Open Storage Summit -- Day 2

主持人：

欢迎各位参加 Supermicro 开放存储峰会！本次峰会我们将探讨适用于生成式 AI（Generative AI）解决方案的存储技术。我是 Rob，担任董事总经理兼首席分析师。当前，企业级 AI 正迎来关键转折点 —— 各机构纷纷采用生成式 AI 系统，这类系统能够自主决策、执行多步骤任务，并以复杂方式实现交互。而支撑这些能力的基础设施正面临极限挑战。这并非传统 AI 工作负载的又一次演进，生成式 AI 带来的是全新类别的挑战，在存储基础设施领域尤为突出。

这类生成式 AI 系统对存储提出了极高要求：需具备海量、高速、低延迟的存储能力，同时搭配可扩展、灵活的架构 —— 既要能无缝管理各类数据类型，与计算层和网络层深度集成，又要满足严苛的性能、安全性、数据治理及成本效益要求。

生成式 AI 的核心，是围绕大型语言模型（Large Language Models, LLM） 或生成式 AI 模型 (Generative AI Models) 构建的智能框架，可称之为整个系统的 “大脑”。但在 “大脑” 之外，还需一套复杂的支撑系统才能实现自主性。以键值缓存（Key-Value Cache, KV Cache） 为例，它相当于 AI 的 “短期记忆”，负责存储即时结果和记忆令牌（Memory Tokens）。借助键值缓存，AI 能够保留上下文信息、提升响应速度，在长时间对话或复杂工作流中保持逻辑连贯，无需每次都从头重新处理所有数据。

此外，还有控制流层（Control Flow Layer） —— 它负责协调智能体（Agent）的规划、

信息检索与任务执行流程。在众多企业场景中，这意味着要运用检索增强生成（Retrieval Augmented Generation, RAG） 技术，并集成向量数据库（Vector Database）、知识图谱（Knowledge Graph）等工具，让 AI 的推理过程更具语义性、因果性，且能适应新的输入数据。

生成式 AI 的另一重要功能是工具调用（Tool Invocation）——AI 智能体可调用外部 API 或内外部服务，实时完成任务。这类系统还依赖反馈循环（Feedback Loops） ，能根据实际业务结果和目标持续优化决策能力。

若将上述所有能力与模型托管（Model Hosting）、非结构化数据向量化服务（Vectorization Services for Unstructured Data）、可复用模块管理目录（Aging Catalogs for Managing Reusable Modules）等功能相结合，你将看到真正可扩展的多智能体生态系统（Multi-Agent Ecosystem） 的基础架构 —— 这类系统不仅能推理、行动、自适应，更重要的是，还能解释自身的决策逻辑（这一点至关重要）。

本次会议汇聚了行业专家，将深入解析如何构建具备高韧性的存储基础设施，以规模化支撑这类生成式 AI 系统 —— 从技术考量到实操建议，我们都会全面覆盖。现在，让我们开始深入探讨。首先，有请 AMD 的 Kevin，感谢您的参与；还有 SanDisk 的 Praveen，感谢您的到来；以及 DDN 的 Blanche，还有 Supermicro 的 Vince，非常感谢各位的参与。

我认为，Supermicro 开放存储峰会系列此次聚焦生成式 AI，无疑是今年最热门的话题之

一。首先，我想请各位先 “退一步”，从各自企业的视角出发，谈谈对生成式 AI 的看法，尤其是生成式 AI 场景下的存储需求。Kevin，我们先从您开始吧。

AMD 代表：

大家好，我是 Kevin，任职于 AMD，负责嵌入式存储部门的管理工作。非常荣幸能与各位交流，分享 AMD 对当前 AI 如何塑造数据、影响存储技术，以及存储需求发生哪些变化的观点，同时也会介绍 AMD 如何构建整体基础设施，以支撑这个全新的 AI 时代。

首先要问的是：如今存储对 AI 为何如此重要？因为我们正迈向生成式 AI——AI 不再只是提供答案，而是能自主学习、自适应，并在学习后采取行动。这就要求系统必须快速访问数据、高效利用数据，并实时更新数据。要理解这一点，我们可以看看 AI 数据路径中的三个关键阶段：

