



White Paper

Paras and ONTAP: Unlocking the AI Value Chain for Enterprises

Architecture Overview

Madhurima Agarwal, NetApp
Shivaram K R, Curl
May 2020 | WP-7324

In partnership with



Abstract

This whitepaper shows how the network infrastructure and data provisioning capabilities of NetApp® and the artificial intelligence (AI) automation capabilities of Curl can improve the efficiency of the development and implementation of AI processes, thereby creating cost savings.

TABLE OF CONTENTS

1	Executive Summary	4
1.1	Scope of the Document	4
1.2	Objectives	4
2	NetApp Architecture	5
2.1	Data Fabric	5
2.2	NetApp ONTAP AI	5
2.3	ONTAP Software	6
3	Paras Architecture	6
3.1	Paras Pods	6
3.2	Use of Kubernetes and NVIDIA	7
4	Use Cases	8
4.1	Introduction	8
4.2	Datasets	8
4.3	Workflow	9
5	Performance Improvements	11
6	Additional Capabilities with NetApp Products and Concepts	13
6.1	FlexVol	13
6.2	FlexGroup	13
6.3	Quality of Service	13
6.4	NetApp Volume Encryption	14
6.5	Data Protection	14
6.6	NAS	14
6.7	SAN	14
6.8	Storage Efficiency	14
6.9	ONTAP FabricPool	15
6.10	NetApp Trident	15
6.11	Edge to Core to Cloud	17
7	Benefits of Deploying Paras in a NetApp Environment	18
7.1	Technical and Operational Benefits	18
7.2	Business and Financial Benefits	19
8	Conclusion	19

Where to Find Additional Information	19
Acknowledgements	20
Version History	20

LIST OF FIGURES

Figure 1) ONTAP AI solution rack-scale architecture.	5
Figure 2) Paras system modules.	7
Figure 3) Paras, NVIDIA, and ONTAP integration.	8
Figure 4) Paras model-wise performance on a NetApp system.	11
Figure 5) Paras model-wise performance on a number of batches on a NetApp system.	12
Figure 6) Paras total training time on a NetApp system.	12
Figure 7) FlexGroup volume.	13
Figure 8) Kubeflow.	17
Figure 9) NetApp solution for the deep learning pipeline.	18

1 Executive Summary

1.1 Scope of the Document

This document describes how NetApp® ONTAP® AI is integrated with the Paras AI solution. The integration of these solutions benefits businesses that need AI solutions that are computationally heavy but require ease of data connectivity.

The document also describes the basic hardware and software structure for the storage-compute environment, future work, and the benefits of this environment.

1.2 Objectives

Design an Automated ML System

Massive amounts of data are being generated in many enterprises, but this data is often not being used effectively. Data is only as effective as the insights derived from it. Artificial intelligence (AI) is a powerful tool for deriving actionable intelligence from data.

AI is fueling the fourth industrial revolution, which is transforming processes in every industry and affecting every vertical. This revolution is increasing efficiency and creating new opportunities with a positive effect on both the bottom and top lines.

Building AI models to effectively mine data requires large data science teams, structured data, and massive computation resources. Setting up these teams is costly, and it typically takes several weeks or months to build an AI model. Data collected from various sources has to be structured and made available seamlessly. This collection can be challenging because of the huge quantities of data. The training of AI models requires massive computation abilities, such as heavy GPUs to train deep learning (DL) models. If the wrong architecture is used for training and inference of the models, the results can be costly.

Today there are a huge number of automated systems with a large amount of data flowing at a very rapid pace. The data generation dynamics are changing quickly, and it is difficult to build models manually to forecast, understand, or find anomalies in the data. What is needed is a system that automatically builds the model in real time for various data-science-related activities.

Paras is an automated machine learning platform from Curl Analytics that makes predictions based on data provided to it by the end user. Paras is capable of performing classification, regression, forecasting, image analytics, speech analytics, text analytics, and more for many types of businesses. Paras can build machine learning (ML) pipelines and can select and build ensembles of models to provide accurate models automatically.

Paras uses the most recent advances in technologies such as deep learning and reinforcement learning to build models and pipelines. It is highly modular and uses Kubernetes Orchestrator to connect with various Dockerized microservice pods. Each pod has a different task, and each pod is managed by a resource manager for efficient resource utilization. Paras enables business experts to build AI models on their own. It reduces model development time by five times and provides highly accurate output.

Outline of the System Deployed in the NetApp Environment

To solve the AI requirements of enterprises, Curl Analytics and NetApp have worked together on a unique solution that can greatly reduce the effort and cost of AI adoption. Our solution not only reduces the cost of data science teams, it also reduces cloud and data integration costs. Paras is an automated ML engine that can generate AI models automatically. It provides data seamlessly from varied sources, enabling you to create an AI solution with ONTAP and NetApp Trident to train models in a hybrid cloud ecosystem. The solution reduces cloud costs and offers high reliability and security of data across all AI processes, making it possible to build AI models up to five times faster and at one-tenth the cost.

The ONTAP solution includes capabilities such as data storage, various ways of data connectivity and data movement, and computation hardware. It has one NVIDIA DGX for performing AI model training. NetApp Trident provides seamless data movement from NetApp storage to Kubernetes Orchestrator, which is accessible by the NVIDIA DGX GPU for training the AI models.

2 NetApp Architecture

2.1 Data Fabric

The data fabric delivered by NetApp is an architecture and set of data services that provides consistent capabilities across a choice of endpoints spanning on-premises and multiple cloud environments. The data fabric simplifies and integrates data management across the cloud and on premises to accelerate digital transformation.

2.2 NetApp ONTAP AI

This proven NetApp ONTAP AI architecture, powered by NVIDIA DGX supercomputers and NetApp cloud-connected storage with ONTAP data management software, has been developed and verified by NetApp and NVIDIA. It provides a prescriptive architecture that enables your organization to:

- Eliminate design complexities
- Independently scale compute and storage
- Start small and scale seamlessly
- Choose from a range of storage options for various performance and cost points

ONTAP AI integrates NVIDIA DGX-1 servers and NetApp AFF A800 systems with state-of-the-art networking. ONTAP AI simplifies AI deployments by eliminating design complexity and guesswork. Your enterprise can start small and grow nondisruptively while intelligently managing data from the edge to the core to the cloud and back.

Figure 1) ONTAP AI solution rack-scale architecture.

