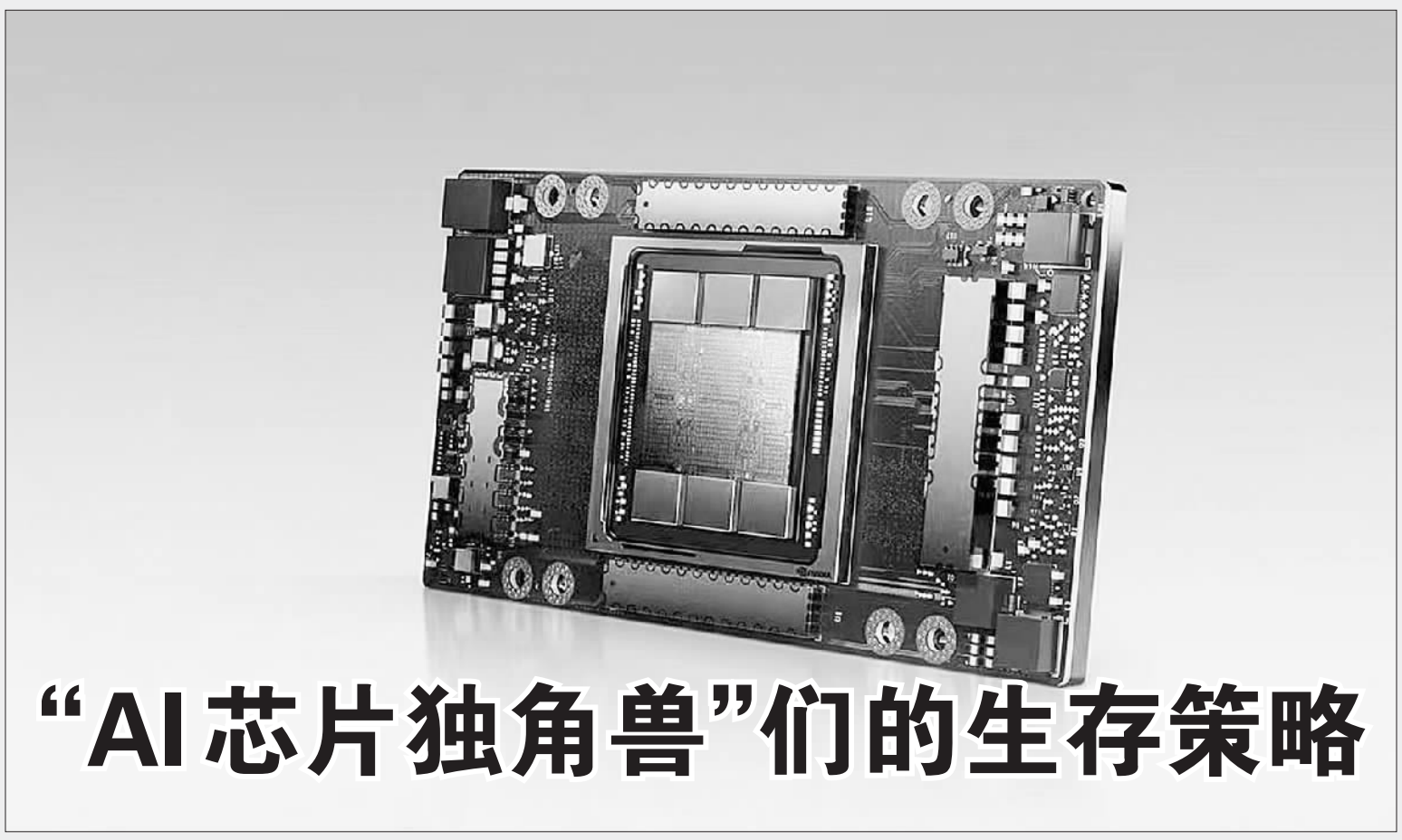

“AI芯片独角兽”们的生存策略

本报记者 王信豪

近日，全球的“AI芯片独角兽”们接连发布公司和产品的进展。Groq宣称，其推出的Groq Chip推理效率是英伟达H100的10倍，并在社交媒体上引发热议；被誉为“小英伟达”的Astera Labs于当地时间3月20日在美国纳斯达克上市，当前总市值达104.3亿美元。

