Using Hammerspace with NVIDIA GPUDirect Storage

TECHNICAL BRIEF

Summary

GPUs are the undisputed engines of modern AI development, and key players in other compute-intensive applications. A single AI training run may require processing Petabytes of data and take weeks or even months to complete. Optimizing the data path from storage to GPU is critical to maximizing GPU utilization and minimizing pipeline duration, whatever the application.

GPUDirect Storage (GDS) is NVIDIA's answer to this challenge. It leverages Remote Direct Memory Access (RDMA) technology to streamline the path between storage and GPU.

Hammerspace software is certified with GDS, enabling development teams to leverage the global view and parallel performance of the Hammerspace data platform to make any storage GPUDirect Storage, accelerate their applications, and make the most of their investment in GPUs.

GPU Direct Storage Architecture

GPUs are tailor-made for solving problems that benefit from massively parallel processing. Due to their specialized architecture, they are faster and more efficient than general-purpose CPUs for these tasks. By some estimates[1], GPUs are ~10-1000x faster, and use ~10-100x less power compared to CPUs completing the same tasks.

Because of their processing speed, it can be a challenge to provide data to GPUs fast enough to keep them running at maximum utilization. In a typical GPU-equipped server, every bit of data flowing between the storage and the GPUs must be handled by a system CPU and run through a buffer in RAM. This has several impacts:

- Additional latency from copying data into and out of the buffer, as well as buffer allocation and de-allocation activity
- Potential throughput restriction when the CPU has less PCIe bandwidth than the GPU

1 https://cset.georgetown.edu

Table of Contents

- The Challenge – Accelerating GPU Workloads
- GPUDirect Storage Architecture and Benefits with Hammerspace
- Hammerspace GPUDirect Storage Testing and Compatibility
- GPUDirect Storage Design Considerations

HAMMERSPACE

- Application throttling, as CPU cycles used to move data are not available for applications
- Memory dedicated to buffering data is not available for applications

GPUDirect Storage avoids these problems by providing a DMA path between GPUs and storage, bypassing the CPU. More CPU cycles, RAM, and PCIe bandwidth are thus available for other tasks, and storage performance (both latency and throughput) is improved.

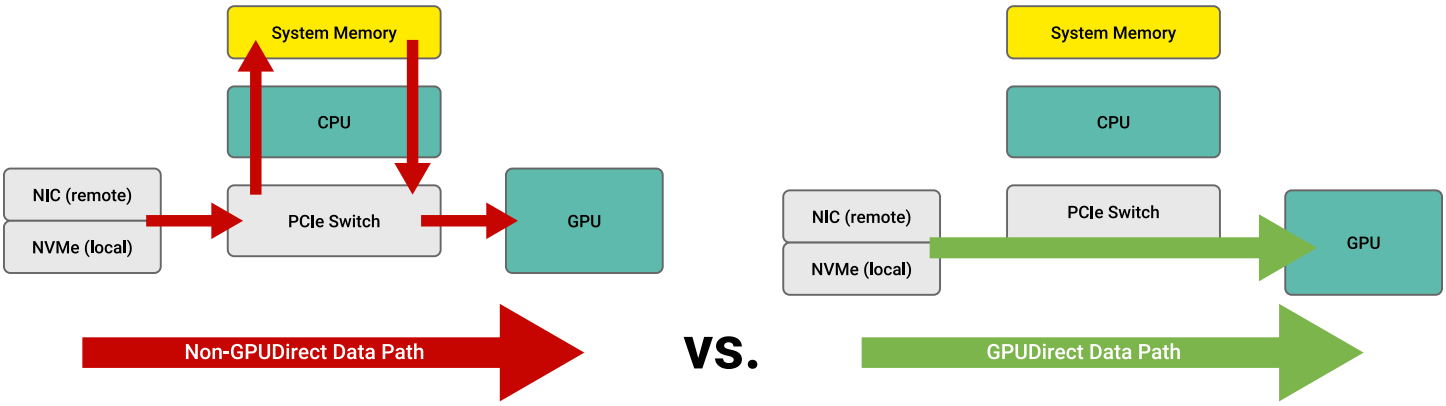


Figure 1 - GPUDirect Data Path

GPU Direct Storage and NAS

Note in Figure 1 above that storage utilized by GDS may consist of locally attached NVMe storage devices, or remote storage connected over the network. For network attached storage, RDMA over a lossless network provides the shortest and most efficient data path, since RDMA provides the benefits of CPU offload and lower latency on the storage server as well.

