共封装光学器件（CPO）手册——以光为介质实现下一代互连技术扩展

横向扩展与纵向扩展的CPO技术、CPO总拥有成本与功耗预算、DSP收发器与LPO/NPO/CPO对比、台积电COUPE工艺、MZM/MRM/EAM调制器深度解析、CPO核心企业与供应链布局

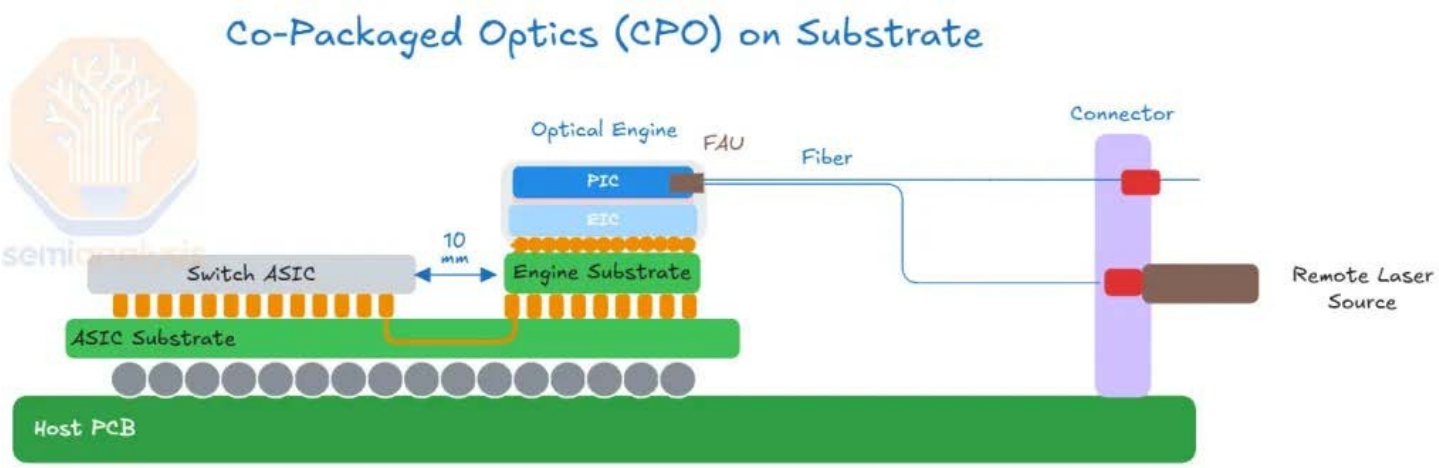
DYLAN PATEL、DANIEL NISHBALL、MYRON XIE 及另外3位作者 2026年1月2日 · 付费内容

共封装光器件（CPO）技术长期以来被寄予厚望，有望彻底改变数据中心互联格局，但该技术历经漫长周期才得以面世，真正具备部署条件的成熟产品直至2025年才问世。在此期间，可插拔收发器凭借其相对成本效益、部署便捷性以及基于标准的互操作性，始终满足网络需求并保持着主流地位。

然而，人工智能工作负载带来的高网络需求意味着此次情况不同。人工智能网络带宽的发展路线图表明，互连速度、覆盖范围、密度和可靠性要求很快将超越收发器所能提供的水平。固态光子学（CPO）将带来一定效益，为横向扩展网络提供更多选择，但它将成为纵向扩展网络的核心技术。**在本十年后期及之后，CPO将成为纵向扩展网络带宽增长的主要驱动力。**

当前基于铜缆的扩展解决方案（如NVLink）可提供高达7.2 Tbit/s的单GPU带宽——鲁宾架构时代将提升至14.4 Tbit/s。然而铜缆链路的传输距离上限仅为两米，这意味着扩展域的规模最多只能覆盖一两个机架。此外，通过铜缆提升带宽的难度正日益增加。 在Rubin架构中，NVIDIA将通过双向SerDes技术使每铜缆通道带宽再翻倍。但依靠开发更高速SerDes来提升铜缆带宽的扩容路径充满挑战，进展缓慢。**而CPO技术不仅能实现同等甚至更优的带宽密度，更能提供多元化的带宽扩容路径，同时支持更大规模的扩展域。**

要理解CPO技术诞生的动因，首先需审视光通信中使用收发器时存在的诸多低效与权衡。收发器虽能延长链路距离，但网络交换机或计算托盘前面板上用于插拔收发器的插槽，通常距离XPU或交换机ASIC有15-30厘米之遥。 这意味着信号必须先通过LR串解码器在15-30厘米距离内进行电传输，再由收发器内的数字信号处理器（DSP）恢复并调理电信号，最后转换为光信号。 采用CPO方案时，光引擎直接部署于XPU或交换机ASIC旁，从而省去DSP环节，并可使用低功耗SerDes将数据从XPU传输至光引擎。相较于DSP收发器，该方案可降低50%以上的数据传输能耗——许多方案甚至致力于将每比特能耗降低80%。



来源：SemiAnalysis

尽管像英伟达和博通这样的横向扩展型CPO解决方案正获得更多关注，并受到终端客户的密切关注，但主要超大规模企业已开始规划其纵向扩展型CPO战略，并向供应商作出承诺。 例如Celestial AI预计到2028年底可实现10亿美元营收规模——我们认为这主要得益于其与亚马逊Trainium 4协同推出的纵向扩展CPO解决方案。

专注于CPO的企业现已超越论文、试点项目和演示阶段，正就光端口架构等关键产品决策展开布局，以解决大规模量产难题。对于扩展型CPO**而言，问题已不再是"是否采用"和"为何采用"，而是"何时实现"和"如何**推进"——如何推动这些系统进入量产阶段，以及何时能确保激光器等关键组件的供应链稳定。

制造商能够扩大足够的生产规模。

本文将深入探讨CPO的优势与挑战、CPO架构的工作原理、当前及未来的CPO产品、专注于CPO的企业、CPO相关组件及其各自的供应链。本文旨在为从业者、行业分析师、投资者以及所有对互连技术感兴趣的人士提供指南。

目录与阅读指南

本文分为五个部分——读者可根据自身兴趣或关联度选择重点阅读章节。

在**第一部分：CPO总拥有成本（TCO）分析**中，我们将首先探讨采用CPO如何改变横向扩展与纵向扩展网络的总拥有成本。我们认为，在横向扩展网络中采用CPO时，总拥有成本、可靠性及设备供应商议价能力将是主要考量因素。 我们将探讨CPO技术是否已具备在横向扩展场景中的成熟应用条件，并结合现有解决方案可靠性数据进行论证，**例如Meta**公司在ECOC 2025大会上发布的**CPO横向扩展交换机研究成果**。

在**第二部分：CPO 介绍与实施**中，我们将深入探讨 CPO 的工作原理。 本节将探讨市场从**铜缆到共封装铜缆**的演进历程，从**数字信号处理器（DSP）光模块到线性可插拔光模块（LPO）再到CPO**的发展轨迹，并剖析推动CPO应用的驱动力与核心论据。同时将深入分析**串行器/解串器（SerDes）的扩展极限，以及宽带I**/O作为SerDes替代方案的潜力——尤其当其与CPO协同应用时所展现的优势。

在**第三部分：推动CPO进入**市场中，我们将阐述助力CPO获得市场认可并实现商业化的关键技术。首先探讨主机与光引擎封装技术，详细解析**台积电COUPE**工艺及其成为首选集成方案的原因。**光纤连接单元（FAU）、光纤耦合**技术以及**边缘耦**合器与**光栅耦合器的**对比将得到全面阐述。 我们将涵盖调制器类型，包括**马赫-曾德尔调制器（MZM）、微环调制器（MRM）和电吸收调制器（EAM）**。本节将以阐释光纤光子学被广泛采用的核心原因收尾——即通过光纤光子学实现**带宽扩展的多重路径**：增加光纤连接数量、采用波分复用技术

波分复用（WDM）和高阶调制。

在**第四部分：当前与未来的CPO产品**中，我们将分析现今市场上可用的CPO产品及其相关供应链。我们将从**英伟达和博通的解决方案**开始，随后探讨主要CPO企业。 我们将深入**剖析Ayar Labs、Nubis、Celestial AI、Lightmatter、Xscape Photonics、Ranovus及Scintil等企业的**解决方案，详细阐述各供应商的技术方案，并针对每家公司的技术路线进行利弊权衡。

最后，在**第五部分：英伟达CPO供应链**中，我们将详细描述英伟达CPO生态系统的供应链，列出**激光光源、ELS模块、光纤接头（FAU）、FAU点灯工具、FAU组装、切换盒、MPO连接器、MT插芯、光纤以及电光测试等**关键供应商，以此为本报告作结。

第一部分：CPO总拥有成本（TCO）分析

今年早些时候NVIDIA GTC 2025大会上最受瞩目的议题之一，是詹森宣布推出公司首款支持CPO技术的横向扩展网络交换机。

值得注意的是，在纵向扩展领域，英伟达仍坚持推进铜缆技术，并竭力避免转向光纤方案——这一策略将持续至2027年乃至2028年。

让我们从探讨这些新型CPO交换机的总体拥有成本入手，分析横向扩展型CPO技术能带来的成本与能耗节约效益。

英伟达在GTC 2025主题演讲中发布了三款采用不同CPO交换机ASIC的横向扩展交换机。尽管其具备TCO、功耗和部署速度优势，但这些优势尚不足以促使客户立即转向全新的部署模式，因此我们预计首批CPO横向扩展交换机的采用率将较为有限。具体原因如下：

Nvidia CPO Roadmap			
Switch Model	**Quantum 3450 CPO**	**Spectrum 6810 CPO**	**Spectrum 6800 CPO**
Launch Date	2H 2025	2H 2026	
Networking Standard	InfiniBand	Ethernet	
Switch ASIC	Quantum-3	Spectrum-6	
Throughout per Package	28.8 Tbps	102.4 Tbps	
Number of Switch Packages	4	1	4
Switch Aggregate Bandwidth	115.2 Tbps (not all-to-all)	102.4 Tbps	409.6 Tbps (not all-to-all)
SerDes speed (Gb/s uni-di)	200 Gbps	200 Gbps	
Optical Connectivity	DR Optics	DR Optics	
Physical MPO Ports	144	128	512
Bandwidth and Logical Port Configurations Available	144 Ports of 800G	512 Ports of 200G 256 Ports of 400G 128 Ports of 800G	512 Ports of 800G
Bandwidth per Optical Engine (OE)	1.6 Tbps	3.2 Tbps	
Number of OEs	72	32	128
External Light Sources (ELSs)	18	16	64
For Spectrum CPO there are 36 OEs on the package, but only 32 OEs are enabled			

来源：SemiAnalysis

典型AI集群网络架构与TCO对比

典型的AI集群包含三种主要网络架构：后端架构、前端架构和带外管理架构。其中使用最频繁且技术要求最高的当属后端架构。该架构用于GPU之间的横向扩展通信，支持GPU在集体操作中相互通信并交换数据，从而实现训练与推理的并行化处理。后端网络通常采用InfiniBand或以太网协议。

由于其高要求特性，后端网络在总网络成本和功耗中占据主导地位：在采用英伟达X800-Q3400后端交换机部署于InfiniBand的3层GB300 NVL72集群中，其网络成本占比达85%，网络功耗占比达86%。基于CPO的交换机和网络解决方案可同时应用于后端与前端网络，但我们认为当前阶段的部署重点仍应放在后端网络。读者可查阅更多关于后端网络拓扑结构、端口配置、交换机及收发器等详细信息。

在我们2024年的《光学怪兽》文章中，以及在我们的[illegible]均有提及

模型。若需了解网络总拥有成本，可参阅我们的《AI Neocloud解剖与实践指南》文章。

英伟达的光学噩梦——NVL72、InfiniBand横向扩展、800G与1.6T 加速

DYLAN PATEL 与 DANIEL NISHBALL · 2024年3月25日

阅读全文 →

AI Neocloud 操作指南与架构解析

迪伦·帕特尔与丹尼尔·尼什鲍尔 · 2024年10月3日

阅读全文 →

缩小范围来看——网络成本是AI集群总成本中仅次于AI服务器本身的第二大组成部分。在采用GB300 NVL72集群搭配三层InfiniBand网络的配置中，网络成本占集群总成本的15%；若升级为四层网络，该比例将升至18%。 光收发器是该成本的重要组成部分，在采用相对昂贵的Nvidia LinkX收发器的三层网络中，其成本占比高达网络总成本的60%。同时，三层网络中光收发器消耗的功耗占网络总功耗的45%。

GB300 NVL72 Cluster Cost and Power Budget Breakdown, per Rack-Scale 72-GPU Server

	2-Layer Network (up to 10,368 GPUs)		3-Layer Network (up to 746,496 GPUs)		4-Layer Network (up to ~53.7M GPUs)	
Item	Cost	Power (W)	Cost	Power (W)	Cost	Power (W)
Server	$4,357,943	142,000	$4,357,943	142,000	$4,357,943	142,000
Optical Transceivers	$363,125	4,600	$499,925	6,184	$568,325	6,976
Switches	$208,681	5,214	$312,681	8,014	$416,681	11,606
Fiber, Cables, Software, Others	$17,723	0	$24,338	0	$30,953	0
Networking	$589,529	9,814	$836,944	14,198	$1,015,959	18,582
All Others	$269,152	281	$269,152	281	$269,152	281
Total Power	**$5,216,624**	**152,095**	**$5,464,039**	**156,479**	**$5,643,054**	**160,863**

Power for a Neocloud Giant InifiniBand cluster using X800-Q3400 Back-end switches with 144 ports of 800G
Source: SemiAnalysis Networking Model

来源：SemiAnalysis AI网络模型

人工智能集群中的GPU数量越多，就越可能需要增加网络层数。从两层网络扩展到三层及以上网络，意味着更高的成本和更大的功耗预算。CPO技术既能在保持层数不变的情况下降低功耗和成本，又能通过在固定层数的网络中扩展可连接的GPU数量，从而减少总功耗和成本需求。

CPO横向扩展功耗预算

今年早些时候，在GTC 2025大会上，英伟达首席执行官黄仁勋强调，仅收发器本身就消耗了巨大功耗，这是推动CPO技术发展的关键驱动力。 根据上表中的机架功耗预算数据，在三层网络架构下，一个由20万个GB300 NVL72（每机架含72个GPU封装和144个计算芯片）组成的GPU集群将消耗435兆瓦关键IT功耗，其中光收发器单项功耗就高达17兆瓦。 显然，通过大幅减少收发器数量可节省海量能耗。

通过对比单个800G DSP收发器的功耗与CPO系统中光引擎及激光源（按800G带宽计算）的功耗即可轻松验证： 800G DR4光收发器功耗约为16-17瓦，而我们估算Nvidia Q3450 CPO交换机中光引擎与外部激光源的功耗仅为每800G带宽4-5瓦，实现73%的能耗降低。

这些数据与Meta在ECOC 2025会议上发表的论文中所呈现的数据非常接近。 在本报告中，Meta展示了800G 2xFR4可插拔收发器功耗约为15W，而博通Bailly 51.2T CPO交换机内部的光引擎和激光源在传输每800G带宽时功耗仅约5.4W，实现了65%的节能效果。

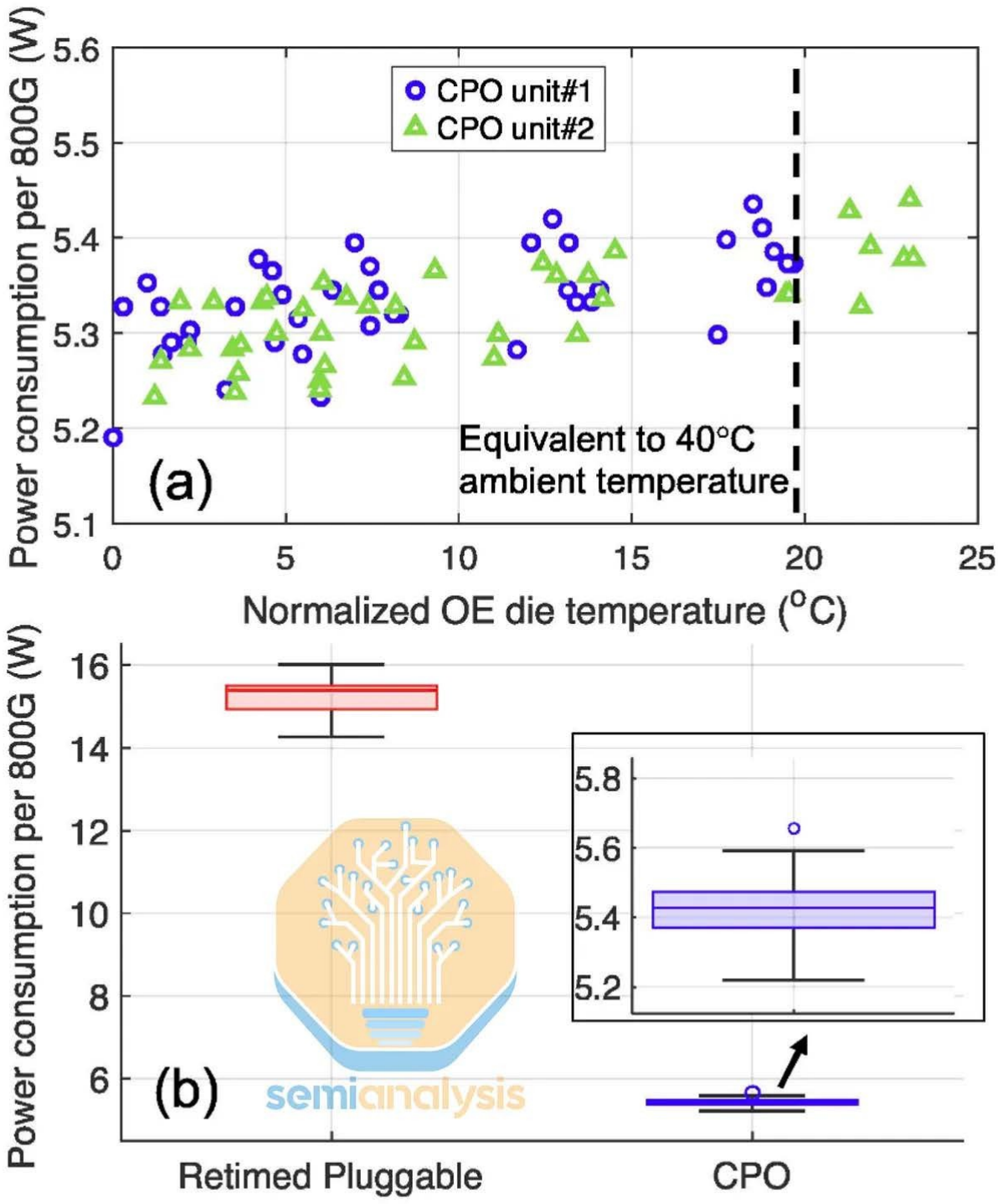


Fig. 1: a) Optics power consumption at different OE die temperatures. b) Power consumption comparison of 2x400Gbps FR4 pluggable and CPO. Inset: magnified CPO power.

来源：Meta

让我们将分析扩展到集群层面。以基于三层网络构建的GB300 NVL72集群为例，可见将后端网络中的DSP收发器替换为LPO收发器后，可降低36%的总收发器功耗和16%的总网络功耗。 若全面过渡至CPO技术，相较于DSP光模块可实现更显著的节能效果——收发器功耗降低84%——但部分节能效益会被新增的光学引擎（OEs）和外部光源（ELSs）所抵消。

交换机，其总功耗现已增加23%。在下例中，CPO场景下的光收发器功耗仍维持在每台服务器1000W的基准值，因为我们假设前端网络仍将采用DSP收发器。

GB300 NVL72 Cluster Power Budget: Traditional vs CPO Network, per Rack-Scale 72-GPU Server								
	DSP Transceivers (3-Layer Network)		LPO Transceivers (3-Layer Network)		CPO (3-Layer Network)		CPO (2-Layer Network)	
Item	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)
Server	142,000	0%	142,000	0%	142,000	0%	142,000	0%
Optical Transceivers	6,184	0%	3,935	-36%	1,000	-84%	1,000	-84%
Switches	8,014	0%	8,014	0%	9,884	23%	6,336	-21%
Fiber, Cables, Software, Others	0	0%	0	0%	0	0%	0	0%
Networking	14,198	0%	11,949	-16%	10,884	-23%	7,336	-48%
All Others	281	0%	281	0%	281	0%	281	0%
Total Power	**156,479**	**0%**	**154,230**	**-1%**	**153,165**	**-2%**	**149,617**	**-4%**

Neocloud Giant InfiniBand cluster using Nvidia x800-Q3400 Back-end switches with 144 ports of 800G and Nvidia LinkX Transcievers. CPO used only in Back-end network.
Source: SemiAnalysis Networking Model

来源：SemiAnalysis人工智能网络模型

采用英伟达的CPO横向扩展交换机意味着默认使用高基数网络，尽管这种机制对终端用户而言是"抽象化"的——因为在CPO交换机内部即可完成端口切换，而使用高基数非CPO交换机时则需通过配线架或八爪鱼线缆在交换机外部进行切换。 相反，这些NVIDIA CPO交换机呈现出极高的端口数量——例如Quantum 3450提供144个800G端口，Spectrum 6800则提供512个800G端口。 我们使用"默认"一词是因为英伟达的非CPO InfiniBand Quantum Q3400交换机同样提供144个800G端口，但其其他InfiniBand交换机（如QM9700）仅提供32个800G端口——唯有前者具备这种"高基数集成"特性，能提供大量有效端口。 如此高的端口密度有望帮助客户将三层网络扁平化为两层网络，同时省去部署交换盒、配线架或笨重八爪鱼线缆的麻烦，这将成为关键卖点。在两层网络架构下，相较传统DSP收发器，收发器功耗降低84%，交换机功耗下降21%，整体网络功耗可减少48%。

Spectrum 6800交换机凭借其在两种可用逻辑配置下均具备的大量端口——512个800G端口——相较于Spectrum 6810（提供128个800G端口、256个400G端口或512个200G端口）实现了这一特性。 采用Spectrum 6810的128个800G端口方案时，两层网络最多可连接8,192个GPU；而Spectrum 6800在512个800G端口配置下，可连接131,072个GPU。

顺带一提，使用交换机

L层网络中k个端口的计算方式如下：

$2(\frac{k}{2})^L$ $2(\frac{512}{2})^2$

其魔力在于端口数量k会随层数呈指数级增长。 因此，对于双层网络，通过将每个端口的带宽减半（例如将800G端口切分为两个400G端口）来实现逻辑端口数量翻倍——无论是采用内部切换（如Spectrum 6800所示）、分支电缆还是双端口收发器——所支持的主机数量将增加四倍！

本节讨论的节能效果——三层CPO网络节省23%，降至两层CPO网络可节省48%——看似惊人，但关键在于：三层网络中，网络功耗仅占集群总功耗的9%。因此，对于横向扩展网络而言，转向CPO技术带来的实际效益最终会被大幅稀释。 三层网络采用CPO可降低23%的网络功耗，但仅带来2%的集群总功耗节省；而两层网络虽能降低48%的网络功耗，集群总功耗节省却仅为4%。

GB300 NVL72 Cluster Power Budget Breakdown, per Rack-Scale 72-GPU Server

	2-Layer Network (up to 10,368 GPUs)		3-Layer Network (up to 746,496 GPUs)		4-Layer Network (up to ~53.7M GPUs)	
Item	Power (W)	%	Power (W)	%	Power (W)	%
Server	142,000	93%	142,000	91%	142,000	88%
Optical Transceivers	4,600	3%	6,184	4%	6,976	4%
Switches	5,214	3%	8,014	5%	11,606	7%
Fiber, Cables, Software, Others	0	0%	0	0%	0	0%
Networking	9,814	6%	14,198	9%	18,582	12%
All Others	281	0%	281	0%	281	0%
Total Power	**152,095**	**100%**	**156,479**	**100%**	**160,863**	**100%**

Power for a Neocloud Giant InifiniBand cluster using X800-Q3400 Back-end switches with 144 ports of 800G
Source: SemiAnalysis Networking Model

来源：SemiAnalysis人工智能网络模型

GB300 NVL72 Cluster Power Budget: Traditional vs CPO Network, per Rack-Scale 72-GPU Server

	DSP Transceivers (3-Layer Network)		LPO Transceivers (3-Layer Network)		CPO (3-Layer Network)		CPO (2-Layer Network)	
Item	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)
Server	142,000	0%	142,000	0%	142,000	0%	142,000	0%
Optical Transceivers	6,184	0%	3,935	-36%	1,000	-84%	1,000	-84%
Switches	8,014	0%	8,014	0%	9,884	23%	6,336	-21%
Fiber, Cables, Software, Others	0	0%	0	0%	0	0%	0	0%
Networking	14,198	0%	11,949	-16%	10,884	-23%	7,336	-48%
All Others	281	0%	281	0%	281	0%	281	0%
Total Power	**156,479**	**0%**	**154,230**	**-1%**	**153,165**	**-2%**	**149,617**	**-4%**

Neocloud Giant InfiniBand cluster using Nvidia x800-Q3400 Back-end switches with 144 ports of 800G and Nvidia LinkX Transcievers. CPO used only in Back-end network.
Source: SemiAnalysis Networking Model

来源：SemiAnalysis AI网络模型

观察集群总资本成本时，情况类似。

CPO横向扩展总拥有成本（TCO）

让我们简要聚焦于收发器与CPO解决方案的成本细节对比。首款Nvidia CPO交换机——Quantum X800-Q3450 CPO将采用72个光引擎，每个运行速率为1.6Tbit/s； 后续版本的Quantum CPO交换机预计将升级为36个光引擎，每个引擎速率达3.2Tbit/s，单价约1000美元（含FAU），这意味着每套系统光引擎总成本高达3.6万美元。

为便于比较，不妨估算采用传统光收发器模块的总成本：非CPO版本的X800-Q3400配备72个OSFP插槽，并通过
采用1.6T双端口收发器实现144个800G端口。假设成本为
通用型1.6T DR8收发器单价1,000美元，该交换机所需收发器总成本将达72,000美元，是实现同等带宽的CPO交换机所需光引擎和ELS组件预估成本（35,000-40,000美元）的两倍。但此计算未计入交换机供应商的利润空间。 若按60%毛利率计算，最终用户的光引擎成本将升至8万至9万美元——高于同等收发器的成本。 此外，光纤切换器等其他组件同样会叠加此类利润层级。这解释了为何根据收发器采购成本及交换机厂商利润率的不同，迁移至CPO方案的成本节约可能并不显著。

如下表所示，在三层网络中将收发器替换为CPO时，CPO组件的额外利润导致交换成本增加81%，这抵消了因不采购收发器而节省的86%成本。 尽管CPO方案的总网络成本仍比DSP收发器方案低31%，但与电源成本类似的情况再次出现：由于服务器机架在集群总拥有成本中占据主导地位，导致集群总成本仅下降3%。

将网络从三层简化为两层可实现更显著的成本节约——集群总成本最高可降低7%，收发器成本下降86%，网络总成本减少46%。

GB300 NVL72 Cluster Capital Costs: Traditional vs CPO Network, per Rack-Scale 72-GPU Server

	DSP Transceivers (3-Layer Network)		LPO Transceivers (3-Layer Network)		CPO (3-Layer Network)		CPO (2-Layer Network)	
Item	Cost	Δ Cost (%)	Cost	Δ Cost (%)	Cost	Δ Cost (%)	Cost	Δ Cost (%)
Server Cost	$ 4,357,943	0%	$ 4,357,943	0%	$ 4,357,943	0%	$ 4,357,943	0%
Optical Transceivers	$ 499,925	0%	$ 434,146	-13%	$ 71,525	-86%	$ 71,525	-86%
Switches	$ 312,681	0%	$ 312,681	0%	$ 564,821	81%	$ 359,965	15%
Fiber, Cables, Software, Others	$ 24,338	0%	$ 24,338	0%	$ 24,338	0%	$ 17,723	-27%
Networking Cost	$ 836,944	0%	$ 771,164	-8%	$ 660,684	-21%	$ 449,213	-46%
All Others	$ 269,152	0%	$ 269,152	0%	$ 269,152	0%	$ 269,152	0%
Total Cost	**$ 5,464,039**	**0%**	**$ 5,398,259**	**-1%**	**$ 5,287,779**	**-3%**	**$ 5,076,308**	**-7%**

Neocloud Giant InfiniBand cluster using Nvidia x800-Q3400 Back-end switches with 144 ports of 800G and Nvidia LinkX Transcievers. CPO used only in Back-end network.
Source: SemiAnalysis Networking Model

来源：SemiAnalysis AI 网络模型

那么——既然CPO方案一方面仅能带来最高7%的成本节约和4%的功耗降低，另一方面却引发了现场维护难度增加、可靠性与影响范围的担忧（无论是否合理），以及失去多供应商议价能力的顾虑，为何GPU云仍选择采用它？ 简而言之，该技术尚未广泛普及——我们预计短期内超大规模企业对横向扩展CPO系统的采用曲线不会陡峭。

适用于纵向扩展网络的**CPO**方案

相反，我们认为用于扩展的CPO才是杀手级应用。如前所述，主要超大规模供应商已承诺在本十年末前部署基于CPO的扩展解决方案。

当前，基于铜缆的现有扩展范式正面临极限挑战：铜缆传输距离受限——单通道200Gbit/s速率下最长仅能达到两米，且单通道带宽翻倍难度日益增加。光子芯片（CPO）可有效解决这些问题，它不仅能满足带宽密度要求，更能提供多重扩展路径以支撑未来带宽需求，同时将系统规模扩展至更广阔的维度。

一旦将CPO部署用于纵向扩展网络，纵向扩展域将不再受互连距离的限制。原则上，客户能够将纵向扩展域扩展至任意规模。当然，若希望将纵向扩展域维持在单层扇出网络中以实现全互连，则其规模将受限于交换机的基数。

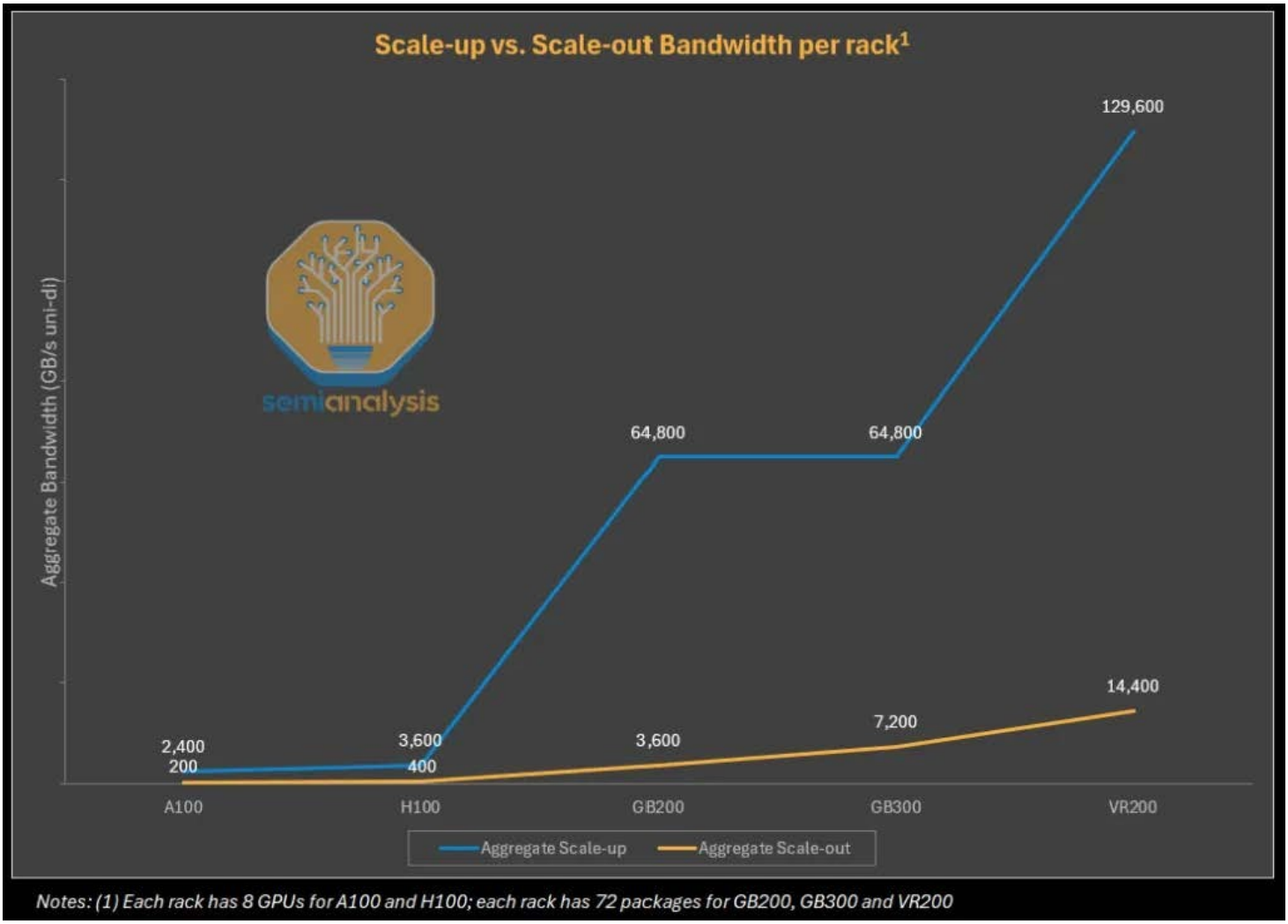
横向扩展与纵向扩展的总可寻址市场

规模扩展型结构的网络要求远比后端横向扩展网络更为严苛。GPU间或与交换机间的连接需要更高的带宽和更低的延迟，以实现GPU的互联互通，从而使其能够协同共享内存等资源。

例如，Nvidia Blackwell平台上的第五代NVLink可为每块GPU提供900GB/s（7,200Gbit/s）的单向带宽。这相当于后端扩展网络中每GPU 100GB/s（800Gbit/s）带宽的9倍（以GB300 NVL72搭配CX-8网卡为例）。这也催生了对更高岸线带宽密度的需求

带宽密度，这正是推动GPU SerDes线路速度发展的动力。

同样重要的是要认识到，随着纵向扩展领域的规模扩大以及纵向扩展互连速度的提升，纵向扩展互连（最终包括纵向扩展CPO）的总可寻址市场（TAM）已远超横向扩展网络。CPO的TAM很可能由纵向扩展而非横向扩展网络应用主导。



来源：SemiAnalysis

纵向扩展场景中的铜缆与光纤对比：全球规模、密度与传输距离

目前，扩展型网络完全采用铜缆运行是有充分理由的。在现行可插拔架构中，若要通过光收发器匹配NVLink带宽，不仅成本和功耗将高得令人望而却步，还会引入额外延迟。此外，计算托盘的面板空间可能根本不足以容纳所有这些收发器。铜缆在实现低延迟、高吞吐量的连接方面表现卓越

连接场景。然而如前所述，铜缆传输距离的局限性制约了"世界规模"——即单个扩展域内可连接的GPU数量上限。

扩大纵向扩展世界规模是计算扩展的关键路径。在当前基于推理的模型扩展和测试时间计算体系下，单一纵向扩展域内增加更多计算能力、内存容量及内存带宽已变得至关重要。

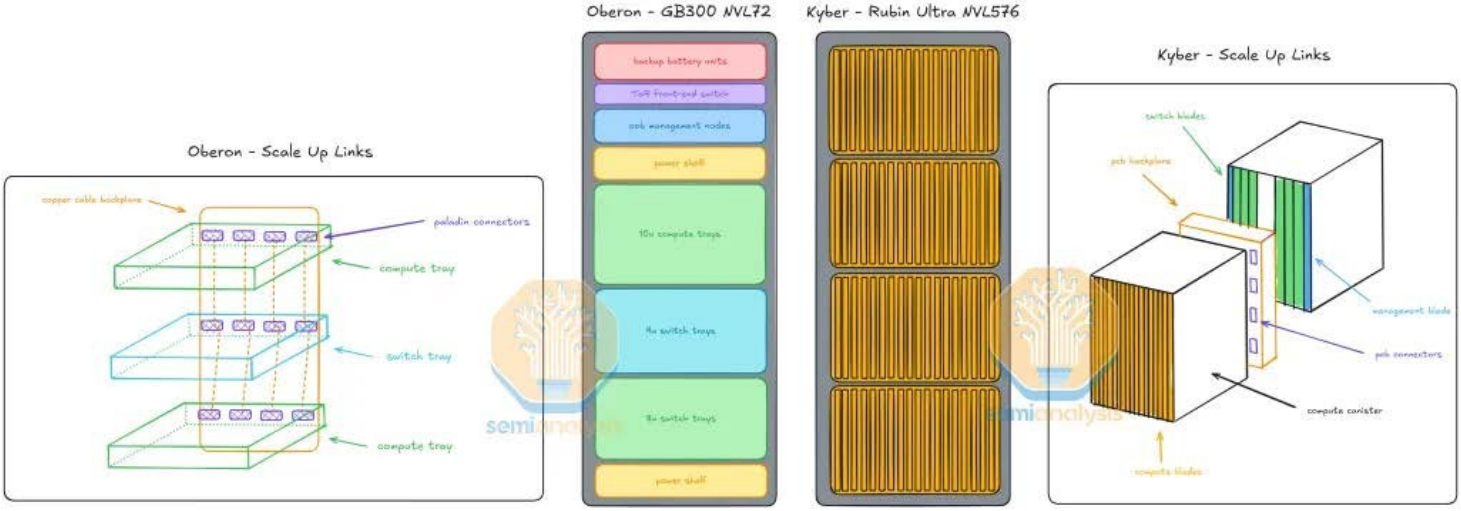
英伟达的GB200系统实现了性能飞跃，其世界规模从仅8个互联GPU扩展至72个全互联GPU，通过更复杂的集体通信技术实现了海量吞吐量提升——这些技术在规模扩展网络中难以实现。

在铜缆架构下，该方案仅能部署于单机架空间内，对供电系统、热管理及制造工艺提出严苛要求。该系统的复杂性至今仍令下游供应链在产能爬坡方面举步维艰。

英伟达将继续坚持使用铜制材料。他们还需进一步提升其纵向扩展型系统规模，以保持对AMD等竞争对手及正在追赶其纵向扩展网络的超大规模企业的领先优势。因此，英伟达被迫采取极端措施，在单个机架内拓展纵向扩展领域。

在GTC 2025大会上展示的Nvidia为Rubin Ultra设计的极端Kyber机架架构，可扩展至144个GPU封装单元（576个GPU芯片）。该机架的密度是原本就极为密集的GB200/300 NVL72机架的4倍。鉴于GB200在制造和部署方面已相当复杂，Kyber架构将这一复杂性提升到了全新高度。

光学技术实现了相反的方案：通过跨多个机架扩展来增大世界规模，而非在有限空间内堆砌更多加速器——后者面临供电和热密度方面的挑战。当前可通过可插拔收发器实现此方案，但光收发器的成本及其高功耗特性仍使其难以普及。



来源：SemiAnalysis

铜缆与光纤在扩展中的对比：带宽扩展

在铜缆上扩展带宽也日益困难。通过Rubin芯片，英伟达采用创新的双向SerDes技术实现了带宽翻倍——该技术使发送和接收操作共享同一通道，从而实现每通道224Gbit/s发送+224Gbit/s接收的全双工通信。要在铜缆上实现每通道"真正"的448Gbit/s仍是一项充满挑战的壮举，其上市时间尚不确定。 相比之下，CPO技术提供了多重带宽扩展路径：波特率提升、DWDM技术、增加光纤对数以及调制方式革新——本文后续将对此进行详细探讨。

CPO何时能登上主流舞台?

那么，既然CPO是解决方案，为什么英伟达最初只将其应用于横向扩展交换机，而未用于Rubin Ultra？这要归因于供应链不成熟、制造挑战以及客户在部署方面的犹豫。Quantum和Spectrum CPO交换机的推出，旨在加速供应链建设，并获取更多关于数据中心可靠性和可维护性的实际应用数据。

在此期间，Meta在欧洲光通信大会（ECOC）期间发布的CPO可靠性数据提供了一些有价值的信息。Meta与博通合作开展了这项研究，博通也~~发布了若~~干实用幻灯片。该研究中，Meta在15台Bailly 51.2T CPO交换机上进行了规模可观的测试，累计运行达1,049k个400G端口设备小时，并公布了最大非零KP4前向纠错（FEC）分桶值：

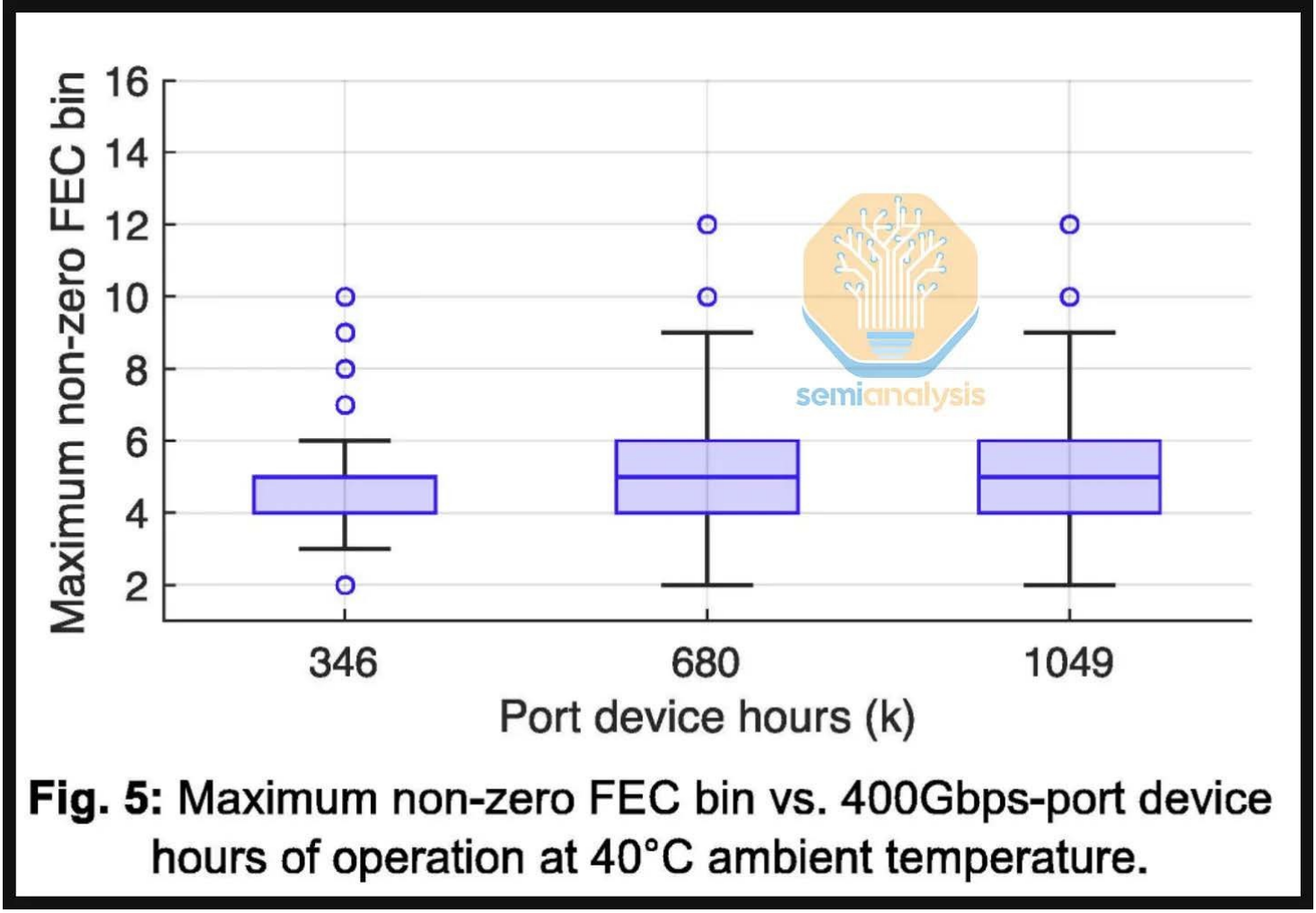
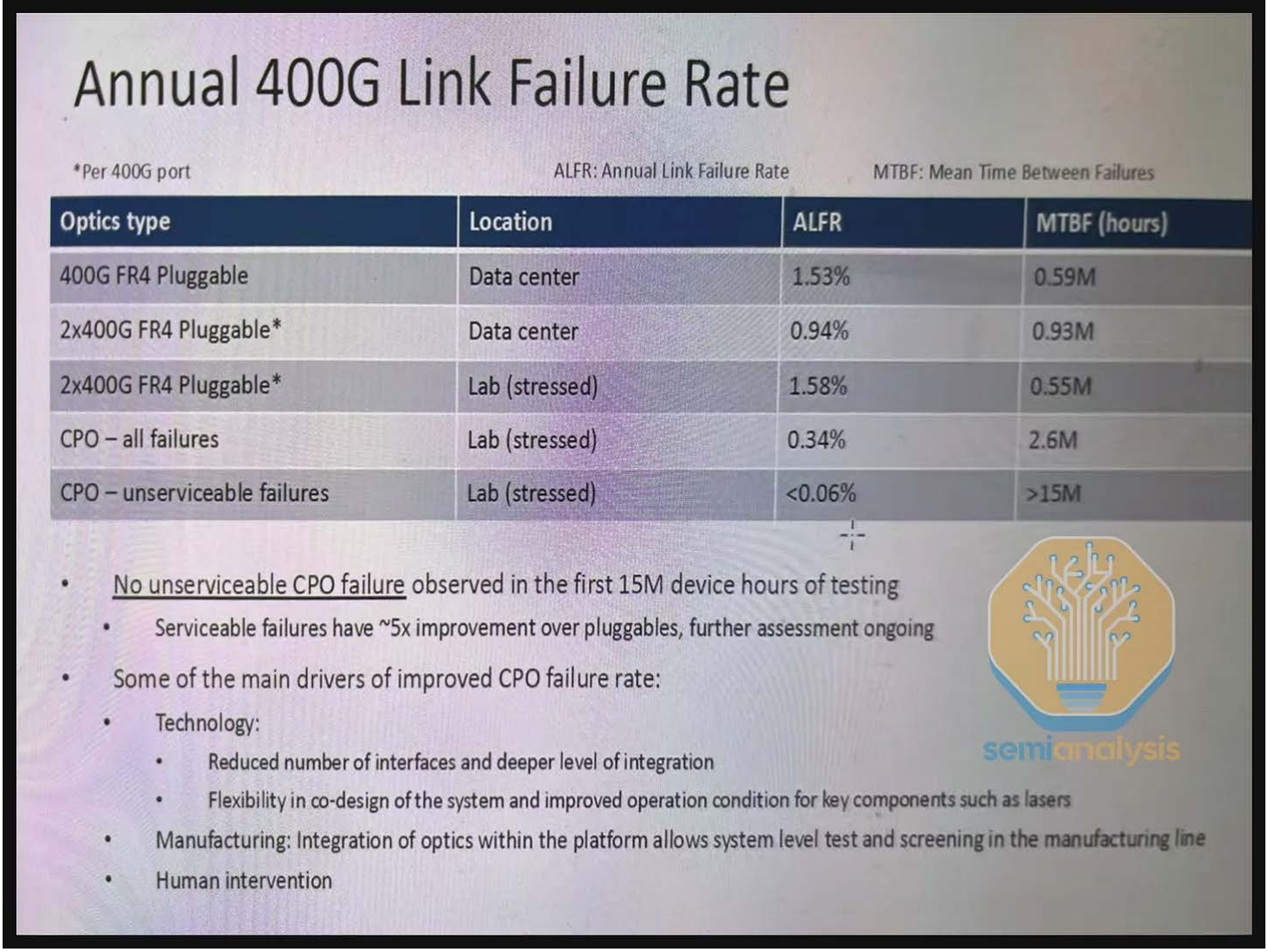


Fig. 5: Maximum non-zero FEC bin vs. 400Gbps-port device hours of operation at 40°C ambient temperature.

来源：Meta

论文同时说明测试期间链路未出现故障或不可纠错码字（UCW），仅在总计104.9万400G端口设备小时的测试周期中观察到一次FEC分箱值>10的情况。

但Meta并未止步于此。在ECOC会议上发表该论文时，他们展示了扩展至1500万400G端口设备小时的测试结果。 新数据表明：前400万设备小时内未出现UCW；400G 2xFR4收发器平均故障间隔时间(MTBF)为0.5-100万设备小时（全球2xFR4平均值为55万），而CPO的MTBF达260万设备小时。



来源：Meta

虽然1500万端口设备小时看似是个大数目，但这是以400G端口小时为单位计算的。因此——一台51.2T交换机运行一小时相当于128个400G端口小时。15台51.2T交换机累计1500万400G端口小时，相当于7812个墙钟小时，约合325天。 事实上，这个1500万小时的数据常被简化为"小时"或"设备小时"，省略了"端口"的限定。 尽管在400万端口设备小时内实现零故障和零UCW的统计数据极具参考价值——但在行业转向CPO横向扩展交换技术并投入数十亿美元前，仅凭15台CPO交换机在实验室环境下测试11个月的数据远远不够。

在动态现场环境中运行数千台横向扩展交换机是完全不同的挑战，这些交换机在生产环境中的表现仍有待观察。 生产环境的温度波动可能远高于实验室环境，导致组件性能或耐用性出现意外变化。Meta在Llama 3论文中指出，数据中心1-2%的温度波动就足以引发功耗异常——此类波动是否会以难以预料的方式影响整个网络架构？

One interesting observation is the impact of environmental factors on training performance at scale. For Llama 3 405B , we noted a diurnal 1-2% throughput variation based on time-of-day. This fluctuation is the result of higher mid-day temperatures impacting GPU dynamic voltage and frequency scaling.

During training, tens of thousands of GPUs may increase or decrease power consumption at the same time, for example, due to all GPUs waiting for checkpointing or collective communications to finish, or the startup or shutdown of the entire training job. When this happens, it can result in instant fluctuations of power consumption across the data center on the order of tens of megawatts, stretching the limits of the power grid. This is an ongoing challenge for us as we scale training for future, even larger Llama models.

