HPE PERFORMANCE WHITE PAPER

PEAK:AIO AI DATA SERVER

Anton Gavriliuk, Hewlett Packard Enterprise operated by Sophela
Mark Klarzynski, PEAK:AIO

This document provides a comprehensive overview of the testing environment, employed technologies, and attained results during the execution of performance tests. The primary objective was to validate the compatibility of the HPE ProLiant DL380 Gen11 with PEAK:AIO's AI Data Server software.

The ultimate aim was to evaluate the solution's performance potential within GPU cluster environments and to substantiate PEAK:AIO's assertion that this solution serves as a replacement for conventional storage systems and parallel file systems in AI installations.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The following paper highlights the performance results of PEAK:AIO's AI Data Server software in collaboration with HPE's ProLiant DL380 Gen11.

The experimentation revolved around employing a single HPE ProLiant DL380 Gen11 as the foundational hardware for PEAK:AIO. Despite deploying only one storage node, the paper magnifies a substantial increase in performance metrics compared to those published by prominent storage vendors utilizing multiple nodes.

The tests were carried out over four weeks and specifically focused in on the performance advantages presented by the PEAK:AIO AI Data Server, a software solution tailored for the AI domain.

This software holds the ambitious intention of replacing conventional and outdated storage hardware solutions. Impressively, PEAK:AIO, when coupled with HPE's latest hardware generation, exhibited HPC-level performance through a singular node. It's worthwhile noting that PEAK:AIO is focused on AI implementations and does not promote HPC, despite these remarkable performance feats.

The tests were conducted by Anton Gavriliuk, technical consultant, Hewlett Packard Enterprise operated by Sophela.

THE EVOLVING MARKET

A new era of AI dominance is sweeping across various industries, driven by remarkable advancements in deep learning. However, for numerous new-starts, research teams and even large organizations, breaking into this arena poses significant challenges when it comes to selecting the infrastructure to kick-start or scale an AI project.

Whether you are training generative AI models, creating medical breakthroughs, researching scientific problems, or modelling financial markets, GPUs are evolving and driving faster and faster results, providing that the data throughput can keep up. However, today's GPU driven AI systems consume and analyze data at much higher rates than many legacy storage solutions can deliver, resulting in low utilization of expensive GPU resources and dramatically extended training and project times.

Defining GPU servers and networking options is straightforward. Yet, when it comes to storage, things get complex. A lack of new, AI specific storage can drive us towards choosing between traditional feature-rich enterprise storage or HPC-focused parallel file systems, predominantly because they can deliver the performance and are a known solution. However, both options are costly, demanding to manage, and not designed for AI-scale solutions.

WHY STORAGE MATTERS

If your aim is to center your efforts on AI innovation, strategic use of budget to gain optimal outcomes and ensuring your resources remain sharply focused on your project is of utmost significance. The last thing any scaling AI innovation needs is to waste budget and resource on unnecessary storage features or storage administration.

PERFORMANCE

Whatever your AI project, it will demand performance, and as you scale, it will demand more. Implementing the correct solution at the start ensures you can continue to focus on innovation and not drift towards a storage challenge.

Here are some specific examples of how storage performance can impact AI outcomes:

- A study by Google found that using a true high-performance storage device can reduce the training time of an AI model by up to 80%.
- A study by IBM found that ultra-fast data access can improve the performance of an AI application by up to 50%.
- Generative AI is placing increasing demands on data speed, whether training or to effectively chat, generate text or translate languages in real-time.
- In various studies, it was found that a 1 millisecond delay in accessing data can lead to a decrease in the accuracy of an AI model image classification (from 2% upwards)

AI BUDGET FOCUS

Traditional storage solutions which deliver AI level performance are often re-purposed from the HPC and Enterprise world, accompanied with a matching price tag. In contrast, PEAK:AIO a software package written explicitly for AI by a globally esteemed software-defined storage team, is priced for AI.

Recognizing that the majority of AI projects start small and required funds directed towards resources, innovation, and GPUs, PEAK:AIO focused on delivering the highest levels of performance even when starting at less than 100TBs. In combination with HPE, this is achieved at a viable AI price point. Reducing the cost of entry on average by 10:1 without compromising on AI performance.

Typically, clients adopt what is initially seen as an affordable hybrid approach, utilizing flash-based storage to cater to high-performance requirements and employing more budget-friendly network-attached storage (NAS) for retaining data. This approach, however, gives rise to complex and laborious data pathways, necessitating data movement between different tiers before commencing AI training. Importantly, adopting such a data flow diverts resource and momentum away from the projects core innovation and goal.

The PEAK:AIO AI Data Server, is a simple software package that converts a straightforward server, such as the HPE ProLiant DL380 Gen11, into an AI focused ultra-performance AI-NAS solution. By doing so, delivering the economics and stability of large-scale HPE production hardware with no lock into propriety storage hardware. And as this document highlights, the solution provides the performance needed for even the most demanding AI applications. Providing a full AI performance All-flash solution at a price never seen before within the market.

PERFORMANCE SUMMARY

The below is a simple performance matrix providing a summary of the following work and results within the document.

