

HAMMERSPACE

激活您服务器中未被开发的Tier 0级存储

利用您已拥有的当前最高性能NVMe存储加速AI工作负载

By Floyd Christofferson

V. 1.0 - March 2025

概述

当前存储厂商之间正在展开一场激烈的竞赛，他们争相标榜自己的存储系统是应对人工智能工作负载持续增长（性能和扩展性）的理想解决方案。但问题远不止于哪家存储供应商能提供最高性能，能以加速GPU服务器的数据吞吐量这么简单。更严峻的挑战在于，随着AI模型复杂度提升且推理会话需要处理更长的上下文序列，工作负载经常会导致GPU服务器内存和主系统内存达到饱和状态。由此带来的直接后果是数据传输至外部存储时的频繁换页操作和I/O等待时间增加。这种日益增长的外部存储网络流量严重影响GPU计算工作负载，并导致本地数据中心和云端数据中心的运营成本上升。

核心症结在于：即使厂商不断宣传其最新存储性能提升方案，或推荐在GPU服务器上安装专属客户端软件来实现智能缓存、令牌优化等应对措施，根本问题仍然存在——外部存储与GPU服务器之间的数据传输通道始终受制于网络带宽瓶颈，这与存储类型无关。

虽然通过增强智能调度来优先处理需要传输至GPU的数据子集是重要且有效的优化手段，但即便采用最先进的优化方法，为满足GPU运算需求而必须突破网络瓶颈的数据量仍在持续攀升，这正是问题的核心所在。

值得庆幸的是，针对这些挑战的卓越解决方案其实早已存在于每台运行标准Linux发行版的本地或云端GPU服务器中，这就是长期被忽视的GPU服务器内置本地NVMe硬盘空间。

本白皮书将阐述各类规模的组织机构，如何通过激活这些闲置资源，将其转化为无缝集成、数据受保护且性能卓越的共享Tier 0存储环境，从而显著加速GPU工作负载并提升GPU利用率，最终实现可量化的效益和成本节约。

激活后，Tier 0 是目前为止AI工作负载性能最强的存储类型。当这些容量在集群中的多台GPU服务器间完成整合后，企业能以显著更低的成本充分释放GPU投资潜力，无论是在本地还是云端。

Hammerspace Tier 0存储解决方案是目前人工智能工作负载领域性能最强的共享存储方案。基准测试显示，在同等客户端连接数情况下，Tier 0方案可支持的GPU数量是使用Lustre方案的16.5倍。当这种存储能力聚合到GPU服务器集群时，机构无论部署在本地还是云端，都能以显著降低的成本从GPU投资中获得更大的效益。

尤为重要的是，这项创新技术无需更换现有基础设施，即可与任何厂商的本地或云端GPU（或CPU/ARM）服务器协同工作，通过激活已支付了购买费用的服务器内部NVMe硬盘空间，来实现性能飞跃。

需要再次强调：这种**超高速本地NVMe硬盘容量**——当前大部分处于未充分利用状态——**属于沉没成本**，因为它已包含在本地和云端GPU服务器的购置成本中。对于电力供应接近饱和的数据中心而言，这种方案不会增加额外电力需求的特点也构成一项净收益。

通过Hammerspace技术，将本地NVMe硬盘激活为Tier 0共享存储，其利用了Linux内核的智能功能，且该功能已集成于所有标准发行版中。这意味着服务器无需安装任何专属客户端软件，也无需对出厂配置进行任何修改。正如后文所述，Tier 0是一项具有颠覆性价值的解决方案，它能立即为各种规模的部署环境创造价值——无论是仅有数台GPU服务器的小型集群，还是拥有数千块GPU的超大规模环境。

更重要的是，利用本地Tier 0共享存储，可大幅降低甚至完全免除对超高性能网络基础设施和/或额外外部高性能存储系统的采购成本。当企业试图通过使用昂贵的方案来缓解网络对GPU满载运行的影响时，Tier 0方案能从根本上解决成本问题。

通过Hammerspace实现的Tier 0共享存储不仅具备上述优势，还能提供全局多协议文件/对象访问能力、无中断数据编排、数据保护等企业级数据服务。这些功能通过统一混合云环境中任意厂商存储资源的全局命名空间来实现，从而加速AI投资的成果转化。最终，Hammerspace使客户能够简化多厂商/多站点基础设施的管理复杂度，并显著降低运营成本。