Meta

即便看似寻常的问题——例如数据中心内的灰尘——也令技术支持人员头疼不已，他们可能耗费大量时间清理光纤端口。当然，CPO交换机采用LC或MPO类型的前插式连接器，但CPO交换机机箱内部的灰尘问题又该如何解决？ 0.06%的不可修复故障率看似诱人，但此类故障的波及范围可达64个800G端口。本文虽聚焦于基于FR光学的CPO交换机，但下一代CPO交换机将采用DR光学技术。这些仅是已知的未知因素，实际部署中可能还会出现更多未知的未知问题。

事实上，这些成果通过提供切实可靠的数据，在说服业内人士方面产生了深远影响。我们的目的并非制造恐惧、不确定性或怀疑（FUD），而是呼吁开展更大规模的现场测试，使行业能够快速识别并解决突发问题，从而为CPO技术的广泛应用铺平道路——尤其是在扩展型网络领域。

归根结底，英伟达此次推出的横向扩展型CPO产品，实为大规模部署前的预演与试探。我们认为纵向扩展架构的部署规模与影响力将远超横向扩展，因为纵向扩展在总拥有成本（TCO）及性能/TCO效益方面具有更显著的优势。

此外，在横向扩展型CPO领域，Rubin Ultra计划于2027年推出（我们认为实际可能延迟至2027年底），而供应链尚无法准备好向GPU需求市场交付数千万台CPO终端设备。即便如此，这个时间表对英伟达而言仍过于雄心勃勃。正因如此，费曼架构世代似乎将成为CPO技术注入英伟达生态系统的核心节点。

现在让我们深入探讨CPO的核心理念、技术考量、面临的挑战以及当前生态系统的现状。

第二部分：CPO 导论与实施

CPO究竟是什么？为何引发业界如此热议？

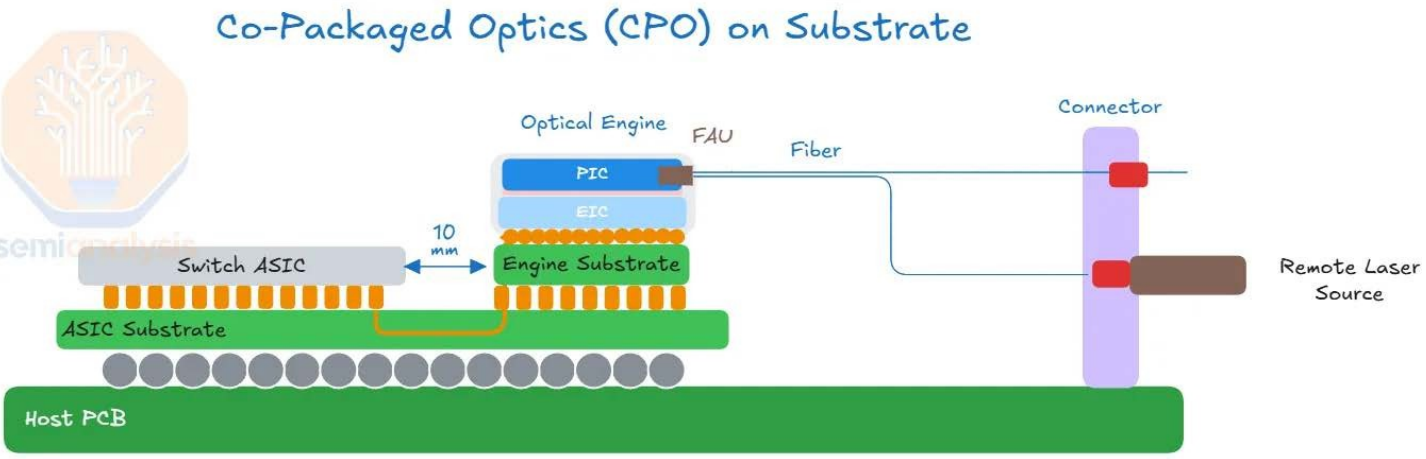
CPO技术将光引擎直接集成于高性能计算或网络专用集成电路（ASIC）的同一封装或模块内。这些光引擎负责将电信号转换为光信号，从而实现光链路上的高速数据传输。

当数据传输距离超过数米时，必须采用光链路进行通信，因为铜缆传输的高速电信号无法跨越数米距离。

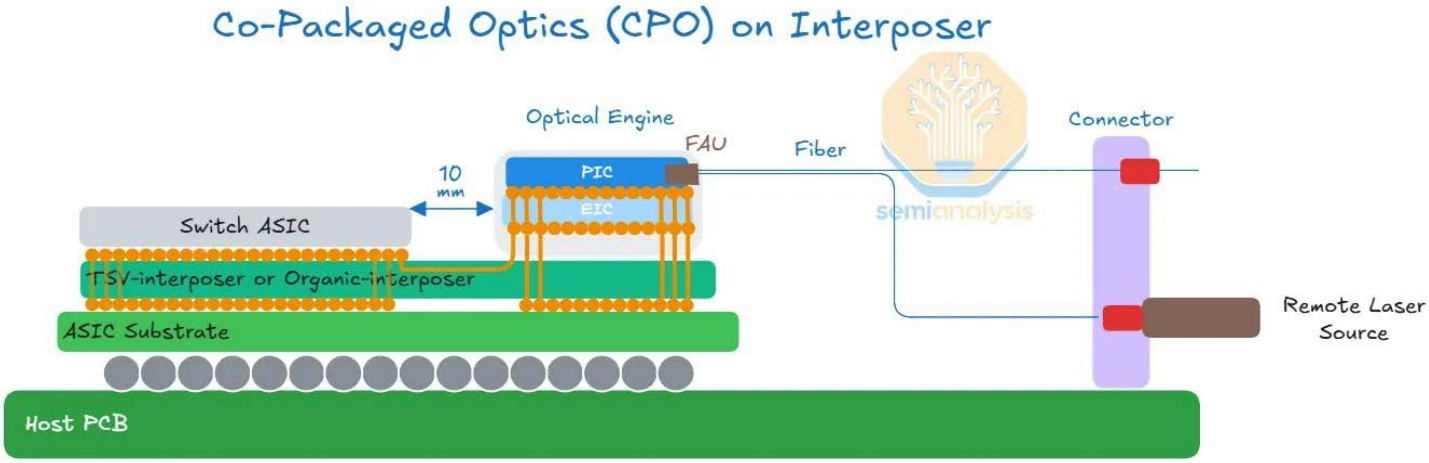
如今，大多数电光转换都是通过可插拔光收发器实现的。在这种情况下，电信号会从开关或处理芯片出发，穿越数十厘米甚至更长的PCB电路板，最终抵达机箱前面板或后面板上的物理收发器插槽。可插拔光收发器就安装在该插槽内。 收发器接收电信号后，由光数字信号处理器（DSP）芯片重新整形，再传输至光引擎组件进行电光转换。光信号经光纤传输至链路另一端，由另一收发器执行逆向转换流程，将光信号转为电信号，最终返回目标硅芯片。

在此过程中，电信号在抵达光链路前需穿越较长距离（至少对铜缆而言），并经历多次转换点。这导致电信号逐渐劣化，需要消耗大量能量并采用复杂电路（串行器/解串器）进行驱动与恢复。 为改善这一状况，我们需要缩短电信号的传输距离。由此催生了"**共封装光器件**"的概念——即将原本置于可插拔收发器中的光引擎与主芯片进行共封装。 此方案将电路走线长度从数十厘米缩短至数十毫米，因光引擎与XPU或交换机ASIC的物理距离大幅缩短。通过最小化电气互联距离并缓解信号完整性挑战，该方案显著降低功耗、提升带宽密度并减少延迟。

下图展示了CPO的实现方案，其中光引擎与计算芯片或交换芯片集成在同一封装内。初期光引擎将部署在基板上，未来则会转移至中介层。

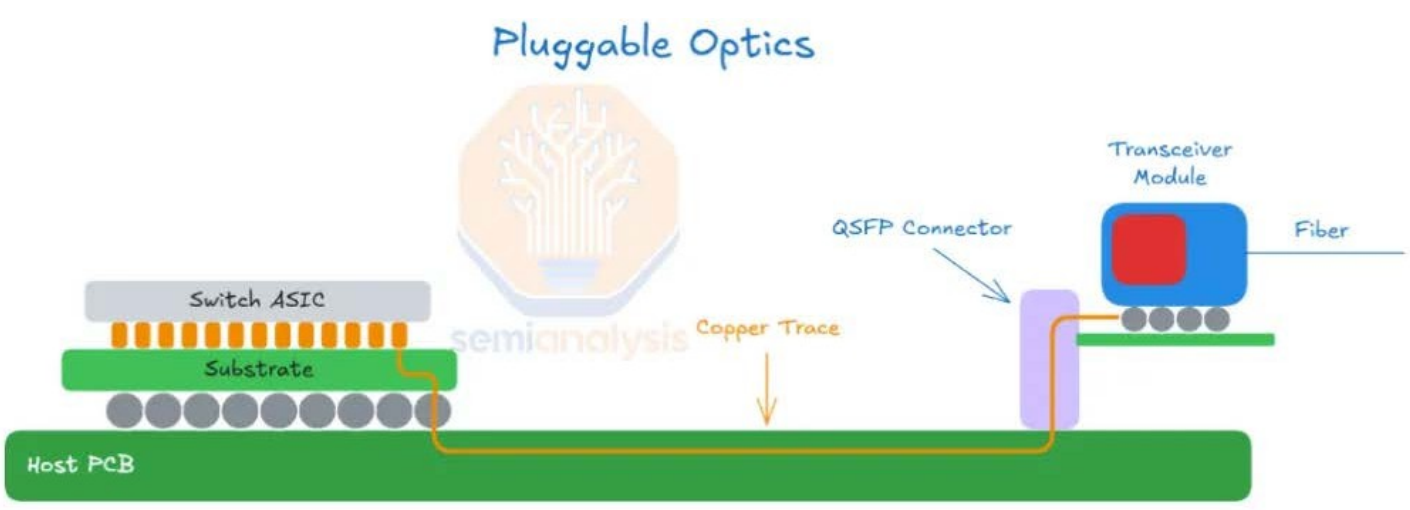


来源：SemiAnalysis



来源：SemiAnalysis

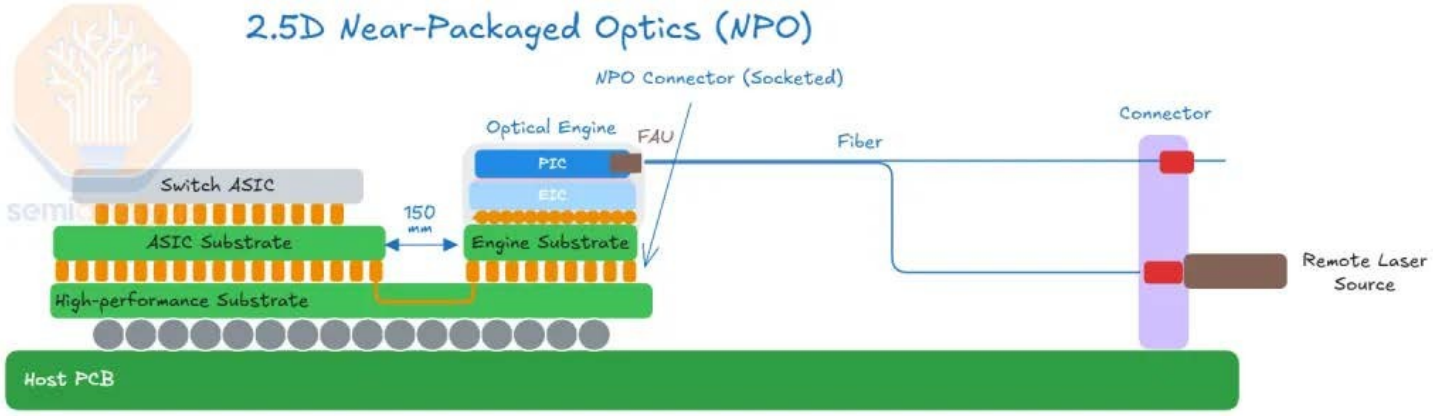
如今，如下方示意图所示的前插拔式光模块解决方案已广泛应用。 该图的核心要点在于说明：电信号需穿越长达15-30厘米的铜迹线或跨接线缆，才能抵达收发器中的光引擎。如前所述，这同样要求采用长距离（LR）串解码器（SerDes）来驱动可插拔模块。



来源：SemiAnalysis

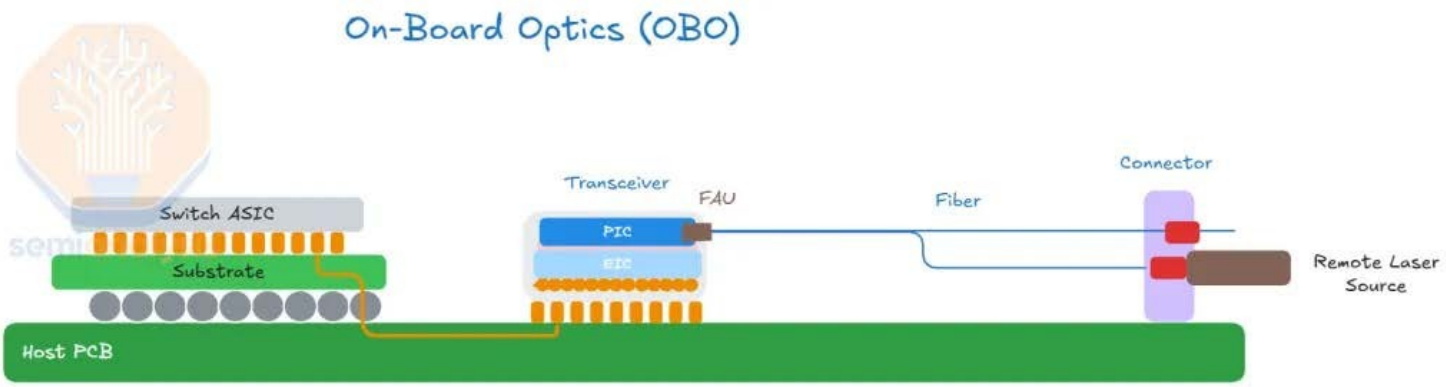
此外，还存在介于CPO与传统前插式光模块之间的中间实现方案，例如近封装光模块（NPO）和板载光模块（OBO）。

近年来，NPO作为迈向CPO的中间过渡方案应运而生。NPO存在多种定义：其特征在于光学引擎（OE）不直接安装在ASIC基板上，而是与另一基板共同封装。该光学引擎仍保持插拔特性，可从基板上拆卸。电气信号仍将通过XPU封装上的串行器/解串器（SerDes）经铜通道传输至光学引擎。



来源：SemiAnalysis

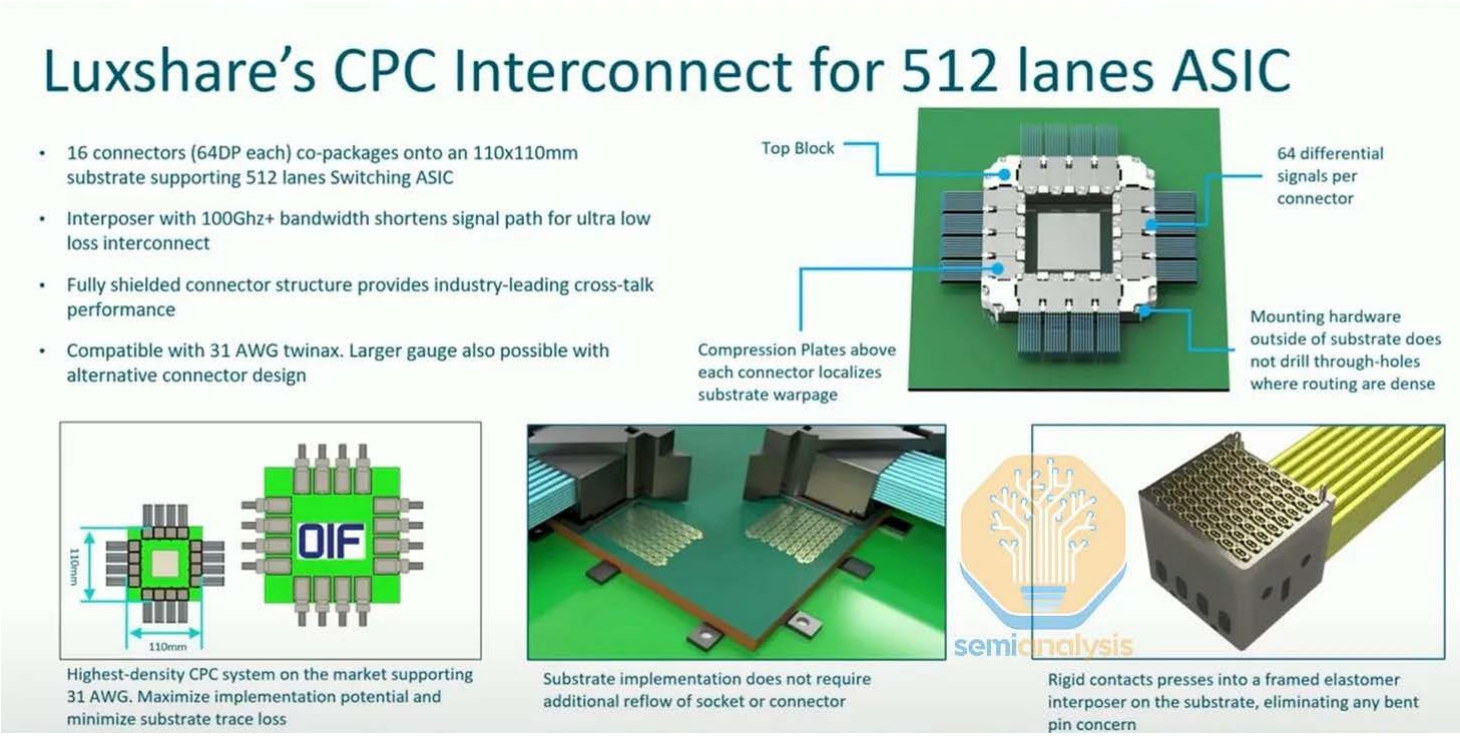
还有一种方案是板载光学器件（OBO），它将光学引擎集成到机箱内的系统PCB上，使其更接近主机ASIC。然而，OBO不仅继承了CPO的诸多挑战，在带宽密度和功耗节约方面带来的优势也更少。 我们认为OBO是"两全其美的反面典范"，它既保留了CPO的复杂性，又继承了前插式光模块的部分局限性。



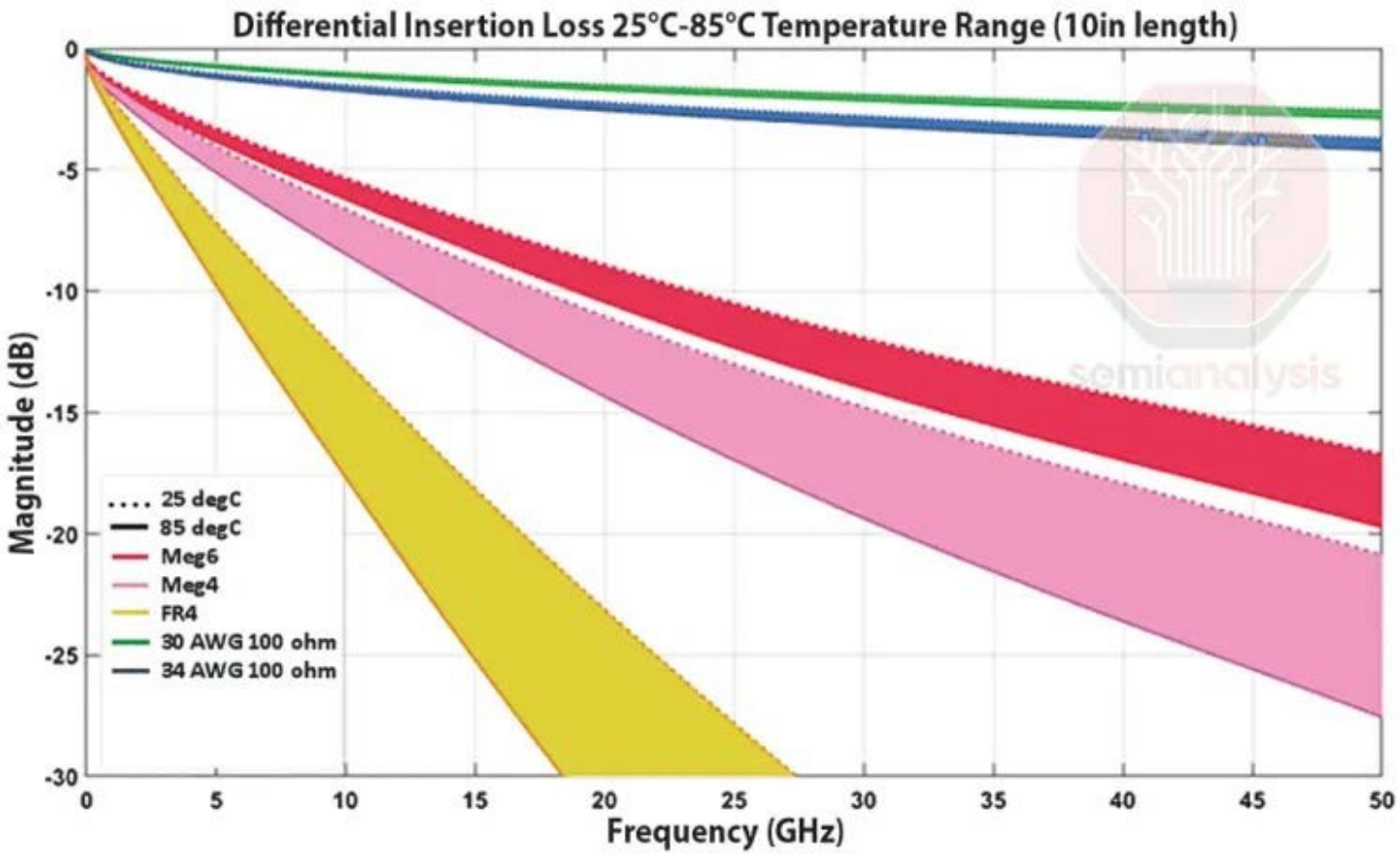
来源：SemiAnalysis

共封装铜光纤

CPO的另一种替代方案是"共封装铜线"(CPC)。CPC采用直接从基板连接器引出的铜线缆。CPC使用的线缆与飞线相同，目的也相同：绕过PCB走线。CPC将飞线技术进一步延伸，使插座直接始于封装基板本身。 所用线缆为双轴线缆（Twinax），其绝缘性能优异可有效抑制串扰，与传统电路走线相比能显著降低插入损耗。尽管该方案仍采用铜质材料，但在信号完整性方面具有关键优势。CPC技术有望为部署448G串行器-解串器（SerDes）提供可行路径，从而实现封装外互连的进一步扩展。



来源：立讯精密



来源：立讯精密

CPC技术的挑战在于封装基板的复杂性增加。基板必须为数千条电缆分配电源和信号。尽管存在这一挑战，CPC仍比CPO简单得多——后者仍需克服供应链多个环节的制造障碍。 我们认为CPC技术在某些短距离应用中具有独特优势，例如机架级扩展连接（下文将详述）。通过规避损耗性CCL走线，CPC有望成为实现448G传输速率的关键技术。鉴于该带宽信号在PCB传输中会产生不可接受的衰减，CPC技术正被广泛研究以支持448G应用。

CPO技术市场化进程中的历史障碍：为何直到现在才实现？

尽管技术上具有优势，但由于存在若干推高成本的挑战，CPO在实际应用中的普及仍十分有限。这些挑战包括：封装工艺复杂（成本甚至高于器件本身）、制造难度大、可靠性与良率问题，以及光电元件高度集成导致的热管理难题。另一个阻碍是行业标准尚未统一。此外，客户还担忧其可维护性问题。

了从传统可插拔光模块向CPO解决方案的过渡。

另一项关键客户担忧在于，采用CPO技术可能意味着他们将丧失成本控制能力。相较于少数几家交换机供应商，向数量庞大的收发器厂商施压降价显然更为容易。

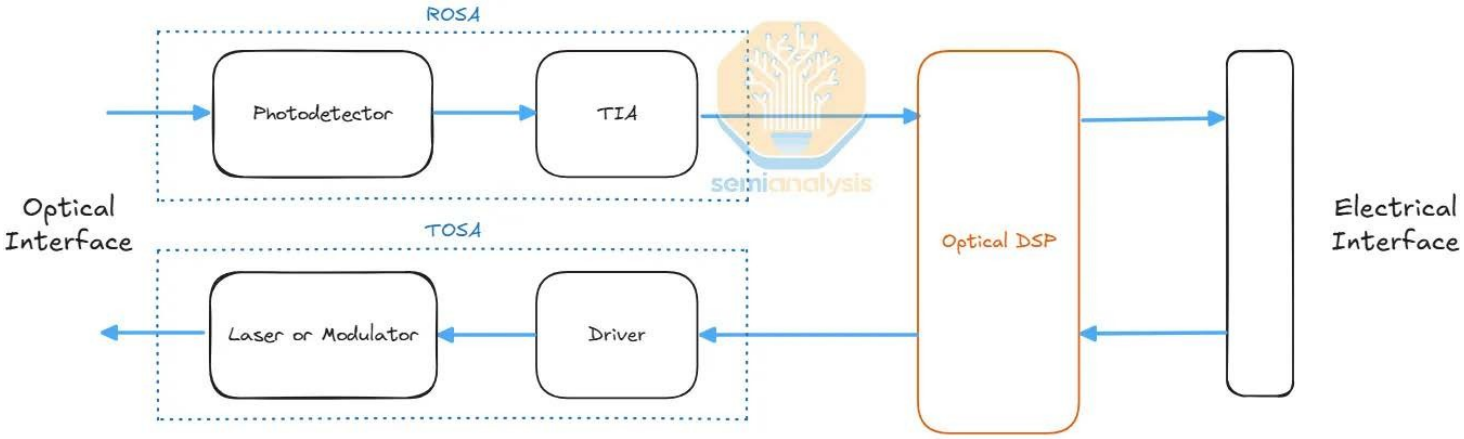
与此同时，可插拔光模块——即CPO将取代的现有技术——仍在持续改进，其性能足以满足几乎所有应用需求，且能显著降低终端用户的顾虑。

在第二部分的剩余内容中，我们将深入探讨采用CPO的驱动力。首先阐述SerDes扩展已趋于饱和，使得宽I/O配合CPO等其他接口类型成为必要选择，随后将探讨制造考量与市场推广策略。我们将重点解析关键CPO组件，包括光引擎、光纤耦合器、外部激光源及调制器。 最后，我们将阐述基于CPO技术实现带宽扩展的发展路线图。

超越基于DSP的收发器：从LPO到CPO

DSP收发器同时处理光信号的发送与接收，其内部包含负责电光转换的"光引擎"(OE)。该光引擎由驱动器(DRV)和调制器(MOD)组成，用于发送光信号；同时配备跨阻放大器(TIA)和光电探测器(PD)，用于接收光信号。

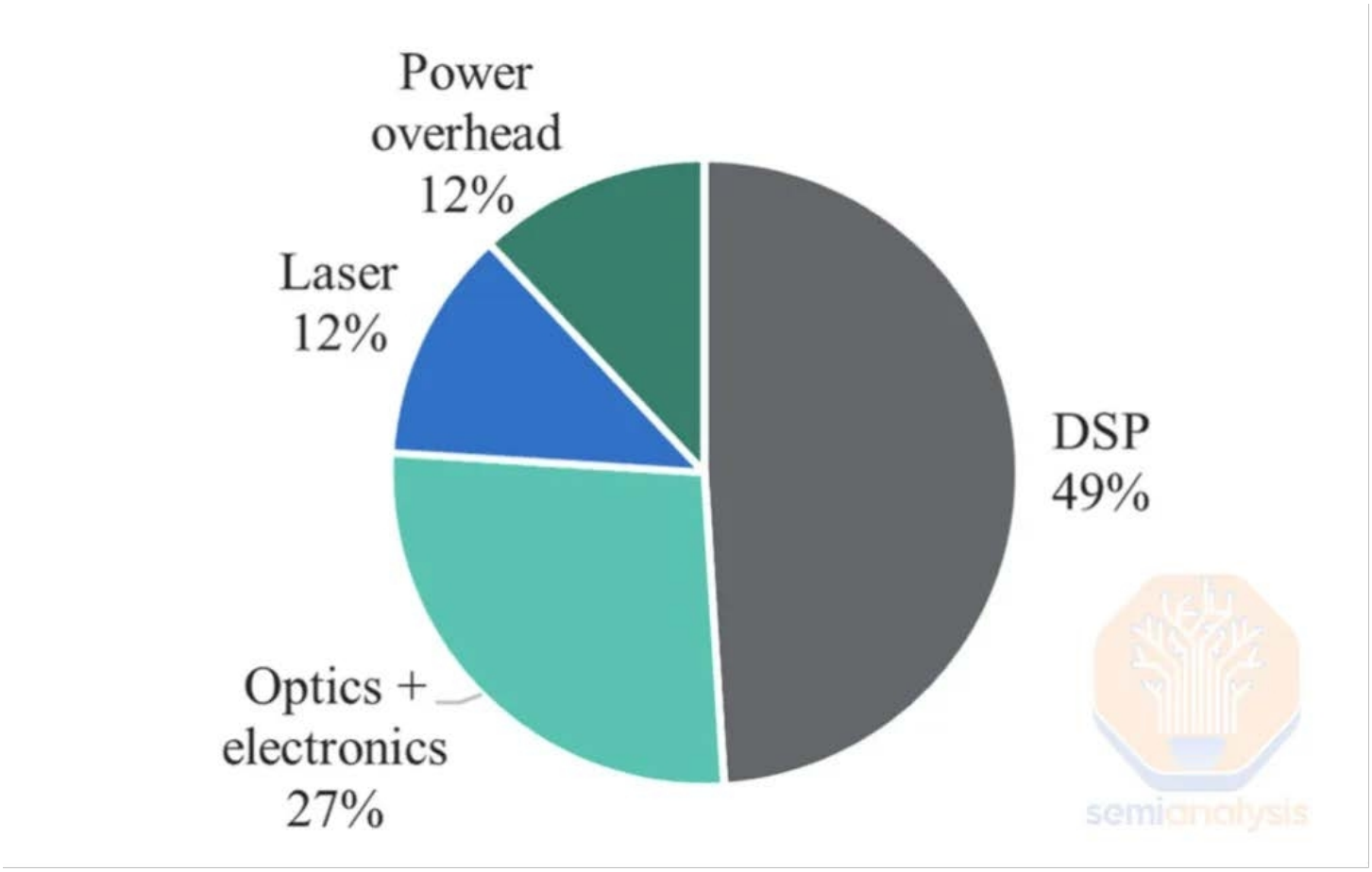
另一个重要组件是光DSP芯片，它有时会将驱动器和/或TIA集成到单个封装中。从主机开关或处理芯片传输的高频电信号需要通过损耗性铜迹线传输较长距离，才能到达服务器机箱前端的收发器。DSP负责对该信号进行重新定时和重新调理。 它通过执行误差校正和时钟/数据恢复，补偿信号从开关或ASIC芯片经基板等传输介质传播过程中产生的电气劣化与衰减。在调制方面，对于PAM4调制（四电平脉冲幅度调制），DSP将二进制信号映射为四个独立幅度电平，从而增加每信号位数，实现更高比特率和更宽带宽。



Source: SemiAnalysis

DSP芯片是收发器中功耗最高且最昂贵的组件之一，甚至可能是最耗电的部件。以800G SR8收发器为例——DSP几乎占用了

~50%的模块总功耗，这正是业界极力寻求替代DSP方案的核心原因。



来源：Radha Nagarajan博士等：《数据中心应用低功耗数字信号处理技术的最新进展》

构建一个采用双层InfiniBand网络的18k GB300集群，需配备18,432个800G DR4收发器和27,648个1.6T DR8收发器。DSP带来的额外成本与功耗需求将显著增加整体拥有成本。 按800G DSP功耗6-7W、1.6T DSP功耗12-14W估算

仅后端网络部分的DSP功耗就将达到480kW，相当于每台服务器机架约1.8kW。若采购自高端品牌供应商，收发器成本可能占集群总拥有成本近10%。 因此——考虑到DSP占典型收发器功耗的50%及物料清单成本的20-30%——许多人视DSP为成本与能效领域的头号公敌。

GB300 NVL72 Cluster Cost and Power Budget Breakdown, per Rack-Scale 72-GPU Server						
	2-Layer Network (up to 10,368 GPUs)		3-Layer Network (up to 746,496 GPUs)		4-Layer Network (up to ~53.7M GPUs)	
Item	Cost	Power (W)	Cost	Power (W)	Cost	Power (W)
Server	$4,357,943	142,000	$4,357,943	142,000	$4,357,943	142,000
Optical Transceivers	$363,125	4,600	$499,925	6,184	$568,325	6,976
Switches	$208,681	5,214	$312,681	8,014	$416,681	11,606
Fiber, Cables, Software, Others	$17,723	0	$24,338	0	$30,953	0
Networking	$589,529	9,814	$836,944	14,198	$1,015,959	18,582
All Others	$269,152	281	$269,152	281	$269,152	281
Total Power	**$5,216,624**	**152,095**	**$5,464,039**	**156,479**	**$5,643,054**	**160,863**

Power for a Neocloud Giant InifiniBand cluster using X800-Q3400 Back-end switches with 144 ports of 800G
Source: SemiAnalysis Networking Model

来源：SemiAnalysis AI网络模型

反DSP运动

DSP的高成本和高功耗特性促使业界寻求能够绕过DSP的中间环节的技术。 针对DSP的首轮攻势是线性可插拔光模块（LPO）——该方案试图彻底移除DSP，让交换机的串解码器直接驱动收发器中的发射/接收光元件。然而正如DSP先知Loi Nguyen在2023年接受我们采访时准确预言的那样，LPO至今仍未实现大规模应用。

Marvell的DSP困境？由博通、英伟达、Arista Networks、微软、Meta、Macom等引领的网络领域板块级变革

DYLAN PATEL · 2023年3月8日

阅读全文 →

CPO将LPO概念推进到新阶段，将光引擎与计算或交换芯片封装在同一封装内。其关键优势在于，由于主机与光引擎间距极短，收发器中的DSP已不再必要。 CPO的突破性还在于：通过淘汰功耗高、占面积大的长距离SerDes，转而采用短距离SerDes，甚至在宽I/O接口场景下采用时钟前传式宽D2D SerDes，从而大幅提升芯片封装密度。

官方引述的说法是，CPO技术过去二十年间始终近在咫尺——但为何迟迟未能普及？为何业界仍倾向于采用可插拔DSP收发器？

可插拔收发器的核心优势在于其卓越的互操作性。凭借OSFP和QSFP-DD等标准封装形式，并遵循OIF规范，客户通常可独立于交换机和服务器供应商选择收发器厂商，从而获得采购灵活性和更强的议价能力。

另一个巨大优势在于现场可维护性。收发器的安装与更换极为简便，只需远程操作人员拔出交换机或服务器机箱中的插头即可。相比之下，采用CPO方案时，光引擎的任何故障都可能导致整台交换机瘫痪。 即便可维修的故障也可能面临复杂的排查与修复过程。激光器作为最常见的故障点，当前多数CPO方案已采用可插拔外部激光源以提升可维护性与可替换性，但其他不可插拔CPO组件的故障风险仍令人忧心。

为何选择CPO？I/O挑战、带宽密度与瓶颈问题

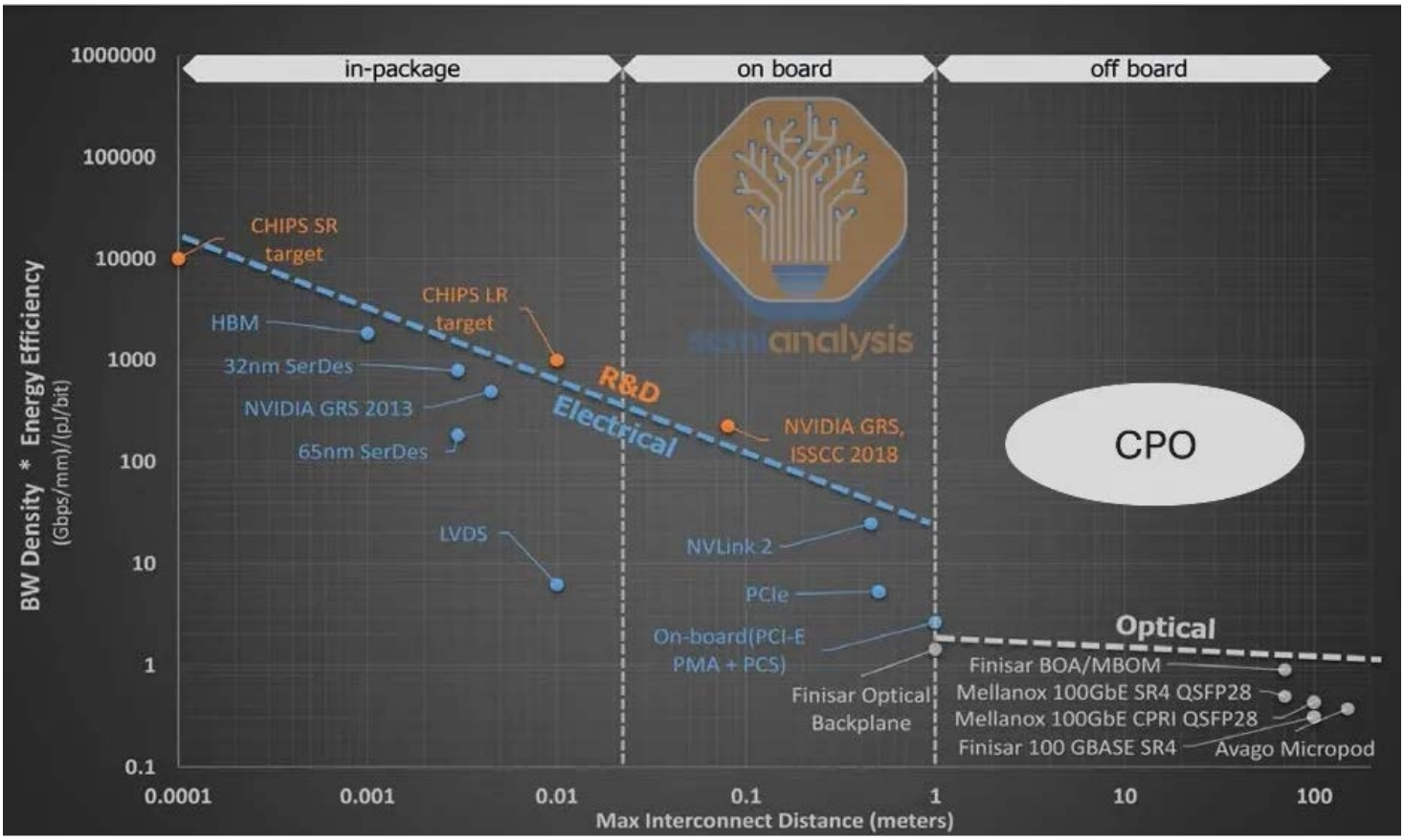
除淘汰耗电且昂贵的DSP器件、最大限度减少或消除LR SerDes使用外，采用CPO的另一重大优势在于其在能耗水平下能实现更高的互连带宽密度。

带宽密度衡量单位面积或通道传输的数据量，反映有限空间用于高速数据传输的利用效率。能效则量化传输单位数据所需的能量消耗。

因此，在评估特定互连结构的客观质量时，互连带宽密度与能耗之比成为至关重要的性能指标（FoM）。当然，最优互连方案还需满足距离和成本参数的约束条件。

观察下图可发现明显趋势：随着距离增加，电气链路的该性能指标呈指数级恶化。此外，从纯电气接口转向需要光电转换的接口时，会引入

效率大幅下降——可能达到数量级的降幅。这种下降源于信号从芯片传输至前板收发器所需的能量消耗，而驱动光DSP则需要更多能量。基于CPO的通信性能指标曲线明显优于可插拔模块。如下方图表所示，在相同传输距离范围内，CPO技术在单位面积能耗下提供更高的带宽密度，客观上成为更优的互连方案。



来源：G·基勒，DARPA 2019，SemiAnalysis

该图表也印证了"能用铜缆就用铜缆，必须用光纤才用光纤"的行业箴言。在可行的情况下，短距离铜缆通信更具优势。英伟达秉承这一理念，其机架级GPU架构专为突破机架内密度极限而设计，旨在最大化通过铜缆联网的GPU数量。 这正是GB200 NVL72采用纵向扩展网络架构的逻辑基础，而英伟达在Kyber机架中将这一理念推向极致。然而——随着光纤光栅（CPO）技术的成熟，其性能曲线将逐步覆盖纵向扩展场景，从每TCO性能角度考量，其应用价值终将显现，这只是时间问题。

GB200硬件架构 - 组件供应链与物料清单

DYLAN PATEL, WEGA CHU, AND 4 OTHERS · 2024年7月17日

阅读全文→

NVIDIA GTC 2025 - 推理引擎打造，Vera Rubin，Kyber，CPO，Dynamo推理，Jensen数学，Feynman

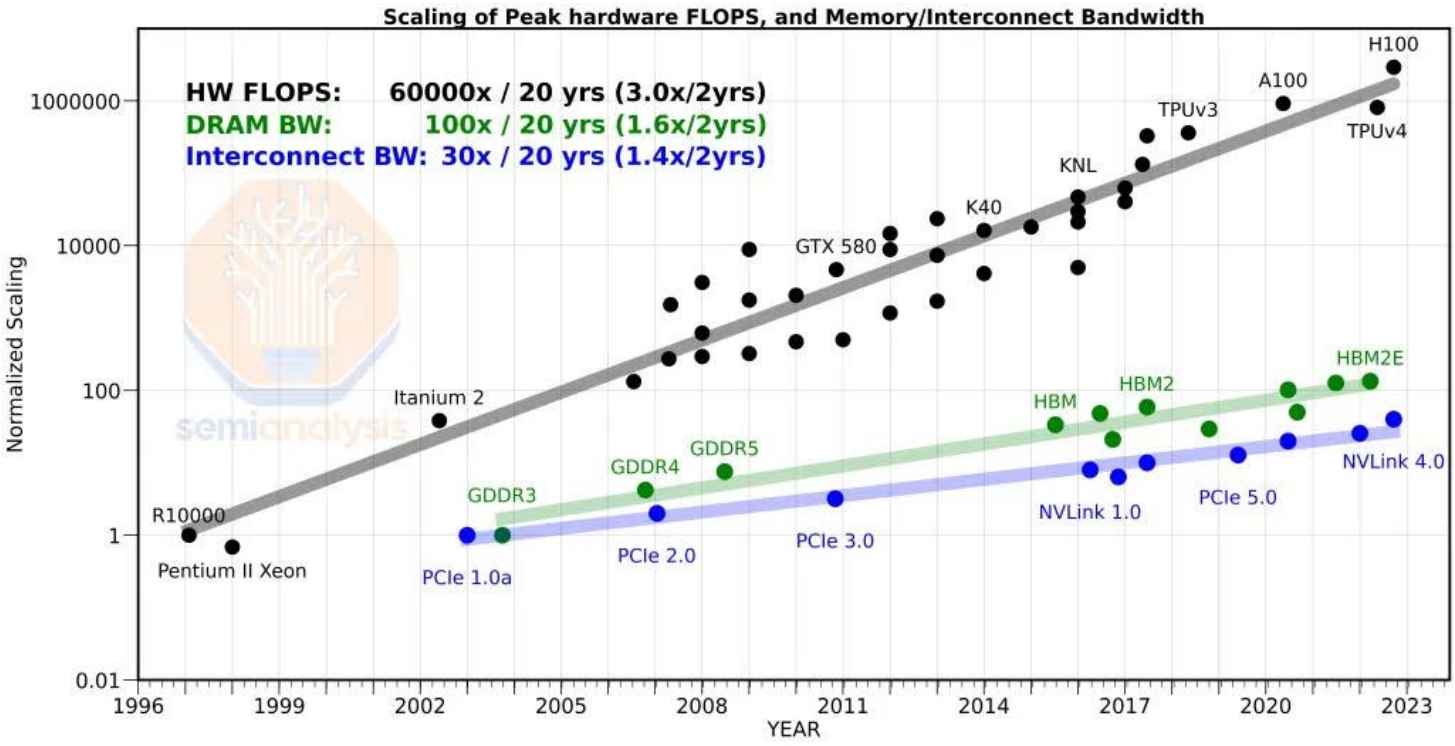
迪伦·帕特尔、谢迈伦及5位作者 · 2025年3月19日

阅读全文 →

输入/输出（I/O）的减速带与路障

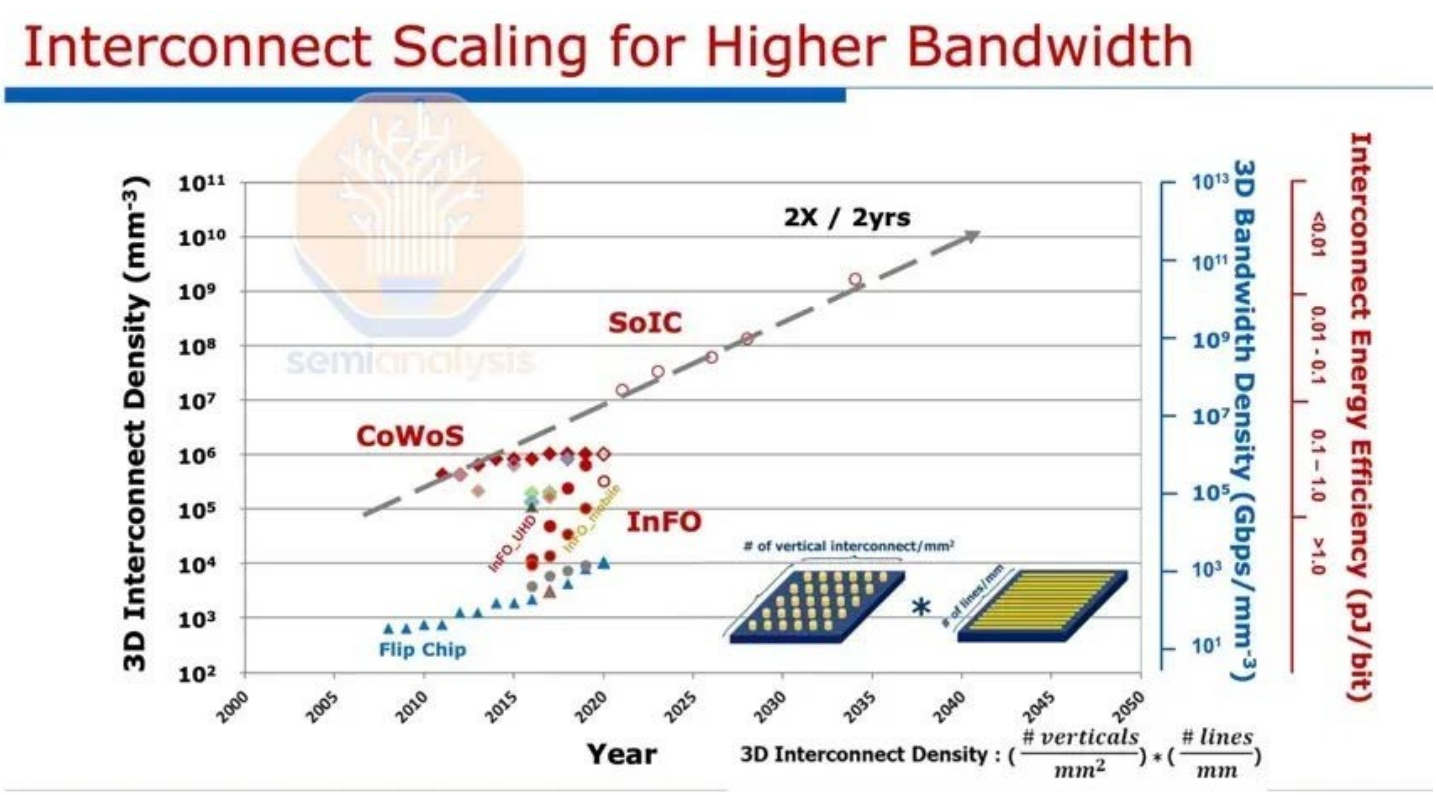
尽管晶体管密度和计算能力（以FLOPs为代表）实现了良好扩展，I/O的扩展速度却明显滞后，导致整体系统性能出现瓶颈：由于数据需通过有机封装基板上有限数量的I/O接口传输，芯片外I/O的可用岸线资源极为有限。

此外，提升单个I/O接口的信号传输速度正变得日益困难且耗能巨大，进一步制约了数据流动。这正是过去数十年间，相较于其他计算技术趋势，互连带宽提升如此缓慢的关键原因。



来源：Amir Gholami

由于单个倒装芯片BGA封装中凸点数量的限制，面向HPC应用的封装外I/O密度已趋于平稳。这成为突破带宽瓶颈的关键制约因素。

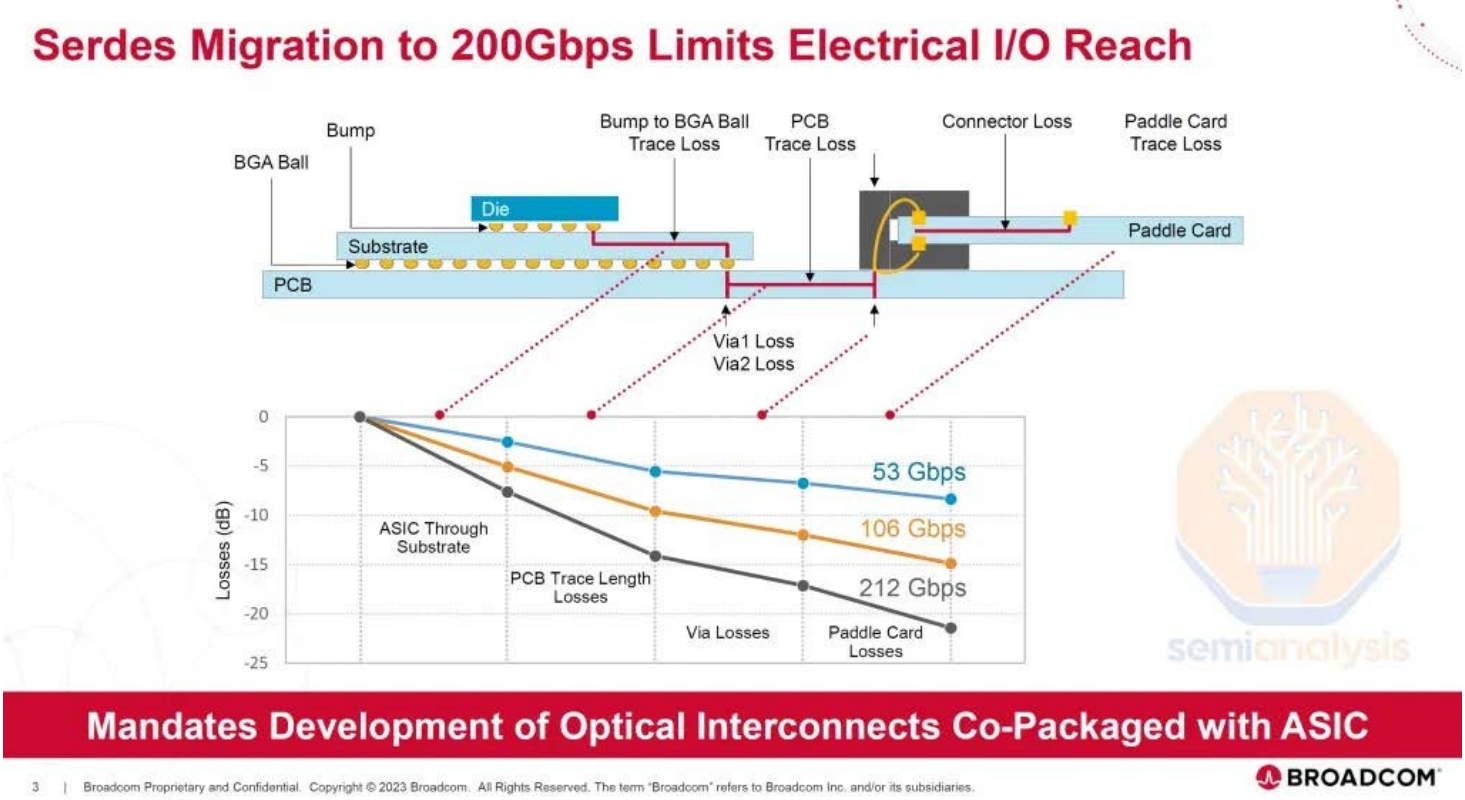


来源：台积电

电气SerDes扩展瓶颈

在I/O接口数量有限的情况下，实现更高逃逸带宽的方法是提升每个I/O信号的传输频率。目前，英伟达和博通在串行器/解串器（SerDes）IP领域处于领先地位。英伟达在Blackwell芯片中搭载了224G SerDes技术，正是这项技术赋予了其NVLink闪电般的传输速度。 同样，自2024年末起，博通已在光DSP中开始提供224G SerDes样品。业内AI浮点运算性能领先的两家企业同时主导高速SerDes IP领域绝非偶然。这印证了AI性能与吞吐量之间的根本关联——最大化数据传输效率与提供原始计算能力同样关键。

然而，在理想传输距离下提供更高线路速率正变得日益困难。如下图所示，随着频率增加，插入损耗随之上升。我们观察到，在更高的串行器/解串器（SerDes）信号传输速率下，尤其是信号路径延长时，损耗会显著增加。



来源：博通

SerDes技术正趋于瓶颈期。更高传输速率仅能在极短距离内维持，且需额外添加信号恢复组件——这反过来又增加了系统复杂度、成本、延迟及功耗。实现224G SerDes技术始终面临巨大挑战。