Notes: While the figures compare to leading multi-node solutions, a single, off-the-shelf, HPE ProLiant 380 Gen11 was used for the PEAK:AIO AI Data Server software (creating a single storage node).

GPUDirect® tests were conducted outside of the HPE due to the connectivity of an NVIDIA® DGX A100. The results match those found within the HPE l, the only difference seemingly the data path.

PEAK:AIO Protect Internal to DL380 Gen11	Max GB/sec
Raw drive maximum potential read (7.5GB/sec per drive)	120GB/sec. (16 drives)
Raw drive maximum potential write (2.4GB/sec per drive)	38.4GB/sec
PEAK:PROTECT RAID0 Read	117GB/sec (97.5% of max)
PEAK:PROTECT RAID6 Read	119GB/sec. (99% of max)
PEAK:PROTECT RAID6 Write	18GB/sec (47% of max)

RDMA NFS Performance (Single Storage Node, 5 Clients)	
PEAK:PROTECT RAID6 Seq. Read	118GB/sec
PEAK:PROTECT RAID6 Seq. Write	18GB/sec
PEAK:PROTECT RAID6 Random Read 4K IOPS	620K IOPS

NVMe-oF Performance (Single Storage Node, 5 Clients)	
PEAK:PROTECT RAID5s Random Read 4K IOPS	9.5Million IOPS

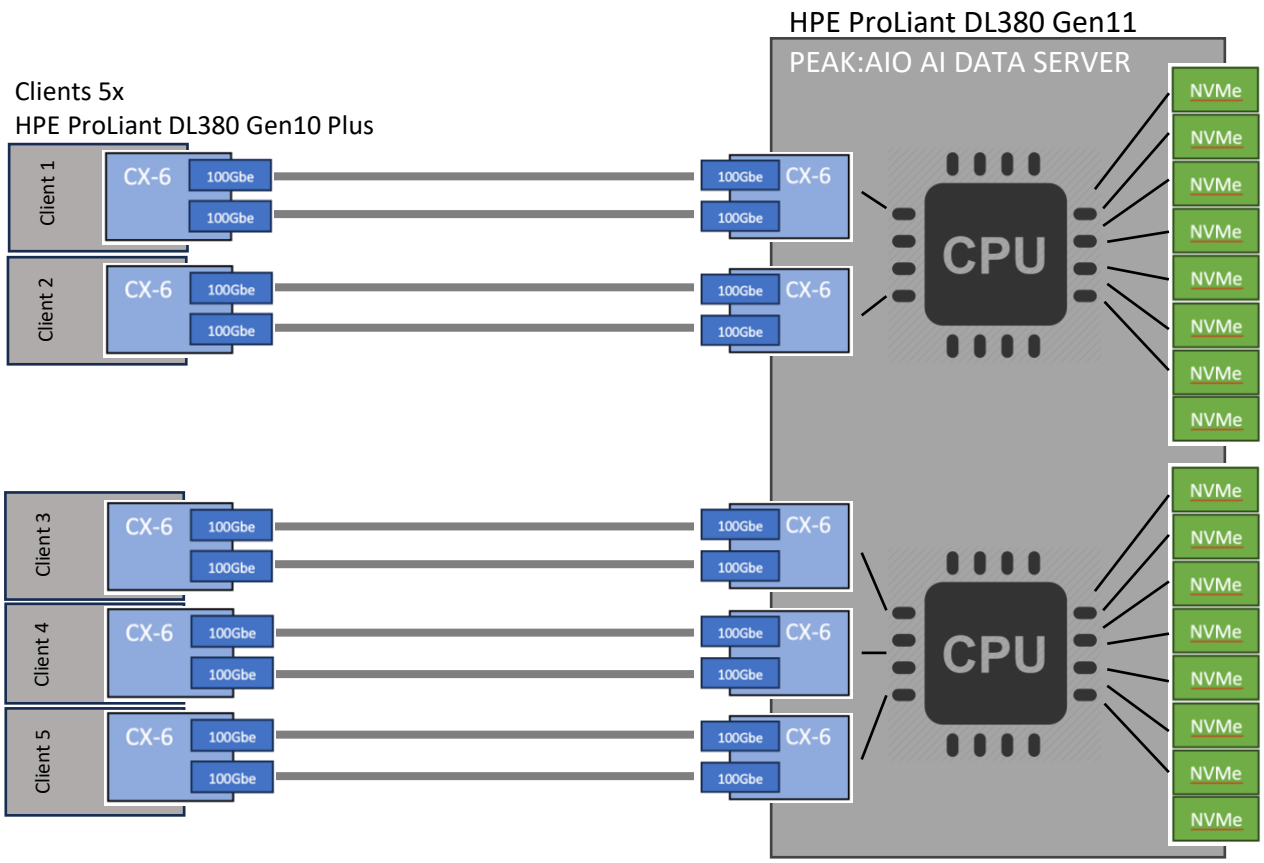
RDMA NFS Performance (Single Storage Node, 5 Clients)	GPUDirect® (GDSIO)
PEAK:PROTECT RAID6 Seq. Read	118GB/sec
PEAK:PROTECT RAID6 Seq. Write	18GB/sec

TEST ENVIRONMENT

The test environment was constructed and housed within HPE's laboratory, testing performed by Anton Gavriliuk. The hardware was housed within HPE's Customer Innovation Center. Remote access to PEAK:AIO was provided to configure and tune.

