Technical White Paper

Edge to Core to Cloud Architecture for AI

Key Considerations for the Development of Deep Learning Environments

Sundar Ranganathan, NetApp

August 2018 | WP-7271

Abstract

As organizations embrace AI-powered innovations and deploy deep learning applications, it is crucial to design a robust infrastructure that takes advantage of the compelling features of both on-premises and cloud deployment options. You must consider several technical requirements before you design and deploy an AI-enabled application: the application type, required performance, data lake size and growth, data backups and archives, and TCO, to name a few. This paper describes these considerations and the corresponding NetApp® solutions that enable seamless data management and deployment options across the edge environments, on-premises data centers, and the cloud.

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

Artificial intelligence (AI) solutions and deep learning (DL) are everywhere these days. But the algorithms that they rely on must be properly trained on large datasets, and the rate at which DL datasets are growing can be astounding. For example, the amount of data that is collected by autonomous cars can scale to hundreds of terabytes (TB) each day. Examples from Internet of Things (IoT) applications are equally staggering.

In addition, training datasets for DL models must be managed across multiple physical locations throughout their lifecycle, from ingestion to transformation, exploration, training, inference, and archiving.

For these reasons, depending on the use case, DL benefits from a holistic architecture with data management that extends across edge, core, and cloud environments. A smart data architecture relieves data scientists of tedious data housekeeping tasks so that they can focus on their core mission of developing AI applications that deliver business value.

This paper describes an architecture for DL environments that span edge, core, and cloud environments. This architecture enables data management across all environments and multiple deployment options at each stage of the data pipeline.

2 Progression of the Data Pipeline

Traditional big data analytics are batch-processed, MapReduce-based jobs that use the Hadoop Distributed File System (HDFS) and data lakes based on either the open-source ecosystem or one of the commercial Hadoop distributions. However, analytics systems have evolved from the batch processing of relatively static datasets to newer approaches, such as Amazon Kinesis, Kafka®, and Apache Spark™, that can process streaming data in real time.

In the newer approaches, a second layer of a Spark-based real-time analytics has emerged above the data lake. This change in architecture means that analytics are being applied on the data streams as they are received. In addition, in-memory caching engines have brought in the ability to host the entire dataset in the cache. The infrastructure requirements change as the architecture moves from a batch-processed MapReduce platform to one real-time analytics platform.

Furthermore, another class of emerging compute applications layer on the same dataset – high-performance computing (HPC) simulations. HPC simulations are starting to converge onto the data pipeline as they take advantage of GPU accelerated computing and operate on top of datasets in industry verticals like manufacturing, supply chain, and logistics.

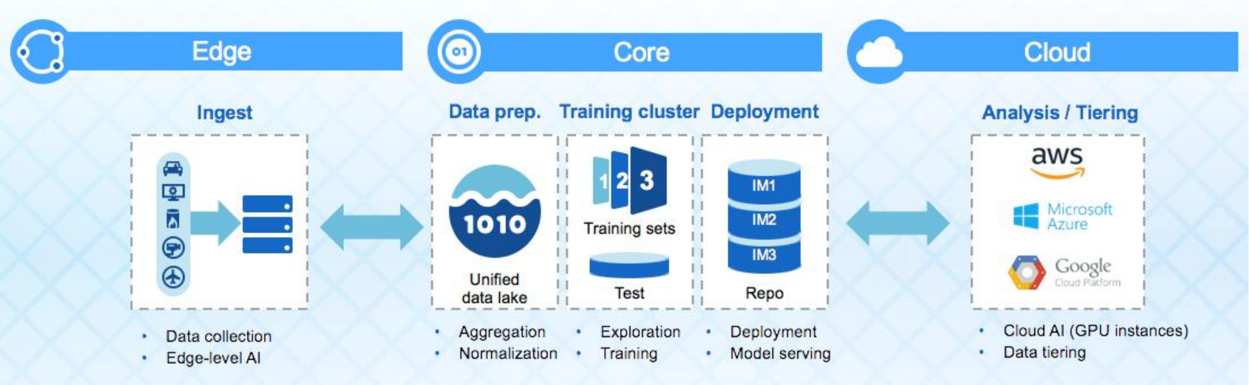
The emergence of smart IoT endpoints has resulted in an explosion of streaming data and the development of data analytics services, such as Amazon Kinesis and Spark Streaming, that add new components for handling the onslaught of data.

