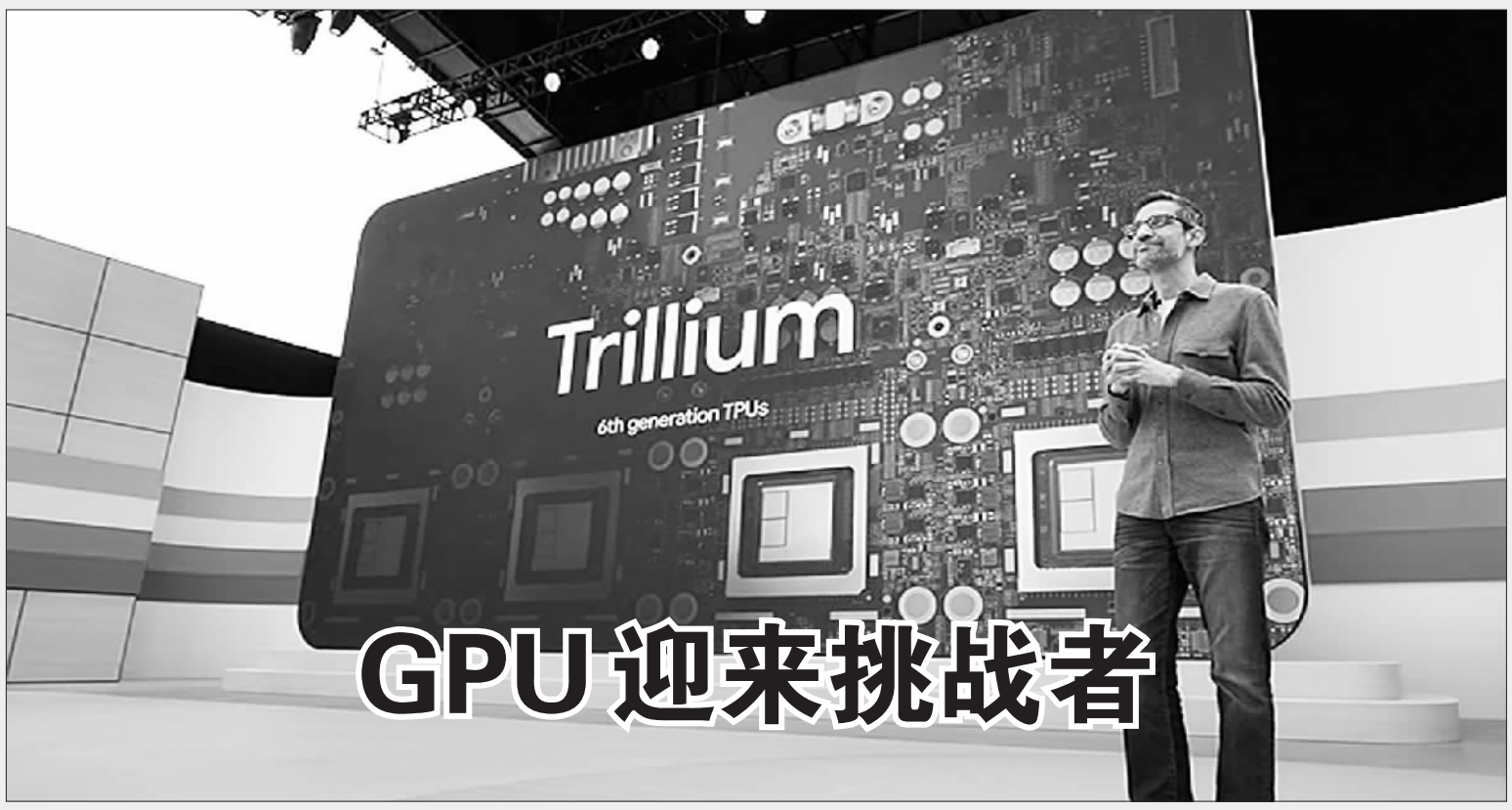

GPU迎来挑战者

本报记者 杨鹏岳

在全球科技产业将目光紧紧锁定人工智能领域之际，TPU（Tensor Processing Unit，张量处理器）正在悄悄成长为AI时代的“弄潮儿”。在GPU用于大模型训练、推理存在高能耗、高成本等问题的当下，由AI算力需求增长带来的问题，或许可以从TPU身上找到答案。

