SOLIDIGM™

Solution Brief

Solidigm™ SSDs for NVIDIA and VAST Data SuperPOD™ Solutions

As artificial intelligence rapidly advances to fuel humanity's aspirations, computing power has had to grow as well. Clusters of thousands of GPUs are emerging everywhere, powered by high-throughput, low-latency networks and deep learning models. NVIDIA's DGX SuperPOD™ provides a reference architecture for data center architecture for AI that supports high-performance needs.

This fast-growing segment prompts deep contemplation among AI architects. With a high-performance architecture like SuperPOD, NVIDIA came up with Magnum IO GPUDirect® Storage which allows for a direct data path between GPU and local or remote NVMe/NVMe-oF storage.

One of the most important questions is this: What kind of storage devices can keep AI accelerators (GPUs, CPUs, and others) and network devices running at full capacity without idle time? Table 1 summarizes each phase of the AI project cycle and their respective I/O characteristics and consequent storage requirements.

Impact of AI Project Cycle Phases			
AI Phase	I/O Characteristics	Storage Requirements	Impact
Data ingestion and preparation	Reading data randomly; writing preprocessed items sequentially	Low latency for small random reads; high sequential write throughput	Optimized storage means the pipeline can offer more data for training, which leads to more accurate models
Model development (training)	Random data reads	Scalability in multi-job performance and capacity; optimized random reads; high sequential write performance for checkpointing	Optimized storage improves utilization of expensive training resources (GPU, TPU, CPU)
Model deployment (inference)	Mixed random reads and writes	Self-healing capabilities to handle component failures; non-disruptive expansion and upgrades; same features as training stage if the model undergoes ongoing fine-tuning	Business requires high availability, serviceability, and reliability

Table 1. I/O characteristics and consequent storage requirements of AI project cycle phases

NVIDIA GPUDirect Storage for storage scalability

Recognizing the importance of storage to AI and increasing sizes of the datasets, NVIDIA is driving several initiatives to accelerate storage performance and capacity scaling. This section will describe these initiatives and the impact they have on storage requirements.

1. Network File System over RoCE and NVIDIA GPUDirect Storage

NVIDIA has extended NFS over RDMA to enable GPUDirect Storage (GDS). As shown in Figure 1, GDS accelerates GPU-enabled workloads by bypassing the CPU and main memory altogether, using RDMA to transfer data between the storage system and GPU memory directly.