A HPE ProLiant DL380 Gen11 was used for the PEAK:AIO AI Data Server with 16 x Samsung 3.2TB SFF NVMe drives, direct attached via 4 lanes each to enable PEAK:PROTECT RAID to control the drives directly without any controller. The solution housed five NVIDIA® ConnectX®-6 dual 100Gbe port HBA's for external connectivity allowing PEAK:AIO to take advantage of its RDMA capabilities for both NFS and NVMe-oF.

The five clients were HPE ProLiant DL380 Gen10 Plus with a single NVIDIA® ConnectX®-6 dual 100Gbe port HBA and Ubuntu operating system. The ports were direct attached for the trials.



Hewlett Packard Enterprise

PEAK AIO

INTERNAL RAID PEAK:PROTECT

Standard *fio* commands were utilized to ascertain the performance of the NVMe PEAK:PROTECT N+2 (RAID6) configuration. Cache was intentionally excluded with the *direct=1* flag.

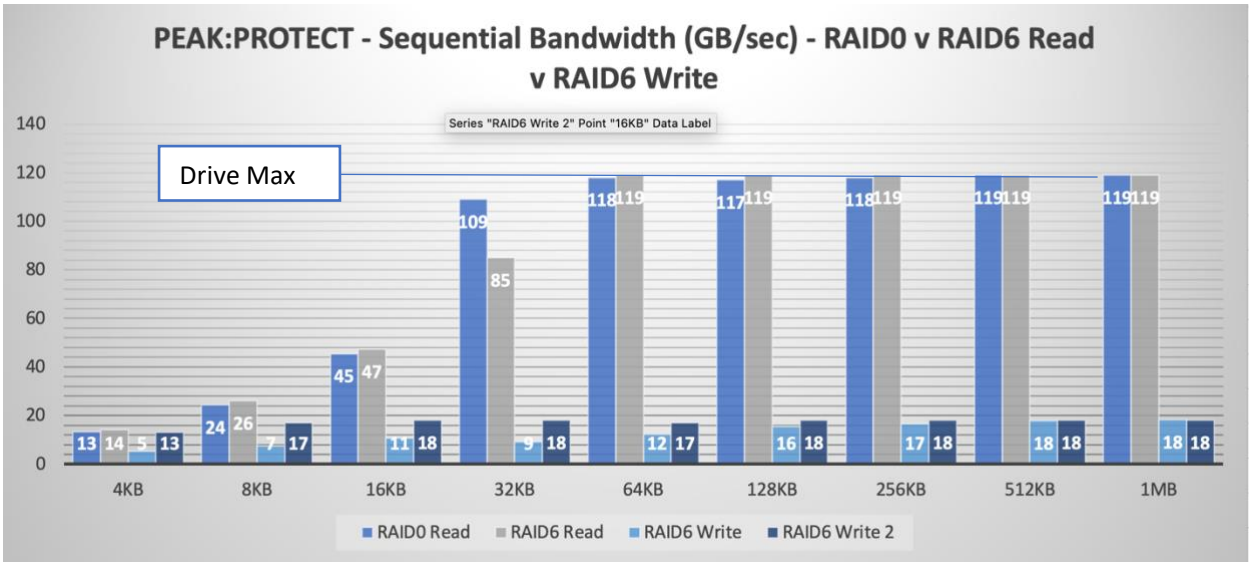
PEAK:AIO claim that their PEAK:PROTECT N+2 RAID6 implementation has been meticulously optimized for AI workloads, offering the distinctive advantage of equivalent read performance to RAID0, and can scale to the maximum of the drives capabilities. This contrasts with conventional RAID6 setups that usually encounter a performance penalty equivalent to that of two drives. Nevertheless, the subsequent tests, as showcased below, substantiate this claim by consistently attaining maximum performance across a configuration of 16 Samsung PM1735a NVMe drives.

Additional observations include:

At a block size of 32KB, a minor anomaly was detected. PEAK:AIO expresses confidence that, given more dedicated optimization time for the HPE hardware, this anomaly could be rectified.

