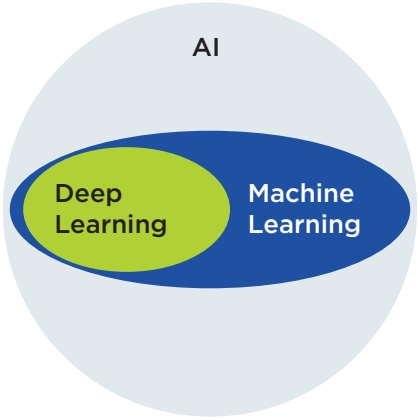
Accelerate AI Initiatives with a Powerful, Scale-Out Architecture

Streamline AI with best-of-breed technology from Nutanix, NVIDIA, and Mellanox

Enterprises across diverse industries including manufacturing, financial services, retail and healthcare are turning to digital transformation and artificial intelligence (AI) to gain greater insights from their business data to increase customer engagement, improve the efficiency of business processes, optimize the supply chain—and enable new business opportunities. AI success often means identifying promising ideas and quickly scaling projects from proof of concept to pilot to production.



AI encompasses all the techniques used to enable computers to perform tasks that would otherwise require human intelligence, including the machine learning and deep learning technologies that are currently hot topics in both the IT and popular press.

Machine learning (ML) is rooted in statistics and mathematical optimization, and has applications in prediction, analytics, and data mining. Machine learning approaches may include mathematical and statistical approaches such as decision trees and clustering algorithms in addition to deep learning.

Deep learning (DL) is the area of ML utilizing a variety of computationally intensive neural network architectures. Enterprises apply deep learning to detect patterns, understand behaviors, and make predictions based on large volumes of data such as data gathered from IoT.

The large volume and variety of data required, combined with the computational intensity of deep learning in particular, create unique infrastructure requirements that can be challenging for teams tackling AI projects to address. Architecting an AI infrastructure solution, deploying the necessary hardware and software, and identifying and eliminating bottlenecks can be an expensive, time-consuming and error-prone process.

"Even with the best data scientists and machine learning engineers, and the best tools, you won't get the best results on large data sets without the right infrastructure."

– InformationWeek

To enable your team to focus on its AI goals rather than the complexities of infrastructure, Nutanix has partnered with NVIDIA and Mellanox to design, test, and validate a reference architecture capable of taking on the world's toughest deep learning problems. This solution is quick to deploy, simple to scale, and delivers the data services necessary to meet the needs of software developers and data scientists.

NUTANIX
YOUR ENTERPRISE CLOUD

AI ARCHITECTURE COMPONENTS

Nutanix All Flash Appliance

Predictable scaling from a software-defined, distributed system designed to self-heal in the event of failure.

The Nutanix 3060 offers a blend of compute performance and storage capacity. A 2RU footprint provides:

- 4 independent compute nodes
- Up to 6 direct-attached SSDs per node, 24 devices per appliance.
- Up to 20 cores/socket and 768GB memory per node.
- Edge solutions available for inferencing.

NVIDIA DGX-1

The industry's leading AI platform, DGX-1 is optimized for the most complex AI problems. A 2U footprint delivers:

- 8 NVlink-connected GPUs
- NVIDIA GPU Cloud (NGC), with a full catalog of optimized AI software tools
- Container registry: NVIDIA-tuned, containers for top algorithms.

THE ADVANTAGES OF A BEST-OF-BREED ARCHITECTURE

The Nutanix AI approach combines the proven hyperconverged infrastructure technology in Nutanix Enterprise Cloud with NVIDIA GPU computing and low latency, 100Gb Ethernet switching from Mellanox. The result is a balanced best-of-breed solution designed to eliminate bottlenecks. The architecture supports large and diverse datasets, delivering data efficiently to NVIDIA GPUs, and scaling out as your needs grow.

Deep learning is computationally intensive. IT can require the parallel processing capabilities of multiple GPUs during the training of inference models. A large dataset may be needed for initial training of a model. As new data comes in, the model is periodically re-trained to further refine it, so the training process is iterative and the total AI workload grows over time. Advantages of the Nutanix solution include:

- **Ease of deployment and management.** comprehensive tools streamline deployment and ensure that firmware, patching and software upgrades are seamless. Easy-to-manage data protection and DR protect your AI operations.
- **Superior performance and scaling.** AI deployments can grow rapidly as you move into production. Datasets grow, new data sources are added, and algorithms increase in complexity. With Nutanix, you can start small and scale out as your needs grow without worrying about bottlenecks or having to re-architect.
- **Ease of DevOps.** AI training is an iterative process. Applications need to be updated regularly as model continue to be trained. Nutanix Calm makes it simple to manage application lifecycles and push out changes regularly without the manual processes that result in mistakes.
- **Built-in data security.** To ensure the security of sensitive data, many architects find they have no choice but to deploy dedicated infrastructure for AI. Nutanix software builds in security features including two-factor authentication and data-at-rest encryption in a hardened security framework that has been certified to ensure compliance with the strictest standards.

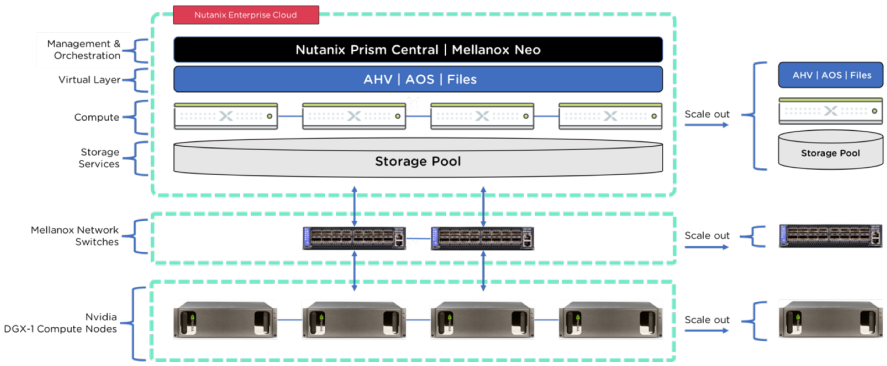
ARCHITECTURE DETAILS

The reference architecture combines the following elements as illustrated in the figure below:

- Nutanix 3060-G6 servers provide high-performance storage and data services, and CPU compute.
- NVIDIA DGX-1 provide the GPU computing necessary for complex AI algorithms.
- Mellanox 100GbE Network switches provide high throughput and low latency to pipeline data efficiently between Nutanix and NVIDIA.

All components scale out, ensuring that CPUs, storage capacity and I/O, network connectivity, and GPUs can all be expanded quickly as needed.

NUTANIX™
YOUR ENTERPRISE CLOUD



GETTING STARTED

To find out more about the Nutanix approach to AI, download the detailed reference architecture at www.nutanix.com/ai, or contact Nutanix at info@nutanix.com, follow us on Twitter @nutanix, or send us a request at www.nutanix.com/demo to set up your own customized briefing.

Mellanox

Mellanox delivers low latency, high-throughput Ethernet switching The Mellanox SN2100offers a compact 1RU footprint with:

- 16 ports (operating as 25, 40, 50 or 100GbE)
- Throughput of 3.2Tb/s
- 2.38Bpps processing capacity
- Scale-up to 128 ports

NUTANIX™
YOUR ENTERPRISE CLOUD

T. 855.NUTANIX (855.688.2649) | F. 408.916.4039
info@nutanix.com | www.nutanix.com | @nutanix