



NVIDIA Nemotron 3 Ultra Hardware Requirements: VRAM and Cost

Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-07-22 · nemotron 3 ultra · nvidia nemotron · gpu requirements · vram requirements · llm hardware requirements · dgx spark · dgx station · nvfp4 quantization · self-hosting llm cost · inference cost per token

Original article: <https://gpusmith.com/articles/en/nvidia-nemotron-3-ultra-hardware-requirements>



In this article

1. Executive Summary
 2. Introduction and Background
 3. Local and Workstation-Class Deployment
 4. Single-Node Datacenter Deployment: Minimum GPU Requirements by Platform
 5. Multi-Node and Enterprise-Scale Deployment
 6. API and Usage-Based Access: Inference Cost per Token
 7. Comparative Context and Market Positioning
 8. Data Analysis and Evidence
 9. Case Studies and Real-World Examples
 10. Amazon SageMaker JumpStart: Managed Enterprise Deployment
 11. CoreWeave GB200 NVL72: Rack-Scale Inference Capacity
 12. DGX Spark Four-Node Cluster: The Documented Prosumer Path
 13. DGX Station with MoE Expert Offloading: Single-Box Deployment
 14. Independent Consumer Hardware Testing: The Strix Halo Reality Check
 15. Implications and Future Directions
 16. Frequently Asked Questions (FAQs)
 17. Conclusion
-

Executive Summary

NVIDIA's **Nemotron 3 Ultra** (formally **NVIDIA-Nemotron-3-Ultra-550B-A55B**) is a 550-billion-parameter Mixture-of-Experts (MoE) model with 55 billion parameters active per token, released on **June 4, 2026** as the flagship of the Nemotron 3 family (huggingface.co) (research.nvidia.com). It uses a hybrid Mamba-2, Mixture-of-Experts, and Attention architecture NVIDIA calls LatentMoE, supports context windows up to 1,048,576 tokens, and ships in two primary checkpoints: a BF16 (16-bit floating point) release and a quantized NVFP4 (4-bit floating point) release that roughly halves the hardware footprint (research.nvidia.com). Hardware requirements differ sharply by precision. NVIDIA's own model card lists a minimum of **8x GB200, B200, GB300, or B300 GPUs, 16x H100, or 8x H200** for the BF16 checkpoint, while the NVFP4 checkpoint drops that floor to **4x GB200, B200, GB300, or B300, or 8x H100** (huggingface.co) (build.nvidia.com).

At the weight level, community benchmarking site llmrun.dev measures the BF16 checkpoint (560.5B parameters counting embeddings) at 1,233.2 GB of [VRAM](#) in full precision, dropping to 369.9 GB at a 4-bit Q4_K_M quantization, with 481+ GB recommended once KV cache and system overhead are included (llmrun.dev). NVIDIA's official NVFP4 checkpoint measures approximately 352 GB on disk, matching a real-world Reddit hardware audit (reddit.com). No single consumer or even most professional GPUs can hold these weights; llmrun.dev states plainly that "no single GPU has enough memory" and multi-GPU or cluster configurations are required (llmrun.dev).

On the cost side, cloud GPU rental prices as of July 2026 range from roughly \$2 to \$7 per GPU per hour for H100 and H200 units and \$3.75 to \$10 per GPU per hour for B200 units, depending on provider and commitment level (getdeploying.com) (getdeploying.com). Using NVIDIA's own published throughput benchmarks of 310.8 tokens per second per GPU on a 4x B200 agentic workload (docs.nvidia.com) combined with Lambda's published four-GPU B200 SXM6 pricing, this report calculates an illustrative self-hosted compute cost near \$6 per million output tokens at full utilization, roughly 2.7 times the \$2.20 per million output tokens that DeepInfra lists on OpenRouter for the same model (openrouter.ai). For local and workstation use, NVIDIA's own deployment cookbook recommends a four-node DGX Spark cluster (each node \$4,699, 128 GB unified memory) or a single DGX Station (starting around \$94,000, 748 GB of coherent GPU-plus-CPU memory) as the only sanctioned non-datacenter paths (marketplace.nvidia.com) (exxactcorp.com).

This report walks through every documented deployment tier for Nemotron 3 Ultra: local and workstation-class hardware, single-node datacenter GPU configurations, multi-node clusters, and usage-based API access, quantifying VRAM, GPU count, node topology, and dollar cost at each tier and cross-checking NVIDIA's official specifications against independent community benchmarks and cloud pricing data gathered in July 2026.

Introduction and Background

Nemotron 3 Ultra sits at the top of NVIDIA's third-generation Nemotron family of open-weight language models, released alongside a mid-sized "Super" (120B total, 12B active) and smaller "Nano" (30B total, 3B active) sibling (docs.nvidia.com). Unlike a conventional dense transformer, Nemotron 3 Ultra interleaves Mamba-2 state-space layers, Mixture-of-Experts feed-forward blocks, and select attention layers, a design NVIDIA calls a hybrid Mamba-Transformer, further augmented with a Latent Mixture-of-Experts (LatentMoE) routing scheme; NVIDIA's TensorRT-LLM guide notes "Nemotron v3 Ultra additionally uses a Latent Mixture-of-Experts (LatentMoE) design and ships with built-in Multi-Token Prediction (MTP) layers" (nvidia.github.io). The model also ships with built-in MTP layers, which let inference engines draft several tokens per forward pass for native speculative decoding, and it was pretrained using an NVFP4 (4-bit floating point) quantization-aware recipe intended to preserve accuracy while cutting memory and compute cost (research.nvidia.com).