超高速本地NVMe硬盘，虽长期处于未充分利用状态；但实则已是沉没成本，因其已包含在GPU服务器的采购价格中，无论是本地部署还是云端部署。

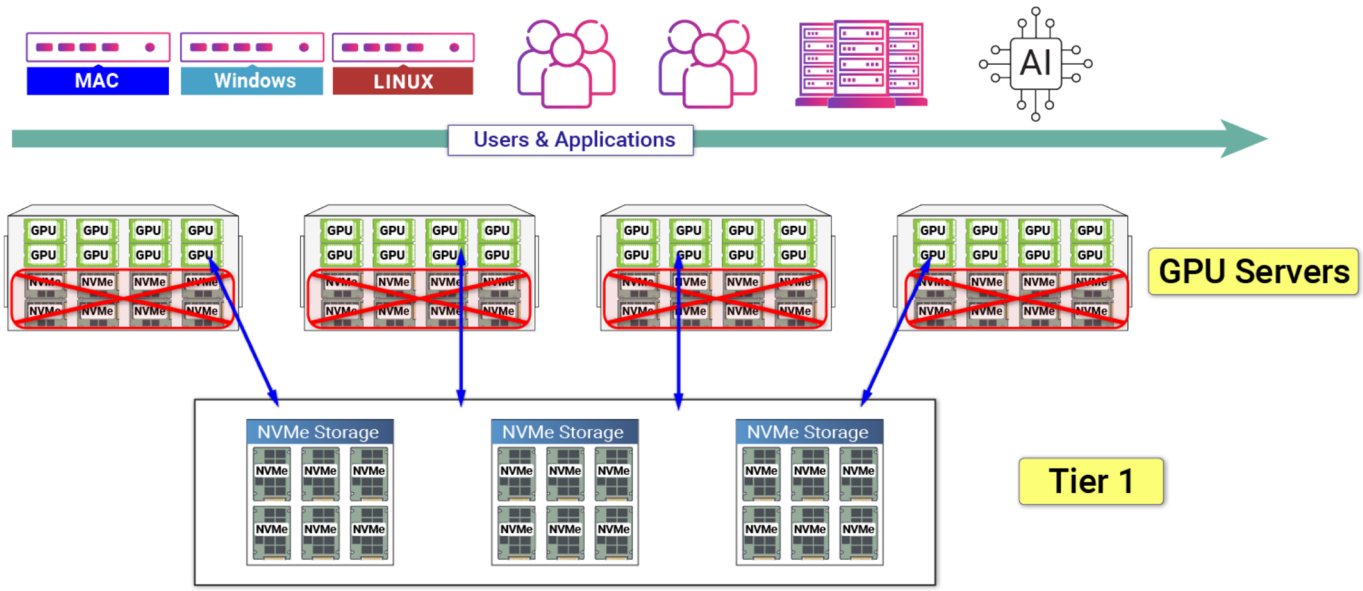




激活Tier 0：GPU服务器本地NVMe硬盘

每台GPU服务器内部都潜藏着一项未被开发的资源，能彻底改变GPU计算型工作负载的性能格局。几乎所有GPU服务器——无论是本地还是云端——都运行着标准Linux发行版，这些系统已内置Hammerspace多年来为开源社区贡献的NFS与pNFS v4.2高性能协议。

此外，这些服务器至少配备2块（通常为8块）高性能本地NVMe硬盘。部分厂商提供的服务器甚至支持最多16块本地NVMe硬盘。不同制造商在服务器中预装的NVMe硬盘数量与容量存在差异：有些允许仅部分插槽安装磁盘，而另一些（包括云端GPU服务器）则默认全数配置。



上图：在GPU集群中，GPU服务器中的NVMe容量大多未被充分利用。GPU所需的数据必须通过网络从外部存储读取或写入外部存储，这会减慢工作负载并降低GPU的利用率。

问题在于：尽管部分应用能有限利用本地NVMe硬盘作为高性能缓存或临时暂存空间（尤其适用于单台服务器或较小训练数据集），但在**两台及以上GPU服务器集群或大型训练数据集场景**中，这类NVMe硬盘往往沦为孤立容量，难以充分使用。原因如下：

- **容量孤立**：本地NVMe硬盘被隔离在各台服务器内部，GPU服务器无法跨集群共享这些存储容量。
- **缺乏共享访问**：若不桥接这些磁盘，以创建跨服务器共享存储环境的方法，本地存储上的数据必须手动复制，导致额外的开销、延迟和操作复杂性。
- **有限数据保护**：由于GPU服务器的NVMe硬盘缺乏跨集群的内置冗余，若服务器崩溃或出现其他故障，将面临数据丢失或服务中断的高风险。

大容量SSD与大型GPU服务器集群正在改变行业格局

随着SSD硬盘容量增大及GPU服务器集群规模扩大，行业格局正在发生根本性转变。截至2025年3月，GPU服务器厂商交付的服务器普遍配备30TB至122TB本地NVMe硬盘容量。这些配置基于八盘位和十六盘位的服务器架构，单个SSD的容量在4TB到8TB之间。

当前企业级SSD的存储密度正在快速提升，30TB和60TB硬盘已成为市场主流。根据制造商的技术路线图，预计122TB硬盘将于2025年第四季度上市。

这意味着即使只采用当前服务器标配的较小容量硬盘，一个八节点集群在默认配置下即可获得接近1PB的超高速NVMe硬盘存储。对于可配置的GPU服务器，若将同规格八节点集群升级至30TB硬盘，可用容量将提升至3.8PB。

相比较于提供给GPU的“每GB/秒/价格/性能”，外部存储成本比Tier 0高约10倍。

Tier 0创新的颠覆性意义在于：这种极致性能可完全依托现有GPU服务器和电力预算实现，无需额外配置带有配套交换机基础设施、占用机架空间、增加电力消耗等成本的外部高性能存储系统。**从单TB成本来看，完全利用GPU服务器已内置NVMe硬盘的成本，仅相当于通过外部存储实现同等容量所需投入的一半。但真实的性价比优势体现在了性能成本上：在200Gb/s网络环境下，为GPU提供数据的综合成本分析显示，外部Tier 1 NVMe存储每GB每秒的综合成本约为内部Tier 0存储的10倍。**该Tier 1存储的综合成本计算已包含硬盘及机箱、电力/冷却、机架空间、200Gb/s网络等要素。若升级至成本更高的800Gb/s交换机网络，Tier 0存储的性价比优势将更加显著*。

下文将详细阐述，激活如此大规模的高性能存储容量（结合AI集群中持续增长的GPU服务器部署量），正在推动企业重新评估其IT架构布局，以及对外部高性能存储和网络基础设施的投资策略。

*企业级SSD的价格差异很大，具体取决于容量、采购数量、制造商和驱动器类型。在本文中，我们基于12.8 TB的硬盘，按每TB 80美元的价格进行了比较计算，我们认为这是典型的实际平均价格。

