# Hammerspace与NVIDIA GPU直连存储的融合应用

技术综述

## 综述

GPU是现代AI开发无可争议的算力引擎，也是其他计算密集型应用的核心组件。单次AI训练任务可能需要处理PB级数据，耗时数周甚至数月。无论何种应用场景，优化从存储到GPU的数据路径对最大化GPU利用率与缩短训练周期都至关重要。

NVIDIA推出的GPU直接存储（GPUDirect Storage, GDS）技术通过远程直接内存访问（RDMA）革新存储与GPU间的数据传输路径。Hammerspace软件现已通过GDS认证，使开发团队能够：

- 利用全局数据视图与并行性能，将任意存储升级为支持GDS
- 加速AI应用训练流程
- 充分释放GPU硬件投资价值

## GPU直连存储架构

GPUs专为大规模并行计算而生，其定制化架构在处理特定任务时比通用CPU快10-1000倍，功耗降低约10-100倍*。

GPU的极致算力对数据供给速度提出严苛挑战。在典型GPU服务器中，存储与GPU间的数据流必须经过：

- CPU处理
- RAM缓冲区中转

目录


- 挑战——加速GPU工作负载
- GPU直连存储架构与我们的优势
- Hammerspace GPU直连存储测试与兼容性
- GPU直连存储的设计考量








HAMMERSPACE

* https://cset.georgetown.edu

**而这会导致多种瓶颈现象产生：**

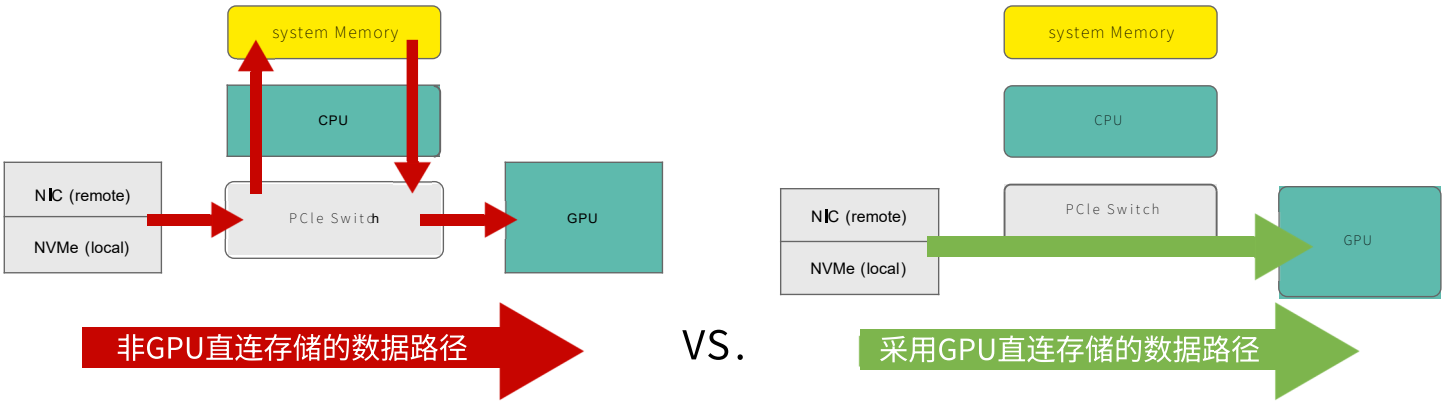
**1. 额外延迟：数据在缓冲区反复拷贝，以及缓冲区的分配与释放操作；**

**2. 吞吐量限制：当CPU的PCIe带宽低于GPU时形成传输瓶颈；**

**3. 应用节流——用于移动数据的CPU在该周期内，无法提供给应用程序所使用；**

**4. 专用于缓冲数据的内存无法供应用程序使用**

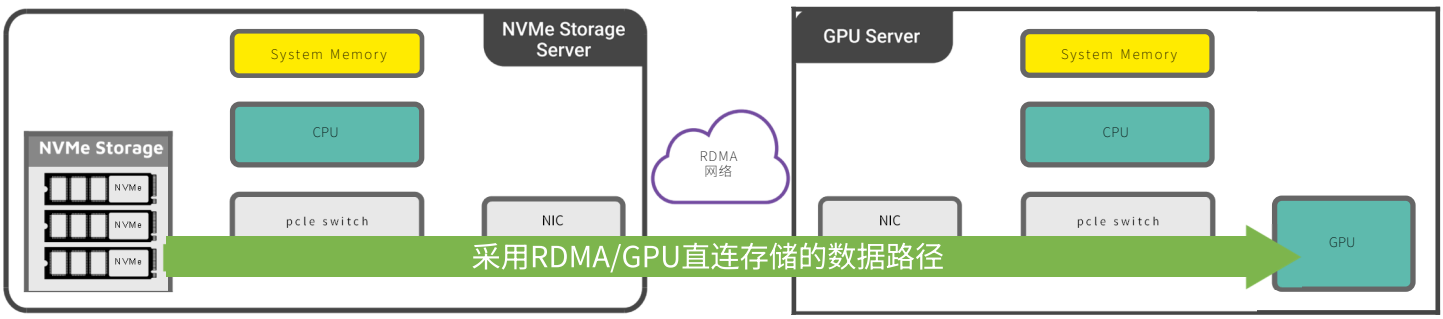
