GPU太烧钱 巨头自研AI芯片渐成趋势

本报记者 李玉洋 上海报道

身处AI(人工智能)的“iPhone时刻”,英伟达的地位依旧稳固。

英伟达于近日发布的2024财年三季报显示,其营收达到创纪录的181.20亿美元,同比增长206%,环比增长34%;净利润再创新高,达到92.43亿美元,同比增长1259%,环比增长49%。

同时,英伟达还预计下个财季营收将达到200亿美元,这份强劲的收入展望,表明了支撑AI繁荣的芯片需求仍然旺盛。目前,英伟达的GPU占据了AI芯片绝大部分的市场份额,且供不应求,高端GPU更是一卡难求。

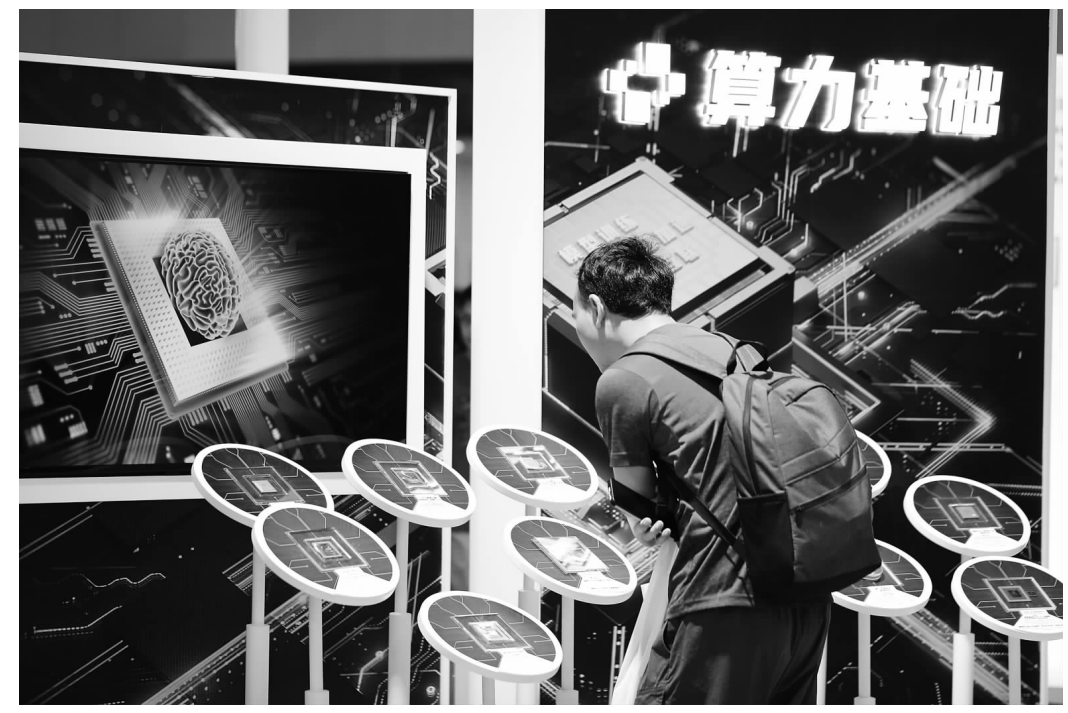
而一股自研AI芯片的风潮正在科技巨头之间兴起,不管是为了降低成本,还是减少对英伟达的依赖,包括谷歌、亚马逊、阿里巴巴、腾讯、Meta等大厂先后入局造芯,下场自研AI芯片。《中国经营报》记者注意到,这份巨头自研AI芯片的列表如今又增加了新成员——微软,在11月中旬举行的开发者大会上,微软推出自家的AI芯片Maia 100;而风头正盛的OpenAI也被爆出正在探索研发自己的AI芯片。

“他们的算力需求比较大,英伟达的卡对他们来说太贵了,而且专门为大模型设计的芯片,能效比会比N卡更高。”一名算力提供商员工表示,做AI芯片太烧钱了,很多AI芯片公司扛不住了,大模型会加速淘汰。