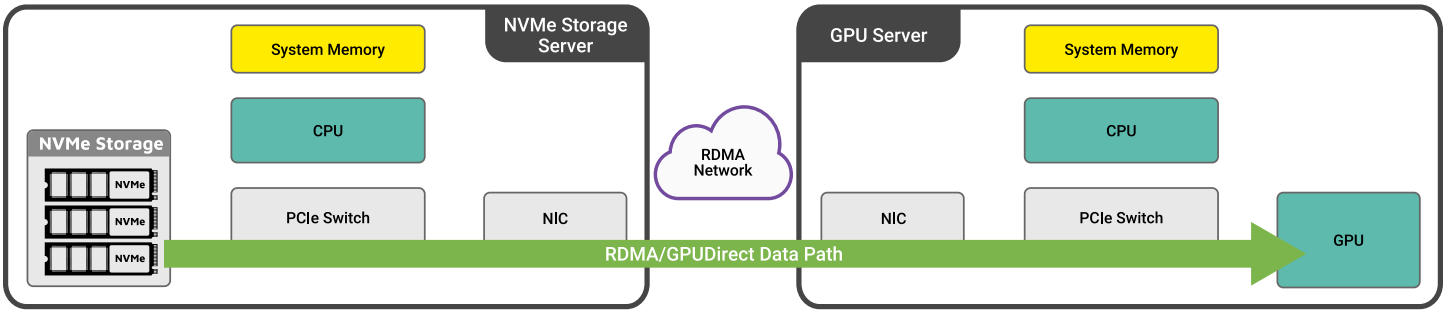


Figure 2 - GPUDirect with Remote Storage

Hammerspace Testing with GPU Direct Storage

Recently Hammerspace conducted functional testing with GDS in a Hammerspace global data environment. The purpose of this lab testing was to verify functionality and compatibility only, not characterize performance, so a minimal system was used. Performance testing remains a project for the future.

Hammerspace Architecture Overview

In a Hammerspace cluster, Anvil nodes (typically two in a production implementation) are the metadata servers, providing control plane functions only. No data I/O flows through Anvil nodes. DSX data services nodes can provide access to block, file, and object storage and handle all I/O operations, including advanced functions such as data orchestration. DSX nodes may be scaled out as needed. If storage is RDMA capable it does not need to be located behind a DSX node but may be accessed by clients directly at the direction of the Anvil.

Together the Anvil and DSX nodes and storage operate as a single parallel scale-out cluster. For complete details, refer to the Hammerspace Technology White Paper.

