英伟达计划明年推出新AI芯片

近日,英伟达创始人兼首席执行官黄仁勋表示,使用NVIDIA NIM将AI模型部署在云、数据中心或工作站上的开发者,可将模型部署时间从以前的数周缩短至几分钟。和硕、劳氏公司、西门子等客户均在使用。

此外,被英伟达寄予厚望的Blackwell芯片已开始投产,预计将在2025年推出Blackwell Ultra AI芯片。

据悉,英伟达的第一款Blackwell芯片名为GB200,宣称是目前"全球最强大的芯片"。目前,供应链对GB200寄予厚望,预估2025年出货量有机会突破百万颗,将占英伟达高端GPU出货量的近40%至50%。

NIM大幅加快模型部署

6月2日晚间,黄仁勋重点介绍了NVIDIA NIM。NVIDIA NIM是一种推理微服务,可通过经优化的容器形式提供模型,旨在助力各种规模企业部署AI服务。

不过,严格来说,NVIDIA NIM并非新品,最早露面是在今年3月。英伟达在6月2日晚间宣布,全球2800万开发者皆可下载NVIDIA NIM,将AI模型部署在云、数据中心或工作站上,构建Copilot(微软的AI助理)、ChatGPT聊天机器人等生成式AI应用。下月起,NVIDIA开发者计划的会员可免费使用NIM,在其选择的基础设施上进行研究、开发和测试。

根据英伟达的说法,新的生成式AI应用正变得越来越复杂,通常需要使用具有不同功能的多个模型来生成文本,如图像、视频、语音等。而NVIDIA NIM提供了一种简单、标准化的方式,将生成式AI添加到应用中,可使模型部署时间从以前的数周缩短至几分钟。

黄仁勋还透露,包括Cadence、Cloudera、Cohesity、DataStax、NetApp、Scale AI和新思科技等在内的近200家技术合作伙伴正在将NIM集成到他们的平台中,以加快生成式AI的部署。"每个企业都希望在其运营中融入生成式AI,但并非每个企业都拥有专门的AI研究团队。NVIDIA NIM可被集成到任意平台中,任何地方的开发者都可以访问,并且可以在任意环境中运行。"黄仁勋称。

记者了解到,NIM是预先构建的,目前有近40个模型可作为NIM的端点供开发者体验;开发人员可以开源社区平台Hugging Face访问适用于Meta Llama 3模型的NVIDIA NIM微服务,使用Hugging Face推理端点访问和运行Llama 3 NIM。

值得注意的是,英伟达还透露了一批大客户的使用情况,如电子制造商Foxconn正在使用NIM开发针对特定领域的大语言模型(LLM),用于智能制造、智慧城市和智能电动汽车;和硕正在将NIM用于一个当地的混合专家(MoE)模型;劳氏公司正在用NVIDIA NIM推理微服务来提升员工和客户的体验;西门子正在将其运营技术与NIM微服务整合,用于车间AI工作负载;还有数十家医疗保健公司正在部署NIM,为包括手术规划、数字助理、药物发现和临床试验优化等在内的应用领域的生成式AI推理提供支持。

Blackwell芯片开始投产

除了上述产品,黄仁勋还在演讲中透露,英伟达Blackwell芯片已开始投产,并将在2025年推出Blackwell Ultra AI芯片。

今年5月,黄仁勋在财报电话会上称,预计今年Blackwell架构芯片将为公司带来大量收入。英伟达对Blackwell芯片寄予厚望,还是与市场强劲需求有关。从最新披露的财报数据来看,2025财年第一财季,英伟达实现营收260亿美元,较上年同期增长262%。其中,数据中心业务营收226亿美元,与上年同期相比增长427%,是业绩收入的"大头"。

据英伟达首席财务官科莱特·克雷斯解读,数据中心业务的增长源自Hopper架构GPU(例如H100)出货量的增加;该季度的重要亮点之一就是Meta宣布推出Llama 3开源大模型,使用了近2.4万块H100GPU。