GPU直连存储通过提供GPU与存储间的DMA路径（绕过CPU）规避这些问题。更多CPU周期、内存与PCIe带宽因此可用于其他任务，同时存储性能（包括延迟与吞吐量）也能够得到提升。




*是否采用了GPU直连技术的数据路径对比*

## GPU直连存储与NAS

请注意上图中，GPU直连存储所使用的存储设备可包含两类：本地连接NVMe存储设备和通过网络连接的远程存储。对于网络存储，基于无损网络的RDMA能提供最短且最高效的数据路径，因为RDMA技术既能实现CPU卸载（offload），又能降低存储服务器的延迟。




*上图 - GPUDirect with Remote Storage*

## Hammerspace与GPU直连存储的测试

Hammerspace在自身的全局数据环境中完成了GPU直连存储（GDS）的功能性测试。此实验室测试仅用于验证功能与兼容性，未涉及性能评估，因此采用了最小化系统配置。

## Hammerspace架构概述

在Hammerspace集群中：

- Anvil节点（生产环境通常部署两个）作为元数据服务器，仅提供控制平面（control plane）功能，不参与数据I/O传输。
- DSX数据服务节点支持块存储、文件存储与对象存储访问，处理所有I/O操作及数据编排等高级功能，可按需横向扩展。若存储支持RDMA，则无需通过DSX节点中转，客户端可直接在Anvil节点指导下访问存储。

Anvil节点、DSX节点与存储设备共同构成单一并行横向扩展集群。完整技术细节请参阅《Hammerspace Technology White Paper》技术白皮书。

