AI就要“无底洞式”投入？变化正在发生

作者：赵广立 来源：中国科学报

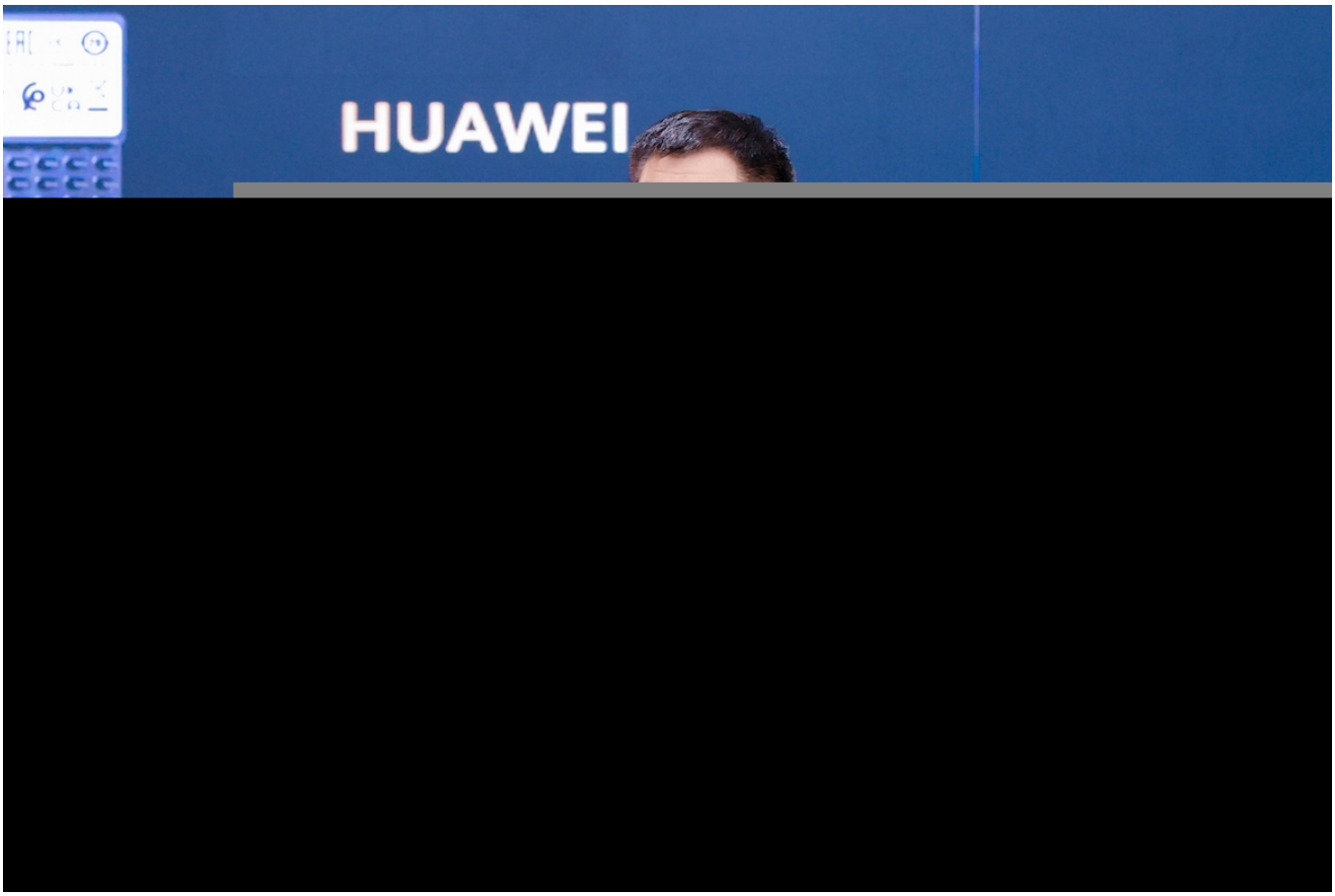
本文原地址：https://www.iikx.com/news/topnews/35450.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

在今年的世界人工智能大会(WAIC 2025)上，2024年诺贝尔物理学奖得主、被誉为“神经网络之父”的杰弗里·辛顿(Geoffrey Hinton)指出：“人类智能或许难以超越机器智能，因为机器具备永久记忆。”

这一观点引起华为公司副总裁、数据存储产品线总裁周跃峰的高度关注。他表示：“这说明，AI的能力不仅依赖于计算，也依赖于存储。”

作为信息系统的“记忆体”，存储产品在AI系统中扮演关键角色。周跃峰强调，随着AI大模型走向实际应用，人们日益意识到“存力”——即存储系统的性能与优化——在整体AI架构中的重要性。



周跃峰。华为供图，下同

AI SSD：为提升AI训推效率而生

“这些年AI虽然火热，但也遇到了许多难题，真正的应用落地、创造价值的案例较少，许多都卡在了训练、推理乃至微调环节。”上海人工智能研究院副院长杨浩在近期一次公开分享中提到，AI训练难、推理难、成本高，“很多企业承担不起”。

对于许多中小规模机构或企业而言，他们能触及的“AI集群”，大多只是一台训推一体机。

周跃峰向《中国科学报》介绍说，最常见的8卡的训推一体机，往往它的显存或高带宽存储(HBM)大小是一定的，遇到参数量大的模型，想去做推理就推不动。他举例说道，要训练一个参数量为671B的模型，需要的原始语料数据要超过3.5PB;拿它来做微调，需要计算集群系统的显存容量超过13TB。

但现实问题是，一台8卡训推一体机的数据存储上限还在GB级别，仅能支持在32B参数规模的模型微调。

类似训推一体机这样的“小盒子”，毕竟资源有限，怎么办?

