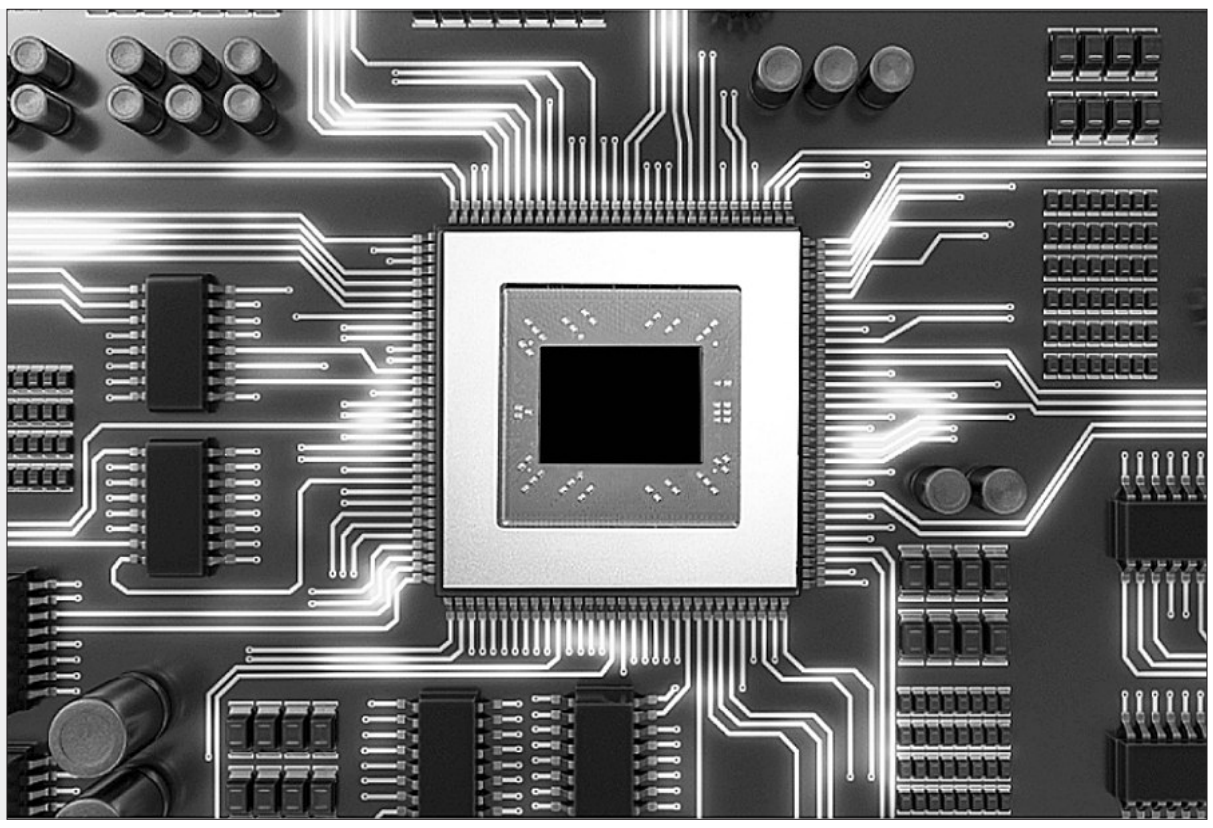
人工智能大模型需要怎样的芯片？

本报记者 张心怡

AI对话机器人ChatGPT的走红，让“大模型”的热度从学术界、产业界一路烧到了大众媒体。信息显示，ChatGPT是一款由大型语言模型驱动的聊天机器人，在它背后起作用的是OpenAI最强大的大型语言模型GPT-3，参数量高达1750亿。

大模型的崛起，也为AI技术的地基——芯片带来了重要的商机与严峻的技术挑战。据Lambda实验室测算，如果采用英伟达V100 GPU和当时最便宜的云服务进行计算，GPT-3训练一次需要355个GPU年（一块GPU运行355年的运算量）、花费460万美元。大模型需要怎样的AI芯片，AI芯片企业又该如何应对？带着这样的疑问，记者采访了有关专家和企业家。



