Excelero

BREAKING THE AI STORAGE PERFORMANCE BARRIER

PERFORMANCE BRIEF

NEW DATA DEMANDS REQUIRE NEW STORAGE SOLUTIONS

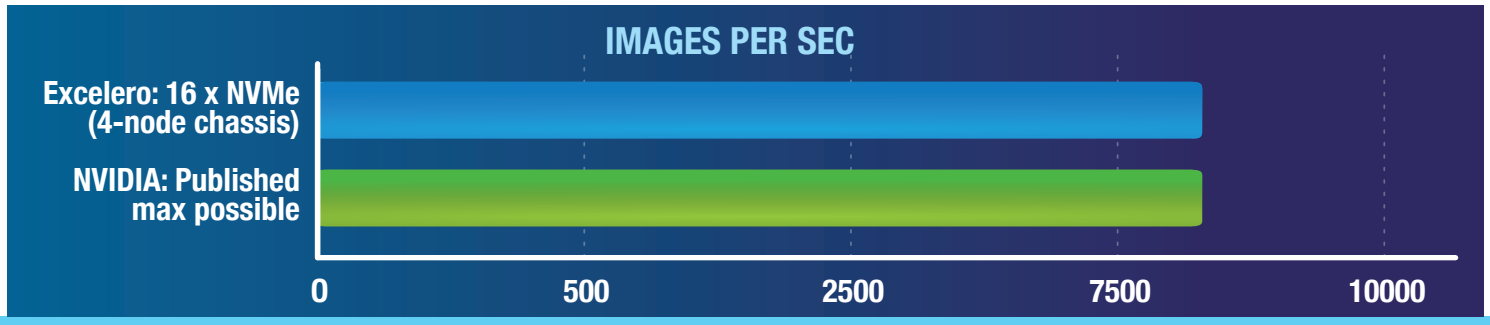
The ecosystem around GPU computing is rapidly evolving to increase the efficiency and scalability of GPU workloads. Yet, many organizations find their storage systems cannot keep up with the performance demands of their GPU systems. This is particularly true when the training phase of Deep Learning is run on traditional, controller-based storage systems.

GPU COMPUTING INTRODUCES A NEW SET OF DEMANDS ON DATA STORAGE THAT DID NOT EXIST WHEN TRADITIONAL STORAGE SYSTEMS WERE CONCEIVED:

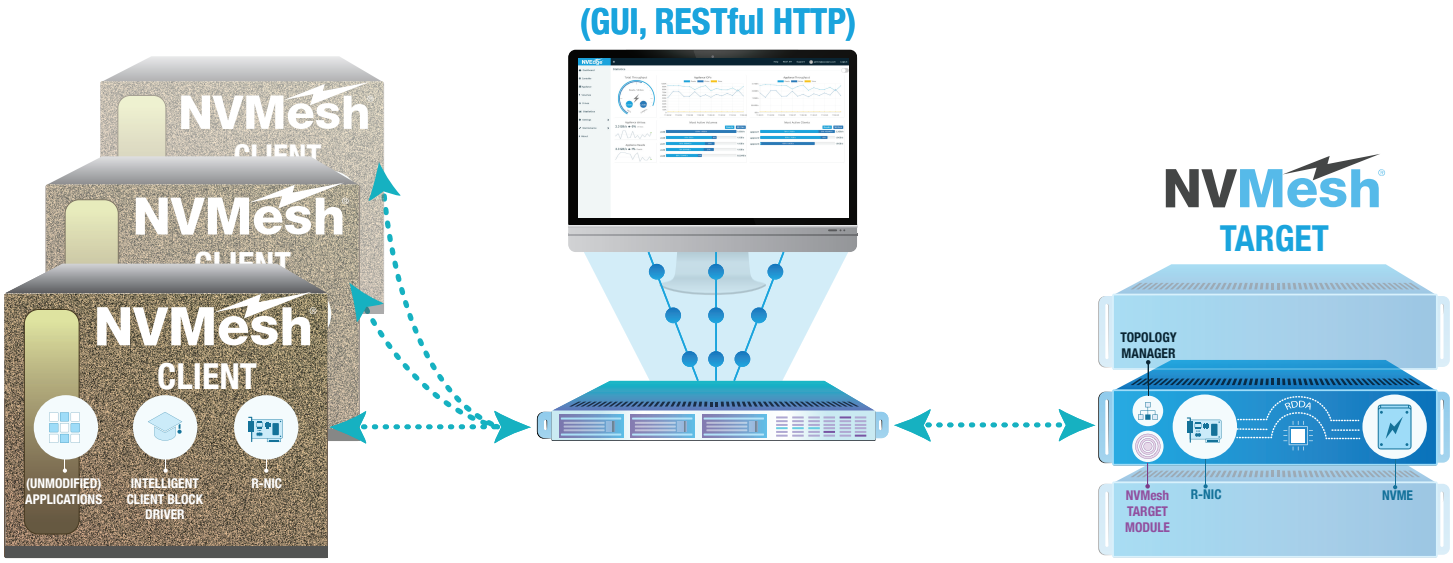
- Deploying multiple GPUs per server is a common approach to increasing computational power density, but this creates an unusually high demand on storage performance from a single client.
- Highly parallelized data access patterns, common during the training phase, require the storage system to provide a unique mix of high throughput and low access latency while simultaneously supporting a high number of small I/O operations per second.

