Open Storage Summit -- Day 1

主持人：

大家好，欢迎参加 “AI 工作负载分层存储” 主题分享，本环节是超微开放存储峰会（Supermicro Open Storage Summit）的组成部分。我是罗布·斯特雷奇（Rob Stretch），担任董事总经理兼首席分析师。人工智能革命不仅在改变企业的运营方式，更在重塑企业基础设施的核心根基 —— 而其影响最深远的领域，莫过于数据存储领域。

随着人工智能模型规模不断扩大、复杂度持续提升，并深度融入企业业务流程，存储系统面临的需求正以前所未有的速度加剧：从训练大规模多模态模型，到每日处理数十亿次低延迟推理请求，AI 工作负载对存储的需求已远超 “容量” 本身。它们需要混合环境下的高性能、高效率、高灵活性，以及智能调度能力。这一趋势正促使存储的设计、部署与使用方式发生根本性变革。

新一代存储解决方案正为满足 AI 需求量身打造 —— 它们需支持高吞吐量数据管道、GPU 加速计算集群，以及以模型为核心的应用架构的动态演进需求。如今，企业已不再以 “PB（拍字节）” 为单位规划存储，而是着眼于 “EB（艾字节）”；且越来越多企业选择结合本地部署、边缘计算与云环境的混合架构。

推动这一需求激增的因素有二：一方面，生成式 AI 的爆发产生了海量数据（文本、图像、代码、视频等），且需要实时访问大规模训练数据集；另一方面，行业整体趋势也在推动变革 —— 随着预训练模型投入生产，“推理” 环节的重要性日益超过 “训练”。如今，

企业优化基础设施时，不仅要考虑实验需求，更要兼顾大规模生产环境下的稳定性、响应速度。推理正迅速成为主流 AI 工作负载，并推动存储领域的重心转移。

与此同时，各厂商正竞相推出能跟上这一趋势的存储架构：无论是为推理引擎提供超低延迟的 NVMe 网络（NVMe Fabrics）、针对 AI 工作流程优化的超融合平台，还是能预判并适应动态使用模式的智能数据放置技术，创新节奏都在不断加快。

与此同时，可持续发展的迫切需求也催生了新方案 —— 在能源效率、工作负载分配、环境影响缓解等方面，行业正探索全新路径。这些挑战并非 “渐进式改进”，而是 “变革性突破”，并催生了一系列新的市场动态：无论是传统基础设施领军企业，还是云原生创新企业，都在竞相探索 “AI 就绪型存储” 的未来形态。从超大规模服务商、半导体巨头，到企业级 IT 核心厂商，所有人都在重新思考自己在新生态中的角色。

在今天的对话中，我们将深入探讨这一融合趋势的核心：AI 如何重塑存储的未来？企业在扩大 AI 业务规模时需考虑哪些因素？最具前瞻性的厂商正如何为未来布局？我们将围绕推理、数据本地化、可持续发展，以及新一代智能架构展开讨论。因为在 AI 时代，存储不再是后端辅助功能，而是推动创新、提升性能与构建竞争优势的战略核心。现在，我非常荣幸地邀请我们的嘉宾 panel 登场。

接下来，我们将深入探讨一系列精彩话题。首先有请各位嘉宾：来自英伟达（NVIDIA）的约翰（John）、来自 WEKA 的凯文（Kevin）、来自 Scality 公司的乔治奥（Georgia）、来自京瓷（Kyocera）的安德斯（Anders），以及来自超微（Supermicro）的艾伦（Alan）。

欢迎各位的到来！

首先，我希望听听各位嘉宾的看法：从各自企业的视角出发，如何看待分层存储（tiered storage）？以及这种存储方式正如何改变 AI 基础设施的实际形态？约翰，我们先从你开始吧。

英伟达代表：

好的，罗布。我先从 AI 领域中训练与推理环节的存储分层说起。首先是 AI 训练 —— 我们如今看到的 AI 模型（比如能处理查询的 ChatGPT 等模型），正是通过训练环节构建的。

通常，训练这些模型时，首先会用到数据湖（data lake），其中存储了大量数据，载体通常是硬盘（可能是硬盘与闪存的混合，但以硬盘为主），存储类型可以是文件存储或对象存储。这类存储的核心特点是 “容量大”，性能需达标，但无需顶尖水平。

当为训练准备数据时，会将数据迁移到高速文件存储中 —— 存储类型通常是文件存储（也可能是对象存储），但载体必须是闪存，因为需要提供极高的读取性能：训练用的 GPU 集群会反复从这里读取数据。

而在训练过程中，还需要 “检查点（checkpoints）” 功能：检查点用于保存训练进度，以防意外情况发生（比如故障）—— 此时无需重新启动整个训练任务，只需从最近的检查点恢复即可。要知道，有些训练任务可能持续数月，因此检查点至关重要。这类存储需要极高的读写性能（以写入性能为主），但偶尔故障恢复时也需读取数据；传统上使用文件存储，