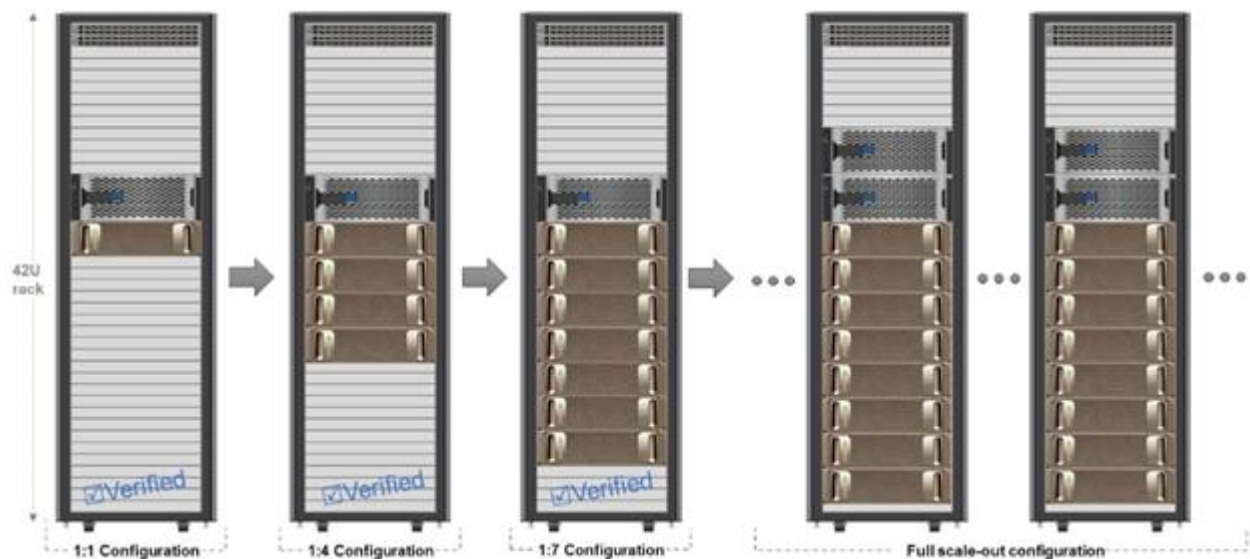


Figure 1 shows the scalability of the ONTAP AI solution. The AFF A800 system has been verified with seven DGX-1 servers and has demonstrated enough performance headroom to support even more

DGX-1 servers without affecting storage throughput or latency. By adding more network switches and storage controller pairs to the ONTAP AI cluster, the solution can scale to multiple racks to deliver extremely high throughput and to accelerate training and inferencing. This approach gives you the flexibility to alter the ratio of compute to storage independently according to the size of your data lake, your DL models, and the performance metrics that you need.

The number of DGX-1 servers and AFF systems that you can place in a rack depends on the power and cooling specifications of the rack that you use. Final placement of your systems is subject to computational fluid dynamics analysis, airflow management, and data center design.

2.3 ONTAP Software

NetApp ONTAP is data management software that has different versions for edge devices, on-premises storage arrays, and cloud storage.

With ONTAP, you can perform the following tasks:

- Manage data efficiently
- Seamlessly move data from one location to another
- Keep data secure in storage devices and when it's moved from one location to another
- Protect and back up data as required
- Gain insights from data

We briefly describe the capabilities of ONTAP and related products and their potential uses with respect to Paras in section 6, "Additional Capabilities with NetApp Products and Concepts".

3 Paras Architecture

This section describes the architecture and capabilities of Paras and NetApp in detail.

Paras is an automated machine learning engine that can be used for AI-based predictions. Paras takes in data from various sources such as a data warehouse, a database, flat files, and more. The data can be multivariate or univariate. Using the front end or settings file, the data can be specified for forecasting, classification, or other ML activity. Calibration and learning happen automatically.

The following steps are needed to create and deploy a trained model:

- Create additional features to enhance the data feature set.
- Estimate the characteristics of the given data set.
- Determine which models to test on the data set from the universe of models.
- Select the best model or models for the given dataset using various validation techniques.
- Deploy the trained model with one click.

Paras has a curated set of hundreds of models of type, time series, machine learning, and deep learning to facilitate model development. In addition, the platform is domain agnostic and has been tested on data from various industries such as finance, manufacturing, health care, economics, energy generation, and more.

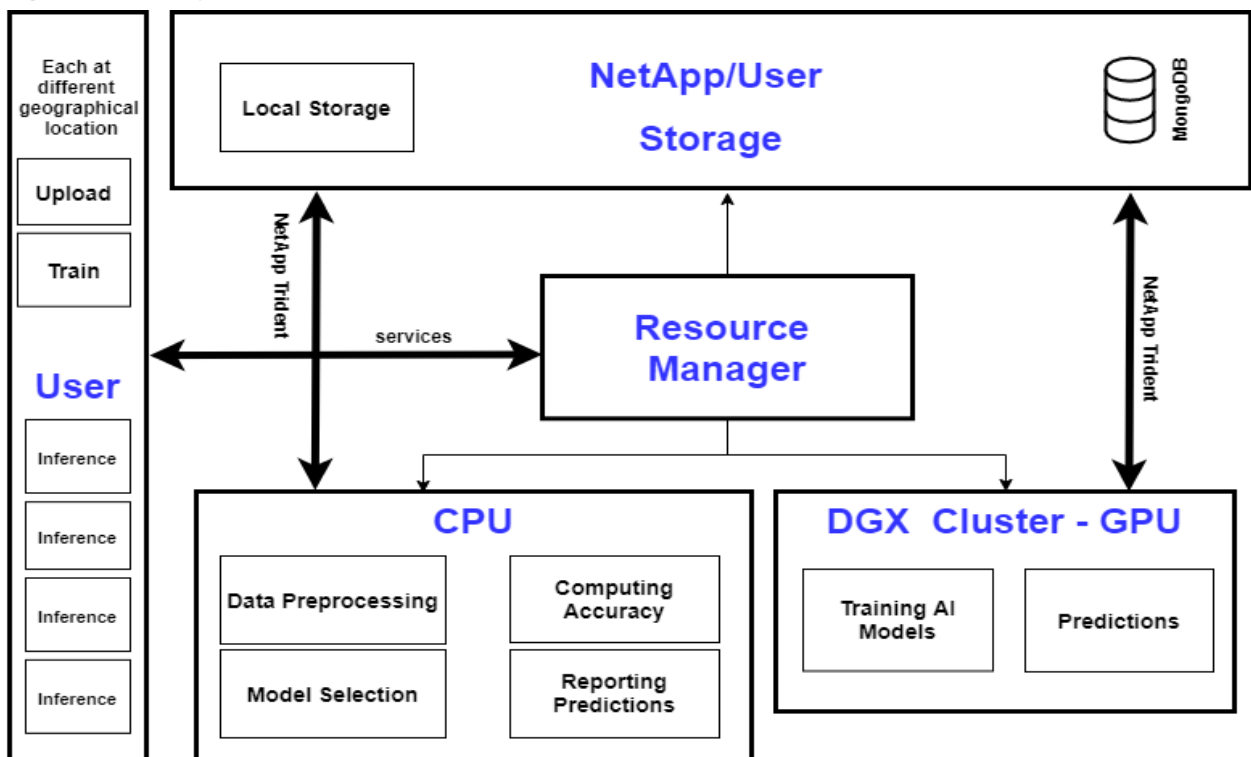
3.1 Paras Pods

Paras has the following pods, which are microservice enabled:

- **Data loading pod.** Data loading can be time consuming if the dataset is too large. This pod creates a database in MongoDB and a local file storage system in NetApp storage or the client cloud. All the images and metadata are stored by using this pod as a service.