算力需求：对于AI芯片的要求全面拉升

大模型对计算的要求主要体现在三个方面，一是算力，二是互联，三是成本。

4年的时间、1500倍的参数量提升，大型语言模型展现出强劲的扩张态势。2018年，OpenAI推出第一代生成式预训练语言模型GPT-1，拥有1.17亿个参数。2019年推出的GPT-2拥有12亿参数，是GPT-1的10倍多。一年后面世的GPT-3，参数量达到1750亿个，是GPT-2的100多倍，预训练数据量高达45TB。ChatGPT正是基于GPT-3.5——一个基于GPT-3的微调版本。

与参数量一起飙涨的，是大模型的算力需求。燧原科技创始人、董事长兼CEO赵立东向《中国电子报》表示，以ChatGPT为代表的生成式AI模型为了实现高质量的内容生成，具备两大特性：参数规模巨大的模型、海量的训练数据集。因此，大模型在底层算子支撑、系统架构方案、软件生态支持方面都和原先的决策式AI模型有着显著的区别，需要芯片厂商打造更加领先的系统级软硬件方案，并在技术和产品层面打破原有的路线与壁垒。

“从算力芯片角度，主要有三方面的需求：一是分布式计算能力，包括数据并行、模型并行、流水并行等分布式计算方案，计算效率尤其关键。二是大容量高带宽的内存方案，在每个AI芯片内部有效提升数据处理能力和算力利用率，结合HBM以及CXL等新型存储技术，进一步提升本地存储能力和算力利用率。三是更高的单芯片计算能力，以降低整体系统复杂度，并降低TCO成本。”赵立东说。

昆仑芯科技负责人也向记者表示，伴随着ChatGPT的迭代，大模型算法对算力的要求不断提高，而算力的核心就是人工智能芯片。“大模型对计算的要求主要体现在三个方面，一是算力，二是互联，三是成本。大模型的热潮也将加速AI芯片技术的创新和进步，包括芯片架构、算法优化、功耗控制等方面的提升。AI芯片公司可以在这些方面进行研发和创新，不断提高产品竞争力。”

除了在已有的AI芯片路径持续迭代调优，获得更优的算力、效率、功耗、成本，大模型强烈的高算力需求，也呼唤芯片电路与架构层面的进一步创新。北京大学集成电路学院研究员贾天宇向《中国电子报》记者指出，大模型技术的出现和普及，将进一步推动AI芯片的发展，拉升多方应用产业对于AI芯片的需求，对于提升AI芯片的重要性和不可或缺性有着积极的意义。但同时也应认识到，支持大模型计算的AI芯片算力需求高、设计复杂度高，其设计要求和门槛也随之提升。

“由于传统芯片技术发展的限制，当前芯片设计的电路与架构面临着存算力瓶颈、能效瓶颈、设计复杂度瓶颈等多方面挑战。如何解决算力需求和芯片设计瓶颈之间的矛盾，还需要多方的创新和技术进展。”贾天宇说。

技术路径：通用与定制的平衡

大模型作为一个趋势场景，其需求定义清楚了，设计和实现技术就会水到渠成。

在AI芯片的发展过程中，通用性和定制化像是两个互相拉扯的作用力，衍生出一系列的芯片架构。其中，GPU和ASIC分别是通用性和定制化的代表架构，也有着各自的优势和局限性。

“GPU架构提供了大量数据并行结构，因此能够提供大量的AI并行计算，适用于AI训练等。ASIC等定制化AI芯片针对特定的应用场景进行芯片优化，能够取得更高的计算能效，但通用性相对较弱。”贾天宇表示。

好在，随着芯片技术的发展，通用性与定制化已不再壁垒森严。一方面，英伟达在GPU架构中引入了Tensor Core（张量计算核心），一种专门针对深度学习应用而设计的专用ASIC单元，使GPU更加适合深度学习。另一方面，定制化芯片也逐步增加了通用计算单元，并引入可编程或部分可编程的架构，增强芯片的场景覆盖能力。

