面向大规模智算集群场景光互连技术白皮书

（2025年）

发布单位：中国移动

编制单位：中移智库、中国移动通信研究院、中国移动云能力中心、中国移动设计院

前 言

当前，智算集群已成为支撑人工智能大模型训练、自动驾驶算法迭代等前沿领域的核心基础设施，并以惊人的速度从万卡向十万卡级规模演进。随着单节点算力突破每秒百亿亿次，这类超大规模集群的极致计算能力对互连链路带宽、延迟和功耗提出了极其严苛的要求。传统基于铜介质的电互连方案，正面临“带宽墙”、“延迟墙”及“功耗墙”等三重严峻挑战：单通道速率难以突破400Gbps，传输延迟高达数微秒，单机架互连功耗占比更是超过40%，这一系列瓶颈已成为制约超大规模智算集群算力释放的核心障碍。

相较于传统可插拔光模块等设备级光互连技术，芯片级光互连正在开辟全新的技术路径和产业赛道。它通过先进封装将光引擎与电芯片合封在一起，把电信号的传输距离从米级大幅压缩至毫米级，从而改写了物理层互连架构，实现50%以上的系统能效提升。由此构建的“芯片—设备—集群”一贯式全光互连架构，已被业界广泛认定为下一代智算基础设施的关键技术。

本白皮书系统性剖析芯片级光互连技术的核心原理和架构设计，深入探讨光源、调制器等关键器件的技术发展路径。同时，全面梳理芯片级光互连在国内外的产业现状，客观研判未来演进趋势和技术挑战。期望通过产学研用多方协作，加速芯片级光互连技术从实验室原型走向规模化商用落地，推动我国智算基础设施在硬件架构层面实现跨越式升级，为数字经济的高质量发展筑牢坚实的算力基石。

编写说明

牵头编写单位：

中国移动通信集团有限公司

联合编写单位（排名不分先后，按汉语拼音排序）：

北京凌云光通信技术有限责任公司

烽火通信科技股份有限公司

飞腾信息技术有限公司

光本位智能科技（上海）有限公司

华为技术有限公司

昆仑芯（北京）科技有限公司

沐曦集成电路（上海）股份有限公司

摩尔线程智能科技（北京）有限责任公司

锐捷网络股份有限公司

上海曦智科技有限公司

上海图灵智算量子科技有限公司

苏州盛科通信股份有限公司

苏州奇点光子智能科技有限公司

无锡芯光互连技术研究院有限公司

新华三技术有限公司

中兴通讯股份有限公司

目 录

1. 下一代智算集群提出近乎严苛的互连需求

1.1. 大模型的巨量迭代引发智算集群架构变革

实现通用人工智能（AGI, Artificial General Intelligent）已成为大模型未来发展方向的广泛共识。大模型技术总体仍遵循扩展法则（Scaling Law），参数已迈向万亿甚至十万亿规模，对智能算力的需求呈现爆炸式增长。如下图所示，模型参数规模的增长速度约每两年400倍，其算法结构在原有Transformer的基础上，引入扩散模型、专家系统（MoE, Mixture of Expert）等，使模型泛化能力增强，并具备处理10M+超长序列能力，推动芯片算力（FLOPS）约每两年3倍的提升，需要至少百倍规模的集群演进速度来支撑大模型的发展，但芯片间的互连能力提升缓慢，只有约每两年1.4倍，远落后于模型规模和算力的演进速度。

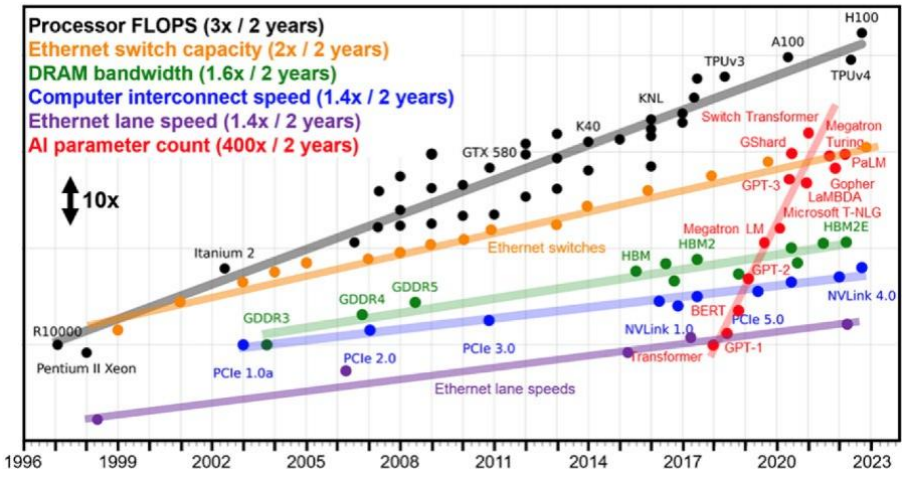

图 1-1 智算场景中各技术领域扩展趋势[1]

超大模型的训练过程尤其是张量并行（TP, Tensor Parallelism）、专家并行（EP, Expert Parallelism）等模式依赖集群内GPU芯片之间频繁的数据交互。然而，互连速率的提升已严重滞后于算力的快速演进，导致显著的通信开销，这直接限制了集群有效算力随GPU数量的线性增

长，已成为制约集群规模扩展和性能提升的关键瓶颈，如下图所示。在此背景下，仅仅依靠IB（InfiniBand）或RoCE（RDMA over Converged Ethernet）等传统网络技术来满足模型性能指标已十分困难，需构建具备高带宽、低延迟特征的GPU卡间互连技术体系，以扩大节点规模，大幅降低通信时间占比，最终实现集群算效的显著提升。

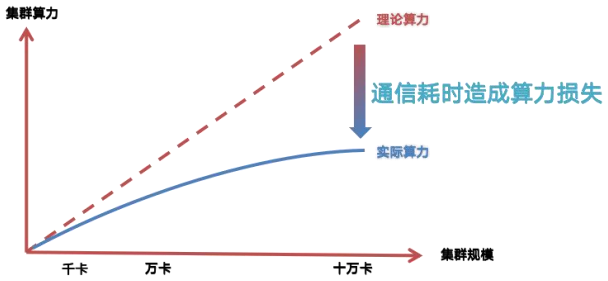

图 1-2 算力随着卡数规模扩大难以线性扩展

同时，全球智算中心规模触达十万卡级别，智算集群架构正经历一场根本性变革，从传统单机八卡向超节点演变。超节点并非简单的硬件堆叠，是一种通过极致性能的高速互连技术，将数十乃至上千颗GPU芯片集成于单个或多个机柜的集群系统，突破传统设备算力瓶颈，显著降低多芯片并行计算的通信损耗，实现大模型训练与推理效率的飞跃。

1.2. 大规模智算集群呼唤“光进电退”技术

目前，超节点智算集群展现出三大技术特性，一是互连性能高，GPU之间具有超低时延超高带宽（百纳秒级，TB/s级）且无收敛的互连能力；二是算力密度高，由单个或多个机柜构成，包含32个以上甚至到千卡的GPU数量，不断逼近电互连物理部署极限；三是能效PUE高，超节点单机柜功率可达40kW以上，采用液冷为主、风冷为辅的散热方案，配合柜级集中电源供电，在提供更高供电效率的同时大幅降低数据中心PUE。

为了实现更高的集群算效水平，互连技术方案的演进迫在眉睫。在超节点设备的互连选择上，当前主要存在两种路径：基于铜缆和基于光纤的传输方式。尽管铜缆作为目前的主流方案，相较于传统的可插拔光模块与光纤组合，拥有技术成熟度、成本、可靠性以及部署维护便捷性等多方面优势。通常在小于2米短距离和低于800Gbps的非超高速组网场景中，铜缆凭借这些优势依然能满足绝大多数应用需求。特别是无源直连铜缆（DAC, Direct Attach Cable），凭借其极低的成本和超高的平均无故障时间（MTBF, Mean Time Between Failures），成为当前主流选择。

然而，在高速传输场景下，铜缆面临着距离受限、功耗激增、速率瓶颈和布线困难等严峻挑战，已然逼近其性能极限。随着超节点集群规模继续扩展至256节点乃至千卡级别，且单通道传输速率迈向800Gb/s，铜缆的固有物理局限性正日益凸显，已成为制约智算集群互连性能与扩展潜力的严峻挑战。

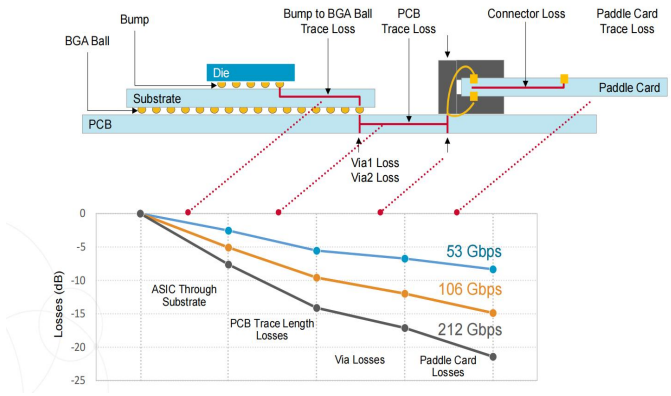

图 1-3 不同速率的电信号在服务器内不同位置的损耗情况[2]

首先，铜缆的局限性体现在其距离限制。受限于信号衰减，铜缆的有效传输距离极其有限。例如，在极短的10厘米PCB走线中，100Gbps的速率就足以造成超过15dB的插入损耗，导致信号失真率突破5%。当GPU跨越多机柜时，距离超过10米的情况下，信号衰减与功耗问题更为

突出。其次，功耗激增是另一核心痛点。在800Gbps及以上的高速传输场景下，电流通过铜线产生的巨大热量不仅大幅推高了数据中心的运营成本，也显著增加了系统的散热复杂性。再者，铜缆面临着传输速率瓶颈。受限于“趋肤效应”和PCB走线的寄生电容、电感，其中长距离传输的单通道速率难以突破200Gbps,且多通道并行会导致严重的串扰，进一步限制了电互连的带宽密度。最后，布线困难成为规模化部署的巨大障碍。随着智算集群规模呈指数级扩张，所需的铜缆数量几何级增长，使得布线难度与成本显著提高，严重制约集群快速扩展和高效运维。这四大固有物理局限，使得铜缆已无法满足未来高算力密度和大规模扩展的智算集群的严苛需求。

为跨越基于电信号铜缆传输的固有物理极限，新一代光互连技术正快速登上历史舞台。以近封装光学（NPO, Near Package Optics）、共封装光学(CPO,Co-Packaged Optics)、以及光输入/输出(OIO,Optical Input Output）为代表的创新方案成为替代铜缆方案的优秀选择。这些技术的核心在于最大程度地缩短电信号与光引擎（OE, Optical Engine）之间的距离，实现在芯片层面即完成光电转换，从根本上规避了传统可插拔光模块的高成本与易故障问题，同时继承了光纤传输的技术优势。

功耗显著降低。NPO、CPO等技术将光引擎与GPU封装在同一基板甚至同一芯片上，将电信号路径缩短至厘米甚至毫米级别，大幅减少了传输过程中的中继损耗，并降低了SerDes接口的性能要求，从而系统性地降低了整体功耗。

