美股半导体与硬件板块

投资者调研：RAISE 全球 AI 峰会核心观点汇总
2026 年 7 月 12 日 美东时间 21:15:43

我们全程参与上周 2026 RAISE 全球人工智能峰会，整理核心行业观点如下。

AI 算力需求

海量 Token 消耗型智能体工作流、实体代码生成应用爆发，催生前所未有的 AI 推理基础设施需求。当前机器学习算力需求同比增速接近 10 倍。与互联网泡沫时代不同，当下企业布局 AI 基础设施均具备实打实营收支撑，大量公司年营收达数亿乃至数十亿美元。尽管需求前景极度向好，但行业高度依赖私人信贷、债务融资扩产，或成为本轮资本开支周期的潜在风险。

AI 智能实习生替代传统对话机器人
头部 AI 实验室逐步淘汰传统聊天机器人，转向原生 AI“智能实习生”产品，内置代码工具与自主研究能力。先进模型可自主辅助工程开发、软件编写、科研流程，实现自我迭代加速。这套正向循环飞轮，将主流大模型迭代周期从以往 6-9 个月压缩至月度更新，创新速度大幅提升。

硬件解耦架构成为主流
云厂商快速落地解耦式推理架构，充分释放现有英伟达 GPU 集群效率。不再单一依赖 GPU 算力集群，厂商针对不同负载搭建专用基础设施，覆盖低延迟推理、实时工具调用、智能体任务执行等场景。随着 AI 负载类型持续多元化，这套架构转型的战略意义持续凸显。

CPU 需求强势复苏
GPU 仍是模型训练主力平台，但伴随智能体 AI 规模化落地，CPU 迎来需求大幅回暖。复杂推理链路、调度管理层、强化学习环境、企业级 AI 智能体均重度依赖 CPU 运算，通用计算芯片的市场空间持续扩容。

存储芯片供给紧缺
数据中心将消耗全球约 70% 存储芯片，供给缺口将传导至全电子产业链。头部云厂商已签订多年长协锁定产能；行业供需格局显示，至少 2028 年新增产能大规模投产后，供给紧张局面才会实质性缓解。

云厂商自研专用芯片提速
传统 CPU、GPU 架构受网络、功耗、能效瓶颈约束愈发明显，各大云厂商加速自研适配 AI 负载的专用芯片，针对内部运行的各类模型、推理任务、智能体工作流做定制优化。谷歌最新芯片路线图清晰体现专业化思路：推出 TPU 8i 低延迟推理芯片、TPU 8t 大规模训练芯片两套独立硬件平台，代表全行业向负载专用化算力架构转型。

KV 键值缓存成为推理核心瓶颈
多轮次智能体工作流成本持续走高，云厂商将优化上下文内存管理作为核心攻坚方向。核心

优化方案：将 KV 缓存数据从高价 GPU 显存转移至低成本存储介质，减少重复计算、提升硬件利用率。该方案缓存命中率可超 95%，首 Token 生成延迟降低 20 倍，推理成本显著下降。同时闪存芯片整体紧缺，搭载机械硬盘的混合存储架构重新大规模用于海量 AI 数据集存储。

推理延迟构建核心竞争壁垒

企业场景中，推理响应速度直接决定用户使用意愿与工作效率。更低延迟减少流程卡顿、降低上下文切换成本，支撑用户运行更复杂的代码编写、数据分析任务。Cerebras 硬件运行主流大模型可达每秒 750 Token，速度较标准 GPU 集群高出一个数量级。

边缘 AI 与轻量化小模型

中心云基础设施承载压力持续加大，行业加速向边缘计算、专用轻量化小模型（SLM）迁移。企业广泛部署行业定制小模型，以极小硬件开销实现垂直领域专业能力。英伟达 RTX Spark 超级芯片可让 AI 智能体直接在消费级、商用 Windows 设备本地运行，无需依赖云端算力。