展望448G串解码器技术，其传输距离能否突破数厘米仍存在较大不确定性。英伟达通过双向串解码技术在Rubin平台实现了每电气通道448G的连接能力。要实现真正的448G单向串解码器，仍需进一步研发。 我们可能需要转向更高阶调制方案，例如PAM6或PAM8，而非自56G SerDes时代以来广泛采用的PAM4调制。若使用每信号编码2位的PAM4方案实现448G传输，将需要244Gbaud的波特率，这因过高的功耗和插入损耗而难以实现。

SerDes扩展瓶颈成为NVLink扩展的障碍

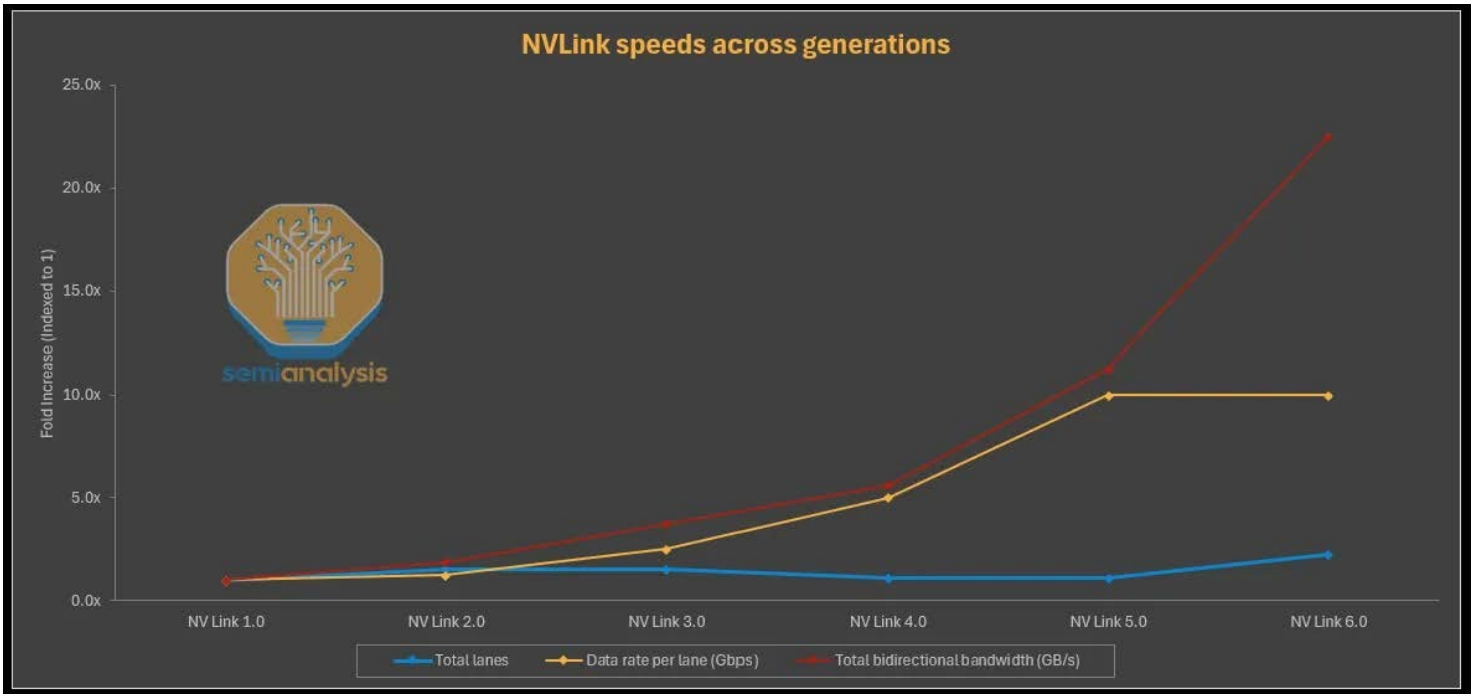
在NVLink协议中，NVLink 5.0的带宽较NVLink 1.0提升了11倍以上。然而，这种增长并非源于通道数量的显著增加——通道数仅从NVLink 1.0的32条略微提升至NVLink 5.0的36条。 带宽提升的核心驱动力在于串行器/解串器（SerDes）通道速度的十倍增长——从20G提升至200G。但在NVLink 6.0中，英伟达预计将维持200G SerDes规格，这意味着必须实现通道数量的倍增——而该方案确实做到了这一点。

通过巧妙运用双向SerDes技术，在保持物理铜线数量不变的前提下，有效实现了通道数量的倍增。然而，进一步提升SerDes速度或突破海岸线资源限制以增加通道数量将日益困难，最终逃逸带宽将陷入瓶颈。

提升逃逸带宽对处于技术前沿的企业至关重要，因为吞吐量是其核心竞争力。对于英伟达而言，其NVLink纵向扩展架构是重要护城河，这一瓶颈可能使AMD等竞争对手及超大规模企业更容易迎头赶上。

NVLink Speeds								
	Calculation	NVLink 1.0	NVLink 2.0	NVLink 3.0	NVLink 4.0	NVLink 5.0	NVLink 6.0	NVLink 6.0 vs. NVLink 1.0 Multiple
Year		2014	2017	2020	2022	2024	2026	
Link Number	(a)	4	6	12	18	18	18	4.5x
Lane Number per Link	(b)	8	8	4	2	2	4	0.5x
Total Lane	**(a)*(b)=(c)**	**32**	**48**	**48**	**36**	**36**	**72**	**2.3x**
Data Rate per Lane (Gb/s)	**(d)**	**20**	**25**	**50**	**100**	**200**	**200**	**10.0x**
Total Bidirectional Bandwidth (GB/s)	**((c)*(d))/8**	**160**	**300**	**600**	**900**	**1,800**	**3,600**	**22.5x**

来源：英伟达、SemiAnalysis



来源：英伟达、SemiAnalysis

解决这一困境的方案——或者说必要的折衷方案——是尽可能缩短电气I/O距离，并将数据传输卸载至尽可能靠近主机ASIC的光学链路，从而实现更高带宽。正因如此，CPO被视为互连技术的"圣杯"。无论通过基板还是中介层，CPO都能在ASIC封装上实现光通信。 电信号仅需在封装基板内传输几毫米距离，或

理想情况下，应通过更高品质的中介层实现更短的传输距离，而非通过数十厘米的损耗性覆铜层压板（CCL）。

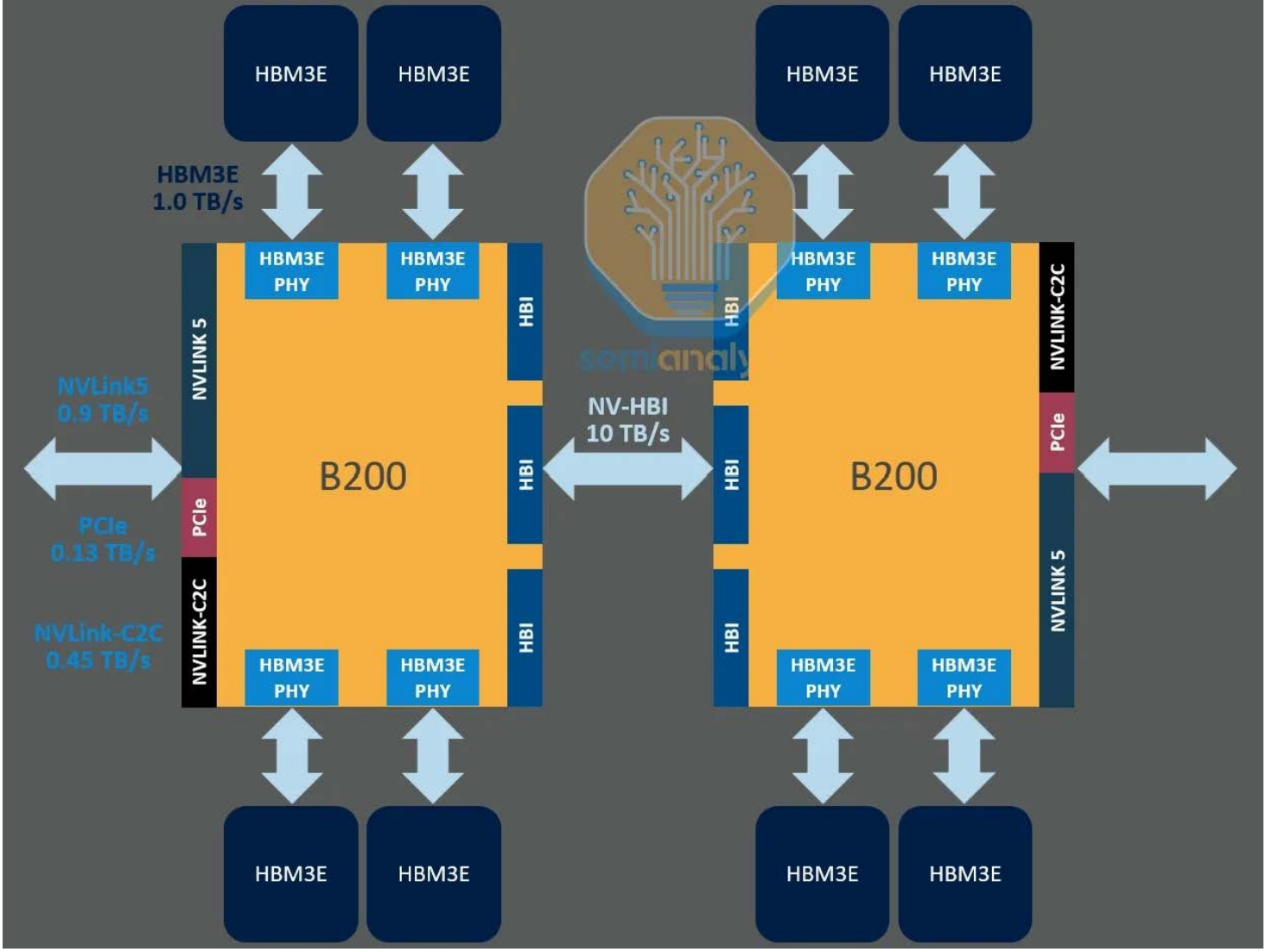
相反，SerDes可针对短距离传输进行优化，其所需电路远少于同等长距离的SerDes。这不仅简化了设计流程，还降低了功耗和硅片面积。这种简化使得高速SerDes更易实现，并延长了SerDes的扩展路线图。然而，我们仍受制于传统带宽模型——带宽密度仍与SerDes速度成正比增长。

为实现更高的带宽密度，在极短距离内采用宽I/O物理层是优于串行器/解串器接口的选择，其单位功耗带宽密度更优。但宽I/O技术需以更复杂的封装工艺为代价。不过对于CPO而言，这点差异可忽略不计：其封装技术本就高度先进，集成宽I/O物理层几乎不会增加额外封装复杂度。

宽I/O与串行器/解串器

当无需驱动电信号传输至较长距离时，我们完全可以摆脱串行化接口，转而采用宽接口——其在短距离内能提供更优的海岸线密度。

UCIe接口便是典型案例。UCIe-A可提供高达约10 Tbit/s/mm的岸线密度，专为先进封装设计（例如通过中介层实现2mm以下连接距离的芯片级封装）。在光罩尺寸芯片的长边上，其封装外带宽可达330 Tbit/s（41TByte/s）。 若两侧边缘同时使用，双向带宽可达660 Tbit/s。相比之下，Blackwell架构仅提供23.6 Tbit/s的封装外带宽，相当于约0.4 Tbit/s/mm的海岸线密度，两者差距显著。



来源：SemiAnalysis

Shoreline Density Comparison					
Location	PHY Interface	Interface Type	Edge Dimensions (mm)	Bi-directional B/W (TB/s)	B/W Density (Tbps/mm)
North+South	HBM3E PHY	Wide	101.2	8.0	0.6
East + West	NVLink5	Serialized		1.8	
	NVLink-C2C			0.9	
	PCIe Gen 6			0.3	
	Total off-package B/W		**65.0**	**3.0**	**0.4**
	Nvidia HBI	**Wide**	**32.5**	**10.0**	**2.5**

来源：SemiAnalysis

当然——这并非同等条件下的比较，因为这些离封装物理层器件需要驱动长距离传输。若真要说的话，这恰恰印证了核心观点：采用CPO技术后，传输距离不再是瓶颈，因为信号不再需要通过电信号进行长距离驱动。 当带宽密度达到10 Tbit/s/mm时，瓶颈已不再是电气接口，而是链路的其他环节——即另一端光纤能承载多少带宽。

达到这种限制是远离当今现实的终极状态，而OE

将不得不与主机共享一个中介层。在 e i n tps er is lf

在路线图上的实现难度甚至远超在基板上可靠集成OE。基板上物理层性能固然较弱，UCIe-S技术仅能提供约1.8Tbit/s/mm的海岸线密度。但相较于我们预估的224G串行器解串器约0.4Tbit/s/mm的水平，这仍属显著提升。

然而，尽管宽带接口具有诸多优势，博通和英伟达仍在其路线图中坚持采用电信号收发器（SerDes）。主要原因在于他们认为SerDes仍有扩展空间，且需要针对铜缆进行设计——尤其在光纤技术普及缓慢的背景下。此外，混合封装铜缆与光纤的解决方案似乎更可能长期存在，这迫使他们必须同时优化两种技术。

此策略旨在避免不同方案需多次流片测试的繁琐流程。

链路弹性

链路弹性和可靠性是推动CPO技术发展的另一重要因素。在大型人工智能集群中，链路停机时间是影响整体集群可用性的关键因素，即使链路可用性和稳定性获得微小提升，也能为基础设施投资带来显著回报。

如今，在接近100万个链接且采用可插拔模块的大型AI集群中，每天可能发生数十次链接中断。其中部分属于因元件故障或硬件质量导致的"硬性"故障，而多数则是源于可插拔解决方案固有复杂性和变异性的多样化根本原因所引发的"软性"故障。 故障模式呈现长尾分布，包括但不限于信号完整性问题与波动、连接器及焊线质量、元件与引脚污染、噪声注入及其他瞬态效应。这些故障与元件失效关联性较弱——因链路故障退回的光模块中，80%最终检测结果为"未发现故障"。

通过以下方式，CPO显著降低了大规模AI网络中高速信号路径的固有复杂性和变异性：

- 大幅减少光接口元件数量。在光子层面和芯片/封装层面的高度集成，降低了关键高速组件的复杂性，并提升了系统级可靠性与良率。同时减少了电光接口数量，从而

最大限度地减少每个接口处发生的功率损耗

- 显著提升主机ASIC（如交换机）与光引擎之间主机电气接口的信号完整性。通过采用设计规则和制造公差明确可控的一级封装工艺，大幅降低光引擎的插入损耗、反射及其他非线性损伤
- 减少交换机高速信号路径中端口间性能波动，该波动会增加DSP信号处理、主机与模块均衡、主机与模块固件及链路优化算法的开销与复杂度。所有可插拔模块方案及主机串解码器均需针对端口间性能差异进行设计，此类差异将导致系统复杂度提升并增加故障点。

消除光链路配置中的"人为因素"。CPO交换机或光引擎出厂时已完成全套组装和测试，确保"已知良好"状态，无需在现场进行大量操作即可完成交换机内光模块的配置。这可避免安装差异、损坏、污染以及系统与光模块间的兼容性问题。

第三部分：CPO市场化进程与部署挑战

CPO光引擎制造考量与市场推广策略

目前CPO尚未实现足以支撑大规模应用的量产规模。博通是唯一一家推出搭载CPO技术的商用系统的供应商，其Bailly和Humboldt交换机曾实现量产，但如今英伟达也加入了竞争行列。这些产品的出货量仍处于极低水平。CPO引入了诸多新型制造工艺，带来了显著的可制造性挑战。
鉴于供应链尚不成熟且可靠性数据匮乏，客户对采用该技术持谨慎态度实属情理之中。

要使CPO技术获得广泛应用，行业领军企业必须投资于这些产品的运输，推动供应链建立可扩展的制造和测试流程。
英伟达正率先投身其中，其目标是完善供应链体系，

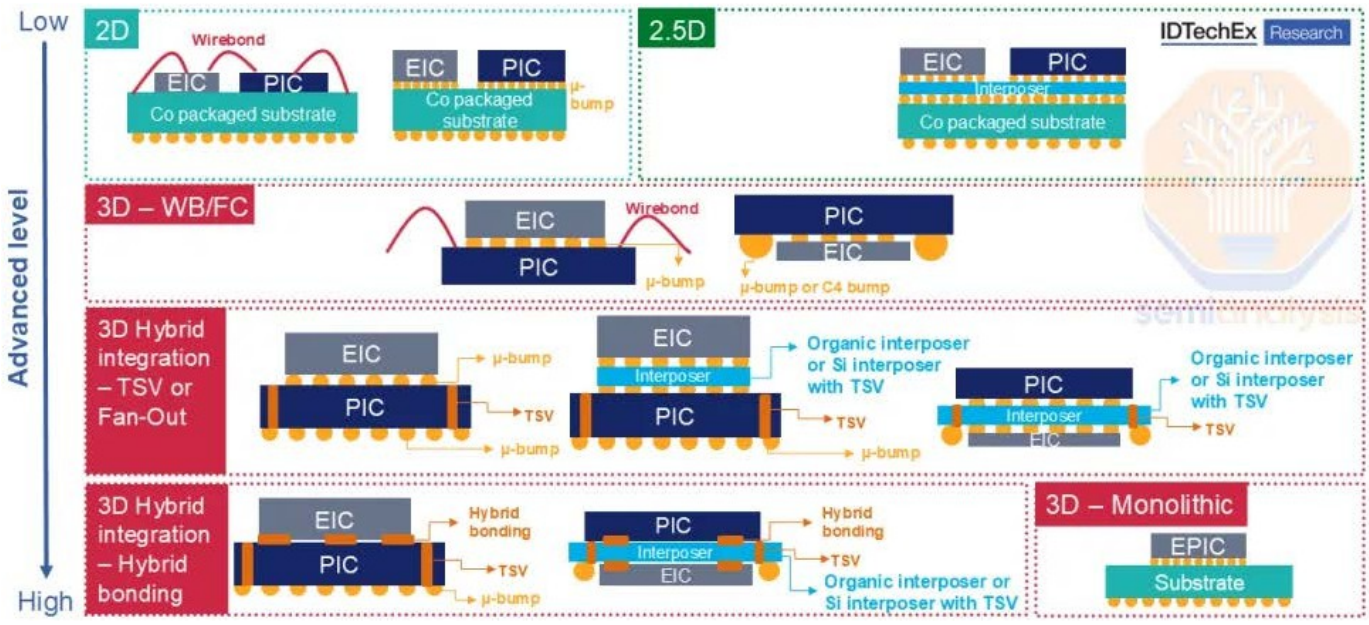
将成为"杀手级"应用：大规模网络。关于CPO有几个关键组件和考量点需要关注，这些都将影响性能和可制造性。具体包括：

1. 主机与光引擎封装
2. 光纤与光纤耦合
3. 激光源与波长复用
4. 调制器类型

主机与光引擎封装

顾名思义，“**共封装**光学器件”本质上是封装与组装方面的挑战。

光引擎同时包含光学部件与电子元件。光探测器和调制器作为光学部件，集成于"PIC"（光子集成电路）内部。 驱动器和跨阻放大器则是"EIC"（电集成电路）中的电子电路。PIC与EIC必须集成才能使光学引擎正常工作。目前存在多种封装方法实现这种PIC-EIC集成。



来源：ID TechEx

光引擎可通过将PIC与EIC集成于同一硅晶圆实现单片化。就寄生效应、延迟及功耗而言，单片集成是最优雅的解决方案。

寄生效应、延迟和功耗方面最为优雅。Ayar Laser公司正是采用此方案

第二代TeraPHY芯片组（尽管其下一代芯片组转向台积电COUPE工艺）。格罗方德、Tower和Advanced Micro Foundry等代工厂可提供单片CMOS和SiPho工艺。 然而，由于光子工艺无法像传统CMOS那样缩放，单片工艺在35纳米以下的尺寸节点便难以推进。这限制了EIC（光子集成电路）的能力，尤其在CPO系统预期更高通道速率的背景下。尽管单片集成具有固有的简洁性与优雅性，但这使其成为缩放过程中的关键障碍。正因如此，Ayar Labs也正将路线图转向异构集成光子芯片（OEs），以实现更进一步的缩放。

异构集成正成为主流方案，其工艺流程包括采用SiPho工艺制造光子集成电路（PIC），并通过先进封装技术将其与CMOS晶圆上的电子集成电路（EIC）实现集成。现有封装方案种类繁多，更先进的封装技术能带来更高性能。其中，3D集成技术在带宽和能效方面表现最为优异。 EIC与PIC通信面临的核心挑战在于寄生参数对性能的侵蚀。缩短走线长度能显著降低寄生参数，从而提升耦合效率：从带宽和功耗角度看，3D集成是实现CPO性能目标的唯一途径。

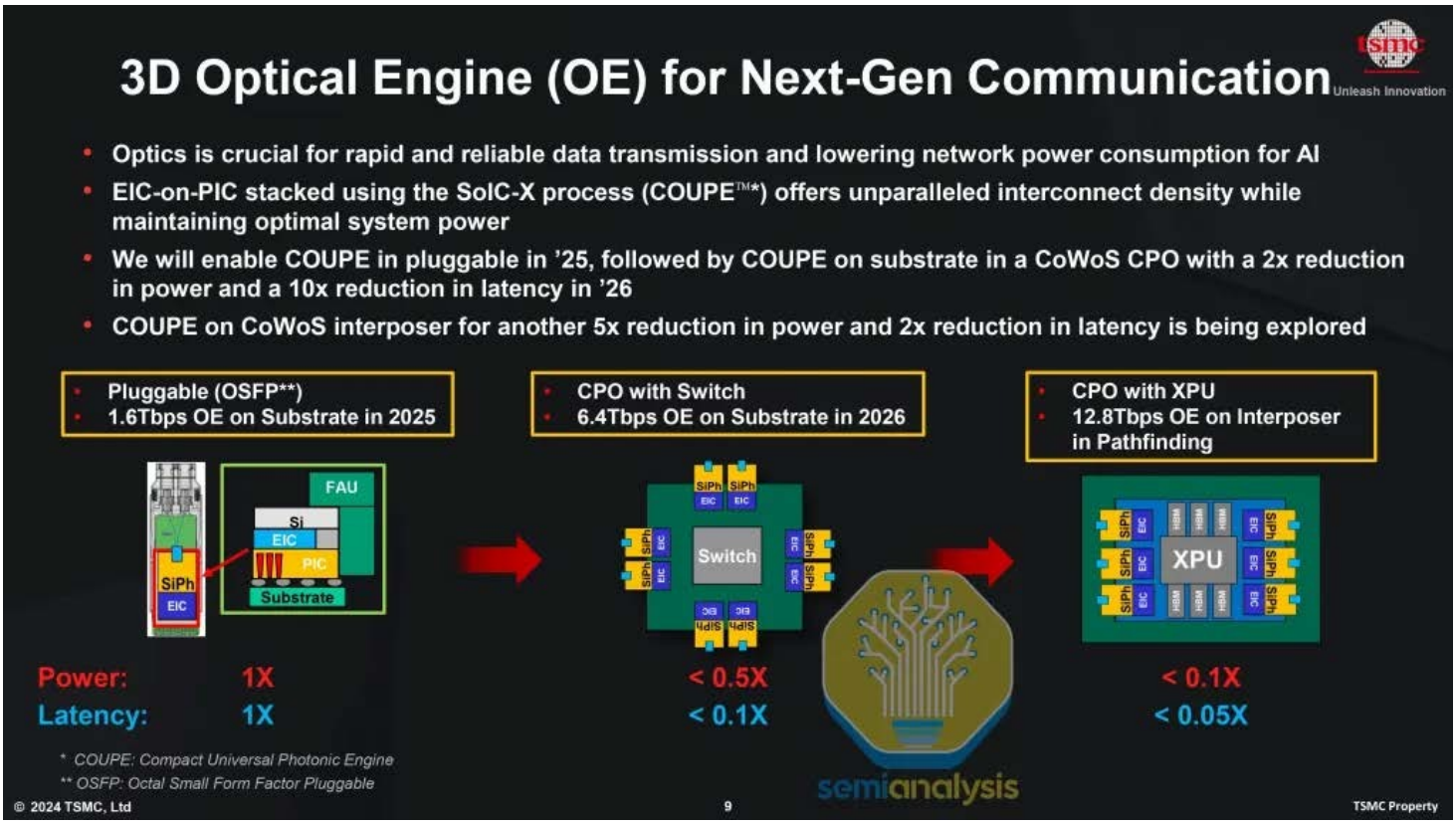
台积电COUPE正成为首选的集成方案

台积电正作为下一代光电元件制造商的首选代工合作伙伴，在无晶圆厂巨头与初创企业中加速布局。首批搭载CPO终端的高产量产品以"COUPE"（紧凑型通用光子引擎）之名面世，涵盖EIC与PIC的制造，以及台积电COUPE解决方案下的异构集成。 英伟达在GTC 2025大会上隆重展示了其COUPE光引擎，这些将成为首批量产的COUPE产品。博通公司尽管现有光引擎产品线仍与其他供应链伙伴合作，但未来路线图也将采用COUPE方案。如前所述，此前依赖格罗方德Fotonix平台生产单片光引擎的Ayar Labs，如今同样将COUPE纳入其产品规划。

与在传统CMOS逻辑领域的主导地位不同，台积电此前在硅光子学领域的布局有限，该领域主要由格罗方德和台积电的合作伙伴塔半导体主导。然而近年来，台积电在光子技术能力方面正迅速迎头赶上。台积电还凭借无可争议的工艺优势，在光子芯片制造领域占据领先地位。
光子技术领域的差距。台积电还凭借其无可争议的

台积电在EIC组件的前沿CMOS逻辑技术方面实力雄厚，同时拥有领先的封装能力——作为唯一成功实现大规模晶粒到晶圆混合键合的代工厂，台积电已批量交付多种AMD混合键合芯片。混合键合是连接PIC和EIC的更高效方案，但成本显著更高。 英特尔正致力于开发类似技术，但在开拓这项技术过程中面临重大挑战。

总体而言，尽管台积电此前在独立硅光子领域实力较弱，如今已然成为CPO领域的重要参与者。与其他主要玩家一样，台积电致力于尽可能掌控价值链的各个环节。采用台积电COUPE解决方案的客户，实际上承诺将使用台积电制造的光子芯片，因为台积电不为其他代工厂的硅光子晶圆提供封装服务。 许多专注于CPO领域的公司确实已果断转向采用台积电的COUPE方案，将其作为未来几年的核心市场解决方案。



来源：台积电

芯片制造：台积电提供全面的芯片制造解决方案。EIC芯片采用N7工艺节点制造，集成高速光调制器驱动器和电致导电放大器（TIAs），并配备加热器控制器以实现波长稳定等功能。而PIC芯片则基于SOI N65工艺节点制造，台积电为光子电路设计、光子

布局设计与验证，以及光子电路仿真建模（涵盖射频、噪声及多波长等维度）。

EIC与PIC采用台积电SoIC键合工艺进行封装。 如前所述，更长的走线长度意味着更多寄生参数，从而降低性能。台积电的SoIC作为无凸点接口，在非单片结构下实现了最短走线长度，因此是异构集成EIC与PIC的最佳性能方案。如下图所示，在等功耗条件下，基于SoIC的光电器件（OE）的带宽密度是采用凸点集成OE的23倍以上。

TABLE I. COMPARISON OF 3D INTEGRATION OF MULTIPLE WAVELENGTH SYSTEM WITH μBUMP AND SOIC

		μbump	SoIC
Pitch		1x	< 0.25x
Bond array per MRM unit		4 x 3	5 x 4
MRM density		1x	> 10.7x
Iso-power	Speed	1x	2.2x
	BWD	1x	> 23x

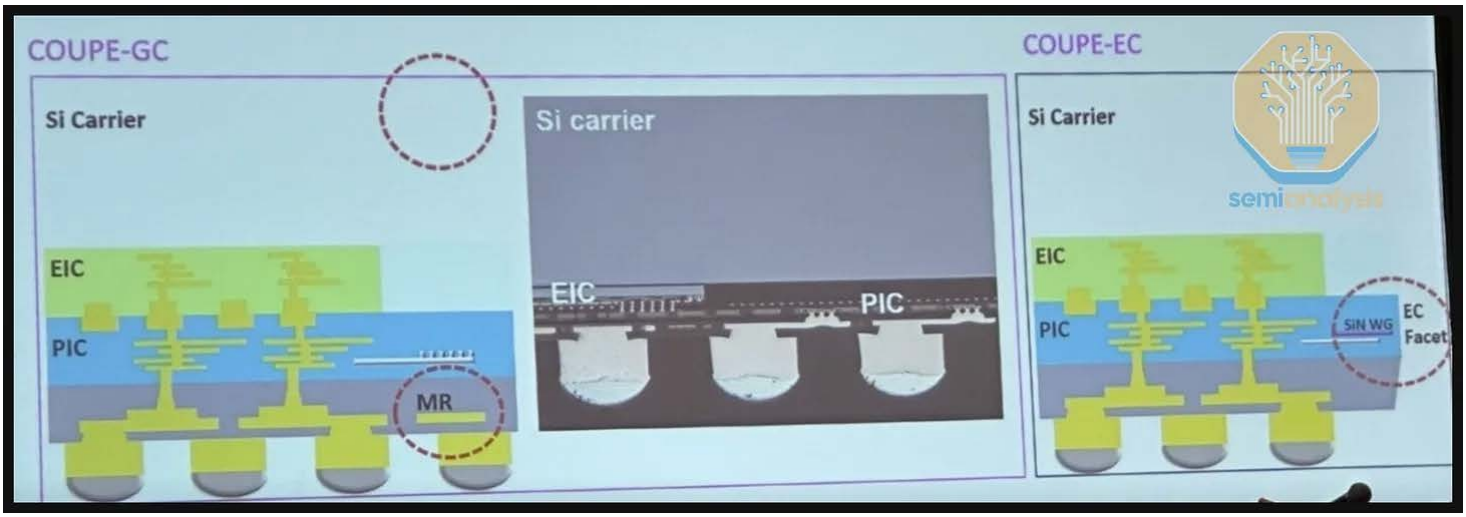
来源：台积电

COUPE支持整个光学引擎的设计与集成流程。在光I/O领域，其支持微透镜设计，可实现晶圆级或芯片级微透镜集成，并提供涵盖反射镜、微透镜、光栅耦合器（GC）及反射器的光I/O路径仿真。在3D堆叠领域，其支持3D平面布局规划、SoIC-X/TDV/C4凸点布局实现、接口物理验证，以及高频通道模型提取与仿真。为保障无缝开发流程，公司提供完整的COUPE设计验证PDK与EDA工作流，助力设计师高效实现技术方案。

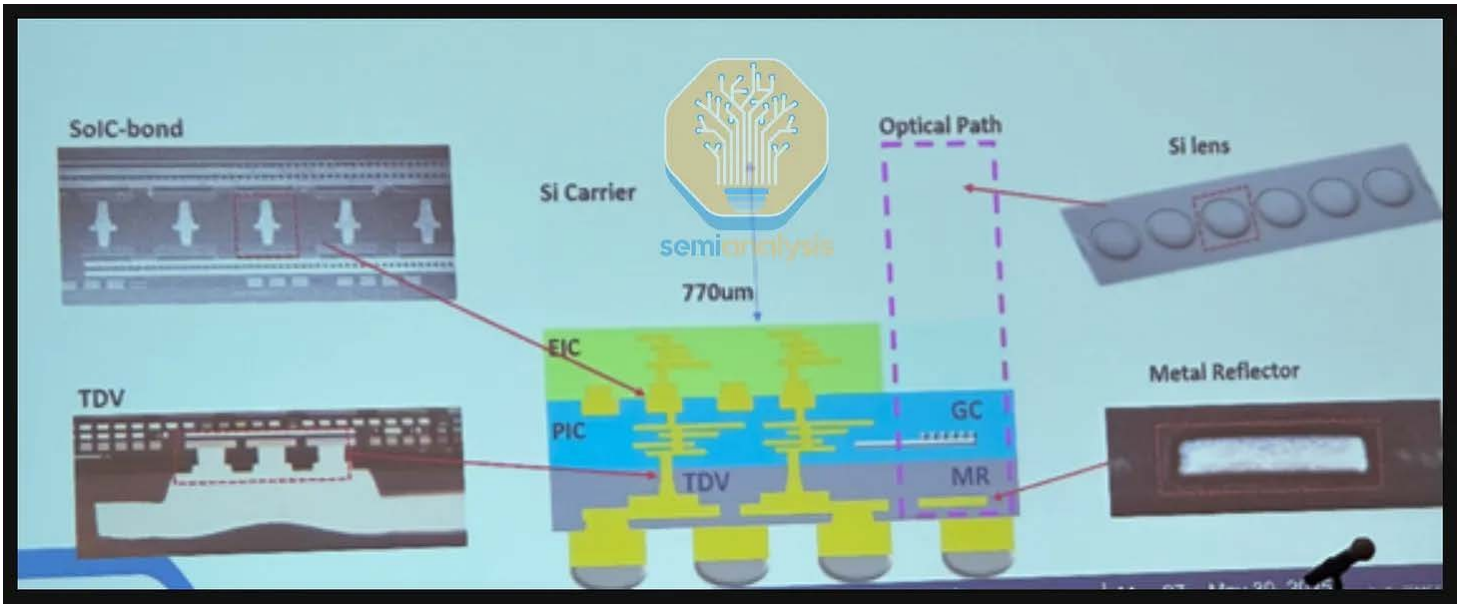
耦合：如后文将详细阐述，主要存在两种耦合方式——光栅耦合（GC）与边缘耦合（EC）。COUPE采用基于PIC无凸点堆叠结构的通用EIC实现GC与EC双重功能。但COUPE-GC结构将采用硅透镜（Si透镜）与MR（金属反射器），而COUPE-EC

将唯一具有EC端面（用于终止EC至光纤）。在of GC, he Si lens 的情况下

则基于770微米硅载体（Si-carrier）设计，MR直接置于GC下方，并配置优化光学性能所需的介质层。随后通过晶圆对晶圆（WoW）键合技术将Si-carrier与晶圆对晶圆（CoW）晶圆结合。

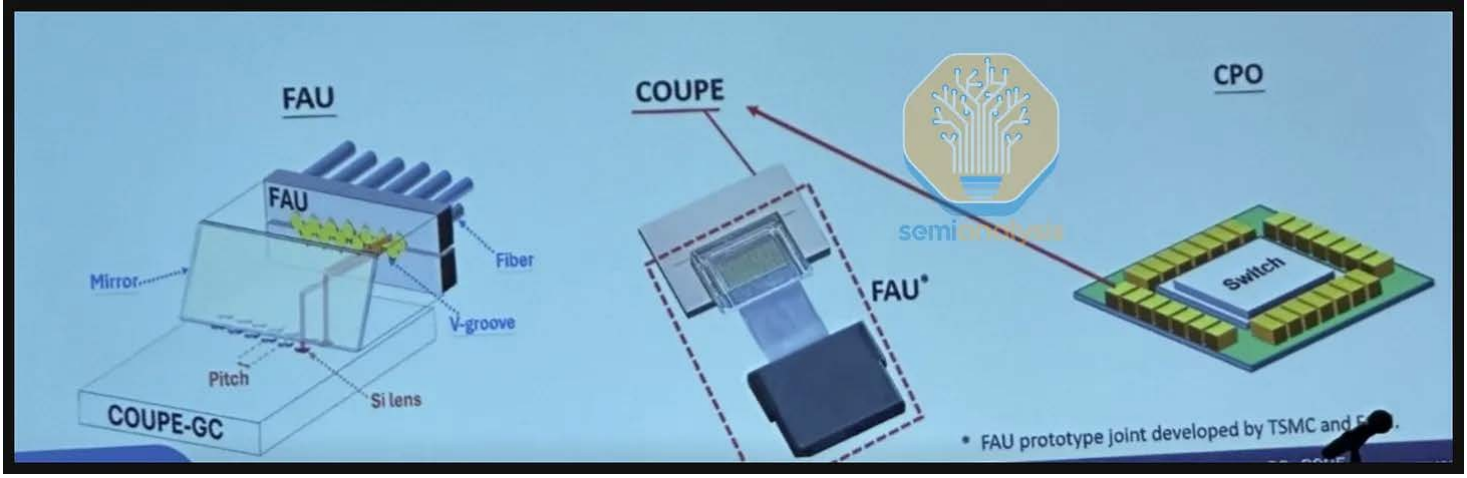


来源：台积电



来源：台积电

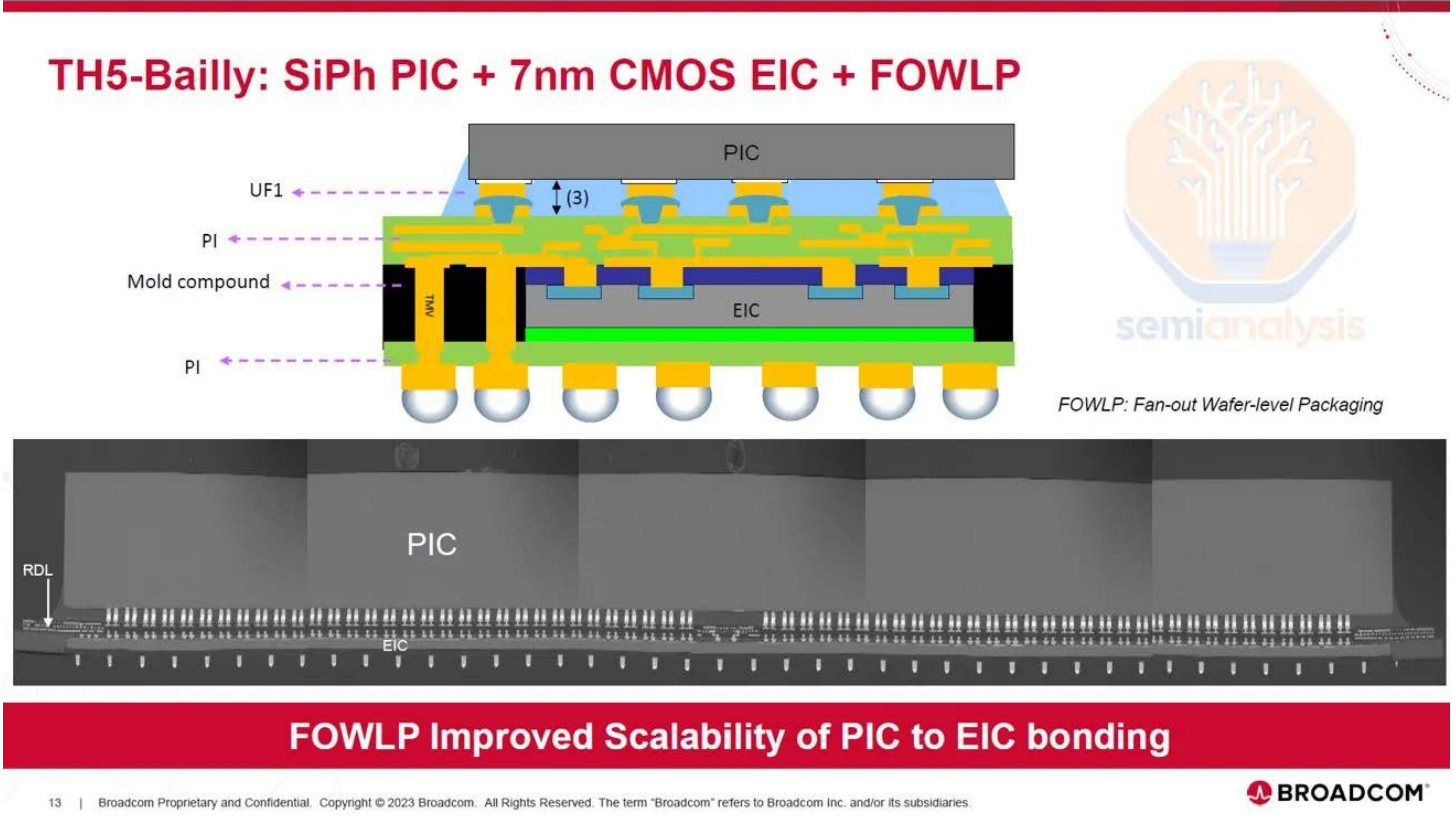
光纤耦合单元（FAU）：FAU需根据COUPE的光路进行协同设计。其设计目的是将硅透镜的光信号以低插入损耗耦合至光纤。随着输入输出端口数量增加，制造难度随之提升，但若行业能遵循特定标准，则可缩短开发周期并降低成本。总体而言，每个组件均需优化设计以实现最佳光学性能。



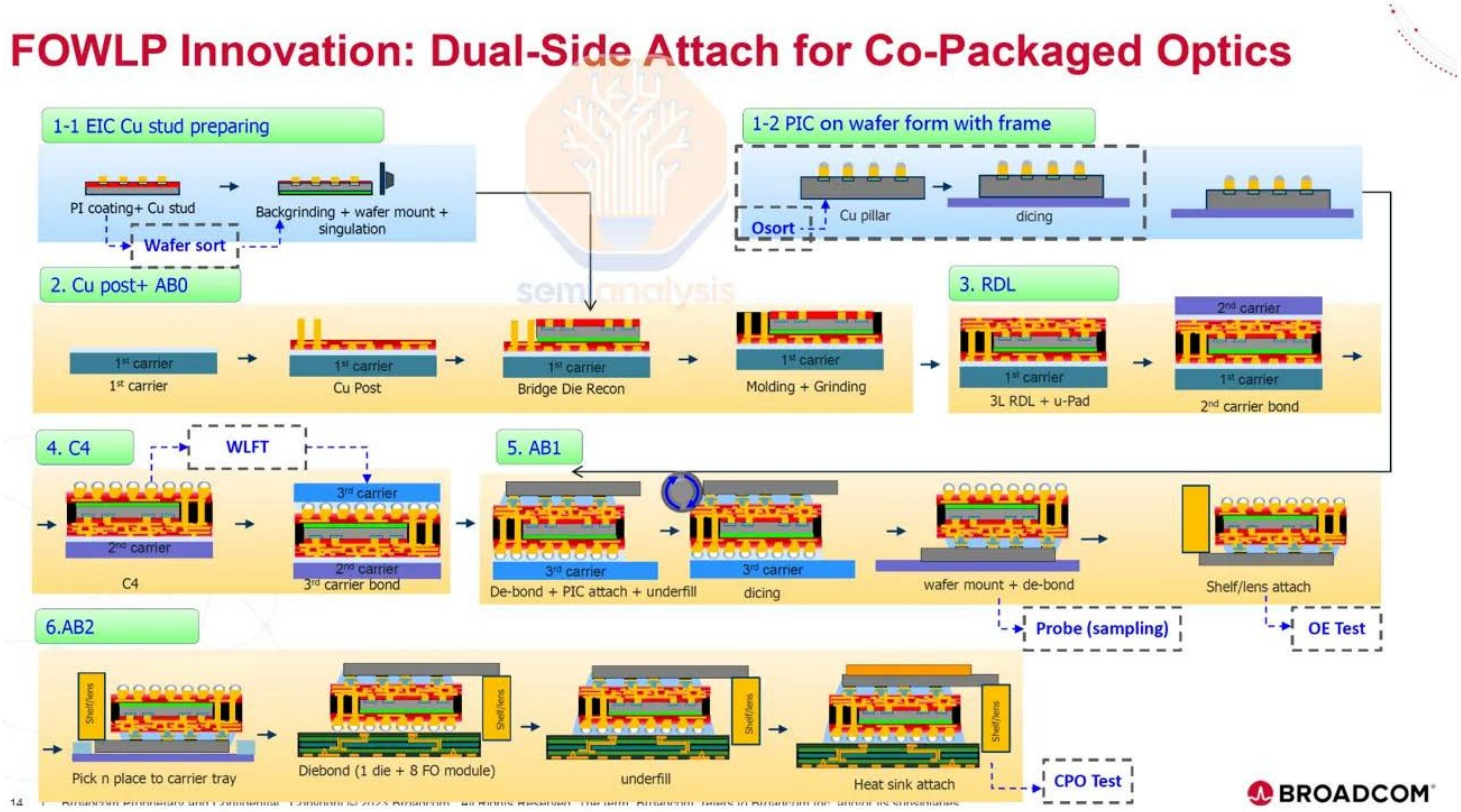
来源：台积电

产品路线图：COUPE的首批迭代产品将采用基板上光学引擎（OE）方案，最终目标是实现将OE置于中介层（interposer）之上。中介层能提供远高于基板的I/O密度，从而在OE与ASIC物理层之间实现更高带宽，单个OE可达12.8Tbit/s带宽，相当于约4 Tbit/s/mm。集成中介层的挑战在于缩放中介层尺寸（其成本高于封装基板），以适配光学引擎。

正因如此，博通正将其CPO解决方案转向台积电的COUPE工艺——尽管此前已采用新世科开发的扇出型晶圆级封装（FOWLP）技术迭代了多代CPO产品。值得注意的是，博通已承诺在未来的交换机和客户加速器产品路线图中全面采用COUPE工艺。 我们了解到，由于寄生电容过高，FOWLP技术无法实现每通道100G以上的扩展——因为电信号必须通过通模孔（TMV）才能到达EIC。为保持技术路线的竞争力，博通必须转向性能更优、可扩展性更强的COUPE技术。 这凸显了台积电的技术优势，使其即便在传统弱项的光学领域也能赢得订单。



来源：博通



来源：博通

将光学元件封装至主机

光学元件（OEs）本身放置在基板上，随后基板通过倒装芯片技术与主机封装体进行键合。为实现光学元件的共封装，需要大量封装区域。这要求大幅扩大封装基板或介质板，具体取决于其放置位置。对于英伟达的Spectrum-X前s

切换ASIC封装，基板尺寸将达110mm×110mm。作为参考，Blackwell封装尺寸为70mm×76mm，而该芯片本身已属大型芯片。

此外，在基板上附加更多元件会带来良率挑战。以Spectrum-X为例，需先将36个已知良品的光电元件进行倒装芯片键合至基板，随后才能进行"基板上"步骤的互连模块键合，最终完成共封装（CoWoS）组装。

同样地，对于中介层而言，制造更大尺寸中介层的成本高昂，且需要键合更多元件，这带来了良率方面的挑战。

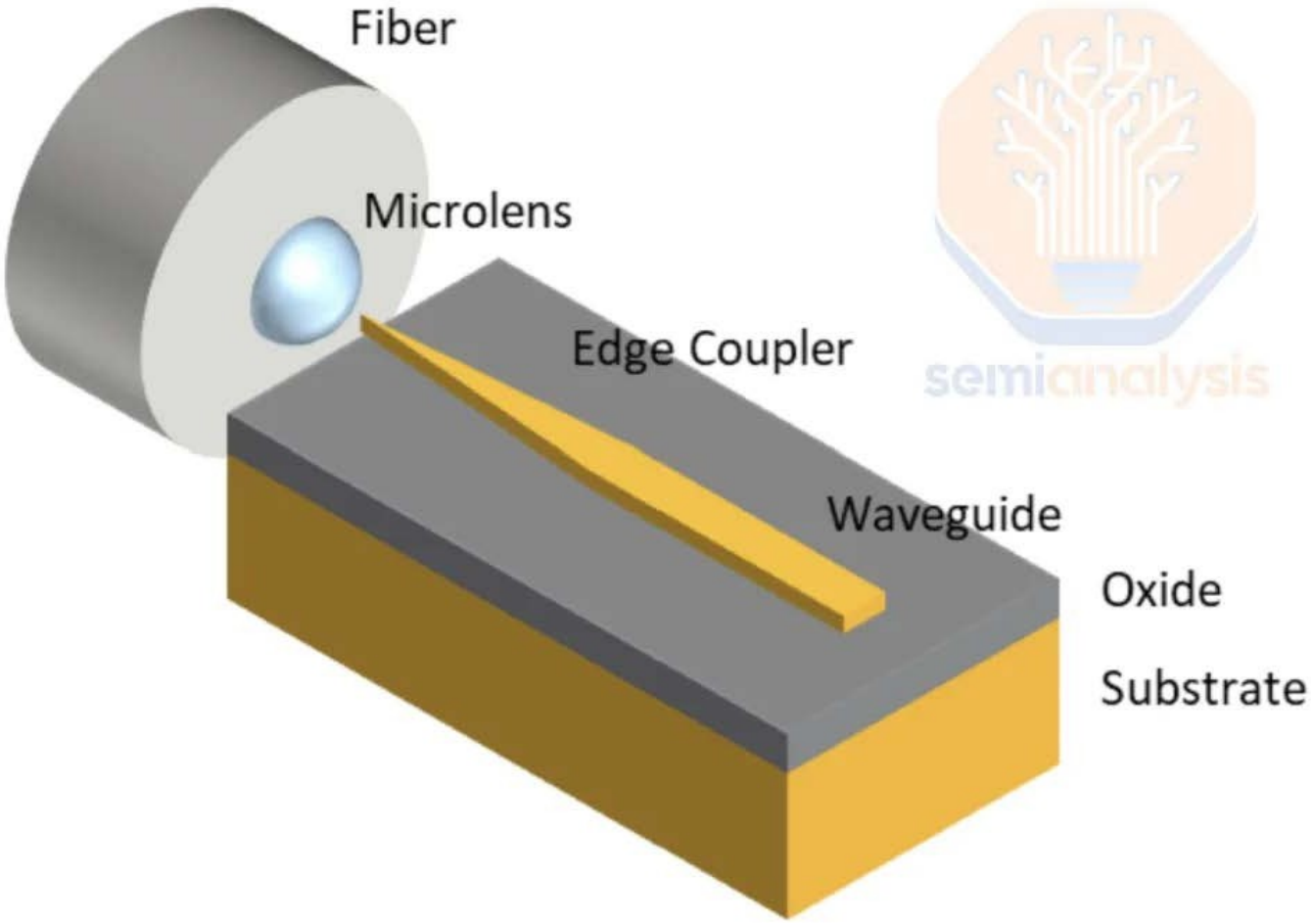
此外，翘曲问题加剧了这些挑战，随着中介层/基板尺寸的缩小，翘曲问题变得更加明显。

光纤耦合器与光纤耦合器件

光纤从光电转换器（OE）中引出以传输数据。一条光通道由两根光纤或一对光纤（发送加接收）组成。 光纤耦合——将光纤与芯片波导精确对准以实现顺畅高效的光传输——是CPO中至关重要且极具挑战性的步骤，光纤阵列单元（FAU）在CPO中被广泛应用以辅助该过程。主要有两种实现方式：边缘耦合（EC）和光栅耦合（GC）。

边缘耦合

边缘耦合将光纤沿芯片边缘对齐。从下图可见，光纤端面必须与芯片抛光边缘精确对准，以确保光束准确进入边缘耦合器。 光纤末端的微透镜将光线聚焦并引导至芯片，使其进入波导。波导锥体逐渐扩宽，实现平滑的模式过渡，从而减少反射和散射以确保耦合效率。若缺少该透镜和锥体结构，光纤端面与波导端面之间的界面将产生显著的光学损耗。



Ansys

边缘耦合器因其耦合损耗低、工作波长范围广且对偏振不敏感而备受青睐。然而它也存在若干缺点：

1. 制造工艺更为复杂，需要进行底切和深蚀刻；
2. 因属一维结构，光纤密度受限；
3. 与芯片堆叠工艺不兼容（因TSV需减薄）；
4. 在结构尺寸、机械应力、翘曲及光纤处理方面存在机械可靠性挑战；
5. 热可靠性较低；
6. 整体生态系统兼容性不足。

