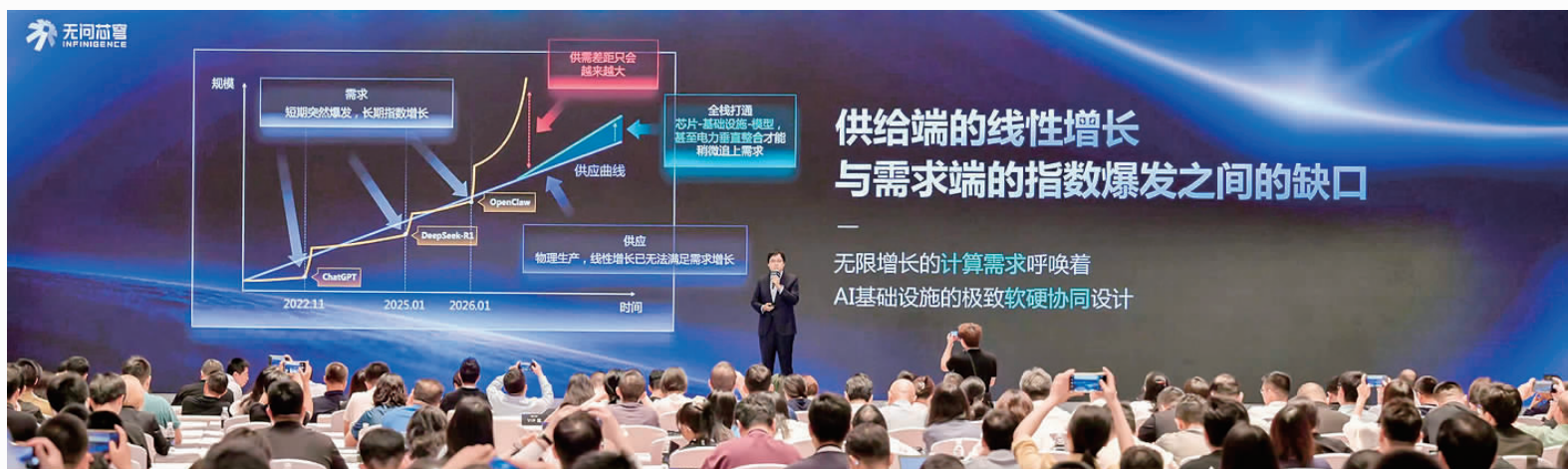


规模与效率双升级

无问芯穹亮出 AI 原生 Agentic Infra 新布局

文 / 记者 王永娟 图 / 资料

7月20日,2026世界人工智能大会上,扎根徐汇的科创企业无问芯穹现场发布面向 AGI 时代的 AI 原生基础设施完整体系 Agentic Infra,推出“前店后厂一中心”战略布局,同步上线面向大规模具身任务的云边端一体化管理与调度平台与配套技术框架。企业联合创始人兼 CEO 夏立雪现场解读整套技术方案,立足基础设施,锚定“规模与效率”,破解行业算力供需失衡、资源价值转化低效难题,为数字与物理世界的 Agentic AI 的持续进化提供支撑性力量。



算力集散中心: 全域异构算力统一调度 实现智能资源规模最大化

作为整套体系的底层支撑,“算力集散中心”指的是无问芯穹 Agentic Infra 自主式基础设施平台,重点解决算力碎片化、跨区域集群协同难等行业痛点。依托计算通信流水编排、自适应通信调度、异构算力调度三大自研技术,平台已完成三类生产级场景落地验证:联合中国电信搭建新疆哈密、广东深圳超 4000 公里跨域训练链路,集群性能提升 165%;打通四代不同型号 GPU 集群训练百亿参数大模型,协同效率提升 41%;打通私有集群与公有云混合算力,补齐企业算力错配短板,整体性能提升 68%。目前平台已整合全国超 37000P 算力,适配 16 类主流芯片。

