Open Storage Summit -- Day 3

主持人：

欢迎参加 Rob 主持的 Supermicro 开放存储峰会。今天，我们将探讨大规模 AI 推理（AI inference at scale） 及相关存储考量。随着企业持续推进 AI 进程，一场根本性变革正在发生 —— 焦点正从训练大型模型转向在生产环境中高效运行这些模型。这场变革的核心在于推理，以及实现推理高效规模化的需求。

推理绝非仅为单个模型提供服务。它关乎在边缘、数据中心和云环境中，实时运行成千上万甚至数百万个由模型驱动的决策。而每一次低延迟响应的背后，都有一个 “隐形功臣”——存储基础设施（storage infrastructure） 。与训练流程不同，推理工作负载高度依赖海量、多样的非结构化数据。这些数据包括日志、图像、视频帧、传感器输出等，通常以文件和对象数据的形式存储，并需输入到以 GPU 为核心的计算环境中。

要实现此类数据的规模化输入，就需要具备灵活性、高吞吐量且高度集成的存储解决方案。这也是部署新型存储协议和数据服务的重要意义所在 —— 它们能支持在分布式推理框架中高效访问非结构化数据，减少瓶颈，进而实现规模化的实时性能。我们还看到分层存储架构（tiered storage architectures） 日益兴起，这种架构能智能平衡性能、容量与成本，确保推理流程顺畅且无需过度配置。

其中，对象存储（object storage）提供了大规模的可扩展性和持久性，而并行文件系统（parallel file systems）则能确保对性能敏感型工作负载的低延迟访问。企业正寻求 “便

捷方案”，以部署经验证的 GPU 优化基础设施，并围绕这些新兴数据流程构建集成解决方案。显然，存储已不再是 “事后考量”。

在本次会议中，您将直接听取行业领导者分享下一代推理基础设施的构建经验，并了解如何扩展非结构化数据流程，以匹配 AI 的运行速度。现在，会议正式开始。今天与我一同出席的嘉宾包括：来自 NVIDIA 的 Harish、来自 Hammerspace 的 Tony 、来自 Cloudian 的 Peter、来自 Solidigm 的 Ace，以及最后一位 —— 来自 Supermicro 的 Tian。欢迎各位的到来！

那么，为了开启讨论，我们不妨先明确各位在大规模推理领域的当前进展。不如从 NVIDIA 开始吧，Harish，您先来分享一下，谢谢。

英伟达代表：

谢谢 Rob，也感谢 Supermicro 提供这个绝佳的机会，让我能在存储峰会上发言；同时也感谢主办方组织本次会议。我是 Harish，任职于 DGX 产品管理团队，负责该团队的存储项目管理工作。我非常荣幸能与大家探讨 “AI 工厂（AI factories)” 以及其新定义 —— 我们如今看到的 AI 工厂所承担的扩展角色。

早期，AI 工厂的重心主要放在训练上：客户部署所有基础设施，主要用于基础模型的训练、微调，然后在其他地方部署这些模型。但如今，AI 工厂的应用范围已得到扩展，不仅包括推理（inference）场景下的生产级服务，还涵盖了数据预处理（data prep)。为实现大规

模推理，我们已向市场、合作伙伴及客户推出两大类技术，这让我们倍感振奋。

第一类是一组服务，我们称之为 NeMo Retriever 服务（NeMo Retriever Services），其核心是生成洞察，并连接生成式 AI 引擎（generative AI engines）与企业数据。第二类是一组用于通过聚合推理（aggregated inference）实现生成式 AI 纵向扩展和横向扩展的技术。接下来，我们将深入探讨这两类技术。

首先来看 NeMo Retriever：据预测，到 2028 年，企业数据（尤其是非结构化企业数据）将从 140 字节（原文：fourteen zero bytes，表述可能存在偏差，按原文直译）增长到 30 字节（原文：30 bytes，表述可能存在偏差，按原文直译）。其中大部分数据将以文档、非结构化视频、图像、日志文件等形式存在。企业数据文件中包含的数据实际上非常复杂，涵盖多种文件类型和多种数据模态 —— 这些文件内部可能包含图像、图表、示意图、文本等。

要从数万亿个文件中提取这些数据的价值，并将其用于生成洞察，是一个极为复杂的过程。NVIDIA 推出了以 “NeMo Retriever” 为名的一组技术，旨在通过 GPU 加速所有必要的处理流程 —— 这些处理针对的是将作为生成式 AI 和视觉语言模型（vision-language models）上下文数据的数据。

实际上，NeMo 是 “NVIDIA Inference MicroService”（NVIDIA 推理微服务）的缩写，它基于云原生架构，易于使用。这些 NeMo 服务会从各类企业数据文件中提取所有数据、生成嵌入向量（embeddings），并将嵌入向量存储在向量数据库（vector database）中，