全球晶圆厂（GFS）在今年的VLSI大会上展示了其标志性45纳米"Fotonix"平台上实现的单片集成SiN边缘耦合器，该器件支持32通道且间距达127微米。

光栅耦合（GC）

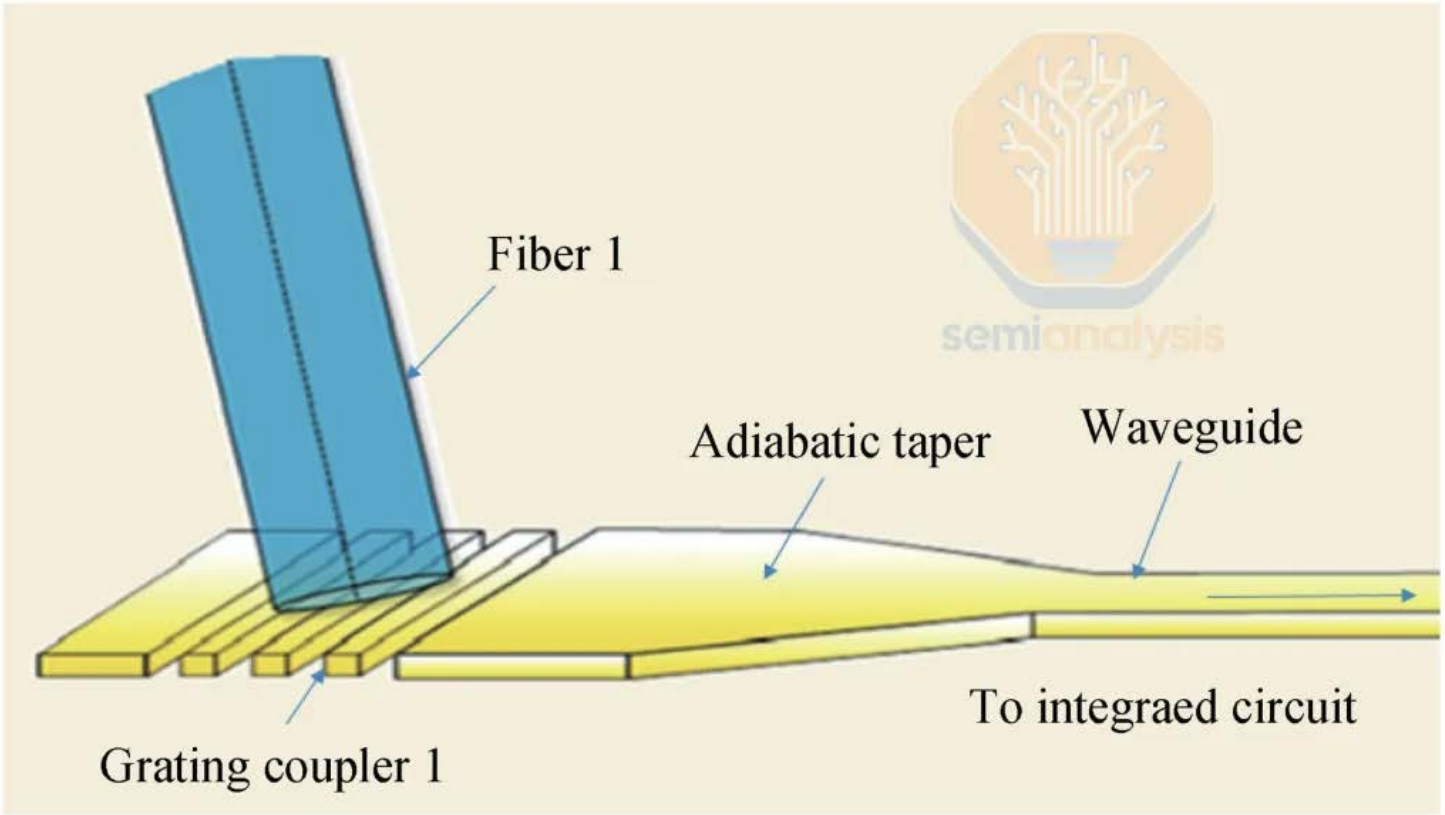
在光栅耦合器（GC）中，光线从顶部进入，光纤以微小角度置于光栅上方。当光线抵达光栅时，其周期性结构会将光线散射并向下弯曲至波导中。

光栅/垂直耦合的主要优势在于可容纳多排光纤，从而实现单个光引擎搭载更多光纤。此外，光栅耦合器无需置于基板底部，使得光引擎能够部署在中介层上。

最后，GC无需极高精度定位，且可通过简单的两步蚀刻工艺更易于制造。其缺点在于单偏振光栅耦合器仅适用于有限波长范围，且对偏振方向极为敏感。

英伟达倾向于采用GC技术，因其具备多重优势——相较于EC技术，它能实现更高2D密度、占用更小空间、更易于制造，并能简化晶圆级测试流程。然而该公司也意识到GC存在若干缺陷——通常会引入更高光损耗，且光带宽较EC更窄（后者通常能覆盖更宽的光谱范围）。

台积电同样明显更倾向于GC技术，其COUPE平台也支持该技术。



来源：《半导体杂志》

激光类型与波分复用（WDM）

将激光器集成到CPO中主要有两种方式。

第一种方案——片上激光器，将激光器与调制器集成于同一光子芯片上，通常通过将III-V族（InP）材料键合到硅基底实现。尽管片上激光器简化了设计并降低了插入损耗，但仍存在若干挑战：

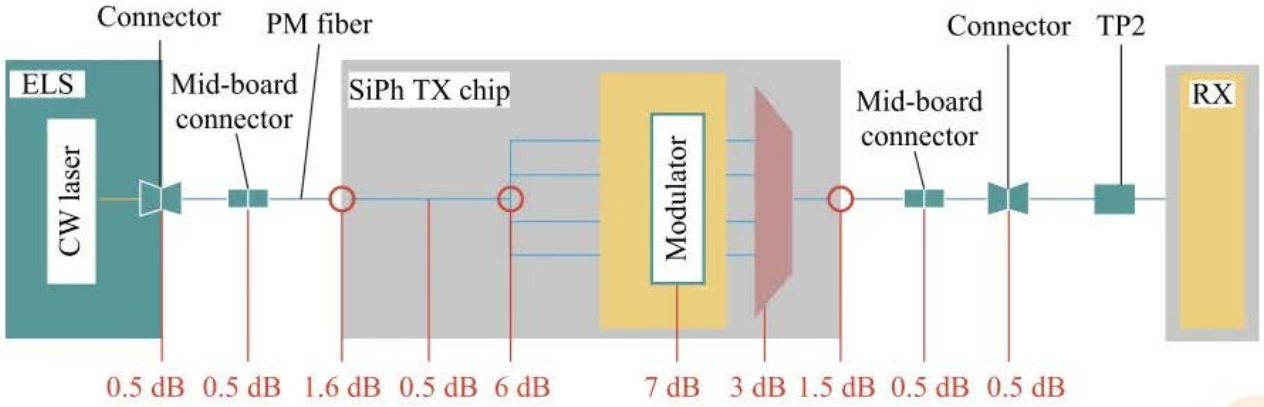
1. 激光器素来是系统中最易失效的元件之一——若集成至CPO引擎，其故障将引发连锁反应，导致整个芯片失效；
2. 激光器对热敏感，若将其放置在共封装光学元件上，将暴露于高温环境——因其紧邻系统最热部件（主硅片），这只会加剧问题；
3. 片上激光器通常难以提供足够高的功率输出。

业界达成共识的解决方案是采用外部光源（ELS）。激光器置于独立模块中，通过光纤连接至光学引擎。这类激光器通常采用OSFP等可插拔封装形式。此方案在激光器故障的常见场景下，能简化现场维护流程。

ELS方案的弊端在于功耗更高。如下方示意图所示，基于ELS的系统中，输出功率会因连接器损耗、光纤耦合损耗及调制器效率等因素在多个环节衰减。因此系统中每台激光器必须提供24.5 dBm的光功率，才能补偿损耗并确保传输可靠性。 高输出激光器在热应力下会产生更多热量并加速退化，其中激光器和热电冷却器占
约70%的ELS功耗。尽管激光器设计、封装和光路方面的渐进改进有所帮助，但激光器高功耗需求的问题尚未完全解决。

在今年的VLSI大会上，英伟达重点介绍了其生态系统中的多家激光器合作伙伴：Lumentum提供单高功率DFB激光器，Ayar Labs提供DFB阵列，Innolume提供量子点锁模梳，Xscape、Enlightra和Iloomina则提供泵浦非线性谐振梳。

英伟达同时探讨了探索垂直腔面发射激光器（VCSEL）阵列作为潜在替代方案的可能性。尽管单纤数据速率较低且可能存在散热问题，但VCSEL在功率和成本效率方面具有优势，适用于"宽带低速"应用场景。不过我们认为这并非英伟达当前的优先事项。



Element	Value	
Optical loss per fiber	21.6 dB	
TP2 requirement from 800G pluggable MSA*	0.2 dBm	
Minimum-optical power per fiber	**21.8 dBm**	
Optical power per fiber	**23.0 dBm**	**1.2 dB margin**
Optical power per laser	**24.5 dBm**	**1.5 dB coupling loss**
*TDECQ=1.4 dB		

CPO现状、挑战与解决方案

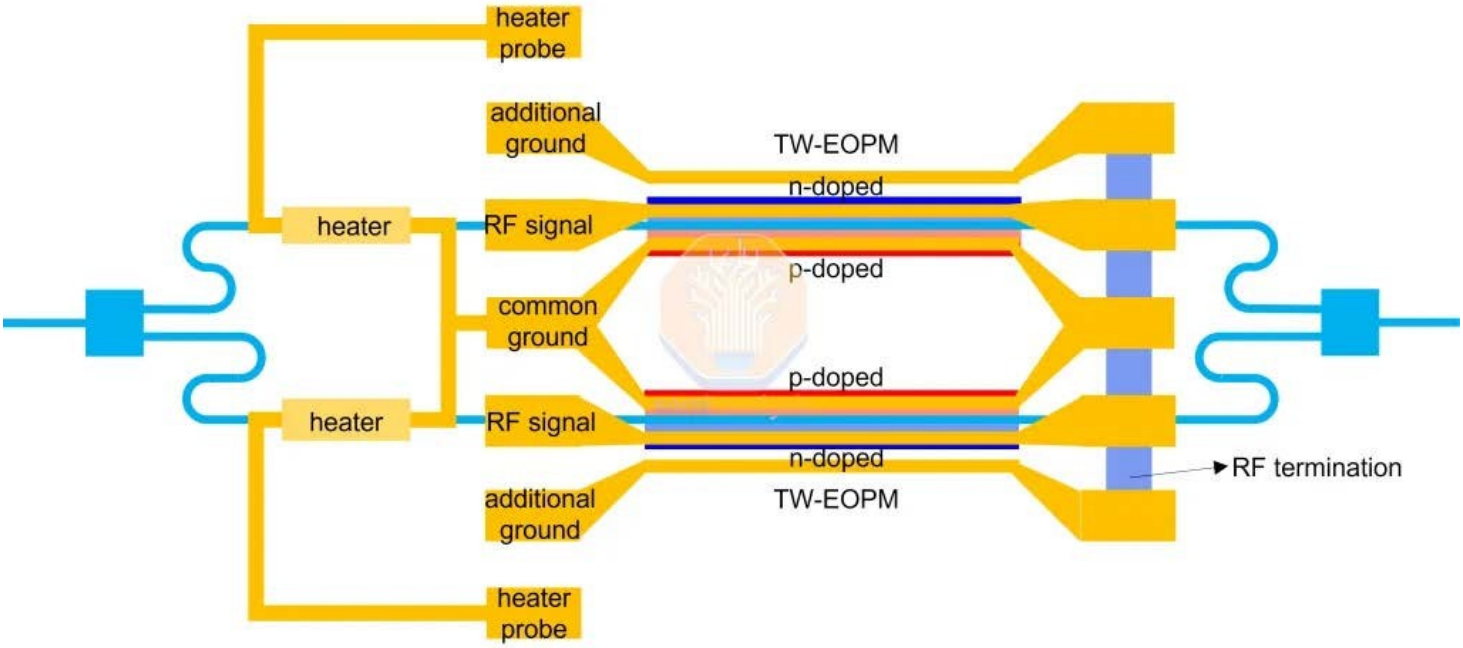
波分复用（WDM）是指通过同一根光纤传输多个不同波长（或称波长）的光信号。WDM主要分为两种常见类型：粗波分复用（CWDM）和密集波分复用（DWDM）。 CWDM通常承载较少通道且间隔较宽（典型间隔20纳米），而DWDM则以极窄间隔（常小于1纳米）密集排列大量通道。CWDM较宽的通道间隔限制了其容量，而DWDM的窄间隔可容纳40、80甚至超过100个通道。WDM技术至关重要，因为当前多数CPO方案的实现受限于可连接至光引擎的光纤数量。有限的光纤对意味着必须最大化利用每对光纤。

调制器类型

当激光进入PIC时，会经历调制阶段（由驱动器驱动），将电子信号编码为激光波长。该过程主要采用三种调制器：马赫-曾德尔调制器（MZM）、微环调制器（MRM）和电吸收调制器（EAM）。 每个独立波长（单个光通道上的特定波长）均需配备独立调制器。

马赫-曾德尔调制器（MZM）

MZM通过将连续波光信号分裂为两条波导臂来编码数据，这两条臂的折射率可通过施加电压进行调节。当臂重新结合时，其干涉图案会调制信号的强度或相位。



来源：Luceda Academy

在三种调制器中，MZM最易实现且热敏性低，减少了精密温控需求。其高线性度支持PAM4和相干QAM等先进调制格式（尽管QAM不适用于HPC/AI工作负载）。 MZM的低频移特性可提升高阶调制与长距离传输的信号完整性。该技术还支持更高通道带宽：单通道200G传输已实现，采用非相干PAM调制时单通道400G传输亦被认为可行。

然而MZM的缺点在于：

- **大型器件尺寸以毫米为单位**（相较于微米级别的MRM），因其需要两个波导臂和一个组合波导（因此包含在OE **PIC中**）。MZM**尺寸约为100**纳米。

区域，消耗更多芯片面积并限制了前沿信息收录的密度：

因此通道）包含在OE PIC中。MZM尺寸约为

~12,000平方毫米^2，EAM尺寸约为250平方毫米^2（5×50毫米），而MRM尺寸介于25平方毫米^{2}至225平方毫米^{2}之间（直径5-^{15平方}毫米^2）。这是MZM的关键缺陷之一，可能限制其缩放能力。然而，若将调制器周边的驱动器及光电控制电路纳入考量，形成完整的PIC/EIC组合时，MZM尺寸劣势的显著性可能有所降低。

- **高功耗**问题在于相位移过程需要消耗大量能量。其偏置条件（即启动电压）也高于MRM（后者在亚伏特级电压下工作）。不过像Nubis这样的公司正致力于开发巧妙的设计方案，以弥补MZM在功耗方面的劣势。

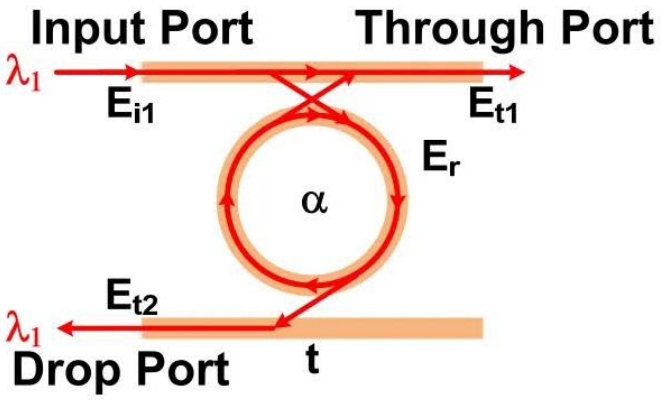
在初创企业生态中，Nubis是主要采用MZM技术实现CPO解决方案规模化的企业之一。由于MZM器件体积庞大且波长数量有限，该技术尚未在初创生态中广泛普及。

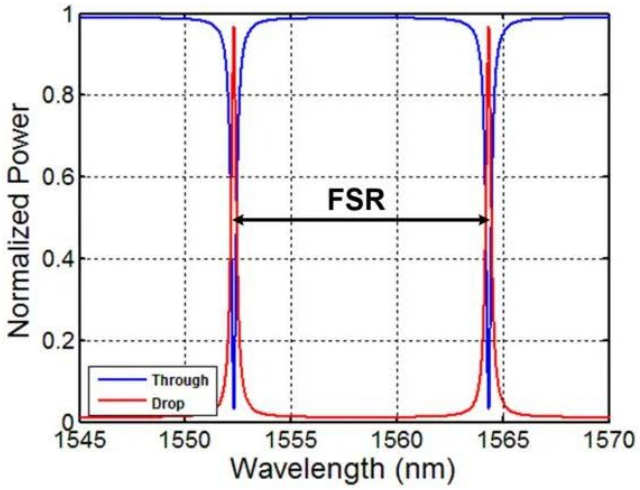
微环调制器（MRM）

MRM采用紧凑型环形波导，与一条或多条直线波导耦合。电信号改变环形波导的折射率，从而改变其共振波长。通过调节共振状态使其与输入光对齐或错位，MRM可调制光信号的强度或相位，从而实现数据编码。

光源从输入端口进入环形腔——对于大多数波长的光，环形腔不会产生共振，因此光会直接穿过设备，从输入端口到达输出端口。若波长满足共振条件，光将在环形腔内发生建设性干涉，从而被引导至分光端口。 如下图归一化功率曲线所示，特定波长的光会在分光端口产生传输功率的尖锐峰值，同时对应地导致直通端口的传输功率骤降。此效应可用于调制目的。

Ring Resonator Filter





- Ring resonators display a high-*Q* notch filter response at the through port and a band-pass response at the drop port
- This response repeats over a free spectral range (FSR)

7

来源：萨姆·帕勒莫，德克萨斯农工大学

光学引擎通常采用多个MRM，每个环形器均可调谐至不同波长，从而实现基于环形器本身的波分复用（WDM），无需额外设备即可完成波分复用功能。

MRM具有以下关键优势：

- 体积极为紧凑（尺寸级为数十微米），可实现远高于MZM的调制器密度。 MZM尺寸约为12,$^{000mm^2}$，EAM约为$^{250mm^2}$（5×50mm），而MRM尺寸介于$^{25}mm^2$至225$^{mm^2}$之间（直径5-15mm^2）；
- 环形器特别适用于波分复用（WDM）应用（包括8波长或16波长的密集波分复用DWDM），并具备内置复用/解复用功能；
- MRM可实现高能效（单位比特功耗更低）；
- 最后，环形天线具有低啁啾效应，这能提升信号质量。

然而，MRM也存在若干挑战：

- MRM的温度敏感度可达MZM和EAM的10-100倍，需要设计和制造难度极高的精密控制系统；
- 其非线性特性使PAM4/6/8等高阶调制方案复杂化；
- MRM的灵敏度与严苛的温度控制公差使得标准化困难重重，因每种设计均有精确要求。

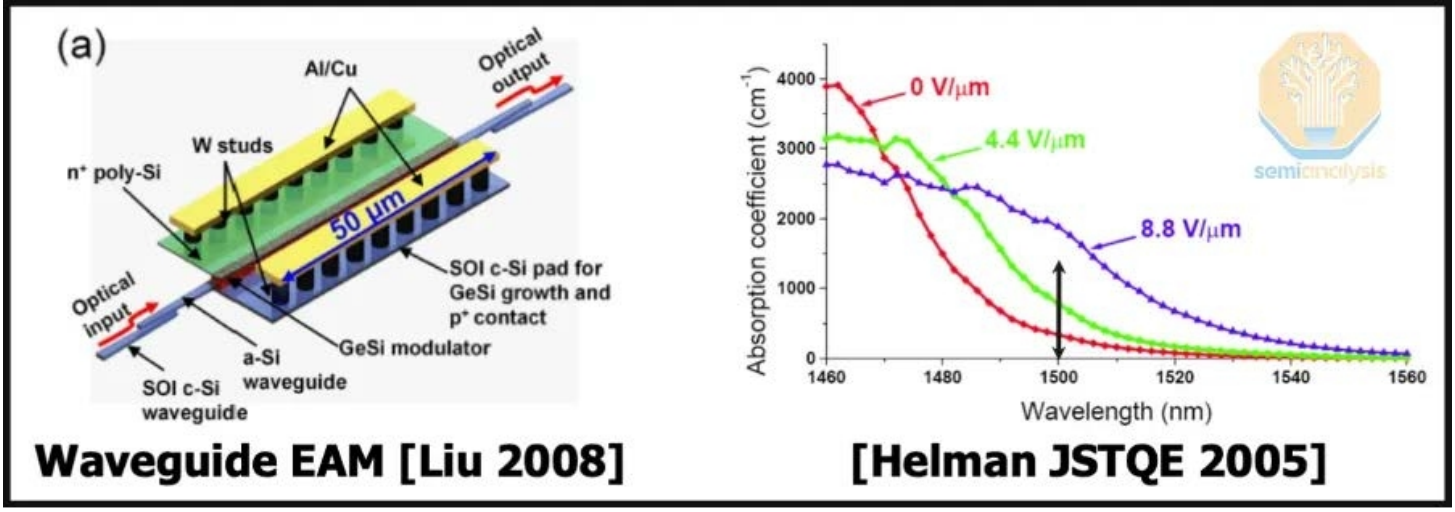
在解决方案供应商中，英伟达对MRM展现出明显偏好。该公司宣称率先设计并将MRM应用于CPO系统，认为其核心优势在于体积紧凑和低驱动电压，有助于降低功耗。但MRM技术也以控制难度高著称，设计精度成为成功实施的关键——这恰恰是英伟达的优势所在。

在制造工艺方面，台积电先进的CMOS技术专长非常适合制造具有高精度和高品质因子的MRM。此外，Tower公司也为其光子节点带来了强大的制造能力。

MRM的实现虽具挑战性但完全可行，其潜在带宽密度可超越MZM。正因如此，台积电、英伟达及Ayar Labs、Lightmatter、Ranovus等众多光处理器公司都聚焦于此技术路线。

电吸收调制器（EAM）

电致变色材料通过改变其吸收光的能力来调制信号，这种能力取决于施加的电压。更具体地说，当低电压或无电压施加于电致变色材料时，该器件允许大部分入射激光光线通过，使其呈现透明或"开启"状态。 当施加更高电压时，GeSi调制器的带隙会向高C波段范围（1500nm以上）偏移，从而提高该波段的吸收系数，并衰减（"关闭"）通过邻近波导的光信号。此现象被称为弗兰茨-凯尔迪什效应。 这种在"开"与"关"状态间的切换可调制光强，从而将数据有效编码至光信号中。



来源：德克萨斯农工大学，刘2008，赫尔曼2005

如今，采用电吸收调制激光器（EML）进行调制的收发器也遵循相同原理。 通过将连续波（CW）分布式反馈（DFB）激光器与基于InP的EAM耦合，可构建单个离散EML以调制单条通道。例如，800G DR8收发器在8条独立光纤通道上使用8个EML，每通道采用PAM4调制（2位/信号）并以约56 GBaud速率传输。 与基于GeSi的调制器不同，InP调制器的带隙对应于O波段（1310nm），该波长是所有数据通信DR光器件的标准波长，从而实现了高度的互操作性。

砷化铟调制器存在若干缺点，使其在CPO中的应用不够理想。砷化铟晶圆尺寸通常较小（3英寸或6英寸），且良率较低——这两点因素导致基于砷化铟的器件单价远高于硅基器件，后者可采用8英寸或12英寸工艺制造。此外，砷化铟与硅的耦合难度远高于锗硅与其他硅器件的耦合。

与磁阻调制器（MRMs）和微腔调制器（MZIs）相比，电致发光调制器（EAMs）具有以下优势：

- 显然——无论是电致发光分子（EAMs）还是磁致发光分子（MRMs），都具备控制逻辑和加热器以抵御温度波动，但EAMs在温度敏感性方面具有根本性优势。 相较于MRM，EAM在50°C以上具有显著更优的热稳定性，而MRM对温度极为敏感。MRM典型稳定性为70-90 pm/°C，意味着2°C温差将导致0.14nm的共振位移——远超MRM性能崩溃的临界值0.1nm。 反之，EAM可承受高达35°C的瞬时温度变化。这种耐受性对Celestial AI的技术尤为关键——其EAM调制器置于中介层电源内。

在高XPU功率计算引擎之下，其散热量达hundreds of watts of

电源模块中。EAM还可耐受约80°C的高环境温度范围，这适用于紧邻XPU而非置于其下方的芯片应用场景。

- 与MZI相比，EAM尺寸更小且功耗更低，因为MZI相对较大的尺寸需要高电压摆幅，需要放大器将SerDes信号放大至0-5V摆幅。 马赫曾德调制器（MZM）尺寸约为 12,000平方毫米，电子调制器（EAM）约为 250平方毫米（5x50 毫米），磁阻调制器（MRM）尺寸在 25平方毫米至 225平方毫米之间（直径 5-15平方毫米）。 MZI 还需消耗更多功率来驱动加热器，以使如此大的器件保持所需的偏置。

另一方面，采用GeSi EAM进行CPO存在若干局限：

- 基于硅或氮化硅的物理调制器结构（如MRM和MZI）被认为具有远高于锗硅基器件的耐用性和可靠性。 诚然，鉴于锗基器件的加工与集成难度，许多人担忧其可靠性。但Celestial公司指出，基于锗硅的电致发光调制器（本质上是光探测器的反向结构）在可靠性方面具有确定性——这得益于当今收发器中光探测器的广泛应用。
- GeSi调制器的带边自然位于C波段（即1530nm-1565nm）。 设计量子阱以将该特性转移至O波段（即1260-1360nm）是极具挑战性的工程难题。这意味着基于GeSi的EAM调制器很可能仅适用于封闭式CPO系统，难以融入开放式芯片生态系统。
- 相较于成熟的O波段连续波激光器生态系统，围绕C波段激光器构建生态可能面临规模不经济问题。多数数据通信激光器针对O波段设计，但Celestial指出1577nm XGS-PON激光器存在庞大生产规模，这类激光器通常应用于家庭及企业光纤接入场景。
- 硅锗EAM插入损耗约为4-5dB，而MRM和MZI均为3-5dB。MRM可直接实现多波长复用，而EAM需额外配置复用器才能实现CWDM或DWDM，这会略微增加潜在损耗预算。

总体而言，在当前的CPO实现中，EAMs尚未得到广泛应用，Celestial AI作为少数积极探索此方法的公司之一尤为突出。

OE路线图——扩展OE规模

当前可用的光引擎通常提供1.6T至3.2T的总带宽。 英伟达的Quantum CPO包含1.6T引擎，其Spectrum版本计划采用3.2T规格。博通展示了适用于Bailly平台的6.4T光引擎，但其体积庞大（宽度达英伟达产品的2-3倍），且需占用两个FAU插槽，因此带宽密度可能与英伟达产品相当。Marvell的光引擎同样存在此类情况。

6.4T OE 需要 2 个 FAU，因此占用较大空间。据我们所知，Marvell 的 OE 近期也不会应用于任何生产系统。

Optical Engine Specifications					
Name	Company	Year	Total number of lambdas	Lambda Speed - unidirectional (Gbps)	Bandwidth - unidirectional (Tbps)
Odin 8	Ranovus	2021	8	50	0.4
Odin 32	Ranovus	2022	32	50	1.6
Tomahawk 3.2T OE	Broadcom	2022	32	100	3.2
XT1600	Nubis	2023	16	100	1.6
XT3200	Nubis	2025	16	200	3.2
TeraPHY (1st Gen)	Ayar Labs	2023	64	32	2.0
OCI chiplet	Intel	2024	64	32	2.0
Bailly 6.4T OE	Broadcom	2024	64	100	6.4
3D SiPho Engine	Marvell	2024	32	200	6.4
TeraPHY (2nd Gen)	Ayar Labs	2025	64	64	4.0
Nvidia 1.6T OE	Nvidia	2025	8	200	1.6
Nvidia 3.2T OE	Nvidia	2026	16	200	3.2

来源：SemiAnalysis

正如我们所讨论的，Nvidia Spectrum-X光子交换机中采用的3.2T光引擎方案，其岸线带宽密度并未超越基于长距离SerDes驱动的可插拔模块。换言之，光引擎密度必须实现数倍级提升，才能带来显著的性能优势并推动客户采用。 这意味着需要同时提升主机芯片与光引擎EIC之间的电接口性能，并扩大光纤传输带宽。

但如果我们能够自由设计下一代互连技术，那么解锁当前及未来世代更大带宽的方法会有哪些？

提升带宽的关键途径

让我们探讨基于共封装光引擎的带宽扩展关键方案：

1. 继续采用基于电气SerDes的物理层设计：通过使用短距离（SR）SerDes替代长距离SerDes，实现更简化的设计实施、更小的面积占用和更低的功耗。但最终仍将受限于电气接口处的SerDes速度——该领域的发展空间已趋近极限。 此处的思路是采用过渡方案，使硅片设计师无需重构I/O架构。此外，采用电信号SerDes可灵活兼容现有可插拔光模块和/或铜缆，实现同一硅片的多模态适配。
2. 采用宽I/O物理层接口（如UCIe），以较低波特率（例如56G）配合NRZ调制运行。这种方案对光引擎的EIC要求较低，甚至可能消除昂贵的混合键合需求——因低速运行时寄生效应影响较小。然而，低信号速率意味着光引擎输出的光纤数量更易成为瓶颈。 波分复用技术通过使每根光纤并行承载多个数据流，有效缓解了这一问题。
3. 采用宽I/O物理层（如UCIe），再通过EIC串行化至较少的光纤通道。继续使用高波特率与PAM4调制以最大化每条光通道的速度，必要时通过WDM方案添加多波长，使每对光纤承载多个波长以进一步提升带宽。

解决电信号传输后，下一个挑战在于光纤能承载多少逃逸带宽。光纤总带宽取决于三大关键因素：1）光纤数量（决定光通道数）；2）每通道速率；3）每根光纤承载波长数——三者均构成可扩展的维度。

近期行业将概念划分为两大路径：**"快速窄带"与"慢速宽带"**。**"快速窄带"**设想每个光纤接入单元（FAU）配备较少光纤（最多不超过两位数），并在每对光纤上实现高速传输；而**"慢速宽带"**则基于大幅增加光纤对数量（可能采用更精细的间距）的构想，同时降低单对光纤的传输速率。

1. **更多光纤对。**光纤密度受光纤间距限制，而单个光纤阵列单元（FAU）内的光纤总数则受制于可制造性——即在良率成为问题之前可实现的最大数量。 目前光纤最小间距为127微米（μm），意味着每毫米最多容纳8根光纤。行业正致力于实现80微米间距及多芯光纤技术，以进一步提升单位面积的光纤承载量。然而增加光纤数量将带来制造难题：

   A) 光纤对准仍依赖大量手工操作，易导致良率损失，且每增加一根需对准的光纤都会降低FAU整体良率；虽然Ficontec等公司提供自动化工具，但其吞吐量仍较低；

   B) 耦合方式的选择同样关键：边缘耦合限制光纤阵列仅能单排排列，而光栅耦合技术可支持多排布局。目前我们所见最大规模的光纤阵列是Nubis公司的36芯二维FAU。

2. **每条通道的速度**。影响通道速度的因素主要有两个维度：

   A) **波特率**：定义每秒传输的符号数量；当前先进系统运行在100Gbaud，而行业正推动实现200Gbaud。然而更高的波特率对调制器提出更高要求，需在更高频率下切换；在各类调制器中，MZM在该指标上表现最优，且实现200Gbaud的技术路径相对清晰。

   B) **调制方式**：决定单符号承载的比特数。当前主流方案为NRZ（单符号1比特）和PAM4（通过4种幅度实现单符号2比特）。 研究正向PAM6（约每符号2.6比特）和PAM8（每符号3比特）拓展。更高阶调制方案可通过结合多幅值与多相位光信号实现。DP-16QAM采用两个正交平面，各含4种幅值与4种相位，共256种信号组合——每信号传输8比特。

3. **波分复用（WDM）**。光纤可同时承载多波长光信号。例如，一根承载8个波长（每个波长传输200Gbit/s数据）的光纤，总传输容量可达1.6Tbit/s。当前商用DWDM解决方案通常提供8波长或16波长配置。 研究人员正探索宽谱、带宽复用及交织技术以提升波长数量。扩展波长数量的关键挑战在于开发可靠激光源，实现多通道光信号的高效稳定生成。Ayar

实验室的 超新星光源配备了一台可产生16种[illegible]的激光器。

由Sivers公司提供）。Scintil公司的晶圆级InP激光器同样可提供多达16个波长，而Xscape Photonics正致力于开发具备64波长调谐能力的梳状激光器。在调制器领域，MRM调制器最适合处理多波长信号，并具备内置的复用（mux）与解复用（demux）功能。

下表概述了将光引擎扩展至12.8T及更高容量的多种方案。

Approaches to Scaling Optical Engines to 12.8T and Beyond															
Description	Fiber pitch	# fibers / mm per row	# fibers per row	Implied FAU edge (mm)	Number of rows	Total fibers in FAU	# Optical lanes	# laser fibers	Modulation scheme	Lane Speed (Gbps) uni-directional	# Wavelengths /lambda per channel	Bandwidth per optical lane (Gbps)	Total # modulators ie. lambdas	Bandwidth density (Gbps/mm)	Bandwidth density (Tbps/mm)	Total OE BW (Tbps)
			(a)		(b)	(c) = (a) x (b)	(d)	(e) = (d) / 4		(f)	(g)	(h) = (f)x(g)	(i) = (d) x (g)			(j) = (d) x (h)
Micro Ring Modulator																
Scaling fibers (50G MRM)	0.127	7.9	9	1.1	1	9	4	1	NRZ	50	1	50	4	200	0.2	**0.2**
	0.127	7.9	18	2.3	2	36	16	4		50	1	50	16	800	0.8	**0.8**
	0.127	7.9	27	3.4	3	81	36	9		50	1	50	36	1,800	1.8	**1.8**
	0.127	7.9	36	4.6	4	144	64	16		50	1	50	64	3,200	3.2	**3.2**
Scaling fibers (200G MRM)	0.127	7.9	9	1.1	1	9	4	1	PAM4	200	1	200	4	800	0.8	**0.8**
	0.127	7.9	9	1.1	2	18	8	2		200	1	200	8	1,600	1.6	**1.6**
	0.127	7.9	18	2.3	1	18	8	2		200	1	200	8	1,600	1.6	**1.6**
Scaling colors (50G MRM)	0.127	7.9	18	2.3	1	18	8	2	NRZ	50	4	200	32	1,600	1.6	**1.6**
	0.127	7.9	18	2.3	1	18	8	2		50	8	400	64	3,200	3.2	**3.2**
	0.127	7.9	18	2.3	1	18	8	2		50	16	800	128	6,400	6.4	**6.4**
Scaling colors (200G MRM)	0.127	7.9	18	2.3	1	18	8	2	PAM4	200	4	800	32	6,400	6.4	**6.4**
	0.127	7.9	18	2.3	1	18	8	2		200	8	1,600	64	12,800	12.8	**12.8**
Scaling colors and fibers (50G MRM)	0.127	7.9	18	2.3	2	36	16	4	NRZ	50	4	200	64	3,200	3.2	**3.2**
	0.127	7.9	18	2.3	2	36	16	4		50	8	400	128	6,400	6.4	**6.4**
	0.127	7.9	18	2.3	2	36	16	4		50	16	800	256	12,800	12.8	**12.8**
Scaling colors and fibers (200G MRM)	0.127	7.9	18	2.3	2	36	16	4	PAM4	200	4	800	64	12,800	12.8	**12.8**
	0.127	7.9	18	2.3	2	36	16	4		200	8	1,600	128	25,600	25.6	**25.6**
Mach Zender Modulator																
Scaling fibers (200G MZM)	0.127	7.9	9	1.1	1	9	4	1	PAM4	200	1	200	4	700	0.7	**0.8**
	0.127	7.9	12	1.5	2	24	11	3		200	1	200	11	1,400	1.4	**2.1**
	0.127	7.9	12	1.5	3	36	16	4		200	1	200	16	2,100	2.1	**3.2**
	0.127	7.9	16	2.0	5	80	36	9		200	1	200	36	3,500	3.5	**7.1**
Scaling fibers (400G MZM)	0.127	7.9	12	1.5	1	12	5	1		400	1	400	5	1,400	1.4	**2.1**
	0.127	7.9	12	1.5	2	24	11	3		400	1	400	11	2,800	2.8	**4.3**
	0.127	7.9	16	2.0	3	48	21	5		400	1	400	21	4,199	4.2	**8.5**
	0.127	7.9	16	2.0	5	80	36	9		400	1	400	36	6,999	7.0	**14.2**
	0.127	7.9	16	2.0	8	128	57	14		400	1	400	57	11,199	11.2	**22.8**

来源：SemiAnalysis

CPO应用进程与部署挑战

英伟达首批CPO产品将面向后端横向扩展交换机，其InfiniBand CPO将于2025年下半年上市，以太网CPO交换机则计划于2026年下半年推出。我们认为这一初始阶段主要作为市场测试和供应链成熟度提升的准备期。预计2026年总出货量将在1万至1.5万台之间。

要使CPO技术在部署中取得更快速、更广泛的普及，必须具备更具说服力的理由。可能的推动因素有二：一是采用CPO能带来显著的总体拥有成本优势；二是当驱动信号从交换机ASIC传输至机箱前面板所需的电气长距离串行器解串器（LR SerDes）遇到速度或距离瓶颈时。

抵消TCO优势的两大因素正是数据中心运营商抵触CPO系统的根源：缺乏互操作性与可维护性

挑战。

CPO的挑战不仅限于封装内部，更延伸至整个系统。光纤管理、前面板密度、外部激光器等关键环节均存在挑战。要实现CPO技术落地，芯片厂商需提供客户可直接部署的端到端解决方案。这延续了当前行业趋势，尤以Nvidia为代表——其专注于通过系统设计实现性能扩展。

专有解决方案与行业标准之争

CPO技术推广面临的核心挑战在于实现互操作性，同时克服行业对成熟且高度互通的可插拔光模块模式的根深蒂固依赖。

互操作性主要包含三类关键要素：(1) 电气互操作性，(2) 光学互操作性，(3) 机械互操作性。对于可插拔模块，互操作性通常由OIF（光互连论坛）负责处理。

1. 通常由OIF（光互连论坛）负责规范。
2. 通常由IEEE（有时也由OIF）负责处理。 IEEE通过其IEEE 802.3标准发挥核心作用，该标准定义了以太网物理介质相关层（PMD）。这些规范涵盖关键参数，包括调制格式、通道速率、通道数量、传输距离、介质类型以及光信号波长。通过遵循标准化PMD规范，不同厂商的收发器可实现互换操作，确保在多厂商生态系统中实现真正的即插即用兼容性。
3. 而多源协议（MSA）通常负责处理特殊需求。MSA定义了特定解决方案，确保在官方IEEE标准之外实现多厂商互操作性。

通过结合OIF、IEEE标准和MSA规范，可插拔收发器实现了广泛的互操作性，并构建了强大的多供应商生态系统。对于CPO：

1. CPO模块必须满足电气兼容性要求，否则将无法与先进的SerDes进行通信。
2. 光互操作性至关重要，因为它能确保与集群中其他标准可插拔模块兼容。
3. 需注意的是，CPO技术仍处于"狂野西部"阶段，部分解决方案和架构决策正导致完全专有化的形态

。这正是OIF高密度互连新标准（如CPX范式）试图解决的问题。

一旦(1)+(2)+(3)得到解决，CPO在操作层面将与可插拔设备高度相似，这将有助于推动其广泛应用。

然而，目前CPO尚未像可插拔模块那样广泛采用标准规范，也无法像光收发器那样确保充分的互操作性。 部分挑战在于供应商正倾向于推广系统级解决方案，而非单纯向设备制造商销售芯片。这是因为CPO的挑战已延伸至封装之外，涉及整个系统。光纤管理、前板密度、调制器架构及外部激光器等环节均是亟待解决的关键难题。为推动CPO普及，英伟达等企业需率先提供端到端解决方案。

实现这一目标的一种方法是在组件层面采用标准化解决方案，即封装在同一封装内的光电元件（OEs）遵循标准化光纤接口，并与符合以太网标准或多波长、速度及调制技术多方协议（MSA）的光子元件（如激光器、调制器和光电二极管）对接。这将实现真正的互操作性，使客户能够自由组合不同供应商的产品，而无需从单一供应商处采购全部设备。 若光电元件支持插拔式设计，运营商在设备故障时可轻松更换。此类标准化还将催生更具竞争力且稳健的多供应商市场，从而降低客户成本。客户亦可根据交货周期、价格或独特功能选择不同供应商，摆脱对单一供应商的依赖。

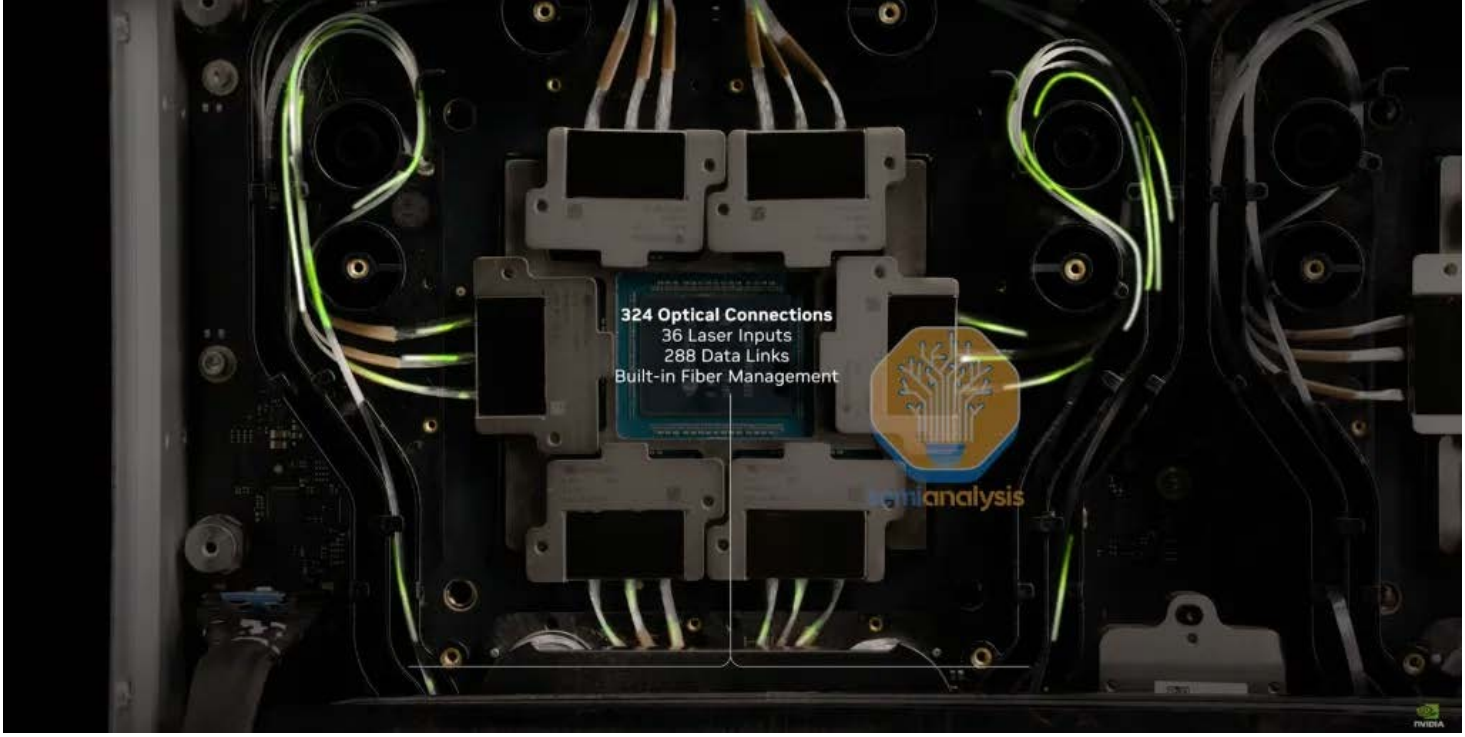
可维护性与可靠性

在CPO技术得以广泛应用之前，必须解决的一个关键问题源于光纤耦合技术。在可插拔光模块中将光纤连接至硅光子器件相对简单，而将光纤耦合至CPO的光引擎则面临更大挑战。 在CPO系统中，光纤必须以亚微米级精度精确对准，才能将光耦合到芯片上极其微小的波导（宽度和高度通常均小于微米级）；而可插拔模块则采用预对准的标准化连接器，操作更为便捷。 此外，基于CPO的系统需在狭窄且热活性强的开关机箱内完成光纤耦合，而可插拔模块的光纤耦合则远离设备主体封装。

与光通信相比，铜基电气通信存在插入损耗和信号完整性较差的问题，但在其他方面铜材通常表现可靠。光器件本身具有温度敏感性。 工作温度的变化会改变激光波长、降低元件效率并影响可靠性，通常需要专门的温度控制或校准机制。此外，光子元件也会随时间自然退化，因老化、污染或机械应力逐渐丧失光学效率，进一步挑战长期可靠性。环境因素如灰尘、湿度和机械干扰对光系统的影响尤为显著，其脆弱性远高于铜基电气链路。

这些环境与操作挑战直接延伸至物理光纤本身。光纤性能对物理干扰极为敏感，尤其易受弯曲影响——弯曲不仅会增加光插入损耗，还会加速断裂与故障。在包含多个光纤阵列的机箱中（通常每个引擎配备两个阵列：一个用于连接器，另一个从ELS连接光源），每个阵列的布线都需严格遵循拓扑考量。 每条独立光纤链路均需定制专属长度，既要考虑面板至各光引擎的距离差异，还需满足相邻阵列施加的布线约束条件。

从下方Quantum-X CPO交换机的特写图像中可见，从光端机（OEs）引出的光纤带需通过光纤盒进行布线，以实现对光纤的规范管理。 光纤单元（FAU）采用可拆卸设计，便于更换损坏部件。但交换机内部更复杂的光纤布线结构意味着，光纤/FAU的更换远比更换损坏的可插拔收发器困难得多——后者只需在面板前端进行热插拔即可完成。 对于CPO交换机，工程师需进入机箱内部，拆卸损坏的FAU，再通过卡匣正确安装新FAU。操作过程中必须避免干扰其他光纤。尽管英伟达强调CPO的可靠性优势，其可维护性仍是值得深入探讨的另一关键要素。



来源：英伟达

第四部分：当前与未来的**CPO**产品

在本部分，我们将首先介绍当前已上市或即将面世的CPO产品，从英伟达和博通的产品组合开始，进而阐述多家专注于CPO技术公司的解决方案。 **我们将涵盖英特尔CPO技术、联发科的CPO研发进展，以及Ayar Labs、Nubis、Celestial AI、Lightmatter、Xscape Photonics、Ranovus和Scintil等企业的**解决方案，详细剖析各供应商的技术方案，并评估其核心方案的优劣势。 最后我们将回归供应链议题，涵盖光引擎、外部激光源等关键CPO组件的制造、测试与组装环节。

英伟达**CPO**

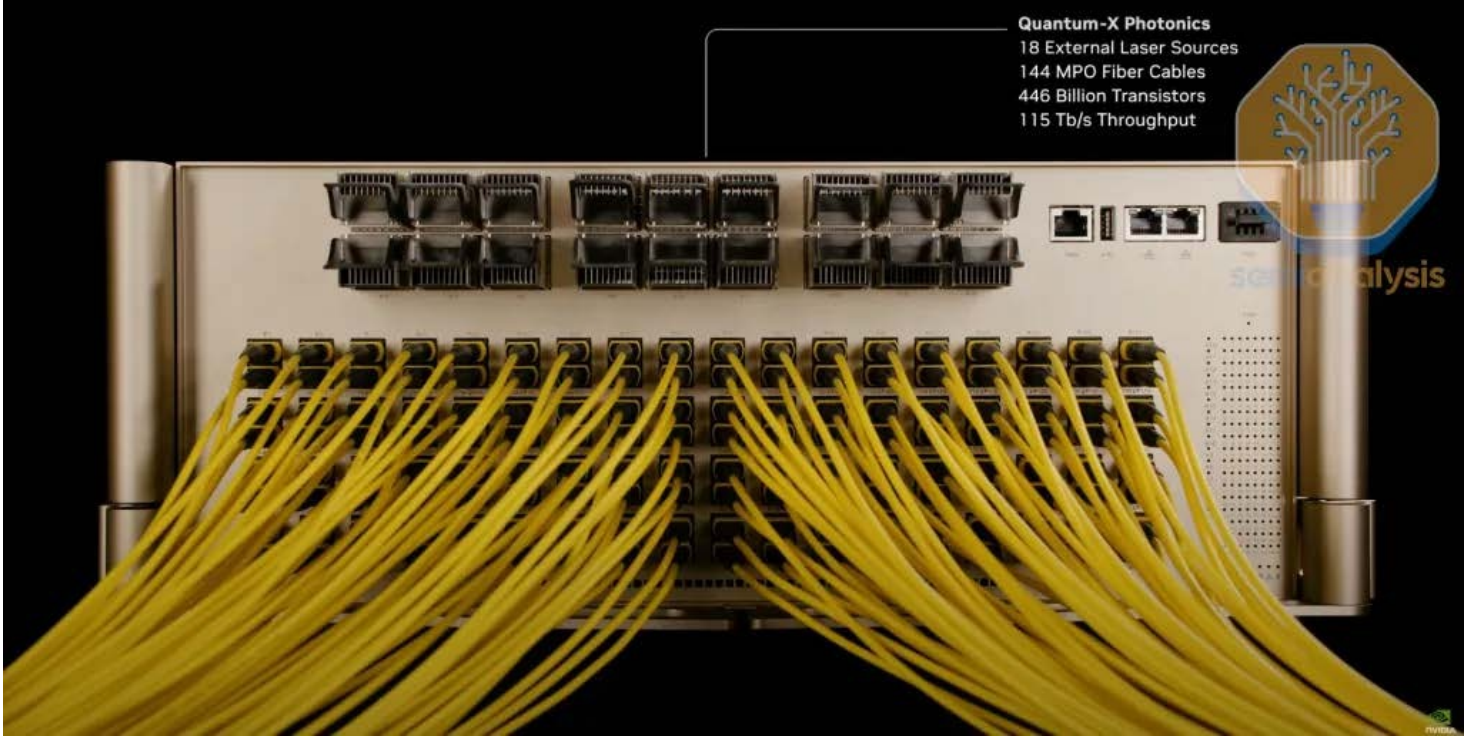
在GTC 2025大会上，英伟达首次推出基于CPO技术的横向扩展网络交换机。本次共发布三款CPO交换机，我们将逐一介绍，但首先呈现一张汇总所有关键规格的清晰表格：

Nvidia CPO Roadmap			
Switch Model	**Quantum 3450 CPO**	**Spectrum 6810 CPO**	**Spectrum 6800 CPO**
Launch Date	2H 2025	2H 2026	
Networking Standard	InfiniBand	Ethernet	
Switch ASIC	Quantum-3	Spectrum-6	
Throughout per Package	28.8 Tbps	102.4 Tbps	
Number of Switch Packages	4	1	4
Switch Aggregate Bandwidth	115.2 Tbps (not all-to-all)	102.4 Tbps	409.6 Tbps (not all-to-all)
SerDes speed (Gb/s uni-di)	200 Gbps	200 Gbps	
Optical Connectivity	DR Optics	DR Optics	
Physical MPO Ports	144	128	512
Bandwidth and Logical Port Configurations Available	144 Ports of 800G	512 Ports of 200G 256 Ports of 400G 128 Ports of 800G	512 Ports of 800G
Bandwidth per Optical Engine (OE)	1.6 Tbps	3.2 Tbps	
Number of OEs	72	32	128
External Light Sources (ELSs)	18	16	64
For Spectrum CPO there are 36 OEs on the package, but only 32 OEs are enabled			

