英伟达AI芯片上新传递两大信号

本报记者 王信豪

美国当地时间11月13日，在2023年全球超算大会（SC23）上，英伟达推出了新一代GPU H200。当日，英伟达的股价在截至收盘时间上涨0.6%，为486.2美元，数据显示，英伟达创下了近7年来最长的连涨时间纪录。这一次，英伟达将AI芯片的发展方向引向了存储和推理能力。

英伟达的新品，与早些时候英特尔和AMD透露的AI芯片发布计划呼应，AI芯片从以算力为中心的竞争向更加多样化的竞争发展。

AI芯片竞争焦点转向存储

想要通过生成式AI和HPC应用创造智能，就必须使用大型、快速的GPU显存来高速、高效地处理海量数据。

相比于前代H100，H200的性能提升了60%~90%，但是单看算力这一指标，H200的算力与H100基本相同，对比两者的产品规格表可以发现，实现算力不变而性能飞跃的关键点就是内存。

英伟达在发布H200时表示，该产品是全球首款搭载HBM3e的GPU，而H100中使用的仍是HBM3。据了解，HBM3e能够为H200提供传输速度达4.8TB/秒的141GB显存。与前代产品、常被其他竞争者视作“计量单位”的A100相比，其容量也翻了近乎一倍，带宽增加了2.4倍。

事实上，英伟达在AI芯片方面的挑战者AMD在几个月前就强调了AI芯片存储容量的重要性。AMD CEO苏姿丰表示，AMD即将于今年12月推出的纯GPU产品Instinct MI300X使用具备192GB显存的HBM3，带宽为5.2GB/秒。值得注意的是，AMD产品具备8个HBM3显存堆栈，而英伟达产品仅有6个。同时，AMD首个AI加速器产品MI300A也拥有多达13颗小芯片，共包含1460亿个晶体管，配置128GB的HBM3内存，相比前代的MI250，MI300A的性能提高8倍，效率提高5倍。

英特尔CEO帕特·基辛格曾在9月举行的英特尔On技术创新峰会上公布了英特尔AI芯片的路线图：到2024年，英特尔将推出采用5nm制程的Gaudi3，其下一代AI芯片代号则为Falcon Shores。Gaudi2于2022年推出，从产品规格来看，Gaudi2所使用的是6个HBM2堆栈共96GB。英特尔方面表示，Guadi3的算力将是前代产品Gaudi2的2倍，网络带宽、HBM容量是Gaudi2的1.5倍。

可以看出，在各芯片企业的推动下，AI芯片的竞逐正从初期的算力指标竞争，进一步延展到存储领域，HBM（高带宽存储）几乎成为未来AI芯片当中必备且必争的存储器。英伟达超大规模和高性能计算副总裁Ian Buck表示：“想要通过生成式AI和HPC应用创造智能，就必须使用大型、快速的GPU显存来高速、高效地处理海量数据。”

半导体行业专家盛陵海告诉《中国电子报》记者：“从技术角度来讲，存储性能是提高AI训练能力的瓶颈，如果存储能力跟不上算力，整个模型的训练便难以高效运行；从企业角度来讲，发展存储性能也是性价比相对较高的方式之一。”据了解，当前最高规格高带宽存储器的HBM3e主要由SK海力士、三星以及美光提供。

围绕模型推理提升AI芯片性能

面向训练需求，半导体厂商推出了多块芯片互连的解决方案，用以支持更大参数的模型训练。

大模型训练和推理的效率与效果是衡量GPU性能的重要指标，各大厂商也围绕着大模型训练与推理不断提升自身实力。

H200的性能提升着重体现在模型推理上面。英伟达方面表示，H200在700亿参数的大语言模型Llama 2上的推理速度比H100提高了近一倍，功耗也会降低50%。同时，英伟达H200和H100都基于Hopper架构，互相之间具备强兼容性，可实现一定程度上的替换，同时英伟达透露，下一代采用全新架构的GPU B100也将于2024年推出，进一步强化训练和推理的性能。

面向训练需求，半导体厂商推出了多块芯片互连的解决方案，用以支持更大参数的模型训练。英特尔的GAUDI 2 HLBA-225支持8块Guadi2共同运行；AMD推出Instinct平台同样搭载8块MI300X，被苏姿丰称作是“人工智能推理和训练的终极解决方案”；英伟达依靠自身NV Link和NV Switch的高速互连技术，推出了HGX H200的服务器主板。英伟达称，客户可选择搭载4块或8块H200，配合Grace Hopper芯片可为各种应用工作负载提供最高的性能，包括针对1750亿参数以上超大模型的LLM训练和推理。

“从训练来看，虽有目前常说的‘百模大战’，但是其参数量和精度等还需要进一步优化，同时，专业大模型的成熟度也不比通用大模型，故而当前的主要诉求仍在大模型训练上。”盛陵海说，“与此同时，未来的推理需求则逐步增加，因为人工智能的应用落地，最终还是要看推理能力。”