网络瓶颈是核心问题

此处关键在于：无论外部存储速度多快，其实际性能始终受限于服务器与存储之间的网络连接速度。这一事实并非新发现，在所有的IT环境中历来如此。

但对于以GPU为核心的AI训练和推理工作负载而言，这一限制的影响远比传统企业IT场景更为严峻。因为另一个现实是：若需要更快的网络通道来维持GPU满载运行，网络基础设施的额外成本将急剧攀升。随着GPU集群规模扩大导致数据量激增，以及更复杂模型和更长上下文长度引发的流量暴增（导致服务器系统内存饱和），网络瓶颈问题只会愈发严重。

即使采用昂贵的400Gb/s InfiniBand交换机和市面最快的存储设备，通过网络写入数据所需时间仍比直接写入配备八块硬盘的GPU服务器本地Tier 0存储多约2.5倍。这种延迟意味着宝贵的GPU算力可能因等待数据到达而被浪费——即便已使用缓存或其他token优化技术也不例外。

若网络速度为100Gb/s，向外部Tier 1存储写入相同数据量的耗时将比服务器内部Tier 0存储多出近10倍。

虽然通过不同数据保护方法、亲和性调度（智能地将高频访问数据尽可能靠近所需GPU）等技术可优化数据流，上述数值存在一定变量空间。但核心结论不变：任何需要跨网络传输的数据访问速度，必然低于直接利用服务器内置的本地NVMe存储。如下表所示，这一性能差异始终存在。

1. Tier 0 vs. Tier 1 - 本地部署（90%的网络效率）

假设Checkpoint大小为500GB

	存储类型	最佳吞吐量	有效吞吐量	写入500GB所需时间
Tier 0	Local NVMe (8-drive GPU server)	112 GB/s	112 GB/s ✅	~4.5 sec ✅
Tier 1	100 Gb/s Network	12.5 GB/s	11.25 GB/s	~44 sec (~2.1 min)
	200 Gb/s Network	25 GB/s	22.5 GB/s	~22 sec (~1 min)
	400 Gb/s Network	50 GB/s	45 GB/s	~11.1 sec

Note: *External Tier 1 storage throughput adjusted assuming 90% practical network efficiency.*

云实例场景下的性能差异更为显著：在基于云端的8 GPU H100实例中，500GB检查点文件写入本地NVMe存储仅需约4.5秒。若使用AWS的高性能EBS（io2）存储执行相同任务，则耗时约139秒（约2.3分钟），**时长增加约31倍**。若通过性价比更高的AWS EBS（gp3）存储写入检查点，耗时将达约9分钟，**延迟扩大近124倍**。

2. Tier 0 vs. Tier 1 - 云端存储（90%的网络效率）

假设Checkpoint大小为500GB

	存储类型	最佳吞吐量	有效吞吐量	写入500GB所需时间
Tier 0	Local NVMe (8-drive GPU server)	112 GB/s	112 GB/s ✅	~4.5 sec ✅
Tier 1	AWS EBS (gp3)	1 GB/s	0.9 GB/s	~555.6 sec (~9.3 min)
	AWS EBS (io2)	4 GB/s	3.6 GB/s	~138.9 sec (~2.3 min)
	AWS S3 (via high-speed network)	1 GB/s	0.9 GB/s	~555.6 sec (~9.3 min)

Note: *External Tier 1 storage throughput adjusted assuming 90% practical network efficiency.*

Checkpoint的操作场景尤其具有代表性，因其需保存模型当前状态以保护长时间运行的任务免受节点故障影响。问题在于，部分旧版训练框架中，GPU在等待checkpoint写入存储时必须处于闲置状态。在云端GPU工作负载中，这种闲置成本尤为高昂——无论GPU是否使用，用户都需持续支付费用。

新版训练框架，可通过checkpoint异步操作缓解此问题，最大限度减少GPU停机时间。但Tier 0与外部存储的性能差异不仅限于checkpoint的场景。前文示例中的性能差异同样影响着GPU与外部存储之间的所有数据传输，且数据量往往远超500GB的checkpoint文件。

需要明确的是：即使采购最昂贵的外部存储和800Gb/s网络设备，虽可实现接近Tier 0 NVMe存储的吞吐量，但每个网络端口成本高达数千美元，更遑论额外的电力需求及其他间接成本。

通过Tier 0，这样的极致性能可以利用已内置在GPU服务器的NVMe硬盘容量，并基于已有的网络基础设施实现。

而通过Tier 0方案，企业可利用GPU服务器内已配置的NVMe硬盘资源（已支付成本）及现有网络基础设施即可实现同等性能。即使在现价约200美元/端口的100Gb/s网络环境下，采用Tier 0仍能显著提升工作负载速度。