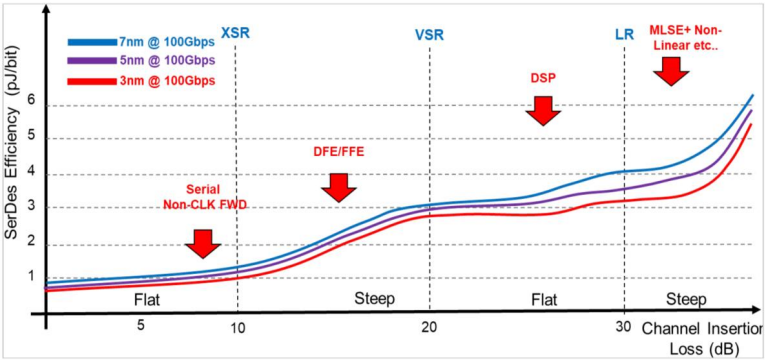

图 1-4 不同接口设计的SerDes功耗[3]

带宽密度显著提升。通过缩短电信号传输路径，这些技术能支持更高的单端口传输速率，同时在同一封装体内集成多个光通道，使得带宽密度达到百Gbps/mm²至Tbps/mm²，远超铜缆互连方案。此外，连接距离得到极大扩展。光信号的低损耗特性使其能够轻松覆盖数据中心内数百米甚至10公里以上的距离，彻底打破了铜缆在远距离传输上的桎梏。

更为重要的是，光互连在信号完整性上展现出压倒性优势。多根铜缆并行传输时固有的串扰和反射问题，需依赖复杂的均衡算法进行补偿，而光信号在传输过程中几乎不受电磁干扰，其传输损耗比电信号低4至5个数量级，且与传输频率无关，从根本上保障了信号纯净度。

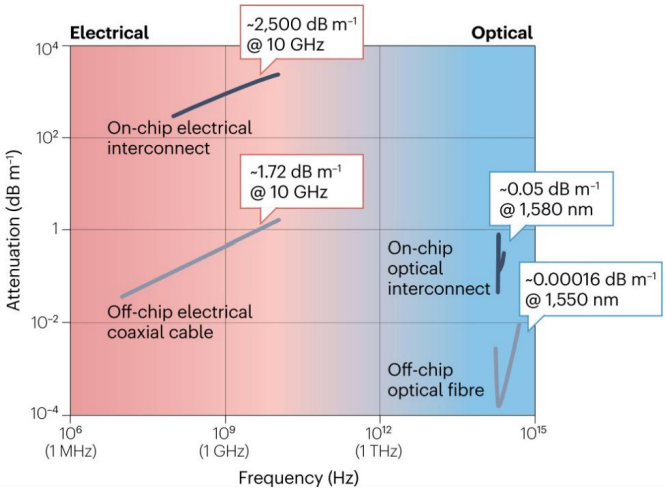

图 1-5 不同速率下光和电信号分别在芯片内和芯片外的损耗情况[4]

在空间利用方面，光互连在空间占用和重量上也展现出较大优势。一束直径仅6mm的光纤即可容纳12根光纤，提供超过19.2Tbps的总传输带宽，而提供同等带宽的铜缆束直径将超过50mm，重量更是光缆的8倍。这种极致的轻量化与小型化设计，极大地简化了大规模集群的布线难度，降低了数据中心的运营成本，并为未来更高密度的集成提供了宝贵的物理空间。

图 1-6 铜缆和光纤的部署对比[5]

尽管面临初期成本高和技术门槛高等挑战，但光互连技术所带来的低损耗、长距离、高带宽密度、高信号完整性以及低空间占用等核心优势，使其成为突破超节点规模和算力极限的关键支撑。通过将光电转换技术集成到芯片级别，光互连不仅拓展了传输距离，降低了系统功耗，更通过光信号的长距离传输解决了单节点规模扩大的空间限制问题。“光进铜退”已成为智算集群的必然趋势，是实现未来算力跨越式发展的核心驱动力。

此外，光技术的引入已拓展到交换层，即光交换技术（OCS，Optical Circuit Switching）。为解决传统电交换机多次光电转换导致的高能耗和微妙级延迟瓶颈，OCS直接在光域完成信号路由，最高可达纳秒级切换速度，较电交换快2-3个数量级。纯光交换中微镜反射型（MEMS，

Micro-Electro-Mechanical Systems）做为其中一种比较成熟的技术，已经实现了商业化应用。

2. 极致化需求驱动光互连技术革新

根据不同应用场景，光互连技术主要分为数据中心间（Data Center Interconnect，DCI）与数据中心内两大类。数据中心内聚焦短距传输场景（数米至数百米），核心诉求是高带宽密度、低延迟及低功耗，常用多模光纤，精准适配机柜内/跨机柜互连需求。**本白皮书重点探讨数据中心内光互连技术的分类、器件与技术趋势。**

2.1. 业界存在两大类光互连技术

光互连技术是通过应用光电转换与融合技术，取代电信号在传统数据传输场景中的主导角色，甚至直接替代芯片上的电IO功能，最终实现信号在传输过程中远距离、低功耗、高密度的目标。其中，实现光电转换的光引擎（Optical Engine，OE）是光互连技术的核心。根据应用场景、光引擎与xPU芯片的距离以及封装集成程度的差异，业界衍生出许多技术范畴，我们将其主要分为两大类：设备级光互连和芯片级光互联。

如下图所示，在未来十万卡级以上的智算中心集群设计中，设备级光互连主要有两大技术，一是以光交换技术为主，主要应用于交换设备间网络连接中，提供超高端口密度、极高速率（无带宽瓶颈）、连接距离从米级到百公里级；二是以可插拔光模块技术为主，主要应用于超节点设备间网络连接中，提供较高速率、千卡及以上规模、公里级别长距离连接；芯片级光互连主要以共封装光学为主，主要应用于超节点内并进一步下探到芯片内场景，提供超高带宽密度（可达Tbps/mm²级）、超低时延、千卡以下互连规模、公里距离之内的连接，要求高可靠性。

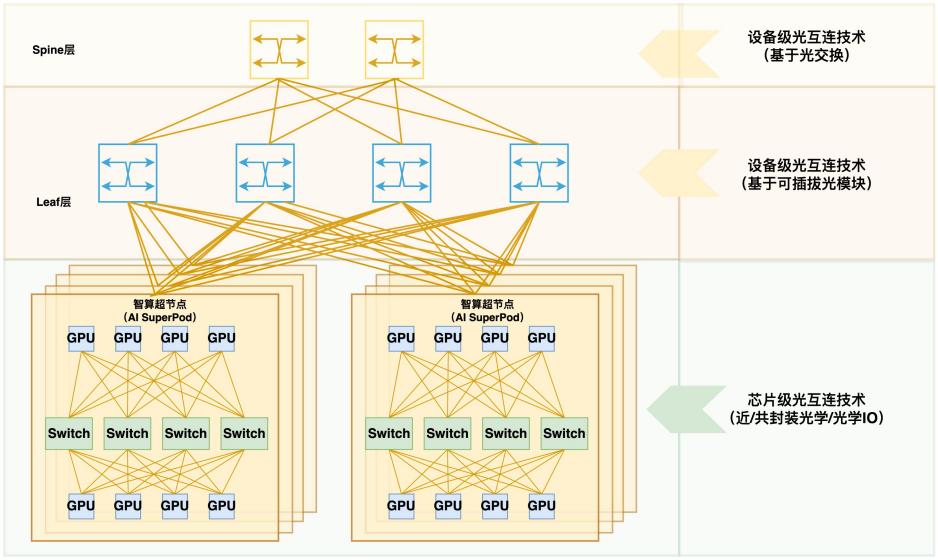

图 2-1 十万卡级智算中心集群光互连架构设计

2.1.1. 设备级光互连：光交换机的演进与应用

随着智算集群规模持续扩展，电交换芯片逐渐显现瓶颈。单芯片容量受制于集成电路工艺的发展，使得电交换芯片在制程工艺、转发架构与缓存设计等方面面临诸多挑战，交换芯片更新迭代速度明显放缓，网络规模难以快速扩展；高速SerDes和复杂转发架构导致功耗和延迟不断上升，信号完整性问题也需要依赖复杂DSP补偿。

光交换为突破电交换的限制提供了新的路径：

一是，其在光层面直接完成端口间的切换，无需O-E-O转换，彻底绕开了制程、缓存和SerDes衰减等物理瓶颈，可支持极高传输速率与超大规模集群部署。光交换天然具备速率和协议无关的特性，从400G到800G乃至1.6T均可平滑支持，在速率升级时无需更换交换设备，极大降低了系统演进的复杂度和成本。

二是，光交换通过端到端光路直通，避免了复杂的包解析与缓存转发，延迟大幅降低，功耗显著优于电交换。其大规模端口集成能力，使得数百乃至数千端口的互联成为可能，从而支撑大规模GPU集群的灵

活组网需求。

三是，通过集中化的控制与软件编排，光交换还能够支持拓扑重构、故障绕行和网络切片，提升算力利用率与网络鲁棒性。在运维方面，自动化光路配置减少了人工布线带来的潜在错误，进一步增强了网络的可用性和可靠性。

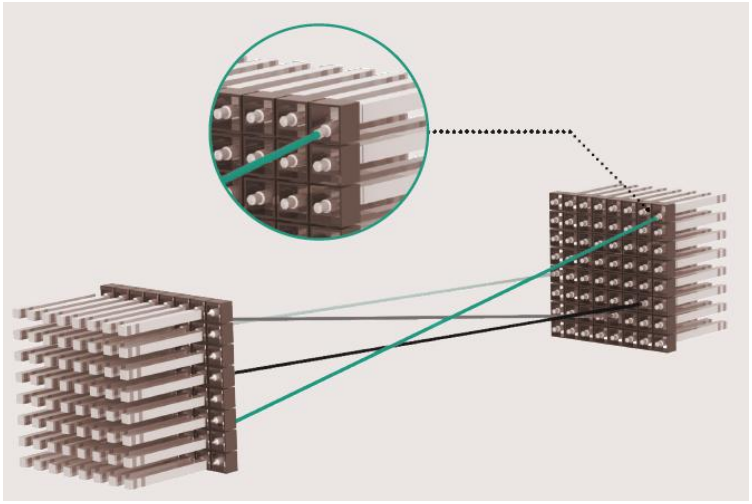

图 2-2 光交换机内部架构示例

2.1.2. 设备级光互连：可插拔光模块的演进与应用

可插拔光模块已广泛应用在传统数据中心、电信网络以及智算中心大规模连接中，具备灵活性高与兼容性强等特征。其将光引擎（OE, Optical Engine）集成在可插拔模块中，如下图所示，通过PCB（PCB, Printed Circuit Board）板级走线与有独立基板的xPU（GPU, NPU, Swtich, etc）相连。目前市场主力产品的速率已达800G，未来采用硅光技术可达1.6T水平，封装向高密度QSFP-DD/OSFP等演进。但面向智算未来高速率1.6T/3.2T以上的互连场景下，可插拔光模块将面临信号完整性恶化、依赖数字信号处理器（DSP，Digital Signal Processor）进行复杂信号补偿导致的系统功耗高、传输时延高等难题。