以便进行后续排序和优化，从而生成更优结果。这样，当任何生成式应用需要查询数据时，都能从企业数据中找到或提取正确的上下文，进而为您生成贴合上下文的、有意义的洞察。

在整个流程中，存储扮演着关键角色。我建议所有观众和合作伙伴访问 “nvidia.com”（原文：belly dot and video dot com，推测为网址表述偏差，应为 nvidia.com），了解更多关于 NeMo Retriever 与存储的关联信息：向量数据库对所有嵌入向量的摄入，以及索引创建，都是以存储为核心的操作，需要高性能存储 —— 这也正是高性能文件访问和高性能对象访问至关重要的原因。

此外，在 AI 工厂方面，我们正与多家高性能对象存储供应商合作，以扩大客户的选择范围：除了并行文件系统和 NFS 文件系统，还将高性能对象存储纳入其中，使其成为 AI 工厂的坚实基础。

过去几年，随着模型规模不断扩大，一些挑战逐渐显现。我还记得 2018 年 BERT 模型推出时的场景，而如今最先进的模型（state-of-the-art models）规模已是 2018 年（原文：2080，应为 2018 年笔误）推出的模型的数千倍。我们看到生成式应用中出现了许多推理智能体（reasoning agents），这些智能体在协同完成各类任务时，会生成比以往多 100 倍的 “推理令牌（reasoning tokens）” 和 “思考令牌（thinking tokens）”。同时，模型支持的上下文长度也大幅增加：Llama 模型刚开源时，第一代模型的上下文长度仅为数千个令牌，但如今最先进的 Llama 模型（原文：lama for models，应为 Llama 模型）支持的上下文长度已高达 1000 万个令牌。

这些变化都增加了部署的复杂性，也对基础设施和模型管理提出了更高要求。NVIDIA 内部以及 Kashmir（推测为某研究机构）和其他公司都开展了大量研究，这些研究成果最终促使我们推出了 NVIDIA Dynamo—— 这是 NVIDIA 推荐的聚合推理解决方案，旨在助力生成式 AI 实现扩展。

接下来，我们简要介绍 Dynamo 的架构：它采用高度模块化设计，且基于开源（open source）理念 —— 您可以根据需求灵活使用其任意组件。Dynamo 包含 KV 缓存管理器（KV Cache Manager），该管理器能跨多个内存层级管理 KV 缓存状态，无论 KV 缓存状态存储在 GPU 内存、系统内存、本地存储还是网络存储中。它通过智能缓存策略，将最常使用的 KV 缓存状态保留在 GPU 内存中，而将不常使用的 KV 缓存状态卸载到存储中，从而实现高效管理。

此外，我们还推出了一个名为 Excel 的库，它是 “NVIDIA Inference Transfer Library”（NVIDIA 推理传输库）的缩写。该传输库能帮助 KV 缓存管理器在不同内存层级间迁移 KV 缓存状态。我们鼓励所有存储合作伙伴开发自定义后端，以与 Excel 集成 ——Excel 同样基于开源理念，因此如果您有需求，欢迎与我们联系，我们很乐意提供支持。

主持人：

非常好。现在我们请 Tony 发言，Tony，麻烦为我们介绍一下 Hammerspace，以及贵公

司在推理领域的定位和作用。

Hammerspace 代表：

好的。Hammerspace 是一家软件公司，也是 AI 数据平台领域快速崛起的领导者。如果允许我这样说，我们正在实现一场 “认知飞跃”，打破当前的行业现状 —— 我们的客户正选择 Hammerspace，而非传统存储方案。

我们的核心优势包括：基于标准的并行文件系统（支持 NFS 4.2，同时也支持 S3，因为我们发现客户对 S3 的需求正不断增长），但所有功能都运行在高性能并行文件系统之上；我们将对象视为文件，并提供配套的数据服务；我们支持多站点数据访问 —— 无论您的 GPU 资源位于全球何处，都能获取分布式数据；我们还具有极强的生态兼容性 —— 刚才各位提到的技术栈集成需求，在 Hammerspace 数据平台中都能通过可编程方式实现。

数据平台的核心价值就在于生态协同，因此我们正与行业领先的第三方解决方案合作，将其整合到我们的数据平台中 —— 无论是软件层面、解决方案层面，还是系统层面。而 Supermicro 正是我们在这一领域的战略合作伙伴。

我们的客户包括 Meta（Meta 正将我们的方案用于训练和推理），同时我们也与 NeoClouds 合作，并重点服务企业客户 —— 因为企业领域将是 AI 下一波大规模增长的

核心市场。我们提出了 “AI Anywhere（AI 无处不在）” 的理念，接下来我会详细介绍：您的数据是分布式的，而 Hammerspace 所做的，就是消除 “数据引力（data gravity）” —— 这并非夸张说法，而是我们实际能实现的目标。