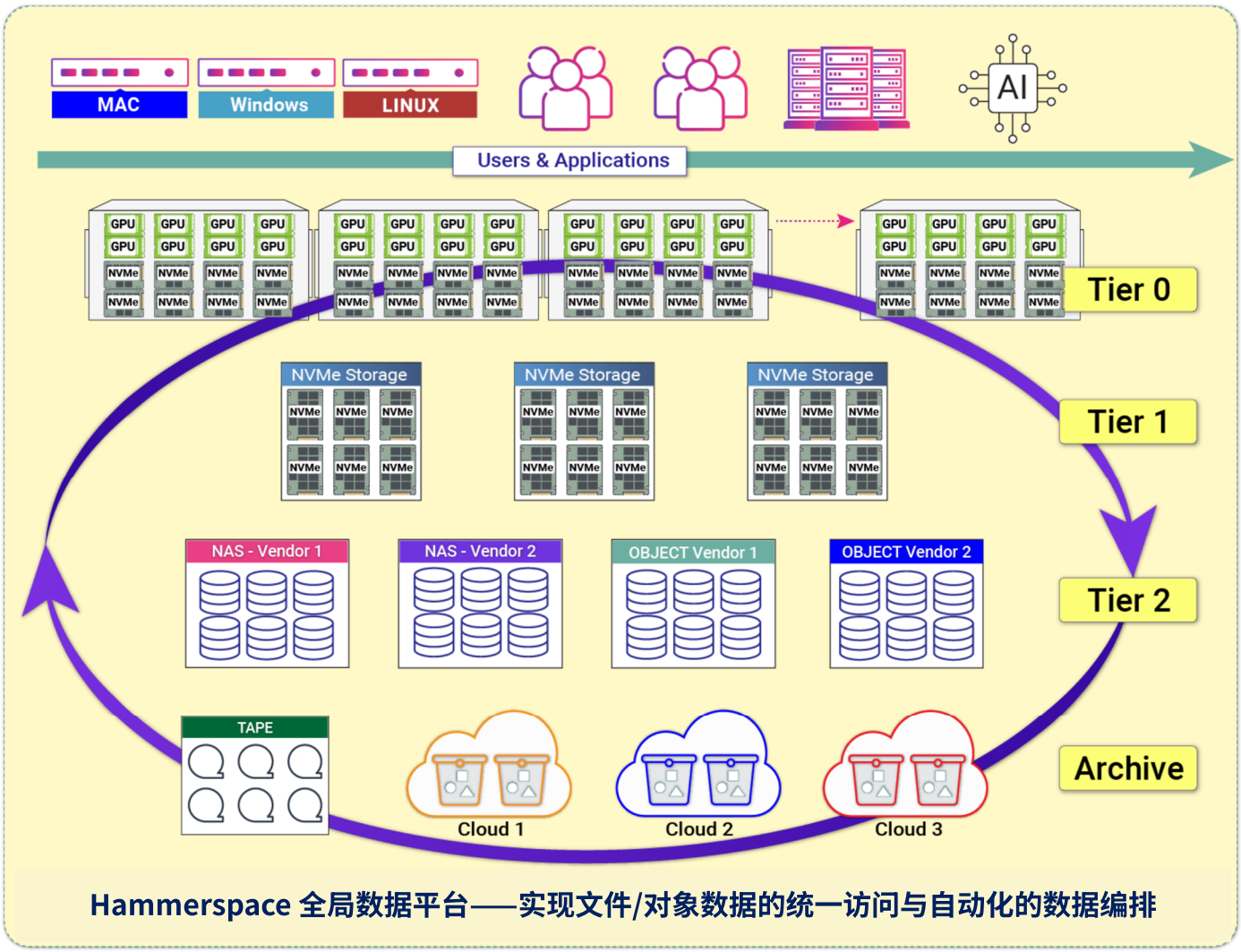
在需要数百个GPU的大规模部署中，若采用外部方案，昂贵网络端口和存储设备的叠加成本将呈指数级增长。简言之，为AI项目额外投入大量资金构建外部网络和存储（尤其当GPU服务器已内置高性能本地NVMe硬盘时），将直接降低AI投资的回报率。

通过Tier 0方案（特别是在大型GPU集群与大容量硬盘配置下），**企业可实现双重优势：既能使用市场上性能顶尖的存储方案，确保GPU数据供给和工作负载流畅运行，又能充分利用已投资的NVMe硬盘资源和现有网络基础设施。**

如何激活GPU服务器中的Tier 0存储容量

在考虑Tier 0存储时，首先需要明确的问题是：既然激活GPU服务器中的本地NVMe硬盘——能够带来如此显著的性能提升和成本优势，那为什么这个办法尚未被广泛采用？为何存储行业主要聚焦于外部存储系统的性能，或是为GPU服务器开发专有客户端软件来优化数据供给？

简单来说，要安全有效地将本地NVMe硬盘容量，激活为集群范围内共用的超高速Tier 0共享存储，必须借助Hammerspace全局数据平台软件所独有的技术能力。



上图: Hammerspace可激活GPU服务器内未被充分利用的NVMe硬盘容量，使其在多层级、多厂商存储环境中成为无缝集成的Tier 0存储层级。Hammerspace数据编排功能可自动化完成数据定位和数据保护，以最大化GPU利用率。

Hammerspace开发的数据平台旨在消除存储孤岛现象，并实现跨平台的数据自动化编排，通过创建高性能的并行文件系统，提供跨任何厂商多种存储类型的统一文件/对象访问能力——包括服务器本地NVMe硬盘、异地数据中心存储及云端存储。

为满足超大规模AI工作负载所需的扩展性与性能，Hammerspace创新性地利用Linux标准pNFS v4.2的并行特性，以极致性能实现数据向GPU的自动化+智能编排。作为Hammerspace多年前为Linux内核贡献的标准协议，pNFSv4.2客户端已预装在全球所有数据中心，并内置于各类Linux发行版中已有五年以上。

这使得Hammerspace成功实现Meta等客户所需的线性扩展能力：支撑Llama 2、3及最新Llama 4大语言模型（LLMs）在成千上万GPU集群上的训练——无论是本地数据中心还是云端皆可部署。更值得注意的是，Hammerspace达成这些突破无需安装任何专有客户端软件，也无需对任何厂商的服务器和存储进行修改*。

解锁Tier 0存储的核心优势

1. 数据的保护与编排

Hammerspace通过将GPU服务器集群中的本地NVMe硬盘容量聚合为无缝的全局命名存储空间，激活并保护这些存储资源。该命名存储空间可整合来自任何厂商的多种存储层级（包括跨多站点和云端存储）。

