PNY GRAPHICS | HPC | AI | PRO

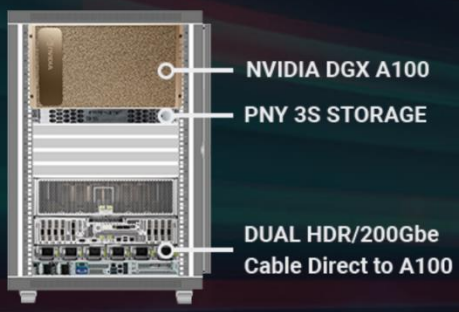
PNY 3S STORAGE SERVERS

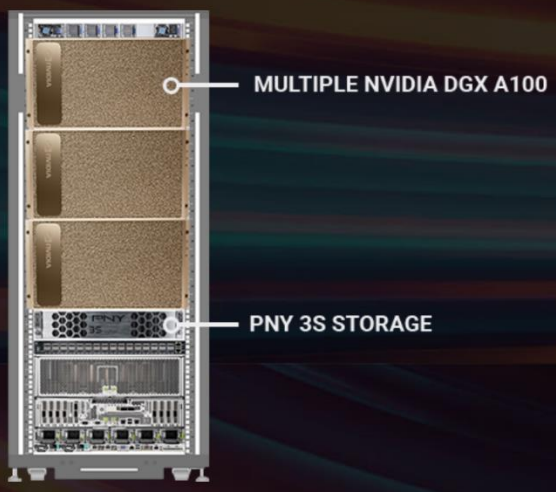
AI optimised storage for deep learning acceleration and inference

SIMPLE, FAST, AFFORDABLE AI OPTIMISED STORAGE...

...FOR NVIDIA BASED SOLUTIONS

Simple Connectivity, Flexible Design and Solutions, one DGX configuration, or a PNY AI CLUSTER up to 10 nodes, no need for multiple storage nodes or controllers, everything needed is contained and automated within the single appliance





PNY GRAPHICS | HPC | AI | PRO

Contents

Executive Summary

NVIDIA's DGX range has helped shape the AI landscape and changed future possibilities for organisations and research teams throughout the world. As the market has scaled, so has the need for simplified and ratified solutions which will partner the NVIDIA ecosystem and accelerate artificial intelligence (AI) and machine learning (ML) projects, as well as the ultimate stage of inference.

This reference design details a unified solution based upon PNY's AI Optimised Storage, NVIDIA DGX A100's and NVIDIA Mellanox Quantum InfiniBand and Ethernet switches.

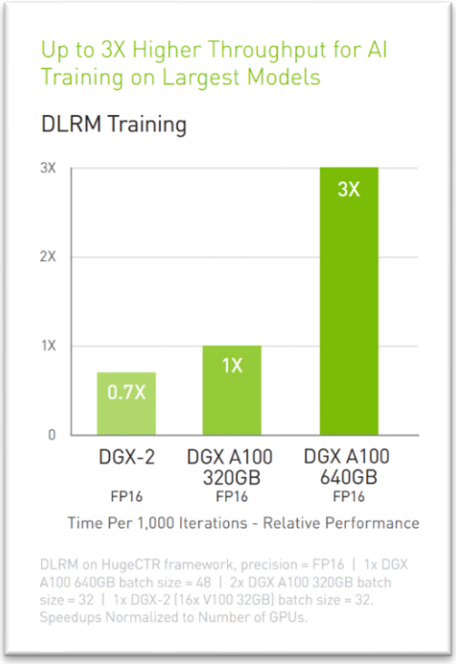
The operation and performance of the detailed solution were validated using standard benchmark tools as well as NVIDIA specific tools.

The PNY AI Optimised Storage clearly demonstrates top-tier performance and ensures maximum NVIDIA DGX A100 GPU efficiency by sustaining the demanding workloads.

Introduction

The NVIDIA DGX A100

The DGX A100 is the universal system for all AI workloads, offering unprecedented compute density, performance, and flexibility in the world's first 5 petaflop AI system. The NVIDIA DGX A100 features the world's most advanced accelerator, the NVIDIA A100 Tensor Core GPU, enabling enterprises to consolidate training, inference, and analytics into a unified, easy-to-deploy AI infrastructure.



