Excelero

INSTADEEP™ POWERS AI AS A SERVICE WITH SHARED NVME

NVMesh® feeds unlimited streams of data to GPU-based Systems with local performance

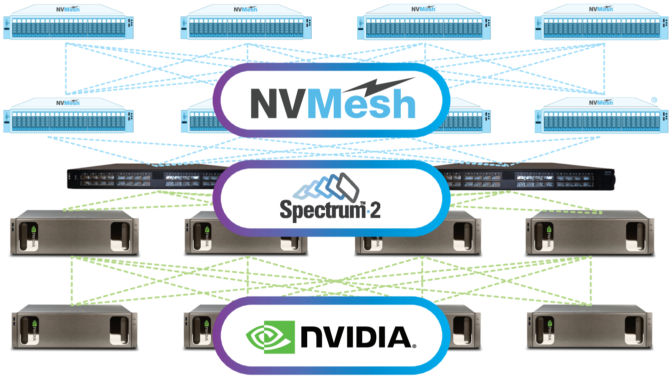
CASE STUDY

InstaDeep™ Ltd. is a global AI innovator, headquartered in London with additional offices in Paris, Tunis, Nairobi and Lagos. The company delivers AI-powered decision-making solution systems for the enterprise across a wide array of industries including Logistics, Manufacturing, Oil and Gas, Financial Services and Mobility. With full end-to-end in-house expertise from machine intelligence research through to execution and business deployment, InstaDeep provides a competitive advantage to customers in an AI-first world: the company's AI solutions allow businesses to unlock data insights, realize value, and increase efficiency and speed across the organization.

InstaDeep chose Excelero's NVMesh® on Boston Ltd. Flash-IO Talyn storage to provide GPUs with access to a scalable pool of high-performance NVMe at data center scale, ensuring full utilization of the GPU processing power. AI and ML applications running on GPU-based systems benefit significantly from NVMe storage that can feed virtually any size GPU farm with far greater performance. Excelero's NVMesh delivers low-latency (25µs), high bandwidth distributed block storage for AI and ML workloads. NVMesh enables shared NVMe across any network and supports local or distributed file systems. GPU-based systems benefit from the performance of local NVMe flash with the convenience of centralized storage while avoiding proprietary hardware lock-in and maximizing the overall GPU ROI.

Benefits of NVMesh for ML/AI:

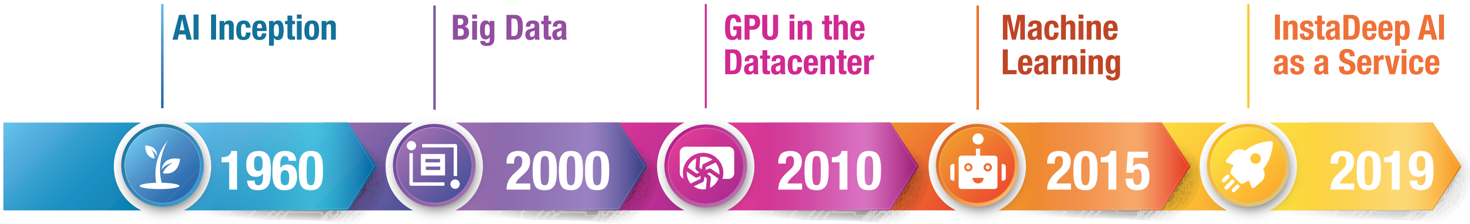
- Access remote storage at local speed
- Distributed, scalable storage infrastructure
- Shared storage resources across multiple GPU servers
- Full CPU offload on both ends
- Exceed performance limits of local flash on GPU servers
- Eliminate the need to copy data locally
- Datasets can be larger than what can fit inside the DGX



UNCAGING GPUS AND NVMe PERFORMANCE FOR AI AND MACHINE LEARNING

AI and ML use has exploded over the past few years as four key technology evolutions have made it far easier to capture, store and process data into insights that can help enterprises outsmart the competition:
1) New sensor technologies have proliferated that capture images, temperature, heartrate, and more – adding even more data volumes.
2) Big Data analytics and Data Lakes for storing these massive volumes of data arose, so that teams could analyze and mine that data to turn it into valuable business or research insights.
3) The rise of powerful GPU technologies that lower the cost of massive compute on those data sets have made parallel processing faster and much more powerful.
4) Next-gen storage options such as NVMe flash media have swept the storage industry and are well-suited to these new computational engines, although they harken back in time to the days when direct attached storage (DAS) models were new. DAS is fast, but often underutilized and hence costly.