根据英伟达数据,在没有以Transformer模型为基础架构的大模型之前,算力需求大致是每两年提升8倍;而自利用Transformer模型后,算力需求大致是每两年提升275倍。

“自研AI芯片是一个产业走向成熟绕不开的趋势,但凡哪个厂商AI运算的体量大幅度提升,就需要自家的芯片来支撑,这样才能达到最高的优化。”研究机构Omdia AI行业首席分析师苏廉节对记者表示,未来AI芯片市场不再由GPU独霸,“我觉得比较有趣的是英伟达自己怎么看这个趋势”。

而英伟达方面在业绩说明会上证实,其正在为中国开发新的合规芯片,但同时表示这些不会对第四季度的收入作出实质性贡献。“明年年初才能下单,我们在等价格,下单后啥时候能到货还不好说。”对此,记者从国内一家服务器公司的技术研发人员口中得到了一些侧面印证。



一股自研AI芯片的风潮正在科技巨头之间兴起。 视觉中国/图

洗牌避免不了

芯片可能会在未来很长一段时间都是AI竞争的核心,包括国家、巨头、初创公司之间。

“目前,全球几个云大厂基本上都在自研AI芯片,比如亚马逊的Trainium和Inferentia、谷歌的TPU、百度的昆仑、华为的昇腾、阿里的玄铁,现在微软也开始做自己的AI芯片。”苏廉节表示,未来云大厂会把自家的模型部署在自己的芯片上,这是产业发展绕不开的趋势。

然而,他还指出:“这不代表英特尔、AMD、寒武纪这种芯片厂商就没有出路,毕竟不是所有的开发者都会选择使用云厂商的芯片。”

同时,作为云大厂,微软也没有把自己的AI芯片Maia 100的使用权赋予所有人。微软首席执行官萨提亚·纳德拉(Satya Nadella)表示,Maia将首先为微软自己的人工智能应用程序提供支持,再提供给合作伙伴和客户。

据了解,微软Maia 100采用台积电5nm工艺,拥有1050亿个晶体管,比AMD的MI300X GPU的1530亿个晶体管少约30%。Maia将用于加速AI计算任务,服务于商业软件用户和使用Copilot的个人用户,以及希望制作定制人工智能服务的开发者。

今年6月,AMD对外展示可与英伟达H100一较高低的Instinct MI300X GPU。根据爆料信息,其显存容量提升到192GB,相当于H100 80GB的2.4倍,同时HBM内存带宽高达5.2TB/s,Infinity Fabric总线带宽也有896GB/s,同样超过H100。

最新消息显示,AMD将于北京时间12月7日凌晨2点举办一场专门针对AI的特别活动,不出意外的话就是宣布正式发售MI300X。在第三财季业绩发布会上,AMD指出其AI方面的进展,预计第四财季数据中心GPU收入约为4亿美元,到2024年将超过20亿美元,MI300系列产品有望成为AMD历史上在最短时间内达到销售额上10亿美元的产品。

而英伟达则在11月13日发布新一代AI处理器H200,旨在培训和部署各种人工智能模型。H200是当前H100的升级产品,集成了141GB的内存,在用于推理或生成问题答案时,性能较H100提高了60%至90%。

英伟达透露,搭载H200的系统将于2024年第二季度由英伟达的硬件合作伙伴和主要的云服务提供商提供,包括亚马逊云AWS、谷歌云和微软云Azure等。英伟达在上个月还表示,将从两年一次的发布节奏转变为一年一次,明年将发布基于Blackwell架构的B100芯片。