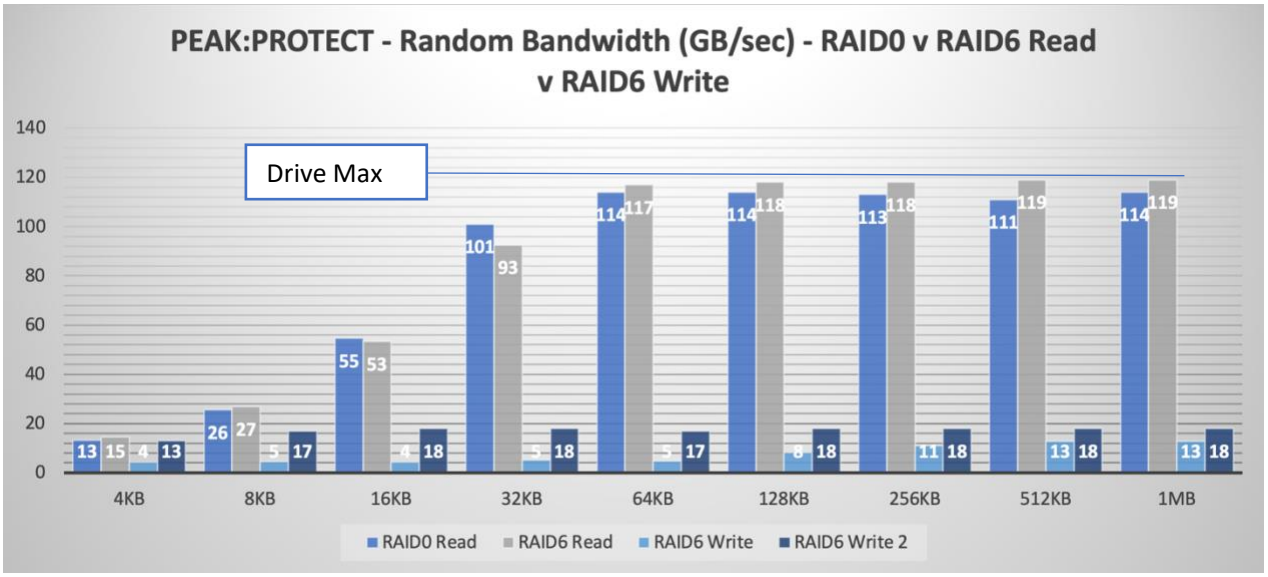
Ordinarily, 3.2TB SFF NVMe drives exhibit relatively lower write performance in comparison to their larger counterparts (2.4GB/sec v 3.8GB/sec). In the case of PEAK:AIO's external testing (external to HPE), it is evident that while the PEAK:PROTECT RAID6 configuration maintains a maximum write performance of 18GB/sec, the larger drives of 7.69TB and above scale to 18GB/sec rapidly and at a much lower block size, this is shown as 'write 2' within the following graphs.

The following graphs show sequential and random bandwidth, comparing PEAK:PROTECT RAID0 with RAID6 with sequential and random reads and writes.



Hewlett Packard Enterprise

PEAK AIO



Summary:

16 x PM1735a Gen 4 NVMe	Max GB/sec
Raw drive maximum potential read (7.5GB/sec per drive)	120GB/sec
Raw drive maximum potential write (2.4GB/sec per drive)	38.4GB/sec
PEAK:PROTECT RAID0 Read	117GB/sec
PEAK:PROTECT RAID6 Read	119GB/sec
PEAK:PROTECT RAID6 Write	18GB/sec

EXTERNAL RMDA NFS PERFORMANCE

Given the previous test highlighted the maximum possible performance from this single node was 119GB/sec, the goal was to obtain results as close as possible.

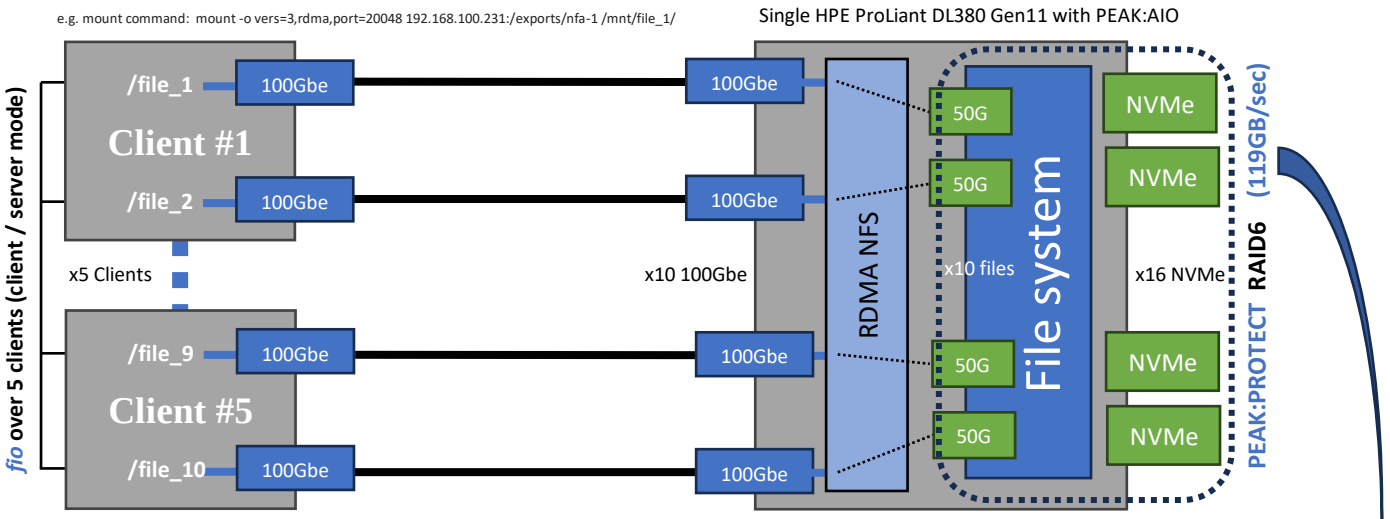
TEST ONE – RDMA NFS, Multiple files

In the first test phase, we generated ten 50GB files which were exported to the five clients over 100Gbe links utilizing the RDMA NFS protocol.