如今部分客户也开始采用对象存储，但载体均为闪存。

一个重要变化是：随着模型规模扩大和 GPU 集群升级，检查点的体积越来越大，生成频率也越来越高 —— 集群规模越大，检查点生成越频繁，恢复需求也越频繁。以上是训练环节的存储特点：硬盘与闪存结合，训练数据需侧重读取性能，而检查点则需兼顾读写（以写为主）。

再来看推理环节，这一领域的需求正在不断演进。传统观点认为，AI 推理不需要存储 —— 模型完全加载到内存中，用户提一个问题、得到一个答案，全程依赖内存，与存储无关。但这种思路已经过时了 —— 这里的 “过时”，指的是 18 个月前的认知。

如今，推理环节对存储的需求主要由两方面驱动：

第一是 RAG（检索增强生成，Retrieval-Augmented Generation）。简单来说，用户的查询会触发语义搜索，实时检索相关文档（甚至可能是几小时前刚生成的文档）；这些文档会作为查询的一部分输入大语言模型（LLM），最终也可能成为回答的一部分。通过 RAG，大语言模型的回答会更贴合需求、更实时、更少出现 “幻觉”（hallucination，指模型生成虚假信息），准确性也会显著提升。

第二是 KV 缓存（KV Cache，Key-Value Cache）。当查询长度增加时（比如使用推理模型、生成式 AI 或 RAG 时），输入序列的上下文长度会越来越长，计算耗时也随之增加。因此，在很多场景下，我们会将查询和上下文文档对应的 “KV 缓存” 存储起来 —— 如

果后续有用户提出相同或相似的问题（且用到相同上下文 / 文档），就可以直接重新加载 KV 缓存，无需重新计算。这就需要高速闪存存储来保存 KV 缓存，以便快速读取复用。

综上，推理环节的存储呈现出分层结构：从 GPU 上的 HBM（高带宽内存，High-Bandwidth Memory），到 CPU 内存、本地闪存、网络闪存，再到网络硬盘，每一层都存储推理环节的不同数据。KV 缓存可以从 GPU 的 HBM 一直延伸到网络闪存；而查询历史、日志文件（以及用于未来再训练或新模型训练的历史数据），则通常存储在硬盘和对象存储中。

以上就是我们观察到的 AI 训练与推理环节的存储分层情况。

主持人：

非常精彩。接下来，凯文，我们想听听 WEKA 公司的视角 —— 你们的切入点与英伟达有所不同，所以想了解你们如何看待这一趋势的发展方向。

WEKA 代表：

好的，没问题。约翰提到的一个核心点很关键：AI 的演进速度极快。有人说 “AI 领域的一周，堪比其他行业的一年”，这种说法毫不夸张。我们看到的趋势是：AI 正从 “大语言模型训练” 向 “推理” 演进，再到约翰提到的 “智能体（Agents）” 阶段。这一转变

意味着，我们过去为 “训练” 构建的基础设施，已无法完全满足新技术的需求。

核心问题在于：AI 演进如此之快，如何确保技术架构具备 “未来适应性”？如何在选择技术栈时，既满足当下的工作负载需求，又能灵活适配未来的技术创新？这就对 “数据处理” 提出了全新要求 —— 数据是 AI 的核心，无论计算、网络还是存储，最终都围绕 “数据获取与使用” 展开。

当前面临的关键挑战包括：

实时响应能力：AI 对延迟的要求越来越高，需确保数据访问的实时性；

数据量爆炸式增长：存储规模需从 GB（千兆字节）、TB（太字节）无缝扩展到 EB（艾字节）级别；

工作负载不可预测性：18 个月前的认知与现在已天差地别，未来需求更难预判，因此前期的 “数据架构决策” 至关重要，必须具备可演进性。

基于这些挑战，我们开发并迭代了自家产品 —— WEKA 神经网格（Neural Mesh by WEKA）。这款产品的设计初衷，就是适配现代 AI 工作负载与高性能计算负载的需求，同时确保能随问题规模扩大和 AI 创新节奏同步扩展。

其核心优势在于 “软件定义（software-defined）” —— 可以将其理解为 “软件定义的存储 fabric”：通过该技术连接 AI 计算、网络与实际数据，实现三者的智能管理与动态平衡，从而随业务需求灵活扩展。

该产品的关键特性是 “面向服务的微架构能力”：不仅具备高可用性和容错性，更重要的是能实现 “从本地到云端再到边缘” 的无缝扩展 —— 因为 AI 的未来发展方向，就是 “数据在哪里、业务需求在哪里，AI 就在哪里”，因此存储必须具备高度可移植性。

具体来看，传统部署模式包含计算基础设施（支持多场景、多工作负载及多种接口协议），而我们的 “神经网格” 软件运行在通用硬件上（包括 x86 架构和 ARM 架构，目前已有在 Grace Hopper 架构基础设施上运行的案例）。核心在于 “软件定义”：可在不同架构、本地环境和云端灵活部署，同时从软件层面覆盖整个存储分层 —— 上至内存和 GPU 级存储，下至 EB 级硬盘与对象存储，实现全分层的统一管理。

