AFFORDABLE DISAGGREGATED STORAGE FOR KUBERNETES AI/ ML/DL DISTRIBUTED WORKLOADS

EXECUTIVE SUMMARY

Artificial intelligence (AI) implementations based on deep learning (DL) continue to gain traction thanks to their ability to solve real-world problems. Running a distributed model training phase is key in achieving timely and accurate results and reducing training times decreases time to value.

Many implementations include copying raw datasets to drives local to the GPUs employed for training, often repetitively, as the data is updated between training cycles. This is wasteful in terms of the cost of purchasing the drives and refreshing them as their endurance is quickly eroded and precious time is wasted synchronizing the multiple copies of data required for distributed training. Finally, this process leaves little capacity for use as local scratch space.

Disaggregated storage can alleviate these shortcomings if it can meet the performance requirements in a cost-effective way. Excelero NVMesh software defined storage can facilitate this by providing a Kubernetes (K8s) CSI driver that is easy to integrate with most workflows. By leveraging NVIDIA® Mellanox® NVMe Software-defined Network Accelerated Processing (SNAP) technology, a K8s CSI is not required since the remote drive is emulated to the workers as a local NVMe drive. This makes deployment even easier by avoiding the need for any local agent.

This paper is based on a reference architecture with the functional elements and demonstrates the performance benefits it brings to a typical AI training session.

NVIDIA AND EXCELERO PARTNERSHIP

NVIDIA is the leading supplier of end-to-end Ethernet and InfiniBand intelligent interconnect solutions and services for servers, storage, and hyper-converged infrastructure. NVIDIA intelligent interconnect solutions increase data center efficiency by providing the highest throughput and lowest latency, delivering data faster to applications, and unlocking system performance.

Excelero delivers low-latency distributed block storage for web-scale applications. Founded in 2014 by a team of storage veterans and inspired by the Tech Giants' shared-nothing architectures for web-scale applications, the company has designed a software-defined block storage solution that meets the low-latency performance and scalability requirements of the largest web-scale and AI applications.

NVIDIA and Excelero have fostered a productive partnership that provides end-users the most efficient usage of their storage media across Ethernet and InfiniBand networks.

HIGHLIGHTS

- Efficient storage disaggregation supports AI/ML/DL distributed workloads using:
  - Excelero NVMesh® Technology
  - NVIDIA Mellanox BlueField® Data Processing Unit (DPU) with NVMe SNAP™
- Software-defined storage virtual volumes exposed to bare metal K8s workers as local NVMe drives using NVMe SNAP means:
  - K8s CSI is not required
  - Clientless solution for ease of deployment
- Single shared volume that is easily exposed to a distributed set of workers without any performance penalties
- Simple RoCE configuration
- Fastest time to ROI in the market

DL CHALLENGES ADDRESSED

To meet the challenges of DL, NVIDIA and Excelero offer the market a highly efficient, hardware-accelerated, and tightly integrated and disaggregated software-defined storage solution that combines Excelero NVMesh with NVIDIA Mellanox ConnectX®-5 SmartNIC network adapters, NVIDIA Mellanox BlueField® Data Processing Units (DPUs) with NVIDIA Mellanox NVMe SNAP technology and NVIDIA Mellanox Spectrum® switches. Instead of using vendor lock-in strategies with patented and tightly coupled dedicated hardware and software, which tend to be static and difficult to scale or commercial-off-the-shelf (COTS) network infrastructure that have limited offloads and accelerators and drop packets, they are replaced with NVIDIA Mellanox Ethernet Storage Fabric (ESF) components optimized for high-bandwidth and low latency and offer a predicable storage aware fabric that boosts efficiency. The joint solution runs workflows that utilize standard GPU such as NVIDIA V100 GPUs and storage software running on COTS servers.