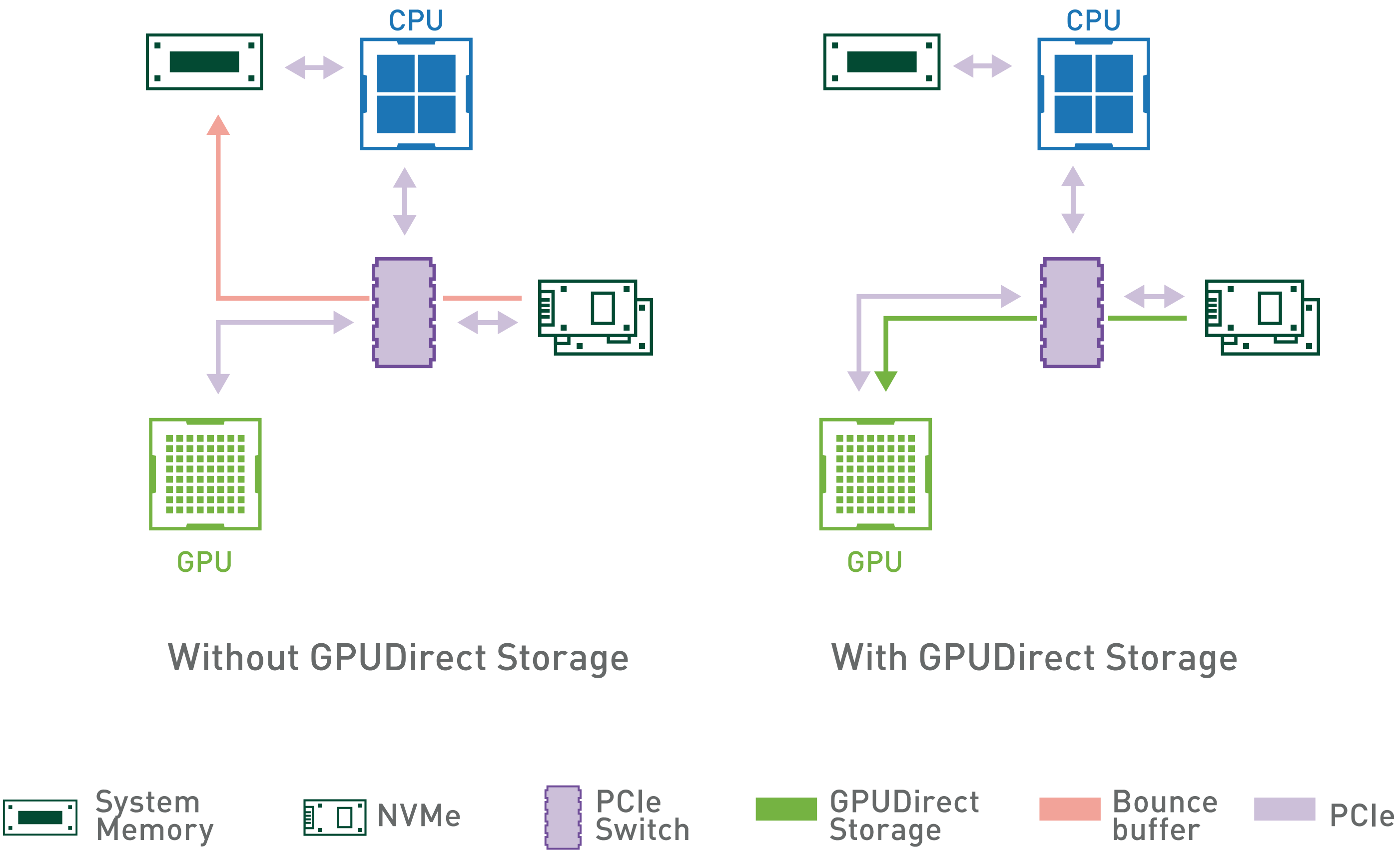


Figure 1. NVDIA GPUDirect Storage[1] Copyrighted image used under permission by NVIDIA

While the PCIe specification includes provisions for P2P (peer-to-peer) communication, practical implementations beyond FPGA solutions have been relatively rare in the real world. In the absence of GPUDirect Storage (GDS), GPUs accessing NVMe storage typically required data to be copied to system memory (bounce buffer), albeit with the aid of Direct Memory Access (DMA) to alleviate CPU involvement. GDS eliminates the need for a bounce buffer. When GPUs access NVMe storage with GDS, data doesn't need to be copied to system memory. This eliminates consumption on the memory bus and memory space, offering significant optimizations for performance and memory resource utilization.

2. Choosing memory management

In an excellent memory experiment reported by NVIDIA,[2] cuDF CSV reader's original implementation (shown by the green line at the bottom of Figure 2) experienced faults during data movement from system memory to the GPU, data transfer from storage to memory, and unpinned data movement through a CPU bounce buffer. The revised bounce buffer implementation, now included with RAPIDS (shown by the yellow line), utilizes the best-available memory management and explicit data movement. Reading data from a preloaded page cache is depicted in dotted red. GPUDirect Storage, shown by the blue, outperforms all of these methods, limited only by NVMe drive speeds.

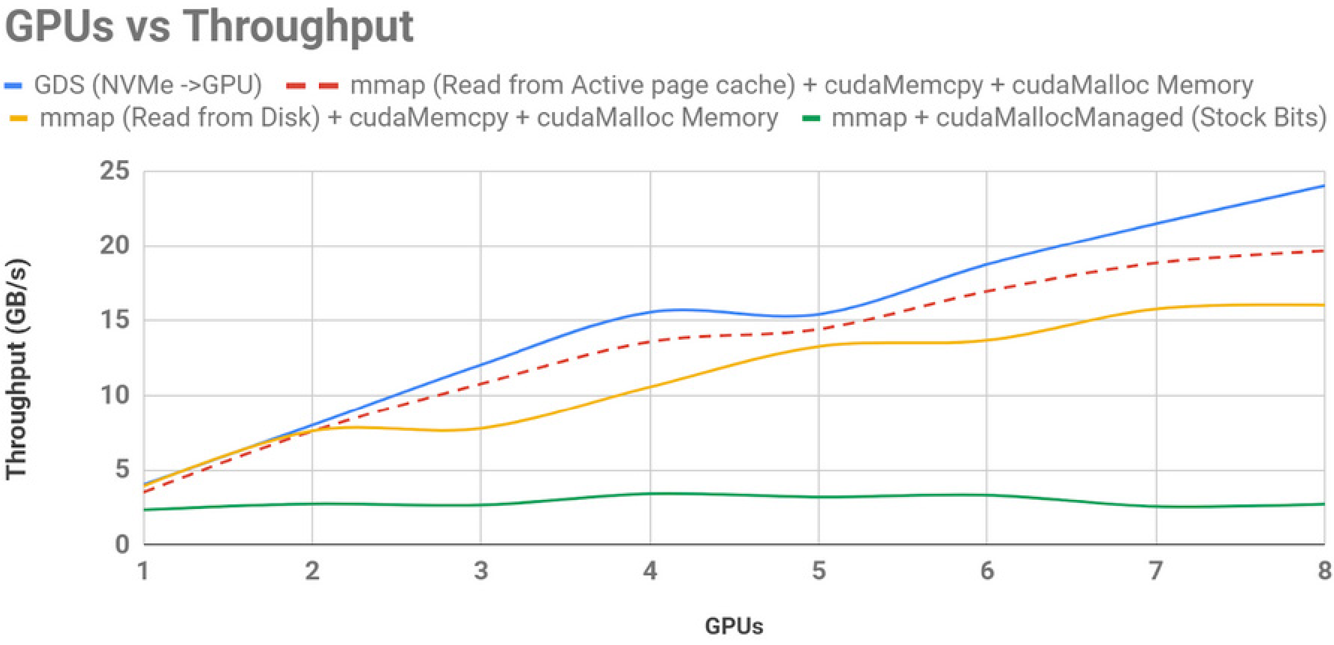


Figure 2. NVDIA GDS benefits[3] Copyrighted image used under permission by NVIDIA

In practical terms, the choice between NFSoRDMA and GPUDirect Storage depends on specific data processing requirements. If data needs to be preprocessed by the CPU before reaching the GPU (e.g., for tasks like decompression), NFSoRDMA is a reliable choice. On the other hand, if data from shared storage can be directly processed by the GPU, or if GPU-processed data can be stored in shared storage without CPU intervention (e.g., checkpoint operations), GPUDirect Storage becomes an exciting implementation.