但仅实现全局文件系统跨所有存储类型仍不够。Hammerspace使Tier 0存储真正高效的核心能力在于：能自动化执行本地或远程Tier 0——与其他存储层级之间的无中断的文件级数据编排。

与传统分层存储管理（HSM）系统或其他单一解决方案采用的简单策略不同（这些策略通常仅支持有限选项的单维度指令），Hammerspace允许管理员创建全面的服务级别目标，这些目标可根据业务逻辑进行精细调整，例如数据持久性、数据布局、工作流自动化及其他数据服务。

多种目标策略均可自定义，可适配几乎无限量的业务规则、场景及工作流，并能应用于全部数据或智能筛选的存储类型子集。

仅覆盖全局文件系统中的所有存储类型是不够的。Hammerspace核心优势在于能够自动化实现——无中断文件级数据编排，跨不同存储类型和地理位置进行数据管理。

*如需了解更多关于pNFSv4.2与Flex Files的行业领先性能的相关信息，请参阅白皮书：“How to Adapt Legacy Infrastructure to Power GPU-Based AL/DL Workloads Using Parallel NFSv4.2”

目标策略可基于文件/对象颗粒度进行应用，包括任意厂商（包括云端）的任何存储类型，确定应如何管理数据、目录和/或存储桶。

如上文示例中checkpoint文件写入本地第0层存储时，可在Hammerspace中设置服务级别目标（SLO）来定义：该检查点的保护机制、所需数据持久性等级、存储位置等策略。

例如，当文件刚写入第0层存储时，Hammerspace即可异步将该文件复制或迁移至下游对象存储，或其他任意存储目标。对于具有较大第0层存储容量的集群，还可将该文件复制到集群中其他节点以实现快速恢复。通过简单的声明式目标策略，即可确定满足性能和/或持久性要求的任意数量文件需求。

Hammerspace并非创建需要维护的分叉文件副本，而是生成共享文件元数据的同一文件实例。这些元数据通过全局统一命名空间进行管理，可供所有用户/应用程序访问。

2. 数据编排为所有工作负载激活第0层存储

Hammerspace数据编排不仅可被用于保护checkpoint文件。借助其通过pNFSv4.2与Flex Files实现的极致性能，目标策略还可自动执行第0层存储的数据移动，实现GPU集群间的并行化工作负载。

需注意：连接计算节点的"横向（东西向）"网络速度远快于服务器与外部存储间的"纵向（南北向）"网络。Hammerspace通过将数据预存至临近计算节点的第0层存储，结合pNFSv4.2并行I/O加速技术，形成双重加速优势，使第0层存储的激活可为任何高性能工作负载带来净收益。

此外，网络具备双向传输能力。Hammerspace可在后台将检查点文件移出第0层存储（GPU节点发送端），传输至外部存储或其他第0层存储节点，同时不会影响GPU节点接收端，从同一第0层存储读取训练数据所需的带宽。

3. GPU资源定位：实现云GPU集群的突发工作负载

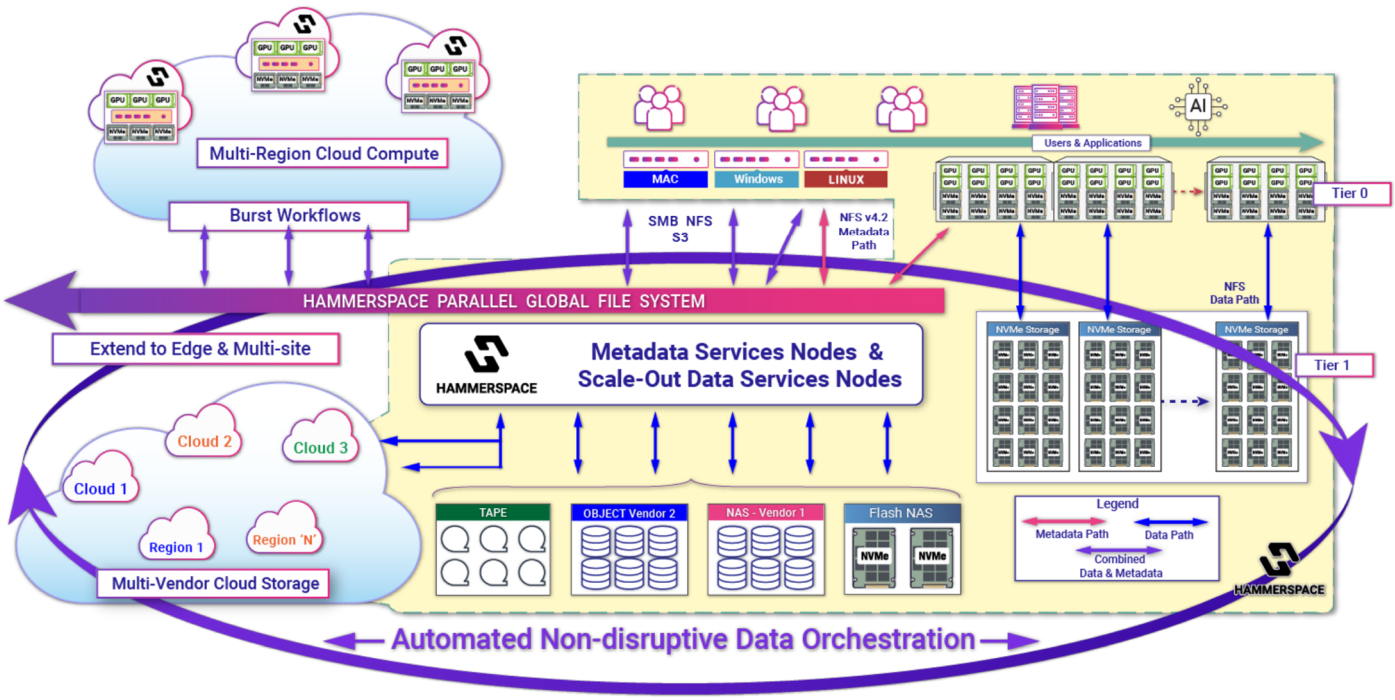
对于开展AI计划的组织来说，本地GPU资源的可用性仍是一项重大挑战。即使具备强大的采购能力，实际使用需求或数据中心电力也可能使购买方案成为非最优选择。因此众多组织正采用第三方GPU即服务供应商或租用云端GPU资源。但这两个方案均存在复杂性，需要人工将文件/对象推送至远程GPU资源，并在新的存储孤岛环境中管理工作负载。

