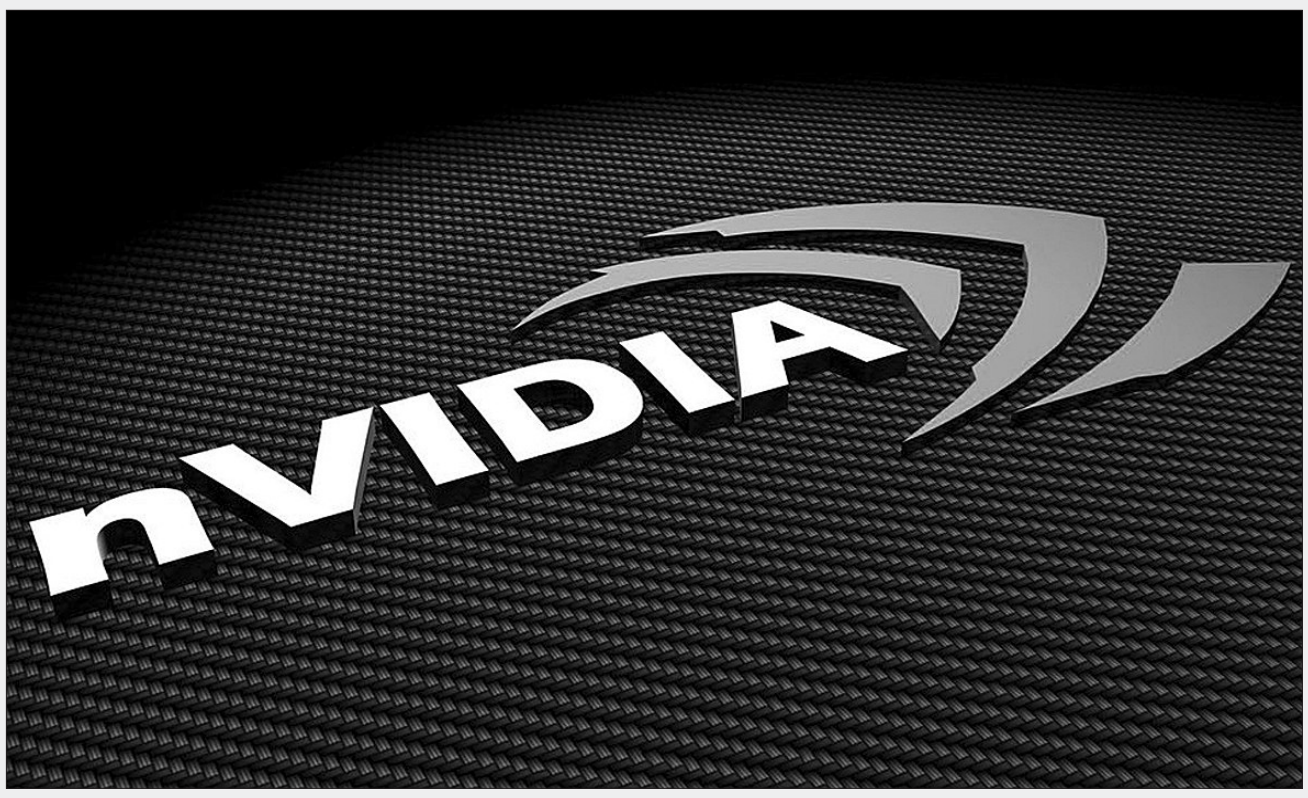
英伟达：云与AI交相激荡

特约撰稿 李佳师

自今年以来，英伟达GPU芯片紧俏，其GPU在市场上的价格也一涨再涨，从2022年12月开始，其A100价格累计涨幅达68.75%，H100的最新价格超过了4万美元。即便如此，客户依然追着英伟达下单。

不过，尽管市场上英伟达GPU量价齐升，英伟达CEO黄仁勋却忙着去卖云服务。继今年3月英伟达与微软、谷歌、甲骨文等云计算巨头合作托管其云服务NVIDIA GDX-Cloud之后，6月底黄仁勋又宣布与云计算巨头Snowflake合作，售卖英伟达的软件和模型服务。最近，有媒体透露，英伟达正计划向云服务商Lambda Labs投资3亿美元，Lambda Labs最早业务是销售GPU驱动的计算机，后转型为GPU云服务器租赁，目前主要面向企业出租带有英伟达芯片的服务器，从而与亚马逊、微软、谷歌等云提供商展开竞争。