正进入主流市场

TPU芯片正在逐渐走出谷歌公司，获得更大范围的市场青睐，进入AI算力主流芯片市场。

由于入局早、算力强，由英伟达提供的GPU芯片几乎已成为各大企业训练、推理模型，处理AI相关算力需求的标配。但当前，TPU芯片也正在逐步进入AI算力主流芯片市场。

TPU由谷歌率先推出。从初代芯片推出至今，TPU的应用范围正逐步扩大，谷歌以外的市场也逐渐打开。

最初，TPU是谷歌专为加速机器学习和深度学习任务而设计的专用芯片，特别是针对深度学习模型的训练和推理。2013年，谷歌开始研发TPUv1，这是全球首款专为AI打造的加速器。2017年，谷歌推出Cloud TPU，用于处理云端计算任务。

自2022年年底生成式人工智能获得产业界广泛关注以来，TPU在生成式人工智能领域的应用范围也逐步拓宽。例如，2023年12月，谷歌推出的多模态通用大模型Gemini的三个不同版本，该模型的训练大量使用了Cloud TPU v5p芯片。

谷歌曾表示，TPU是其推出许多服务的最大功臣之一，要是少了它，如即时语音搜寻、相片物件辨识及互动式语言翻译，还有最先进的Gemini、Gemma、Imagen模型等都无法顺利问世。

今年5月，谷歌又发布了第六代TPU芯片Trillium。据悉，Trillium能在单个高带宽、低延迟Pod中扩展为多达256个TPU的集群，相较于前代产品，Trillium在适配模型训练方面的功能更强。

如今，TPU芯片正在逐渐走出谷歌公司，获得了更大范围的市场青睐。一如，7月30日苹果公司发布的一篇研究论文称，苹果在训练Apple Intelligence生态中的人工智能模型AFM时，使用了谷歌的两种TPU芯片。

更具有价格优势

GPU带来的高昂算力成本，使一众全球顶尖的科技企业望而生畏，而TPU帮谷歌大幅降低了算力成本。

以英伟达产品为代表的GPU在算力基础设施市场“一骑绝尘”的情况下，TPU何以崭露头角，又何以赢得苹果等全球知名企业的青睐？

市场分析师表示，以GPU为代表的通用计算架构和针对特定领域的DSA（Domain Specific Architecture，面向特定领域）计算架构是目前两大主流AI芯片设计思路。但在市场容量巨大的应用领域，计算芯片的架构演进总呈现出从通用型向专用型转变的规律。

随着大模型训练的计算量和复杂度的指数级增长，传统GPU成本昂贵、算力利用率低、能耗大的局限日益凸显。

今年1月，扎克伯格发帖公布了一组数字：到2024年年底，Meta“大规模的计算基础设施”将包括高达35万张英伟达H100显卡，为Meta在人工智能领域的突破提供强大支持。如果将其他GPU计算在内，Meta的计算基础设施将拥有“相当于近60万张H100的算力”。这将是一笔巨大的开销。

与此同时，马斯克表示，2024年特斯拉仅仅在英伟达的人工智能芯片上就将耗费超过5亿美元，特斯拉未来还需要价值数十亿美元的硬件才能赶上最大规模的竞争对手。

高昂的算力成本，使一众全球顶尖的科技企业望而生畏。在此背景下，作为AI专用芯片之一的TPU被业界期待能够从新的技术路线上另辟蹊径。在这方面，谷歌已经提供了成功经验。据谷歌副总裁兼工程院士Norm Jouppi透露，TPU的出现足足让谷歌省下了15个数据中心的建设成本。

探索市场新可能

谷歌通过谷歌云平台向外部客户提供基于TPU的算力服务。其他TPU企业也在寻找落地机会。

从TPU产品逻辑来看，作为一种专用集成电路(ASIC)，TPU专为单一特定目的而设计，用以运行构建AI模型所需的独特矩阵和基于矢量的数学运算，而GPU的设计初衷是处理图像信息。因此，从架构设计的角度来看，相比于适合处理高度并行任务的GPU，TPU更适用于处理矩阵乘法等神经网络算法。

TPU架构AI芯片公司中昊芯英创始人兼CEO杨龚轶凡在接受《中国电子报》记者采访时表示：“TPU具备处理AI大模型的天然优势。它专为神经网络结构而设计，在相同制造工艺、能耗和芯片尺寸条件下，性能优于GPU3~5倍。在个别应用场景中，TPU对既有的算法和框架进行优化，性能和功耗表现均优于GPU，更适用于深度学习模型的大规模部署。从长远来看，TPU比GPU在AI赛道上更具竞争力。”

在TPU芯片领域，市场正在进行不同的尝试。

一方面，“先行者”谷歌探索出了自己的专属路线。

咨询公司D2D Advisory首席执行官Jay Goldberg直言：今天只有两家公司具备用以训练人工智能模型的成熟芯片研发体系，一个是英伟达的GPU，另一个是谷歌的TPU。但区别于英伟达，谷歌并不会以独立产品的形态单独出售自己的TPU芯片，而是通过谷歌云平台向外部客户提供基于TPU的算力服务。

