AI催生巨量Token消耗 算力租赁供不应求

打破Chatbot聊天对话框,AI(人工智能)正走向物理世界。

2026年春节期间,阿里、腾讯、字节等AI大厂纷纷加入红包大战,全民AI应用热情被点燃。国外,以Open Claw、Cowork为代表的应用,引发了新一轮桌面Agent(智能体)热潮。

AI群星闪耀,胜负未分,Token(词元)却已狂飙。

摩根大通研报分析认为,这些大型互联网平台投入真金白银进行推广,是为加速用户从传统搜索转向聊天机器人,并培养智能体商业习惯。这一进程实际上推升了推理量,从而加快Token的消耗速度。

一边是Token巨量消耗,另一边是内存等硬件价格狂飙,加剧了算力租赁需求。算想科技CEO(首席执行官)曾宇近日透露,公司算力租赁从2025年初的2000卡GPU(图形处理器)规模,到如今已迈入万卡GPU规模。

AI算力需求持续增长,但政策对数据中心能耗约束趋严,PUE(电源使用效率,是衡量数据中心能源效率的核心指标)成为衡量绿色算力的一个关键指标。相比传统风冷PUE普遍在1.5以上,液冷方案PUE可降低至1.1—1.2,强劲驱动力正推动液冷市场实现数量级的规模扩张。

新安股份硅基新材料研究院院长刘继近期在接受采访时表示,在浸没式液冷技术选择路径上,是多种介质并存的格局,氟化液、硅油、合成油都有各自适用场景。其中,硅油冷却液从性能、安全环保与成本等方面综合考虑,有相对优势,未来有望迎来规模化应用。


AI生成

Token使用量指数级增长

“不要小看Open Claw,AI自动运行,可能一个小时就能把你的Token消耗完。”税友股份亿企赢首席产品官周源表示,AI大模型以及应用的发展,带来更多内存和存储的需求,对Token的消耗量也在持续增加。

事实上,2025年,国内外科技巨头日均Token使用量呈现指数级增长。

据火山引擎数据,截至2025年12月,字节跳动旗下豆包大模型日均Token使用量突破50万亿,较上年同期增长超过10倍,相比2024年5月刚推出时的日均调用量增长达417倍。据谷歌披露,截至2025年10月,其各平台每月处理的Token用量已达1300万亿,相当于日均43.3万亿,而一年前月均仅为9.7万亿。

一场基于Token用量飙升的算力连锁反应正在发生。

云服务厂商开始上调其云服务价格。2026年1月23日,亚马逊云宣布上调其EC2机器学习容量块价格约15%;1月27日,谷歌云正式官宣涨价,自2026年5月1日起,对GoogleCloud、CDN Interconnect、Peering以及AI与计算基础设施服务进行价格调整。

据报道,国内云厂商虽在通用计算领域保持价格稳定,但在高性能AI算力板块,受限于芯片供应与电力、散热等基础设施瓶颈,实际成本压力持续上升。

2026年1月,有业内人士透露,部分头部云服务商正在评估对专属AI集群或预留型算力产品进行结构性调价。若GPU资源持续紧缺,国产云平台或将在保障客户权益的前提下,适度跟进国际定价趋势,推动行业进入“价值导向”新阶段。

算力租赁行业预期增强

除了云计算厂商价格上涨,内存价格上涨正在让更多AI公司从自建算力转向算力租赁。

信达证券研报分析表示,AI大模型训练和推理需求的快速增长是此轮存储行业复苏的核心动力,高性能存储产品需求呈现爆发式增长,其中HBM(高带宽内存)作为AI服务器的核心配套器件,成为头部厂商的业绩增长支柱。

算想科技CEO曾宇向透露,随着内存和硬盘价格上涨,显卡GPU的价格也水涨船高,如今公司买的服务器价格比2025年初涨了30%左右。对应算力租赁新客户的价格也随行就市,根据硬件市场价格波动,算力租赁价格也会上涨20%—30%。

“现在只要有算力服务器,基本不愁租。很多和AI相关的企业原本想自己买服务器,但现在价格波动太大,大多数公司都希望避开这个高点,选择算力租赁。”曾宇表示,市场对算力的需求仍在增加,算力租赁目前基本处于供不应求的状态。