为什么黄仁勋要在英伟达GPU奇货可居时急着去卖云服务？



目前英伟达正与三星等其他代工厂洽谈合作，但这些代工厂是否能满足英伟达需求仍是未知数。

三大原因威胁未来发展空间

手里拥有GPU这个“印钞机”，黄仁勋没有开足马力拼命“印钞”，而是去卖GDX Cloud和AI软件，其中有三个关键原因。

一是英伟达GPU产能受限，卖多少、以什么样的节奏卖GPU取决于台积电。目前英伟达的H100和A100全部代工都交给了台积电，但在台积电的产能分配中，英伟达并不在前五中。

根据研究机构2023年1月数据，台积电的最大客户是苹果公司，另外九家客户对台积电营收贡献排序为：联发科、AMD、高通、博通、英伟达、Marvel、意法半导体、亚德诺、英特尔。AIGC时代到来，GPU爆火，英伟达已经向台积电紧急追加订单。据消息人士透露，目前台积电5纳米制程产能利用率推高至接近满载，正在以超级急件生产英伟达GPU，其订单已排至年底。

或许人们会问，为何英伟达不将一些GPU代工订单分给三星？事实上，英伟达究竟能产出多少GPU取决于台积电有多少CoWoS的产能。

目前台积电给英伟达的3~4纳米制程的产能基本管够，但CoWoS先进封装产能奇缺，即便是三星能够提升3纳米晶圆的良率，但尚无法提供CoWoS。CoWoS是台积电的一种“2.5D”封装技术，能将计算、内存等晶片堆叠到硅中介层（硅转接板），通过硅中介层上的高密度布线实现晶片互连，再安装到基板上进行封装。这种封装方式提升了芯片的连接速度，并降低了功耗，为目前AI服务器芯片厂商主要采用的封装方式。按照TrendForce集邦咨询的预测，台积电的CoWoS月产能将在2023年年底达到1.2万片，即便是台积电把CoWoS产能都给英伟达（事实上是不可能的），也只能达到月产1.2万片。据了解，目前英伟达正与三星等其他代工厂洽谈合作，但这些代工厂是否能满足英伟达需求仍是未知数。

二是巨头们纷纷自研AI芯片。微软、亚马逊、甲骨文、特斯拉等都是英伟达GPU的大客户，每一个巨头都给出了数以万颗计的订单，但这些“大金主”，早就在自研AI芯片，这意味着今年他们是金主，明年就有可能不是了。事实上，马斯克在从英伟达手里买了1万颗H100后没几天，就在宣布成立生成式人工智能公司xAI的同时，xAI将自研AI芯片。

微软是英伟达GPU的最大买主之一，为支撑Open AI模型训练所搭建的超算中心，微软下单就是几万颗A100。最近有消息称，微软自研AI芯片即将推出，微软从2019年开始启动代号为“雅典娜”（Athena）的AI定制处理器研发计划，该芯片基于台积电的5纳米工艺制造，目前芯片已经在微软和OpenAI的特定员工手中进行测试，性能与英伟达GPU相近的芯片，成本仅为后者的1/3。

此前亚马逊是英伟达GPU的大买主，亚马逊云科技（AWS）的自研芯片已经在加速落地并集成在其云服务中。目前AWS的自研芯片包括Nitro、通用处理器芯片Graviton、用于机器学习的训练和推理芯片Inferentia和Trainium。在训练方面，相较于通用GPU，AWS Trainium的Trn1实力在训练吞吐率上，单节点吞吐率可提升1.2倍，多节点集群的吞吐率可提升1.5倍。推理芯片需要权衡吞吐率和延迟之间的平衡，AWS最新的推理芯片Inferentia2可以实现吞吐率提升3倍，而成本只是通用GPU的1/4。

谷歌从2012年开始采购英伟达GPU，三年后谷歌开始自研AI芯片TPU，目前谷歌TPU已经进化到第四代。最近谷歌披露其用于训练人工智能模型使用的超级计算机的最新细节，谷歌称，这些系统基于第四代TPU芯片，速度和能耗效率均高于英伟达A100，性能是英伟达A100的1.7倍。谷歌同时透露，其正在研发能够与H100相抗衡的AI芯片。

这些云巨头通过自研芯片+AI模型提供人工智能算力与模型服务，大幅降低用户使用大模型的门槛，显著降低使用算力与模型服务的成本，具有巨大的竞争优势，蚕食英伟达的领地。