除了披露芯片量产进度,英伟达此次还推出了一系列采用NVIDIA Blackwell架构的系统。

据悉,这些系统搭载了Grace CPU以及NVIDIA网络和基础设施,用于助力企业建立AI工厂和数据中心。其中,NVIDIA MGX模块化参考设计平台加入了对NVIDIA Blackwell产品的支持,包括专为主流大语言模型推理、检索增强生成和数据处理提供卓越性能打造的NVIDIA GB200 NVL2平台。

英伟达强调,GB200 NVL2适合用于数据分析等新兴领域,借助NVLink-C2C互连技术带来的带宽内存性能及Blackwell架构中专有的解压缩引擎,较使用X86 CPU时的数据处理速度可最多提速到18倍,能效提高8倍。"新一轮工业革命已经开始,众多企业和地区正在与NVIDIA合作推动价值万亿美元的传统数据中心向加速计算转型,并建造一种新型数据中心来生产新的商品。"黄仁勋称。

英伟达方面表示,目前已有超过25家合作伙伴的90多套已发布或正在开发中的系统使用了MGX参考架构,开发成本较之前最多降低了四分之三,开发时间缩短到六个月,较之前减少了三分之二。另外,英伟达还透露,比亚迪电子、西门子、泰瑞达和Alphabet旗下公司Intrinsic等全球10多家机器人企业正在将NVIDIA Isaac加速库、基于物理学的仿真和AI模型集成到其软件框架和机器人模型中,以此提高工厂、仓库和配送中心的工作效率。

机器人时代已经来临

"机器人技术和物理人工智能的时代已经到来,它们正在各地被广泛应用,这并非科幻,而是现实,令人感到振奋。"演讲中,黄仁勋谈及了机器人的未来发展。

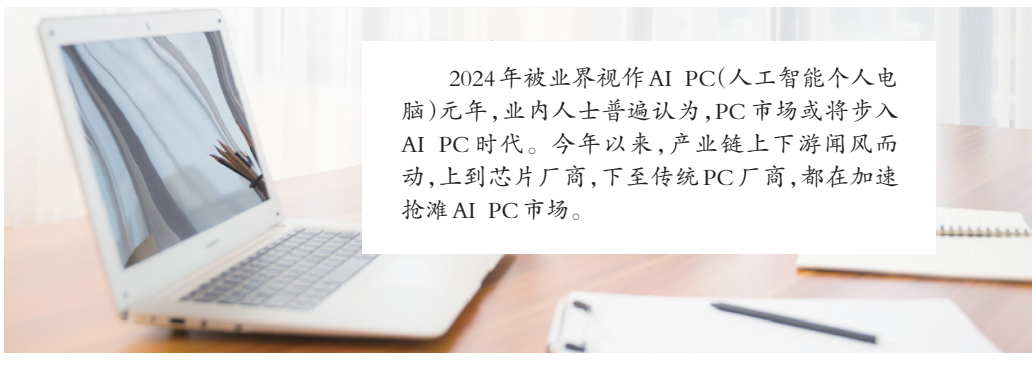
在黄仁勋看来,物理人工智能正引领人工智能领域的新浪潮,它们深谙物理定律,并能自如地融入我们的日常生活。

展望未来,机器人技术将不再是一个遥不可及的概念,而是日益融入我们的日常生活。

"机器人时代已经来临,这是人工智能的下一波浪潮。我们将见证一个更为激动人心的时刻——制造会走路、四处滚动的计算机,即智能机器人。"黄仁勋说。

综合《每日经济新闻》《深圳商报》 作者:杨卉 陈小慧

AI电脑成新动能 产业链"抢滩登陆"



2024年被业界视作AI PC(人工智能个人电脑)元年,业内人士普遍认为,PC市场或将步入AI PC时代。今年以来,产业链上下游闻风而动,上到芯片厂商,下至传统PC厂商,都在加速抢滩AI PC市场。

企业纷纷入场布局

根据行业研究机构IDC给出的定义,只要处理器中含有NPU就属于AI PC。"与传统PC相比,AI PC能够完成更多的运算,在更多场景辅助用户的工作、生活、学习、创意等。"IDC中国高级研究经理陈舒歆表示,更强大的辅助功能会在诸多场景带来新的用户需求。