With AI applications and DL models, we have yet another architectural layer to account for. This evolution of the infrastructure requirements calls for decoupling data and compute, because the data that is being generated must be applied across various compute analytics engines. With this decoupled approach, compute cycles are offloaded from data management tasks. Furthermore, the goals around data protection and compliance can be met independently of the compute ecosystem.

3 The Edge to Core to Cloud Pipeline

At a high level, an end-to-end DL model deployment consists of three stages through which the data travels: edge (data ingest), core (training clusters, data lake), and cloud (data archival). This movement of data is very typical in applications such as IoT, where the data spans all three phases of the data pipeline. Figure 1 illustrates the stages of the data pipeline.

Figure 1) Edge to core to cloud data pipeline.



3.1 Edge

With the growth of IoT applications and consumer devices, a constant and significant amount of data is being generated and ingested at the edge. As an example, the edge can consist of an army of sensors that gather raw data, ranging from a few gigabytes (GB) to a few terabytes (TB) a day, depending on the application.

Moving this volume of raw data from thousands of edge locations over the network at scale is impractical and is prone to performance issues. Shifting analytics processing and inference at the edge is an ideal way to reduce data movement and to provide faster results. In such advanced solution cases, the data that is fed into the data lake is being operated upon from an analytics and an AI/DL perspective, and the trained AI/DL models are pushed back to the edge for real-time inferencing.

3.2 Core

The core is the heart of the data pipeline, and it demands high I/O performance. The core can be logically divided into four stages: data preparation, exploration, training, and inference.

- **Data preparation.** The data that is ingested from the edge has to be preprocessed (normalized and transformed) so that subsequent stages receive high-quality data in a format that the model can accept. This step is critical because the DL models are only as effective as the data that they are trained with. This step is also the stage in which the data is labeled so that the models can classify, predict, or analyze the phenomenon of interest with high accuracy. The nature of computation in this stage is suitable for enterprise-grade server central processing units (CPUs).
- **Exploration.** The data in the exploration phase is what data scientists use to formulate a recipe for success before the learnings can be taken into the training stage. This step includes but is not limited to experimenting with small to large datasets, deeper models, and different DL frameworks (such as Caffe2, PyTorch, MXNet, and TensorFlow).
- **Training.** Training is the process of feeding millions of example objects that enable the DL models to make predictions in identifying or in classifying the objects. This step is where the preprocessed data is fed to the neural networks for model training. The computation is highly parallel in nature; because graphics processing units (GPUs) are highly effective with parallel computations, they are beneficial to use in this stage. Depending on the dataset size and the depth of the neural network, the model training could run for hours to days. To maintain high system performance, the GPU caches should always be fed with data. Data systems that deliver high raw I/O bandwidth are a key requirement.
- **Inference.** The trained models are tested and deployed directly into production. Alternatively, they could be fed back to the data lake for further adjustments of input weights, or in IoT applications, the models could be deployed to the smart edge devices.

3.3 Cloud

The benefits of the cloud can be leveraged in several ways. You can use GPU instances for computation, and you can use cloud for cold storage tiering and for archives and backups. In many AI/DL applications, the data might span across the edge and/or the core and/or the cloud. As a result, you must orchestrate data across these environments.

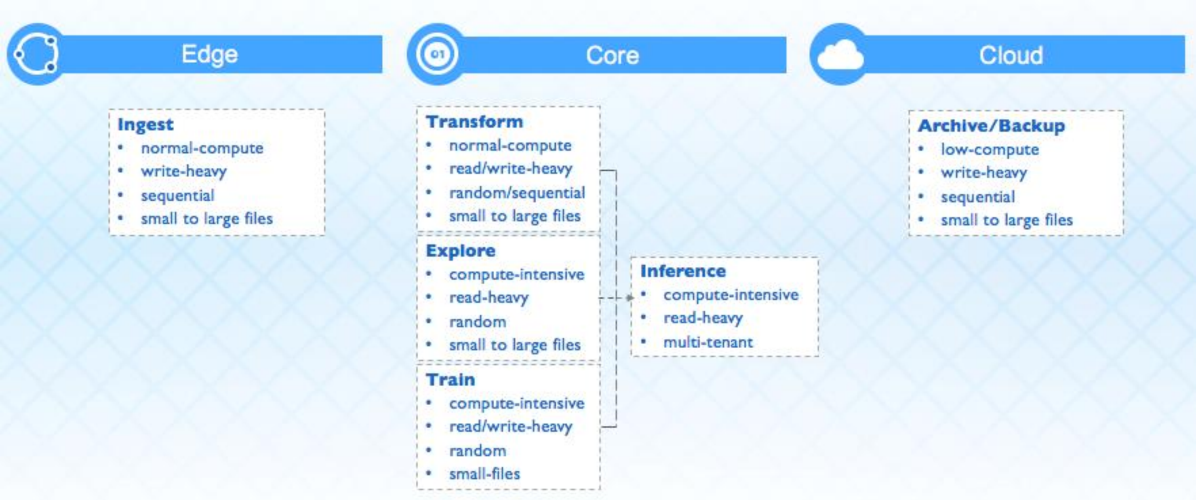
A solution ecosystem in the cloud offers an ideal environment to begin development of AI workflows, to run proofs of concept (POCs), and to form a foundation to expand upon. Organizations that expect to scale their investments in AI leverage on-premises solutions to be more cost-effective while using the cloud for test/dev and data lifecycle management. Data backups and archives are a key component in managing the lifecycle of data management in AI applications, for which petabyte-scale datasets are common. In such cases, the data must be moved between the cloud and the on-premises data center, and it is vital to account for data egress costs from the cloud.