Employing *fio* in the client-server mode facilitated the execution of these tests and the aggregation of resultant data. It's important to note that the previously discussed section demonstrated the maximum achievable performance of 119GB/sec. (only 1GB/sec less than the raw NVMe maximum)

fio bypassed client cache with the direct=1 flag and showed a consistent 118GB/sec throughout the test.

Diagram and screenshots below:



Tests demonstrated **118GB/sec** (internal 119GB/sec max)

fio configuration

```
root@dl141:/home/anton# cat read-256k.fio
[global]
bs=256k
iodepth=1
direct=1
ioengine=libaio
group_reporting
time_based
runtime=300
numjobs=32
name=nfs-read-256k
rw=read

[job1]
filename=/mnt/nfs-1/fio.bin:/mnt/nfs-2/fio.bin
root@dl141:/home/anton#
```

```
filename=/mnt/nfs-1/fio.bin:/mnt/nfs-2/fio.bin
root@dl141:/home/anton#
root@dl141:/home/anton# numactl --cpunodebind=0 --membind=0 /home/an
hostname=dl145, be=0, 64-bit, os=Linux, arch=x86-64, fio=fio-3.35-10
hostname=dl144, be=0, 64-bit, os=Linux, arch=x86-64, fio=fio-3.35-10
hostname=dl143, be=0, 64-bit, os=Linux, arch=x86-64, fio=fio-3.35-10
hostname=dl142, be=0, 64-bit, os=Linux, arch=x86-64, fio=fio-3.35-10
hostname=dl141, be=0, 64-bit, os=Linux, arch=x86-64, fio=fio-3.35-10
<dl144> job1: (g=0): rw=read, bs=(R) 256KiB-256KiB, (W) 256KiB-256Ki
<dl144> ...
<dl143> job1: (g=0): rw=read, bs=(R) 256KiB-256KiB, (W) 256KiB-256Ki
<dl143> ...
<dl142> job1: (g=0): rw=read, bs=(R) 256KiB-256KiB, (W) 256KiB-256Ki
<dl142> ...
<dl144> Starting 32 processes
<dl143> Starting 32 processes
<dl142> Starting 32 processes
<dl141> job1: (g=0): rw=read, bs=(R) 256KiB-256KiB, (W) 256KiB-256Ki
<dl141> ...
<dl141> Starting 32 processes
<dl145> job1: (g=0): rw=read, bs=(R) 256KiB-256KiB, (W) 256KiB-256Ki
<dl145> ...
<dl145> Starting 32 processes
Jobs: 160 (f=320): [R(32)][98.1%][r=118G][r=462k IOPS][eta 00m:26s]
```

Screenshots courtesy Anton Gavriliuk

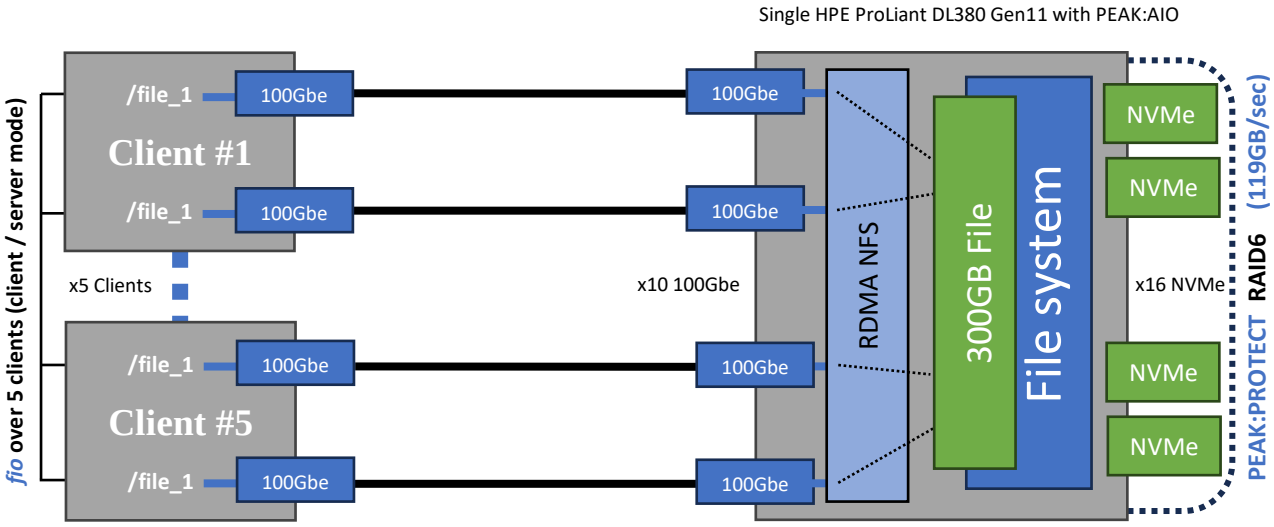
TEST TWO – RDMA NFS, Single File

In this second test phase, we generated a single 300GB file which was exported as a single share to the five clients over 100Gbe links utilizing the RDMA NFS protocol.