AI PC市场高预期的背景下,吸引了英特尔、英伟达、AMD、高通等芯片厂商,以及联想、惠普、华为、苹果等PC厂商纷纷入场布局。例如,今年年初,英伟达在2024国际消费电子展(CES)上推出了三款面向消费者的全新显卡,包括RTX 4080 SUPER、RTX 4070 Ti SUPER和4070 SUPER GPU(图形处理器)芯片产品。英伟达表示,这是公司为AI PC设备专门推出的新的GPU芯片。

下游厂商中,联想也推出了一系列AI PC产品。此外,戴尔、惠普、华硕等头部PC厂商也相继发布了AI PC产品。

德邦证券发布研报称,2024年将成为AI PC量变的关键节点。据IDC预测,今年AI PC在中国的渗透率将从2023年的8.1%上升至54.7%,其有望在换机潮中加速渗透PC市场。

A股上市公司也颇为看好AI PC未来的发展,纷纷透露相关布局动作。

万祥科技近日在回复投资者提问时表示,AI PC是PC未来重要的发展趋势,AI PC的发展将带来PC能耗的提升,这将对PC的电池电芯提出更高的要求。公司产品主要围绕消费电子、智能穿戴设备的电池电芯业务。AI PC带来的PC换机需求、能耗和电芯需求的提升,都将为公司产品需求提供保障。

PC市场将重回增长轨道

事实上,PC行业此前长期处于低谷之中。据IDC发布的数据显示,截至2023年第四季度,全球PC出货量已经连续8个季度出现同比下降,市场在需求疲软和依赖大幅促销的情况下复苏缓慢。

不过,随着AI PC对PC行业的变革带动,PC市场萎缩似乎已经触底,将重回增长轨道。Canalys统计数据显示,今年一季度,全球PC出货量一改连续几个季度下滑的颓势,出货量达到5720万台,同比增长3.2%。

"未来,随着产业链上游厂商的不断推广,AI PC会快速替换部分原有PC。"陈舒歆表示,AI PC可以应对的场景很多,不论是B端还是C端都能提供较大的辅助作用。在工作、教育、学习、娱乐、生活、智慧出行等层面,都会为AI PC的需求量增长提供动能。

在一些传统PC厂商眼中,AI PC已被视作未来的业绩增长点。惠普CEO恩里克·洛雷斯预计,在2024财年下半年,惠普约10%的PC销量将来自AI PC,AI将在2025年和2026年对PC销售产生更大影响。

此外,尽管AI PC对于PC行业的整体带动尤为重要,但业内人士也认为,AI PC目前尚处于发展阶段,技术、应用、生态的最佳实践都还处在探索期,未来还有很大发展空间。

"AI PC处于刚起步阶段。"巨丰投顾高级投资顾问王旭表示,目前整个市场还没有开展大规模的相关布局。

一家上市公司近日在接受机构调研时表示,AI PC目前处于前期推广阶段,暂未出现必须通过更换PC才能满足包括生成式人工智能(AIGC)等功能需求的消费者体验产品。

据《证券日报》 作者:冯雨瑶

▶ 科工前沿

伏羲气象大模型升级 可服务新能源等行业

近日,在上海科学智能研究院(以下简称"上智院")和复旦大学联合举办的"走进智能气象"主题活动暨智能气象创新生态联盟成立仪式上,面向产业应用的伏羲系列气象大模型2.0(以下简称"伏羲"2.0)发布。

相较于去年推出的1.0系列,"伏羲"2.0的中期天气预报大模型和次季节大模型,面向新能源、航空运输等行业取得显著进展。

生态环境部应对气候变化司一级巡视员蒋兆理介绍,当前大数据、云计算、人工智能等新兴技术正在加速向经济社会和公共治理的各个领域融合渗透。"联盟只是一个开始,我很期待未来看到新兴技术在应对气候变化的更多领域发挥作用,助力经济动能转化、推动减污降碳深度融合、支持实体经济发展。"

