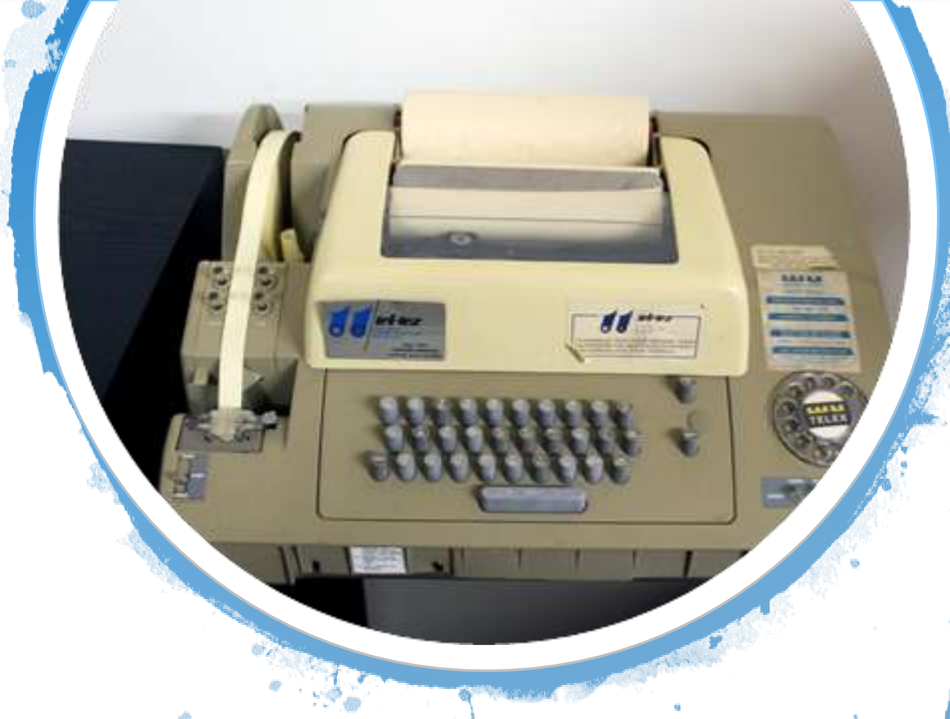
Microsoft

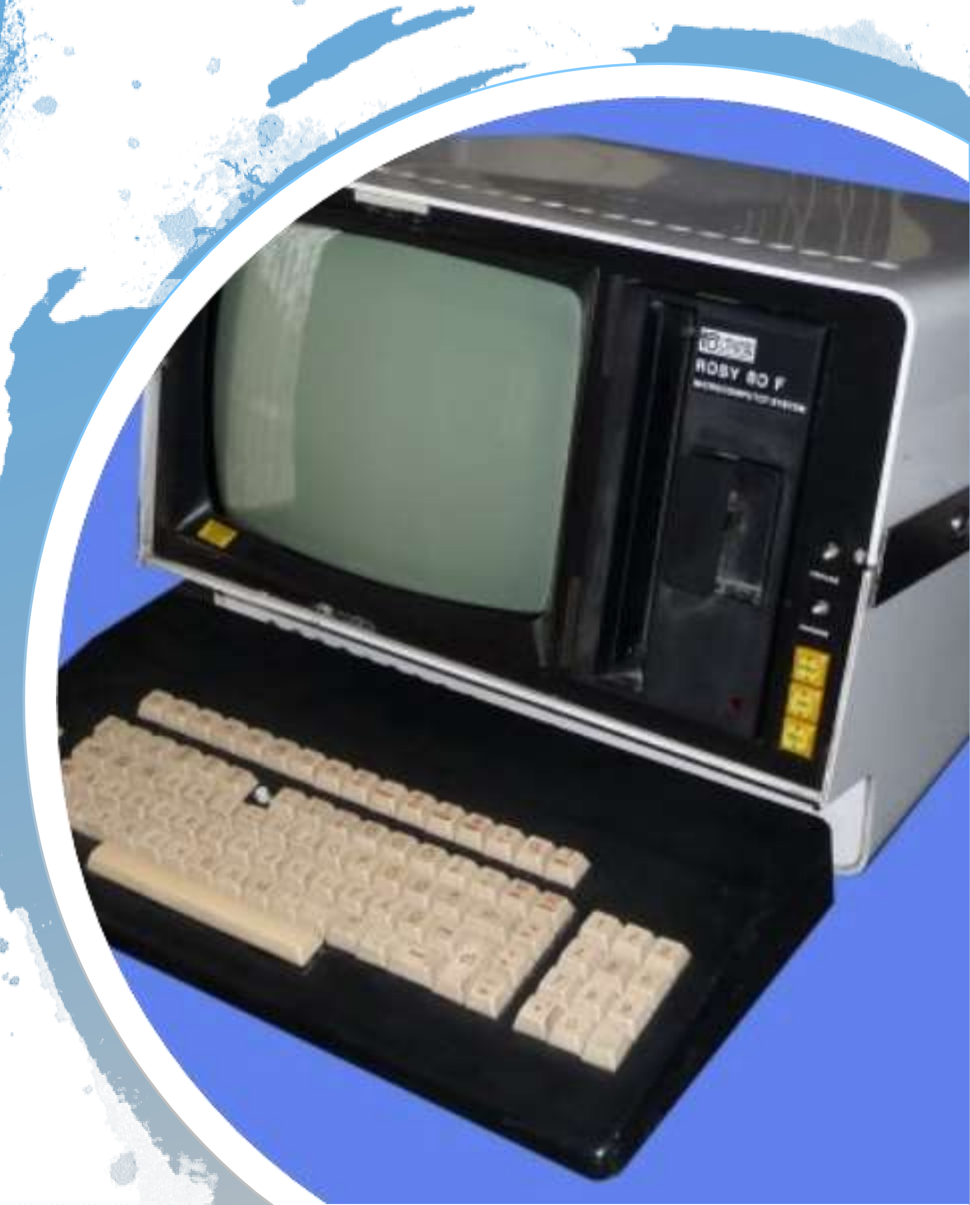
Supercomputing on demand with GPU

Gabor Varga

National Technology Officer



By Jamie - Flickr: Telex machine TTY, CC BY 2.0,
https://commons.wikimedia.org/w/index.php?curid=19282428



Can you recognise these machines?

Hint: neither is a computer

The PC as convergence point in the 90’s

- PCs became the unified platform three decades ago
- It was much easier to develop an application on a PC rather than design electronics from scratch
- Even industrial process control and ATMs run on PCs
