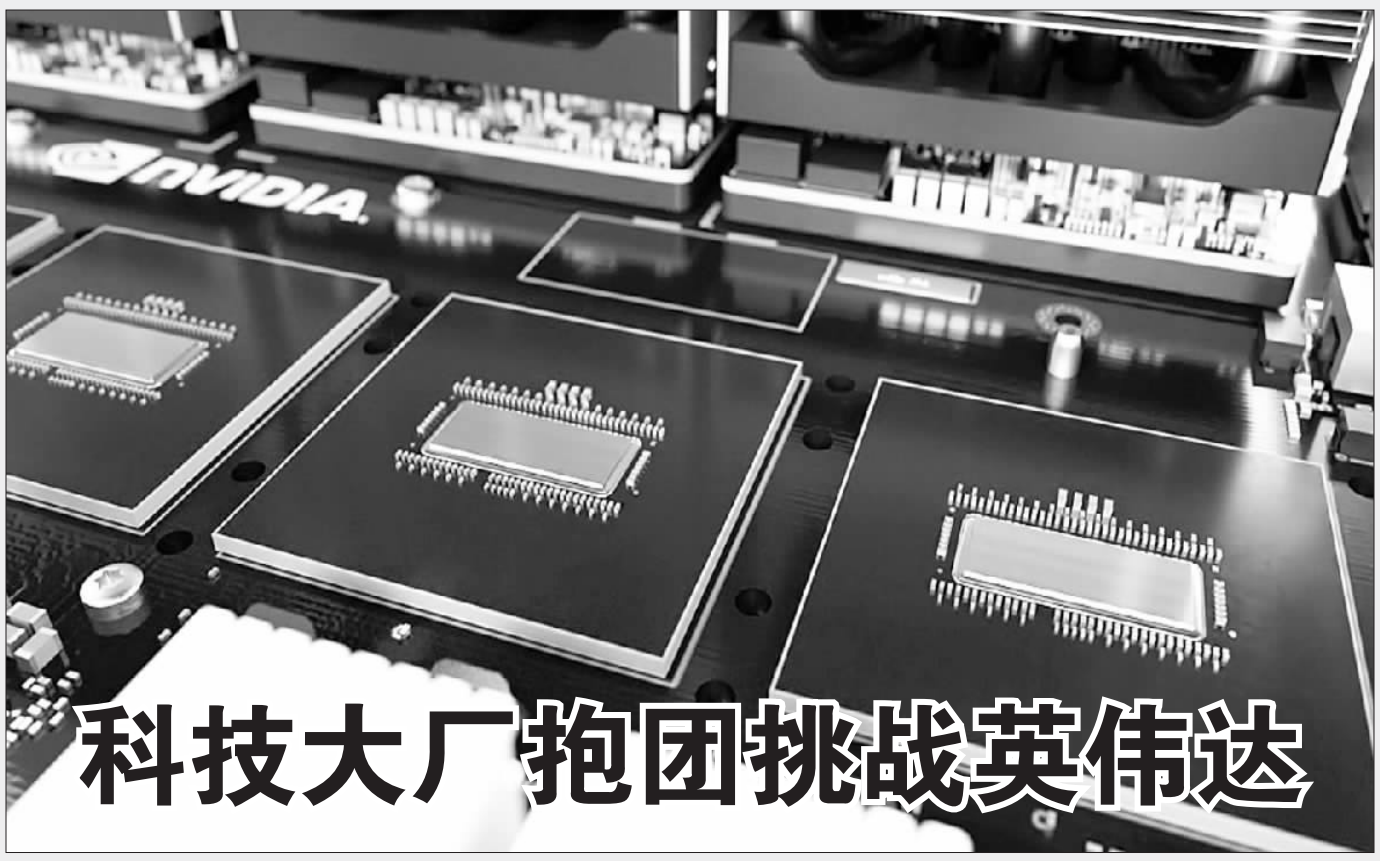

科技大厂抱团挑战英伟达

本报记者 许子皓

近日，英特尔、谷歌、微软、Meta等八家科技巨头联合宣布建立“超加速器连接推广小组”（UALink Promoter Group，简称UALink），意在制定行业标准，指导数据中心AI加速器芯片之间连接组件的发展。在业内专家看来，此举背后的真正深意，在于挑战英伟达的NVLink技术和其在AI加速器领域的霸主地位。