在当前的AI芯片市场格局之下，英伟达乘上东风先拔头筹，AMD和英特尔紧追不舍，谷歌和微软等云服务商纷纷加入自研行列。在越发激烈的竞争中，新入局的“AI芯片独角兽”正在孵化自己的力量，摸索出合适的生存和盈利路径。

产品定位：训练还是推理？

对于已经具备一定规模的云服务商而言，它们是拥有更多选择的一方。如谷歌、微软等具有深厚软件开发技术和资金支持的企业还可选择自研算力芯片。

记者整理了关注度较高的15家国外芯片初创企业后发现，推理是更受AI芯片初创企业青睐的应用场景。

OpenAI CEO山姆·奥特曼在2023年年底陷入“宫斗风波”时传出将投资一家AI芯片公司，后又爆出将花费5100万美元购买Rain AI公司基于RISC-V架构的NPU用于边缘侧应用的人工智能推理；Etched.ai针对大语言模型推出的ASIC芯片专注于AI推理；MatX在官方网站上表示“我们专注于低成本进行大模型预训练和推理”，同时补充道，“推理优先”。

推理成为大多数初创企业的共同选择，这背后是对训练和推理两种不同场景市场增量的考量。

在训练方面，芯片企业的下游客户，即通过购买GPU或算力芯片进行大模型训练的AI市场存在饱和风险。

对于新的AI企业来说，参与大模型竞争的门槛正在提高。在通用大模型“狂野生长”的过程中，训练数据不断膨胀，参数量级也越来越大，这也意味着训练需要AI企业筹备更多数量的算力芯片，“万卡起步”的算力门槛将导致未来的大模型格局走向寡头竞争的收敛阶段——能否像埃隆·马斯克一样投入5亿美元用于购入上万块英伟达H100对自家的大模型或聊天机器人进行训练？这是所有AI企业在入局之前都需要评估的问题。

对于已经具备一定规模的云服务商而言，它们是拥有更多选择的一方。如谷歌、微软等具有深厚软件开发技术和资金支持的企业还可选择自研算力芯片，且更加适配自家AI产品。

相比于训练，AI芯片初创企业在推理上的机会更多。在从“炼大模型”向“用大模型”的转变过程里，使用8块H100或MI300进行推理的性价比较低，也存在延迟和能源消耗等问题，这些都成为下游云服务商在推理环节关注的重点。小体量的芯片初创企业可以通过这些痛点打开突破口，从而找到在激烈竞争中的一席之地。

当然，并非所有企业都只聚焦于推理，部分企业正在尝试用其他方式来解决大模型训练所带来的问题。相较于GPU这种已经成熟的解决方案，几家企业在训练上呈现的思路更具想象力。

Cerebras Systems推出一款体积巨大的芯片WSE-3。据了解，WSE-3拥有超过4万亿个晶体管和46225mm²的硅片面积，相比于通过NVLink连接8块或者更多的H100，保持完整性可以降低互连成本和功耗。

Extropic希望通过热力学和信息技术来构建AI超级计算机，目前该计算机已经进入到硬件组装阶段。Lightmatter推出了光子处理器Envise。相较于传统的硅基芯片，光子处理器可以在高功耗和高性能中达成平衡。“人类正在为AI的发展投入大量能源，而这种能源消耗正在迅速增加，芯片技术也到了无法解决这个问题的地步。”Lightmatter在官网上表示。尽管在概念上天马行空，但是两家企业的产品距离落地还有一段时间。

