doi:10.13756/j.gtxyj.2024.240062.
专题:数据中心内光交换
唐雄燕,魏步征,沈世奎,等.智算数据中心光电交换技术综述[J].光通信研究,2024(5):240062.
Tang X Y, Wei B Z, Shen S K, et al. Overview of Optoelectronic Switching Technology in Intelligent Computing Data Centers [J]. Study on Optical Communications, 2024(5):240062.

智算数据中心光电交换技术综述(特邀)

唐雄燕,魏步征,沈世奎,王创业,王泽林,张 贺,王光全,张晨芳
(中国联通研究院,北京 100048)

摘要:近年来,由于人工智能现象级应用的出现,智算数据中心(AIDC)和超算数据中心(SCDC)网络演进成为研究热点。传统的 3 层全连接电交换架构已经不能很好地满足 AIDC 或 SCDC 高带宽、低时延、低功耗和低成本的需求,需要一种更为高效和易于扩容的电光融合交换或光交换方案来逐步替代传统纯电路包交换的策略。文章对近年来基于数据中心的光电融合交换技术方案进行了综述性介绍,对行业中率先进行光交换技术尝试的头部互联网企业的部署经验进行了分析,并结合电信行业需求,给出了 AIDC 和 SCDC 交换节点的推荐部署和演进方式。

关键词:智算数据中心交换架构;光电融合交换;光电路交换
中图分类号:TN29 **文献标志码**:A

Overview of Optoelectronic Switching Technology in Artificial Intelligent Data Centers

TANG Xiongyan, WEI Buzheng, SHEN Shikui, WANG Chuangye,
WANG Zelin, ZHANG He, WANG Guangquan, ZHANG Chenfang
(China Unicom Research Institute, Beijing 100048, China)

Abstract: In recent years, due to the emergence of phenomenal applications of artificial intelligence, the evolution of Artificial Intelligent Data Center (AIDC) and Super Computing Data Center (SCDC) network has become a hot research topic. The traditional three-layer fully connected electrical switching architecture can not well meet the requirement of high bandwidth, low delay, low power consumption and low cost of the AIDC or SCDC, which requires a more efficient and easy expansion way of the electrical and optical fusion switching or optical switching scheme to gradually replace the traditional pure circuit package switching strategy. In this paper, the photoelectric fusion switch technology solutions based on data center in recent years are summarized. The deployment experience of leading Internet enterprises that take the lead in optical switch technology is analyzed. Combined with the needs of the telecom industry, the recommended deployment and evolution strategy of the switch nodes of AIDC and SCDC and super computing center are given.

Key words: AIDC switching architecture; optoelectronic convergence switching; optical circuit switching

0 引 言

2023 年 12 月 29 日,国家发展改革委、国家数据局、中央网信办、工业和信息化部以及国家能源局联合印发了《深入实施“东数西算”工程 加快构建全国一体化算力网的实施意见》,从通用算力、智能算力和超级算力一体化布局,东中西部算力一体化协同,算力与数据、算法一体化应用,算力与绿色电力一体化融合,算力发展与安全保障一体化推进等 5 个方面统筹出发,推动建设联网调度、普惠易用和绿色安全的全国一体化算力网。传统数据中心组网主要需求包括[1-6]:

① 高带宽和低延迟:数据中心组网需要提供高带宽和低延迟的网络连接,以满足大规模数据传输和实时应用的需求,包括快速的数据存取、实时数据备份和恢复以及远程访问等。

② 冗余和高可靠性:数据中心组网需要具备冗余和高可靠性,以确保数据中心的连续性和可用性,包括冗余网络链路、备份设备和冗余电源等,保证其可以应对网络故障、硬件故障或电力故障等情况。

③ 可扩展性:数据中心组网需要具备可扩展性,以适应不断增长的数据量和业务需求,包括快速部署新设备、添加新网络节点和调整网络容量等,以满足不断变化的需求。

收稿日期:2024-03-20; 修回日期:2024-04-10; 纸质出版日期:2024-10-10
基金项目:国家重点研发计划资助项目(2022YFB2804102)
作者简介:唐雄燕(1967—),男,湖南永州人。教授级高工,博士,主要研究方向为光通信。
通信作者:魏步征(1991—),高级工程师,博士。E-mail:weibz5@chinaunicom.cn
© Editorial Office of *Study on Optical Communications*. This is an open access article under the CC BY-NC-ND license.

④ 虚拟化和软件定义：数据中心组网通常采用虚拟化和软件定义网络（Software Defined Network，SDN）等技术，提供灵活性、管理简化和资源优化的特性，包括虚拟网络、虚拟机迁移和网络切片等功能，以提高网络资源利用率和性能。

随着由数十亿到数万亿个参数组成的大模型（Large-Scale Model，LSM）的出现，需要极其庞大的参数量和计算能力支撑[7]。LSM的出现主要得益于计算硬件的不断发展和可用性的提高以及对更复杂任务和更大规模数据的需求[8-9]。人工智能计算（Artificial Intelligent Computing，AIC）利用人工智能和机器学习技术来提升计算系统的智能化和自动化能力[10]。AIC在各个领域取得了显著的进展，并在人工智能应用中发挥了重要作用。

在LSM/AIC的背景下，数据中心组网还面临一些新的挑战：

① 高带宽和低延迟：智算数据中心（Artificial Intelligent Data Center，AIDC）之间的通信带宽呈现层级式递增趋势，接入交换机单端口速率将从10/25 Gbit/s向25/100 Gbit/s过渡，汇聚交换机单端口速率将从40/100 Gbit/s向100/200 Gbit/s扩容，而数据中心间出口路由器则会持续向400/800 Gbit/s，甚至1.6 Tbit/s演进。通常为了满足实时数据传输和处理的要求，确保训练过程的高效性，需要端到端ms级时延，分解到端侧、交换节点的时延要求会进一步提升，目前主流交换技术的极限为ns级，结合缓存和调度，能够实现节点间ns级切换已属不易。

② 可扩展性和可重构性：LSM的深度学习模型和AIC的计算需求可能会发生较大的变化，以现在超算数据中心（Super Computing Data Center，SCDC）的基本配置来看，一个架顶（Top of Rack，ToR）交换机下挂10个服务器，每个服务器含有8个100 Gbit/s出口的图形处理器单元（Graphics Processing Unit，GPU），而一个机架集群至少含有上百个机柜，按100个计算，总出口容量将达到800 Tbit/s。此规模的节点需要数据中心组网具备弹性和可扩展性。网络架构和资源配置应能够根据需求进行自动调整和适应，以满足不同规模和复杂度的计算任务。同时还需要具备可重构性，即可根据需求重新配置网络拓扑和资源，以适应不同的应用场景，此灵活性可让数据中心快速适应不断变化的工作负载和需求。

③ 分布式计算：LSM深度学习模型的训练和推理通常需要分布式计算资源，涉及多台服务器的协同工作。数据中心网络需要支持分布式计算框架，使各个节点之间能够高效地进行通信和数据交换。

④ 网络虚拟化和软件定义智能管理：LSM和AIC对资源的灵活配置和利用率的最大化提出了要求，数据中心组网正朝着网络虚拟化和软件定义的方向发展，以实现更高的灵活性和可管理性。通过虚拟化网络功能和使用SDN技术，数据中心可以更加灵活地配置和管理网络资源，实现网络资源的动态分配和优化。

⑤ 能耗：随着数据中心规模的不断扩大，能源消耗的增长十分显著。2021年12月，国家发改委四部门发布新政提出，到2025年，全国新建大型、超大型数据中心平均能源使用效率（Power Usage Effectiveness，PUE）值降至1.3以下，国家枢纽节点还将进一步降至1.25以下。谷歌公司的数据中心机房PUE年平均值达到1.21，美国惠普的新一代数据中心机房夏季PUE值可以达到1.6～1.7。我国与世界先进数据中心机房的能耗水平还存在差距，新型数据中心组网需要考虑资源利用效率并降低能耗，以降低运营成本。