最初，我们的产品主要针对 “模型训练加速”（这是 AI 发展前期的核心需求），通过多种优化手段提升训练效率；而在同一代码库和设计框架下，我们也能快速迭代，适配推理环节的新兴需求 (比如提升 “令牌生成速度”（time to token）和 “预填充速度”（pre-filling speed））。我们认为，这种 “覆盖 AI 全流程” 的能力，将是未来存储产品的核心竞争力。

主持人:

我完全认同。正如你所说，存储需要支持文件服务、管理海量数据，且必须延伸到边缘端 —— 因为推理环节不会只局限在数据中心，而是会走向分布式部署。我知道英伟达在这方面也有相关技术，我们之后可以再深入探讨。

乔治奥，接下来想请你聊聊对象存储（object storage）在这一体系中的角色 —— 它显然是分层存储的重要组成部分，想听听你的见解。

Scality 代表：

谢谢罗布。我的观点会基于约翰和凯文的分享展开，整体方向是一致的。Scality 是一家软件定义的对象存储公司，专注于大规模部署场景 —— 支持从 TB 到 EB 级存储，目前已服务多家 EB 级客户，涵盖金融服务、政府、服务提供商等领域，而 AI 正是这些客户的核心优先级需求。我们的业务范围包括公有云、私有云，以及备份与容灾备份（immutable backup），AI 数据湖是我们的核心业务方向之一。

正如大家所见，AI 正在快速演进，而如今的对象存储已远不止 “存储数据” 这一单一功能。从宏观层面看，存储分层可分为两类：一类是约翰提到的 “全闪存高性能层”，另一类是 “容量层”（闪存与硬盘的混合）。在 AI 数据流程中，首先是数据收集、过滤与清洗环节 —— 这一步我们主要使用对象存储的 S3 协议（注：亚马逊 S3 兼容协议，对象存储常用协议），对性能要求 “达标即可”，无需顶尖水平，但需满足三大核心需求：大容量、多数据中心支持、高安全性（毕竟这是企业的核心数据）。

而在高性能层面，我们需要 “极致性能 + 低延迟”（延迟低于 1 毫秒）—— 这对应我们的 Ring XP 产品（注：Scality 旗下高性能对象存储产品），主要用于训练、微调与推理环节。若进一步细化存储分层，从下至上依次是：

硬盘层：提供基础容量；

闪存层：容量低于硬盘层，但可扩展至数百 TB；

冷存储层：包括磁带（tape）和云存储（很多人会忽略磁带，但它仍是分层存储的重要组成部分）；

超高性能层：支持 GPU 直连（GPU Direct）—— 这是我们与英伟达合作的重点方向之一。

以上所有层级都可通过对象存储实现。我们与 WEKA 公司也有深度合作：WEKA 的产品作为 “T0/T1 层”（注：近计算端的高性能存储层），负责 GPU 直连和文件接口，再接入我们不同安全级别的后端存储层。

以超微（Supermicro）、WEKA 与 Scality 的联合客户为例 —— 这是一家基于 AI 的遗传学公司，他们 90% 的容量存储使用 WEKA + Scality 的组合：WEKA 提供高性能支持，Scality 提供容量扩展；若需要更多容量，会进一步扩展到对象存储，同时支持向云端和边缘端的分层扩展。由此可见，对象存储已覆盖 “高性能 - 容量 - 归档” 的全分层需求。

主持人：

我认为这一点至关重要 —— 正如凯文之前提到的，“提前搭建适配未来的架构” 是核心。我们希望观众能从本次分享中收获的，正是 “明确未来方向，以及如何从分层角度规划下一代存储架构”。安德斯，接下来想请你聊聊京瓷（Kyocera）在这一领域的定位，以及你

们如何推动相关创新。

京瓷代表：

谢谢罗布。我是安德斯·格雷厄姆，担任京瓷（Kyocera）市场营销与业务发展高级总监。京瓷是全球领先的闪存（flash memory）与存储设备供应商，前身为东芝（Toshiba）存储业务部门，2019 年独立并更名。我们在日本拥有全球最大规模的晶圆厂（fabs），同时生产内存与 SSD（固态硬盘），其中 SSD 是我负责的核心业务 —— 受 AI 推动，SSD 领域正迎来爆发式增长，主要集中在两大方向：

第一是高性能需求。在 AI 生命周期的多个环节（尤其是训练与数据输入阶段），对吞吐量的需求持续攀升。因此，PCIe 5.0 接口的 SSD（注：第五代 PCI Express 接口固态硬盘，高吞吐量特性）已成为主流；同时，新形态因子（form factor）也在快速普及：比如 E3.s 形态 —— 专为高效闪存设备设计，仅 2.5 英寸大小，目前在高性能场景中增长迅速；此外，E1.s 形态也被广泛用于英伟达参考架构（NVIDIA Reference Architectures）中的 “近存储”（near storage）场景（即靠近 GPU 的存储）。

京瓷在 PCIe 5.0 SSD 领域的布局如下：当前有 CD8P 系列硬盘，下一代 CD9P 系列即将推出；企业级产品方面，现有 CM7 系列硬盘（最高吞吐量），下一代 CM9 系列（采用 BigSAS Flash 技术，将替代 CM7）也即将发布；E1.s 形态产品则有 XTA 系列。