面对大厂：竞争还是合作？

在目前的市场环境中，不论是与大厂直接竞争，还是成为供应链的一环，初创企业必须体现出自己的差异化特性才能保证存活，换言之，企业需要不断创新。

AI芯片初创企业面临的第二个问题是如何衡量与英伟达等大厂之间的关系。

上文提到，许多企业瞄准推理领域，一个有趣的现象是，英伟达的产品成了各企业对标的主要对象。

Etched.ai的ASIC芯片Sohu专为大模型推理设计。Etched.ai负责人表示：“通过将Transformer架构刻录到Sohu中，我们正在创建世界上最强大的Transformer推理服务器。”Etched.ai官网显示，在同样使用8块芯片的前提下，Sohu的推理效率比H100和A100都要高。

Groq推出的LPU（语言处理器）宣称其推理性能是H100的10倍，且成本是H100的十分之一。

d-Matrix的产品Corsair在与英伟达的对比中，不论是数据吞吐量、时延，还是成本方面都具有更好的表现。据了解，Corsair使用PCIe5对8块Chiplet进行互连，拥有约1300亿个晶体管，且Chiplet之间的带宽达8TB/s，最终可节约90%左右的成本。“我们所有的硬件和软件都是为了加速Transformer模型和生成式AI构建的。”d-Matrix首席执行官兼CEO表示。

在参与竞争之外，也有企业选择成为大厂的合作伙伴，在供应链中担任其中一环。

成功上市的Astera Labs的产品聚焦于连接数据和内存的器件。Astera Labs创始人之一的Jitendra Mohan认为，随着AI和机器学习的发展，除了算力，数据连接也将是关键问题。Astera Labs官网的自我介绍为“专为AI和云基础设施构建的连接”，其主要产品包括Aries PCIe/CXL智能定时器、Leo内存控制器，以及Taurus有源智能电缆模块，可帮助企业连接芯片、存储器和服务器，从而构建GPU算力集群。也因此，英特尔、谷歌、亚马逊等芯片和云服务商都将成为其潜在客户。

在目前的市场环境中，不论是与大厂直接竞争，还是成为供应链的一环，这些初创企业必须体现出自己的差异化特性才能保证存活，换言之，企业需要不断创新。

事实证明，当前的AI芯片领域正在涌现出更丰富的设计思路。Etched.ai的Sohu选择将Transformer架构刻录在芯片上（Etched意为“蚀刻”），Groq通过SRAM和TSP（张量流处理器）来提升推理效率。新的设计理念层出不穷，而差异化的创新不能止步于此，Sohu作为ASIC，能否适应Transformer架构的优化升级，而Groq的芯片如何处理好此前饱受争议的成本问题，还需时间和市场进一步检验。

开发生态：自研还是联合？

如果说产品质量决定了AI芯片初创企业能否站稳脚跟，那么开发生态的完整度和牢固程度就将决定企业能否长远发展。

如果说产品质量决定了AI芯片初创企业能否站稳脚跟，那么开发生态的完整度和牢固程度就将决定企业能否长远发展。

CUDA一直被视作英伟达的护城河，CUDA在长年被使用的过程中已经暗中抬高了开发者的迁移门槛。在3月英伟达禁止第三方硬件兼容CUDA之后，其市场垄断地位得到巩固。

面对当前的生态竞争，一方面，初创企业正在尝试自研，d-Matrix推出开源软件栈Aviator。据了解，Aviator使用开源软件，使用户能够轻松部署模型，并将系统软件集成到推理服务器中，用于进程生成和扩展通信。Modular发布Mojo编程语言，可支持CPU、GPU、TPU和ASIC等多种硬件的程序编写。

另一方面，AMD和英特尔的做法也能提供参考思路。

AMD在生态上的考量是“方便开发者迁移和使用”。ROCm平台是AMD开发的开源计算平台和生态系统，旨在为开发者提供一种跨平台的编程模型。ROCm提供了与CUDA相似的API与功能函数库，使得为英伟达GPU编写的代码只需稍加修改便可运行在AMD的GPU上，从而降低开发者将程序从CUDA转移至ROCm的迁移成本。