数据湖（Data Lake）阶段：将数据加载到训练系统中，例如文档、图片、视频等序列数据。此阶段要求存储具备高速加载能力和高带宽。

数据准备阶段：为模型训练预处理数据，包括数据清洗、标注、缓存和转换。这一阶段的操作大多在内存中进行，因此需要高速 IO、出色的内存带宽，同时依赖 CPU 进行多任务协调。若此阶段速度缓慢，整个训练过程的成本会显著增加，且会出现延迟。

推理（Inference）阶段：这正是生成式 AI 的核心应用场景 ——AI 持续接收新输入并快速决策。例如我们现在使用的聊天机器人，需要实时提供答案、即时响应，不能让用户等待；

再比如车辆的环境目标检测，同样需要无延迟的响应。这就要求存储具备极低的延迟和庞大的内存吞吐量，且数据需能直接访问 GPU 内存，无需等待 CPU 响应。

那么 AMD 能提供哪些支持呢？我们从 CPU 开始：AMD EPYC（霄龙）CPU 支持大量 PCIe 通道、高核心数和强大的内存控制能力，非常适合数据预处理、存储 IO 管理，以及支持 NVMe 或高速网络。同时，我们还为存储解决方案提供了高耐用性和可靠性保障。

其次是我们的 GPU 产品系列：最新款 GPU 同时适用于模型训练和推理，支持从存储到 GPU 内存的直接数据路径（Direct Data Path） —— 这意味着数据传输速度更快、延迟更低，且能减少对 CPU 资源的占用。

除此之外，我们还有智能网卡（Smart NIC）和数据处理单元（DPU）：这些专用芯片能在数据传输过程中保障安全性，同时卸载部分存储任务，让大型系统（尤其是 AI 集群）运行更流畅。

此外，AMD 的边缘计算（Edge）和客户端（Client）产品组合也能为 AI 基础设施提供支持。但 AMD 的优势不仅在于硬件性能，我们还拥有 ROCm 软件栈—— 它如同 “粘合剂”，将所有硬件组件整合起来，让开发者能轻松利用 AMD 的全栈产品组合，构建高效的 AI 基础设施。

现在，让我们简要看一下第一页幻灯片。AMD 在 AI 存储的所有阶段都占据着有利地位，但我们并非孤军奋战。今天，我们与合作伙伴携手，从边缘到云端、从研发到实际部署，共

同打造经过测试、优化且可扩展的解决方案。

总而言之，生成式 AI 不仅关乎智能模型，更离不开支撑这些模型的高速、智能的基础设施。无论你是在训练大型模型，还是运行实时推理任务，AMD 及其合作伙伴都能帮助你实现更快的数据传输和更快速的决策。非常感谢大家。

主持人：

接下来，我们不妨深入到技术栈的更底层。Praveen，能否请您分享一下 SanDisk 在生成式 AI 领域的方案或愿景？

SanDisk 代表：

谢谢 Rob，也感谢 Kevin 刚才的精彩介绍。Rob，感谢您为本次讨论搭建了清晰的框架。我认为，要理解生成式 AI 的重要性及其未来发展方向，首先需要明确其核心内涵 —— 我准备了三张幻灯片，但会重点讲解第一张，因为它能帮助我们建立关键上下文。

什么是生成式 AI（Generative AI）？它之所以被称为 “生成式”，核心在于具备智能体属性（Agency）：生成式 AI 系统能够自主行动、自主决策，在极少人工干预的情况下完成任务。它能 “思考”、能推理、能独立行动、能规划。若深入剖析其运作机制，会发现它

本质上是将特定用例作为推理工作负载（Inference Workload） —— 基于过往的训练数据集做出决策。

需要注意的是，推理技术近年来已发生显著演进：从最初的 “感知（Perception）”（主要用于图像标注和分析），到 “生成式 AI”（仅能提供单次响应），再到如今的 “复杂推理（Complex Reasoning）”（即生成式 AI 的核心能力），未来还将迈向 “物理 AI（Physical AI）” 和机器人技术。我们能清晰看到，推理的复杂度正呈指数级增长 —— 从 “单次响应” 到 “多步骤响应”，复杂度甚至提升了 100 倍。