来源：SemiAnalysis

Quantum-X Photonics

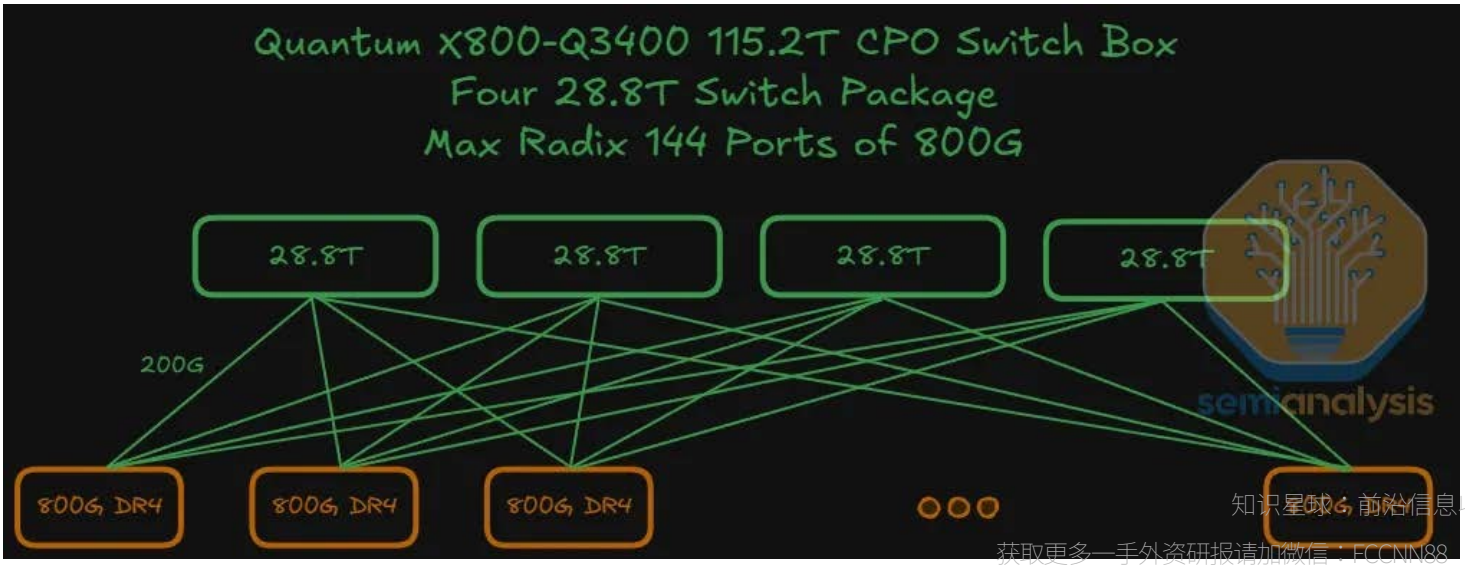
首款将于2025年下半年上市的CPO交换机是Quantum X800-Q3450。该设备配备144个物理MPO端口，可支持144个800G逻辑端口或72个1.6T逻辑端口，总带宽达115.2T。其如意大利面怪物般的复杂结构令笔者们胃中翻腾。



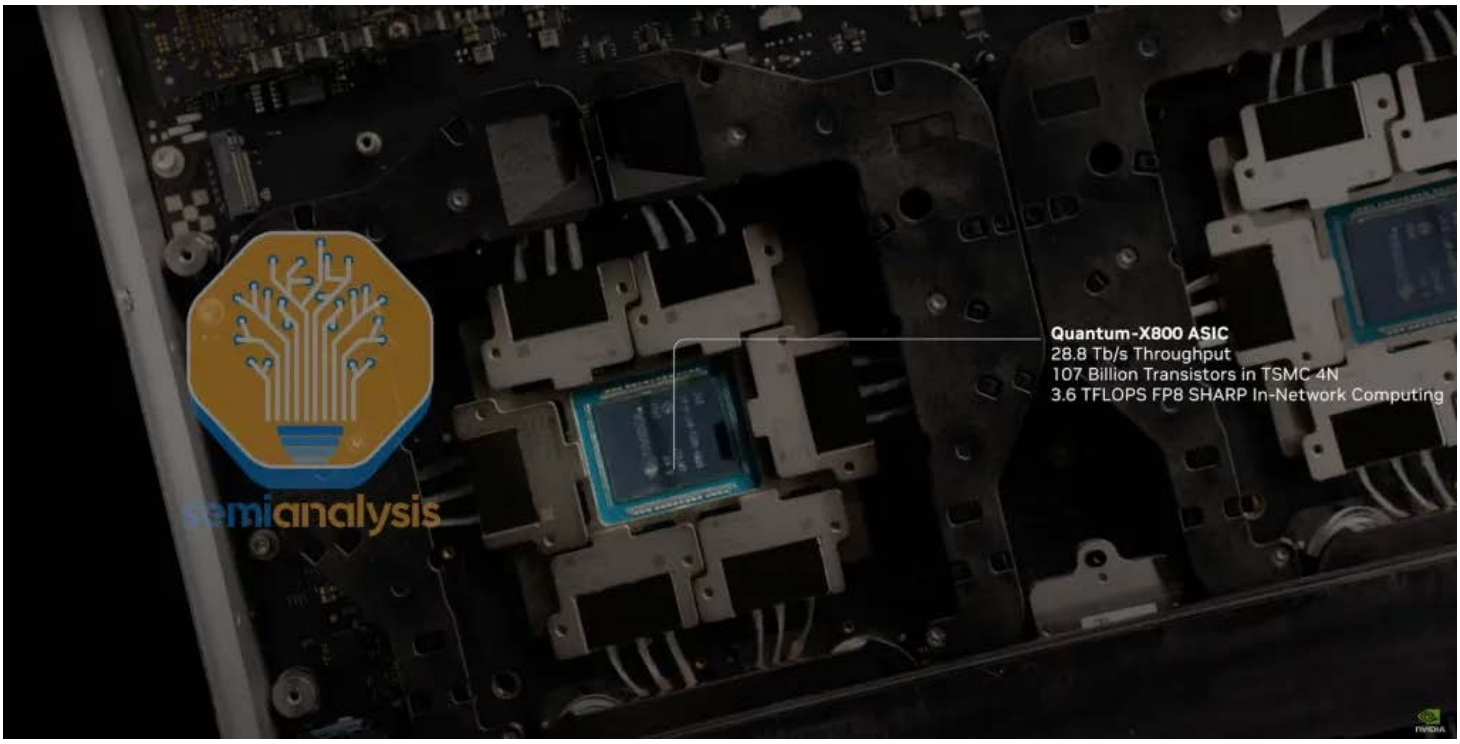
来源：英伟达

Quantum X800-Q3450通过采用四颗Quantum-X800 ASIC芯片实现高基数与高聚合带宽，每颗芯片带宽达28.8 Tbit/s，并采用多平面配置。 在此多平面配置中，每个物理端口均连接至四块交换ASIC芯片，通过四条不同的交换ASIC芯片将数据分散至全部四条200G通道，从而实现任意物理端口间的通信。

对于三层网络的最大集群规模而言，这种方案与理论上使用四倍数量的28.8T交换机箱（配备200G逻辑端口规格）所达到的效果相同——两者均支持最大746,496个GPU的集群规模。 区别在于：使用X800-Q3400交换机时，数据切换可在交换机内部高效完成；而采用独立28.8T交换机构建同等网络时，则需铺设更多光纤连接至更多目的地。

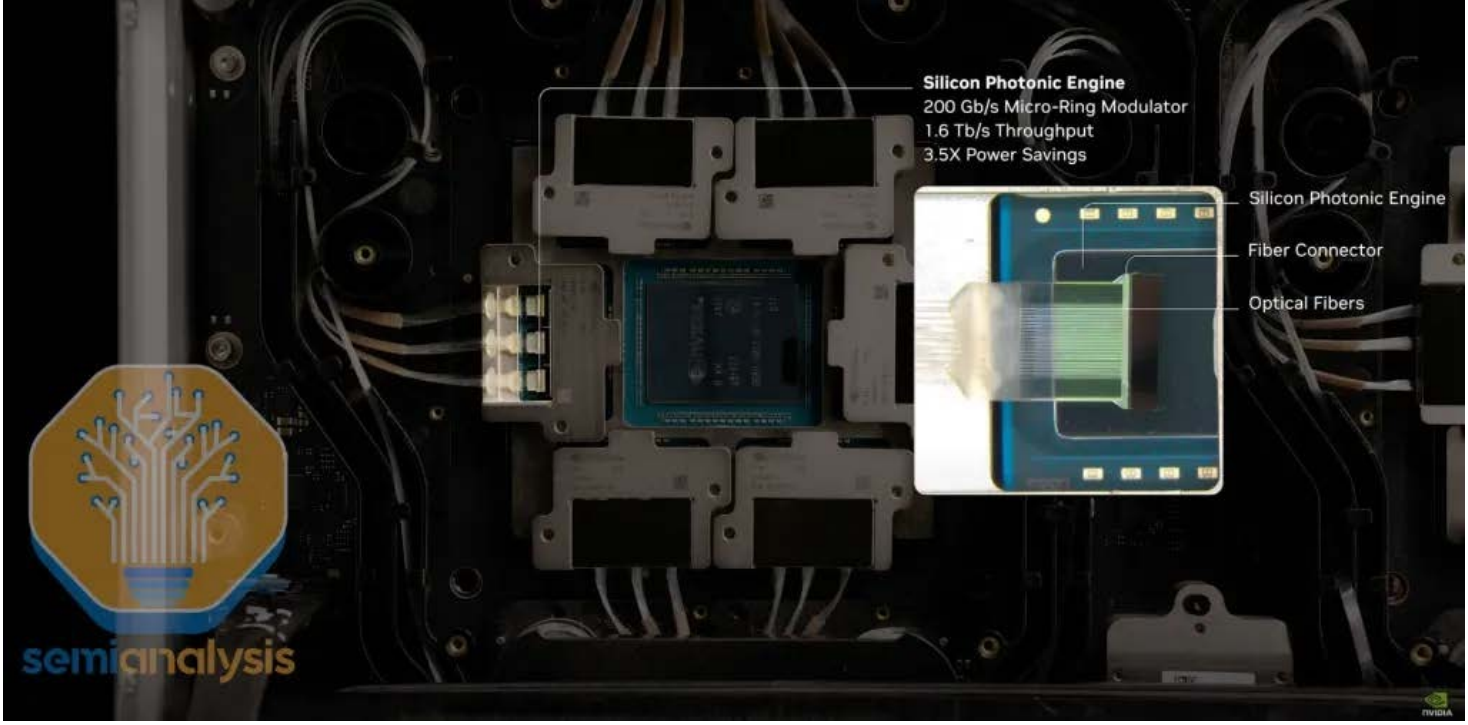


Quantum-X800-Q3450中的每个ASIC周围均配备六个可拆卸光子子组件，每个子组件内置三个光子引擎。每个光子引擎提供1.6 Tbit/s带宽，因此每个ASIC共搭载18个光子引擎，实现单ASIC总计28.8 Tbit/s的光子带宽。 需注意这些子组件可拆卸，因此严格意义上这属于"非完全可拆卸式"（NPO）而非"完全可拆卸式"（CPO）。虽然可拆卸光学引擎会带来些许额外信号损耗，但我们认为实际性能影响有限。



来源：英伟达

每个引擎配备8个电光通道，电侧由200G PAM4串解串器驱动，光侧则采用8个微环调制器（MRM），通过PAM4调制实现单调制器200G传输速率。该设计方案是本次发布会的重大突破之一：证明英伟达与台积电已实现200G MRM量产出货。 该性能已与当前最快的MZM调制器持平，并打破了MRM仅限于NRZ调制的行业认知。英伟达达成此里程碑堪称卓越的工程成就。

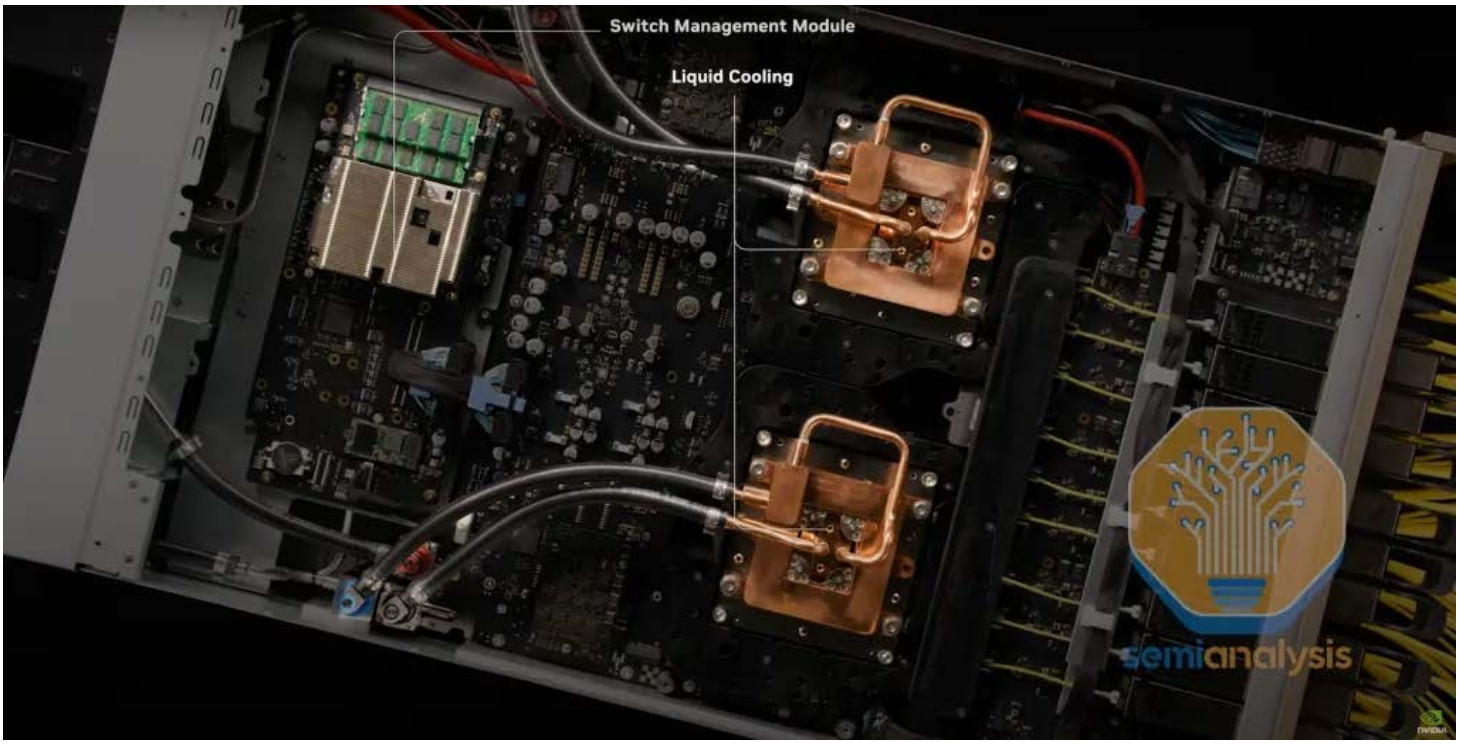


来源：英伟达

每个光引擎集成了基于成熟N65工艺节点的光子集成电路（PIC）和基于先进N6节点的电子集成电路（EIC）。PIC采用较旧工艺节点，因为其包含调制器、波导和探测器等光学部件——这些器件无法从缩放中获益，且通常在较大几何尺寸下表现更佳。 相较之下，EIC包含驱动器、电致导电放大器（TIAs）及控制逻辑，这些元件能显著受益于先进工艺节点带来的更高晶体管密度和更优功耗效率。这两枚芯片通过台积电的COUPE平台进行混合键合，实现了光电领域间超短距离、高带宽的互联。

来源：英伟达

两块铜冷板置于Quantum-X800-Q3450的ASIC芯片上方，作为闭环液冷系统的一部分，可高效散发每个交换ASIC芯片产生的热量。 连接冷板的黑色管路循环冷却液，有助于维持热稳定性。该冷却系统不仅对ASIC芯片至关重要，对相邻的温度敏感共封装光学元件同样不可或缺。



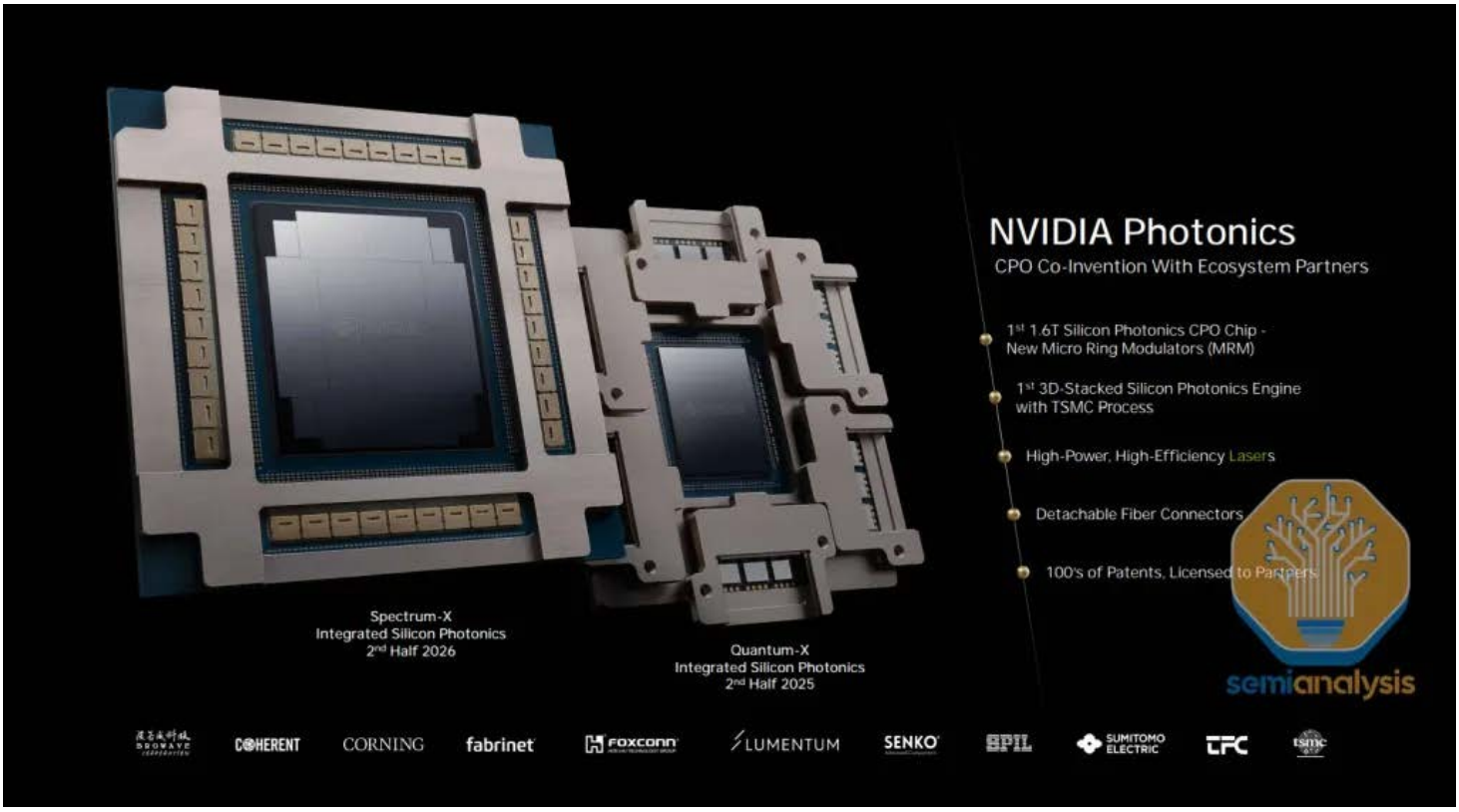
来源：英伟达

Spectrum-X Photonics

Spectrum-X Photonics计划于2026年下半年发布，将推出两款独立的交换机配置产品：一款是以太网Spectrum-X版本的X800-Q3450 CPO交换机，提供102.4T聚合带宽； Spectrum 6810提供102.4T聚合带宽，其更大规格的Spectrum 6800则通过四枚独立Spectrum-6多芯片模块（MCM）实现409.6T聚合带宽。

Quantum X800-Q3450 CPO交换机采用四组独立开关封装，通过多平面架构连接至物理层。每组开关封装均为单片芯片，内含28.8T交换ASIC芯片及所需的串行器/解串器（SerDes）等电子元件。相比之下，Spectrum-X Photonics的交换芯片采用

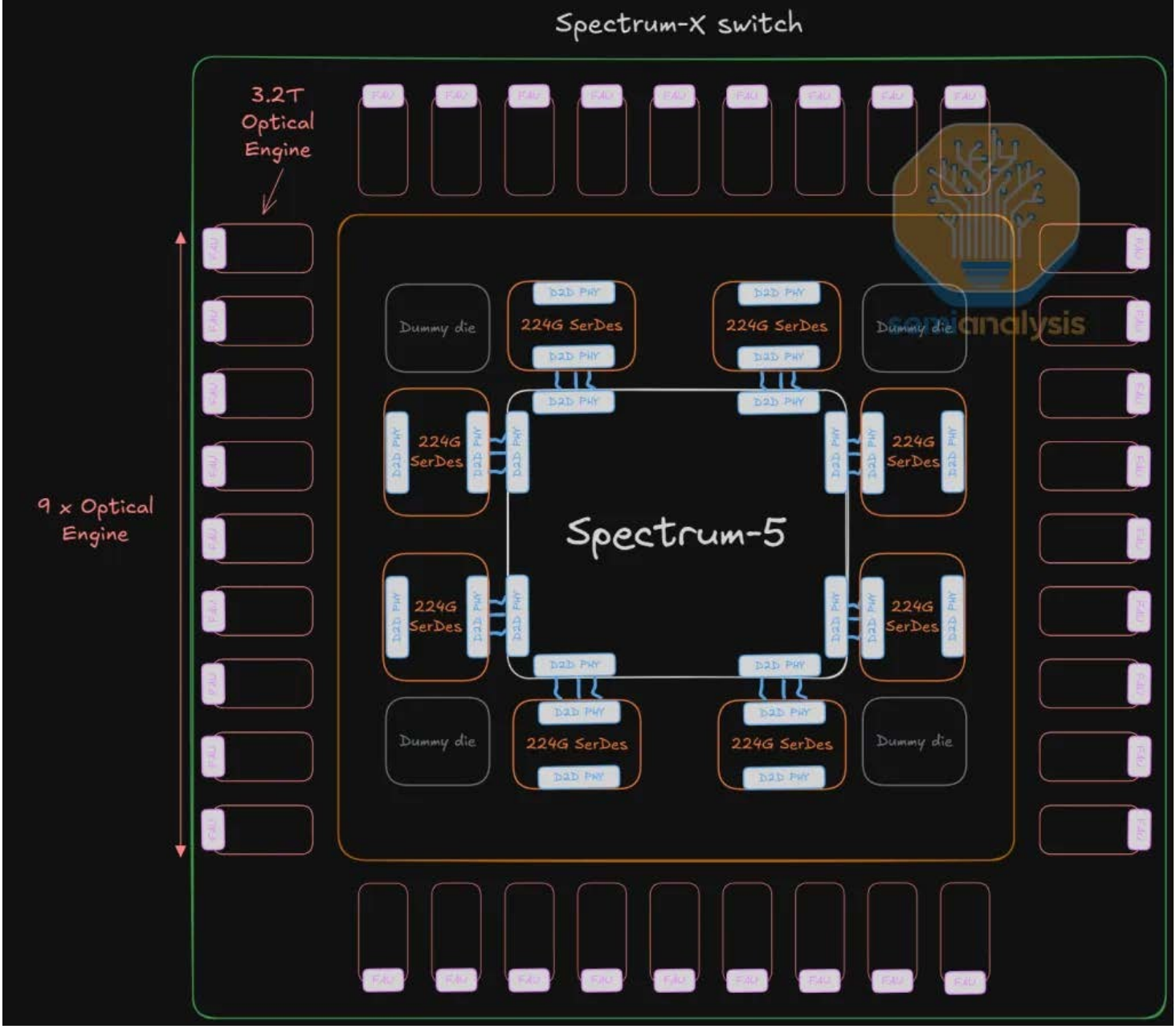
多芯片模块（MCM），其光罩尺寸更大：中央为102.4T交换ASIC，外围环绕八个224G串解码器I/O芯片片——每侧各两个。



来源：英伟达

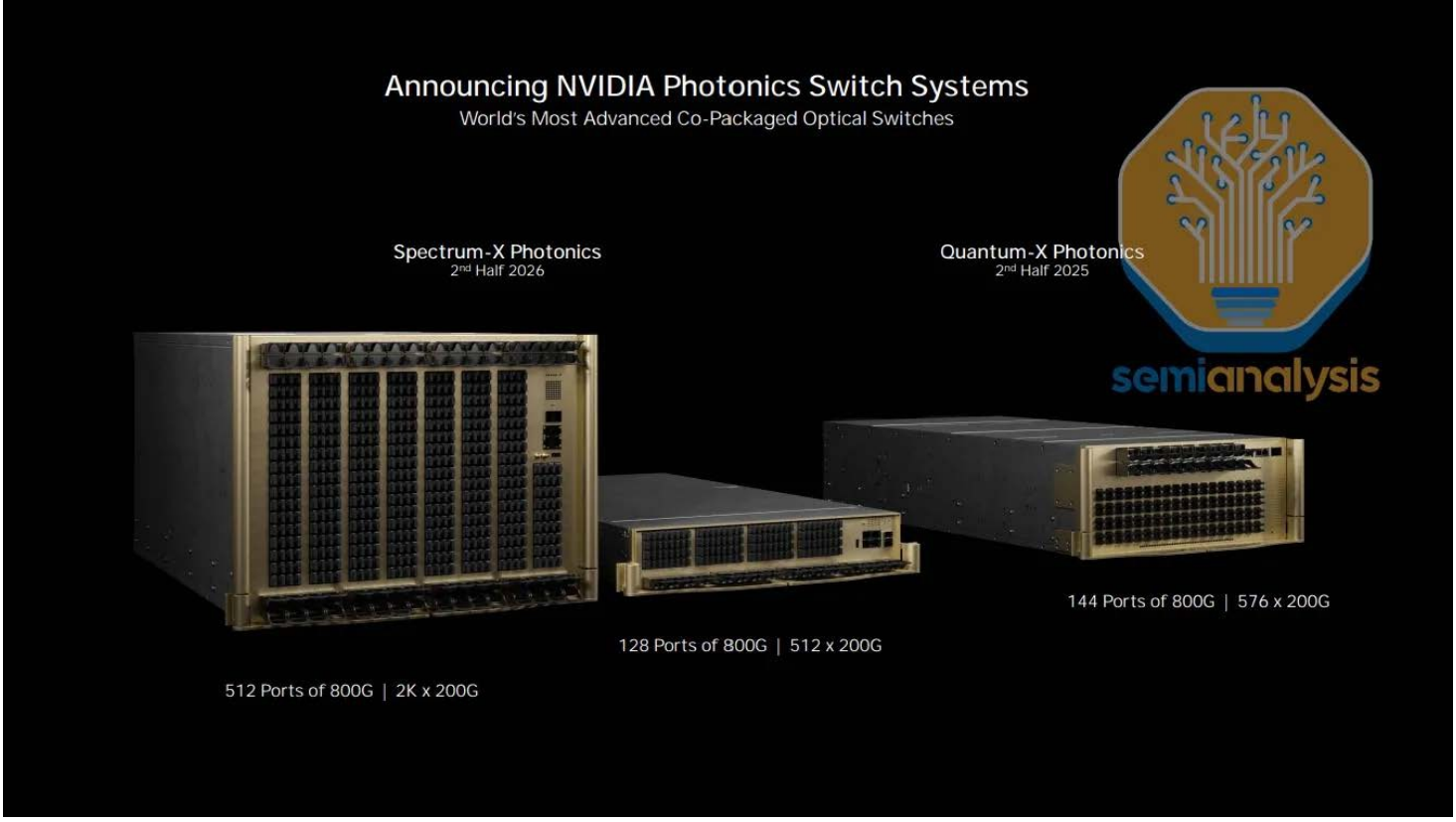
每个Spectrum-X光子学多芯片模块交换机封装将集成36个光引擎于单个102.4T交换机封装中。该封装采用英伟达第二代光引擎，带宽达3.2T，每个光引擎包含16条200G光通道。 需注意仅有32个光引擎处于工作状态，其余4个作为冗余设计以应对光引擎故障。这是因为光引擎通过焊接固定在基板上，难以实现快速更换。

每个I/O芯片提供12.8T的总单向带宽，包含64条SerDes通道，并分别与4个光电接口（OE）相连。正是这种设计使Spectrum-X能够提供远超Quantum-X Photonics的总带宽，同时为SerDes通道提供了更宽的岸线和更大的面积。



来源：SemiAnalysis

Spectrum-X 6810交换盒采用单个上述交换芯片封装，可提供102.4T聚合带宽。更大的Spectrum-X 6800交换盒型号采用高密度机箱设计，通过集成四个上述Spectrum-X交换芯片封装实现409.6T聚合带宽，这些芯片封装还以多平面配置连接至外部物理端口。



来源：英伟达

与四颗ASIC 115.2T Quantum X800-Q3450类似，Spectrum-X 6800采用内部分线设计，将每个端口物理连接至全部四颗ASIC芯片。

来源：SemiAnalysis

博通**CPO**交换机产品组合

Broadcom CPO Roadmap			
Switch Model	**Humbolt**	**Bailly**	**Davisson**
Launch Date	2022	2024	2026
Networking Standard	Ethernet	Ethernet	Ethernet
Scale-out or Scale-up	Scale-out	Scale-out	Scale-out
Switch ASIC	Tomahawk 4	Tomahawk 5	Tomahawk 6
Switch ASICs per Package	1	1	1
Throughput per Package	25.6 Tbps	51.2 Tbps	102.4 Tbps
Number of switch packages	1	1	1
Switch Aggregate Bandwidth	25.6 Tbps (half-electrical)	51.2 Tbps	102.4 Tbps
SerDes speed (Gb/s uni-di)	100G	100G	200G
Optical Connectivity	DR Optics	FR4 Optics	DR4 Optics
Bandwidth and Logical Port Configurations Available	256 Ports of 100G 128 Ports of 200G 64 Ports of 400G	128 Ports of 400G 64 Ports of 800G	128 Ports of 800G 64 Ports of 1.6T
Bandwidth per Optical Engine (OE)	3.2 Tbps	6.4 Tbps	6.4 Tbps
Number of OEs	4	8	16

来源：SemiAnalysis

博通是最早推出真正支持CPO的系统厂商之一，因此被视为CPO领域的领导者。其第一代CPO设备"洪堡"主要作为概念验证平台，代号"TH4-洪堡"，是一款25.6Tbit/s以太网交换机，其总容量均分为传统电连接和CPO两部分。其中12.8Tbit/s由四个3.2Tbit/s光引擎处理，每个引擎提供32条100Gbit/s通道。这种铜缆与光纤的混合设计具有显著的应用场景。在一种方案中，机架顶部（ToR）交换机依赖于
用于短距离铜缆连接至邻近服务器的电气接口，而
其光纤端口则向上连接至下一层交换设备。在另一种场景中，汇聚层采用电接口实现机架内各交换机的互联，同时通过光纤链路延伸至上下层的交换设备。

TH4-Humboldt: First Generation System

Product Features:

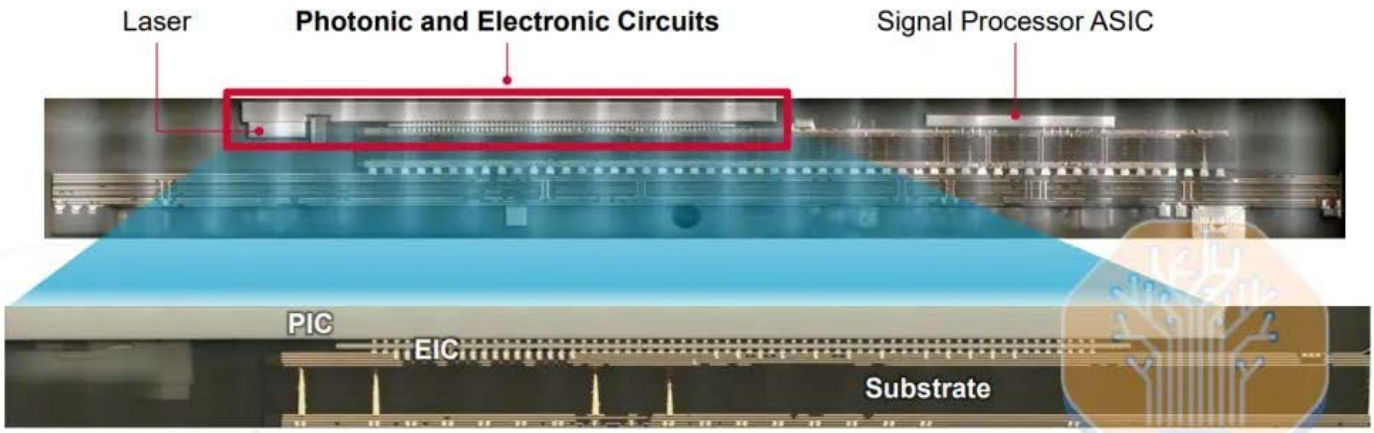
- 25.6T Ethernet Switch
- Half CPO, Half Electrical connectivity
- Four 3.2T optical engines (32x100Gbps DR connectivity)
- Optical engine is a PIC bonded to a SiGe EIC
- Each optical engine has ~ 250 optical components

SiGe dissipates additional 3 pJ/bit power consumption compared to CMOS solutions

来源：博通

在此设计中，博通采用硅锗（SiGe）EIC芯片，但在下一代产品（即Bailly系列）中转为CMOS工艺。

TH4-Humboldt: SiPh PIC + SiGe EIC + TSV



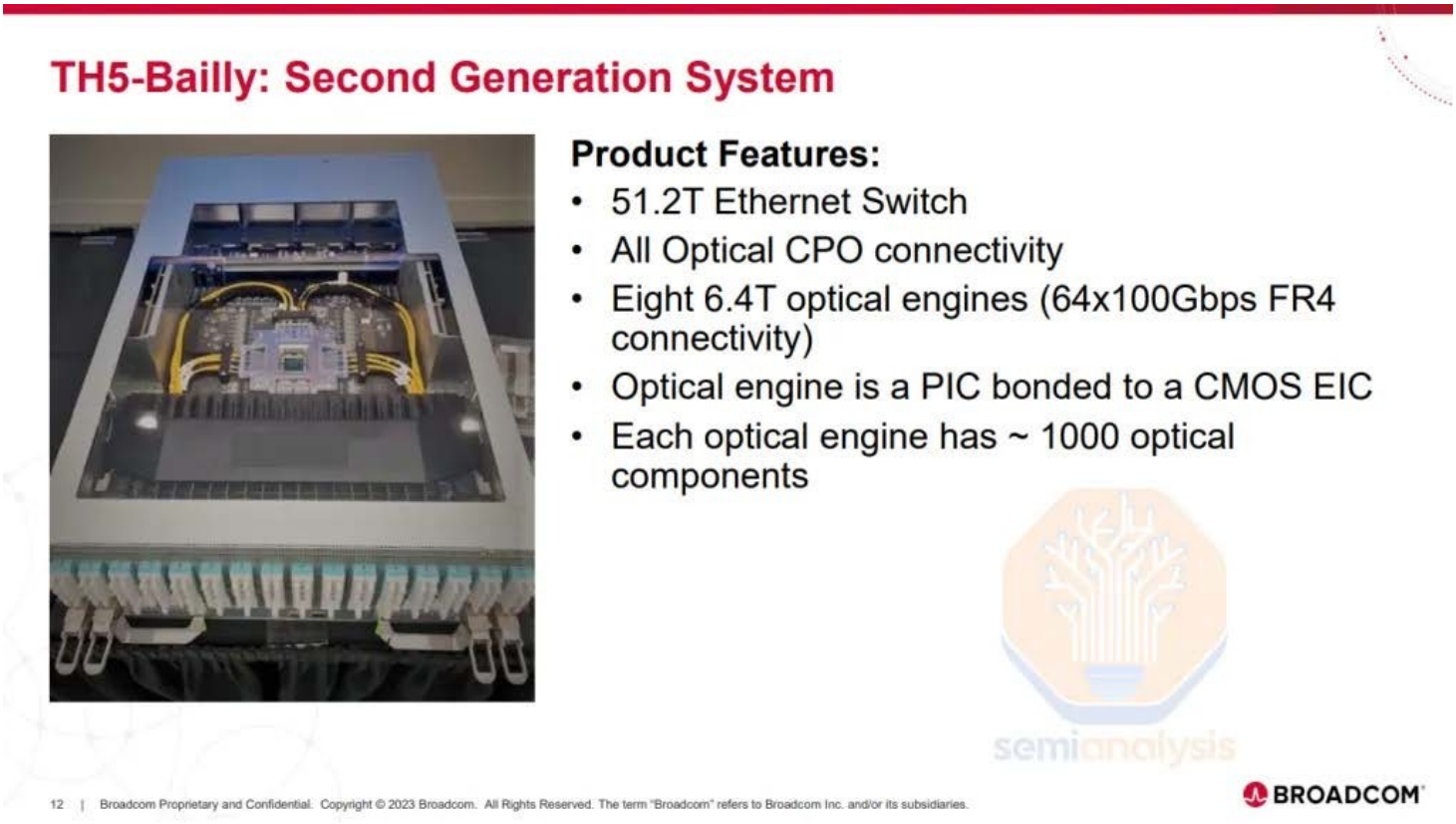
来源：博通

博通第二代CPO设备Bailly是一款51.2 Tbit/s以太网交换机，

——不同于其半光纤前代产品——完全依赖光纤I/O。它由八个

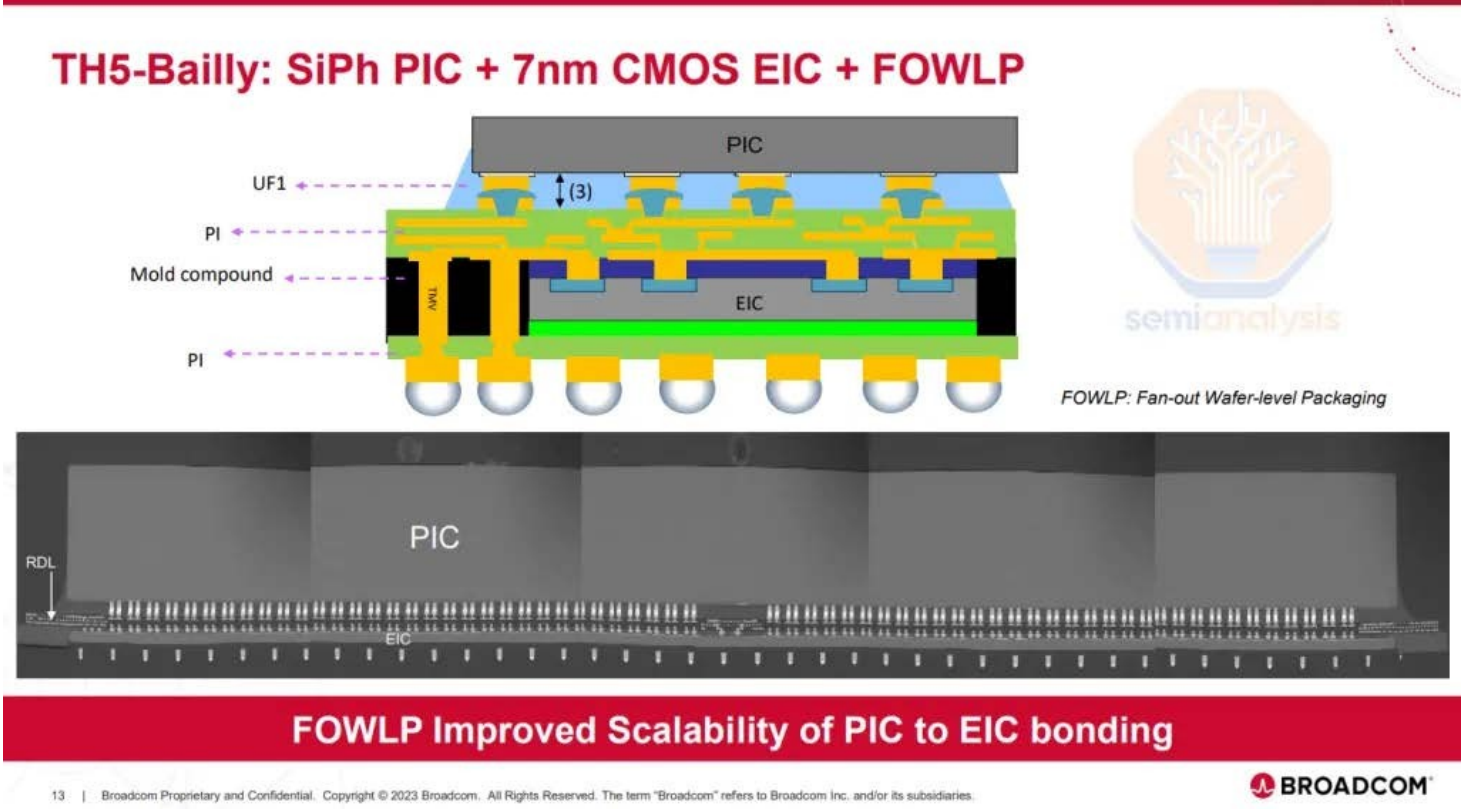
6.4Tbit/s光引擎，每个提供64条100 Gbit/s通道。另知晓

该变化在于，其不再采用硅锗（SiGe）光电集成电路（EIC），而是改用7纳米CMOS光电集成电路。转向CMOS光电集成电路的设计使更复杂的集成方案成为可能，新增的控制逻辑进而实现了更高的通道数量——从原先的光学引擎32通道扩展至新型光学引擎的64通道。



来源：博通

从第一代到第二代的另一重大转变是工艺从TSV过渡到扇出型晶圆级封装（FOWLP）。该设计中，EIC采用模塑通孔（TMV）将信号路由至PIC，同时通过铜柱
凸点连接基板。采用FOWLP的主要原因在于该技术已在移动手机市场得到验证，并获得OSAT厂商广泛支持，从而具备更强的可扩展性。ASE/SPIL是该FOWLP工艺的OSAT合作伙伴。

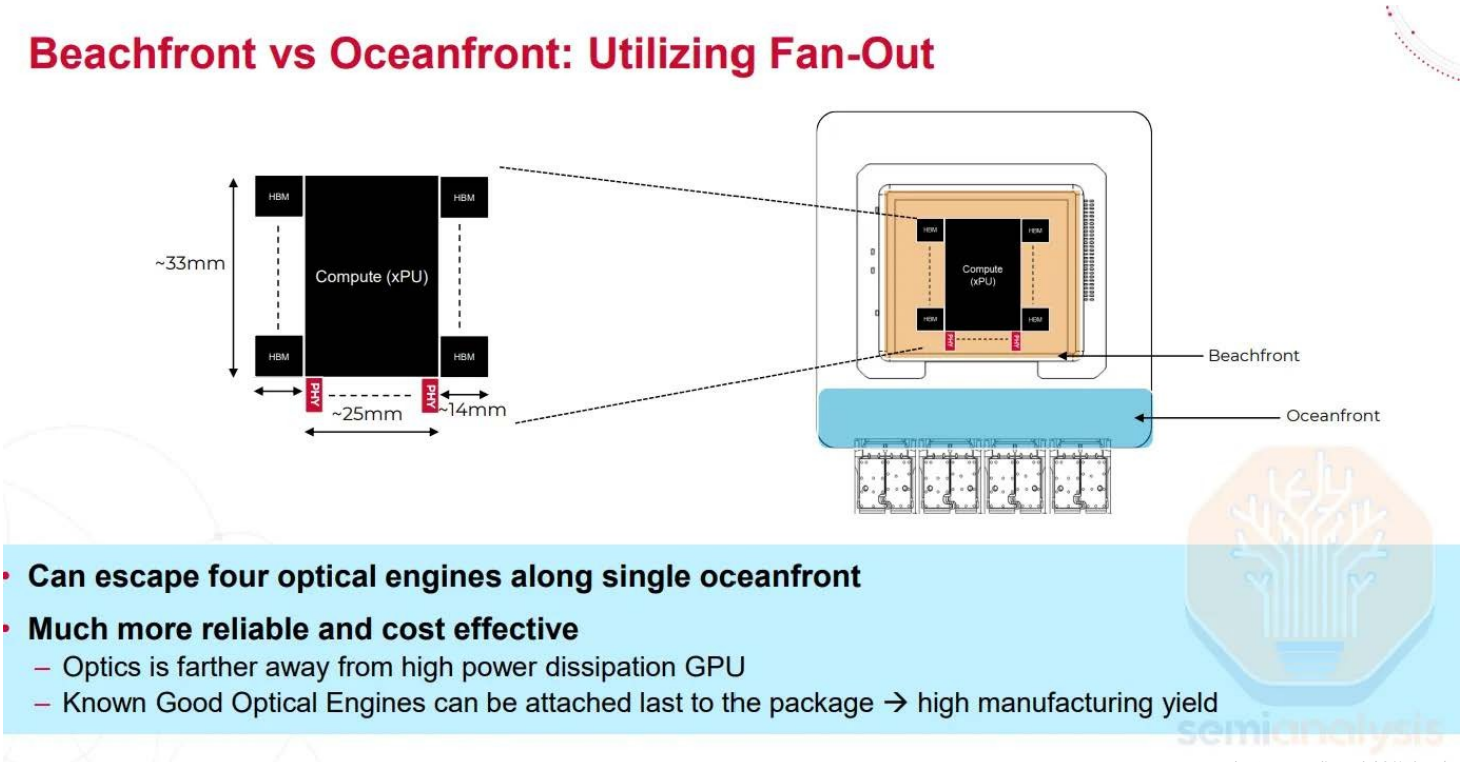


来源：博通

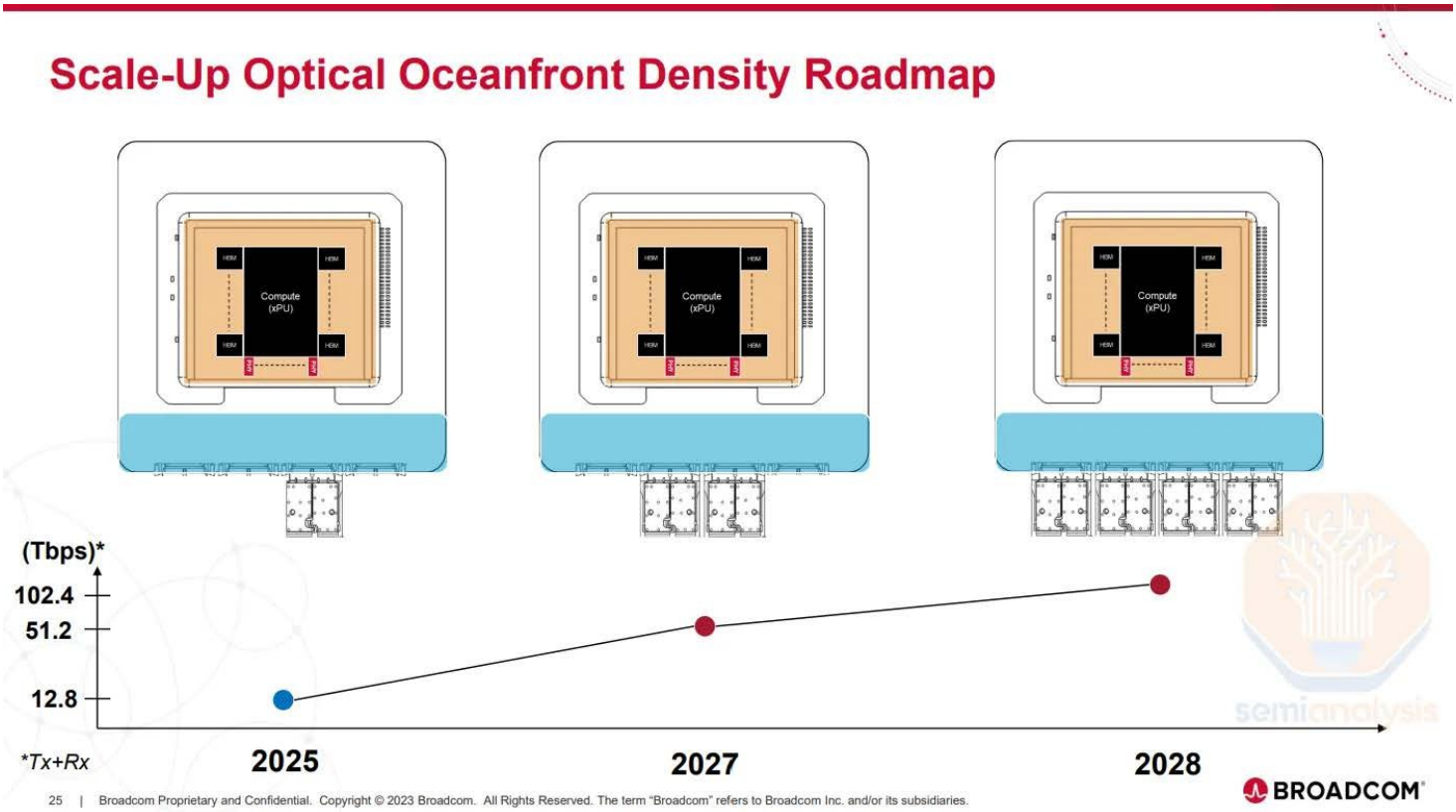
博通在Hot Chips 2024大会上展示了一项实验性设计，该设计将

6.4 Tbit/s 光学引擎集成于单逻辑芯片封装中，该封装包含两个HBM堆栈和一个SerDes模块。他们提出采用扇出布局方案，将HBM置于基板东西两侧，从而在同一封装内为两个光学引擎预留空间。 从CoWoS-S向CoWoS-L的演进，意味着基板边长将突破100毫米。由此可容纳多达四个光引擎

引擎，实现51.2 Tbit/s带宽。



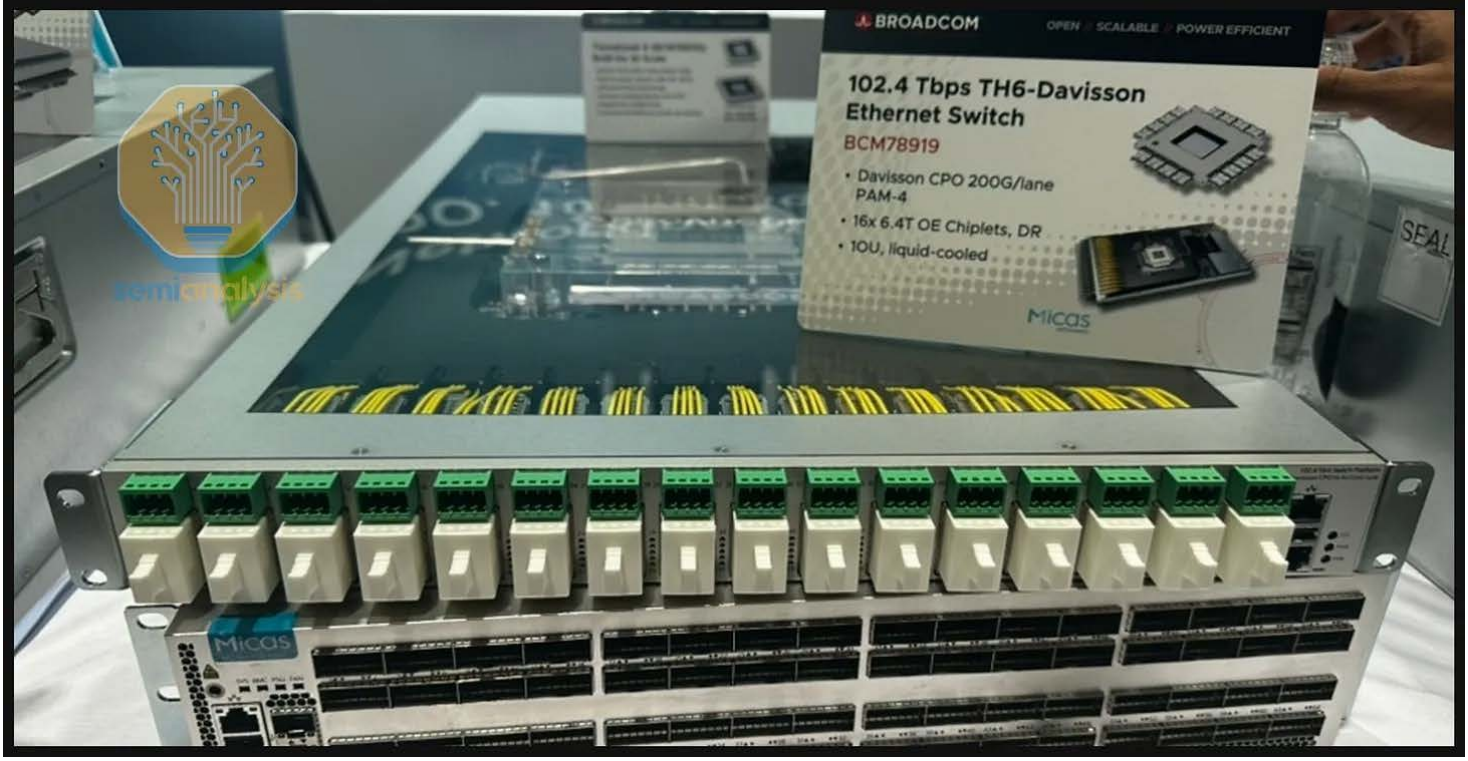
来源：博通



来源：博通

今年，博通将推出基于Tomahawk 6的Davisson CPO交换机，该产品集成了十六个6.4T光电器件（OEs）。该交换机ASIC采用台积电N3工艺节点制造，单封装提供102.4 Tbit/s带宽。博通通过Micas和Celestica等合同制造商（CMs）进行盒装组装。