关键在于，我们通过将数据与物理层解耦来实现这一目标，并确保整个过程具备高度智能 —— 包括可控性、安全性和针对性（即对数据的文件级精细控制，包括数据的编排位置和迁移方向），最终实现 “在需要的时间将数据送达需要的位置”。

此外，我们还能充分利用现有基础设施：无论您的数据存储在 20 个不同的存储系统中，还是希望将 GPU 服务器内的存储作为高性能层级以获取极致性能；无论您的数据分布在全球 6 个不同地点，还是 GPU 位于数据中心且希望弹性扩展至公有云或 Neo Clouds 以补充本地 GPU 资源；无论您面临训练所需的海量数据，还是推理流程所需的持续数据流 —— 这正是 Hammerspace 的价值所在。

同时，我们还支持全面的数据生命周期管理。人们常关注 “最后一英里” 的速度，但数据管理需要覆盖整个流程，而这正是 Hammerspace 的独特优势：从数据摄入、预处理到后续所有环节，Hammerspace 能编排并整合数据，将数据实时送达 GPU 服务器以支持推理等工作负载。核心在于，我们无需手动进行数据暂存（staging）和去暂存（de-staging），而是通过编排和整合实现这些流程的自动化。

不妨这样理解：数据生命周期就像一场接力赛，最终的速度取决于所有选手的配合，而非仅靠最后一棒。数据始终处于流动状态，而选择何种路径至关重要 —— 是穿越复杂曲折、

缓慢痛苦的 “迷宫”，还是走规划清晰、路径明确的 “坦途”？这是一种全新的思维方式，但理解其价值的客户已从中受益。接下来，我将话筒交给我的同事 —— 来自 Cloudian 的 Peter。

主持人：

Peter，麻烦为我们介绍一下 Cloudian 及其对象存储在这一领域的定位。目前我们已经有了不错的整体认知，但还想进一步了解 Cloudian 在整个推理流程中的角色，谢谢。

Cloudian 代表：

谢谢 Rob。Cloudian 是一家对象存储供应商，其解决方案已在全球数千个站点部署。我们的核心目标是：接收来自任意来源的数据，并满足各类存储需求。

从历史来看，对象存储的重心多集中在数据长期留存、数据保护等领域；而现在，我们正探索如何优化这些能力，以适配 AI 工作负载，并满足今天讨论的各类需求。我们的重点在于与 NVIDIA 的合作，共同推动我们眼中的行业变革 —— 即 NVIDIA 的 GPU Direct for Object Storage（GPU 直接访问对象存储） 技术。这项技术近期刚推出，其核心是建立

从存储到 GPU 内存的直接并行通道，从而在使用对象存储的同时，实现数据访问的加速——这是一个许多人尚不熟悉或未曾预料到的领域。

我们已公布了性能测试结果：在 6 节点系统中，吞吐量可超过 200GB/s。同时，该技术还能减轻传统 CPU 的负担。我们的目标不仅是成为一家对象存储供应商，更是成为 AI 数据平台领域的对象存储供应商。

需要强调的是，正如 Tony 刚才提到的，S3 API 已成为行业标准。我们 100% 严格遵循 S3 API，因此无论您的应用是在云端还是本地部署，都能直接使用相同的 API——事实上，我们认为，您可以基于相同的软件开发工具，在对象存储平台上构建应用，从而实现上述高性能。我们正推动这项技术纳入 S3 API 标准，相信这将彻底改变人们使用对象存储的方式。

凭借这一性能优势，我们得以实现 Harish 刚才提到的场景：将 NeMo 模块（尤其是 Retriever 模块）直接嵌入存储平台。其核心思路是：在数据所在位置，为这些微服务能力提供加速支持。我们认为，对象存储是能支持各类 AI 工作负载的平台，因此将这些能力直接集成到我们的 HyperStore 平台中，既能满足 AI 的数据提取需求，也能满足数据检索需求。

针对今天讨论的核心——大规模推理，我们的关注点在于：随着客户推理能力的不断扩展，我们的性能也能实现相应的规模化提升，以满足长期需求。我们将此称为 “Cloudian 新型数据中心 AI 平台（Cloudian New Data-Centric AI Platform）”。

若您关注开放工作流的四个关键阶段，就能理解我们的定位：

数据摄入（Ingestion）：接收来自服务器、传感器、云端等各类来源的原始数据（其中大部分为非结构化数据），并进行本地存储，这是第一步；

数据处理与留存（Processing & Preservation）：利用 NeMo Retriever 等工具为数据处理提供支持，并实现数据的长期留存，为后续环节做准备；

数据消费（Consumption）：为数据消费方提供支持，无论是模型训练、微调，还是本次重点讨论的推理。

这就是我们在该领域的定位。

主持人：

非常好。我认为这些方案需要协同作用，才能为企业提供简单、经济高效的解决方案 —— 毕竟这也是所有企业关注的重点。接下来，我们请 Ace 发言，从推理的角度分享一下 Solidigm 的观点和定位。麻烦您为我们介绍一下，谢谢。