The practical consequence of this architecture is that Nemotron 3 Ultra behaves very differently from a same-sized dense model when it comes to hardware planning. Because only 55 billion of its 550 billion total parameters activate per forward pass, its compute cost per token resembles a much smaller model, but its memory footprint, every parameter must still be resident on a GPU or reachable over a fast interconnect, remains that of a genuinely enormous model. NVIDIA's own TensorRT-LLM deployment guide is explicit on this point, noting that developers must "ensure you have sufficient GPU memory for the full parameter weights even though only a fraction of parameters are active per token" (nvidia.github.io).

NVIDIA's own sources describe pretraining scale in two slightly different ways worth noting honestly rather than glossing over. The official model card states the base checkpoint "was pre-trained for approximately 20T tokens" of crawled and synthetic code, math, science, and general-knowledge data (build.nvidia.com), while NVIDIA's developer blog separately frames Nemotron 3 Ultra as "building on a 10T token pre-training foundation, Nemotron 3 Ultra adds 212B new tokens" targeting legal, wiki-derived factual, and refreshed GitHub data through September 30, 2025 (developer.nvidia.com); this report cannot fully reconcile the "20T total" and "10T foundation plus 212B" framings from public NVIDIA materials alone and presents both figures

as published rather than silently picking one. Post-pretraining stages included supervised fine-tuning, reinforcement learning with GRPO, and a novel Multi-Teacher On-Policy Distillation (MOPD) stage that trains the model against feedback from more than ten domain-specialist teacher models. The model, its training data, and its recipes are released under the Linux Foundation's OpenMDW-1.1 license, which NVIDIA frames as a more permissive successor to prior Nemotron license terms (huggingface.co).

Because the weights, datasets, and fine-tuning recipes are all open, Nemotron 3 Ultra is being evaluated by a wide spectrum of users, from hyperscale cloud providers running it at scale for millions of agentic tool calls to individual hobbyists on Reddit trying to squeeze a runnable quantization onto a single high-memory workstation. Those two audiences face almost entirely different hardware and cost questions, and this report addresses both, working from NVIDIA's own model cards and deployment guides outward through independent benchmarks, cloud pricing trackers, and firsthand community reports gathered as of July 2026.

Local and Workstation-Class Deployment

The honest starting point for anyone evaluating Nemotron 3 Ultra for local use is that, as of July 2026, there is no consumer-grade, single-GPU, or even single-workstation path that NVIDIA officially sanctions for the full model. NVIDIA's own deployment cookbook index lists only two non-datacenter targets for Nemotron 3 Ultra: a clustered **DGX Spark** configuration and a single **DGX Station** (docs.nvidia.com).

DGX Spark, NVIDIA's smallest AI desktop appliance, pairs a GB10 Grace Blackwell Superchip with 128 GB of coherent unified memory and roughly 1 petaFLOP of FP4 compute in a 150mm-by-150mm enclosure, retailing at **\$4,699** through NVIDIA's marketplace (marketplace.nvidia.com). A single Spark's 128 GB is far short of the roughly 352 GB the NVFP4 checkpoint occupies on disk, so NVIDIA's official cookbook instead documents deployment "across a 4x DGX Spark cluster with vLLM, including tensor parallelism, RoCE networking, MTP speculative decoding, and NVIDIA AIPerf benchmarking," which pools four units into 512 GB of aggregate unified memory (docs.nvidia.com). At list price, four Spark units cost **\$18,796** before networking hardware, a meaningful capital outlay but one still well below datacenter GPU pricing.

DGX Station takes a different approach: a single desktide unit built around one GB300 Grace Blackwell Ultra Desktop Superchip, offering 252 GB of HBM3e GPU memory plus 496 GB of LPDDR5X CPU memory bridged by NVLink-C2C at 900 GB/s, for **748 GB of coherent memory** in total, enough to hold the NVFP4 checkpoint's roughly 352 GB of weights with substantial headroom (nvidia.com). Because the GPU's own HBM3e (252 GB) alone is smaller than the model, NVIDIA's cookbook for this path explicitly relies on "coherent CPU memory for selective MoE expert offloading and a FlashInfer TensorRT-LLM NVFP4 backend," meaning inactive experts are held in system memory and pulled across the NVLink-C2C interconnect on demand rather than all 550 billion parameters residing in GPU memory simultaneously (docs.nvidia.com). System integrator Exxact lists a DGX Station starting at **\$94,011.50**, and other OEM channel partners have listed configurations from roughly \$97,000 to \$99,500 (exxactcorp.com) (servethehome.com).

Below the DGX tier, the model becomes genuinely impractical. Community quantizer Unsloth published dynamic GGUF quantizations for llama.cpp-family runtimes with the guidance to "run 2-bit on 200GB RAM, 3-bit on 256GB, 8-bit on 600GB," while its maintainer noted on Reddit that the most aggressive quantization the team could produce without breaking usability was "190GB 1-bit dynamic" (reddit.com) (reddit.com). A

separate independent hardware audit on r/LocalLLaMA testing a 128 GB AMD Strix Halo (Ryzen AI MAX+ 395) unified-memory mini PC found "I did not find a direct GGUF / llama.cpp route for Ultra during this scan," concluding the full model is a "watchlist item" rather than a runnable target on that class of hardware, while the mid-sized Nemotron 3 Super (120B total, 12B active) did run directly at roughly 17.9 tokens per second ([reddit.com](#)). One commenter on that thread summarized the consensus bluntly: "Ultra-size models are not meant to run on 128GB devices" ([reddit.com](#)). Other users estimated a workable 4-bit build could be assembled for roughly \$2,000 to \$2,500 in used server hardware, accepting throughput "around 10 tokens per second," a workable but far from snappy interactive pace ([reddit.com](#)). A different commenter on the same thread called the 3-bit path "almost no one owns and with effectively zero context," a pointed reminder that quantized VRAM figures alone do not guarantee a usable context window ([reddit.com](#)).