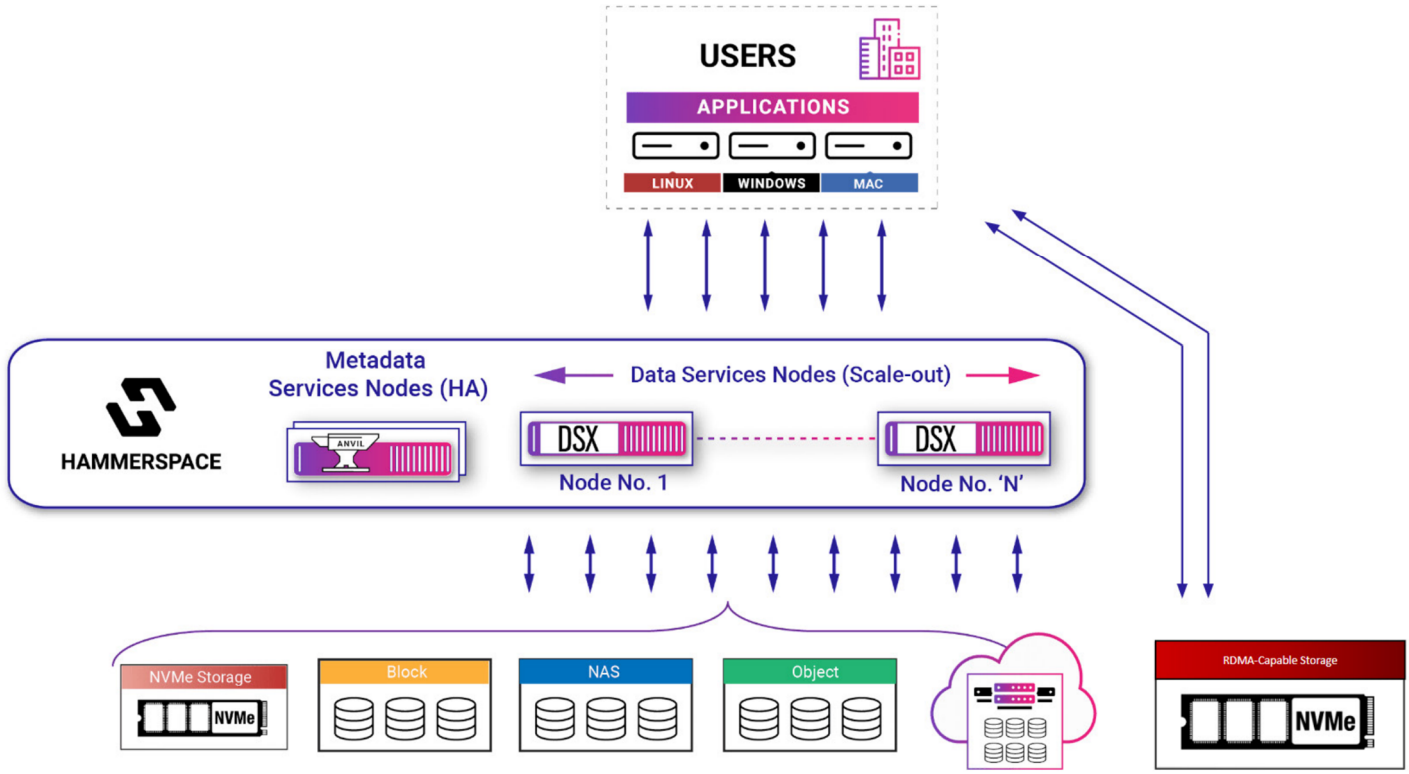



*Figure 3 - Hammerspace High-Level Architecture*

# Hammerspace与GPU直连存储（GDS）的集成方案

Hammerspace系统在支持GDS时无需额外配置要求，仅需满足RDMA基础条件即可。系统可自动检测ConnectX-4（及更高版本）RNIC网卡并启用RDMA支持。此后，Hammerspace将响应NFSv4.2客户端请求——优先提供RDMA数据布局，同时保留标准数据布局作为备选。详细RDMA配置方法请参阅《Hammerspace RDMA技术白皮书》。

如下图所示，Hammerspace与GDS兼容的存储配置方案包括：

- **支持RDMA的NFSv3 NAS存储**
- **连接至DSX节点的块存储**（通常为NVMe闪存）
- **由DSX节点代理的非RDMA NAS存储**

Hammerspace可使**任意类型存储兼容GDS**进行使用——即使原生不支持GDS或RDMA的存储设备。通过DSX节点的Portal功能，可直连非RDMA存储，但此方式无法获得最佳性能。推荐方案为：