- **Preprocessing pod.** Loads the data and performs the calculations required to preprocess the data for training. All the necessary calculations are stored in the MongoDB database so that it can be used for training and predictions.
- **Training pod.** Trains the AI model by loading the data using the GPU capacity provided by NetApp. The data is preprocessed and sent in batches to the model for training. This pod trains about 20 or more deep learning models and stores the performance of each model in the database for the user to view as required.
- **Model selection pod.** Selects the best model based on the accuracy of each model and stores the information.
- **Inference pod.** Employed by the user to get real-time predictions. This pod is very interactive, with many features that allow the user to analyze the work and predictions.
- **Resource Manager.** Responsible for sequentially allocating tasks to each pod. All the pods are microservice enabled, and first-come, first-served logic is used. Resource Manager is currently also the link between the front end and the core engine; it serves as a web framework application (Figure 2).

Figure 2) Paras system modules.



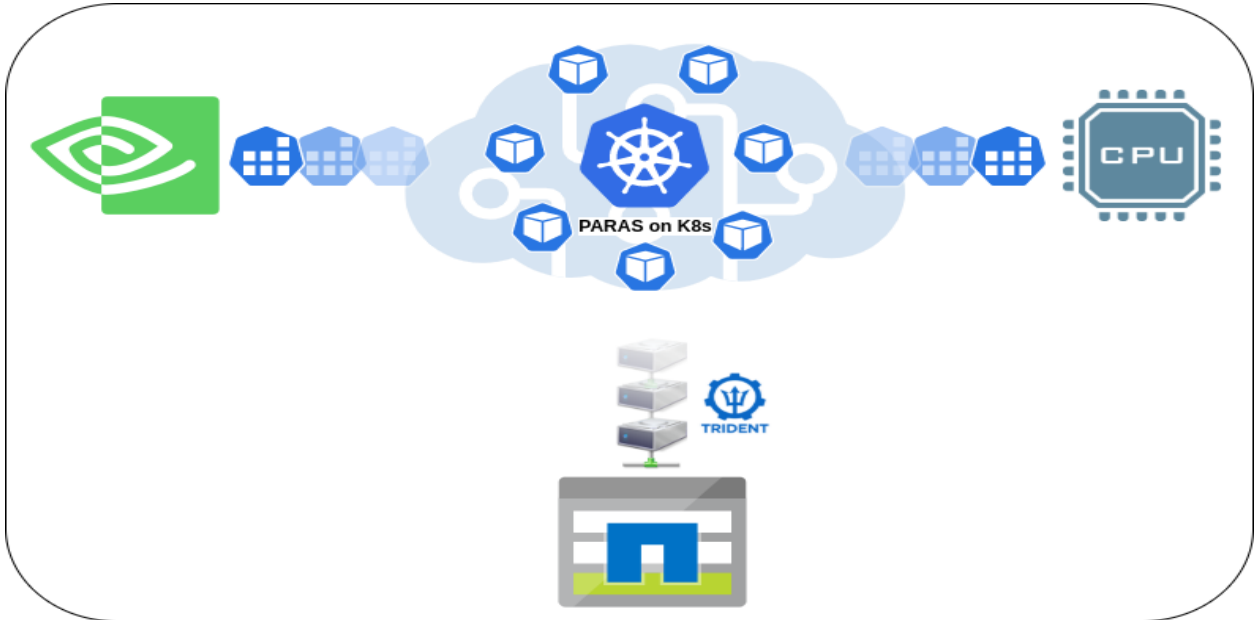
3.2 Use of Kubernetes and NVIDIA

Kubernetes offers an easy way to deploy any machine learning, allowing developers to automate the deployment process with very little effort. The Paras AI solution has multiple services and is scaled across multiple nodes. Kubernetes helps Paras pods to interact with each other and to allocate resources to each pod: CPU, GPU, and NetApp storage.

NVIDIA DGX is built for deep learning, receives jobs from Kubernetes, and allocates GPUs to the Paras Training pod. Kubernetes can access GPUs with the help of the NVIDIA Docker device plug-in. NVIDIA DGX supports out-of-box Ubuntu 18.04 LTS.

The end user uses CIFS/NFS to upload the dataset to NetApp enterprise storage, and NetApp Trident provisions the volume, which is directly mounted into Paras pods. Paras pods can seamlessly access data such as file system data (Figure 3).

Figure 3) Paras, NVIDIA, and ONTAP integration.



4 Use Cases

4.1 Introduction

Paras was tested on the NetApp ONTAP platform on two medical datasets that are available for public use (described in the next section). Paras is compatible with the NetApp system and runs the complete flow on the ONTAP solution provided by NetApp. The steps used to train the model with the datasets and inference the results using the trained model and new data are described in section 4.3.

4.2 Dataset

The following medical dataset was chosen to demonstrate Paras:

Name of dataset: Blood Cell Images

Image: Augmented images of four blood cell types

Number of classes: 4 (Eosinophil, Lymphocyte, Monocyte, and Neutrophil)

Number of Images: 12,500 (approximately 3,000 for each class)

Dimension of image: 640 x 480

Business value: The diagnosis of blood-based diseases typically involves identifying and characterizing a patient's blood samples. Often the diagnosis of the blood cell reports involves intervention of an expert doctor to identify which blood cell image that report shows (Eosinophil, Lymphocyte, Monocyte, or Neutrophil) and then diagnosing it. Identifying the type of cell in a few seconds means that the doctor could potentially diagnose the patient and start treatment much faster.

4.3 Workflow

The Classification module of Paras is employed for the use cases discussed. The Classification module can handle both single-label and multilabel classification. Single-label classification means that one image belongs to only one category; multilabel means that a single image can belong to multiple categories. For example, a person's brain MRI scan might detect multiple types of hemorrhage; that would be classified as multilabel. If only one type of disease is present in one image, then it belongs in the single-label category.

Here is the workflow for the Blood Cell Disease Detection dataset:

1. Upload the images (in a Zip folder) with metadata (in CSV):
 - a. Create a new project based on the type of problem statement. In this case, it is detection of blood cell disease.
 - b. Upload the images in a Zip folder with a CSV file that contains the image names as one column and the class or category of the image as the other column.
 - c. Verify that the CSV file is uploaded. The Image and Class tab is automatically populated with the column names provided in the CSV. Select the correct combination accordingly.
 - d. Save the project with a project name and the dataset name. Now you can start training.