NVIDIA DGX A100 features eight NVIDIA A100 Tensor Core GPUs, which deliver unmatched acceleration, and is fully optimized for NVIDIA CUDA-X™ software and the end-to-end NVIDIA datacentre solution stack. NVIDIA A100 GPUs bring a new precision and provides 20X higher floating operations per second (FLOPS) for AI compared to the previous generation.

Working at such high levels of performance has demanded a new generation of storage, optimised to deliver ultra-low latency and able to sustain the incredible bandwidth needed to maintain the GPUs running at full.

The PNY 3S-Storage Solution

The need for a change

Today's AI servers consume and analyse data at much higher rates than many traditional storage solutions can deliver, resulting in low GPU utilisation and dramatically extending training times decreasing productivity. In many ways, this fast growth has caught the storage world by surprise.

The standard approach for a traditional vendor to supply the performance needs has been to shoehorn Enterprise class storage along with their unnecessary features into the AI solution. This simply leads to significant expense which dramatically impacts the budget for GPU resource.

A second approach is to utilise complex and expensive parallel filesystems designed for HPC environments which are driven by often hundreds of servers and not small, often isolated pockets of DGX servers. Such solutions not only have been designed for large scale, but also require advanced storage administration and skills, or expensive vendor supported solutions.

PNY Storage Evolution

PNY, NVIDIA's global partner, has been working with PEAK:AIO to develop solutions from the ground up for AI workloads and optimised for the NVIDIA DGX range of AI appliances. Delivering ultra-low latency and tremendous bandwidth at a price which allows more investment to be made on GPU resource and less on expensive, slower storage.

The PNY AI Optimised Storage creates a central pool of ultra-low latency NVMe which can be shared amongst one or multiple DGX servers. Providing each DGX with the ideal level of resource without the need for upfront over investment. Simply connected via NVIDIA compatible InfiniBand / Ethernet, the unique RDMA protocol ensures the NVMe resource is seen and performs as if it were internal to the DGX.

Blisteringly Fast Performance for AI Workflows, and More Budget for GPU's

Ensuring your project's funds are better spent on GPUs and your team are more productive by taking full advantage of the DGX capabilities.

Start Small. Scale Only When Needed.

With many new AI projects and inference solutions requiring only limited amounts of storage, the PNY 3S-Storage range starts from 30TB, while still delivering full performance. With a 1U delivering up to 150TB and a 2U capable of housing a massive 360TB, starting small and scaling as needed is simple. And should a project require larger capacities, additional expansion units are available.

Flexible Fault Tolerance

PNY appliances feature variable RAID protection which provides various levels of resilience. With the recommended configuration being RAID5, with RAID 1, 10, 5 and 6 all being supported.

The Connectivity

Keeping it simple and compatible

Maintaining focus on NVIDIA, the PNY 3S-Storage solution is based upon the NVIDIA Mellanox CX-6 HDR/200GbE range for maximum compatibility. Each 1U or 2U has 2 x HDR/200GbE QSFP56 Ports.

The Protocols and Filesystems

The PNY 3S-Storage supports both Block storage as well as a Network Filesystem with multiple protocol options:

NFS-over-RDMA

The Network File System (NFS) has traditionally been a popular protocol choice for accessing files remotely over networks within enterprise environments due to its simplicity. For AI however, traditional, or standard enterprise level NFS has historically been too slow for such high-performance requirements.

PNY and PEAK:AIO have focused on an evolved generation of NFS which takes advantage of the advances within RDMA support by NVIDIA and Mellanox.

With the addition of RDMA support, PNY's 3S-Storage can now deliver much greater performance capable of sustaining the demands of multiple DGX hosts.

The PNY 3S-Storage supports RDMA NFS over InfiniBand and Ethernet (RoCE). The latter being a popular choice for newer start projects due to the wider use of Ethernet.

NFS-over-TCP

Where RDMA may not be an option, whether through network limitations or none RMDA supporting Linux hosts, the PNY 3S-Storage provides an enhanced variation of TCP based NFS which can still provide full bandwidth performance and maximum compatibility.

NVMe-over-Fabric (NVMe-oF) provides ultra-low latency to centralised NVMe flash. Block storage tends to be used with local filesystems such as XFS and therefore generally does not support data sharing and mainly used within single or non-data sharing DGX configurations. Due to the nature and design of NVMe-oF, it will offer the lowest latency and highest performance, however, lacks cluster scalability from the host perspective.