As an AI innovator, InstaDeep offers an AI as a Service solution that democratizes the many benefits of AI and ML. It lowers the threshold to make meaningful use of AI workloads and puts them in reach of a wider range of organizations that may not have the needs or means to run their own AI stack.



InstaDeep AI as a Service

When InstaDeep decided to build an AI as a Service offering, it had a handful of essential requirements. Firstly, the data center infrastructure needed to scale modularly, as the company began offering the cutting-edge service to a few key clients, planning on rapidly expanding to the entire global client base. Secondly, the infrastructure had to be flexible to meet performance requirements for a wide range of workloads, as today the infrastructure is used by multiple scientists who run workloads for many different clients. Finally, for the service to be attractive for customers, and a sound business move for InstaDeep, the infrastructure needed to be highly efficient – the GPUs especially would need to deliver the highest return on investment (ROI).

When deploying its first NVIDIA® DGX-1™ system, the InstaDeep infrastructure team learned that the local DGX storage would be too limited; the DGX only had 4TB of local storage while customers' workloads require 10s to 100s of terabytes (TBs). The InstaDeep team investigated external storage options and noticed that with traditional arrays they would get much more capacity but the performance ultimately would hinder AI workloads since applications needed to move data to and from the DGX systems, interrupting the workflow and impacting system efficiency.

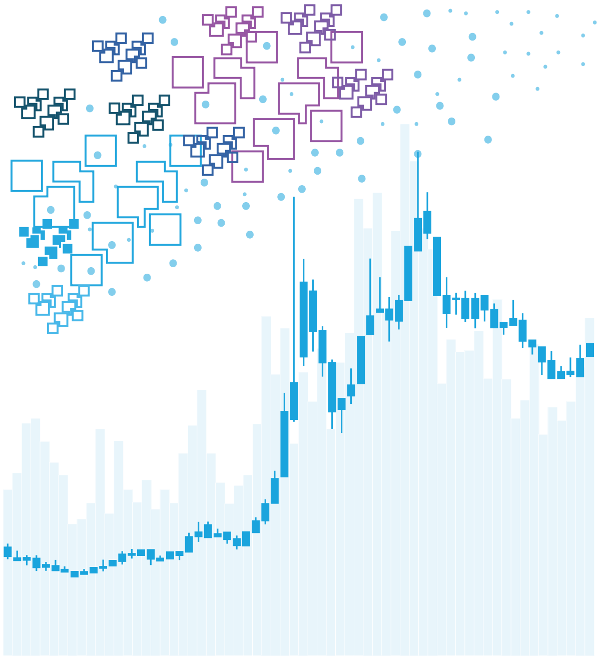
Therefore, InstaDeep chose Excelero's NVMesh on Boston Flash-IO Talyn storage to provide GPUs with access to a scalable pool of high-performance NVMe to ensure full utilization of the GPU processing power. The Talyn system included a 2U Boston Flash-IO Talyn server with Micron NVMe flash and Excelero NVMesh software that provides access to up to 100TB external high-performance storage. Leveraging the Mellanox 100GB Infiniband network cards in the DGX, the GPUs use the NVMe storage with local performance. The ability to choose any file system to run on NVMesh was an immense benefit. Early tests quickly showed that external NVMe storage with Excelero gives equal or better performance than local cache in the DGX.