第二是容量需求。正如之前讨论的，AI 数据摄入阶段需要海量数据湖，这一领域的创新同样活跃。我们即将推出的 LC9 系列产品，就采用了 BigSAS 技术（注：京瓷自研下一代闪存技术）。这里我想简单提一下底层闪存技术：3D NAND（三维闪存）已成为市场主流，但它存在一个核心问题 —— 随着层数增加，成本曲线（cost curve）也在上升，开发更高层数的 3D NAND 成本越来越高。

为此，京瓷推出了 CBA 技术（CMOS Bonded to Array，CMOS 与存储阵列直接键合），该技术有两大优势：

横向缩减：去除存储阵列中的部分外围电路，实现芯片的横向尺寸缩小，提升 NAND 芯片的效率，从而在相同层数下实现更高密度；
独立优化：将 CMOS 与存储阵列分离，可针对两者分别进行热管理优化，既提升了能效，也增强了性能（在相同功耗下，性能提升约 30%）。

具体到容量层面：我们目前已实现 2TB 的 QLC（四级单元）芯片（采用 Fixate 技术），16 颗芯片堆叠可实现单封装 4TB 容量，而一块 PCB（印刷电路板）上集成 32 个封装，即可实现 128TB 的 SSD 容量 —— 这使得大容量硬盘的生产变得简单。未来，我们还计划推出 32 颗芯片堆叠的方案，实现 256TB 甚至更高容量的 SSD。当闪存密度达到这一水平时，从硬盘（HDD）迁移到闪存（SSD）的投资回报率（ROI）将显著提升，具备更强的商业可行性。以上就是京瓷在 AI 存储领域的核心布局与创新方向。

主持人：

我认为这是一个非常清晰的视角。接下来，艾伦，想请你从超微（Supermicro）的角度，谈谈如何将这些优秀技术整合并推向市场。

超微代表：

好的，谢谢罗布。在过去 12 个月里，AI 工作负载的爆发式增长给 AI 基础设施与数据中心带来了巨大挑战：对更强算力、高性能存储、高速低延迟网络的需求显著提升，同时数据中心的规模也在不断扩大。超微一直与合作伙伴紧密合作，提供端到端优化解决方案，覆盖整个数据路径，以满足市场需求。

超微近期刚刚发布了新的数据中心构建框架 —— 数据中心积木式解决方案（DCBBS，Data Center Building Block Solution）。这是一套完全集成、均衡设计的模块化解决方案，包含数据中心构建所需的全部核心要素：

硬件层面：机架级与存储级积木模块，涵盖计算、存储、网络设备；

环境管理层面：包括热管理与液冷系统；

服务层面：设计服务与专业服务。

AI 基础设施架构的设计核心是 “性能、容量、可扩展性与成本效益的平衡”。在应用层，超微的 AI 服务器全面兼容英伟达 AI 产品（如 HGX H100 系列、NVL 系列）；同时，我们也推出了基于 Grace CPU（英伟达 Grace 处理器）的服务器（本次分享现场也有演示机）；存储系统方面，我们有基于 BlueField-3（英伟达 BlueField-3 DPU）的 AJ Buff 系列产品。此外，超微产品与英伟达网络设备（如 CX 系列网卡、InfiniBand 与以太网的 Spectrum 交换机）实现无缝兼容。

在存储层（即分层存储），正如各位之前提到的：WEKA 的 Neural Mesh 负责数据平台管理，运行在超微的 Packaged Scale（注：超微模块化存储平台）或 “闪存即服务（Flash as a Service）” 架构上；容量层则与 Scality 代表的对象存储协同，兼顾性能与成本效益。同时，我们还提供 NVIDIA-Certified Platforms（NCP，英伟达认证平台）参考架构 —— 客户可直接采购并快速部署，性能与质量均有保障。

在整个架构中，京瓷（Kyocera）的 SSD 扮演着关键角色 —— 其高容量、高性能特性为存储层提供了核心支撑（正如安德斯之前介绍的）。超微解决方案的核心目标，是为客户提供 “快速、便捷的大规模 AI 基础设施搭建方案”：客户只需从单一供应商（超微）采购少数几种模块化产品，即可快速部署。例如，我们曾为客户构建包含 6144 台 HGX 系统的 AI 集群，仅用了 122 天 —— 这就是模块化方案为客户带来的核心价值。

主持人：

谢谢艾伦。接下来，我们想更深入一层，帮助观众理解底层逻辑 —— 不同规模的企业面临不同的存储需求，但有一点是共识：存储规模将持续增长，且越来越多企业正尝试部署 RAG（检索增强生成）。约翰，我们从你开始，聊聊 RAG 对基础设施的影响 —— 英伟达在数据流转层面有深入布局，想请你分享一下这方面的见解。

