



HAMMERSPACE

在Checkpointing机制中 使用Hammerspace Tier 0技术后的收益详解



Hammerspace Asia

是我们的微信服务号

发布内容偏重于

公司新闻与行业方案

扫码可关注



订阅号 “焊宙尔空间”

发布内容偏重于

技术细节

扫码可关注



引言

在高性能计算（HPC）和人工智能（AI）应用中，检查点（Checkpointing）是指将应用程序状态周期性保存至持久化存储的过程，从而在发生故障时无需从头重启计算任务即可恢复。然而，检查点机制可能带来显著性能损耗，尤其是在通过网络向共享存储写入海量数据时，主要面临以下挑战：

- ◆ **GPU闲置时间**：检查点操作过程中，GPU需等待所有数据写入共享存储后才能恢复工作，导致昂贵算力资源的利用率下降。
- ◆ **共享存储瓶颈**：多节点同时向共享存储系统写入数据时，可能耗尽网络及存储带宽，延长检查点完成时间。
- ◆ **数据丢失风险**：若仅依赖无冗余设计的本地存储，节点在检查点安全存储前发生故障将导致数据丢失。

本系列白皮书定义了Tier 0存储层级的概念，阐述其在检查点工作流中的应用方法，并量化了Hammerspace Tier 0科技，可以为检查点工作机制带来的核心收益。

Hammerspace Tier 0的优势体现在：

- **检查点完成时间大幅缩短：从分钟级降至秒级**
- **GPU有效算力显著提升：等效于新增数百个GPU的计算资源**
- **成本效益革命性突破：节省数千万硬件扩容开支**

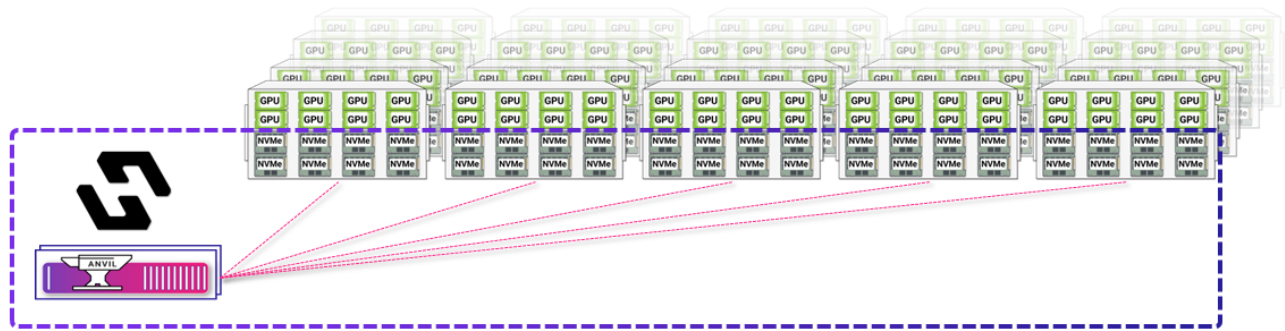
Tier 0是一种新型的超高速共享存储层级，通过将GPU服务器本地NVMe存储空间纳入Hammerspace并行全局文件系统后，其可以被转化为共享存储资源池的一部分。