1. 在DSX节点内部署NVMe存储
2. 通过策略或卷组，将必要数据从非GDS存储提升至DSX本地NVMe存储

此时DSX的NVMe存储将作为**高速缓存层**，加速对非GDS存储的访问。

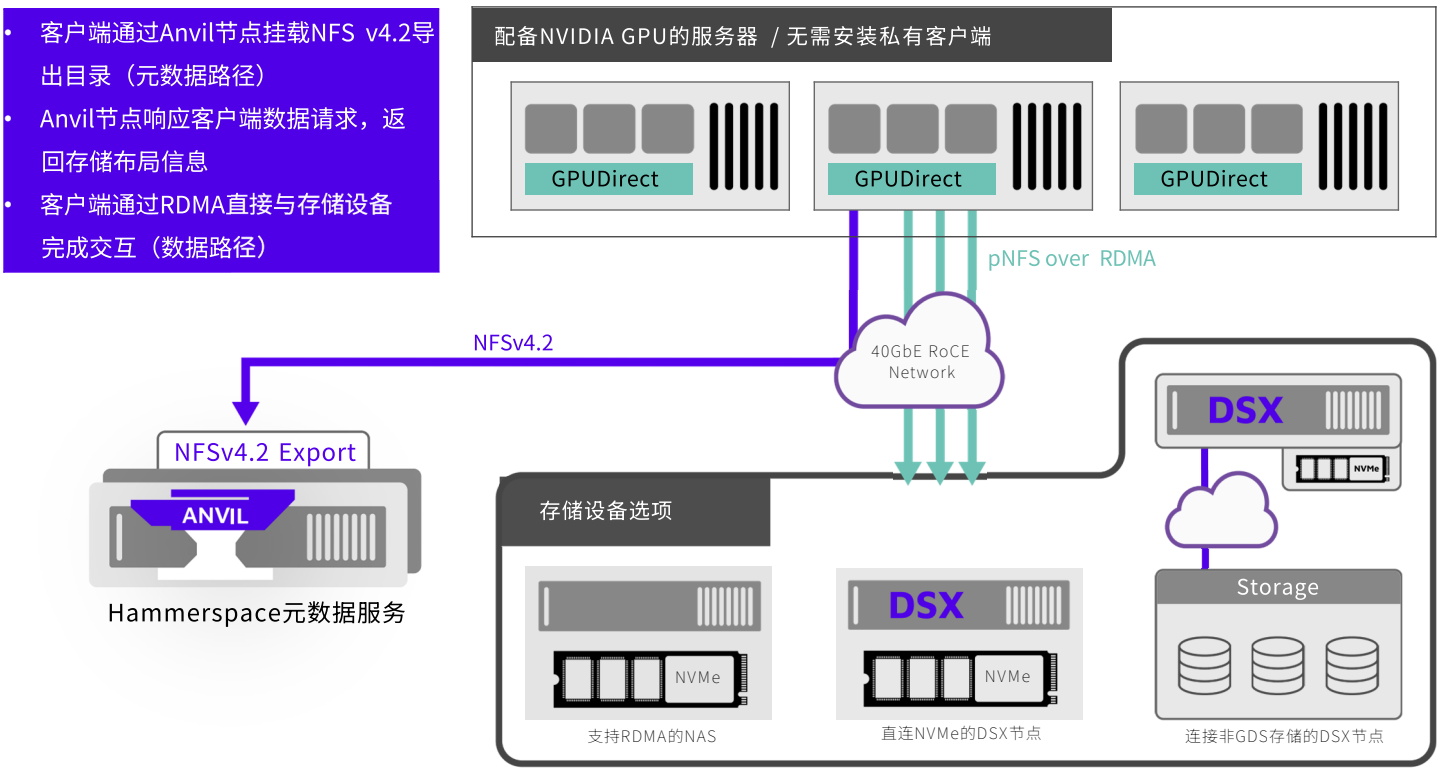



*Figure 4 - Hammerspace with GPUDirect Storage*

# 测试环境配置

如下图所示测试架构，本次功能验证在非生产环境中执行，具体配置如下：

