Western Digital.

Technical Brief

NVIDIA® GPUDirect Storage

Benchmarking GPUDirect and the Western Digital OpenFlex™ Data24 NVMe-oF™ Storage Platform

Abstract

This paper demonstrates the architectural viability of high-performance disaggregated storage infrastructure using RDMA with RoCE v2 in a ML/AI training environment. The architecture demonstrates simplicity whilst allowing a flexible approach to the linear scaling of GPU's, Performance and Storage.

Artificial Intelligence

The rapid adoption of Artificial Intelligence (AI) in recent years represents a paradigm shift in the technology industry. This surge is attributed to significant advancements in computing power, data availability, and algorithmic innovations. As AI integrates deeper into various sectors, industries are compelled to adapt, not only to leverage its potential but also to remain competitive and relevant.

AI's transformative impact stems from its ability to analyze vast amounts of data and learn from it, offering unprecedented insights and automation capabilities. This has opened new frontiers in areas such as healthcare, where AI assists in diagnosing diseases more accurately and quickly; in finance, through the automation of complex trading strategies and risk assessment; and in manufacturing, where predictive maintenance and optimized production processes have become the norm.

One of the most visible effects of AI's proliferation is the disruption of traditional business models. Companies are increasingly relying on AI to enhance customer experiences, streamline operations, and drive innovation. The trend towards personalization in services, powered by AI's data processing capabilities, is a stark example of this shift. Businesses that fail to integrate AI into their model's risk falling behind in an increasingly data-driven world.

General Considerations in Deploying Machine Learning Models

Computational Resources: ML models, especially deep learning, require significant computational power. GPUs or TPUs are often preferred for their parallel processing capabilities. The choice depends on the model complexity and real-time processing needs.

Data Storage and Management: Efficient data storage is crucial. This includes considerations for data volume capacity, flexibility, resilience and performance that complements the GPU's. High-speed storage solutions and databases capable of handling large datasets and ensuring data quality are important.

Scalability: The infrastructure should support scaling up or down based on demand. This involves using cloud services or scalable on-prem solutions to handle varying loads and data growth.

Security and Privacy: Secure storage and transmission of data, adherence to data privacy laws (like GDPR), and implementing robust access controls are essential to protect sensitive information.

Monitoring and Maintenance: Continuous monitoring of model performance and data drift is required. Infrastructure should support easy deployment of updates and maintenance.

Integration Capabilities: The ability to integrate with existing systems and data pipelines is important for seamless operations.

Cost Efficiency: Balancing computational requirements with cost, especially when using cloud services, is vital for sustainable operations.

The Challenges of accessing cloud-based Machine Learning

In recent years, cloud adoption for ML/AI tasks has surged, driven by its scalability, flexibility, and the promoting of its cost-effectiveness. Cloud platforms offer the vast computational resources and specialized hardware that enable businesses to develop, train, and deploy models more efficiently. This shift also democratizes access to advanced AI capabilities, allowing even small organizations to leverage sophisticated technologies. However, there are various concerns that are starting to persist which (in some instances) can be a catalyst for moving to on prem solutions.

Cost

Subscription and Usage Fees: Cloud-based ML services often come with subscription costs, and the pricing models can be complex, involving pay-per-use or tiered services which can escalate quickly with increased usage and in turn contribute to budget unpredictability.

To help place costs into context, the below is an estimated costing involved in the training of 70B parameter LLM model in the cloud.

- It will take approximately 24 days to train such a model.
- To meet the 24 days, this would require 128 current generation GPUs.
- Those GPU could be housed in 16 servers (8 GPU per server).
- If a single GPU server rental is priced conservatively at $30 per hour, that equates to $480 per hour for the 16 GPU Servers.
- That's $11,520 per day, or $276,480 for the 24 days.

This is a base cost for a single model. It does not account for any quality issues in the training cycle and assumes constant 100% utilization of those GPUs, with no network latency or additional bottlenecks. Accumulative post training inference and fine-tuning costs are not factored.

These costs can in theory be significantly offset by using a suitable (if available) 3rd party pre-trained model. Such a model still requires company specific fine tuning and so subscription costs would remain accumulative although time to value would be well within the 24 days of the previous example.

Data Transfer Costs: Moving large datasets into and out of the cloud can incur substantial costs.

Operational Costs: Ongoing operational costs for cloud services may include data storage, compute time, and additional services like data transformation or transfer.

The combination of these factors leads to a situation where even experienced ML practitioners can struggle to accurately forecast the costs associated with cloud-based ML projects.

Performance

Latency: Network latency can affect the performance of cloud-based ML models, especially in real-time applications.

Computational Limits: Some ML tasks may require more computational power than what is allocated in certain cloud tiers, leading to performance bottlenecks.

Resource Contention: Shared resources in the cloud can sometimes lead to contention, impacting performance if the cloud provider does not adequately isolate workloads.

Availability and Reliability

Downtime Risks: Cloud providers can have outages, affecting the availability of ML services.