Hammerspace通过将其全局文件系统与数据编排能力扩展至本地数据中心和多个云环境，彻底消除这些问题。只需数分钟即可在云虚拟机实例中启动新的Hammerspace集群，并能根据任务需求弹性扩展规模。

当云端的Hammerspace新实例激活后，数分钟内便会自动与现有本地Hammerspace集群同步文件系统的元数据。此后，所有云端存储（包括云GPU集群内的Tier 0存储）均无缝成为本地Hammerspace全局文件系统的延伸。

Hammerspace消除了这些问题，可以将其全局文件系统和数据编排功能扩展到本地数据中心与一个或多个云端站点之间。

Hammerspace策略（Objectives）可自动编排数据在云与本地之间的流动，如同管理任意本地数据中心的存储层级一样，对云端GPU的Tier 0存储进行相同的管理。此多站点能力可桥接多达16个本地或云端Hammerspace集群，构建跨越任何厂商现有或新增存储的统一高性能并行文件系统，实现真正全局化的数据环境。



上图: Hammerspace全局数据平台通过在GPU服务器内激活Tier 0存储层，提供统一的文件/对象访问和覆盖本地及云端资源的全局数据编排功能，实现对数据的高效管理。

用户和应用程序仍通过标准的SMB、NFS及S3挂载点访问文件系统，无需安装客户端软件或改变现有工作流程。Hammerspace会在后台处理数据放置与其他数据服务，并跨越所有数据孤岛自动执行工作流程，全程都不会中断用户与应用程序的读写操作。

为最大化地节省成本，当突发工作负载完成后，基于云的Hammerspace集群可立即销毁以减少云服务费用。数据编排已作为后台操作，透明完成，确保GPU生成的输出数据获得自动保护，并被精准放置到所需位置——无论存储类型或地理位置。

4. 基于标准的统一文件/对象访问

如前所述，Hammerspace无需在用户电脑或应用服务器上安装专有客户端软件，也无需改造现有存储系统。用户甚至可能察觉不到Hammerspace的存在。终端用户和应用程序看到的仍然是标准的NFS、SMB和/或S3协议访问的全局文件系统共享空间——与以往完全一致，只是现在他们可以通过统一入口访问所有存储类型、孤岛和位置的全局数据命名空间。

用户和应用，仅能通过全局文件系统的共享空间，查看到其权限范围内的数据。需要访问S3存储桶的应用会看到包含任何存储位置数据的S3存储桶；需要通过NFS或SMB访问文件层级结构的应用，则会看到以标准文件/文件夹结构呈现的相同数据。

期望使用S3的应用程序，可以看到包括任何存储上的全部或部分数据的S3存储桶。

期望通过NFS或SMB共享查看文件层次结构的应用程序，可以看到以标准文件/文件夹结构呈现的相同数据。

数据当前具体存储于何种介质完全由Hammerspace的服务级别目标进行动态管理，这个过程对用户和应用完全透明。事实上，Hammerspace整合了其向Linux内核贡献的关键功能——作为pNFSv4.2的Flex Files功能实现的实时数据迁移特性，**使得即使在文件处于打开状态并被持续读写时，也能在后台透明完成跨存储层的数据编排。整个过程不会中断应用程序或用户工作流。**

Hammerspace通过标准协议提供对全类型存储数据的统一访问能力，以及文件/对象颗粒度的后台透明化数据编排能力，这两者在AI工作负载中尤为重要。对于拥有大量待激活非结构化数据的企业来说，无需将整个存储卷迁移到新存储池，Hammerspace仅需将任务所需的特定文件/对象精准编排至计算资源，既不会产生数据分叉副本，也不会干扰其他用户或应用。

彻底告别需要执行成本高昂的数据迁移、将数据副本推送到全新存储设备（这些存储设备也许最终又会形成新数据孤岛）的时代。

关于超融合基础设施（HCI）与基于CPU的服务器

尽管本文的重点讨论是基于GPU的工作负载（这些负载是当前AI场景最迫切的需求），但事实是：**基于CPU的服务器**通常也配备了同样超高速的本地NVMe硬盘，这些硬盘可被激活为**Tier 0存储**。这一特性同样适用于低功耗的ARM CPU。

此类服务器常会被与**超融合基础设施（HCI）**发生混淆。HCI通常将计算与存储集成在单一设备中，主要用于处理**结构化数据、虚拟机和其他面向块的I/O模式**，并不适合处理CPU密集型的非结构化数据工作负载。虽然HCI系统可采用分布式存储架构和其他分层解决方案，但其对块存储场景的专注，使其难以高效处理非结构化数据。

然而，结构化数据仅占企业数据总量的约20%，且年增长率仅约20%。相比之下，**非结构化文件/对象数据**占企业数据的80%，并以每年约65%的速度增长。

由于存在大量**CPU密集型场景**需要处理海量非结构化数据，这样的场景同样也需要通过Hammerspace激活CPU服务器中本地NVMe硬盘的Tier 0存储，获得与上述GPU计算负载相同的性能与成本优势。例如：