综上所述，在LSM和AIC背景下的数据中心组网面对新挑战和新需求，需要网络技术的发展和创新，以不断适应LSM计算和数据处理的要求。

1　光电融合交换网络基础架构

光电融合交换多应用在以脊－叶（Spine-Leaf）网络架构为主的数据中心中，脊（Spine）交换机具有高吞吐量、低延迟且端口密集的特点，其与每个叶（Leaf）交换机都有直接的高速（40～400 Gbit/s）连接。Leaf交换机与传统ToR交换机非常相似，其通常是24或48端口1、10或40 Gbit/s的接入层连接。面对成倍增长的通信容量，Leaf或Spine层交换机可以完全或部分由光交换机替代，同时在各层交换机上或统一调度层部署适应于光电融合交换的路由策略和交换机制，来满足光交换技术的波长交换特点。为了满足以上需求，国内外高校、科研机构和商业公司相继推出不同的解决方案，从不同的维度提升交换节点性能。

在光电融合交换方案中，光交换功能模块的主要方案分为光电路交换、光突发交换（Optical Burst Switching，OBS）和光包交换（Optical Packet Switching，OPS）3种。光电路交换的交换粒度较粗，导致包交换完成时间长，带宽利用率低，但实现复杂度

低，易于部署。相比之下，OPS 得益于较细的交换粒度，可以快速完成带宽利用率较高的业务流交换。然而，其实现的复杂性和较高的控制开销限制了其在未来 LSM 光交换 AIDC 的实际应用。作为一种折衷，OBS 的特点介于光电路交换和 OPS 技术之间。

2 基于光电路交换的方案

2.1 Helios 方案

Helios 是 2010 年前后由美国加利福尼亚大学研究团队提出的一种基于模块化数据中心混合光电交换机的体系架构，是最早提出光电融合交换的几种方案之一，其网络架构如图 1 所示[11]。

Helios 作为一个由节点（Point of Delivery，PoD）交换机和核心交换机（Core Switch，CS）组成的 2 级多根树，其 CS 由传统的电交换机和基于微机电系统（Micro-Electro Mechanical System，MEMS）的光电路交换机（Optical Circuit Switch，OCS）组成。每个 PoD 内部都有许多服务器通过铜缆连接到 PoD 交换机。PoD 交换机包含多个光收发器连接到核心交换阵列，其中一半的上行链路连接到电包交换机（Electrical Packet Switch，EPS），每个 EPS 也需要一个光收发器，另一半上行链路通过无源光波分复用（Wavelength Division Multiplexing，WDM）器连接到光交换机，称为超级链路。超级链路的容量大小受到 WDM 和波长数量的限制。PoD之间通过CS连接，电交换部分为PoD交换机

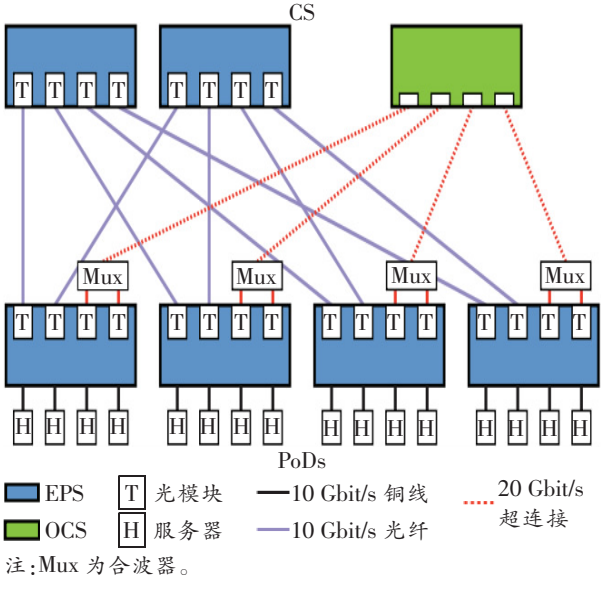


图 1 Helios 网络架构[11]

Figure 1 Helios network architecture[11]

间通信的突发部分提供全对全带宽，光交换部分负责处理需要高带宽且持续时间长的流量。

通过实验对比和评估 Helios 与电交换网络在传统多根树拓扑结构中的性能，结果表明，结合 WDM 收发器的 MEMS 光交换机给每个端口提供可扩展带宽的成本和功耗明显低于电交换机，例如一个含有 48 个 10 GE 端口的电交换机每端口耗电量 12.5 W（不含光模块功率），而 Helios 只有 240 mW。在 PoD 交换机间通信稳定的情况下，Helios 可以提供与非阻塞电交换相当的性能（除去内部信号处理时间，交换时延约 30 ms），且成本相比电分组交换降低 1/3，约四千万美元[11]。

但 Helios 方案也存在一定缺陷：一是端口规模有限，无法大规模组网；二是交换配置时间长，光带宽利用率低；三是 MEMS 插损较大。

2.2 C-Through 方案

莱斯大学、卡耐基梅隆大学和匹兹堡英特尔实验室的研究团队同期提出一种集成了光电路和分组交换的光/电混合网络架构，并通过构建一个名为 C-Through 的原型系统验证了该网络的可行性，其网络架构如图 2 所示[12]。

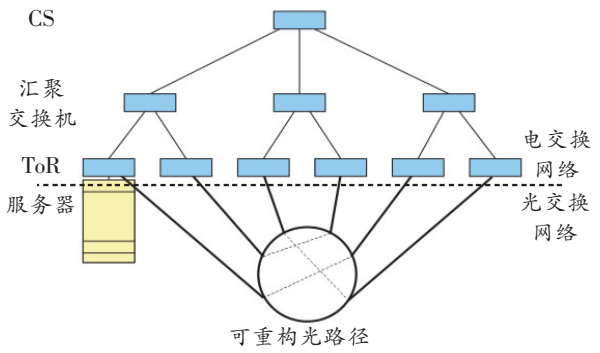


图 2 C-Through 网络架构[12]

Figure 2 C-Through network architecture[12]

该网络 ToR 的电交换网络使用传统的树状分层结构，底部的光交换网络通过 MEMS 光交换机与 ToR 交换机连接。由于每个机架在一个时刻最多只能有一条连接到其他机架的高带宽光链路，需要通过光网络的重新配置来匹配不同的机架，这一过程需要几 ms，在此期间快速路径是不可用的。因此为了确保对延迟敏感的应用程序能正常运行，保留了电交换网络，任何节点都可以在任意时刻通过分组交换链路与其他节点通信。

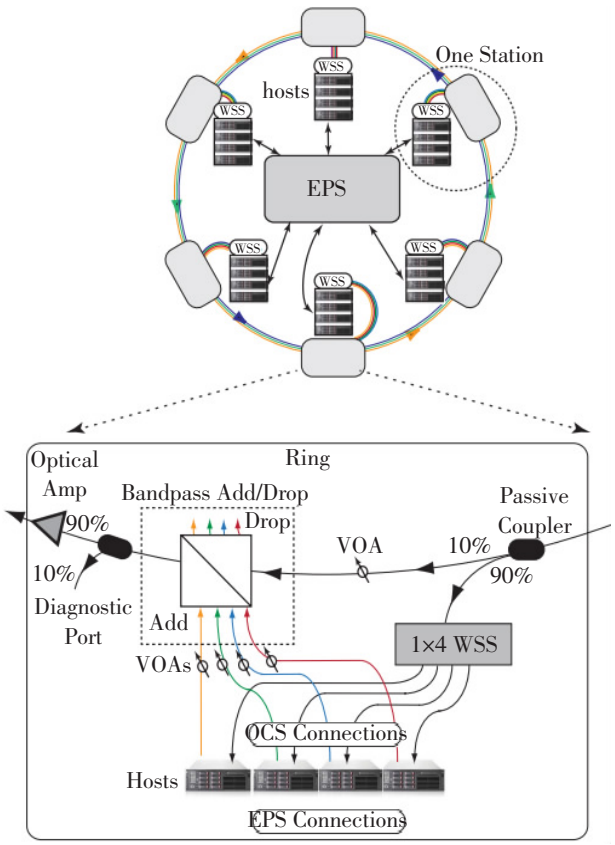
尽管 C-Through 能够在提供高通信带宽的同时保持网络的低复杂度，但在实际部署中却遇到了较多的挑战。主要原因在于数据中心的实际流量远

比初始设计架构时的假设流量复杂，需要从应用层、任务调度层和网络层多个维度解析应用的流量需求，目前还未有关于该问题解决方案具体配置细节的报道。

Helios 和 C-Through 的关键区别是，Helios 在交换机上实现流量估计和流量解复用功能。这种方法使流量控制对终端主机透明，但需要修改所有交换机。C-Through 的优点是，通过缓冲主机中的数据可批处理流量，并在光链路可用时有效地填充光链路。