Tier 0中的数据还可以通过Hammerspace数据保护策略实现安全防护，例如：

- ◆ **跨GPU服务器节点的数据镜像**
- ◆ **确保数据始终在不同存储层级保留三副本**

同时支持数据在存储层级间、跨站点及云端环境间的无缝编排迁移。



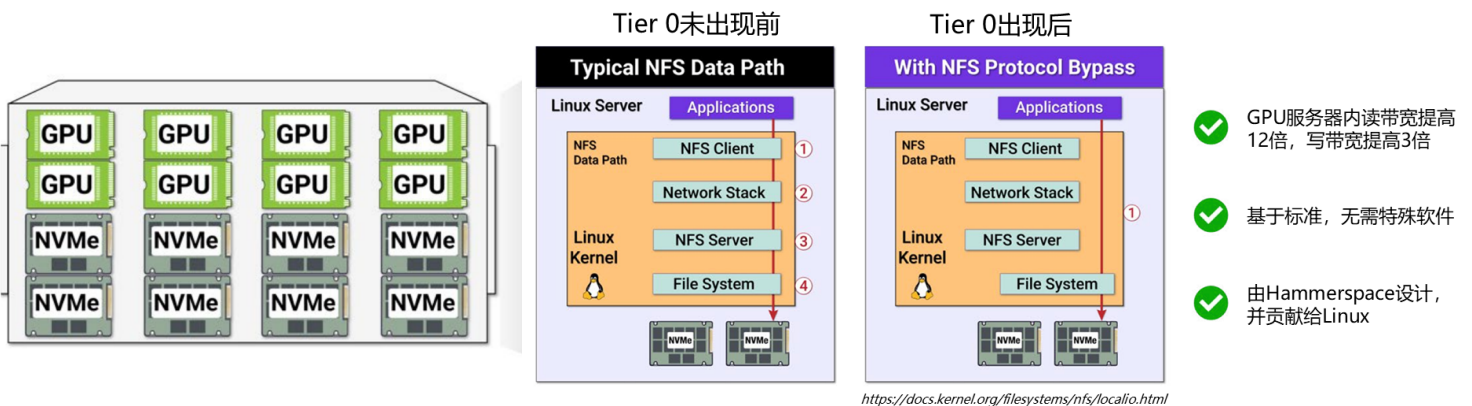


- ✓ 降低成本与能耗
- ✓ 减少GPU闲置时间
- ✓ 更快实现设备价值

上图-Hammerspace的Tier 0 是一种新型超高速共享存储层，通过整合GPU服务器本地的 NVMe存储资源，将其作为『并行全局文件系统』的组成部分。

Hammerspace Tier 0利用超大规模NAS架构——一种基于标准协议的并行文件系统架构——并通过对[Linux内核的更新](#)，绕过NFS客户端和服务器的网络堆栈和网络适配器硬件（如下面图所示）。

实际上，这种在GPU和本地NVMe存储之间创建了一个“捷径”或者说更直接的数据路径，可以减少延迟、增加带宽，并如本文档稍后的分析所示，可以将检查点操作速度加快10倍至100倍。



上图——Hammerspace的Tier 0技术，借助Linux内核的更新,通过创建GPU与本地存储间更直接的数据传输路径，实现了延迟优化



基于Hammerspace Tier 0技术的检查点工作流程示例

Hammerspace的Tier 0技术，能够以本地NVMe访问速度，将数据直接传输至GPU，是AI/HPC等极端高性能应用场景的理想解决方案，并为检查点操作提供多重优势。

使用Hammerspace Tier 0的检查点工作流程示例如下：

步骤1 – 检查点启动：应用程序触发检查点操作，此时每个节点在Hammerspace全局共享存储系统中创建文件。根据Hammerspace中预设的服务等级目标（SLO），系统会选择与每个节点本地NVMe绑定的存储节点来承载文件，从而实现绕过NFS协议和网络协议栈（如上文所述）

步骤2 – 本地写入完成：写入完成后，GPU立即恢复计算任务，无需等待

步骤3 – 异步复制：Hammerspace检测到新检查点文件后，根据预设策略开始将其复制到指定的存储层级

步骤4 – 数据可用性：检查点文件现已安全存储在多个位置，确保在需要时可用于恢复

使用Hammerspace Tier 0的优势体现在

与外部存储方案相比，该工作流程具备以下优势：

优势 #1：将检查点写入本地NVMe存储，绕过网络协议栈

- **本地速度优势：**计算节点直接将检查点数据写入自身高速本地NVMe存储，消除网络延迟和拥塞
- **非冗余写入：**初始写入为非冗余本地操作，支持内存数据快速转储至磁盘，无需网络传输或数据复制开销