Due to the superior capabilities of GPUDirect Storage, its ecosystem continues to expand. NVIDIA's official website lists nine categories of file systems, which are more than sufficient to meet diverse needs:

1. DDN EXAScaler
2. WekaFS
3. VAST NFSoRDMA
4. EXT4 via NVMe or NVMoF drivers from MLNX_OFED
5. IBM Spectrum Scale (GPFS)
6. DELL Technologies PowerScale
7. NetApp/SFW/BeeGFS
8. NetApp/NFS
9. HPE Cray ClusterStor Lustre.

These options cover a wide range of requirements and scenarios, ensuring flexibility and performance for various storage needs in AI and GPU-intensive computing environments

NVIDIA SuperPOD reference architecture and Solidigm QLC

The NVIDIA DGX SuperPOD represents a cutting-edge AI data center infrastructure, featuring powerful NVIDIA DGX A100 and H100 systems. The term "POD" in DGX SuperPOD stands for "point of delivery," and the DGX A100 system is recognized as the world's first AI system capable of delivering a remarkable 5 petaflops of computational power.[4]

This reference architecture design introduces the following key components:

- **Powerful nodes:** These nodes are equipped with numerous GPUs, a substantial memory capacity, and high-speed interconnections between the GPUs.
- **Low-latency, high-bandwidth interconnect:** The system incorporates a fat tree HDR InfiniBand interconnect, ensuring minimal latency and maximum data transfer rates.
- **Storage hierarchy:** As shown in Figure 3, the architecture includes a storage hierarchy designed to deliver peak performance for various dataset structures, catering to diverse AI workloads.

The storage fabric is specifically engineered to provide high-throughput access to shared storage, offering the following capabilities:

- **Single node bandwidth:** It offers individual node bandwidth exceeding 40 GBps, enabling rapid data access.
- **RDMA communication:** Leveraging RDMA (Remote Direct Memory Access) communications, it achieves fast, low-latency data movement, which is essential for AI workloads.
- **Peak I/O performance:** The system can accommodate the training of deep learning models that demand peak I/O performance, surpassing 16 GBps (2 GBps per GPU) directly from remote storage.

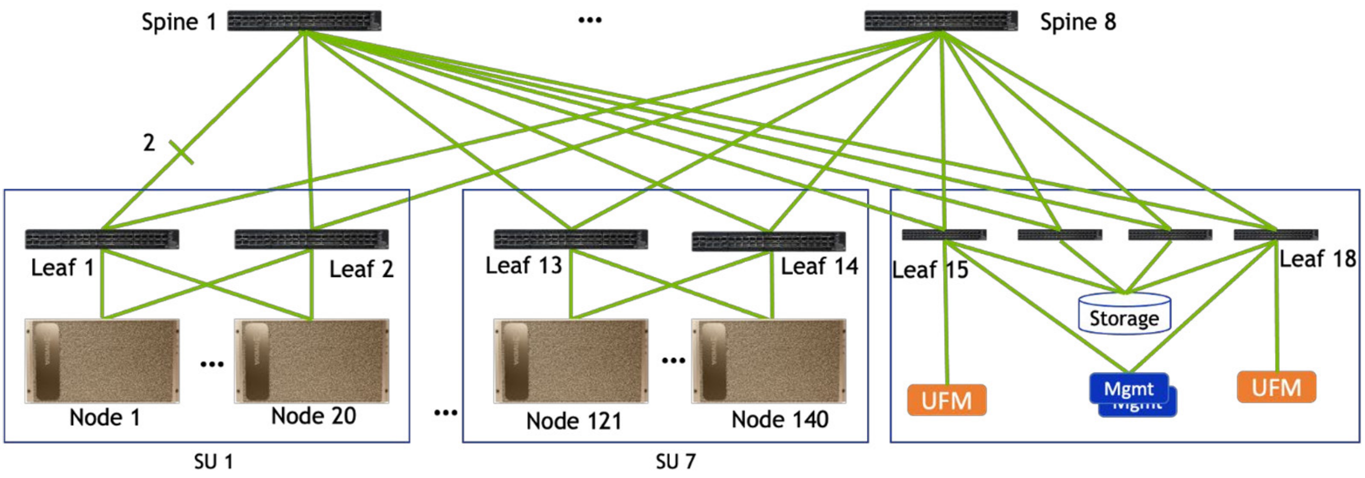
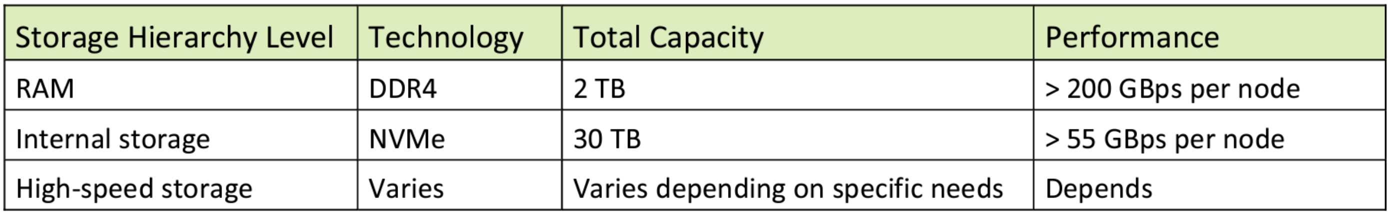


