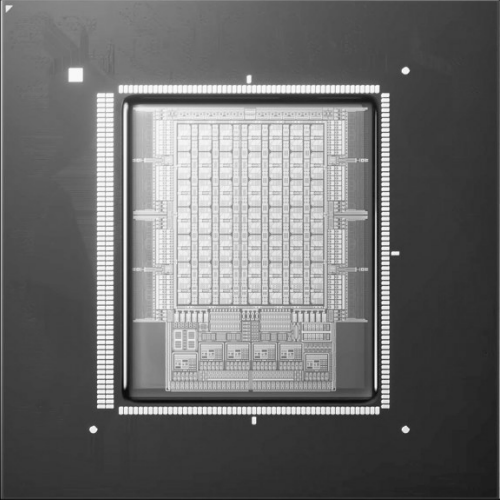


CPU 重回焦点位



88 NVIDIA Custom Olympus Core With Spatial Multithreading

PCIe Gen 6 CXL 3.1

164 MB L3 Cache

3.4 TB/s Core-to-Core Bisection Bandwidth

Up to 1.5 TB LPDDR5X Memory

NVLink-C2C 1.8 TB/s Coherent CPU-CPU and CPU-GPU Interface

图为英伟达发布的Vera CPU

本报记者 姬晓婷

为期三天的台北国际电脑展(COMPUTEX)落幕。在智能体应用的推动下,今年的COMPUTEX受到了比以往更多的关注。而在这场活动上唱“重头戏”的几大芯片巨头,不约而同地把CPU托举到了焦点位。CPU,成为智能体时代芯片厂商争夺的关键。

四家公司,同一个信号

6月1日,英伟达GTC台北大会召开。会上,英伟达创始人兼首席执行官黄仁勋重磅发布了Vera CPU——首款专为智能体打造的CPU。黄仁勋说:“AI智能体将成为计算资源的最大用户,而Vera正是专为在超大规模基础设施中运行智能体AI而生。”

同一天,高通CEO安蒙在COMPU-TEX开幕主题演讲中,将2026年定义为“智

能体之年”。他明确指出:“你需要一个非常强大且高能效的CPU来负责任务的编排规划。”

6月2日,英特尔CEO陈立武在COM-PUTEX主题演讲中发布了至强6+处理器。这是基于Intel 18A制程的首款数据中心CPU,搭载288颗能效核。

6月3日,Arm宣布Oracle Cloud Infra-

AI智能体将成为计算资源的最大用户,大幅提升数据中心对CPU的需求。

structure加入其AGI CPU生态系统。Arm云AI事业部执行副总裁Mohamed Awad在博客中提到:在3月24日举办的Arm Every-where活动上,Arm曾指出,智能体AI将大幅提升数据中心对CPU的需求。仅两个月后,市场发展速度已超出Arm的预期。

短短几天,四家全球芯片头部企业释放出同一个信号:CPU的作用不容小觑。

智能体时代,CPU是GPU利用率的瓶颈,直接影响Token吞吐量、延迟和用户体验。

径上,CPU如果效率不够高,将阻止Agent执行下一步。”

因此,智能体时代,CPU是GPU利用率的瓶颈,直接影响Token吞吐量、延迟和用户体验。所以,智能体CPU要尽可能降低延迟,尽可能具有交互性。

而实现这一目标最简单的做法便是提高CPU在计算系统中的占比。Creative Strate-gies首席执行官兼首席分析师Ben Bajarin指出:“在模型训练时代,AI部署中CPU与GPU的配比大致接近1:4,而智能体推理则将这一比例改写为接近1:1(甚至GPU占比更低)。”

英特尔至强6+是基于Intel 18A打造的首颗数据中心CPU,可用于机架级AI基础设施。

SN-50 RDU(可重构数据流单元)紧密结合,专门用于AI推理。

英特尔还与富士康、西门子、日立、Echo Neurotechnologies和Greenstone Bio-sciences等行业领导者展开战略合作,聚焦提供基于英特尔处理器和定制芯片的垂直行业整合解决方案。

高通的打法与前三家不同。安蒙在演讲中强调的是“端云协同”——智能体AI不是只在云端运行,也不是只在终端运行,而是两者协同。他举了一个例子:在编程场景中,采用分布式智能体AI架构,可节省约140万Token,成本降低60%。

为此,高通在COMPUTEX上发布了数据中心业务新品牌Dragonfly(高通飞龙),正式进军数据中心市场。安蒙表示,高通的产品组合已覆盖“计算连续体”的各个层级——从可连接智能体的最小的可穿戴设备,到高性能的数据中心。