英伟达代表：

好的，罗布。正如我之前简要提到的，RAG 是推动推理环节存储需求的两大核心因素之一。传统推理环节中，大语言模型（LLM）的训练数据是 “固定时间点” 的（可能是 3 个月、6 个月或 1 年前）—— 如果仅使用 LLM 本身，其回答只能依赖训练时的最新数据，无法涵盖实时信息。

而 RAG（检索增强生成）的核心价值，在于能将 “额外文档”（企业内容、网页信息、法律文件、财务报告、股市数据、PDF 文件等）的索引与嵌入（embedding）导入 LLM。正如幻灯片所示（注：此处为会议现场提及，翻译时保留场景描述），这一嵌入流程不仅用到了英伟达的多个组件，还依赖 WEKA 的文件存储、Scality 代表的对象存储，以及京瓷（Kyocera）等厂商的闪存设备。这些文档会被建立索引，支持相似性搜索（similarity search）—— 也就是说，RAG 激活了企业文档与网页内容中的价值，且这些内容可以是 “几分钟前刚生成的实时信息”。

在实际流程中，这些实时内容会与用户查询一同输入 LLM：从 RAG 数据库（含向量数据

库，负责相似性搜索）中提取最相关的内容，作为查询的一部分传入 LLM，最终生成更贴合需求、更准确的回答。但这一过程也显著增加了存储需求 —— 正如艾伦提到的，部署 RAG 的客户通常会采用英伟达参考架构（如 MGX、HGX、NVL 系列），超微也提供相应的服务器产品；文档存储通常采用文件存储或对象存储（或两者结合），而向量数据库则需要高速闪存存储（以支撑索引与检索效率）。综上，RAG 在激活内容价值的同时，也推动了 AI 推理环节的分层存储需求。

主持人：

我认为这一点很关键 —— 随着企业将 RAG、智能体（Agents）与传统 AI / 机器学习工具结合，基础设施的复杂度也在提升。凯文，接下来想请你从 WEKA 的角度，深入聊聊 RAG 流程及 AI 面临的挑战 —— 尤其是在 RAG 效率优化方面，你们有哪些观察？

WEKA 代表：

好的，罗布。正如约翰提到的，推理环节的架构与训练环节有本质区别：训练环节更关注 “GPU 利用率、高吞吐量、长批次任务”，以及 “通过检查点实现容错”，对实时性与延迟的要求相对较低；而推理环节的核心是 “低延迟” —— 包括 “首次令牌生成时间”（time to first token），以及智能体工作流程中产生的 “多轮交互（multi-turn）” 需求。多轮交互会产生大量中间数据与中间令牌，因此推理环节需要 “实时突发式数据访问”

（包括不同层级的 RAG 数据）。

我们从中总结出两个核心认知：

效率平衡：存储不仅要满足 “高吞吐量”，更要兼顾 “低延迟”，二者需协同优化；

架构复杂度：传统认知中，推理流程是 “用户输入提示→GPU 服务器运行 LLM→输出答案”，但实际流程远复杂于此 —— 正如约翰展示的 RAG 工作流，需与向量数据库、图数据库等多系统交互。这种复杂度要求从架构层面进行协同设计。

这也是为什么我们与英伟达（NVIDIA）、Scality、超微（Supermicro）等厂商开展深度合作的原因 ——AI 技术动态变化，必须通过跨企业协作、大量测试与调优，才能实现系统兼容。例如，英伟达的 AI 推理工具（如 TensorRT-LLM）与我们的存储软件需深度集成；同时，我们也更多参与开源社区 —— 因为这是一个生态共建的过程，需要全行业共同推动技术落地。

主持人：

我完全认同 “生态协同是解决方案核心” 这一观点。接下来，乔治奥，想请你从对象存储的角度进一步探讨：对象存储在 AI 领域的应用场景非常广泛，你认为它的核心优势是什么？

Scality 代表：

从开发者视角来看，这一点很清晰。开发者既聪明又 “追求效率”（我自己也是开发者，所以可以这么说）——“追求效率” 意味着他们希望自动化一切，避免手动操作，且不关心数据存储的底层细节。

正如大家提到的 “生态”——AI 领域有上千种开源工具，它们有一个共性：都需要 “存储桶（bucket）”、对象存储端点（object endpoint）、访问密钥（access key）和密钥（secret key），有了这些就能直接运行。因此，若为开发者提供 “本地对象存储”，他们会青睐这种简单性：只需自动创建一个存储桶，输入密钥，即可访问 EB 级存储，无需关心底层是闪存、硬盘还是磁带 —— 所有操作都通过统一 API（应用程序接口）完成，底层的 “数据迁移、生命周期策略、安全性管理” 都由存储厂商负责。

这种 “简单性” 本质上是 “本地部署与云端体验的统一” —— 过去开发者习惯在云端原型开发，现在本地部署也能获得相同的便捷性，这就是对象存储在 AI 领域的核心优势。

主持人：

我认同 “简单性” 是关键。安德斯，接下来想请你聊聊 SSD 的技术进步对 AI 的推动作用 —— 毕竟，若无法为英伟达 GPU 提供持续的数据输入，AI 就无法高效运行。想请你分享一下 SSD 的技术突破，以及这些突破如何支撑 AI 发展。

