编辑 赵天舒 美编 李烝 责校 杨志强 电话:64101946 bbtzbs@163.com

AI芯片下半场:英伟达不再一家独大

ChatGPT爆火迄今,英伟达被公认为本次全球AI淘金浪潮的最大"卖铲人",也是各大媒体和社交平台上讨论度最高的一家AI芯片公司。不过,随着AI热潮持续升温,越来越多厂商也开始在AI芯片领域发力:前有英特尔、AMD等半导体巨头公布新一轮的AI芯片研发计划,后有OpenAI、微软等下游客户推动自研芯片,以打破英伟达的垄断。

微软 | 推出加速AI计算任务的Maia芯片,以及采用Arm CPU架构的Cobalt芯片

AMD | 发布加速卡Instinct MI300X,高带宽内存(HBM)密度最高可达英伟达H100的2.4倍

SK | 推出其最新的人工智能芯片X330

微软自研

当地时间11月15日,微软在西雅图举行的Ignite开发者大会上,推出两款定制芯片,以应对不断增加的大模型训练成本挑战,并试图降低提供AI服务的成本。微软表示,新发布的芯片不会出售,仅供支持自己的产品,并作为微软云Azure云计算服务的一部分。

微软最新推出的两款芯片分别为用于加速AI计算任务的Maia芯片,以及采用Arm CPU架构的Cobalt芯片。Maia芯片旨在运行大型语言模型。微软云和人工智能部门执行副总裁Scott Guthrie表示,希望通过优化这款AI芯片来提供更快、成本更低、质量更高的解决方案。据悉,这两款芯片都将于2024年上市。

微软没有透露这些芯片相对于传统芯片具体有多少能力提升的技术细节。但微软Azure硬件系统和基础设施副总裁Rani Borkar表示,两款芯片均采用台积电5纳米制造技术制造。此外,Maia芯片的构建方式与英伟达使用的网络连接技术也有所不同,Maia芯片与标准以太网电缆连接在一起。

事实上,微软推定制芯片早有征兆。2010年时,微软便希望能自研AI硬件。据The information报道,微软至少从2019年开始,便在研发代号为"Athena(雅典娜)"的新AI芯片组,目的是为ChatGPT等大语言模型的训练及推理提供英伟达芯片之外的替代方案。另据Tom's Hardware消息,Athena使用的是台积电5nm工艺,专为大语言模型训练设计。

曾有知情人士透露,在开发Athena期间,微软为了满足OpenAI的需求,已经向英伟达订购了至少数十万块GPU。而进入9月,随着ChatGPT热潮趋于平缓,不断有市场消息传出微软开始下调英伟达H100显卡的订单量。在微软10月的财报电话会上,"削减成本"一事也被反复强调。

不过,微软推出自研AI芯片并不是希望取代英伟达等厂商。微软表示,明年将为Azure客户提供云服务,这些服务就是运行在英伟达和AMD的最新旗舰芯片上。微软正在AMD的最新旗舰芯片上测试OpenAI最先进的模型GPT-4。

纷纷入局

在芯片供不应求和AI应用需求激增的背景下,微软对定制芯片计划的迅速推进,被外界普遍解读为拥有资源的云计算巨头的不二选择。彭博报告显示,随着OpenAI的ChatGPT等服务的涌入,在未来十年内,生成式人工智能市场有望从2022年的400亿美元激增至1.3万亿美元。

在此之前,亚马逊和谷歌等科技巨头也已经使用了自研芯片,并部分向客户提供。另一个芯片大厂AMD也在不久前推出了最新加速卡Instinct MI300X。在发布会现场,PPT上专门打出一行字——大语言模型专用,这被业界视为直接向英伟达宣战。

据悉,MI300X的高带宽内存(HBM)密度最高可达英伟达H100的2.4倍,高带宽内存带宽最高可达H100的1.6倍,显然MI300X能运行比H100更大的AI模型。

而紧随微软之后,韩国无线运营商SK电信周四在首尔举办科技峰会SK Tech Summit 2023。在这场峰会上,SK集团下属的半导体初创公司Sapeon Inc.宣布,该公司推出了其最新的人工智能芯片X330。

资料显示,Sapeon公司总部位于美国加州圣克拉拉。据报道,韩国巨头企业SK电信、SK海力士、SK Square等是该公司的股东。2020年,SK电信首次开发了Sapeon X220,这是韩国国内首款用于提供高速低功耗AI服务所需的数据中心的芯片。两年后,SK电信将Sapeon业务分拆出来,使其成为总部位于加州的一个独立实体企业,以加速人工智能芯片的商业化。

