Azure GPU Instance Overview for M-Star CFD

This document provides a technical and financial overview of the premier Microsoft Azure GPU virtual machine (VM) families recommended for running large, GPU-native M-Star CFD simulations. It includes hardware specifications, architectural communication limits, and high-performance storage recommendations for both single-node and multi-node scaling.

> **Note on Azure terminology and architecture.** Azure exposes GPU compute primarily through the **ND-series** ("network-dense") VMs for scale-out HPC/AI, with each VM occupying a full physical 8-GPU node. A defining architectural feature is that **Azure uses NVIDIA Quantum InfiniBand for node-to-node communication**, giving each GPU its own dedicated RDMA link — a meaningful advantage for multi-node M-Star scaling. Prices below are approximate on-demand U.S. (East/Central) rates verified July 2026 and will vary by region, commitment (Reserved/Spot), and quota; treat them as planning figures, not quotes.

Executive Summary

Azure offers two practical tiers of GPU compute for M-Star CFD:

- **ND-series (8-GPU nodes) for production and scale-out.** The `ND96isr_H100_v5` (8x H100, 640 GB VRAM, ~$98.32/hr) is the workhorse for large single-node runs and — via its 400 Gb/s per-GPU Quantum-2 InfiniBand fabric — for efficient multi-node scaling. The older `ND96amsr_A100_v4` (8x A100, ~$27–33/hr) remains the most broadly available node and a reasonable fallback when H100 quota is scarce, though it is being phased toward end-of-life. At the top end, `ND GB200 v6` (Grace Blackwell) places up to 72 GPUs in a single NVLink domain for the very largest models.
- **NC-series (1–4 GPUs) for development and smaller runs.** These have no InfiniBand and cannot scale across VMs, but they are the cost-effective entry point for UDF development, validation cases, and small-to-mid production. Highlights: `NC24ads_A100_v4` (1x A100 80GB, ~$3.67/hr) for the cheapest iteration, and `NC40ads_H100_v5` (1x H100 NVL 94GB, ~$6.98/hr) for current-generation single-GPU work.

Key decisions for M-Star:

1. **Match the tier to the task.** Iterate on the Giesekus UDF, dissolution kernels, and reproducibility cases on a single-GPU NC VM; reserve full ND nodes for grids that need aggregate VRAM or true multi-node scaling.
2. **Exploit InfiniBand for multi-node.** ND-series scale-out is graceful because each GPU has a dedicated RDMA link; place VMs in a single Scale Set with proximity placement to let Azure auto-configure the fabric.
3. **Use Spot for checkpointed batch work.** Spot commonly saves 60–80% but sessions can be evicted on ~30 seconds' notice as demand shifts — pair it with frequent solver checkpointing, and use on-demand for interactive development.
4. **Plan for quota lead time.** Every subscription starts at zero GPU vCPU quota per region; ND H100 quota approval commonly takes 1–4 weeks, and reserved instances give better capacity assurance than on-demand.

---

1. Instance Specification Summary

ND H100 v5 (NVIDIA H100) — *The Scalability Sweet Spot*

- **Primary Model:** `Standard_ND96isr_H100_v5`
- **GPUs:** 8 x NVIDIA H100 (80GB SXM5 per GPU / 640GB total VRAM)
- **vCPUs:** 96 vCPUs (Intel Sapphire Rapids)
- **System Memory:** ~1,900 GiB *(Meets the 1.5–2x total GPU memory recommendation)*
- **Local Storage:** ~28 TB local NVMe SSD (scratch space)
- **Networking Technology:** NVIDIA Quantum-2 CX7 InfiniBand (400 Gb/s per GPU, GPU Direct RDMA)
- **On-Demand Price:** Approximately $98.32 per hour (U.S. East/Central)
- **Availability:** Available in ~24 regions but capacity-constrained; requires an Azure quota increase request (commonly a 1–4 week wait). Reserved instances give better capacity assurance.

ND A100 v4 (NVIDIA A100) — *The Legacy Budget Option*

- **Primary Model:** `Standard_ND96amsr_A100_v4` (80GB variant; `ND96asr_A100_v4` is the 40GB variant)
- **GPUs:** 8 x NVIDIA A100 (80GB or 40GB SXM4 per GPU)
- **vCPUs:** 96 vCPUs (AMD EPYC 7V12 "Rome")
- **System Memory:** ~1,800 GiB (80GB variant) / ~900 GiB (40GB variant)
- **Local Storage:** ~10.7 TB local NVMe SSD
- **Networking Technology:** NVIDIA Mellanox HDR InfiniBand (200 Gb/s per GPU, GPU Direct RDMA)
- **On-Demand Price:** Approximately $27–33 per hour depending on variant (typically accessed via Reserved or Spot at a steep discount)
- **Availability:** Broadest availability among the 8-GPU nodes (~29 regions), but being phased toward end-of-life for new workloads.