With this set of hardware and software in place, training DL can be done with the highest efficiency on GPUs at any scale. NVMesh technology enables setting up the training environment with minimal initial investment, with a small set of data on minimal media, and capable of linear performance and cost growth to any scale. NVIDIA networking elements provide the ESF network to support this linear growth while maintaining the highest performance.

With this combination, data scientists running DL environments can focus on the task at hand and not the infrastructure required to support it. Training results are readily available, enabling a quick turnaround.

EXCELERO NVMESH TECHNOLOGY

Excelero delivers low-latency distributed block storage for web-scale applications. NVMesh technology enables shared NVMe across any network and supports any local or distributed file system. The solution features an intelligent management layer that abstracts underlying hardware with CPU offload, creates logical volumes with redundancy, and provides centralized, intelligent management, and monitoring. Applications can enjoy the latency, throughput, and IOPs of a local NVMe device with the convenience of centralized networked storage while avoiding proprietary hardware lock-in and reducing overall storage TCO.

NVMesh features a distributed block layer that allows unmodified applications to utilize pooled NVMe storage devices across a network at local speeds and latencies. Distributed NVMe storage resources are pooled with the ability to create arbitrary, dynamic block volumes that can be utilized by any host running the NVMesh block client.

Being a 100% software-based solution, NVMesh was built to give customers maximum flexibility in designing storage infrastructures. With Excelero MeshConnect™ technology, customers can choose the network fabric and protocol that best meets their performance or efficiency requirements. NVMesh supports the widest selection of supported protocols and fabrics, including TCP/IP, InfiniBand, RDMA over Converged Ethernet (RoCE) v2, RDMA, and NVMe over Fabrics (NVMe-oF). Excelero MeshProtect™ offers flexible protection levels for differing application needs, including mirrored and parity-based redundancy.

MeshInspect™ provides performance analytics for pinpointing anomalies quickly and at scale. NVMesh is deployed as a virtual, distributed non-volatile array and supports both converged and disaggregated architectures, giving customers full freedom in their architectural design.

NVIDIA BLUEFIELD DPU AND NVME SNAP

Modern data-center servers are challenged by heavy compute and storage workloads that consume valuable CPU cores for processing and steering network traffic, resources that should be used for critical application processing. The BlueField DPU offloads critical network, security, and storage tasks from the CPU and is the best solution for addressing performance, networking efficiency, and cyber-security concerns in the modern data center.

The BlueField DPU is a System-on-a-Chip (SOC) and is at the heart of the software programmable SmartNICs and Storage Controllers. BlueField is a single piece of silicon, which integrates 64-bit Arm® multi-core processing power with ConnectX network adapter and a high-speed memory controller to enable enhanced and flexible software programmability.

NVIDIA NVMe SNAP enables hardware virtualization of NVMe storage. The NVMe SNAP framework enables customers to easily integrate networked storage solutions into their cloud or enterprise server deployments. NVMe SNAP brings virtualized storage to bare-metal clouds and makes composable storage simple. It enables the efficient disaggregation of compute and storage to allow fully optimized resource utilization.

NVMe SNAP logically presents networked storage, such as NVMe-oF, as a local NVMe drive. This allows the host OS/Hypervisor to use a standard NVMe-driver instead of a remote networking storage protocol. The host benefits from the performance and simplicity of local NVMe storage, unaware that remote Ethernet or InfiniBand connected storage is being utilized and virtualized by NVMe SNAP. Furthermore, SNAP applies sophisticated logic and data protection mechanisms (mirroring, compression, data-de-duplication, thin-provisioning, encryption, etc.) to the network storage that it virtualizes and displays it as local NVMe pools.

NVMe SNAP empowers customers with the freedom to implement their own storage technology and solutions on top of the NVMe SNAP framework which runs on the BlueField DPU SOC controller. SNAP achieves both performance and software transparency by leveraging BlueField's embedded hardware storage acceleration engines along with integrated programmable Arm cores. This powerful combination is agile yet completely transparent to host software, allowing SNAP to be integrated into almost any storage solution that can expose storage using the NVMe over Fabrics protocol, such as Excelero NVMesh.