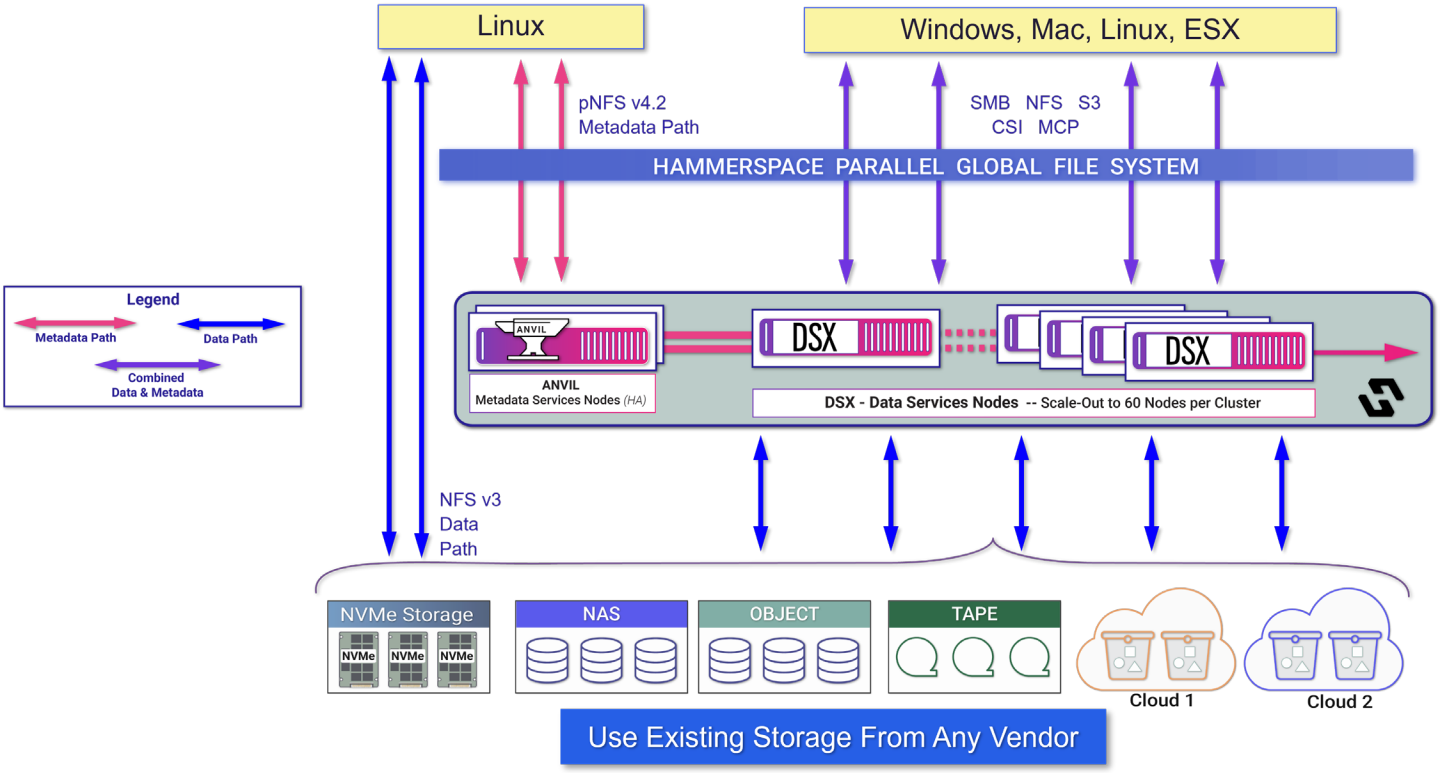


Figure 3 - Hammerspace High-Level Architecture

Hammerspace with GPUDirect Storage

Note that there are no GDS-specific requirements or configuration on the Hammerspace systems beyond what is required for RDMA. Hammerspace automatically detects ConnectX-4 (and greater) RNICs and enables RDMA. Hammerspace will then respond to NFSv4.2 client requests by offering both RDMA and standard data layouts, with RDMA being preferred. The Hammerspace RDMA Tech Brief has more information on leveraging RDMA with Hammerspace.

There are multiple storage configurations that may be used with Hammerspace and GDS, illustrated in Figure 4 below:

- Any RDMA-capable NFSv3 NAS storage
- Any block storage (typically NVMe flash) attached to a DSX node
- Any non-RDMA NAS storage fronted by a DSX node

Hammerspace can make any storage available for use with GDS, even storage that does not work on its own due to a lack of GDS or RDMA support. The Portal function of a DSX may be used as a direct passthrough to non-RDMA storage, but this will not provide the best performance. The preferred method is to have some NVMe storage within the DSX node and use objectives or volume groups to promote the required data from the non-GDS capable storage into the DSX's local NVMe storage. The DSX NVMe storage then acts as a fast cache, accelerating access to the non-GDS storage.