AI带动FPGA市场爆发式增长

本报讯 记者**沈丛**报道：11月14日，英特尔FPGA中国技术日在北京举办。会上，英特尔可编程方案事业部副总裁兼网络业务部总经理Mike Fitton表示，在人工智能等新兴市场的带动下，FPGA市场将迎来爆发式增长，市场规模将从2022年的90亿美元增长到2028年的130亿美元，年复合增长率达到7%。

随着人工智能的快速发展，FPGA作为一种灵活度高、可配置性强的芯片，逐渐成为人工智能应用领域的明星。被誉为“万能芯片”的FPGA，因其适用于各种人工智能应用场景，正逐渐受到业界的热烈追捧。

英特尔可编程方案事业部FPGA芯片产品营销高级总监Lux Joshi表示，当今更智能、更互联的市场对FPGA技术提出了三点需求。首先，FPGA技术必须具备应对各项标准不断变化的更高工作负载的能力，包括加快创新步伐、快速集成新功能，以及支持任何制程节点、代工服务或IP的能力。其次，在数据激增的背景下，FPGA需要具备更高级别的存储方案和带宽。对FPGA的要求包括拥有更高的内存带宽、集成内存的新选项以及支持全新的内存接口。最后，随着FPGA设计和工作负载加速的复杂性日益增加，未来需要更简便的FPGA开发、AI和工作负载加速工作流，并建立开放式的加速生态系统。

在英特尔FPGA中国技术日的现场，各种融合FPGA技术的AI应用纷纷亮相，向观众展示了FPGA在AI领域的无限可能。

在一款静脉成像仪的展台前，记者看到，通过仪器的照射，能够清晰地看到皮下血管的分布。展台工作人员告诉《中国电子报》记者，该技术利用近红外光进行照射，通过FPGA采集传感器接收的图像，并利用FPGA的高速并行特性，采用大窗口多尺度卷积神经网络，对皮下血红蛋白吸收后的光强数据进行计算。随后，再经过一系列图像增强算法处理，通过DLP模组将图像投射到手臂上，实现增强静脉可视化的效果。

随后，记者在工业缺陷检测实训平台的展台中看到，在一个微型检测仪中放入测试产品后，该测试仪能自动分析出检测产品的缺陷并发出口令，机械臂会将其放置在预定的位置。重庆海云捷迅科技有限公司客户成功部总监柴广龙介绍，工业缺陷检测实训平台以高端铝材表面缺陷检测需求为背景，深度融合FPGA、边缘计算、人工智能与深度学习等关键技术，可以为高校的FPGA、人工智能等专业教学提供实训产品。工业缺陷检测实训平台能够将工业相机、光源、微型传输机、机械臂等模块集成为一体，并采用FPGA进行模型推理，利用深度学习算法和铝片表面缺陷数据集来实现缺陷识别。该平台可检测出铝片表面的划痕、针孔、褶皱、脏污等缺陷类型，并支持用户通过二次开发增加其他缺陷种类。

2030年RISC-V芯片全球出货量将超160亿颗

本报讯 近日，RISC-V International的首席执行官Calista Redmond在RISC-V峰会上表示，未来数年里，RISC-V将以40%年复合增长率（CAGR）攀升，预计到2030年可能会有超过160亿个RISC-V芯片。目前RISC-V芯片大约有10亿个，这意味着接下来的十几年里出货量将大幅度增加。

RISC-V作为一种开放标准指令集架构，根据开源协议可以免费试用，基本指令集具有32位固定长度自然对齐指令，并且支持可变长度扩展，如今已应用在各种小型嵌入式系统到大规模的机架式并行计算机上。Calista Redmond表示：“目前在世界各地已经有数十亿个RISC-V内核，有分析人士甚至指出，很难找到任何不包括RISC-V架构的新设计，RISC-V是我们这个时代最深刻的技术革命。”

RISC-V目前广泛应用于微控制器，比如高通将RISC-V用于旗下移动系统芯片上。RISC-V现在能迅速发展，是因为不少企业选择利用RISC-V开发人工智能（AI）和高性能计算解决方案，随着时间的推移，RISC-V技术可能还会继续扩展，比如应用到GPU领域。

Calista Redmond指出，RISC-V架构处理器如果想要与Arm、x86架构处理器竞争，还需要一个更为强大的软件和硬件生态系统，而两者现在发展的速度都很快。据了解，RISC-V得到了来自世界各地4000多家软件开发公司的支持，业界也推出了很多针对软件和硬件设计人员的主板产品。

（微 文）

DDR3内存市场涨幅超过10%

本报讯 根据集邦咨询的最新报道，DDR3内存市场近期出现了显著反弹，涨幅超过10%。机构预测2023年第四季度合同价格将上涨10%~15%，而明年第一季度将继续保持上涨趋势。

业内人士普遍认为，DDR3内存本次反弹的最大原因是三星、SK海力士和美光等全球领导者持续减产，并严格控制出货量，以实现涨价的目标。另一个主要原因是三星、SK海力士和美光等公司正积极投入人工智能应用领域，并将主要产能转移到生产高带宽内存（HBM）和DDR5内存，在短期内造成了DDR3供应紧张。