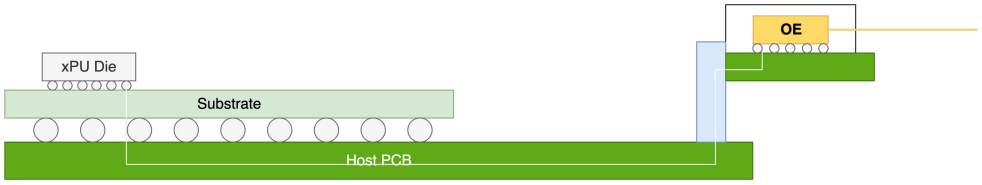

图 2-3 可插拔光模块示例

为解决DSP带来的功耗、时延等难题，2022年Macom联合英伟达推出线性直驱可插拔光模块（LPO, Linear Pluggable Optics）方案，如下图所示，相较于传统可插拔光模块，LPO直接去除了DSP芯片，保留发射端高线性度的驱动芯片（Driver），以及接收端高线性度的跨阻放大器（TIA, Transimpedance Amplifier），从而构建一个纯模拟的、“线性直驱”的光信号处理通道，实现功耗和时延的降低。虽然去除了DSP，但是DSP的功能并未消失，而是将部分功能转移到了xPU芯片中。这意味着xPU的SerDes必须具备更强的线性驱动能力和信号处理能力。

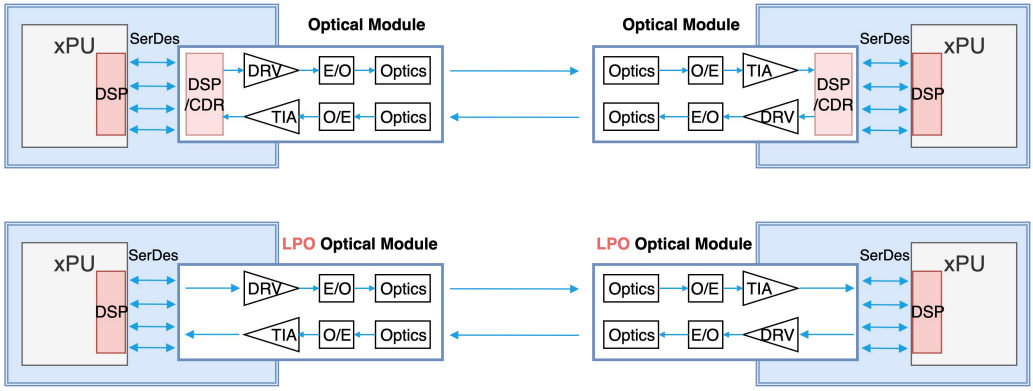

图 2-4 传统可插拔光模块（上图）与LPO（下图）的对比

2.1.3. 芯片级光互连：从近封装到光学I/O

随着专家模型的大EP（Expert Parallelism）架构发展趋势，更大规模、更高带宽密度和极低时延成为智算集群的主要需求。如下图所示，在规模方面，当前Scale-Up单层规模以32卡或64卡为主，需要进一步提升至256卡甚至千卡，高速传输的距离从板级、柜内扩展到柜间；

在带宽密度方面，当前国内单通道带宽以200Gbps为主，需要进一步向800Gbps甚至1.6Tbps迈进，带宽密度要求提升至百Gbps/mm^{2}到TGbps/mm^2；在时延方面，当前卡间数据传输时延为微秒级，需要进一步缩短至百纳米甚至十纳秒级。目前可插拔光模块的互连延迟和带宽瓶颈已无法满足大规模智算集群互连需求。

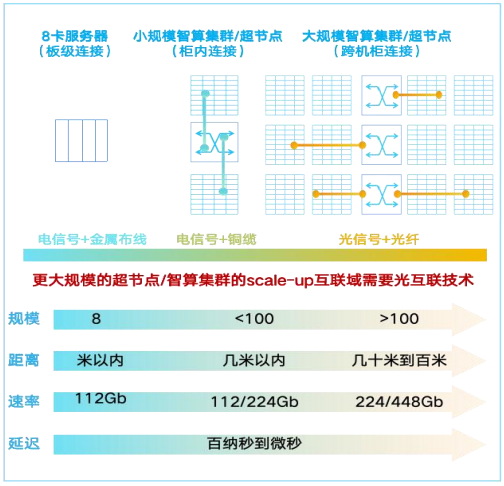

图 2-5 大规模智算集群的互连性能需求

芯片级光互连技术通过将电信号传输路径缩短至厘米到毫米级（即加速卡内部），相较于基于电互连与可插拔光模块的互连方案，可实现超高带宽密度、超低时延及高能效的智算集群互连能力。根据应用场景、光引擎与 xPU 芯片的距离及封装集成度，可将该技术分为近封装光学（NPO，Near Packaged Optics）、共封装光学（CPO，Co-packaged Optics）及光学IO（OIO，Optical Input/Output）三类。

- **近封装光学（NPO）**

NPO的核心思想是将光引擎（OE）与封装后的xPU芯片相邻布局于同一块高性能PCB基板上，通过极短的高性能电气链路与GPU相连，形成一个集成度较高的系统，GPU与OE的间距通常在数厘米以内，同时确保信道损耗≤13dB。相较于传统可插拔光模块，互连密度提高了2-3

倍，是光互连向高集成度发展的过渡阶段技术，为进一步向CPO演进奠定基础。

因NPO将GPU与光引擎物理分离，避免了GPU在工作时的高温热量直接冲击对温度敏感的光器件，从而导致波长漂移和系统性能下降，因此散热设计更简单、高效，系统更加稳定。同时，由于光引擎未和GPU共同封装，在可维护性方面具备一定优势，如果光部分失效，只需更换光引擎模块即可，避免了大量的维护成本；因此，NPO目前是国内GPU芯片厂家选择的主要技术路径，但仍需要在集成度、带宽密度、延迟和能效方面进一步优化。

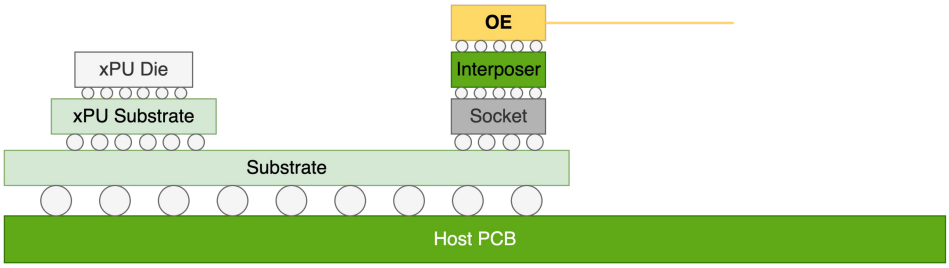

图 2-6 近封装光学（NPO）结构

- 共封装光学（CPO）

CPO技术通过将OE与电芯片共同封装在同一芯片基板或中介层上，实现系统的高集成度，使电信号只需传输几毫米。

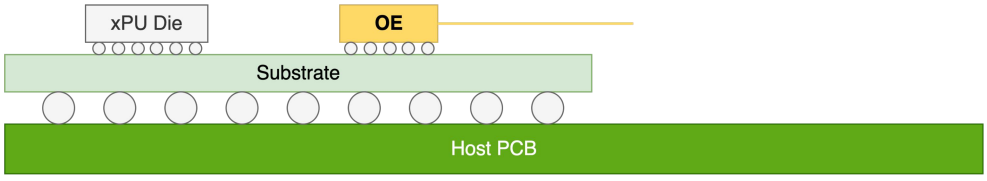

图 2-7 共封装光学（CPO）结构

CPO技术极大地提升了互连带宽密度并能够显著降低系统误码率和设备功耗，同时也能够大幅节省设备（如交换机）面板的空间，克服面板IO密度的限制。目前作为可插拔光模块的一种替代技术，CPO

可实现整机设备功耗降低50%左右，如下图所示。

51T Switch Box Total Power			
	Bailly CPO	Pluggable LPO	Pluggable w/DSP
8x OE Chiplets	241	630	1024
16x PLS	118		
ASIC, CPU, Other	975	975	975
Total (Watts)	1,334	1,605	1,999

图 2-8 交换机设备功耗分析[6]

由于光引擎和电芯片紧密共封装，任何子模块的故障都可能导致整个封装体的更换，对良率和可维护性方面提出了极高要求。因此，基于CPO技术的产品处于发展初期，主要应用场景是智算中心的交换设备。但凭借其在超高带宽、低功耗、低延迟、高密度互连等方面的巨大潜力，CPO有望进一步下探至GPU算力芯片，实现算力芯片的直接出光，构建更高效的端到端光互连链路。

- 片间光学互连（OIO）

相比NPO/CPO是突破可插拔光模块的性能限制，OIO技术目标是为了取代计算芯片上电IO方案，通过先进封装以芯粒形式与计算芯片集成，比CPO的互连性能更优。其核心理念是彻底摒弃传统的铜线电气I/O，消除了板级电气走线的瓶颈，将带宽密度提升至1Tbps/mm²（3D封装）并将延迟降低至纳秒级，能效提升相比CPO低一个数量级。

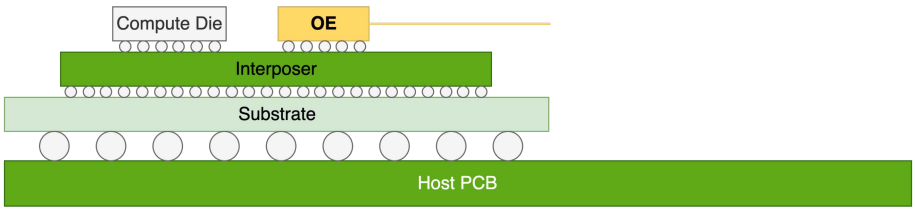

图 2-9 光学IO（OIO）结构

OIO技术可以在计算资源池化领域发挥更大的作用，如应对计算芯片显存容量和带宽扩展受限的双重挑战，依托其显著传输性能和距离，

打破单芯片显存物理边界，将多节点独立显存整合为共享显存池，通过光域直连实现池化显存的低时延调度与高带宽访问，成为未来新数据中心架构革新的关键驱动力。

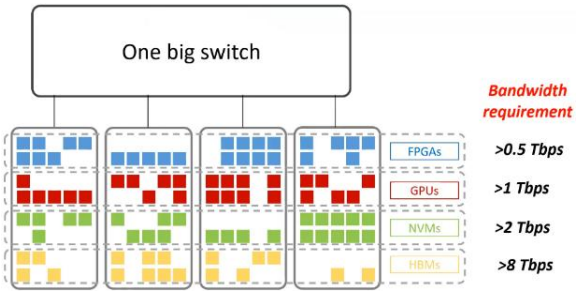

图 2-10 非聚合数据中心（Disaggregated DC）的互连带宽要求[7]

2.1.4. 新型光互连技术具备巨大潜力

可插拔光模块、NPO、CPO和OIO四大技术在带宽密度、时延、能耗、兼容性等方面表现各异，共同构成了覆盖数据中心内不同需求场景的光互连技术体系（见下表），**其中芯片级光互连展现出更能精准匹配智算集群未来演进需求的潜力，后续将聚焦该类技术展开具体分析。**

表 2-1 传统电交换和光交换（OCS）对比分析

分析维度	传统电交换	OCS
原理	传统电交换采用队列存储-转发方式，根据报文头信息将不同的数据包转发至不同的输出端口	内部采用MEMS技术（或压电陶瓷等其他技术），直接将光信号折射到对应的输出端口
工艺制程	要求高，依赖先进工艺，51.2T需要3nm制程	要求低，180nm成熟工艺
信号转换	O-E-O转换	无需光电转换
功耗	20w/端口	<1W/端口
带宽速率	取决于端口速率	速率和协议无关，从400G到800G乃至1.6T均可平滑支持
通信模式	Any2Any	点到点

