Solution Brief

Powering the Deep Learning Platform with ONTAP AI and allegro.ai

Run deep learning models faster and more efficiently

Key Benefits

Accelerate Time to Completion
Streamline the flow of data reliably and speed up training and inference with a solution that spans from the edge to the core to the cloud.

Increase GPU Utilization
Exceed raw I/O bandwidth needs and maintain high GPU throughput for accelerated DL and AI training and inference.

Simplify Deployment
Eliminate design complexities and guesswork with a validated reference architecture.

The Challenge

Artificial intelligence (AI) has taken the world by storm. Fueled by the continuous generation of data, advanced modeling solutions that are developed with AI are driving innovations across all industries. Advanced computer vision solutions, developed through an area of AI known as deep learning (DL), now enable computers to gain understanding from digital images and from videos. Computer vision technologies have broad applications in every industry, from autonomous vehicles to drones, security, logistics, and more.

Computer vision poses unique challenges that span both data science and software engineering. Training neural networks requires constant slicing and dicing of datasets and addressing inherent biases. These issues create significant challenges for development, for QA, and for deployment. Working with huge datasets creates additional challenges in the physical and logical manipulation of the datasets that are required throughout the lifecycle of computer vision and other sensor-based solutions.

The Solution

NetApp and allegro.ai have partnered to create a unified solution that handles all aspects of the lifecycle of your dataset. With this joint solution, you can optimize your data and model management, from the physical layer up to the application layer.

allegro's functional modules enable you to tightly manage every piece of your DL solution, end to end, for its entire life span. allegro gives you the data management, automated annotation, training and model management, and continuous learning capabilities that you need to power your DL platforms. With the NetApp® ONTAP® AI proven architecture, powered by NVIDIA DGX supercomputers and NetApp cloud-connected, all-flash storage, you can fully realize the promise of AI and DL. ONTAP AI enables you to simplify, to accelerate, and to integrate your data pipeline with a NetApp Data Fabric that spans from the edge to the core to the cloud.

allegro.ai's sophisticated, robust DL platform combined with ONTAP AI helps your data scientists create DL models at scale and offers dramatic performance improvements over CPU-only solutions. With allegro and ONTAP AI, you can accelerate time to completion, increase GPU utilization, and simplify deployment.

Accelerate Time to Completion

The time that it takes to train DL models is an important factor when you are building AI pipelines. If a model takes too long to be trained, you lose valuable time before you can put it into production. To speed up training time, you can add more GPUs. Often, however, the data I/O becomes the bottleneck, causing significant delays in overall time to completion.

NetApp®

allegro.ai Technology Partner

The ONTAP AI high-throughput, all-flash storage and NVIDIA GPUs enable you to train more models in less time while extracting maximum performance from your hardware. With end-to-end NVMe performance, from back-end storage to front-end connectivity, the NetApp AFF A800 all-flash storage system can feed data to NVIDIA DGX-1 systems 6 to 9 times faster than competing solutions can.

Compared with solutions that use hyperscale cloud storage, ONTAP AI improves time to completion and processing across the entire DL pipeline. With allegro running on ONTAP AI, you can be confident that your DL pipeline uses the full performance bandwidth of the NVIDIA DGX supercomputers to accelerate insights and to enable sophisticated model building.

Having a large cache size also plays an important role in overall time to completion when you run DL models. With ONTAP AI, your organization can use large pools of flash storage for data caching, dramatically accelerating decoding and training for allegro workloads.

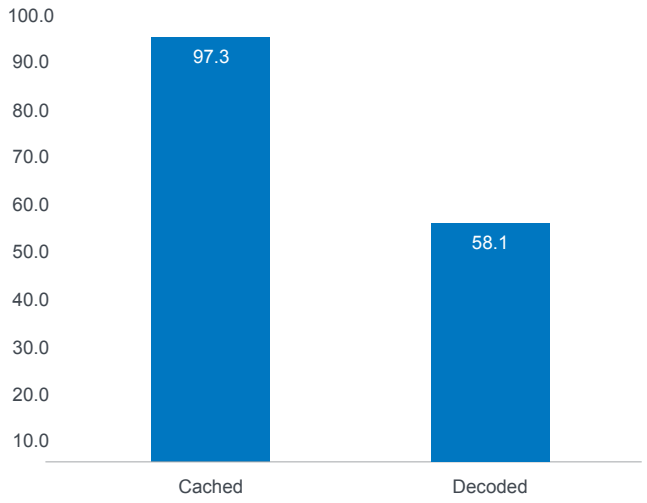

Figure 1) Average GPU utilization percentage during caching and decoding while running the video experiment.

Increase GPU Utilization

When it comes to DL training and modeling, you can spend a lot of time on data access and on writing metadata. And while that data I/O occurs, GPUs sit idle. That idle time is even more significant when you have a big GPU cluster with several servers—each with multiple GPUs.

ONTAP AI high throughput and data I/O speeds can help you cut down on GPU idle time by keeping the GPUs engaged more often. ONTAP AI is designed around the AI and DL processing power of NVIDIA GPUs, giving you much higher GPU utilization for training and for inference workloads than competitive solutions provide. At the same time, allegro provides high-performance software caching optimization that also significantly improves throughput and data I/O speeds. The combination and integration of both components deliver superior performance.

During internal experimentation of allegro on ONTAP AI and with all allegro software caching features turned on, GPU utilization jumped to 97.3% on cache-based testing. That jump was from a baseline of 58.1% GPU utilization with decoding-based testing, without ONTAP AI and with allegro software caching disabled.

Simplify Deployment

With ONTAP AI, you get a prescriptive architecture that eliminates design complexities and enables independent scaling of compute and storage. By combining NVIDIA DGX-1 servers, NVIDIA Tesla V100 GPUs, and NetApp AFF A800 storage with state-of-the-art networking, ONTAP AI simplifies your AI deployments with a simple design and no guesswork. Start small and grow nondisruptively while intelligently managing data from the edge to the core to the cloud and back.

About allegro.ai

allegro.ai is a pioneering deep learning computer vision platform that provides a complete product lifecycle management solution for AI development and production, focused on the domain of computer vision. allegro's platform enables you to increase your development velocity and the quality and scalability of your deep learning initiatives across domains including autonomous vehicles, robotics, security cameras, logistics and many others. **Learn more**.

About NetApp

NetApp is the data authority for hybrid cloud. We provide a full range of hybrid cloud data services that simplify management of applications and data across cloud and on-premises environments to accelerate digital transformation. Together with our partners, we empower global organizations to unleash the full potential of their data to expand customer touchpoints, foster greater innovation and optimize their operations. For more information, visit **www.netapp.com**. #DataDriven