NVMesh FEATURES FOR GPU

- NVMesh unifies remote NVMe devices into a logical block pool that performs the same as local NVMe flash
- NVMesh allows full utilization of the IOPs and bandwidth capabilities of NVMe drives across a network
- DGX-1 and 2 can use their massive network connectivity to access remote NVMe logical volumes, with redundancy if desired
- MUCH faster than local SATA SSDs
- Larger shared pools than possible within the platform
- Other GPU optimized systems can access remote NVMe at local latencies and bandwidth
- Random IO characteristics of NVMe preserved, achieving 10's of millions of potential IOPs at very low latencies

FEEDING THE GPU BEAST

The biggest advantage of modern GPU computing is also creating its biggest challenge: GPUs have an amazing appetite for data. Current GPU servers can process tens of gigabytes of data per second. NVIDIA's latest DGX-2™ system has as many as 16 GPUs, but by far not enough local storage. The DGX-1 has a theoretical limit of 7.8GB/s bandwidth, but with only 4 SATA SSDs it is limited to about 2.2GB/s. Theoretically, it can process 2 million random IOPs but local storage only provides 400K IOPs. The latest NVIDIA DGX-2 has 30TB (8 x 3.84TB) local NVMe but is not optimized to use it efficiently. Other brand GPU servers typically feature few PCIe lanes for local flash (NVMe or other), meaning even the lowest latency option for these servers is a severe bottleneck or is simply too little capacity for the GPUs. Starving the GPUs with slow storage or wasting time copying data wastes expensive GPU resources and affects the ROI.



Fortunately, NVIDIA's DGX nodes also have massive network connectivity. They can ingest as much as 48GB/s of bandwidth via 4-8 x 100Gb ports – playing a key part in the solution: Excelero's NVMesh enables customers to maximize the utilization of their GPUs leveraging the massive network connectivity of the DGXs and the low-latency and high IOPs/BW benefits of NVMe in a distributed and linearly scalable architecture.

Optimal Approach for Specific AI, ML Uses

NVMe flash offers great benefits for specific AI use cases like training a machine learning model, and checkpoints. Machine learning involves two phases - training a model based on what is learned from the dataset, and running the model. Training of a model is the most resource hungry stage. Hardware used for this phase, incorporating high-end GPUs or specialized system-on-chips (SoCs, is expensive to buy and operate so it should be always busy for best ROI.

The modern datasets used for model training can be huge – for example, MRI scans can reach terabytes each and a learning a system may use tens or hundreds of thousands of images. Even if the training itself runs from RAM, the memory should be fed from non-volatile storage; this storage has to support very high bandwidth. In addition, paging out the old training data and bringing in new data should be done as fast as possible to keep the GPUs from being idle necessitating low latency. The only protocol allowing for both high bandwidth and low latency like this is NVMe.

Another common use is checkpoints. If a training process is long, the system can choose to save a snapshot of the memory into non-volatile storage to allow restart from that snapshot in case of a crash. NVMe storage is very suitable for this kind of usage.

However there are limitations of how many local NVMe drives can be used – and such limitations frequently are a result of amount of PCIe lanes allocated to NVMe drives since the GPUs/SoCs also require PCIe lanes. Additionally, for checkpoint usage many local NVMe drives need to be synchronized to allow for a usable snapshot. Both of these use cases do best with distributed NVMe storage perhaps incorporating a shared/distributed file system as well.

Summary

The incredible capability of GPUs and the rise of affordable compute power, challenges IT teams to think at data center scale – leveraging the ability to apply AI, ML and Deep Learning techniques to large data pools, while making sure the entire system is scale-out, highly performant and efficient. The only storage that is fast enough to keep up with these GPUs is local NVMe flash, since GPUs, networking and NVMe are all competing for valuable PCIe connectivity, hence one of them must compromise and settle for less.

Excelero's NVMesh eliminates any compromise between performance and practicality, and allows GPU optimized servers to access scalable, high performance NVMe flash storage pools as if they were local flash. This technique ensures efficient use of both the GPUs themselves and the associated NVMe flash. The end result is higher ROI, easier workflow management and faster time to results.