京瓷代表：

好的，罗布。我们结合 AI 数据处理生命周期来谈 SSD 的作用 —— 约翰之前已经梳理过这个周期，我简要补充：

首先是数据摄入阶段。如前所述，AI 需要处理超大规模数据集，而我之前提到的 “高容量 SSD”（如 128TB 产品）将在这一阶段发挥关键作用：从机架级密度来看，高容量 SSD 能降低 “每 GB 功耗”；从系统级来看，能降低 “每 IO（输入输出）功耗”。随着存储架构师开始规划如何将高容量 SSD 融入基础设施，长期来看，这一趋势将显著提升数据摄入效率。

其次是 AI 训练阶段。核心需求是 “避免 GPU 闲置”，因此需要极高的吞吐量 —— 这正是当前 PCIe 5.0 SSD 的核心价值。专为闪存设计的新形态因子（如 E3.s）能提升机架级效率，未来这一领域仍有创新空间。此外，训练环节中的 “检查点” 功能也至关重要：训练过程中，模型会分阶段处理数据，检查点用于验证模型准确性，因此需要极高的吞吐量来保障检查点读写，避免训练中断。

最后是推理阶段。正如之前讨论的，RAG（检索增强生成）对延迟要求极高 —— 而 SSD 在性能（低延迟）与容量上的双重突破，正为推理环节提供关键支撑。综上，SSD 的技术进步（高容量、高吞吐量、低延迟）贯穿 AI 数据处理全生命周期，是推动 AI 发展的核心

基础设施之一。

主持人：

我认为大家都会认同 “所有组件需协同工作，没有哪一个比另一个更重要”—— 它们是一个有机整体。艾伦，你之前在介绍中提到 “完全集成、均衡的解决方案”，想请你从超微的角度解释一下：“完全集成” 和 “均衡” 具体意味着什么？我认为这是很多客户关心的核心问题。

超微代表：

好的。当我们提到 “完全集成、均衡设计”，实际上包含多个核心要素：

首先是 “完全集成” 的含义：

架构统一性：整个架构需是 “无孤岛” 的 —— 所有组件（计算、存储、网络）必须协同工作，具备完全兼容性，不能出现 “部分组件脱节” 的情况；

端到端优化：优化范围不仅限于硬件，还包括整个架构中的软件栈 —— 例如，超微的解决方案中，不仅有自家服务器，还整合了英伟达（GPU / 网络）、WEKA（存储管理）、斯卡拉特（对象存储）以及京瓷（SSD）的技术，所有环节都需深度优化、完美协同；

可扩展性：架构需采用模块化设计，确保能随业务需求灵活扩展。

其次是 “均衡设计” 的含义：

资源优化：整个架构中，不能出现 “部分组件过度利用，而部分组件利用不足” 的情况 —— 需确保算力、存储、网络资源的配比合理；

能效均衡：电源与液冷系统需高效协同 —— 超微在液冷领域处于领先地位，我们的 DLC 2.0（Direct-to-Chip Liquid Cooling，芯片直冷）技术能实现高效热管理；

成本与性能均衡：不能为了追求性能而忽视成本 —— 需在 “高性能” 与 “经济性” 之间找到最佳平衡点。

超微的解决方案正是围绕以上要素构建的，这也是我们 “完全集成、均衡设计” 的核心内涵。

主持人：

我完全理解了。回到底层技术，我想聊聊 KV 缓存（Key-Value Cache）—— 这是一种非常智能的提升运行效率的方式。约翰，想请你从英伟达的角度，解释一下 KV 缓存的工作原理及其在 AI 中的作用。

英伟达代表：

好的，罗布。很多人可能不知道，推理环节实际上分为两个阶段：大家通常看到的是 “用户提问题→得到答案”，但背后的 GPU 运算分为 “预填充（prefill）” 和 “解码（decode）” 两步 —— 预填充阶段处理用户输入的提示（prompt），解码阶段生成回答。这两个阶段对 GPU 的需求不同，有时会将它们分配到不同 GPU 上（即 “解耦（disaggregation）”）。

更重要的是，很多时候会出现 “重复查询” —— 比如不同用户询问同一家公司的信息、同一种医疗方案、某只股票的价格、某个新闻事件或某段代码。此时，若能将 “提示对应的 KV 缓存” 存储起来，当再次出现相同或相似提示（及上下文）时，就能直接复用该 KV 缓存，无需重新计算。对于长提示而言，这种复用能节省大量时间、算力与能源 —— 这就是 KV 缓存的核心价值：存储与复用。

正如我之前提到的，RAG 推动了推理环节的存储需求；而 KV 缓存则是提升推理效率的关键 —— 企业需要大规模、高效的 AI 推理，而 KV 缓存能同时满足 “快速” 与 “高效” 的需求。提示越长、上下文包含的文档越多，复用 KV 缓存节省的成本与时间就越多。但这需要 “高速存储” —— 可以是高速文件存储或对象存储，但载体必须是闪存（正如凯文、乔治奥和安德斯之前提到的）。