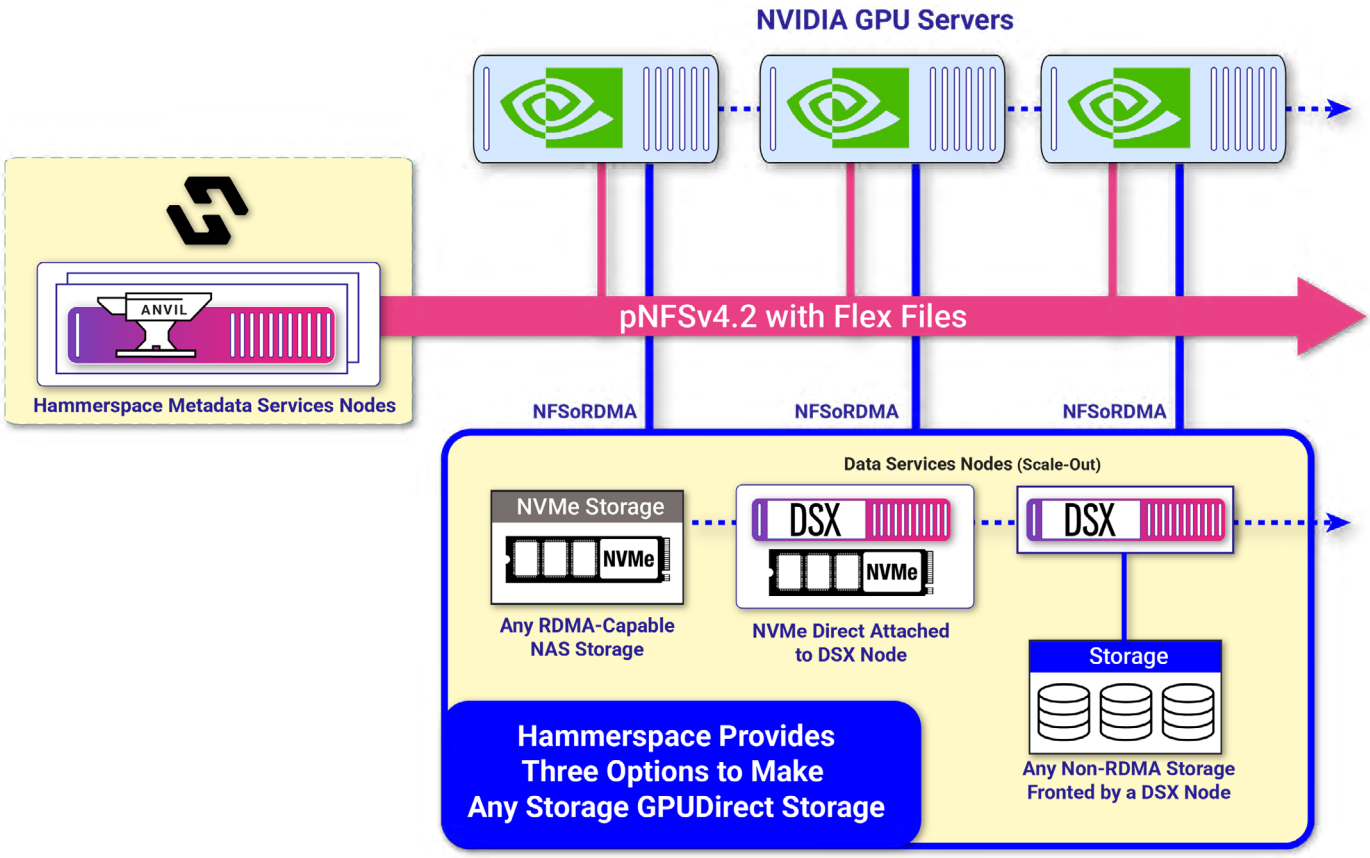


Figure 4 - Hammerspace with GPUDirect Storage

Test Configuration

Figure 5 below illustrates the configuration used for testing. Since this was a functional test in a non-production environment, the Hammerspace cluster consisted of a single Anvil node and a single DSX node, both running on Supermicro hardware. Storage was provided by Micron NVMe drives (one for boot, three for data). The GPUDirect Storage client was a Supermicro server with 7x NVIDIA L40 GPUs and a single NVIDIA A100 GPU. Testing was performed with one of the L40 GPUs and the A100 GPU, making sure the GPU in use was on the same PCIe switch as the ConnectX-6 RNIC to minimize internal latency. Components were connected with 40Gb Ethernet.