表 2-2 光互连技术对比分析

分析维度	可插拔光模块(含LPO)	NPO	CPO	OIO
定义	OE独立封装，通过标准接口与设备主板热插拔	OE与封好的芯片通过高速基板封装，电互连距离缩短至厘米级	OE与未封好的电芯片实现芯片基板封装，电互连距离缩短至毫米级	OE与计算芯片通过实现芯粒级封装，终极目标省去电IO
带宽密度	50-100 Gbps/mm²	100-200 Gbps/mm²	200-500 Gbps/mm²	500-1000 Gbps/mm²
延迟	20-50 ns	10-20 ns	5-10 ns	<5 ns
能耗	15-20 pJ/bit	8-12 pJ/bit	5-8 pJ/bit	<5 pj/bit
灵活性	极高 支持热插拔	中 OE与主板绑定	低 OE与芯片绑定	极低 OE与芯粒集成
接口兼容性	支持多种接口	适配SerDes接口	适配SerDes接口	适配UCIe接口
可维护度	低 无需停机拆板	中 需拆板	高 拆封芯片	极高 整体更换芯片
主要应用场景	中长距互连 （公里级）	中短距互连 （百米级）	中短距互连 （百米级）	片间直连 （10米内）

2.2. 芯片级光互连三大技术路线场景互补

2.2.1. 芯片级光互连技术的组成原理

从器件构成上来看,相较于采用分立式器件的传统可插拔光模块，主流芯片级光互连技术由于硅光的引入，除激光器外，大部分已实现了多种光电器件的硅基集成。其技术方案构成主要分为三大关键组件：激光器（外置或与光引擎耦合）、光引擎、光纤及连接器。无论与电芯片的距离与集成度如何，实现高效光电转换的光引擎和激光器都是芯片级光互连方案的主要研究对象。

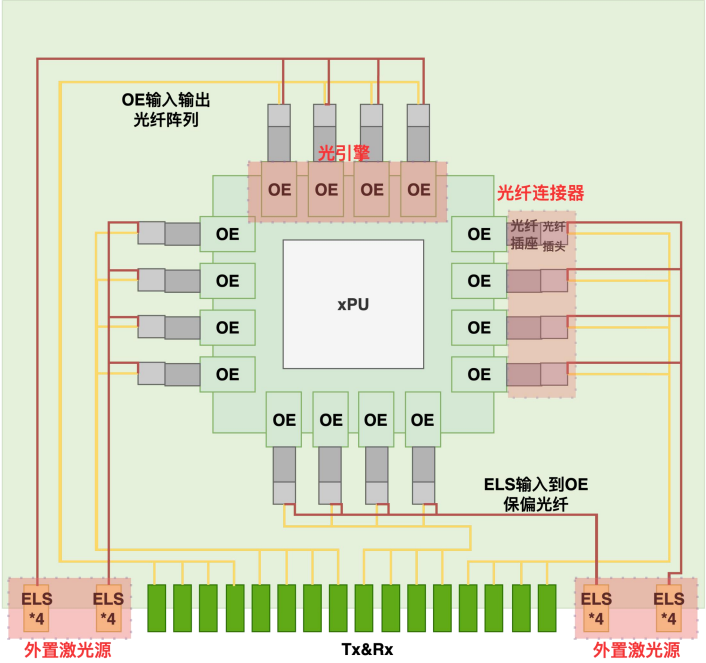

图 2-11 芯片级光互连的组件构成（以基于硅光技术的CPO设备为例）

如下图所示，光引擎由光集成电路（PIC, Photonic Integrated Circuit）和电集成电路（EIC, Electronic Integrated Circuit）组成。其中PIC主要包含调制器（MOD, Modulator）和探测器（PD, Photo-detector），基于硅光子或III-V族化合物材料实现光信号的调制、探测、解调和滤波等功能。其中，调制器负责将光信号调制成与电接口匹配的带宽能力，多采用硅光调制器，包括马赫-曾德尔调制器（MZM, Mach-Zehnder Modulator）、微环调制器（MRM, Micro Ring Modulator）等方案；探测器负责在收端将光信号转换成电信号；传统可插拔光模块中常采用分立的PIN或者雪崩光电探测器，在芯片级光互连中，集成于硅光芯片上的锗硅探测器（Ge-Si, Germanium-Silicon）成主流方案。EIC主要由驱动电路（DRV, Driver）、跨阻放大器（TIA, Transimpedance Amplifier）等组成，提供光调制器的驱动与控制，接收端信号的放大、均衡以及功耗管理等功能。

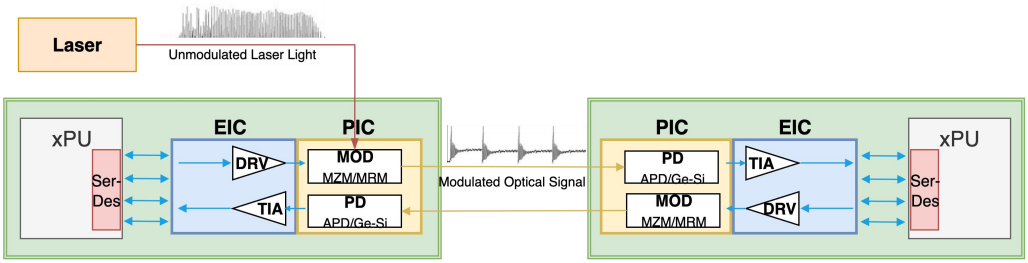

图 2-12 光引擎光电转换的原理

激光器（Laser）负责提供连续的、高品质的光源，而调制器则将电信号编码到光信号上。与传统光模块将激光器和调制器封装在同一个发射光组件（TOSA， Transmitter Optical Subassembly）内不同，该方案通常将调制器集成到硅光芯片上，而将激光器作为独立的外置光源（ELS, External Laser Source）。这种架构通常以可插拔模块的形式存在，如下图所示，可减少散热影响，增强系统稳定性。外置激光器方案与光引擎的耦合带了新的挑战，业界也有基于直接调制光源的技术方案，可解决光源与调制器分离带来的光效率问题，但也面临传输距离以及速率性能受限等难题。

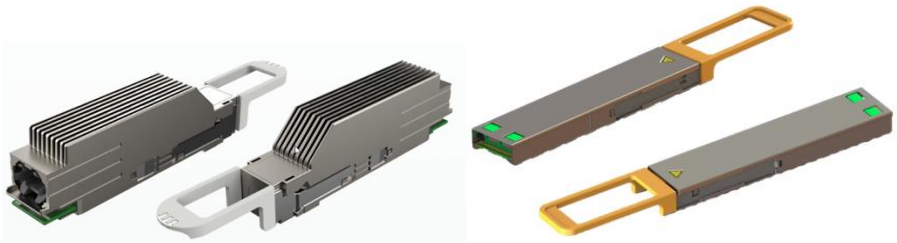

图 2-13 左图：博通自定义的ELS模块；右图：符合OIF ELSFP规范ELS模块

与传统设备内部无光纤布线设计不同，基于芯片级光互连技术的设备内引入了额外的光纤及光纤连接器。如下图所示，以基于硅光技术的CPO交换设备为例，光引擎紧密围绕ASIC芯片放置，设备内部的光互连路径包含两条：从ELS到光引擎，以及光引擎到机箱前面板。其中后者为业界主要研究方向，其连接方法和类型会影响信号、热量和布线密度的设备设计。

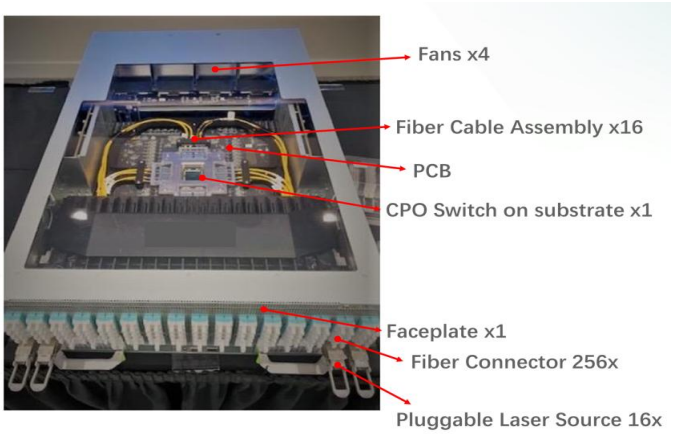

图 2-14 基于硅光技术的CPO交换设备示例（博通CPO交换机）

2.2.2. 三大技术路线并驾齐驱，硅光或成未来主流

行业内已提出并应用了多种芯片级光互连（如CPO）的实现方案，这些方案可按材料分类，也可按激光器的放置位置分类，而材料与激光器位置往往密切相关。业界目前有三大主要技术路线：其中基于硅光的集成方案通常采用外置激光源，属于间接调制（即需要一个独立的调制器来对光进行编码）；而基于垂直腔面发射激光器（VCSEL, Vertical-Cavity Surface-Emitting Laser）的方案则是由电子设备直接改变其注入电流来调制光源，无需额外的独立调制器；Micro-LED则摒弃传统激光器，采用Micro-LED作为光源，采用阵列形式，单个芯片可集成数十至数百个，满足高聚合速率需求。

目前产业主线多以硅光集成为核心，采用MZM或MRM等调制方式，并配合外置激光器实现高速信号中短距（~几百米）传输；VCSEL阵列则在短距互连（~几十米）中有成熟应用，但在高温稳定性和更高速率下仍面临一定挑战；Micro-LED作为一种新兴技术，主要聚焦于柜内短距高速链路（~数米内）中的应用，展现出高响应速度、高密度阵列集成及低功耗的特性，但其在高速调制（如100Gbps以上）的稳定性以及与电芯片异质集成适配性等方面仍存在问题。

● 外置激光源+硅光光引擎

硅光集成方案是利用现有CMOS（Complementary Metal Oxide Semiconductor）工艺进行光器件（包括调制器、探测器、光波导等）开发和集成的技术。根据调制器的不同，硅光方案可进一步分为两类：一类采用MZM调制器，另一类采用MRM调制器。MZM在硅光可插拔光模块市场中应用广泛，经过大量部署验证了其可靠性。基于MZM的芯片级互连方案借助这一优势，通过高度集成进一步提升了密度。MRM方案则提供了另一种可能，能够进一步降低调制器的功耗，并提高集成密度。MZM与MRM相比，MRM具有小尺寸及低驱动电压的优点，而MZM则有较宽的可操作光波长范围及较佳的热稳定性，相关比较如图2-13所示。

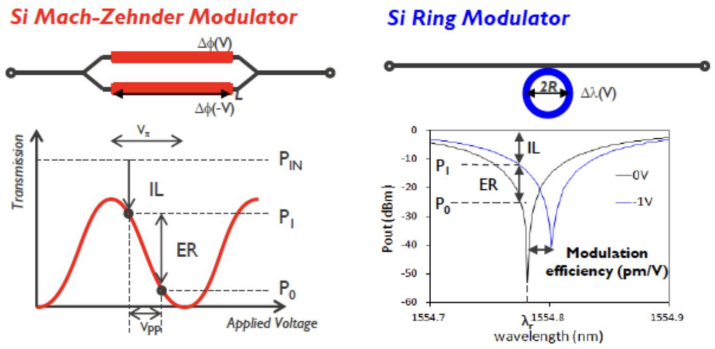

图 2-15 左图：MZM调制器 右图：MRM调制器[8]