Employing *fio* in the client-server mode facilitated the execution of these tests and the aggregation of resultant data.

fio bypassed client cache and used the same variables as the previous test.

Diagram and result:



Within this configuration, tests demonstrated a consistent 88GB/sec to 94GB/sec, which is obviously less than the 119GB/sec potential. example *fio* output:

All clients: (groupid=0, jobs=5): err= 0: pid=0: Tue Aug 8 12:05:03 2023
read: IOPS=360k, BW=88.0Gi (94.5G)(25.8TiB/300006msec)

Following in-depth diagnostics conducted within HPE, Anton meticulously examined the deceleration observed when handling a single, larger file. This deceleration stemmed from the necessity for the large file to traverse NUMA nodes while being constantly accessed by all clients over all links, consequently limited to the CPU-to-CPU performance. A comprehensive internal assessment was undertaken, pinpointing the inherent hardware limitation of CPU-to-CPU memory reads, which caps at 96GB/sec.

Considering the dynamic scenarios encountered in real-world environments, where diverse workloads are distributed across multi-client clusters, the practicality of exclusively utilizing a single file for the entire workload is minimal. Given this context, the notable outcome is the discernible uplift, with a ceiling of 94GB/sec per file.

This limitation is perceived as a considerably positive achievement, again, reflecting the optimization efforts made to ensure robust performance within intricate operational contexts.

NVMe-oF – EVOLUTION AND USE WITHIN AI

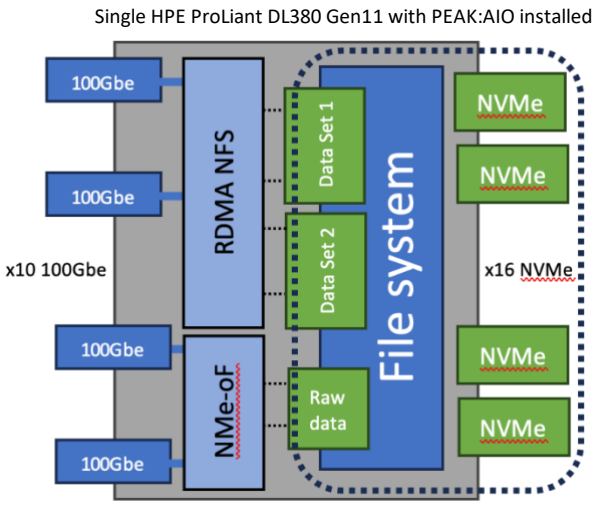
The evolution of storage technology has been marked by leaps in performance, and none have been as impactful as the transition from conventional drive interfaces to NVMe. NVMe introduced a paradigm shift from the likes of SAS-based flash drives, propelling storage capabilities to unprecedented heights. With this leap, NVMe not only brought enhancements in throughput and IOPS but, more crucially, it dramatically reduced latency—a game-changing achievement for data-intensive applications.

The logical progression was the natural evolution of NVMe into a more versatile and scalable form—enter NVMe over Fabric (NVMe-oF).

NVMe-oF encapsulates the blazing performance of NVMe while liberating it from the constraints of local servers. This decoupling of the NVMe drive from a single server ensures numerous servers can seamlessly harness the power of a centralized NVMe repository. For instance, Server_1 could be allocated 100TB of NVMe, while Server_2 benefits from 20TB and so forth.

Yet, amidst its advantages, NVMe-oF comes with caveats, particularly in the context of AI configurations. Given that NVMe operates as "block" storage, it inherently inhibits data being shared across multiple servers. Sharing a dataset between two GPU servers necessitates additional layers like distributed or shared file systems, eroding many of the inherent benefits of the NVMe architecture.

However, where NVMe-oF finds a compelling role within the realm of AI is coexisting with RDMA NFS. A tangible instance of this was found in a real-world use case involving PEAK:AIO. A drug discovery initiative employed multiple GPU servers to leverage AI algorithms for novel therapeutic approaches, utilizing shared datasets to extract valuable insights—perfectly aligned with RDMA NFS. Additionally, a singular GPU server separately processed and analyzed distinct genetic data, which was subsequently amalgamated with the shared datasets. This highly I/O-intensive operation was dramatically improved (20-fold) by using a dedicated NVMe-oF volume.



While NVMe-oF, in isolation, may have limited use in AI applications, it can be a significant advantage when paired with a high-performance sharable filesystem. It introduces accelerated options for diverse projects—be it as scratch space for analytical work or dedicated image classification tasks. PEAK:AIO uniquely addresses this versatility, allowing volumes to be exported as either NVMe-oF or NFS (RDMA or TCP), effectively merging the best of both worlds into a unified solution.

