如何驯服AI大模型“能耗巨兽”

本报记者 张心怡

AIGC（生成式AI）及其背后的大模型，是不折不扣的“能耗巨兽”。在部署大模型的过程中，AI工作负载带来的功耗和成本挑战，已然成为产业链的“阿喀琉斯之踵”。近期，蚂蚁集团联合多所高校发布的《围绕绿色计算发展机遇的一项调查》（以下简称“调查”）指出，绿色措施、节能人工智能、节能计算系统和可持续发展的人工智能是构建绿色计算的四个关键，强调了人工智能的节能降耗对于计算产业的意义。与此同时，恩智浦、英特尔、英伟达等半导体企业也在积极应对大模型带来的能耗挑战，从架构创新、软硬件协同方案、网络平台等多个层面，为大模型部署提质降耗。

为“大脑”分担任务 半导体厂商创新方法论

大模型作为近年来最重要的新兴计算场景，对芯片的算力和内存需求都提出了极大挑战。

大模型作为近年来最重要的新兴计算场景，对芯片的算力和内存需求都提出了极大挑战。北京大学集成电路学院研究员贾天宇向《中国电子报》记者表示，在传统摩尔定律难以为继的背景下，半导体企业需要通过利用架构设计、制造工艺、先进封装等多个层级协同的设计手段来满足大模型的计算需求。例如，通过采用异构计算、存算一体、三维堆叠等先进计算理念降低芯片的能耗，提高单一芯片的性能和能效。此外，针对大模型的大算力需求，芯片的可扩展性也变得尤为重要。重视多芯片的系统级扩展和互联技术，也成为半导体企业在技术研发过程中不可忽视的重要内容。

“在控制大模型带来的能耗和成本方面，半导体企业可以通过优化计算架构、提升能效等措施，降低大模型的能耗和成本。但针对大规模的大模型训练，需要与软件生态、数据中心建设等多环节配合，共同为企业和开发者提供更具成本效益的解决方案。”贾天宇说道。

如贾天宇所说，架构创新素来被视为实现芯片技术突破的发动机。在采访中，多家半导体企业负责人都向记者阐述了基于架构创新和优化为大模型增效降耗的思路。

长期以来，人工智能及其热门应用常常被喻为“大脑”，比如自动驾驶被喻为“车轮上的大脑”。但是，人脑并不是仅仅依靠大脑工作，脑干、小脑也承担了许多任务，比如控制心跳和体温，保持身体的稳定性、协调性等。如果大脑进行所有的决策和控制，人体就很难负担大脑需要的能量。

比照包含大脑、小脑、脑干的人脑系统，恩智浦半导体执行副总裁兼首席技术官Lars Reger为记者描述了一种更加高效节能的计算体系：以AI算力芯片为“大脑”，进行高效能的计算和加速；以网关处理器为“小脑”，实现不同功能区的互联和集成，以及异构网络中的数据处理；以MCU、感知芯片、联网芯片作为“脑干”，实现感知和实时任务处理。

“有的车厂CEO告诉我们，由于能耗的问题，他们不得不降低汽车的续航，这主要是因为现有的技术架构不够高效。所以在开发系统时，我们要确保在正常运行的情况下，不需要时刻激活大脑的功能，只要小脑就可以了。比如我们在开车的时候，更多的是基于平时的规则和训练下意识地驾驶，只有遇到挑战的时候才需要用到大脑。”Lars Reger向《中国电子报》记者表示。

据悉，恩智浦已经将这套计算理念集成到了智能驾驶的技术架构中，架构中“小脑”和“脑干”部分是由恩智浦的技术来保障的。S32G作为网关处理器，扮演小脑的角色；S32K等用于车身各部分控制的MCU、S32R等传感芯片、S32Z等域控制和区域控制芯片以及以太网连接产品，共同构成“脑干”，以应对能耗对续航的局限和挑战。此外，《中国电子报》记者获悉，恩智浦即将推出最新的5纳米级旗舰产品，能够把数据从车辆传输到控制单元。

