Solution Brief

NetApp ONTAP AI

Simplify, accelerate, and integrate your data pipeline for deep learning with NetApp and NVIDIA

Key Benefits

Simple to deploy

- Get going faster by eliminating design complexity and guesswork
- Speed innovation and experimentation
- Streamline deployment with enterprise-grade data services and simple technology refreshes

Deliver the performance and scalability your business needs

- Start small and grow nondisruptively
- Accelerate results with a high-performance solution
- Handle more than 400 billion files with a single namespace

Build an integrated data pipeline

- Intelligently manage your data with an integrated pipeline, from edge to core to cloud
- Backed by AI expertise and single-point-of-contact support
- Accelerate cloud integration with the NetApp® Data Fabric

AI Infrastructure Challenges

Artificial intelligence (AI) and deep learning (DL) enable enterprises to detect fraud, improve customer relationships, optimize the supply chain, and deliver innovative products and services in an increasingly competitive marketplace. Yours may be one of the many organizations that are leveraging new DL approaches to drive digital transformation and gain a competitive advantage. To wring maximum benefit from DL, you must first address several key challenges.

Do-it-yourself integrations are complex. Assembling and integrating off-the-shelf DL compute, storage, networking, and software components can increase complexity and lengthen deployment times. As a result, valuable data science resources are wasted on systems integration work.

Achieving predictable and scalable performance is hard. DL best practices suggest that organizations should start small and scale as they go. Traditionally, compute and direct-attached storage have been used to feed data to AI workflows. But scaling with traditional storage can lead to disruption and downtime for ongoing operations.

Disruptions impact OpEx and the productivity of data scientists. DL infrastructure is complex, involving numerous hardware and software interdependencies. Keeping a DL infrastructure up and running requires deep, full-stack AI expertise. Downtime or slow AI performance can set off a chain reaction that impacts developer productivity and causes operational expenses to spin out of control.

The Solution

Now you can fully realize the promise of AI and DL by simplifying, accelerating, and integrating your data pipeline with the NetApp ONTAP® AI proven architecture, powered by NVIDIA DGX supercomputers and NetApp cloud-connected all-flash storage. Streamline the flow of data reliably and speed up training and inference with the Data Fabric that spans from edge to core to cloud.

"Deep learning is revolutionizing almost every market we work in. NetApp ONTAP AI, powered by NVIDIA DGX supercomputers and NetApp all-flash storage, is simplifying and accelerating the data pipeline for deep learning."

Monty Barlow, Head of Artificial Intelligence
Cambridge Consultants

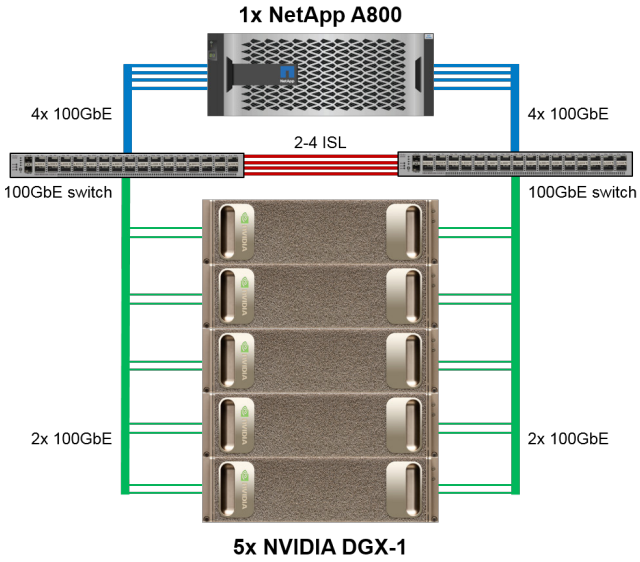

Figure 1) The NetApp ONTAP AI proven architecture.

Simplify design and deployment

The rapid pace of AI innovation makes designing an effective AI infrastructure challenging. But with ONTAP AI you can eliminate guesswork and get started faster with a validated reference architecture that detangles design complexity. Trident, NetApp's storage provisioner for Kubernetes, further accelerates your ONTAP AI deployment by seamlessly moving your NVIDIA GPU Cloud (NGC) container images onto NetApp's enterprise-grade flash storage.

DL training routines demand massive amounts of compute power. Faster image training can cut down on overall compute costs while accelerating AI innovation and productivity. Just one DGX-1 server provides over 1 PFLOPS of AI computing power, the equivalent of an entire data center of traditional CPU-based servers. Investing in state-of-the-art compute demands state-of-the-art storage that can handle thousands of training images per second. You need a high-performance data services solution that will keep up with your most demanding DL training workloads.

ONTAP AI testing using ImageNet data with a NetApp AFF A800 system and NVIDIA DGX-1 servers in a 1:4 storage-to-compute configuration achieved training throughput of 23,000 training images per second (TIPS) and inference throughput of 60,000 TIPS. With this configuration, you can expect to get over 2GBps of sustained throughput (5GBps peak) with well under 1ms of latency while the GPUs operate at over 95% utilization. A single AFF A800 system supports throughput of 25GBps for sequential reads and 1 million IOPS for small random reads at under 500-microsecond latencies for NAS workloads. These results demonstrate available performance headroom that can support many more DGX-1 servers as requirements increase.