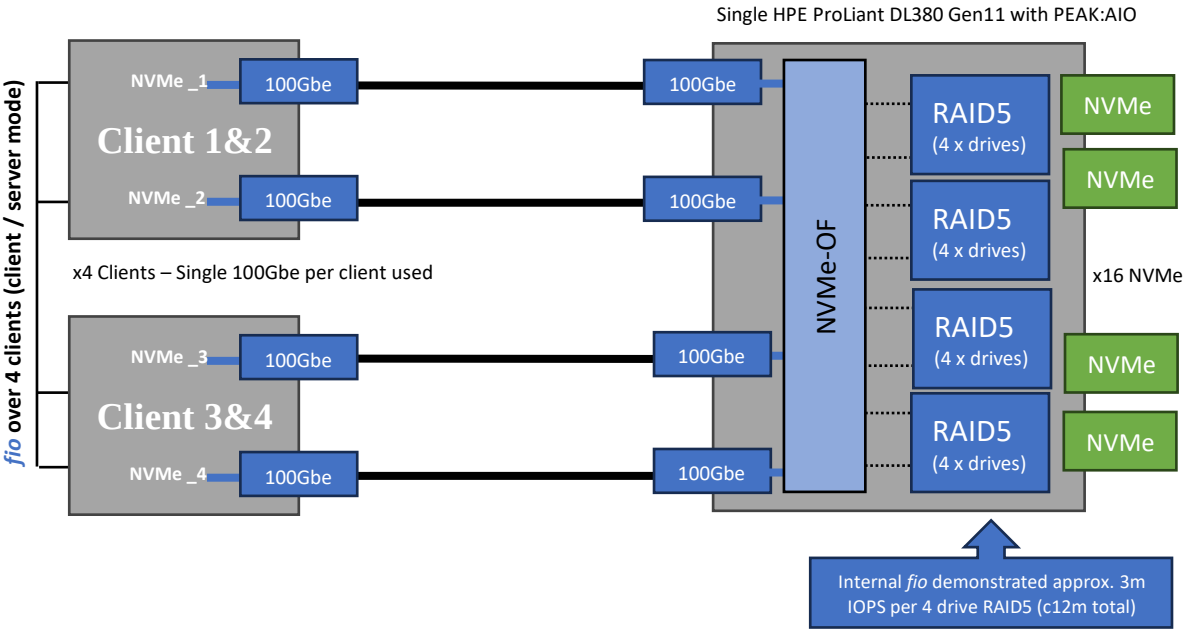
IOPS

In the RDMA NFS test, the AI Data Server demonstrated impressive bandwidth capabilities. However, it's in the realm of small file IO where shared filesystems face limitations and NVMe-oF excels, particularly IOPS at the standard 4KB scale. This prompted the creation of a specific configuration and subsequent 4K IOPS tests for accurate assessment.

RAID arrays distribute data across multiple drives, favouring aggregate throughput, yet this distribution can restrict the performance of small, random IOPS due to the need for multiple disk accesses per operation. Unlike bandwidth, IOPS scaling is not directly proportional to the number of drives in a RAID, often plateauing after a certain drive count.

For this HPE hardware test, PEAK:AIO was configured with 4 x RAID5s, connected to 4 clients. Although each client utilized a single 100Gbe link (due to cable speed limitations), standard *fio* commands for 4K random read IOPS were employed. The results exhibited a variation ranging from 9 million to 9.5 million.

While the internal performance reached 12 million IOPS, the confidence persists that utilizing ConnectX-6 cards at full 200Gbe capacity could yield even higher results. Nevertheless, achieving 9 million IOPS from a single node equipped with 16 gen 4 drives within a RAID5 configuration is a significant milestone.



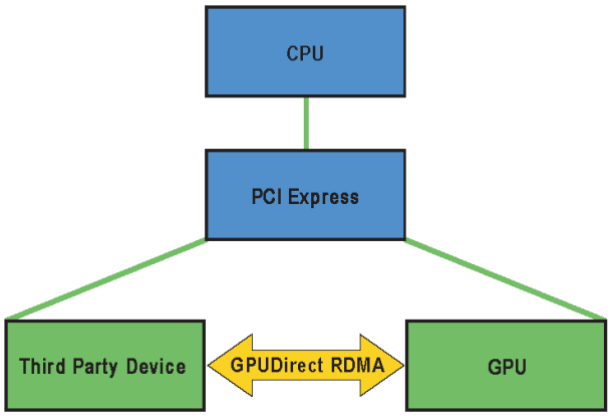
Single client *fio* command and results showing 2.4 million IOPS

```
root@dl143:~# nvme list
Node                  SN                   Model                                    Namespace Usage                      Format           FW Rev
--------------------- -------------------- ---------------------------------------- --------- -------------------------- ---------------- --------
/dev/nvme0n1          3520181254           pAio-grupa-3/slice-3                     1          1.10  TB /   1.10  TB      4 KiB +  0 B   23.09
root@dl143:~# fio --name=raw-raid5 --rw=randread --bs=4k --filename=/dev/nvme0n1 --direct=1 --numjobs=24 --iodepth=64 --exitall --group_reporting --ioengine
=libaio --time_based --runtime=300
raw-raid5: (g=0): rw=randread, bs=(R) 40968-40968, (W) 40968-40968, (T) 40968-40968, ioengine=libaio, iodepth=64
...
fio-3.35-102-g62f35
Starting 24 processes
Jobs: 24 (f=24): [r(24)][44.7%][r=9456MiB/s][r=2421k IOPS][eta 02m:46s]
```

NVIDIA® GPUDirect®

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 CPUs are a bottleneck to the GPUs 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 PEAK:AIO enabled GPUDirect®, unlike traditional storage solutions, data can be directly read and write to/from GPU memory, completely bypassing the CPU, eliminating unnecessary memory copies, decreasing CPU overheads and reducing latency, resulting in significant performance improvements.

