Western Digital.

Technical Brief

NVIDIA® GPUDirect® Storage

Benchmarking GPUDirect and the Western Digital OpenFlex™ Data24 4200 Series NVMe-oF™ Storage Platform

The Future of Memory and Storage (FMS) 2024 - Tornado Demonstration

Abstract

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

Benchmark

This benchmark was designed to showcase the high sustainable throughput that can be achieved when correctly architecting a disaggregated storage solution for the purposes of high-performance computing (such as Machine Learning). In this example the aim was to demonstrate a visual rendering of a 2011 Oklahoma Tornado in its formation stage.

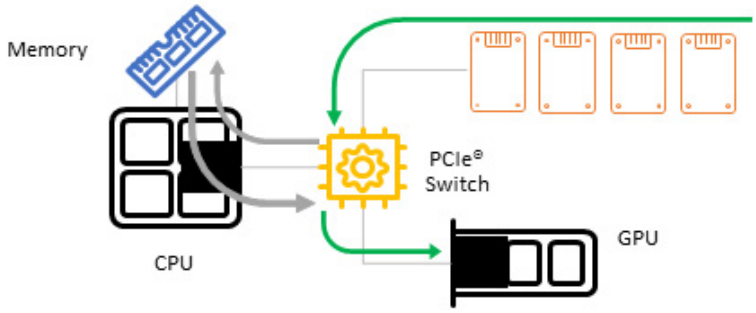
This simulation (mathematically derived from initial conditions just before the tornado formed) includes 250 billion grid points, each with over a dozen attributes such as rain, hail, pressure, and wind speed. This detailed visualization, showing 6000 simulation steps, provides unprecedented insight into the tornado's dynamics with a storage requirement of 5.9TB of data.

The Benchmark Enablers

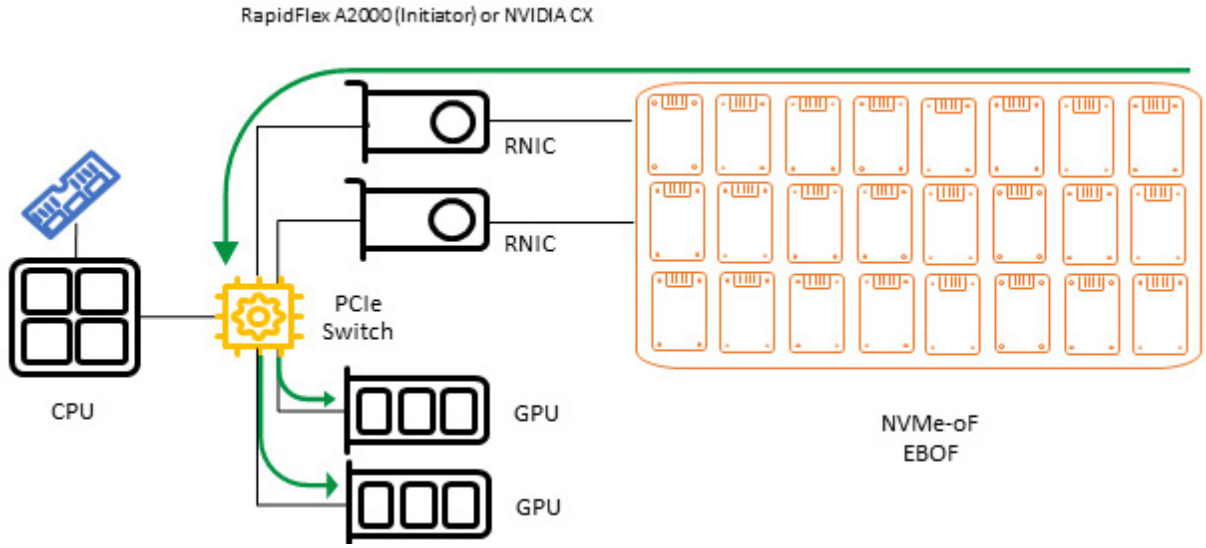
NVIDIA GPU Direct Storage (GDS)

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.

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.



Gdsio Utility

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. In this instance, Gdsio allows for synthetic benchmarking of the solution.

NVIDIA IndeX®

NVIDIA IndeX is an advanced volumetric visualization tool designed to handle massive datasets with high fidelity. IndeX leverages GPU acceleration to provide real-time interactive visualization of 3D volumetric data, making it indispensable for industries such as oil and gas exploration, medical imaging, and scientific research. Traditional visualization tools often struggle with the sheer size and complexity of modern datasets, leading to slower rendering times and less interactive user experiences. IndeX overcomes these limitations by utilizing NVIDIA's GPU technology to deliver high-performance rendering and data processing, ensuring users can interact with their data in real-time.