该公司在声明中表示,X330芯片的计算性能是上一代X220的4倍,能效是其2倍多。Sapeon计划先为主要客户进行测试,并在2024年上半年开始批量生产该芯片。

地位仍在

对于行业来说,越来越多的供应自然是好事,因为AI芯片的缺货已经持续了一段时间。许多行业人士预计,AI芯片短缺至少会持续到明年。"目前英伟达订单能见度已至2024年,高端芯片非常紧缺。以现在的排产进度,就连A800/H800都要到今年底或明年才能交货。短期内,从其受追捧程度来看,唯一影响英伟达高端GPU销量的或许只有台积电的产能。"有芯片从业人士对北京商报记者表示。

更有创始人和投资者表示,即使AI芯片有着落的公司,他们仍然需要等待数周才能使用到。一位AI初创公司的CEO说:"即使你已经预付了费用,也不意味着GPU会在第二天或下周送到你那里,你只能等。"

"英伟达不会永远在大规模训练和推理芯片市场占据垄断地位。"这是特斯拉CEO马斯克对社交问答网站和在线知识市场Quora的首席执行官Adam D'Angelo一条推文的回应,后者写道:"人工智能热潮被低估的一个原因是GPU/TPU短缺,这种短缺导致了产品推出和模型培训的各种限制,但这些都不明显。相反,我们看到的是英伟达的股价飙升。一旦供给满足需求,事情就会加速发展。"

不过,摆脱对英伟达的依赖似乎不是那么容易。除了定制芯片的真实水平需要做进一步考察,微软在会上发布的其他系列更新仍少不了英伟达的身影。比如微软宣布,其Azure将为客户企业提供英伟达新的生成式人工智能模型器。在创道咨询合伙人步日欣看来,这意味着二者之间的合作关系还将扩大,也表明微软并不想完全破坏与英伟达之间的默契。

曾有人坦言,英伟达与其他芯片厂商的差距,是院士与高中生的差别。就如同英伟达CEO黄仁勋所言,英伟达"一直在奔跑",想要超越巨人的其他芯片厂商只能奋力狂奔。据Khaveen Investments测算,英伟达数据中心GPU 2022年市占率高达88%,AMD和英特尔瓜分剩下的部分。

全联并购公会信用管理委员会专家安光勇也指出,这些企业的发展情况各有差异,但总体来说它们的市场份额还无法撼动英伟达的主导地位。痛点方面,包括芯片制造成本、芯片能效比、产品性能和稳定性等方面的挑战。

北京商报记者 方彬楠 赵天舒

聚焦 Focus

不服被列"守门人" Meta、TikTok上诉

大型科技公司正在挑战欧盟委员会的数字"守门人"标签。在期限届满前,TikTok和Meta纷纷就"守门人"地位向欧洲普通法院提出上诉。当地时间周三,美国科技公司Meta表示,其平台服务Messenger和Marketplace被欧盟委员会列入"守门人"名单,打击范围有过大之嫌,已正式提起上诉。

今年9月,欧盟根据《数字市场法》(DMA)首次指定了6家科技企业"守门人"——谷歌、亚马逊、苹果、字节跳动、Meta、微软,一共涉及22个由"守门人"提供的核心平台服务。

据悉,"守门人"企业指那些提供社交网络、搜索引擎等核心平台服务的大企业,其市值至少为750亿欧元或年营业额75亿欧元,还需在欧盟每月至少有4500万终端用户,每年有1万名商业用户。

DMA为"守门人"规定了一系列细化义务,包括不得滥用优势地位义务、数据信息保护义务、广告信息披露义务、举报信息处理义务、消费者权益保护义务等。如果相关企业违反相应要求,将面临高达其前一财政年度全球年营业额10%的罚款,屡次违规将可能被处以高达全球总营业额20%的罚款。

Meta旗下的Facebook、Instagram、Marketplace和WhatsApp有资格符合DMA义务要求。DMA的目的是在大型科技公司和较小的竞争对手之间创造公平的竞争环境。

据了解,如果不同意欧盟"守门人"的认定,企业必须在11月16日之前向总部位于卢森堡的欧盟普通法院提出上诉,预计欧盟普通法院将在几个月内对上诉做出裁决。

Meta发言人表示:"这一上诉旨在澄清有关在DMA法规下指定Messenger和Marketplace的具体法律要点。这不会改变或减损我们遵守DMA的坚定承诺,我们将继续与欧盟委员会开展建设性合作,为合规做好准备。"

