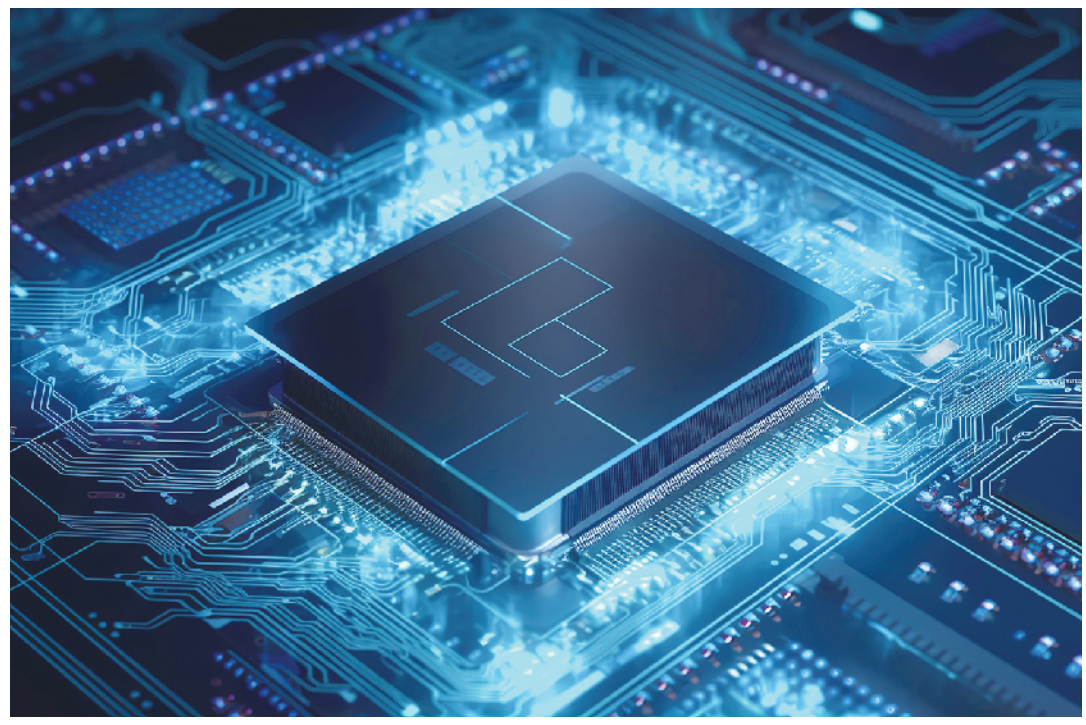
# 巩固霸主地位 英伟达推出“超级芯片”

美东时间8月8日,英伟达升级产品,推出容量和速度双双大幅提升的超级芯片GH200 Grace,力求巩固AI芯片领域的霸主地位。

该平台依托于搭载全球首款HBM3e处理器的新型Grace Hopper超级芯片(下称GH200),专为加速计算和生成式AI时代而打造。

新平台旨在处理世界上最复杂的生成式AI工作负载,涵盖大型语言模型、推荐系统和矢量数据库,将提供多种配置。英伟达称,GH200将于2024年第二季度投产。



## 新平台内存容量提高3.5倍

新平台采用的GH200可以通过英伟达的NVLink技术连接其他GH200芯片,共同部署在生成式AI的大模型上。这种技术让GPU能充分访问CPU内存,在双芯片配置时提供合计1.2TB的快速内存。

英伟达的超大规模和高性能计算主管Ian Buck介绍,GH200比英伟达基于H100的数据中心系统配备了更多内存和带宽。它采用英伟达的Hopper GPU,并结合基于Arm架构的英伟达Grace CPU。

相比前代平台,新GH200 Grace Hopper平台的双芯片配置将内存容量提高3.5倍,带宽增加三倍,一个服务器就有144个Arm Neoverse高性能内核、8petaflops的AI性能和282GB的最新HBM3e内存技术。

HBM3e是一种高带宽内存,带宽达每秒5TB。该内存比当前的HBM3快50%,可提供总共每秒10TB的组合带宽,使新平台能运行比前代大3.5倍的模型,同时通过快三倍的内存带宽提高性能。

当地时间8月8日,在计算机协会(ACM)计算机图形和交互技术特别兴趣组织(SIGGRAPH)的年度大会上发表主题演讲时,英伟达CEO黄仁勋将GH200称为“世界上最快的内存”。

