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
- 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

Build an integrated data pipeline

- Intelligently manage your data with an integrated pipeline, from edge to core to cloud
- Backed by AI expertise and simple support options

Unify AI workloads

- Eliminate infrastructure silos
- Flexibly respond to business demands

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 systems and NetApp cloud-connected all-flash storage. Streamline the flow of data reliably and speed up analytics, training, and inference with your data fabric that spans from edge to core to cloud.

NetApp ONTAP AI is one of the first converged infrastructure stacks to incorporate NVIDIA DGX™ A100, the world's first 5-petaflop AI system, and NVIDIA Mellanox® high-performance Ethernet switches to unify AI workloads, simplify deployment, and accelerate return on investment.

"Deep learning is revolutionizing almost every market we work in. We're applying deep learning in diverse markets, driving forward the art of the possible. NetApp ONTAP AI, powered by NVIDIA DGX systems and NetApp all-flash storage, is simplifying and accelerating the data pipeline for deep learning."

Tim Ensor, Director of Artificial Intelligence
Cambridge Consultants

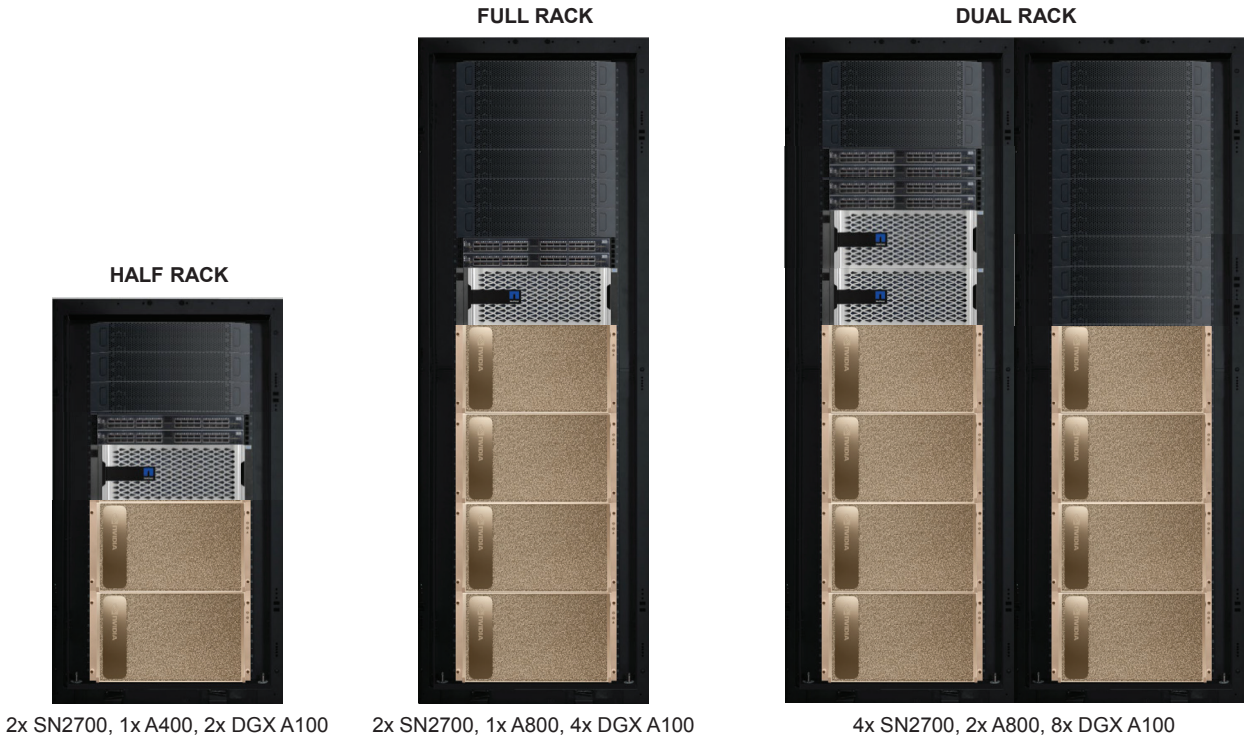

Figure 1) ONTAP AI architectures using DGX A100; 2-, 4-, and 8-node configurations.

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.

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.

Built using the new NVIDIA® Ampere architecture, the DGX A100 system delivers up to 6 times the training performance of the prior generation, delivering the equivalent of a data center of compute infrastructure for analytics, training, and inference, now consolidated in a single system. Compared with CPU systems, DGX A100 requires 1/25th the space and 1/20th the power, while costing 1/10th as much.

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 keeps up with your most demanding DL training workloads.

You can expect to get more than 2GBps of sustained throughput (5GBps peak) with well under 1ms of latency, while the GPUs operate at over 95% utilization. A single NetApp AFF A800 system supports throughput of 25GBps for sequential reads and 1 million IOPS for small random reads at latencies of less than 500 microseconds for NAS workloads.

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 1:4 configuration and beyond.