“过去被认为只具备专用性的ASIC或DSA（领域专用架构），现在不仅含有用于AI加速计算的专用单元，还含有与英伟达GPU中CUDA Core类似的通用计算单元，同样可以实现对各种指令的处理。因此，无论是GPU、ASIC，还是DSA架构，云端AI芯片企业在架构设计中需要关注的是通用和专用计算单元的搭配，以便应对AI计算持续演变带来的挑战。”昆仑芯科技相关人员向记者表示。

“鉴于大模型对于大算力的显著需求，以及模型训练算子的多样性，具有大算力、通用性的芯片将是大算力应用的首选。在现存的技术方案中，GPU是能够提供算力和开发生态的选择。然而，由于GPU的功耗过高，类GPU架构的定制化大算力AI芯片也将存在市场，满足对于芯片计算能效的提升需求。”贾天宇指出。

而通用与定制的“配比”，要根据具体场景设计。昆仑芯科技相关人员表示，在通用性、易用性和性能之间实现平衡，需要在实际设计中结合需求。除了单一技术或者算力指标，更要注重产品的综合竞争力是否均衡。百度的AI应用场景，为昆仑芯提供了验证和调优机会。就大模型而言，昆仑芯在产品定义上已经做了布局，昆仑芯2代AI芯片相较昆仑芯第1代产品大幅优化了算力、互联和性能，在百度内外部的大模型场景中都有落地，昆仑芯在研的下一代产品将为大模型和AIGC等应用提供更佳的性能体验。

“硬件和场景是双轮驱动的，场景催生新的技术方案，新的技术促使场景更好地发展。大模型作为一个趋势场景，其需求定义清楚了，设计和实现技术就会水到渠成。”昆仑芯科技相关人员告诉记者。

另外，无论是通用芯片还是定制芯片，抑或是通用、专用计算单元兼而有之，设计环节之后的制造、封装环节，也将作用于AI芯片的性能提升。

“无论GPU还是定制化AI芯片路线，Chiplet、3D堆叠等先进集成与封装技术将成为进一步提升大模型计算能力的重要技术手段，也将在未来AI芯片发展中起到重要作用。”贾天宇表示。

国内企业：需进一步增强软件及系统能力

国内AI芯片企业需要在软件、系统和生态层面进一步向国际领先企业看齐。

虽然ChatGPT还没有进入盈利阶段，但英伟达已经成为第一波吃到红利的企业。从2023年第一个美股交易日至今（截稿前最后一个交易日2月24日），英伟达的股价增幅超过60%，为处于下行周期的半导体产业增添了一丝亮色。在北京时间2月23日凌晨的财报发布中，英伟达创始人兼首席执行官黄仁勋表示，从初创公司到大型企业，对于生成式AI的多功能性与能力的兴趣越来越浓厚。英伟达将帮助客户从生成式AI和大型语言模型技术的突破中获取优势。

英伟达在AI芯片的先发优势和占比优势，固然有硬件性能的原因，但更关键的是软件生态的加持。在21世纪初，GPU的并行计算能力引起了学术界和产业界的关注。但是，开发者想要调用英伟达GPU的计算能力进行图形处理以外的任务，必须编写大量的底层语言代码，这对于习惯高级语言的程序员极其不便。2006年，英伟达推出CUDA平台，支持开发者用熟悉的高级程序语言进行编程，灵活调用GPU的算力。自此，GPU的使用范围不再局限于显卡，而是扩展到所有适合并行计算的领域。GPU与CUDA组成的软硬件系统，形成了英伟达的产品壁垒。

近年来，国内AI芯片企业在架构创新、算力性能、平台方案等领域涌现出一系列成果，但仍然需要在软件、系统和生态层面进一步向国际领先企业看齐。赵立东表示，针对大模型对于AI芯片的需求，芯片厂商一方面通过拆解大模型的系统级需求，快速迭代下一代芯片，从底层提升性能和支持效率。另一方面，要基于既有的芯片打造系统级方案，通过软件升级解决大模型加速遇到的内存容量小、通信占比高等核心痛点问题。