Depending on the scale of the AI/DL workflows and the criticality of your organization's AI/DL strategy, you should evaluate performance and TCO requirements before you determine your infrastructure deployment options (cloud-only, on-premises-only, or hybrid architectures).

4 Data Characteristics

The first step toward determining the right deployment model is to understand the nature of data at each stage of the pipeline. The DL data pipeline should be capable of delivering smooth data flow while delivering the required performance for all data access patterns. Figure 2 highlights the characteristics of data traffic at each step from the edge to the cloud.

Figure 2) Data characteristics across the pipeline.



At the edge, the data traffic predominantly comprises small to large files that are sequentially written. Low latency is a requirement when you are looking at real-time DL inference models at the edge. The ability of the infrastructure to handle and to scale data movement back and forth from the edge to the core speeds up the overall workflow.

Traditional Big Data architectures have encouraged all data to be transferred into a single data lake (HDFS or S3). As the emerging dataset capacities have grown exponentially and varying types of data have emerged from Digital Transformations (DX), it is no longer optimal to use a single data lake for all data types. Rather, the data lake of the modern multi-ecosystem architecture is a federation across multiple data sources, chosen carefully to be the optimal placement for the corresponding type of data.

For example, time-series data may be placed in graph-databases or other NoSQL databases. Video, image, and audio files may be placed in a general-purpose shared filesystem such as NFS. HPC data may be placed in massively parallel filesystems such as Lustre and GPFS, log data in Splunk or Hadoop, unstructured data in HDFS or S3, and structured data may be in a variety of traditional RDBMS, newer NoSQL databases, or in business intelligence (BI) systems. The modern data pipeline architecture builds a federated data lake that is unified across various data types and data sources.

The data needs to be cleansed and transformed before it can be used for exploration and model training. In the transform stage, several processes must run to cleanse and to format the data based on the model’s input format. As the amount of data grows, it can stress storage systems that cannot accommodate the growth and the associated performance needs of transforming the data quickly. The explore phase is compute-intensive and consists predominantly of random read data traffic of varying file sizes. The data and the training processes are validated at this stage before the actual training can start. Because the access patterns can vary, this step demands a robust storage system.

The training stage demands very high compute cycles and the data is moved in small chunks to the GPU caches for computations, so the data streams are predominantly random small file reads in nature. This stage demands that massively parallel data streams be accessed quickly to achieve high GPU utilization and high system performance. This demand translates to high parallel bandwidth requirements, running into tens of gigabytes per second (GB/s) at latencies of less than 1ms.