集邦咨询表示，自9月以来，DDR3价格稳步上升。其中，4Gb容量的产品累计涨幅接近10%，而2Gb容量的产品累计涨幅达到了14%。预计本季度合约价格将强劲增长10%~15%，并预测明年第一季度将持续走强，未来可能会再上涨5%~10%。

DDR3相关公司对市场发展持乐观态度。钰创公司认为，随着库存消化接近尾声，“周期性底部已经结束”，并逐步迎来复苏的曙光。公司对明年全球DRAM市场的显著增长持乐观态度。

（吉 文）

韩国10月内存芯片出口增长1%

本报讯 韩国科学技术信息通信部11月14日发布的数据显示，10月韩国信息和通信技术（ICT）产品出口同比下降4.5%，至171亿美元，连续16个月下滑。不过，这是自去年9月以来的最小同比降幅。

占ICT总出口额一半左右的芯片出口额为89.7亿美元，同比减少4.7%，创下今年以来最小降幅。其中，内存芯片出口额同比增长1%，为16个月以来首次增长。多芯片封装引领反弹，增长12.2%，而DRAM芯片出口额降幅一年多来首次收窄至个位数。

半导体是韩国经济出口的支柱，韩国央行预计明年韩国经济将增长2.2%，如果贸易状况继续改善的话。这将比今年1.4%的预期增长有所加快。

10月，韩国出口总额自去年年底以来首次出现增长，这是韩国经济增长前景和全球经济需求复苏的一个积极迹象。日前公布的ICT出口分项数据证实了一种观点，即半导体将引领韩国出口增长，并支撑经济增长势头。

（韩 宇）

台积电先进封装明年月产能较原目标增加约20%

本报讯 由于英伟达、苹果、AMD、博通、Marvell等重量级客户近期大幅追单，台积电CoWoS先进封装需求将迎来爆发。

据称，台积电为应对上述五大客户需求，已经在努力加快CoWoS先进封装产能扩充的脚步，预计明年月产能将比原目标再增加约20%，达3.5万片。

业界人士分析称，台积电五大客户大追单，表明AI应用已经遍地开花，各大厂商对于AI芯片的需求都出现了大幅度增加的情况。

目前CoWoS先进封装技术主要分为三种——CoWos-S、CoWoS-R、CoWoS-L，其中CoWoS-L是最新技术之一，结合了CoWoS-S和InFO技术的优点，使用中介层与LSI（本地硅互连）芯片提供灵活的集成方案，可用于芯片到芯片的集成。

公开资料显示，英伟达是目前台积电CoWoS先进封装的主要大客户，几乎包下了六成相关产能，包括H100、A100等AI芯片都有应用，而且AMD最新AI芯片产品目前也正处于量产阶段，预计明年上市的MI300芯片将采用SoIC及CoWoS等两种先进封装结构。

除此之外，AMD旗下赛灵思也一直是台积电CoWoS先进封装的主要客户。随着未来AI需求量持续增加，赛灵思、博通等公司同样也开始对台积电追加CoWoS先进封装产能。

（微 言）

半导体厂商积极布局氮化镓业务

本报讯 记者**沈丛**报道：近日，欧洲功率半导体大厂英飞凌宣布成功收购GaN Systems，并表示此次收购已经获得所有必要的监管部门审批。

集邦咨询数据显示，2022年GaN Systems在氮化镓领域的市场份额为12%，排名第五，而英飞凌在氮化镓领域的市场份额较低，仅归属于“其他”这个门类。收购GaN Systems之后的英飞凌在氮化镓领域的市场份额将提升至15%，排名跃升至第四。

据了解，此次收购GaN Systems并不是英飞凌第一次在氮化镓领域发力。英飞凌曾在2015年以30亿美元收购氮化镓企业International Rectifier（IR），但并没有使氮化镓业务获得突破性的进展，且始终难以与硅和碳化硅业务匹敌。

Yole数据显示，预计到2026年全球氮化镓元件市场规模将增长到423亿美元，即突破人民币千亿元，年复合增长率约为13.5%。

安森美总裁Hassane El-Khoury在接受《中国电子报》记者采访时表示，安森美未来可能会通过收购的方式拓展氮化镓业务，并密切关注氮化镓在汽车和工业市场的应用前景。

随着氮化镓市场的不断增长，意法半导体、Wolfspeed等硅和碳化硅的头部企业也有可能像英飞凌一样通过并购来拓展氮化镓业务。

据悉，氮化镓目前主要市场在于100W以内的快充领域，相比较于其他功率器件市场比较有限。距离氮化镓在汽车、工业功率器件中大规模应用还很远。

专家指出，氮化镓在过去十年没有迎来爆发式增长，主要是受到市场需求、技术难度、战略规划和市场竞争等因素的制约。目前氮化镓还没有找到其不可替代且市场规模庞大的应用领域，需要等到这样的应用领域出现后，氮化镓才能迎来真正的拐点。

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