NVMe-oF is often used for new projects which mainly consist of a single DGX server and is an ideal new start protocol. It is also useful to provide for example a small scratch pool or ultra-high-performance capacity to a DB application or standalone project which does not require data sharing.

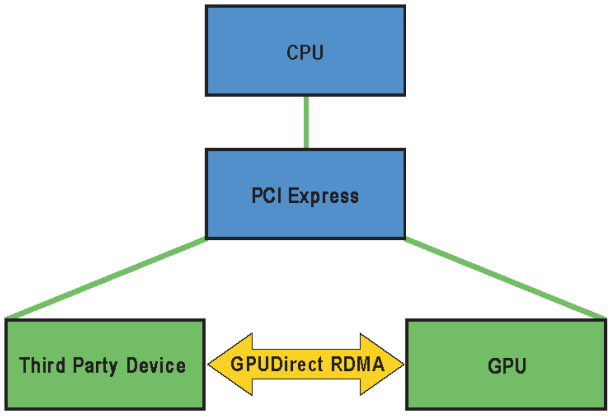
The PNY 3S-Storage incorporates a unique block to network filesystem conversion, allowing users to benefit from NVMe-oF in the early stages and convert to i.e. RDMA NFS as needed.

NVIDIA GPUDirect Storage Support

Enhancing Data Movement and Access for GPUs

Whether you are exploring mountains of data, researching scientific problems, training neural networks, or modelling financial markets, you need a computing platform with the highest data throughput. GPUs consume data much faster than CPUs and as the GPU computing horsepower increases, so does the demand for IO bandwidth. Meaning that within super computers such as the NVIDIA DGX range, the CPU's are a bottleneck to the GPU's should data need pass through or be handled by them.

NVIDIA GPUDirect® is a family of technologies, part of Magnum IO, that enhances data movement and access for NVIDIA data centre GPUs. With PNY enabled GPUDirect, unlike traditional storage solutions, data can be directly read and write to/from GPU memory, completely bypassing the CPU, eliminating unnecessary eliminating unnecessary memory copies, decreasing CPU overheads and reducing latency, resulting in significant performance improvements.



PNY and PEAK:AIO have support GPUDirect for both RDMA NFS and NVMe-oF

Technology Requirements

The following details the hardware and software used for all the testing described in the Solution Verification Section.

Hardware Requirements

The below table details the hardware components that were used to verify this solution.

Product	Quantity	Notes
NVIDIA DGX A100 System	4	Each DGX Containing 8 x A100 GPU's
PNY 3S-1050-0907	4	Each 1U containing 9 x 7.69TB NVMe Flash Drives
NVIDIA QM8700 Switch	2	Mellanox InfiniBand HDR Switch for Compute Network
NVIDIA SN3700 Switch	2	Mellanox Ethernet 200GbE Switch for Storage Network

Table 1 Hardware Requirements

Software Requirements

The below table details the software components that were used to verify this solution.

Software	Version	Notes
NVIDIA DGX A100 OS	5.0.2	Default NVIDIA Ubuntu
Docker	19.03.14	
OFED (Mellanox)	MLNX_OFED_LINUX-5.4-1.0.3.0	Mellanox OFED
GPUDirect	1.0 GA	
PNY 3S-Storage OS	1.0.444	PEAK:AIO

Table 2 Software Requirements

Solution Overview

The key design focus of the PNY PEAK:AIO implementation was to deliver the performance associated with large scale, enterprise and HPC class storage, with a much simpler implementation and price point. Removing the complexity and need to over invest in storage or storage administration and support.

The PNY solution is specifically designed to scale from a project early stage to a larger POD or edge solution. Given most projects start with a single DGX, it was essential that the PNY solution delivered the performance even at a small entry point, as low as 30TB.

The solution architecture tested and shown below is a typical four DGX POD and a consistent NVIDIA storage reference architecture.

Although the PNY solution is fully compatible with lower latency InfiniBand HDR, a 200GbE network was utilised within this test for the storage network.

This configuration was used to test and benchmark, RDMA NFS and NVMe-oF with one to four DGX's.

The PNY does not require a controller node and each NVMe storage node acts as the entire appliance, reducing cost and complexity.

The tests clearly show the performance characteristics required to ensure a productive and efficient solution demanded across a representative set of AI workloads.

Hardware Configuration

The below diagram details the hardware and its connectivity which was used to verify this solution.

Four PNY 1U's solutions where connected utilising NVIDIA Mellanox CX-6 ports, in this test 200GbE was used, however, HDR InfiniBand is equally supported.

