“AI算力荒”解困的短、中、长策

AI算力不够，已经是老大难问题。ChatGPT等大语言模型，掀起了新一轮“大炼模型”和“炼大模型”的热潮，又让本就不足的算力雪上加霜。

适用于AI计算的GPU供不应求，买不到卡的企业和科研机构嗷嗷待哺，买到了卡的企业不得不面对涨价，也被架在成本的火上烤。

目前，英伟达的GPU是AI计算最主流的硬件。有业内人士估算过，GPT-4模型仅满足日访问量的计算需求，就需要六万张英伟达A100，每一张价格在60-100万人民币，而A100和更强大的H100，这两款芯片此前都被列入了美国的禁止出口清单里。还好在英伟达的游说下，又能够在2023年3月1日之前继续给大陆提供A100产品。

现在缓冲期已经到了，AI算力的局面是蜡烛两头烧，一边是越来越少的海外高性能芯片，一边是越来越多的大模型训推需求，究竟怎么办呢？

我知道很多普通网友很急，觉得又被卡脖子了，但大家确实不用那么急，为AI算力荒解困，业内其实已经探索出了短策、中策和长策。今天就来讲讲，如何见招拆招。

短策：开源节流，过紧日子

实事求是来说，最关键的AI芯片GPU被国际厂商垄断，市场占有率达到80%以上。而国产厂商虽然也有相应的产品，但要么还没有量产，无法满足规模应用的需求；要么性能跟海外先进产品的差异很大，实用中大概只能达到60%左右的水平。而中美博弈，短期内是不会有方向性的转变的，未来高性能芯片的封禁可能是常态。

所以结论就是，放弃幻想，接下来要准备过AI算力的紧日子了。

目前业内的应付办法有两种：

一是开源。

对于N卡，继续买，抓紧囤货。

国内头部互联网公司，尤其是已经推出了大模型的企业，都会进行20%左右的战略备货，储备了万片级别的英伟达A100芯片，所以算力基础都不差。某一线云厂商透露，现在自家有10万片的A100，能够满足好几个客户复现GPT的AI算力需求。

此前国内区块链火爆，矿机厂商和数字货币贩子也买了大量英伟达GPU用来“挖矿”，听说也被一些AI公司紧急收了过来。而且，虽然缓冲期已到，但只要交付模式上商务合规，还是有可能继续用到先进产品的。

对于国产芯，加快上马，落地部署。

目前，国内的头部科技公司，已经开始筹备或落实，将寒武纪MLU370/590、燧原、百度昆仑芯、阿里平头哥等，部署到算力集群中，尽管占比还比较少，但国产芯的使用和适配已经开始了，随着合规及产能提速，也能满足AI并行计算的需求。

芯片的国产化替代，这一步迟早要走，大模型成了那个提前上马的变量。

二是节流。

既然AI基础设施跟大模型建设热潮之间有剪刀差，咱能不能把钱花在刀刃上呢？还真能。

OpenAI选择训大语言模型来实现通用人工智能，超大规模参数来达到“智能涌现”，堪称为“败家子儿式创新”。微软公司的博客中透露，2019年微软宣布向OpenAI投资10亿美元，为了让OpenAI能训练出越来越强大的模型，将28.5万个CPU和10000个GPU联接起来，造了一个超级计算集群。

背靠家大业大的微软，这么烧无可厚非。但放到中国语境下，或许我们还可以想一想，等这波GPT式热度消退，基础模型已经成型，那些烧钱打造的算力基础设施该何去何从？百亿、万亿参数的大模型，部署到工厂、矿区、城市之类的行业场景，是不是也有相应的算力支撑？

冷静下来后，为AI算力“节流”，才是大模型真正落地的必经之路。

节流，有两个办法：一是大模型“瘦身”，通过剪枝让模型稀疏化、知识蒸馏对模型进行压缩，通过权重共享来减少参数量……总之，一旦一种技术路线被证明有效，那么很快就会有多种技术手段对其进行优化，让模型成本大幅下降。

