

三条路径解读“光进铜退”

——访曦智科技创始人、CEO 沈亦晨

本报记者 姬晓婷 实习生 刘璐

当大模型参数突破万亿，以铜为介质构建的算力互联方案同时撞上了带宽、延迟和功耗“三堵墙”，而“光”正成为化解上述“三堵墙”的破局之道。以光为介质构建的算力互联基础设施，将给光互连、光交换、光计算三条赛道带来新的想象空间。

那么，“光进铜退”的进展究竟如何？近日，曦智科技创始人、CEO 沈亦晨对“光进铜退”的三条路径进行了细致解读。

光互连：NPO 将迎落地元年

如今市面上的数据中心九成以上依旧采用铜线连接。然而，随着数据中心内部数据传输需求不断提升，铜导线带宽不足、延迟高、功耗大等核心痛点愈加明显。在沈亦晨看来，算力中心互连“铜转光”是必然趋势。不仅如此，算力光互连未来的市场体量庞大，至少是传统光纤传输行业的10倍，成长空间十分广阔。由于当前行业尚未形成统一标准，不同AI芯片厂商需求各有差异，定制化落地机会充足，创新企业更容易打造独有的竞争优势；相比传统赛道，这一赛道比拼的并非基础代工制造，而是一整套芯片系统综合设计能力，技术壁垒更高。

可插拔光模块是数据中心互连网络“光代铜”的第一种形态。光模块作为电信号与光信号之间的转换器，连接服务器板卡与光纤，使数据传输的距离得以延长。但光模块只是计算互连网络“光代铜”的“初级形态”。传统可插拔光模块置于服务器远端，需搭载博通、迈威尔等企业供应的DSP芯片进行信号纠错。该芯片成本高、功耗大，还会增加传输时延。

CPO(光电共封装)因具备超低功耗、高带宽密度、能实现信号完整性等特征，是行业中公认的能够解决光模块方案现存问题的解决方案。它将光学引擎与交换或计算芯片(如ASIC、GPU)封装在同一基板上，将电信号传输距离从厘米级缩短到毫米级，大幅减



少长距离电线带来的电力损耗与发热，能效比传统方案提升数倍；超短电路路径最大程度减少高频信号衰减与失真，保障极高传输速率下的稳定性。

但受封装良率、供应链等因素制约，CPO方案暂未实现大规模量产。

在产业链探索CPO的过程中，作为由光模块方案向CPO过渡的近封装光学(NPO)将有望在未来两年内实现批量应用。NPO的设计思路是将光电转换的芯片和计算芯片放在同一个板卡上，在省去DSP芯片的同时大幅缩短铜线传输距离。沈亦晨表示，未来12~24个月，主流的互联网大厂将逐步开始采用NPO方案，待放量后海内外NPO光芯片市场规模可达百万至千万颗级别。

光交换：国产方案后发跟进

如果说光互连解决的是“芯片之间传数据太慢”的问题，那么，光交换解决的则是“大规模集群怎么灵活调度”的问题。电交换需要反复将光信号转换成电信号，处理后再转回光信号，延迟更长，功耗更高。光交换可省去光电转换环节，传输速度更快、能耗更低。

国内外光交换赛道的发展存在明显代差。作为在光交换赛道布局较早的海外领军企业，谷歌在2012年便推出了微秒级光路交换机(OCS)原型Mordia，现已将OCS光交换技术深度整合进TPU超级计算机集群中。相比海外，国内缺少垂直整合光交换技

术的领军企业，整体落后近十年。

今年WAIC上，曦智科技与合作伙伴共同推出的基于光互连光交换方案的光跃超节点实现了数千卡落地，成为国内首个国产、商业化的光互连光交换千卡算力集群。

在沈亦晨看来，未来，光交换方案存在巨大发展潜力。“在大模型发展初期，由于市场上应用的模型变化很快，电交换是一种更灵活的部署方案，但电交换方案的效率相对更低。”但现在，模型快速迭代的阶段已经过去，市场主流模型逐渐确立，已经进入更高效的产出Token的新阶段。在这个阶段，用更经济的方式产出更多高质量Token是AI基础设施厂商的核心目标，更看重AI基础设施的效率和成本。而光交换方案在实现高效率、低成本方面具有天然优势。一方面

能够减少数据中心一半的光模块用量，另一方面能够实现更低的时延。“虽然采用光交换方案会使数据中心丧失一定的灵活性，但随着更多为大模型推理定制的算力中心方案出现，光交换会有更大的市场。”沈亦晨说道。

光计算：进入市场渗透期

光计算依靠光子替代电子完成AI矩阵运算，具备低功耗、高并行的算力优势。从整体来看，我国在光计算研发投入、产品成熟度、商业化落地进度上均领先海外市场。近两年国内新增20余家光计算相关创业公司，今年WAIC大会期间，曦智、光本位、每刻深思等多家企业展出光计算硬件，路线各有不同：曦智主打硅光集成路线，光本位发力光子存内计算，每刻深思深耕空间光计算芯片。

沈亦晨给光计算芯片划定了产业发展的不同阶段：2017年以前属于实验室理论研究阶段；2017年至2026年为产品突破阶段；2027年后，光计算芯片将进入市场渗透期，将在适合部署的垂直领域实现批量落地。

沈亦晨认为，相比大模型训练，光芯片将优先在推理领域部署。在推理赛道跑通量产与商业化闭环、积累成熟方案后，再逐步探索训练场景。

基于光计算芯片本身具备的低延时特征，该类芯片将在如视觉智能、大模型、具身智能、AI for Science、金融量化、太空算力等对时延敏感的垂直场景优先布局。

下一步，光计算芯片在商业推广过程中将面临生态难题。沈亦晨表示，光计算芯片真正实现商用，需要配套规模庞大的软件团队、构建全栈系统开发能力、具备针对行业场景的专属算法，并投入不亚于传统GPU厂商的研发人力。但同时他也认为，光计算真正的市场空间并不是分流目前GPU的应用市场，而是会涌现出新的、针对光计算硬件开发的杀手级应用，为光计算开辟新的商业空间。目前来看，光计算芯片要实现规模化商用仍需两到三年时间。

稳增长 强创新

促融合 优治理 防风险

确保实现“十五五”良好开局

中国电子报 宣

公益广告