计算机行业跟踪周报

构建数据库的“CUDA”，英伟达存储变革下软件重构

增持（维持）

2025 年 12 月 07 日

证券分析师 王紫敬
执业证书：S0600521080005
021-60199781
wangzj@dwzq.com.cn
证券分析师 王世杰
执业证书：S0600523080004
wangshijie@dwzq.com.cn

投资要点

- **AI 推理带来新的 GPU 存储架构：**AI 训练需要大数据块（10MB-1GB），少并发，总存储容量相对较低（1-10TB）。而 AI 推理完全不同，需要小数据块（几 KB 或更小），高并发（数千条），大存储容量（PB 级）。传统架构以 CPU 为中心，CPU 的串行任务特点无法满足高并发需求，导致 AI 推理出现瓶颈，GPU 未能充分利用。GPU 地位亟需提升，把控制路径和数据路径都放在 GPU 里，硬件方面，GPU 直连 SSD 增加存储量和传输速率，软件方面，SCADA 软件架构控制存储 IO。

- **底层硬件变革带动软件重构，GPU-Native 数据库呼之欲出：**架构层面发生变化。从“以 CPU 为中心”到“以 GPU 为中心”，GPU 成为主计算单元。传统数据库以 CPU 为中心进行设计，数据库软件需要围绕 GPU 的数据获取和处理能力重新设计。核心组件层面升级改造。例如存储引擎的革新、数据布局优化、查询执行引擎的重构等。 **GPU 直连 SSD 技术将使得数据库从一个在通用操作系统上运行的、管理磁盘文件的应用程序，演变为一个直接调度和管理 GPU 和存储硬件的“数据中心级操作系统内核”。GPU-Native 数据库市场空间打开。**

- **产业进展逐步加快：硬件方面：**2025 年 8 月，闪迪与 SK 海力士签署谅解备忘录，共同制定 HBF 技术规范并推动标准化进程。双方目标在 2026 下半年发布 HBF 样品，首批搭载 HBF 的 AI 推理系统预计于 2027 年初面世。2025 年 9 月 2 日在东京都内举行的面向 AI 市场的技术说明会上，铠侠表示将与英伟达合作，开发可直接连接到 GPU 并进行数据交换的 SSD。**软件方面：**Hammerspace 已经通过更快更可扩展的元数据读取功能以及在 GPU 服务器直连存储驱动器中的更优数据放置策略，加速了其数据编排平台软件的性能表现。Cloudian HyperStore：通过 RDMA over S3 技术，实现对象存储与 GPU 内存的直接数据传输，使基于 S3 接口的向量数据库性能提升 8 倍。

- **投资建议：**随着 AI 推理的爆发，GPU 地位需要进一步提升，取代 CPU 成为数据流的核心，在硬件架构上直连 SSD，快速传输数据，充分发挥 GPU 的并发性能。在 GPU 直连 SSD 的创新架构下，软件生态也需要发生重大变化。核心技术软件数据库将不再以 CPU 为中心，转而以 GPU 为中心，针对 GPU 直连 SSD 新型架构做重构，满足 AI 推理旺盛的需求。数据库产业有望迎来新机遇。

- **相关标的：**【星环科技】，达梦数据，海量数据，MongoDB，Snowflake 等。

- **风险提示：**技术发展不及预期；AI 产业发展不及预期。

行业走势

相关研究

《商业航天：奇点时刻，航天强国》

2025-12-03

《十五五规划说明中，为何没有重点提及人工智能？》

2025-11-12

内容目录

图表目录

1. AI 推理时代来临，GPU 直连 SSD 存储新架构出现

颠覆 CPU 主导时代，GPU 全面接管存储 IO。

AI 推理与训练的 IO 需求差距很大。AI 训练依赖海量数据的批量传输，单轮数据块尺寸通常在 MB 级以上，控制路径的延迟占比极低；而 AI 推理完全不同，LLM 推理的 KV 缓存访问粒度仅 8KB-4MB，向量数据库检索、推荐系统的特征读取更是低至 64B-8KB，但需要支持数千条并行线程的并发请求。LLM 推理的存储需求已突破 10TB 级，向量数据库和推荐系统的存储规模更是达到 1TB-1PB，**这种“小块高频”的访问模式，让传统存储架构不堪重负。**