Hammerspace集群：控制节点：单个Anvil节点+存储节点：单个DSX节点

硬件平台：超微（Supermicro）服务器

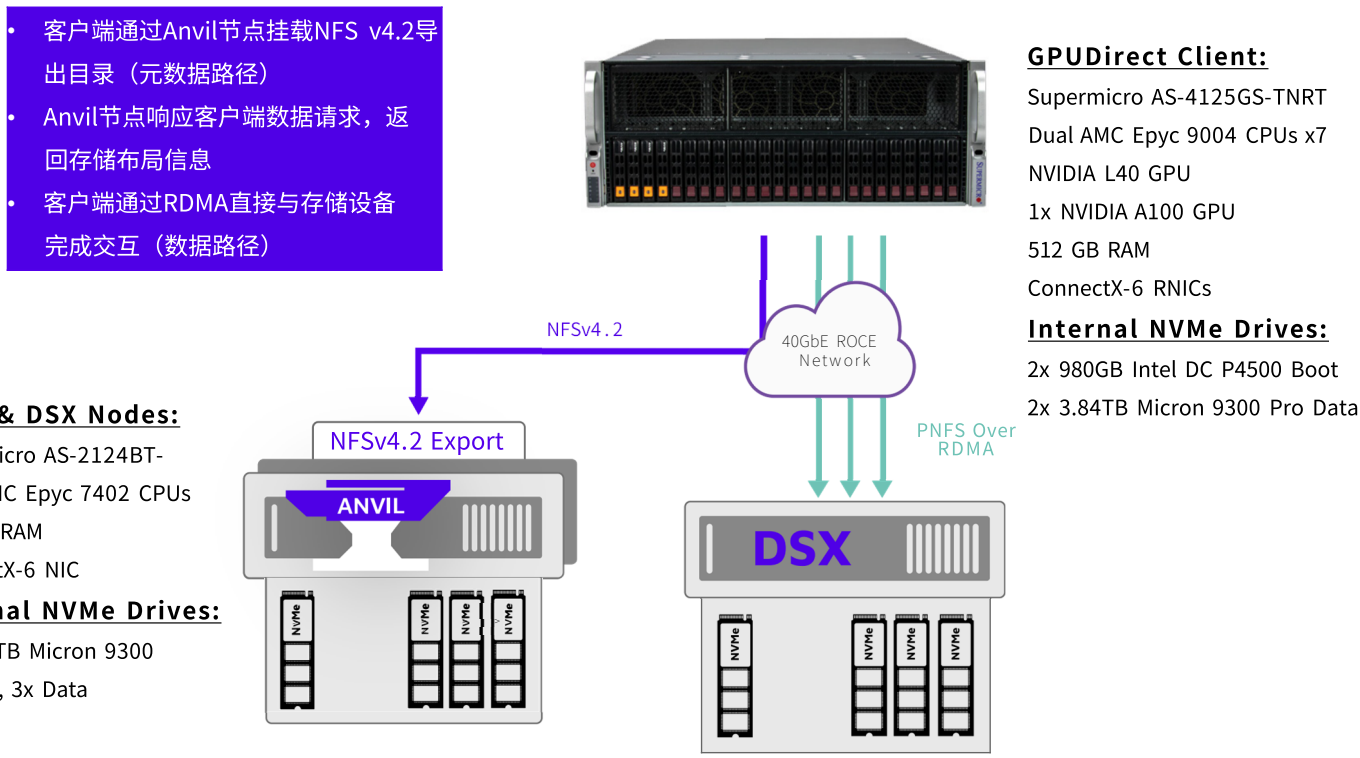
存储设备：类型：美光（Micron）NVMe驱动器，配置：1块用于系统启动，3块用于数据存储

GDS客户端：服务器：超微（Supermicro）

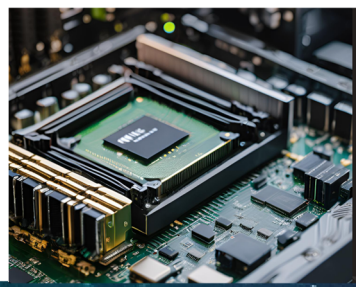
GPU配置：7块NVIDIA L40 GPU + 1块NVIDIA A100 GPU