Solidigm 代表：

好的，谢谢 Rob。我来自 Solidigm，我们主要生产 SSD（固态硬盘）—— 与其他嘉宾分享的精妙解决方案相比，这听起来可能相对简单，但我认为您可能有点低估我们，也有点高估这个 “简单” 了。

主持人：

或许您确实有点低估自己了。

Solidigm 代表：

其实，这意味着我们的工程师团队每天都在实验室里与物理定律 “博弈”，致力于研发性能更强、耐用性更高（higher endurance）、能效更佳的 NAND 闪存 —— 这些都是客户告诉我们他们关心的核心特性。而这些产品最终会被集成到其他嘉宾所讨论的优秀解决方案中，为其提供支持。

可能有些观众对 “Solidigm” 这个名字不太熟悉 —— 我们是一家相对年轻的公司，成立约 3 年半，但我们的技术传承更为悠久：Solidigm 由 SK 海力士（SK Hynix）控股，源自 2021 年 SK 海力士对英特尔（Intel）存储业务的收购。因此，我和许多同事都在行业内深耕多年，只是曾效力于不同公司。

如今，我们高度聚焦 AI 领域 —— 因为 AI 正推动全球数据量呈爆炸式增长，而所有这些数据都需要存储载体。我们常说 “无基础设施，无数据（No data without

infrastructure）”，而我想补充的是 “无效率，无基础设施（No infrastructure without efficiency）”。因为我们不断听到关于数据中心能耗的 “惊人预测”：数据中心空间有限、冷却面临挑战，构建高效数据中心需要考虑诸多因素。而很多人可能没有意识到，如果存储未经过优化，会占据大量空间和能耗 —— 如果仍在使用传统存储，那么优化的空间可能非常大，而 Solidigm 的产品正是为此而生。

我们拥有完整的产品系列，从 AI 应用角度来看，以下两款产品能很好地体现客户的选择方向，它们对应 AI 集群中存储的两个核心应用场景：

左侧：直连存储（Direct Attached Storage，DAS）：即 GPU 服务器（如 Harish 提到的 DGX 服务器）内部的存储。这类场景需要高性能 SSD，通常采用 PCIe Gen5（并将向 Gen6 过渡）接口，基于 TLC（每单元 3 比特）闪存 —— 这是高性能存储介质。这类 SSD 的核心作用是确保为 GPU 持续供数，使其保持最高利用率，而这也正是 Hammerspace 擅长支持的场景（顺便恭喜 Hammerspace 去年营收增长 10 倍）。

右侧：高密度存储：采用 QLC（每单元 4 比特）闪存，目前多为 PCIe Gen4 接口。这类产品的核心优势是效率和密度 —— 以我展示的这款硬盘为例，它的尺寸仅相当于一副扑克牌，但容量可达 122TB，且仍在不断提升。这意味着完成相同存储需求所需的硬盘数量更少，进而节省空间、降低能耗和冷却成本 —— 这些都是实实在在的优势，也与 Cloudian 的 HyperStore 方案所追求的价值高度契合。

我认为各位嘉宾在 AI 领域的目标高度一致。接下来，我想简要谈谈 “为何存储对推理至关重要” —— 这也为后续讨论埋下伏笔。

很多人会说 “更好的 SSD 能带来更好的 AI 结果”，这确实没错，但具体如何体现？核心在于理解在运行神经网络的过程中，内存与存储的交互机制 —— 我们在这方面投入了大量时间和精力：运行工作负载、分析其特性、理解 AI 工作负载的 I/O 特征，进而构建能优化数据存储位置的解决方案，确保数据在需要访问时已准备就绪。

从方案角度来看，主要有两类方向：

主流方向：包括数据摄入和归档 —— 如果您的推理流程需要从 SSD 中提取输入数据，就要确保数据能快速获取；同时，在流程结束后，也要确保能高效存储所有输出数据。

RAG 扩展（RAG Scaling）：这是另一个重要领域，企业正迅速采用 RAG 技术 —— 他们青睐 RAG 能引入模型训练时未涉及的额外数据，从而为问题提供更优质的答案。但问题在于，如果将所有 RAG 数据都存储在内存中，成本会迅速攀升。因此，现在出现了将部分 RAG 数据从内存卸载到 SSD 的方案 —— 通过这种方式，企业能根据自身应用或业务问题，调整参数以实现更优的总拥有成本（TCO）。

此外，还有两类新兴方向：

KV 缓存卸载（KV Cache Offload）：Harish 刚才提到了 Dynamo 带来的令人兴奋的可能性，我们正密切关注这一领域，并在实验室中开展相关测试 —— 这是一个充满活力的领域。

