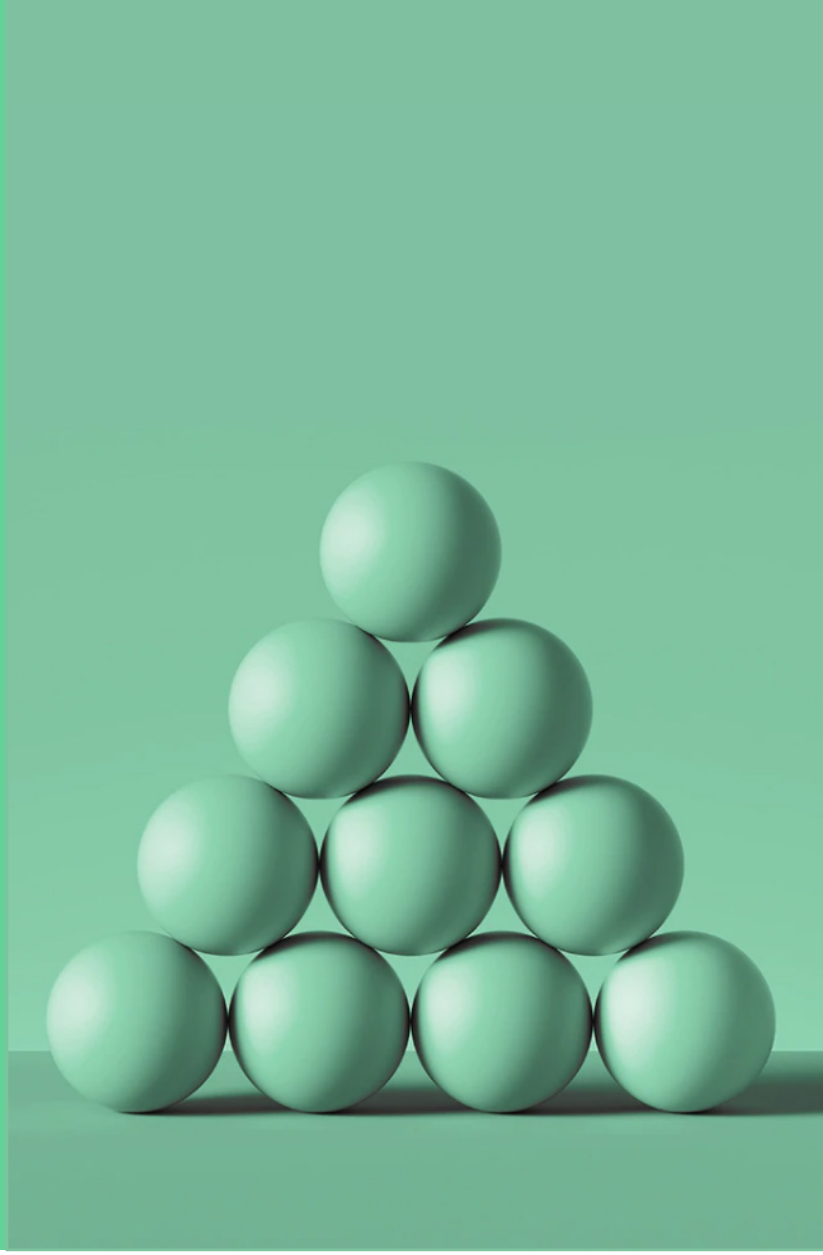
SOLUTION BRIEF

NetApp EF-Series AI

Accelerate time to insight with fast streaming of data to AI applications



NetApp | NVIDIA

AI infrastructure challenges

Artificial intelligence (AI) and deep learning (DL) enable enterprises to detect fraud, to improve customer relationships, to optimize the supply chain, and to deliver innovative products and services in an increasingly competitive marketplace.

To obtain the maximum benefit from DL, you must first overcome 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.