测试用GPU：1块L40与1块A100（确保GPU与ConnectX-6 RNIC网卡处于同一PCIe交换机以最小化内部延迟）

网络连接协议：40Gb以太网




*上图- GPU直连存储测试配置*

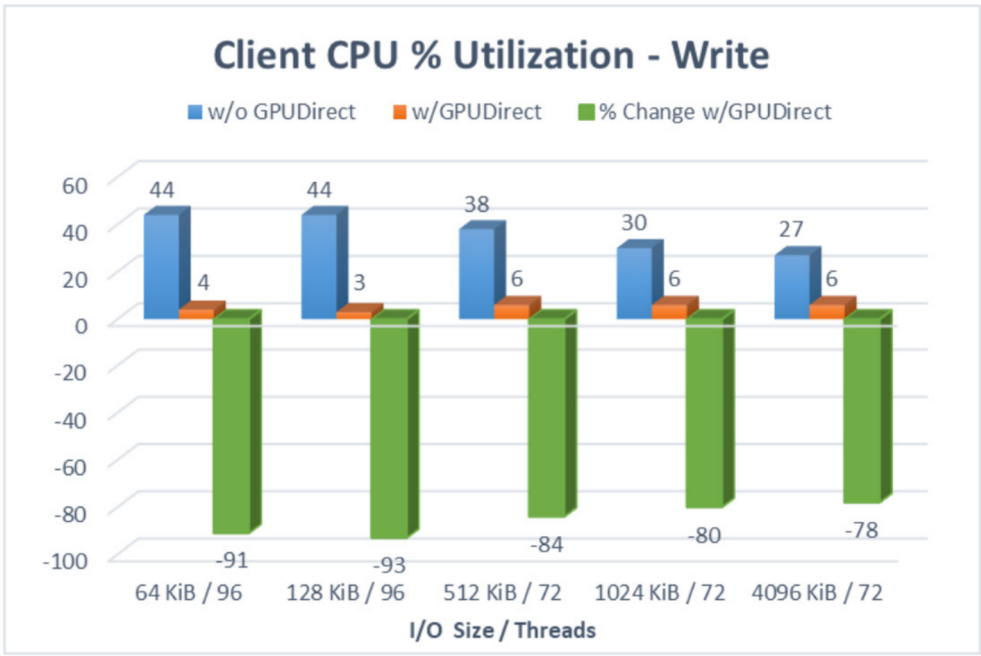


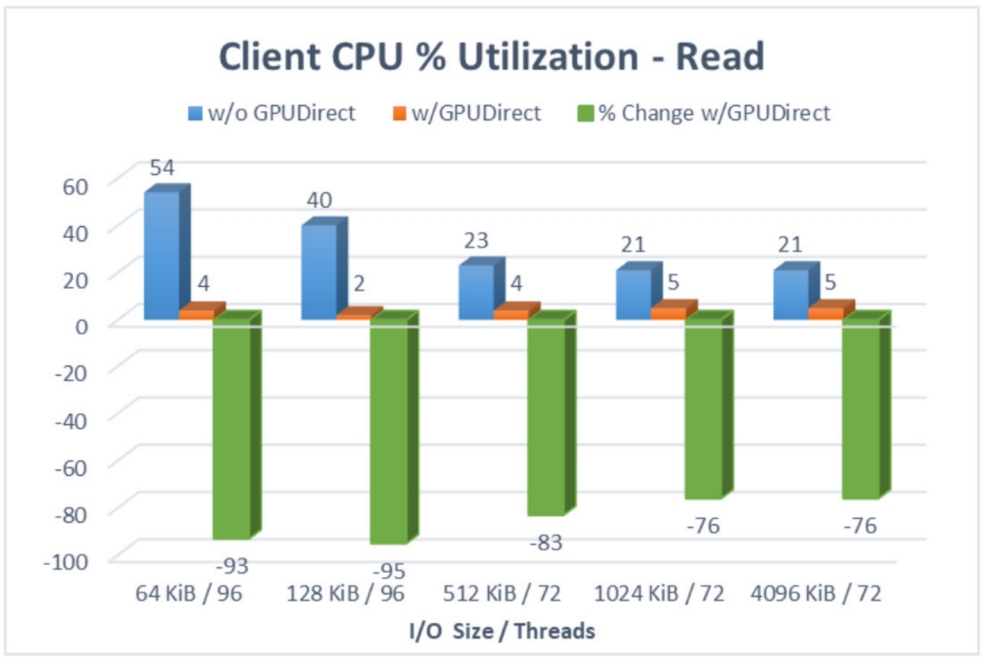
如需了解更多-可联系
ralphie.yu@hammerspace.asia
或 18810311799
手机・微信同号
HAMMERSPACE

## 测试结果

NVIDIA的GDS功能测试软件工具（版本1.8.0.36）在GPUDirect客户端运行，提供测试中的I/O操作。在读写过程中测试客户端的CPU利用率，通过多次运行测试覆盖多种I/O大小和线程数。在启用和禁用GDS的情况下运行相同测试，以验证GDS是否按预期工作，并量化其对降低CPU利用率的效果。

如下图所示，启用GDS后CPU利用率显著降低，最高降幅达95%。在未启用GDS的测试场景中，平均CPU利用率高达54%——且此时系统负载并不算特别高。可以预见，若采用更多并行数据流、更小I/O尺寸或更高速网络，存储流量可能会使CPU不堪重负，导致应用性能下降。GDS的技术优势在此场景下尤为突出。




*上图 - 客户端的CPU利用率降低百分比（启用GDS后）*

## GPU 直连存储的设计考量

尽管GDS效果显著，但它并非万能的解决方案。企业需投入资源才能充分释放其价值，且某些场景可能存在更简单的替代方案。设计考量可分为三大类：应用层、基础设施层和组织层。以下部分内容改编自《NVIDIA GPUDirect Storage设计指南》，开发者可参考该文档获取更多技术细节。