模型权重卸载（Model Weight Offload）：通过特定技术，无需将整个模型始终保存在内

存中，只需保留必要部分，其余部分存储在磁盘中。这一技术的价值在于：能在内存较小的 GPU 上运行原本无法支持的复杂模型。

以上这些，都是高性能存储与内存协同工作，为推理提供更优结果和新可能性的具体体现。

主持人：

我完全认同您的观点。接下来，我们请 Tian 也加入讨论。对我而言，核心问题是 “如何在有限空间内实现更多功能”，而您刚才提到的 SSD 正从多个维度解决这一问题。Tian，麻烦您为这一部分内容收尾，为我们介绍 Supermicro 在整个生态中的定位和作用。

超微代表：

谢谢 Rob。感谢所有合作伙伴今天出席会议，分享深刻见解和前沿技术。我认为，可以这样定义 Supermicro 的角色 —— 我的职责就是向大家展示，Supermicro 如何将所有这些技术整合为一个完整的解决方案。因为归根结底，任何希望实现大规模推理的企业，都需要一套完整的解决方案 —— 仅靠某一个组件无法取得成功，关键始终在于整体集成方案（total integrated solution） 。

我想通过 Supermicro 的发展历程来阐述这一点，相信大家最终会理解其相关性。Supermicro 的基因中，始终蕴含着 “整合”“合作” 和 “整体解决方案思维”。我们的旅程始于早期的主板制造 —— 在很多人看来，主板只是组件级产品，但实际上，它的设计初衷就是服务于整体解决方案。

经过卓越的工程研发和对 “通过整合实现品质” 的理解，我们的业务逐步扩展到服务器领域，并在该领域取得了优异成绩。近年来，AI 热潮为数据中心带来了前所未有的挑战 —— 包括规模扩展、设备需求激增等一系列问题。

接下来，我想深入解释 “整体集成方案” 的含义。传统上，人们认为技术栈的顶层是网络，下方是相互连接的服务器机架，再往下是核心组件系统。但这种认知并未反映全部事实 —— 在构建技术栈的同时，还需考虑一系列其他挑战、服务、集成技术和相关工作。

如左侧所示，Supermicro 拥有丰富的经验和成熟的服务，能够帮助客户应对整个技术栈的挑战，提供完整解决方案。

主持人：

完全同意。这也为我们接下来的讨论做好了铺垫。那么，我们还是从您开始 ——NVIDIA 与

整个存储基础设施领域有众多合作伙伴，众所周知，NVIDIA 在行业内拥有广泛的客户群体，也见证了各类应用场景。能否分享一下，NVIDIA 目前如何与整个存储行业开展合作？

英伟达代表：

当然可以。这是一个很好的问题，谢谢。NVIDIA 始终秉持 “生态共赢” 的理念 —— 帮助合作伙伴成功，就是我们的成功。我们将这一理念同样应用于存储合作伙伴，并与包括行业领导者和新兴企业在内的所有合作伙伴，通过两种核心方式开展紧密合作：

第一种方式是认证计划（certification programs），针对不同市场领域制定了明确的认证体系：

DGX 认证计划：面向所有 DGX 超级组件和基础组件产品线；

Neo Cloud 认证计划（NCP）：面向希望大规模部署基础设施的 Neo Cloud 从业者；

NVIDIA 认证存储计划：面向 OEM 服务器和存储合作伙伴 —— 非常荣幸，Supermicro 是该计划的核心合作伙伴。

第二种方式是技术协同研发：在我们开发新技术、识别存储合作伙伴的集成机会时，会积极向所有存储合作伙伴开放 NVIDIA 的新兴技术 —— 甚至在技术尚未完全成熟时，就邀请存储生态伙伴早期参与，以便他们能尽早开展与 NVIDIA 技术的集成工作，开发更高价值

的产品和解决方案，最终服务于客户。

主持人：

这一点至关重要。我也非常认同您提到的 “开源理念” 和 “生态开放”—— 大家都在讨论 NeMo、Dynamo 等技术。接下来，我们再次请 Ace 发言 —— 您之前提到 “为 GPU 供数” 和 “高性能存储的关键作用”，而 “推理卸载” 以及 “数据如何从内存迁移” 是我们刚才略作提及的话题。能否深入探讨这一点？因为对于企业而言，在采购 AI 基础设施（尤其是推理相关设施）时，这是架构设计中至关重要的考量因素。

Solidigm 代表：

当然可以。这一领域正涌现大量创新，有许多令人兴奋的工作正在推进。本质上，我们观察到的现状是：随着 RAG 数据的爆炸式增长和模型复杂度的提升，目前已难以（甚至无法）将所有数据都存储在内存中 —— 即使技术上可行，经济成本和技术挑战也难以承受。因此，核心问题变成：如何设计一个协同工作的系统，以解决这些经济和技术问题，同时实现更优结果？

我们已开展了相关研究（相关内容已发布在官网和 GitHub 上，欢迎大家查阅和试用）。我们的研究始于一个核心问题：在推理流程中，有多少数据可以迁移到存储中？这样做能带来哪些收益？