Figure 3. NVIDIA DGX A100 Reference Architecture Storage Fabric Topology[5] Copyrighted image used under permission by NVIDIA

1. Storage architecture

Source code, container definition scripts, data processing scripts, and other critical files including training datasets need to be stored. An optimized platform for critical files should use a highly available NFS appliance with a DGX SuperPOD configuration. The appliance will host the home file system for the users and provide a place for administrators to store any data necessary for managing the system.

Training performance can be limited by the rate at which data can be read and reread from storage. The key to performance is the ability to read data multiple times. The closer the data is cached to the GPU, the faster it can be read. See Table 2. Storage architecture considers the hierarchy of different storage technologies to balance the needs of performance, capacity, and cost.

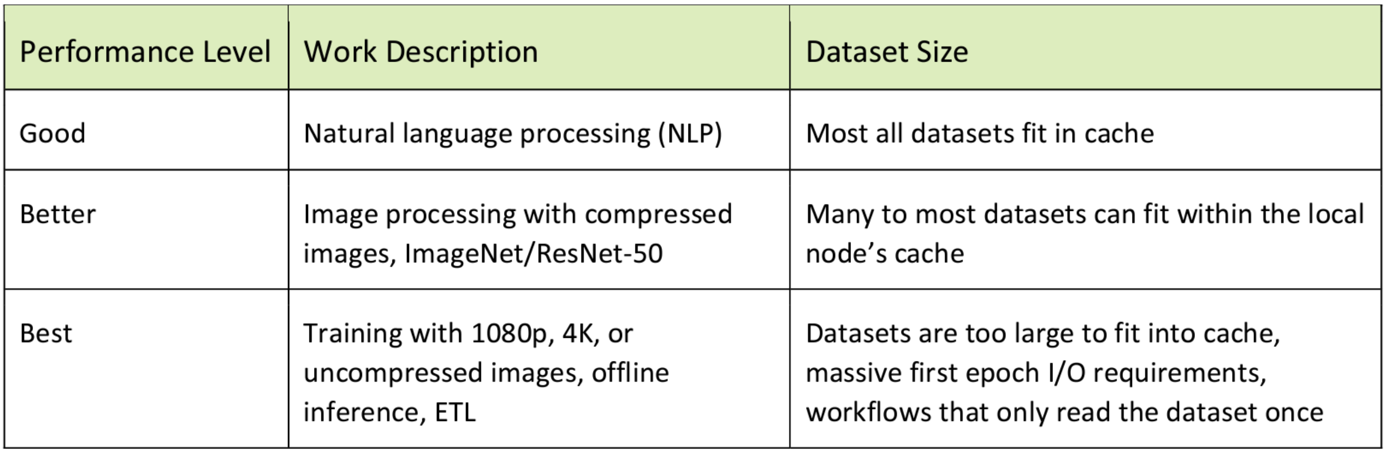


Storage Hierarchy Level	Technology	Total Capacity	Performance
RAM	DDR4	2 TB	> 200 GBps per node
Internal storage	NVMe	30 TB	> 55 GBps per node
High-speed storage	Varies	Varies depending on specific needs	Depends

Table 2. DGX SuperPOD storage and caching hierarchy[6] Copyrighted image used under permission by NVIDIA

Caching data in local RAM provides the best performance for reads. This caching is transparent after the data is read from the file system. But the size of RAM is limited and this approach is less cost-effective than others. Local NVMe storage is a more affordable way to provide caching close to the GPUs. However, manually replicating datasets to the local disk can be tedious. Furthermore, considering modern AI development involves teamwork and not individual delivery, network storage is a must in most cases.

The concept of layered or hierarchical storage has its roots in the computer model proposed by John von Neumann, known as the "stored-program" computer model. Within the DGX SuperPOD environment, the storage performance requirements are detailed in Table 3.



Performance Level	Work Description	Dataset Size
Good	Natural language processing (NLP)	Most all datasets fit in cache
Better	Image processing with compressed images, ImageNet/ResNet-50	Many to most datasets can fit within the local node's cache
Best	Training with 1080p, 4K, or uncompressed images, offline inference, ETL	Datasets are too large to fit into cache, massive first epoch I/O requirements, workflows that only read the dataset once

Table 3. Storage performance requirement[7] Copyrighted image used under permission by NVIDIA

High-speed storage serves as a shared repository of an organization's data, offering accessibility to all nodes within the system. It must be finely tuned to handle small, random I/O patterns efficiently, providing both high peak node performance and robust aggregate filesystem performance to accommodate a diverse range of workloads that an organization may encounter. High-speed storage should be capable of facilitating efficient multithreaded reads and writes from a single system, although most deep learning (DL) workloads tend to be read-dominant.