- The Cloud offers a similar convergence point today for different chip architectures

BrayLockBoy - Own work - CC BY-SA 4.0 through Wikipedia

Trivia:
What do the Cloud and the youngest British royal baby Archie have in common?

Answer: Their names skyrocketed in popularity very soon after their birth. Just about anything in computing was named or renamed cloud in the past decade.

Microsoft Azure

Trust
Protect your business

Open and Hybrid
Build freely, deploy consistently in the cloud & on-prem

App Innovation
Accelerate innovation with the cloud

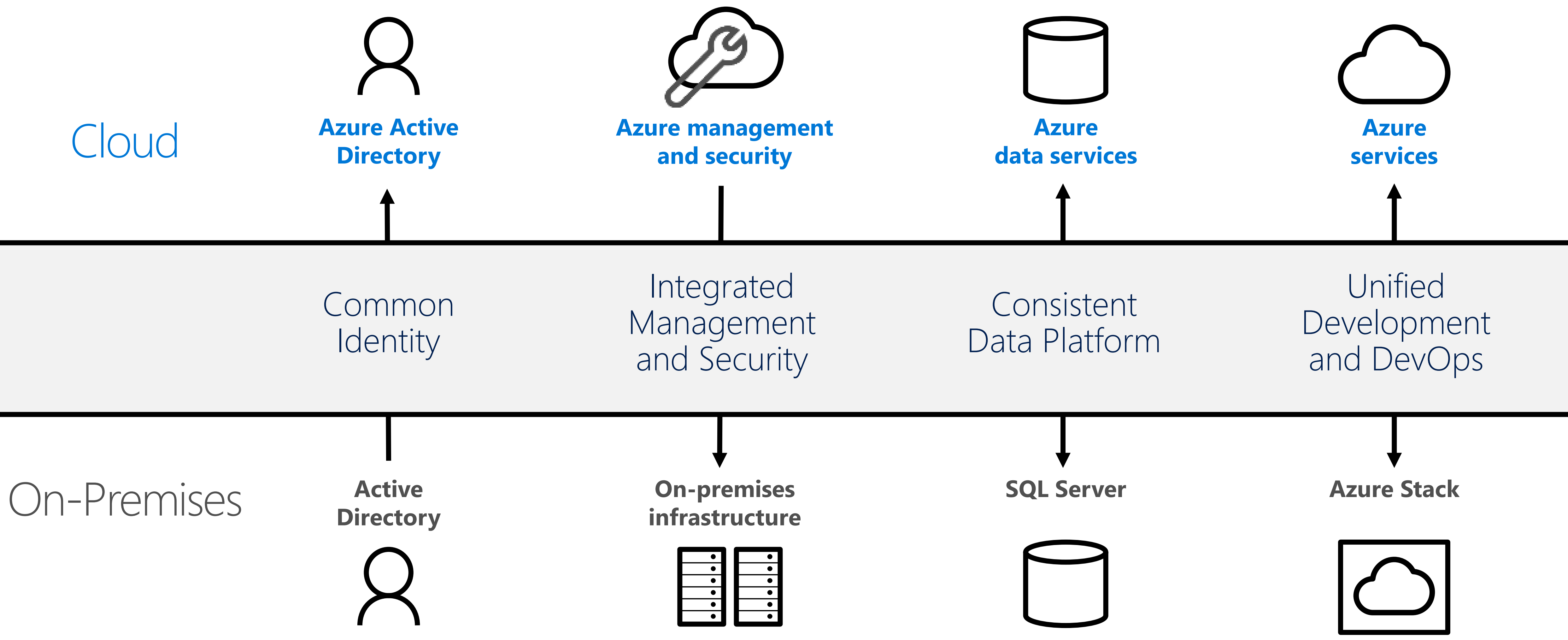
Data-Driven Intelligence
Power decisions & apps with insights