Despite the hardware barrier, community interest in local quantizations has been substantial. llmrun.dev's tracker shows the official BF16 checkpoint has accumulated "67.2K downloads" and "16.6K quant downloads," with independent GGUF conversions available from repositories including unsloth, AesSedai, and ddh0, plus a 4-bit MLX-format build for Apple Silicon from mlx-community ([llmrun.dev](#)). That level of community tooling activity, despite the model's impracticality on nearly all locally available hardware, mirrors the broader pattern seen with the Nemotron 3 Super and Nano siblings, which NVIDIA positions as the more deliberately local-friendly members of the family.

Table 1 below breaks out the full community GGUF quantization ladder for the BF16 checkpoint as measured by llmrun.dev, illustrating just how much precision reduction is required before the model becomes even theoretically local, and confirming that even the most aggressive practical quantization remains far beyond consumer GPU memory.

Table 1. Community GGUF quantization tiers for the BF16 checkpoint (llmrun.dev estimates)

Quantization	VRAM Required	Downloadable File Size	Community Assessment
IQ2_XXS (smallest practical tier)	169.6 GB	154.14 GB	Still "exceeds the unified memory of most consumer Macs" (llmrun.dev)
Q4_K_M (recommended default)	369.9 GB	336.31 GB	"Offers the best balance of quality and VRAM usage" (llmrun.dev)
Q5_K_S (higher quality)	423.9 GB	not separately published	"Provides better quality if you have the VRAM" (llmrun.dev)
BF16 (full precision)	1,233.2 GB	1,121.05 GB	Full-precision baseline; "the full-precision BF16 version is 1121.05 GB" (llmrun.dev)

No tier in *Table 1* fits inside a single consumer or most professional GPUs. llmrun.dev's own compatibility checker states flatly that "no single consumer GPU has enough VRAM to run NVIDIA Nemotron 3 Ultra 550B A55B BF16 at Q4_K_M (369.9 GB)" and that "multi-GPU or professional hardware is required" for any of the tiers above ([llmrun.dev](#)).

Single-Node Datacenter Deployment: Minimum GPU Requirements by Platform

For anyone deploying inside a datacenter or cloud GPU instance, NVIDIA publishes concrete minimums that differ by both GPU generation and checkpoint precision. Precision is the single largest lever on hardware requirements: the BF16 checkpoint needs roughly twice the GPU count of the NVFP4 checkpoint at every tier.

For the BF16 checkpoint, NVIDIA's Hugging Face model card lists a minimum of **8x GB200, B200, GB300, or B300; 16x H100; or 8x H200** (huggingface.co), and NVIDIA's own recommendation frames a single node of **8x B200 ("approximately 1.5 TB aggregate HBM") as the practical minimum, fitting the BF16 weights plus KV cache with headroom, while H100, H200, GB200, and GB300 configurations of eight or more GPUs are described as requiring multi-node orchestration via Ray v2** (huggingface.co) (huggingface.co).

For the NVFP4 checkpoint, the requirement drops substantially. NVIDIA's build.nvidia.com model card, mirrored on the equivalent docs.api.nvidia.com reference page, lists a minimum of **4x GB200, 4x B200, 4x GB300, 4x B300, or 8x H100** (build.nvidia.com) (docs.api.nvidia.com), and its quick-start guide recommends "single-node: 4x B200 (fits NVFP4 weights plus KV cache with headroom)" as the practical baseline (build.nvidia.com). NVIDIA's TensorRT-LLM deployment guide corroborates this and explains why H100 needs twice as many GPUs as Blackwell-class chips for the same checkpoint: Hopper "lacks a native NVFP4 tensor-core GEMM," so the NVFP4 weights are dequantized on the fly through a W4A16 fallback path, which "requires a minimum of 8x H100 (fewer may suffice on the higher-memory H200) and delivers somewhat lower throughput than Blackwell" (nvidia.github.io) (nvidia.github.io).

NVIDIA's NIM (NVIDIA Inference Microservices) support matrix goes a level deeper, publishing the exact tensor-parallel (TP) and pipeline-parallel (PP) topology validated for every supported GPU and precision combination. On B200, for instance, "BF16 TP=8, PP=1 on a single node (8 GPUs); NVFP4 TP=2, TP=4, or TP=8 on a single node" are all validated configurations, while H100 in BF16 requires "TP=8, PP=2 across two nodes (16 GPUs total)" with inter-node InfiniBand or RoCE networking, and GB200 in BF16 needs "TP=4, PP=2 across two NVL trays of four GPUs" (docs.nvidia.com) (docs.nvidia.com) (docs.nvidia.com). *Table 2* below consolidates these NVIDIA-published minimums across GPU platforms and both checkpoints.

Table 2. Minimum validated GPU count by platform and precision (NVIDIA official specifications)

GPU Platform	Per-GPU Memory	BF16 Minimum GPUs	NVFP4 Minimum GPUs	Topology Notes
H100 (Hopper)	80 GB HBM3	16 (TP8, PP2, 2 nodes)	8 (TP8, PP1, single node, W4A16 fallback)	Hopper lacks native NVFP4 tensor cores; runs dequantized (nvidia.github.io)
H200	141 GB HBM3e	8 (TP8, PP1, single node)	4 or 8 (TP4 or TP8, single node)	Higher memory than H100 may allow fewer GPUs for NVFP4 (docs.nvidia.com)