AI, ML, DL, AND STORAGE

AI is a broad term encompassing different steps, disciplines, and methods. Very broadly, with AI there are two main steps: Training and Inference.

Training is the part where a machine learning (ML) model is fed data so that it can "learn" something about that data. At a very high level, most of the training is done by ML methods. DL, a subset of ML, is used for

some training aspects as well. ML is generally where data scientists program a model and then feed it lots of data and may then iterate on the model to perfect it. DL is unique in that it often involves neural networks that "train themselves" to find patterns and relationships without any a priori knowledge such as data labeling. DL often goes over the data multiple times.

The goal of the training phase is to teach "the computer" to recognize something of interest. Popular examples for this are persons, cats or traffic signs in the image recognition domain.

But how can a computer learn to recognize a cat? The answer is trivially simple on the one hand, but at the same time can present a major challenge without the proper storage system. The computer (or rather the GPUs) need to see lots and lots of cats in different colors and angles and in different poses and sizes. And when each image has been processed once, the whole dataset is typically processed again and again in a different order, with a different image rotation factor, stretched or in other variations. The more varieties and distinctions all improve the accuracy of recognition.

Also, the data scientists continuously work on improving the model generation of the existing training dataset, which makes training an endless process that is continuously running to improve the company' s problem-solving capabilities and keeping them ahead of the competition.

In clusters of multiple GPU servers, a certain GPU server typically doesn't read the same subset of objects that it read in the first pass again, due to the random reordering of data. This means a cache on the GPU server would not be effective and thus the storage system itself must be able to provide data at the speed at which the GPUs can process it - very, very fast. Due to the nature of the access consisting of lots of small file reads, NVMe has proven to be the only technology today that fits for these requirements, providing both the ultra-low access latency and the high number of read operations per second.

Another important observation for GPU server clusters is that the dataset should be shared between the GPU servers, as they typically all work with the same dataset.

The fact that this phase is mostly about repeated readings, it's an important property that allows optimizations of the solution design.

STORAGE FOR SCALABLE DL TRAINING

The most important properties that an optimal storage solution for training should have:

- Low latency and high read operations per second: This makes NVMe the ideal storage access technology from a protocol and drive aspect, as it provides both, high access performance and affordable capacity at scale
- Data protection: The training phase is business critical for modern companies, which requires the storage system to be fault tolerant. The Excelero MeshProtect feature MeshProtect feature protects the data across multiple NVMe drives and across multiple servers with different options, including distributed erasure coding for logical volumes
- Scalable capacity: A BlueField DPU in each server and Excelero elastic NVMe technology allows NVMe-oF connectivity to remote elastic storage without the need to replace or add local drives to the workers when storage capacity scale-up is required. As Excelero patented remote direct drive access (RDDA) technology combined with the very low latency of the NVMe-oF over RDMA enables access to remote NVMe drives at the same performance that an internal GPU server drive would have, the solution is not bound by the performance and capacity limitations of internal drives of GPU servers
- Short copy time: A single logical training volume shared among all workers allows a very short phase of updating the dataset in addition to drastically reducing write endurance requirements from NVMe drives
- Scalable performance: Similar to the capacity scale approach, more drives and more servers can seamlessly be added to Excelero NVMesh storage to increase performance along with the NVIDIA ConnectX, BlueField, and Spectrum network infrastructure that deliver unprecedented performance and efficiency at massive scale

Given that Excelero NVMesh software defined storage combined with BlueField and NVMe SNAP technology allows access to remote NVMe drives at the same performance that an internal GPU server drive would have, customers are free to use NVMe drives in dedicated servers. This allows to flexibly design and scale the capacity and performance of the storage solution independent of the number of GPU servers and without the need for additional caching or data copying into the GPU servers. A block diagram of the architecture can be seen in figure 1.