2.3　Mordia 方案

Mordia 是 2013 年由加州大学圣地亚哥分校和谷歌的研究团队提出的光电路交换原型，其网络架构如图 3 所示[13]。



注：Optical Amp 为光放大器；Bandpass Add/Drop 为带通滤波器；VOA 为可调节光衰减器；Passive Coupler 为无源耦合器；WSS 为波长选择开关；Diagnostic Port 为诊断接口。

图 3　Mordia 网络架构[13]

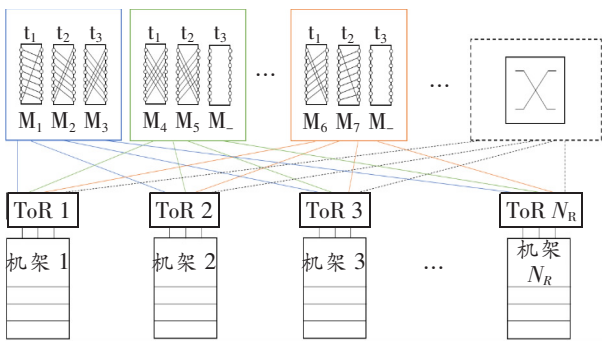
Figure 3　Mordia network architecture[13]

Mordia 光电路交换原型是一个 24 端口 OCS，由一个通过 6 个站点传输 $N(N=24)$ 个波长的环组成，利用 6 个 1×4 交换机和旁路端口构建单个24×24 端口交换机，支持输入到输出端口映射的任意重新配置，为了解决 MEMS 光交换机链路切换速度慢的问题，采用了具有 μs 级配置延迟的 WSS 交换机，切换时间为 11.5 μs。每个波长都是连接输入输出端口的单独通道，每个输入端口都被分配了特定的波长，该波长不被其他输入端口使用。输出端口可以调谐以接收环中的任何波长，并从任何输入端口传输数据包。每个源 ToR 交换机以自己的波长传输，每个站点将 4 个波长的子集转发给与其相连的 ToR，来自每个端口的流量在返回到源之前传输整个环。

Mordia 是对 Helios、C-Through 和光交叉结构（Optical Switching Architecture，OSA）等研究的补充，该结构采用的调度方式都是热点调度（HotSpot Scheduling，HSS）方法，这种被动策略通过测量机架间的流量需求矩阵来识别流量热点，重配置光交换机为流量矩阵热点建立光链路，从而最大化总体吞吐量。但是这种方法可能存在局部最优等问题，导致电路非饱和，在当前配置的持续时间内剩余容量反而被浪费掉。Mordia 采用的流量矩阵调度（Traffic Matrix Scheduling，TMS）主动调度方法克服了这一问题，提高了网络利用率。但是该结构的扩展性较差，考虑 WDM 最多只能扩展到 44 个端口，超过 88 个端口就需要扩展成多环，但是这种堆叠体系结构是阻塞的，不能实现任意的输入输出端口映射。虽然可以通过可调谐激光器引入新的自由度，但代价是额外的光学和算法复杂性。

2.4　RotorNet

RotorNet 是加州大学圣地亚哥分校的研究团队提出的一种可扩展的、基于 OCS 的低复杂度光数据中心网络，其互连架构如图 4 所示[14]。



注：N_R 为第 R 个 ToR；t_1~t_3 为交换矩阵；M_1~M_7 为匹配组；M_为备用匹配组。

图 4　RotorNet 互连架构[14]

Figure 4　RotorNet interconnect architecture[14]

RotorNet 采用传统的基于分组交换的 ToR 交换机实现与服务器的电互连，ToR 交换机之间则通过 Rotor 交换机实现光互连。每个 ToR 交换机连

接到一组 Rotor 交换机，可以在指定的时间间隔内提供任意一对 ToR 之间的直接连接。这种结构不需要通过重新配置光交换机来匹配网络流量，且允许完全分散的控制平面，可以最大化网络吞吐量。每个交换机都通过一组固定的静态配置轮循，这些配置在所有端点之间提供统一的带宽。这种设计消除了集中的控制平面，因为轮循交换机调度不需要需求估计、调度分配或全网同步。

RotorNet 作为一种运行开环切换调度的设计，将光电路交换机的控制与网络的其他部分解耦，极大地简化了网络的控制和部署，同时带来更高的可扩展性。通过模拟不同通信类型的实验表明，对于数据中心流量模式，RotorNet 提供的吞吐量是理想的电交换网络的 70%～95%，但是成本更低，与成本大致相同的胖树（Fat-Tree1）结构相比，RotorNet 在最坏情况下能提供 1.6 倍的吞吐量，在数据中心流量模式下提供 2.3 倍的吞吐量，在均匀流量下提供高达 3 倍的吞吐量。

3 基于 FSO 交换的组网方案

3.1 ProjecToR

ProjecToR 是 2016 年由微软研究院和亚利桑那大学研究团队提出的一种基于自由空间光（Free Space Optical，FSO）的可重构光互连架构，其架构如图 5 所示[15]。

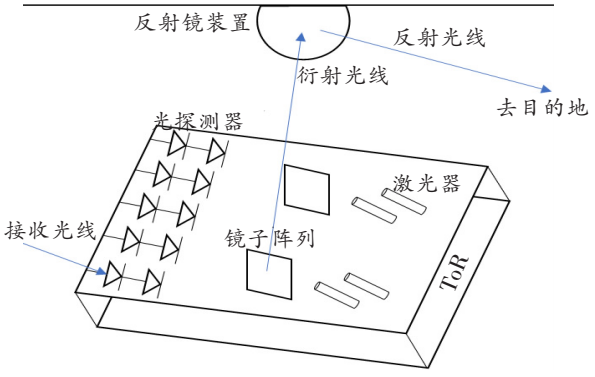


图 5 ProjecToR 互连架构[15]

Figure 5 ProjecToR interconnect architecture[15]

ProjecToR 使用数字微镜设备（Digital Micromirror Device，DMD）和镜面组件组合作为发射器实现高敏捷性和高扇形输出，使所有机架对建立直接链接，并能在 12 ms 内重新配置这些链接。DMD 包含成千上万个可以独立调整开关状态的微镜，可以通过对微镜的精确控制调整衍射光的方向，与基于 MEMS 的光交换机相比，DMD 的特性使其适用于超高端口数光交换机。但是由于 DMDs 的角度范围有限，限制了物理空间的覆盖范围。为了不使其扇出优势失效，在数据中心上方悬挂一个球状镜像组件，这个多面镜负责将源 ToR 交换机通过 DMD 衍射的激光反射到目的 ToR 交换机。

ProjecToR 采用光学装置的优势之一是，DMD 和镜面组件只引导光线，与传统的有线拓扑结构相比，无需改变互连中除了收发器之外的其他部分就可以扩展到更高的带宽。该结构支持更灵活的配置，具有建立非对称链路的能力，使吞吐量提高了 45%。原型的实验结果表明，基于 DMD 的 FSO 通信能提供与光纤和电缆相当的吞吐量，可以覆盖长距离并在接收器之间快速切换。大规模模拟表明，与全对分、电交换网络和 FireFly 相比，ProjecToR 可以将流完成时间（Flow Completion Times，FCT）提高 30%～95%，组件成本将比全对分网络便宜 25%～40%[16-17]。

3.2 FireFly

FireFly 是 2013 年前后由布鲁克大学和卡内基梅隆大学研究团队提出的一种基于 FSO 学的可重构无线数据中心结构，其架构如图 6 所示[18-19]。

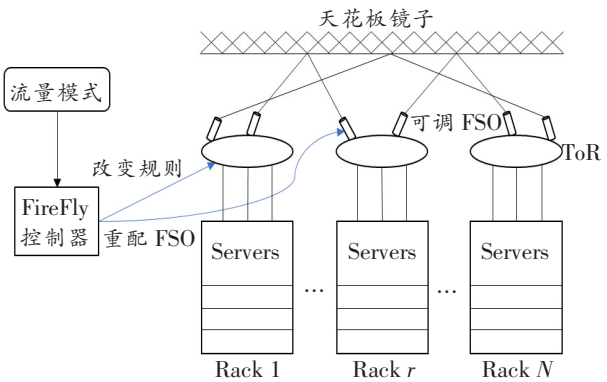


图 6 FireFly 光互连架构[18-19]