The screenshot displays the Paras Classification module interface. At the top, there are tabs for 'DENOISE', 'OBJECT DETECTION', and 'CLASSIFICATION'. Below these, there are sections for 'NEW' and 'EXISTING' projects. The main configuration area includes fields for 'Select Image Folder' (Image_Folder.zip), 'Select CSV File' (blood.csv), and 'Multi Label' (checked). A 'CSV File Configuration' section allows selecting the 'Image Column' (Images) and 'Class Column' (Category). The 'Project Name' is 'Blood Disease Detection', the 'Dataset Name' is 'Blood Cell', and the 'Image Type' is 'JPEG'. A 'TRAIN' button is visible. Below the configuration, a table shows the CSV file content:

Images	Category
..72_3507.jpeg	EOSINOPHIL
..29_7135.jpeg	EOSINOPHIL
..11_9187.jpeg	EOSINOPHIL
..13_6624.jpeg	EOSINOPHIL
..34_720.jpeg	EOSINOPHIL

2. Train the AI model:
 - a. Paras preprocesses the data completely, using the information according to the metadata (in CSV) provided by the user.
 - b. The images are standardized, brought to the same size, and so on. Paras uses multiprocessing techniques to save time.
 - c. Paras has 20+ models to train in the image processing pipeline. All the models are trained and validated.
 - d. Paras uses different model selection techniques and stores the best model.
 - e. In this case, Paras was able to handle 600,000 images in 250GB of training data. Paras fully utilized the hardware capacity provided by NetApp. The CPU capacity was 400GB memory and 24 cores.
 - f. All the images were trained batchwise. DGX was fully utilized for training.
 - g. The user can view the training results.

DENOISE OBJECT DETECTION **CLASSIFICATION**

NEW EXISTING

Select Image Folder Select the Image Folder
 Select CSV File Select the CSV File
 Multi Label

CSV File Configuration
 Select the Image Column Class
 Select the Class Column

Project Name Dataset Name Image Type Select the Image Type

TRAIN

Accuracy Report	Support	Precision	Recall	Accuracy	F1-Score
	1979	0.9922014583741756	0.9921862874355649	0.9959537087185106	0.992192800402039

Rows per page: 5 1-1 of 1

3. Perform inference:

- With the Inference tab or an existing tab, a lab technician uploads the image of a radiograph.
- The lab technician clicks the Test button and sees the predictions by the saved AI model.
- A doctor is informed about the report provided by the lab technician. The doctor can review the radiograph along with the predictions and save the report.
- The doctor can generate a PDF of the report to give to the patient.

DENOISE OBJECT DETECTION **CLASSIFICATION**

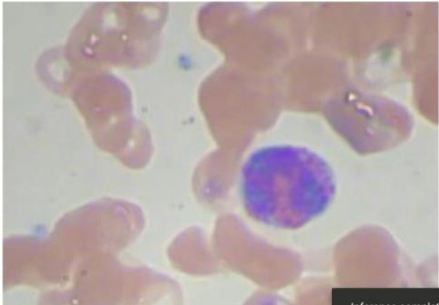
NEW EXISTING

NEW **REVIEW**

Upload Image Choose file
 Select the Project Blood_Cell
CLASSIFY

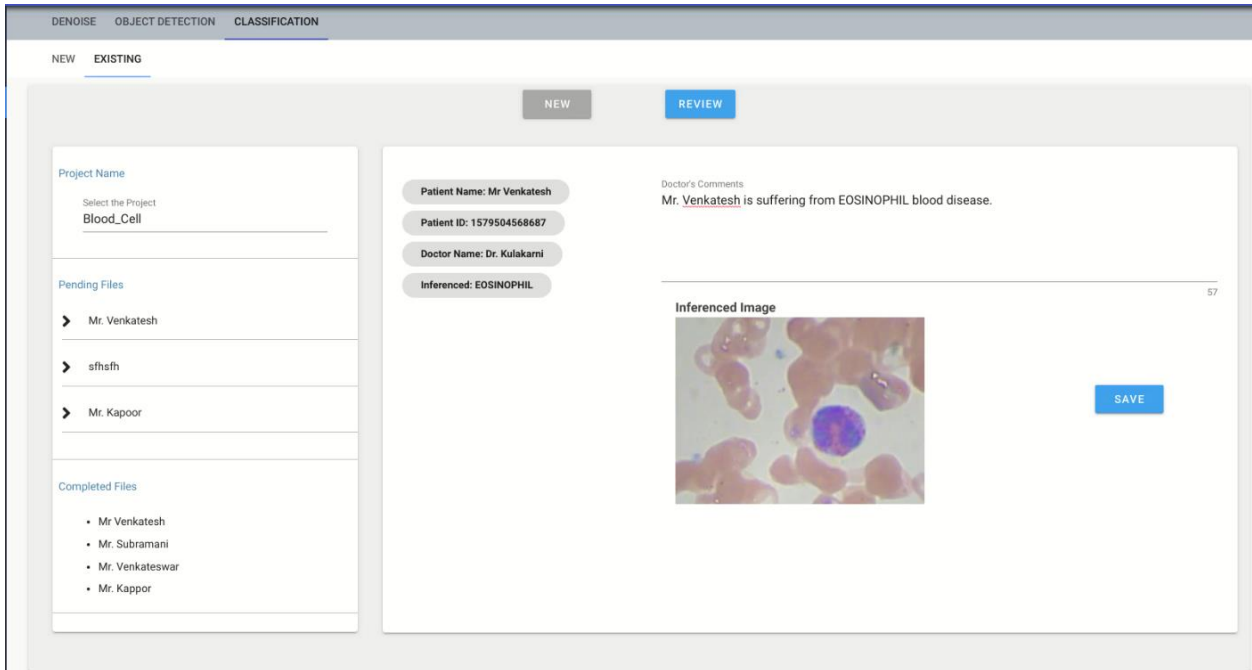
Patient Information
 Patient Name Mr Venkatesh
 Doctor Name Dr. Kulakarni

Input Image



Classified Image is: EOSINOPHIL

Inference completed! **CLOSE**



These steps describe Paras from an end-user perspective. The user interface makes it very user friendly, especially for end users without an IT background.

5 Performance Improvements

Paras was separately tested on the NetApp platform and also on a more typical setup (a single Titan GPU with SSDs). A NetApp system containing one DGX was used in the experiments. The results shown in this section were produced using the same standards. The experiment was performed with only one epoch using the same models separately on both systems with various datasets.

Figure 4 shows the time taken by each model for training in one time period. It clearly shows that the NetApp system is 2 to 5 times faster than a typical setup.

Figure 4) Paras model-wise performance on a NetApp system.