图1：不同 AI 任务对存储的要求有很大差异

Area	Usage model	Applications	Access granularity	Total size /worker
Training	Checkpoint save/restore	LLM pretraining, fine tuning	10MB – 1sGB	1-10TB
Inference	KV context caching across queries, docs	LLM inference	8KB – 4MB	>10sTB
	LLM+GNN, GNN+LLM	Contextual LLMs	512B – 8KB	5TB – 400TB
	Vector database	Dynamic Index build	64B – 4KB	6.4Gb – 20TB
		LLM RAG doc retrieval	512B – 8KB	400GB – 1PB
Predictive AI		Graph RAG	64B – 8KB	400GB – 1PB
		Recommenders	64B – 4KB	5TB – 400TB
	GNN induced subgraphs	eCommerce, fraud, social networks	512B – 8KB	>2TB
	Anomaly detection	eCommerce, fraud, social networks	512B – 8KB	>10TB
	Relational graphs	Data Science Automation	8B – 4KB	>100sTBs

数据来源：英伟达，东吴证券研究所

AI 工作负载正在根据其 I/O 模式（访问粒度和强度）分化为两大类，这正在推动存储评估指标从传统的“每 TB 成本”（TB/TCO）转向新兴的“每 IOPS 成本”（IOPS/TCO）。

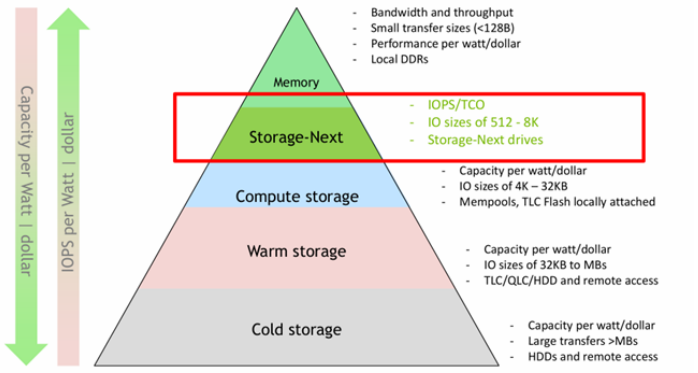
工作负载分化：

第一类：训练 (Training)。 以 LLM 预训练为代表，其特点是大块顺序 I/O（10MB-1GB）。这类应用更关注存储的吞吐量和总容量，因此传统的 TB/TCO（每 TB 成本）指标依然适用。

第二类：推理 (Inference) 和 预测式 AI (Predictive AI)。 包括 LLM 推理、RAG、向量数据库、推荐系统和图计算。这类应用的共同特点是极小的随机 I/O（访问粒度低至 8B、64B、512B）和极大的数据集（高达 1PB 或数百 TB）。

IOPS 成为存储新的挑战。对于推理和预测式 AI，性能瓶颈不再是存储容量或顺序吞吐量，而是系统处理海量、高并发、小 I/O 请求的能力，即 IOPS（每秒读写操作次数）。例如，RAG 检索、图谱遍历、推荐系统都需要极低延迟地从庞大的数据集中随机读取微小的数据块。

