AI+5G叠加赋能 半导体大厂为终端大模型筑基

本报记者 张心怡

大模型正在全面拥抱智能终端。在近期于上海召开的进博会上，记者看到了大模型和AIGC被装入手机、笔记本电脑等终端，无须云端参与，也可以轻松完成文字生图、线稿渲染、旋律续写、文字扩写、照片扩充、音质提升乃至编程等操作，在提升消费者与终端交互体验的同时，也为设计师、艺术家、工程师等职业的工作流程带来极大便利。与此同时，5G与AI的结合，使云端、边缘云的训练和计算资源都能够服务于终端，进一步提升了大模型在终端上的执行效率和能力上限。

在大模型走进终端的背后，高通等半导体厂商正在围绕AI算力的提升、功耗和成本控制，以及5G赋能等大模型部署的关键需求，提供芯片底座支撑。

提算力降能耗 半导体大厂备战端侧大模型

在进博会的高通展台上，记者看到了一系列手机大模型的应用demo（演示）。用数据线将键盘（手卷琴或电子琴）与手机连接，随手弹一段旋律，大模型会将旋律续写为一段乐章；用手机前置摄像头瞄准电竞选手，大模型会根据微表情评估选手的情绪能量值，并生成包含激动、专注、压力等关键指标的五维雷达图。这些功能不需要云端的参与，仅凭手机中的AI大模型就可以实现。

笔记本电脑作为生产力与娱乐性兼具的便携设备，也成为大模型的入驻目标。高通在骁龙峰会期间展现的AI PC应用demo令现场听众心潮澎湃，包括起草完整的电子邮件、转写会议记录，以及文本生成、图像生成等。几个月前还只能在数据中心实现的AI功能，已经集成到PC，让每一个用户都能拥有个人数字助理。

而支撑这些功能的计算中枢，是高通第三代骁龙8移动平台（以下简称“骁龙8 Gen3”）和高通面向AI PC推出的骁龙X Elite芯片平台。据工作人员介绍，骁龙8 Gen3能够在手机端运行100亿参数的大模型，骁龙X Elite支持在PC端运行超过130亿参数的生成式AI模型。

在大模型部署过程中，有一项挑战贯穿始终——随着AI任务越来越复杂多样，计算系统该如何兼顾性能和功耗？尤其对于手机、笔记本电脑等强调续航能力的移动终端，性能的提升必须与可接受的功耗并肩而行，且相应的指标会越来越严苛。

在这种趋势下，通用计算单元搭档专用计算单元的异构计算模式，成为终端芯片供应商的着眼点。具体来看，CPU、GPU等通用计算单元，能够处理多样化的计算任务；而NPU等专用计算单元，专门处理和加速AI工作负载。相比通用计算架构，这种“通用+专用”的架构，能够更好地平衡高性能、实时性和低功耗需求，提供AI计算的更优解。

这也是为什么骁龙8 Gen3和骁龙X Elite都采用了以“CPU+GPU+NPU”为主力单元的异构计算模式。

其中，高通Hexagon NPU是AI引擎的核心，在加速AI负载的同时进一步释放CPU和GPU的算力，提升多任务运行的实时响应能力和能效表现。针对业界关心的性能与功耗的平衡问题，高通从骁龙Gen2开始就在Hexagon NPU中采用了业界首创的微切片推理技术。传统的神经网络推理过程，是将整个神经网络加载到NPU中，运行一层网络并写入内存，再运行下一层网络。这种逐层运行和写入的方式，一方面会造成大量能耗，另一方面只能一次激活一个加速器，拖慢了推理过程。而微切片推理技术，将神经网络切割成大量切片，标量、矢量和张量加速器同时运行多个切片，再读取到内存中，从而在发挥所有加速器能力的同时消除了大部分的内存读写过程，使NPU能够以更低的功耗更快地完成推理。

全新骁龙X Elite的NPU进一步提升了处理能力，并引入了更多的降耗技术。新一代Hexagon NPU的张量加速器将矩阵处理速度提升了2.5倍，进一步提高了标量和矢量加速器的时钟速度，共享内存规模也增加一倍，能够容纳参数更为庞大的神经网络。同时，新一代NPU增加了全新的供电系统，能够按照工作负载适配功率，还为张量加速器增加了独立的电源传输轨道，以实现更加理想的能效表现。

