AWS GPU Instance Overview for M-Star CFD

Executive Summary

For most large M-Star CFD workloads, `p5.48xlarge` is the best default choice because it combines 8x H100 GPUs, 640 GB of total VRAM, strong intra-node NVLink bandwidth, and 3.2 Tbps EFA networking at a materially lower hourly cost than Blackwell-based P6. If a simulation fits within a single P5 node, it generally offers the strongest price-to-performance balance for production runs.

Choose `p6-b200.48xlarge` when the model exceeds 640 GB of aggregate GPU memory or when wall-clock speed is valuable enough to justify roughly 2x the hourly cost. Avoid `p4d.24xlarge` for new multi-node deployments unless budget or regional availability is the controlling constraint; its older A100 platform and much weaker 400 Gbps network fabric make it the least attractive option for tightly coupled distributed solves.

For development, UDF validation, and smaller production cases, `p5.4xlarge` is the most practical single-GPU option because it gives access to the same H100 architecture at about one-eighth of the full-node cost. Use `g6e` only when cost sensitivity outweighs peak solver throughput, and pair any serious production deployment with checkpointing, FSx for Lustre, and advance AWS quota planning.

This document provides a technical and financial overview of the premier AWS GPU instances 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.

Pricing verified July 2026 (us-east-1, On-Demand, Amazon Linux). AWS list prices change without notice — confirm live rates on the EC2 pricing page before budgeting.

---

1. Instance Specification Summary

P5 Instances (NVIDIA H100) — *The Scalability Sweet Spot*

- **Primary Model:** `p5.48xlarge`
- **GPUs:** 8 x NVIDIA H100 (80GB SXM5 per GPU / 640GB total VRAM)
- **vCPUs:** 192 vCPUs (AMD EPYC, 96 physical cores)
- **System Memory:** 2048 GiB *(Meets the 1.5–2x total GPU memory recommendation)*
- **Local Storage:** 8 x 3.84 TB NVMe SSD (~30.4 TB local scratch space)
- **Networking Technology:** Elastic Fabric Adapter (EFAv2), 3,200 Gbps
- **On-Demand Price:** ~$55.04 per hour (~$40,179/mo). Spot as low as ~$18.83/hr when available.
- **Availability:** 6+ regions; requires proactive AWS Service Quota increase requests. New accounts default to 0 vCPUs for P instances.

P4 Instances (NVIDIA A100) — *The Legacy Budget Option*

- **Primary Model:** `p4d.24xlarge`
- **GPUs:** 8 x NVIDIA A100 (40GB or 80GB SXM4 per GPU)
- **vCPUs:** 96 vCPUs
- **System Memory:** 1152 GiB

- **Local Storage:** 8 x 1000 GB NVMe SSD
- **Networking Technology:** Elastic Fabric Adapter (EFAv1), 400 Gbps
- **On-Demand Price:** Starts around $11.80–$32.77 per hour depending on region/OS (typically accessed via Capacity Blocks or Spot). Reflects the June 2025 price reduction (up to 33% off P4d/P4de).
- **Availability:** Broadest availability among the 8-GPU nodes, but still subject to placement groups and quota restrictions.

P6 Instances (NVIDIA Blackwell B200) — *Maximum Absolute Solver Speed*

- **Primary Model:** `p6-b200.48xlarge`
- **GPUs:** 8 x NVIDIA B200 (~180GB SXM6 per GPU / ~1,432GB total VRAM)
- **vCPUs:** 192 vCPUs (Intel, 96 physical cores, ~2.4 GHz)
- **System Memory:** 2048 GiB
- **Local Storage:** 8 x 3800 GB NVMe SSD (~30.4 TB)
- **Networking Technology:** Elastic Fabric Adapter (EFAv4), 3,200 Gbps
- **On-Demand Price:** ~$113.93 per hour (~$83,171/mo). Spot as low as ~$21.59/hr when available. Realistic all-in cost (storage + egress + networking) runs closer to $120–130/hr.
- **Availability:** Cutting-edge hardware; constrained to ~4 regions (N. Virginia, Ohio, Oregon, Mumbai). Savings Plans now available (previously Capacity Blocks only at launch). Quota approval takes ~3–7 business days with written business justification.