Token消耗已实现约100倍的量级增长,这意味着,算力基础设施未来的市场增长空间不可限量。

其一,高效能比。数据中心的电费 and 散热是硬约束。在同等性能下,功耗越低的CPU越有竞争力。

其二,算力与GPU的协同能力。在智能体时代,需要“CPU+GPU”异构计算的协同配合,两者之间的数据传输效率至关重要。

(上接第1版)即“用数百乃至数千条低速并行光通道替代单一超高速链路,在采用更简单、廉价、低速组件的同时实现更高总带宽。”据介绍,该方案的核心能力在于将数百个光发射器集成至服务器数据处理器和存储组件近端,其带来的优势包括:通过配置多重冗余通道实现单个故障发射器无感失效;每个发射器处于相对较低的开关频率(典型值约1GHz),较超高频激光发射器显著降低功耗并减少废热;同时该架构具有原生并行特性,无须复杂且高成本的串行化与解串行化方案。

Dominik Schulten指出,当前“低速宽带”方案在数据中心领域尚属未经验证的技术路径,需要与数据通信设备制造商展开深度合作,具体包括集成适配高频数据发射器的驱动电路、优化微型封装设计方案以及确保与商用光连接器、线缆及光纤的互操作性。

总体而言,MicroLED在能效上取得了“颠覆性突破”,解决了CPO长期面临的“功耗墙”问题,但单通道速率过低被普遍认为当前最大的技术瓶颈。正如何军所言,单通道速率过低意味着实现1.6T总带宽需要数百个并行通道,这对巨量光纤的封装精度、成本和良率构成巨大挑战。

锁定AI数据中心短距互联 规模化应用或在2028年前后

尽管如此,对于MicroLED光互联CPO最先落地的应用场景,产业链企业还是高度一致地指向了AI数据中心内部的超短距高速互连。

在雷曼光电技术研发中心高级总监屠孟龙看来,最先落地的场景是“AI数据中心机柜内/板间/芯片间≤50米超短距高速光互联”。何军则进一步细化为“共封装光学(CPO)模块和主动式光缆(AOC)”,传输距离锁定在10米以内的机柜内部或相邻机柜间互连。

“MicroLED的最大优势是极低功耗(低于1pJ/bit)和高热稳定性(耐125℃),而AI数据中心正面临严峻的‘功耗墙’挑战,传统铜缆在高带宽下传输距离不足2米且发热严重。”屠孟龙认为,在这一场景中,MicroLED“单通道速率较低”的短板,可通过“宽而慢”的数百通道并行架构来弥补,实现总带宽1.6T以上。

关于规模化应用的时间表,各家企业给出了较为一致的预判,但态度谨慎。

何军向记者表示,综合多方机构研判:2026年是MicroLED光互联CPO产品落地元年(已有多家厂商推出评估套件和样品),2027—2028年将逐步进入量产爬坡阶段,规模化应用预计在2028年前后实现。需要说明的是,这一进程高度依赖于发光芯片性能提升和光学耦合精度等工程瓶颈的突破进度。

屠孟龙分析,从客户接受度上,预计2027年规模化导入,2028年全面接受,成为AI数据中心短距互联主流。从成本竞争力上,预计2027年是成本拐点(持平),2028—2029年形成显著优势(综合成本低30%以上)。黄少华则认为,“预计还需要3~5年才能看到有竞争优势的情况”。

此外,晓明科技显示产品线副总裁陈黎喧补充道,MicroLED产品是一种互补的光通信方案,在近程通信中有其优势,但可能仍需要2~3年进行产业链的方案打通。

值得注意的是,并非所有企业都聚焦于难度最高的CPO。黄少华坦言:“从目前项目的成熟度来看,CPO难度还是相当大,光模块是相对可能的落地应用。”

从“单点突破”到“系统协同” 光学耦合效率是最大技术难点

虽然前景可期,但MicroLED CPO要真正实现商业化,仍面临技术与非技术两大维度的严峻挑战,产业链在多个环节存在明显短板。

在技术挑战方面,光学耦合效率与芯片良率是核心卡点。何军指出,最大难点在于光学耦合效率。MicroLED发光角度达120~150度,呈朗伯体发散,而光纤接收角极窄,耦合精度需控制在±1~2微米内,否则效率损失可达30%以上。其次是芯片良率,光通信对可靠性要求远超显示,芯片良率不足直接影响通道数,制约总带宽。

屠孟龙则总结道,巨量转移+异质集成+光电耦合的三重精度与良率瓶颈,是当前最核心卡点。此外,艾迈斯欧司朗新技术高级总监Norwin von Malm提到,当前MicroLED CPO面临的最大的技术挑战在于,“低速宽带”方案在数据中心领域尚属未经验证的技术路径。

非技术挑战更为关键,且短期内难以解决。首先是成本,当前MicroLED CPO方案成本是铜缆的5~10倍。其次是产业链配套不成熟。

陈黎喧向记者表示,目前全产业链成熟度还不高,虽然经过显示行业的耕耘,MicroLED技术在AR眼镜等领域已经得到了商业化应用,但移植到光通信领域,还需要克服诸如发光角、耦合效率、信噪

