SOLUTION BRIEF

ONTAP AI

Simplify, accelerate, and integrate your data pipeline for ML and DL with NetApp and NVIDIA

NetApp | NVIDIA

AI infrastructure challenges

Artificial intelligence (AI), machine learning (ML), 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 might be one of the many organizations that use new AI approaches to guide digital transformation and to gain a competitive advantage. To obtain the maximum benefit from AI, you must first overcome several key challenges.

Do-it-yourself integrations are complex. Assembling and integrating off-the-shelf ML and 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 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 affect the productivity of data scientists. ML and DL infrastructure involves numerous hardware and software interdependencies. Keeping an infrastructure up and running requires deep, full-stack AI expertise. Downtime or slow AI performance can set off a chain reaction that reduces developer productivity and causes operational expenses to spin out of control.

The solution

Now you can fully realize the promise of AI, ML, and DL. Simplify, accelerate, and integrate your data pipeline with the NetApp® ONTAP® AI proven architecture that's 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.

Key benefits

Reduce risk with flexible, validated solutions

- Get going faster by eliminating design complexity and guesswork.
- Streamline configuration and deployment with available preconfigured solutions.

Deliver the right performance and scalability

- Start small and grow nondisruptively.
- Speed 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.
- Deploy a solution that's backed by AI expertise and simple support options.

Unify AI workloads

- Eliminate infrastructure silos.
- Flexibly respond to business demands.

NetApp ONTAP AI is one of the first converged infrastructure stacks to incorporate the NVIDIA DGX A100 system, the world's first 5-petaflops AI system, and NVIDIA Mellanox high-performance Ethernet switches. You get unified AI workloads, simplified deployment, and fast return on investment.

"Deep learning is revolutionizing almost every market we work in. We're applying it 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 with DGX A100; two-, four-, and eight-node configurations.

Reduce risk with flexible, validated solutions
The rapid pace of AI innovation makes designing an effective AI infrastructure challenging. With ONTAP AI, you can eliminate guesswork and get started faster by using a field-proven reference architecture. Or by choosing a preconfigured integrated solution that is easy to procure and to deploy, you can eliminate design and management complexity.

The ONTAP AI integrated solution is available in four preconfigured options with capacity expansion and optional advanced software. This integrated solution further reduces complexity by including on-site installation and comprehensive support with a single number to call, from incident reporting through to resolution.

Deliver the right performance and scalability
DL training routines demand massive amounts of compute power. Faster image training can cut down on overall compute costs while speeding up AI innovation and productivity.

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

With NetApp all flash storage, you can expect to get more than 2GBps of sustained throughput (5GBps peak). Further, there's well under 1 millisecond 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.

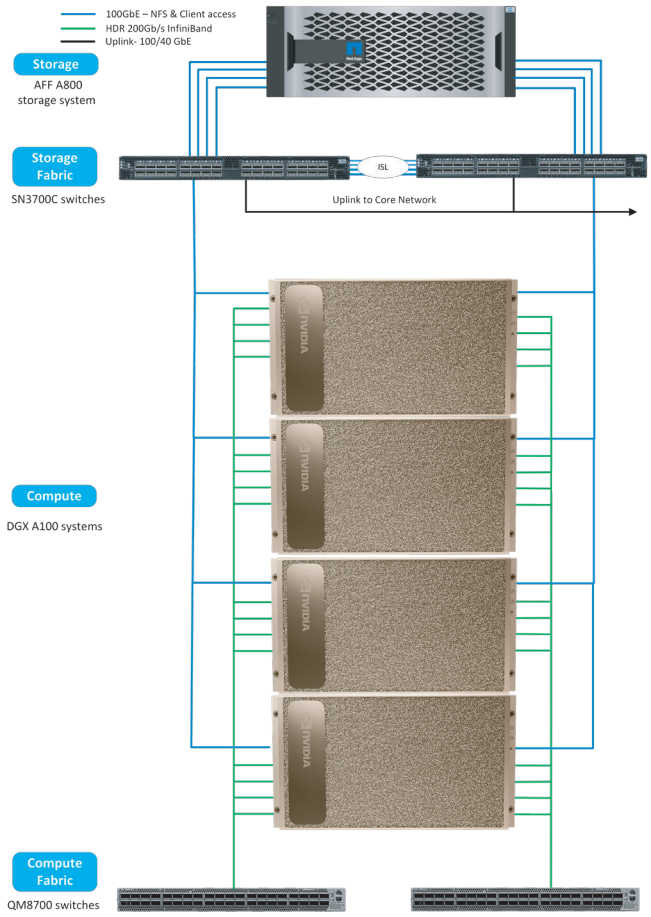

Figure 2) ONTAP AI four-node configuration with Mellanox Spectrum 100GbE switches.