三是英特尔、AMD等直接竞争对手在紧锣密鼓地出牌抢英伟达生意。对于这个正在开启的大模型时代，每家计算芯片厂商都不可能放过此次机遇。

在最近举行的AMD新品发布会上，AMD推出了用于训练大模型的GPU Instinct MI300X。AMD CEO苏姿丰称，MI300X是AMD真正为生成式人工智能设计的产品，比起英伟达的H100芯片，MI300X提供了2.4倍多的内存和1.6倍多的内存带宽。苏姿丰还发布了“AMD Instinct Platform”，集合了8个MI300X，可提供总计1.5TB的HBM3内存，该平台对标英伟达的CUDA。

在为英伟达贡献将近一半收入的中国市场（2023财年中国市场营收占英伟达全球市场份额的47%），同样有竞争对手，中国的芯片厂商、互联网企业正在加速研发AI芯片。不仅如此，英特尔也在最近针对中国市场推出云端AI训练芯片——Habana Gaudi 2。英特尔在发布现场称，该芯片在一些关键性能指标上胜过英伟达A100，Gaudi2运行ResNet-50的每瓦性能约是英伟达A100的2倍，运行1760亿参数BLOOMZ模型的每瓦性能约达A100的1.6倍。

产能限制加上对手们的快速追赶，对英伟达的未来收入构成威胁，英伟达盯上云服务就很好理解了。

既要大模型也要云红利

产能限制加上对手们的快速追赶，对英伟达的未来收入构成威胁，英伟达盯上云服务就很好理解了。

其实，一直以来英伟达就是“左手硬件、右手软件”，也正是因为有软件平台CUDA这一推手，才让英伟达的GPU在AI时代风生水起。

云服务是先进模式，从大型机、小型机、PC、互联网迭代到今天的按需提供，整个计算产业都进入了云时代，微软、亚马逊、谷歌等云计算巨头皆因提供有规模效应的云计算服务而“躺着赚钱”，黄仁勋不可能不动“云”的心思。以云的方式提供GPU的AI算力服务、AI模型服务就成为英伟达的必须选择。

“计算平台正经历60年来最根本的变革，有两个动力正在推动计算发生变革，一方面，你不能再不断购买CPU，因为CPU扩展已经到了尽头，购买再多的CPU，也得不到相应的吞吐增加，需要加速计算。另一方面，计算机的整个操作系统都发生了深刻变革。”黄仁勋说。

自今年以来，英伟达频繁投资云服务企业。最近，有媒体爆出英伟达或将以3亿美元收购Lambda Labs。Lambda创立于2012年，早年业务重点是销售GPU驱动的计算机，后转型为GPU云服务器租赁，公司的年收益也从千万美元规模上升至数亿美元。Lambda labs的规模虽然不大，但其号称能提供全世界价格最低的NVIDIA A100、H100算力资源。而一个多月前，英伟达投资了Lambda的竞争对手CoreWeave。CoreWeave专门为企业级GPU加速工作负载提供云服务，其官网称，为计算密集型用例（机器学习和人工智能、视觉特效和渲染以及像素流）构建云解决方案，算力比大型通用公有云快35倍，比传统云服务提供商便宜80%。

除了投资，英伟达频频与云计算巨头联手卖自己的云服务。在今年3月举行的2023年GTC大会上，英伟达宣布与微软Azure、谷歌Cloud等多家云服务供应商合作，推出NVIDA DGX Cloud。用户无须采购和拥有服务器，通过浏览器就可以获得云服务商合作托管的DGX Cloud计算、AI框架、预训练模型服务，NVIDA DGX Cloud的收费标准是每个实例每月36999美元起，每个实例包括8个Nvidia H100或A100 80GB GPU，每个GPU节点内存达640GB，计算资源专用，不和云中另外的租户共享。

6月底，英伟达宣布与数据云Snowflake合作，为这家基于云计算的数据仓库公司的客户提供生成式人工智能技术。Snowflake的用户可以在数据不离开平台的前提下，直接利用英伟达的预训练AI模型，在云平台上对自己公司的数据进行分析，开发针对自己数据的AI应用，这招非常有杀伤力，因为目前企业对采用大模型的最大忧患就是数据安全。在与Snowflake的合作中，英伟达除了提供GPU芯片，另外一个重点是人工智能软件、模型服务。