黄仁勋说,为了满足生成式AI不断增长的需求,数据中心需要有针对特殊需求的加速计算平台。新的GH200 Grace Hopper超级芯片平台提供了卓越的内存技术和带宽,以此提高吞吐量,提升无损耗连接GPU聚合性能的能力,并且拥有可以在整个数据中心轻松部署的服务器设计。

黄仁勋表示,在AI时代,英伟达的技术可以替代传统数据中心,投资800万美元的新技术可以取代用旧设备建造的1亿美元设施,而且用电量可以减少20倍。“这就是数据中心在向加速计算转变的原因。你买的越多,越省钱。”

## 降低企业启动AI项目门槛

为了加速各类企业定制生成式AI,英伟达推出了AI Workbench。

据称,它为开发人员提供了一个统一、易于使用的工具包,可以在个人电脑或工作站上快速创建、测试和微调生成式AI模型,然后将其扩展到几乎任何数据中心、公有云或NVIDIA DGX Cloud。

AI Workbench主要降低企业启动AI项目的门槛。通过在本地系统上运行的简化访问界面,它允许开发人员从流行库(如Hugging Face、GitHub和NGC)中微调模型,使用自定义数据。然后,这些模型可以在多个平台之间共享。

全球各地企业都在竞相寻找合适的基础架构并构建生成式AI模型和应用,尽管现在已经有数以千计的预训练模型可用,但是使用许多开源工具进行定制可能仍具挑战性且耗时。

“为了使这种能力普惠,我们必须使其能够在几乎所有地方运行。”黄仁勋称,“让所有人都能参与生成式AI。”

借助AI Workbench,开发人员只需点击几下就可以定制和运行生成式AI。它允许他们将所有必要的企业级模型、框架、软件开发工具包和库汇集到一个统一的开发者工作区中。

据称,包括戴尔、惠普、Lambda、联想和Supermicro,都正采用AI Workbench,因为它可以将企业生成式AI能力带到开发人员希望工作的任何地方,包括本地设备。

黄仁勋展示了AI Workbench和ChatUSD如何将所有这些功能结合在一起:允许用户从GeForce RTX 4090笔记本电脑启动项目,并随着项目变得更加复杂而无缝扩展到工作站或数据中心。

黄仁勋表示,用户可以提示模型生成一张玩具黄仁勋在太空中的图片,但初始模型提供的结果不适用,因为它从未见过玩具黄仁勋,这时候用户可以用八张玩具黄仁勋的图片微调模型,然后再次输入提示,获得正确的结果。

综合

# 多维度演进 脑机接口探寻应用扩围

由脑科学研究延伸出的脑机接口技术,其产业化落地正呈现加速推进态势。

政策层面在持续加注:“十三五”规划中首次将脑科学列入国家重大科技项目,并在“十四五”规划中重点支持类脑计算与脑机融合计算研发。地区上,北京、上海、浙江、广东等为代表的省市都在积极从政策层面具体推进。

头豹研究院指出,脑机接口行业起步于上世纪70年代,经历了前期的理论探索期、科学论证期,目前已进入成果落地时期,对脑机接口的研究取得了一定成效。

目前市面上的脑机接口公司主要分为两大类技术路线:侵入式和非侵入式,前者以马斯克主导推动的Neuralink为代表,后者是目前大多数厂商的选择方向。因选择的路线差异,其落地场景和进展有所不同。

近日柔灵科技运营总监张新闻表示,公司目前聚焦于非侵入式脑机接口路线,对产业化落地分为短期、中期、长期目标进阶推进。短期内聚焦脑电设备,中期主打肌电产品,长期则迈向侵入式脑机接口设备和应用。

“我们认为,在未来10-15年后,从事脑机接口行业的公司都将殊途同归,往侵入式脑机接口方向发展。”张新闻表示,目前看,侵入式脑机接口可以落地到疾病治疗等场景,但随着技术本身的不断发展,不排除能实现类似电影《黑客帝国》中的场景。