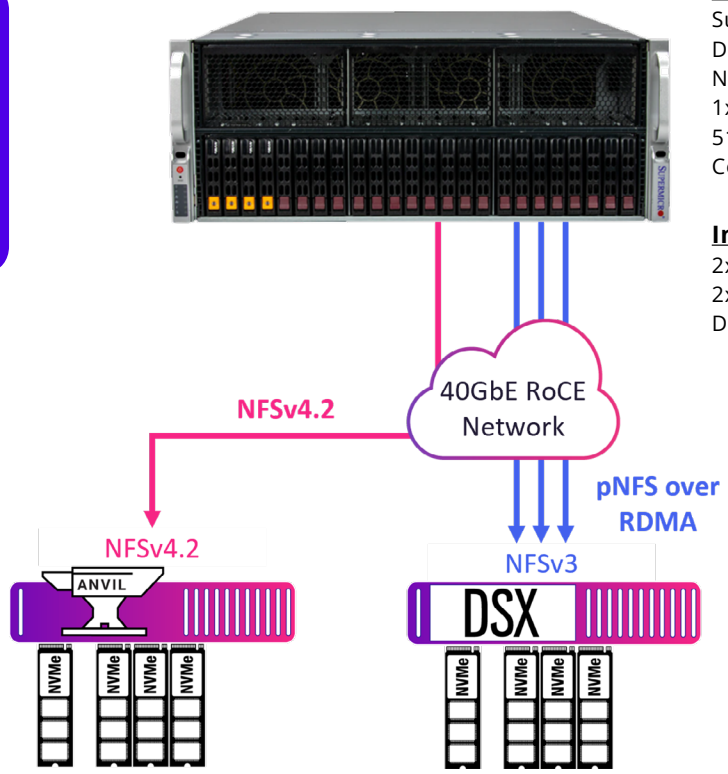
Client connects to Anvil, mounts NFSv4.2 export (metadata path)

Anvil responds to client data requests with storage layouts

Client interacts directly with DSX storage via NFS over RDMA(data path)

Anvil & DSX Nodes:
Supermicro AS-2124BT-HNTR
Dual AMC Epyc 7402 CPUs
512 GB RAM
ConnectX-6 NIC

Internal NVMe Drives:
4x 3.84TB Micron 9300 Pro
1x Boot, 3x Data



GPUDirect Client:
Supermicro AS-4125GS-TNRT
Dual AMC Epyc 9004 CPUs x7
NVIDIA L40 GPU
1x NVIDIA A100 GPU
512 GB RAM
ConnectX-6 RNICs

Internal NVMe Drives:
2x 980GB Intel DC P4500 Boot
2x 3.84TB Micron 9300 Pro Data

Figure 5 - GPUDirect Storage Test Configuration

Test Results

NVIDIA's GDS Functional Test Software tool version 1.8.0.36 was run on the GPUDirect client to drive I/O for the tests. CPU utilization on the client was measured during both reads and writes, with several runs using a variety of I/O sizes and thread counts. Identical tests were run both with GDS enabled and disabled to confirm that GDS was working as expected and to measure the benefit in reduced CPU utilization.

As the graphs in Figure 5 clearly show, enabling GDS resulted in a dramatic reduction in CPU utilization, up to 95%. In the non-GDS test scenarios, the average CPU utilization was as high as 54%, and not under a particularly heavy load. One can easily imagine with more parallel streams, smaller I/O sizes, and faster networking, it would be possible to swamp the CPU with storage traffic, causing application execution to suffer. The benefit of GDS is very apparent.