Figure 6 FireFly optical interconnect architecture[18-19]

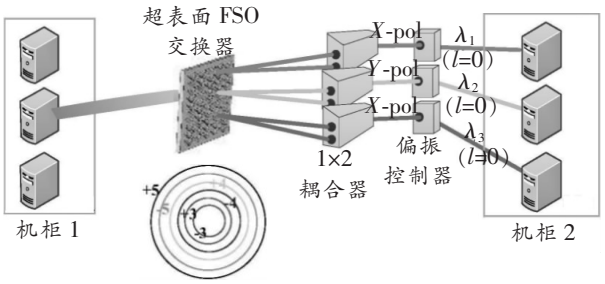
FireFly 使用传统的线缆进行机架内连接，每个 ToR 交换机都配备了可调的 FSO 设备，通过可重新配置的无线链路连接到其他 ToR 交换机。为了保证 FSO 设备之间不互相遮挡，利用机架上方的空间建立无阻碍光路，需要在天花板部署反射镜。FSO 系统的两个光纤端点直接与自由空间链路耦合，不需要经过任何光/电转换，因此节省了电力和成本。源 ToR 交换机通过光纤将激光束发送到自由空间，经过天花板镜面反射进入目标区域，激光束再被目的 ToR 交换机接收到光纤中。为了使激光束从光纤进入自由空间时的发散最小化，并使激光束在接收端点附近聚焦回光纤，在收发端设置了设

计合理的透镜来准直光束。

与静态有线拓扑相比，树形结构由于过载而性能不佳，全等分带宽设计的成本高且扩展性差，FireFly 更灵活且消除了布线成本，可扩展性更高。与光/电混合架构相比，FSO 的使用避免了这种光学设计所带来的布线复杂性，通过在每个机架上部署多个 FSO 设备，FireFly 可以创建更丰富的机架级拓扑。研究结果表明，FireFly 的总成本比 Fat-Tree 低 40%～60%，性能接近全等分带宽网络。

3.3　基于智能超表面的 FSO 交换方案

重庆邮电大学团队 2022 年前后提出了一种基于智能超表面的 FSO 交换方案，其架构如图 7 所示[20]。



注：X-pol、Y-pol 分别为 X 和 Y 方向偏振控制器；λ_1～λ_3 分别为光波长。

图 7　基于智能超表面的 FSO 交换架构[20]

Figure 7　FSO switching architecture based on intelligent metasurface[20]

该方案下的网络架构中包括 P 台服务器、M 个机柜、偏振控制器、1×2 耦合器和超表面 FSO 交换器，每个机柜内有 P/M 台服务器，每台服务器具有一个偏振控制器和一个 1×2 耦合器。为了提高服务器短距高速密集信息交换能力，实现高带宽传输速率和低网络能耗，该重构系统利用基于长短期记忆(Long Short Term Memory，LSTM)网络的流量预测模块、重构触发判决模块和拓扑生成模块及基于系统的重构方法能提升网络带宽，实现更高的网络资源利用率。

超表面作为一种支持几何相位调控的手性材料，能够在超小尺寸、自由空间以及可见光谱区内实现任意偏振态入射光束的轨道角动量(Orbital Angular Momentum，OAM)空分复用，还能实现多条波束并行相位调控和调向。因此，针对现有数据中心 FSO 互连系统网络单一维度复用的局限性，探索支持 WDM、偏振分复用(Polarization Division Multiplexing，PDM)、OAM 多维复用的、基于超表面的 FSO 交换，可以实现大容量数据中心内服务器间短距光互连。

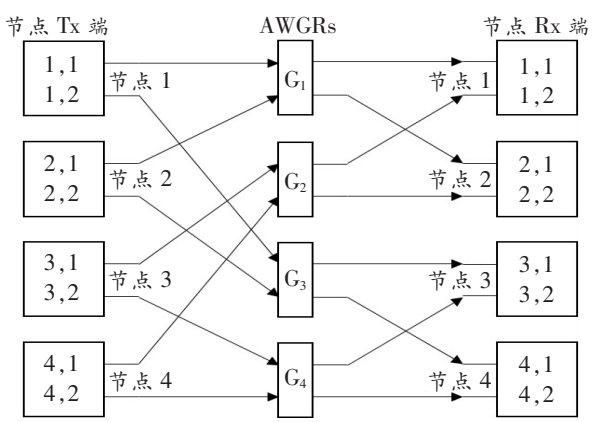
以上 3 种基于空间光的光电融合交换策略存在一个共性的缺点，即光路需要无遮挡。室内环境能最大程度减轻路径气候因素带来的不稳定性，但依然面临遮挡和持久工作带来的转向性部件老化问题，这是 FSO 的固有缺陷，需要设计冗余备份链路保障数据中心工作故障率。

4　动态可重构(光电路交换/OBS/OPS 混合)光电融合交换方案

4.1　Sirius

Sirius 是 2020 年由微软研究院研究团队提出的一种基于可调谐激光器的 ns 级粒度重构光交换网络，4 节点 Sirius 互连架构如图 8 所示[21]。

这个结构用可调谐激光器和无源光栅的组合作为高基数全光开关替换了电开关，它可以直接连接数千台服务器，形成一个高性能网络集群，或者通过一个扁平和无缓冲的核心连接所有机架交换机，从而扩展到整个数据中心。为各节点的收发器配备可调谐激光器作为物理层交换机，通过改变传输数据的波长来指示目的地地址。节点既可以是服务器，也可以是机架交换机。节点通过上行链路连接到由单层阵列波导光栅路由器(Arrayed Waveguide Grating Router，AWGR)或光栅组成的无源核心网。每个上行端口都配有包含可调谐激光器的收发器，并通过光纤连接到光栅，因此节点可以通过改变激光的波长将数据发送到相连光栅输出的所有其他节点。



注：Tx 和 Rx 分别为发送和接收；G_1～G_4 均为 AWGR。

图 8　4 节点 Sirius 互连架构[21]

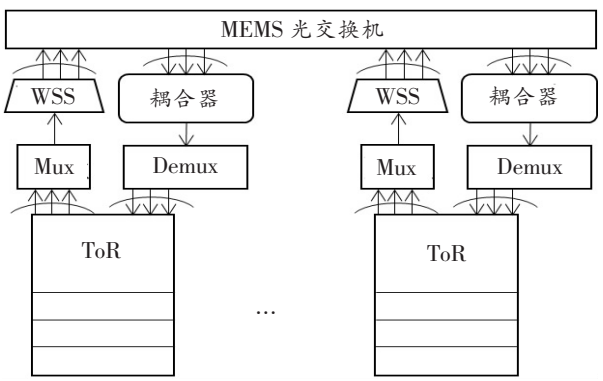
Figure 8　Four node Sirius interconnect architecture[21]

Sirius 提供了一个平坦的全光网络，提供了高效的无阻塞连接和超快速重构。首先，所有机架之间的带宽都是统一的，与理想的电交换网络的性能

非常接近，而功耗和成本显著降低，无源的网络核心不依赖于互补金属氧化物半导体（Complementary Metal Oxide Semiconductor，CMOS）组件，不需要跨代升级，具有良好的未来扩展潜力。通过基于机架的部署，Sirius 可以连接多达 25 600 个机架，这是当今大型数据中心规模的 6 倍。其次，因为允许在 ns 级的时间尺度上重新配置，其模拟了电交换网络的逐包交换，表明了用全光核心支持广泛工作负载的可行性。相比之下，以前的光交换架构交换的粒度从 μs 到 ms 不等，且对延迟敏感的工作负载通常依赖于单独的电交换网络。第三，通过消除光网络内部的缓冲并仔细管理节点本身的缓冲，可实现非常低且可预测的延迟。Sirius 的循环调度消除了调度平面，但是任何一对节点之间的流量都需要通过中间节点进行额外的跳转，可能对网络的吞吐量和延迟产生不利影响，负载均衡技术能保证任何流量模式下的最坏吞吐量不低于理想的非阻塞网络的 1/2。