此外，据报道日本NTT公司正采购博通的TH6裸片，并采用非博通来源的专有光电器件和光学解决方案自主构建CPO系统。此举拓展了基于TH6的CPO系统潜在商机，并推动形成更开放的供应商生态体系。

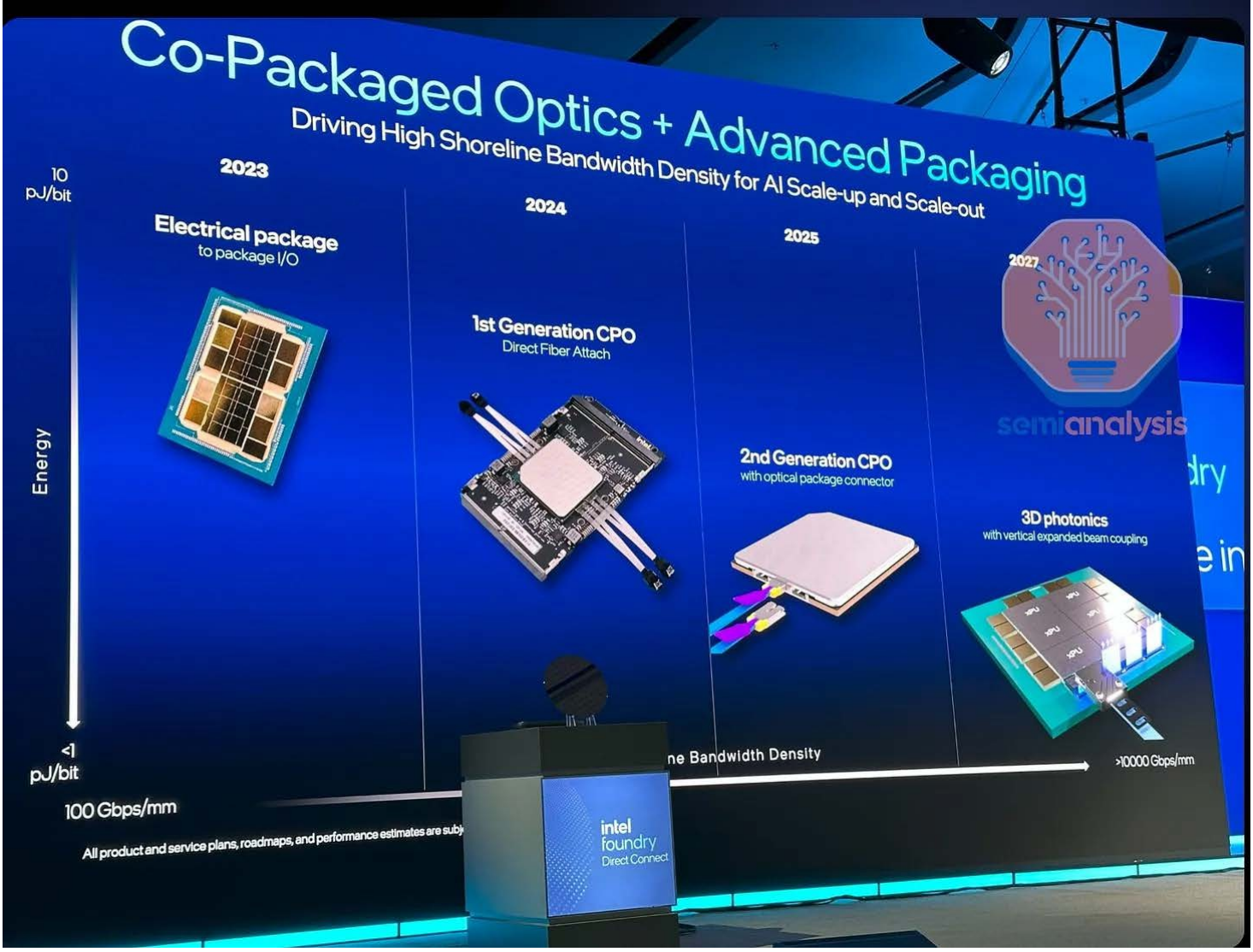


来源：SemiAnalysis

随着我们在扩展型网络架构中看到CPO的更大价值，我们认为博通交付的首个量产CPO系统将应用于其客户的人工智能专用集成电路（AI ASIC）。博通在CPO领域的经验使其成为吸引人的设计合作伙伴，尤其对于那些在中长期ASIC路线图中规划采用CPO的客户而言。 我们了解到，这正是OpenAI选择博通的关键因素。值得注意的是，作为博通最大的ASIC客户，谷歌却是最不愿在其数据中心部署CPO的超大规模企业。谷歌的基础设施理念更注重可靠性而非绝对性能，而CPO的可靠性问题对其而言是不可逾越的障碍。我们预计谷歌短期内不会采用CPO技术。

博通未来几代CPO终端设备也将转向台积电的COUPE平台——这明确表明COUPE平台提供的特性为带宽扩展开辟了路径。这不仅意味着光电器件封装方式的转变，更标志着博通此前采用的边缘耦合器与磁致伸缩光栅（MZM）方案的终结。从实现角度看，这两种方案虽更简便，但如前所述其可扩展性较弱。 COUPE平台倾向于采用光栅耦合器和磁阻耦合器，这与现有方案形成根本性转变。尽管博通拥有最丰富的CPO技术积累，但技术路线的调整意味着其部分技术领域需重新起步。关键在于台积电能提供多少支持来简化博通的设计流程。

英特尔首席产品官路线图



英特尔

英特尔在今年的英特尔代工直通大会上公布了其CPO路线图，并概述了四阶段发展规划

2023年：英特尔提出先进电学封装间I/O互联概念，作为光集成技术的先导。该里程碑聚焦于实现芯片封装间的高带宽短距离电学直连（绕过传统PCB走线），以支持多芯片系统。通过建立封装级I/O基础设施为光子集成铺路，该架构后续可扩展光通道功能。

2024年：英特尔展示了其首款支持直接光纤连接的CPO解决方案。该方案中，光引擎芯片直接与光纤耦合，无需任何外部连接器，从而简化了连接链路。 在2024年OFC光通信大会上，英特尔展示了与概念性至强CPU协同封装的4 Tbit/s（双向）光计算互连（OCI）芯片，通过单模光纤链路实现零误码传输，提供64条通道（每通道32 Gbit/s）。该光接口在

~5皮焦/比

2025年：英特尔第二代CPO解决方案采用可拆卸光学封装连接器替代永久性光纤尾纤。英特尔工程师研发出可插入封装侧面的玻璃光学桥，内置嵌入式3D波导和机械对准功能，用于将封装内光子器件与标准光纤连接器对接。这种光学封装连接器设计实现了模块化组装，标志着向更易连接、更便于维护的封装形式转型。

2027年：英特尔计划在三维集成光子学领域实现突破——通过垂直扩展光束耦合技术实现光子元件的垂直堆叠。 在设想的第三代设计中，光输入输出将通过短距离自由空间或玻璃内光路在芯片层间（例如光子中介层与逻辑芯片之间）垂直传输。英特尔计划通过封装内的垂直光耦合技术，在十年后期进一步消除电气瓶颈，实现超高带宽小芯片互联架构。

联发科CPO计划

作为定制ASIC设计公司，联发科正致力于将CPO能力整合至其设计平台，旨在提供能与定制加速器无缝协作的PIC/EIC设计方案。 该公司认为在200G/通道的代际中，近封装铜（NPC）技术搭配>900微米光纤间距可成为有效方案；当数据速率提升至200-300G区间时，采用>400微米高密度间距的CPC技术可能更具优势。 然而，当速率达到每通道400G及以上时，转向CPO架构——采用约130微米的更密集光纤间距和更紧凑的互连IP——将成为必然选择。

聚焦CPO的企业

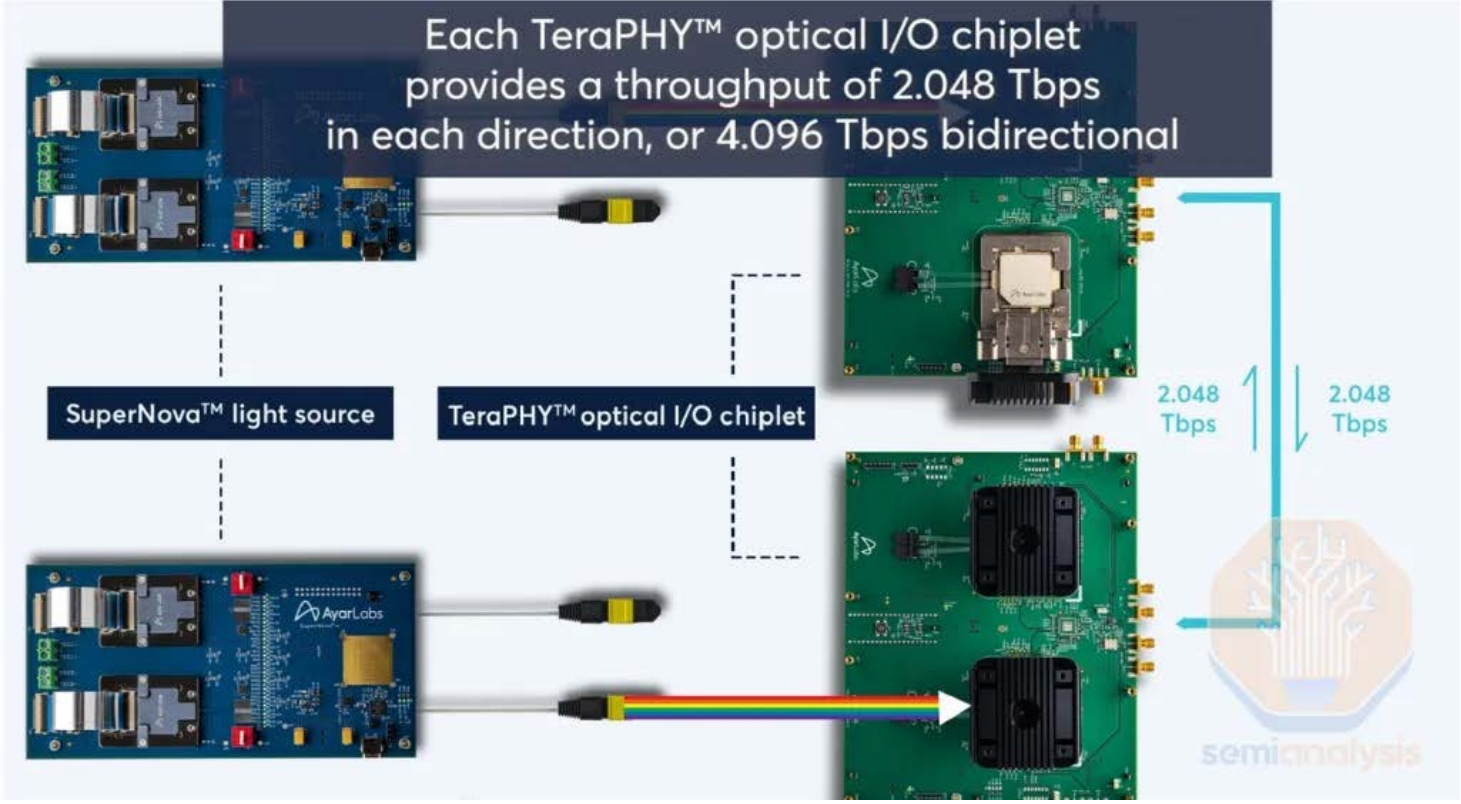
英伟达、博通和美满电子正各自开辟道路，打造专属解决方案，而多家专注于光子芯片（CPO）的企业则在探索另一套技术路径。这些公司面临的核心问题在于如何与主流交换芯片及GPU/ASIC供应商竞争——尤其考虑到多数现有厂商已宣布或展示了专属方案。AMD仍是例外：尽管已知其内部正在开发光子IP，但尚未展示任何产品。

对于Ayar Labs、Lightmatter、Celestial AI、Nubis和Ranovus等原始设备芯片供应商而言，挑战在于超越现有市场参与者，并提供足够具有吸引力的解决方案以实现集成。Ayar Labs、Celestial AI和Ranovus提供完整的"端到端"系统，这意味着客户必须采用其完整的端到端解决方案。 相比之下，Nubis专注于更开放、基于标准的解决方案，旨在简化实施流程并降低采用门槛。 此外，某些企业的产品路线图中还包含更具颠覆性的方案——例如Lightmatter的光学中介层和Celestial AI的光子桥技术。这些方案需要对封装和主机硅片设计进行根本性重构才能释放全部潜力，但同时也伴随着成本上升和重大不确定性，尤其在与CMOS硅片无缝集成及大规模量产方面存在挑战。

让我们开始逐一了解这些公司的架构和市场推广计划。

Ayar Labs

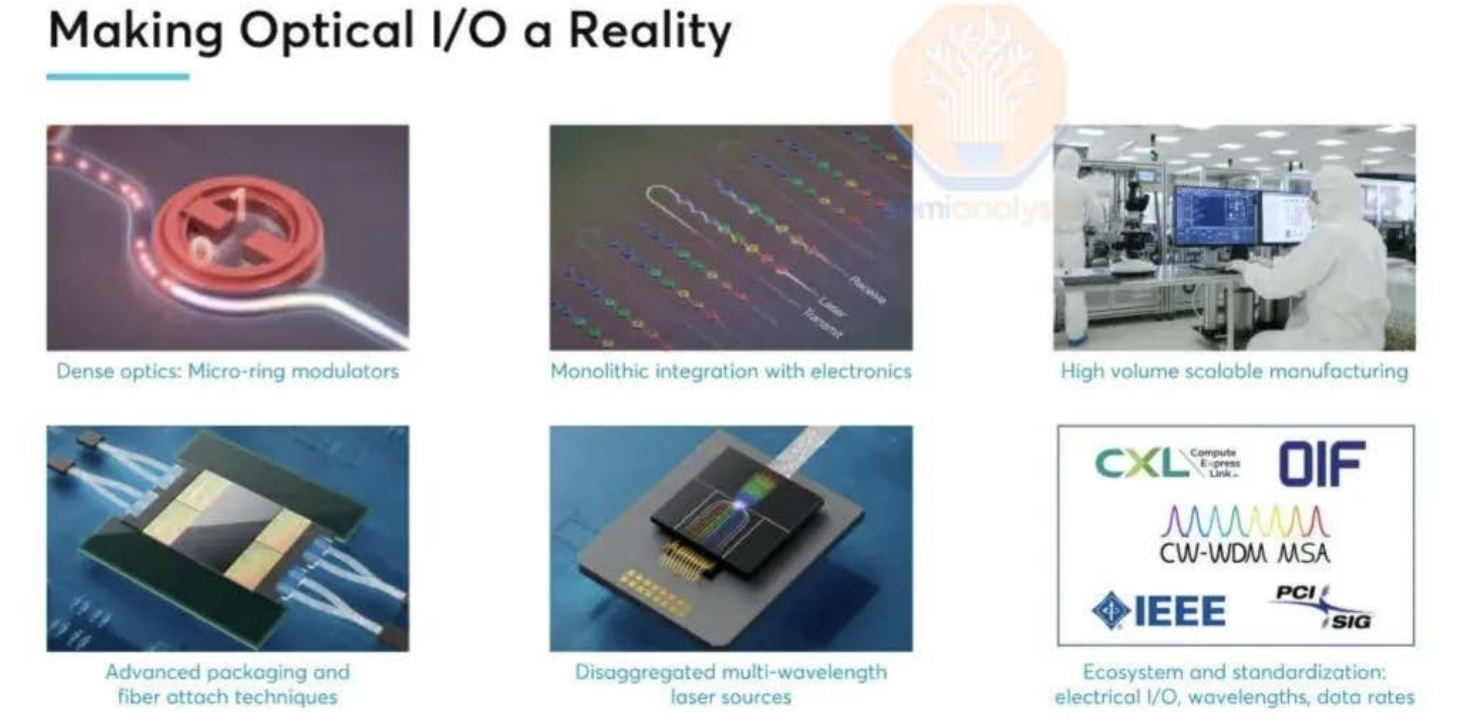
Ayar Labs的产品是其TeraPHY光引擎芯片，可封装为XPU、交换机ASIC或内存芯片。第一代TeraPHY在仅消耗10W功耗的情况下，可提供2Tbit/s的单向带宽。 第二代TeraPHY提供4Tbit/s单向带宽，作为全球首款UCIe光重定时芯片片，可在芯片片内部完成电光转换，实现主机信号的光学传输。UCIe接口的标准化特性使其能轻松集成至主机芯片，这一特性将显著提升产品对客户的吸引力。



Ayar Labs

Ayar Labs采用GlobalFoundries的45纳米工艺制造了前两代TeraPHY芯片，作为集成电子元件与硅光子学的单片解决方案

而第三代TeraPHY则采用台积电COUPE工艺。这种将环形调制器、波导和控制电路紧密集成的方式有效降低了电损耗。然而前两代产品采用的成熟单片节点限制了EIC性能，这也是早期TeraPHY采用低调制速率的原因。



来源：Ayar Labs

在代号"Eagle"的4Tbit/s单向第二代产品中，TeraPHY集成了八个512Gbit/s I/O端口，每个端口采用32Gbit/s NRZ×16波长架构，由MRM调制。外部激光源"SuperNova"由瑞典Sivers公司提供。 该激光器通过DWDM技术将16个波长（"颜色"）聚合至单根光纤。 每个端口通过一对单模光纤分别实现发送（Tx）和接收（Rx）功能，这意味着每个4T芯片需连接总计24根光纤——其中16根用于收发，8根用于激光输入。该公司在封装工艺中采用边缘耦合（EC）技术，同时亦支持光栅耦合（GC）方案。

在提升单芯片带宽方面，该公司指出随着连接器技术的进步，光纤密度（目前为每芯片24根）在未来几年内有望实现翻倍。此外，通过提高单波长数据速率，端口/光纤带宽同样可实现倍增，这将推动近期路线图中整体带宽实现4倍扩张。

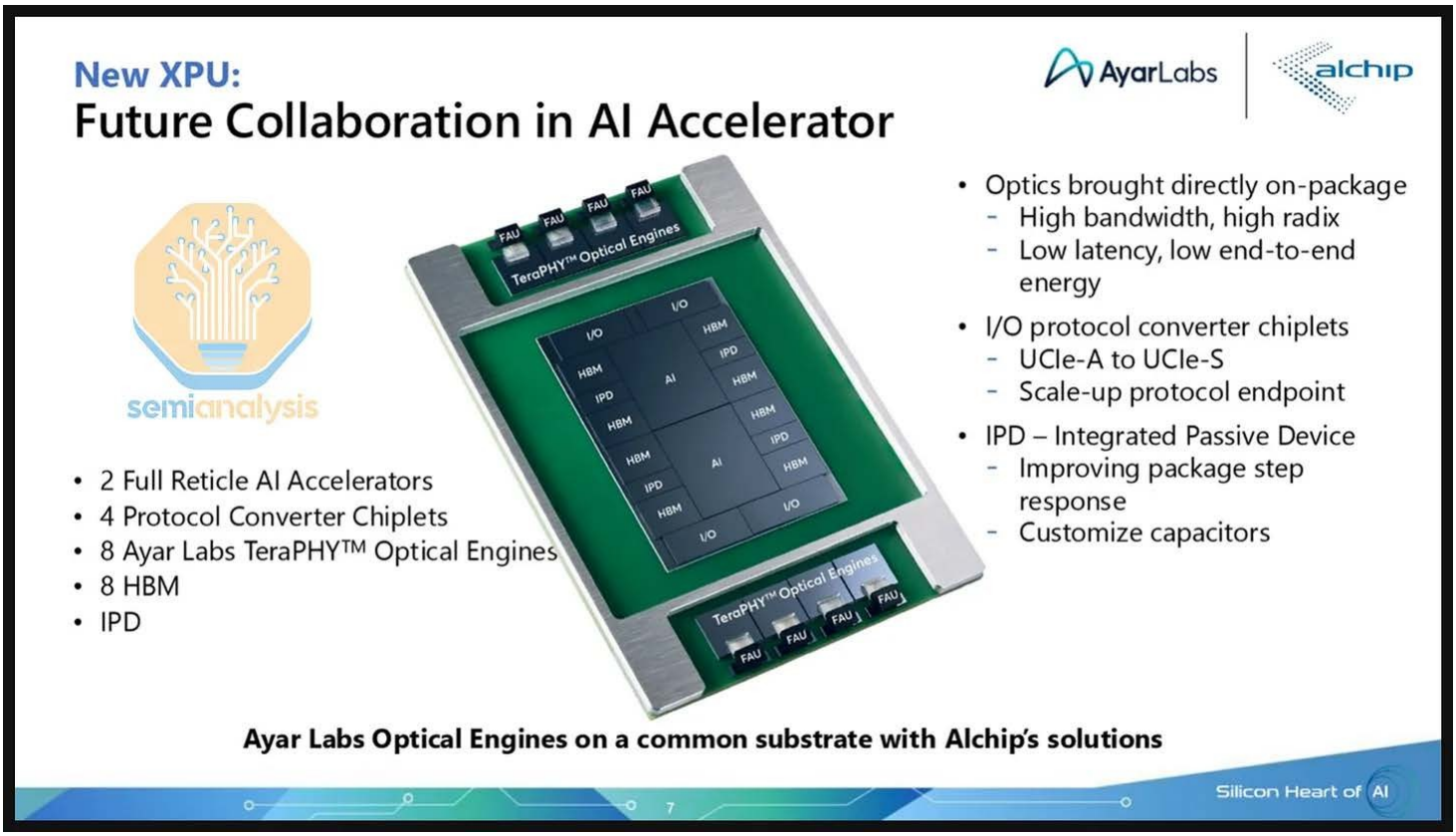
SuperNova激光器符合MSA（多源协议）规范，可与其他CW-WDM标准光学部件互操作。

Figure 3. Ayar Labs' 16-wavelength SuperNova™ light source

来源：Ayar Labs

Ayar第三代TeraPHY芯片转向采用台积电COUPE工艺，单个光引擎单向带宽达13.5 Tbit/s，较第二代提升逾3倍。通过8个光引擎协同工作，可为下文所述Ayar Labs与Alchip合作的XPU解决方案提供约108 Tbit/s的整体封装扩展带宽。 该约13.5+Tbit/s的性能通过采用PAM4调制技术实现，每个波长可提供约200Gbit/s带宽。

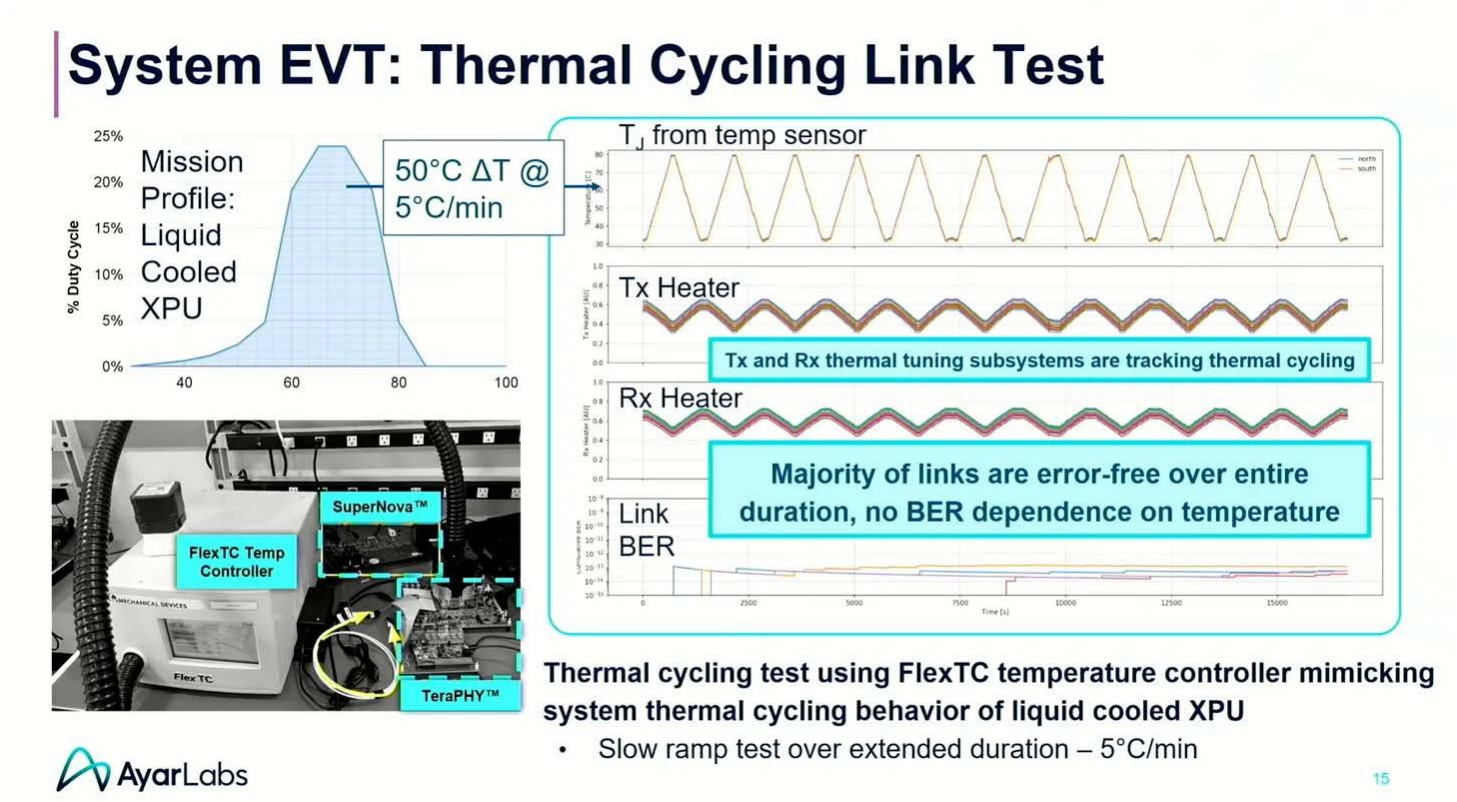
尽管Ayar Labs尚未披露确切的端口架构（即DWDM波长数量、每个FAU的光纤数量等），但其采用双向光链路的设计意味着：发射与接收最多需要约64根光纤，连接外部激光源则最多需额外数十根光纤。 然而——Ayar的战略始终聚焦于WDM技术，这意味着每组光纤单元（FAU）的总光纤数量可能低至32根。与前两代产品相同，第三代TeraPHY仍采用微环调制器技术，在保持光芯片小型化的同时，通过CWDM或DWDM技术为未来带宽扩展提供支持。



Ayar Labs 与 Alchip

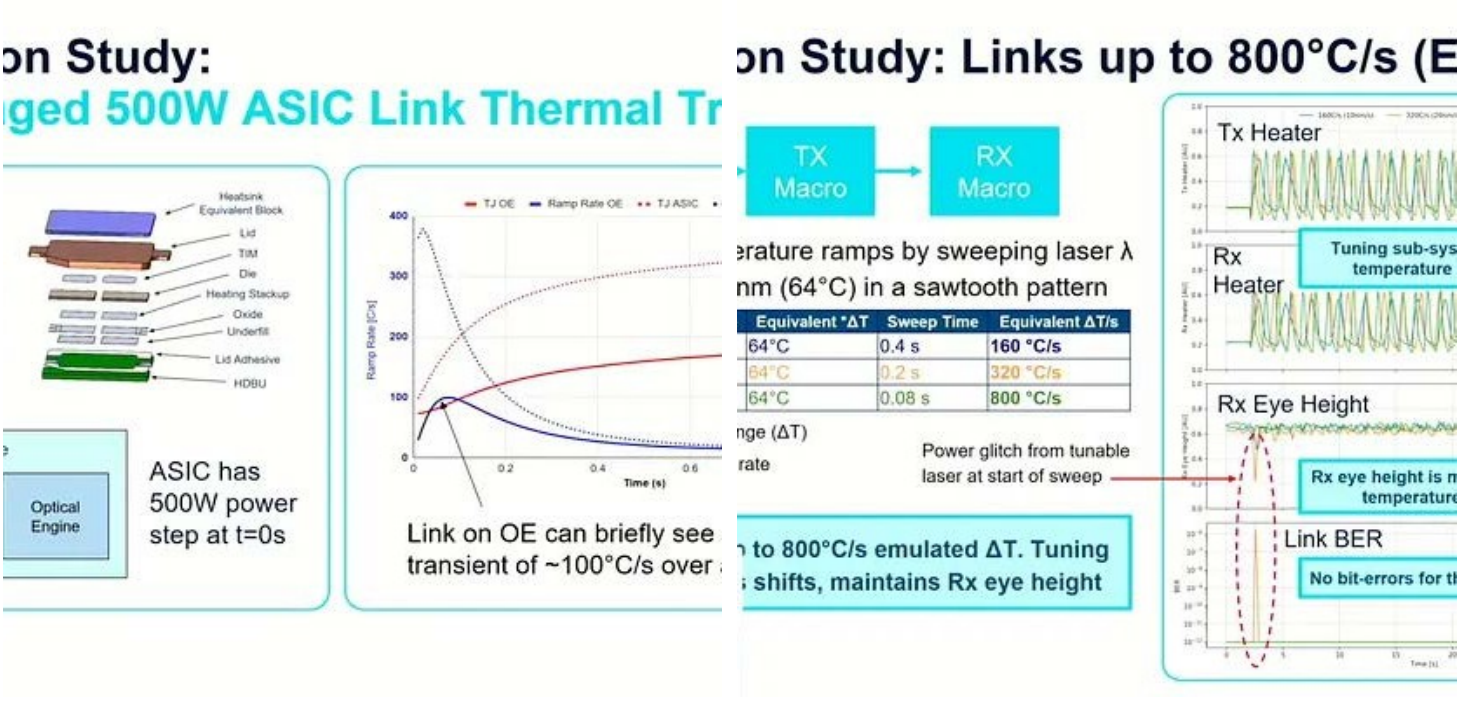
Ayar Labs还与Alchip和GUC建立合作，将其芯片集成到Alchip和GUC的XPU解决方案中。上图展示了一款搭载两个光罩尺寸计算芯片和8个TeraPHY光引擎的XPU，可实现高达108 Tbit/s的单向带宽。

在Hot Chips 2025大会上，Ayar Labs公布了低速热循环链路测试结果——在约5°C/分钟的速率下持续热循环超过四小时，全程链路误码率表现优异。



来源：Ayar Labs

然而，研究MRM在温度急剧变化时的抗扰性，与证明该连接在宽温度范围内长期稳定性同样重要。在同一场Hot Chips演讲中，Ayar解释了团队如何通过扫描激光波长来模拟快速温度变化，而非采用能在封装内实现0至500W阶跃的ASIC芯片。 控制电路会检测环形谐振是否发生漂移——这种漂移可能由输入激光波长变化或环形温度变化引起，因此控制电路以相当于温度变化的速率扫描激光波长。 例如20nm/s的扫描速率可模拟0.2秒内64°C的温度变化，相当于320°C/s。该研究表明在高达800°C/s的温度变化下仍无比特误码。



Ayar Labs

Ayar Labs拥有众多战略支持者，包括格罗方德、英特尔资本、英伟达、AMD、台积电、洛克希德·马丁、应用材料和唐宁。

Nubis

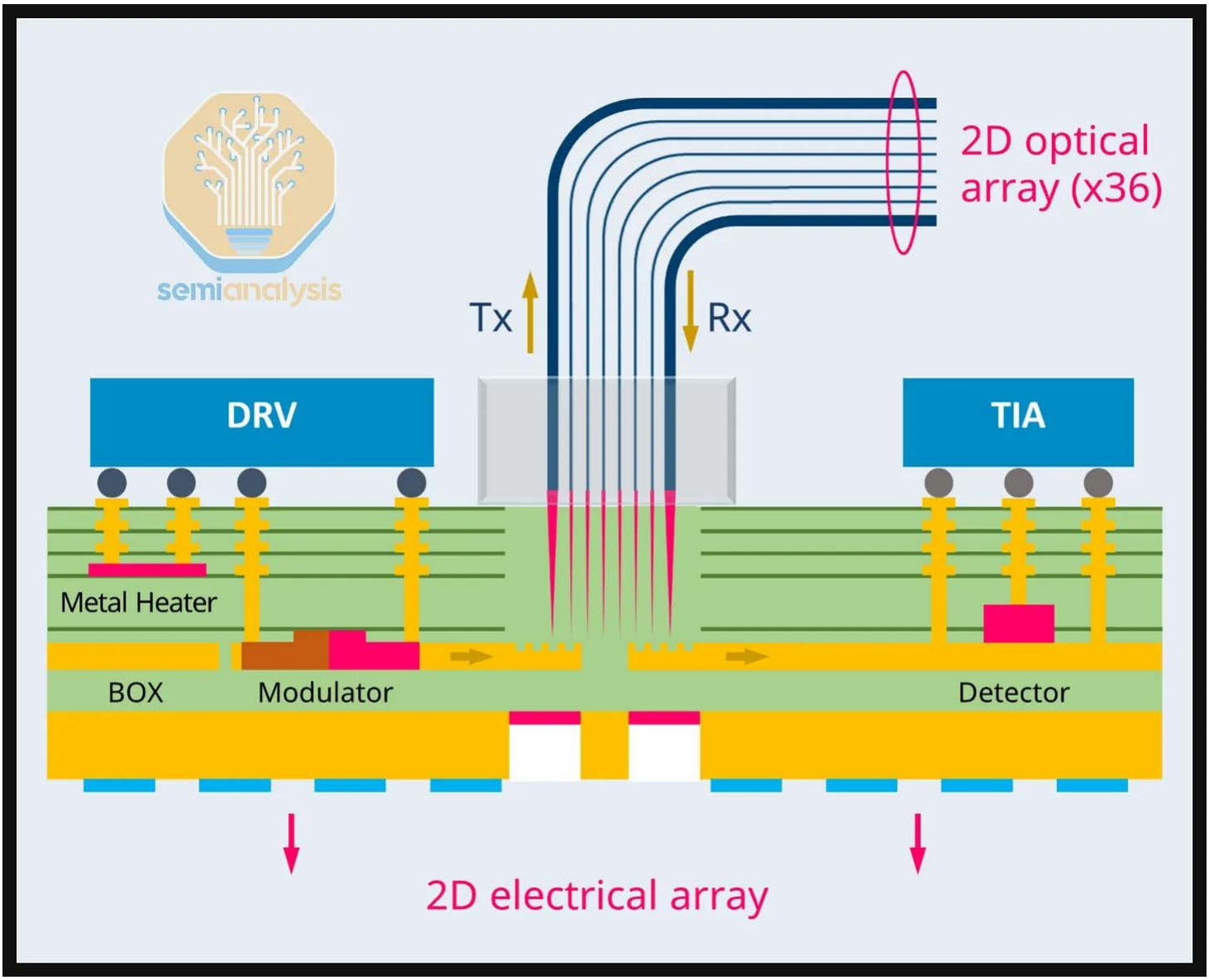
Nubis公司于2025年10月被Ciena收购。与Ayar类似，Nubis提供可集成至客户主机芯片的光学引擎芯片，但侧重于单波长连接。该公司专注于互操作性——涵盖协议与机械（即可插拔）层面——这决定了其技术选择。 Nubis还肩负着更宏大的使命：全面攻克I/O瓶颈问题，其解决方案同时涵盖光纤与铜缆技术。

其现有光引擎产品为Vesta 100 1.6T NPX光引擎。该插拔式模块提供16条100G通道，实现1.6T双向带宽，模块尺寸为6x7毫米。与其他公司不同，Nubis主要采用MZM调制器，这主要得益于该调制器的互操作性、可靠性和成熟度。 另一项关键设计决策是Nubis模块兼容符合IEEE/OIF标准的电接口，因为他们认为多数ASIC开发者将继续采用这些技术。

Nubis的核心差异化在于其光纤耦合方式。该公司采用表面耦合技术，具体而言是利用薄玻璃片辅助光纤的布线与对准。与边缘耦合技术（即光纤连接至芯片边缘）不同，Nubis的二维光纤阵列方案通过将光纤从硅光子芯片顶部连接至阵列边缘实现。

边缘连接光纤，Nubis的二维光纤阵列方案则是从硅光子芯片顶部连接光纤。

观察下图：PIC（底部绿色部分）包含调制器、光探测器及波导，其顶部安装有EIC。红色极柱为光纤，而容纳光纤的块状结构为玻璃块（FAU），用作光纤固定座。 FAU顶部设有激光钻孔，确保光纤精确定位。通过采用二维光纤阵列，该设计可将36根光纤（16根发射、16根接收、4根激光）连接至PIC，无需采用波分复用技术即可在更少光纤上承载更多波长。这使Nubis FAU成为当前出货密度最高的同类产品之一。

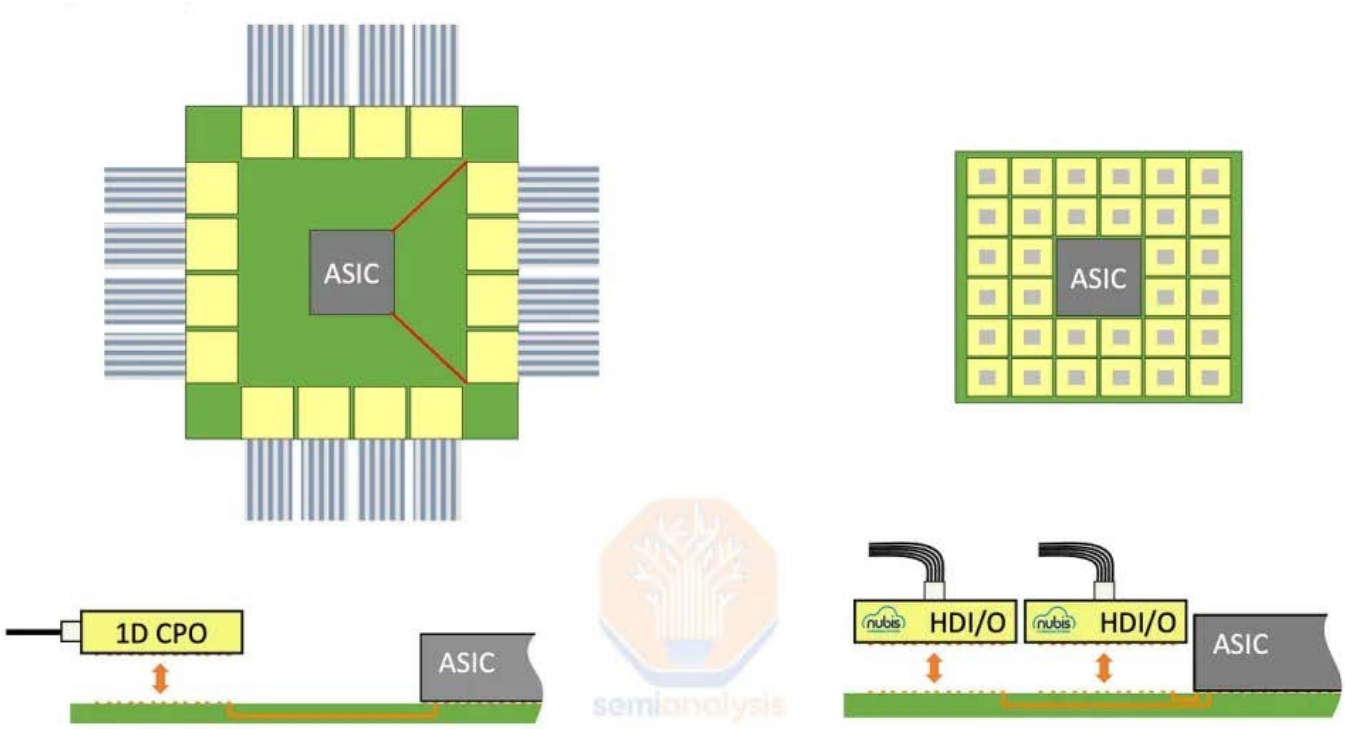


Nubis

尽管台积电等公司已将二维光纤阵列纳入路线图，且其在垂直耦合方面具有关键优势，但目前除Nubis外尚无厂商实现量产，这使其独树一帜——尽管其他厂商计划后期转向二维阵列。

该方案采用住友电气研发的特殊光纤FlexBeamGuidE，在实现90度弯曲时仍能保持高可靠性与低损耗特性。

使用二维阵列而非边缘耦合的另一优势在于，物理层面上可连接的光纤数量限制更少。如下方示意图所示，采用Nubis的二维光纤阵列结构，可在ASIC周围布置多排光引擎，只要封装允许，即可提升带宽密度。



来源：Nubis

2025年4月，Nubis宣布推出新一代PIC产品——一款每通道支持16×200G的硅光子集成电路，其单向海滩式封装密度达0.5Tbps/mm（与电气主机接口密度匹配）。 此外，Nubis宣布与Samtec建立合作伙伴关系，将提供32通道200G（6.4T）光模块样品，该模块可与Samtec Si-Fly HD共封装铜连接器实现即插即用兼容。

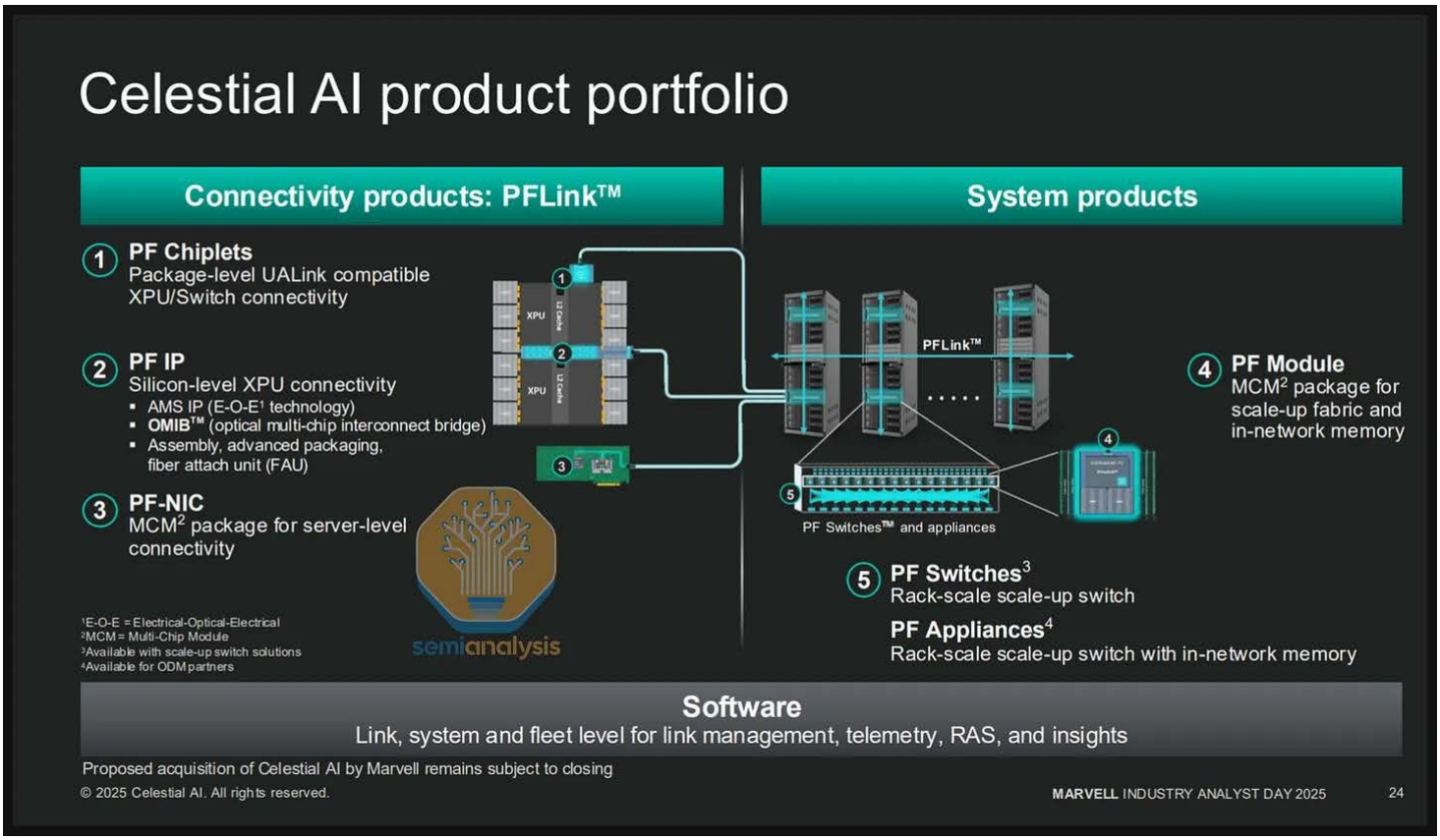
相较于其他CPO方案，该方案实现了铜缆与光纤的通用封装尺寸；长期来看，这将为CPO部署构建开放可插拔生态系统。

最后，在铜缆领域，Nubis也在OFC展会上宣布并展示了其线性再驱动芯片Nitro，该芯片适用于有源铜缆（ACC），可将200G铜缆传输距离延长至数米。此项技术是与安费诺合作开发的，后者将

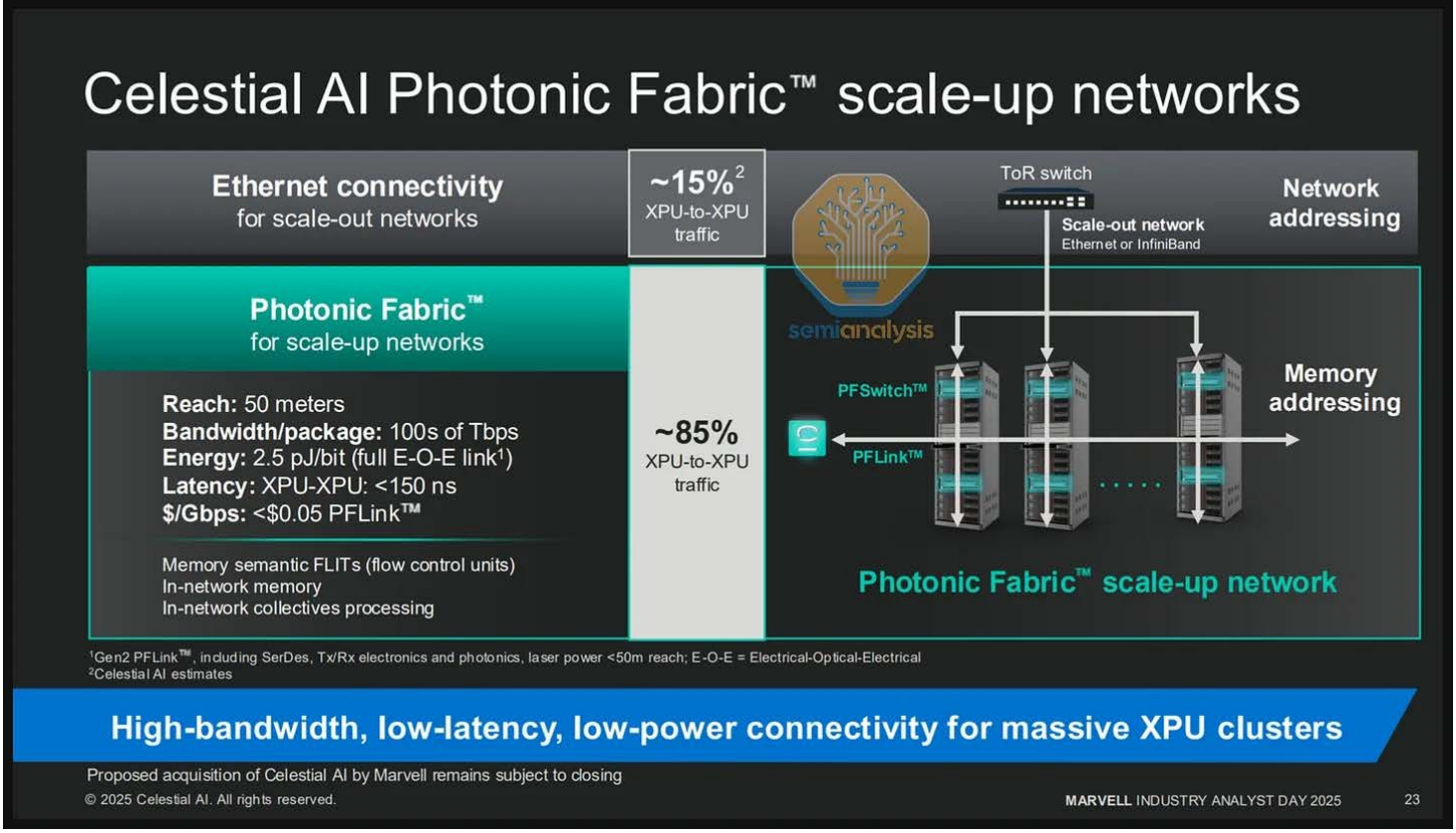
基于Nitro线性重驱动器构建ACC。

Celestial AI

Celestial AI是一家专注于人工智能规模化网络光互连解决方案的知识产权、产品与系统公司。 该公司的核心技术目标是将光子器件（调制器、光电探测器、波导等）集成到中介层中，并通过接口（带光纤耦合器的光纤耦合器）连接外部世界。下图清晰展示了Celestial AI基于光子学的互连解决方案套件，该公司将其命名为"光子织构™"（PF）。