最终，我们基于开源组件构建了一套方案（任何人都可以使用），该方案主要实现两个目标：

模型权重从内存迁移到存储：例如，在白皮书（white paper）中，我们介绍了如何在 L40S GPU 上运行 700 亿参数的 Llama 模型 —— 如果没有 SSD 的支持，这种配置是无法实现的。这是模型与 GPU 协同作用的结果，而数据向存储的迁移是关键前提。

RAG 数据从内存迁移到存储：在生成嵌入向量并填充向量数据库时，会同时构建索引 —— 我们利用微软的 DiskANN 技术，将向量数据索引也迁移到存储中。

测试结果显示了两大优势：

内存占用降低：这是该方案的核心目标。在针对 1 亿个向量数据集的测试中，通过将数据卸载到磁盘，DRAM（动态随机存取存储器）的使用率降低了约 57%。

性能提升：这一点有些反直觉，甚至令我们感到惊讶 —— 将部分 RAG 数据迁移到存储后，性能反而有所提升。通常，人们会认为 “从内存迁移到存储会导致性能下降”，这是常见的权衡取舍。但我们发现，DiskANN 方案使用的索引算法效率极高 —— 在将数据从内存迁移到存储后，不仅性能未受影响，在部分场景下甚至实现了性能提升。

最终结果是：不仅每秒查询数（Queries Per Second，QPS）有所增加，每美元每秒查询

数（QPS per dollar） 也得到提升 —— 这正是客户最关心的核心指标。

如前所述，我鼓励大家查阅相关资料。我们还开发了一个演示案例（demo），以交通安全为应用场景：上传一个十字路口的视频，然后可以实时观察整个流程 —— 第一个模型是视觉语言模型（vision-language model），生成每个视频帧的文本摘要；第二个模型生成嵌入向量并填充向量数据库；第三个模型最终生成报告，内容包括 “十字路口的实时情况” “存在的安全隐患”“改进建议” 等。在我们构建的控制面板上，有一个开关可以切换 “内存运行” 和 “存储运行” 模式，大家可以亲自观察两种模式的结果差异。欢迎大家体验。

主持人：

非常好。对我而言，核心在于 “平衡” —— 在规模化场景下平衡性能与效率。正如您所说，700 亿参数的模型在推理时，可能会用到小型语言模型（SLM）、大型语言模型（LLM），甚至在生成式 AI（generative AI）场景下用到多个模型。但归根结底，所有这些都需要 “数据供能”。Tony，Hammerspace 的发展速度非常快，在文件存储领域正与大型企业竞争。能否分享一下，Hammerspace 的竞争优势是什么？以及 Supermicro 在未来合作中扮演的角色？

Hammerspace 代表：

当然可以。主要有以下几点：

第一，我们拥有高性能、可轻松实现大规模扩展的文件系统。许多客户重视 “基于标准” —— 因此，遵循行业标准至关重要。

但真正的核心优势在于 “AI Anywhere（AI 无处不在）” 理念：客户的数据是分布式的，他们希望利用我们的多站点能力、全局命名空间（global namespace）、数据编排（data orchestration）以及混合云支持 —— 几乎所有客户都采用某种混合云模式，这一点至关重要。此外，我们还支持多协议环境：除了 NFS，还支持 SMB（因为许多数据源仍在使用 SMB）和 S3；同时，我们还能从 Dropbox、OneDrive 等平台提取和整合数据。

因此，核心价值在于：无需手动进行数据暂存和去暂存，而是通过自动化 AI 流程实现数据整合。同时，我们还能对数据进行智能识别 —— 并非所有数据的价值都相同，我们会在流程中对数据上下文进行判断。

这些能力使我们区别于 “同质化产品” —— 我们拥有丰富的功能，这也是我们能在竞争中胜出的原因。

此外，我们采用软件定义（software-defined） 模式 —— 这既是优势，也可能给客户带来挑战：因为客户需要完整的解决方案。而 Supermicro 作为完整解决方案提供商，与我

们开展合作，能为客户提供 “交钥匙” 系统（turnkey systems）。

我们之所以重视与 Supermicro 的合作，还因为 Supermicro 在 AI 领域拥有深厚的积淀 —— 几乎所有 AI 项目中都能看到 Supermicro 的身影。他们理解 AI、精通 AI，在行业内享有良好声誉。与 Supermicro 合作，能显著加速我们的市场渗透 —— 这是一次非常成功的合作。

主持人：

我完全理解。我认为企业确实需要 “便捷方案”，而正如您所说，你们在并行文件系统领域的工作（以及基于 Linux 的标准支持），确实为企业部署提供了更多便利。接下来，我们将话题转向对象存储 —— 这并非突兀的跳转。Peter，我们都曾在另一家大型企业工作，那家公司也涉及对象存储业务。但为何对象存储会成为 AI 未来发展的核心？另外，Cloudian 并非硬件公司，而是软件公司 —— 这一属性如何与当前行业趋势（尤其是 AI 推理领域的趋势）相契合？