40

announced regions

Consistent identity, apps, data & management

Open and Hybrid



Cloud	**Azure Active Directory**	**Azure management and security**	**Azure data services**	**Azure services**
	Common Identity	Integrated Management and Security	Consistent Data Platform	Unified Development and DevOps
On-Premises	**Active Directory**	**On-premises infrastructure**	**SQL Server**	**Azure Stack**

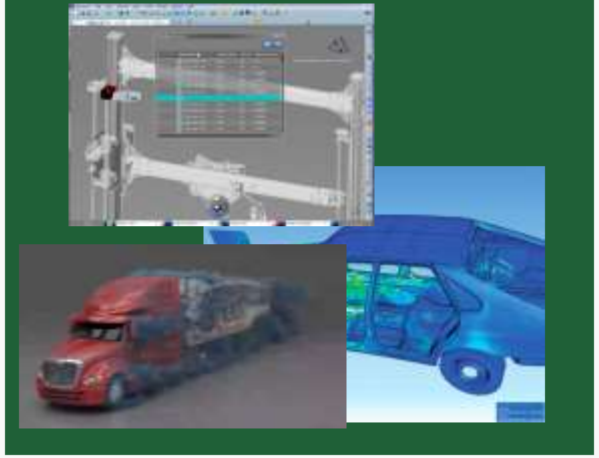
Microsoft Open Source Fun Facts

1. The Vice President of the Apache Foundation works in the Azure Compute team
2. Microsoft has the most contributors on GitHub
3. Joined Linux Foundation
4. The Windows team has a Docker committer
5. The co-creator of Kubernetes is the development manager of Azure Resource Manager and Azure Container Service

Azure GPUs

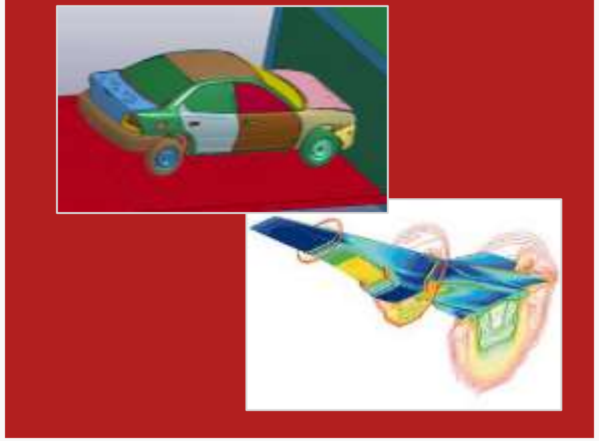
Broad Range of GPU Scenarios

Visualization



Rendering

HPC/Simulation



Deep-Learning/AI

Media, Entertainment & Gaming

Healthcare & Research

Financial Services

Oil & Gas and Sciences

Manufacturing

Automotive

Aerospace

Retail

In Preview

http://aka.ms/AzureNDv2GPUs

ND v2 – Volta Generation GPU Compute

- Excellent for accelerating machine learning and HPC workloads
- Volta SXM GPU instances – 8X NVIDIA V100 GPUs interconnected with NVLink mesh
- Tensor Core technology to deliver over 100 TFLOPS of deep learning performance
- Skylake based processor with premium storage support (SSD backed)
- Specs:
    - 640 NVIDIA Tensor Core
    - FP64 - 7.8 TFLOPS of double precision floating point performance
    - FP32 – 15.7 TFLOPS of single precision performance
    - GPU Memory 16 GB
    - 300 GB/s GPU interconnect through NVLink