硅光技术方案因集成度高、调制速率高，光源外置稳定性高，可覆盖其他短距方案等特性，成为CPO的主流方案。随着研究的深入，硅光技术有望成为OIO中最核心的光学解决方案。采用MRM的硅光集成方案实现OIO的第一步，可利用多个波长携带信号，提高带宽密度。

目前此路线面临光链路效率与系统协同性上的挑战。一是外置激光器耦合损耗与对准难题，易因偏移导致功率衰减，激光器需提升输出功率增加整体功耗；二是单个光源故障可能影响多通道工作；三是光源参数与硅光引擎的驱动需求适配依赖定制化调试，缺乏统一标准

导致集成成本高。未来产业可通过采用晶圆级光学技术集成微透镜阵列，并结合先进封装方案，将系统损耗降低；光源侧可采用量子点光频梳激光器，减少光纤用量并降低功耗并通过标准化统一光源电气与机械参数，进一步优化能效与互操作性。

- **基于VCSEL的光引擎方案**

VCSEL方案依托垂直出光结构带来的光路设计灵活性，以及高密度阵列支持多通道并行传输的能力，可满足智算集群柜内/间的短距传输需求。凭借成本优势与低功耗特性，在光模块领域已应用多年。但基于VCSEL的芯片级互连方案目前仍处研发阶段，核心瓶颈在于砷化镓材料与硅基工艺存在晶格失配，异质集成良率低，难以实现与电芯片的深度共封装，更合适应用于NPO互连方案。

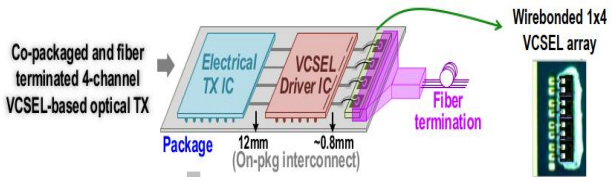

图 2-16 基于VCSEL的光引擎示例[9]

基于VCSEL的芯片级互连方案研究趋势主要聚焦于性能瓶颈突破，如推动单模化以解决带宽限制，业界通过光子晶体结构设计、氧化限制层精度优化，结合PAM4高阶调制技术，已实现单通道200Gbps速率原型，同时抑制杂模提升信号完整性；通过低损耗硅基波导与VCSEL的异质集成可降低信号损耗，使传输距离延伸，进一步提升方案能效与可靠性。

- **基于Micro-LED的光引擎方案**

在光互连领域中，Micro-LED作为新型光源阵列逐渐受到关注。与硅光和VCSEL相比，Micro-LED的突出特点在于其天然适合构建二维高密度阵列，能够实现多通道并行和空分复用，在有限封装岸线上实现

超过Tbps/mm^{2}的带宽密度。在功耗方面，研究表明其链路能效有望达到亚pJ/bit量级，适用于机柜内的10米级短距连接。工艺路径上，Micro-LED通常基于氮化镓（GaN, Gallium Nitride）外延，在蓝宝石或GaN衬底上制备微米级发光单元，并通过异质集成与CMOS电路键合，为短距互连带来一种能效与密度兼具的潜力方案。

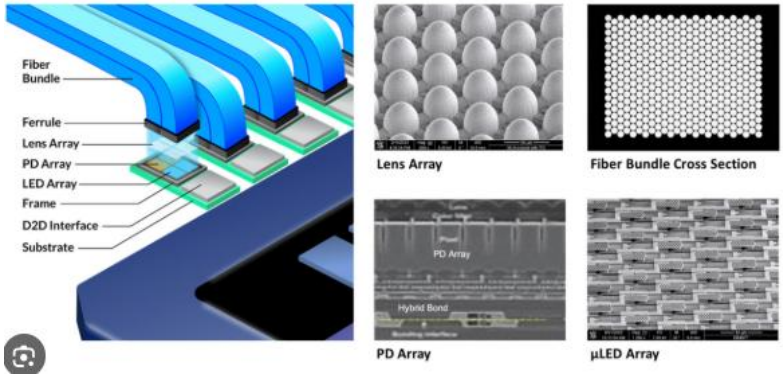

图 2-17 AVICENA MicroLED的光引擎方案示例

基于Micro-LED的光互连方案技术趋势着重于优化驱动电路、改进量子阱材料结构、延长激光源使用寿命以适配大规模集群的高可靠性需求。

总体来看，硅光方案因其性能优、CMOS工艺集成高等特性已形成较成熟的产业和标准牵引，VCSEL阵列依托既有的短距应用在NPO方案中仍具竞争力，Micro-LED阵列提供了能效和并行密度上的新路径。三者将在智算互连不同场景中形成互补，共同推动短距至中短距光互连技术的迭代升级。

3. 前瞻性芯片级光互连生态迎来关键窗口期

3.1. 国际产业由巨头牵引率先打通产业链

(一) 标准化工作进展

OIF CPO工作组从2020年起组织芯片、光器件、封装和系统厂商共同制定标准，旨在满足高带宽、低功耗互连的需求，并为产业界提供统一的技术规范。已发布的CPO相关标准与项目包括：

- 《Co-Packaging Framework Document》，该文档对光电合封系统框架进行了说明和定义；
- 《Implementation Agreement for a 3.2Tb/s Co-Packaged (CPO) Module》定义了用于以太网交换机的 3.2T CPO 模块，光口 FR4 和 DR4 、电接口32xCEI-112G-XSR 、光机械模块规格、电气规格以及通过增强现有 OIF CMIS 规范来实现的控制和管理接口等；
- 《 External Laser Small Form Factor Pluggable(ELSFP) Implementation Agreement》定义了前面板可插拔外部激光光源规格，以及对机械、热、电气和光学参数的互通性，标准的功率范围和光纤结构等进行了定义。

(二) 产品/技术方案进展

整体而言，国际厂商已经形成了从技术验证到商用量产的闭环。以GPU/ASIC芯片的供应商为例，他们同时具备计算芯片与光引擎的设计能力，可以内部整合后直接交由台积电、格罗方德（GF, GlobalFoundries）负责光互连芯片制造，再组装成完整共封装产品。国际领先CPO厂商主要由芯片巨头和光引擎企业构成。其中，芯片巨头以博通、英伟达、英特尔为代表，光引擎初创企业以AyarLab、Lightmatter、Avicena为代表。

● 博通产品及技术方案

博通从2021年开始布局CPO，技术积累深厚。2024年实现了第二代51.2T CPO交换机的技术展示。其系统中含有8个光引擎，分布在交换机芯片的四个方向。每个光引擎内置64路收发器，每路收发器通道的带宽为100Gbps，边缘带宽密度为500Gbps/mm^2。博通采用的是行波马赫曾德调制技术（Traveling-wave Mach-Zehnder Modulator,TWMZM）方案。未来，博通新一代交换机容量将达到102.4Tbps。

图 3-1 博通CPO交换芯片样例[10]

● 英伟达产品及技术方案

英伟达在GTC2025大会首次展示了CPO交换机产品，如下图所示。

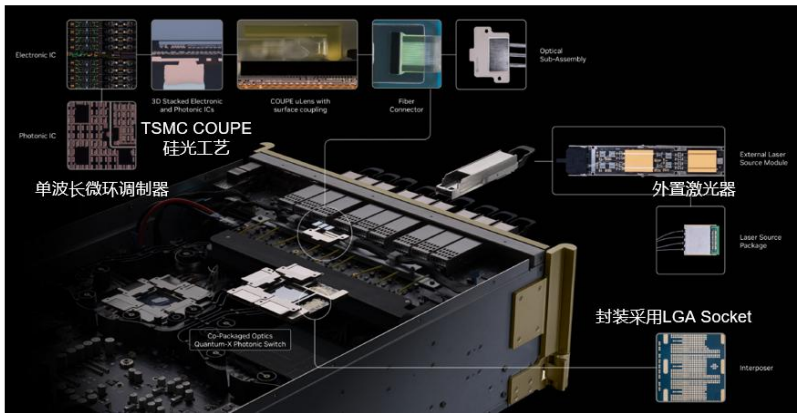

图 3-2 英伟达CPO交换机展示的技术细节[11]

英伟达的硅光芯片采用了单波长的微环调制器(microring modulator,MRM)，单通道信号速率为200Gbps。每个光引擎含8个通道，

总速率为1.6Tbps。单颗交换芯片配置6个模组，每三个光引擎组合成一个模组，交换芯片总带宽为28.8Tbps。交换机中含有四颗交换芯片，交换设备总带宽为115.2Tbps。电芯片采用TSMC 6nm工艺，包含2.2亿个晶体管，通过将EIC减薄后混合键合到PIC上，如下图所示。由于键合后的EIC-PIC厚度比较薄,可在最上层增加一层硅表面用于光学耦合对准。

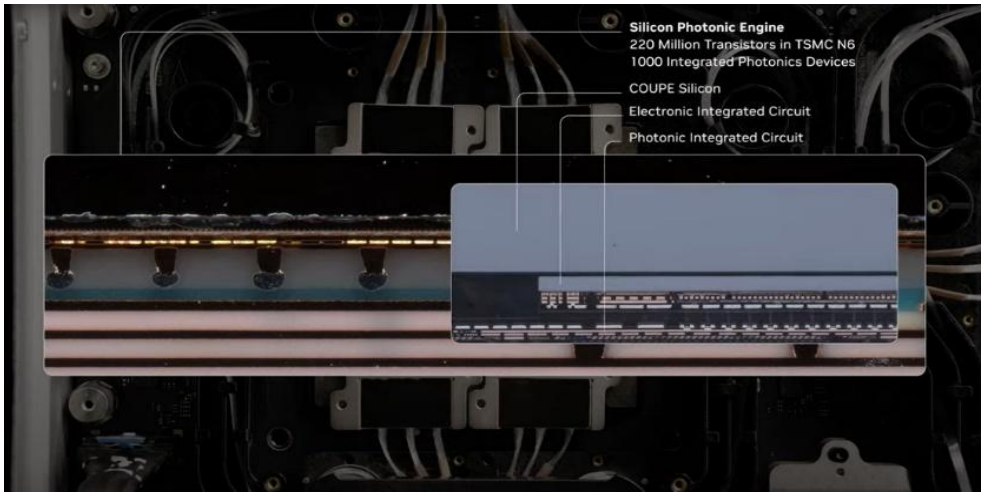

图 3-3 英伟达CPO交换机光引擎截面图[12]

- **英特尔产品及技术方案**

英特尔凭借其成熟的硅光子工艺平台与深厚的可插拔光学的技术积淀，在光互连领域持续引领技术演进。早在2020年，公司便基于硅光工艺搭建起采用微环调制器的CPO样机,依托该调制器低功耗特性奠定技术基础；2023年，通过与光互联独角兽Ayar Labs的生态合作，成功展示将2颗 4Tb带宽TeraPHY OIO芯片嵌入FPGA的异构集成方案，实现高带宽芯片级互连；2024年公布的CPO最新进展，传输速率达 4×64Gb/s，且 具备1.3pJ/bit的低功耗，方案优势显著，为下一代数据中心的高带宽、低能耗连接提供关键支撑。