对于该项技术路线,则需要不断从电极、算法、芯片、系统等多维度演进。

## 技术演进

脑机接口技术演进正进入加速阶段。

5月25日,Neuralink宣布获得FDA临床许可。华鑫证券指出,Neuralink采用的侵入式方案,通过神经手术机器人,将柔性电极植入至大脑皮层中,植入后不可移除。此前2022年初,FDA基于安全等因素拒绝了公司的临床申请,此次获批意味着在安全性等方面已经获得更多的证据。除Neuralink之外,Precision Neuroscience等多家脑机接口公司也计划在2023-2024年申请人体临床。虽然从第一例临床到产品上市到大规模临床应用,仍有较长的时间验证,但万里征途已踏出关键一步。

相比需要进入人体内的侵入式技术,非侵入式技术的落地探索看起来更快。

张新闻表示,作为一家商业公司,在运营过程中需要考虑落地周期、商业化变现等问答,目前侵入式产品还面临着安全、伦理、临床验证等门槛,其落地周期相对较长,因此从非侵入式着手,再逐渐进入侵入式技术路线,是一种相对现实的考虑。柔灵科技目前对侵入式技术路线有推进相关技术储备和合作,只是暂时还未体现在产品层面中。

“侵入式脑机接口,是直接把电极深入到人体大脑中,涉及到生物兼容性、安全性等问题,研发和产品化都需要较长的周期。”张新闻分析道,但优势也很明显,因为可以直接在大脑层面采集相关信号,其采集数据的量级和精确度都不是非侵入式可以比拟。

由此可以看出,非侵入式路线的优势在于,只需要在皮肤表面采集脑电信号,因此安全性可以保障,当然弊端是采集到的脑电信号相对微弱。

“这决定了两条路线的应用路径差异。”他续称,比如侵入式路线可以实现未来对视觉重建、声音重建,治疗帕金森症、阿尔茨海默症等难题的辅助解决;非侵入式路线很难对这类神经退行性病变产生明显治疗效果,但可以通过材料、算法等技术创新手段,把收集到的微弱信号转化成有效信息,再进一步解决如睡眠监测、睡眠质量干预等难题。

张新闻表示,非侵入式技术路线主要的技术门槛包括三方面:材料技术、信号处理技术、算法。“这基本上是明牌,就看谁做得更扎实、水平更高。”

算法层面,基于前面两个环节中,把更多的有效信号收集和分析,也能够辅助让算法水平得到保障。这三方面核心技术,将确保侵入式产品的准确性、有效性。

## 落地渗透

随着技术不断演进,其商业化也在持续找到落脚。

国信证券研报指出,医疗影像、娱乐内容等都是脑机接口的落地方向,且这项技术有望成为元宇宙入口的终极形态。

张新闻告诉记者,目前柔灵科技的脑机接口技术分为两条产品线:脑电产品、肌电产品。前者的主要场景目前落地在睡眠监测和干预;后者以2B市场为主,提供一种新型交互方式和设备。

“通常来说,非侵入式脑机接口技术路线的应用场景主要包括:脑电领域有注意力监测、睡眠监测和干预、疲劳度监测等解决方案,肌电领域有手势控制、智能义肢等方案。”他续称,柔灵科技目前选择以睡眠为主要落脚,是因为关注到这对个人生活将带来巨大帮助。

张新闻表示,面向空间计算时代,柔灵科技认为神经3D肌电交互手环将为下一代计算平台提供一种全新交互方式,因此正在与各类型AR/VR厂商洽谈合作。“只需要一个简单手势,比如大拇指稍微挪动,就可以被精准识别到信号,用来控制设备。”

也由此,虽然技术框架要求一致,但肌电产品线对跨人跨天的准确性、鲁棒性、延时率,数据库体量大小等,都有相比脑机线产品更多的技术挑战。

## 应对挑战

头豹研究院指出,目前中国脑机接口行业尚面临诸多痛点,但在利好政策等因素影响下,行业未来将在材料端、技术端、道德风险端、产业端有更进一步的研究与发展。

国信证券也具体指出,目前脑机接口发展面临的技术挑战包括:脑电信号采集方法还有待改进;软件系统稳定性和自适应性较差,信号处理方式和信息转换速度有待提升;侵入式脑机接口对植入芯片的软硬件要求较高;如何把信号精确地传送到脑内相关技术有待探索等。