Cloudian 代表：

好的。我认为人们最常问的问题是 “对象存储为何能适配 AI 领域”，答案其实很简单：对象存储是数据的 “归宿” 。所有人都明白数据对 AI 工作负载（无论是数据提取、模型训练还是推理）的重要性。而人们面临的最大问题之一，就是需要访问各类数据 —— 包括摄入的原始数据（如视频流、传感器数据、文档等非结构化数据），这些数据如今都存储在对象存储中。

因此，对象存储自然成为 AI 工作负载的起点。而现在，我们正将各类流程（如 Harish 提到的数据价值提取、模型训练支持）整合到工作负载中 —— 尤其是在推理领域，对象存储已具备五年前不具备的性能，且在本地部署场景下的性能表现，也优于云端对象存储。这种本地部署的系统，能够支持大规模 AI 推理。

在 AI 解决方案领域，我始终认为：“说能解决问题很容易，难的是在规模化场景下不出现故障”。规模化的定义虽因人而异，但核心在于：Cloudian 作为软件公司，旨在将我们的技术能力扩展到多个平台。而与 Supermicro 服务器环境的合作至关重要 —— 随着系统规模扩大，我们需要满足工作负载的存储需求；同时，随着对 I/O 能力需求的提升（以支持高性能交互），我们也高度依赖 Solidigm 等公司提供的硬件支持。

尽管我们是软件公司，但我们非常依赖硬件合作伙伴。我们正与今天在座的所有企业合作，共同打造能支持对象存储各类功能的 AI 解决方案。

主持人：

我完全同意。当您审视数据（例如从 “青铜级” 到 “黄金级” 的 medalla 策略）时，会发现数据在整个生命周期中会以不同形式被使用 —— 尤其是在 AI 领域。

Cloudian 代表：

所有环节都需要协同作用。

主持人：

Tian 在您看来，面对如此完善的生态系统，Supermicro 如何为客户整合这些解决方案？此外，企业在推进 AI 进程（尤其是推理相关进程）时，应考虑哪些关键因素？

超微代表：

谢谢 Rob。这是一个很好的问题。我认为您刚才提到的几个关键点（以及其他嘉宾强调的内容）都指向一个核心 ——合作带来整合，整合催生解决方案。Supermicro 的核心角色，就是作为 “纽带”，将所有合作伙伴和技术整合为一套集成解决方案。

幻灯片中展示的是一个典型的 AI 基础设施示例（可根据需求扩展）。从中可以看到，大家通常会考虑的组件包括：网络栈，以及与之连接的各类服务器机架。但整合的核心在于理解

技术间的关联 —— 例如，应用层、GPU 层如何通过高速网络与高性能存储层交互；此外，还需考虑通过独立网络连接的容量存储层。

这些都是 Supermicro 正在推进的工作。我们希望提醒所有正在考虑大规模 AI 推理的企业和观众：在推进项目前，务必退后一步，从整个生态系统的角度进行设计 —— 这样才能避免因后续重新架构而导致的过度支出，确保数据在整个流程中高效流动，而这正是大规模推理的核心需求。

主持人：

我完全认同这一点。接下来，我想向所有嘉宾提出一个共同问题 —— 大家可以准备一下，我们先从 Harish 开始。“智能体（Agents）” 是当前的热门话题，今年甚至被称为 “智能体之年”，所有人都在讨论各类智能体应用。而智能体的构建，在很大程度上依赖于推理（以及其他类型的 AI 技术）。那么，在推进智能体相关项目的过程中，企业应如何为推理环节做好准备？作为智能体的核心构建模块之一，推理环节需要关注哪些要点？各位对客户有何建议？

英伟达代表：

这是一个重要的问题。很高兴您先邀请我回答 —— 这正是 NVIDIA 对 “AI 工厂扩展角色” 的核心思考方向。

AI 工厂的核心组成部分包括：高性能计算、高性能网络、高性能存储 —— 这些组件需在与今天在座的领先合作伙伴的协作下，实现良好整合与测试，从而形成一个通用基础。基于这个通用 AI 工厂，可运行各类工作负载：无论是包含多个智能体的生产级推理工作负载，还是大规模训练任务，都能在同一 AI 工厂中运行。

从 NVIDIA 的收购动作中，也能看出我们在软件基础设施领域的布局 —— 例如，Run:AI 是一套优秀的工作负载管理技术套件，支持在同一基础设施上运行各类工作负载，包括训练、数据预处理和推理。