ND GB200 v6 (NVIDIA Grace Blackwell) — *Maximum Absolute Solver Speed*

- **Primary Model:** `ND GB200 v6` (rack-scale, sold at node/rack level — not per single GPU)
- **GPUs:** NVIDIA GB200 Grace Blackwell Superchips. Each VM = 2 Grace CPUs + 4 Blackwell GPUs (each GPU ~192GB HBM3e); racks scale to **72 GPUs in a single NVLink domain** via the GB200 NVL72 design.
- **vCPUs / CPU:** NVIDIA Grace ARM64 CPUs (custom Microsoft server), 2 per VM
- **System Memory:** Unified coherent Grace+Blackwell memory via NVLink-C2C
- **Local Storage:** High-density local NVMe (rack-class)
- **Networking Technology:** NVIDIA Quantum InfiniBand, scaling out to tens of thousands of GPUs
- **On-Demand Price:** Roughly $10.50–$27 per GPU-hour depending on billing model and provider tier; whole-VM/rack cost scales accordingly. Sold at the Superchip/rack level, not as single-GPU slots.
- **Availability:** Generally available (GA reached in 2025) but highly constrained; restricted to specific high-density, liquid-cooled HPC regions and typically reserved for the largest workloads.

Consolidated Pricing (U.S. East/Central, verified July 2026)

VM	GPUs	Total VRAM	On-Demand ($/hr)	On-Demand ($/mo, 730 hr)	Spot ($/hr)	1-Yr Reserved
`ND96amsr_A100_v4`	8x A100 (80GB)	640 GB	~$27–33	~$19,700–24,100	60–80% off	~35% off
`ND96isr_H100_v5`	8x H100 (SXM5)	640 GB	~$98.32	~$71,774	~$18.17 (≈82% off)	~35% off (~$63.40/hr)
`ND GB200 v6`	GB200 (rack-scale)	~192 GB/GPU	~$10.50–27 /GPU-hr	scales by rack	limited	reserved/contract

ND GB200 v6 is sold at the Superchip/rack level (4 GPUs per VM within a 72-GPU NVLink domain), not as single-GPU slots; whole-VM cost scales accordingly. Regional surcharges of ~20–33% apply outside U.S. regions.

> **Spot pricing caveat:** Spot rates and instance sessions are subject to real-time demand. Azure can **evict a running Spot VM with as little as 30 seconds' notice** if on-demand demand rises or the spot price exceeds your cap — the session simply terminates. For M-Star on Spot, implement frequent solver checkpointing so a simulation can resume after eviction; treat Spot as a cost-reduction tool for restartable batch work, not for interactive or time-critical runs. Reserved instances are a billing commitment and improve capacity assurance, but do not guarantee Azure will have capacity available in a given region.

1b. Smaller Instances — Single- and Dual-GPU Options (NC-Series)

The ND-series 8-GPU nodes above are sized for large production runs, but many M-Star workflows do not need a full node. **Development, UDF debugging, mesh/case setup validation, small-to-mid grid production runs, and single-GPU benchmarking** are far better served by the **NC-series**, which exposes 1, 2, or 4 GPUs per VM at a fraction of the cost. These are the natural entry point for iterating on the Giesekus UDF, dissolution kernels, or reproducibility test cases before committing to a full ND node.

> **Key architectural caveat.** NC-series VMs **do not have InfiniBand** — there is no low-latency RDMA scale-out fabric between separate NC VMs. This makes them unsuitable for multi-node M-Star runs. However, the multi-GPU NC sizes *do* retain **on-board NVLink** between the GPUs inside a single VM, so intra-VM domain decomposition (e.g., 2x or 4x GPU on one VM) still scales well. Treat NC-series as "single-node, up to 4 GPUs" and reserve ND-series for anything spanning physical servers.

NCads H100 v5 (NVIDIA H100 NVL, 94GB) — *Best Single/Dual H100 for Iteration*

- **Models:** `Standard_NC40ads_H100_v5` (1 GPU) / `Standard_NC80adis_H100_v5` (2 GPU)
- **GPUs:** 1–2 x NVIDIA H100 NVL (94GB HBM3 per GPU — note the *larger* 94GB VRAM vs. the 80GB SXM5 in the ND nodes)
- **vCPUs / CPU:** 40 (1 GPU) or 80 (2 GPU) — AMD EPYC Genoa
- **System Memory:** 320 GiB (1 GPU) / 640 GiB (2 GPU)
- **Local Storage:** ~3.5 TB (1 GPU) / ~7.2 TB (2 GPU) NVMe temp disk

- **Interconnect:** NVLink between the two GPUs on the dual-GPU size; **no InfiniBand**
- **On-Demand Price:** ~$6.98/hr (1 GPU) / ~$13.96/hr (2 GPU); Spot drops to roughly $1.40–$2.80/hr
- **Best For:** Current-generation single-GPU M-Star production, UDF development against H100 hardware, and 2-GPU runs that fit in 188GB combined VRAM.