GPU Platform	Per-GPU Memory	BF16 Minimum GPUs	NVFP4 Minimum GPUs	Topology Notes
B200	180 to 192 GB HBM3e	8 (TP8, PP1, single node)	4 (recommended minimum; 2 or 8 also validated)	NVIDIA's own recommended baseline is 4x B200; NVFP4 TP=2, TP=4, or TP=8 are all validated on a single node (docs.nvidia.com)
B300	288 GB HBM3e	4 or 8 (single node)	4 (recommended minimum; 2 or 8 also validated)	Matches B200 topology at higher per-GPU memory (nvidia.github.io)
GB200 (Grace Blackwell)	192 GB per GPU	8 (TP4, PP2, 2 NVL trays)	2 or 4 (single node)	Superchip pairs Grace CPU with Blackwell GPUs (docs.nvidia.com)
GB300 (Grace Blackwell Ultra)	288 GB per GPU	4 (single node)	2 or 4 (single node)	Same silicon family as DGX Station's GPU (docs.nvidia.com)

Table 2 makes the core planning tradeoff visible at a glance: choosing NVFP4 over BF16 roughly halves GPU count at every tier, and choosing Blackwell-class silicon (B200, B300, GB200, GB300) over Hopper (H100, H200) roughly halves it again for the same precision, since Hopper lacks native 4-bit tensor cores. NVIDIA's own storage guidance for the NIM container reinforces the scale involved even before considering compute: "the container image is approximately 38 GB, and the model cache ranges from approximately 330 GB for an NVFP4 profile to approximately 1.1 to 1.7 TB for a BF16 profile," meaning fast local NVMe storage sized in the multiple-terabyte range is itself a deployment prerequisite (docs.nvidia.com).

Independent community measurement broadly confirms NVIDIA's figures from the weights side. Model-hardware tracker llmrun.dev calculates that the 560.5-billion-parameter BF16 checkpoint "requires 369.9 GB of VRAM to load the model weights" at a 4-bit Q4_K_M community quantization, and recommends "481+ GB" once KV cache and framework overhead are added, concluding flatly that "no single GPU has enough memory" and multi-GPU or cluster setups are mandatory (llmrun.dev) (llmrun.dev) (llmrun.dev). The same tracker notes that only three known systems, the NVIDIA DGX H100, NVIDIA DGX A100 (640 GB configuration), and a 512 GB Mac Studio (M3 Ultra), have enough unified or aggregate memory to run even that reduced quantization, and estimates DGX H100 throughput around 47 tokens per second for the 4-bit build (llmrun.dev).

Multi-Node and Enterprise-Scale Deployment

Beyond a single 4x or 8x GPU node, enterprises running Nemotron 3 Ultra at scale, or fine-tuning it, typically move to multi-node clusters. NVIDIA's model card recommends Ray v2 as the multi-processing backend for these configurations, with the vLLM serving example spanning "2x 4x GB300" nodes at tensor-parallel size 4 and pipeline-parallel size 2, orchestrated through a `--distributed-executor-backend ray` flag documented identically on NVIDIA's NIM reference page (docs.api.nvidia.com). For H100-class hardware, NVIDIA's NIM

support matrix documents a 16-GPU, two-node BF16 topology (TP=8, PP=2) requiring "inter-node InfiniBand or RoCE networking" (docs.nvidia.com), underscoring that network fabric, not just GPU count, is part of the hardware bill of materials for BF16 deployments on older silicon.

NVIDIA's NIM container also documents per-GPU runtime tuning as a practical node configuration reference. For a 2-GPU B200 or GB200 NVFP4 profile, it recommends capping concurrency at `--max-num-seqs 64`; for the 8-GPU, 2-node H100 BF16 profile, `--max-num-seqs 512`; and for memory-tighter 8-GPU H200 BF16 and 4-GPU B300/GB300 BF16 profiles, it recommends lowering `--gpu-memory-utilization` to 0.98 or 0.99 respectively while capping `--max-model-len` at 131,072 tokens, trading part of the model's native 1M-token context ceiling for stability at that GPU count. NVIDIA notes that "profiles not listed in this table start successfully with default settings and do not require additional NIM_PASSTHROUGH_ARGS," meaning these tuned flags apply to the more memory-constrained configurations rather than as a universal requirement (docs.nvidia.com).

For serving software, NVIDIA validates three inference engines for Nemotron 3 Ultra: vLLM, SGLang (recommended container `lmsysorg/sclang:v0.5.11`), and TensorRT-LLM (`nvcr.io/nvidia/tensorrt-llm/release:1.3.0rc21`), each with distinct configuration flags for tensor parallelism, expert parallelism, KV cache dtype, and MTP speculative decoding (nvidia.github.io). NVIDIA's Dynamo framework additionally supports disaggregated serving, splitting prefill and decode across separate GPU pools; its published disaggregated 1P1D (one prefill, one decode) configuration on 8x B200 total (4 prefill plus 4 decode) measured "61.6 user output tok/s and 231.1 system output tok/s/GPU" on an agentic workload trace at concurrency 32, without MTP enabled (docs.nvidia.com).

Enterprises without in-house GPU infrastructure can reach the same topologies through managed cloud services, including Amazon SageMaker JumpStart, Google Cloud, Microsoft Foundry, Oracle Cloud, and a long list of specialized inference providers such as CoreWeave, Together AI, Fireworks AI, and Baseten. AWS's own announcement recommends provisioning "sufficient service quota for GPU instances (for example, `ml.p5en.48xlarge`, `ml.p5.48xlarge`, or `ml.g7e.48xlarge`)" before deployment, and cautions that "GPU instances like `ml.p5en.48xlarge` can cost several dollars per hour" and will continue billing until the SageMaker endpoint is explicitly deleted (aws.amazon.com) (aws.amazon.com). At the largest scale, GB200 NVL72 rack-scale systems are available directly from cloud providers: CoreWeave lists an on-demand GB200 instance (a four-GPU, two-Superchip unit) at **\$42.00 per hour**, equivalent to \$10.50 per GPU per hour, with the provider's documentation noting "each instance is comprised of 2 GB200 or GB300 Superchips, where each Superchip has 1 Grace CPU and 2 Blackwell GPUs" (coreweave.com) (coreweave.com). CoreWeave's comparable eight-GPU HGX H100 node, the older-generation topology H100 requires under NIM's BF16 matrix, lists at **\$49.24 per hour**, or \$6.16 per GPU per hour on-demand (coreweave.com).