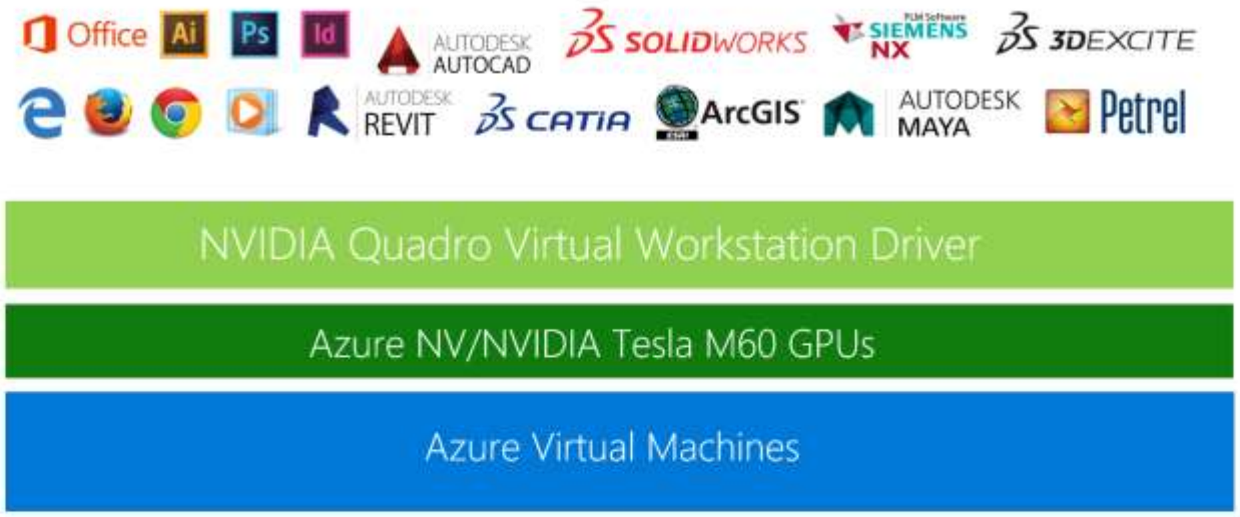
	ND40s_v3
Cores	40 cores
GPU	8 x V100 SXM
Memory	768 GB
Local Disk	~1.3 TB SSD
Network	Azure Network + NVLink GPU interconnect

In Preview

http://aka.ms/AzureNVv2GPUs

NV v2 – Updated GPU Visualization Platform

- Get faster results for the your graphic intensive 2D and 3D applications with visualization optimized GPU instances featuring NVIDIA Tesla M60 GPUs
- Broadwell based CPU processor with doubled memory from previous generation (up to 448 GB)
- Premium storage support (SSD backed)
- Grid license included with each GPU instance
- Specs:
  - 2048 NVIDIA CUDA cores per GPU
  - 36 H.264 1080p30 streams
  - GPU Memory 8 GB/GPU



NVIDIA Quadro Virtual Workstation Driver

Azure NV/NVIDIA Tesla M60 GPUs

Azure Virtual Machines

	NV6s_v2	NV12s_v2	NV24s_v2
Cores	6	12	24
GPU	1 x M60	2 x M60	4 x M60
Memory	112 GB	224 GB	448 GB
Local Disk	~700 GB SSD	~1.4 TB SSD	~3 TB SSD
Network	Azure Network	Azure Network	Azure Network
GRID Licenses	1	2	4