IndeX's capabilities are driven by its ability to harness the parallel processing power of GPUs, enabling it to manage and render large-scale volumetric data efficiently. This capability is valuable in applications that require high-resolution visualization, such as seismic interpretation and reservoir simulation in the oil and gas sector. By providing detailed and accurate visual representations of subsurface structures, IndeX helps geoscientists make more informed decisions. In the medical field, IndeX facilitates the visualization of complex anatomical structures from imaging modalities like MRI and CT scans, aiding diagnosis and treatment planning.

The real-time rendering capability of IndeX is also crucial for scientific research, where large datasets from simulations and experiments need to be visualized and analyzed promptly. Researchers can interactively manipulate and explore their data, allowing faster hypothesis testing and discovery. IndeX's scalability ensures it can handle the growing data volumes generated by advanced scientific instruments and simulations, providing researchers with the tools to visualize and interpret their data effectively. By integrating seamlessly with existing workflows and supporting various data formats, IndeX enhances productivity and accelerates the pace of discovery across multiple disciplines.

OpenFlex Data24 4200 Series NVMe-oF Storage Platform

With up to 368TB[1] of low latency Dual Port SSDs in a 2U 24-bay platform, Western Digital's OpenFlex Data24 4000 series NVMe-oF storage platform extends the high performance of NVMe flash to shared storage. Similar to the original OpenFlex Data24 and the OpenFlex Data24 3200 series, it provides low-latency sharing of NVMe SSDs over a high-performance Ethernet fabric to deliver similar performance to locally attached NVMe SSDs. Unsurpassed connectivity in its class, using Western Digital RapidFlex NVMe-oF controllers, allows up to six hosts to be attached without a switch, like a traditional JBOF.

OpenFlex Data24 4000 series uses Western Digital's RapidFlex A2000 Fabric Bridge devices to provide 12-ports of 100GbE which can 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 4000 series offers the flexibility of connecting to either RoCE or TCP host ports for optimum usage.

OpenFlex Data24 4000 series enables PCIe Gen4 performance throughout the chassis, bringing the full performance capability of each SSD to the Ethernet fabric. PCIe Gen4 SSDs from Western Digital and 3rd parties are supported.

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 4000 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.

For this benchmark, 24 x 15.362[1] TB Dual ported Western Digital Ultrastar® DC SN655 NVMe™ enterprise SSDs were chosen. These drives offer high-capacity, cost-optimized, read-intensive performance for data-intensive applications whilst allowing for 368TB raw capacity.

OpenFlex Data24 4000 Series NVMe-oF Enclosure

OpenFlex Data24 4000 Series Raw Performance		
Specification	**RoCE**	**TCP**
Random Read (4kB)	27 M IOPs	21 M IOPs
Random Write (4kB)	3 M IOPs	3 M IOPs
Sequential Read (32kB)	135 GB/s	113 GB/s
Sequential Write (128kB)	92 GB/s	86 GB/s
Average Random Read Latency (4kB)	96.17 µs	101.9 µs
Average Random Write Latency (4kB)	25.16 µs	53.43 µs

The Architecture

For this testing, the OpenFlex Data24 4000 and GPU server are connected through a 200GbE switch using the NVMe-oF RoCEv2 protocol with matched MTUs of 5000. The GPU server uses 3 Mellanox® CX7 RNICs with 2x 200 GbE per RNIC. The OpenFlex Data24 4000 is available with 12x 100GbE ports.

Each CX7 port has 2 IP addresses, allowing a single CX7 to map to four ports on the Data24. This provides connectivity to all 4 PCIe lanes on each dual-ported drive. The 6x 200 GbE links equal the bandwidth potential of 12x 100GbE links for a non-blocking network architecture.

Each NVIDIA H100 is connected via a PCIe Gen5 x16 slot, which can theoretically achieve 64GB/s of bandwidth bidirectionally. Each 200GbE and 100GbE RNIC port can theoretically reach 25 GB/s and 12.5 GB/s respectively.