API and Usage-Based Access: Inference Cost per Token

Not every organization needs to own or rent raw GPUs. Nemotron 3 Ultra is available as a pay-per-token API through OpenRouter, which aggregates several third-party inference providers. As of mid-July 2026, DeepInfra lists the model at **\$0.50 per million input tokens and \$2.20 per million output tokens**, Together AI at \$0.60

input and \$3.60 output, and Venice at \$0.625 input and \$3.125 output, all per million tokens ([openrouter.ai](#)). Once prompt-caching discounts are factored in, OpenRouter's trailing 7-day weighted-average pricing across all providers works out to \$0.308 per million input tokens and \$2.81 per million output tokens, reflecting a DeepInfra cache-hit rate of 49.1 percent and a Venice cache-hit rate of 88.0 percent ([openrouter.ai](#)). Best-in-class provider throughput on the aggregator was measured at 159 tokens per second with a 0.49-second median latency to first useful output ([openrouter.ai](#)).

It is possible to sanity-check self-hosted compute economics against these API rates using NVIDIA's own published throughput benchmarks. NVIDIA's Dynamo deployment recipe documentation measured a 4x B200 aggregated vLLM deployment (NVFP4 plus FP8 KV cache, MTP with one speculative token) at **310.8 system output tokens per second per GPU** on an agentic workload trace, and 201.4 tokens per second per GPU on a shorter chat-style trace ([docs.nvidia.com](#)). Combining that measured throughput with Lambda's published four-GPU B200 SXM6 instance pricing of \$6.79 per GPU per hour, or \$27.16 per hour for the full four-GPU node, yields an aggregate system throughput of 1,243.2 tokens per second and an implied compute-only cost of roughly **\$6.07 per million output tokens** at continuous full utilization; the same node running the lower-throughput chat trace (805.6 tokens per second aggregate) implies roughly **\$9.37 per million tokens** ([lambda.ai](#)). On 8x H200 (CoreWeave on-demand price of \$50.44 per hour for the full node), Dynamo's measured 27.4 tokens per second per GPU on the agentic trace implies roughly **\$63.93 per million tokens**, over an order of magnitude more expensive per token than the B200 configuration, illustrating how strongly GPU generation and quantization affect self-hosted unit economics ([docs.nvidia.com](#)) ([coreweave.com](#)).

Table 3 below summarizes these calculations alongside the OpenRouter API rates. All self-hosted figures are illustrative, derived by combining NVIDIA's own published per-GPU-hour throughput benchmarks with the cited hourly rental prices under an assumption of continuous, fully utilized operation; they exclude engineering, redundancy, storage, and networking overhead and should be read as directional rather than as a quote.

Table 3. Illustrative cost per million output tokens: self-hosted compute versus managed API (July 2026 data)

Deployment	Configuration	Measured System Throughput	Node Rental Cost	Implied Cost per 1M Output Tokens
Self-hosted, agentic workload	4x B200, NVFP4, MTP	310.8 tok/s/GPU (1,243.2 tok/s total)	\$27.16/hr (lambda.ai)	approximately \$6.07
Self-hosted, chat workload	4x B200, NVFP4, MTP	201.4 tok/s/GPU (805.6 tok/s total)	\$27.16/hr (lambda.ai)	approximately \$9.37
Self-hosted, agentic workload	8x H200, NVFP4, MTP	27.4 tok/s/GPU (219.2 tok/s total)	\$50.44/hr (coreweave.com)	approximately \$63.93
Managed API, list price	DeepInfra via OpenRouter	not applicable, per-token billing	not applicable	\$2.20 (list) (openrouter.ai)
Managed API, effective average	OpenRouter, all providers, 7-day weighted	not applicable, per-token billing	not applicable	\$2.81 (effective) (openrouter.ai)

The pattern in *Table 3* is counterintuitive to a common assumption that self-hosting is automatically cheaper than API access at scale. Even in the best-case, highest-throughput, fully-utilized self-hosted scenario (4x B200 on an agentic trace), the implied per-token compute cost still runs more than double the list price DeepInfra charges through OpenRouter, and roughly 2.16 times the platform's own effective blended rate. This gap exists because managed API providers pool demand from many simultaneous customers, achieving batching efficiency and hardware utilization that a single dedicated node running one workload rarely matches outside of a benchmark trace; it narrows or reverses only at high, sustained request volumes, or where data residency, customization, fine-tuning, or air-gapped requirements make self-hosting a functional necessity rather than a pure cost optimization.

Comparative Context and Market Positioning

Nemotron 3 Ultra's hardware demands should be read against its own performance claims and against sibling and rival models. NVIDIA reports that Nemotron 3 Ultra "achieves 5.9x, 4.8x, and 1.6x higher inference throughput compared to GLM-5.1-754B-A40B, Kimi-K2.6-1T-A32B, and Qwen-3.5-397B-17B respectively" on an 8,000-token input, 64,000-token output benchmark setting, while reaching "on-par accuracies compared to other state-of-the-art open LLMs across a diverse set of benchmarks" (research.nvidia.com) (research.nvidia.com). NVIDIA's developer blog separately claims a broader "5x higher throughput compared to other open models in its class," attributing this partly to the NVFP4 pretraining recipe, which "delivers up to 5x higher throughput per GPU at the same interactivity compared to BF16 on Blackwell," and partly to the model completing agentic benchmarks like SWE-bench using fewer total tokens, which NVIDIA states "lowers the cost for agentic tasks by up to 30%" relative to token-hungrier competitors (developer.nvidia.com) (developer.nvidia.com) (developer.nvidia.com). These are vendor-published figures rather than independently reproduced benchmarks and should be read with that caveat. An independent third-party benchmark aggregator tells a more mixed story: llmrun.dev's tracked LiveBench scores place the BF16 checkpoint's coding result at "71.3 (#42/75)," its math result at "54.5 (#72/75)," and its reasoning result at "37.5 (#70/75)" among tracked models, a meaningfully more middling placement than NVIDIA's own "on-par accuracies" framing suggests, and a useful reminder to weigh vendor-selected benchmark comparisons against independent leaderboards (llmrun.dev) (llmrun.dev) (llmrun.dev).