4.2 Proteus

Proteus 是 2010 年由伊利诺伊大学厄巴纳－香槟分校和美国 NEC 实验室研究团队提出的一种具有拓扑可延展性的数据中心全光互连架构，其架构如图 9 所示[22]。



注：Demux 为分波器。

图 9 Proteus 光互连架构[22]

Figure 9 Proteus optical interconnect architecture[22]

Proteus 利用 MEMS 的可重构性实现动态拓扑结构以满足流量需求。将 $S=k$ 个 ToR 交换机分别连接到 MEMS 光交换机的 k 个端口，k 表示 ToR 交换机的度，则每个 ToR 交换机都可以同时与其他 ToR 交换机通信，通过 MEMS 的配置决定连接哪一组 ToR 交换机。对于未连接到 MEMS 的 ToR 交换机，可以从已连接的 k 个 ToR 交换机中选择一个作为中间节点转发数据。由于每个 ToR 交换机都有 k 度，将光纤的 WDM 与 WSS 相结合，ToR 交换机将多个端口输出的多波长信号通过多路复用器复用到一条光纤上，再通过 $1\times k$ 的 WSS 将这些波长拆分并送入 MEMS 光交换机。

相对于现有的解决方案，在数据中心网络的设计中，一个可分配网络和按需运行时拓扑可重构的组合是一个更灵活的结构。Proteus 避免了使用 ToR 交换机以外的电气设备，实现了高能效并简化了布线，可以在网络中节省 50%以上的电力。当服务器或 ToR 交换机发生变化时，光互连仍然可以保持不变，不需要重新布线，因此更容易升级和扩展。使用更多的 MEMS 和 WSS 端口，可以构建 ToR 交换机数量更多的拓扑结构，也可以进行异构互连。但由于 Proteus 使用 MEMS 交换机，因此对于业务的流量特点仍然有较高的要求。

4.3 OSA

光交换架构（Optical Switching Architecture，OSA）是 2014 年由 Chen K 研究团队提出的一种动态可重构的数据中心网络全光互连架构，如图 10 所示[23]。

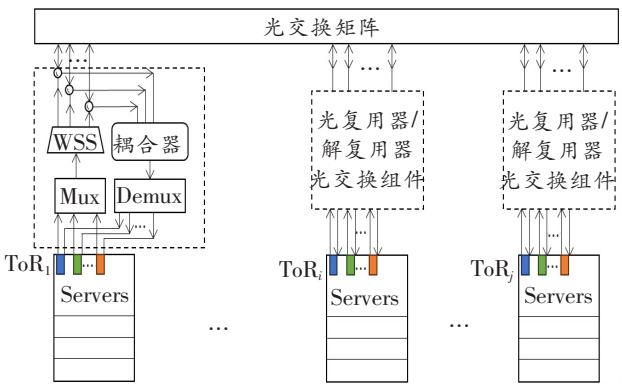


图 10 OSA 光互连架构[23]

Figure 10 OSA optical interconnect architecture[23]

OSA 可以动态改变其拓扑结构和链路容量，从而实现高度灵活性以适应动态流量模式。不同之处在于，OSA 采用多光链路逐跳拼接，为老鼠流和突发通信提供整体连接，并处理具有高扇入/出的工作负载，而现有的单跳电/光架构无法通过其光互连有效地解决这些问题。为了有效利用昂贵的光端口，OSA 还引入了循环器，这是一种能同时在两个方向上传输的双向功能组件，它能使光交换机端口的使用量增加一倍。

OSA 利用 MEMS 的可重构性来实现柔性拓扑，将 ToR 交换机连接到 MEMS 的每个端口，每个 ToR 交换机都可以同时与其他 ToR 交换机通信，MEMS 通过配置来决定连接哪一组 ToR 交换机。给定由 MEMS 光交换机连接的 ToR 交换机拓扑，

使用这种电路的逐跳拼接来实现网络范围的连接，为了到达没有直接连接的远程 ToR 交换机，需要选择一个已连接的 ToR 交换机作为第一跳，数据包经过光/电转换读取包头信息后，再将其路由到目的 ToR 交换机。ToR 交换机之间循环使用同一组波长，每个端口都在一个固定的波长发送和接收流量，以保证源 ToR 交换机的所有波长进行多路复用，并在解复用后发送到目标 ToR 交换机的各个端口。

在 OSA 中，一个 ToR 交换机可以同时连接到多个 ToR 交换机，任意一对远程 ToR 交换机之间通过逐跳电路拼接存在多跳连接。此外 OSA 还允许动态调整链路容量。与现有的混合架构不同，OSA 避免使用 ToR 交换机以外的电子元件。OSA 提供了比 Helios 或 C-Through 更大的灵活性，能够满足更大的非平均流量需求，性能与非阻塞网络相似。通过与抽象混合架构模型的粗略定量比较，表明 OSA 实现了更高的对分带宽。但当 ToR 交换机的通信对等点数量大于 4 个时，一些流必然会使用多跳路径，从而导致性能下降。

研究结果表明，OSA 可以提供高对分带宽，是非阻塞网络的 60%～100%，并且在真实和合成流量模式下都优于混合结构约 80%～250%。对于数据中心部署来说，OSA 是传统过载网络更好的选择。

4.4　Lotus

Lotus 是 2021 年由杜克大学和西安电子科技大学研究团队提出的一种应用于大规模分布式机器学习的新拓扑架构，如图 11 所示[24]。

为了提高机器学习的性能，通常需要使用参数梯度同步算法来使多个计算节点并行处理数据。典型的同步算法包括基于参数服务器的同步（Parameter Server Synchronization，PSS）、基于网格的同步（Mesh Synchronization，MS）和基于环的同步（Ring Synchronization，RS）。同步算法具有不同的通信特性，对网络架构提出了不同的要求，传统的数据中心网络很难满足这些需求，因此提出了一种基于 AWGR 的面向机器学习的光/电混合架构 Lotus。

Lotus 是一个多层网络，这些层分别对应于交换机、组和系统。每组交换机有两种类型：域内直连交换机（Direct Switch，DS）和域间交换机（InterDomain Switch，IS），DS 与计算节点、DS 与 IS 之间通过电链路连接，每个交换机组内使用完全二部图来提高对分带宽和可伸缩性。不同分组间的间接交换机通过光链路连接构成系统，光链路提供的高带宽可以快速传输组内节点产生的聚合流量。组之间的数据传输只能在一跳中完成，因为每个组与任何其他组之间都有链路。

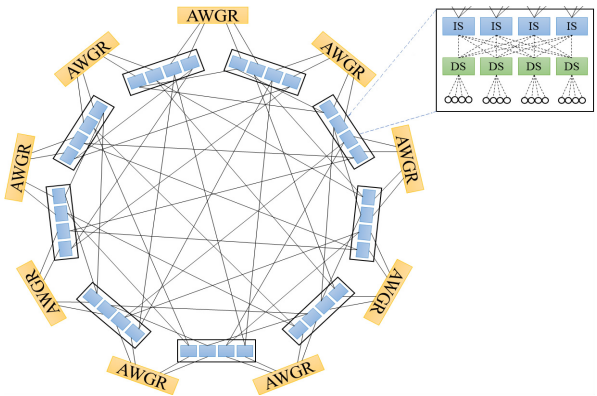


图 11　Lotus 互连架构[24]

Figure 11　Lotus interconnect architecture[24]

通过仿真模拟比较流量场景下不同网络的性能，结果表明，Lotus 比 Dragonfly 和 3D-Torus 减少了高达 50%的延迟。在平均吞吐量方面，Lotus 的性能分别比 3D-Torus 和 Dragonfly 高 5 倍和两倍。测试使用同步算法完成训练任务的性能，结果表明 3D-Torus 在 PSS 下消耗的时间最多，网络吞吐量也最低。Lotus 和 3D-Torus 在 RS 下的性能优于 Dragonfly。MS 下 Lotus 和 Dragonfly 的吞吐量高于 3D-Torus[25]。

由于 3D-Torus 中的交换机端口数量最少，因此其成本和功耗也最低，但随着规模的增大，网络直径限制了其扩展性。Lotus 和 Dragonfly 的网络直径不会随着计算节点数量的增加而增加，但需要更高的成本和功耗来构建网络。