来势汹汹且有备而来

UALink的成员包括英特尔、谷歌、微软、Meta、AMD、慧与、博通和思科。其中，英特尔、AMD和博通是全球领先的芯片制造商；微软和谷歌是云服务提供商，此二者和身为社交媒体巨头的Meta都在自研芯片；慧与是专注企业业务的云和设备公司；思科是网络解决方案供应商。

在业内人士看来，八家公司明显有备而来，意欲通过全产业链联合，对抗英伟达在AI加速器领域的统治。

UALink在宣布成立的同时，就已开发出了一项新的行业标准：致力于推进数据中心内连接的大规模AI系统的高速和低延迟通信。据悉UALink计划在今年第三季度成立一个官方行业联盟，并向加入该联盟的公司提供UALink 1.0技术。该技术将能够在单个集群中连接多达1024个AI加速器，通过将大量加速器连接在一起，共同完成大规模计算任务。在第四季度，UALink还将发布其互联技术的第一轮迭代版本UALink 1.1。

为了在AI领域不落人后，微软、谷歌、Meta等企业目前已经花费了数十亿美元购买英伟达的GPU。由于只要使用英伟达的GPU就必须使用英伟达专有的NVLink连接技术，这让各大企业非常被动。因此UALink的组建，被认为是各大企业想联合起来摆脱对英伟达的依赖。

AMD数据中心解决方案总经理Forrest Norrod表示，行业需要一种标准，允许创新不受任何一家公司的束缚，快速推进。博通数据中心解决方案事业部副总裁Jas Tremblay表示，开放的生态系统协作对于通过各种高速和低延迟解决方案实现网络扩展至关重要。

据悉，首批UALink产品将在未来几年内推出。

针锋相对 剑指英伟达

UALink“步步为营”的操作，剑指英伟达的NVLink技术。NVLink是英伟达在2014年发布的一种总线及其通信协议，采用点对点结构、串行传输，用于中央处理器（CPU）与图形处理器（GPU）之间的连接，也可用于多个图形处理器之间的相互连接。其特点包括高速、低延迟和高带宽，为多个GPU之间提供了直接连接，从而显著提升系统的性能和可扩展性。英伟达凭借NVLink技术约占全球AI数据中心市场80%~95%的份额，实现了高度垄断。如今，英伟达已经推出了第五代NVLink和NVLink Switch 7.2T。

NVLink的连接协议并不是第一次被挑战。2019年3月英特尔在其Interconnect Day 2019上推出了Compute Express Link（CXL）协议，作为一种开放性互联协议，CXL旨在让CPU与GPU、FPGA或其他加速器之间实现高速高效的互联，满足高性能异构计算的要求。但CXL目前尚未达到NVLink在GPU间直接连接的性能水平，尤其是在带宽和延迟方面。其他像PCIe、QPI和InfiniBand都有各自的优势，但在带宽、延迟以及与GPU间连接方面不及NVLink。

英伟达2024第一财季财报显示，公司期内实现营收为260.44亿美元，同比上涨262%，远高于市场预期的247亿美元。其中，数据中心业务营收为226亿美元，与上年同期相比增长427%。这些数据证明，英伟达依然拥有市场掌控权。

专家认为，英伟达在技术实力和生态建设上仍然具有明显优势，UALink的建立，短期内对英伟达造不成太大威胁。但英伟达的NVLink并不向行业开放，而UALink是完全开源的，不仅适用于大型企业，也为行业中的所有人提供了一个在规模和创新方面与英伟达竞争的机会。

第一季度PC GPU出货量达7000万颗，同比增长28%

本报讯 市场调查机构Jon Peddie Research近日发布的最新报告显示，2024年第一季度全球基于PC客户端的GPU出货量为7000万颗，同比增长28%，环比下降9.9%。其中台式机显卡同比下降了7%，笔记本电脑显卡同比增长了38%。