值得一提的是,近日,有“英国英伟达”之称的AI芯片独角兽Graphcore称,由于美国出台的出口管制新规限制了公司向中国的技术销售,公司将停止在华销售。

“很遗憾,这意味着我们将大幅缩减在华业务。”Graphcore发言人在电子邮件中表示。但该公司拒绝透露受影响的员工人数。此前,Graphcore CEO奈杰尔·图恩(Nigel Toon)曾把中国视为一个潜在增长市场,尤其是英伟达被限制向中国销售产品之后。今年10月,图恩表示,来自中国的销售额可能占其公司业务的20%至25%。

如今,Graphcore已经陷入困境。该公司最新提交的文件显示,2022年营收下降46%,亏损扩大11%至2.046亿美元。去年10月,Graphcore披露需要筹集资金以维持运营。在那之后,Graphcore再未宣布任何融资消息。美国对华禁令无疑使其雪上加霜。

对于Graphcore的最新动态以及AI芯片行业的发展现状,记者联系Graphcore方面,该公司表示:“不方便评论。”国内AI芯片初创公司壁仞科技同样也是保持沉默。

对此,资深行业分析师张慧娟表示:“芯片可能会在未来很长一段时间都是AI竞争的核心,包括国家(可参考美国对高端AI芯片出口中国的限制)、巨头、初创公司之间。但行业发展还处于初期,未来肯定会有一轮洗牌。”

“另外,OpenAI最近的宫斗,体现了人才对于AI的重要性。所以,算力(芯片)和人才,会是很长一段时间内AI的竞争焦点。”张慧娟说。

非云端是突破口

需要指出的是,边缘和端侧没有特别明确的区分或界定。

上述算力提供商员工表示,英伟达十几年建立的CUDA生态确实不容易被撼动,但是大模型也带来了很多机会。

“从今年市场所看到的融资信息来看,AI芯片赛道整个都很冷,现有的市场有点饱和,最关键的是AI还没有大量变现,很多公司部署了AI不是拿来赚钱,主要是拿来降低内部运作成本。”苏廉节表示,“生成式AI软件的投资有显著的提升。”这里的软件指的是大模型,代表公司为百川智能等。

苏廉节还指出:“未来的AI芯片市场,可能不再由GPU独霸,比较有趣的是英伟达自己怎么看这个趋势,可以从他们现在积极部署的新市场,去看看他们的心态。”

对于新增长战略,英伟达在Q3财报中强调了三大要素:CPU、网络、软件和服务。据悉,Grace是英伟达第一款数据中心CPU,Grace和Grace Hopper将在明年全面投产,英伟达CEO黄仁勋指出,英伟达可以帮助客户建立一个AI工厂并创收。

NVIDIA Quantum Infiniband是英伟达推出的网络解决方案,其具有高带宽、低时延、高可靠、易部署的特点,可实现多台DGX计算机的高效互联,InfiniBand同比增长了5倍。

将GPU、CPU、网络、AI企业软件等作为增长引擎,是英伟达的远期展望和技术路线,对于处于其他阶段的玩家不一定适用。

苏廉节认为,边缘设备或许是AI芯片赛道上其他玩家的突破口。“比如智能汽车、机器人、无人机,现在的大模型都在云端,但它渐渐会走向终端,量就会很大。”他说。

张慧娟也持有类似观点。“GPU在云端大算力有绝对优势,但是端侧、边缘侧,有着非常丰富的应用类型,所以这也是其他AI芯片厂商寻求突破的点。且从市场规模来讲,边缘端增长空间很大,这也是吸引AI芯片厂商进入的一个原因。”

她还提到,国内AI芯片厂商,除了少数几家进入云端外,大部分集中在端侧,也有一些从端侧跨进边缘,“国内芯片厂商在边缘侧突破比较大的有瑞芯微、全志,高通、NXP、瑞萨、ADI等国际芯片厂商也在大力做边缘AI。”

需要指出的是,边缘和端侧没有特别明确的区分或界定。“所以,很难说这家厂商只做边缘或只做端侧。其实英伟达也有边缘AI方案,比如智能驾驶方案,也很强。英伟达、英特尔等巨头号称从云到边缘到端都做,主打一个端到端全覆盖。”张慧娟说。