Four DGX's used HDR InfiniBand for compute fabric

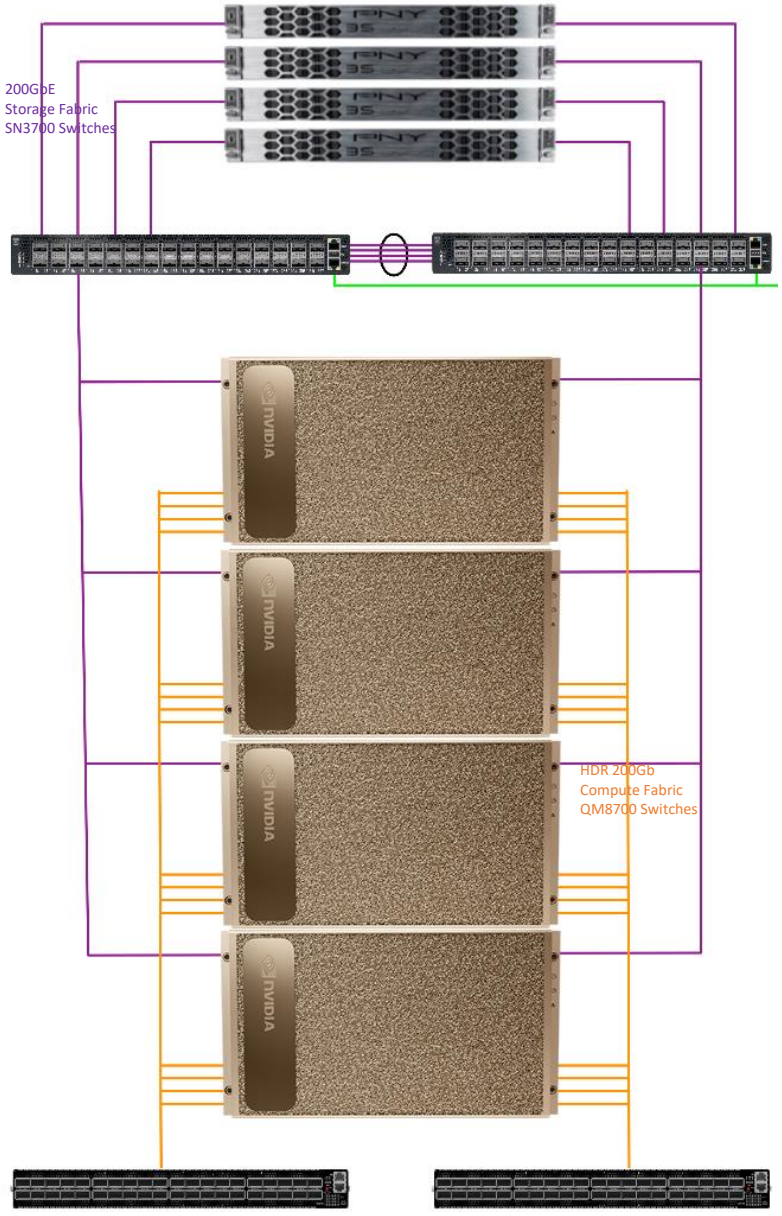


Figure 1 Solution Configuration

Testing Methodology and Results

This solution's configuration was tested by measuring each benchmark as defined further below while scaling from one to four DGX A100 systems against PNY 3S-Storage appliances.

Test	Source
NVIDIA NCCL all_reduce_perf test	GitHub - NVIDIA/nccl-tests: NCCL Tests
FIO	1. fio - Flexible I/O tester rev. 3.27 — fio 3.27-24-gd3dac-dirty documentation
GPUDirect Test - GDSIO tool	https://docs.nvidia.com/gpudirect-storage/configuration-guide/index.html
MDtest – Metadata Performance	GitHub - hpc/ior: IOR and mdtest
MLPerf – ResNet 50	GitHub - mlcommons/training: Reference implementations of MLPerf™ training benchmarks

NCCL Tests: all_reduce_perf

The NVIDIA Collective Communications Library (NCCL) tests the maximum scalability across multiple DGX A100 systems. With a single system, the bottleneck is the bandwidth of the NVIDIA NVLink GPU interconnect.