最近加州大学伯克利分校打造的icuna（小羊驼）模型，就只用8张A100训练了一天时间，将130亿参数模型的训练成本，从1000美元降低至300美元。所以，模型“瘦身”可以有效减少单个模型的算力资源消耗。

二是硬件“压榨”，通过端到端优化，从AI芯片中“压榨”出更多性能，把有限的硬件用到极致，也是一种节流。

举个例子，主流的大模型，包括ChatGPT、GPT-4，都是以Transformer架构为主，微软通过ONNX开源推理引擎的优化，可以将大语言模型的推理性能提高17倍。某国产芯片厂商针对Transformer结构特性进行优化，将芯片性能提升到原本的五倍以上，压缩显存30%以上。资源利用率更高，相当于在AI训练和推理时单位部署成本更低了。

总的来说，面对短期内“AI算力荒”，我们只能接受现实，正视差距，广积粮食，开源节流。

承认这一点没有什么好憋屈的，毕竟中国AI从零起步，到今天能跟no.1站在同一张牌桌，这才是我们熟悉的故事。

中策：兼容并包的全国算网

一双眼睛全盯着高性能GPU，会发现差距简直无从弥补，还在越拉越大。英伟达、英特尔、AMD等已经将AI芯片支撑推进到了4nm，而光刻机禁运，制程追不上，国内14nm制程将将量产，巧妇难为无米之炊。

但换个角度，可能就柳暗花明又一村。

大家可能还记得，去年东数西算工程正式启动，新型国家算力网络成了新的热点，我们也做过很多报道和分析。

当时我们就提到：实现先进算力的一体化、集约化、多样化供给，是“全国算力一盘棋”的题中之义。而这只是全国一体化大数据中心协同创新体系中的一环。

今天看来，通过几年时间，构建数网、数纽、数链、数脑、数盾，对于AI大模型的数据、算力、联接、商业化等多种挑战，是一种持续释放影响的“中策”。

本质上说，AI模型的训练推理是CPU+加速芯片。GPU的高并行性，可以成规模地处理AI工作负载，为深度学习加速，在进行模型的训练和推断时会更具有效率优势。英伟达的A100，在AI推理时吞吐量是CPU的249倍。

但这并不意味着，CPU不能做并行计算，加速芯片没有其他选择。

生成式AI的模型训练通常是在云端完成的，云端芯片以CPU+GPU异构计算为主。一些小型的模型是完全可以CPU训练的，可能训练速度慢一点，但确实可以用。

此外，ASIC芯片也很适合AI计算，目前还没有明显的头部厂商，国产厂商还有机会，很多企业开始推出自研的ASIC加速芯片。比如谷歌的TPU、英特尔的DPU、国内寒武纪的NPU、地平线的BPU等。

模型训练好之后，需要结合数据计算“推理”出各种结论。手机人脸识别认出“你是你”这个环节就是“端侧推理”，iPhone将相册上传到云端进行用户行为分析就是“云端推理”。

相对模型训练而言，推理阶段处理的是小批量数据，这时候GPU并行计算的性价比就不那么明显了，尤其是在边缘和终端大规模部署AI算法，是难以承受如此高的成本的。FPGA、ASIC等加速芯片，协助CPU来满足推理的计算需求，是具有竞争优势的。

这跟算网有什么关系呢？

划重点，在全国一体化算力网络体系的各种政策文件中，“算力多元化”的出现频率是非常高的。

多元化，一方面体现在多种计算架构，支持CPU、GPU、ASIC、FPGA等多种芯片的混合部署，充分发挥不同体系架构的优势。

另一方面，体现在多种算力，模型训练、边缘推理、数值模拟的不同场景需要不同的算力，AI算力、通用算力、高性能算力等综合配给，才能很好地支撑各类行业AI应用。

正如微软Azure高性能计算和人工智能产品负责人Nidhi Chappell所说，“让更大的模型训练更长的时间，意味着你不仅需要拥有最大的基础设施，还必须能够长期可靠地运行它”。