Data Lock-In: There's a risk of becoming dependent on a single cloud provider, making it difficult to switch services without significant migration costs.

Service Level Agreements (SLAs): The guarantees provided by cloud providers in their SLAs may not always align with the needs of all ML applications.

Data Security and Privacy

Data Protection: Ensuring the security and privacy of sensitive data when using cloud-based ML services can be a concern, along with compliance to regulations like GDPR or HIPAA.

Access Control: Managing access to ML models and datasets in the cloud requires robust identity and access management systems.

Integration and Complexity

Integration with Existing Systems: Integrating cloud ML offerings with existing on-premises systems can be complex and may require significant architectural changes.

Learning Curve: There is often a learning curve associated with using cloud-based ML services, which can delay adoption and require training for the existing workforce.

Scalability

Auto-scaling Features: While cloud services generally offer good scalability, configuring auto-scaling correctly to handle variable workloads without incurring unnecessary costs can be challenging.

Geographical Limitations

Data Sovereignty: Legal and regulatory constraints may limit the geographical locations where data can be stored and processed, affecting the choice of cloud providers and services.

Vendor Lock-In

Proprietary Technologies: Cloud providers may use proprietary technologies that make it difficult to move ML applications between different clouds or to on-premises environments without significant rework.

Ecosystem and Support

Community and Support: The ecosystem of tools and community support varies between cloud providers, which can affect the ease of use and problem-solving resources.

OpenFlex Data24 NVMe-oF Storage Platform

Western Digital's OpenFlex Data24 3200 series NVMe-oF storage platform extends the high performance of NVMe flash to shared storage. Similar to the original OpenFlex Data24, It provides low-latency sharing of NVMe™ SSDs over a high-performance Ethernet fabric to deliver similar performance to locally attached NVMe SSDs.

Western Digital RapidFlex™ NVMe-oF controllers, allows up to six hosts to be attached without a switch, like a traditional JBOF.

OpenFlex Data24 3200 series uses Western Digital's RapidFlex C2000 Fabric Bridge Adapters to provide 2, 4, or 6-ports of 100GbE which can now connect to RDMA and/or TCP configured host ports. While RoCE (RDMA over Converged Ethernet) connections have historically been preferred in data centers, TCP offers greater ease-of-use and is sometimes preferred. OpenFlex Data24 3200 series offers the flexibility of connecting to either RoCE or TCP host ports for optimum usage. NVMe-over-Fabrics, or NVMe-oF, is a networked storage protocol that allows storage to be disaggregated from compute to make that storage widely available to multiple applications and servers. By enabling applications to share a common pool of storage capacity, data can be easily shared between applications or needed capacity can be allocated to an application to respond to application needs.

OpenFlex Data24 3200 series NVMe-oF storage platform can also be used as a disaggregated storage resource in an open composable infrastructure environment using the Open Composable API. The platform can also be specified with just two RapidFlex adapters for simpler environments and as a direct replacement for SAS external storage.

OpenFlex Data24 3200 NVMe-oF Enclosure

RoCE Specification Data				
RoCE		128K Bandwidth	4k IOPS	4K QD1 Latency
6 x 100GbE	Read	71.47 GB/s	16.76 M	83.6 us
	Write	66.52 GB/s	6.16 M	27.8 us

GPUDirect Benchmarking with the OpenFlex Data24 NVMe-oF Storage Platform

Aims of this benchamark were as follows:

- To showcase OpenFlex Data24 storage platform as a valid on-premises disaggregated storage solution in the use case of AI model training.
- In doing so, demonstrate that well architected NVMe-oF based solutions can provide sufficient bandwidth and the low latency required to saturate GPU's.
- Introduce the OpenFlex Data24 3200 as a high performant SSD cache tier that allows processed training set data to be staged from a data lake (for example) storage tier.

Positioning

This benchmark was not meant to address the infrastructure requirements of Large Language Models (LLM). These LLMs can contain upwards of billions of parameters and in turn cost billions of dollars to implement.

This example looked to address those companies who associate with the below:

- Require accelerated GPU learning capabilities whilst understanding the need to keep GPU utilization high.
- Wary of renting cloud-based ML infrastructure due to budget limitations, cost and or project duration uncertainty.
- May be exploring hybrid cloud model where on prem is anticipated to be cost advantageous (over cloud) for the compute intense model building phase.
- Have modest capital budget to invest in a scalable on-prem ML architecture.
- Are seeking an architecture that can scale predictably in capacity, compute, performance and cost.
- Do not want to be tied into a single vendor solution.
- Exploring specialized Domain (task) Specific Models some 10-100x smaller than the large foundational models – generally using private / enterprise data.
- May have a previously trained model and wish to move to Edge based mainstream inference and fine tuning to accelerate time to value.

The key considerations of such a solution:

- Selection of a suitable server (CPU and PCIe® architecture) that will allow for GPU scale without becoming the bottleneck.
- It is not uncommon to see servers allocating only 1-2 NVMe SSD slots per GPU.
- In this benchmark, it takes at least 10 NVMe SSD to near saturate the bandwidth of a single A100 GPU.