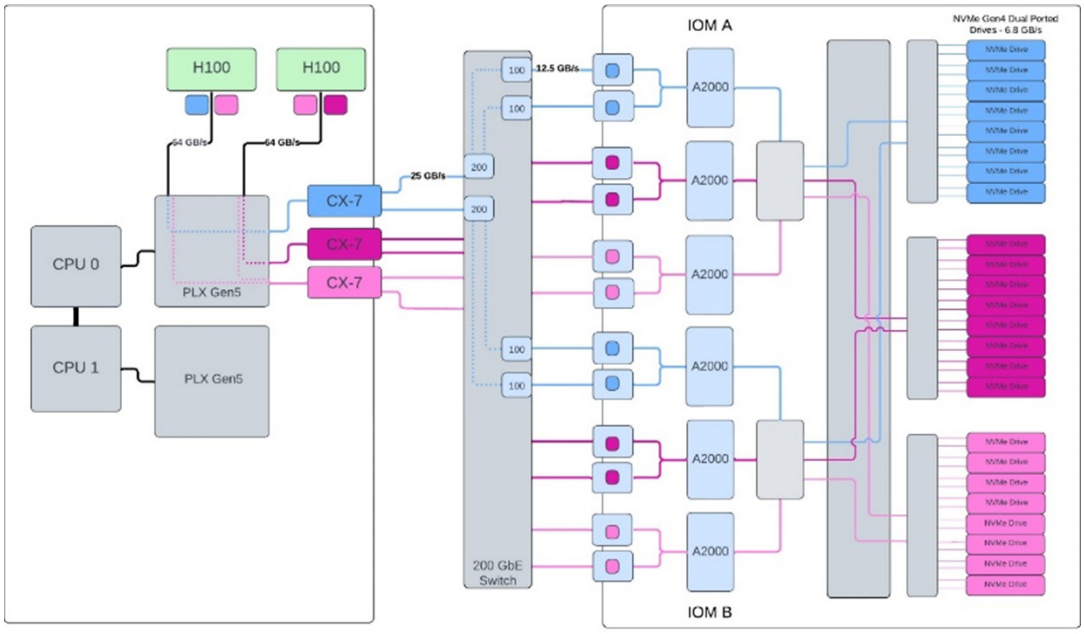
A critical design consideration is to ensure a non-blocking architecture. This requires that the GPUs, RNICs, and NVMe-oF drives are all mapped physically on the same CPU, NUMA, and PLX switch. Using the PLX switch provides superior peer-to-peer communication (performance) when comparing data traversing a standard CPU root complex. This allows the configuration to take full advantage of GPUDirect.

[1] 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.

As seen in this implementation, a mirrored configuration on the second CPU, NUMA, and PLX switch would allow a predictable compute scale and a theoretical doubling of performance.

This introduces scale considerations such the number of GPU's that can be assigned to a PLX switch along with the required number of RNIC's needed to connect to the disaggregated storage The RNIC's need to capable of providing the aggregated throughput requirements to those GPU's.

An understanding the performance of a chosen NVMe SSD and the quantity of those of NVMe drives (mapped to the GPU) is required in order to saturate the bandwidth of a single H100 GPU.

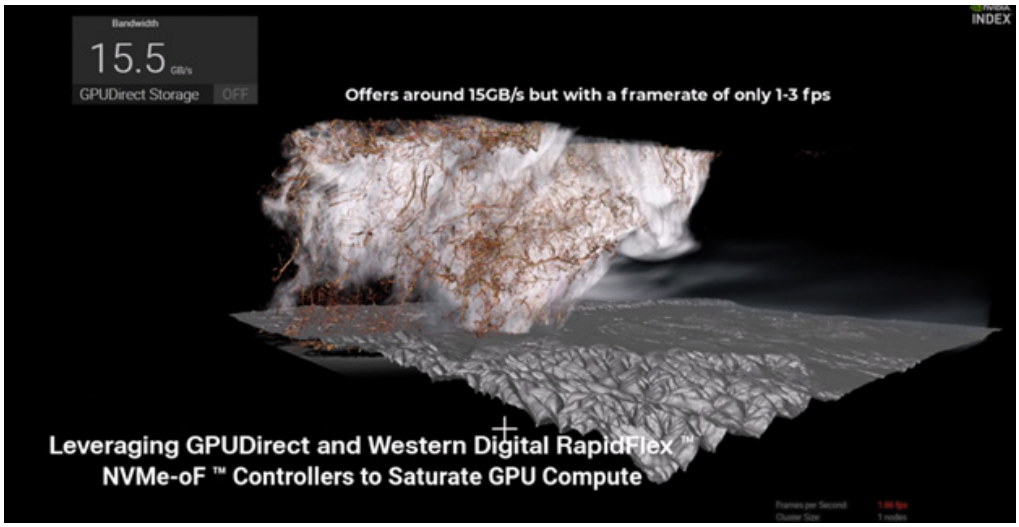


In AI training clusters, the combination of Data24 4000 and GPUDirect can allow for faster training times by reducing the bottlenecks associated with data loading. The efficient data paths ensure that GPUs can continuously receive data without interruption, maintaining high processing speeds and improving overall system efficiency. This setup is also advantageous for real-time analytics and other applications that demand rapid data access and processing, providing a significant performance boost to various computational workloads.