Based on extensive customer deployments with the world's leading AI-driven organizations, and leveraging its award-winning NVMesh Elastic NVMe software, Excelero developed a ground-breaking AI storage solution that provides market-leading performance and shatters the AI storage performance barrier.

This document describes a series of internal benchmark tests we ran to demonstrate how a single NVMe storage appliance with Excelero NVMesh meets the performance requirements to maximize GPU utilization as defined by NVIDIA.



BENCHMARK ARCHITECTURE



To demonstrate the new possibilities enabled by Excelero's AI storage solution, an NVIDIA DGX-1 was connected to a standard 2U/4-node NVMe storage appliance. Each node in the appliance was equipped with two Intel Xeon SP Silver CPUs, 96GB RAM, one Mellanox ConnectX-5 VPI 100GbE/EDR InfiniBand adapter and a total of only 16 NVMe drives (4 per node).

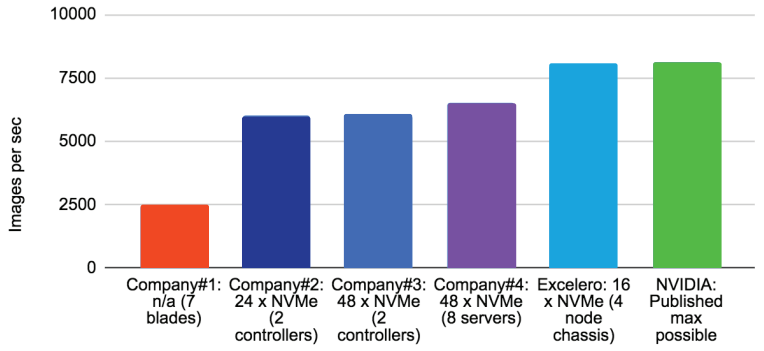
NVMesh software was installed on the storage appliance (target) and the NVIDIA DGX-1 (client). Leveraging MeshProtect, a cross-node, RAID-based data protection feature, all 16 NVMe drives were seamlessly combined into a single, high-capacity, high-performance logical volume with RAID-10 data protection. Next, the logical volume was attached to the DGX and formatted with an XFS file system. And finally, the file system was populated with the data required for use by the training applications.

Training data sets are often of static nature: they are usually updated only periodically. Excelero's AI storage solution enables customers to manage training data access for a fleet of GPU servers and deliver the highest possible local file system performance at any scale. For this purpose, it integrates with the customer's existing data pipeline orchestration tools and the NVMesh API to manage multiple read-only attachments and single read-write attachments as needed. This simple approach eliminates the complexities and costs associated with parallel file systems in favor of modern orchestration capabilities deployed as a cost-effective alternative to more traditional approaches.

BENCHMARK RESULTS

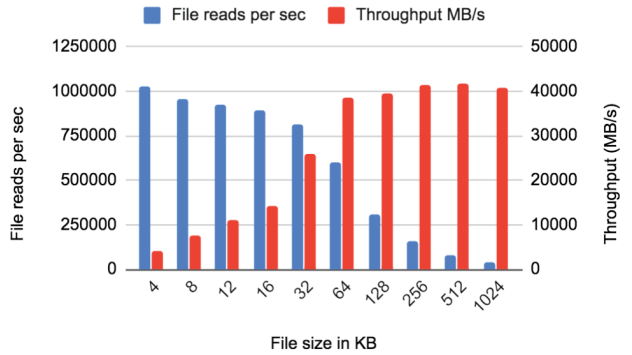
The combination of ResNet50 and TensorFlow benchmarks trained on the ImageNet 2012 dataset are a common means for establishing image training performance baselines for a given infrastructure deployment. Extensive published results exist online from multiple vendors. As *Benchmark Results 1* shows, one NVMe storage appliance with Excelero's NVMesh already provides enough performance for the full GPU potential of this model training test case.

ResNet50 Training with NVMesh vs Classic Solutions



BENCHMARK RESULTS 1

NVMesh XFS File Read Performance



BENCHMARK RESULTS 2

In practice, the application frameworks, the computational effort per input object, and the object size will differ from this ResNet50/TF/ImageNet test case. Many applications need much higher data rates and workloads often require more GPU servers connected to the same storage. Fortunately, this benchmark does not at all drive the Excelero solution to its limits and there is a lot of performance capacity left for additional GPU servers to be connected to the same storage system.

We ran a second test to demonstrate the capabilities of the Excelero AI Storage solution to achieve much higher access rates for various input object sizes, including the classic challenge of managing lots of tiny file accesses (4K to 64K). This second test used the file system test tool, mdtest. *Benchmark Results 2* shows that NVMesh is capable of delivering over a million 4KB file accesses per second (open, read, then close) and nears the DGX limit of ~40GB/s network throughput with just 64KB input file sizes. As such, the Excelero AI storage solution exceeds – by far – any available parallel file system's capabilities for the same on such little infrastructure.

CONCLUSION

The simplest solution to a challenge is usually also the best. The results presented in this document demonstrate that the Excelero AI storage solution eliminates complexity by avoiding the administrative overhead of a parallel file system. More importantly, the solution enables users to finally unleash the full performance potential of their GPU servers by closing the significant gap inherent to traditional storage architectures. Excelero's AI storage solution with NVMesh leverages standard servers and can be scaled seamlessly to any number of GPU servers and/or storage servers. It is available as a turn-key solution through a global network of partners and perfectly tailored to meet any customer's demands with ease.