Across multiple systems, the bottleneck should be interconnecting fabric and reach close to maximum bandwidth. Figure 2 shows the single system inter-GPU bandwidth reaches NVLink interconnect capabilities. Multi system inter-GPU bandwidth reaches aggregate bandwidth of all InfiniBand ports assigned to the test

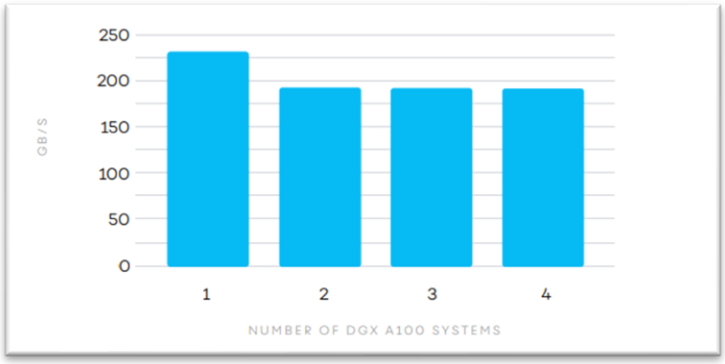


Figure 2 NCCL Test Results

The results of the NCCL test show a single DGX A100 system's inter-GPU bandwidth reaching the maximum throughput of the internal NVLink. Scaling to four, the inter-GPU bandwidth across multiple systems approaches the aggregate bandwidth of the eight InfiniBand network adapters allocated for compute per DGX A100

FIO Bandwidth

Flexible IO tester (FIO) is an open-source synthetic benchmark tool consistently used within the storage industry to profile storage performance. FIO will generate a wide range of workloads and a full performance profile will generally consist of tests with a block, data or file size ranging from 4KB to 1MB. The latter generally used to demonstrate bandwidth capability and an indication maximum performance.

The below FIO test was ran with the variables:

ioengine: libaio, iodepth: 64, numjobs: 18, direct=1, blocksize=1024K, r/w 100% Seq

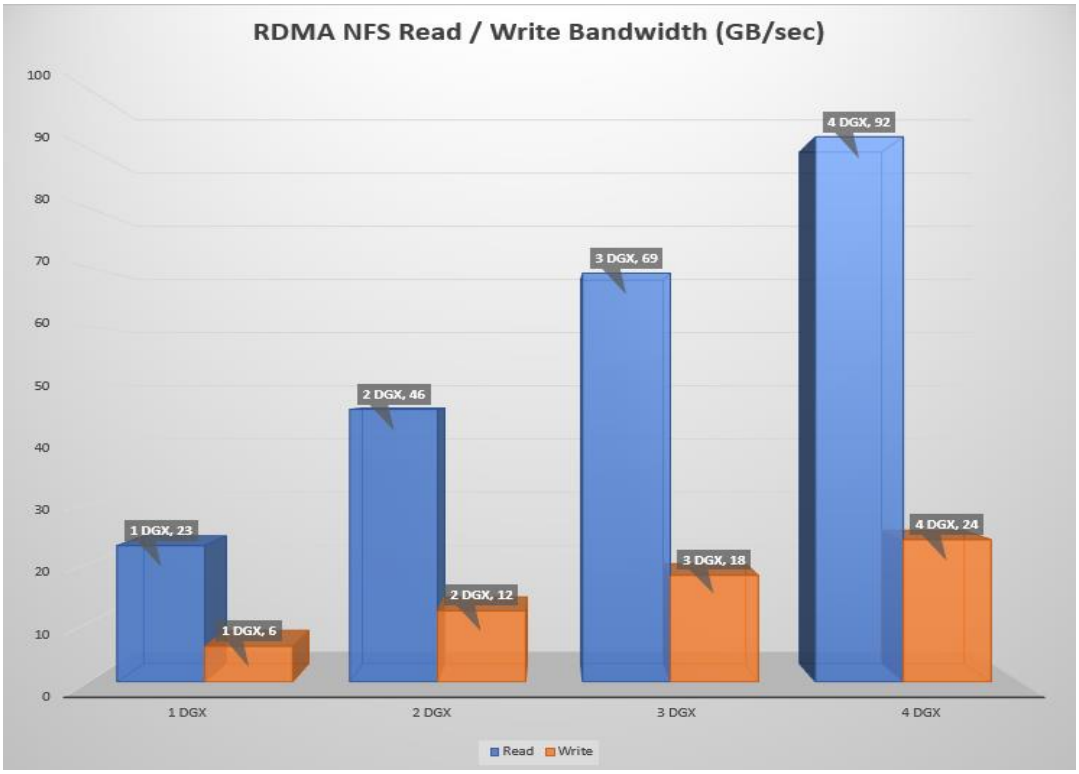


Figure 3- FIO Performance Results

GPUDirect Bandwidth

NVIDIA® GPUDirect® Storage (GDS) is the newest addition to the GPUDirect family. 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 utilisation load on the CPU.