4.5　ReSAW

由北京邮电大学和中国联通光交换团队于 2022 到 2023 年提出的 ps 级可重构光数据中心光交换系统（ReSAW）主要解决同机房内、不同机架间的流量调度任务[16-17]。与传统流量结构不同，基于 AIDC 和 SCDC 机房主要面对半数以上的东西向流量。这意味着同一个数据中心内，甚至同一个机房内的交互流量将面临巨大的交换压力。在这种情况下，北邮团队提出了使用具有流量分类功能的传统 ToR 交换机出彩光，配合 AWGR，将识别到的相同目的地的长报文通过光学路径转发，而类似控制面和短消息报文，则维持原来的电交换策略，该方案的拓扑示意图如图 12 所示。

假设机房内有 I 个机架，每个机架内含有 M 个服务器。如果报文目的地是同一个机架内的不同服务器，则通过 ToR 交换机直接进行架内转发。这里

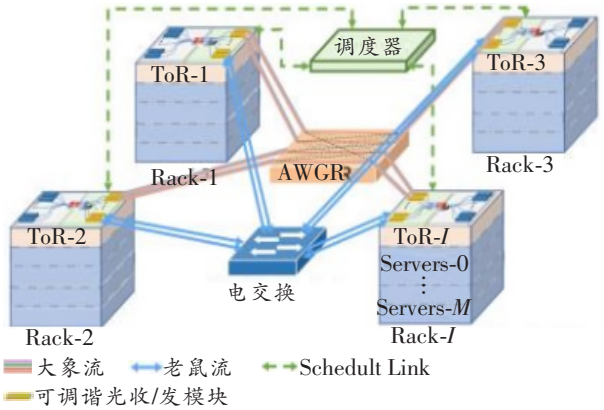


图 12 ReSAW 拓扑图[16-17]

Figure 12 ReSAW topology diagram[16-17]

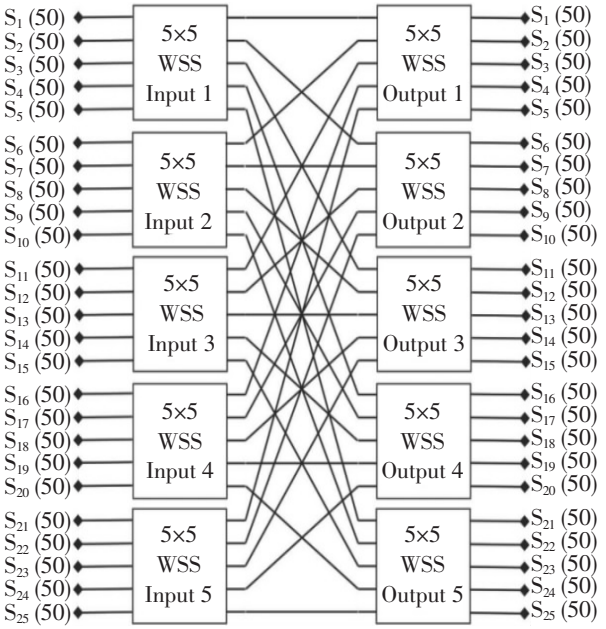
的 ToR 交换机具备流量目的地识别功能，可以将不同机架方向的流量分为所谓“大象流”和“老鼠流”，分别代指长包和短包。识别出的流量需要一个统一的调度器来进行端口波道和时隙配置。按照配置的 25 Gbit/s 端口速率，实现了一个时钟周期(2.56 ns)内 1.6 ns 的同步精度和极低的抖动(40.04 ps)。核心交换功能通过服务器产生随机流量进行测试，测试结果显示，当 ToR 交换机遇到老鼠流时，光口处于关断状态，AWGR 无流量经过；当某时刻出现大象流时，开启光通道交换，在 AWGR 中监测到去往不同机架的流量。

该方案具有与现网使用环境适配的优点，不需要大规模改造数据中心连接架构，同时可以对电交换机减配。但由于可调谐彩光模块的波长切换和稳定时间对时延影响巨大，所以需要快速可调谐光模块满足使用需求。

4.6 基于 Spanke 架构的全光交换数据中心方案

苏州大学光交换研究团队于 2022 年公开了面向基于 Spanke 架构全光交换数据中心的 Ring 业务部署方法，其中涉及光交换的内容[26]。从已经公开的文献中，仅可知该架构根据网络中服务器、WSS 模块的实时连接情况为业务建立光通道，能够灵活分配网络资源、有效降低部署所有业务所需总时间以及缩短业务平均等待时延，有效缓解了 Ring 业务部署过程中存在的光通道波长竞争以及资源动态分配的问题。

Spanke 架构拓扑如图 13 所示，全光交换网络包括输入级(Input)和输出级(Output)两个层级，每个层级为多个 WSS 模块组成的阵列，输入级中的每个 WSS 模块与输出级中的每个 WSS 模块通过唯一的链路连接，服务器通过光收发器与 WSS 模块建立光路连接；其中，所述服务器包括源节点和目的节点，源节点通过光收发器与输入级 WSS 模块建立光路连接，目的节点通过光收发器与输出级 WSS 模块建立光路连接。没有更多的数据对该架构性能做说明。



注：S_1~S_{25} 分别为 1~25 号交换机。

图 13 Spanke 架构拓扑图[26]

Figure 13 Spanke topology diagram[26]

5 光电融合交换的商用案例

5.1 谷歌的光电交换技术方案

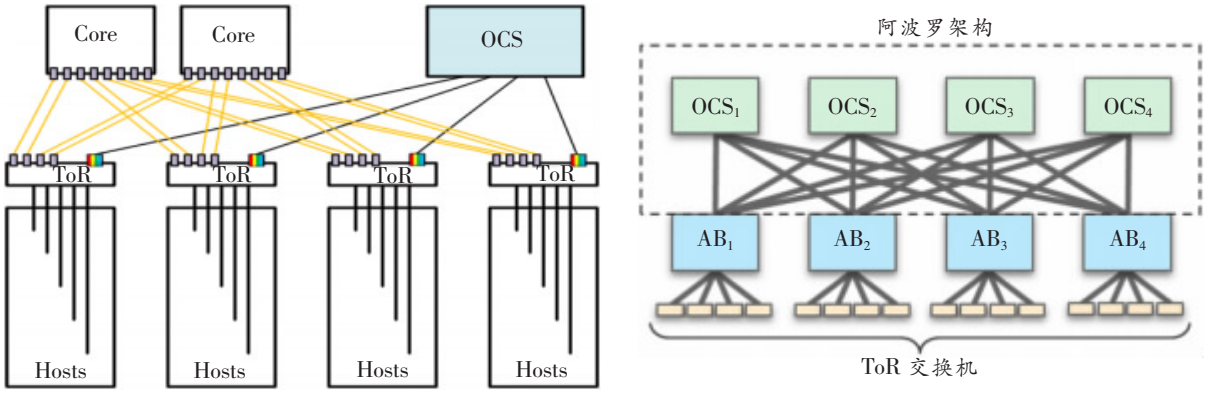
谷歌在建设数据中心的过程中引入了 OCS 形成新的解决方案。目前 OCS 在谷歌基础设施中主要有 Jupiter 数据中心和张量处理器(Tensor Processing Unit，TPU)数据中心两大应用场景，其中后者为专注于人工智能算力的数据中心。

在初代 Jupiter 的基础上，通过引入 OCS 取代 Spine 层传统电交换机，将网络逻辑拓扑由 CLOS 架构演进到 Aggregation 块的直接光互联，其整体架构如图 14 所示。由于 OCS 采用光交换，对传输的速率无感，通过进一步引入 WDM 和环行器等技术可以实现在单根光纤上传输通道数的增加以及 Tx/Rx 双路信号，提升单光纤的数据传输速率，实现整个 Jupiter 网络互联带宽的数倍增长。基于以上的技术，Jupiter 现超过 6 Pbit/s 带宽容量，即相对于初代实现约 5 倍带宽提升的同时，电力消耗减少了 41%，成本降低 30%。

5.2 部分厂商的 OCS 商用方案

目前已有 Polatis、Coherent 和光迅等多家公司

推出了商用的 OCS 产品，如图 15 所示。3 款商用方案支持的最大端口数分别为 576、300 和 400。Polatis 采用了 Directlight 技术降低插入损耗，而 Coherent 选择了液晶 WSS 的方案，具有超低驱动电压，光迅的技术方案则未公布细节。