建议大家深入了解 Run:AI 和 AI 工厂的相关信息 —— 这些都是构建 AI 工厂的优秀平台和软件基础。

主持人：

非常好。接下来，我想请 Peter 回答同样的问题 —— 为了推进智能体相关项目，企业在推理环节应做好哪些准备？您有何建议？

Cloudian 代表：

我认为，首先要理解 “智能体 AI（Agent AI）” 的本质 —— 它是一种以目标为导向的解决方案，通常包含多个推理步骤，可能需要调用多个 LLM（大型语言模型）。同时，您会发现，智能体需要访问多个数据集、多种信息（无论是企业自定义或专有文档，还是其他数据集或数据库中的信息）—— 这些数据通常存储在不同位置（如对象存储的存储桶中）。

因此，对象存储是存储这些信息的理想选择，能确保智能体的推理模型随时访问所需数据。在我看来，企业需要做好准备，以支持各类数据类型 —— 我们正处于多模态数据时代。尤其是在构建智能体 AI 时，您可能无法提前预知需要访问哪些数据集；但如果这些数据集已存在（例如存储在对象存储的存储桶中），就能随时满足智能体的需求。

因此，在我看来，对象存储为企业进入智能体 AI 时代提供了一个 “AI 就绪” 的起点。

主持人：

我同意。接下来，我们请 Tony 发言 ——Hammerspace 认为企业应如何为智能体时代

的推理环节做好准备?

Hammerspace 代表:

很高兴您让他们先发言，因为我完全同意他们的观点。智能体时代的核心在于 “整合”——这正是数据平台的价值所在。将存储系统称为 “数据平台” 往往只是口头上的说法，但真正的价值在于：能与技术栈上层的应用、工具整合，通过协同作用，在工作流程中为不同资源（而非单一资源）提供数据推理支持。

关键在于：通过自动化实现这一整合，并赋予数据上下文信息。在 Hammerspace 的方案中,我们的数据编排(data orchestration）能力能将数据部署到最高性能的存储层(如 Tier 0 层)，或利用现有基础设施实现高性能访问（如 Tier 1 层)。例如，我们之前讨论的 “降低能耗”——利用 GPU 服务器内部存储（无需额外外部存储)，通常不会增加能耗；而这一切决策，都贯穿于整个数据流程。

这种 “端到端流程优化”“与上层技术栈整合”“在数据层引入智能” 的能力，正是 Hammerspace 在智能体时代的核心聚焦方向——无论涉及生成式智能体（generative agents）还是物理 AI（physical AI)，这些能力都能在编排层和数据流程中发挥作用。

主持人：

我完全同意。接下来，我们请 Ace 发言 —— 数据最终需要 “落地” 存储，无论是缓存还是长期存储。那么，从推理角度来看，企业在向智能体时代迈进的过程中，应关注哪些要点？您对客户有何建议？

Solidigm 代表：

当谈到智能体 AI 时，我想到 OpenAI 首席执行官 Sam Altman 几个月前在播客中说过的一句话：“AI 很难，而能源是最难的部分（AI is hard and energy is the hardest part）”。确实，有很多挑战，但能源消耗无疑是最受关注的问题之一。

我想补充的是，随着智能体 AI 时代的到来，挑战只会增加 —— 因为在智能体工作流程中，一个人类生成的初始提示，会迅速被 AI 智能体转化为数百甚至数千个增量提示，进而生成大量增量数据和需要处理的令牌（tokens）。

AI 智能体的能力令人惊叹，其应用前景也令人兴奋，但支持这些能力需要大量资源 —— 因此，高效利用资源至关重要。而在边缘场景中，挑战更为严峻：我们都认同 “将计算靠近数据源” 的理念（从延迟、安全性等角度来看，这有无数好处），但边缘场景往往存在核

心数据中心没有的限制（如空间、能耗、冷却等）。

因此，研发下一代更高耐用性、更高密度的存储（减少硬盘数量，节省空间和能耗），并将其与行业领先的软件（今天在座的企业都提供了优秀的软件方案）相结合，对于实现智能体 AI 的高效规模化至关重要。

主持人：

我完全同意。如今的存储已不再是 “传统存储” 了。Tian，您之前提到了 Supermicro 的参考架构（reference architectures），现在请您为我们总结一下：客户在向智能体时代迈进、推进推理项目时，应重点考虑哪些因素？毕竟，Supermicro 为客户提供了 “便捷方案” —— 您有何建议？

超微代表：

谢谢 Rob。在我们迈向智能体时代的过程中，显然很多人会关注边缘场景和随之而来的复杂性。因此，我给所有正在推进这一进程的企业的最佳建议是：退后一步，全面思考—— 理解各类技术的整合方式、各自的优势，以及如何通过协同作用实现更大价值；同时，做好架

构设计，确保数据流程顺畅，不损失数据传输速度。

归根结底，AI 不仅需要从整体角度思考，还需要利用其优势改善我们的工作和生活 ——这才是核心目标。

主持人：

我完全同意。今天的讨论非常精彩。总而言之，推理是智能体构建的核心支柱之一，而构建坚实的推理架构，就像为房屋打下稳固的地基。各位嘉宾分享了许多值得企业参考的要点。感谢大家今天的参与！