Benchmark Results

GDSIO Disabled

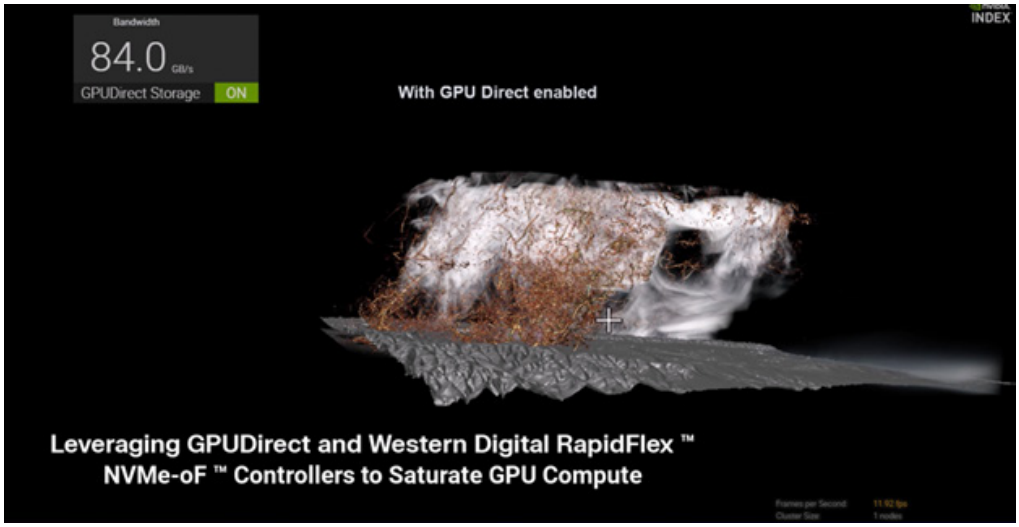
With GDISO disabled, we experienced a throughput of around 15.5GBs. The bottleneck in this instance is the CPU complex. Whilst the CPU are 4th Gen Intel® Xeon® Scalable Processors (8462Y), they are unable to offer the required throughput. The observed 1.5 frames per second (fps) rate is far too slow to make interactive navigation of the model a productive experience.



GDSIO Enabled

With GDSIO enabled, we see an immediate increase in throughout to 84GB/s. This is a 460% increase in performance. Throughput was observed to reach 90GB/s although the 84GB/s can be considered sustained throughput. Frame rate has also increased from 1.5 fps to around 15fps, much improving the usability of the simulation. In total the solution was seeing a throughput of 6TB approximately every 60 seconds.

Both of the H100 used to drive the simulation were observed to reaching peak multiprocessor utilization. The GPU were the bottleneck in this instance whereas (in this configuration) the Data24 4200 had around 8% read throughput performance capability in reserve.



Conclusion

The performance, time, and cost benefits that a well-configured, non-blocking architecture can offer GPU-accelerated workloads are well demonstrated in this project. Put simply, driving GPUs to their maximum throughput or processing capability drives more efficient outcomes and return on investment.

Western Digital's architecture supports Open Composable Infrastructure (OCI), and the OpenFlex Data24 4000 platform leverages this OCI approach by disaggregating data storage using NVMe-over-Fabrics (NVMe-oF). This decoupling of the storage resources from the GPU server not only helps free up the servers' resources (releasing those resources from traditional lockstep upgrades), but in doing so, also allows a fine-tuning of NVMe Drive mapping to GPUs. This precise drive matching to GPU requirements allows GPU capability, performance, and data capacity needs to be closely addressed, which in turn offers the predictable scale and flexibility required for those resources.

As the data is no longer siloed, it becomes an accessible networked storage resource, shareable amongst multiple GPU servers as needed, further increasing flexibility.

The Western Digital OpenFlex Data24, combined with NVIDIA GPUDirect technology, demonstrates a formidable capability in handling AI and other GPU-accelerated workloads. By enabling direct data paths between GPU memory and NVMe storage, the Data24 significantly reduces latency and maximizes bandwidth, ensuring efficient data handling and optimal GPU utilization. This integration allows faster, more effective processing of large-scale datasets, making the Data24 an invaluable asset in modern data-intensive environments.

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