行业内的共识是：2025 年将是生成式 AI 推理的关键元年。因为此前我们严重低估了生成式 AI 推理所需的计算和存储资源规模。

若进一步拆解推理过程，会发现它主要分为两个阶段：预填充（Pre-fill）和解码（Decode）。预填充阶段的核心是接收输入令牌（Input Tokens），在名为 “KV 缓存（KV Cache）” 的中间缓冲区中构建上下文；解码阶段则负责输出令牌（Output Tokens），并将这些令牌重新写入 KV 缓存，以持续保留整个交互过程中的上下文 —— 本质上，这是一个自回归（Auto-regressive）过程。

为何这一点至关重要？因为随着用例复杂度提升（如多令牌、多步骤、多用户、多租户、多模型场景），KV 缓存的复杂度也急剧增加 —— 无论是超大规模企业，还是企业级 RAG 场景（需对数据进行向量化处理），我们都能看到数据集的 “爆炸式增长”，而这些数据集越来越难以完全存储在 HBM（高带宽内存）或系统内存中。

同时我们也意识到：与其在用户返回或任务重启时重新构建 KV 缓存，不如将 KV 缓存、数据湖或向量湖（Vector Lake）在存储中进行 “换入换出（Swap In/Out）” —— 这种方式的速度要快得多。

总而言之，存储是生成式 AI 的关键赋能者：借助高效存储，我们能缩短 “首次令牌输出时间（Time to First Token）”、降低延迟，同时避免因重复构建 KV 缓存导致的 GPU 资源浪费。

以上是整体概述。另一个关键趋势是：随着越来越多企业转向生成式 AI 推理工作负载，他们正将大量核心数据迁移到更大规模的数据湖中 —— 这直接导致数据中心的 “数据热度（Data Temperature）” 上升：传统基于 HDD 的数据湖和归档存储，正逐步被基于闪存（Flash）的数据湖取代，这也是我们第一张幻灯片想要强调的核心观点。

第二张幻灯片则想传递一个重要信息：我们认为生成式 AI 场景下的 SSD 需求主要分为两类：左侧是计算密集型（或性能密集型）SSD，右侧是容量密集型 SSD—— 两者同样重要，而 SanDisk 的产品组合能同时覆盖这两类需求。

计算密集型 SSD 的核心需求是高随机读取性能、高 IOPS（每秒输入输出操作） 以及高顺序写入性能；容量密集型 SSD 则主要面向顺序工作负载—— 例如 KV 缓存、向量湖的 “换入换出”，或训练过程中的快照（Snapshot）操作。

SanDisk 在这两个领域均拥有强大的产品组合：计算密集型 SSD 紧邻 GPU 部署，而数据湖则通过网络与存储端相连。后续 Blanche 应该会详细介绍存储端的相关内容。

主持人：

好的，接下来我们请 Blanche 加入讨论。正如我们之前所说，存储是所有数据的核心，而数据又是 AI 的核心 —— 没有数据，AI 就无从谈起。即便在 10 年前传统机器学习（ML）时代，这一点也同样成立。如今 AI 已演进到生成式阶段，与以往相比有了显著差异。DDN 在该领域深耕多年，能否分享一下 DDN 的定位，以及您对生成式 AI 的看法？

DDN 代表：

非常感谢。Rob，很高兴能参与此次讨论。正如您所说，从某种程度上看，生成式 AI 的用例与 DDN 近 30 年前成立时的定位（高性能计算 HPC 领域）高度契合。

多年来，DDN 持续提升存储软件的能力，使其更可靠、更稳健、更具扩展性。当 AI 浪潮来袭时，我们先聚焦于模型训练，随后转向推理，如今又拓展至生成式 AI 领域。我们发现，传统存储软件正面临新的要求：需进一步提升性能，以充分发挥 GPU 的巨大投入价值，最大化 GPU 的利用率。

最初是训练场景，如今在生成式 AI 时代，情况发生了有趣的变化 —— 正如 Praveen 和 Kevin 此前提到的，“数据热度” 正在上升。我更倾向于用 “冰山” 来比喻这一变化：过去，冰山顶端是 “热数据”，其余均为 “冷数据”；而在生成式 AI 时代，整座冰山（即所有数据）都可能被随时调用 —— 无论是 RAG 推理，还是其他场景，都可能在任意时刻提取数据。这意味着我们需要能主动管理所有数据集。

正是得益于 AI 革命早期与合作伙伴的协作，DDN 有机会将存储软件升级至 ExaScale 级别，近期又推出了 Infinity 平台。该平台充分利用了 “大规模异构数据集” 的特点 —— 涵盖不同协议、不同介质类型、不同数据大小，并通过单一控制台将它们整合起来，确保能以应用所需的速度为其提供数据，从而保证 GPU 持续高效运行。