GPUs, NICs, and/or, NVMe all need to reside on the same server PCIe switch.

- A PCIe switch provides superior peer-to-peer communication compared to standard CPU root complexes.
- The ability to provide high performance storage networking that has the capability to not only meet (and exceed) the ever-increasing performance demands of GPUs; but also provide scalable capacity to meet learning model data set sizes.
- The solution is agnostic – to the GPU server, network, and SSD.

The Benchmark Software enablers:

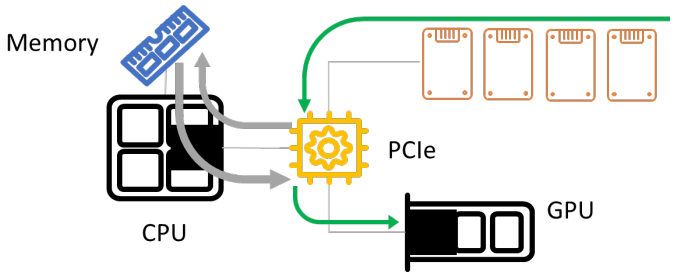
- **NVIDIA GPU Direct Storage (GDS)**

  o GDS enables a direct data path for direct memory access (DMA) transfers between GPU memory and storage, which avoids a bounce buffer through the CPU. This direct path increases system bandwidth and decreases the latency and utilization load on the CPU. This is particularly beneficial in high-performance computing and complex data processing tasks.

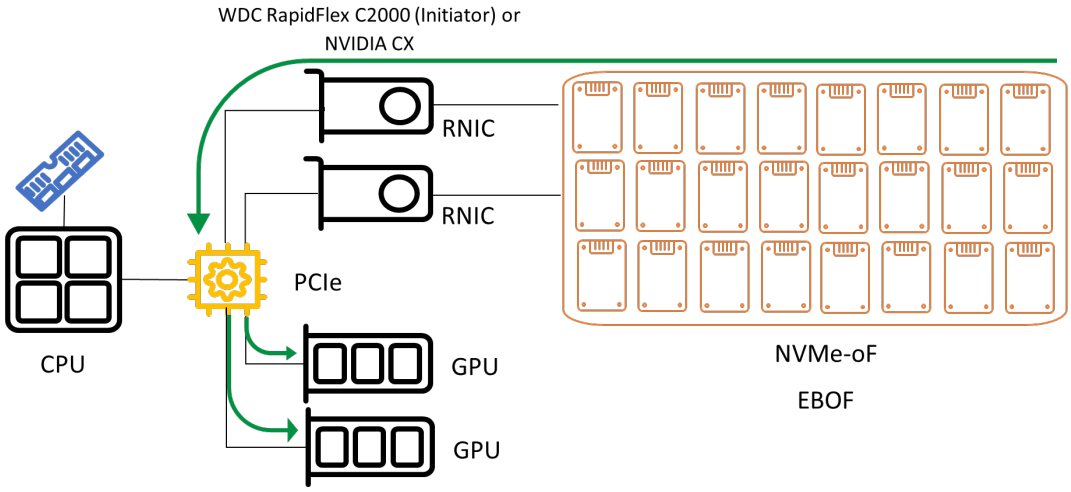
- **Gdsio Utility**

  o The gdsio utility is like several disk/storage IO load generating tools. It supports a series of command line arguments to specify the target files, file sizes, IO sizes, number of IO threads, etc. Additionally, gdsio includes built-in support for using the traditional IO path (CPU), as well as the GDS path - storage to/from GPU memory.

Without GDS, GPUs directly read training / inference data from local SSDs via the CPU complex which significantly limits GPU performance potential and scale:



With GDS, GPUs instead of traversing the CPU complex, have a direct path for data exchange. Western Digital RapidFlex adapters make the disaggregated storage (provided by the OpenFlex Data24) look like local NVMe storage. This SSD caching tier allows for linear performance and storage scale.