此外，推理环节的核心指标与训练不同：训练更关注 “吞吐量”，而推理则更关注 “延迟” —— 用户希望尽快得到 “首次令牌”，并获得更高的 “每秒令牌数（tokens per second）”。因此，英伟达正开发相关工具，简化 KV 缓存在不同存储层级（内存、文件存储、对象存储）中的存储与复用流程 —— 这也是当前推理技术的前沿方向之一。

主持人：

我认同这一点。凯文，WEKA 与英伟达合作紧密，在 KV 缓存方面也有相关布局 —— 想请你从 WEKA 的角度，分享一下 KV 缓存的应用实践。

WEKA 代表：

好的，罗布。这正是 AI 创新与技术响应能力的体现 —— WEKA 一直专注于 “为 AI 打造高速存储”，并与英伟达保持深度合作。我们近期在产品中新增了 “WEKA 增强型内存网格（WEKA Augmented Memory Grid）” 功能，其设计灵感正是源于 KV 缓存的价值。

正如约翰所说，KV 缓存的核心是 “令牌（token）” —— 从用户输入的提示生成令牌，再到向用户输出令牌，整个过程的核心是 “如何快速实现令牌的存取”，以支撑多轮交互、推理与智能体工作流程。这需要与 AI 推理服务器和引擎进行深度集成 —— 这也是我们与英伟达合作的重点领域：当英伟达开发新的推理工具（如支持 KV 缓存的优化库）时，我们会同步迭代存储软件，实现协同优化。

“增强型内存网格” 的核心目标是 “大幅提升令牌经济性（token economics）” —— 即

如何以更低成本交付更多令牌。同时，正如艾伦提到的，我们希望基于 “已认证、预设计的技术”（如英伟达的 NCP 架构），通过复用现有基础设施来实现这一目标。

从部署方式来看，WEKA 的神经网格（Neural Mesh）有两种模式：

独立部署：在超微服务器（包括 Grace Hopper 架构服务器）上构建独立集群，搭配京瓷的 SSD 等设备；

融合部署：将存储软件直接部署在搭载英伟达 GPU 的 AI 计算服务器中 —— 得益于 “软件定义” 架构，两种模式均可灵活实现。

实践中，若采用独立部署，我们能优化令牌交付效率；若基于英伟达 NCP 架构，还能将同一架构复用为 “推理优化架构” —— 即我们本周初刚发布的 “神经网格 Axon（Neural Mesh Axon）” 融合方案。

这种方案能带来多重价值：

延迟优化：在多轮交互或智能体工作流程中，显著降低 “首次令牌生成时间”；

架构简化：为运维团队提供更简洁的推理架构；

规模扩展：将 KV 缓存从 GPU 的 HBM（GB 级）扩展到服务器内存（TB 级），再通过英伟达 BlueField 技术实现 “低延迟网络扩展”，最终达到 PB 级 KV 缓存 —— 不仅能提升单服务器效率，还能覆盖整个数据中心的 GPU 集群；

吞吐量提升：最终实现整个数据中心级的 “令牌吞吐量” 提升。

我们认为，这种技术协同若没有 “开放合作” 的理念是无法实现的 —— 正是超微、英伟达、京瓷、斯卡拉特等企业的协作，才让这些创新成为可能。

主持人：

这一点被多次提及，足见其重要性。乔治奥，接下来想请你展望未来：从对象存储的角度，你认为它在 AI 领域的发展方向是什么？

Scality 代表：

这是我非常关注的话题。我认为对象存储在 AI 领域的未来发展主要有两个方向：

第一个方向是 “性能提升”。正如我们之前讨论的存储分层，对象存储正不断向 “更高性能层” 渗透。举个例子：我有一辆跑车，朋友想让我改装它以提升速度，建议我拆掉所有座椅 —— 我当然不会这么做，但这说明 “极致性能需要精简冗余”。亚马逊的 S3 Express（注：亚马逊高性能对象存储服务）正是采用这一思路：精简 S3 API 的功能，移除训练、元数据（metadata）、安全等非核心模块，从而提升性能。但我认为这种精简有些过度 —— 有些功能其实是必要的。因此，未来会出现 “中间形态的存储桶”：其数据布局专为 GPU 直接访问优化，同时通过 “NVMe over Fabrics” 技术实现 “GPU 直连闪存”（跳过

软件栈，降低延迟）。这意味着，闪存上的数据布局需针对 GPU 访问优化 —— 我们正在 Ring XP 产品中开发相关功能，核心是 “优化数据布局 + 移除非必要功能”，以在 “性能” 与 “功能完整性” 之间找到平衡。

第二个方向是 “功能融合”。很多企业为了不落后于 AI 浪潮，快速推进 AI 部署，但忽略了 “安全性” 等关键需求 —— 而对象存储的高安全性正是其核心优势之一，未来会在 AI 领域得到更广泛应用。此外，向量数据库（vector database）也是一个重要方向：当前的向量数据库大多仅支持基础 S3 API，安全性较弱；未来，向量数据库功能将与对象存储深度融合，成为对象存储 API 的原生功能 —— 这样既能保留对象存储的高安全性（如访问令牌、统一命名空间），又能实现向量数据库的功能，为 AI 提供更安全、更集成的存储方案。