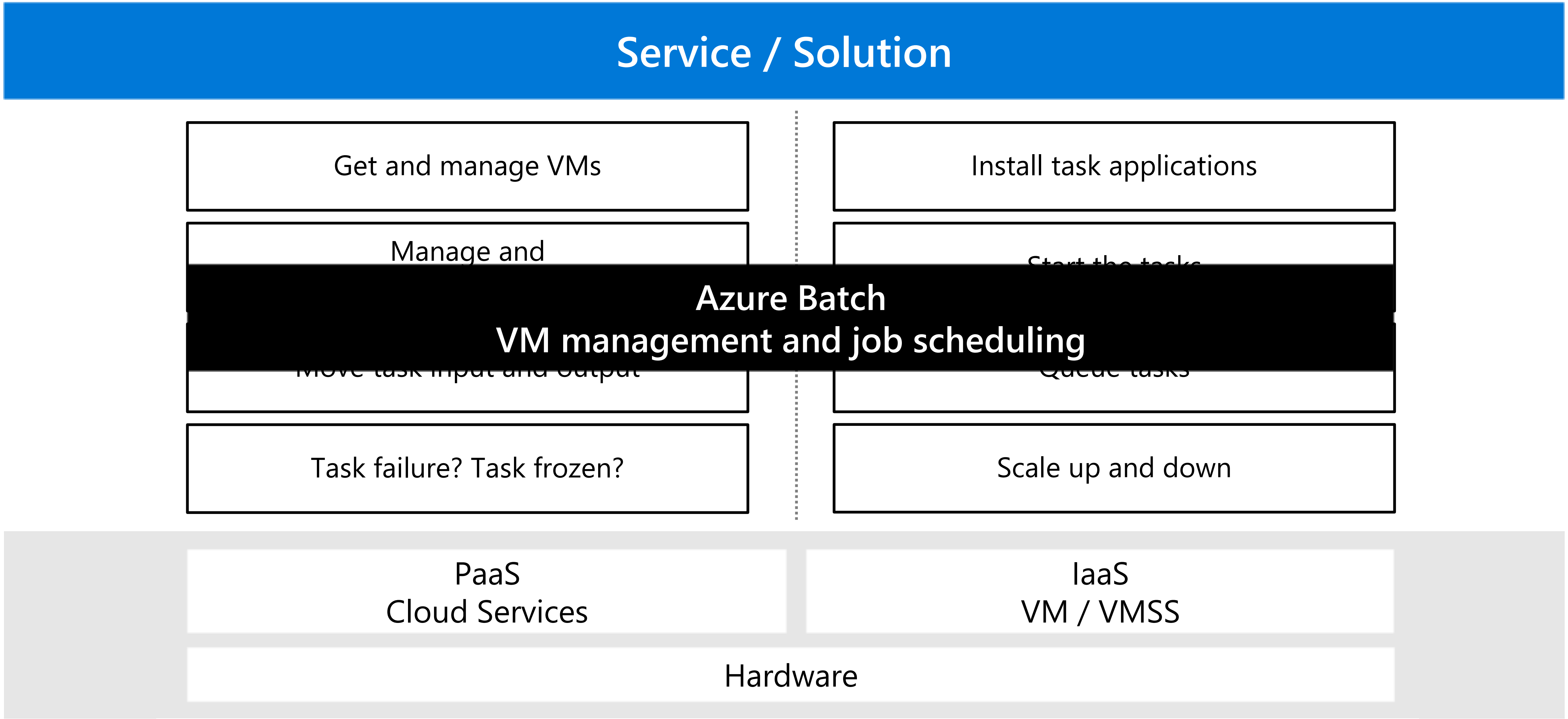
Full Lineup of GPU Families

GPU Accelerated Compute Family			
	NC	NC v2	NC v3
Cores	6, 12, 24	6, 12, 24	6, 12, 24
GPU	1, 2, or 4 K80 GPU	1, 2, or 4 P100 GPU	1, 2, or 4 V100 GPU
Memory	56/112/224 GB	112/224/448 GB	112/224/448 GB
Local Disk	~380/~680/~1.5 TB SSD	~700/~1.4/~3 TB SSD	~700/~1.4/~3 TB SSD
Network	Azure Network + InfiniBand (largest size only)		

	GPU Accelerated Deep Learning Family		GPU Visualization Family	
	ND v1	ND v2	NV v1	Nv v2
Cores	6, 12, 24	40	6, 12, 24	6, 12, 24
GPU	1, 2, or 4 P40 GPU	8 V100 SXM GPU	1, 2, or 4 M60 GPU	1, 2, or 4 M60 GPU
Memory	112/224/448 GB	768 GB	56/112/224 GB	112/224/448 GB
Local Disk	~700/~1.4/~3 TB SSD	~1.3 TB SSD	~380/~680/~1.5 TB SSD	~700/~1.4/~3 TB SSD
Network	Azure Network + InfiniBand (largest size only)	Azure Network + NVLink GPU interconnect	Azure Network	Azure Network

Azure Batch

Azure Batch



Service / Solution

Get and manage VMs

Install task applications

Manage and

Start the tasks

Azure Batch
VM management and job scheduling

Move task input and output

Queue tasks

Task failure? Task frozen?