跟着最新算法走 随时提供软硬件协同方案

由于无须添加更多的训练数据，基于提示的微调能节省大量的时间和成本，以实现更加绿色的计算。

随着业界对大模型的研发和部署逐步深入，新的压缩、调优方式层出不穷，从算法和框架层面实现对大模型的提效降耗。但算法、框架的更新，需要半导体厂商及时跟上，提供相应的软硬件协同方案，才能实现新算法、新框架的部署。

比如，大模型的调优方式，正在从完全微调走向更加简捷、低功耗的微调方式。蚂蚁集团联合多所高校发布的《围绕绿色计算发展机遇的一项调查》中提到，传统的调优方式会微调所有模型参数，使通用大型语言模型适应特定的目标任务，这一过程称为完全微调。然而，当大模型的参数规模从数百万、数千万向数亿乃至数万亿规模发展时，完全微调会带来更长的程序代码运行时间和高昂的存储成本。为了解决这个问题，更加简捷的微调方法已成为大模型的研究热点。比如PEFT（高效参数微调）仅更新模型参数的子集或附加的模块，使大模型适配下游任务，以降低微调所需的计算和存储资源；基于提示的微调则训练大模型根据特定提示或指令生成响应，引导大模型做出更理想的决策和行动。由于无须添加更多的训练数据，基于提示的微调能节省大量的时间和成本，以实现更加绿色的计算。

围绕基于提示的微调等最新微调方式，英特尔基于AI加速引擎和配套的软件工具，进一步减少了微调所需的工作量。在第四代英特尔至强可扩展处理器中内置了矩阵乘法加速器AMX，能够更快速地处理BFloat16（BF16）或INT8数据类型的矩阵乘加运算，从而提升模型训练和推理的性能。尤其对于ChatGLM-6B等在开源微调代码中支持CPU自动混合精度的大模型，开发者在启动微调时加入CPU自动混合精度的使能参数，就可以直接利用矩阵乘法加速器提升大模型的微调计算速度。

英特尔院士、大数据技术全球CTO戴金权向《中国电子报》记者表示，解决大模型功耗和成本压力的关键，是根据不同大模型的最新技术需求，提供软硬件协同的支撑方案。

“我们可以通过提供跨平台的XPU架构和开放的AI软件架构及优化，为预训练、微调、推理等不同的AI和大模型应用，提供高效的开发部署平台。”戴金权告诉记者。据他介绍，英特尔在硬件架构方面，正在基于Gaudi AI加速器、数据中心Max GPU、内置AMX加速的第四代至强处理器，以及边缘和客户端处理器为用户提供XPU异构计算支持。在软件架构层面，英特尔和社区在PyTorch、HuggingFace Transformers等开源项目开展合作，加入了在Intel XPU平台上的众多优化。同时，英特尔的开源软件工具IPEX、BigDL、INC、OpenVINO等，为大模型提供了专门的支持和加速。

加速与降耗并行 打造更加节能的基础设施

大模型的到来，也在倒逼算力系统的创新，为基础软硬件的各个节点带来新的市场机会。

数据中心是AIGC和大模型主要的基础设施之一，也是节能技术的重点发力领域。随着全球的热点IT应用地区都在强调绿色数据中心，如何在降低能耗的同时释放更高效能，成为英伟达、AMD等数据中心芯片供应商的必答题。

围绕数据中心的减碳需求，AMD宣布了到2025年为人工智能训练和高性能计算应用程序带来30倍能效提升的目标。据AMD测算，30倍的能效提升将在2025年节省数十亿千瓦时的电力，使系统在5年内完成单次计算所需的电力降低97%。这一方面需要核心制程的提升，另一方面需要架构的改进和技术的创新来提高算力。比如基于一颗第三代AMD EPYC服务器处理器和四个AMD Instinct MI250x GPU的加速节点，AMD实现了在2020年的基准水平之上提高6.79倍能效。