Infrastructure Overview

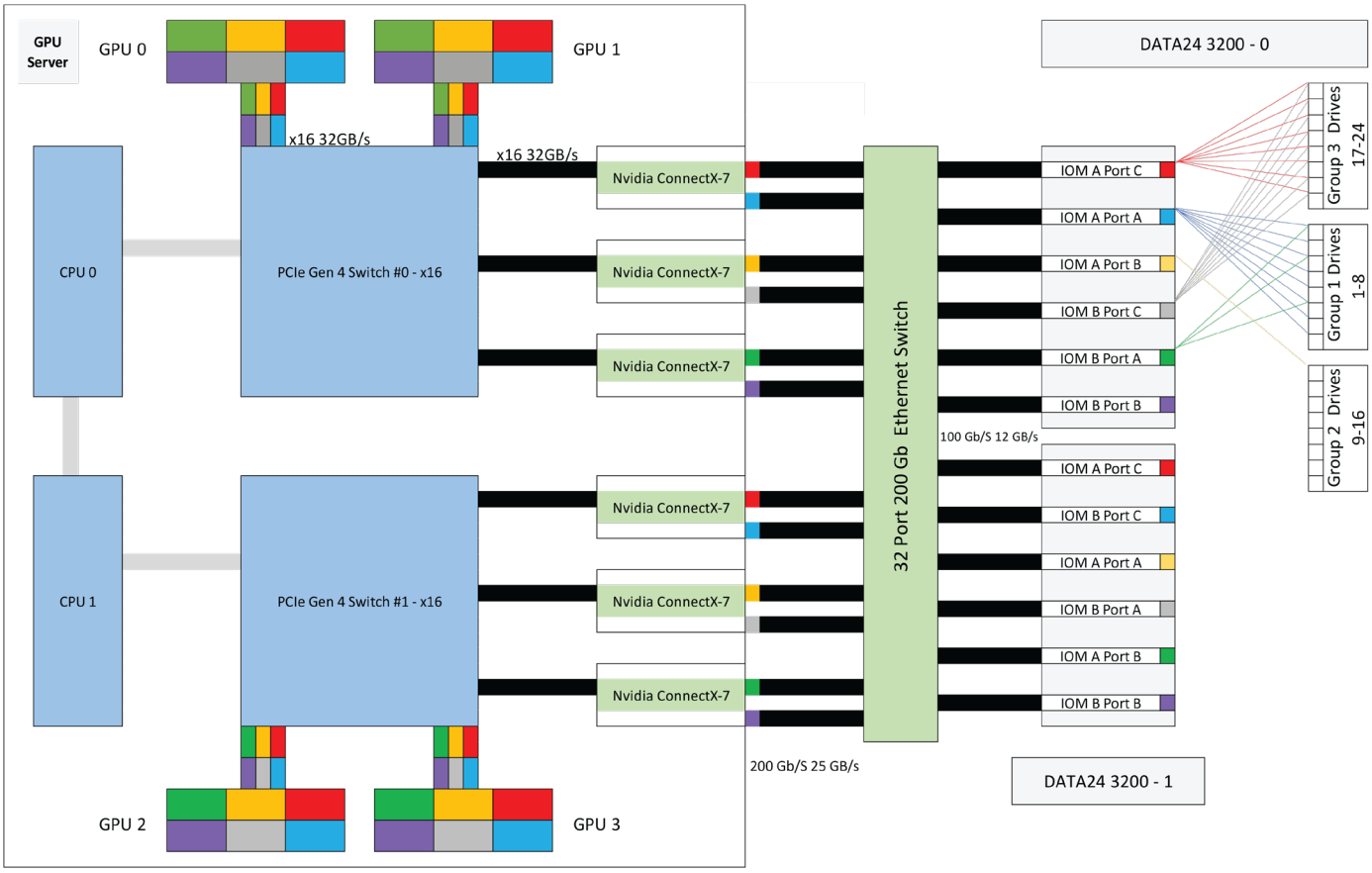
- GPU Server Software
  - RHEL 9: 5.14.0-70.70.1.el9_0.x86_64
  - Mellanox® OFED: 5.8-3.0.7.0
  - Nvidia Driver: 535.86.10
  - CUDA: 12.2.1
  - GDS: 2.17.3
  - Libcufile: 1.7.1.12
- GPU Server Hardware
  - FW/Redfish/BIOS/CPLD: 01.02.61/1.9.0/1.4B/F1.0C.08
  - Dual Intel® Xeon® Gold 6348 CPU 26 Cores @ 2.60GHz
  - 512GiB Memory
  - 8-Bay NVMe with no RAID Controller - Root Complex Connected
  - 16-Bay Drive with no RAID Controller
  - PCIe: (12) x16 Gen4
    - 10 PCIe switched (2 Switches)
    - 2 Root Complex Connected
  - 4 x NVIDIA A100 80GB PCIe GPU
    - 2 on Numa Node 0 (PCIe Switch 0)
    - 2 on Numa Node 1 (PCIe Switch 1)
  - 6 x ConnectX®-7 (ConnectX-6 also tested latterly)
    - 3 on Numa Node 0 (PCIe Switch 0)
    - 3 on Numa Node 1 (PCIe Switch 1)
    - FW: 22.35.2000
- Ethernet Switch
  - NVIDIA SN3700 32 Port 2000Gb (Spectrum 2).
  - 24 x 200Gb Direct Attach Copper (DAC) cables.
  - Configured with RDMA with RoCE v2 and appropriate lossless settings.
- Storage: Western Digital OpenFlex Data24 3200 Series
  - The Data24 3200 is using a Gen3 PCIe architecture and a total of 6 front end ethernet AIC ports, each running at 100Gb/s (12.5GB/s).
  - This gives a theoretical performance threshold of 75GB/s per chassis[1]
  - Network Storage Protocol: RDMA with RoCE v2
- Drives
  - 24 x 15.36[2] TB Dual ported Western Digital Ultrastar® DC SN655 NVMe™ enterprise SSDs. These drives offer high-capacity, cost-optimized, read-intensive performance for data-intensive applications.
  - Offering 368TB raw capacity.