Scale up and down

PaaS
Cloud Services

IaaS
VM / VMSS

Hardware

Azure Batch

Batch pools

Configure and create VMs to cater for any scale: tens to thousands

Automatically scale the number of VMs to maximize utilization

Choose the VM size most suited to your application

Batch jobs and tasks

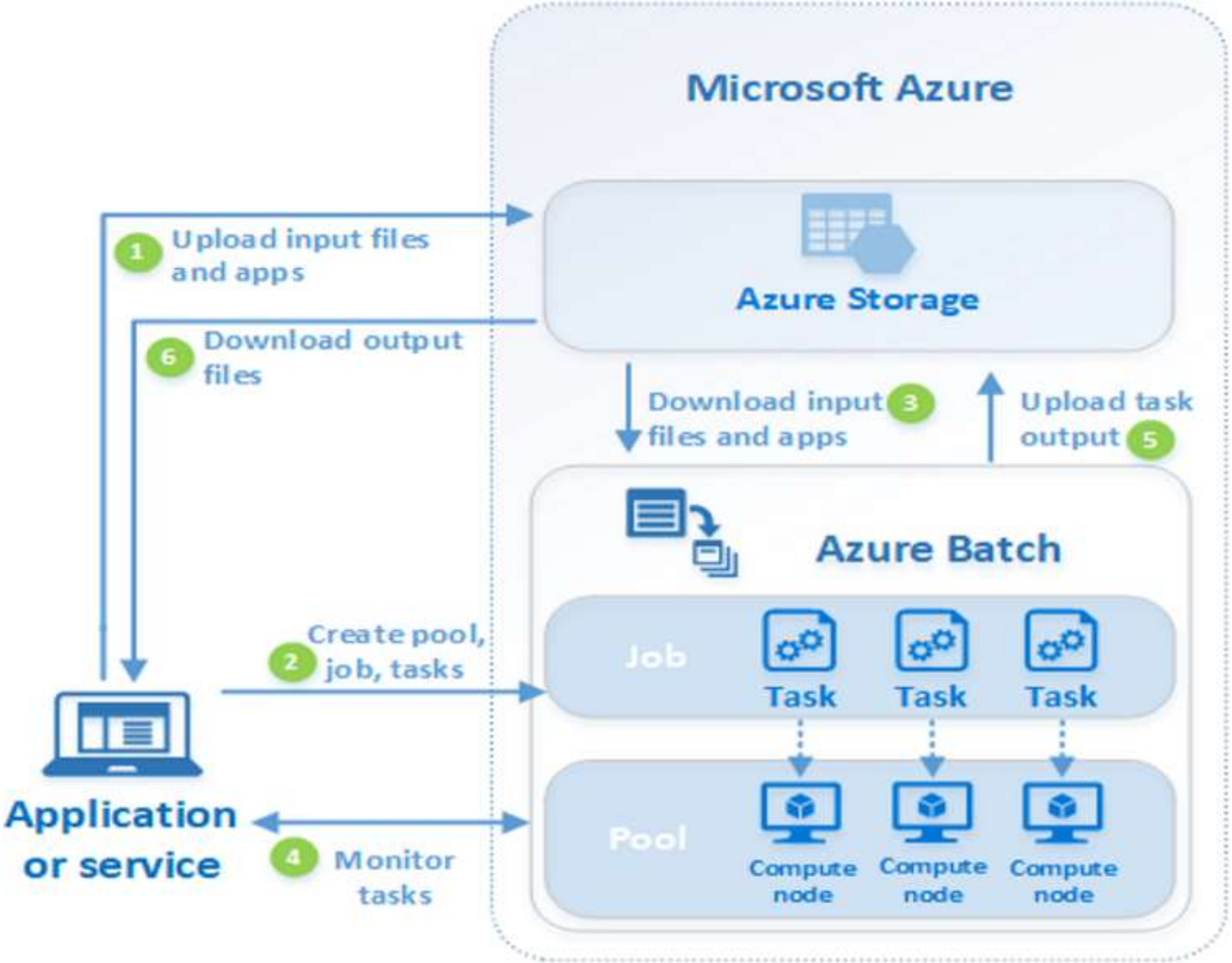
Task is a unit of execution; task = application command line (EXE, BAT, CMD, PS1, etc.)

Jobs are created and tasks are submitted to a pool. Next, tasks are queued and assigned to VMs