行业巨头真金白银的投资,进一步强化了市场对算力租赁行业的预期。

2026年1月26日早间(美国太平洋标准时间),英伟达宣布向CoreWeave投资20亿美元,以加速这家数据中心公司在2030年前增加超过5GW AI计算能力的进程。据了解,CoreWeave是一家美国的AI云端运算企业,公司成立于2017年,当前业务重心转向云计算及GPU算力租赁服务。

开源证券分析师认为,AI算力租赁是AI云基础设施IAAS(基础设施即服务)层核心组成部分,英伟达向CoreWeave追加20亿美元投资再次验证AI云基础设施高景气度。

算力市场有结构性矛盾

尽管眼下算力租赁火热,但行业内仍呈现分化趋势。

在IDC(互联网数据中心)行业深耕数十年的曾宇透露:“尽管传统算力租赁依然是主流模式,但我们更看好未来的方向。我们从液冷设计、ODM定制服务器,到云算力调度平台,从B端客户服务再到面向C端的边缘云算力应用,正在构建一个完整的闭环生态。我们要做的不是简单地出租算力,而是打磨产品,真正让用户感受到边缘算力的价值。”

此外,在曾宇看来,表面火热的算力市场,暗藏结构性矛盾。大量所谓“智算中心”由旧数据中心改造而来,单机柜功率提升后,空间与电力配套严重浪费。“很多项目根本转不动,卖不掉算力,又改不回通算。”

当AI市场正在从模型侧的角逐走向应用端的竞速,对算力的需求也在调整。“之前大家聚焦模型端的竞争,更需要的是训练算力,这类算力需要英伟达H100或者H200这类训练性能较强的GPU卡。但现在随着AI应用爆发,更多需求反而是体现在推理算力上。相对而言,推理算力可以有更高性价比的芯片选择,在算力布局上,并不完全依赖数据中心,边缘算力服务器节点也可以满足市场需求。”

爱芯元智创始人、董事长兼执行董事仇肖莘近日表示,AI的价值主战场正在端侧与边缘侧全面展开。

事实上,边缘计算是“云-边-端”协同体系的关键环节,市场正高速增长。根据灼识咨询的数据,全球边缘及终端AI推理芯片市场规模已于2024年达到3792亿元,预计至2030年将扩增至16123亿元,年复合增长率达27.3%。

不同于数据中心集约式布局,边缘算力的布局往往更灵活和机动。将市场聚焦边缘计算算力租赁,新一年,曾宇有一个非常清晰的目标:2026年达到5万卡GPU,并织就一张覆盖全国的边缘计算网。

曾宇透露,算想科技计划后续在人口密度大的城市周边布局算力节点。“哪里人多,哪里数字经济活跃,我们就去哪里,因为未来人多的地方AI应用的需求量大,推理算力的消耗也会更大,对边缘计算的需求也随之增长。”

绿色算力处于重要位置

随着更高功率算力集群的全面部署,数据中心的能耗密度呈指数级跃升。据高盛预测,到2027年,AI服务器单个机架的功率密度将是5年前普通云服务器的50倍。

微软CEO纳德拉直言:“供电能力是当前最大的瓶颈,甚至超过芯片。”英伟达首席执行官黄仁勋更是表示:“电力的可用性而非GPU,将决定AI的扩展规模和速度。”

绿色算力,被放在一个更加紧迫且重要的位置上。

随着“东数西算”工程全面展开,国家清晰规定新建大型及以上数据中心的PUE 需小于或等于1.25,改造后的存量数据中心PUE要小于或等于1.5。工业和信息化部发布的《新型数据中心发展三年行动计划》说明,全国数据中心的PUE普遍要降低至1.5以下,先进算力中心则会达到1.1左右。

相比传统风冷方案PUE普遍在1.5以上,液冷方案PUE可降低至1.1—1.2,也由此催生了液冷服务器的需求爆发。

刘继表示,随着人工智能发展,加之芯片功率越来越高,浸没式液冷服务器是未来行业趋势,行业增长规模也是指数级的。浸没式液冷未来将有非常大的发展空间。

曾宇透露,目前算想科技液冷服务器占比不到5%,大多数以风冷服务器为主。“但眼下就如同当年光伏发电取代煤炭发电的节点,后续,我们预计液冷服务器占比将达到60%—70%。”他强调,之前液冷服务器布局的成本相对较高,但2025年公司在液冷服务器设计布局上做了一些技术突破,平抑了部分成本,2026年预计将加大液冷服务器布局力度。