以"伏羲"系列气象大模型为核心,当天成立的联盟将通过深度的产学研融合,围绕"新型电力系统与低碳转型、气候风险与韧性城市、金融市场投资与风险管理"这三大领域打造智能气象技术的产业生态。

例如,在新型电力系统与低碳转型领域,"伏羲"模型将助力优化可再生能源的并网和调度,提高电网的稳定性和可靠性,推动能源结构向低碳化转型;在气候风险与韧性城市领域,"伏羲"模型将增强城市对极端天气的预警和应对能力,为城市规划、基础设施建设提供气象风险评估,提升城市应对气候变化的韧性;在金融市场投资与风险管理领域,"伏羲"模型将为金融机构提供更精准的天气预报,优化天气衍生品的设计与定价,助力金融市场更好地应对和管理气候风险。 据《第一财经日报》 作者:金叶子

新成像技术问世 可"透视"晶体结构

近日,美国纽约大学研究人员开发了一项创新技术。该技术使人能够以前所未有的方式窥视晶体结构,仿佛赋予人眼X射线般的超能力。这项名为"晶莹剔透法"的新技术,将透明粒子、显微镜与激光技术相结合,使科学家能够看到构成晶体的每个单元,并据此创建出动态三维模型。

研究人员致力于开发一种方法,以可视化晶体内部的构建块。他们首先创建了透明的胶体颗粒,并添加了染料分子来做标记,从而在显微镜下可用荧光区分每个颗粒。

但仅靠显微镜还不够,研究人员转向了共聚焦显微镜成像技术。该技术利用激光束扫描材料,从染料分子中产生特定的荧光。这不仅能够揭示晶体的每个二维平面,还能将这些平面堆叠起来,构建出三维数字模型,并精确确定每个粒子的位置。这些模型可以旋转、切片和拆解,从而揭示晶体内部任何潜在缺陷。

在静态晶体中,他们用该技术观察了晶体孪生现象。此外,这项技术还允许科学家在晶体变化时对其进行可视化。例如,当晶体熔化时会发生什么?粒子会重新排列吗?缺陷会移动吗?在一项实验中,研究人员熔化了一种具有矿物盐氯化铯结构的晶体。他们发现,缺陷是稳定的,并未如预期那样四处移动。

为了验证静态和动态晶体的实验,研究人员使用计算机模拟来创建具有相同特征的晶体,并证实这一方法可准确捕捉晶体内部情况。这一突破性技术有望为构建更优质的晶体和开发与光相互作用的光子材料铺平道路。

据《科技日报》 作者:张佳欣

声明

遗失我公司公章(编号:3702140026539)一枚,声明作废。

油猴(青岛)汽车服务有限公司和阳路分公司
2024年6月5日

遗失我公司发票专用章(编号:3702131296993)一枚,声明作废。

山东珊旭川建筑工程有限公司
2024年6月5日

遗失我公司财务专用章(编号:3702131296992)一枚,声明作废。

山东珊旭川建筑工程有限公司
2024年6月5日

遗失我公司法人(马卫燕)章(编号:3702131296994)一枚,声明作废。

山东珊旭川建筑工程有限公司
2024年6月5日

公告

青岛福安安全技术服务有限公司:

本委受理的陈旭贵与你单位劳动报酬等争议一案已处理终结。现依法向你单位公告送达青黄劳人仲案字〔2024〕第827号仲裁裁决书,请自本公告发布之日起30日内到本委(地址:青岛市黄岛区水灵山路188号8号楼308室,联系电话:0532-58953781)领取仲裁裁决书,逾期不领取,即视为送达。

特此公告

青岛市黄岛区劳动人事争议仲裁委员会
2024年6月5日

公告

陈旭贵:

本委受理的你与青岛福安安全技术服务有限公司劳动报酬等争议一案已处理终结。现依法向你单位公告送达青黄劳人仲案字〔2024〕第827号仲裁裁决书,请自本公告发布之日起30日内到本委(地址:青岛市黄岛区水灵山路188号8号楼308室,联系电话:0532-58953781)领取仲裁裁决书,逾期不领取,即视为送达。

特此公告

青岛市黄岛区劳动人事争议仲裁委员会
2024年6月5日