Any application, any execution time; run applications unchanged

Automatic detection and retry of frozen or failing tasks

How Azure Batch works – intrinsically parallel example



Azure Batch Supported development technologies

Command-Line

Azure CLI

Azure PowerShell

Languages

.NET

Java

Node.js

Python

REST

Batch Service REST API

Batch Management REST API

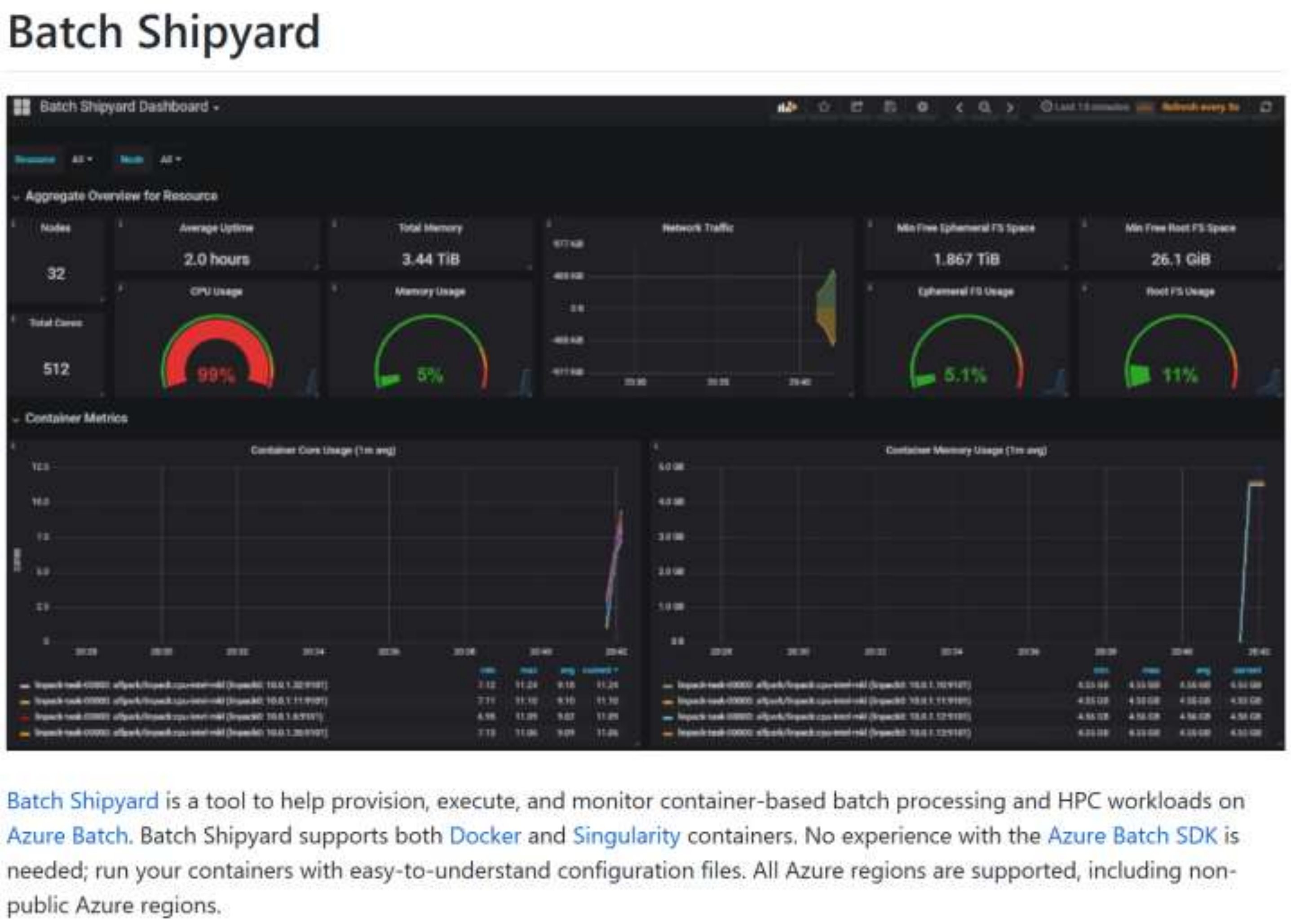
Azure Batch Step-by-Step Tutorials and Training

- **Learn how to run compute-intensive workloads on Batch.**
  - Parallel file processing with .NET SDK
  - Parallel file processing with Python SDK
  - Scene rendering with Arnold
  - Parallel R simulation
  - Tutorial: Trigger a Batch job using Azure Functions

- **Free Pluralsight Video Training**
  - Microsoft Azure Batch, Getting Started

Containers in Azure Batch

- Batch Shipyard is a tool to help provision, execute, and monitor container-based batch processing and HPC workloads on Azure Batch. Batch Shipyard supports both Docker and Singularity containers.
- https://github.com/Azure/batch-shipyard
- Container Runtime and Image Management
- Data Management and Shared File Systems
- Monitoring
- Open Source Scheduler Integration



Batch Shipyard

Batch Shipyard is a tool to help provision, execute, and monitor container-based batch processing and HPC workloads on Azure Batch. Batch Shipyard supports both Docker and Singularity containers. No experience with the Azure Batch SDK is needed; run your containers with easy-to-understand configuration files. All Azure regions are supported, including non-public Azure regions.