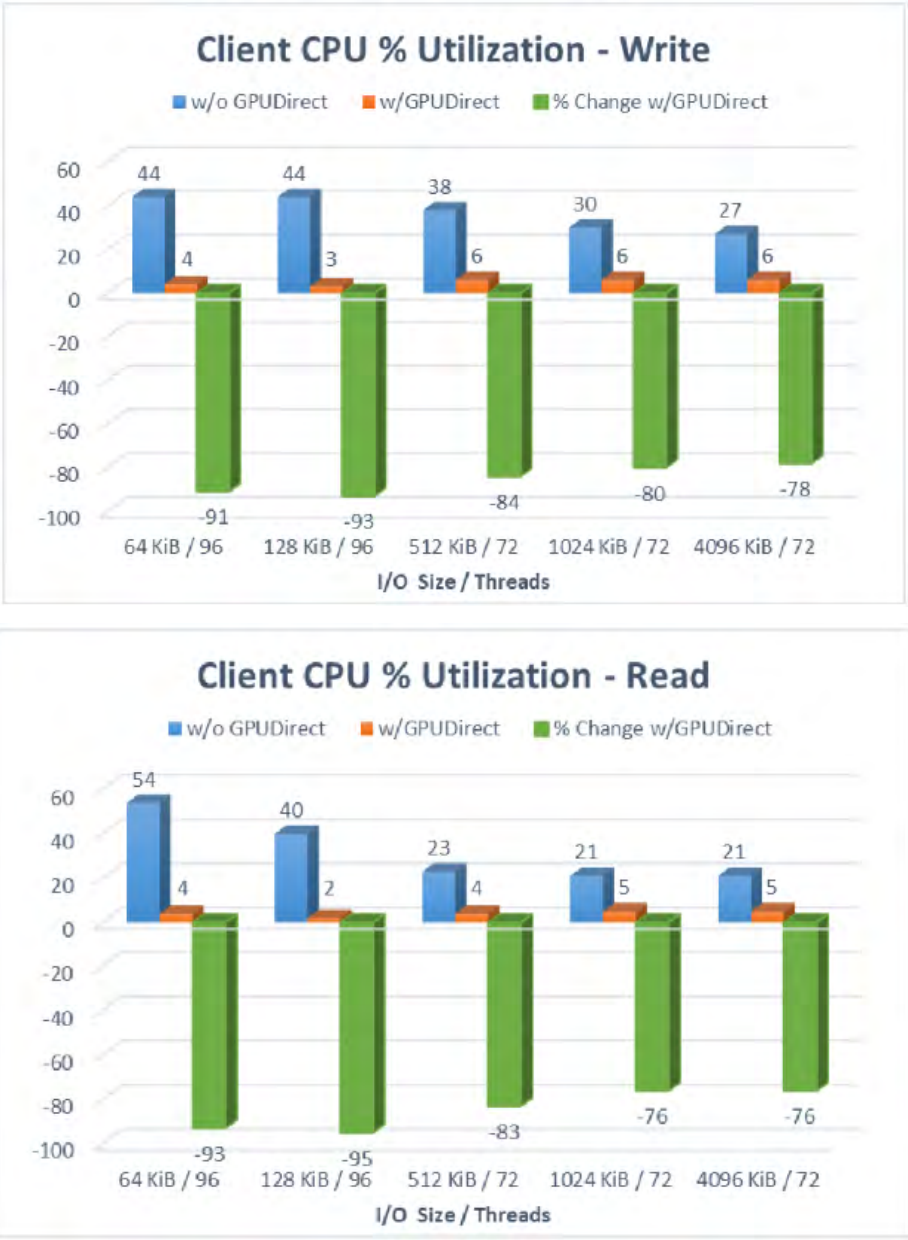


Figure 6 - Client CPU % Utilization Reduction w/GDS

GPU Direct Design Considerations

As impressive as GDS can be, it is not magic. It will take commitment for an organization to experience the benefits, and for some use cases there may be simpler alternatives. Considerations may be broadly grouped into three categories: Application, Infrastructure, and Organization. Some of the points below are paraphrased from the NVIDIA GPUDirect Storage Design Guide, which contains additional valuable detail for application developers.