这一过程中，软件的任务变得更加复杂：一方面，需管理硬件基础设施 —— 确保 CPU 和 GPU（其性能正日益强大）持续处于忙碌状态，同时高效利用容量不断增大、支持 PCIe 5.0 高吞吐量的 SSD；另一方面，需为 GPU 工作负载和生成式 AI 应用提供支撑 —— 例如，当你发起一个查询时（如 Rob 您提到的 RAG 场景），需以高性能、低延迟的方式完成响应。而当规模扩大后，挑战会进一步加剧：数据类型多样（图像、视频等）、协议异构（S3、文件系统等）、部署位置灵活（云端或本地）—— 如何驾驭这种复杂性和多样性，并实时满足需求，正是 DDN 面临的核心挑战。

若观察一个典型的生成式 AI 或 RAG 流程，就能发现存储软件的价值所在：它能加速 GPU 运行、确保 GPU 高利用率，从而帮助用户实现最佳投资回报率（ROI）并降低延迟。

目前，DDN 拥有市场上性能最高、延迟最低的对象存储（Object Store） —— 其延迟比 AWS 低约 100 倍。这带来的直接好处是：执行列表（List）命令时，你能看到比高延迟存储解决方案多得多的内容。

这是高性能存储平台价值的一个具体体现。此外，DDN 还能无缝处理多协议场景、加速网络、支持 SQL 查询以减少数据往返传输 —— 所有这些都旨在最大化基础设施利用率。我们经过长期思考和研发，已构建起稳健的平台，且已得到客户的广泛认可。例如，某客户采用 DDN 解决方案部署了包含 10 多万个 GPU 的大型集群，充分验证了该方案的规模化能力。当然，这一切都离不开今天在座合作伙伴的支持 —— 生态系统的协作至关重要。

在研发和优化能力的过程中，我们曾在内部运行过一个 RAG 流程，这里想分享一个案例：右侧图表显示，我们首先在 AWS 云端运行一个标准 RAG 应用；随后，我们思考如何进一步加速 —— 下一步自然是引入 GPU，这也是图表中间灰色柱状图所示的基准线。仅将基础设施从 CPU 迁移到 GPU（未做任何调优），就能看到显著的性能提升；而当我们用 DDN Infinity 对象存储替换 AWS 对象存储（未做其他任何更改）时，延迟大幅降低，最终实现了 22 倍的应用加速。这一案例充分证明：设计并研发符合生成式 AI 等高性能应用需求的存储管理和数据智能软件，具有多么强大的价值。

主持人：

没错，这一点确实至关重要。在生成式 AI、RAG 和推理场景中，大量涉及非结构化数据 —

— 关键在于如何让 GPU 快速获取这些数据。毕竟，企业采购 GPU 的成本很高，它们就像工厂一样，必须保持持续运转，而这离不开高速存储的支撑。Vince，从 Supermicro 的产品组合来看，您如何看待生成式 AI 存储的发展方向？

超微代表：

谢谢 Rob。正如大家所讨论的，AI 行业正蓬勃发展，深刻改变着每个人的生活。其中，大型语言模型（LLM）的发展是一个里程碑，它改变了人们的生活和工作方式；而生成式 AI 的核心能力在于 “目标驱动”—— 能够自主决策。可以预见，未来将涌现出更多令人兴奋的新应用。

正如其他嘉宾所提到的，生成式 AI 的复杂性极高：其部署场景多样（如超大规模 AI 即服务、企业非规划设施、边缘端等），对性能、规模和数据传输能力的需求也在不断提升。作为基础设施提供商，Supermicro 的价值正体现在这里。

过去几年，大家可能听过 Supermicro 的 “数据中心构建块解决方案（Data Center Building Block Solutions）”—— 我们与技术合作伙伴紧密协作，并深度参与客户支持。我们不仅提供出色的系统构建块，还会根据客户需求持续投入，将能力拓展至数据中心级别。

要运行超大规模 AI GPU 数据中心，以下是关键构建块：计算、存储、网络架构，以及关键的后备电源设施和冷却能力。对于想要部署这类数据中心的企业而言，这些都可能是挑战。