At the same time, NVMesh protects the data from hardware failures and enables shared access to logical volumes from all GPU servers.

Some companies use this concept as the basis for parallel file systems like SpectrumScale or BeeGFS. But given the significant overhead of parallel file systems when working with lots of small files and given

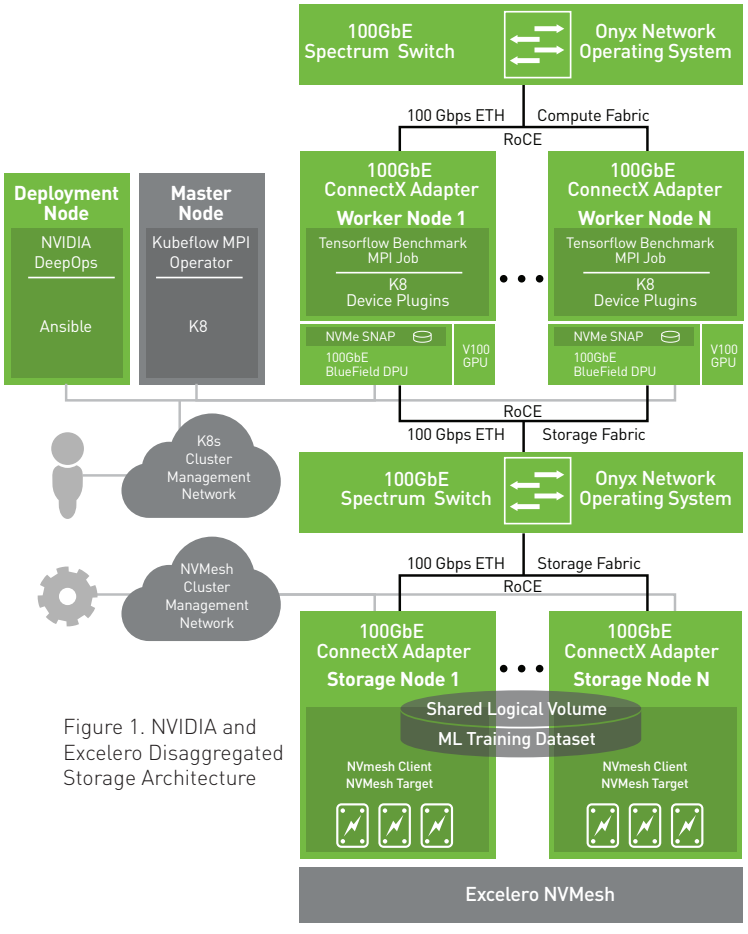

Figure 1. NVIDIA and Excelero Disaggregated Storage Architecture

that the training dataset is typically read-only (except for special occasions when new data is added), our customers recently try to avoid a parallel file system completely to meet their performance demands.

In this case, the NVMesh logical block volumes with the training data are simply mounted with a high-performance local file system like XFS and then read-only access granted to all the GPU servers . On the GPU server, a standard NVMe block device presents the block storage to the file system layer. The block device is implemented by the Bluefield SNAP with which it is translated into NVMe-oF for network transport to the NVMesh system. The NVMesh solution furnishes the I/O from the NVMe drives on the storage server. This allows a level of access performance which is far beyond the performance that customers typically achieve with parallel file systems, this reduces management effort and costs for the company, as less hardware is required to meet the demands.

