEE/ACM Transactions on Networking, 2015, 23(2): 491-504.

[19] QIAO L, TANG W J, CHU T. 32 × 32 silicon electro-optic switch with built-in monitors and balanced-status units[J]. Scientific Reports, 2017, 7: 42306.

[20] JIANG J, GOODWILL D J, DUMAIS P, et al. 16 × 16 silicon photonic switch with nanosecond switch time and low-crosstalk architecture[C]//Proceedings of the 45th European Conference on Optical Communication (ECOC 2019). London: IET, 2019: 1-4.

[21] TUCKER R S. Optical fiber telecommunications V B[M]. Amsterdam: Elsevier, 2008: 695-737.

[22] OTTINO A, BENJAMIN J, ZERVAS G. RAMP: a flat nanosecond optical network and MPI operations for distributed deep learning systems[J]. Optical Switching and Networking, 2024, 51: 100761.

[23] WANG C, YOSHIKANE N, ELSON D, et al. Modoru: Clos nanosecond optical switching for distributed deep training[J]. Journal of Optical Communications and Networking, 2024, 16(1): A40-A52.

[24] LIN J M, SHEN G X, ZHAI Z W, et al. Delivering distributed machine learning services in all-optical datacenter networks with torus topology[C]//Proceedings of the 2021 Asia Communications and Photonics Conference (ACP). Piscataway: IEEE Press, 2021: 1-3.

[25] LI W Z, YUAN G J, WANG Z, et al. Fast and scalable all-optical network architecture for distributed deep learning[J]. Journal of Optical Communications and Networking, 2024, 16(3): 342-357.

[26] KHANI M, GHOBADI M, ALIZADEH M, et al. SiP-ML: high-bandwidth optical network interconnects for machine learning training[C]//Proceedings of the 2021 ACM SIGCOMM 2021 Conference. New York: ACM Press, 2021: 657-675.

[27] YE T, LEE T T, GE M, et al. Modular AWG-based interconnection for large-scale data center networks[J]. IEEE Transactions on Cloud Computing, 2018, 6(3): 785-799.

[28] YIN Y W, PROIETTI R, NITTA C J, et al. AWGR-based all-to-all optical interconnects using limited number of wavelengths[C]//Proceedings of the 2013 Optical Interconnects Conference. Piscataway: IEEE Press, 2013: 47-48.

[29] GUO Y Z, XUE X W, GUO B L, et al. AWGR-based all-optical switching network for distributed machine learning[J]. Optics Express, 2025, 33(1): 829-841.

[作者简介]

叶通（1976– ），男，博士，上海交通大学副教授，主要研究方向为数据中心光网络、光交换与光互联架构、光网络性能分析。

胡卫生（1964– ），男，博士，上海交通大学教授，主要研究方向为光交换结构、全光通信网等。