图2：需要新的存储层级出现



数据来源：英伟达，东吴证券研究所

传统架构中，CPU 同时掌控存储 IO 的控制路径（发起请求、调度资源）和数据路径（传输数据），GPU 仅作为“辅助加速器”被动接收数据。

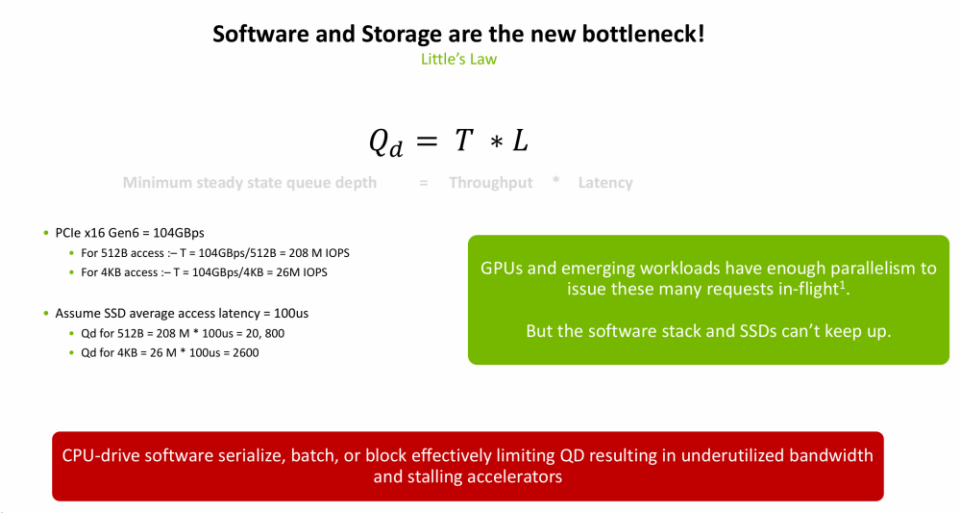
当前以 CPU 为中心的数据加载架构（将 GPU 视为“卸载设备”）已成为 GenAI 工作负载的瓶颈。

AI 工作负载的极端并行需求： 根据利特尔定律，为了充分利用现代硬件（如 PCIe Gen6）来处理 AI（如 RAG）的 512B 小 I/O，系统必须维持一个高达 20,000+的队列深度(Qd)。

GPU 并非瓶颈： GPU 的并行架构（如上一张 PPT 所提）有能力 发出如此海量的并发 I/O 请求。

真正的瓶颈是 CPU 软件栈： **问题的根源在于传统的、由 CPU 驱动的软件栈。**这个软件栈（即上一张图的 "Current Approach"）习惯于"串行化" (serialize) 或"批处理" (batch) I/O，这会人为地压低系统实际的队列深度 (Qd)。

图3：软件和存储是新的瓶颈

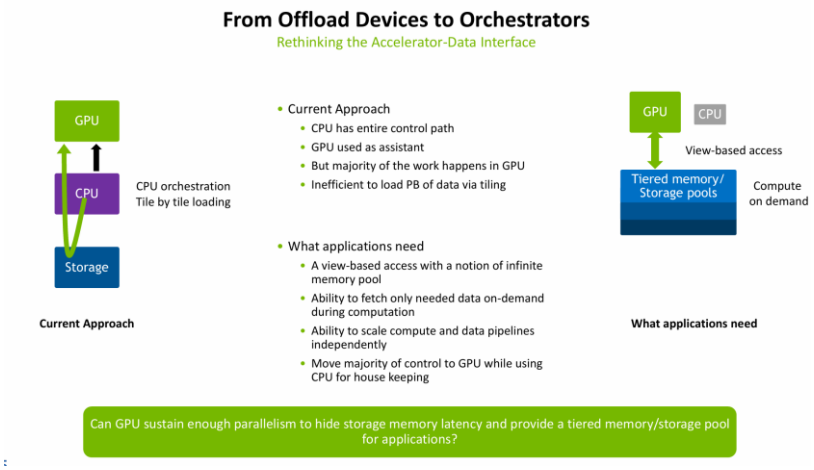


数据来源：英伟达，东吴证券研究所

GPU 地位提升，实现控制权的反转。GPU 成为“编排器”，取代 CPU 成为数据访问的控制中心。CPU 被“降级”，仅负责辅助性的“内务管理” (house keeping)。

数据访问模式从 CPU“推送” (push) 数据块，转变为 GPU“拉取” (pull) 数据。GPU 只在计算需要时才“按需” (on-demand) 从一个统一的分层存储池中抓取它需要的数据。