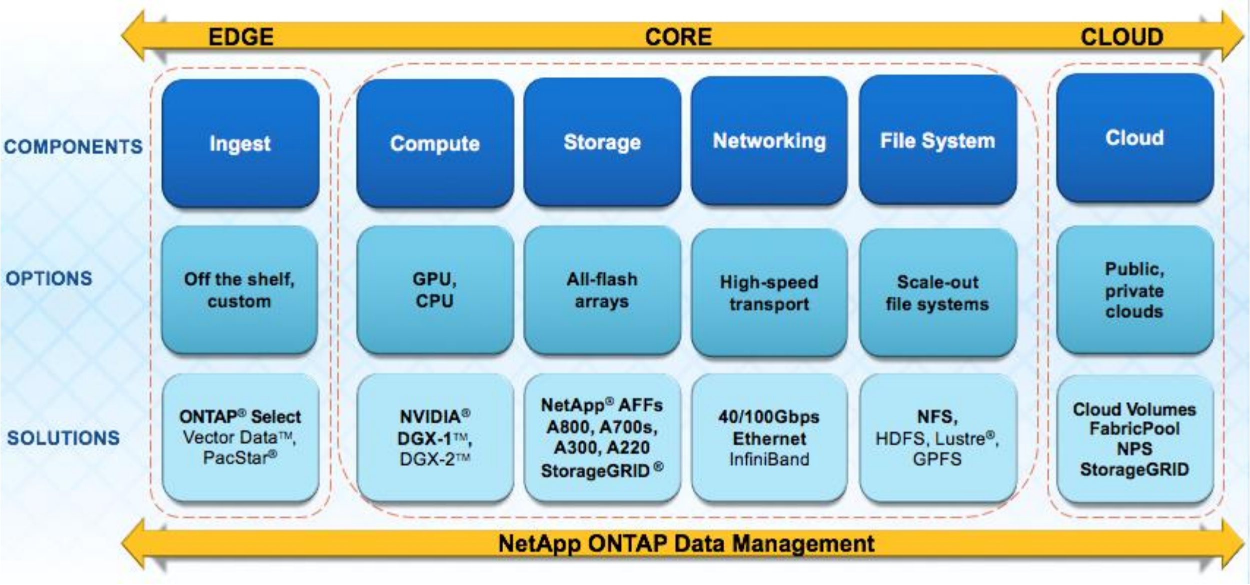
The deployment stage is when the resulting trained models are put into a DevOps-style repository and are subjected to evaluation testing. The data might be fed back to the training clusters for additional iterations, or the models could be deployed back to the edge devices in mature IoT implementations. When the models are trained to achieve the desired prediction accuracy, they can then be exposed to real-world data. This step is called inferencing.

To accomplish high productivity, data must move seamlessly and efficiently through each stage of the pipeline in order. Because high-quality data is an important factor in the success of training models, data growth can scale fast and unpredictably. It is vital to be able to use storage systems with a single namespace, enabling efficient and easy data management to cope with unpredictable scaling.

5 Data Pipeline Components

The data pipeline consists of many key components that deserve close attention so that you can design and implement a holistic environment for DL workloads. NetApp® has a robust line of products and services that manage data across the three realms of the AI infrastructure. Figure 3 presents an overview.

Figure 3) Data pipeline components with recommended solutions.



COMPONENTS	Ingest	Compute	Storage	Networking	File System	Cloud
OPTIONS	Off the shelf, custom	GPU, CPU	All-flash arrays	High-speed transport	Scale-out file systems	Public, private clouds
SOLUTIONS	ONTAP® Select Vector Data™, PacStar®	NVIDIA® DGX-1™, DGX-2™	NetApp® AFFs A800, A700s, A300, A220 StorageGRID®	40/100Gbps Ethernet InfiniBand	NFS, HDFS, Lustre®, GPFS	Cloud Volumes FabricPool NPS StorageGRID

5.1 NetApp Edge Solutions

NetApp has partnered with key hardware ecosystem vendors to deliver a solution at the edge. Ruggedized and small form factor hardware platforms from Vector Data™ and PacStar® are optimal solutions when they run the NetApp software-defined storage solution ONTAP® Select. These hierarchical edge devices are designed to aggregate data from sensors in offline environments. They also seamlessly move data to the on-premises data center or to the cloud through a WAN connection by using ONTAP features such as NetApp SnapMirror® technology.

5.2 NetApp Core Solutions

The core is the heart of the workflow, where data is prepared and models are trained. Key high-level components include compute, storage, file systems, and networking.

Compute

DL model training demands high computation cycles for parallelization. The computation in this stage can be broadly classified into two categories:

- **Forward pass.** Input is passed through the neural network for processing.
- **Backward pass.** The input weights are updated based on errors in the forward pass category.

A key point to note is that both these passes are predominantly matrix multiplication operations that can be executed in parallel. GPUs are several times more efficient in parallel computation than are enterprise-grade CPUs. GPUs consist of thousands of simpler cores, and CPUs consist of a few cores that are designed to implement instruction-level parallelism. CPUs can be used in the data preparation stage, where they are suitable for data normalization and transformation.

NVIDIA® DGX-1™ servers are being broadly adopted for DL workloads because they provide a rich software stack and support for development in addition to best-in-class hardware (GPU, CPU, and memory). With the enormous number of compute cycles that are required in the DL space, several organizations are looking to be more vertically integrated, leveraging purpose-built hardware implementations.

