How AI Inference Is Creating New Memory Demand

AI 推理如何创造新的内存需求

KV cache offloading and agentic AI are its key drivers

KV 缓存卸载和智能 AI 是它的动力因素

By TrendForce

Test-time Scaling "Thinking" = 5X Tokens Per Year
推理时扩展（Test-time Scaling）"思考" = 每年 Token 增加 5 倍
Model weights
KV Cache
SSD POD Demand Driven by KV Cache Offloading
KV Cache 卸载驱动 SSD POD 需求
Agentic AI: More CPUs, More RAM, GPU Still Key
智能 AI：更多的 CPU、更大的内存，GPU 仍是关键

To address the surging KV cache storage demands of the AI inference era, Nvidia has introduced the CMX Context Memory Storage Platform in January 2026, leveraging BlueField-4 DPUs to create a new tier between GPU memory and shared storage.

为了应对在推理时代激增的 KV 缓存存储需求，Nvidia 于 2026 年 1 月推出了 Context Memory Storage Platform（CMX 上下文内存存储平台），利用 BlueField-4 DPU 在 GPU 内存和共享存储之间创建一个新层级。

This platform, dedicated to the storage of KV cache data, enables AI clusters to effectively scale context windows, cut GPU idle time, and boost token throughput for long-context inference workloads. It marks a key shift in how AI infrastructure is being built to support agentic AI and reasoning models.

该平台专门用于存储 KV 缓存数据，使 AI 集群能有效扩展上下文窗口、减少 GPU 闲置时间、并提升长上下文推理工作负载的 token 吞吐量。这标志着 AI 基础设施建设方式的一个关键转变，旨在支持智能 AI 和推理模型。

Related report: Sector Shifted Industry Analysis - 2026

相关报告：行业板块转移分析 - 2026

Test-time Scaling "Thinking" = 5X Tokens Per Year

推理时扩展（Test-time Scaling）"思考" = 每年 Token 增加 5 倍

According to Nvidia's public data, the average output token count per question has surged at a rate exceeding 5x per year since the reasoning era began in late 2024. Reasoning models, such as OpenAI's o1/o3 and DeepSeek R1, require far more tokens per query than standard LLMs, fundamentally reshaping inference economics. This explosion in token generation creates directly more demand on memory and compute infrastructure.

根据 Nvidia 的公开数据，自 2024 年下半年开启推理时代以来，每个问题的平均输出 token 数量以超过每年 5 倍的速度飙升。诸如 OpenAI 的 o1/o3 和 DeepSeek R1 等推理模型，在每个查询上需要的 token 数量远多于标准 LLM，从根本上重塑了推理经济学。这种 token 生成的爆发式增长直接对内存和计算基础设施产生了更多需求。

Total Size of Model Weights = Parameters × Bytes per parameter

模型权重总大小 = 参数量 × 每个参数的字节数

KV Cache

KV cache stores the key-value vectors generated during the inference prefill stage, avoiding redundant computation during the decoding stage. It is a critical component for conversational context retrieval and long-context inference. Mathematically, KV cache memory consumption scales with:

KV 缓存（KV cache）用于存储推理预填充阶段生成的键值向量，以避免在解码阶段进行冗余计算。这是对话上下文检索和长上下文推理的关键组件。从数学角度来看，KV 缓存的内存消耗与以下因素成正比：

The formula for calculating the total size of KV cache is:

计算 KV 缓存总大小的公式为：

Total Size of KV cache (Bytes)

KV 缓存总大小（字节）

= 2 x number of layers x number of KV heads x head dimension

= 2 x 层数 x KV 头数 x 每个头的维度

x sequence length x batch size x precision (Bytes)

x 序列长度 x 批次大小 x 精度（字节）

SSD POD Demand Driven by KV Cache Offloading

KV Cache 卸载驱动 SSD POD 需求

As KV cache memory footprint scales dramatically with conversation length and batch size, management and storage requirements quickly exceed what GPU memory alone can handle. While storing KV cache in shared storage is technically feasible, the system must discard its cache if it exceeds the system's total cost of ownership (TCO).

随着对话长度和批次大小（batch size）的增加，KV cache 的内存占用量急剧膨胀，其管理与存储需求很快就超出单纯 GPU 内存的承载能力。

To address this, Nvidia released Dynamo, a framework that intelligently offloads KV cache to storage tiers, offering infrequently accessed KV cache to lower-tier NVMe SSD storage, and moves only the data required for the next GPU SSD, ensuring this data remains reusable during the inference cycle rather than being discarded or recomputed.

为了解决这一问题，Nvidia 于 2025 年 3 月发布了 Dynamo，这是一个能够智能地将 KV 缓存卸载到存储层级的框架，将不常访问的 KV 缓存提供给更低层级的 NVMe SSD 存储，仅将下一个 GPU 所需的数据移回。这确保了数据在推理周期内可复用，而无需被丢弃或重新计算。

Dynamo 引入 Nvidia 全新 2 级 KV 缓存层级，以优化 SSD 容量分配。（图片来源：Dynamo Memory Storage Platform）。随着时间，这不仅会显著降低推理成本，还可实现更快的响应时间。

Note: An SSD POD is a standalone storage unit comprising multiple SSD racks, dedicated to storing offloaded KV cache. The POD's capacity is optimized to maximize the number of tokens served, balancing storage costs against the local SSD TCO for storing KV cache, while offering lower latency than accessing cache from HDDs.