近日,联想集团董事长兼CEO杨元庆也表示,目前大模型的用户规模还比较小,大多数大模型都在算力较强的公有云上训练。“未来,随着用户规模扩大,无论是出于数据安全和隐私保护的考虑,还是更高效率、更低成本响应用户需求的考虑,大模型的计算负载将逐渐由云端向边缘侧和端侧下沉,越来越多的人工智能的推理任务将会在边缘和设备端进行。”

杨元庆指出,要构建和优化大模型,支持更多生成式人工智能的应用,不仅需要提升云端的算力,在边缘和端侧也需要更强大算力的配合,形成“端-边-云”混合计算架构下更平衡的算力分配。

鸿蒙生态圈加速形成:已有170万人参与相关培训

本报记者 秦枭 泉州 厦门报道

继华为发布HarmonyOS(鸿蒙)NEXT开发者预览版(即鸿蒙原生应用)全面启动后,美团、百度、工商银行、中国电信等60多家企业纷纷表示将成为具备独立设计和开发鸿蒙应用能力的开发者。

随着越来越多企业级开发者加入鸿蒙原生应用开发,鸿蒙生态正在快速形成。根据Counterpoint发布的最新数据,华为HarmonyOS在2023年第一季度的中国市场份额已达到8%,仅次于安卓和iOS。有业内人士认为,鸿蒙将很快替代安卓。实际上,鸿蒙的初心和野望不止互联网,而是定位于万物互联时代的操作系统。

华为终端云服务总裁朱勇刚此前强调,“独木不成林。我们希望能够和所有的合作伙伴一起共同灌溉,共同施肥,让这片林向下扎到根,向上捅破天;让我们的鸿蒙生态,立足于中国这片沃土,从松山湖这个宝地出发,真正走向全球,成为一个全球的生态。”

近日,《中国经营报》记者走访了厦门、泉州等地,了解正在扩容的鸿蒙“朋友圈”。

正在成为确定性的选择

记者在华为应用商店搜索发现,目前带有鸿蒙版三个字的应用正在逐渐增多,而且已经有不少应用被打上了“HMOS”角标,代表应用含有鸿蒙原生服务。对比安卓发现,其安装包较小,页面相对简洁。

一位开发者对记者表示,传统厚重的App,整体体验好,功能齐全,但开发成本高、周期长,且存在搜索、安装、升级、卸载等一系列需要用户主动关注的显性操作,这些显性操作给用户带来了实质性的使用成本。轻量化、可快速达成消费者意图、可独立执行、完成单一功能的程序实体正成为新的趋势,例如小程序、AppClip快应用等。

阿拉丁研究院发布的《2021年度小程序互联网发展白皮书》显示,小程序远超App数量,大型应用开发者普遍向用户提供轻量化程序实体。在很多特定的使用场景下,小程序等轻量化程序实体的使用占比已超过App,成为面向用户的主要触达方式。

实际上,能够让更多的互联网企业开发鸿蒙的原生应用,只是鸿蒙的一部分构想。目前,智能设备数量正在持续高速增长。操作系统也面临设备底座从手机单设备到全场景多设备的转变。

工业和信息化部发布的数据显示,截至2022年年底,我国移动物联网终端用户数达到18.45亿户,比2021年年底净增4.47亿户,占全球总数的70%。GSMA(全球移动通信系统协会)预测,到2025年,全球物联网终端连接数量将达246亿个,其中消费物联网终端连接数量将达110亿个。国际数据公司IDC预计到2025年,中国物联网的IP连接数将达到102.7亿个。

产业观察家王振涛对记者表示,不同设备类型意味着不同的传感器能力、硬件能力、屏幕尺寸、操作系统和开发语言,还意味着差异化的交互方式。同时跨设备协作也让开发者面临分布式开发带来的各种复杂性挑战,例如跨设备的网络通信、数据同步等。若采取传统开发模式,适配和管理工作量将非常巨大。而鸿蒙目前做的是,无论是小到128KB内存的设备,还是达到GB级别内存的智能终端,都能被鸿蒙拉到同一平台来构建系统,进而实现智能设备之间的无缝互联。