英特尔在去年9月联合Arm、富士通、谷歌、Imagination、高通和三星等公司共同成立了UXL基金会（Unified Acceleration Foundation，统一加速基金会），以联盟化的形式构建开放生态，此举也被产业界视作想要摆脱英伟达CUDA生态垄断的联合行动。

“该基金会的目标是围绕开放标准和开源软件将加速器生态系统联合起来，以便开发人员可以构建能够针对多供应商、多架构系统的应用程序——现在和将来。如果您在编写软件时不需要考虑目标处理器，那么我们已经完成了我们的工作。”UXL生态系统副总裁兼基金会指导委员会主席Rod Burns表示。

据悉，该基金会建立在oneAPI的项目规范之上，oneAPI是英特尔推出的开发者接口。“该规范和项目由英特尔为基金会提供，涵盖了开发人员编写代码时所需的基础知识。这些项目将在UXL基金会开放治理的原则下运作，这意味着所有贡献都得到平等对待，基金会成员在项目的未来方面也拥有公开提案和讨论的发言权。”Rod Burns补充道。

4月10日凌晨，随着英特尔发布Gaudi 3，AI芯片的竞争更加激烈，头部公司“神仙打架”，初创企业大浪淘沙，就连英伟达CEO黄仁勋每天都在“担心公司会不会倒闭”。面对更加复杂的环境，“AI芯片独角兽”们也在凭借自身韧性不断探索，求生、求变、求富。

Arm推出新一代AI加速器Ethos-U85及全新物联网参考设计平台

本报讯 记者**姬晓婷**报道：4月9日，Arm宣布推出Ethos-U85神经网络处理器（NPU），其MAC（乘加累计运算）单元可从128个扩展到2048个，算力可实现在1GHz频率下达到4TOPs，能够为工厂自动化、商用零售、智能家居等高性能边缘AI应用提供支持。此次发布会同时发布了全新物联网参考设计平台——Arm Corstone-320。

当前微处理器被部署到包括工业机器视觉、可穿戴设备和消费者机器人等在内的诸多高性能物联网系统中，且各场景对人工智能功能提出了更高的要求。Ethos-U85由此推出，能够与Armv9 Cortex-A CPU相结合，加速处理机器学习（ML）任务，提供更高能效的边缘推理能力，从而满足上述场景对人工智能功能的更高需求。

据了解，Ethos-U85性能相较前代产品提升了4倍，能效提高了20%，可支持Transformer架构和卷积神经网络（CNN）以实现AI推理。Transformer架构适用于视觉和生成式AI用例，对理解视频、填充图像的缺失部分，分析来自多个摄像头的数据以进行图像分类和目标检测等任务非常有效。全新Ethos-U85 NPU支持了TensorFlow Lite和PyTorch等AI框架。Ethos-U85提供了与前几代Ethos U系列相同的一致工具链，合作伙伴能够利用现有的投资，实现开发体验的无缝衔接。

Arm Corstone-320物联网参考设计平台集成了Arm最高性能的Cortex-M CPU——Cortex-M85、Mali-C55 ISP和全新的Ethos-U85 NPU，为语音、音频和视觉等广泛的边缘AI应用提供所需的性能。该平台的软硬件结合特性将使开发者能够在物理芯片就绪前便启动软件开发工作，从而加速推进产品进程，为日益复杂的边缘AI设备缩短上市时间。

Arm此次发布的最新架构主要面向边缘侧，关于Arm此次发布的产品将在家电厂商探索结合人工智能技术新品中发挥什么作用，Arm物联网事业部业务拓展副总裁马健告诉《中国电子报》记者，第一，能够提供更自然的交互能力。传统智能家居，要使用其智能功能的话，需要用户下载新的APP，这对于用户而言，使用过程不够友好。而搭载大模型的家用电器，其自然语言交互能力得到了极大的提升，通过采用支持Transformer架构的Ethos-U85和Corstone-320参考设计，可以比较轻松地将由支持自然语言交互的大模型压缩的小模型部署到各种家电中，这对用户来说门槛会更低，也会更友好。