Storage

Keeping the GPUs waiting for data slows the performance of the system and prolongs model training. To keep the GPU caches loaded with fresh data, you need a storage system that can handle the high bandwidth requirements for model training. NetApp all-flash storage systems are well suited to meet the parallel performance requirements of the DL data pipeline while facilitating seamless scaling as the data lake grows. A single NetApp AFF A800 system supports 25GB/s of sequential read throughput and 1 million IOPS for random reads at sub-500µs latencies. The AFF A700s, AFF A300, and AFF A220 systems offer lower starting points in terms of performance, capacity, and cost. As the data lake scales in capacity, ease of data management is important. NetApp ONTAP FlexGroup volumes provide a single namespace and enable scaling from terabytes to 20PB in capacity.

File Systems

Depending on the DL workflow, the characteristics of data streams can vary. In many cases, the data traffic in DL consists of millions of files (images, video, audio, text files). NFS is perfectly suited for delivering high performance across a diverse range of workloads; NFS handles both random and sequential I/O well and can scale up and out seamlessly.

Although Lustre® and General Parallel File System (GPFS) are high-performance file systems that are built for scale-out, both are ideally suited for cluster computing applications such as high-performance computing (HPC). These file systems can be deployed for the DL development lifecycle, but their performance is not optimized for data streams that consist of small files.

HDFS is a distributed and scalable file system that was written for the Hadoop framework. This file system achieves reliability by replicating the data across multiple hosts (a default value of 3).

In DL workflows, the data growth can be massive and sudden. Maintaining multiple copies of large datasets is not scalable; HDFS is also not optimized for performance of both random and sequential I/O.

Networking

Having support for high-speed transport links is essential to prevent any networking bottlenecks in the infrastructure. Multisystem scaling of DL workloads requires the network transport to provide extremely low-latency and high-bandwidth communication. The login servers, data management tasks, and, optionally, storage can communicate over 10Gb Ethernet links. Compute servers and storage traffic can use 40/100Gb Ethernet or 100Gb InfiniBand (IB) depending on the system architecture and performance requirements. Because of its broad industry adoption, Ethernet is a popular choice. Creating discrete virtual LANs (VLANs) to separate compute data traffic, storage data traffic, and data management traffic is an efficient way to improve the overall system performance.

Remote direct memory access (RDMA) enables the network adapter to transfer data directly to and from application memory without involving the operating system, thus saving CPU cycles. This technology permits high-throughput, low-latency data networking, which is useful in DL environments. RDMA over Converged Ethernet (RoCE) is the most widely deployed implementation of RDMA over Ethernet, and it leverages new Converged Enhanced Ethernet (CEE) standards. CEE networks use Priority Flow Control (PFC) which provides the ability to optionally allocate bandwidth to a specific traffic flow on the network. For example, you can prioritize RoCE over all other traffic allowing 40/100GbE links to be used for both RoCE and traditional IP traffic (such as the NFS storage traffic) at the same time.

- Storage-side networking – Using a fast network to route storage traffic is crucial in ensuring the GPU caches are loaded with new data quickly, consequently maintaining high GPU utilizations and accelerating training runs. While storage systems supporting 40GbE are popular, only tier-1 storage vendors push the envelope by supporting 100GbE networking for storage traffic (AFF A800 supports 100GbE).
- Compute-side networking – For compute cluster interconnect, Ethernet is the popular choice due to wide industry adoption. The NVIDIA DGX-1 server uses 100Gb IB ports to provide RDMA communication between GPUs of two different DGX-1 servers without involving CPU or system memory. To enable seamless adoption among enterprise data centers, these ports can be configured to carry RoCE traffic among DGX-1 servers.

5.3 NetApp Cloud Solutions

Public and private clouds complement the on-premises AI/DL infrastructure with commoditized storage and compute as a service. Depending on the deployment and application requirements, the cloud can play either a central role or a supporting role in the workflow:

- AI applications that run on the cloud can use GPU instance types to run off-group or in-group model training and NetApp Cloud Volumes ONTAP for file services. Cloud Volumes ONTAP is a highly available storage solution on public clouds that supports grow-as-you-go file shares that use NFS, CIFS, or iSCSI file services.
- On-premises data centers can be used for DL model training, and cloud can be used for data tiering and lifecycle management. FabricPool can manage automatic data tiering of cold data to object storage services such as Amazon S3 and back.
- If your organization prefers tighter control of your data, you can use near-cloud solutions such as NetApp Private Storage (NPS). You can use GPUs on public clouds for training runs, and your data can reside in private data centers that are linked to public clouds through low-latency networks.
- You can use the cloud for backups and archives while managing model training on-premises.