液冷冷却液成为数据中心散热核心材料,氟化液与有机硅油等冷却介质主导浸没式液冷市场。2026年2月初,有机硅龙头新安股份携手算想科技在杭州“中国数谷·未来数智港”落地首个商用浸没式硅基液冷算力项目,目前该项目已正式投入运行。

据了解,英伟达就采用陶氏化学的有机硅冷却液为浸没式液冷冷却介质,为高功耗GPU降温。而在国内硅基冷却液赛道,润禾材料、新安股份等企业已推出相关产品并实现销售。

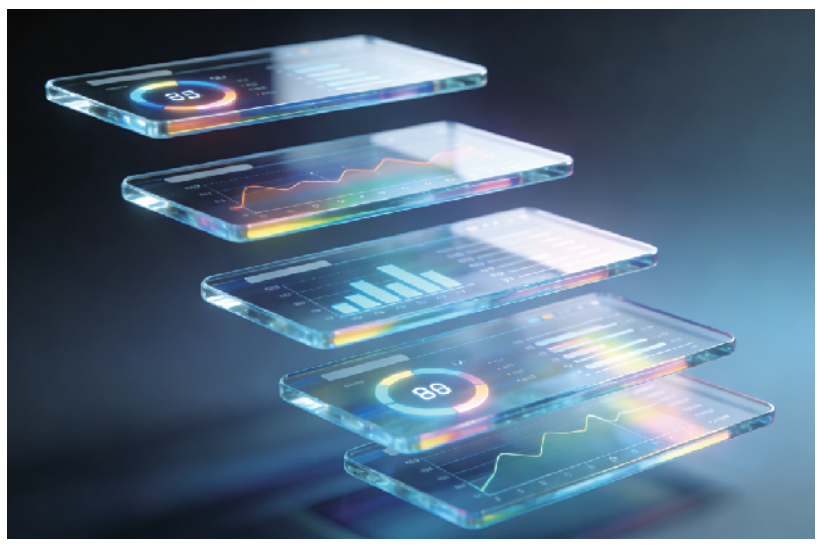
新安股份落地首个商用浸没式硅基液冷算力项目,采用高功率密度浸没式液冷架构,单机柜功率密度达210千瓦,可支持多卡并行计算需求,能够面向多类型算力客户提供服务。

“该项目验证了硅基液冷材料在真实负载环境下的可靠性与经济性,为后续规模化复制提供了成熟样板。公司在硅油领域拥有产业与技术基础,依托有机硅材料研发与工程化能力,推动有机硅从传统工业应用向算力基础设施等新兴应用场景延伸。未来公司将面向超算中心、分布式算力中心等场景提供解决方案,推动有机硅业务向高附加值终端应用升级。”刘继表示。

据《每日经济新闻》 作者:叶晓丹

一小块玻璃能存200部电影

玻璃数据存储技术领域取得重要突破



最新一期《自然》杂志发表的研究显示,美国微软公司在玻璃数据存储技术领域取得重要突破,首次实现在普通硼硅玻璃上长期稳定存储数据。这一进展使利用低成本、易获取的日常玻璃器皿(如耐热炊具)作为超长期存储介质成为可能。

此项研究基于微软自2019年启动的“硅计划”项目。此前,该团队仅能在昂贵的特制熔融硅玻璃上存储数据,而本次突破将介质拓展至广泛应用的硼硅玻璃,大大降低了材料成本与获取门槛。同时,研究团队改进了数据编码与读取方法,提升了技术的实用性。

实验过程中,研究团队在一块面积120平方毫米、厚2毫米的硼硅玻璃片上,以3.13兆字节/秒的速度,将4.8太字节数据(约相当于200部4K高清电影)分层写入,共计301层。尽管这一写入速度目前显著低于传统硬盘或固态硬盘,但其核心优势在于数据保存极端持久。通过加速老化测试验证,存储于玻璃中的数据预计可完整保存超过一万年,而目前常见硬盘的平均寿命仅为十年。