Jon Peddie Research表示，未来5年内，独立显卡在PC市场的渗透率预计达到22%，到2026年，GPU总装机量将接近30亿的安装基数。在2024年第一季度中，GPU和PC的整体连接率（包括集成和独立显卡、台式机、笔记本电脑和工作站）为113%，相比上一个季度略有下降，跌幅为0.6%，其中台式机显卡部分下降了14.8%。

在2024年第一季度里，AMD的GPU出货量环比下降13.6%，英特尔的GPU出货量环比下降9.6%，英伟达的GPU出货量环比下降7.7%。从市场份额来看，AMD的整体市场份额相比起上一个季度下降了0.7个百分点，英特尔增长了0.3个百分点，英伟达增长了0.4个百分点。

GPU一直是PC市场的先行指标，因为在PC供应商发货之前就会安装到系统中。Jon Peddie Research表示，大多数厂商预计下一个季度平均下降7.9%。此外，Jon Peddie Research预计英伟达2024年将出货200万颗数据中心GPU。 （微 文）

第一季度全球NAND Flash营收为147.1亿美元，环比增长28.1%

本报讯 近日，TrendForce集邦咨询发布的最新研究显示，2024年第一季度全球NAND Flash市场营收环比增长28.1%，达147.1亿美元。主要受益于AI服务器厂商自今年2月起扩大采用Enterprise SSD，大容量订单开始涌现，以及PC、智能手机客户为应对NAND Flash价格上涨而主动提高库存水平。

从头部厂商的营收及份额排名来看，三星依然稳居第一，营收为54亿美元，环比增长28.6%，市场份额为36.7%，环比增加了0.1个百分点。而三星的增长主要得益于消费级买家持续提高库存水位，以及Enterprise SSD订单开始复苏。尽管消费级产品订单转趋保守，但第二季度受益于Enterprise SSD出货量扩大，NAND Flash合约价持续上涨，三星第二季度营收有望再环比增长20%。

排名第二的是SK集团，其第一季度营收环比增长31.9%，达32.7亿美元，市场份额为22.2%，环比增加了0.6个百分点。而其营收的增长主要受益于智能手机、服务器订单增长，以及收购的Solidigm拥有独特的Floating Gate QLC技术，大容量Enterprise SSD订单动能续强。预计第二季度SK集团出货的NAND Flash位元增长幅度有望优于其他供应商，营收预计也将继续环比增长20%左右。

铠侠（Kioxia）第一季度营收18.22亿美元，环比增长26.3%，市场份额为12.4%，环比减少了0.2个百分点，排名第三。这主要是由于SK海力士第一季度的出货仍受到了此前的减产策略影响，出货位元仅环比增长了7%，但仍受益于NAND Flash均价上涨，带动第一季营收增长。随着第二季度供应的位元逐步上升，在提供客户更具有弹性的议价空间下，铠侠有望进一步扩大Enterprise SSD出货量，第二季度的营收预计环比增长20%。

美光第一季度营收为17.2亿美元，环比大涨51.2%，是前五厂商当中增幅最大的厂商，带动其市场份额也增长到了11.7%，环比增加了1.8个百分点。这也使得美光的营收及市场份额超过西部数据，排名第四。

排名第五的西部数据（Western Digital）第一季度营收为17.1亿美元，环比增幅仅2.4%，市场份额为11.6%，环比减少了3.9个百分点。这主要是由于消费类市场需求自今年2月起明显萎缩，影响到了出货位元。进入第二季度，受限于整体消费性市场仍未回暖，加上PC及智能手机全年展望保守，故西部数据欲加速Enterprise SSD产品开发，以扩大未来成长动能。然而，企业级产品验证时程较长，对带动短期出货动能成长有限，预期西部数据第二季度营收可能持平。

TrendForce表示，观察第二季度的NAND Flash市场趋势，随着PC及智能手机客户的NAND Flash库存水位升高，加上今年消费终端订单增长仍未优于预期，品牌厂商备货转趋保守。与此同时，受惠于大容量Enterprise SSD订单翻倍，带动第二季度NAND Flash产品均价续涨15%，预估第二季度全球NAND Flash营收有机会再环比增长10%。 （吉文）

我国科学家研制出世界首款类脑互补视觉芯片“天眸芯”