A high-performance network file storage solution becomes a game changer when data sets grow beyond the capacity of local caches, when there is an overwhelming demand for first epoch I/O, when multiple training tasks run concurrently, or when numerous data scientists collaborate.

2. Storage evolution to all flash

The evolution of high-performance network file storage solutions is enabled by innovations in network technology, more sophisticated data protection, and storage devices. Storage device innovation can be facilitated by upgrading from HDDs to TLC SSDs, and then one more step forward to SLC/TLC + QLC SSDs. Once performance and the reliability of mechanical hard drives with moving parts are considered, they become less appealing. Research by USENIX found that the average annual return rate for NAND solid-state drives falls between 0.07% and nearly 1.2%, while mechanical hard drives range from 2% to 9%.[8]

Regarding performance, the typical sequential read speed of a QLC SSD is about 28 times the sustained transfer rate of an HDD, while random read performance can be 4,700 times that of an HDD. The average latency of QLC (read/write) is 3% to 4% of an HDD. Considering these factors, it's apparent HDD-based network storage has the risk of being a bottleneck in AI development impacting XPU utilization.

Opting for an all-flash solution, particularly one built with TLC NAND flash memory, represents the current mainstream choice. However, TLC NAND flash has limitations in terms of cost optimization, and off-the-shelf TLC offerings typically max out at 15TB. This is where QLC NAND flash comes into play.

3. Benefits of QLC storage

As compared to TLC, QLC offers similar attributes while bringing its own unique characteristics to the table. In terms of quality, this includes factors like mean time between failures (MTBF) and annualized failure rate (AFR). In terms of reliability, it encompasses aspects like Uncorrectable Bit Error Rate (UBER) and data retention. Regarding environmental factors, it considers operating vibrations and temperature ranges. QLC data center drives are on par with TLC drives in all these respects.

More importantly, QLC provides equivalent read performance to TLC. Where QLC excels is in storage density and price per TB. Even in the more traditional U.2 form factor, commercial QLC drives already offer capacities of up to 60TB, as seen in products like the Solidigm D5-P5336. This translates to 2 to 4 times the capacity of mainstream TLC drives. Additionally, QLC delivers a price advantage per TB compared to TLC.

To harness the strengths of both high-performance NAND (SLC/TLC) and high-capacity QLC, organizations can employ a solution that uses SLC/TLC as a write cache and metadata storage space, benefiting from its superior write performance and longevity to handle the write-intensive phase of AI training. Data can then be asynchronously synchronized or transferred to the larger-capacity QLC volume layer. For AI training read operations, it's possible to bypass SLC/TLC and directly access data from the QLC volume layer, achieving read performance on par with a TLC-only solution for the read-intensive AI training phase.

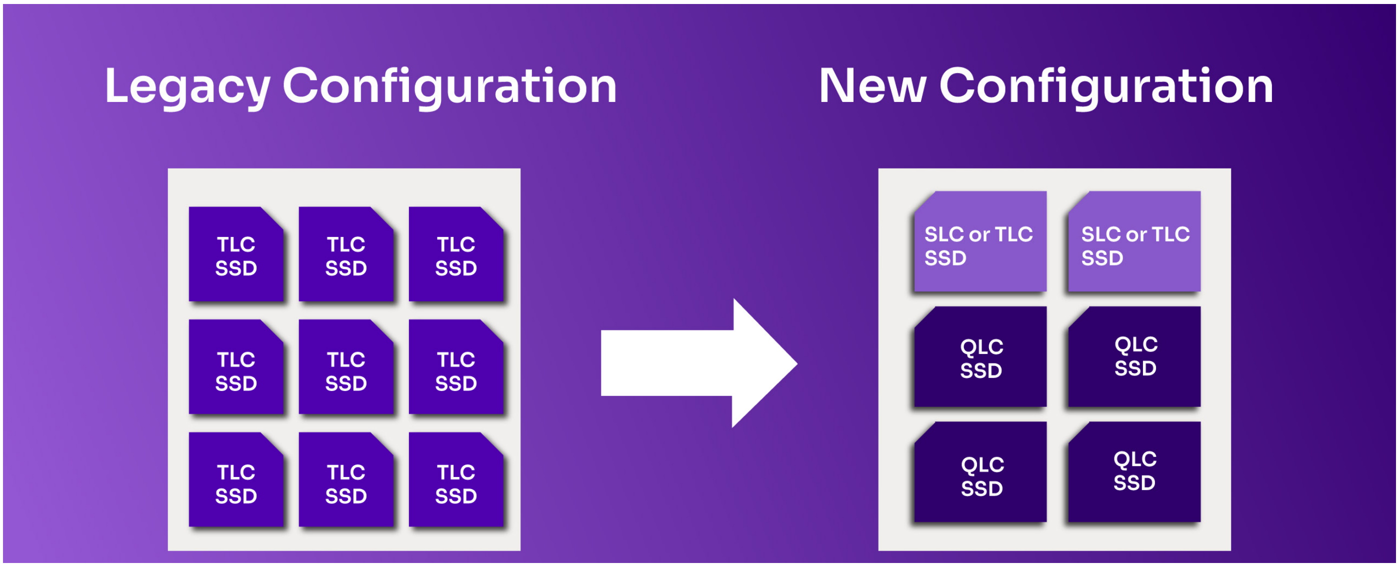


Figure 4: TLC-only solution transitions to SLC/TLC+QLC

4. QLC-based cache solution

The industry already boasts highly mature cache software solutions such as bcache (which is part of Linux kernel) and openCAS.[9] These widely deployed cache software solutions, however, have not delved deeply into the need to aggregate multiple small random write operations into a single large sequential write. This need existed even before the emergence of QLC technology but has largely gone unnoticed in the industry.