With the NetApp rack-scale architecture, your organization can scale from tens of terabytes to tens of petabytes with all-flash storage. And with NetApp ONTAP FlexGroup, up to 20PB 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 uses 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.

Unify AI workloads

Now your organization can eliminate silos of infrastructure that either are underused or starve AI workloads. With ONTAP AI, you get a universal AI infrastructure solution that's built on DGX A100 systems. The solution consolidates analytics, training, and inference onto one platform that flexibly responds to meet your business demands. You also get better TCO than from legacy architectures.

NetApp and NVIDIA: Driving innovation together

At the heart of ONTAP AI is the DGX A100 system, a universal building block for data center AI that supports training, inference, data science, and other high-performance workloads. Each DGX A100 system is powered by eight NVIDIA A100 Tensor Core GPUs and dual 2nd Gen AMD EPYC™ processors. Each system also integrates the latest high-speed NVIDIA Mellanox 100/200Gb Ethernet- and InfiniBand-capable ConnectX-6 adapter interconnects.

Multiple smaller workloads can be accelerated by partitioning the DGX A100 system into as many as 56 instances per system, using new NVIDIA Multi-Instance GPU (MIG) technology. This acceleration enables your organization to allocate GPU performance efficiently in ONTAP AI. Your data science teams across the enterprise can iterate faster, automate reproducibility, and deliver AI projects up to 3 months sooner—with higher quality.

NetApp AFF systems keep data flowing to ML and DL processes with the industry's fastest and most flexible all-flash storage, which features the world's first end-to-end NVMe technologies. The AFF A800 system can feed data to DGX systems up to four times faster than competing solutions do.[1]

The ONTAP AI solution comes integrated with Mellanox Spectrum Ethernet switches. These switches provide the low latency, high density, high performance, and power efficiency that AI environments demand.

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

A data fabric enabled by NetApp offers best-in-class data management and cloud integration to help you accelerate DL while you manage and protect your critical data. ONTAP gives you an unparalleled 22:1 overall data-reduction ratio and up to 54% lower TCO compared with direct-attached storage.

The DGX A100 system is powered by the NVIDIA DGX software stack, which includes optimized software for AI and data science workloads. You get maximized performance that enables your enterprise to achieve a faster return on your investment in AI infrastructure.

The NetApp AI Control Plane helps simplify AI data management by integrating Kubernetes and Kubeflow with a data fabric enabled by NetApp. This integrated solution gives you optimal data availability and portability from edge to core to cloud. Enhancing the AI Control Plane is the NetApp DataOps Toolkit, a Python library that makes it easy for your data scientists and data engineers to perform numerous data management tasks. For example, they can provision a new data volume, clone a data volume instantaneously, and create a NetApp Snapshot™ copy of a data volume for traceability and baselining.

The right tools are critical to success. That's why ONTAP AI is validated with the leading machine learning operations (MLOps) software, including Domino Data Lab, Iguazio, and more. Your teams can use tools that are familiar to maximize the value of your AI environment and to accelerate time to insight.

Solution components

- NVIDIA DGX A100 systems
- NetApp AFF A-Series storage systems with ONTAP 9
- NVIDIA Mellanox Spectrum SN3700C, NVIDIA Mellanox Quantum QM8700, and/or NVIDIA Mellanox Spectrum SN3700-V
- NVIDIA DGX software stack
- NetApp AI Control Plane
- NetApp DataOps Toolkit

Reference architectures

NetApp has released the following reference architectures, based on ONTAP AI, that are targeted for 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: Solution Design

---

About NetApp

In a world full of generalists, NetApp is a specialist. We're focused on one thing, helping your business get the most out of your data. NetApp brings the enterprise-grade data services you rely on into the cloud, and the simple flexibility of cloud into the data center. Our industry-leading solutions work across diverse customer environments and the world's biggest public clouds.

As a cloud-led, data-centric software company, only NetApp can help build your unique data fabric, simplify and connect your cloud, and securely deliver the right data, services and applications to the right people—anytime, anywhere. www.netapp.com

About NVIDIA

The invention of the GPU in 1999 by NVIDIA 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.

NetApp

PNY Technologies Europe
9 rue Joseph Cugnot, 33708 Mérignac cedex | France
T +33 (0) 5 56 13 75 75 | pnypro@pny.eu

For more information visit: www.pny.eu