论文合著者、微软合伙研究经理理查德·布莱克指出,该成果解决了迈向商业化的一项关键障碍——存储介质的成本和可用性,同时实现了并行高速写入及长期稳定性验证。

玻璃存储技术的主要应用场景并非日常计算设备,而是面向需要永久或超长期保存数据的档案管理领域,例如文化遗产、科学资料、法律文书等数字资产的归档。此前,微软已提出利用类似技术在挪威全球音乐库中永久保存音乐作品的构想。

同期,其他研究机构在替代性长期存储技术上也取得进展。例如,有团队开发出基于DNA的数据存储方案,能在特定条件下将海量信息保存数万年,展现了生物介质在超高密度存储方面的潜力。这些技术共同指向一个方向:为人类不断增长的数字遗产寻找能够跨越千年的可靠存储载体。

从人类诞生起,我们就一直在留下信息。我们将信息留在龟甲上、岩壁上、竹简上……随着越来越多信息被存入数据空间,我们更加迫切地需要可靠的存储载体来抵御时间的侵蚀。现在,成本低廉的硼硅玻璃可以担此大任。未来,一只玻璃杯,就能成为一座博物馆。

据《科技日报》 作者:张梦然

科工前沿

全球首个!大规模“量子芯网”成功构建

量子密钥分发为通信安全提供了有效解决方案,我国科学家通过芯片化集成,让这项技术所需硬件更小型、更易应用。北京大学王剑威教授、龚旗煌院士与常林研究员团队,成功研制出全球首个基于集成光量子芯片的大规模量子通信网络,取名为“未名量子芯网”。这一网络能让20个芯片用户并行通信,组网能力可达3700公里,在芯片用户规模与组网能力上均达国际领先水平。相关成果日前发表于国际学术期刊《自然》。

“过去,实现量子通信的设备通常较为庞大和复杂,成本也比较高昂。这次,我们把这些关键器件集成到了小小的芯片上。”王剑威介绍,团队研发出了两款核心芯片——一款是网络中心的“光源芯片”,能为整个网络提供一个统一的节拍,确保所有用户通信的精确同步;另一款是用户手上的“通信芯片”,它单片集成了激光器、调制器、衰减器等全部关键功能模块,相当于把过去一整套体积庞大的发送设备,集成到了一个指甲盖大小的芯片上。

“更让人惊喜的是,这是国际上20余年来首次展示基于光量子芯片的量子密钥分发网络。实验表明,这两款核心芯片在制造时展现出了很高的一致性和良品率。这意味着,它们未来可以进行低成本、大规模的批量生产,也为构建更长距离、更多用户的量子密钥分发网络奠定了技术基础。随着小型化量子通信芯片进一步发展,未来它们有望广泛部署在通信网络节点和安全基础设施中,为各类终端提供安全保障。”王剑威说。

《自然》期刊4位匿名审稿人一致认为:“这是量子芯片和量子网络领域的重大突破,所展示的量子芯片网络具备显著的大规模扩展能力,无疑将对未来的量子通信产生深远影响。”

“这是集成光量子技术推动量子通信发展的范例。”龚旗煌认为,这项成果为构建大规模量子通信芯片网络提供了可行方案,对促进量子通信系统的小型化、实用化与规模化发展具有重要意义。 **据《光明日报》 作者:晋浩天**

什么是量子密钥分发?

假设你要给同事发加密文件,常用办法是压缩打包、设个密码,然后发过去。可你有没有想过——密码本身怎么安全地传过去?

量子密钥分发解决的就是这个问题。它不传文件,只传密码——而且是一种“看一眼就自毁”的密码。

这是如何做到的?它把密码编在一串光子里发出去。光子的特性是:只要被“偷看”,状态就会改变。这就像在黑板上写一串数字,有人偷偷进来看了一眼——他离开时,黑板已经被擦得干干净净。你同事进来只看到一块空黑板,你们一对就知道:“有人来过!”

因此,量子密钥分发的精髓是:不再赌攻击者的算力不够强,而是赌物理定律不会撒谎。偷看必留痕。

有了它,远隔千里的人也能确信:手里的密码只属于彼此,从没被任何人“看过”。 **据《光明日报》 作者:晋浩天**