注：Core 为核心交换机；AB 为汇聚交换机。

图 14　谷歌 OCS 整体架构

Figure 14　The overall architecture of Google OCS

图 15　Polatis、Coherent 和光迅的 OCS 产品示意图

Figure 15　Products of OCS from Polatis, Coherent and Accelink

5.3　英伟达的 GPU 光互联方案

英伟达在 2022 年光纤通讯展览会（Optical Fiber Communication Conference，OFC）上展示了未来 GPU 光互联架构的设想。

随着 GPU 间带宽的急剧增长，电互联距离急剧降低，同时噪声会越来越大，从而影响信号传输质量，英伟达认为硅光互联是 GPU 的互联目标架构。表 1 给出了中介层（InterPoser，IPoser）、印刷电路板（Printed Circuit Board，PCB）、光电合封装（Co-Packaged Optics，CPO）、电缆（Electrical Cables，ECable）和有缘光缆（Active Optical Cables，AOC）在功耗、成本、密度和距离 4 个维度的对比。

表 1　不同 GPU 互联方式对比

Table 1　Comparison of different GPU interconnection schemes

	IPoser	PCB	CPO	ECable	AOC
功耗/J/b	10^{-13}	5×10^{-12}	10^{-12}	5×10^{-12}	10^{-11}
成本/美元/bit	10^{-15}	10^{-13}	10^{-10}	10^{-10}	10^{-9}
密度/b/s-mm^2	10^{13}	5×10^{11}	2×10^{12}	5×10^{10}	10^{11}
距离/m	0.005	0.500	100.000	5.000	100.000

GPU 和 NVSwitch 架构使用光引擎将电信号转换为光信号以光子连接 GPU 的 NVSwitch 网络的框图如图 16 所示。每个光引擎有 24 根光纤，单纤速率 200 Gbit /s，总共 4.8 Tbit/s 的速率。每个 GPU 都有一对这样的接口，为其提供进出 NVSwitch 结构的双向带宽。因此，具有 6 个光引擎的 NVSwitch 的原始速率为 28.8 Tbit/s，去掉编码开销后的速率为 25.6 Tbit/s。

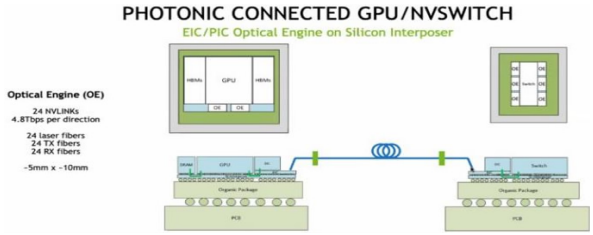


图 16　英伟达 GPU 交换设想架构

Figure 16　GPU switch target architecture from NVIDIA

同时，英伟达曝光了基于CPO方案的GPU互联架构图，如图 17 所示。采用这种架构可以将点对点

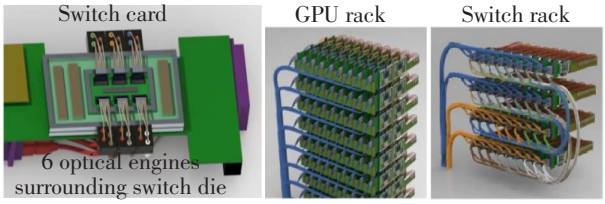


图 17 交换卡、GPU 架和交换架概念图

Figure 17 Schematic diagram of switch card, GPU rack and switch rack

的 GPU 互联时延降低 24.91%[27]。

6 不同光电融合交换技术方案对比

通过以上章节的介绍，给出一个横向的方案对比，从交换类型、控制方式、核心光学器件、连接方式、可扩展性和商用化水平几方面进行对比，如表 2 所示。

表 2 不同光电融合交换技术方案对比

Table 2 Comparison of different photoelectric fusion switching technology solutions

架构名称	交换类型	控制方式	光学设备	连接方式	可扩展性	商用化水平
Helios	EPS/OCS	分布式	WDM	光纤	中	商用设备
C-Through	EPS/OCS	集中式	MEMS	光纤	低	商用设备
Mordia	OCS	集中式	WSS	光纤	低	实验原型
RotorNet	OCS	分布式	MEMS	光纤	高	实验原型
Sirius	OCS	分布式	AWGR、Tunable Laser	光纤	高	实验原型
Proteus	OCS	集中式	WSS、MEMS	光纤	低	商用设备
OSA	OCS/OPS	集中式	WSS、MEMS	光纤	低	商用设备
Lotus	OCS/OPS	分布式	AWGR	光纤	中	软件仿真
ReSAW	OPS/OBS	分布式	AWGR、Tunable Laser	光纤	高	实验原型
Spanke	OCS	分布式	WSS	光纤	高	实验原型
ProjecToR	OCS	分布式	DMD	空间光	高	实验原型
FireFly	OCS	分布式	SM/GM	空间光	低	实验原型

注：Tunable Laser 为可调谐激光器；SM/GM 分别为从时钟和主时钟。

从光学设备的角度分析，MEMS 交换机的响应时间通常较长，一般在 ms 级别以上。其可以支持大量的通道数，通常在数十到数百个通道之间。其具有广阔的频谱宽度和高带宽容量，适用于需要高度并行的光通信应用。而 WSS 交换机响应时间通常较短，一般在 μs 级别。其通常提供较少的通道数，通常在数个到数十个通道之间。其适用于需要频谱选择性或波长级别的光信号调控和重新分配的应用。

AWGR 作为一种常见的光学分波器和光路选择器，也可以用作光网络中的交换设备。其响应时间通常在 ns 级别，可以实现快速的波长选择和光信号交换，因此适用于需要快速调整光信号路径的应用。

FSO 架构中用到的 DMD 具有高空间分辨率和像素密度，能够提供更高精度且更细致的光控制，通常从数百万到数千万个微镜像素。其响应速度通常以 μs 级或 ns 级为单位，适合于需要快速切换和重构光路的应用场景。

以上所有方案均是设备级或网络架构级的设计，还有一种思路是直接对芯片进行光/电互联设计，将芯片内部的电信号直接在板卡上转化成光信号与其他板卡上的芯片进行光互联。这种思路在高性能的服务器内部或存储设备间呈现一定潜在应用价值，例如 xPU（GPU、TPU、数据处理单元（Data Processing Unit，DPU）、中央处理器（Central Processing Unit，CPU））间或 xPU 与存储器间。硅光是实现这种思路的最佳选择，利用成熟的 CMOS 工艺，将光源、调制器、波导和探测器等光学芯片组件进行集成，利用片上接口，实现高速传输，实现设备小型化和低功耗要求。

7 光电融合交换技术标准化工作

由于符合电信级业务可靠性和高转发效率的光交换或光电融合交换产品还未大规模报道，因此在电信领域的标准化进展相对较为缓慢。无论从产业上游的芯片、器件、设备制造商还是下游的头部互联网公司（Over the Top，OTT）或电信运营商，现象级应用还未形成规模。在中国通信标准化协会（China Communications Standards Association，CCSA）和国际电信联盟（International Telecommunication Union Telecommunication Standardization Sector，ITU-T），已经开始出现关于光电融合架构中某些关键器件的标准化建议，包含可调谐激光器和 AWGR

等。

7.1　CCSA行业标准

2023年12月，CCSA TC6通过了《平面光波导集成光路器件 第4部分：阵列波导光栅路由器（AWGR）》和《$N\times N$阵列光交换矩阵开关》两个项目的立项申请，同意制定基于AWGR的光电融合交换方案中AWGR器件和基于$N\times N$全连接交叉矩阵光器件标准。《平面光波导集成光路器件 第4部分：阵列波导光栅路由器（AWGR）》定义了光交换系统中适用于可调谐激光器的AWGR相关内容，尤其是通道波长的选择、差损的限制和通道串扰等几个关键指标。该标准是国内首个关于数据中心光交换器件的标准。《$N\times N$阵列光交换矩阵开关》规范了大端口阵列光交换矩阵开关的技术要求。