NC A100 v4 (NVIDIA A100 PCIe, 80GB) — *The Budget Development Workhorse*

- **Models:** `Standard_NC24ads_A100_v4` (1 GPU) / `Standard_NC48ads_A100_v4` (2 GPU) / `Standard_NC96ads_A100_v4` (4 GPU)
- **GPUs:** 1–4 x NVIDIA A100 PCIe (80GB HBM2e per GPU)
- **vCPUs / CPU:** 24 / 48 / 96 — AMD EPYC Milan
- **System Memory:** 220 / 440 / 880 GiB respectively
- **Local Storage:** 960 GB (1 GPU) up to 4x960 GB NVMe
- **Interconnect:** NVLink bridges across the GPUs within the 2- and 4-GPU sizes; **no InfiniBand**
- **On-Demand Price:** ~$3.67/hr (1 GPU) / ~$7.35/hr (2 GPU) / ~$14.69/hr (4 GPU); Spot as low as ~$0.68/hr for a single GPU
- **Best For:** The cheapest way to run real M-Star cases on datacenter-class hardware. Ideal for the bulk of UDF debugging, small validation cases (Taylor-Couette, power-number sweeps), and cost-sensitive parameter studies. **Note:** Azure is no longer deploying net-new NC A100 v4 capacity (favoring the H100 v5 line), so availability may tighten over time.

NCC H100 v5 (Confidential Computing) — *Niche: Sensitive Customer Data*

- **Model:** `Standard_NCC40ads_H100_v5` (1 GPU, 94GB H100 NVL)
- **Distinguishing Feature:** Trusted Execution Environment (TEE) spanning both CPU and GPU, enabling secure offload of models and data.
- **On-Demand Price:** ~$8.90/hr (a premium over the standard NC40ads for the confidential-compute capability)
- **Relevance to M-Star:** Only worth considering if running a customer's proprietary geometry or formulation data under strict confidentiality requirements. For general R&D, the standard NC40ads is the better value.

Selecting a Smaller Instance for M-Star

Use Case	Recommended Instance	On-Demand ($/hr)	On-Demand ($/mo, 730 hr)	Spot ($/hr)
UDF debugging, cheapest iteration	`NC24ads_A100_v4` (1x A100 80GB)	~$3.67	~$2,681	~$0.68
Single-GPU current-gen production	`NC40ads_H100_v5` (1x H100 94GB)	~$6.98	~$5,095	~$1.40
Mid-size case, 2-GPU intra-VM scaling	`NC80adis_H100_v5` (2x H100 94GB)	~$13.96	~$10,191	~$2.80
4-GPU single-node, budget-conscious	`NC96ads_A100_v4` (4x A100 80GB)	~$14.69	~$10,724	~$2.72

Use Case	Recommended Instance	On-Demand ($/hr)	On-Demand ($/mo, 730 hr)	Spot ($/hr)
Confidential customer data	`NCC40ads_H100_v5` (1x H100, TEE)	~$8.90	~$6,497	~$1.64

> *Spot pricing on all NC sizes commonly saves 60–80% and is well-suited to checkpointed or restartable M-Star runs, at the cost of a 30-second eviction warning. For interactive UDF development, use on-demand; for overnight batch sweeps, Spot is highly cost-effective.*

---

2. Communication & Memory Bandwidth Architecture

Because M-Star CFD is a GPU-native solver utilizing highly parallel lattice or particle-tracking methodologies, performance is often heavily constrained by data-transfer speeds. Evaluating communication layers across the tiers below defines how efficiently a simulation scales. The single most important Azure-specific point is the **InfiniBand scale-out fabric**, which substantially narrows the node-to-node communication bottleneck for multi-node runs.