•**其他AI场景**：基于CPU的AI推理、预测性分析，或直接处理非结构化数据的机器学习流程。

•**非AI场景**：需要高性能且能受益于Tier 0超低延迟与高吞吐的应用，例如：

- **实时分析与数据流处理**

快速处理日志数据、传感器输入或点击流数据，即时生成洞察。

- **高性能搜索与索引**

快速索引和检索大规模文档库或数据集（如Elasticsearch、Splunk）。

- **视频与媒体处理**

需要快速频繁数据访问的低延迟编辑、转码、渲染与内容分发。

- **金融交易与建模**

依赖实时大数据访问的风险分析、算法交易与金融建模。

- **基因组学与生物信息学**

需即时数据检索的基因组测序、影像数据处理或生物信息学工作负载。

基于CPU的Tier 0存储层的优势

与上文提到的GPU案例相同，基于CPU的Tier 0存储层具备以下核心优势：

- 更低成本：激活服务器内超高性能的本地NVMe硬盘，相比通过昂贵网络连接外部存储阵列，显著降低成本。
- 性能跃升：与外部存储相比，Tier 0存储层可大幅提升性能并降低延迟，直接加速相关应用。

Hammerspace为CPU工作负载提供的额外性能优势

除了通过pNFSv4.2与Flex Files协议实现极致性能，并激活CPU服务器内的Tier 0存储层外，Hammerspace还利用与标准Linux的深度集成，为CPU工作负载提供另一项直接性能优势。

传统方案的性能瓶颈

HCI系统和其他部分厂商解决方案采用基于轮询（polling）的I/O处理机制。这种基于用户空间（user-space）的解决方案要求CPU持续检查（轮询）传入数据——即使没有数据传输时也会持续消耗大量CPU资源。这种过高的CPU负载会导致延迟增加和性能下降。

Hammerspace的突破性方案

Hammerspace利用基于Linux内核的标准NFS协议，通过中断请求（IRQ）方法加速I/O处理，显著降低CPU开销和延迟。与轮询机制需要持续占用CPU不同，**Hammerspace采用的内核级IRQ方法仅在数据实际到达时触发CPU响应，实现更高效的事件驱动处理。**相比依赖CPU密集型轮询循环的用户空间解决方案，该方案实现了更快速、更低延迟的I/O以及更高的CPU使用效率。

未来展望：即将登陆您的Linux发行版

Hammerspace近期向Linux内核上游贡献了另一项关键创新——Linux内核中的NFS协议旁路（NFS Protocol Bypass），这将进一步大幅提升Tier 0本地NVMe硬盘的性能。该功能在Linux社区被称为LOCALIO，通过增强NFS客户端和服务器的智能检测能力，当它们运行在同一主机时自动跳过内核中NFS协议栈的冗余路径，直接访问本地文件系统和存储。该功能采用稳健实现的方式，即使NFS客户端和服务器运行在容器中也能正常工作。

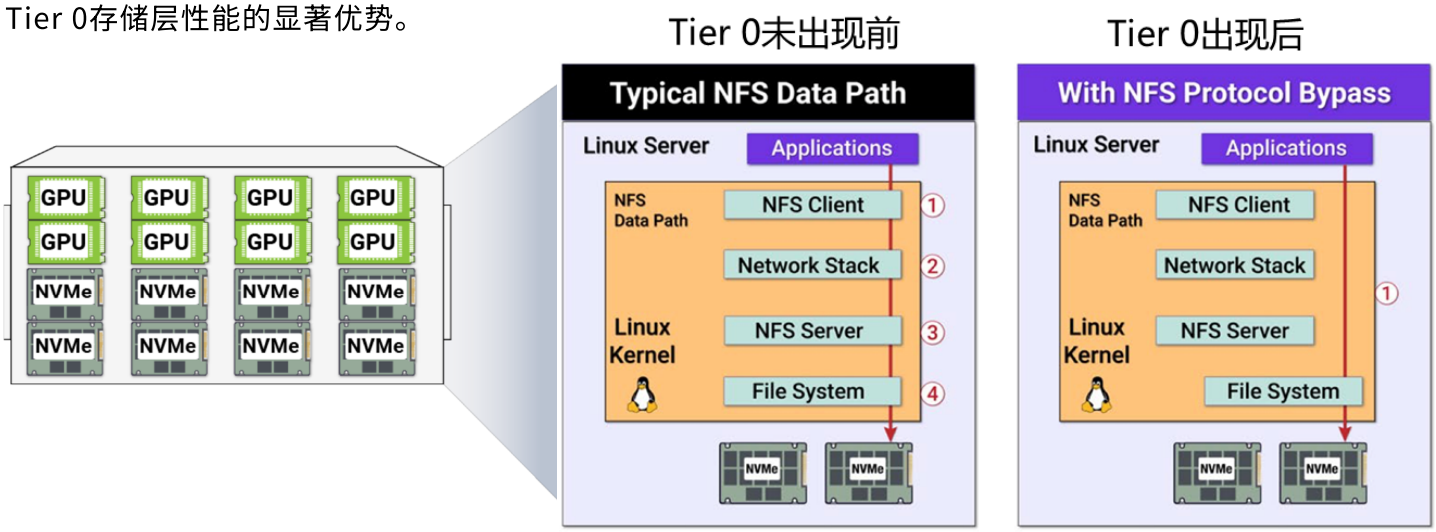
版本演进

- Linux内核6.12（2024年11月发布）：首次实现LOCALIO功能
- Linux内核6.14（2025年3月24日发布）：性能增强与稳定性优化

初步测试结果显示，绕过NFS协议对降低服务器内CPU利用率具有显著效果。此外，当前测试数据表明，在使用LOCALIO时，Tier 0存储层的读取性能实现了突破性的提升。