第二，云、边、端的协同。大模型之所以智能，是由于参数量大。当前多数大模型被部署在云端。家电厂商收集的用户数据等，都集中在云端训练和优化。但同时，为了增强用户体验，家电终端侧、家庭网关边侧，也会部署支持自然语言响应的模型，因此，为了支持更丰富的用例和功能，云、边、端的协同至关重要，而Arm的技术横跨云、边、端，便可很好地充分发挥出优势作用。

马健还补充道，Arm在数据中心、云、边、端都有强有力的产品和统一的工具链支持，因此更具有生态优势。

铠侠计划在2031年量产1000层3D NAND

本报讯 近日铠侠公司CTO宫岛英史在第71届日本应用物理学会春季学术演讲会上表示，该公司将在2030—2031年间推出具有1000层堆叠的3D NAND闪存技术，并对存储级内存业务进行重组，以应对日益增长的数据存储需求。

铠侠与西部数据的合作在NAND闪存技术领域已经取得了显著成果，目前他们最先进的产品是218层堆叠的BICS8 3D闪存，这种闪存能够实现3200MT/s的I/O速率，为数据存储提供了更高的速度和效率。

在提升3D NAND闪存堆叠层数的过程中，技术难度也随之增加。特别是在蚀刻垂直通道的过程中，随着深宽比的增高，这一工艺的难度和成本都在不断上升。铠侠在BICS8中采用了双堆栈工艺，通过分开蚀刻两个NAND堆栈的垂直通道，降低了整体的难度。这种工艺的采用为未来千层堆叠NAND闪存的发展奠定了基础，有望包含更多个NAND堆栈，进一步提升存储密度。

宫岛英史还指出，铠侠在业务丰富程度上相较于同时运营NAND和DRAM的竞争对手处于劣势，因此有必要培育新型存储产品业务，如存储级内存（SCM）。在AI热潮的推动下，DRAM与NAND之间的性能差距正在扩大，而SCM的出现正好可以填补这一空白。

为了加强在SCM领域的研究和开发，铠侠于4月1日将“存储器技术研究实验室”重组为“先进技术研究实验室”，并将研究重点放在MRAM、FeRAM、ReRAM等新型内存技术上，预计在2~3年内实现出货。此前铠侠在SCM领域主要聚焦于XL-FLASH闪存方案，并在2022年推出了支持MLC模式的第二代XL-FLASH，显示出该公司在新型存储技术上的持续投入和创新能力。

（铠 文）

三星宣布完成16层混合键合HBM内存技术验证

本报讯 三星电子近日透露，公司完成了采用16层混合键合HBM内存技术验证，已制造出基于混合键合技术的16层堆叠HBM3内存样品，该内存样品工作正常，未来16层堆叠混合键合技术将用于HBM4内存量产。

据悉，混合键合技术作为一种新型的内存键合方式，相较于传统的键合工艺，具有显著优势。它摒弃了在DRAM内存层间添加凸块的烦琐步骤，直接通过铜对铜的连接方式实现上下两层的连接。这种创新方式不仅提高了信号的传输速率，更好地满足了AI计算对高带宽的迫切需求，同时也降低了DRAM层间距，使得HBM模块的整体高度得到缩减。

混合键合技术的成熟度和应用成本一直是业界关注的焦点，为了解决这一问题，三星电子在HBM4内存键合技术方面采取了多元化的策略。除了积极推进混合键合技术的研究与应用，三星电子还同步开发传统的TC-NCF工艺，以实现技术多样化，降低风险，并提升整体竞争力。在混合键合技术之外，三星还在不断探索和优化TC-NCF工艺。据悉，三星的目标是在HBM4中将晶圆间隙缩减至7.0微米以内，以进一步提升HBM4的性能和可靠性。

业内专家表示，三星在16层混合键合堆叠工艺技术方面的突破，将有力推动HBM内存技术的发展，为未来的计算应用提供更为强大的内存支持。

（星 译）

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