5.4 Data Backups and Archives

In large-scale AI deployments, the data lakes and the data growth can be staggering. AI applications in industry verticals like healthcare and oil and gas exploration rely on a large amount of data collected over a long period of time. Only a small percentage of this data is hot, and the large volume of cold data needs to be archived securely while keeping costs to a minimum. NetApp StorageGRID® Webscale offers a compelling data archival solution; it helps you manage archive data in a single namespace across datacenters, supports widely used protocols such as Amazon S3 and Switft, and automatically migrates data to a cloud or between clouds.

The desire to access cold data instantly is becoming a business concern, as reworking older data is sometimes more cost-effective than acquiring new data. So, choosing an archival solution which can be deployed with various infrastructures is important. Here are few options to manage hot and cold data –

- Hot data in AFFs and cold data tiered to StorageGRID Webscale using FabricPool
- Hot data in AFFs and cold data tiered to StorageGRID Webscale using third-party data management software (via FPolicy)
- Hot data in AFFs and cold data tiered to Amazon S3/Azure Blob using FabricPool

5.5 ONTAP—Built for the Modern Enterprise

NetApp ONTAP is the industry's leading enterprise data management software. You can flexibly deploy storage on your choice of architectures—software-defined storage (edge), engineered systems (core), and the cloud—while unifying data management across all of them.

You can start small and grow with your business by scaling your storage environment, growing to as many as 24 nodes in a cluster (100s of TB to 100s of PB). In scale-out scenarios, ONTAP FlexGroup volumes enable easy data management in a single namespace logical volume of 20PB, supporting more than 400 billion files. For cluster capacities that are greater than 20PB, you can create multiple FlexGroup volumes to span the required capacity. Also, ONTAP reduces the overall storage costs by leveraging leading data-reduction technologies to minimize the storage footprint and to maximize the effective storage capacity.

ONTAP facilitates seamless data movement between architectures to place it in the optimal environment for high performance, high capacity, and cost efficiency. You can use ONTAP Select at the edge, ONTAP 9 at the core, FabricPool to move data to the cloud for tiering, and Cloud Volumes ONTAP for a managed, high-performance file system on the public cloud.

In section 6, we develop these ideas and build out sample architectures for various deployments of the data pipeline.

6 Types of Deployment Architectures

Among other reasons, the AI infrastructure deployment type depends on the stage of an organization's AI/DL projects, on the extent of investment, and on other strategic factors. Regardless of the architecture, choosing a storage solution that can efficiently and seamlessly manage your data across deployment borders is crucial to achieving efficient scalability.

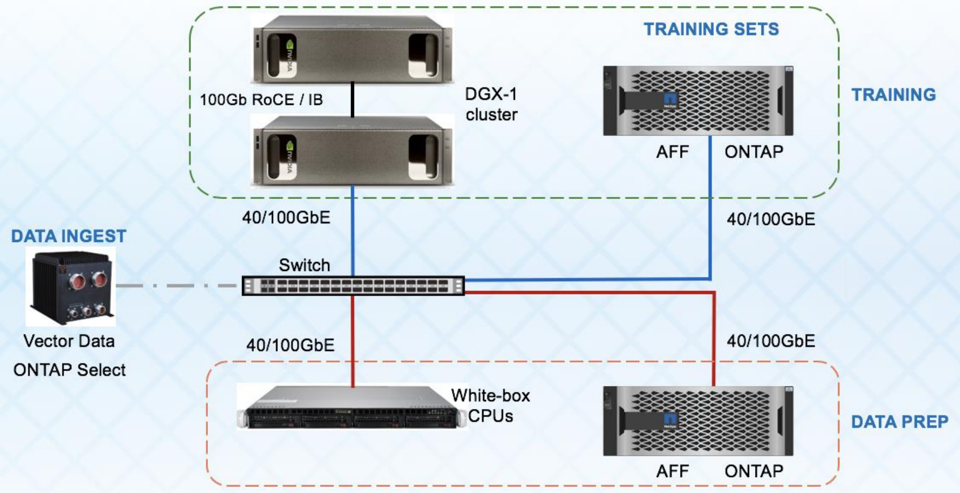
This section describes three sample architectures, weaving in the NetApp solutions that are optimally suited in each case.

Note: Some of the components of the architectures are included for the purpose of illustration, and there could be alternative approaches to meeting your infrastructure needs.