One recurring theme across community deployment reports is that NVFP4, not BF16, is the practical default. The same checkpoint runs across NVIDIA's Hopper, Blackwell, and even Ampere architectures using "specialized NVFP4 quantization kernels," per NVIDIA's developer blog, meaning a single quantized artifact is portable across a heterogeneous GPU fleet rather than requiring separate builds per architecture (developer.nvidia.com). Within the Nemotron 3 family itself, the "Super" variant, 120 billion total parameters with 12 billion active per NVIDIA's own count, offers a middle path for organizations that find Ultra's hardware floor prohibitive (build.nvidia.com): NVIDIA's own TensorRT-LLM guide lists Super's BF16 minimum at just 4x H100 or H200, or 2x B200/GB200, and its NVFP4 minimum at 2x H100/H200 or a single B200/GB200 (nvidia.github.io), and Super's official NIM model card lists a minimum requirement of "8x H100-80GB" for its own baseline configuration (build.nvidia.com). One Reddit user framed the tradeoff succinctly when discussing

enterprise adoption: "Nemotron specifically is for companies to fine tune and train," positioning it less as a chat assistant to run casually and more as a foundation for downstream, infrastructure-backed customization ([reddit.com](https://www.reddit.com)).

Data Analysis and Evidence

Aggregating the cloud GPU rental market provides useful bounds for budgeting a Nemotron 3 Ultra deployment. Tracking service getdeploying.com reports **48 providers offering H100** at prices ranging from \$0.34 to \$14.90 per GPU per hour, averaging \$3.76 per GPU per hour on-demand across 142 tracked on-demand listings (getdeploying.com) (getdeploying.com). The same tracker lists **34 providers for H200**, ranging from \$1.00 to \$13.78 per GPU per hour with an on-demand average of \$4.29 (getdeploying.com), and **25 providers for B200** across 113 tracked listings, ranging from \$2.69 to \$16.11 per GPU per hour with an on-demand average of \$6.97 (getdeploying.com) (getdeploying.com). The newer **B300** tier, tracked across 16 providers, ranges from \$3.27 to \$18.00 per GPU per hour (getdeploying.com). Reserved and spot pricing run meaningfully lower across all four tiers; on H100, for example, spot listings average \$1.68 per GPU per hour against a \$3.76 on-demand average (getdeploying.com).

Pricing trends over the trailing twelve months add helpful context for budgeting multi-year deployments. getdeploying.com reports that H100 on-demand pricing "has decreased by about 4% since July 2025, dropping from \$3.89 to \$3.72/hr per GPU," that B200 on-demand pricing "has decreased by about 7% since July 2025, dropping from \$7.69 to \$7.19/hr per GPU," while H200 on-demand pricing "has remained stable, averaging around \$4.36/hr per GPU" over the same window (getdeploying.com) (getdeploying.com) (getdeploying.com). The modest downward drift on Hopper and early Blackwell pricing suggests self-hosted compute costs for Nemotron 3 Ultra are more likely to improve than worsen over a typical multi-year hardware amortization window, though this trend applies to raw GPU rental rather than to the model-specific throughput improvements NVIDIA and third parties continue to publish for newer inference engine releases.

Applying these ranges to the GPU counts in *Table 2* produces the following approximate hourly cost bands for a single node running Nemotron 3 Ultra continuously. An 8x H100 NVFP4 node (the minimum viable H100 configuration) costs roughly \$16 to \$119 per hour at average on-demand and peak listed prices respectively (8 GPUs times \$3.76 to \$14.90 per hour), broadly consistent with CoreWeave's own listed HGX H100 node price of \$49.24 per hour (coreweave.com). A 4x B200 NVFP4 node, NVIDIA's own recommended baseline, costs roughly \$27 to \$64 per hour (4 GPUs times \$6.79 published by Lambda, or up to the \$16.11 peak market rate). Lambda's own multi-GPU "1-Click Cluster" product, aimed at production deployments rather than single-node testing, prices a 16-GPU HGX B200 cluster at \$9.86 per GPU per hour, dropping to \$8.87 per GPU per hour at 256 or more GPUs, reflecting the volume discounts typical of committed-capacity cloud contracts (lambda.ai).

Community-reported hardware audits add a valuable reality check to vendor specifications, since they surface configurations that technically load the model but perform poorly. The llmrun.dev tracker's benchmark for a Q4_K_M quantization on a DGX A100 (640 GB configuration) estimated roughly 29 tokens per second, while a 512 GB Mac Studio (M3 Ultra) managed only around 2 tokens per second at the same quantization, illustrating that raw memory capacity alone does not guarantee usable throughput; memory bandwidth and interconnect matter just as much as total capacity (llmrun.dev) (llmrun.dev). This mirrors NVIDIA's own benchmark data in *Table 3*, where 8x H200 delivered dramatically lower per-GPU throughput (27.4 tokens per second per GPU on

the agentic trace) than 4x B200 (310.8 tokens per second per GPU) despite H200 offering more total GPUs and more total memory, a gap attributable to Blackwell's native NVFP4 tensor cores and tighter tensor-parallel topology at TP4 versus TP8, as detailed in the API and Usage-Based Access section above.