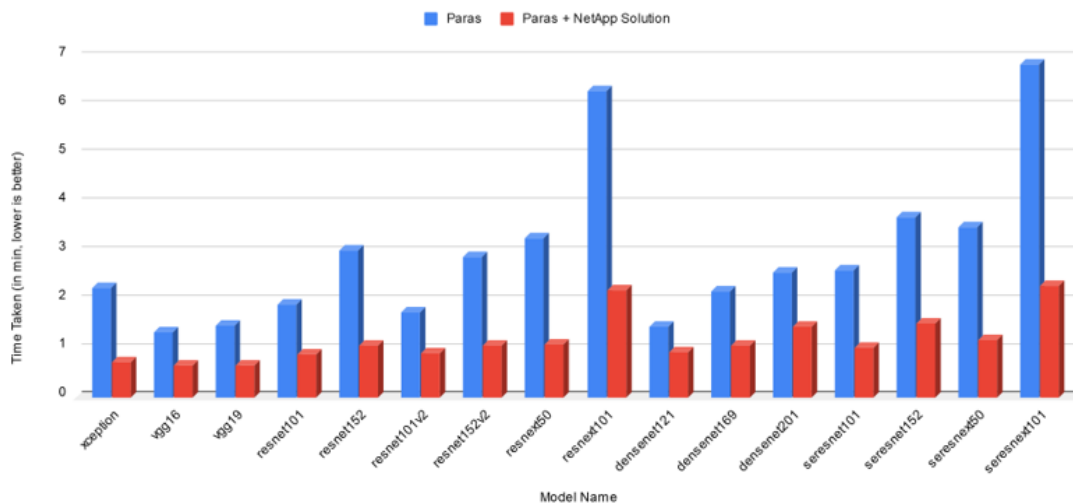


Figure 5 shows the maximum batch size that can be used during training. Using the ONTAP solution, it is possible to train with a batch size 16 to 32 times greater than normal, which can greatly increase speed.

Figure 5) Paras model-wise performance on a number of batches on a NetApp system.

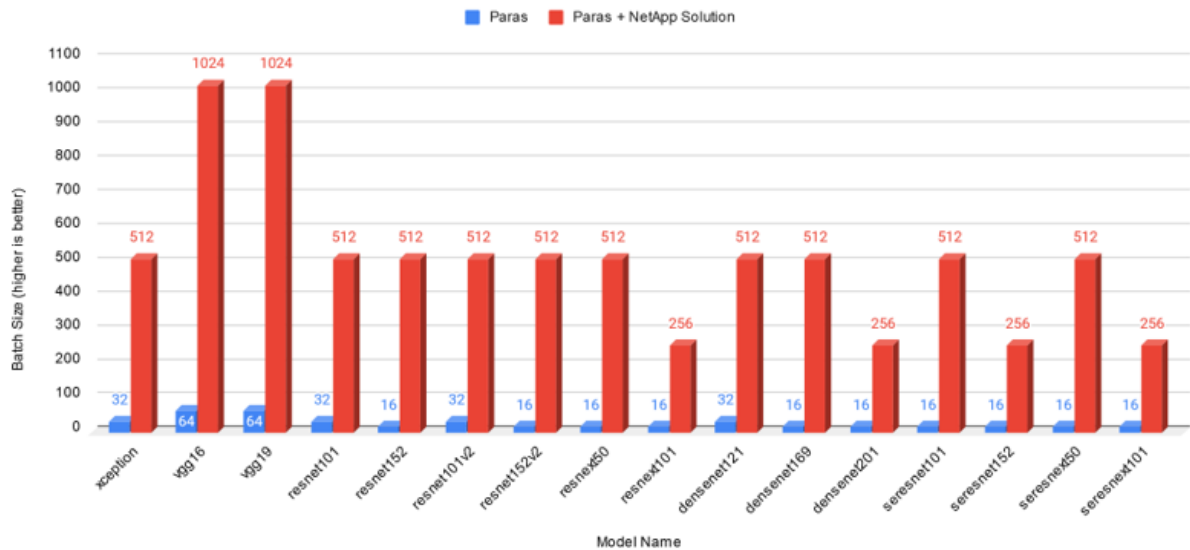
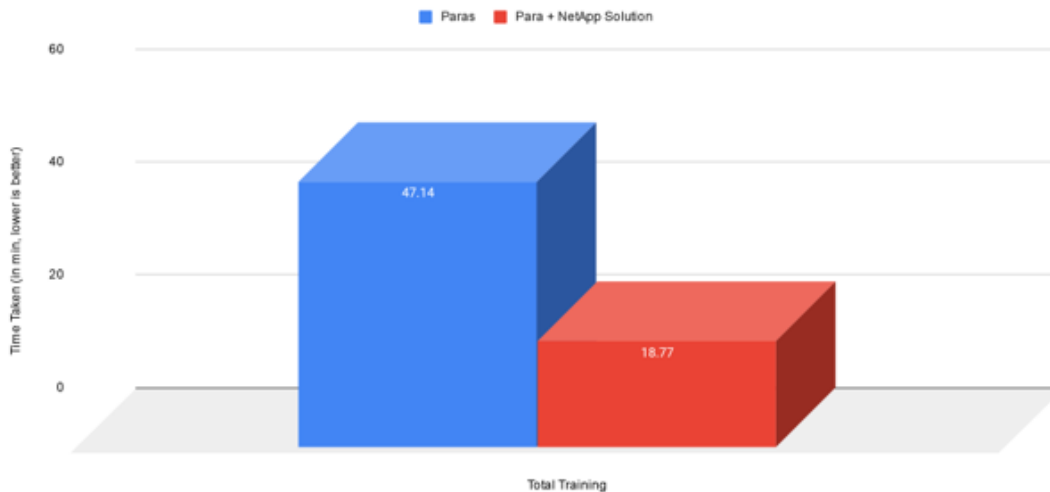


Figure 6 shows that the total time taken using a single DGX during training is about one-third of that of a typical setup on average.

Figure 6) Paras total training time on a NetApp system.



In conclusion, NetApp can process about 16 to 32 times more images at a time than a typical solution. NetApp is three times faster when a single DGX instance is used for training, making it possible to train multiple deep learning models and get the best output in 1 to 2 days, further increasing the time savings for Paras.

6 Additional Capabilities with NetApp Products and Concepts

Data management is crucial for your enterprise IT operations so that you can use appropriate resources for your applications and datasets. NetApp ONTAP includes the following features to streamline and simplify your operations and reduce your total cost of operation.

6.1 FlexVol

A FlexVol® volume is a data container associated with a storage virtual machine. It gets its storage from a single associated aggregate, which it might share with other FlexVol volumes. It can be used to contain files in a NAS environment, or LUNs in a SAN environment.