NetApp's rack-scale architecture allows organizations to start with an AFF A400 and grow as needed to scale from hundreds of terabytes to tens of petabytes with all-flash storage. 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 your data fabric to unify data management across the data 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 our proven support model to help troubleshoot and provide guidance.

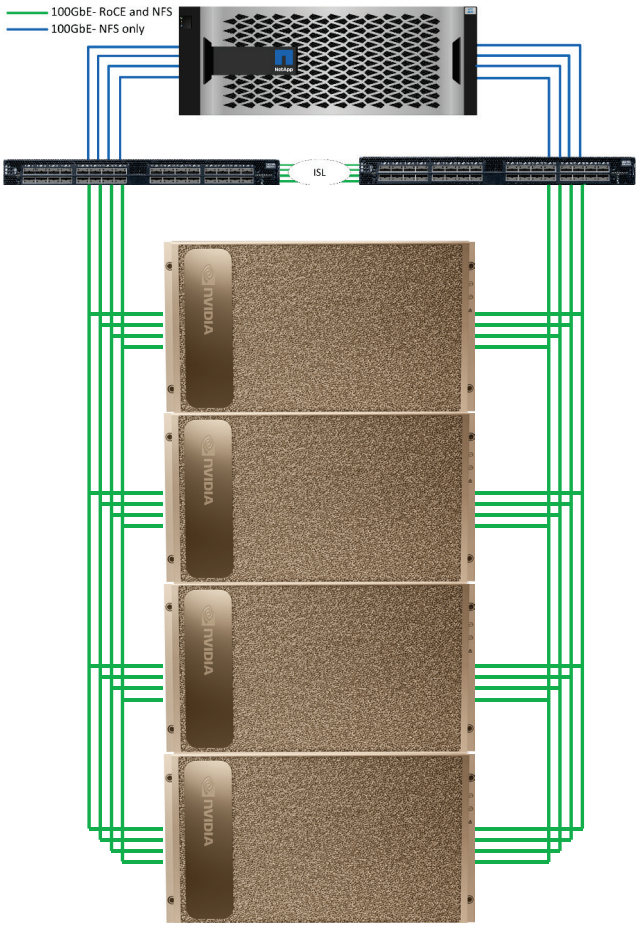

Figure 2) ONTAP AI 4-node configuration with Mellanox Spectrum 100GbE switches.

Unify AI Workloads

Now organizations can eliminate silos of infrastructure that are either underutilized or starve AI workloads. With ONTAP AI, enterprises get a universal AI infrastructure solution built on DGX A100 that consolidates analytics, training and inference onto one platform that flexibly responds to business demands with better TCO than legacy architectures.

NetApp and NVIDIA: Driving Innovation Together

At the heart of ONTAP AI is DGX A100, a universal building block for data center AI, supporting DL training, inferencing, data science, and other high-performance workloads from a single platform. Each DGX A100 system is powered by eight NVIDIA A100 Tensor Core GPUs and integrates the latest high-speed NVIDIA Mellanox HDR interconnects.

Multiple smaller workloads can be accelerated by partitioning the DGX A100 into as many as 56 instances per system, using new multi-instance GPU technology. This acceleration enables organizations to allocate GPU performance very efficiently in ONTAP AI, empowering data science teams across the enterprise to iterate faster, automate reproducibility, and deliver AI projects up to 3 months sooner with higher quality.

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 DGX systems up to 4 times faster than competing solutions.[1]

The solution comes integrated with Mellanox Spectrum Ethernet switches, which provide the low latency, high density, high performance, and power efficiency demanded by AI environments.

A data fabric enabled by NetApp 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.

DGX A100 is powered by the NVIDIA DGX software stack, which includes optimized software for AI and data science workloads, delivering maximized performance, enabling enterprises to achieve a faster return on their investment in AI infrastructure.

The NetApp AI Control Plane is NetApp's full-stack AI data and experiment management solution. It provides extreme scalability, streamlined deployment, and nonstop data availability—when and where you need it.

The NetApp AI Control Plane integrates Kubernetes and Kubeflow with a data fabric enabled by NetApp, which provides uncompromising data availability and portability from edge to core to cloud.

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

Solution Components

- NVIDIA DGX A100 systems
- NetApp AFF A-Series storage systems with ONTAP 9
- NVIDIA Mellanox Spectrum SN2700
- NVIDIA DGX software stack
- NetApp AI Control Plane

Reference Architectures

NetApp has released the following reference architectures based on ONTAP AI targeted at use cases in specific industries:

- ONTAP AI Reference Architecture for Healthcare: Diagnostic Imaging
- ONTAP AI Reference Architecture for Autonomous Driving Workloads: Solution Design
- ONTAP AI Reference Architecture for Financial Services Workloads

About NVIDIA

NVIDIA's 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 **www.nvidia.com**.

About NetApp

NetApp is the leader in cloud data services, empowering global organizations to change their world with data. Together with our partners, we are the only ones who can help you build your unique data fabric. Simplify hybrid multicloud and securely deliver the right data, services, and applications to the right people at the right time. Learn more at **www.netapp.com**.

© 2020 NetApp, Inc. All Rights Reserved. NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners. SB-3939-0520