该机构认为,在市场化落地时,对基础理论研究和工程实现都提出了极高要求,因此也短期内限制了其应用范围的拓展。

从发展阶段来看,脑机接口仍需要持续不断演进,各种产业角色的参与并融合推进,叠加新技术路径加入,也有望为产业本身迭代打开更多空间。 供稿:《21世纪经济报道》

## ITMT快报

### 上半年天翼云收入同比增长63.4%

近日,中国电信股份有限公司公布了2023年中期业绩。财报显示,2023年上半年,中国电信经营收入2607亿元,同比增长7.6%,其中服务收入2360亿元,同比增长6.6%,持续10年保持增长的同时高于行业增幅。

上半年中国电信税息折旧及摊销前利润(EBITDA)733亿元,同比增长5%。净利润202亿元,同比增长10.2%,保持双位数增长且高于收入增幅。每股盈利0.22元,同比增长10.2%;中期股息0.14元,同比增长19.3%。

2023年上半年,中国电信资本开支416亿元,其中移动网占比42.4%,产业数字化占比28.4%,宽带网占比16.9%,运营系统和基础设施占12.3%。自由现金流达到176亿元。

具体来看,中国电信基础业务收入稳健增长。2023年上半年,中国电信移动通信服务收入1016亿元,同比增长2.7%;固网及智慧家庭收入620亿元,同比增长3.6%。结构逐步优化,5G套餐用户渗透率73.4%,千兆宽带渗透率20.3%。其中,移动ARPU为46.2元,同比增长0.4%;宽带综合ARPU48.2元,同比增长2.1%。

围绕云计算、AI、大数据、5G/6G和量子等重点领域,中国电信上半年加大研发投入的同时壮大研发团队,研发费用同比增长27%,研发人员数较去年末增长21%。IT系统及业务平台自研技术占比约40%,自研清单成果较去年末增长124%。

值得关注的是,中国电信天翼云上半年保持高速增长。上半年天翼云收入459亿元,同比增长63.4%。目前,天翼云保持着公有云IaaS、IaaS+PaaS份额国内市场三强和政务公有云基础设施第一、全球运营商云第一。国资委40个行业领域公有云中天翼云占60%,市场地位保持领先。 供稿:《21世纪经济报道》

### 二季度中国学习平板出货量同比增36.6%

据IDC学习平板季度跟踪报告显示,2023年第二季度,中国学习平板市场出货量约103万台,同比上升36.6%;上半年出货量约220万台,同比上升37.2%。

2022年第四季度以来,学习平板市场已进入第二阶段:消费者对学习平板的认知度逐步提升,需求也在不断扩大。与此同时,市场上不断涌入新玩家,学而思、有道等厂商步入该赛道,竞争愈发激烈。目前来看,学习平板未来发展仍展现提升势头,消费者对于产品的需求越发多样化。

数据显示,小度G系列产品帮助百度稳居学习平板整体市场份额第一。亲民的价格,加上AI文心大模型的加持,使百度G系列产品在1000元至2000元价位段处于市场主导地位。

科大讯飞作为人工智能领域的产业先锋,将AI广泛应用于学习平板,使其T系列产品在6000元以上价位段取得了第一的市场份额。此外,2023年第二季度推出的C系列新品也使科大讯飞不仅在1000元至2000元价位段市场取得一席之位,该产品也使科大讯飞的出货量升至市场第二。

IDC认为,市场上不断涌入新玩家,学而思、有道等厂商步入该赛道,竞争愈发激烈。同时,尽管学生平板10寸至11寸产品市场份额依然占据主导地位,但值得注意的是,15寸以上产品的市场份额在日益提升,2023年上半年15寸以上产品的市场份额已高达16.5%。大屏化是学生平板未来发展趋势。 综合

### 苹果头显获新专利 可投影虚拟妙控板

根据美国商标和专利局(USPTO)最新公示的清单显示,苹果获得一项Vision Pro头显配件相关的专利,该公司或考虑为这款头显设计一个专门的虚拟“妙控板”。

根据专利描述,苹果指出,当前头显用户与虚拟对象的交互方式并不直观,且操作过程也相对繁琐。为解决这一问题,苹果设想在Vision Pro头显的底部安装投影系统,该系统将虚拟对象投影到物理环境中,在任何物理表面上投影出全息影像,用户即可在物理环境中感知到虚拟对象。

此外,苹果在专利中还提到投影系统的另一项功能,即在物理表面上投射出一个“妙控板”,方便用户与虚拟元素进行交互,为虚拟现实和增强现实应用带来更加便捷的控制方式。 综合