Machine Learning on Azure

Domain specific pretrained models

To simplify solution development

Vision

Speech

Language

Web search

Decision

Familiar data science tools

To simplify model development

Visual Studio Code

Azure Notebooks

Jupyter

Command line

Popular frameworks

To build advanced deep learning solutions

PyTorch

TensorFlow

Scikit-Learn

ONNX

Productive services

To empower data science and development teams

Azure Machine Learning

Azure Databricks

Machine Learning VMs

Powerful infrastructure

To accelerate deep learning

CPU

GPU

FPGA

From the Intelligent Cloud to the Intelligent Edge

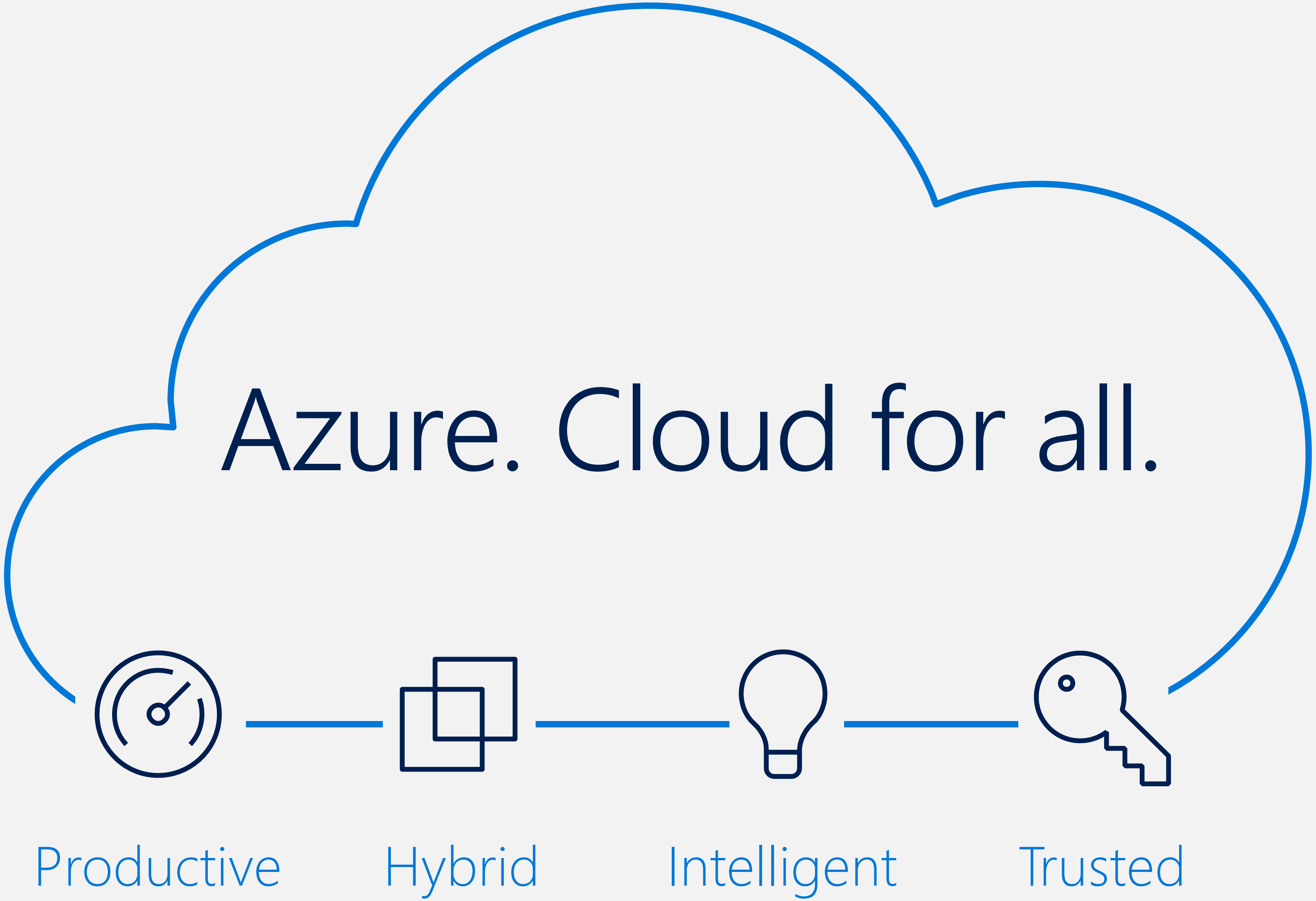
Hardware accelerated models

General availability

Deploy with
Azure Machine Learning

FPGAs

Specialized, hardware-accelerated,
deep learning



Azure. Cloud for all.

Productive

Hybrid

Intelligent

Trusted

Next Steps:

Learn more about Azure and create your free Azure account
https://azure.microsoft.com

Explore Azure Batch
https://docs.microsoft.com/en-us/azure/batch/

Microsoft