本报讯 记者**张心怡** 实习记者**夏冬阳**报道：近日，清华大学精密仪器系类脑计算研究中心团队研制出了世界首款类脑互补视觉芯片“天眸芯”，基于该研究成果的论文《面向开放世界感知具有互补通路的视觉芯片》作为封面文章，登上5月30日的《自然》杂志，标志着我国在类脑计算和类脑感知方向上取得突破。

据悉，该芯片可在极低的带宽和功耗代价下，实现每秒10000帧的高速、10bit的高精度、130dB的高动态范围的视觉信息采集，不仅为智能革命的发展提供了强大技术支持，还为自动驾驶、具身智能等重要应用开辟了新的道路。

视觉感知作为智能无人系统获取信息的核心途径发挥着至关重要的作用，但在面对驾驶中的突发危险、隧道口的剧烈光线变化和夜间强闪光干扰等极端场景时，传统视觉感知芯片由于受到“功耗墙”“带宽墙”的限制，往往出现失真、失效或高延迟的问题，严重影响了系统的稳定性和安全性。

为更好地应对上述问题，清华大学精密仪器系类脑计算研究中心团队聚焦类脑视觉感知芯片技术，提出了一种基于视觉原语的互补双通路类脑视觉感知新范式。论文通信作者、清华大学精密仪器系教授施路平介绍：“该范式借鉴了人类视觉系统的基本原理，将开放世界的视觉信息拆解为基于视觉原语的信息表示，并通过有机组合这些原语，模仿人类视觉系统的特征，形成两条优势互补、信息完备的视觉感知通路。”同时，基于“天眸芯”，团队还自主研发了高性能软件和算法，并在开放环境车载平台上进行了性能验证。在多种极端场景下，该系统实现了低延迟、高性能的实时感知推理，展现了其在智能无人系统领域的巨大应用潜力。

AMD发布全新云端AI加速芯片路线图

本报讯 6月3日，AMD董事长兼CEO苏姿丰在Computex 2024展会的开幕主题演讲中，公布了全新云端AI加速芯片路线图，今年将推出全新Instinct MI325X。同时，AMD还发布了代号为“Strix Point”的第三代AI PC芯片“锐龙AI 300系列”，以及AMD Ryzen 9000系列桌面处理器（Granite Ridge）。

据介绍，MI325X将延续CDNA3构架，采用第四代高带宽内存（HBM）HBM3E，容量大幅提升至288GB，内存带宽也将提升至6TB/s，整体的性能将进一步提升，其他方面的规格则基本保持与MI300X一致，便于客户的产品升级。

苏姿丰指出，MI325X的AI性能提升幅度为AMD史上最大，相较竞品将有1.3倍以上的提升，同时更有性价比优势，MI325X将于今年第四季度开始供货。

另外，AMD还将在2025年推出新一代的MI350系列，该系列芯片将采用3nm制程，基于全新的构架，集成288GB HBM3E内存，并支持FP4/FP6数据格式，推理运算速度较现有MI300系列芯片快35倍。

“锐龙AI 300系列”采用全新的Zen5 PU架构，GPU内核也升级为RDNA3.5架构，NPU也是全新的XDNA2架构，号称是“面向下代AI PC/Copilot+ PC的世界一流处理器”。锐龙AI 300系列首发只有两款型号，“锐龙AI 9 HX 370”和“锐龙AI 9 HX 365”都定位高端市场。

全新发布的AMD Ryzen 9000系列桌面处理器，基于Zen5构架，第一批产品将于2024年7月推出。苏姿丰强调，Ryzen 9000系列是继Ryzen 7000“Raphael”和Ryzen 8000“Hawk Point”系列之后，AM5插槽的第三个系列，将配备两个Zen5小芯片，每个小芯片最多有8个核心，最高16个内核和32线程，与Ryzen 7000系列类似。AMD还继续支持SMT（同时多线程功能）。根据AMD官方测试，其Zen 5内核面向PC平台的IPC性能相比Zen 4平均提升了约16%。 （苏 言）