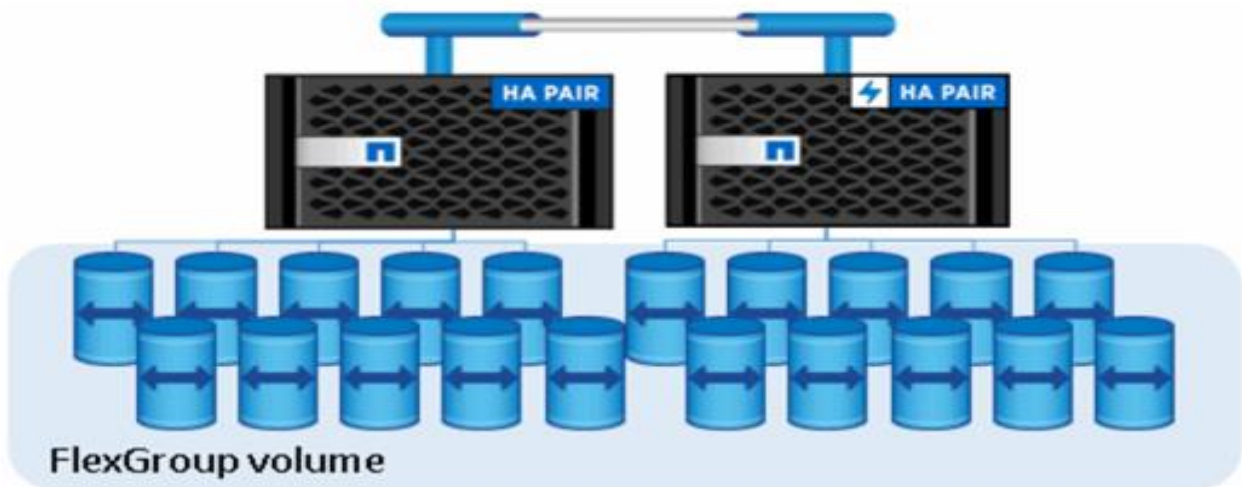
It enables the user to clone the volume, balance loads, reduce network latency, apply quotas to users, move the volume between aggregates and between storage systems, and more.

6.2 FlexGroup

A NetApp ONTAP FlexGroup volume is a high-performance data container that can scale linearly to up to 20PB and 400 billion files, providing a single namespace that simplifies data management. A FlexGroup volume is a scale-out NAS container that provides high performance along with automatic load distribution and scalability. A FlexGroup volume contains several constituents that automatically and transparently share the traffic.

This feature is extremely useful to store very large datasets in a single namespace, and the data lake can be virtually mapped to one single storage volume (Figure 7).

Figure 7) FlexGroup volume.



6.3 Quality of Service

Storage quality of service (QoS) can help you manage risks associated with meeting your performance objectives. You use storage QoS to limit the throughput to workloads and to monitor workload performance. You can reactively limit workloads to address performance problems, and you can proactively limit workloads to prevent performance problems. Granular QoS controls help maintain performance levels for critical applications in highly shared environments.

This feature enables a dataset to be fed to the DGX for training or inference at a reliable and sustained pace. It can also help to estimate the optimal ratio of compute to storage.

6.4 NetApp Volume Encryption

ONTAP offers native volume-level encryption with both onboard and external key management support. This feature ensures that sensitive data is accessible only to authorized users.

6.5 Data Protection

ONTAP provides built-in data protection capabilities with common management across all platforms. The two most common features are NetApp Snapshot™ copies and NetApp FlexClone® volumes.

Snapshot

A Snapshot copy is a read-only image of a FlexVol volume, a FlexGroup volume, or an aggregate that captures the state of the file system at a point in time. ONTAP maintains a configurable Snapshot copy schedule that creates and deletes Snapshot copies automatically for each volume. You can also create and delete Snapshot copies manually.

You can use Snapshot technology while applications are running and create copies in less than a second, regardless of volume size or level of activity on your NetApp system.

You can create up to 1,023 Snapshot copies per volume instantly as online backups for user-driven recovery.

NetApp Snapshot software is the foundation for NetApp SnapManager®, SnapMirror®, SnapRestore®, and SnapVault® software. Together they create unified, all-risks protection for your valuable data.

FlexClone

A FlexClone volume is a point-in-time, writable copy of the parent FlexVol volume or FlexGroup volume. FlexClone volumes can be managed similarly to regular FlexVol volumes, with a few important differences. For instance, the changes made to the parent FlexVol volume after the FlexClone volume is created are not reflected in the FlexClone volume.

These features protect the data from drive failures and provide security for confidential data. By taking backups of the datasets, it is possible to recover the ones that were deleted. Important datasets like medical data can be stored permanently in SnapVault.

6.6 NAS

NetApp network-attached storage (NAS) solutions simplify data management and help you keep pace with growth while optimizing costs. Our NAS solutions give you nondisruptive operations, proven efficiency, and seamless scalability in a unified architecture.

6.7 SAN

Storage area networks (SANs) are the most common storage networking architecture used by enterprises for business-critical applications that need to deliver high throughput and low latency. A SAN is block-based storage, leveraging a high-speed architecture that connects servers to their logical disk units (LUNs). A LUN is a range of blocks provisioned from a pool of shared storage and presented to the server as a logical disk. The server partitions and formats those blocks—typically with a file system—so that it can store data on the LUN just as it would on local disk storage.

Depending on the requirements, it is up to the user to choose the storage type.

6.8 Storage Efficiency

ONTAP offers a wide range of storage efficiency technologies in addition to Snapshot copies. Key technologies include thin provisioning; inline and offline deduplication; inline and offline compression; and

FlexClone volumes, files, and LUNs. Like Snapshot copies, these technologies are built on the ONTAP WAFL® (Write Anywhere File Layout) system.

Thin Provisioning

A thin-provisioned volume or LUN is one for which storage is not reserved in advance. Instead, storage is allocated dynamically, as it is needed. Free space is released back to the storage system when data in the volume or LUN is deleted.

Deduplication

Deduplication reduces the amount of physical storage required for a volume (or all the volumes in an AFF aggregate) by discarding duplicate blocks and replacing them with references to a single shared block. Reads of deduplicated data typically incur no performance charge. Writes incur a negligible charge, except on overloaded nodes.

Compression

Compression reduces the amount of physical storage required for a volume by combining data blocks in compression groups, each of which is stored as a single block. Reads of compressed data are faster than in traditional compression methods because ONTAP decompresses only the compression groups that contain the requested data, not an entire file or LUN.

Data Compaction

Data compaction reduces wasted space inside storage blocks, and deduplication significantly increases effective capacity.

6.9 ONTAP FabricPool

FabricPool is a hybrid storage solution that uses an all-flash (all-SSD) aggregate as the performance tier and an object store as the external capacity tier. Data in FabricPool is stored in a tier based on whether or not it is frequently accessed. Using FabricPool helps you reduce storage cost without compromising performance, efficiency, or protection.

This feature provides automatic tiering of cold data to public and private cloud storage options, including Amazon Web Services (AWS), Microsoft Azure, and NetApp StorageGRID® object-based storage.

Some of the datasets, trained models, or inferred results that are not used frequently but are important enough to be retained, rather than deleted, can be moved to the cloud or to StorageGRID, where it is cheaper to store the data.