这位发言人还声称,不会对Facebook、Instagram和WhatsApp的这一认定提出质疑。但Marketplace是一种消费者对消费者的服务,因此它不属于在线中介服务的定义,而Messenger只是Facebook的聊天功能。

Meta认为,Messenger和Marketplace被列入"守门人"名单是错误的,并对欧盟的决定表示不服。

除了Meta外,TikTok也对此表示不满。"对在这一决定之前没有进行市场调查感到失望。"最新声明中,TikTok坚称,"我们的平台远非'守门人',在欧洲运营仅五年多,可以说是更根深蒂固的平台业务最有能力的挑战者。"

TikTok还认为,它被指定为"守门人"是基于其母公司字节跳动的全球市值,而母公司的全球市值主要基于不在欧洲运营的业务线的业绩,它本身不符合欧洲经济区每年75亿欧元收入的法律门槛。

这并不是科技公司第一次公开反抗欧盟委员会和欧盟执行机构的数字规则。德国在线零售商Zalando和美国科技巨头亚马逊已将欧盟委员会告上法庭,指控它们受到《数字服务法》的不公平待遇。

去年12月,欧盟委员会就投诉称,Meta将Facebook Marketplace与其庞大的社交网络捆绑在一起,在分类广告领域削弱竞争对手。此后,Meta已经卷入了与欧盟的纠纷。Meta正在欧盟抗辩这些指控,该公司已就英国竞争监管机构提起的一起类似案件达成和解。

根据规定,DMA要求微软、苹果、Alphabet的谷歌、亚马逊、Meta和字节跳动的TikTok允许第三方应用或应用商店在它们的平台上运行,并让用户更容易从默认应用切换到竞争对手的应用。

在欧盟规定的22项核心平台服务中,由谷歌提供的服务数量最多,包括其安卓操作系统、地图和搜索服务等。谷歌的一位发言人表示,不会对这一决定提出上诉。

与此同时,欧盟反垄断监管机构正在调查微软的必应和苹果的iMessage是否应该遵守新规定。据报道,苹果公司正在研究向第三方应用程序商店开放iOS和允许绕过官方应用商店直接安装应用程序的方法,以遵守DMA规则。但是,该公司认为iMessage没有达到DMA每月4500万活跃用户的门槛,因此不必与其他消息服务进行交互操作。不过,尽管苹果尚未披露官方数据,但外界估计iMessage在全球可能拥有10亿用户。

北京商报综合报道

·图片新闻·

热浪席卷巴西 能源需求创新高

14日,人们在巴西里约热内卢乌尔卡海滩上纳凉。新华社/图

南半球的巴西即将入夏,一股热浪近日给多个城市带来高温,巴西国家电力系统也在经受历史性考验。巴西气象研究机构预测,本次热浪将至少持续至17日。

13日,受到席卷全国大部分地区的热浪影响,巴西圣保罗气温达到37.7摄氏度,是巴西国家气象研究所自1943年有记录以来第二高温。14日,里约热内卢记录了自2014年以来的最高体感温度,达到58.5摄氏度。

报道称,在巴西国家气象研究所的地图上,巴西三分之二的领土被涂成橙色和红色,这两种颜色代表着"危险"和"极大危险"。根据该机构的说法,"极大危险"级别是指"发生伤害和事故的可能性很高,对身体健全甚至生命构成威胁"。在13个州总共约5000个市镇中,共计有1413个处于风险区,当气温连续数日高出平均水平5摄氏度时,就会被算作风险区。

受热浪影响,相关电器使用增多,导致城市对能源的需求加大。根据巴西国家电力系统运营商的数据,本次热浪导致电力需求在13日和14日连续创下历史新高,14日的需求一度达到约101.5千兆瓦。今年11月的预计用电量比去年同月提高了11%。

巴西国家自然灾害监测预警中心气象学家乔瓦尼·多利夫表示,在巴西每年春夏"过渡"期间,热浪现象很常见。随着南半球夏季的临近,当地会更多地暴露在阳光下。现在是少雨季节,云层较少,也有利于出现高温。由于全球变暖和厄尔尼诺现象,今年的热浪变得更加强烈。

据介绍,本次热浪气温比同时期历史平均水平至少高出5摄氏度。多利夫说,地球正在变得越来越热,之后这种极端天气将会越来越频繁。

据新华社