With added capacity, a large amount of DRAM is needed to maintain the FTL. QLC manages the challenge effectively by increasing the indirection unit (IU) sizes to 16KB or 64KB, exceeding the default 4KB of traditional file systems. This solution not only reduces the amount of DRAM inside the SSD but also allows for it to meet power budgets and reduce cost.

5. Cloud Storage Acceleration Layer (CSAL) software

While QLC SSDs can accommodate any read-intensive workload without modification, software innovations can allow expansion of QLC's use cases to random mixed-access workloads and can enhance the endurance of the drives significantly. Deploying the Solidigm Cloud Storage Acceleration Layer (CSAL) is one way.[10] CSAL has a native feature that place several smaller writes into the cache layer, then merges those writes into a big sequential write and moves them to the QLC capacity layer, improving time-to-solution without further software modifications.

CSAL is an innovative, open-source storage acceleration layer software designed to optimize QLC storage. It utilizes a log structure-based cache, in contrast to traditional options such as Open CAS .[11] The I/O flow difference is shown in Figure 5. Solidigm has built a reference storage design,[12] which is a turnkey solution to help you analyze your storage workload and design the best-fit storage framework for your application with CSAL.

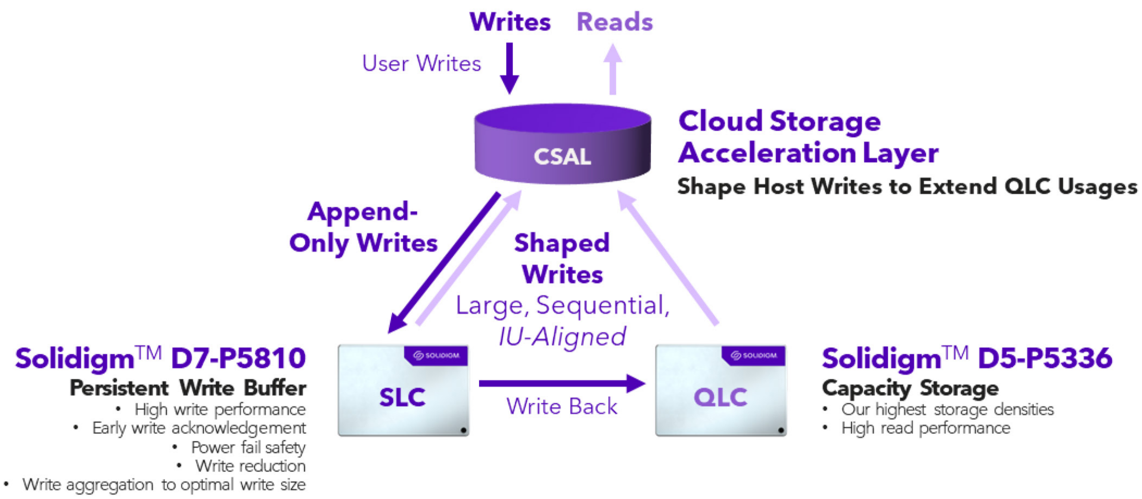


Figure 5. Set-associated cache vs. Log-structured cache

VAST Data Platform for SuperPOD with QLC

In the prior section, we explained how the combination of SLC + QLC with CSAL can meet AI storage requirements. In this section, we present a VAST Data storage solution that incorporates QLC technology into AI network file storage solutions using similar concepts.

1. VAST Data case study

VAST Data has consistently proven its pioneering and leading position in the storage industry.[13] As a well-established vendor, it has also achieved remarkable success in AI storage. The diagram below illustrates its architectural logic.