[1] Synthetic benchmarking with FIO demonstrates steady state sequential (128k) Read at 71.47GB/s and writes at 66.52 GB/s with 15.36TB SN655 drives.
[2] One gigabyte (GB) is equal to one billion bytes and one terabyte (TB) is equal to one trillion bytes. Actual user capacity may be less due to operating environment.

Logical Configuration Overview

The diagram below shows (via color paths) the physical and logical mapping from the drives to their respective GPU.

- Data24 backend architecture.
  - o 3 AIC (100Gb/s) port connections are available from each IOM to their associated NVMe drives.
  - o Each AIC is physically mapped to 8 drives.
  - o 12 drives are mapped to each GPU (4 drives from each Group).
  - o The SN655 drives are dual ported, hence the need to map two paths from each drive (i.e. green / purple) to the GPU to allow full bandwidth from the drive.
- The 32 port Ethernet Switch
  - o Configured with appropriate lossless settings to support RDMA with RoCE v2
- The GPU Server OS was used for logical volume management.
  - o This was to facilitate simplicity in the benchmarking.
  - o Drives were presented as single devices and were both physically and logically mapped to the GPU's as per the color instances in the diagram below.
  - o Using a volume manager in conjunction with GPU Direct, does introduce considerations that could potentially affect the GPU Direct's underlying model.



Note: For simplicity, the drive mapping from the Data24 3200-1 is not shown in the above diagram as it is identical to that of the Data24 3200–0.

Scaling Drives per GPU

In the example below, GDSIO was used to explore the number of drives required to reach a plateau in in GB/s to a single GPU.

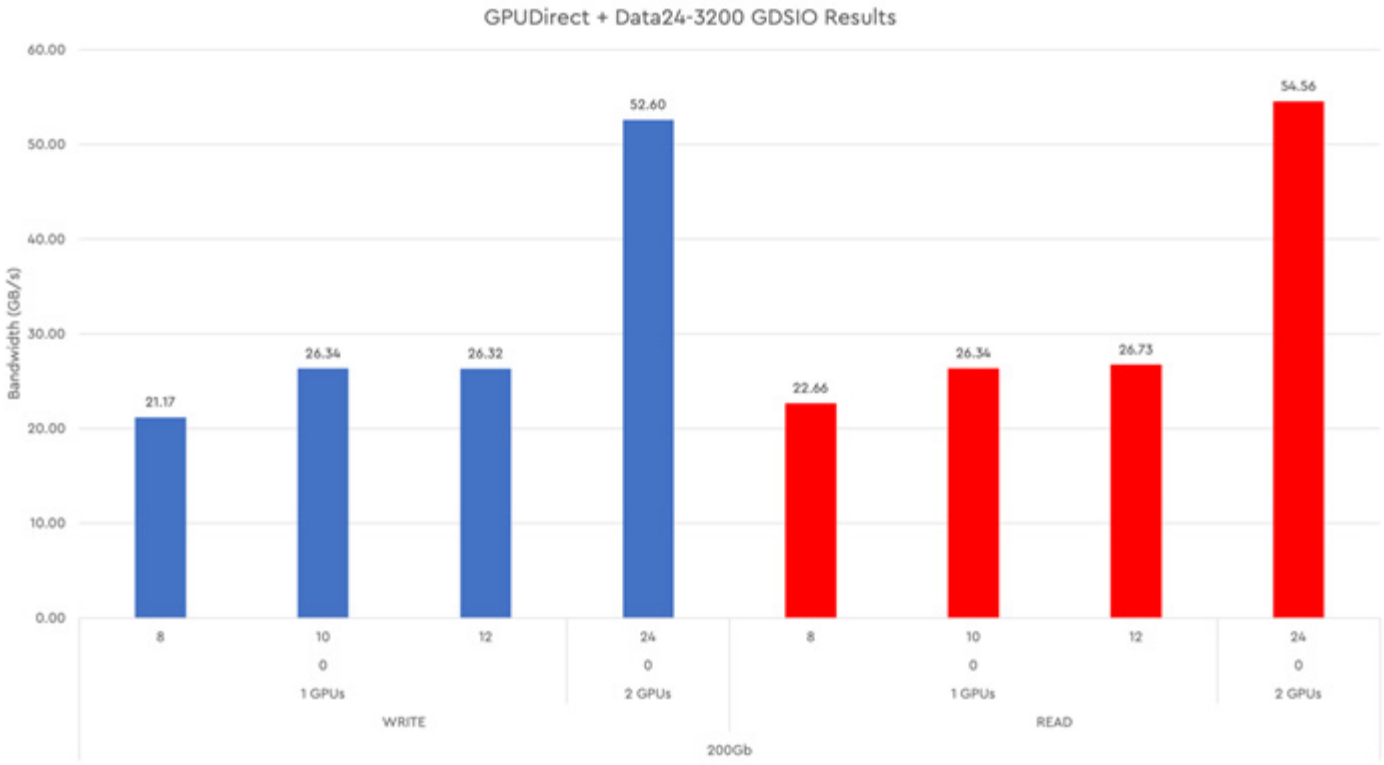
The NVIDIA A100 PCIe GPU uses the PCIe Gen4 architecture which in turn allows for a theoretical 32GB/s across the PCIe link in a symmetrical configuration.