Multi-Token Prediction itself accounts for a meaningful share of Blackwell's throughput advantage, and its benefit scales with GPU generation. On the same 4x B200 agentic configuration, disabling MTP drops system throughput from 310.8 to 175.9 tokens per second per GPU, while on the chat trace it falls more modestly from 201.4 to 181.3 tokens per second per GPU; on 8x H200, the MTP-off penalty is smaller in absolute terms, from 27.4 to 26.5 tokens per second per GPU on the agentic trace and from 46.8 to 43.0 on the chat trace (docs.nvidia.com) (docs.nvidia.com). The pattern suggests MTP speculative decoding delivers its largest wins on already-fast Blackwell hardware and a smaller, though still positive, boost on Hopper, reinforcing that speculative decoding and quantization are complementary rather than substitute levers for hardware planning.

Case Studies and Real-World Examples

Amazon SageMaker JumpStart: Managed Enterprise Deployment

AWS added one-click deployment support for Nemotron 3 Ultra to Amazon SageMaker JumpStart shortly after the model's June 2026 release, targeting enterprises that want managed infrastructure without directly provisioning GPUs. AWS's own guidance requires customers to confirm "sufficient service quota for GPU instances (for example, ml.p5en.48xlarge, ml.p5.48xlarge, or ml.g7e.48xlarge)" before deployment, all of which are 8-GPU H100 or H200-class instances aligned with the single-node minimums NVIDIA specifies (aws.amazon.com). Because SageMaker bills for endpoint uptime rather than per-token usage, AWS explicitly warns customers to "remember to delete your endpoint when finished to avoid ongoing charges," a reminder that reflects how easily an idle 8-GPU H100 or H200 endpoint, at several dollars per GPU per hour, can accumulate unplanned cost (aws.amazon.com).

CoreWeave GB200 NVL72: Rack-Scale Inference Capacity

CoreWeave offers on-demand GB200 NVL72 instances specifically suited to the model's Grace Blackwell-class minimums. Each billing unit pairs two GB200 Superchips, providing four Blackwell GPUs, at \$42.00 per hour on-demand, or \$10.50 per GPU per hour, positioning it between B200 (\$6.97 average on-demand per getdeploying.com) and B300 (\$3.27 to \$18.00 range) in per-GPU cost, while offering the tighter NVLink-C2C coherent CPU-GPU memory architecture that NIM's support matrix validates for both 2-GPU and 4-GPU NVFP4 topologies, per this report's Single-Node Datacenter Deployment section above (coreweave.com).

DGX Spark Four-Node Cluster: The Documented Prosumer Path

NVIDIA's own cookbook repository documents a specific, reproducible path for running Nemotron 3 Ultra outside a traditional datacenter: a four-node DGX Spark cluster connected over RoCE (RDMA over Converged Ethernet) networking, using vLLM with tensor parallelism spread across the cluster and MTP speculative decoding enabled, benchmarked with NVIDIA's own AIPerf tooling (docs.nvidia.com). At \$4,699 per unit, the

complete four-node hardware cost of \$18,796 sits well below even a single datacenter-grade H100 8-pack purchased outright, though it requires careful RoCE network configuration and accepts throughput well below what a dedicated 4x B200 node delivers. This is a real, NVIDIA-documented deployment target rather than a hypothetical configuration, and it represents the closest thing to a "prosumer" path NVIDIA officially supports for the full Ultra checkpoint.

DGX Station with MoE Expert Offloading: Single-Box Deployment

A newer entry in NVIDIA's cookbook catalog, published July 6, 2026, documents deploying Nemotron 3 Ultra on a single GB300-based DGX Station using vLLM's coherent CPU memory support for "selective MoE expert offloading" together with a FlashInfer TensorRT-LLM NVFP4 backend (docs.nvidia.com). This is architecturally distinct from the multi-GPU node configurations elsewhere in this report: rather than sharding the model's experts across multiple GPUs, DGX Station keeps active experts in its 252 GB of GPU HBM3e and swaps inactive ones in from its 496 GB of CPU memory over a 900 GB/s NVLink-C2C link, a technique made viable specifically because Nemotron 3 Ultra activates only 55 of its 550 billion parameters per token (nvidia.com). It represents the single-box, non-clustered alternative to the DGX Spark cluster approach, at roughly five times the capital cost but without the RoCE networking complexity of a multi-node setup.

Independent Consumer Hardware Testing: The Strix Halo Reality Check

Outside NVIDIA's own reference architectures, independent hobbyists have systematically tested what Nemotron 3 Ultra can and cannot do on affordable unified-memory hardware. One widely discussed Reddit post benchmarked a Beelink GTR9 Pro (AMD Ryzen AI MAX+ 395, 128 GB unified memory) and, after initially misjudging which Nemotron 3 variant was runnable, concluded that the 550-billion-parameter Ultra checkpoint has no direct GGUF or llama.cpp path at that memory tier, while the 120-billion-parameter Super variant ran directly via llama.cpp's Vulkan backend at "292.51 pp512 / 17.94 tg128" tokens per second (prompt processing and generation, respectively) (reddit.com). The thread's most upvoted correction, from a different commenter, stated plainly that "Ultra-size models are not meant to run on 128GB devices," a conclusion this report's independent research corroborates: even NVIDIA's own smallest officially documented path (the 512 GB aggregate DGX Spark cluster) requires four times that memory footprint (reddit.com).