6.1 On-Premises Deployment Model

Figure 4 shows a sample architecture with a Vector Data device that runs NetApp ONTAP Select for data ingest, white-box CPUs for data preparation, and NVIDIA DGX-1 servers for model training. It also includes NetApp AFF systems that run ONTAP for data management. The white-box CPUs and DGX-1s are connected to the switch through 10GbE, and 100GbE links connect the two AFF systems. Multiple DGX-1 servers can be connected for sharing memory space through InfiniBand or through 100Gb RoCE links (cluster interconnect). Although not shown in Figure 4, FabricPool can be used for data tiering into the cloud.

Figure 4) On-premises deployment model.



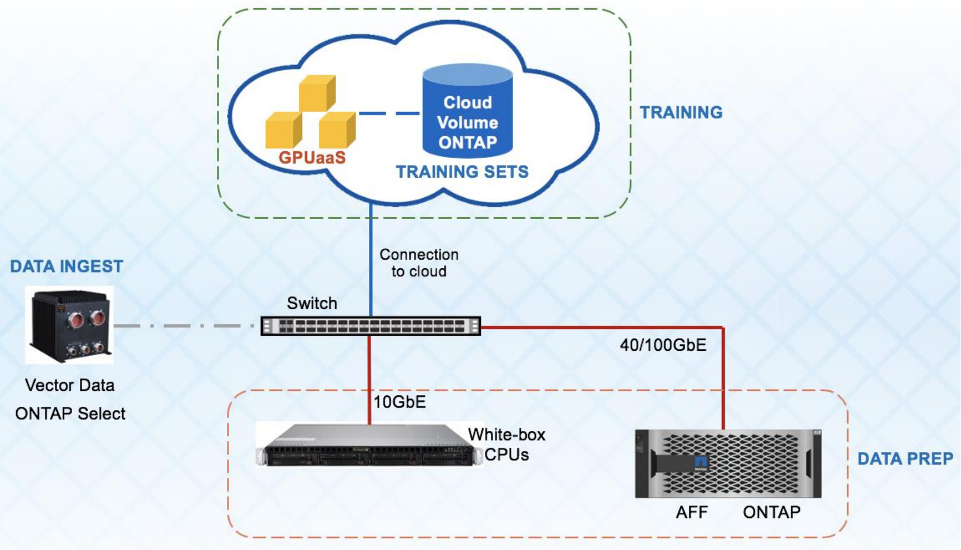
Depending on the availability of hardware and your organization’s appetite for CAPEX, you can use white-box GPUs or other implementations for compute as alternatives to the DGX-1 server.

6.2 Hybrid Cloud Deployment Model

An AI solution ecosystem that includes GPU instances on the cloud presents an ideal starting point for model training. In this architecture, GPUs on the cloud are used for model training, NetApp Cloud Volumes are used for storing training sets and model serving, and an on-premises AFF and CPU architecture can be used for data prep (Figure 5). Cloud Volumes provides highly available file services on public clouds and can scale from 0TB to 100TB in less than 10 seconds. Cloud Volumes can be natively provisioned and used through the management console of the cloud provider, making it easier for users to consume these services.

The NetApp Cloud Sync feature can be used for data transfers between on-premises and cloud storage. This feature provides a path if you prefer to start small in the cloud and transition to an on-premises deployment model as your data lake scales. This architecture can be modified to a cloud-only deployment model with the complete AI workflow managed in the cloud.

Figure 5) Hybrid cloud deployment model.