图4：GPU 地位提升



数据来源：英伟达，东吴证券研究所

通过硬件 GPU 直连 SSD 和 SCADA 软件架构实现 GPU 地位的提升。

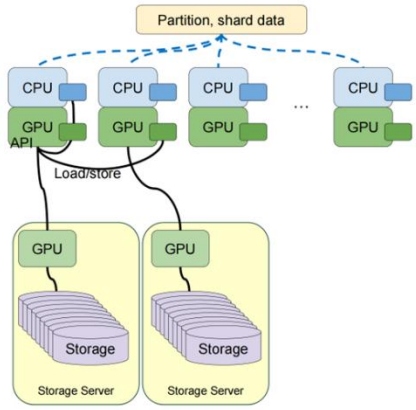
GPU 直连 SSD 允许 GPU 绕过 CPU 和系统内存，直接、高效地从固态硬盘读取和写入数据，是通过 NVMe-of、RDMA、GPUDirect Storage 等技术协议实现的一条优化的直接数据通路。

SCADA 是一个用于解决 AI I/O 瓶颈的、可扩展的、生产级的软件架构。

SCADA 通过两个关键技术解决了“CPU 软件栈”瓶颈：服务器端：使用 uNVMe (用户态驱动) 绕过内核，实现极致的 IOPS。客户端：GPU 应用线程成为数据请求的发起者。传输中：“数据路径"协议（可能基于 RDMA 和 GPUDirect）允许数据从服务器存储直接流向 GPU 显存，最小化 CPU 负载和延迟。

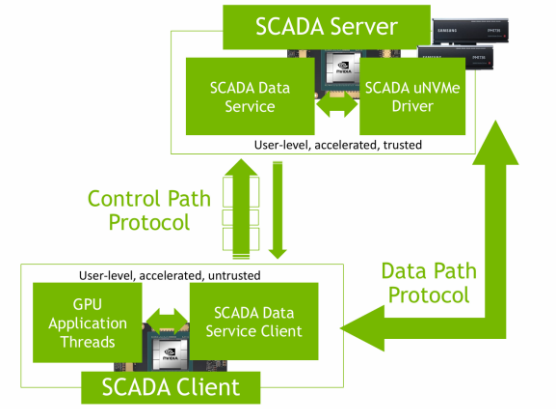
GPU 地位的提升。这个架构实现了“GPU 作为 I/O 编排器”的愿景。GPU 应用线程（通过 SCADA Client）发起请求，数据（通过 Data Path）直接流入 GPU，**CPU 在数据流中被彻底旁路。**

图5：GPU 直连多个 SSD



数据来源：OCP，东吴证券研究所

图6：GPU 成为核心，CPU 被跳过



数据来源：英伟达，东吴证券研究所

2. 存储架构变化带来数据库架构的变化

架构层面发生变化。从“以 CPU 为中心”到“以 GPU 为中心”，GPU 成为主计算单元。传统数据库以 CPU 为中心进行设计，数据库软件需要围绕 GPU 的数据获取和处理能力重新设计。CPU 的角色退化为任务调度器、事务协调器和元数据管理器。存储层级的虚拟化与重构。

核心组件层面升级改造。

存储引擎的革新。传统的、基于系统内存(DRAM)的缓冲池(Buffer Pool)管理机制效率降低。新的缓存管理器需要直接管理 GPU 显存和直连 SSD 之间的数据流动。

数据布局优化：为匹配 GPU 的 SIMD(单指令多数据)架构，数据在 SSD 上可能更倾向采用纯列式或混合式存储格式，并原生支持 Apache Arrow 等零拷贝内存格式，方便 GPU 直接消费。

查询执行引擎的重构。算子的 GPU 原生实现:扫描(Scan)、连接(Join)、聚合(Aggregation)、排序(Sort)等核心算子需要深度重写为 GPU 内核，并能直接从 SSD 流式消费数据。异步、流水线执行:查询计划被组织成高效的 GPU 内核流水线，当前步骤在 GPU 计算时，下一步所需的数据已通过直连路径在后台从 SSD 预取，实现计算与/O 的