而 Supermicro 能为客户带来的价值是：在简化部署流程、缩短上线时间的同时，帮助客户降低超过 20% 的总体成本。

今天在座的 AMD、SanDisk、DDN 都是我们的合作伙伴，我们共同打造了经过验证的参考架构。以往，人们可能认为推理工作负载比训练更 “轻量化”，但生成式 AI 的应用和能力意味着，推理场景也可能需要超大规模、大容量的架构 —— 这正是我们合作展示的重点。

以 AMD 为例，他们近期推出了 MI 350 Instinct GPU 架构，Supermicro 与 AMD 紧密合作，率先将该架构推向市场。考虑到 “数据直达 GPU” 的需求，我们还拥有丰富的网络交换机产品组合，包括支持 800G 带宽的交换机 —— 其在编排、架构管理和性能优化的流量路由方面具备先进能力。

最重要的是存储领域：我们整合了合作伙伴的优势技术，提供预验证、全认证、即发即用的解决方案。例如，通过与 DDN 合作，我们能为客户部署超大规模、具备土壤抗性（Soil Resistance)的 AI 存储系统;具体到本次展示的方案,是在基于 AMD 的 Supermicro 服务器上运行 DDN Infinity 数据智能平台,按机架规模设计,单个构建块即可提供 60PB 容量。

面对数据增长和性能需求，这套集成解决方案不仅能提供强大性能，还兼具易用性和高成本效益、低运维投入的优势。同时，它采用开放架构，几乎具备无限扩展能力，可支持不断增长的计算需求。无论客户拥有何种基础设施、何种需求，Supermicro 都将与合作伙伴携手，

设计、构建并提供最佳解决方案。

主持人：

我认为这正好呼应了我们开篇时的讨论 —— 即 “你需要从某个起点出发，并能随时间逐步扩展解决方案”，而不是被单一方案限制。因为生成式 AI 系统的负载往往是 “不均衡的”：可能这里有一个 RAG 任务，而推理过程又可能调用某个 API 获取数据，进而触发另一个推理任务并返回结果 —— 所有这些环节都需要协同工作。

从技术栈角度，我们来进入问答环节吧。我脑子里有几个问题想请教各位，比如 Kevin，您之前提到 AMD 正大力推进 “全栈 AI 解决方案 / 产品组合” —— 能否帮我们理解，“全栈（Full Stack）” 对 AMD 而言具体意味着什么？

AMD 代表：

关于 “全栈”，结合之前的幻灯片来看，它不仅仅包含 CPU 和 GPU，还涵盖边缘端和客户端的硬件 —— 这些是支撑基础设施的 “构建块”，这一点大家都在讨论。但更重要的是软件栈（Software Stack） —— 它就像 “粘合剂”，将所有组件整合在一起，让工程师和开发者能协同发力，不断优化解决方案。

另外，我认为还有一点非常重要 —— 正如本次圆桌讨论所体现的：若没有生态合作伙伴（如系统厂商、软件供应商）的协作，“全栈” 是无法实现的。因此，“全栈” 本质上是整个社区共同协作、共同投入，以推动全新 AI 时代发展的过程。

主持人：

我认为 “生态协作” 确实是关键 —— 毕竟并非所有 AI 工作负载都是相同的。Blanche，DDN 在该领域深耕多年，早于 “AI”“机器学习（ML）” 等术语流行的时代就已涉足相关领域，且贯穿整个技术栈。存储作为其中的关键环节，DDN 如何确保它不会成为客户的瓶颈？又如何致力于提供 “端到端最佳成果”（而不仅仅是存储层面的优化）？

DDN 代表：

这是个很好的问题。随着规模扩大和应用复杂度提升（例如生成式 AI 场景中，从单个租户到上千个并发租户），挑战会急剧增加 —— 而 DDN Infinity 平台的精细化多租户架构（Granular Multi-Tenancy Architecture） 正是应对这一挑战的关键。这种架构在行业内独树一帜：你可以创建租户和子租户，并为每个租户、子租户设置服务级别协议（SLA）和服务质量（QoS）。