Using fully preconditioned drives, a maximum bandwidth of 26.34 GB/s is reached for Writes and Reads at 10 drives. There is no appreciable increase in performance when scaling to 12 drives per GPU. Whilst not shown in the diagram, this remained the case to the maximum tested 16 drives per GPU.

To allow for overall architectural simplicity, it was concluded that 12 drives per GPU per NUMA node switch presents an optimal configuration. This allows for the performance capabilities of the Data24 architecture, whilst offering additional storage capacity and composable storage flexibility. Configuration options are then realized for when outright performance may be a secondary consideration to the allocation of GPU resources for multiple concurrent training / inference uses.

In turn, utilizing the remaining 12 drives in the Data24 whilst mapping to a second GPU (connected to the same NUMA node switch) demonstrates a linear doubling of performance (52.6 GB/s Writes and 54.56 GB/s Reads)

It is worth noting that SSD model choice and its associated performance characteristics may impact benchmark results. Every bottleneck in the IO path should be considered.



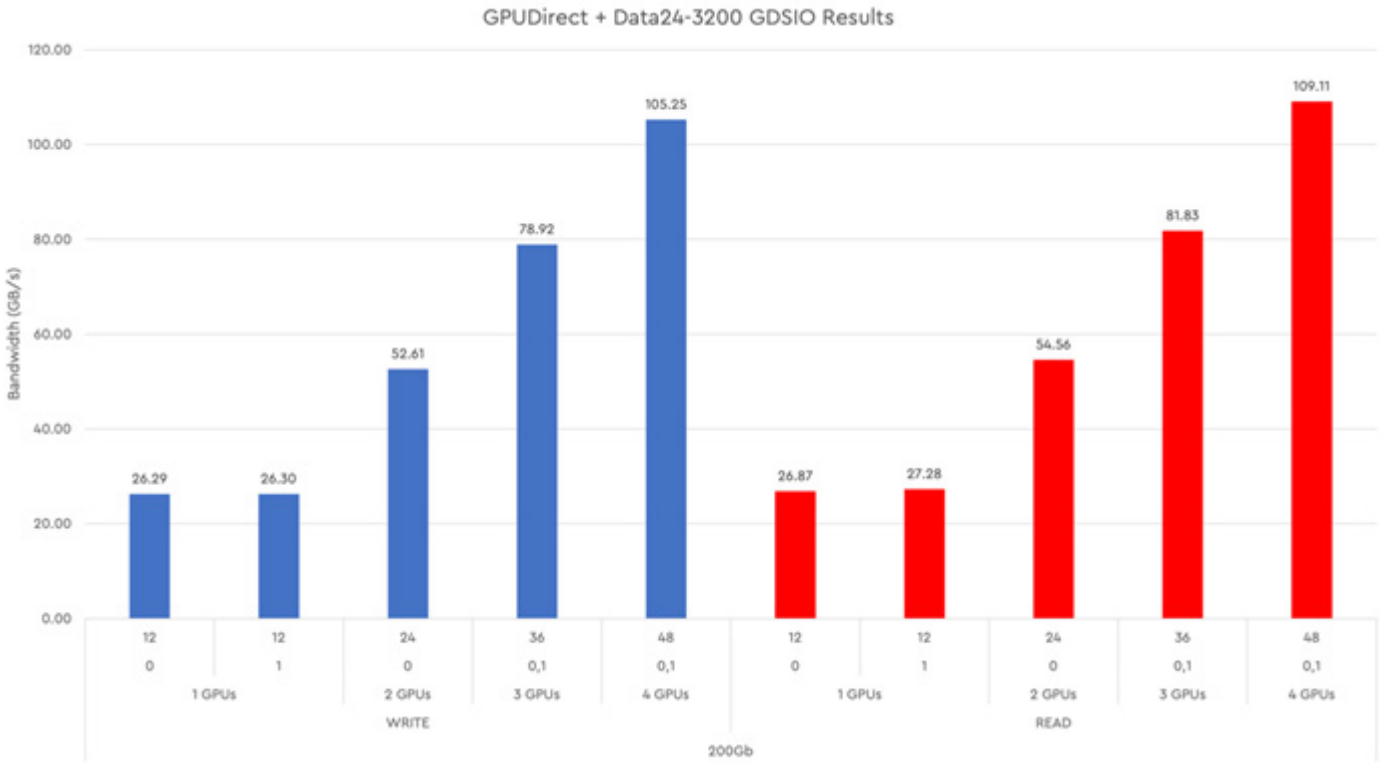
Scaling Data24 per GPU Server

The GPU Server in this benchmark allows for 2 GPU per NUMA node switch and so therefore 4 GPU per chassis whilst allowing or RNIC slot requirements.

