SOLUTION BRIEF

PAVILION

Pavilion HyperParallel Data Platform™ for NVIDIA

Dramatically accelerate GPU based workloads with Pavilion

As organizations move into a data centric world, NVIDIA GPUs power an increasing range of critical applications, such as AI/ML, deep learning, Big Data analytics, HPC, video, and more. GPU-based systems are ideal for these types of applications as the data sets used in them can easily be hundreds of terabytes or larger in size, with millions of files. GPU based systems solve the challenge of processing these massive data sets, which cannot be done efficiently using traditional CPUs.

The key for GPUs to solve the processing bottleneck for these types of workloads is that they need to be fed data fast enough. When connecting to traditional storage systems, GPUs often become IO bound, limiting performance. Unlike CPUs, GPUs process data faster than traditional storage can provide it to them, slowing time to results.

This can have a significant impact on application performance. Analytics used for fraud detection may not identify suspicious activity quickly enough resulting in a loss, a trend in consumer patterns may not be identified by a retailer negatively impacting inventory control, or facial recognition software may not identify an individual in time for law enforcement to act. These situations can be avoided if the GPUs powering those applications are fed data fast enough.

The Pavilion HyperParallel Data Platform™, the most performant, dense, scalable, and flexible storage platform for GPU and CPU based applications, enables customers to get faster results, with larger data sets, which can have a material impact on results for their organization. Stock analysts make better trading decisions, medical researchers make discoveries faster, law enforcement respond to threats quicker, and businesses obtain a competitive advantage that would not otherwise be possible.

Performant Storage for GPU Powered Workloads

The Pavilion HyperParallel Data Platform delivers universally unrivaled performance for block, file, and object workloads. By leveraging the unique architecture of the Pavilion HyperParallel Data Platform and advanced technolgies, such as NVMe-oF and RDMA, Pavilion delivers unrivaled performance and ultra low latency, with a remarkably small footprint.

Recognizing the importance of high performance and low latency when ingesting massive data sets to train AI models, perform analytics, run HPC workloads, and more, NVIDIA has developed Magnum IO GPUDirect Storage, which enables highly effecient I/O to occur between GPU memory and external storage arrays. While this moves data directly to the GPU, bypassing the CPU, the storage still needs to be performant enough to provide that data.

By leveraging Magnum IO GPUDirect Storage, Pavilion provides significantly greater performance than without GPUDirect Storage.

> "Performance improvement for file storage of almost 4x by using GPUDirect Storage"
>
> Michael Kagan, NVIDIA Chief Technology Officer, SuperComputing 20, comparing the Pavilion HyperParallel Data Platform with and without GPUDirect Storage.

Benefits

Shatter Expectations

- Dramatically accelerate results
- Reduce AI model training times
- Perform Big Data Analytics on larger data sets, faster
- Eliminate wasted GPU cycles
- Fully saturate a DGXA-100
- Reduce management overhead

Accelerating Magnum IO GPUDirect Storage

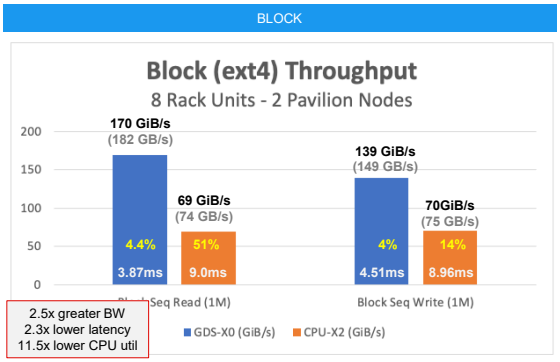
- Support for block and file workloads with Magnum IO GPUDirect Storage
- 4RU High density storage solution
- High throughput for both reads and writes
- 191GB/s read and 118GB/s write file performance from 2 arrays*
- 182GB/s read and 149GB/s write block performance from 2 arrays*
- No requirement for host agents
- Ultra-low latency

