Understanding Azure Machine Learning Infrastructure

David Tucker

TECHNICAL ARCHITECT & CLOUD CONSULTANT

@_davidtucker_   www.davidtucker.net

Azure Machine Learning Solution Components

Workflow

Data Pipeline

Infrastructure

Overview

Reviewing compute approaches for machine learning

Understanding use cases for compute approaches

Deploying an inference web service with GPU support

Exporting data from Azure ML workspace

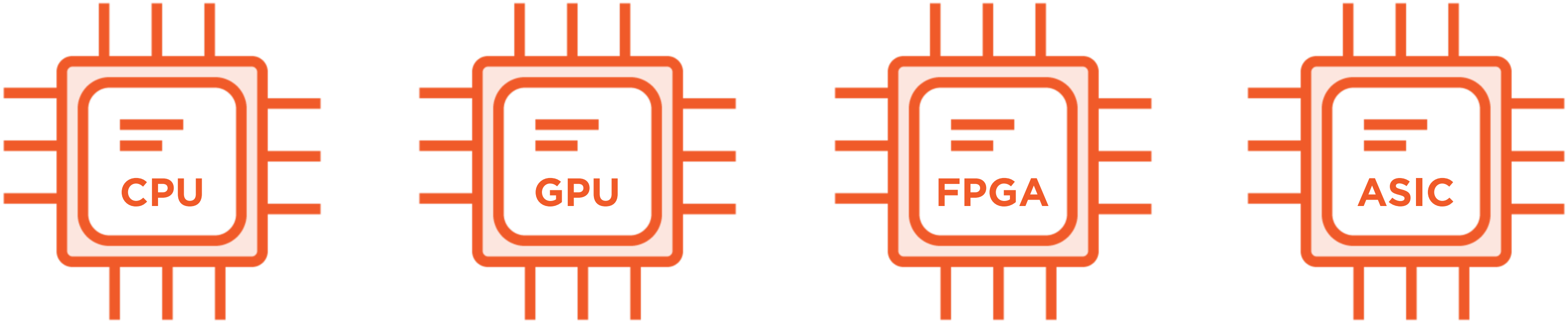
Decommission Azure ML infrastructure and resources

Infrastructure for Machine Learning

Machine Learning Infrastructure

Determining the right infrastructure approach for machine learning requires an understanding of the phase, cost constraints, and desired speed for the activity being performed.

Compute Approaches for Machine Learning

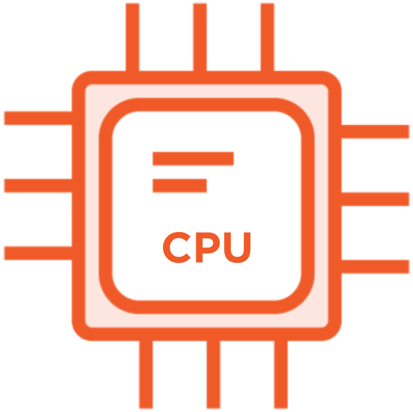


CPU
Compute

GPU
Compute

Reconfigurable
FPGA

Custom
ASIC



Perform general purpose compute workloads

Most cost efficient approach for workloads

Will generally perform slower than other options especially for training

Can be leveraged for inference and training

Supported by Azure ML Compute

CPU VM Series

DC-Series

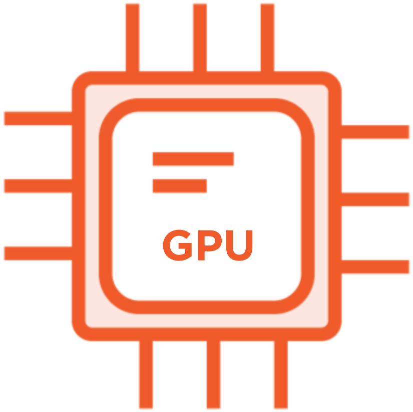
Uses secure enclave to protect sensitive data for processing

F-Series

Optimized for a high compute to memory ratio

D-Series

General purpose compute series



Perform parallel computing tasks which are ideal for machine learning

More expensive than CPU-based compute

Does not enable specific customizations for your model

Can be leveraged for inference and training

Supported by Azure ML Compute

GPU N-Series Family

NC-Series

HPC and ML workloads using Tesla V100 GPU

ND-Series

Training and inference using Tesla P40 or V100 GPU’s

NV-Series

Remote visualization workloads using the Tesla M60 GPU

FPGA

Stands for field programmable gate array

Reconfigurable integrated circuit based on model that is being deployed

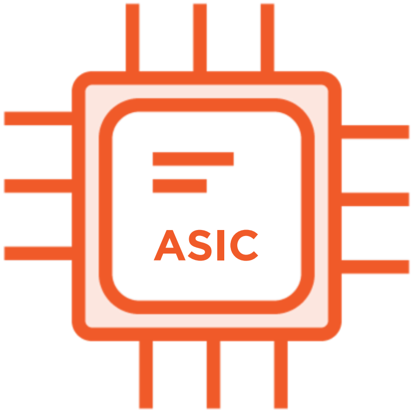
Pricing is similar to GPU approach

Primarily leveraged for inference

Supported by Azure ML Compute

> “The **PBS Family** of Azure VMs contains Intel Arria 10 FPGAs. The PB6 VM has six vCPUs and one FPGA, and it will automatically be provisioned by Azure ML as part of deploying a model to an FPGA. It is only used with Azure ML, and it cannot run arbitrary bitstreams.”

Azure Documentation



Stands for application-specific integrated circuit

Custom chips designed for specific machine learning tasks

Most efficient approach for machine learning

Does not enable specific customizations for your model

Can be leveraged for inference and training

Not supported by Azure ML Compute

Deploying a Model with GPU Support

Demo

Upload exercise files for the module

Creating compute infrastructure for GPU support

Configuring deployment for GPU support

Deploying model to Azure Kubernetes Service (AKS)

Validate deployed endpoint

Decommissioning our Resources

Demo

Exporting needed data from Azure ML workspace

Delete Azure ML workspace

Verifying the deletion of resources

Summary

Summary

Reviewed compute approaches for machine learning

Understood use cases for compute approaches

Deployed an inference web service with GPU support

Exported data from Azure ML workspace

Decommissioned Azure ML infrastructure and resources