而高通新一代处理器搭载的CPU和GPU，也显著提升了AI推理性能和能效表现。以骁龙X Elite为例，在AI算力方面，X Elite搭载的Oryon CPU将AI推理性能提升了5倍，并针对时延敏感型工作负载进行了优化；高通Adreno GPU的AI计算性能提升了50%。在能效方面，Oryon CPU的单线程性能超越了ARM架构竞品，且实现相同水平性能时可以减少30%的能耗；对比x86架构竞品，Oryon CPU实现相同性能时可以减少70%的能耗。Adreno GPU也在性能和功耗之间寻求平衡。在面向PC的热门3D图形基准测试中，Adreno GPU的性能达到x86架构集成GPU竞品的2倍多，且在实现竞品峰值性能同等水平时可以减少74%的功耗。

性能决定大模型跑得多快，功耗决定大模型运行多久。基于性能和功耗平衡兼顾的原则，高通正在为端侧大模型筑基赋能，让AIGC更加触手可及。

5G加持 混合AI架构释放终端能力

虽然端侧AI优势明显、端侧大模型的布局热度正酣，但终端芯片受到体积和功耗的限制，对于更大的模型和调优等重度负载并非最优解，更加高阶或复杂的AI用例往往诞生于云端大模型。

那么，当终端用户偶尔需要超越端侧大模型的功能时，该如何获取？高通在《混合AI是AI的未来》白皮书中，提出了在云端和边缘终端之间分配并协同处理AI工作负载的混合AI架构。这种架构可以根据模型和查询需求的复杂度等因素，在云端和终端侧之间分配处理负载。

5G或未来6G的连接，是实现混合AI架构的关键。当终端大模型判断出现了终端无法处理的工作时，可以通过5G连接，将数据上传到云端进行运算，之后再将结果推回到终端。5G或6G的高速连接，能够保障数据处理和结果反馈的实时性。如此一来，云端、边缘云的训练和计算资源，都将为终端大模型服务，保障了用户体验。

虽然6G时代尚未到来，但行业已经行至5G第二个阶段的演进——5G Advanced（以下简称“5G-A”）。相比5G，5G-A可实现网络能力的10倍提升，支撑10Gbps体验、全场景物联、通信感知一体、L4级别自动驾驶网络、绿色ICT等业态，进一步开发和释放5G的潜能。

最新一代骁龙8 Gen 3集成的骁龙X75是业界首个5G Advanced-ready调制解调器及射频系统，面向智能手机、移动宽带、汽车、计算、工业物联网、固定无线接入和5G企业专网等细分领域，提升网络覆盖、时延、能效和移动性体验。在11月8日于乌镇举办的“世界互联网大会领先科技奖”颁奖典礼上，高通5G Advanced-ready调制解调器及射频系统X75获得“世界互联网大会领先科技奖”。

内生AI被视为5G-A的关键技术之一，使网络更智能、灵活地按需提供功能。骁龙X75是业界首次在调制解调器及射频系统中采用专用硬件张量加速器，将AI性能提升至前代产品的2.5倍以上。骁龙X75还引入了5G AI套件，为连接速度、移动性、链路稳健性等5G性能和特性的优化提供更大空间。

5G-A的加持，除了让手机、PC等终端能够高效运行大模型，也为大模型进入XR（虚拟现实、增强现实等）、汽车等终端带来更多可能。5G-A提供的万兆体验，能够更有力地支撑VR/AR和车联网等需要高带宽、低时延的应用。例如微软基于云流媒体和第二代骁龙XR2芯片平台，在XR终端侧运行先进的算法，实现了将Windows体验安全传输到Meta Quest设备，用Azure远程渲染MR终端内容等用例，更有效地融合了XR云端和客户端内容。

混合AI架构以及5G+AI的协同工作，也为开发者在本地和云上的无缝开发大模型相关应用带来便利。今年9月，微软将AI助手Copilot接入Windows11，将生成式AI引入更多应用程序。微软Windows和设备副总裁帕万·达武鲁里在2023骁龙峰会上表示，终端侧AI促进了客户端和云端之间的平衡，将使开发人员能够使用本地和云计算编写混合AI应用。“未来，我们需要一个能够模糊云端和边缘端芯片界限的操作系统，在合适的地方使用合适的芯片，从而实现最佳体验。这将让Windows成为一个跨应用、内容和图形的编排系统。”

万物互联，万物智能。AIGC提升了终端的智能上限，而5G-A和未来的6G拓展了智能的触达范围，两者将一起为终端的使用和交互体验带来更多可能。

AI大模型进入生态竞争

（上接第1版）

而在训练数据方面，GPT-4在训练数据的广度和深度方面都具备显著的先发生态优势，也可以在调用时实时搜索Web数据，这是Grok所不具备的。Grok的差异化优势在于它可以更有效地访问X平台（即之前的Twitter）的信息，这赋予了Grok实时社交平台数据洞察及多样化的生成风格。