Adding an additional Data24 allowed the second NUMA node to be benchmarked. In essence there is a split configuration of 1 x Data 24 for a pair of GPUs per NUMA node.

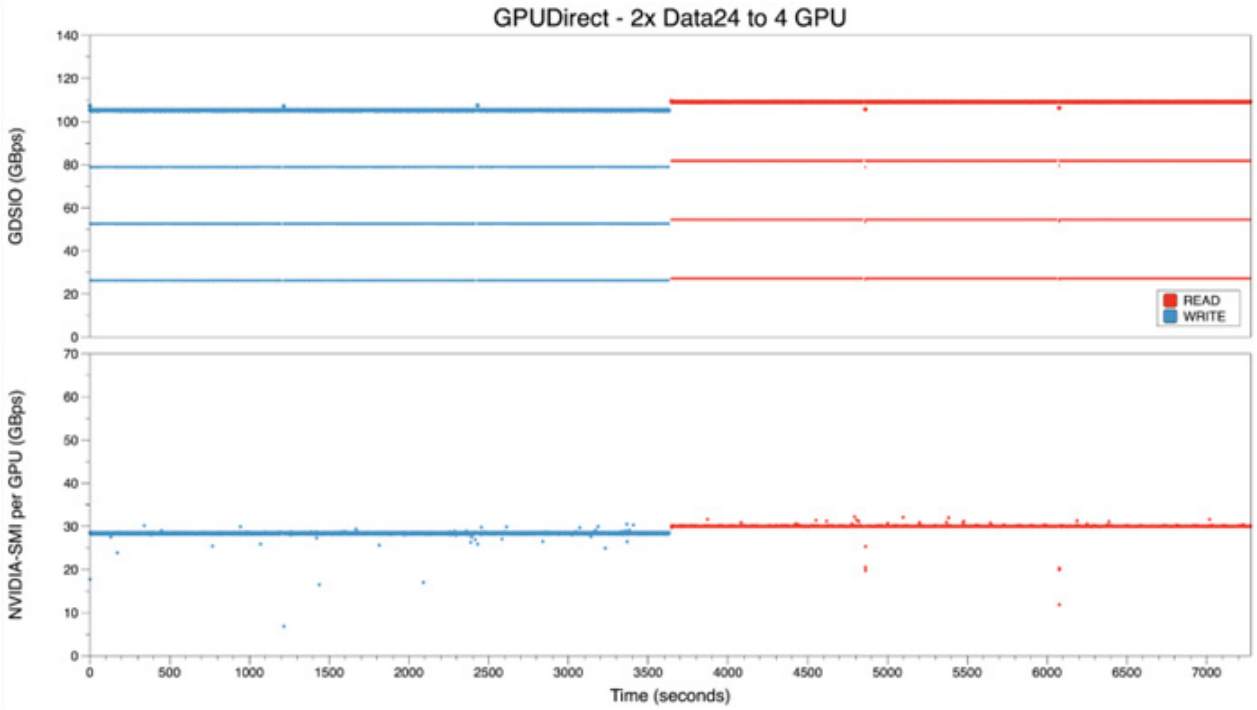
The results below demonstrate the linear scaling of performance to 105.25 GB/s Writes and 109.11 GB/s Reads for the GPU chassis.

It is worth noting that in all benchmarks, the Fabric to Data24 connectivity was set to 100Gb/s (12.5GB/s) and the GPU Server (ConnectX-7) to Fabric connections set to 200Gb/s (25GB/s). A latter phase of benchmarking was to swap the CX-7 for CX-6 and configure both sides of the network to 100Gb/s. and re-run all benchmarks. These tests demonstrated no appreciable performance difference (0.019%) at max scale.