7.2　ITU-T标准

2023年11月至12月，在ITU-T SG15全会上，通过了对ITU-T G.671《Transmission Characteristics of Optical Components and Subsystems》的修订，引入“$N\times N$ Arrayed Waveguide Grating Routers (AWGRs)”。该修订从2023年2月SG15 Q6中间会开始提出，从交换系统开始介绍，引入AWGR在下一代交换系统中的作用。同年7月，Q6中间会再次讨论，将快速可调谐激光器引入关键器件修订讨论。最终，在全会通过了G.671修订提议，并提出加注附录，详细说明可调谐激光器和AWGR在交换系统中配合使用实现交换功能的方案。至此，在ITU-T的首个关于光电交换关键核心器件的国际标准诞生。需要注意，该标准提出了快速可调激光器的指标建议，为了实现ns级切换和ms级端到端时延，必须对可调谐激光器相关指标做出规定。ITU-T在5G前传方向存在基于可调谐机制的光模块标准G.698.4，为该AWGR标准提供了制定依据。

8　结束语

本文从数据中心演进需求入手，分析了AIDC和SCDC应用场景下数据中心通信容量面临成倍增长，时延要求从ms级到μs级变化，传统电交换机PUE指标过高带来的压力，需要一种新的光电融合交换技术来满足AIDC和SCDC演进需求。通过分析数据中心组网架构，找到适合光交换机部署的位置，在保留传统电交换机的同时，分别采用了OCS、自由空间和可重构等技术方案来提升带宽、时延和功耗等指标。通过不同方案的对比，给出了光电融合交换技术的部署建议。然而，光电融合交换商业前景尚不明朗，使用光电交换技术的企业相对较少，产业化程度低。综合来看，光电融合交换是未来数据中心的一种演进手段，需要在真实业务环境下验证其性能，找到适合的应用场景，完善产业链，建立商业模式。

参考文献：

[1] Al-Fares M, Loukissas A, Vahdat A. A Scalable, Commodity Data Center Network Architecture[C]// Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication. Seattle, WA, USA: ACM, 2008:63－74.

[2] Cisco Systems, Inc. Cisco Data Center Infrastructure 2.5 Design Guide [EB/OL]. (2011-11-02) [2024-03-20]. https://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCI_SRND_2_5a_book/DCInfra_3a.html.

[3] 王甫涵，郝祥勇，蔡轶，等. 全光数据中心互联的混合放大技术研究[J]. 光通信研究，2023(3):6－9.
Wang F H, Hao X Y, Cai Y, et al. Research on Hybrid Amplification Technology of All-optical Data Center Interconnection[J]. Study on Optical Communications, 2023(3):6－9.

[4] Juniper Networks, Inc. Documentation Archives (Portable Libraries) [EB/OL]. (2024-02-08) [2024-03-20]. http://www.juniper.net/techpubs/software/erx/junose61/swconfig-routing-vol1/html/ip-jflow-stats-config2.html.

[5] Sun Microsystems, Inc. SUN™ Datacenter Switch 3456 System Architecture Massively Scalable Infini-Band Switch Architecture for Petascale Computing White Paper [EB/OL]. (2007-11-01) [2024-03-20]. https://www.cs.rpi.edu/～chrisc/COURSES/HP-DC/SPRING-2008/papers/ds3456_wp.pdf.

[6] 工业和信息化部. 新型数据中心发展三年行动计划（2021-2023年）[EB/OL]. (2021-07-04) [2024-03-20]. https://www.gov.cn/zhengce/zhengceku/2021-07/14/content_5624964.htm.
Ministry of Industry and Information Technology. Three-year Action Plan for the Development of New data Centers (2021-2023) [EB/OL]. (2021-07-04) [2024-03-20]. https://www.gov.cn/zhengce/zhengceku/2021-07/14/content_5624964.htm.

[7] Abadi M, Barham P, Chen J, et al. TensorFlow: A System for Large-scale Machine Learning[DB/OL]. (2016-05-27) [2024-03-20]. https://arxiv.org/abs/1605.08695.

[8] 张镭. 深度学习在自然语言处理中的应用：从词表征到 ChatGPT[M]. 北京：人民邮电出版社，2023.
Zhang L. Application of Deep Learning in Natural Language Processing: From Word Representation to ChatGPT[M]. Beijin: Posts and Telecom Press, 2023.

[9] Nath S, Marie A, Ellershaw S, et al. New Meaning for NLP: The Trials and Tribulations of Natural Language Processing with GPT-3 in Ophthalmology[J]. The British Journal of Ophthalmology, 2022, 106(7): 889－892.

[10] Asim M, Wang Y, Wang K, et al. A Review on Computational Intelligence Techniques in Cloud and Edge Computing[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2020, 4(6): 742－763.

[11] Farrington N, Porter G, Radhakrishnan S, et al. Helios: a Hybrid Electrical/Optical Switch Architecture for Modular Data Centers[C]//Proceedings of the ACM SIGCOMM 2010 Conference. New Delhi, India: ACM, 2010: 339－350.

[12] Wang G, Andersen D G, Kaminsky M, et al. C-Through: Part-Time Optics in Data Centers[J]. ACM SIGCOMM Computer Communication Review, 2010, 40(4): 327－338.

[13] Farrington N, Forencich A, Porter G, et al. A Multiport Microsecond Optical Circuit Switch for Data Center Networking[J]. IEEE Photonics Technology Letters, 2013, 25(16): 1589－1592.

[14] 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. Los Angeles, CA, USA: ACM, 2017: 267－280.

[15] Ghobadi M, Mahajan R, Phanishayee A, et al. ProjecToR: Agile Reconfigurable Data Center Interconnect [C]//Proceedings of the 2016 ACM SIGCOMM Conference. Florianopolis, Brazil: ACM, 2016: 216－229.

[16] Zhao Z, Xue X, Guo B, et al. ReSAW: a Reconfigurable and Picosecond-synchronized Optical Data Center Network based on an AWGR and the WR Protocol[J]. Journal of Optical Communications and Networking, 2022, 14(9): 702－712.

[17] Dang D, Guo B, Li W, et al. AWGR-based 25 Gb/s Optical-electrical Switching System Featuring Dynamic Bandwidth Allocating[C]//49th European Conference on Optical Communications (ECOC 2023). Hybrid Conference, Glasgow, UK: IET, 2023: 1745－1748.

[18] Hamedazimi N, Gupta H, Sekar V, et al. Patch Panels in the Sky: A Case for Free-space Optics in Data Centers[C]//Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks. College Park Maryland, USA: ACM, 2013: 2535771.

[19] Hamedazimi N, Qazi Z, Gupta H, et al. FireFly[J]. ACM SIGCOMM Computer Communication Review, 2015, 44(4): 319－330.

[20] 尹欣，侯维刚，郭磊. 一种自由空间光数据中心网络架构、拓扑重构系统和方法：中国，CN114745618A[P]. 2022-07-12.
Yin X, Hou W G, Guo L. Free Space Optical Data Center Network Architecture, Topology Reconstruction System and Method: China, CN114745618A[P]. 2022-07-12.

[21] Ballani H, Costa P, Behrendt R, et al. Sirius: a Flat Datacenter Network with Nanosecond Optical Switching[C]//Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication. Virtual Event, USA: ACM, 2020: 782－797.

[22] 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. Monterey California, USA: ACM, 2010: 1868447.

[23] 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, 2014, 22(2): 498－511.

[24] Lu Y, Gu H, Yu X, et al. Lotus: A New Topology for Large-scale Distributed Machine Learning [J]. ACM Journal on Emerging Technologies in Computing Systems, 2021, 17(1): 1－21.

[25] Alam S, Athanassiadou T, Robinson T W, et al. First 12-cabinets Cray XC30 System at CSCS: Scaling and Performance Efficiencies of Applications [C]//CUG 2013. Napa Valley, CA, USA: CUG, 2013: 285583551.

[26] 李泳成，廖晶晶，沈纲祥. 面向基于 Spanke 架构全光交换数据中心的 Ring 业务部署方法：中国，CN115175027B[P]. 2023-06-23.
Li Y C, Liao J J, Shen G X. Ring Service Deployment Method for All-optical Switching Data Center based on Spanke Architecture: China, CN115175027B[P]. 2023-06-23.

[27] Anderson E, González J, Gazman A, et al. Optically Connected and Reconfigurable GPU Architecture for Optimized Peer-to-peer Access[C]//Proceedings of the International Symposium on Memory Systems. Alexandria Virginia, USA: ACM, 2018: 257－258.