英伟达也将加速计算作为降低功耗的主要策略。加速库是英伟达加速计算的核心，目前英伟达面向计算机视觉、数据处理、机器学习和AI等领域布局了300个加速库和400个AI模型。

除了提升计算单元的能效，计算单元构成的AI集群，也对数据中心的整体功耗有着重要影响。在计算单元互联以构建AI集群，以及集群、设备互联构建AI计算网络的过程中，会产生大量的网络数据。若采用传统的以太网架构，会导致数据流的拥塞和延迟，使系统无法有效利用GPU，从而增加了大模型训练的时间和成本。

在网络平台层面，英伟达推出专门面向AI负载的以太网架构Spectrum-X。该架构基于内置Spectrum-4 AISC芯片的交换机与Blue-Field DPU，提升AI集群的资源利用和数据传输效率。在对GPT-3的训练中，NVIDIA Spectrum-X网络平台相比传统以太网网络架构，实现了1.7倍的加速效果，尤其针对数据中心常用的功率封顶措施，Spectrum-4 ASIC能够简化网络设计，提高了每瓦的性能，帮助数据中心控制网络功率预算。

从计算架构的革新，软硬件方案的更新，到基础设施算力和互联方式的迭代，半导体厂商正在从多个维度缓解大模型带来的能耗压力。而大模型的到来，也在倒逼算力系统的创新，为基础软硬件的各个节点带来新的市场机会。

微缩工艺和先进封装 助推芯片制程前行

本报记者 沈丛

自AI快速走热的这一年多来，不断增长的算力需求对本就陷入瓶颈期的芯片制程技术带来了挑战。微缩工艺可以通过不断缩小晶体管尺寸来提高芯片集成度和性能，但随着制程技术越来越接近物理极限，微缩工艺的发展空间越来越小。Chiplet等先进封装技术，可以通过将多个芯片集成在一起提高算力和性能，同时还可以降低成本和功耗。因此，在AI时代，先进封装技术越来越受到关注。未来，芯片制程技术需要在微缩工艺和先进封装之间取得平衡，以满足不断增长的算力需求。

自2012年以来，深度学习被广泛应用，AI算法的网络结构持续高速增长，单一的AI算法对算力的需求增加了30万倍。高速扩张的算力需求，使多次被预言放缓乃至完结的摩尔定律，重新获得了生命力。台积电（中国）有限公司副总经理陈平在日前举办的2023中国临港国际半导体大会上表示，随着对算力的需求越来越高，业内对先进制程芯片越发热衷。

OpenAI CEO奥特曼曾预测，对于AI时代的摩尔定律来说，集成电路上可以容纳的晶体管数目在大约每18个月会增加一倍。其发展周期与此前摩尔定律中的18～24个月相比，略微超前。随着AI时代的到来，摩尔定律的演进反而有所提速。

陈平认为，在关注芯片制程缩小的同时，也要关注芯片的算力和能效比，包括新型晶体管和材料、光刻技术和DTCO（设计与工艺协同优化）的进步、电路和架构的创新、先进封装和STCO（系统工艺协同优化），以及软件优化等。这些因素的协同作用将推动半导体技术不断进步，实现更高性能、更低功耗和高能效比的芯片设计。

中国半导体行业协会集成电路设计分会理事长、清华大学教授魏少军认为，除了缩小芯片制程尺寸外，还可以利用三维混合键合技术对存储器晶圆和逻辑电路晶圆进行异质集成，从而提升芯片的算力效率。这种集成方式对于逻辑电路的晶圆没有代工厂及工艺节点的限制要求，具有更高的灵活性和适应性。而存储器晶圆由DRAM晶圆厂制造，保证了存储器的品质和性能。混合键合晶圆加工则在晶圆代工厂制造完成，实现了工艺的高效整合。这种集成方式将不同工艺的晶圆优势结合起来，提升了芯片的性能和功能，满足了人工智能等领域对于高算力和低能耗的需求。