“要对标国际领先的AI芯片厂商，需要在三个层面开发优化：一是芯片升级，在算力、内存、微架构等层面针对大模型计算做优化；二是软件升级，从传统的单卡以及以单机多卡为主的支持能力拓展至万卡级别大集群支持，有效提供面向大模型支持的分布式计算、混合并行、内存优化等整体软件方案；三是系统方案，以AI芯片为核心，结合计算、存储、网络打造深度优化的系统级方案，面向大模型提供极致的性能和成本优势。”赵立东说。

据介绍，燧原科技已经基于千卡训练集群进行大模型训练，并将推理产品通过云服务商，为内容生成模型开发商提供算力支撑。基于系统级大集群大模型，燧原将持续创新迭代，重点聚焦生态建设，满足应用开发者对AI算力的强劲需求。

昆仑芯科技相关人员也表示，具体到软件生态，AI算法和应用开发者在构建AI应用和业务的过程中，需要一套成熟的编程语言，以及完善的软件工具集来快速迭代开发任务。昆仑芯SDK可以提供从底层驱动环境到上层模型转换等全栈的软件工具，已经适配百度飞桨、PyTorch、TensorFlow等主流框架和服务器，逐渐完善生态建设。“要实现像OpenAI的ChatGPT这样的大规模深度学习模型，需要大量的数据和算法优化，以及相关领域的专业知识。因此，要更好地实现ChatGPT的商用落地，需要相应的技术团队和研究机构，与AI芯片企业协同推进。”

产业观察

台积电全球建厂利弊难判

陈炳欣

2023年，台积电面临的挑战显然要超过去年。这一点从近期网上有关台积电的负面消息增多就可见一斑。此类消息归纳起来大致可分为两个方面。

一方面是全球半导体进入下行周期，台积电亦不得不面临客户削减订单、产能利用率不足的挑战。比如苹果公司日前就下修了给台积电的晶圆投片量，下修数量达到12万片，影响包含N7、N5、N4和N3等多条生产线。在日前召开的法说会上，台积电将2023年的资本支出由最初估计的400亿美元下调为320亿~360亿美元，显示出台积电管理层对2023年市场运行情况的谨慎态度。

另一方面则与台积电赴美国与欧洲建厂有关。2020年以来，台积电开始加快在全球建设生产基地的步伐。2020年5月，台积电宣布在美国亚利桑那州设厂，初期投资120亿美元，去年12月增加到约400亿美元，制程工艺也提升到3nm。2021年10月，台积电宣布在日本九州岛熊本县建设一座晶圆厂，投资86亿美元，制程工艺22nm/28nm。同时台积电还有在欧洲德国建厂的计划。但近期有关台积电在美建厂面临文化冲突的消息频频见诸报端，如“部分美国员工被分派多项任务时做不好，有时甚至拒绝任务分派”，“部分台积电员工抱怨外派到亚利桑那州新厂将比美国同事承担更多责任”等。另有消息称，台积电欧洲建厂的计划遭遇阻滞，将延后两年。

客观来说，第一个挑战对台积电影响并不大。市场的周期性变化是半导体行业的常态，数十年间半导体产业一直在供需不足与过剩之间波动。本次市场波动或许更加剧烈，台积电的产线或许会出现短期利用率下降的情况，但是凭借台积电庞大的生产能力、尖端的工艺水平以及对产业生态的掌控能力，并不会对公司的运营造成重大影响。最新消息显示，台积电近期获得来自英伟达、AMD与苹果的急单，第二季度产能利用率或将满载，显示出台积电巨大的市场掌控力。

但是，全球建厂的情况却有所不同。从台积电成立伊始，其生产基地就集中在中国台湾地区，聚集了台积电4座12英寸晶圆厂、4座8英寸晶圆厂和1座6英寸晶圆厂，另有4座后段先进封测厂，最新规划的2nm工厂也在此建设。这一点与英特尔等习惯于全球布局的做法不同。也就是说，台积电将生产基地从区域集中模式转向全球化布局模式，属于公司战略层面的一次重大调整。