To address the rather moderate requirements for a shared writable file system to store the generated model, a simple highly available NFS service on the NVMesh servers can be used.

using the ImageNet dataset, runs on top of the NVIDIA networking solution and Excelero NVMesh SDS. Native K8s is used as the platform with Horovod distributed training framework for TensorFlow and NCCL, Kubeflow MPI operator to orchestrate the workload across the cluster worker nodes. Two HPE Apollo 6500 Gen10 servers were used as worker nodes to showcase the distributed nature of the solution, each with a single V100 GPU, a ConnectX-5 100GbE network adapter, and a Spectrum 2700 100Gb Ethernet fabric switch, which is dedicated to training related traffic between the worker nodes. An NVIDIA Mellanox 100GbE Bluefield SmartNIC running NVMe SNAP on the same worker nodes and an additional Spectrum 2700 provides the interface to the storage fabric. Between the BlueField DPU and Excelero NVMesh cluster, NVMe-oF over RoCE is used for high-speed, low-latency communication composed out of a set of storage nodes. The NVMesh hardware employs four SuperMicro servers utilizing Intel Xeon E5-2660 processors with 128GB of RAM each and Intel P3600 1.2TB NVMe drives. A high capacity ImageNet training dataset is maintained and updated in a single shared logical NVMesh volume exposed to the GPU workers and emulated as a local NVMe drive using NVMe SNAP on the

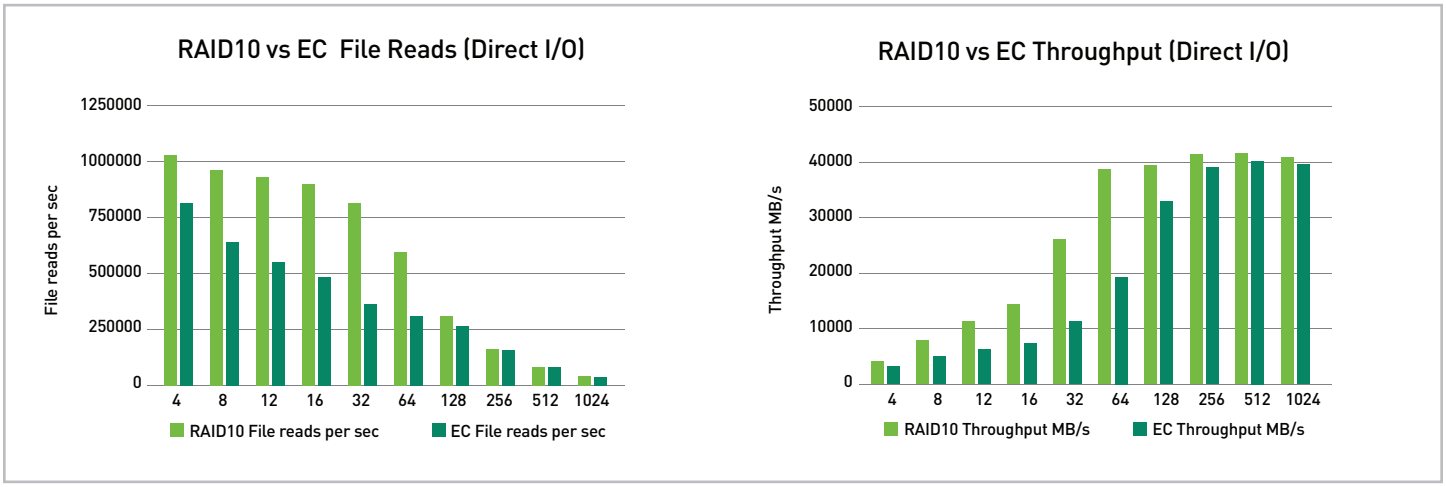

Figure 2. Direct I/O Comparison of Reads and Throughput

The result is a modern storage system that takes full advantage of the latest hardware technologies, while at the same time providing data protection and unmatched performance at arbitrary scale through Excelero software defined storage and patented RDDA technology.

THE RESNET-50 IMAGENET BENCHMARK

A ResNet or residual neural network is a neural network meant to mimic patterns found in the cerebral cortex. By introducing skips in a deep neural network, i.e. one with many layers, it is believed that some of the issues found in regular deep neural networks are resolved. ResNet-50 is a commonly used neural network of 50 layers often used for training on images. The act of training with such a neural network with images from the ImageNet dataset is a typical benchmark used to validate the AI training system's capabilities.