Application Considerations

To benefit from GDS, applications must have specific characteristics and be architected in a way that enables them to take advantage of GDS.

- Is I/O a significant bottleneck? Typically, workloads that benefit from GDS stream large amounts of data and perform small amounts of computation on the data. On the other hand, if time spent on computation is far greater than time spent moving data, GDS is probably not a fit. A good example is AI model training vs. inferencing. Training requires large quantities of data and may be an excellent fit for GDS acceleration, while inferencing requires less data and more computation, and is usually not a fit.
- Does the application perform CPU-based pre-processing on the data? Remember, GDS provides a direct path from storage to GPU memory. Applications that require data to be processed through the CPU before being handed to a GPU will see no benefit from GDS.
- Does the application use, or can it be re-written to use, the CUDA APIs required for GDS?

Infrastructure Considerations

- Are NVIDIA GPUs that support GDS in use or planned? GDS is an NVIDIA-proprietary technology supported by a subset of their GPUs. Other GPUs won't work.
- Is high-speed networking that supports RDMA in place or planned? Are servers equipped with RDMA-capable NICs (RNICs) and drivers? Like anything, a GPU data pipeline is only as fast as its slowest link.
- Can the storage environment deliver the required performance, both IOPS and throughput? Storage devices may be directly attached to GPU-accelerated servers, but scale-out NAS using a high-performance parallel file system like Hammerspace provides superior scalability and flexibility.

Organizational Considerations

- Does the GPU strategy center around NVIDIA? Using GDS requires NVIDIA GPUs and APIs. If a vendor-neutral GPU strategy is desired, GDS is not a fit.
- Is there a cross-functional commitment to performance? If infrastructure responsibilities are spread across multiple teams, it will take coordination to tune the entire stack, from application to server to network to storage. GDS benefits may not be realized, and will not be maximized, if the environment isn't tuned end-to-end.

Alternatives

Just because the time, situation, or strategy aren't right for an organization to deploy GDS today doesn't mean it can't be accommodated in the future, or that nothing can be done to enhance data delivery to GPUs right now. Hammerspace customers who wish to leave the door open to GDS should deploy their DSX nodes with ConnectX-4 (or later generation) RNICs, and keep in mind the network infrastructure considerations above during upgrade cycles.

Equipping DSX nodes with ConnectX RNICs enables the use of RDMA with Hammerspace immediately or in the future. This makes it possible to create a parallel file system that delivers supercomputing-class performance using industry standard hardware and protocols – with or without GPUDirect Storage. Performance scales linearly, limited only by the available servers, storage, and networks.

Hammerspace customer Meta is delivering tens of Petabytes of data to thousands of GPUs at 7.2TB/s (yes, big 'B' Bytes) for a large language model training application. This application does not use GDS or RDMA, instead relying on massively parallel scale-out approach, demonstrating that there is more than one way to ensure the GPU data pipeline is always kept full.

Additional Benefits with Hammerspace

Using Hammerspace unlocks performance potential at levels that formerly required exotic parallel file systems. For the right workloads, RDMA and NVIDIA GPUDirect Storage will increase performance even further. By mounting Hammerspace shares with NFSv4.2, servers may access multiple storage targets, directly and in parallel. Thanks to Hammerspace's contributions to the open-source community, this parallel performance is standards-based and available to everyone through the Linux NFS client. No special client software needs to be purchased, installed, or managed.

In addition to performance, Hammerspace brings broad data orchestration capabilities that enable on-premises, hybrid and multi-cloud workflows across any type of storage from any manufacturer.

Conclusion

Hammerspace is now certified with NVIDIA GPUDirect Storage, providing customers with another option to derive the maximum benefit from their investment in GPU-accelerated servers. GDS is an excellent choice when it aligns with the strategy, capabilities, and goals of the organization. For workloads or situations where GDS is not a fit, Hammerspace's ability to deliver extreme storage performance through a global parallel file system assures maximum GPU utilization and minimum pipeline runtimes.

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