Consolidated Pricing (us-east-1, verified July 2026)

Instance	GPUs	Total VRAM	On-Demand ($/hr)	On-Demand ($/mo, 730 hr)	Spot Floor ($/hr)	3-Yr Savings Plan*
`p4d.24xlarge`	8x A100	320 / 640 GB	~$32.77	~$23,922	50–70% off	up to ~33% off
`p5.48xlarge`	8x H100	640 GB	~$55.04	~$40,179	~$18.83 (≈66% off)	up to ~45% off
`p6-b200.48xlarge`	8x B200	~1,432 GB	~$113.93	~$83,171	~$21.59 (≈81% off)	up to ~57% off (family-locked)

* Savings Plan discounts are approximate and depend on commitment term, upfront choice, and whether the plan is instance-family-locked. 1-year plans typically fall between On-Demand and the 3-year figures shown.

> **Spot pricing caveat:** Spot rates and instance sessions are subject to real-time demand. AWS can reclaim a running Spot instance **with as little as two minutes' notice** if On-Demand demand rises or the Spot price exceeds your bid — the session will simply terminate. For M-Star runs on Spot, implement frequent solver checkpointing so a simulation can resume after an interruption; treat Spot as a cost-reduction tool for interruptible work, not for time-critical or non-checkpointed runs. Spot availability for P-family instances is structurally scarce because AWS prioritizes On-Demand and Reserved purchasers for allocation.

---

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 three tiers defines how efficiently a simulation scales.

Bandwidth Metric	`p4d.24xlarge` (8x A100)	`p5.48xlarge` (8x H100)	`p6-b200.48xlarge` (8x B200)
Local GPU Memory Type	HBM2	HBM3	HBM3e
Per-GPU Memory Bandwidth	1.55 TB/s	3.35 TB/s	8.0 TB/s
Aggregate Node Memory Bandwidth	**12.4 TB/s**	**26.8 TB/s**	**64.0 TB/s**
Card-to-Card Interconnect	NVLink 3rd Gen	NVLink 4th Gen	NVLink 5th Gen
Aggregate Card-to-Card Bandwidth	**4.8 TB/s**	**7.2 TB/s**	**14.4 TB/s**
Node-to-Node Network (EFA)	400 Gbps (EFAv1)	3,200 Gbps (EFAv2)	3,200 Gbps (EFAv4)
Aggregate Network Bandwidth	**50 GB/s**	**400 GB/s**	**400 GB/s**

Architectural Takeaways for M-Star Scaling

- **Local Memory Bandwidth (HBM):** Crucial for localized solver performance. Upgrading from the P4 generation to a P6 provides an immense **5.1x increase in aggregate memory throughput** ($12.4\ \mathrm{TB/s} \rightarrow 64.0\ \mathrm{TB/s}$), massively decreasing the wall-clock time required for dense grid iterations.
- **Intra-Node Scaling (NVLink):** Within a single node, GPUs communicate directly across hardware NVSwitches rather than over the host PCIe bus. The $7.2\ \mathrm{TB/s}$ (P5) and $14.4\ \mathrm{TB/s}$ (P6) fabrics ensure that parallelizing a model across all 8 internal GPUs incurs almost zero communication lag.
- **Multi-Node Scaling Drop-off (EFA vs NVLink):**
  - **On P4d nodes**, the network fabric ($50\ \mathrm{GB/s}$) is **96x slower** than internal NVLink ($4.8\ \mathrm{TB/s}$). Multi-node M-Star runs will hit severe communication bottlenecks when exchanging transient boundary data between separate servers.
  - **On P5 and P6 nodes**, AWS narrowed this architectural delta to an **18x to 36x drop**. This massive network expansion to $400\ \mathrm{GB/s}$ enables highly linear and efficient multi-node scaling.

---

3. Smaller Instances (Fewer GPUs)

Not every M-Star case needs a full 8-GPU node. For model development, UDF debugging, mesh/domain sizing studies, or smaller production grids, single-GPU and reduced-GPU options substantially cut cost.

Single-GPU P5 (H100) — `p5.4xlarge`

- **GPU:** 1x NVIDIA H100 (80GB SXM5) — same architecture as the full node.