HPE with PEAK:AIO support GPUDirect® for both RDMA NFS and NVMe-oF

NOTE: GPUDirect® testing and compatibility was determined outside of HPE due to the dependency on connection to NVIDIA® GPU servers, PEAK:AIO in this case used NVIDIA® DGX A100 servers as the clients. This information is therefore included for completeness and prospective only.

GPUDirect® COMPATIBILITY

GDS installation: The servers have been configured following NVIDIA's guidelines. For more details, see documentation at GPUDirect® Storage on the NVIDIA® Docs Hub.

GPUDirect® supports PEAK:AIO's RDMA NFS and NVMe-oF as shown with the NVIDIA® tool *gdscheck -p*

```
GDS release version: 1.0.0.82
 NVIDIA_fs version:  2.7 libcufile version: 2.4
 ============
 ENVIRONMENT:
 ============
 =====================
 DRIVER CONFIGURATION:
 =====================
 NVMe               : Supported
 NVMeOF             : Supported
 SCSI               : Unsupported
 ScaleFlux CSD      : Unsupported
 NVMesh             : Unsupported
 DDN EXAScaler      : Unsupported
 IBM Spectrum Scale : Unsupported
 NFS                : Supported
 WekaFS             : Unsupported
 Userspace RDMA     : Unsupported
 --Mellanox PeerDirect : Enabled
 --rdma library        : Not Loaded
(libcufile_rdma.so)
 --rdma devices        : Not configured
 --rdma_device_status  : Up: 0 Down: 0
```

```
=====================
 CUFILE CONFIGURATION:
 =====================
 properties.use_compat_mode : true
 properties.gds_rdma_write_support : true
 properties.use_poll_mode : false
 properties.poll_mode_max_size_kb : 4
 properties.max_batch_io_timeout_msecs : 5
 properties.max_direct_io_size_kb : 16384
 properties.max_device_cache_size_kb : 131072
 properties.max_device_pinned_mem_size_kb :
33554432
 properties.posix_pool_slab_size_kb : 4 1024 16384
 properties.posix_pool_slab_count : 128 64 32
 properties.rdma_peer_affinity_policy : RoundRobin
 properties.rdma_dynamic_routing : 0
 fs.generic.posix_unaligned_writes : false
 fs.lustre.posix_gds_min_kb: 0
 fs.weka.rdma_write_support: false
 profile.nvtx : false
 profile.cufile_stats : 0
 miscellaneous.api_check_aggressive : false
```

```
=========
 GPU INFO:
 =========
 GPU index 0 NVIDIA A100-SXM4-40GB bar:1 bar size (MiB):65536 supports GDS
 GPU index 1 NVIDIA A100-SXM4-40GB bar:1 bar size (MiB):65536 supports GDS
 GPU index 2 NVIDIA A100-SXM4-40GB bar:1 bar size (MiB):65536 supports GDS
 GPU index 3 NVIDIA A100-SXM4-40GB bar:1 bar size (MiB):65536 supports GDS
 GPU index 4 NVIDIA A100-SXM4-40GB bar:1 bar size (MiB):65536 supports GDS
 GPU index 5 NVIDIA A100-SXM4-40GB bar:1 bar size (MiB):65536 supports GDS
 GPU index 6 NVIDIA A100-SXM4-40GB bar:1 bar size (MiB):65536 supports GDS
 GPU index 7 NVIDIA A100-SXM4-40GB bar:1 bar size (MiB):65536 supports GDS
==============
 PLATFORM INFO:
==============
```


Hewlett Packard Enterprise

PEAK AIO

The gdsio load generator tool generates various storage IO load characteristics via both the traditional CPU and the GDS data path. Read/write bandwidth test have consistently shown data which matches the NFS RMDA bandwidth (in this case approx. 120GB/sec), however, with GPUDirect® 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.

CONCLUSION

A new era of AI dominance is sweeping across various industries, driven by remarkable advancements in deep learning. However, for numerous new-starts, research teams and even large organizations, breaking into this arena poses significant challenges when it comes to selecting the infrastructure to kick-start or scale an AI project.

Whether you are training generative AI models, creating medical breakthroughs, researching scientific problems, or modelling financial markets, GPUs are evolving and driving faster and faster results, if the data throughput can keep up. However, today's GPU driven AI systems consume and analyze data at much higher rates than many legacy storage solutions can deliver, resulting in low utilization of expensive GPU resources and dramatically extended training and project times.

The tests clearly show performance normally associated with HPC multi-node storage, from a single HPE ProLiant DL380 Gen11 solution. The DL380 Gen11 coupled with PEAK:AIO's AI Data Server software delivers the simplicity and performance at the price of entry needed for evolving AI projects.

ACKNOWLEDGMENTS

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