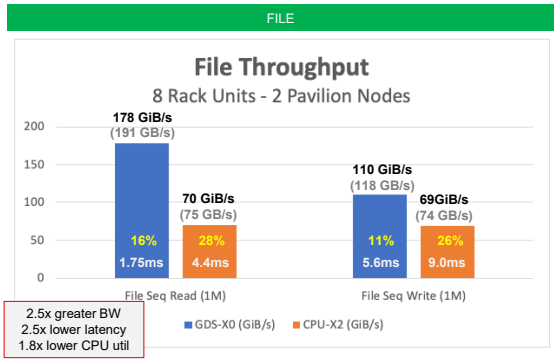
Get more from your array

- Unrivaled performance, density, scalability, and flexibility with up to 120GB/s read and 90GB/s write from each array
- Ultra low latency from 100µs read to 25µs write
- Consistent high performance for block, file, and object workloads
- Deliver data faster to accelerate model training and inference and to analyze larger data sets
- Achieve the highest performance density and lowest latency in the smallest footprint, with unlimited scale
- Eliminate the I/O bottleneck to achieve results faster across every workload

*All numbers were tested with one Pavilion system and 4 GPU's and extrapolated to two Pavilion systems (8RU) and 8 GPUS. The solution is expected to scale linearly. A special configuration and approval from NVIDIA is required to achieve this IO speedup, which is not supported as default GDS configuration in DGX

In testing validated by NVIDIA, the Pavilion HyperParallel Data Platform, using NVMe-RDMA and NVMe-RoCE for block data and NFS RDMA for file data, was shown to shatter performance expectations for DGX systems using NVIDIA Magnum IO GPUDirect Storage, as well as DGX systems without GPUDirect Storage.





Performance testing performed by Nvidia using GDSIO-X0 DGX-A100
Blue = Results with GPUDirect
Orange = Results without GPUDIrect Storage
% in YELLOW = CPU Utilization
ms = latency
All numbers were tested with one Pavilion chassis and 4 GPU's and extrapolated to two Pavilion chassis (8RU) and 8 GPUS. The solution is expected to scale linearly.
A special configuration is required to achieve this IO speedup – not supported as default GDS configuration in DGX

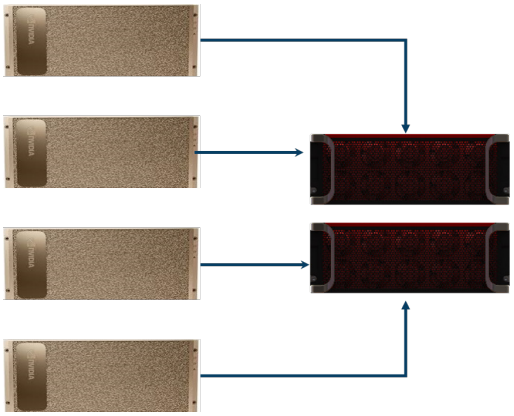
The Pavilion HyperParallel Data Platform enables customers to leverage NVIDIA Magnum IO GPUDirect Storage to saturate a DGXA-100 at near line speed, producing unprecedented results for their most demanding applications. A special configuration and approval from NVIDIA is required to achieve this IO speedup, which is not supported as a default GDS configuration in DGX

NVIDIA environments with Magnum IO GPUDirect Storage receive high performance and low latency equivalent to internal NVMe drives. NVIDIA systems that are not using Magnum IO GPUDirect Storage can still take advantage of the universally unrivaled performance, density, scalability, and fliexibility of the Pavilion HyperParallel Data Platform to achieve unprecedented results.

From customers that have a single workload runnning on an individual DGX system to those running multiple applications across any number of systems, the Pavilion HyperParallel Data Platform delivers universally unmatched storage.

Data Sets of Unlimited Size

Organizations leveraging GPU based solutions are processing data faster than they ever thought possible. Achieving these results require high performance access to massive data sets. When those data sets exceed the internal storage capacity of the server, their ability to achieve those results is impacted.



- Unimagined results with real time analysis and fast model training
- Equivalent performance to internal storage
- Up to 2.2PB of capacity per array
- Unlimited scale across any number of arrays
- High performance, ultra-low latency access to block, file, and object data in any combination

With NVIDIA Magnum IO GPUDirect Storage and the Pavilion HyperParallel Data Platform, organizations take advantage of data sets of unlimited size, across any number of servers, with high performance and low latency equivalent to their internal storage.

Whether customers have adopted GPUDirect storage or not, the Pavilion HyperParallel Data Platform shatters customer expectations and resulting organizational outcomes by revolutionizing data processing for modern AI/ML, HPC, Analytics, Enterprise Edge and other data-driven applications. Pavilion delivers unmatched performance and density, ultra-low latency, unlimited scalability and flexibility, providing customers unprecedented choice and control.