来源：Celestial AI



来源：Celestial AI

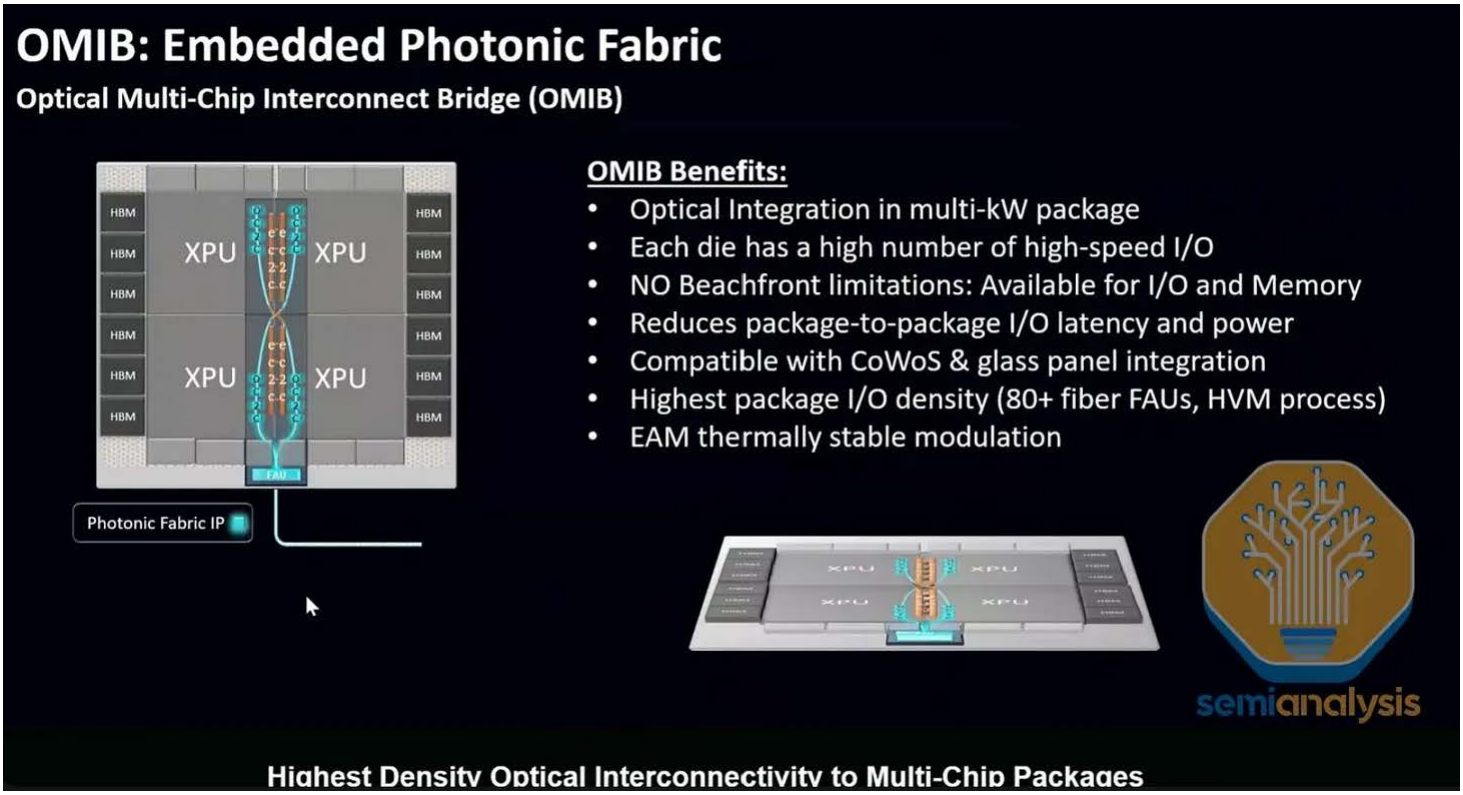
光子织网™（PF）芯片采用台积电5纳米工艺制造，集成UCIe和MAX PHY等芯片间接口，实现XPU间互联、XPU与交换机互联以及XPU与内存互联。客户可将其与自研XPU协同封装，相比基于电信号收发器接口的CPO产品，能提供更高带宽密度和更低功耗。 Celestial AI为客户定制开发这些芯片片，以适配特定的芯片间接口和协议。第一代PF芯片片支持16 Tbit/s带宽，第二代将实现64 Tbit/s带宽。

与传统铜质走线相比，光子芯片具有显著的功耗优势。采用线性SerDes的传统铜缆在224G速率下功耗约为5 pJ/bit。由于需要双端传输，总功耗约达10 pJ/bit。而Celestial AI的解决方案在整个电-光-电链路中仅需约2.5 pJ/bit（外加外部激光器约0.7 pJ/bit）。

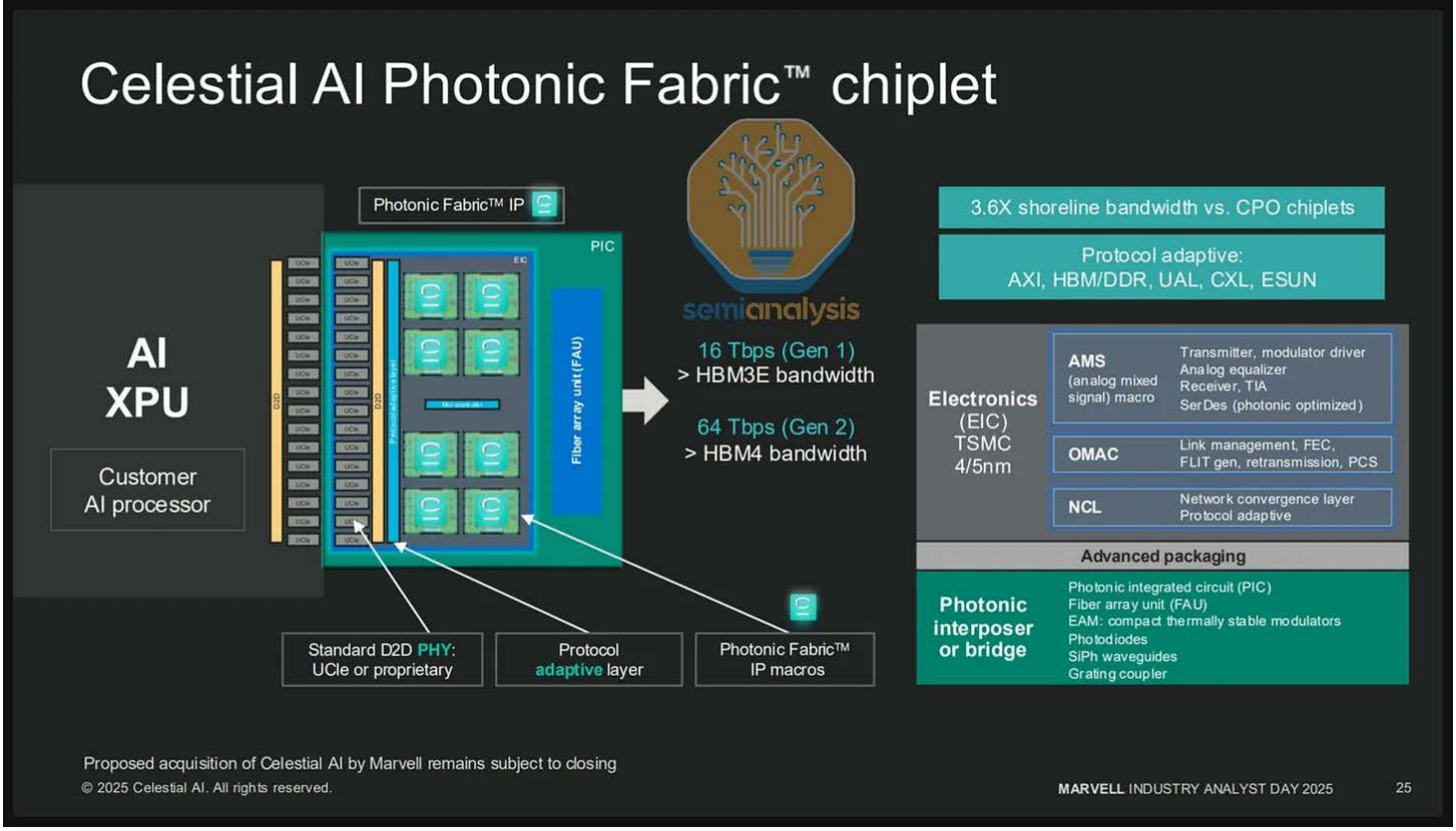
接下来，光子织物™光多芯片互连桥™（OMIB™）本质上是一种CoWoS-L风格或EMIB风格的封装解决方案。它将光子学直接集成到中介层中的嵌入式桥接器上，使桥接器能够将数据直接传输至消费点。由于不受海滩效应限制，其整体芯片带宽高于光子片级芯片。

在采用金属互连的传统中介层或基板中，将I/O接口置于芯片中心并不现实——这会导致布线复杂度过高，且因高密度信号拥堵引发严重的串扰问题。然而借助OMIB光中介层，Celestial AI得以将中介层直接部署于ASIC下方，突破岸线限制，实现更快速高效的数据传输，同时将串扰降至最低。

光互连器使I/O接口能够部署在芯片的任意位置，因为**光**波导在传输距离上几乎不会造成信号衰减，从而消除了传统海岸线布局的限制。 它们还能消除串扰现象——不同波导中的光信号不会像密集铜迹线中的电信号那样相互干扰，因为光信号高度局限于波导芯内，仅在包层外部存在微弱的衰减场。这种从根本上重构I/O设计与布局的方式，充分释放了光学技术蕴含的潜力。

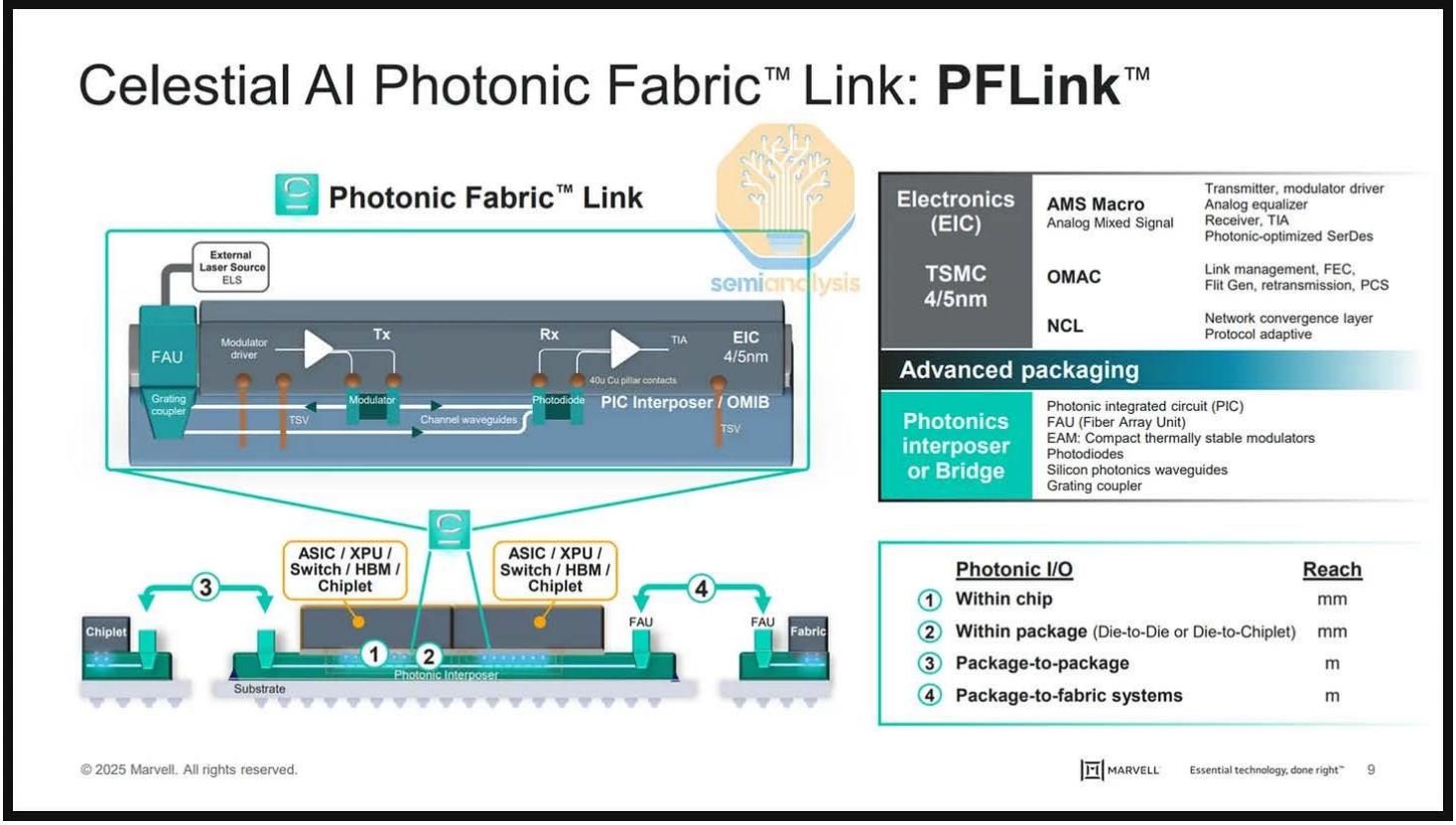


来源：Celestial AI



来源：Celestial AI, Marvell

光学中介层或通过先进封装传输光信号的构想，与Lightmatter的解决方案存在若干相似之处——两者均将光信号路由至逻辑芯片下方，从而规避海岸线限制。但二者存在关键差异： Celestial AI采用类似硅桥的光子桥（可类比CoWoS-L硅桥），而Lightmatter则使用大型多光罩光子中介层，置于多个独立芯片下方。 Lightmatter的技术构想更为宏大——其M1000三维光子超级芯片计划实现400^{0平方}毫米的介质尺寸，同时支持介质内部的光学电路切换，并提供高达114太比特每秒的总聚合带宽。



来源：Celestial AI、Marvell

最后，Celestial AI推出光子结构™内存设备（PFMA），这是一款基于台积电5纳米工艺的高带宽、低延迟可扩展结构，内置网络内存，总带宽达115.2T，可连接16个ASIC，每个ASIC提供7.2T的可扩展带宽。 值得注意的是，PFMA是全球首款将光I/O集成于芯片中心的硅器件，从而为内存控制器释放了珍贵的周边物理I/O资源。这使PFMA成为主机CPU内存与存储之间用于KVCache卸载的"温层"内存。

Celestial AI技术的关键差异化优势在于其采用的电吸收调制器（EAM）。本文第三部分详细阐述了EAM的工作原理，并探讨了其优势与权衡取舍。在此我们重述主要讨论内容，因为理解EAM的利弊是把握Celestial市场战略的核心。

相较于磁阻调制器（MRM）和磁致伸缩调制器（MZI），EAM具备多重优势：

- 显然——EAM与MRM均配备控制逻辑和加热器以抵御温度波动，但EAM对温度的敏感度本质上更低。相较于MRM，EAM在50°C以上环境具有显著更优的热稳定性，而MRM对温度极为敏感。MRM典型稳定性为70-90 pm/°C，这意味着2°C的温度变化会导致

共振位移达0.14纳米，远超with MRM 的0.1纳米共振位移

性能骤降。相比之下，EAM器件可承受高达35°C的瞬时温度变化。这种耐受性对Celestial AI的技术方案尤为关键——其EAM调制器位于中介层内部，紧邻散发数百瓦功率的高功耗XPU计算引擎。 EAM还可耐受约80°C的高环境温度范围，这适用于紧邻XPU而非置于其下方的芯片级封装应用场景。

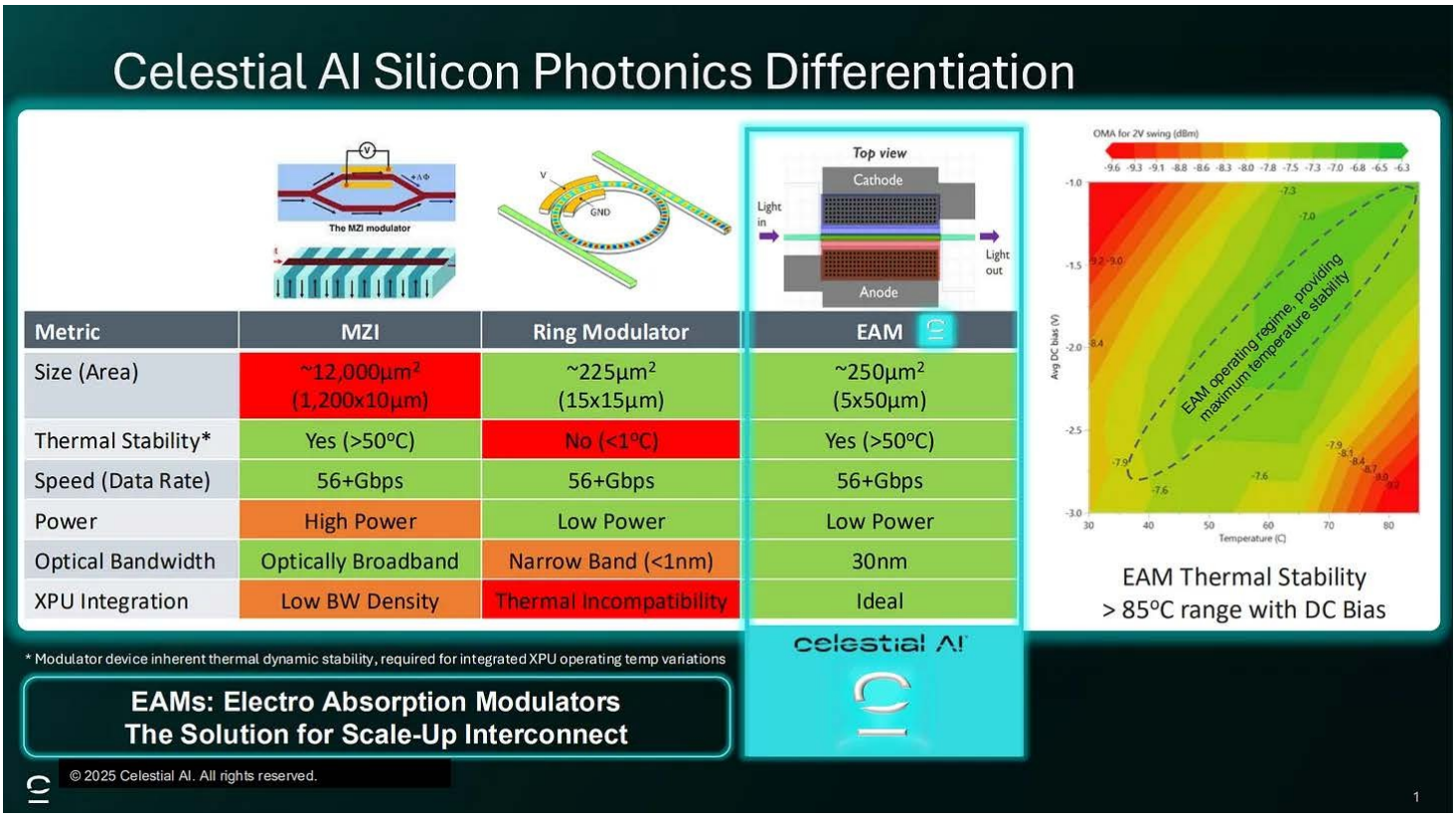
- 与MZI相比，EAM尺寸更小且功耗更低，因为MZI相对较大的尺寸需要高电压摆幅，需要放大串解码器才能实现0-5V的摆幅。 马赫曾德调制器（MZM）尺寸约为 12,000平方毫米，电子调制器（EAM）约为 250平方毫米（5x50 毫米），磁阻调制器（MRM）尺寸在 25平方毫米至 225平方毫米之间（直径 5-15平方毫米）。 MZI 还需消耗更多功率来驱动加热器，以使如此大的器件保持所需的偏置。

另一方面，采用GeSi EAM进行CPO存在若干局限：

- 基于硅或氮化硅的物理调制器结构（如MRM和MZI）被认为具有远高于锗硅基器件的耐用性和可靠性。 诚然，鉴于锗基器件的加工与集成难度，许多人担忧其可靠性。但Celestial公司指出，基于锗硅的电致发光调制器（本质上是光探测器的反向结构）在可靠性方面具有确定性——这得益于当今收发器中光探测器的广泛应用。
- GeSi调制器的带边自然位于C波段（即1530nm-1565nm）。 设计量子阱以将该特性转移至O波段（即1260-1360nm）是极具挑战性的工程难题。这意味着基于GeSi的EAM调制器很可能仅适用于封闭式CPO系统，难以融入开放式芯片生态系统。
- 围绕C波段激光源构建生态系统可能存在规模不经济现象，相较于成熟的O波段连续波激光器生态系统。多数数据通信激光器针对O波段设计，但Celestial指出1577nm XGS-PON激光器产量可观，这类激光器通常应用于家庭及企业光纤接入场景。
- 硅锗EAM插入损耗约为4-5dB，而MRM和MZI均为3-5dB。虽然MRM可直接实现不同波长的复用，

但EAM需搭配独立复用器实现CWDM或DWDM，这会略微增加潜在损耗预算。

总体而言，Celestial AI一直在致力于创新其定制链接——这些链接不依赖任何齿轮箱组件，提供更低的延迟和更高的能效，并能适应不同类型的协议。如前所述，Celestial AI是唯一主要采用EAM进行调制的行业巨头。 这意味着该公司未来需将EAM设计整合至代工厂流程，而其他CPO厂商则可依托台积电COUPE平台——该平台的PDK已包含MRM及相关加热器组件。



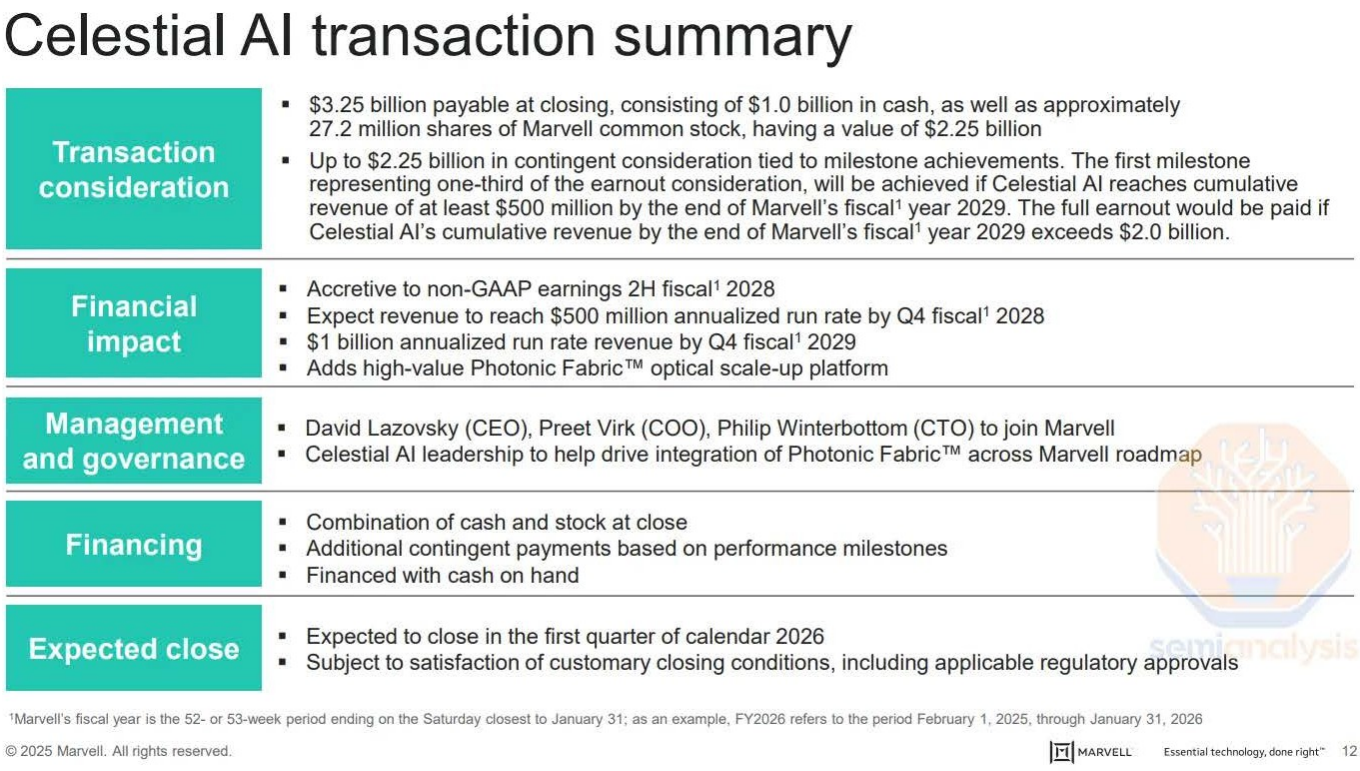
来源：Celestial AI

在短期内，Celestial AI正致力于实现小芯片发布的雄心勃勃的时间表。Marvell在交易摘要中宣布，预计到2028年1月底（即Marvell 2028财年结束时，简称F1/28），Celestial带来的年化营收将达到5亿美元。 在巴克莱全球科技大会上，该公司进一步补充称，该营收年化额预计将在2028日历年末翻倍至10亿美元（其中大部分2028日历年业绩将计入Marvell截至2029年1月的财年，即F1/29），这意味着该产品将在两年内（即2027年底前）实现商业可行性。

根据协议条款，Celestial AI还将获得额外22.5亿美元股权支付。

持有人取决于公司实现累计收入

截至2029年1月（即Marvell 2029财年结束时——即F1/29），累计收入需达到20亿美元。实现该全额付款的首个里程碑是：在2029年1月前达成5亿美元的累计收入，以获取三分之一的付款金额。 预计F1/29财年末的10亿美元营收运行率仅为业绩对赌金额的一半——这意味着Celestial需新增订单客户才能达成20亿美元的业绩对赌目标。



来源：Celestial AI、Marvell

关于收购Celestial AI一事，Marvell于2025年12月2日提交了8-K报告，发行了亚马逊认股权证，行权价为87.0029美元，有效期至2030年12月31日。 该认股权证的归属条件为"基于亚马逊在2030年12月31日前直接或间接采购光子结构产品"，强烈暗示AWS的Trainium将成为目标产品——该产品将于2027年末开始量产。 在Marvell行业分析师日活动中，Celestial AI透露某大型超大规模企业已选定其光互连技术用于先进AI系统，该技术将应用于该企业下一代处理器的量产版本。

结合交易摘要中的业绩对价支付时间表和产品收入指引，这表明Celestial AI的目标是在Trainium 4内部署其解决方案。

Item 8.01 Other Events.

On December 2, 2025, the Company announced that it entered into a definitive agreement to acquire Celestial AI Inc. In connection with this acquisition, which is subject to customary closing conditions, the Company entered into a Transaction Agreement with Amazon.com, Inc. ("Amazon", including its affiliates) and related Warrant, under which the Company issued to Amazon a warrant (the "Warrant") to acquire up to 1,045,171 shares (the "Warrant Shares") of Company common stock (the "Common Stock"). The Warrant Shares vest based on Amazon's purchases of photonic fabric products, indirectly or directly, through December 31, 2030.

Subject to certain conditions, including vesting, the Warrant may be exercised, in whole or in part and for cash or on a net exercise basis, at any time before December 2, 2031, at a purchase price per share of Common Stock equal to $87.0029 (the "Exercise Price"). The Exercise Price and the Warrant Shares issuable are subject to customary antidilution adjustments.

The Transaction Agreement includes customary representations, warranties and covenants of the parties and sets forth certain provisions relating to Amazon's equity interest in the Company.

The Warrant and the Warrant Shares have not been registered under the Securities Act, in reliance on the exemption from registration provided by Section 4(a)(2) of the Securities Act and rules and regulations of the U.S. Securities and Exchange Commission promulgated thereunder.

来源：Marvell证券监管文件

让我们通过深入阐述Celestial AI即将面世的首批规模化解决方案来结束本次讨论。这些方案将围绕其16Tbit/s光子链路展开，该链路将连接至小芯片。光栅耦合器将光子阵列单元（FAU）与通道波导相连。一款规模化交换机专用集成电路（ASIC）——可能是Marvell的115.2T规模化ASIC——将通过光子学链路和PF小芯片与XPU实现光连接。尽管Celestial预计其初期市场收入主要来自小芯片业务，但该公司定位为系统公司，并已推出多款基于光学的内存扩展解决方案，这些方案将在首个规模化网络解决方案上市后陆续面世。

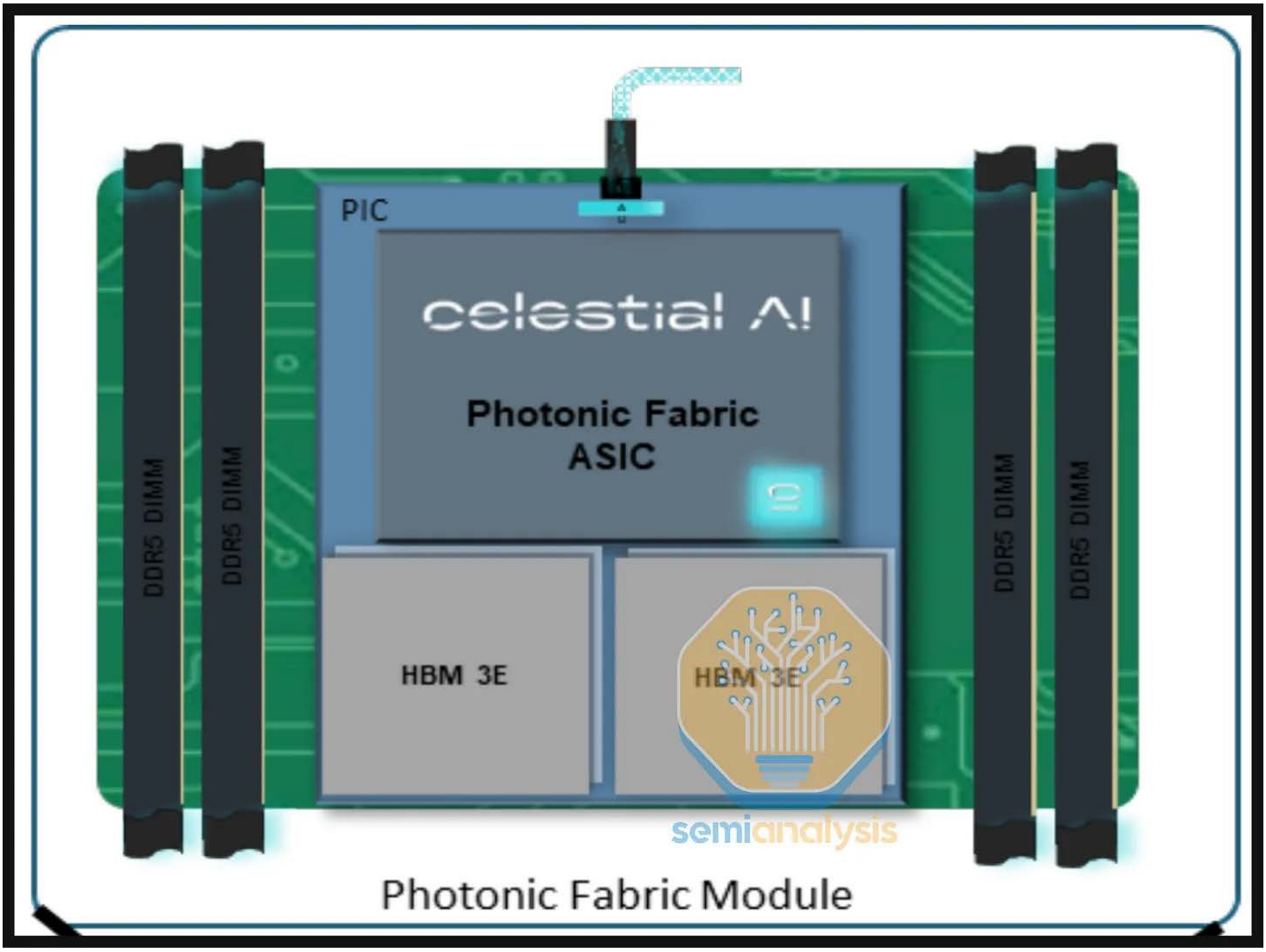
利用光学技术通过多层交换结构实现世界规模扩展并非新概念，尽管其距产品化仍遥遥无期。该方案可采用与GB200的NVL576架构相似的拓扑结构：两层交换结构通过OSFP收发器模块与光纤相互连接。 Celestial AI的多交换层方案与此类似，但省去了实际收发器的使用。

然而，与NVL576概念相比最大的区别在于：这种可扩展ASIC既可充当路由器，又能作为内存端点，而NVSwitch仅能为GPU间的高带宽链路提供路由功能。这一差异至关重要，因为Celestial AI的核心卖点在于其可扩展解决方案能规避"硅海滩"限制——该限制原本会限制连接至XPU的HBM堆栈数量。

为实现这一目标，连接至XPU的HBM堆栈被替换为连接至光子结构（共享HBM池）的芯片组。该共享HBM池即光子结构设备（PFA），这套2U机架式系统由16个

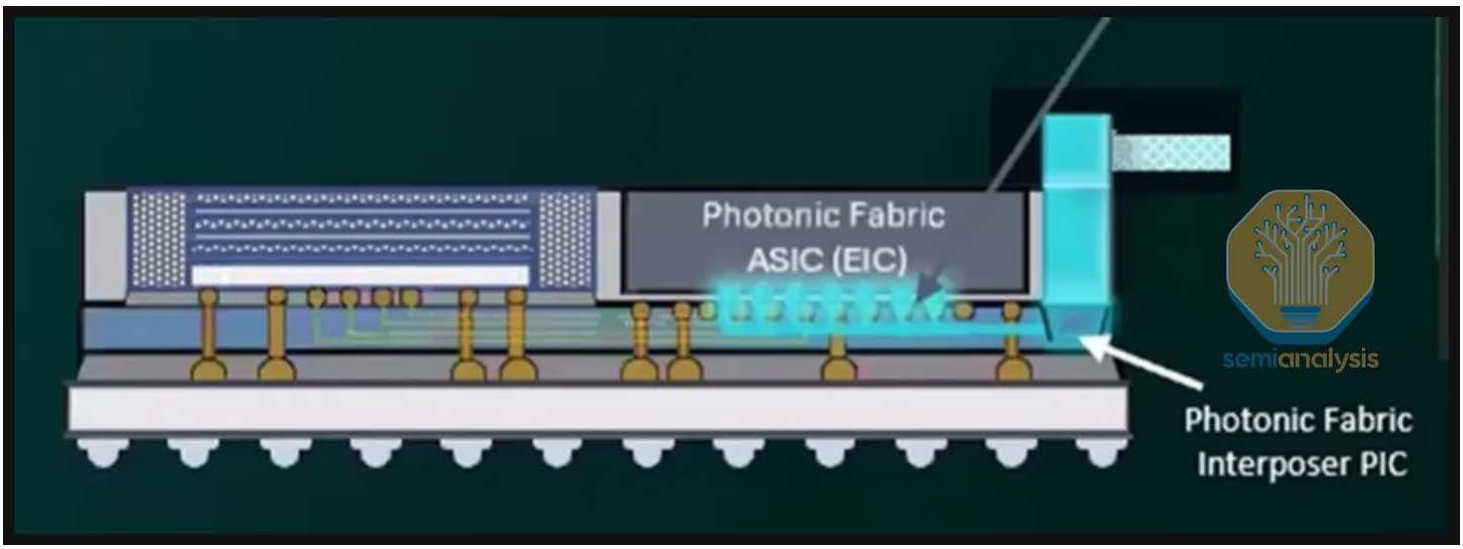
由单端口组成的光纤结构ASIC。

每个ASIC采用2.5层封装，集成两个36GB HBM3E内存和八个外部DDR5内存。



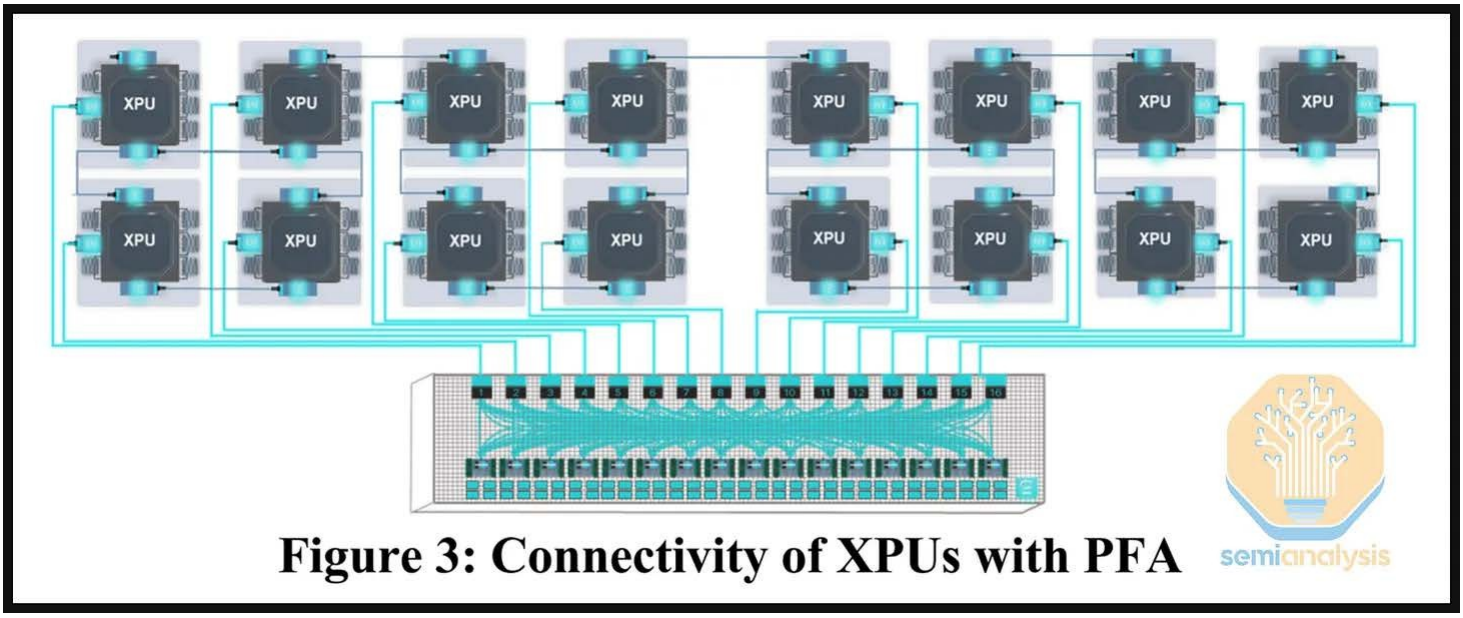
来源：Celestial AI

光电I/O（光子结构IP）被安置在ASIC内部区域而非边缘位置，从而释放边缘区域空间以满足其他应用需求。



来源：Celestial AI

从宏观角度看，每个PFA模块都是一个16基数交换机，最多可支持16个XPUs。 并非让每个XPU向全部16个端口扇出连接，而是通过交换机箱内部实现全互连：连接至每个交换机ASIC的光纤连接单元（FAU）向16个交换机I/O端口进行扇出连接。因此每个XPU仅通过一条光纤链路连接至箱体外部的一个交换机端口。



来源：Celestial AI

通过将内存置于XPU外部并纳入共享交换接口，数据得以聚合，随后所有XPU可在全局聚合通信中从共享内存池中访问数据。

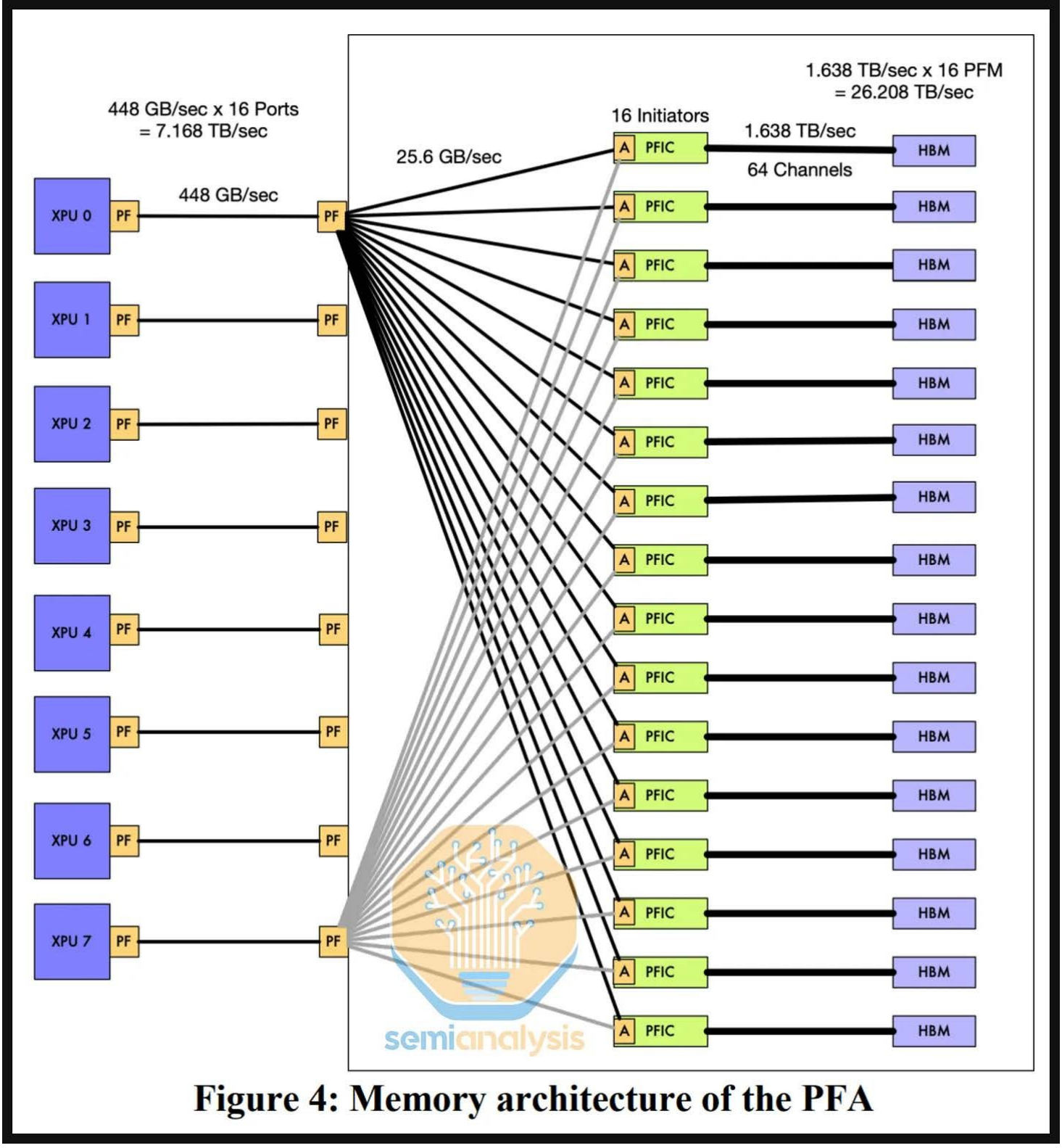


Figure 4: Memory architecture of the PFA

来源：Celestial AI

Lightmatter

Lightmatter以旗下光学中介层产品Passage™ M1000 3D光子超级芯片闻名业界，同时正推出多项解决方案以满足CPO路线图各阶段需求，其中多款芯片小片已在台积电完成流片。

首批上市的解决方案将是面向近封装光学（Near Packaged Optics）的光学引擎。

(NPO)于2026/2027年。在NPO方案中，光学引擎将被 s ol de ed t th

底板采用铜连接线，将XPU上的LR串并转换器与光引擎相连。Lightmatter的光引擎最多支持3个光纤阵列单元（FAU），每个FAU配备40根光纤，总计120根光纤。 NPO战略基于这样的理念：超大规模企业采用CPO的第一步将是先积累NPO的运营经验，这能降低产品风险——因为超大规模企业无需"承诺"采用CPO，可最终选择光纤或铜缆扩展方案来对接XPU或交换机上的LR SerDes接口。

由于Lightmatter的光学引擎解决方案基于台积电COUPE工艺和格罗方德45纳米SPCLO工艺，存在多种可扩展方向。 除通过100Gbaud PAM4实现每通道200Gbit/s（单向）传输外，该方案还支持：- PAM4配合DWDM8实现200Gbit/s传输- PAM4配合DWDM16实现100Gbit/s传输从而达成单纤3.2T传输容量。

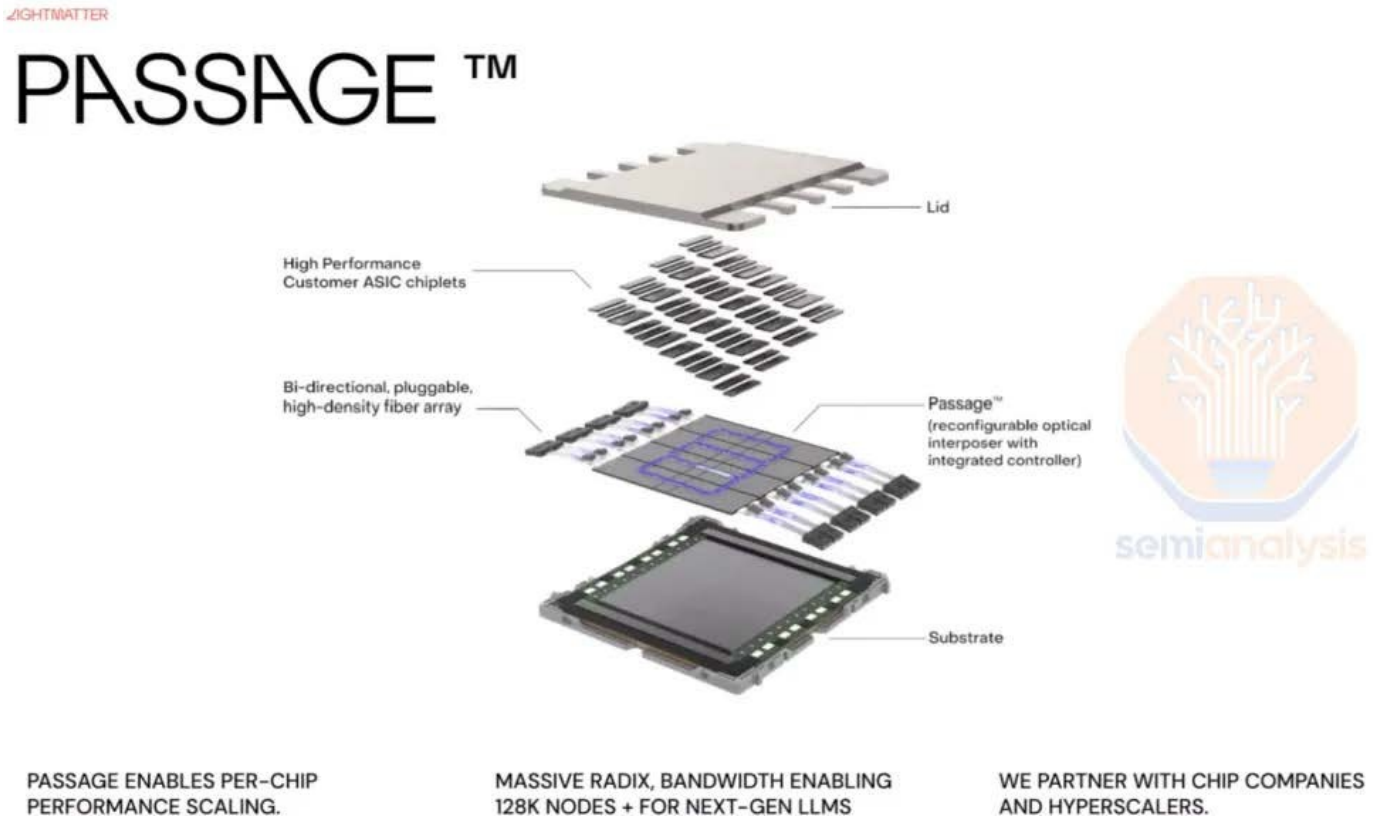
当其他CPO厂商选择采用商用激光器生态系统时，Lightmatter自主研发了名为GUIDE的外部激光器，目前正处于样品阶段。 与其他激光源通过单晶化InP晶圆制造独立激光二极管不同，GUIDE作为业界首款超大规模光子学（VLSP）激光器，开创性地将数百个InP激光器集成于单片硅芯片，可支持高达50Tbit/s的带宽。 Lightmatter宣称其独特的控制技术可管理这些集成激光器，通过超额配置InP激光器数量提升整体可靠性，并通过替换仍可工作的二极管实现"自修复"功能。 英伟达Quantum-X CPO交换机配备144个800G端口，需18个ELS模块，而Lightmatter宣称仅需两台GUIDE激光器即可满足同等带宽需求。

Lightmatter计划遵循COUPE路线图，于2027年和2028年全力推出CPO解决方案，随后在2029年及之后专注于其旗舰产品Passage™ M1000解决方案。

Lightmatter的M1000三维光子超级芯片是一款4000平方毫米的光学中介层，置于主计算引擎下方，负责电信号向光信号的转换。 该芯片已在SC25大会的机架级实测中完成演示，Lightmatter现已将其作为参考设计对外开放。Passage架构通过TSV技术在XPU与光引擎间传输电信号及电源，并采用SerDes实现两者连接。通过将ASIC直接置于光互连器之上，Passage消除了对大型高功耗SerDes的需求，从而

采用1,024个紧凑型低功耗串解码器（SerDes）（传统串解码器的1/8）实现

实现总I/O带宽达114Tbit/s（每路SerDes运行速率为112Gbit/s）。通过将ASIC直接置于光互连器之上，芯片岸线约束问题也得以缓解。



来源：Lightmatter

该系统内置OCS冗余管理机制——当某条通信路径故障时，流量可通过备用路径重新路由，确保大规模系统持续运行。此外，相邻磁贴通过电气缝合技术连接，使其能够借助UCIe等接口实现电子通信。

通道采用直径约15微米的MRM（磁阻模块），每个模块集成电阻加热器，实现56 Gbit/s无归零调制。该模块包含16条水平总线，每条可承载多达16种颜色（波长）。这些颜色将由GUIDE提供，该系统以200 GHz网格频率为每根光纤输送16种波长。

Passage采用256根光纤，每根通过DWDM技术单向承载16个波长（或双向承载8个波长），单纤带宽达1 Tbit/s至1.6 Tbit/s。为提升良率，该方案将芯片连接的光纤数量降至最低，从而降低复杂度与制造难度。此外，其创新的光纤连接系统可实现故障光纤的便捷

从面板上断开并更换，从而提升可靠性和可维护性。下表展示了Passage当前支持的不同模式。

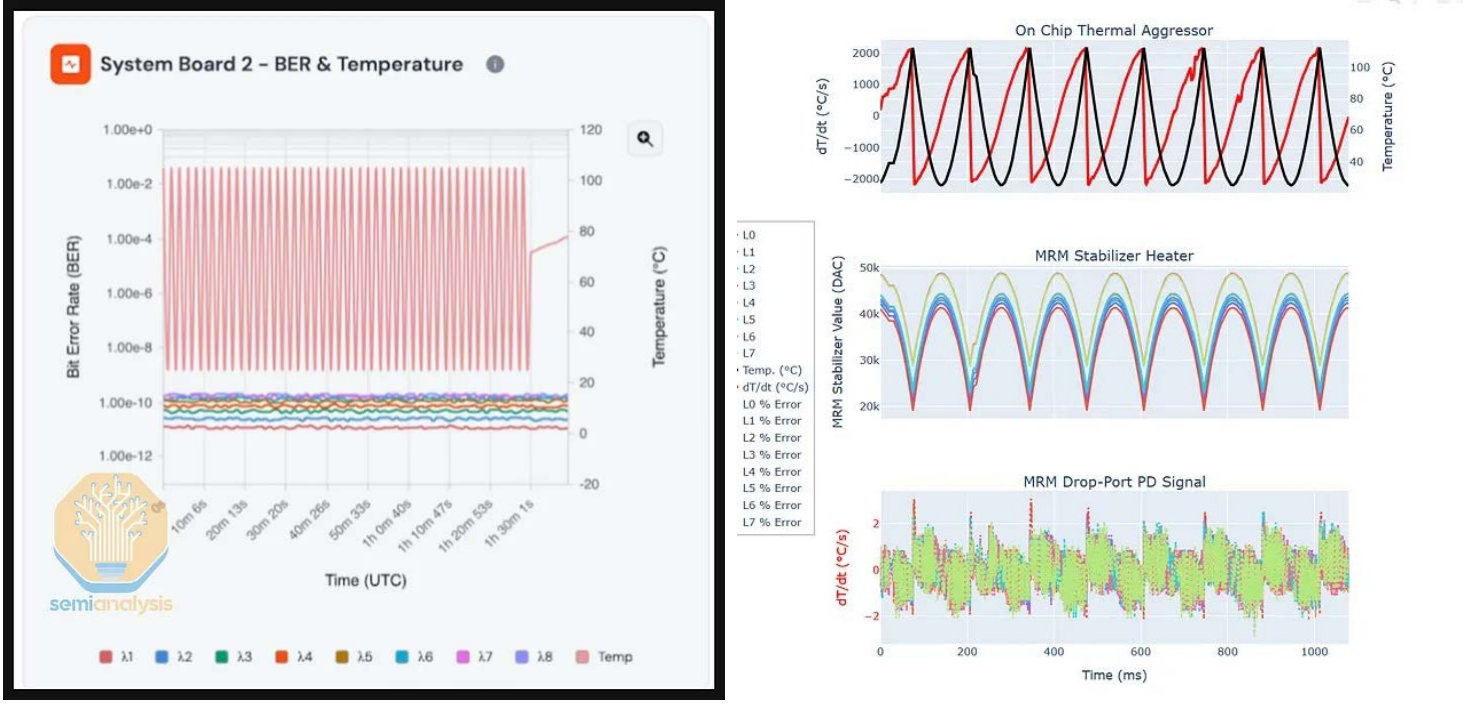
Bandwidth	Modulation	Number of Wavelengths	Transmission Type	CWDM/DWDM
56 Gbps	NRZ	16	Bi-directional	DWDM
56 Gbps	NRZ/PAM4	16	Uni-directional	DWDM
112 Gbps	PAM4	16	Uni-directional	DWDM
224 Gbps	PAM4	4	Bi-directional	CWDM
224 Gbps	PAM4	4	Uni-directional	CWDM