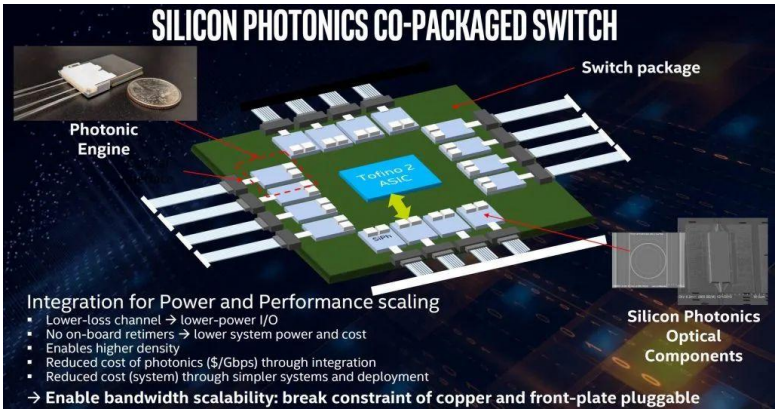

图 3-4 英特尔基于硅光的CPO交换机方案[13]

● Ayar Labs产品及技术方案

Ayar Labs聚焦片间互联场景，结合硅光与Chiplet技术来设计新一代片间互联产品，主要包括TeraPHY（光引擎芯片）和SuperNova（独立激光器），两者常配合使用。其中TeraPHY硅光芯片采用微环调制器，利用多个波长携带信号来提高带宽密度，当前产品通过 8 个光端口实现4096Gbps，每个光口有8个波长，每个波长支持32Gbps，单光口支持256Gbps。

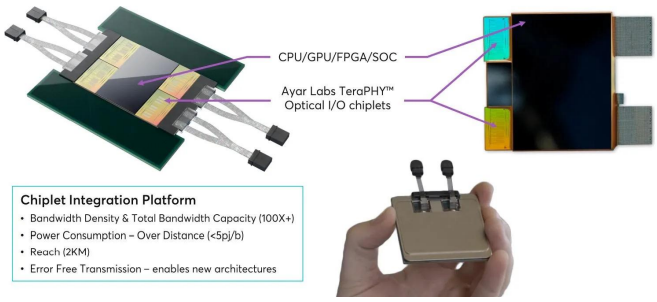

图 3-5 Ayar Labs基于硅光的光引擎和激光方案[14]

Ayar Labs与英特尔在技术研发、产品整合以及投资方面已经展开了合作，Ayar Labs 将其TeraPHY芯片嵌入英特尔的Agilex FPGA产品中，实现了高带宽、低功耗和延时的光FPGA产品。同时，英伟达、AMD也参与了Ayar Labs 2024年的D轮融资。通过应用相关光引擎和激光技术，替代传统电互连，以解决超大规模GPU集群的通信瓶颈。

● Lightmatter产品及技术方案

初创企业Lightmatter推出了基于3D封装技术的CPO产品Passage L200/L200X，其中L200支持56Gbps非归零（NRZ，Non-Return-to-Zero）调制，提供32Tbps的带宽，L200X支持112Gbps四电平脉冲幅度调制（PAM4，Pulse Amplitude Modulation with 4 Levels），提供64Tbps的带宽；接口标准采用UCIe协议，以实现芯片间的灵活通信。光引擎设计方面，Passage PIC与Alphawave Semi公司的EIC产品进行3D集成，可支持320个多协议SerDes。

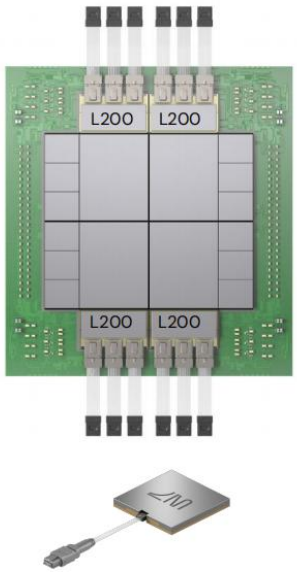

图 3-6 3D CPO光引擎芯片L200示意图[15]

● Avicena产品及技术方案

硅谷初创公司Avicena主推基于Micro-LED的高密度、超低能耗光互连方案，产品线以其模块化的LightBundle平台为核心，定位于Chip-to-Chip/Die-to-Die和板卡间短距互连，尤其面向大规模GPU集群的Scale-Up场景。与基于激光器的硅光方案不同，Avicena使用大量并行μLED光源阵列以“多路低速叠加”的思路实现高总带宽，从而在能耗、成本与可扩展性上提出替代路径。

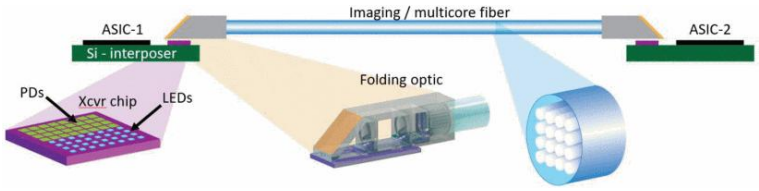

图 3-7 Avicena LightBundle模块架构示意[16]

Avicena的技术主张包括：极高的I/O面密度与极低的每比特能耗。公司公开宣称其LightBundle可实现>1 Tb/s per mm（shoreline）I/O密度，并目标达到sub-pJ/bit级能耗水平；同时系统设计目标覆盖米级到十米级（~10 m）的短距互连，适配数据中心内部及机架间互连需求。该等指标是公司技术路线的核心卖点，用以与传统硅光与VCSEL/激光方案形成差异化竞争。

(三) 产业发展趋势

CPO正从交换侧向算力侧渗透，硅光子集成（如台积电的CoWoS封装）、新材料的应用如薄膜铌酸锂（LNOI）以及片上光源与调制器的异质混合集成等多种技术路线并行发展。科技巨头纷纷发布产品路线图，构建从芯片设计、制造、封装到组装的完整产业链，市场预计将迎来爆发式增长。预判2026-2027年是800G/1.6T CPO商用化关键期，将优先应用于大型云厂商的智算中心，主要面向超大规模模型训练场景。

3.2. 国内处于从研究向应用转化的起步阶段

(一) 标准化工作进展

2022年3月，由中国计算机互连技术联盟（CCITA）联合电子标准

院及多家企业和科研院所共同制订的《T/CESA 1248-2023 小芯片接口总线技术》、《T/CESA 1266-2023 半导体集成电路 光互连接口技术要求》完成标准草案制定，并分别于2023年3月和2023年9月正式实施。

其中，与光互连密切相关的《T/CESA 1266-2023 半导体集成电路 光互连接口技术要求》是我国自主定义的CPO标准，描述CPO模块的设计技术要求，包括概念说明、电学特性、光学特性、数字管理接口、机械结构设计要求等。标准文本结合数据中心交换机芯片和服务器网卡芯片在单链路100Gbps以上的光电融合设计需求，提供了面向交换侧的1.6Tbps以及面向网卡侧的400Gbps的CPO模块标准，并具体列出了并行和波分两种技术路线， 提出了外置光源的规格需求以及CPO模块结构的相应设计要求。

(二) 产品/技术方案进展

整体而言，国内处于从研究向应用转化的起步阶段。国内光互连技术产业分为上中下游，上游包括光器件与光材料企业，代表企业有：中际旭创、仕佳光子、济南晶正等；中游主要包括光电共封、测试、光引擎设计等环节，头部企业包括曦智科技、奇点光子、图灵量子等；下游主要是整机设备商以及云厂商，整机设备商又分为智算服务器和交换机厂商，其中的头部企业有：华为、中兴通讯、新华三等；云厂商主要以阿里云作为头部代表企业。这些企业已在1.6T硅光可插拔模块和CPO样机方面形成布局，逐步积累产业化经验。部分合作厂商及其产品/技术如下：

● 曦智科技公司解决方案

曦智科技作为国内进入CPO技术领域较早的企业，同国内GPU厂商合作研发了国内首个xPU-CPO光电共封装原型系统，采用超短距接口进行数据传输，产品旨在打破国产工艺中因制程限制导致的带宽瓶颈、

充分利用光互连的高密度、低延迟和长距离特性，构建新一代数据中心通信架构。目前进行交换芯片CPO的开发，预计原型机2026年问世。

图 3-8 xPU-CPU光电共封芯片组件（左）及原型系统（右）

目前，曦智与国内领先的GPU及Switch芯片厂商已开展深度合作；同时与下游制造封装厂及设备厂商，聚力攻克封装制造环节的技术难点，以完善产业链生态为目标。

● **奇点光子智能科技公司解决方案**

奇点光子聚焦芯片间互联的光IO芯粒研发和产业化，目标产品是芯片间光互联光电芯片与光学引擎，可用于GPU-GPU间互联的Scale-Up网络，以及GPU-XPU（如CPU、DPU等）之间PCIe/CXL网络，大幅提高超节点规模。公司核心技术包括高速电芯片技术、高速光芯片技术、高速光电芯片联合封装技术、及高效光耦合技术。

目前第一代产品通过先进封装技术将224G/lane电芯片与光芯片3D堆叠在一起形成一个独立工作的芯粒。该芯粒产品可以以NPO、CPO等形式提供。产品具备32通道共计6.4T带宽水平，能效控制在6pj/bit以内，带宽边密度达到700Gbps/mm以上，达到国际领先水平。

● **图灵量子公司解决方案**

图灵量子布局新一代光子集成材料——铌酸锂薄膜（LNOI）光子芯片技术，以更优的非线性指标、更低的波导传输损耗、更高的电光

调制带宽等优势，替代硅光方案，为未来发展102.4T及更高容量CPO系统提供了可行路径。

图灵量子拥有业界领先的GCS-HiCPO（玻璃基异质集成光电共芯封装）光电共芯封装技术方案。通过在GCS玻璃基Interposer上加工TGV（玻璃通孔），可避免出现短路问题，可靠性高，且大大降低信号传输损耗，电信号带宽可提升至200GHz以上；采用可插拔CPO光纤连接器，极大降低SMT焊接难度和CPO整体维护难度；采用铌酸锂薄膜光子芯片技术，相比硅光具有更大带宽，更低功耗，及更远的互连距离。

● **无锡芯光互连技术研究院解决方案**

无锡芯光互连技术研究院由中国计算机互连技术联盟与无锡市锡山区政府共同发起，中科院计算所协助建设，主要研究 Chiplet 芯片架构、3D Logic、芯片光 I/O 等后摩尔时代集成电路新型技术，2021年底开始布局光互连技术解决方案，为光电芯片混合封装设计提供了全面的技术支持。