优势 #2：通过数据编排实现自动化数据保护、关联性配置及迁移

- **策略化目标设定：**借助Hammerspace数据编排功能，管理员可设置策略（称为"目标"），明确规定检查点数据的处理方式
- **关联性策略：**确保检查点文件仅写入节点本地NVMe存储
- **复制策略：**定义检查点数据何时及如何复制到其他存储层级以实现冗余



优势 #3：异步复制和迁移检查点文件

- **数据迁移：**检查点文件关闭后，Hammerspace 的数据迁移器会异步将数据复制到其他节点或存储层级
- **后台操作：**复制过程在后台进行，允许 GPU 在本地写入完成后立即恢复计算
- **数据安全：**检查点数据会安全存储在其他位置用于恢复，且不影响计算工作负载

优势 #4：本地存储成为全局文件系统的一部分

- **统一命名空间：**检查点文件存在于全局共享文件系统中，确保其在集群内可访问且一致性统一
- **逻辑一致性：**数据在物理上从本地 NVMe 移至其他存储，但在逻辑上仍作为同一文件保留在共享系统中

全局命名空间中Hammerspace数据编排的强大能力

即使Hammerspace出于保护目的进行数据镜像或出于分层目的移动数据——该文件始终是同一文件系统下的同一文件。实际移动的仅是文件的实例（从本地存储移至其他位置：其他GPU服务器或外部共享存储节点）。即使在物理层面上文件内容被移至他处，该文件在命名空间的逻辑位置始终保持不变。

关于使用Hammerspace Tier 0进行检查点设置的效益量化分析

本分析量化了使用Tier 0存储（即利用计算节点本地NVMe存储）进行检查点设置的效益，并与依赖外部网络存储系统（即使通过高速800Gb以太网[800GbE]连接）的传统方法进行对比。

计算环境设定：使用NVIDIA A100 GPU进行计算；集群规模：1,000个节点，含8,000个GPU；存储容量：100 PB；聚合吞吐量：1 TB/秒

本分析将估算以下指标：

故障率：基于硬件故障概率模型，推导集群级故障发生频率

最优检查点频率：结合检查点耗时与故障恢复成本，计算经济效益最大化的检查点间隔

年度总GPU小时增益：量化Tier 0缩短检查点时间所释放的GPU有效算力

相关机会成本：将节省的GPU时间换算为算力货币化价值



检查点数据量大小估算

系统配置:

- 计算节点: NVIDIA DGX A100 或 HGX 系统 • 每节点 GPU 数量: 8 个 NVIDIA A100 GPU
- 单 GPU 内存: 80 GB (也提供 40 GB 版本) • 单节点 GPU 总内存: $8 \text{ GPU} \times 80 \text{ GB} = 640 \text{ GB}$
- CPU 内存: 假设每节点 256 GB (可能有所不同)
- 检查点的需要的总内存: GPU 内存 + 相关 CPU 内存

检查点数据量大小 = 600 GB

检查点写入外部存储的时间估算

向外部存储写入检查点时需考虑的因素包括:

- 网络开销: 协议额外开销、拥塞和延迟会降低有效带宽。
- 共享基础设施: 多个节点同时写入检查点可能导致网络和存储阵列饱和。
- 单节点有效带宽: 通常远低于理论最大值。建议保守估计为每节点 1 GB/秒。

将 600 GB 检查点写入外部存储所需的时间: $600 \text{ GB} \div 1 \text{ GB/秒} = 600 \text{ 秒}$

检查点写入Tier 0本地NVMe存储的时间估算

本地NVMe存储设备规格:

- 每节点NVMe设备数量: 8块NVMe固态硬盘 • NVMe接口协议: PCIe Gen5
- 单NVMe设备带宽: 约14 GB/s (实测写入性能) • 节点总聚合带宽: 112 GB/s
- 假设90%效率的有效带宽: 100.8 GB/s

将 600 GB 检查点写入 Tier 0 存储的时间计算: $600 \text{ GB} / 100.8 \text{ GB/秒} = 5.95 \text{ 秒}$ (四舍五入为 6 秒)

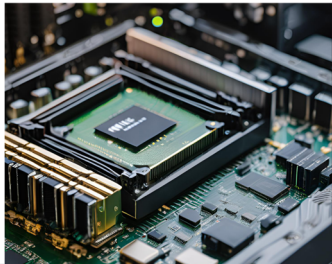


通过Tier 0存储，计算可被节省的GPU工时

基于上述假设并预估GPU服务器的故障率，我们可计算出最优的检查点间隔时间，进而得出每小时因检查点设置与重新计算所产生的总开销。

使用外部网络存储与Tier 0层级存储的对比详见下方汇总表。

参数	外部存储	Tier 0层级存储
检查点数据大小（每节点）	600 GB	600 GB
有效带宽（每节点）	1 GB/秒	100.8 GB/秒
检查点写入时间	600 秒	6 秒
GPU服务器每小时故障率	0.0417 次故障/小时	0.0417 次故障/小时
最佳检查点间隔	54 分钟	17 分钟
每小时检查点次数	1.12	3.54
每小时检查点耗时	212.13 秒/小时	21.21 秒/小时
每小时预期重新计算时间	212.13 秒/小时	21.29 秒/小时
每节点每小时总开销	424.26 秒/小时	42.5 秒/小时
每节点每天总开销	170 分钟/ 天	17.0 分钟/ 天
每节点每年总开销	1032.4 小时/ 年	104.17 小时/ 年
每节点年总节省成本	N/A	928.2 小时/ 年
1,000节点集群年总节省成本	N/A	928,196 小时/ 年
等效新增GPU服务器数量 (总年节省成本 ÷ 8,760 小时/年)	N/A	≈106台服务器
等效新增GPU数量（每服务器8个）	N/A	≈848 GPUs



如需了解更多-可联系

ralphie.yu@hammerspace.asia

或 18810311799

手机 · 微信同号

 HAMMERSPACE



计算使用 Tier 0 的经济效益

上述分析结果表明，在一个包含 1,000 个节点、8,000 个 GPU 的集群中，使用 Tier 0（而非外部存储）相当于额外增加了 106 个 GPU 节点或 848 个 GPU。这等同于每年无需额外投资即可增加 10.6% 的 GPU 计算能力！

这一机会成本的美元价值计算：

- 额外获得的 GPU 数量：848 个
- 单 GPU 成本：2 万至 4 万美元
- 年度 GPU 机会成本：1,700 万至 3,400 万美元

因此，使用 Tier 0 进行检查点存储，**每年可释放 1,700 万至 3,400 万美元的 GPU 机会成本。**

除上述机会成本收益外，Tier 0 还可帮助减少对外部存储容量的需求。在本分析使用的 100 PB 容量基准下，可节省约 4,000 万至 5,000 万美元的成本。您可通过此份[Executive Brief](#)进一步了解 Tier 0 的投资回报率。

结论

分析表明，在大型 GPU 集群的检查点存储场景中，外部共享存储系统无法匹敌本地 NVMe 存储的效率和性能。

通过释放本地 NVMe 存储资源，并将其作为并行全局文件系统的一部分——提供共享存储新层级，Hammerspace Tier 0 实现了：

- **检查点时间的大幅缩短**：从分钟级降至秒级
- **显著的 GPU 时间节省**：相当于新增数百个 GPU
- **可观的财务收益**：避免数千万美元的硬件追加投入

Hammerspace Tier 0 是最大化现代高性能计算（HPC）和 AI 工作负载性能及成本效益的关键。