魏少军认为，为了增强芯片的灵活度，实现算力的合理分配，还可以将软件定义芯片与异质堆叠集成相结合，构建软件定义近存计算芯片技术。软件定义芯片是一种先进的芯片设计技术，通过将任务处理空间并行化，实现硬件资源的时分复用，从而提高了芯片的处理效率和性能。而异质堆叠集成技术则通过将存储单元和计算单元紧密集成在一起，缩短了数据传输距离，降低了数据传输能耗，进一步提升了芯片的性能。这种技术能够更好地满足AI时代对算力和能效比的要求，同时也提高了芯片的安全性。

在人工智能蓬勃发展的背景下，Chiplet逐渐崭露头角，备受业界瞩目。中国半导体行业协会副秘书长兼封测分会秘书长徐冬梅指出，由于人工智能和HPC高性能计算领域需要处理大规模数据和复杂计算，对芯片设计规模的要求极高，因此，这两个领域对于Chiplet技术的需求更为迫切。

随着ChatGPT等高普及度的AI技术不断发展，其背后的芯片需求也日益旺盛。数据显示，到2024年，Chiplet芯片的全球市场规模将达到58亿美元，2035年将达到570亿美元，显示出Chiplet市场的巨大潜力和增长空间。尽管Chiplet技术的发展前景被看好，但它并不能完全取代先进制程技术。对此，陈平表示，尽管通过Chiplet将几个芯片组合在一起可以扩展芯片的功能，但这种组合方式不能完全取代先进制程技术。Chiplet虽然能够实现更复杂的计算和数据处理能力，但并不能改变芯片的品质，也就是指能效比和算力密度。因此，在业界追求更高性能和更低能耗的过程中，仍需不断提升芯片制程，与Chiplet互补提升。

龙芯发布新一代CPU 采用自主指令系统龙架构

本报讯 记者**沈丛**报道：11月28日，龙芯发布3A6000处理器。据介绍，龙芯3A6000采用我国本土设计的指令系统和架构，是我国本土研发、自主可控的新一代通用处理器，可运行多种类的跨平台应用，满足各类大型复杂桌面应用场景。它的推出，标志着我国本土研发的CPU在自主可控程度和产品性能方面达到新高度。

记者在发布会上了解到，龙芯3A6000处理器采用龙芯自主指令系统龙架构，是龙芯第四代微架构的首款产品，主频达到2.5GHz，集成4个最新研发的高性能LA664处理器核，支持同时多线程技术，全芯片共8个逻辑核。综合相关测试结果，龙芯3A6000处理器总体性能达到英特尔酷睿10代四核CPU水平。龙芯表示，下一步将尝试基于成熟工艺、通过设计性能优化，达到英特尔、AMD先进工艺CPU的性能。

在半导体芯片领域，指令系统是一切软硬件生态的起点。其中，ARM和x86最被业界熟知，x86支撑Wintel体系（Windows+Intel），ARM支撑AA体系（Android+ARM）。与此同时，随着物联网、AI新兴领域的兴起，RISC-V和MIPS两大精简指令集架构也频繁地出现在大众的视野内。

在发布会上，龙芯中科董事长胡伟武表示，一直以来，龙芯中科从IP研发、指令系统、软件生态等方面夯实本土信息产业基础。龙芯3A6000走出了一条基于成熟工艺，通过设计优化提升性能的道路。

胡伟武表示，由本次发布的桌面处理器龙芯3A6000、在研服务器处理器龙芯3C6000和移动桌面终端处理器2K3000构成的龙芯“三剑客”将在2023—2024年陆续推出，并在特定的开放市场具有一定竞争力。未来将在此基础上开展结构优化以及工艺升级。

此外，在发布会上龙芯还推出打印机主控芯片龙芯2P0500，这是国内首款基于自主指令系统的打印机主控芯片。