主持人：

我完全认同。

我们正处于 AI 发展的初期阶段 —— 推理（inference）是今年的核心话题，RAG（检索增强生成）快速普及，智能体（Agents）则是今年的热点，预计明年会更加成熟，就像推理在今年的爆发式增长一样。而这些技术都需要全新的存储架构支撑。

安德斯，想请你分享一下：针对 AI 需求，存储架构设计需要考虑哪些关键因素？

京瓷代表：

好的，罗布。我从 SSD 选择与 AI 存储架构适配的角度，分享一些我们的观察与建议：

首先是 AI 数据摄入阶段。如前所述，这一阶段需要超大规模数据集，但并非所有场景都需选择 128TB 的 QLC（四级单元）SSD—— 例如，部分客户仍在使用 30TB 的 TLC（三级单元）SSD（如我们当前的 CD8P 系列），原因是其支持 4KB 的 IO 单元（input/output unit），若数据管理系统依赖 4KB IO，这类 SSD 会更适配。我们即将推出的 CD9P Fixate 系列，也会提供 60TB 容量的 TLC 选项，同样支持 4KB IO。而若数据系统支持 16KB 及以上的 IO 单元，则更适合选择大容量 QLC SSD（如 128TB 产品）。长期来看，我们预计越来越多软件栈会支持更大的 IO 单元，从而充分发挥高容量 SSD 的优势，但当前仍需根据实际 IO 需求选择。

其次是 AI 训练与检查点阶段。核心需求是 “避免 GPU 闲置”，因此需要最高吞吐量的 SSD—— 这一领域主要使用数据中心级或企业级 SSD。例如，15TB 容量的 SSD（支持 3 次 / 天写入（DWPD，Drive Writes Per Day））是当前 AI 训练的 “甜点级” 选择；有时也会使用 12.8TB SSD—— 并非因为其耐用性，而是通过 “超配（over-provisioning）” 可提升随机写入速度。此外，E3.s 形态的 SSD 在训练环节的 “T0 层”（近计算层）中增长迅速，原因是其在 “中等容量” 场景下具备最佳的机架级效率。

需要注意的是，高容量 SSD（如 128TB）目前仍以 2.5 英寸形态为主 —— 虽然 E3.s 能提升单系统的驱动器数量，但 2.5 英寸 SSD 的单盘最高容量仍是 E3.s 的 2 倍，因此在“极致容量”场景下，2.5 英寸仍是主流。不过，开放计算平台（Open Compute Platform）也在探索新形态因子，未来可能会出现更适合 “最大化容量” 的 EDSFF（企业与数据中心标准形态因子）方案。

最后是 AI 推理阶段。核心需求是 “低延迟” —— 当前的数据中心级 SSD 已能满足这一需求，且成本可控。未来，SSD 领域的创新将集中在两个方向：一是 “更高容量”（持续提升单盘容量，降低每 GB 成本），二是 “更高吞吐量”（如 PCIe 6.0 SSD，或其他突破性技术）。随着市场快速成熟，这些创新将在未来 2-3 年内逐步落地。

主持人：

非常感谢。艾伦，最后想请你从超微的角度，分享一下如何将这些不同的存储分层、不同技术整合，为客户提供优化的解决方案。

超微代表：

要为客户提供全面的优化解决方案，作为供应商，我认为有两个核心要素：

第一是 “深度合作”。超微与各位提到的厂商（英伟达、WEKA、斯卡拉特、京瓷）已合作多年，建立了互信关系，形成了 “虚拟团队”—— 从研发、工程、生产到业务层面，我们共享资源、协同推进。超微在市场上已有 30 余年经验，积累了良好的口碑，这为合作奠定了坚实基础。

第二是 “紧密贴合市场与客户需求”。正如安德斯提到的 “未来几年的技术趋势”，我们需要持续关注市场动态，预判客户需求变化，并及时调整方案。例如，当前 AI 工作负载驱动数据量爆炸式增长，我们通过 “分层存储” 满足当前需求；但同时也要思考：这种分层架构能沿用多久？未来客户还会有哪些新需求？如何提前布局？只有持续跟进需求，才能确保解决方案始终贴合客户实际场景。

超微的解决方案正是围绕 “合作” 与 “需求” 两大核心构建的，这也是我们整合各类技术、为客户创造价值的关键。

主持人：

这是一个非常好的总结点。各位嘉宾分享的内容，将帮助观众重新思考自己的存储架构，以及如何规划未来的分层存储策略 —— 因为 AI 基础设施正在变革，存储需要覆盖 “本地、远程” 等不同场景，甚至在远程场景中也需进一步分层，才能让 AI 真正落地。

再次感谢各位嘉宾的精彩分享！本次对话对我而言极具启发，也希望能为屏幕前的观众提供

有价值的参考。感谢大家观看本环节，超微开放存储峰会的更多精彩内容即将呈现，敬请期待！