注：SSD POD 是一个由多个 SSD 机架组成的独立存储单元，用于存储卸载的 KV 缓存。POD 的容量经优化，以最大化所提供的 token 数量，在存储成本与本地 SSD 存储 KV 缓存的 TCO 之间保持平衡，同时提供比从 HDD 访问缓存更低的延迟。（图片来源：TrendForce）

Agentic AI: More CPUs, More RAM

智能 AI：更多的 CPU、更大的内存

The proliferation of AI inference is also accelerating the disbanded of agentic AI workflows, which place heavy demands on CPU, memory, and storage. Unlike chatbots, agentic AI systems must execute tasks, call tools, make decisions, and maintain persistent state across multi-step workflows. This shifts a bottleneck previously dominated by GPU compute toward agent orchestration. As a result, the CPU-to-GPU workload ratio in agentic AI deployments is shifting from the traditional ratio of 1:4 to a tight ratio of 1:1 in some cases, underscoring the growing importance of CPU capacity in AI infrastructure.

AI 推理的普及也在加速分解智能 AI（Agentic AI）的工作流程，这对 CPU、内存和存储提出了高要求。与聊天机器人不同，智能 AI 系统必须执行任务、调用工具、做出决策并在多步工作流中保持状态。这将原先由 GPU 计算主导的瓶颈转向智能体编排。因此，在智能 AI 部署中，CPU 与 GPU 的工作负载比正从传统的 1:4 在某些情况下变为紧张的 1:1，这凸显了 CPU 容量在 AI 基础设施中日益重要的作用。

Related report: SSD Supply Analysis - 2026

相关报告：SSD 供应分析 - 2026

CPU-GPU Ratio Shift in Agentic AI Applications

In 2026, Nvidia introduced the Vera CPU, a custom Arm-based agentic AI workloads. Based on the organization's roadmap, Vera's performance is expected to double that of its predecessor, Grace, while the capacity of the previous generation Grace CPU.

2026 年，Nvidia 推出了 Vera CPU，这是一款基于 Arm 的定制智能 AI 工作负载的 CPU。根据组织路线图，Vera 的性能预计将是其前代产品 Grace 的两倍，其上一代 Grace CPU 容量的三倍。

Note: NVLink has declined to halve the DDC-DRAM latency, which remains significantly higher than local DDR modules, according to SemiAnalysis's latest findings. This adjustment does not reflect a reduction in NVIDIA's overall memory dependence. It is a response to Vera's inability to fully leverage LPDDR5X capacity attributed to NVLink under its supplier's current memory configuration.

注：根据 SemiAnalysis 的最新研究结果，NVLink 拒绝将 DDC-DRAM 的延迟减半，这一延迟仍大于本地 DDR 模块的延迟。这一调整并不反映 NVIDIA 整体内存依赖的减少。这是对 Vera 无法充分利用 NVLink 所限定的 LPDDR5X 容量的回应。

Beyond Nvidia, the broader CPU landscape is also bifurcating. Traditional x86 vendors AMD and Intel launched new products in 2026, with mid-frame Turin for Cleaver Rapids and Granite Rapids-AP for Diamond Rapids. Meanwhile, Arm-based CPU vendors Ampere and Qualcomm are expected to enter mass production late this year or early next year, with their AI CPU roadmaps focusing on inference.

除 Nvidia 之外，更广泛的 CPU 市场也在分化。传统的 x86 厂商 AMD 和 Intel 在 2026 年推出了新产品，Cleaver Rapids 搭载中框 Turin，Diamond Rapids 搭载 Granite Rapids-AP。基于 Arm 架构的 CPU 厂商 Ampere 和 Qualcomm 预计将在今年年末或明年初进入量产，其 AI CPU 路线图聚焦推理。

In sum, we are seeing two new drivers of memory demand generated by AI inference. On one front, inference workloads are driving KV cache offloading to storage, creating new demand for solid state drives. On the other, agentic AI workflows are shifting compute demand toward CPUs and DDR RAM. Both trends point to a rebalancing of infrastructure requirements, with demand in this segment is expected to continue rising.

总结来说，AI 推理带来了内存需求的两大新驱动因素。一方面，推理工作负载正推动 KV 缓存卸载到存储，从而为固态硬盘创造新需求。另一方面，智能 AI 工作流正将计算需求从 GPU 转向 CPU 和 DDR 内存。这两大趋势共同指向基础设施需求的重新平衡，预计该细分市场的需求将持续上升。

关键词：内存需求，KV 缓存卸载，智能 AI，推理经济学，SSD 容量规划，CPU 与 GPU 比例，基础设施重新平衡