There are several storage benchmarking tools and utilities for Linux systems, since GDS is relatively new technology, with support dependencies and a specific set of libraries and APIs that fall outside standard POSIX IO APIs, none of the existing storage IO load generation utilities such as FIO include GDS support.

As a result, the installation of GDS includes the `gdsio` load generator which provides several command line options that enable generating various storage IO load characteristics via both the traditional CPU and the GDS data path.

The below shows a 1MB/sec read/write bandwidth test, which as can be seen for reads maximises the link speeds. However, and more importantly, the data is read directly into the GPU buffer and so although the performance looks similar on graph, the impact of GPU efficiency can be significant.

The PEAK:AIO system coupled with PNY believe GPUDirect will in time be a significant factor for GPU based storage and such placed priority on its integration of co-development.

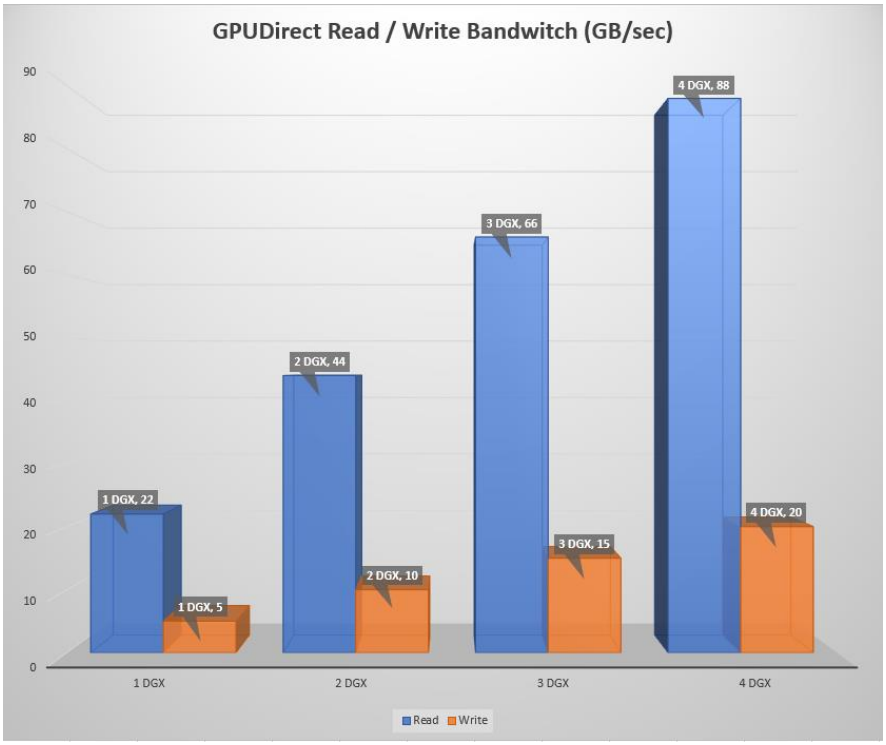


Figure 4 GPUDirect Performance Results

MLPerf Training v0.7 – ResNet-50

MLPerf Training is an industry-standard benchmark suite to test the performance of systems while training models to a target quality metric. The below is based on the standard ResNet-50 neural network, a well-known image classification network, that can be used with the ImageNet dataset. With a quality target of 75.90% classification, this test workload is both sufficiently computationally intensive and I/O-intensive