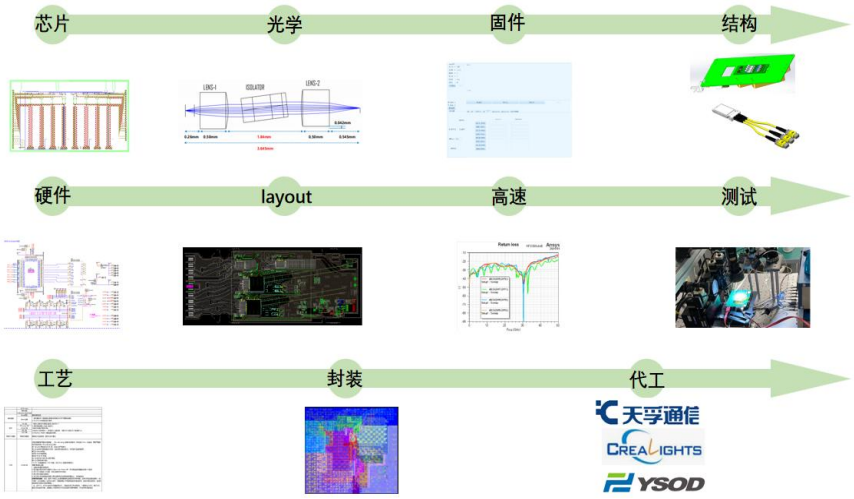

图 3-9 光互连技术解决方案布局

在CPO技术方面，结合可插拔光模块中积累的COB工艺、高精度的SMT贴装工艺和Flip Chip裸片贴装工艺，采用线性直驱技术，开发出4

通道，每通道达到400GBbps的1.6T-CPO首代产品样机。

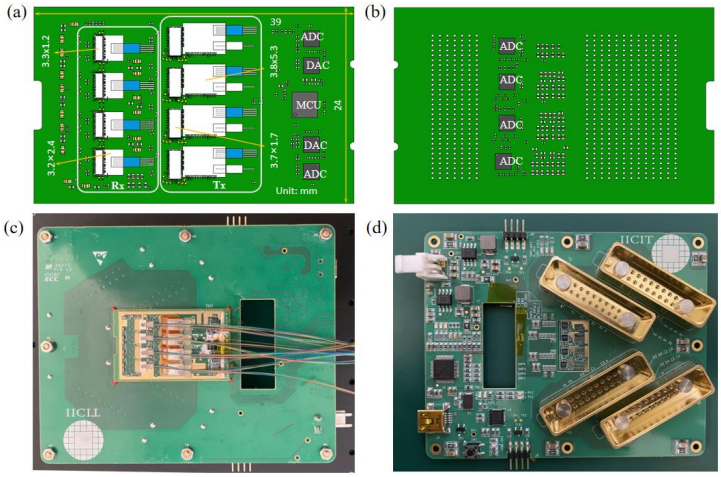

图 3-10 1.6T-CPO首代产品光引擎（LOE）和测试板

● 新华三公司解决方案

新华三聚焦于CPO外置光源的研发，基于IPEC组织的PELS定义，依托其在交换机和光模块领域的深耕经验，投入核心研发力量开发了全新PELS直通型外置光源。

图 3-11 全新PELS外置光源

将传统位于前面板的连接器通道通过Pass-through技术转移到外置光源的前端，使采用高密度直通型PELS外置光源的CPO交换机使用体验更接近传统交换机。

● 清华大学集成光电子实验室解决方案

清华大学集成光电子实验室开发基于Micro-LED阵列的光互连技术，依托其在高速Micro-LED材料外延和芯片工艺领域的深耕经验，演示了多通道并行传输原理样机。

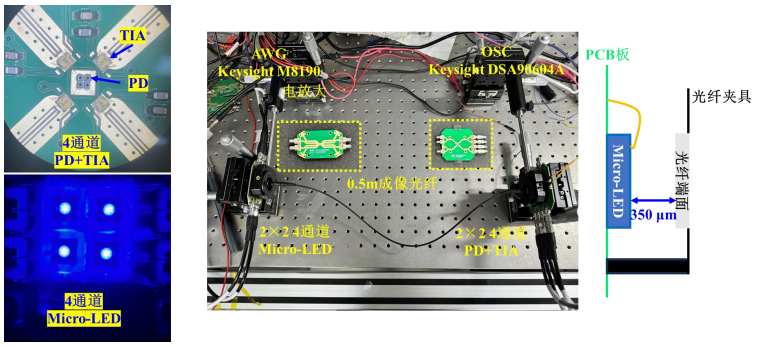

图 3-12 Micro-LED阵列并行传输演示

● 凌云光公司解决方案

凌云光联合HUBER+SUHNER Polatis推出基于OCS的RDCN方案，通过物理层全光线路交换灵活重构网络拓扑，支持8×8至576×576矩阵，让GPU集群真正“跑满算力”，显著提升算力利用率并降低能耗。

在全光互联部署中，凌云光基于DBS（Direct Beam Steering）技术的OCS矩阵光开关产品已累计运行超过188亿端口小时，具备超高可靠性，可满足运营级稳定性需求；同时具备卓越光学性能，典型插损低至2.7 dB、回损优于−50 dB，有效保障链路质量；其超大端口能力最高支持576×576配置，可灵活构建大规模AI集群拓扑。该方案已广泛应用于云计算、AI训练等对性能与可靠性要求极高的场景。

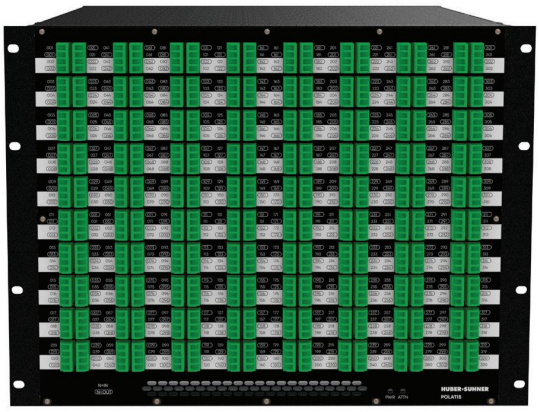

图 3-13 凌云光-HUBER+SUHNER 320x320 HS系列光交换机

(三) 产业发展趋势

我国芯片级光互连技术发展正在起步爬坡期加速蓄力，虽整体产业化进程较国际头部企业晚，但凭借自主技术突破，已展现出强劲后劲。从基础来看，国内已制定相关标准，明确技术路线；企业方面，模块厂家已在可插拔光学领域积累硅光子集成、光电协同封装等经验，多家光引擎初创企业已进入原型验证阶段；产业链上，硅光芯片工艺、高密度封装基板等关键环节正逐步攻克，形成“标准-芯片-设备”的初步闭环。叠加大规模智算集群对高性能互连的刚需拉动，我国芯片级光互连技术正从技术研究向试点商用迈进，未来将在自主可控与规模化应用上的增长潜力持续释放。

4. 规模化应用需跨越技术和产业的双重挑战

设备级光互连和芯片级光互连技术是未来高带宽、低功耗数据中心网络的三大关键技术，然而，这些技术从实验室走向商用落地，仍面临多个方面的挑战。

光互连技术是实现高集成度、低功耗、低成本、小体积的最优互连方案之一，但其产品化受集成光学器件的市场接受度、标准和制造能力的限制，仍面临多个方面的挑战，尤其是在标准、封装工艺、器件性能、散热、仿真测试及良率等维度。

1)芯片级光互连挑战

● 标准化挑战

当前芯片级光互连技术仍处于早期发展阶段，不同厂商提出的解决方案各不相同，尚缺乏统一的接口规范、封装尺寸和散热设计标准，导致产业协同难度较大。尽管IEEE、OIF等标准组织已开始着手制定相关规范，但尚未完全覆盖其中的关键接口，如光引擎、电连接器及光纤接口等，限制了行业的快速推进。此外，光引擎与xPU之间的接口类型（如LGA、BGA等）尚未统一，进一步影响供应链的兼容性与协作效率。与此同时，各家厂商在光路规划、光纤耦合方式（如垂直耦合与平行耦合）等方面也尚未实现标准化，增加了系统集成的复杂性与碎片化风险。

● 封装工艺挑战

CPO涉及到硅通孔（TSV, Through-Silicon Via）、玻璃通孔（TGV, Through-Glass Via）、多层高密度互连基板、Bumping凸点工艺和芯片堆叠等先进封装中的关键技术，每种关键技术都有各自的优缺点，会带来封装的挑战。TGV技术可能会损伤玻璃造成表面不光滑，大多数

TGV加工方法效率低,没法大规模量产,且结构的电镀成本和时间比TSV略高，玻璃衬底表面的黏附性较差，容易导致金属层（RDL, Redistribution Layer）异常，玻璃本身的易碎性和化学惰性给工艺开发带来了难度；TSV的通孔加工、孔填充都有较高的工艺要求，此外还涉及到晶圆减薄，存在潜在的成品率和可靠性的问题。

● 器件性能

光引擎中的光器件技术也有一些需要克服的技术难题。如硅光技术还有一些需要克服的技术难题，一是硅波导的损耗控制难。硅波导的损耗主要源于材料本征吸收、界面散射及工艺缺陷等，需通过优化硅材料掺杂浓度、提升波导抛光精度等方式改进，仍面临工艺成本与性能平衡的挑战；二是波导与光纤耦合难题。硅波导与光纤的模场尺寸失配，会导致单通道耦合损耗达 1-2dB，且密集排布下的串扰风险增加，需依赖超精密对准工艺，否则会降低整体链路效率；三是温度稳定性问题。CPO中xPU芯片与硅光引擎集成度高，芯片运行时的局部温度可达80℃以上，会导致微环调制器折射率变化，引发谐振波长漂移，直接造成信号调制效率下降、误码率升高。

● 系统散热问题

光电共封的高密度集成特性也带来了散热技术的多重挑战。一是是高热密度与共封空间的矛盾。同一封装内部局部热密度极高，而封装内部狭小空间难以容纳传统散热结构，需微型化冷板设计，且要平衡漏液风险与散热效能;二是异构集成材料热失配引发的可靠性风险。片上光源集成也是CPO技术的一个未来发展趋势,但硅基芯片、激光器、封装基板等材料热膨胀系数差异显著，散热过程中温度波动会产生热应力，可能导致基板翘曲，影响光耦合和电子互连的性能，直接影响光信号传输效率。

● 仿真与建模挑战

随着集成密度的不断提高，为提高产品的一次设计成功率，仿真技术在芯片级光互连技术设计阶段的应用将显著提升。仿真建模面临显著挑战，首先是对多物理场仿真的高要求。CPO涉及电、光、热、机械等多种等交叉学科的融合和多层级的跨越，这对仿真工具的协同能力提出了极高要求。此外，目前光器件的建模标准化程度较低，缺乏统一的工艺设计套件（PDK, Process Design Kits）用于系统级设计，限制了设计效率和可靠性。相比之下，光子仿真软件尚不如电子仿真工具成熟，存在工具集成度低、功能分散、学习曲线陡峭等问题，进一步加剧了仿真建模的复杂性。

● 测试与验证挑战

由于光器件和电器件建立在不同的制造工艺上，具有不同的验证要求，因此给测试带来了诸多挑战。一是光引擎被固定封装在电芯片附近，无法独立更换或测试，导致测试点受限，在线调试与故障定位变得困难，显著增加了系统调试和运维的复杂度。二是CPO普遍支持高速率通道，使高速电信号的测试难度和成本大幅上升，测试设备必须具备极高的带宽和精度。三是集成系统中的光路不可见、不可拆卸，光信号的插损、反射和耦合效率等关键参数难以直接评估，对测试方案提出更高要求。目前，自动化测试流程和高精度光学探针技术仍在发展中，尚未能完全满足芯片级光互连产品的量产测试需求。

为推动芯片级光互连技术的商用落地，需要从多个维度系统应对当前面临的挑战。在标准化方面，应加快推动IEEE、OIF等行业组织制定统一的CPO接口与封装标准，解决互通性问题；在工艺封装上，需发展高密度、高精度的2.5D/3D封装技术，并配套高效的散热方案以满足功耗密度要求；仿真建模方面，应建立统一的PDK体系，推动电-光-

