2024年AI芯片创新产品及生态应用揭晓

本报记者 张心怡 姬晓婷

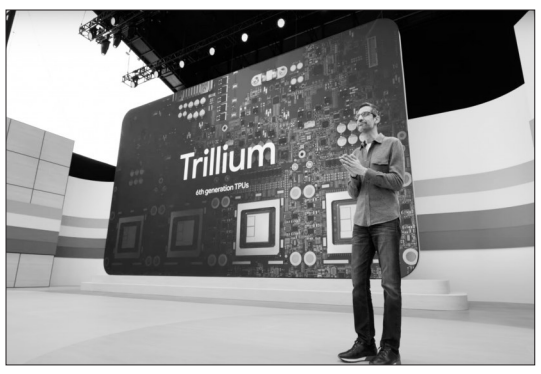
无论是市场动力、技术创新还是热点趋势，2024年的半导体产业都绕不开“AI”这个关键词。从算力芯片，到芯片互联、异构聚合等AI基础设施，再到软件平台、开发者社区与生态合作组织等平台生态建设，以及算力集群等大规模应用——半导体产业链的每一个节点，都在跟随AI的脚步创新发展。1月22日，“中国电子报编辑选择——2024年AI芯片创新产品及生态应用”正式出炉。本次编辑选择结合行业关注、市场热点和企业动态，通过考察技术领先性、市场竞争力、产品应用性、生态构建能力等指标，推选出10个创新产品及生态应用，涵盖优秀产品、创新技术、生态贡献和卓越应用等，以期为行业发展树立典范，为业界人士提供参考，为产业合作提供契机。

⊙优秀产品

第6代TPU Trillium

谷歌公司

Trillium（TPU v6e）是谷歌第6代TPU。相比上一代产品，其训练性能提升超4倍，推理吞吐量提升3倍；单颗芯片峰值计算能力（Int8）提升4.7倍，达到1836TOPs；HBM容量及带宽各提升1倍，分别达到32GB和1640GBps；芯片间互联带宽提高1倍，达到3584Gbps；能源效率提升67%。Trillium支持最多256个v6e芯片训练，以及最多8个芯片的单主机推理。在扩展能力方面，使用由3072个v6e芯片组成的12个计算模块进行部署时，Trillium实现了99%的扩展效率。在跨数据中心网络环境下，使用由6144个芯片组成的24个计算模块对gpt3-175b进行预训练，Trillium展现出了94%的扩展效率。



⊙优秀产品

拥有45TOPS NPU的PC平台骁龙X Elite

高通公司

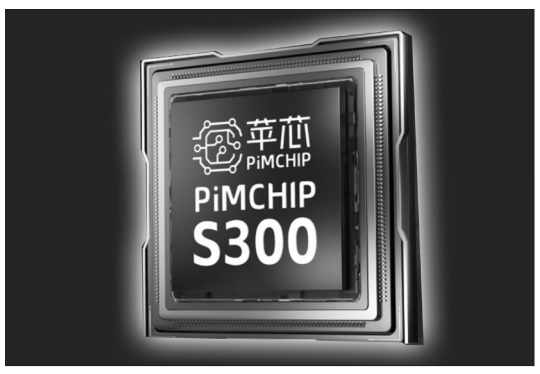
骁龙X Elite是高通专为Windows11 AI+PC打造的PC平台，拥有45TOPS NPU算力。该平台采用定制的集成高通Oryon CPU，拥有12个高性能内核，主频达3.8GHz，双核增强技术可将两个高性能内核提升到4.3GHz，是首个主频达到4GHz以上的ARM架构CPU核心；集成Adreno GPU能够实现每秒4.6万亿次浮点运算的图形性能。骁龙X Elite采用了全新异构高通AI引擎，该平台支持10Gbps 5G下载速度，以及高通FastConnect7800支持的Wi-Fi 7等通信技术。除骁龙X Elite外，骁龙X系列还包括骁龙X Plus、骁龙X Plus（8核）以及骁龙X平台。