- **vCPUs / Memory:** 16 vCPUs / 256 GiB.
- **Local Storage:** 1x 3.8 TB NVMe SSD.
- **Networking:** 100 Gbps baseline (no EFA fabric — single-GPU only, no multi-node scaling).
- **On-Demand Price:** ~$6.88/hr (~$5,022/mo); Spot from ~$2.53/hr.
- **Availability:** Available On-Demand, Spot, or Savings Plans in select regions (London, Mumbai, Jakarta, Tokyo, São Paulo); via Capacity Blocks more broadly. Still requires a P-instance quota increase.
- **Use case:** Ideal for M-Star development, UDF validation (e.g., Giesekus/Taylor-Couette test cases), and single-GPU production runs where the grid fits in 80GB. This is $6.88/hr for the *same* H100 as `p5.48xlarge` at 1/8th the node cost.

> **P4 note:** The P4d family (A100) is **only** offered as the full 8-GPU `p4d.24xlarge` node — there is no single- or reduced-GPU P4d SKU. The `p6-b200` family is likewise only the full 8-GPU node today.

Higher-VRAM H200 variants — `p5e` / `p5en`

- Same P5 chassis but with **H200 GPUs (192GB HBM3e per GPU)** — more than 2x the standard H100's VRAM. Useful when a single-node grid exceeds the H100's 80GB/GPU but you want to avoid multi-node scaling. Available in fewer regions and harder to get quota-approved; check the EC2 console for current rates.

Budget single-GPU — G6e (NVIDIA L40S)

For lightweight development, pre/post-processing, visualization, or smaller LBM grids where H100 throughput isn't required, the L40S-based G-family is far cheaper. The L40S is PCIe-only (no NVLink), so it suits single-GPU work rather than tightly-coupled multi-GPU runs.

Instance	GPU	VRAM	vCPU / RAM	On-Demand ($/hr)	Spot ($/hr)
`g6e.xlarge`	1x L40S	48 GB	4 / 32 GiB	~$1.86	~$1.38
`g6e.4xlarge`	1x L40S	48 GB	16 / 128 GiB	~$3.00	—

G6e also scales up to 8x L40S (384 GB total) in a single node with up to 400 Gbps networking, offering a mid-tier multi-GPU option below P5 pricing. G6e is broadly available On-Demand, Spot, Reserved, or via Savings Plans without the P-family quota friction.

Rule of thumb for M-Star: use single-GPU `p5.4xlarge` for H100-class development and small production; step up to a full `p5.48xlarge`/`p6-b200.48xlarge` node only when the grid demands aggregate VRAM or multi-GPU throughput; and reach for `g6e` when the workload is memory-modest and cost sensitivity outweighs raw solver speed.

---

4. Infrastructure & Deployment Recommendations

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

High-Performance Storage

- **Amazon FSx for Lustre:** Avoid basic EBS volumes for high-transient IOPS. M-Star simulation output datasets are massive; FSx for 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.

Operating System & Image (AMI) Selection

- **AWS Deep Learning AMI (DLAMI) or NVIDIA GPU-Optimized AMI:** Use these pre-built, cloud-optimized Linux distributions (typically Ubuntu or RHEL bases). They come pre-configured with the precise fabric drivers, container toolkits, and validated CUDA/NVIDIA driver stacks necessary to access both NVLink switches and EFA network cards out of the box.

Cost Layers Beyond the Hourly Rate

- Budget for hidden line items that headline hourly rates omit: data egress to internet (~$0.09/GB for the first 10 TB/mo), EBS root volumes, CloudWatch logging, and EFA networking within a placement group. For long-running production work, evaluate 1- and 3-year Savings Plans (roughly 21–27% off flexible, up to ~57% for instance-family-locked commitments).

Architectural Summary Recommendation

- **Single-Node Priority:** If your total lattice/particle grid fits entirely within 640GB of VRAM and requires peak performance for minimal cost, a single `p5.48xlarge` provides the ideal price-to-performance matrix. For grids exceeding 640GB, the `p6-b200.48xlarge` offers ~1,432GB of aggregate VRAM in a single node.
- **Multi-Node Priority:** If simulations demand scaling across multiple physical servers, **avoid the P4d family** due to its $400\ \mathrm{Gbps}$ network bottleneck. Leverage the **P5** or **P6** architectures to maintain highly parallel execution across the $3.2\ \mathrm{Tbps}$ EFA mesh.