因为GPU的光环，很多人只关注到英伟达超级AI计算能力而忽略了英伟达的软件和AI模型服务能力，事实上，英伟达还是一个大模型服务商。在今年三月的GTC大会上，英伟达推出的DGX Cloud就包括了AI超级计算、AI框架与模型服务。目前英伟达有三个人工智能模型服务：语言模型NVIDIA NeMo、图像视频模型NVIDIA Picasso、药物研发模型NVIDIA BioNeMo。

英伟达称其每项模型云服务都包含六个要素：预训练模型、数据处理框架、矢量数据库和个性化服务、经过优化的推理引擎、各种API、以及NVIDIA专家提供的支持，可帮助企业针对其自定义用例调整模型。简而言之，大模型所需的“全栈服务”统统囊括。

在大模型时代，或许并不是每个企业用户都能买得起1000颗GPU（大模型训练起步算力为1000颗），但每月支付36999美元，就可拥有AI算力、AI框架和模型的云服务，却是有可能的，这是一个增长潜力更大的长尾市场。事实上，并不是每一个企业都有必要去训练一个大模型，但每一个企业都需要大模型服务、需要AI。

黄仁勋在接受媒体采访时坦言：“人工智能软件是一个比硬件大得多的市场，基础设施服务市场，硬件销售机会的总额大约在10亿美元；但人工智能在自动化、加速等相关产业，制造业是数十亿美元的市场需求；在医疗领域，药物发现、科学家实验室研发、药物研究等，又是一个数万亿美元的市场，每一个行业的市场都远远大于硬件领域。”

软件市场远比硬件市场大得多，所以英伟达公司将一半的资源用于发展软件平台，现在不仅仅是与AI相关的超级计算，还包括AI框架、模型以及其他软件，英伟达正在给它们加载上“云翅膀”。今年早些时候，英伟达除了让微软Azure托管超级AI计算服务DGX Cloud，还托管了NVIDIA Omniverse Cloud，这包括了设计、开发、部署和管理工业元宇宙应用所需的全栈环境。

应该说，云服务是能够将AI三要素（数据、计算、模型）融合的最佳模式，现在云计算与大模型的叠加效应已经变成了云计算巨头实实在在的收益。日前，微软和赛富时（Salesforce）双双宣布，将对软件中的AI功能收费，其中微软将对其生产力软件中的生成式人工智能功能收取每月30美元的费用，这意味着微软365服务的商业级版本平均月费大幅上涨53%~83%；赛富时宣布向所有用户开放AI产品，并给出了单个产品每个用户每月50美元的定价，赛富时的AI分为Sales GPT和Service GPT两个产品，分别对应销售和客户服务两个职能。

在芯片厂商中，英伟达是一个另类。目前英伟达的营业收入不及英特尔，但市值却是英特尔的8倍，是AMD的5倍，成为市值突破万亿美元的首家半导体公司。这或许与它不只把眼光盯在芯片、盯在成熟市场上有关。

关注变化中的早期市场、在AI领域进行长期布局，是英伟达总踩上AI重要风口的原因。别人只看到了英伟达收获OpenAI的巨大订单，但其实早在2016年，黄仁勋就造访了这家AI创业公司，并捐赠了搭载8颗P100、价值百万的超级计算机DGX-1。最近黄仁勋又在德国与几个欧洲AI初创公司交流，在被问及英伟达的决策为何一直全对时，黄仁勋给出的答案是“靠直觉”。“大量想法同时涌现，相互激荡。在此时，要采取正确的态度，要自信、有学习意愿，并且一开始不能苛求完美。”黄仁勋说。

时下，“云”与“AI”交相激荡。那么，乘上云战车，为客户提供AI算力和模型服务，而不仅仅是硬件GPU，才能挖掘AI时代的更大红利，也因此成了黄仁勋当下最重要的选择。

三星电子下半年将加快拓展高端存储器市场

本报讯 记者**许子皓**报道：7月27日，三星电子公布了2023年第二季度财报。财报显示，三星电子第二季度的销售额为60.01万亿韩元（约合人民币3352.3亿元），同比下降22.28%；营业利润为6700亿韩元（约合人民币37.4亿元），同比下降95.25%，但较第一季度增长4.7%。

三星电子共包括6个部门，分别为DX部门（智能手机和数字电器设备）、VA/DA部门（电视等消费电子）、MX/Networks部门（智能手机等通信设备）、DS部门（半导体）、SDC事业部（显示器）以及Harman（汽车设备）业务。其中，DX部门、MX/Networks部门和DS部门都出现了利润亏损，但三星电子表示，DS部门运营的亏损减少，抵消了第二季度智能手机出货量下降带来的负面影响，使公司运营利润好于上一季度。