“在模型微调的场景，我们建议应该配一些超快性能的AI SSD。”周跃峰介绍说，基于8张算力卡的内存，再搭配2张华为3.2TB的AI SSD，单机可用内存可提升至7TB，单机可微调模型参数可达235B。

据了解，自2024年底开始，各大存储厂商都在加速布局面向AI工作负载需求的SSD产品(SSD即固态硬盘，系常用存储介质)。相较于传统SSD，AI SSD以“性能更强、时延更低、耐用性更优、软件栈更适配”的核心特质，成为支撑AI基础设施的关键存储载体。

做好存算协同，胜过“拼命堆算力”

一个显而易见的现象是，AI对数据有多大依赖，就意味着给存储介质带来多大压力——在训练阶段，AI模型需要从PB级数据集中持续读取和写入数据;在推理阶段，面对多并发请求和多模态输入，系统对数据响应速度、带宽和稳定性的要求也持续攀升。只是，AI在提升“理解”与“生成”能力的同时，底层存储却往往难以支撑这种非线性增长的数据流量，这种不匹配造成了“AI存力瓶颈”，具体表现就是“内存墙”和“容量墙”。

目前，人们广泛使用HBM和DRAM作为AI存储，其中HBM常与GPU配合使用，DRAM则在更广泛的AI应用中作为主内存使用。两种存储介质虽然速度快，但容量受限，难以“吃下”迅速增长的大模型参数和历史KV Cache(缓存)，进而导致模型训练推理无法正常进行。

近期，华为推出了其自研的两类AI SSD：分别对应为助力扩展高带宽内存(HBM)和替代机械硬盘(HDD)，目的就是为了打破AI训推系统中普遍存在的“显存墙”和“内量墙”。

中国电子工业标准化技术协会数据存储专业委员会秘书长孙钢认为，在数据驱动智能的时代背景下，通过软硬件技术协同解决AI训推中的存储瓶颈，是一个“比不断堆叠昂贵算力更为经济可行的选项”。

“ AI时代，从算力驱动到数据驱动，存储产业正在发生跃迁。大家现在耳熟能详的‘ 以存强算 ’，印证了存力发展的必要性，也说明以前简单堆算力‘ 暴力计算 ’的方式既不经济，也不可持续。 ”孙钢表示，这并不是说算力不重要，而是在许多AI应用场景下，更多问题来自于存储瓶颈。



孙钢

从华为等企业已开始推出各类AI SSD来看，AI存储正朝着以技术创新打破“ 性能—容量 ”对立、实现二者平衡的方向发展。从单一介质的技术突破来说，需要一种能打破性能和容量的瓶颈的存储介质出现;从存储架构来说，需要构建一个智能协同的AI存储架构，即通过软件创新建设一个三级缓存架构(HBM-DRAM-AI SSD)，让不同性能、容量的存储高效协同工作。

从“ 堆硬件 ”到“ 提效能 ”：终结 AI“ 无底洞 ”式投入

随着生成式AI向Agentic AI(智能体协同群)迅速发展，AI模型的复杂度呈指数级攀升，Token规模从千亿级向万亿级甚至更高维度突破。持续加码的硬件采购、机房建设、能耗支出，正让不少企业陷入“ 无底洞 ” 式投入的困境。

“ 我们一直在思考，如何帮助企业在有限的AI基础设施资源下让AI训练和推理的性能得以最大化发挥。 ”周跃峰谈到，这一思考不仅道出了企业的普遍焦虑，更指向了Agentic AI时代AI基建的核心命题 ——从“ 堆硬件 ”的粗放式增长，转向“ 提效能 ”的精细化运营。

周跃峰说，这也正是华为的发力方向：无论是AI SSD等硬件产品，还是UCM推理记忆数据管理器等软件方案，其核心逻辑均围绕架构重构、技术创新与策略优化展开，目标就是将企业有限的

AI基建资源，高效转化为能落地的业务价值。

“我们观察到许多此前被认为是‘冷数据’正在被激活或者‘变暖’。孙钢分享说，当数据从冷到温、从温到热，原来的“成本”正在变成“价值”，AI时代存储的逻辑正在发生变化，各类围绕数据存储的创新也正纷至沓来。

更多 科研头条 请访问 https://www.iikx.com/news/topnews/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发