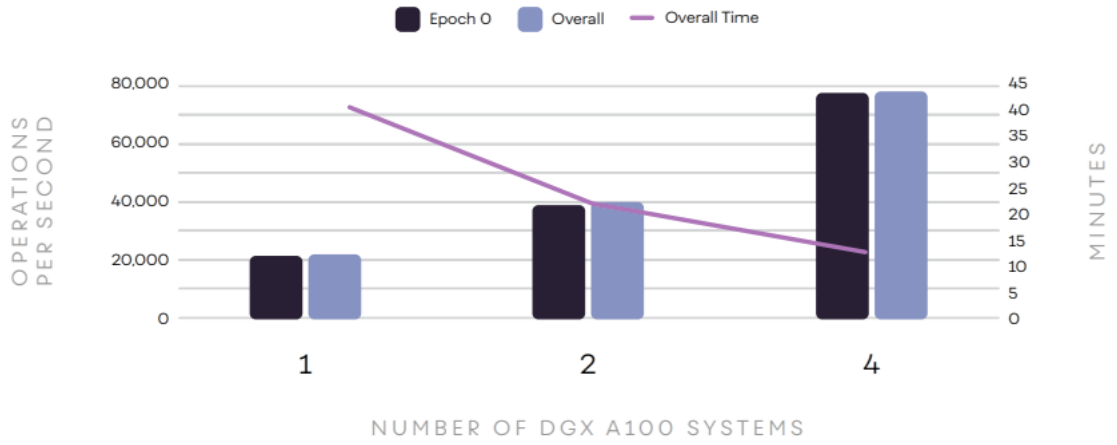


Figure 5 - MLPerf

The MLPerf test results show a close to zero difference between performance during epoch zero compared to the average overall performance of subsequent epochs. Because the first epoch reads data directly from the filesystem, local cache is used and subsequent epochs read data direct from cache, therefore the filesystem is most heavily stressed during the first epoch and has a smaller impact on the later passes.

For further information on MLPerf, see for example https://ieeexplore.ieee.org/document/9001257

Scale and PNY Storage

The PNY solution has been specifically designed to fit within the early stage DGX projects and scale as required. To achieve such, a streamlined software defined storage stack has been developed from the ground up to provide low capacity new-start projects with the performance of a HPC / Enterprise class storage solution.

As well as performance, the solution delivers a simplicity ensuring that the users and teams can focus on the project and the GPU functionality, not storage administration.

Although 75% of PNY solutions are within PODs containing 1-3 DGX servers, a single 1U can perform for four DGX A100's, multiple PNY solutions scale to larger for performance and capacity PODs.

Conclusion

AI is expanding at a tremendous pace and with evolving innovations within DL solutions, is becoming common place in a wide range of industries.

Healthcare is spearheading change with world challenges and implementations are proving invaluable daily. PNY and PEAK:AIO are proud to have played a significant role within many such projects with healthcare solutions around the world already utilised NVIDIA and PNY storage used to solve live changing challenges.

It is clear that there are efficiency and commercial advantages to organisations who invest in AI and turn their data into intelligence which in turn drives newer approaches to business.

While many teams, projects and organisations want to kickstart their AI initiatives, challenges building a scalable and AI-optimised infrastructure can be restricted by funds and breakthrough is often held them back or slowed down. Traditional compute infrastructures are not suitable for demanding AI workloads due to slow legacy CPU architectures. PNY and NVIDIA partnered to architect a scalable and powerful infrastructure that pushes the boundaries of AI innovation and performance. The results show robust linear performance scalability from one to four DGX A100 systems and beyond, allowing organisations to start small and grow seamlessly as AI projects ramp.

The results demonstrate that scaling GPU infrastructure to accelerate time to insights will be well supported by the PNY PEAK:AIO Solution.

Most importantly, PNY and PEAK:AIO have focused on simplicity, AI is often not driven by longstanding IT teams, and the primary design focus has been to provide solutions that do not require specialised storage or filesystem knowledge. A simply plug-n-play appliance for a GPU focused solution.

Acknowledgments

The authors gratefully acknowledge the contributions that were made to this technical report by our esteemed colleagues from NVIDIA and PNY. Our sincere appreciation and thanks go to all the individuals who provided insight and expertise that greatly assisted in the research for this paper

Where to find additional information

To learn more about the information that is described in this document, review the following resources:

Solution	Location
PNY 3S-Storage	https://www.pny.eu/promo/pny3s-storage-servers
NVIDIA DGX A100	https://www.nvidia.com/en-gb/data-center/dgx-a100/
NVIDIA A100 Tensor Core GPU	https://www.nvidia.com/en-gb/data-center/a100/
NVIDIA Mellanox Spectrum SN3700	Spectrum SN3000 Open Ethernet Switches | NVIDIA
NVIDIA Mellanox Quantum QM8700	NVIDIA Quantum HDR 200Gb/s InfiniBand Smart Edge Switches | NVIDIA