对于备受业界关注的存储器业务，三星电子表示，当前客户仍在调整库存，存储器市场仍然疲软，特别是NAND产品的市场价格还在下跌，但由于DDR5和高带宽存储器（HBM）等用于AI技术的高端存储芯片需求持续高涨，DRAM产品的第二季度出货量超出预期。预计今年下半年，存储器市场将逐渐复苏。三星电子将加强在高性能服务器和移动旗舰产品领域的领先地位，加快拓展DDR5、LPDDR5x、HBM3等高端产品的市场份额，计划到2024年，将其HBM产品的生产能力提高一倍。

对于消费电子市场，三星电子表示，智能手机市场预计将在今年下半年恢复同比增长，但主要是在高端市场，而平板电脑市场将基本持平。在可穿戴设备方面，智能手表市场可能会有所收缩，而无线耳机（TWS）市场会略有增长。三星电子将通过推出新一代折叠屏智能手机、高端平板电脑以及可穿戴产品来实现营收增长，同时，还将加强5G芯片的核心技术以保持市场领先地位。

此外，三星电子还在财报中提到了S.LSI和Foundry（晶圆制造厂）业务。三星电子表示，经济放缓和通胀导致智能手机需求疲软，延迟了芯片需求复苏，该业务的业绩低于预期，预计今年下半年，生产线的利用率还将下降，利润还将下滑。但三星电子表示，其目前的GAA工艺3纳米产品研发进展顺利，正在按计划开发3纳米的改进工艺以及先进封装等方案，还将拓展欧洲汽车领域客户。

SK海力士下半年将扩大AI用存储器生产能力

本报讯 记者**许子皓**报道：7月26日，韩国存储器领军企业SK海力士发布了截至6月30日的2023财年第二季度财报。财报显示，SK海力士在2023财年第二季度合并收入为7.3059万亿韩元（约合人民币408.6亿元），相较于第一季度，营业亏损率从67%降低至39%，营业亏损为2.8821万亿韩元（约合人民币161.2亿元）。SK海力士表示，由于当下AI服务器的存储器需求激增，SK海力士将持续扩大高容量DDR5、HBM3 DRAM等DRAM产品的生产能力。但NAND的去库存速度比DRAM缓慢，因此将继续减产NAND产品。

SK海力士表示，随着以ChatGPT为中心的生成式AI市场的扩大，面向AI服务器的存储器需求激增，SK海力士的HBM3和DDR5 DRAM等高端产品销售量增加，因此，第二季度的营业收入比第一季度增加44%，营业亏损减少15%。

据SK海力士透露，虽然当下PC、智能手机市场依旧保持颓势，DDR4等普通DRAM价格仍在下降，但由于AI服务器用的高价格、高配置产品的销售持续增长，DRAM整体平均售价较上一季度有所提升，对于SK海力士的营业收入起到了关键性的带动作用。

SK海力士预测，今年下半年，AI用存储器的强劲需求将得到延续。因此，今后SK海力士将以HBM3、高性能DDR5和LPDDR5 DRAM等AI用存储器以及基于176层NAND闪存的SSD为中心，持续提高销售业绩，从而加速改善下半年业绩。

另外，SK海力士表示，今年将提升第五代10纳米级（1b）DRAM和238层NAND闪存的初期量产良率和质量，以应对即将到来的存储器市场转暖。但SK海力士认为NAND的去库存速度比DRAM要缓慢，因此决定扩大NAND产品的减产规模。

在资本投资方面，SK海力士财务担当副社长（CFO）金祐贤表示：“虽然全公司的投资规模同比减少50%以上的基调没有变化，但SK海力士将持续扩大今后主导存储器市场的高容量DDR5、HBM3 DRAM等产品的生产能力。”

赛迪顾问集成电路高级分析师杨俊刚向《中国电子报》记者表示，当前，存储器市场整体依然处于底部震荡的阶段，服务器DRAM产品占DRAM整体市场的35%左右，企业级SSD产品的市场占比不足20%。但受AI市场的使用需求增长，相较于NAND，DRAM产品销售量会提升快一些。未来，随着AI通用大模型的复杂化和多样化发展，企业对存储器产品的性能要求越来越高，数量的需求越来越大，存储器市场将保持回升态势，企业间的竞争也将更加激烈。