完全重叠

查询优化器的挑战。成本模型剧变:传统的基于 CPU 周期和磁盘寻址的代价模型失效。新模型需纳入 GPU 计算核心占用率、HBM 与 SSD 间的带宽、PCIe 传输延迟等新因素。

数据本地性优化。优化器在生成执行计划时，必须优先考虑数据在 GPU 显存、直连 SSD、网络存储中的位置，尽量将计算调度到离数据最近的处理器上。

为了适应这一新型硬件架构，数据库软件需要深度架构改造、适配结合 GPU 软件技术栈（特定驱动、CUDA 库）、GPU 硬件生态。数据库将与 CXL 内存池化、NVLink 等高速互连技术结合，向存算一体和池化架构演进，成为 AI 时代数据基础设施的核心引擎。

GPU 直连 SSD 技术将使得数据库从一个在通用操作系统上运行的、管理磁盘文件的应用程序，演变为一个直接调度和管理 GPU、SSD 计算和存储硬件的“数据中心级操作系统内核”。

3. 产业进展逐步加快

硬件方面:

HBF 新型存储是未来。2025 年 8 月，被业界誉为"HBM 之父"的韩国科学技术院（KAIST）教授金正浩提出“AI 时代的力量平衡正从 GPU 向存储领域转移。”金教授指出，在人工智能时代存储器件将扮演日益关键的角色，甚至预言英伟达未来可能收购存储企业。他特别强调高带宽闪存（HBF）的战略意义，预计该技术将在 2026 年初取得突破，并于 2027 至 2028 年间正式亮相。

2025 年 8 月，闪迪与 SK 海力士签署谅解备忘录，共同制定 HBF 技术规范并推动标准化进程。双方目标在 2026 下半年发布 HBF 样品，首批搭载 HBF 的 AI 推理系统预计于 2027 年初面世。值得关注的是，在 10 月中旬举办的 2025 OCP 全球峰会上，SK 海力士首次展示了搭载 HBF 技术的"AIN B 系列"存储产品。

铠侠将与英伟达合作，推出直连 GPU 进行数据交换的 SSD 。2025 年 9 月 2 日在东京都内举行的面向 AI 市场的技术说明会上，铠侠 SSD 应用技术部门首席工程师福田浩一表示，“将按照英伟达的建议和要求进行开发”。迄今为止，SSD 一般通过 CPU（中央处理器）与 GPU 连接。铠侠将与英伟达合作，开发可直接连接到 GPU 并进行数据交换的 SSD。英伟达表示，与 GPU 连接的 SSD 需要达到 2 亿 IOPS，将以 2 个 SSD 应对这一需求。计划支持被称为 PCIe（PCI Express）的 SSD 接口的下下代标准 “PCIe 7.0”。

软件方面:

Hammerspace 已经通过更快更可扩展的元数据读取功能以及在 GPU 服务器直连

存储驱动器中的更优数据放置策略，加速了其数据编排平台软件的性能表现。

Cloudian HyperStore: 通过 RDMA over S3 技术，实现对象存储与 GPU 内存的直接数据传输，使基于 S3 接口的向量数据库性能提升 8 倍。

4. 投资建议

随着 AI 推理的爆发，GPU 地位需要进一步提升，取代 CPU 成为数据流的核心，在硬件架构上直连 SSD，快速传输数据，充分发挥 GPU 的并发性能。在 GPU 直连 SSD 的创新架构下，软件生态也需要发生重大变化。核心技术软件数据库将不再以 CPU 为中心，转而以 GPU 为中心，针对 GPU 直连 SSD 新型架构做重构，满足 AI 推理旺盛的需求。数据库产业有望迎来新机遇。

相关标的:【星环科技】，达梦数据，海量数据，MongoDB，Snowflake 等。

5. 风险提示

技术发展不及预期。GPU 直连 SSD 技术尚在研发阶段，如果技术研发进展不及预期，会影响对应软件产业进展。

AI 产业发展不及预期。新的硬件架构主要为了满足 AI 推理大幅增加的需求，如果 AI 应用进展不及预期，会影响产业发展进展。

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn