1. Introduction

Microsoft Azure provides a broad set of GPU-accelerated virtual machines optimized for AI training, inference, graphics rendering, HPC, and data processing. GPU families on Azure include the NC, ND, NV, and NCads/NDas/Lsv3 variants, covering a wide range of NVIDIA GPUs such as A100, H100, L40S, A10, T4, and previous generations.

Azure's GPU portfolio is designed to meet different levels of performance, memory requirements, scalability, and cost, making it suitable for LLM training, fine-tuning, high-throughput inference, and visualization workloads.

2. GPU VM Families on Azure

2.1 NC Series – Compute-Optimized for AI Training

- Best for: Deep learning training, scientific computing.
- GPU models: NVIDIA A100 80GB, V100, K80.
- Examples:
    - Standard_NC24ads_A100_v4 – 1× A100 80GB.
    - Standard_NC96ads_A100_v4 – 4× A100 80GB (multi-GPU node).
- Use cases:
    - Transformer/LLM training (up to 70B parameters).
    - Computer vision training.
    - Reinforcement learning.

2.2 ND Series – Multi-GPU AI Training & Large Model Workloads

- Best for: Scalable multi-node distributed model training.
- GPU models: NVIDIA A100 80GB, H100 (ND H100 preview).
- Examples:
    - Standard_ND96asr_A100_v4 – 8 GPUs A100 80GB.
    - Standard_ND96isr_H100_v5 (preview) – 8 GPUs H100.
- Key features:
    - NVLink connection between GPUs.
    - High-bandwidth InfiniBand HDR.

2.3 NV Series – Visualization & Rendering

- Best for: Graphics rendering, visualization, remote desktops.
- GPU models: NVIDIA L40S, RTX A6000, M60, T4.
- Examples:
    - Standard_NVadsA10_v5 – 1× A10 GPU.
    - Standard_NV32as_v4 – 1× L40S GPU.
- Use cases:
    - 3D visualization.
    - CAD workloads.

    - Media encoding.

2.4 Other GPU-Accelerated Instances

- Lsv3 – High-performance local NVMe storage for data pipelines.
- HB/HC Series – Not GPU-based, but relevant for HPC workloads involving CPU-heavy simulations.

3. Azure GPU Pricing Overview

Azure GPU pricing varies by:

- Region
- GPU type
- VM family
- CPU/RAM configuration
- Networking capabilities
- Purchase model (Pay-As-You-Go, Reserved Instances, Spot)

Example Average Prices (Fake Data for Mock/RAG Use Only)

A100 80GB

- Standard_NC24ads_A100_v4: $3.40/hour
- Standard_ND96asr_A100_v4: $28.50/hour (8 GPUs)

L40S

- Standard_NV24as_v4: $2.10/hour

A10

- Standard_NVadsA10_v5: $1.40/hour

T4

- Standard_NC4as_T4_v3: $0.85/hour

4. Purchase Models & Cost Optimization

4.1 Pay-As-You-Go

- No commitment.
- Highest hourly cost.
- Good for POCs or infrequent training jobs.

4.2 Reserved Instances (1 or 3 years)

- Up to 65% cheaper than PAYG.
- Best for long-running production workloads.

4.3 Spot VMs

- Up to 90% cheaper.
- VM may be evicted at any moment.
- Suitable for:
    - Long-running but restart-tolerant jobs.
    - Ray/TensorFlow/PyTorch distributed training that can checkpoint.

4.4 Savings Plans

- Flexible commitment to spend amount per hour.
- Applies across multiple VM families.

5. Networking Considerations

5.1 High-Performance Networking

- ND and NC series support:
    - InfiniBand HDR networking
    - NCCL-optimized GPU communication
- Essential for:
    - Multi-GPU distributed training (Data Parallel and Tensor Parallel)
    - Multi-node clusters for LLMs >70B

5.2 Bandwidth Requirements for Large Models

- Model parallel training requires:
    - >= 200 Gbps GPU-to-GPU interconnect.
- When using A100/H100 with NVLink:
    - Inter-node latency reduced significantly.

6. Storage Options

6.1 Managed Disks

- Premium SSD recommended for high IOPS.