艾恩科技集团(厦门)有限公司品牌经理吴燕对记者表示:“鸿蒙加快了我们产品的智能化进程,拓宽了智能适配的场景,加入鸿蒙前我们在产品的智能操控上,更多是依靠我们自己的App开发,适配鸿蒙之后,通过华为手机、华为智能音箱,都可以操作自家的产品。”

奥佳华智能健康科技集团股份有限公司集团董事、中国区总经理郭桃花表示,奥佳华产品支持鸿蒙操作系统,它的操控方式除了AI智能语音、专门配备的华为平板和扶手按键之外,还支持HarmonyOS万能卡片,生态用户无须下载App,只要碰一碰卡片区域即可直达按摩操作界面,交互更简单。得益于鸿蒙操作系统的加持,奥佳华AI按摩机器人2.0在硬件配置和系统软件上更全面、更精准、更高效,实现了极简互联、健康数字化,为消费者打造更便捷流畅、鲜活妙趣的高品质全场景智慧健康生活。

不过,在王振涛看来,针对鸿蒙做全新的开发,无疑会增加企业的开发成本。对厂商而言,是否愿意在鸿蒙项目中投入更多资源,根本因素还是在于商业利益。相比对于终端设备的赋能,鸿蒙最大的吸引力在于数亿名华为终端用户,接入鸿蒙就相当于成为数亿终端用户的潜在消费者。

人才是关键

鸿蒙生态的构建,也引得鸿蒙开发者遭到企业的哄抢。京东、网易、美团、WPS、微博等互联网大厂纷纷在一些招聘App上,发布招聘鸿蒙开发者的岗位。

其中美团10个正在招新的职位中,2个岗位是鸿蒙高级工程师(C++)、鸿蒙基建工程师,其余岗位职责或包含鸿蒙系统的设计与开发工作,或表示鸿蒙相关经验优先。

不过,上述开发者对记者表示,从技术上来讲,在鸿蒙系统开发一个专版App,没有难度。鸿蒙系统的分布式架构和原子化编程框架,应用开发者可以只写一次代码,在所有屏幕上都能跑。简单来说,就是让开发者在不同设备的屏幕规格下,从系统底层进行像素转换,自动适配手机、折叠屏、平板、PC、智慧屏、智能手表等不同设备的显示效果,不需要每一个设备都单独打包、手动上传和繁琐的维护。

“不过,鸿蒙生态提出来没几年,目前市场上并没有经验丰富的鸿蒙开发人员。”上述开发者坦承。

厦门城市职业学院人工智能学院副院长邓汉勇对记者表示,目前互联网大厂都在做鸿蒙人才的储备,且工资都比较高。但对于目前尚在学习阶段的学生而言,鸿蒙开发的门槛相对较高。因为纯粹针对一个产品做应用是比较简单的,但鸿蒙开发是一个科学系统且需要大量实践的过程。

根据华为方面的数据,截至今年9月,已有超过170万人参加了鸿蒙学堂的课程学习、线下活动,华为还和全国300多所高校展开了合作。

邓汉勇表示,目前学生对于学习鸿蒙开发的意愿很高,因为学生之前学的是国外的技术或者系统,那么鸿蒙作为一个纯国产化的系统,并且是与国际先进水平相当的操作系统,学生们可以提升自豪感。

实际上,不只是高校,近年来,多地政府、企业也在积极推动鸿蒙生态人才的建设。福建、深圳、江苏等省市已先后发布关于HarmonyOS开发者的人才培养政策,从政策层面为HarmonyOS开发者的成长开辟道路,并希望通过培养HarmonyOS相关人才,助力当地打造更加完备的数字化产业生态,同时,也推动着HarmonyOS人才培养体系在本省的落地生根。