通过计算通信重叠技术将跨域通信效率提升 3 至 4 倍,让训练不再因跨域通信产生额外耗时损耗;配套故障无感训练机制,可实现大规模跨域强化学习连续一周零中断稳定运行。夏立雪表示,依托并行策略、通信融合、智能算子、极致容错等底层技术积累,平台未来可拓展至十万卡级算力规模,充分支撑 Agentic AI 的长期在线进化。

Token 工厂: 全新跨集群异构 PD 分离架构落地 持续挖掘推理降本空间

同时,无问芯穹通过“Token 工厂”AgenticMaaS 大模型服务平台,在规模之上向效率要产能,让 Token 生产以“效”搏大。适配智能体业务长上下文、高频交互、高缓存复用的业务特征,本次大会,企业发布行业首创的新一代推理系统优化成果——跨集群异构 PD 分离架构(Prefill-RelayDecode-MainDecode, PDD)。将传统的 PD 分离链路“P-D”创新解构为“P-RLD-MD”的三级分离式推理架构,实测显示,该架构首 Token 延迟降低 51.5%,单 Token 生产成本下降 37.5%。不仅意味着用户端延迟体验的显著提升,更意味着,每一块芯

片都能被利用起来做自己擅长的工作,全域内的异构算力都将可以被最大化地释放产业价值。

截至今年 7 月,平台单日 Token 调用量较去年 12 月暴涨 40 倍,深度服务 Kimi、智谱等头部大模型企业;同时携手上海移动、AI 社区观猎,落地“天问”模型服务门户、TokenDance 词元跳动两大行业服务平台,面向全行业开放高效推理能力。

AI 生产力商店: 与行业伙伴共探 Infra 标杆方案 以智能蜂群实现基础设施自优化

依托算力集散中心与 Token 工厂,无问芯穹把广域的计算资源高效转换成了高质量 Token。然而当前,要实现相同的业务效果,基于不同的推理路径、模型规模和芯片,Token 成本却天差地别,因此,无问芯穹致力于通过 AI 生产力商店——Agentic Infra 行业解决方案,携手多行业伙伴打通 Token 资源向产业价值转化的通道,覆盖文娱、医疗、法律、能源多个垂直赛道,落地多项标杆项目:联合 Tripo 3D 打造 3D 游戏智能生成方案;为上海瑞金医院搭建医疗大模型专属算力资源与模型管理平台;推出律所私有化 AI 合伙人“律盾芯人”,服务多家国内顶尖律所;联合南网能源探索智算中心能耗智能调度方案。

不同于普通行业解决方案,无问芯穹也将智能体技术反向赋能 AI 基础设施,依旧遵循“用智能体赋能基础设施,提升计算规模和智能效率”的核心目标,打造无问芯穹基础设施智能体蜂

群。包含全局统筹者和用户入口的平台管家智能体系统,可以独立处理 80% 平台日常问题。

布局具身智能: 云边端一体化平台 打通数字与物理世界 AI 边界

本次发布同步延伸至物理世界具身智能赛道,无问芯穹推出国内首个面向大规模具身任务的云边端一体化调度平台,统一管理云端 GPU、边缘算力节点与实体机器人设备,配套两大自研核心技术升级:大规模真机强化学习训练框架 RLInf V0.3、机器人端侧全栈推理优化体系 Mizar-Robo。

RLInf V0.3 具身智能大规模真机、跨域端云协同的强化学习训练,独创 HotSwarm 模式保障真机设备上千次在线上下线,训练 0 中断,同时通过端云协同技术,成功让 100 台真机规模的训练专线需求降至 2Gbps 以下,大幅降低具身智能规模化训练的成本门槛。Mizar-Robo 通过流水线、算子、量化、计算图四层协同优化压缩端侧推理时延。在与自变量机器人 WALL-OSS 模型联合优化中,实现推理性能提升 7.14 倍,端侧推理时延从 500ms 压缩至 70ms,有效降低机器人规模化商用门槛。

“让智能无所不能,是 AGI,而让智能无处不在,是 AGI Infra。”夏立雪表示,无问芯穹将持续以“规模、效率”为两大核心支点,为未来数字与物理世界的所有 AI 生产力的运行构筑基础设施,助力上海人工智能产业高质量发展。