6.2 Azure Blob Storage

- Ideal for storing large datasets and checkpoints.
- Supports:
    - Multi-part uploads
    - High throughput
    - Lifecycle policies for reducing cost

6.3 Ephemeral OS Disks

- Faster start times.
- Good for ephemeral workloads.

6.4 Local NVMe Storage (Lsv3 Series)

- High bandwidth.
- Best for temporary data processing or vector DB storage.

7. Recommended Workload Profiles

7.1 Large Language Model Training

- Recommended GPUs: A100 80GB, H100
- Node types: ND A100 v4 or ND H100 v5
- Parallel methods:
    - Data Parallelism
    - Tensor Parallelism
    - Pipeline Parallelism
- Datasets:
    - Must be streamed from Blob or mounted via Azure Files.

7.2 LLM Fine-Tuning

- GPUs: A100 40GB/80GB, A10, L40S
- Batch size often limited by VRAM.
- Parameter-efficient tuning (LoRA/QLoRA) reduces cost.

7.3 High-Throughput Inference

- GPUs: L40S, A10, T4
- Recommended VM families:
    - NVadsA10_v5
    - NCas_T4_v3
- Best for:
    - Chatbots
    - Embedding generation
    - Vision inference

7.4 Graphics & Rendering

- GPUs: L40S, A6000, A10
- NV series ideal.

8. Region Comparison (Fake Data)

Region	Avg GPU Cost	Availability	Notes
East US	Medium	High	Best for training workloads.
West US 2	High	Medium	Higher demand → higher price.
Central US	Low	Medium	Often best price/performance.

West Europe	Medium-High	High	Good for compliance workloads.
Southeast Asia	High	Low	Limited A100 availability.

9. Benchmark Notes (High-Level)

Training Performance (Fake Data)

- A100 80GB achieves:
  - 320 TFLOPS Tensor performance.
  - 40–60% faster LLM training vs V100.
- H100 (preview) achieves:
  - Up to 3× faster LLM throughput vs A100.

Inference Performance

- L40S offers:
  - 6× throughput of T4 for vision tasks.
- A10 is:
  - 2× faster than T4 for transformer inference.

10. Key Limitations

- Some regions lack A100/H100 availability.
- Spot VMs may be terminated frequently.
- Large multi-node clusters may require quota increases.
- Peak times may cause provisioning delays.
- NV series not recommended for deep learning training.

11. FinOps Recommendations

- Use Spot VMs for:
  - Non-urgent training.
  - Batch inference pipelines.
- Use Reserved Instances for:
  - Always-on inference clusters.
- Enable:
  - Automatic checkpointing.
  - Autoscaling with GPU utilization metrics.
- Monitor with:
  - Azure Monitor
  - Application Insights
  - Prometheus/Grafana

12. Scaling Strategies

Horizontal Scaling

- Add more GPU nodes.
- Useful for embedding generation and distributed inference.

Vertical Scaling

- Switch to a larger VM with more VRAM.
- Useful for training larger models.

Multi-Node ML Training

- Use Azure Machine Learning (AzureML):
    - Orchestrates distributed PyTorch jobs.
    - Supports DeepSpeed, HuggingFace Accelerate, Ray.

13. Use-Case Mapping Summary

Use Case	Recommended GPU	VM Series	Cost Level
LLM Training (Large)	A100/H100	ND v4/v5	Very High
Fine-Tuning	A100/A10/L40S	NC/NV	Medium
Real-Time Inference	L40S/T4	NV/NC	Low–Medium
Vision Inference	L40S/A10	NV	Medium
Rendering	L40S/A6000	NV	Medium
CAD/Visualization	A10/M60	NV	Low–Medium

14. Conclusion

Azure provides one of the most flexible and scalable GPU environments for AI workloads. From small inference deployments to massive LLM training clusters, users can choose between A10, L40S, A100, and H100 GPUs, adjusting compute, networking, and storage options to match performance and cost.

For optimal results:

- Choose the VM family based on workload type.
- Use Spot VMs when possible.
- Scale with AzureML for distributed training.
- Use FinOps principles to monitor and control spend.
- Pick regions based on availability and pricing.