初步测试结果显示，在使用LOCALIO时，Tier 0存储层的读取性能实现了显著提升。早期测试还表明，写入性能也将得到改进，但具体提升幅度仍在量化评估中。当前测试基于最新内核更新持续进行，待结果就绪后将正式发布完整报告。

结合即将推出的GPU集群数据布局亲和性优化等其他增强功能，LOCALIO特性已证明其能够提供远超当前Tier 0存储层性能的显著优势。



上图：Hammerspace工程团队开发了LOCALIO功能，并将其贡献至Linux内核，旨在为基于GPU和CPU的服务器中的本地NVMe硬盘提供更高性能。

尽管LOCALIO已包含在2024年11月发布的Linux 6.12长期支持版（LTS）中，但企业级Linux发行版对新内核版本的适配仍需时间。不过，由于该功能需求强烈，Red Hat Enterprise Linux (RHEL10)计划于2025年5月发布——其内核将基于Linux 6.12，同时集成了6.14版本中对LOCALIO的增强功能。企业用户升级至RHEL10可能需要一定周期，但一旦完成升级，这些性能改进将立即可用于优化Tier 0存储层。

标准化才是核心

Hammerspace长期致力于向Linux贡献关键内核优化，近年来已累计推动**2,400余项功能与补丁**进入开源内核。

核心团队的技术领导力包括但不限于：

•Trond Myklebust（Hammerspace CTO）：

担任Linux NFS客户端内核维护者超过20年，确保Hammerspace的技术开发严格遵循开源社区标准。

•Mike Schnitzer（Hammerspace高级首席软件工程师）：

现任Linux内核上游**设备映射（Device Mapper，DM）子系统**维护者。

•Tom Haynes（Hammerspace软件工程师）：

2018年参与起草**RFC 8435**协议，定义了pNFS Flexible File Layout（弹性文件布局），其核心特性于2019年被纳入**NFS v4.2规范**。

对于开放标准的专注

Hammerspace对开放标准的坚持与其愿景一致：既要突破企业数据环境中专有厂商孤岛（vendor silos）和厂商锁定（vendor lock-in）的限制，同时保持与现有基础设施的兼容性。

用户/应用侧：
客户无需在应用服务器上安装专有软件，也无需修改用户的计算机配置。

存储侧：
Hammerspace兼容绝大多数厂商的存储设备。

更重要的是，Hammerspace在标准Linux中实现的极致性能改进（这些改进已被纳入所有的主流发行版本）已深度集成到Hammerspace软件中，使其能充分发挥这些优势。

总结

各个行业、各类规模的用户都在探索如何改造其IT基础设施，以应对AI技术的爆发式潜力。传统企业IT的线性模式——即应用与数据的一对一绑定、非结构化数据的快速归档——已不再适用。尽管多数机构意识到需要制定AI战略，但AI投资的回报率仍充满不确定性，尤其是对非超大规模企业而言。行业媒体常暗示，启动AI项目的唯一方式是在现有架构之外另建一套专为AI设计的独立IT架构。

此类解决方案主张采购全新的高性能存储、昂贵的高速网络设备，或在客户应用服务器上部署专有客户端软件，以及其他高成本、厂商锁定的方案。这些做法延续了历史遗留问题：专有数据孤岛割裂访问路径，阻碍数据高效利用，导致数据引力（data gravity）、资源利用率低下，以及资本支出（CAPEX）和运营成本（OPEX）的显著增加。

Hammerspace已验证了，其基于标准协议的解决方案能有效打破孤岛惯性，支持包含现有存储和网络基础设施的高性能应用场景。对于计划启动或扩展AI战略的机构而言，这意味着AI相关资源现在可作为现有IT基础设施的增强组件做到无缝集成，而无需独立构建另一个新的孤岛。

Hammerspace全球数据平台已在真实场景中实现这一愿景，通过以下能力覆盖本地和多云环境：

统一文件/对象访问
跨规模（包括极端数据量与性能需求）的高性能数据编排
支持包括HPC和超大规模AI在内的全行业混合云部署

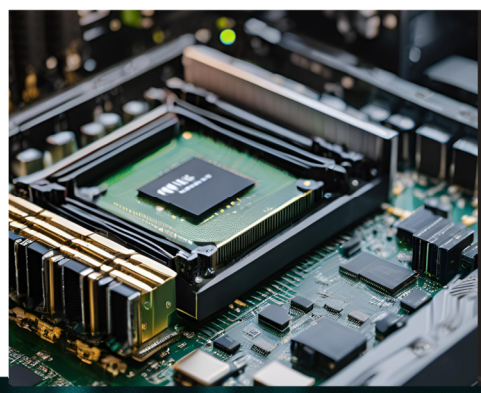
Tier 0的变革性价值

上述这些能力与实战经验使得Tier 0存储层，成为了对基于GPU和CPU的高性能计算负载具有颠覆性影响的解决方案。

Hammerspace可以赋能各行业、不同规模的机构利用Tier 0实现的前所未有性能跃升——轻松超越市场上最快的外部存储层。而通过激活这一未充分利用的极致性能存储层，客户还能同步获得显著的成本节约。**更关键的是，这一性能层可能已存在于您现有的基础设施中，激活它只是我们锦上添花的一项任务。**

而这，也是Hammerspace的创新所在。

对于企业而言，这意味着用户可以专注于人工智能所需的资源是否能够与现有IT基础设施无缝整合，成为其有力补充，而不是作为一个孤立的新孤岛而存在。



如需了解更多-可联系
ralphie.yu@hammerspace.asia
或 18810311799
手机・微信同号
HAMMERSPACE