意法半导体宣布在意大利建8英寸SiC晶圆厂

本报讯 近日，意法半导体（ST）宣布，将在意大利卡塔尼亚新建一座200mm（8英寸）碳化硅（SiC）制造工厂，主要用于SiC功率器件和模块的制造以及测试和封装。

结合在同一地点准备就绪的SiC衬底制造工厂，这些工厂将组成意法半导体的SiC园区，实现该公司建立完全垂直整合的制造工厂的愿景。意法半导体总裁兼首席执行官Jean-Marc Chery表示：“卡塔尼亚碳化硅园区所释放的全面集成能力，将在未来几十年内为意法半导体在汽车和工业客户中的碳化硅技术领导地位做出重大贡献。该项目提供的规模和协同效应将使我们能够更好地利用大批量生产能力进行创新，造福于我们的欧洲和全球客户，帮助他们向电气化转型，寻求更节能的解决方案。”

据了解，该SiC园区将成为意法半导体全球碳化硅生态系统的中心，整合生产流程中的所有步骤，包括碳化硅衬底开发、外延生长工艺、200mm前端晶圆制造和模块后端组装，以及工艺研发、产品设计、芯片、电源系统和模块的先进研发实验室以及完整的封装能力。这将在欧洲率先实现200mm碳化硅晶圆的量产，工艺的每个步骤（衬底、外延和前端以及后端）均采用200mm技术，以提高产量和性能。

据介绍，该SiC工厂计划于2026年开始量产，到2033年达到满负荷生产，满负荷生产时每周可生产多达15000片晶圆。预计总投资约为50亿欧元。意大利政府将在欧盟《芯片法案》框架内提供约20亿欧元的支持。 （文 编）

日月光推出powerSiP创新供电平台可提高AI应用能源效率50%

本报讯 半导体封测厂商日月光半导体近日宣布推出powerSiP创新供电平台，可以减少信号和传输损耗，同时应对电流密度挑战。

日月光半导体表示，powerSiP平台可实现垂直整合的多阶（Multi-stage）电压调节模块（VRM），提供更高的系统效率和更低的功耗，并比传统并排配置缩小25%的面积。powerSiP技术创新可使电流密度从0.4A/mm²增加50%至0.6A/mm²，并将布线功耗从12%降至6%，相较于传统并排配置的布线功耗降低了50%。由于人工智能（AI）市场规模、覆盖范围和影响仍在不断扩大，日月光通过powerSiP持续创新满足数据中心需求、性能预期和功耗改进。

另外，powerSiP是为了应对数据中心内算力（compute power）与冷却这两项最耗能的流程。根据国际能源总署（IEA）的数据，2022年数据中心消耗460太瓦时（TW·h），占全球用电量的2%。到2026年时，这个数字将上升至1000太瓦时（TW·h）。AI依赖强大但耗电的CPU、GPU、内存和磁盘系统，来实现功能、性能和低延迟，不断普及的人工智能使能源消耗激增，为了解决电力转换和冷却方面极端低效的问题，对创新的需求也空前高涨。

日月光研发副总叶勇谊表示：“powerSiP提供了将稳压器直接放置在系统单芯片（SoC）和小芯片（chiplet）下方的选项，垂直整合允许在较短的电力传输路径上提供较大的电流，借此可降低供电网络中的阻抗，进而提高系统性能和功能，同时增加整体效率和功率密度。”

日月光业务与行销资深副总Yin Chang表示：“人工智能正在逐步渗透到我们的生活，并在强大的高效能运算系统支持下，重塑知识工作、企业功能和人类体验，而先进封装对于数据中心运算系统效率优化扮演着关键角色，是我们将powerSiP平台推向市场的驱动力。通过独特的先进封装结构和创新的技术蓝图，powerSiP将持续精进以满足AI应用和HPC运算对于功耗和性能的需求。而日月光powerSiP是一个可根据产业技术蓝图和应用需求扩展的创新供电平台，目前已经上市。” （明 文）

社长：张立 社址：北京市海淀区紫竹院路66号赛迪大厦18层 邮编：100048 每周二、五出版 周二8版 周五8版 零售4.50元 全年定价420元 广告部：010-88558848/8808 发行部：010-88558777 广告许可证：京石工商广登字20170003号 发行单位：中国电子报社 印刷单位及地址：经济日报印刷厂 北京市西城区白纸坊东街2号