比、信号驱动等多方面问题,需要产业链上下游共同努力。黄少华也强调:“整个供应链还不具备完整的方案,挑战还是相对较高。”

此外,生态协同的缺失是普遍共识。何军认为,芯片、CMOS驱动、封装、系统厂商需深度协同,但目前“各环节标准未定、磨合周期长”。当然,针对MicroLED光互联的专用测试设备也很缺乏。

值得关注的是,产业链的明显短板集中在光学耦合组件、多芯光纤配套、先进封装工艺等方面,而这恰恰是部分企业的发力点。

记者了解到,芯视元作为CMOS驱动背板供应商,将其在显示驱动领域积累的高密度阵列、低功耗设计能力向光通信领域做核心延伸。该公司正在重构驱动架构,目前已完成8英寸工艺平台的验证,相关驱动芯片已向多家MicroLED光源厂商送样;下一步将联合光源与封装伙伴,共同推进标准制定与量产导入。

“从市场端看,AI数据中心对CPO的功耗要求极为严苛。MicroLED方案要求每个发光像素都配套独立驱动单元,这正是CMOS背板的核心优势所在。”何军告诉记者,“我们可以提供每平方毫米数千个驱动通道的高密度阵列,同时将驱动功耗控制在pJ/bit级别,直接命中功耗痛点。”

艾迈斯欧司朗积极推动MicroLED技术在光学互连领域的应用转型,具体包括集成适配高频数据发射器的驱动电路,优化微型封装设计方案以及确保与商用光连接器、线缆及光纤的互操作性。该公司同时具备MicroLED与光电二极管(光接收器)的量产能力,能够在全集成光学数据传输系统的开发中发挥作用。

雷曼光电依托自身20多年的MicroLED封装研发经验,十分看重AI算力需求带来的新增长空间,积极布局光互联CPO,目前在无衬底芯片转移和封装上取得良好进展,下一步将与高等院校团队对接最新的技术方案。三安光电则聚焦在自身位于产业链中的发端角色,为客户提供高速的MicroLED光源以及迭代效率和降低发散角来提高光的耦合效率。

替代铜缆方向明确 与硅光CPO互补共存

MicroLED CPO对传统铜缆的替代关系已较为明确,传统铜缆在800G/1.6T高带宽下功耗超过10pJ/bit,传输距离不足2米,而MicroLED CPO可将功耗降至铜缆的5%、至1.6W左右,在机柜内极短距离场景下优势显著。

那么,MicroLED CPO与硅光CPO之间是什么关系?受访企业普遍认为,MicroLED CPO与硅光CPO更多是互补共存。何军形象地比喻道,两者将形成“短距MicroLED+长距硅光”的协同格局:MicroLED专注10米以内的芯片间、板间、机柜内短距传输,主打极致能效;而硅光CPO则覆盖10米以上及长距互联,技术成熟,传输距离远。屠孟龙也持相近观点,认为二者是“短距vs中长距”的互补关系,将长期共存。

放眼当下,今年以来资本市场对MicroLED CPO概念关注度极高。那么,这项技术真正改变数据中心互联格局的机会有多大?是否存在过热风险?

“MicroLED CPO真正改变数据中心互联格局的机会确实存在,且不容忽视,但行业过热背后也需保持清醒。”何军认为,机会集中在10米以内的机柜内/板间互联场景。在800G/1.6T及以上带宽下,铜缆功耗已逼近物理极限,而MicroLED CPO可将功耗降至铜缆的5%~10%,这是刚需级优势。一旦工程瓶颈突破,有望在AI集群等功耗敏感场景占据20%~30%的短距互联份额。

但何军同时指出,最大的风险不是技术路径本身,而是产业链成熟的速度能否跑赢铜缆迭代与硅光降本。如果铜缆在1.6T/3.2T时代通过新材料延缓功耗增长,或硅光大幅降低短距成本,MicroLED的市场窗口可能被压缩。

屠孟龙则判断,MicroLED CPO大概率将会改变数据中心≤50米高速互联格局,成为AI时代主流方案。但最大的风险仍是技术路径不确定性,包括芯片性能与良率、巨量转移、耦合效率等“生死线”,决定其能否从“实验室”走向“量产”。而另一个风险则是市场接受度与生态决定商业化节奏,2027年前难有大规模业绩兑现。他提醒资本市场:“当前处于研发验证期,若技术突破不及预期,估值将大幅回调。”

风口之下,MicroLED光互联CPO凭借其在功耗上的颠覆性优势,为AI数据中心的“功耗墙”困境提供了一个极具吸引力的解决方案。然而,从实验室到数据中心机柜,它仍需跨越单通道速率、光学耦合、产业链协同和成本等多道鸿沟。机会是结构性的,但商业化节奏取决于整个产业链能否在2027—2028年形成有效闭环。在此之前,资本市场在热情拥抱的同时或许需要更多的理性与耐心。