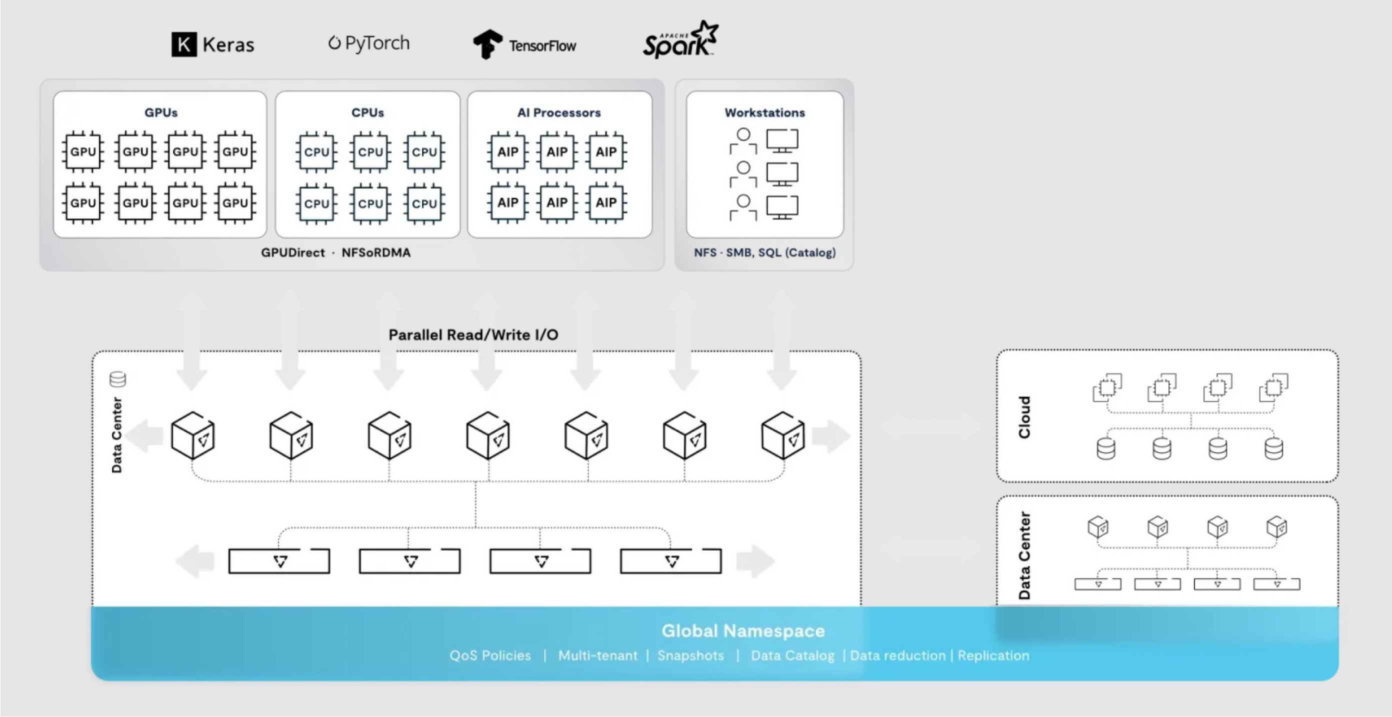


Figure 6. VAST Data Platform Reference Architecture[14] Copyrighted image used under permission by VAST Data

The storage nodes in this system serve two distinct roles. The first caters to computational resources aimed at GPUs, AI processors, and workstations and is fulfilled by service nodes. These service nodes provide semantic interpretation and connection protocol support. GPUs, CPUs, and AI processors are all connected to these service nodes using GPUDirect and NFSoRDMA. Data scientists utilize workstations that connect to the service nodes via NFS, SMB, or SQL semantics.

Going southward from the service nodes leads to the second role—storage nodes. These storage nodes are equipped with SCM (storage-class memory) for write caching and persistent metadata storage. QLC technology offers extensive storage capacity and provides direct read services. Customized network modules offer redundant links to the service nodes.

Importantly, there are no rigid bindings between service nodes and storage nodes; they are mutually visible. Unlike the more common "share nothing" design in most distributed storage systems, VAST's design adopts a "share everything" approach. What sets it apart is the data exchange between storage nodes, which ensures that data traffic remains minimal, steady, and free from peak and off-peak variations. The storage system presents a unified global namespace to the outside world. This architecture is referred to as Disaggregated Shared Everything (DASE).

Inside its storage nodes in Figure 7, the chassis comprises three primary components. The DNode takes responsibility for establishing connectivity with the compute head node, and redundancy can be achieved by incorporating two DNode network modules. In terms of storage, there are two distinct layers. SCM serves as the cache layer, while QLC technology plays the role of hyperscale flash, constructing a high-capacity, persistent storage layer.

Thanks to its large capacity and read performance on par with TLC, QLC boasts high storage density, and provides superior cost optimization per terabyte. This combination of TLC and QLC in hybrid storage leverages their respective strengths, effectively mitigating any shortcomings. The result is an AI storage solution that optimizes performance and cost.