这种架构设计的思路，源自 DDN 在 HPC 领域的经验，并已成功迁移、扩展到 AI 和生

成式 AI 场景 —— 它不仅能很好地支持高性能、低延迟这两大核心需求，还能实现线性扩展，无论是 1 个租户还是 10000 个租户，都能稳定运行。

主持人：

这一点对企业而言尤为重要。例如，某企业可能有一个用于再保险业务的 AI 智能体，同时还有用于共同基金业务的智能体 —— 出于安全考虑，这两类业务的数据必须隔离，不能交叉。因此，这种多租户能力对运行多个智能体的生成式 AI 系统来说，确实是关键需求。

DDN 代表：

完全正确，这一点在本地部署、云端部署以及混合部署场景中都至关重要。另一个关键点是减少数据移动—— 数据的进出（Egress/Ingress）会增加延迟、导致性能瓶颈。DDN Infinity 平台采用了特殊架构，能最大限度减少数据移动；更重要的是，它支持无限标签（Unlimited Tagging） —— 这在行业内是首次实现，你可以为所有数据添加标签，以便在需要时快速查找。再结合内置的 SQL 查询功能，就能进一步减少数据移动，高效完成数据的标记、查找和检索 —— 这对生成式 AI 和 RAG 场景来说非常有帮助。

主持人：

尤其是在 “向量化” 成为核心需求的场景中。Vince，回到生成式 AI 的话题 —— 它的潜力非常巨大。当企业在设计面向当下、同时兼顾未来的基础设施时，有哪些关键因素需要考虑？您有什么建议？

超微代表：

从与客户的沟通中我们发现，他们非常重视部署时间（Time to Deployment） —— 无论是推向市场还是启动服务，都希望越快越好。我们理解这一需求，并会提供相应支持。但同时，我们也需要提醒客户：在实施过程中，要为未来的升级预留空间。

例如，有些客户最初部署了 20-30 个节点的系统，但很快就发现存储空间不足 —— 这正是生成式 AI 用例的特点：数据访问方式发生了变化，数据量也在持续增长。因此，在设计解决方案时，必须考虑升级路径（Upgrade Path） 。

以 SanDisk 为例，他们的硬盘容量路线图持续提升，密度不断增加 —— 这正是预留增长空间的价值所在。从 Supermicro 的基础设施实施角度来看，我们的能力不仅限于单系统，

还能扩展到数据中心级别。例如，在电源方面，我们能帮助客户在提升性能和密度的同时，提供最佳的电源和冷却基础设施解决方案，实现高性能与超高效率的平衡。

主持人：

“效率” 确实是关键。Praveen，我们来聊聊效率 —— 有人说，AI 竞赛的胜负将取决于“功耗”：万亿参数模型在 GPU 上运行需要大量时间，功耗极高；除此之外，还有很多其他高功耗组件。存储多年来也被认为是 “功耗大户” 之一，尽管随着技术发展，其效率已有所提升。SanDisk 如何看待这一问题？在提升 AI 系统效率方面，SanDisk 有哪些举措?

SanDisk 代表：

谢谢，这是个非常好的问题。如今，功耗已成为所有人关注的焦点，尤其是数据中心架构师——GPU 的功耗非常高，甚至超过了 CPU。这意味着：如果企业的预算有限，大量预算会被 GPU 占用，留给存储的预算就会减少。其次，不同地区的电价差异较大，功耗已成为新建数据中心设计或现有数据中心运维的最大约束因素之一。事实上，有数据显示，功耗已占数据中心运营成本的 60% 以上。

因此，功耗已成为衡量存储性能的新指标—— 这也是我们为何推崇 “每瓦 IOPS（IOPS per Watt)”“每瓦吞吐量（Throughput per Watt)” 这类指标的原因。

SanDisk 的解决方案是提供多功耗状态（Multiple Power States） ：我们正与生态伙伴合作，根据工作负载的性能需求变化，实时调整功耗状态 —— 性能需求高时调高功耗，需求低时调低功耗，从而动态优化功耗 footprint。

SanDisk 不仅在性能和容量（如之前提到的路线图）方面领先，还在功耗与性能效率方面持续投入 —— 通过多功耗状态的设计，让存储的性能评估标准从单纯的 IOPS 和吞吐量，扩展到 “每瓦 IOPS” 和 “每瓦吞吐量”。

主持人：

没错，未来甚至可能出现 “每瓦令牌数（Tokens per Watt)” 这样的指标 —— 衡量模型在单位功耗下的效率，以及与 IO 等指标的关联。我们还有时间再问 1-2 个共性问题，我想多听听各位的看法。