“巨量、实时且独特的数据是该模型的基础，可以实时从推文中获取最新知识，主打一个实时性，也就是说，这个模型还在不断学习和更新。同时，Grok有着不走寻常路的幽默模式语言风格，马斯克的个人风格在其中凸显，开发语言是Rust以及深度学习框架新秀JAX，分布式的架构让系统避免了大模型的系统性风险。”王娟表示。

“Grok的定位并不一定是GPT的竞品，GPT定位是全场景的AI平台，Grok更多是适合X的大模型应用平台，两者的定位不一样。Grok是一个不错的补充。”开放原子开源基金会TOC主席谭中意对《中国电子报》记者说道。

GPT生态选择题：开源还是闭源？

无论是Grok与OpenAI的对决，还是国内各种大模型的比拼，想要突出重围，生态建设是重中之重。这也是OpenAI、阿里云、腾讯等厂商着急做模型商店的根本原因。

“OpenAI的模型商店从理念上与国内大部分厂商的战略规划是一致的，都希望通过模型商店的方式打造生态，一方面便于中小厂商引流，另一方面更利于客户选择和开发人员与厂商协作，以此实现围绕GPT大模型更强的黏性，最终推动营收的增长。”戴鲲说道。

但同时他也指出，由于厂商自身的市场定位与发展战略不同，模型的覆盖、模型被集成的机制、面向的客户群体、针对的行业细分与商业场景、对底层云平台的选择、与相关云服务的集成、计量计费、许可与定价模式等各个方面肯定会存在差异。

“虽然都要做模型商店，但OpenAI要做的模型商店和腾讯、阿里等要做的模型商店差异很大。腾讯、阿里云提供的是MaaS服务（模型即服务），它们的模型商店支持多种不同的模型（开源和闭源的模型都包括）；OpenAI则是在其闭源的ChatGPT上提供各种定制化的服务，依赖于其提供的底层模型。”谭中意说道。

就像持续至今的“iOS”和“安卓”之争，大模型的生态建设同样面临开源还是闭源这道选择题。谭中意认为：“OpenAI的ChatGPT是闭源的，Meta的Llama2是开源的。以后的大模型生态，将是开源生态和闭源生态之间的竞争和合作同时并存。”

在国内，百川智能大模型、智源AI大模型、腾讯混元大模型、阿里云通义千问大模型等都宣布加入开源的“大部队”。而华为的盘古大模型、百度的文心一言等则选择了闭源。

戴鲲指出，与传统技术领域不同，大模型的开源包含多种不同层次，涉及模型架构、用于模型预训练的代码与超参数、完成预训练的模型权重与参数、用于模型评估的输入数据预处理代码与模型评估代码、全过程配置与开发文档、API与插件接口、许可证方式等。保持开放的接口与插件体系、搭配开放的文档与有限开源的商业许可是必然的选择，而其他层次的开源与否可以根据市场发展动态选择。

“模型商店带来的将是更加广泛的数据和商业模式。如果开源能够提供闭源所不能替代的活跃度，同时促进开发，当然很好。但如果只是增加了短期的应对负荷和同质化产品竞争，对技术和商业价值都没有太大意义，闭源就很好。”王娟说道。同时她指出，OpenAI至少目前的目的不是纯粹的商业化。腾讯、阿里的模型商店是要用模型盈利和定价带动应用层的配套，以及云和硬件产业链市场。

百川智能创始人、CEO王小川表示，未来开源和闭源会像苹果和安卓系统一样并行发展。大部分服务会依赖开源模型，而闭源会提供特定的增值服务。开源模型提供80%，最后靠闭源提供剩下的20%。

模型生态究竟应该怎么建？

发展至今，无论是通用大模型还是行业垂直大模型，赛道上都已挤满了各类玩家，有互联网科技公司，有AI技术公司，还有手机厂商、家电厂商、金融机构、文娱公司、教育机构等跨界选手。这反映出业界对大模型抱有极大的热情与信心，但同时也表明产业尚未形成一个真正具备吸引力和竞争力的模型生态。

谈及构建模型生态的关键要素，戴鲲表示，模型自身的能力、厂商的平台化能力和生态运营能力，三者缺一不可。首先，模型要有卓越的性能、出色的多模态支持、良好的开放性与快速迭代、良好体验的开发环境、完善的文档与案例等；其次，厂商必须具备平台化能力，比如涵盖公有云、私有云与边缘云在内的服务于ModelOps的全栈云原生能力平台化，围绕模型的人工智能平台与数据管理的全生命周期平台化，面向各行业细分业务场景进行模型定制的平台化，以及涵盖从底层芯片到开发和上层应用的软硬件适配平台化等；此外，厂商还需具备生态运营的能力，比如对国内、国际开源社区与产业联盟的贡献与影响，从模型开发到工程实践对开发人员的有效支持以及企业业务与技术决策者的思想领导力等。