6.10 NetApp Trident

Trident enables microservices and containerized applications to leverage enterprise-class storage services (such as QoS, storage efficiencies, and cloning) to meet an application's persistent storage demands. The Kubernetes volumes are static. Dynamic provisioning of Kubernetes volumes can be performed with the help of NetApp Trident, depending on an application's requirements. It can dynamically provision storage from the following sources:

- NetApp ONTAP data management software (NetApp AFF, FAS, ONTAP Select, and Cloud Volumes ONTAP)
- NetApp Element® software (NetApp HCI and SolidFire®)
- NetApp SANtricity® software (NetApp E-Series and EF-Series)

Trident uses the StorageClass object that was introduced in Kubernetes 1.4 to dynamically provision Persistent Volumes (PVs) when a Persistent Volume Claim (PVC) object is created. A storage class

allows administrators to describe the classes of storage that they offer. A storage class might map to different QoS levels, backup policies, or other storage characteristics. For more information, see the Trident documentation.

This feature is extremely useful because it eliminates the need for a storage administrator to allocate more storage space. When experiments are run, the results accumulate over time and a larger amount of storage space is consumed. A piece of code written in the application itself can automatically allocate more storage space to store the results. This point is also valid for datasets or raw data that comes from an edge device. That is, the size of the data lake expands or shrinks if data is deleted or moved to the cloud.

The following example illustrates automatically importing an ONTAP volume and resizing a NetApp NFS volume, without the involvement of a storage administrator.

The `ontap-nas` driver creates an ONTAP FlexVol volume for each volume. Trident uses NFS export policies to control access to the volumes that it provisions. It uses the default export policy, unless a different export policy name is specified in the configuration.

Importing an ONTAP Volume

Trident 19.04 and later allows you to import an existing storage volume into Kubernetes with the `ontap-nas`, `ontap-nas-flexgroup`, and `aws-cvs` drivers.

Here are some use cases for importing a volume into Trident:

- Containerizing an application and reusing its existing dataset
- Using a clone of a dataset for an ephemeral application

The `tridentctl` client is used to import an existing storage volume or a clone of a storage volume. Trident imports the volume by persisting volume metadata and creating the PVC and PV.

```
$ tridentctl import volume <backendName> <volumeName> -f <path-to-pvc-file>
```

```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: my_claim
  namespace: my_namespace
spec:
  accessModes:
    - ReadWriteOnce
  storageClassName: my_storage_class
```

Resizing a NetApp NFS Volume

Data processing and preprocessing pods use a MongoDB database, which might need to be resized to store learning based on multiple training sets.

Trident supports volume resize for NFS PVs. Specifically, PVs provisioned on `ontap-nas`, `ontap-nas-economy`, `ontap-nas-flexgroup`, and `aws-cvs` back ends can be expanded.

To resize an NFS PV, the administrator first needs to configure the storage class to allow volume expansion by setting the `allowVolumeExpansion` field to `true`:


```

$ cat storageclass-ontapnas.yaml
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: ontapnas
provisioner: netapp.io/trident
parameters:
  backendType: ontap-nas
allowVolumeExpansion: true

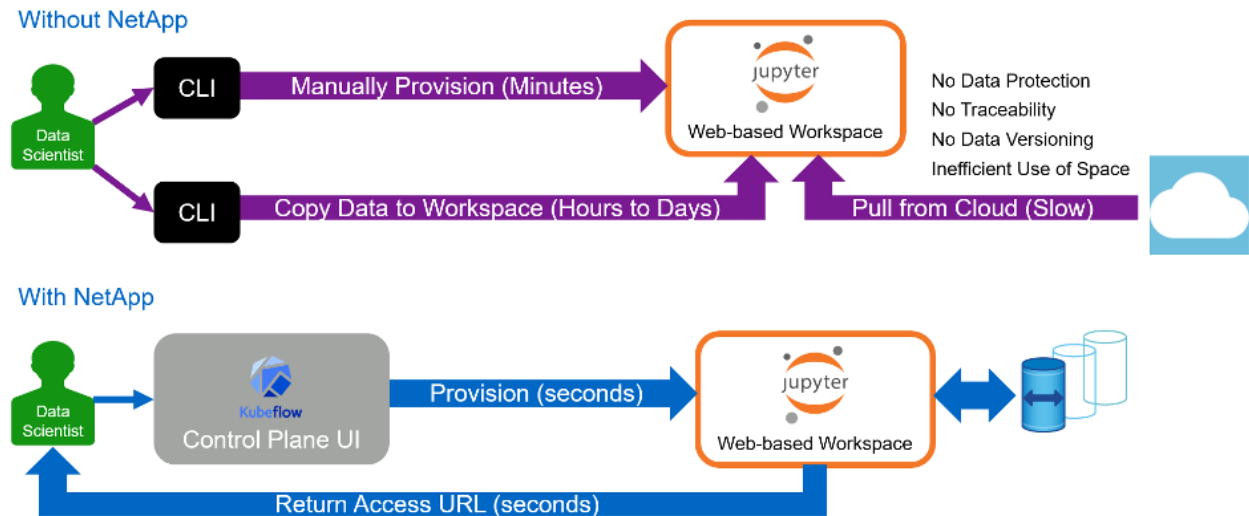
```

Kubeflow

Kubeflow is a component of the NetApp AI Control Plane that enables simplicity of deployment for AI workloads. As a Kubernetes-native framework, it provides a standard and open platform for deploying AI, ML, and DL workloads. It abstracts away the intricacies of Kubernetes, allowing data scientists and developers to focus on what they know best – AI, ML, and DL. With Kubeflow, data scientists no longer need to be Kubernetes administrators; rather, they can define end-to-end workflows by using a simple Python SDK. They don't need to know how to define Kubernetes deployments in YAML or execute kubectl commands.

Additionally, Jupyter Notebooks are included with Kubeflow out of the box. A team lead or administrator can provision and destroy Jupyter Notebook servers for data scientists and developers on demand. When Kubeflow is deployed as part of the NetApp AI Control Plane solution, data volumes, potentially containing petabytes of data, can be presented as simple folders within a Jupyter workspace. Data scientists have instant access to all their data from within a familiar interface (Figure 8).

Figure 8) Kubeflow.