VALIDATION ARCHITECTURE AND TESTING ENVIRONMENT

In this validated reference architecture, the ResNet-50 benchmark,

BlueField DPU, therefore preventing the need for a CSI or any other agent on the worker hosts.

TESTING RESULTS AND COMPARISON

Using the setup, ImageNet TensorFlow benchmark results were recorded. The stable result showed 1250 images/sec utilizing a V100 GPU on a NVIDIA DGX-1™ system with a batch size of 256 and 2454 images/sec was achieved with a batch size of 256 across two worker nodes with V100 GPU each. Network utilization was 4.5 Gbps on the compute fabric and 1.25 Gbps on the storage fabric. The jitter is negligible with results remaining stable for hours. These results compare very favorably with various publicly available benchmark results including results that NVIDIA has published of 1,061 images/sec.

In all test cases the network utilization was very low compared to the system's potential. The same applies for the storage system provided. Repeating the benchmark with local drives to compare against the networked storage rendered similar results, proving there is no performance lost when using SNAP and networked drives. CPU load

was the same for both tests showing that solutions based on SNAP technology do not consume more resources accessing remote volumes than is consumed when accessing local drives.

The graphs in Figure 2 above depict the performance achieved from the same storage system comprising NVMesh software on a Supermicro server with Intel NVMe SSDs using a slightly different test methodology to exhibit performance potential. These numbers were measured when taxing the same NVMesh storage construct using file operations directly instead of via the AI benchmark presented.

The first graph exhibits the performance when using an XFS file system on top of an NVMesh MeshProtect 10 volume with data striped across multiple drives with mirrored copies of data. The second graph depicts the performance rendered with an XFS file system on an NVMesh MeshProtect 60 volume with data spreads across multiple drives using erasure coded dual-failure data protection.

In the demonstrated storage disaggregation setup, the network utility is very low for both the compute fabric and the storage fabric. Such a setup can scale to additional nodes and GPUs while maintaining linear performance. On the compute fabric side, with 100 Gbps networking, there is room for a 20-fold increase, which would suggest that scaling 20X is feasible without the compute fabric becoming a bottleneck. On the storage system side, with the ability to run over a million file operations per second and over 320 Gbps of throughput, the same 20-fold increase can be accommodated as well.

Finally, there is headroom left for more bandwidth on SNAP, which means that the same architecture can be used with more performant NVIDIA A100 GPUs without losing any performance.

CONCLUSION

The combination of NVIDIA Mellanox ESF cutting edge networking technologies and software solutions and Excelero NVMesh SDS make a compelling reference architecture for scalable artificial intelligence. Incorporation of the Bluefield DPU with NVMe SNAP facilitates software deployment without sacrificing performance by saving the need to install and configure the worker nodes agents and tools to allow for the high-speed hardware-accelerated connection to the remote storage.

NVMesh even with a minimal configuration leaves plenty of room for growth in system scale providing enticing economics. From the relatively small scale demonstrated all the way up to data center wide setups. The reasonable cost includes distributed data protection ensuring data center quality data protection is guaranteed.

WANT TO LEARN MORE?

Learn more about the NVIDIA BlueField DPU:
https://www.mellanox.com/products/smartnic

Learn more about how NVIDIA NVMe SNAP technology enables hardware virtualization of NVMe storage:
https://www.mellanox.com/products/software/nvme-snap

Learn more about NVIDIA Spectrum Ethernet Switches:
https://www.mellanox.com/products/ethernet-switches

Learn how an NVIDIA Ethernet Storage Fabric solution can deliver the fastest storage networking solution:
https://www.mellanox.com/ethernet-storage-fabric

Learn more about Excelero NVMesh:
https://www.excelero.com/product/nvmesh/

Learn more about Excelero software-only solutions, like MeshConnect:
https://www.excelero.com/software-only-solutions/