要长期可靠地保障AI算力资源，自然要发挥中国智慧——东方不亮西方亮，黑了南方有北方。通过全国一体化算力网络的建设，充分推动多种架构的落地部署，国产芯片的同步发展。

未来几年算网成型，对于保障算力供给，应对不可抗力，会起到非常关键的作用。

长策：长出那双手

理想化的角度来说，缓解AI算力荒的终极解决思路，肯定是造出对标国际一流水平的自研芯片。但这就像“中国什么时候能有自己的OpenAI”一样，是一个漫长的畅想。

漫长，指的不只是足够长的时间和耐心，给半导体行业足够多的钱，还要能吸纳全球顶尖的技术人才、全球优质的风险投资机构、计算机基础人才的培养、允许失败试错的创新氛围和兜底机制、充分信息化数字化的优质数据基础、繁荣的商业市场……这是一个社会工程。

那么，我们是不是就得一直这么憋屈呢？

当然不是。咱们除了“脖子”，还有“手”啊，就不能用自己的长处，去卡别人的脖子呢？

这双手，可能是新的计算体系。

今天，经典计算的“摩尔定律”已死，英伟达提出的“新摩尔定律”也面对AI算力供需的剪刀差有心无力。

光计算、类脑计算、量子计算等新计算体系，正在成为各国的重点布局方向。以量子计算为例，有望彻底解决经典计算的算力不足问题。

当然，总想着“弯道超车”大概率会翻车，提到这点只是想提醒一下，不要只盯着CPU/GPU这些已经被卡脖子的焦点领域，而忽视了其他路线，将路走窄了。毕竟谁能想到，当年游戏宅们追捧的显卡能卡住今天的AI计算市场呢？

英伟达GPU被发现可以用来跑AI之前，只有游戏发烧友会对N卡津津乐道，这种“无心插柳柳成荫”的结果，恰好说明了多技术路线创新的重要性，或许会在某条路上就发现惊喜。

这双手，也可能是产业生态。

AI本来就是一个工程性、交叉性很强的学科，AI芯片要充分释放能力，除了更高制程的工艺，也离不开深刻理解行业用户的使用习惯，才能把软硬件做到位。

英伟达GPU的主流地位，与CUDA生态有直接关系。而CUDA的护城河正是软件堆栈，可以让研究人员和软件开发者更好地在GPU上编程，构建应用。

如果说AI算力问题，国产硬件的差距是明线，软件生态就是那条更难的暗线。

首先是软件，就拿大模型来说，下接底层算力硬件、操作系统和框架，上接行业应用，需要提供一整套从开发、应用、管理的全流程服务和工程化方法，而目前积累了全面技术栈的只有少数国内头部企业。

其次是生态，CUDA生态经过多年积累，在AI计算的绝对主导地位，而国内几个头部企业都有各自的AI生态。我们就曾遇到过这样的采访对象，一个工业企业的数字化案例中，既有A生态的一些软硬件，又有B生态的一些解决方案。多个生态并存，增加了产业的选项自由和安全感，也难免带来适配上的复杂度，以及一些重复性工作。

国产芯片硬件的突破或在旦夕之间，但软件生态的爆发却需要漫长的时间去酝酿。而一旦生态如同齿轮一样转起来了，吸纳更多产业资源和人才力量，很多软硬件创新都能加速发展。

大语言模型的这波热闹中，我有听到一些声音，说中国AI行业“浑身上下都是脖子”“一卡脖子就翻白眼，一开源就全球领先”。

很能理解大家“怒其不争”的心情，但实事求是地看，中国AI走到今天，靠的从来不是谁的施舍，是真的有一群人，在卡脖子时没有翻白眼，而是与禁令抢时间，与海外合作伙伴想对策，把国产芯片扶上马送一程。

如果说，无需担心“AI算力荒”，这是一种无视现实差距的盲目自信。但也确实不用一提算力、一提芯片，就萦绕着“生于忧患死于安乐”的焦虑气息。

星光不问赶路人，与其花时间自怜自哀，不如在有限的规则里，做力所能及的事。短策、中策、长策久久为功，这才是中国缓解“AI算力荒”的真实选择。