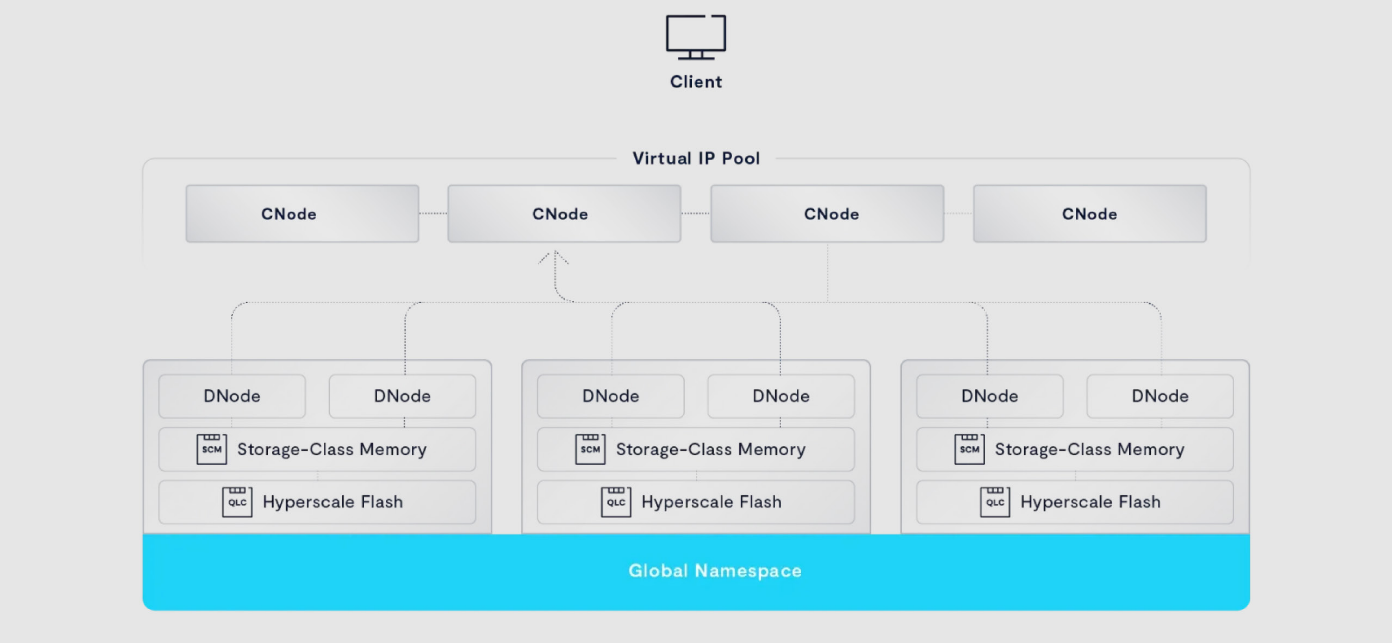


Figure 7. VAST cluster[15] Copyrighted image used under permission by VAST Data

2. Case study conclusions

VAST Data offers an exceptional AI storage solution. Interestingly, they have chosen a hybrid all-flash solution combining SCM and QLC technologies. The key competitive advantage lies in their data placement strategies. For other AI storage providers, leveraging an open-source solution like CSAL can facilitate a faster time-to-solution.

CSAL for cost and performance optimized AI storage solution

The development of AI has surpassed our wildest expectations. There is an increasing urgency to feed computational behemoths with more data at a faster rate, leaving no room for idle processing time and power. When considering performance, capacity, reliability, and cost in tandem, QLC storage technology has stepped into the spotlight.

In addition to being an excellent, cost-effective solution, CSAL can connect storage devices with diverse performance capabilities and capacities to meet the requirements of variety of AI workloads. CSAL has the following key capabilities that are either available today or under development to tune your system for varied AI training needs:

Available features: Shaping small random writes and caching to provide higher bandwidth and lower latency capability.

Features under development: Raid5F to increase system reliability.

CSAL has the potential to integrate the following features to further improve reliability, performance, and store the data that matters: RAID6 for reliability and durability, LRC (local reconstruction codes), compression and deduplication, and similarity data reduction to improve data quality and endurance of the capacity layer. Being open source, community and partners are empowered to integrate features in CSAL that matter most for their deployment.

About the Authors

Sarika Mehta is a Storage Solutions Architect at Solidigm with over 15 years of storage experience. Her focus is to work closely with Solidigm customers and partners to optimize their storage solutions for cost and performance.

Wayne Gao is a Principal Engineer as Storage Solution Architect at Soldigm. Wayne has worked on CSAL from Pathfinding to Alibaba commercial release. Wayne has over 20 years of storage developer experience as a member of the Dell EMC ECS all-flash object storage team and has 4 US patent filings/grants and is a published EuroSys paper author.

Yi Wang is a Field Application Engineer at Solidigm. Before joining Solidigm, he held technical roles with Intel, Cloudera, and NCR. He holds "Cisco Certified Network Professional," "Microsoft Certified Solutions Expert," and "Cloudera Data Platform Administrator" certifications.

[1] Figure 1: https://developer.nvidia.com/gpudirect-storage

[2] https://developer.nvidia.com/blog/gpudirect-storage/

[3] https://developer.nvidia.com/blog/wp-content/uploads/2019/08/GPUDirect-Fig-4.png Figure 4: https://developer.nvidia.com/blog/gpudirect-storage/

[4] NVIDIA DGX A100 Reference Architecture https://images.nvidia.com/aem-dam/Solutions/Data-Center/gated-resources/nvidia-dgx-superpod-a100.pdf

[5] Figure 7, Page #15: https://images.nvidia.com/aem-dam/Solutions/Data-Center/gated-resources/nvidia-dgx-superpod-a100.pdf

[6] Table 7, Page #20: https://images.nvidia.com/aem-dam/Solutions/Data-Center/gated-resources/nvidia-dgx-superpod-a100.pdf

[7] Table 8, Page #20: https://images.nvidia.com/aem-dam/Solutions/Data-Center/gated-resources/nvidia-dgx-superpod-a100.pdf

[8] "A study of SSD Reliability in Large Scale Enterprise Storage Deployments." usenix.org/system/files/fast20-maneas.pdf

[9] https://github.com/Open-CAS/standalone-linux-io-tracer

[10] CSAL introduction https://www.solidigm.com/products/technology/cloud-storage-acceleration-layer-write-shaping-csal.html

[11] Open Cache Acceleration Software https://github.com/Open-CAS/standalone-linux-io-tracer

[12] https://www.solidigm.com/products/technology/csal-based-reference-storage-platform.html

[13] https://vastdata.com/whitepaper

[14] Image: Deeplearning web

[15] Diagram 15 web