Bandwidth Metric	`ND96amsr_A100_v4` (8x A100)	`ND96isr_H100_v5` (8x H100)	`ND GB200 v6` (Blackwell)
Local GPU Memory Type	HBM2e	HBM3	HBM3e
Per-GPU Memory Bandwidth	~2.0 TB/s	3.35 TB/s	8.0 TB/s
Aggregate Node Memory Bandwidth *(8 GPU)*	**~16.0 TB/s**	**26.8 TB/s**	**64.0 TB/s** *(8-GPU equivalent basis)*
Card-to-Card Interconnect	NVLink 3rd Gen	NVLink 4th Gen	NVLink 5th Gen
Per-GPU NVLink Bandwidth	~0.6 TB/s	~0.9 TB/s	1.8 TB/s
Node-to-Node Network (InfiniBand)	200 Gb/s per GPU (HDR)	400 Gb/s per GPU (Quantum-2 CX7)	Quantum InfiniBand (latest gen)
Aggregate Scale-Out Bandwidth per VM	**~1.6 TB/s** (12.8 Tbps)	**~3.2 TB/s** (25.6 Tbps)	**Rack-scale, multi-Tbps**

> *Bandwidth figures combine vendor specs (NVIDIA H100/A100/B200 datasheets) with Azure VM documentation. The Blackwell column reflects per-GPU Blackwell specs presented on an 8-GPU-equivalent basis for consistency across the table; the actual `ND GB200 v6` VM ships 4 GPUs per VM within a 72-GPU NVLink rack domain.*

Architectural Takeaways for M-Star Scaling

- **Local Memory Bandwidth (HBM):** Crucial for localized solver performance. Moving from A100 (~2.0 TB/s per GPU) to Blackwell (8.0 TB/s per GPU) is a **~4x per-GPU memory-throughput increase**,

massively decreasing the wall-clock time required for dense grid iterations.

- **Intra-Node Scaling (NVLink):** Within a node, GPUs communicate directly over NVSwitch fabric rather than the host PCIe bus. NVLink 4th Gen (H100) and 5th Gen (Blackwell, 1.8 TB/s per GPU) ensure that parallelizing a model across all GPUs in the node incurs near-zero communication lag.
- **Multi-Node Scaling — Azure's InfiniBand Fabric:** Dedicated per-GPU RDMA links are what make ND-series scale-out efficient.
    - **On ND A100 v4 nodes**, each GPU gets a dedicated 200 Gb/s HDR InfiniBand link with GPU Direct RDMA — roughly **1.6 TB/s aggregate scale-out per VM**.
    - **On ND H100 v5 nodes**, each GPU gets a dedicated 400 Gb/s Quantum-2 InfiniBand link (~3.2 TB/s aggregate per VM), enabling highly linear multi-node M-Star scaling.
    - **On ND GB200 v6**, up to 72 GPUs sit inside a single NVLink domain (behaving like one enormous GPU) before InfiniBand even comes into play for further scale-out — effectively eliminating the conventional multi-node boundary for models that fit the rack.

---

3. Infrastructure & Deployment Recommendations

To guarantee that hardware performance translates effectively to M-Star environments on Azure, implement the following ecosystem selections:

High-Performance Storage

- **Azure Managed Lustre (AMLFS):** Avoid standard managed disks for high-transient IOPS. M-Star output datasets are massive; Azure Managed Lustre provides the parallel file-system architecture needed to match the ingestion speed required by 8 concurrent high-tier GPUs during checkpoint writing and post-processing exports. **Azure NetApp Files** is a viable alternative for shared scratch where Lustre is overkill.

Operating System & Image Selection

- **Azure HPC / AI Marketplace Images (Ubuntu-HPC or the NVIDIA GPU-Optimized image):** Use the pre-built Azure HPC images, which ship with validated NVIDIA driver stacks, CUDA, the InfiniBand (OFED/Mellanox) drivers, and NCCL pre-configured. These ensure both the NVLink switches and the InfiniBand HCAs are accessible out of the box — critical, since misconfigured InfiniBand is the most common cause of poor multi-node scaling on Azure.

Scale-Out Mechanics

- **VM Scale Sets + Proximity Placement:** Multi-node M-Star runs must place VMs in the same **Virtual Machine Scale Set** so Azure auto-configures the InfiniBand fabric between them with GPU Direct RDMA. Use proximity placement groups to keep nodes physically close and minimize latency.

Architectural Summary Recommendation

- **Single-Node Priority:** If your total lattice/particle grid fits entirely within 640GB of VRAM and you want peak performance for minimal cost, a single `ND96isr_H100_v5` provides the ideal price-to-performance matrix.
- **Multi-Node Priority:** Azure's InfiniBand fabric makes multi-node scaling graceful, with each GPU carrying a dedicated RDMA link. The **ND H100 v5** family (400 Gb/s per-GPU Quantum-2 InfiniBand) is

the practical workhorse for distributed M-Star runs. Reserve **ND GB200 v6** for the largest models that can exploit the 72-GPU single-NVLink-domain rack, where conventional multi-node communication penalties effectively disappear.

- **Budget / Availability Fallback:** The **ND A100 v4** family remains the most broadly available 8-GPU node and, thanks to its 200 Gb/s per-GPU InfiniBand, still scales out respectably — a reasonable choice when H100 quota is unavailable.