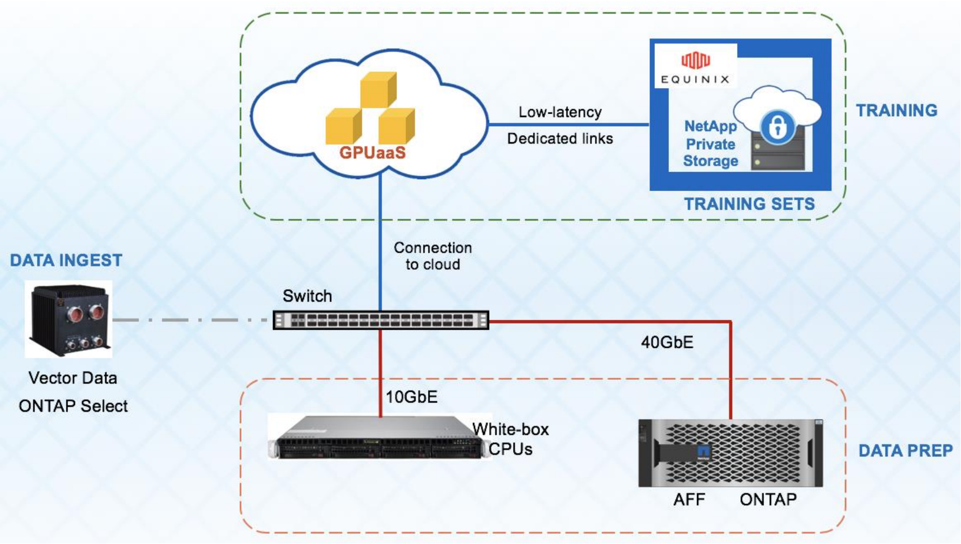
6.3 Near-Cloud Deployment Model

If your organization requires greater control over data, the near-cloud deployment model offers a viable solution (Figure 6). For example, you might need to maintain greater control for compliance reasons, for regional regulations such as the General Data Protection Regulation (GDPR), and for requirements of enhanced security (healthcare, defense, and national security).

In this architecture, GPU instances on the cloud are used for training, and the data is stored in data centers with dedicated high-speed links. Equinix offers such services, and you can use the NetApp Private Storage (NPS) solution to optimize your use of cloud services. NPS provides the freedom to switch between public clouds in seconds and helps you adhere to country-, regional-, and industry-specific data compliance standards.

In a variation of this architecture, all compute-related tasks are performed in the cloud, and data that is housed in the near-cloud data centers uses NPS.

Figure 6) Near-cloud deployment model.



7 The NetApp Solution Suite

NetApp offers a complete suite of products that are designed to meet the requirements of high-performance workflows such as AI/DL, enabling efficient data management. NetApp has the capability and the expertise to handle intelligent data movement from the edge to the core to the cloud. Because the quality and the quantity of data have a direct impact on the model accuracy, AI/DL workflows are inherently bound to consume more data. Following are the key industry-leading products from NetApp that enable you to implement a holistic AI/DL infrastructure.

- **Primary products and features for AI/DL workflows:**
  - **NetApp AFF** systems deliver industry-leading performance, capacity density, and scalability. These systems are well suited for demanding workloads such as AI and DL.
  - **NetApp ONTAP** is the data management software that powers AFF systems. It enables seamless data management capabilities and provides an efficient transition to a cloud-ready data center.
  - **NFS** is the most popular file system for AI/DL workloads; NetApp is the industry leader in NAS. The native NFS service in Azure is implemented by NetApp.
  - **NetApp ONTAP FlexGroup** enables massive scalability in a single namespace to more than 20PB with over 400 billion files, while evenly spreading the performance across the cluster.
- **Products for the edge:**
  - **ONTAP Select** is a software-defined storage solution that can be deployed on your choice of commodity hardware. It combines the best of the cloud in terms of agility, capacity scaling, with the resilience, and locality of on-premises storage. Hardware platforms at the edge from Vector Data and PacStar run ONTAP Select for data management.
  - **NetApp SnapMirror** is an ONTAP feature that replicates volume snapshots between any two ONTAP systems. This feature optimally transfers data at the edge to your on-premises data center or to the cloud. It efficiently transfers only changes, saving bandwidth and speeding replication.
- **Products for the cloud:**
  - **FabricPool** provides automatic storage tiering capability for cold data to object storage, such as Amazon S3, and back. FabricPool is compatible with Amazon S3 for data tiering on the cloud. StorageGRID can also be a FabricPool target for cold data tiering.
  - **NetApp Cloud Volumes ONTAP** is a managed, high-performance file system that enables you to run highly available workloads with improved data security in public clouds.
  - **NetApp Private Storage (NPS)** for cloud provides a near-cloud storage solution with data sovereignty, elastic compute, and multi-cloud access.
  - **NetApp Cloud Sync** service offers a simple and secure way to migrate data to any target, in the cloud or on your premises. Cloud Sync seamlessly transfers and synchronizes your data between on-premises or cloud storage, NAS and object stores.

8 Conclusion

The benefits of AI have become more apparent, and an increasing number of organizations seek to deploy solutions to glean tangible insights from the large amounts of data that they generate. Infrastructure plays an important role in defining success for your AI/DL applications. It is crucial to take a holistic approach to cover all deployment scenarios without compromising on the agility to expand as you need to and while keeping costs in check.

With a cloud-first strategy, NetApp has the right suite of products to simplify data management across edge, core, and cloud environments. NetApp also has real-world expertise in deploying small- to large-scale AI solutions.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

Trademark Information