回顾过去，我们曾构建网格（Grids）等技术，那在当时是最先进的；我们还尝试过分布式网络、分布式处理等多种方式。但如今在 AI 时代，我们看到了一种 “反向趋势”：将 AI 推向数据，数据仍具有 “引力”。各位在与客户合作时，是否发现他们正重新思考存储架构？

例如，过去的架构可能适用于传统应用（这些应用可能仍是生成式 AI 系统的一部分，需通过 API 或数据库调用），但面对生成式 AI，他们需要全新的架构思路。对此，您有什么观察？又有哪些建议可以提供给正在重新规划存储架构的企业？

SanDisk 代表：

我先来说说吧。我认为 AI 正成为 “数据中心架构重构的推动力”。正如 Blanche 所说，减少数据移动至关重要；行业内也有一些相关倡议，例如如何将 IO 管理从 CPU 代码转移到 GPU 代码 —— 因为 GPU 拥有更多核心，能更快地发起 IO 请求。

SanDisk 正与部分行业伙伴合作，探索如何让更多数据靠近 GPU、加速 IO 速度，从而减少数据往返。这仍是一个持续的挑战，需要生态系统共同解决。但可以肯定的是，数据移动、功耗、高性能将持续是数据中心重构过程中的核心关注点。

主持人：

Blanche，您怎么看？

DDN 代表：

Praveen 的观点非常好，尤其是关于数据移动和功耗的部分 —— 功耗已经（或即将）成为数据中心规模的决定性约束因素。回到您最初的问题，Rob：企业对基础设施（这里主要指存储基础设施）的思考方式确实在改变。

过去，存储硬件系统和存储软件往往是绑定的（尤其是本地部署场景）；而现在，我们看到了解耦（Disaggregation） 的趋势 —— 企业不再希望为对象存储、文件存储、块存储分别采购不同的解决方案（这是传统架构的局限），而是需要一个 “一站式存储解决方案”：它能运行在最先进的硬件平台上（搭配最佳服务器、CPU、SSD），并由顶级软件平台统一管理，既能接收 GPU 产生的所有数据，又能精细化管理底层基础设施，最终实现 “顶层最大化 GPU 利用率、底层最大化基础设施效率” 的平衡。

主持人：

非常有道理。既然提到了 GPU，Kevin，您认为企业在重构或优化存储架构时，有什么关键建议可以分享？

AMD 代表：

我认为我们已经讨论了大部分关键方面，但我想强调可扩展性（Scalability） —— 因为数据量正以极快的速度增长。这一点与 “软件定义（Software-Defined）” 的思路密切相关：如果能采用更标准化的服务器或平台，同时通过软件管理所有数据，这将是未来的趋势。越来越多的客户正在寻找这种更开放的架构，以支撑存储需求的持续增长。

主持人：

完全同意。Vince，Supermicro 常被视为企业获取这类解决方案的 “便捷选择”。对于那些正在思考 “我需要做什么才能为生成式 AI 做好准备” 的企业，您有什么建议？尤其是当他们已在构建其他 AI 工具（这些工具将成为生成式 AI 智能体的一部分）时，第一步应该怎么做？

超微代表：

我想分享一些与客户沟通的心得。过去，我们通常将数据中心视为孤立的实体；但在生成式 AI 时代，数据来源几乎是 “无处不在” 的，数据中心也可能具有地域性。例如，你可能在某个区域部署了 GPU 集群和配套的数据存储，当规模扩大后，又会新增多个集群 —— 这就引发了一个行业热议的话题：除了单集群内部的互联，多集群之间的互联（Inter-Rack/Inter-Cluster Connectivity） 该如何实现?

因此，我的建议是：在规划存储架构时，不要局限于 “专用存储”，而应拓宽视野，考虑多集群部署场景，以及跨站点数据管理。基于这一思路，你会设计出完全不同的架构 —— 不仅要考虑数据访问，还要考虑数据流式传输（Data Streaming）的需求。未来，数据规模可能会超越单个数据中心或单个集群的边界。

主持人：

非常有道理。我认为我们可以在此处收尾 —— 各位的建议为企业提供了清晰的起点参考。感谢各位的参与，这是一场非常精彩的讨论。感谢各位观看本次 Supermicro 开放存储峰会 —— 我们深入探讨了适用于生成式 AI 解决方案的存储技术。敬请关注更多后续内容。