而这样的调整对台积电管理层来说绝不轻松。台积电前任董事长张仲谋就不看好在美建厂之举，直言美国毫无优势。台积电管理层在法人说明会上也提到过，受人工成本、许可证、遵守法规及物价高涨影响，赴美设厂成本比在中国台湾地区高出很多。

成本的增加又会对台积电的自由现金流带来挑战。台积电在1995年的时候曾在美国投建8英寸厂Wafertech，成本比中国台湾地区高出5倍。该厂建成后在市场上也不具备竞争力，经历了长期的亏损之后才转为盈利。因此，外界完全有理由猜测，台积电在美欧建设的新工厂可能会像WaferTech一样，长期稀释台积电的利润，拖累其整体经营业绩。

台积电全球建厂的挑战还不止于资金成本，硬件工程师短缺、能源使用成本高企等问题同样存在。半导体专家莫大康指出，台积电制造虽然转向全球布点，但是能不能真正做好全球化布局却是一个挑战。此前，台积电在中国台湾地区深耕数十年，在当地有着得天独厚的区位优势。可是台积电能否在美欧、日本继续获得当地政府的足够支持，却需要打一个问号。私募基金柯克兰资本董事长杨应超更是直言，就经营角度来看，台积电在美国的投资“完全不合理”，“可能是因政治考量而被迫在美国设厂”。

去年第四季度巴菲特旗下公司曾经重砍台积电ADR股票达5176万股、减持幅度高达86%。此举曾令业界人士十分不解：台积电的稳定业绩与盈利能力，何以难得“股神”的青睐？问题的主因或许正在于此——台积电大规模的全球建厂计划存在较大不确定性，目前来看利弊难判。

英特尔公司高级副总裁、中国区董事长王锐：

全球半导体行业长期向好

本报讯 记者**沈丛**报道：近日，在2023英特尔中国战略媒体沟通会上，英特尔公司高级副总裁、中国区董事长王锐表示，虽然如今全球半导体产业处于周期调整阶段，但这并不妨碍市场的长期向好。

“从本世纪开始，半导体行业经历了三次比较大的周期调整，第一次是2000年左右，由互联网泡沫所导致；第二次是2008年，由金融危机所导致；第三次是当下，由疫情和各种其他原因所引起。每一次半导体行业都经历了下滑调整，但是这样的短期下滑并没有影响半导体行业一直增长的趋势。”王锐表示。

王锐表示，在半导体市场的每一次急速下滑后，随之而来的又是一个相对快速的回升，所以并没有感觉半导体行业长期发展有任何危机，而且多家分析机构都预测，到2030年，全球半导体市场规模有望达到1万亿美元。

如今，英特尔在IDM2.0的转型之路上也遇到了很多困难，特别是在周期调整阶段。面对转型挑战，王锐表示，越是危机越要投资创新、投资未来，大型公司的转型历程是长期的，不能看一时的成绩。英特尔通过发展代工业务、引领架构创新、拓展产品布局三大策略来推进。

此外，王锐表示，虽然中国半导体产业因疫情等原因也受到了些许影响，但中国仍是全球第一大半导体消费市场，且国家发改委数据显示，中国数字经济规模已连续多年位居世界第二，新的商业模式、应用场景、增长动能不断涌现，带动了中国经济高质量发展。2023年中国数字经济规模将达到52万亿元，2032年将达到100万亿元。

王锐称，随着数字经济持续壮大，半导体和计算行业将发挥越来越重要的作用。而在中国的高质量发展过程中，数字化和绿色化就好比是双轮驱动，协同呼应、相互促进。构建更美好的未来世界，离不开数字化、绿色化，而这些都以半导体和计算行业作为支撑。

“进入2023年，中国经济春潮涌动。”王锐表示，“我们对数字经济的前景充满信心。随着英特尔中国战略升级，我们将以更强的领导力、更整合的本地运营，更深入本土，为客户、为产业、为社会创造价值。”

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