Deliver the performance and scalability your business needs

ONTAP AI allows you to start small and grow as needed. Add compute, storage, and networking to clustered configurations without disrupting ongoing operations. Start with a 1:1 storage-to-compute configuration and scale out as your data grows to a 1:5 configuration and beyond. NetApp's rack-scale architecture allows organizations to start with an AFF A220 and grow as needed to scale from hundreds of terabytes to tens of petabytes with all-flash. And with NetApp ONTAP FlexGroup, up to 20 petabytes of single namespace can handle more than 400 billion files.

Build an integrated data pipeline that spans from edge to core to cloud

ONTAP AI leverages the NetApp Data Fabric to unify data management across the pipeline with a single platform. Use the same tools to securely control and protect your data in flight, in use, or at rest and meet compliance requirements with confidence. If an issue arises in your DL environment, you can rely on your single point of contact and our proven support model to help troubleshoot and provide guidance.

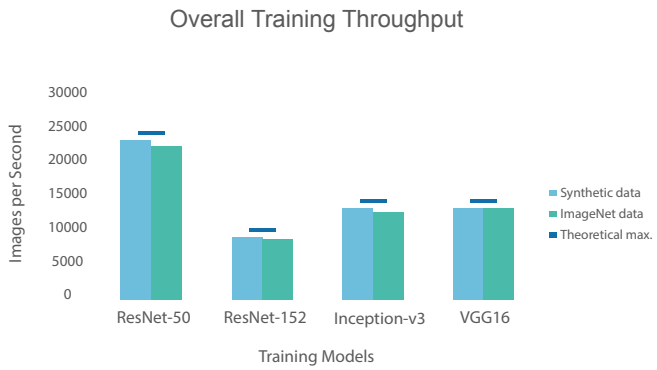

Figure 2) Training throughput for all models.

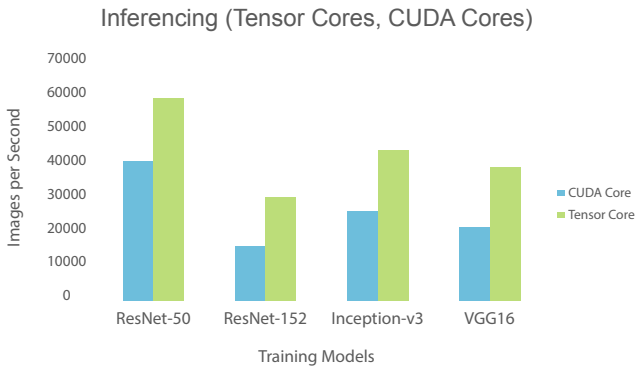

Figure 3) Inference for all training models.

NetApp and NVIDIA: Driving Innovation Together

At the heart of ONTAP AI is the NVIDIA DGX-1 AI supercomputer, a fully integrated hardware and software system that is purpose-built for DL. Each DGX-1 server is powered by eight Tesla V100 Tensor Core GPUs, configured in a hybrid cube-mesh topology using NVIDIA NVLink. With NVLink in the DGX-1, you get an ultra-high-bandwidth, low-latency fabric for the inter-GPU communications that are essential to multi-GPU training, eliminating the bottleneck associated with PCIe-based interconnect. The DGX platform leverages the NVIDIA GPU Cloud Deep Learning Software Stack, which is optimized for maximum GPU-accelerated DL performance.

NetApp AFF systems keep data flowing to DL processes with the industry's fastest and most flexible all-flash storage, featuring the world's first end-to-end NVMe technologies. The AFF A800 is capable of feeding data to NVIDIA DGX-1 systems up to 4 times faster than competing solutions.[1]

1. Read throughput of up to 300GBps per all-flash cluster versus 75GBps from a leading competitor.

NetApp's Data Fabric offers best-in-class data management and cloud integration to help you accelerate DL while managing and protecting your critical data. ONTAP delivers an unparalleled 22:1 overall data-reduction ratio and up to 54% lower TCO compared to direct-attached storage. Leveraging industry-leading data services capabilities, ONTAP helps you manage and protect your data with a single set of tools, regardless of where it resides, and freely move data to wherever it's needed, from edge to core to cloud.

The solution comes integrated with Cisco Nexus 3232C 100Gb Ethernet switches, which provide the low latency, high density, high performance, and power efficiency demanded by AI environments. Now, with ONTAP AI, you can simplify deployment and management with single-point-of-contact support for your NVIDIA, NetApp, and Cisco proven architecture.

Solution Components

- NVIDIA DGX-1 servers
- NetApp AFF A800 storage system
- Cisco Nexus 3232C network switches
- NVIDIA GPU Cloud Deep Learning Software Stack
- Trident, NetApp's open source, dynamic storage provisioner

About NVIDIA

NVIDIA's (NASDAQ:NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing — with the GPU acting as the brain of computers, robots and self-driving cars that can perceive and understand the world. More information at **http://www.nvidia.com/dgx**.

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