Traditionally, compute and direct-attached storage have been used to feed data to AI workflows. But scaling with existing storage can lead to disruption and downtime for ongoing operations. Disruptions affect the productivity of data scientists and data engineers. 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

The computationally intensive algorithms of DL systems are suited to GPU architecture. And because GPUs based on NVIDIA DGX systems can handle these computations, they're now the preferred platform for workloads such as high-performance computing (HPC), DL, video processing, and analytics. These environments need storage and networking infrastructure that can keep GPUs fed with data. They also need dataset access at ultralow latencies with high bandwidth.

NetApp® EF-Series AI tightly integrates NVIDIA DGX™ A100 systems, NetApp EF600 NVMe storage systems, and the BeeGFS parallel file system with state-of-the-art InfiniBand networking. NetApp EF600 AI simplifies artificial intelligence deployments by eliminating design complexity and guesswork. You can start small and scale seamlessly from science experiments and proofs-of-concept to production and beyond.

Figure 1 shows some options in the EF-Series AI family of solutions with DGX A100 systems. The

Key benefits

Deploy easily

- Reduce risk with a flexible, validated solution.
- Get going faster by eliminating design complexity and guesswork.

Deliver the performance and scalability your business needs

- Start small and grow nondisruptively.
- Get faster results with a high-performance solution.

Supercharge BeeGFS storage and metadata services

- Optimize diverse AI workloads within a single storage namespace.
- Get ease of use, straightforward installation, and simple management.

EF600 powered BeeGFS building blocks have been verified with up to eight DGX A100 systems. By adding more of these building blocks, you can support many DGX A100 systems and petabytes of storage capacity. You have the flexibility to alter compute-to-storage ratios independently based on the size of the data lake, the DL models you're using, and the performance you need.

Deploy easily

The rapid pace of AI innovation makes designing an effective AI infrastructure challenging. With EF-Series AI, you can get started faster by using a field-proven, validated reference architecture. Configuration and deployment are streamlined.

Deliver the performance and scalability your business needs

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 previous generation. You get the equivalent of a data center of compute infrastructure for analytics, training, and inference,

Figure 1) BeeGFS building blocks based on EF-Series with NVIDIA DGX A100 systems

now consolidated in a single system. Compared with CPU systems, the DGX A100 system requires 1/25 of the space and 1/20 of the power—at only 1/10 the cost.

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.

The NetApp EF600 all-flash array gives you consistent, near-real-time access to data while supporting any number of workloads simultaneously. For fast, continuous feeding of data to AI applications, EF600 storage systems deliver up to 2 million cached read IOPS, response times of under 100 microseconds, and 42GBps sequential read bandwidth in one enclosure. With 99.9999% reliability from EF600 storage systems, data for AI operations is available whenever and wherever it's needed.

Supercharge BeeGFS storage and metadata services

BeeGFS is a parallel file system that provides the flexibility you need to meet diverse and evolving AI workloads. NetApp EF-Series storage systems supercharge BeeGFS storage and metadata services by offloading RAID and other storage tasks such as drive monitoring and wear detection.

NetApp and NVIDIA: Promoting innovation together

The DGX A100 system is a next-generation DL platform that requires equally advanced storage and data management capabilities. Because it combines DGX A100 with BeeGFS building blocks based on NetApp EF600 systems, this verified architecture can be implemented at almost any scale. You can pair a single DGX A100 with a single BeeGFS building block. Or you can have up to 140 DGX A100 systems with a scalable number of BeeGFS building blocks that present a single storage namespace. With the superior cloud integration and software-defined capabilities of the NetApp product portfolio, NetApp solutions let you span the edge, core, and cloud for successful DL projects.

Solution components

- NVIDIA DGX A100 systems
- NetApp EF600 all-flash storage
- NVIDIA Mellanox Quantum QM8700 switches
- NVIDIA DGX software stack
- ThinkParQ BeeGFS parallel file system

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.

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

NetApp