⊙优秀产品

多模态智慧感知决策AI芯片PIMCHIP-S300

北京苹芯科技有限公司

PIMCHIP-S300是一款基于存算一体技术的多模态智慧感知决策AI芯片。该芯片搭载的SRAM存算一体技术能够使计算直接在存储器内部发生，有效减少了传统架构中数据搬运带来的能耗和延迟问题，核心能效比达27TOPS/W，同时能够实现低功耗运行。该芯片支持音频、视频多模态融合感知，具备超低功耗唤醒、VAD（声音活动检测）、语音识别、运动监测和视觉识别功能，可适用于智能穿戴、智慧家居、无人机等多个新兴行业，为合作伙伴带来高能效、低成本的解决方案，推动其产品的智能化转型。



⊙创新技术

异构GPU协同训练方案HGCT

上海壁仞科技股份有限公司

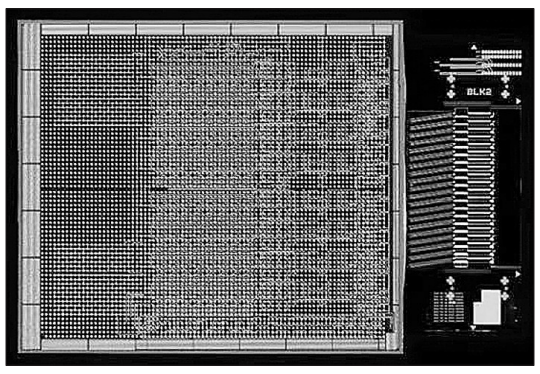
壁仞科技自主原创的异构GPU协同训练方案HGCT，是业界首次实现统一异构通信库支持多种不同厂商、不同型号的GPU进行GDR高速通信，也是业界首次实现四种异构GPU混合训练同一个大模型，将异构算力有效聚合，端到端混训效率达95%~98%，突破了大模型异构算力孤岛难题，有望实现万卡、十万卡异构集群混训。该方案具备普适性、易用性、兼容性，有助于加快国产GPU的落地迁移，有助于应用方最大化异构GPU集群利用效率，助力国产大模型落地。壁仞科技已联合中国移动发布“芯合”异构混合并行训练系统，联合中国电信、中兴通讯等发布“智算异构四芯混训解决方案”。

⊙创新技术

实现光学I/O芯粒的完全集成

英特尔公司

面向大模型和生成式AI的部署需求，数据中心需要指数级提升的I/O带宽和更长的传输距离，以支持更大规模的处理器集群和更高效的架构。在2024年光纤通信大会上，英特尔展示了完全集成的OCI（光学计算互连）芯粒。基于已实际验证的硅光子技术，英特尔在OCI芯粒中集成了包含片上激光器的硅光子集成电路（PIC）、光放大器和电子集成电路。英特尔现场展示的OCI芯粒与自家CPU封装在一起，但它也能与下一代CPU、GPU、IPU等SOC集成。该OCI芯粒可在最长100米的光纤上，单向支持64个32Gbps通道，以满足AI基础设施日益增长的对更高带宽、更低功耗和更长传输距离的需求。

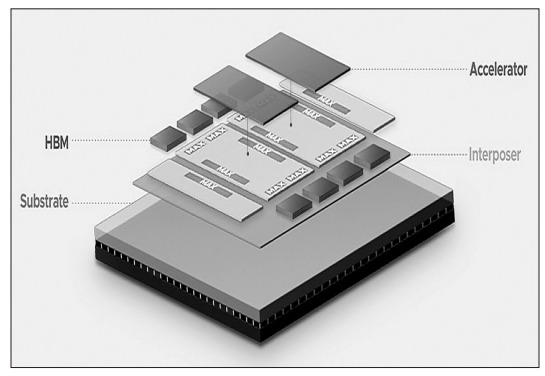


⊙创新技术

用于AI XPU的3.5D面对面（F2F）封装技术

博通公司

2024年，博通公司推出3.5D系统级封装平台技术XDSiP，助力消费级人工智能领域企业开发下一代定制加速器（XPU）。该技术结合了3D硅片堆叠与2.5D封装，能在单个封装器件中集成超6000平方毫米的硅片以及多达12个HBM堆栈。基于创新的面对面（F2F）堆叠方式，3.5D XDSiP直接连接上下层芯片的顶部金属层，提供了密集可靠的连接，将电气干扰降至最低，并具备出色的机械强度。相比传统的面对背（F2B）封装，3.5D XDSiP技术将堆叠芯片之间的信号密度提升7倍，降低了芯片间接口的功耗以及3D堆叠内计算、内存和输入/输出组件之间的延迟，能够实现更小的中介层和封装尺寸。