“头部厂商积累的数据客观上形成模型生态的竞争基础，所以在生态建设方面，字节、腾讯、阿里这些企业的核心竞争力更具优势。”王娟表示。实际上，模型并非越多越好，国内现在大模型很热，已有的大模型愿景大多是做全产业链布局的。很多看上去不错的大模型，实际本身不仅不产生任何价值，还造成了算力、人力和财力的浪费。

根据专家预测，未来几十年与大模型相关的产业形态，首先是有几家提供通用大模型服务的企业，包括百度、阿里等；其次是多家企业提供行业大模型的服务，包括金融、能源、制造等行业；最后是数百家甚至上千家技术企业提供企业内部的私有化大模型服务，用于知识管理、软件开发、供应链等具体场景。每家企业内都会有很多大模型的服务，其中大部分是部署在企业内部的私有化大模型服务，也有少部分是访问公网大模型API服务。

“要建立不错的生态环境，需要卓越的技术能力和商业能力，从国内大模型厂商来看，百度相对而言技术实力比较靠前，腾讯和阿里云也有丰富的应用场景，都有不错的前景。”谭中意表示。不过，对比OpenAI，差距还是非常明显的。比如中文数据集在数量上和质量上还跟英文数据集有很大差距，算力也受到很大限制，架构在大模型上的开发生态才刚刚开始。

“我们需要的大模型是一个能够持续进化的大模型，是一个能在此基础上产生健康生态（开发活跃、良性竞争、技术和商业都兼顾）的大模型，中国的大模型生态应该是闭源和开源互相竞争、互相合作的模式。”谭中意表示。

金砖国家奏响新型工业化和声

（上接第1版）

巴西驻华大使高望表示，数字化将是未来几十年的主要发展方向，巴西已与中国签署了信息通信技术合作谅解备忘录，推动科学技术发展成果加速转化，不断夯实工业基础。阿尔弗雷德·马什希表示，随着数字创新浪潮席卷全球，南非高度重视高速网络、高性能云计算、人工智能等领域的发展，希望依托金砖国家工业革命合作伙伴关系论坛加深与各国之间的联系交流。

在共谋工业绿色可持续发展方面，金砖国家也涌现出诸多生动实践例子。近年来，巴西淡水河谷与中国在可持续开采和低碳解决方案等领域紧密合作，“塞拉多骄阳”太阳能园区和淡水河谷-中南大学低碳和氢冶金实验室便是其中的典范。据介绍，该园区是拉美最大的太阳能项目之一，其装机容量相当于一座80万居民的城市一年的电力消耗。该项目使用的光伏组件和备件均由中国企业提供。

为建立金砖国家大学和科技园合作网络框架，在今年的论坛上，中国—金砖国家新时代科创孵化园启动建设。俄罗斯之家（厦门）协同创新中心、德国史太白（厦门）协同创新中心、新加坡创士锋国际技术转移中心、新西伯利亚大学（厦门）技术转移中心，以及金砖技术交易平台、金砖知识产权运营平台等第一批项目举行了入驻孵化园签约仪式。

在进一步深化金砖国家工业领域合作的过程中，中国企业一直是排头兵。“继2022年自动驾驶车辆出口阿联酋阿布扎比后，金龙的自动驾驶技术进一步获得世界多国认同。无人驾驶小巴车、智慧车辆编队、车路协同系统……未来更多智慧交通有望落地中东市场，为当地的交通建设输出‘中国智慧解决方案’。”厦门金龙联合汽车工业有限公司专业总师蒋金告诉记者。

在非洲地区的朱安恩格矿，华为公司的4G LTE网络解决方案为该矿提供了稳定的连接，连接了260多台设备，包括钻机、挖掘机和卡车，使该矿的生产、安全和安保系统之间可以互相连接。据介绍，在过去五年中，华为为非洲地区的许多采矿企业提供了服务，加速了采矿数字化转型。

为畅通各国信息数据流通“大动脉”，中国联通已经在全球部署超过了60条国际海缆系统、12个国际传输维护中心、22个边境站，以及超过350个产品业务的接入点，其云数据中心等信息基础设施为包括金砖国家全球化布局提供了端到端的综合数字化解决方案。