另一方面，更多芯片从业者仍在探索基于TPU架构的新产品。

今年2月，美国AI芯片初创公司Groq凭借其开发的新型AI处理器LPU（Language Processing Unit）引发关注，使用的TSP（张量流处理器）源头是谷歌研发的TPU。今年4月，英特尔推出了专攻深度学习神经网络推理的类TPU芯片Gaudi 3。此外，国内初创AI芯片企业中昊芯英也已量产TPU芯片，并自研AIGC预训练大模型，正在与行业伙伴进行金融、教育、医疗等垂直领域专业大模型的探索落地。

SEMI预测今年交付中国大陆的半导体设备总额将超400亿美元

本报讯 记者**姬晓婷**报道：今年以来，业内对全球半导体市场发展情况作出了越来越积极的判断。9月11日，SEMI全球副总裁、中国区总裁居龙在2024北京微电子国际研讨会暨IC WORLD大会上表示，从半导体设备投资情况来看，今年第二季度全球半导体市场增长乐观。居龙预测称，2024年中国大陆地区半导体设备交付额预计将在去年基础上再次增长，超过400亿美元，继续保持全球第一的市场地位。

根据SEMI统计的全球半导体产业投资情况，自2020年至今，全球半导体厂房和设备投资持续增长。居龙表示，即便是在整个半导体行业走入下行周期的2023年，半导体工厂和设备投资额也没有减缓。根据统计数据，100多家新的半导体制造工厂在2022年至2026年之间投入运营，这意味着今年与明年的半导体设备投资仍存在较大增长空间。

在全球半导体制造行业积极投资的版图中，中国大陆对半导体设备销售额的贡献率最高。2023年，中国大陆半导体设备销售额达到360亿美元，同比增长28.3%，在全球半导体设备销售市场中居首。

居龙表示，今年上半年半导体设备在中国大陆地区的交货金额已经达到230亿美元，预计今年全年设备交货金额将超过400亿美元，继续保持自2020年以来全球第一的市场地位。

根据SEMI对全球主要半导体设备公司市场份额的统计数据，2024年，中国市场在日本半导体设备企业Tokyo Electron、荷兰半导体设备企业ASML的营收占比有所上升，在美国企业应用材料的营收占比则有所下滑。

关于未来带动半导体市场的几大新技术、新机遇，居龙给出了三个关键词：AI、新能源汽车、先进封装。在AI方面，全球IT行业对计算设施的投资将逐年增加，预计至2027年，包括云端、汽车、消费端、PC等应用市场在内的AI半导体设备营收的年复合增长率将达到31%。在新能源汽车方面，汽车半导体价值规模将持续增长，预计到2026年，汽车半导体市场规模将增长至990亿美元。在先进封装方面，各海外龙头企业均在加大扩产力度，但扩产难度大、周期长，新建工厂普遍需要2至3年才能量产，短期内先进封装产能缺口难以解决，将持续供不应求。

中国工程院院士倪光南：

RISC-V提供硬件定制新路径

本报记者 张心怡

人工智能对性能和功耗的极致要求，使面向特定问题或特定领域的定制化芯片获得计算产业的广泛需要。9月10日，中国工程院院士倪光南在2024奕斯伟计算开发者伙伴大会作致辞时表示，RISC-V为硬件定制化提供了一条新路径。

自2010年在加州大学伯克利分校诞生，RISC-V以开放、简洁、模块化、易扩展和低功耗的综合优势引起全球关注，出货增长和应用拓展呈现星火燎原之势。

倪光南表示，RISC-V已经从一项学术成果发展成为具有全球影响力的开源架构和产业底座，不仅覆盖了从微控制器到高性能计算的广泛应用领域，更在架构设计的简洁性、模块化和易扩展方面展现出独特的优势。

在14年的发展历程中，RISC-V实现了嵌入式系统、物联网、边缘计算等领域的大量应用，为各类计算设备提供了强有力的支撑。如今，RISC-V正在迈进高性能计算领域，向桌面、服务器、人工智能、数据中心等计算领域拓展。

“这一系列的突破标志着RISC-V将迎来新的发展时代，为未来的计算架构开拓新的发展路径。”倪光南说道。

人工智能，是RISC-V重塑芯片产业格局的突破点。当前人工智能对计算架构提出了更高性能、更低功耗、更强的并行计算能力等一系列新要求。后摩尔时代，通用处理器的算力提升和功耗下降速度有所放缓。针对特定问题或特定领域来定义计算架构，成为市场的普遍诉求。

倪光南指出，软件技术的发展使软件定制化较易实现，但硬件定制化仍有待于解决，RISC-V在这方面提供了一条新路径。

“RISC-V在设计思想中就包含了DSA（Domain Specific Architecture，即面向特定领域）的概念。为此，RISC-V架构包含了模块化和自定义扩展指令集功能，并为扩展指令集预留了很大的扩展空间。”