⊙生态贡献

开源软件平台ROCm

超威半导体产品（中国）有限公司

ROCm是一个开源的软件平台，通过支持异构硬件、优化平台工具、提升开发者友好程度以及拓展合作伙伴等方式，助力AI产业生态建设。ROCm既支持AMD的多种GPU架构，还推出了适用于中国本土需求的定制化解决方案；提供了诸多优化的库和工具链，如MIOpen（深度学习库）、rocBLAS（线性代数库）和rocFFT（快速傅里叶变换库）等，为开发者提供了强大的开发基础，加速了算法的实现和优化。该平台同时提供了编译器工具链、性能分析工具、调试工具等多种类型的平台工具，通过提供详尽文档和教程等方式给予开发者支持。

⊙生态贡献

海光产业生态合作组织

海光信息技术股份有限公司

海光产业生态合作组织（以下简称“光合组织”）是海光信息打造的贯通“芯片设计与制造—整机系统—软件生态—应用服务”各个环节的开放创新链和产业链，凝聚了超过4000家上下游合作伙伴。该组织通过共同开展技术攻关、方案优化、应用创新及市场开拓，为千行百业提供了高质量的产品及解决方案。海光信息C86-3G、C86-4G等CPU系列产品，深算一号、深算二号等DCU系列产品，完全兼容x86指令集以及国际主流操作系统和应用软件，为该组织的应用拓展提供了产品基础。

⊙卓越应用

庆阳“东数西算”智能算力枢纽示范工程

上海燧原科技股份有限公司

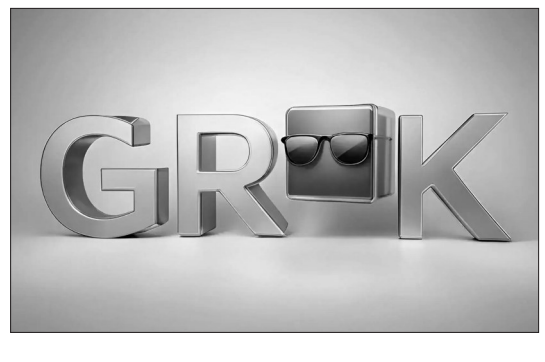
2024年6月19日，由燧原科技支持建设的全国算力枢纽（甘肃·庆阳）首批万P算力上线，首个国产万卡算力集群启动。该算力集群搭建燧原科技新一代人工智能推理加速卡“燧原S60”。在项目建设过程中，燧原科技致力于打造智算中心的新范式，解决“谁来建设”“如何运营”“用户在哪儿”三大核心问题，整合贯通智算产业生态链条，实现了智算中心“建”“运”“用”的商业闭环与可持续发展，助力庆阳充分发挥大模型算力资源优势，发展人工智能技术应用，实现当地产业集聚和人才汇聚。

⊙卓越应用

NVIDIA以太网加速构建全球最大AI超级计算机

英伟达公司

2024年，英伟达宣布人工智能初创企业xAI的Colossus超级计算机集群达到了10万颗NVIDIA Hopper GPU规模。该集群使用了英伟达第一款专为AI打造的以太网网络平台NVIDIA Spectrum-X，是专为多租户、超大规模的AI工厂设计的RDMA（远程直接内存访问）网络。Colossus是世界上最大的AI超级计算机之一，被用于训练xAI的Grok系列大语言模型。xAI和英伟达用了122天就建成了所有配套设施和Colossus超级计算机，从第一个机架落地到开始训练任务，只用了19天，而建造这种规模的系统通常需要数月乃至数年的时间。



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