## 应用层考量

要发挥GDS优势，应用需具备特定架构特性：

- **I/O是否成为主要瓶颈？** 通常受益于GDS的应用需高频传输大量数据且计算需求较低。若计算耗时远高于数据传输耗时（例如AI推理场景），则GDS可能不适用。典型案例如AI模型训练适合GDS加速，而推理通常不适合GDS加速。
- **是否依赖CPU预处理数据？** GDS提供存储到GPU内存的直连路径。若数据需经CPU处理后再传输至GPU，则无法受益于GDS。
- **是否使用或可适配GDS所需的CUDA API**

## 基础设施层考量

- **是否部署或计划部署支持GDS的NVIDIA GPU？** GDS为NVIDIA专有技术，仅部分GPU型号支持。
- **是否具备或计划部署支持RDMA的高速网络？** 服务器需配备RDMA网卡（RNIC）及驱动。GPU数据管道的性能受制于最慢环节。
- **存储环境能否满足性能需求（IOPS/吞吐量）？** 存储可直接供GPU服务器挂载，但采用Hammerspace等并行文件系统的横向扩展NAS更具扩展性与灵活性。

## 组织层考量

- **GPU战略是否以NVIDIA为核心？** GDS需绑定NVIDIA GPU及API，与多厂商中立策略并不兼容。
- **是否有跨部门性能优化承诺？** 若基础设施由多团队分管，需协调优化全栈（应用→服务器→网络→存储）。未实现端到端调优的环境可能无法充分发挥GDS价值。

## 替代方案

若当前因时间、策略等因素暂不部署GDS，未来仍可规划适配。企业亦可立即采取其他措施优化GPU数据供给效率。

Hammerspace对客户的实施建议包括：希望保留GPU直连存储（GDS）扩展能力，应为DSX节点配备ConnectX-4（或更高版本）RNIC网卡，并在升级中遵循该网络要求。为DSX节点部署ConnectX RNIC网卡可即刻或在未来启用RDMA支持。这使得用户能够基于行业标准硬件与协议——构建具备HPC级性能的并行文件系统——无论是否启用GPU直连存储。其性能可线性扩展，仅受服务器、存储与网络硬件限制。

某Hammerspace客户正在为自身的大规模语言模型训练应用实现了：“数据规模：数十PB；传输速率：7.2TB/秒（大写TB表示字节级吞吐）；服务对象：数千块GPU。”该应用未使用GDS或RDMA，而是依赖大规模横向扩展架构，证明存在多种技术路径可确保GPU数据管道持续满载。

# Hammerspace附加优势

采用Hammerspace可释放原本需依赖专用并行文件系统才能实现的性能潜力。对于适用场景，RDMA与NVIDIA GPU直连存储可进一步加速性能。通过NFSv4.2协议挂载Hammerspace共享目录，服务器可直接并行访问多个存储目标。**得益于Hammerspace对开源社区的贡献，这种并行性能基于开放标准，所有用户均可通过Linux NFS客户端免费获得——无需购买、安装或管理专用客户端软件。**除性能优势外，Hammerspace提供跨厂商存储的全场景数据编排能力，支持本地、混合云及多云工作流。

# 结论

Hammerspace已通过NVIDIA GPU直连存储认证，为客户提供最大化GPU服务器投资回报的新选项。当GDS符合企业战略、技术能力与发展目标时，该技术是理想选择。对于不适合GDS的工作负载或场景，Hammerspace可通过全局并行文件系统提供极致存储性能，确保GPU利用率的最大化与训练周期的最短化。

## References


Hammerspace AI Solutions Page
https://hammerspace.com/artificial-intelligence/

Revolutionize AI Workloads with Hammerspace White Paper
https://hammerspace.com/how-hammerspace-optimizes-ai-workloads/

NVIDIA GPUDirect Storage Support Matrix
https://docs.nvidia.com/gpudirect-storage/release-notes/index.html#support-matrix

NVIDIA GPUDirect Storage Best Practices Guide
https://docs.nvidia.com/gpudirect-storage/best-practices-guide/index.html

NVIDIA GPUDirect Storage Installation and Troubleshooting Guide
https://docs.nvidia.com/gpudirect-storage/troubleshooting-guide/index.html

NVIDIA GPUDirect Storage Design Guide
https://docs.nvidia.com/gpudirect-storage/design-guide/index.html

NVIDIA GPUDirect Storage Benchmarking and Configuration Guide
https://docs.nvidia.com/gpudirect-storage/configuration-guide/index.html

NVIDIA CUDA-Enabled Datacenter Products
https://developer.nvidia.com/cuda-gpus