芯粒（Chiplet）及其互联标准的发展，也为集成自定义控制指令集的专用支持硬件模块提供了支撑。“我们相信，RISC-V和Chiplet的完美融合将促进DSA计算架构的创新和发展，再结合软件调优，我们有望进入‘需求定义硬件’的新时代。”倪光南说道。

新思科技发布UCIe IP全面解决方案

本报讯 记者**姬晓婷**报道：9月10日，新思科技推出业界首个完整的UCIe IP全面解决方案，包括控制器、物理层和验证IP，每引脚运行速度达40 Gbps，实现异构和同构芯片之间的快速连接。新思科技表示，在同样的芯片尺寸和能效基础上，40G UCIe IP能够提供比UCIe规范高25%的带宽。

UCIe互连是裸片到裸片连接的行业标准，对于多裸片封装中的高带宽、低延迟裸片到裸片连接至关重要，该技术也助力了人工智能数据中心系统中的更多数据在异构和同构裸片或芯片组之间高效传输。新思科技40G UCIe IP支持有机基板和高密度先进封装技术，使开发者能够灵活地探索适合其需求的封装选项，可实现从早期架构探索到制造的快速异构集成。

新思科技介绍，40G UCIe IP解决方案具有多重优势。

其一，IP集成更简化。为便于使用和集成，该IP加快了裸片到裸片链路的初始化，无须加载固件。

其二，多芯片系统封装的可靠性增强。为了确保芯片、裸片到裸片以及多芯片系统封装层面的可靠性，该解决方案提供了测试和芯片生命周期管理（SLM）功能。此外，监控、测试和修复IP以及集成信号完整性监控器，可实现从设计到现场的多芯片系统封装诊断和分析。

其三，生态系统互操作性强。针对当前全新CPU和GPU的片上互连需求，新思科技40G UCIe IP支持业界广泛的芯片上互连结构，包括AXI、CHI芯片到芯片、streaming、PCI Express和CXL。为了实现成功的互操作性，该IP符合UCIe 1.1和2.0标准。

其四，具备预验证的设计参考流程。新思科技UCIe IP与3DIC Compiler（一个统一的从探索到签收平台）的组合可用于预验证设计参考流程，该流程包括所有必要的设计辅助工具，如自动布线流程、内插研究和信号完整性分析。

该产品预计将于2024年年底上市，适用于多种晶圆代工厂及其工艺。

瓦克成功开发供高性能芯片使用的新型特种硅烷

本报讯 记者**张维佳**报道：记者近日从德国有机硅及多晶硅龙头企业瓦克化学股份有限公司（以下简称“瓦克化学”）获悉，该公司开发出了一种新的、供高集成型存储芯片和微处理器生产使用的前驱物。

据介绍，该前驱物是一种新型硅烷，可在半导体生产中用于化学气相沉积。它能够与晶片表面发生反应，形成极薄的、介电常数低的绝缘层，以屏蔽对安装在密集空间中的导轨及其他元件的电磁干扰，从而确保高集成型芯片无中断地可靠运行。值得一提的是，此类前驱物所应用的高集成型存储芯片多用于人工智能、自动驾驶、云计算等高复杂计算领域。

记者了解到，半导体芯片体积极小，却有着数十亿个晶体管。当前，芯片零部件不断追求微小化，这对晶体管数量和半导体性能提出新的挑战。例如，电路越来越多，切换频率越来越高，芯片性能或受到电磁场影响。

而瓦克化学的该款新型硅烷产品则可以解决此类问题。“这种由硅、碳和氯构成的特种硅烷，是芯片生产过程中使用的重要成膜原料。它能够与受热的超纯硅片的表面进行反应，形成介电常数低的绝缘层，大大减少了导轨中高频率运动的电荷所受的电磁干扰。”瓦克化学相关负责人告诉记者。

据介绍，瓦克化学，创立于1914年，是欧洲规模最大的多晶硅生产商。目前，全球每两块芯片中就有一块用瓦克多晶硅生产而成。同时，瓦克也为半导体行业提供了重要的工艺化学品，如薄膜沉积用化学品，即前驱物。目前，中国已经成为瓦克化学最大的单一市场。2023年，瓦克化学在中国的销售额达到13.6亿欧元。

“中国是全球最大的化学品市场以及全球经济增长的重要引擎。中国市场有着巨大的发展潜力，将为我们提供诸多发展机遇。我们致力于继续投资中国，扩建生产基地以及研发中心来满足日益增长的需求。”瓦克化学总裁兼首席执行官贺达表示。

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