来源：Lightmatter

关于PASSAGE的关键争议之一在于其MRM的热稳定性——由于光学插入器直接位于高温XPU下方。相比之下，其他CPO方案并未将调制器置于XPU正下方，因此更易于热管理。 针对此质疑，Lightmatter公司解释称：PASSAGE系统中MRM的控制回路可承受每秒2000℃的温度波动，工作温度范围覆盖0至105℃——这意味着在10毫秒内可实现60至80℃的温度转换，且不会中断光链路传输。

SC25演示视频展示了25℃至105℃的温度变化曲线，呈现出宽广的工作温度范围，但其中80℃的过渡过程耗时约一分钟——这相当于每秒仅1.33℃的温升幅度。然而在SC25会议上，另一项采用片上热激器装置的独立演示实现了每秒2000℃的速率，此时MRM稳定器加热器使温度偏移值保持在1.33℃/秒。

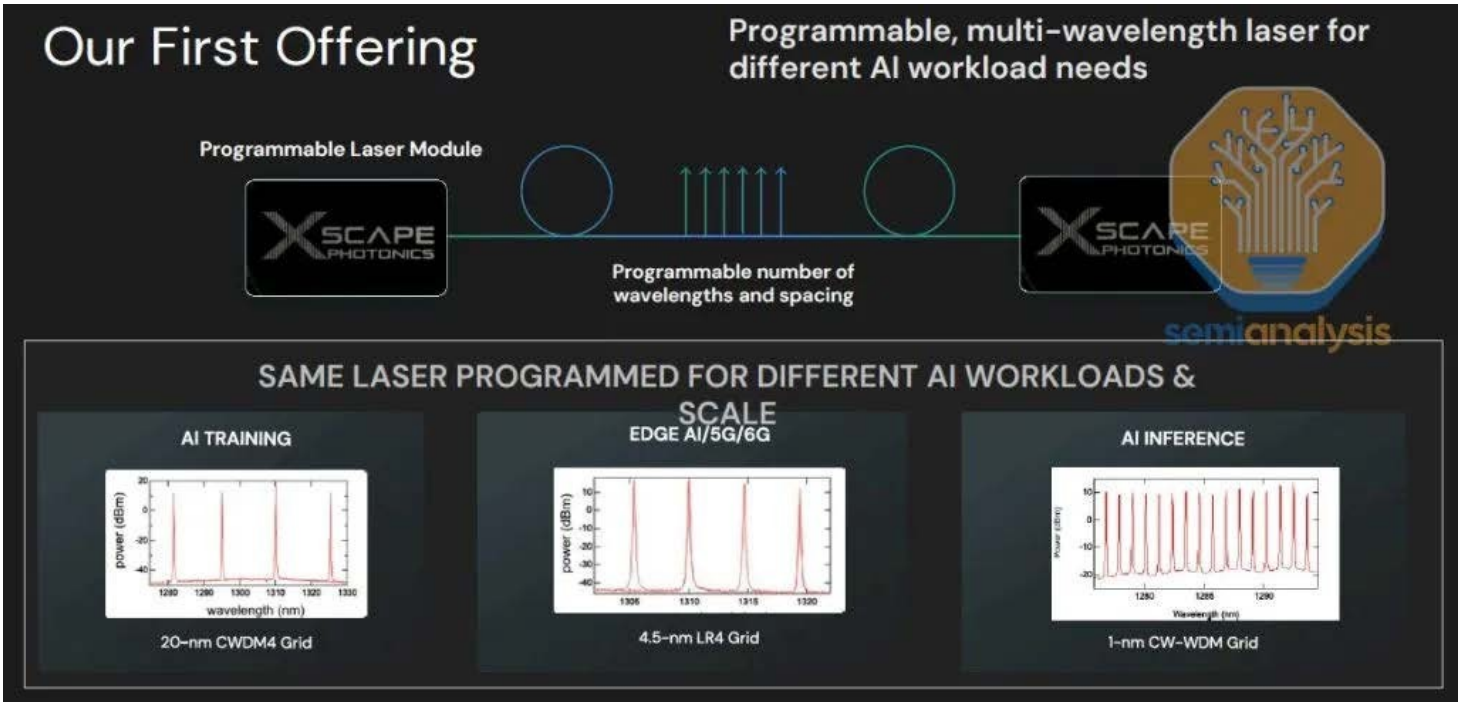
1.33℃/秒的偏移速率，但另一项同样在SC25会议上展示的芯片级热激荡器实验达到了2000℃/秒的速率，而MRM稳定器加热器使MRM本体的温度变化范围控制在更窄的-2至+2℃/秒区间。



来源：Lightmatter

Xscape Photonics

Xscape Photonics是一家创新公司，正在研发名为ChromX的可编程激光器，该产品可提供4至16种波长，未来计划扩展至128种波长。通过提供多达128种不同颜色，ChromX将实现远高于现有激光器的带宽——后者仅能提供4至8种波长。 ChromX 采用外部 III-V 激光器配合片上多色发生器，为波分复用技术生成多重波长。



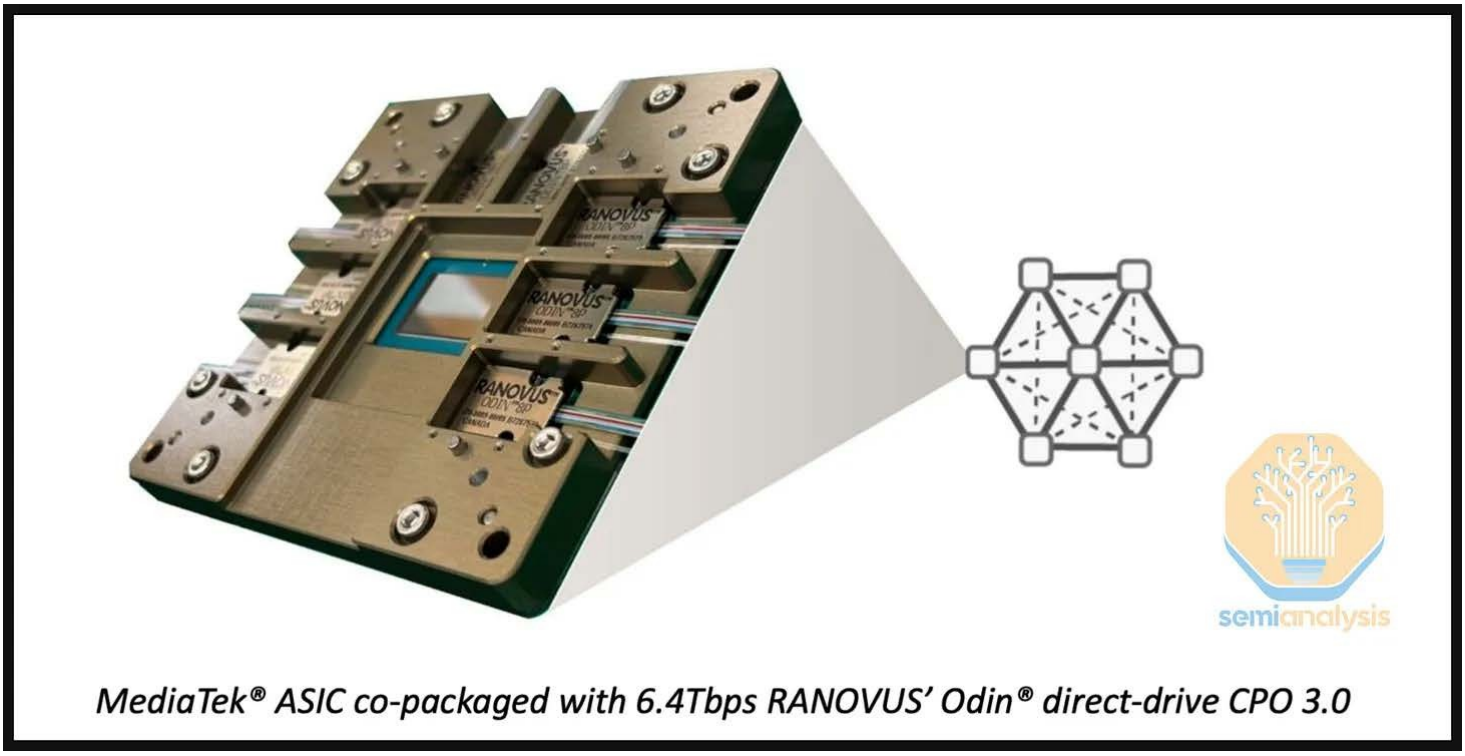
来源：Xscape Photonics

激光器的可编程特性提供了灵活性，能够为不同类型的工作负载提供特定波长，以满足不同的带宽和距离要求。值得注意的是，该解决方案仅需单一激光器，而现有CPO方案则需要多台功率极高且耗电量巨大的激光器。此外，所有波长均通过单根光纤传输，避免了大多数CPO系统因需多根光纤而产生的复杂性，显著减少了光纤耦合问题。

Ranovus

Ranovus专注于光子芯片技术及激光器设计制造。其产品已通过多个工艺节点流片：在格罗方德（GlobalFoundries）实现单片式CPO产品（该产品最初在AMF完成流片，后由GF收购AMF）；同时基于台积电COUPE工艺开发的产品可自然集成不同几何结构的光子集成电路（PIC）与光子互连电路（EIC）。

其奥丁光引擎采用微环谐振器调制器，可通过PAM4调制技术提供高达64条100Gbit/s的通道。



来源：Ranovus

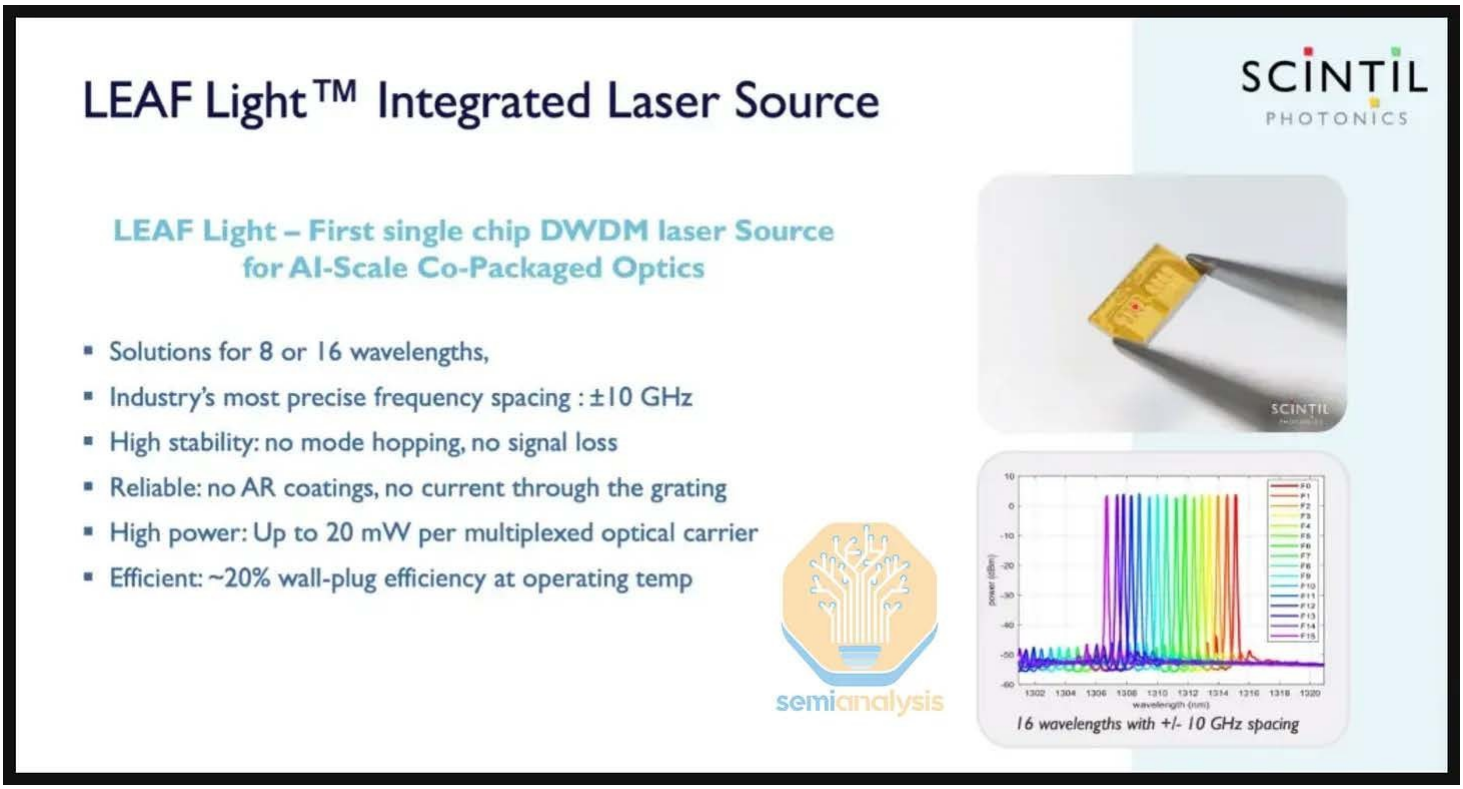
Ranovus的市场策略核心在于提供客户所需的互操作解决方案。当前重点是100G PAM4 DR光模块，但微环谐振器调制技术的应用使其能够灵活转向其他方案，例如56Gbaud NRZ调制。

但采用波分复用技术，将4个波长组合后，每对光纤可传输400G。

Ranovus已在其800G小芯片上实现与AMD的互操作性，并与联发科合作推出Odin直驱CPO 3.0解决方案，为超大规模企业未来的定制硅基XPUs提供小芯片方案。

Scintil

Scintil的主打产品LEAF Light是一款光子系统级芯片（PSoC），可提供裸片形式（KGD）或集成模块，整合8或16颗不同波长的激光器，波长间隔为200 GHz或100 GHz，通过DWDM技术实现单根光纤承载多波长传输。 该公司开发的电子控制技术可在温度变化条件下精确维持100GHz或200GHz波长间隔。其ELSFP模块（类似OSFP规格）遵循OIF定义的封装参考设计，便于客户集成外部激光源。Scintil方案与基于环形调制器的共封装光学器件兼容性良好。

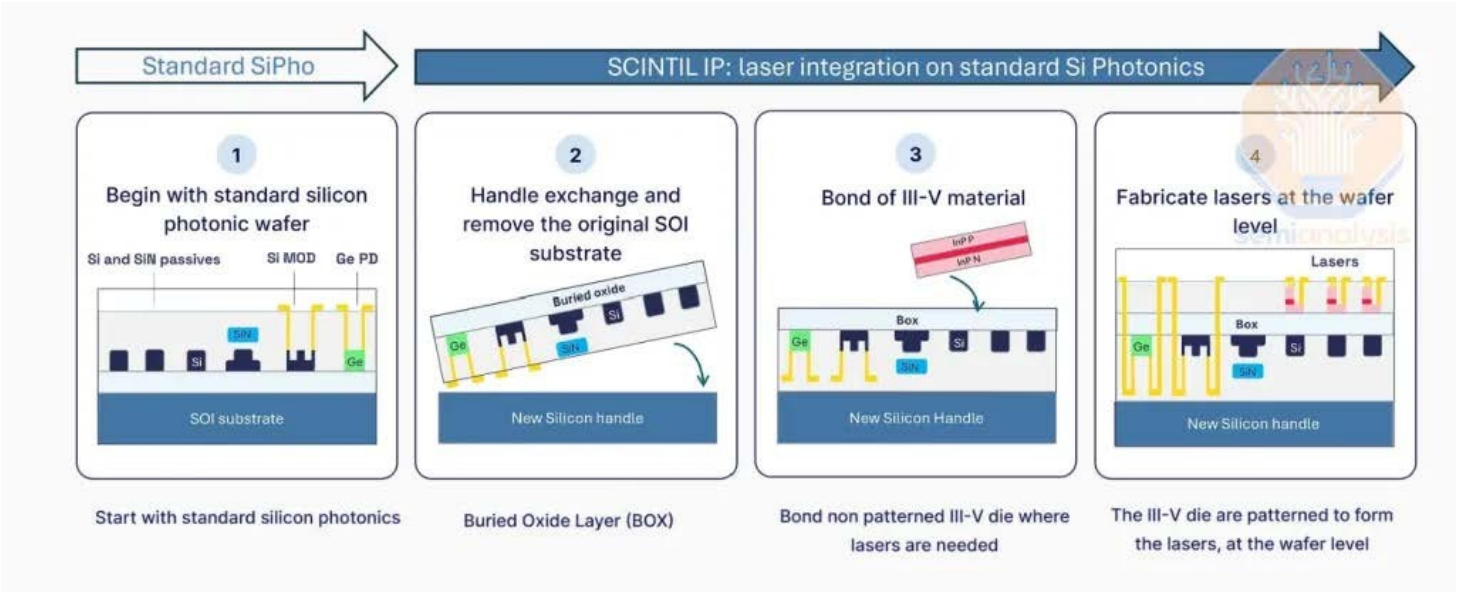


来源：Scintil

Scintil的技术被称为SHIP（Scintil异质集成光子学）。该技术的核心在于通过晶圆级工艺将III-V族激光器集成到标准硅光子器件上。工艺流程始于采用常规代工厂工艺制造的标准硅光子晶圆——其已集成波导、探测器及复用/解复用器。随后将晶圆翻转并键合至新衬底，

允许原始基板被移除，并露出埋入的识ox星id球e la前ye沿r。

未图案化的III-V族材料随后被键合至新暴露的表面。通过光刻技术对III-V材料进行图案化并蚀刻，最终制成集成激光器的单片硅光子芯片。这与传统基于InP激光器（采用电子束光刻机图案化）形成鲜明对比——后者难以实现DWDM所需的精确波长控制，导致难以支持高密度通道。



来源：Scintil

开发DWDM DFB激光器阵列面临巨大挑战，因为每个波长的频率都必须精确生成。为实现100 GHz的通道间隔，需借助硅光子学代工厂和光刻工艺的先进能力，在硅片上精确且可重复地形成光栅图案。此外，由于激光器在晶圆级制造，每片晶圆可集成数百个器件，从而实现大规模可扩展生产。

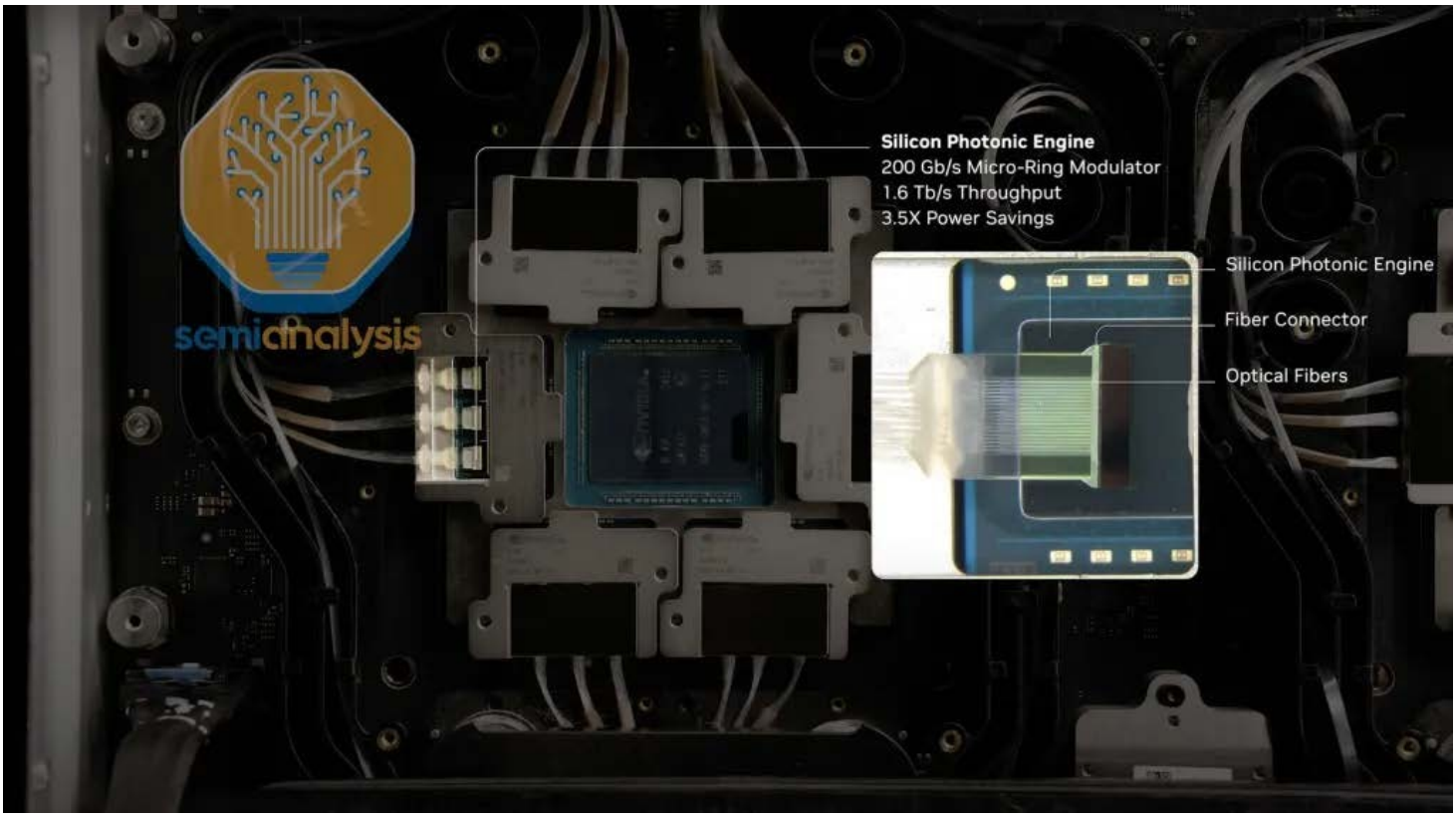
Scintil解决方案的核心优势在于功率效率。该方案可在单芯片上生成并复用多色（8或16色），而采用多路激光器与组合分路器的方案需使用超高功率激光器才能达到目标复用功率密度。 Scintil方案不仅提供卓越的功率效率和更高的带宽密度，还使每个传输比特所需能量减半。相比之下，现有共封装解决方案（包括英伟达当前用于Q3450 CPO交换机的方案）采用单波长高调制速率技术，而非多波长低调制速率方案。

第五部分：英伟达的CPO供应链

前文已阐述关键组件在CPO系统中的作用，本节将聚焦Nvidia供应商链中的具体企业，重点**分析激光源、ELS模块、光纤耦合器（FAU）、FAU点光工具、FAU组装、混光盒、MPO连接器、MT插芯、光纤及电光测试等**关键组件的BOM成本构成。

光学引擎

英伟达的X800-Q3450 CPO交换机具备115.2T总吞吐量，专为横向扩展网络设计。初始版本将采用72个光引擎，每个运行速率为1.6Tbit/s；后续版本预计将过渡至36个光引擎，每个运行速率为3.2Tbit/s，单价约1000美元（含FAU）。因此光引擎总物料成本约为3.5-4万美元（3.2T版本）。



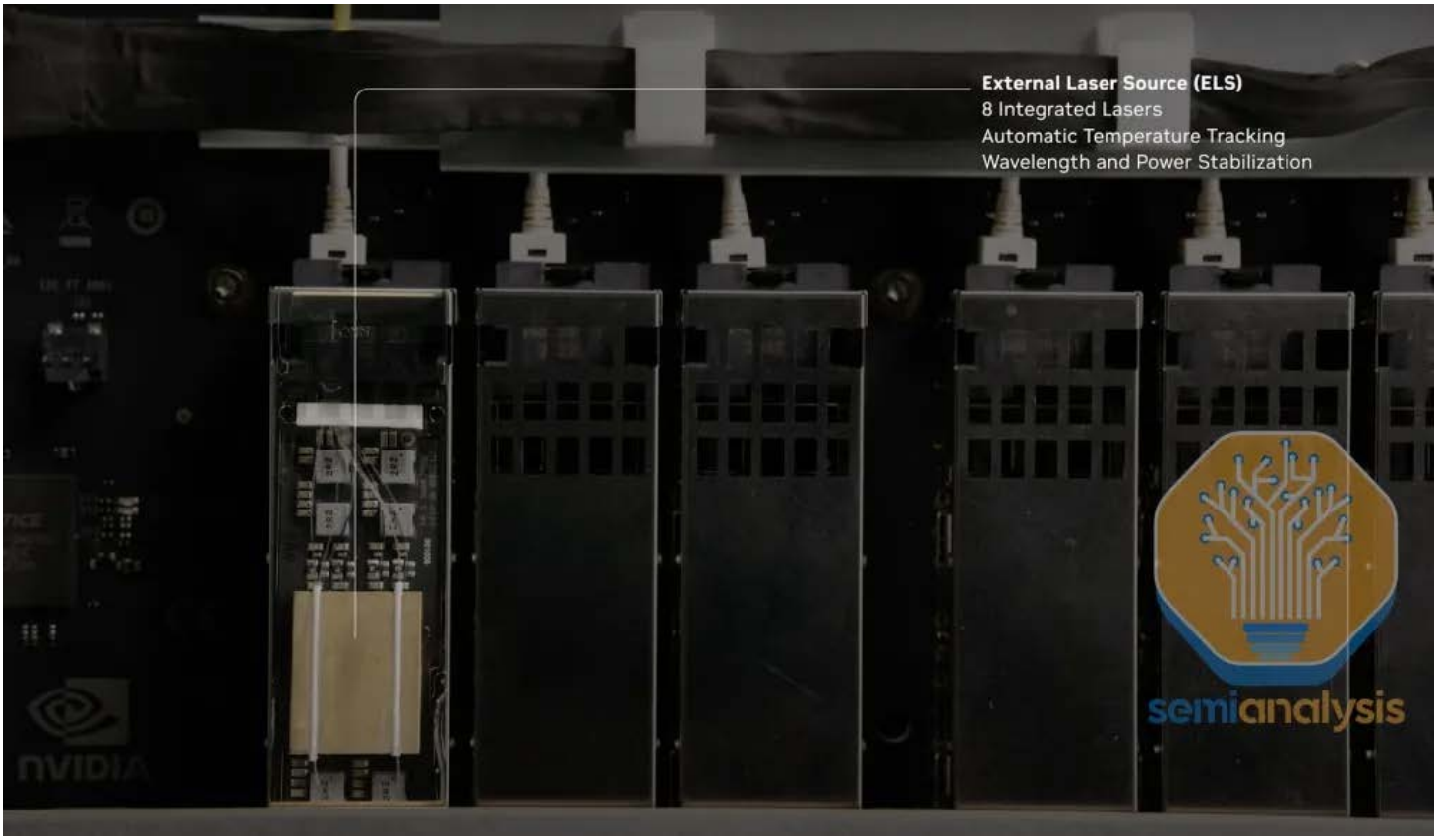
来源：英伟达

外部激光源（ELS）

Nvidia X800-Q3450 CPO交换机采用18个ELS模块作为激光源，每个模块内含8颗连续波（CW）DFB激光芯片。CPO系统需使用

相对更高功率的激光光源，每个CWDFB芯片可输出约350mW功率。

具备连续波激光器单元生产能力的主要行业参与者包括博通（美国）、古河（日本）、Lumentum（美国）、Coherent（美国）、元捷（中国）和世佳（中国）。 Lumentum、Coherent、古河电气和博通通常定价高于中国供应商（元捷和世佳）。 我们预计Lumentum将成为英伟达首批CPO交换机出货的独家供应商，Coherent可能于2026年底成为第二供应商。中国制造商未来或将迎来机遇，因连续波激光器源总体上已趋于标准化和商品化，但构建CPO应用所需的高功率激光器源仍存在一定技术壁垒。



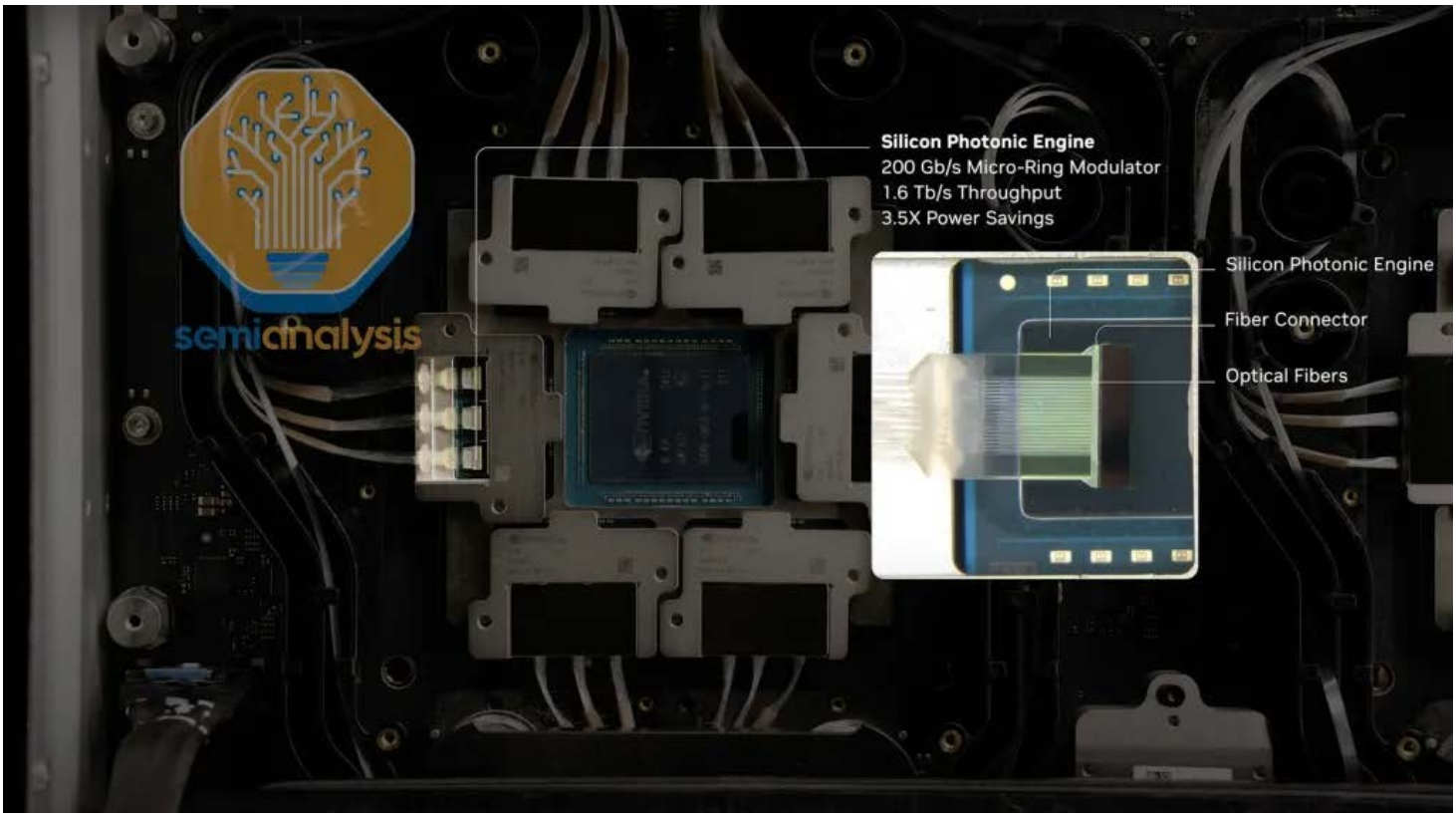
来源：英伟达

光纤连接单元（FAUs）

光纤耦合器（FAU）是连接光纤与光引擎的关键被动元件。优质的FAU及其精确对准对确保最佳光学性能至关重要。当前FAU组装过程面临的挑战在于：用于测量耦合损耗的测试设备尚未实现全自动化。因此测试流程仍高度依赖人工操作，这

导致整体生产速度减缓且成本更高。康宁公司估算，Spectrum X CPO系统中每组FAU的测试平均耗时10-15分钟。

除材料和组件成本外，人工成本在光纤耦合器单元（FAU）成本中占比很大——熟练工人对于确保FAU单元中光纤的高质量组装和对准至关重要。 每台X800-Q3450上的1.6T OE均配备含20根光纤的FAU：8根发射光纤、8根接收光纤及4根外部激光器光纤。每套系统总计1,440根光纤，其中1,152根用于发射/接收。



来源：英伟达

FAU领域的龙头企业包括TFC光电（300394.SH）、森科（9069.JP）和富智康（3363.TW）。 TFC极有可能为X800-Q3450 CPO交换机供应FAU，而森科则被视为Spectrum X CPO和博通Tomahawk 6 CPO系统的最可能供应商。与此同时，富智康则可能更侧重于英伟达的大型CPO解决方案。

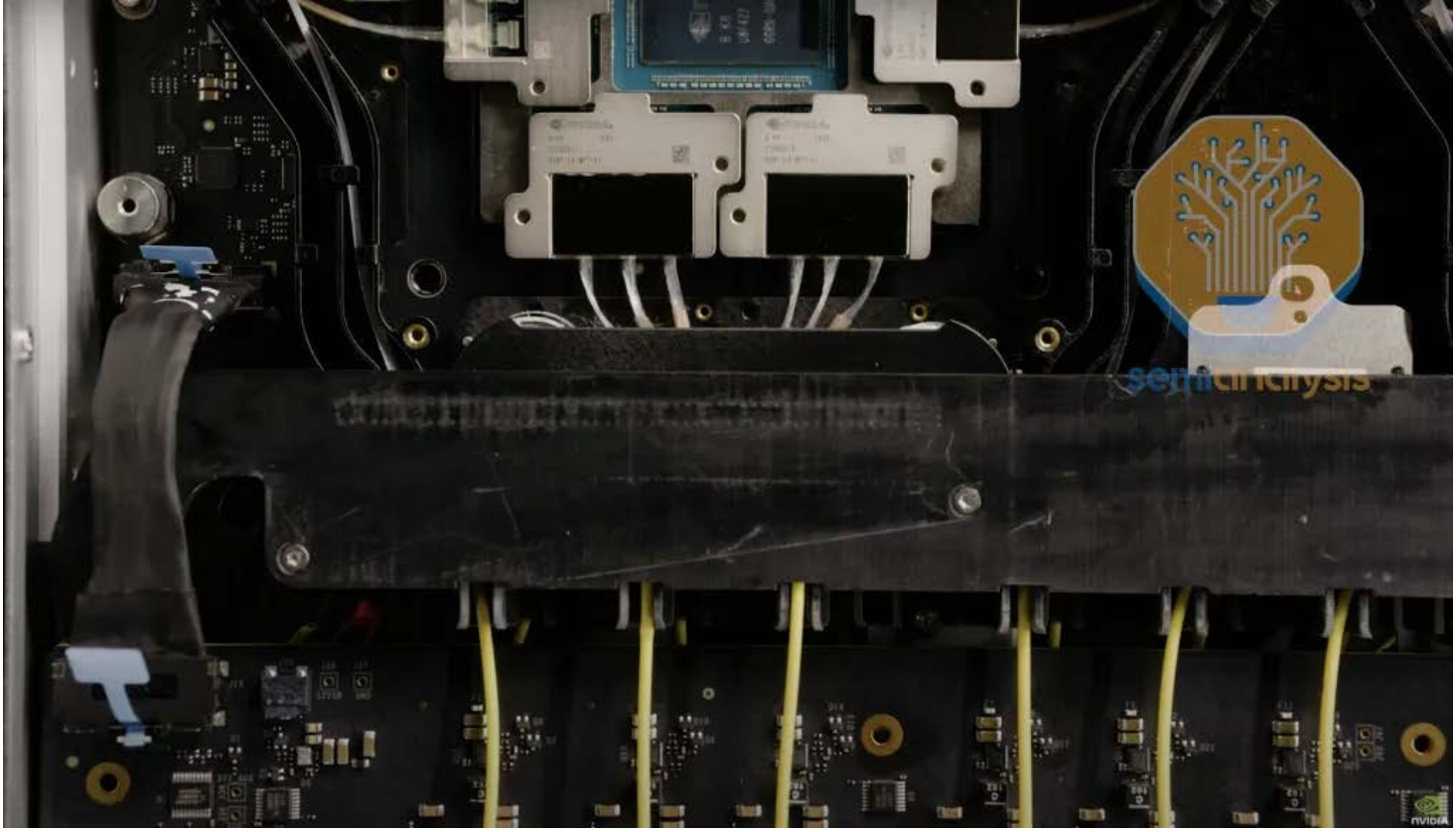
TFC的核心竞争力在于其强大的制造能力以及在中国拥有庞大且成本效益高的技术劳动力资源，正如前文所述，这是维持企业竞争力的关键因素。此外，TFC约三年前开始与英伟达在CPO设计领域展开合作，这一早期伙伴关系现已初见成效，预计TFC将在英伟达当前推进的CPO部署计划中发挥关键作用。

森科拥有其标志性的SEAT（森科弹性平均技术）平台，为CPO系统提供可拆卸的FAU解决方案。该公司正与GFS在边缘耦合技术领域展开深度合作，将森科反射镜直接集成到晶圆沟槽中。这种方法实现了晶圆级边缘耦合测试，而该技术此前一直是光栅耦合的关键优势所在。

该领域其他知名企业包括住友及先进光纤资源公司（AFR）等。据信AFR与博通供应链存在紧密合作关系。

SiPho芯片采用微米级波导通道，每个通道都需要与系统进出光束进行精确对准。因此，在制造过程中需要使用精度极高的耦合设备。

FiconTEC（德国）目前在提供高精度耦合机方面处于行业领先地位。其设备单价可达30万美元以上，但凭借卓越的精度仍备受客户追捧。 与此同时，台湾全环科技也提供自动化光纤连接设备。该公司投入约10%的员工力量研发该领域产品，预计其耦合设备营收将从2026年起实现增长。 最后，台湾GMT Global公司设计了FAU键合、对准和检测设备，通过校正光波长确保光纤耦合器（FAU）的精准传输。该公司计划将设备定价压低至同类日本设备价格的75%左右（日本设备通常售价20-25万美元），预计2025年第四季度将实现批量生产。



来源：英伟达

光纤切换盒

在英伟达X800-Q3450 CPO交换机中，超过1000根光纤从光端机（OEs）中引出，因此需要一个光纤交叉连接器来整理这些光纤并将其路由至目标位置。传统上，这种交叉连接器的对准过程是手动完成的，技术人员需要物理调整光纤位置。然而，一些领先企业已开发出自动化设备，能够以更高精度和效率完成对准操作。

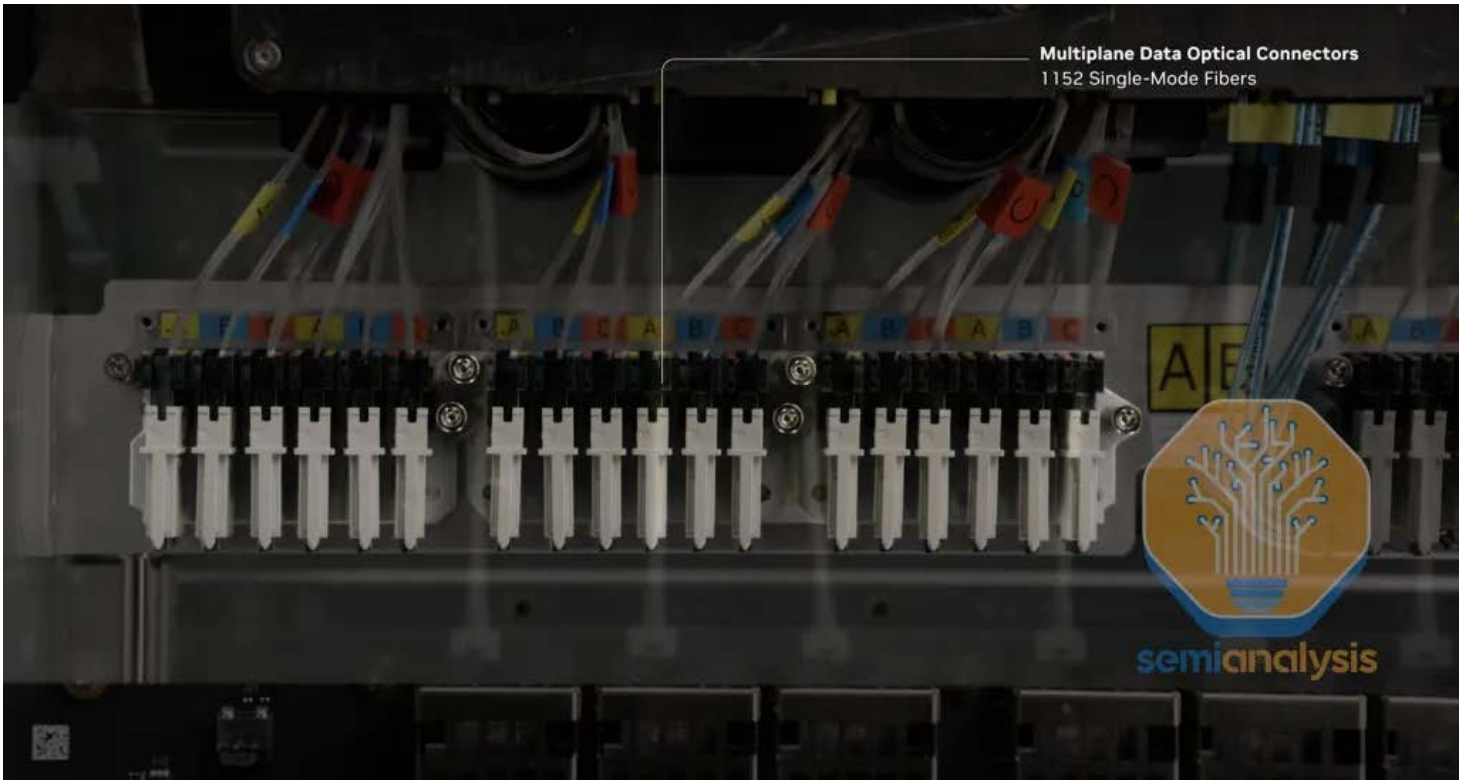
光纤切换盒的价格通常与其管理的光纤数量挂钩。例如，T&S通信公司销售的48芯切换盒售价约150美元，300芯版本约1000美元，500芯型号则约1600美元。 对于配备数千根光纤的X800-Q3450芯片，英伟达采购混排盒的成本将超过3000美元。混排盒的主要物料清单（BOM）组件包括MT插芯和光纤。

T&S通信（300570.SH）是光纤切换盒行业的龙头企业。该公司研发了用于在切换盒内对齐光纤的自动化设备，相关技术已获得专利保护，竞争对手若要绕过这些技术可能需要耗费时间和额外成本。T&S的主要客户是康宁公司，双方常协同合作为客户提供服务。 例如康宁协助客户（英伟达、博通等）设计其CPO解决方案的光纤网络，并将洗盒器制造业务外包

部分内容转至

莫仕是另一家知名企业，同样较早进入市场。但因生产效率较低，其产品价格通常比泰盛高出约20%

。



来源：英伟达

MPO连接器

在交换盒外围设有MPO连接器，用于将交换盒内部的光纤连接至外部端口。 光纤MPO电缆可插入该端口，实现交换机与远端交换机或远端网卡的连接。MPO连接器的制造工艺主要包括注塑成型、真空灌胶以及单元内部MT插芯的光纤穿线。X-800 Q3450 CPO交换机需配备144个MPO连接器

。

能够生产MPO连接器的厂商包括美国康泰克（美国）、泰盛通信（中国）、森科（日本）、博达（中国）和奥普特（中国）。这些厂商还需与光纤网络合同制造商（如康宁）合作，将自身组件纳入整体网络设计方案。

MT套筒

MT套筒是光纤接头盒、光纤切换盒及MPO连接器中的关键组件，用于将多根光纤平行对准。许多能够生产MT接头的公司包括美国康尼康（美国）、T&S（中国）

森科（日本）、福岛（日本）、FOCI（台湾）、住友（日本）和TFC（中国）。MT插芯的制造本身并不特别困难，但要达到所需的精度和坚固性仍需大量工程投入。各公司在模具制造能力上展开竞争，以生产出能在光纤连接过程中最大限度降低插入损耗的插芯。

在这些企业中，美国康尼克公司同样拥有逾三十年的技术研发经验，预计将成为英伟达Q3450 CPO系统的主要供应商之一。福岛公司则具备强大的模具设计与制造能力，能够以具有竞争力的价格生产高质量的MT插针。 与此同时，福晟、泰富和泰盛主要为自家FAU和切换盒生产MT插芯——其逻辑在于尽可能实现垂直整合，以提升质量控制和成本效益。

115.2T Nvidia X800-Q3450 CPO Switch Cost and Power (Current Pricing)

Item	Unit Cost	Quantity	Extended Price	BOM (%)	Power (W)	Extended Power (W)
Switch Chip		4			400.0	1,600
Optical Engines		72			7.0	504
FAUs (included in OE)		72			0.0	0
External Laser Source (ELS)		18			8.0	144
MPO connectors and jumper cables		144			0.0	0
Shuffle box		1			0.0	0
Fibers (per meter)		2000			0.0	0
Others						1,300
Total			**$70,640**	**100%**		**3,548**
Gross Margin			$134,216	60%		
Selling Price			**$176,600**			
3y Service and Warranty			$28,256			
Selling Price with Service			**$204,856**			

Bill of Material estimates are based on current scale of production and could improve as production scales up

来源：SemiAnalysis AI网络模型

制造、组装与测试流程

光电元件制造：如前所述，台积电将在PICs、EICs的制造及CPO系统集成中发挥关键作用。其COUPE平台似乎是下一代CPO终端的首选解决方案。格罗方德和Tower是具备强大硅光子能力的代工厂，但其

缺乏尖端CMOS技术和先进封装工艺，限制了其提供未来更高带宽光电元件的能力。

先进封装：外包半导体封装与测试（OSAT）供应商将聚焦后端工艺，包括光学元件封装、光学元件测试及系统级封装（激光器与耦合器集成测试）。

ASE/SPIL（3711.TW）、Amkor（$AMKR）和顺兴（6451.TW）是此类解决方案的主要供应商。其中，ASE作为英伟达供应链中的关键供应商脱颖而出，包括参与未来的Rubin机架式CPO系统，而顺兴则与博通保持着密切合作关系。

其他值得关注的企业包括Fabrinet（$FN）、TFC光电（300394.SH）和富士康（2354.TW）。 Fabrinet长期担任英伟达内部光模块单元的组装商，目前正积极拓展光电封装、测试及整机系统组装能力。该公司与Micas、富士康共同被视为博通CPO系统组装业务的潜在承接方之一。

过去三四年间，TFC一直与英伟达在芯片级封装（CPO）设计领域保持紧密合作，并将成为英伟达可调式加速器单元（FAU）的主要供应商。该公司持续在中国苏州投资建设先进封装设施，彰显其在CPO供应链中谋求更大份额的雄心。

电光（E/O）测试设备：在测试过程中，上述服务供应商使用电光（E/O）测试工具来确保系统可靠性。然而该行业仍在发展中，尚未就光子引擎的标准化测试方法达成完全共识。各供应商正开发各种解决方案，以在这个新兴领域中占据一席之地。

CPO供应链中的其他关键设备供应商包括是德科技、Ficontec、泰瑞达、爱德万测试、FormFactor、Chroma、安立和Multilane。其中是德科技是该领域的重要参与者——他们以提供高品质（以及高价格！）的高速测试设备而闻名。 例如该公司近期发布了两款用于1.6T光收发器测试的新型示波器。凭借其市场地位，他们将充分受益于CPO技术的发展趋势。

菲康在光子学测试领域根基深厚，现正积极拓展电气测试能力。其核心优势之一在于晶圆级测试技术。

光子测试——该技术提升了传统f ...

光子测试流程。例如，该公司近期推出了一款适用于光子集成器件（PIC）的新型高吞吐量晶圆级测试设备，该设备兼容现有半自动测试设备（ATE）架构。据宣称，该设备是业内首款双面晶圆测试机。除测试设备外，Ficontec还提供光纤耦合器（FAU）组装与耦合设备，我们将在下一节中对此进行更深入的探讨。

泰瑞达同样是该领域的重要参与者，在电气测试领域拥有深厚的历史积淀，但该公司对进军光子测试领域展现出"极大决心"。例如，该公司近期还收购了一家专注于封装光器件测试的初创企业，此举彰显其战略性布局，旨在全面拓展其在封装光器件测试（CPO）领域的能力。

Chroma一直为3D传感和光通信领域使用的激光二极管提供光子测试设备，并有望凭借这一专业技术进军CPO领域。然而，该公司当前的创新步伐落后于部分竞争对手。

Nvidia CPO Supply Chain Market Map					
Laser Source	ELS Module	FAU	FAU Align Tools	FAU Assembly	Shuffle box
LITE	**TFC**	**TFC**	**finconTEC**	**Fabrinet**	**T&S**
COHR	**O-Net**	**Senko**	All Ring	**FOCI**	**Browave**
AVGO	Innolight	FOCI		Foxconn	Molex
Furukawa	Eoptolink	Sumitomo		ASE	
Yuanjie					
MPO Connector	MT Ferrule	Fibers	Foundries	OSAT/ Assembly	E/O Testing
US Conec	T&S	**Corning**	**TSMC**	**Amkor**	**ficonTEC**
Senko	Senko	Sumimoto	Tower	**ASE/ SPIL**	**Keysight**
T&S	US Conec	Nittobo	GF	Fabrinet	Teradyne
Browave	Fukushima			TFC	Formfactor
	FOCI				Chroma
	Sumitomo				Multilane

Vendors highlighted green are key players in the sectors & confirmed to be part of Nvidia's CPO supply chain
For more information on market shares and competitive dynamics, contact sales@semianalysis.com.

来源：SemiAnalysis AI网络模型

向读者推荐**SemiAnalysis**

架起全球最重要产业——半导体与商业之间的桥梁。

推荐

51个赞 · 7次转发

上一篇

客座文章

帕特里克·周

社交网络 @SemiAnalysis

订阅 Patrick

关于此帖的讨论

评论 **重新堆叠**

发表评论...

乔治·约翰逊 1小时前

你对POET Technologies及其作为ELS供应商的角色有何看法？是否了解他们是否仍是Celestial AI的供应商？或者在Marvell收购后，Celestial AI是否会转向更成熟的供应商采购ELS以支持其业务扩张？

赞 回复 分享