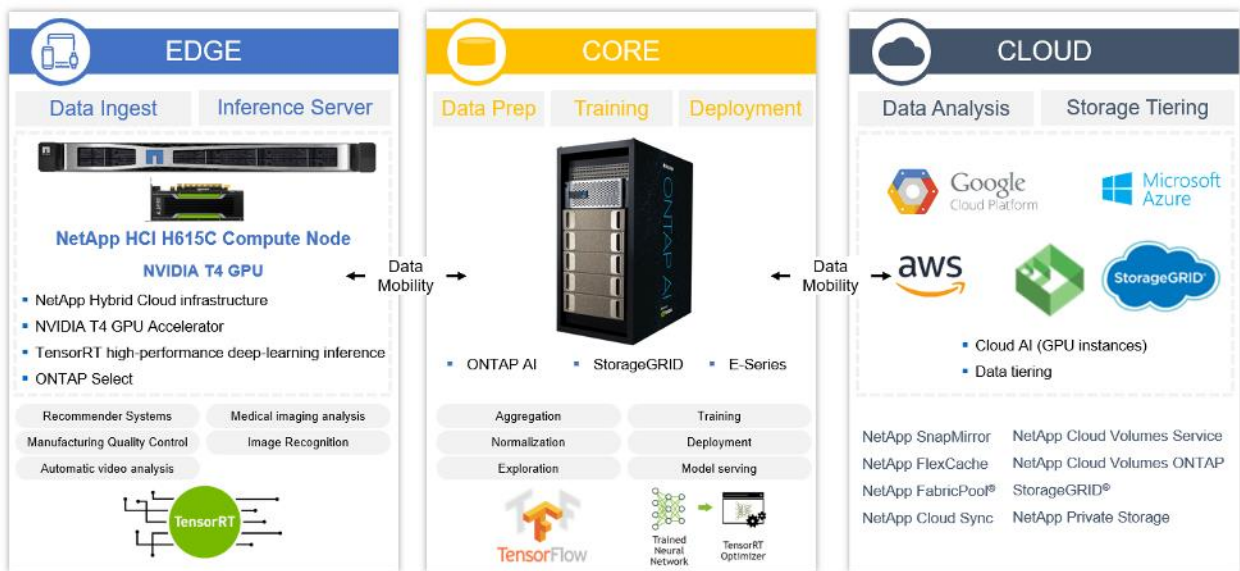
6.11 Edge to Core to Cloud

NetApp HCI is positioned in the NetApp AI portfolio as an inference server for the edge. The NetApp HCI 615c1 compute node provides three NVIDIA T4 GPUs for supporting inference applications. With the NetApp AI portfolio, you can create a deep learning trained model on NetApp ONTAP AI and then access the model for inferencing with FlexCache® in ONTAP Select on NetApp HCI without having to transport the model.

NetApp HCI offers the following competitive differentiation and advantages for AI Inference:

- The NetApp AI portfolio can support the entire spectrum of AI use cases from edge to core to cloud, including ONTAP AI for training and inferencing, Cloud Volumes Service and Azure NetApp Files for training in the cloud, and inferencing on the edge with NetApp HCI.
- A data fabric powered by NetApp allows data to be replicated from core to edge to cloud data centers to move the data closer to where the application needs it.
- With NetApp FlexCache, AI deep learning models trained on NetApp ONTAP AI can be accessed from NetApp HCI without having to export the model.
- NetApp HCI can host inference servers on the same infrastructure concurrently with multiple workloads, either virtual machine or container based, without performance degradation.
- NetApp HCI is NVIDIA NGC2-ready certified for NVIDIA GPU-Accelerated containers.
- A disaggregated architecture allows independent scaling of compute and storage and eliminates virtualization licensing costs and performance tax on independent NetApp HCI storage nodes.
- NetApp Element storage provides QoS per storage volume and allows guaranteed storage performance for workloads on NetApp HCI, preventing adjacent workloads from negatively affecting inferencing performance (Figure 9).

Figure 9) NetApp solution for the deep learning pipeline.



Management of datasets and the inference results can be automated using the NetApp tools just described, which makes it inexpensive, simple, convenient, and time saving when compared to manual data management by a storage administrator.

7 Benefits of Deploying Paras in a NetApp Environment

7.1 Technical and Operational Benefits

ONTAP data management software is a convenient tool to manage a dataset, thanks to features like FlexVol, FlexGroup, Snapshot, SnapMirror, SnapVault, and more.

The benefit of using on-premises storage instead of the cloud is that the data is more secure, traceable, and accountable. You can eliminate reliance on a cloud service provider, removing the risk with associated with bugs in data management software and reducing the possibility that data is damaged or lost.

There are no latency or throughput issues because dependency on the internet is eliminated, unless cold data is moved to the cloud.

7.2 Business and Financial Benefits

The benefits of deploying Paras in a NetApp storage-compute environment are significant. Data management can be automated, eliminating the need for storage administrators and data scientists, saving time and costs. With highly automated data management by ONTAP running on state-of-the-art GPUs and flash storage, the time needed to train a model and perform inference with the results is minimal.

Typically, it takes 2 to 3 months to build a model from a data set. Using Paras on typical infrastructure, the same task can be performed in 2 to 3 weeks, which is, approximately, a 5x improvement. Using the ONTAP solution, Paras can build models 3x faster still. Therefore, the overall improvement in speed is up to 15x, which allows models to be built in a week's time using both ONTAP AI and Paras.

The operational cost for 2 months for a data science team with two people is approximately \$35k (assuming \$50 per hour). A Paras and ONTAP AI solution can be deployed in a week with an estimated cost of around \$7000, assuming the use of ONTAP and other NetApp products. This solution saves \$28k, or ~80% of the total cost of a traditional system. Notably, the total cost of the solution can vary, depending which NetApp products are necessary in your environment.

8 Conclusion

The Paras and ONTAP solution unlocks an end-to-end AI value chain for users, resulting in time and cost efficiencies. You can build a deep learning solution in a couple of days rather than in a few months. and you can also achieve time savings of 15x for model development. In total, you can save more than 80% over the cost of standard AI development and execution.

In a world where data is growing at exponential rate, this Paras and ONTAP solution can provide cost and time efficiencies to keep your business ahead of the competition.

Where to Find Additional Information

To learn more about the information that is described in this document, review the following documents and/or websites:

- NetApp FlexVol technology
<https://library.netapp.com/ecmdocs/ECMP1368017/html/GUID-4DF6A167-6C98-4E48-8F5C-41E73A506139.html>
- NetApp FlexGroup technology
<https://www.netapp.com/us/media/tr-4557.pdf>
- Quality of Service
https://kb.netapp.com/app/answers/answer_view/a_id/1002428/~/faq%3A-storage-quality-of-service-%28qos%29
- SAN
<https://www.netapp.com/us/info/what-is-storage-area-network.aspx>
- Data Protection
<https://library.netapp.com/ecmdocs/ECMP1635994/html/GUID-0F851C3C-78F2-4C29-AB7D-30D4849850CB.html>
- NetApp Trident
<https://www.netapp.com/us/media/ds-netapp-project-trident.pdf>
- Edge to Core to Cloud
<https://www.netapp.com/us/media/wp-7271.pdf>

Acknowledgements

We would like to thank the following people for their invaluable contributions to the creation of this document:

Sachhin Sreedharan, NetApp

Srinivas Venkat, NetApp

Varun Kondagadapa, Curl

Ashwin Ravishankar, NetApp

Swapnik Jakkampudi, Curl

Version History

Version	Date	Document Version History
Version 1.0	May 2020	Initial version written by Shivaram K R.

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

Copyright © 2020 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

Data contained herein pertains to a commercial item (as defined in FAR 2.101) and is proprietary to NetApp, Inc. The U.S. Government has a non-exclusive, non-transferrable, non-sublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b).

Trademark Information

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.