Performance Stability

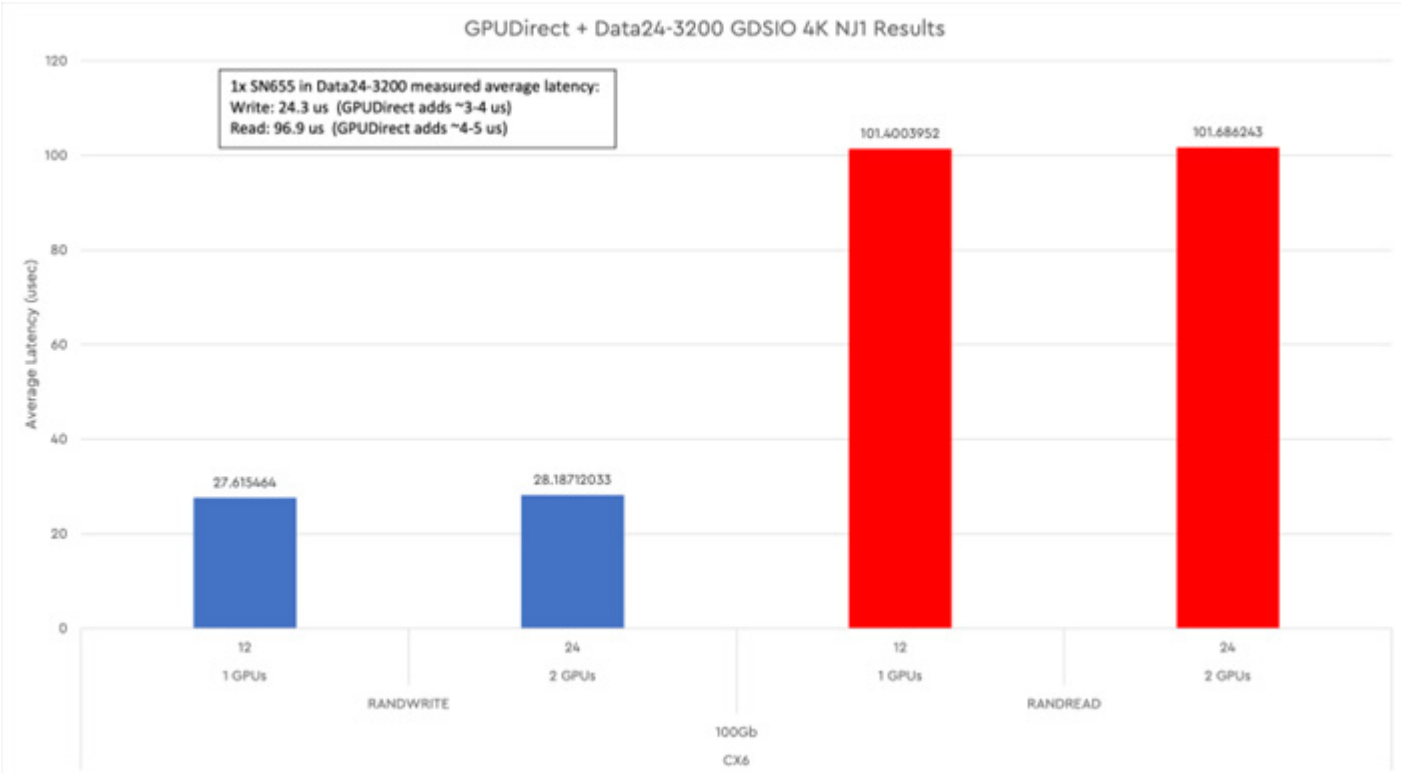
- Time series data is important as it allow for additional analysis beyond that of the average result provide by the benchmark utility.
  - With timeseries data one looks for peaks and troughs that may get lost in the benchmark's final average.
  - Flat plateaus are preferred as they indicate stable performance.
  - Time series data was collected from:
    - nvidia-smi which hows bandwidth from the perspective of the PCIe bus.
    - gds-stats: which shows bandwidth from the perspective of the GPUDirect Storage software.
- In both instances of measurement, it evident that performance is stable both across the duration of testing and as GPU are scaled.



Performance Latency

Western Digital testing of RapidFlex based NVMe-oF products has shown that disaggregation of NVMe to NVMe-oF only adds ~10µ seconds when compared to in-server NVMe drives.

- Sequential 4k tests were used to show the lowest attainable latency in this architecture.
  - 4k Random would show most of the drives with reads around 100µ seconds.



Conclusion

This paper demonstrates the architectural viability of high-performance disaggregated storage infrastructure using RDMA with RoCE v2 in a ML/AI training environment. The architecture demonstrates simplicity whilst allowing a flexible approach to the linear scaling of GPU's, Performance and Storage. The infrastructure is also largely agnostic in terms of hardware. Whilst somewhat dependent on compute, storage capacity and performance requirements, the consumer ultimately has choice over which GPU servers, GPUs RNIC's, Network components and SSD model to incorporate; all within a clearly defined cost model.

Next Steps

This PoC has proven that its underlying architecture along with GDS presents a compelling option for near line saturating of GPUs in the ML space.

There is more to be explored here, however:

- GDSIO, whilst specializing in benchmarking GDS technology, does not simulate the requirements of specific machine learning tasks.
  - o GDSIO benchmarks the capability of the underlying infrastructure.
  - o The GPUs themselves are not placed under load by GDSIO.
- To address this, MLPerf is an option to be used for a secondary set of benchmarks.
  - o MLPerf is a consortium backed benchmarking suite whose objectives are to move to standardized evaluation benchmarks offering direct hardware and software comparability for published results.
  - o The MLPerf Training benchmark suite will allow for a better understanding of the PoC's capabilities in areas such as image classification, object detection, speech recognition, translation, recommendation systems, reinforcement learning, etc.
- The impact of file systems their physical location and associated Volume Managers within the ML infrastructure are also areas for consideration.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
www.westerndigital.com

February 2025 / Revision 02