热多物理场协同仿真工具的开发；在测试验证层面，需设计专用的测试平台与自动化在线诊断机制，以提升测试效率和系统可维护性；针对可靠性问题，可通过引入冗余设计和可替换的光引擎结构来降低运维成本；同时，生态建设方面要加强上下游协同，打通芯片、光器件、封装、测试等环节，共同构建端到端的产业生态闭环。

2)设备级光互连挑战

除芯片级光互连技术外,设备级光互连技术如LPO和光交换虽技术发展早，但仍面临诸多难题。

- **光交换技术**

尽管光交换展现出诸多优势,其在实际应用中仍然存在一些挑战。首先，由于缺乏O-E-O再生功能，链路必须承受更高的光损耗，对光模块提出更高的性能要求。在一些场景下，还需要依赖高性能甚至相干光模块以维持传输质量。其次，光交换只能在端口间建立整路光连接，交换粒度相对较粗，难以实现电交换在包级或子波长级的灵活调度。可靠性也是重要考量，核心光交换设备一旦发生故障，可能导致大范围连接中断，因此OCS必须具备极高的可靠性以降低风险。

此外,传统OCS切换时间通常在毫秒级,适用于静态或低频率网络重配置，业界仍在积极探索新型光开关技术，推动切换时间向微秒甚至纳秒级演进。控制与编排的复杂性也不容忽视，光交换缺乏电交换原生的控制功能，需要依赖额外的软件层实现集中调度和动态拓扑管理。高可靠性、低损耗和大端口数的设计要求也对设备制造提出了更高挑战。

从组网角度看，智算中心引入光交换技术可带来跨代融合组网、降低成本、降低功耗、提升性能等优势，但也引入切换灵活性差，可

重构性受限等现实挑战。同时，随着光电协同网络作为新一代数据中心架构的引入，传统分层协议栈面临新的适配挑战，这一挑战贯穿协议各层，为充分发挥光电协同架构潜力，软件协议栈各层需要适配光电组网拓扑。

目前业界尚未系统覆盖光交换技术的控制接口、路由协议、跨层协同等关键环节，要实现商用落地仍面临很多挑战：1、物理层与控制接口标准化缺失；2、链路层资源调度机制尚未统一；3、路由与拓扑管理协议适配不足；4、传输层与应用层协同接口缺位；5、多厂商互操作性与协议兼容性挑战；6、测试验证与性能评估标准空白。

在光交换机（OCS）组网架构设计中，仿真面临以下挑战：首先是现有的智算业务仿真系统缺少光交换组件，且现有的通信库不具备适配光电协同组网能力，需要仿真系统同时具备光电协同拓扑仿真能力和智算业务仿真能力，对仿真系统提出了更高要求。

总体来看，光交换在突破电交换瓶颈方面展现出巨大潜力，但要实现大规模商用，还需在链路预算、可靠性、控制复杂性、光电协同等方面持续优化与突破。

- **线性直驱（LPO）技术**

LPO光模块依赖于交换机ASIC芯片SerDes来直接驱动光模块的传输器件，如激光器或调制器驱动器，并恢复直接从光接收器TIA接收到的信号，而无需在光模块内额外进行任何数字信号处理。与重定时光模块不同，LPO光模块中没有DSP作为电气和光接口的分界点，因此LPO光模块通道任何部分出现的信号劣化都会传输到链路末端。

LPO通道受到线性效应和非线性效应的双重影响，由光器件的非线性以及链路中大量信号转换/连接引起的反射造成。这些因素会增加噪

声并产生符号间干扰（ISI），从而严重影响信号完整性和整体系统性能，因此必须进行精心设计和测试，以确保链路稳健可靠。由于LPO光模块的性能在很大程度上取决于交换芯片中使用的SerDes，因此SerDes的质量和功能对于管理信号完整性和补偿LPO通道带来的损伤至关重要。这就使得交换机ASIC芯片SerDes必须具有功能强大的均衡器，可以补偿所有损伤和信号劣化，包括反射和噪声，交换机和LPO光模块将形成强耦合关系，交换机ASIC芯片Serdes性能对LPO技术的实际应用将起到决定性作用。

当前LPO技术仍处于探索阶段，兼容性及标准化问题有待持续研究评估。由于线性直驱方案与传统重定时光模块有较大区别，且不同厂商提出的解决方案各不相同，所以LPO光模块需要根据交换机的ASIC芯片能力进行参数调优，并与异厂家不同技术方案的光模块进行互联互通测试。

5. 呼吁产学研擘画一贯式全光互连产业蓝图

在“光进铜退”的发展趋势下，基于先进封装的芯片级光互连技术已然成为全球智算产业的焦点。相较于海外市场已形成了完整生态，国内芯片级光互连产业尚未形成整合态势，缺乏核心牵引力量。从企业参与情况来看，少数企业（如华为等）较早启动硅光技术端到端布局，覆盖芯片、模块及系统环节；其余计算/交换芯片厂商在该技术路线上起步较晚，尚未大规模投入；设备端厂商入局较少，实际参与度有限。在制造端，如硅光芯片流片、异质集成封装等技术协同性与工艺成熟度仍需进一步拉通与培育。当前阶段，产业内相关技术进展主要由光器件厂商推动，光引擎厂商为核心参与主体。

结合国内技术和产业现状，需系统角色入局构建产业平台，整合资源孵化产业链，协同推进芯片级光互连技术实现从0到1的突破。中国移动作为算力网络新发展理念的引领者和实践者，已面向智算产业卡间互连领域和机间互联领域分别提出**全向智感互连技术体系**（OISA, Omni-directional Intelligent Sensing Express Architecture）和全调度以太网技术体系（GSE，Global Scheduling Ethernet），期望以此为基座，与业界共进，推动光互连技术整合，向芯片、设备、集群三层维度深化，构建面向大规模智算集群场景的一贯式全光超节点系统架构。

中国移动一贯式全光互连系统方案按照“三步走”推进，计划如下：

第一阶段以NPO为核心技术，快速开发相关原型设备，采用成熟的光引擎与封装工艺，为高集成度的光互连技术的部署奠定基础；同步推进CPO/OIO技术的光引擎研发，目标实现主流的链路速率与能效，联合产业链伙伴构建系列标准。同时以探索OCS在智算集群光互连中的应

用场景和组网可行性为目标，开展光电协同组网仿真验证，为后续光电协同网络协议栈设计提供仿真基础；

第二阶段在智算场景中开展光互连原型系统的验证，包括链路损耗、带宽、时延、功耗、稳定性等关键指标；同步推进CPO样机验证平台搭建，联合产业链伙伴在冷却、维护、封装和自动测试等方面取得突破，初步引入超节点系统。并开展基于OCS技术的光电协同网络原型验证，构建开放、统一的光电协同系列标准，涵盖物理接口、性能测试、路由协议等方面，初步引入智算中心网络；

第三阶段试点验证基于高集成度光互连技术的超节点方案，可引入OCS等配合技术，实现全光传输路径和多任务训练负载的高速互通，支持1.6T及以上速率、并提升每瓦通信能效，实现一贯式全光互连超节点技术愿景；打通标准和产业链生态，实现从研究探索到产业落地的完整过渡。试点验证基于光电协同方案，构建更低时延、更大规模、更高性能、更低功耗的光电协同网络，满足未来智算部署需求。

缩略语列表

缩略语	英文全称	中文解释
AI	Artificial Intelligence	人工智能
AGI	Artificial General Intelligent	通用人工智能
APD	Avalanche Photo Diode	雪崩光电二极管
ASIC	Application-Specific Integrated Circuit	应用特定集成电路
CMOS	Complementary Metal Oxide Semiconductor	互补金属氧化物半导体
CPO	Co-Packaged Optics	共封装光学
DAC	Direct Attach Cable	无源直连铜缆
DBS	Direct Beam Steering	直连光束扫描
DCI	Data Center Interconnect	数据中心间互连
DSP	Digital Signal Processer	数字信号处理器
DPU	Data Processing Unit	数据处理单元
DRV	Driver	驱动电路
EIC	Electronic Integrated Circuit	电集成电路
ELS	External Laser Source	外置光源
EP	Expert Parallelism	专家并行
GPU	Graphics Processing Unit	图形处理单元

GSE	Global Scheduling Ethernet	中国移动提出的全调度以太网技术体系
IB	InfiniBand	无限带宽技术
MEMS	Micro-Electro-Mechanical System	微机电系统
Micro-LED	Micro Light-Emitting Diode	微型发光二极管
MRM	Micro Ring Modulator	微环调制器
MZM	Mach-Zehnder Modulator	马赫 – 曾德尔调制器
MTBF	Mean Time Between Failures	平均无故障时间
NPO	Near Package Optics	近封装光学
NRZ	Non-Return-to-Zero	非归零编码
NPU	Neural-network Processing Unit	嵌入式神经网络处理单元
OCS	Optical Circuit Switching	光交换
OE	Optical Engine	光引擎
OIO	Optical Input Output	光输入输出
OISA	Omni-directional Intelligent Sensing Express Architecture	中国移动提出的全向智感互连技术体
PCB	Printed Circuit Board	印刷电路板
PIC	Photonic Integrated Circuit	光集成电路
PIN	Positive-Intrinsic-Negative	PIN 二极管

PAM4	Pulse Amplitude Modulation with 4 Levels	四电平脉冲幅度调制
PDK	Process Design Kits	工艺设计套件
PUE	Power Usage Effectiveness	电源使用效率
RDCN	Reconfigurable Data Center Networks	可重构数据中心网络
RDL	Redistribution Layer	重分布层。集成电路封装设计中的一个重要层次，主要用于实现芯片内电气连接的重新分配。
RoCE	RDMA over Converged Ethernet	融合以太网 RDMA 技术
TGA	Through Glass Via	玻璃穿孔技术
TGV	Through-Glass Via	玻璃通孔
TIA	Transimpedance Amplifier	跨阻放大器
TIM	Thermal Interface Material	热界面材料
TOSA	Transmitter Optical Subassembly	发光组件
TSV	Through-Silicon Via	硅通孔
UCIe	Universal Chiplet Interconnect Express	统一芯片互连扩展

VCSEL	Vertical-Cavity Surface-Emitting Laser	垂直腔面发射激光器
xPU	GPU, NPU, Swtich, etc	多种处理单元

参考文献

[1] Nubis Communication: https://doi.org/10.1364/OE.555476

[2] 博通技术报告

[3] D Tonietto, OFC 2024, W4H.2

[4] Nat Rev. Phys. 5, 717–734 (2023)

[5] NVL72 的部署案例

[6] 博通技术报告

[7] Courtesy of Microsoft

[8] Cammarata, S., et al. "Compact silicon photonic Mach–Zehnder modulators for high–energy physics." Journal of Instrumentation 19.03 (2024): C03009.

[9] S. Mondal et al., "18.2 A 4x64Gb/s NRZ 1.3pJ/b Co–Packaged and Fiber–Terminated 4–Ch VCSEL–Based Optical Transmitter," 2024 IEEE International Solid–State Circuits Conference (ISSCC), San Francisco, CA, USA, 2024, pp. 340–342, doi: 10.1109/ISSCC49657.2024.10454455.

[10] 博通技术报告

[11] GTC2025

[12] GTC2025

[13] Intel 官方网站

[14] Ayar Labs 官方网站

[15] Lightmatter 技术报告

[16] Avicena 官方网站