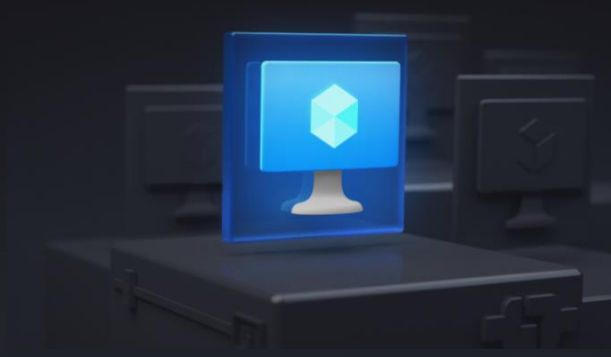
Microsoft Azure



NCas_T4_V3

Azure VM for Machine Learning, Deep Learning, and Data Analytics

Featuring the **NVIDIA® T4 GPU**

GA: 11.20	SKUs: 4	Regions: 6	Local SSD: 2.8TB

A new breed of GPU

The NCasT4_v3-series virtual machines are powered by Nvidia Tesla T4 GPUs and AMD EPYC 7V12 (Rome) CPUs. The VMs feature up to 4 NVIDIA T4 GPUs with 16 GB of memory each, up to 64 non-multithreaded AMD EPYC 7V12 (Rome) processor cores, and 440 GiB of system memory. These virtual machines are ideal for deploying AI services - such as deep learning training & inference, machine learning, and data analytics - using NVIDIA's GRID driver and virtual GPU technology. Featuring multi-precision Turing Tensor Cores and new RT Cores combined with accelerated containerized software stacks from NGC, the Tesla T4 GPU packs a serious punch. This optimizes the NCas_T4 virtual machine for mainstream computing environments to deliver revolutionary performance at scale.

HPC+AI Workflows

The monumental challenges faced by society today require bold solutions. Data-intensive applications and interactive processes have skyrocketed both in scope and distribution, requiring greater overall system responsiveness that, often times, is out on the Edge. The ability to bring highly scalable processing to the point where the data is being generated has become pivotal to generating user engagement for conversational AI, object/identity recognition, and autonomous driving. By harnessing the computational power of the T4 GPU, the NCasT4_v3 virtual machine offers a powerful yet versatile tool for building new machine learning models in a cost-effective way.

NCasT4_v3 Use Cases

Machine Learning • Deep Learning • Data Analytics

NVIDIA Tesla T4 GPU

SPECIFICATIONS	
GPU Architecture	NVIDIA Turing
NVIDIA Turing Tensor Cores	320
NVIDIA CUDA® Cores	2,560
Single-Precision	8.1 TFLOPS
Mixed-Precision (FP16/FP32)	65 TFLOPS
INT8	130 TOPS
INT4	260 TOPS
GPU Memory	16 GB GDDR6 300 GB/sec
ECC	Yes
Interconnect Bandwidth	32 GB/sec
System Interface	x16 PCIe Gen3
Form Factor	Low-Profile PCIe
Thermal Solution	Passive
Compute APIs	CUDA, NVIDIA TensorRT™, ONNX

NVIDIA

NCas_T4_V3 Virtual Machine

Performance & Functional Specifications

SKU & Size	vCPU	Memory: GiB	Temp storage (SSD) GiB	GPU	GPU memory: GiB	Max data disks	Max NICs
Standard_NC4as_T4_v3	4	28	180	1	16	8	2
Standard_NC8as_T4_v3	8	56	360	1	16	16	4
Standard_NC16as_T4_v3	16	110	360	1	16	32	8
Standard_NC64as_T4_v3	64	440	2880	4	64	32	8

Networking:

- FE Eth: 50GB/s (Nic)
- Accelerated Networking: (Available GA release)
- BYOL for NVLINK support

Local Storage:

- 2.8T SSD

Supported Capabilities:

- Premium Storage
- Premium Storage caching

Images to get started:

- Cananonical Ubuntu 18.04 LTS Gen 2
- HPC+AI General Purpose Image (coming soon)

Supported Azure Regions:

- West US 2 (In Preview)
- South Central US – (Q4 CY20)
- West Europe
- Japan East
- Korea Central
- China

AMD EPYC 7V12 CPU ("Rome")

- Second iteration of AMD's EPYC server processor family
- Compatible with the existing workstation and server platforms
- Clock speeds up to 3.1GHz (Boost speeds up to 3.4GHz)
- Full support for 256-bit AVX2 instructions with two 256-bit FMA units per CPU core
- Up to 256MB L3 cache per CPU (up from 64MB)
- 15% increase in instructions completed per clock cycle for integer operations

AMD

Performance Benchmarks:

Scan this QR Code with your mobile device to view the performance benchmarks for the NCas_T4_V3 VM.

Additional Resources:

- https://www.nvidia.com/en-us/data-center/tesla-t4/
- https://docs.microsoft.com/en-us/azure/virtual-machines/nct4-v3-series
- https://github.com/matterport/Mask_RCNN
- https://interestingengineering.com/meet-bert-the-ai-system-that-can-finish-your-sentence

Microsoft Azure