Implications and Future Directions

The clearest implication of Nemotron 3 Ultra's hardware profile is that NVFP4 quantization has effectively become the default deployment path rather than an optional optimization. Every officially documented hardware minimum this report identified, whether NVIDIA's own model card, its TensorRT-LLM guide, its NIM support matrix, or its DGX Station and DGX Spark cookbooks, treats the NVFP4 checkpoint as the primary target and BF16 as a secondary, roughly-double-the-hardware option intended mainly for research, fine-tuning baselines, and cases where maximum numerical precision is worth the added GPU count (build.nvidia.com). As more inference software (vLLM, SGLang, TensorRT-LLM, and Dynamo were all validated with NVFP4 kernels at launch) matures its low-precision kernel support, the gap between BF16 and NVFP4 requirements is

unlikely to narrow; if anything, future Nemotron releases seem likely to ship NVFP4-native by default, following the same "pretrained in NVFP4" approach NVIDIA already used for Nemotron 3 Ultra itself rather than quantizing after the fact (research.nvidia.com).

A second implication concerns the growing bifurcation between datacenter-scale and prosumer-scale hardware paths. NVIDIA's decision to publish official cookbooks for both a four-node DGX Spark cluster and a single DGX Station suggests the company sees meaningful demand for below-hyperscale, above-consumer deployment options, even for a model this large. That said, the community testing summarized in this report's case studies indicates the gap between "documented as possible" and "genuinely convenient" remains wide: a four-node Spark cluster or single Station is a five- or six-figure capital commitment requiring specialized configuration (RoCE networking, MoE expert offloading tuning) that most individual developers and small teams will not undertake. As one Reddit commenter put it while discussing why enterprises might still prefer larger, harder-to-run open models over smaller ones that fit on a single workstation, "they need better models than what will fit on things like 128gb of ram," a framing that suggests managed API access will likely remain the default on-ramp for the majority of Nemotron 3 Ultra's eventual users even as local hardware technically catches up (reddit.com).

Finally, the cost calculations in this report's Table 3 suggest that self-hosted infrastructure investment is best justified by factors other than raw per-token cost, at least at the deployment scales examined. Data residency, regulatory or contractual requirements against sending data to third-party inference providers, the ability to fine-tune or continually post-train the model on proprietary data using NVIDIA's published GRPO and MOPD recipes, and sustained, highly concurrent request volumes that push utilization well above what a single-tenant benchmark trace captures are all more defensible reasons to self-host than an assumption that owning hardware is inherently cheaper. As GPU rental prices and API prices both continue to shift, organizations evaluating Nemotron 3 Ultra should treat the \$6 to \$64 per million token range calculated here as a starting framework to re-run against their own actual concurrency and utilization patterns rather than as a fixed number.

Frequently Asked Questions (FAQs)

What is Nemotron 3 Ultra's total parameter count? Nemotron 3 Ultra has 550 billion total parameters, with 55 billion active per token due to its Mixture-of-Experts architecture (huggingface.co).

What is the minimum GPU memory required to run Nemotron 3 Ultra? There is no single-GPU answer; the model requires multi-GPU configurations regardless of quantization. The NVFP4 checkpoint occupies approximately 352 GB on disk and needs a minimum of four Blackwell-class GPUs (B200, B300, GB200, or GB300) or eight H100s; the BF16 checkpoint occupies over 1.1 TB and needs a minimum of eight Blackwell-class GPUs, sixteen H100s, or eight H200s, as detailed in *Table 2* above (docs.api.nvidia.com) (huggingface.co).

How many GPUs does Nemotron 3 Ultra need? As few as four B200, B300, GB200, or GB300 GPUs for the recommended NVFP4 single-node deployment, up to sixteen H100s for a multi-node BF16 deployment, per NVIDIA's own published minimums summarized in *Table 2* of this report.

Can Nemotron 3 Ultra run on a single GPU or a consumer graphics card? No. Even the most memory-rich single consumer GPUs on the market in 2026 fall far short; llmrun.dev's compatibility checker confirms an NVIDIA GeForce RTX 5090, with 32 GB of VRAM, "requires at least 169.6 GB at IQ2_XXS," meaning even the most aggressive practical quantization exceeds consumer GPU memory by more than five times (llmrun.dev).

Can Nemotron 3 Ultra run on a DGX Spark or other unified-memory device? A single DGX Spark (128 GB) cannot; NVIDIA's own cookbook requires a four-node DGX Spark cluster (512 GB aggregate) for the NVFP4 checkpoint (docs.nvidia.com). A single DGX Station (748 GB coherent memory) can, using MoE expert offloading between GPU HBM3e and CPU LPDDR5X memory (docs.nvidia.com).

Does the 1M-token context window affect hardware requirements? Yes. NVIDIA's NIM documentation notes the model's native context defaults to 262,144 tokens and that extending to the full 1,048,576-token window "requires significantly more KV cache memory, which reduces the number of requests that can be served concurrently," making effective context length a tunable tradeoff against concurrent throughput on a fixed GPU count (docs.nvidia.com).

Which inference engines support Nemotron 3 Ultra? NVIDIA validates vLLM, SGLang, and TensorRT-LLM for Nemotron 3 Ultra, alongside its own Dynamo framework for disaggregated serving; all three general-purpose engines shipped NVFP4 kernel support at the model's June 2026 launch, each with its own tensor-parallel, expert-parallel, and speculative-decoding configuration flags (nvidia.github.io).

How much does it cost to run Nemotron 3 Ultra via API? As of July 2026, list pricing on OpenRouter starts at \$0.50 per million input tokens and \$2.20 per million output tokens through DeepInfra, with a platform-wide effective average (after caching, trailing 7 days) of roughly \$0.308 input and \$2.81 output per million tokens (openrouter.ai).

What does it cost to self-host Nemotron 3 Ultra? This report's illustrative calculations, combining NVIDIA's published throughput benchmarks with published cloud GPU rental prices, put compute-only self-hosted costs between roughly \$6 and \$64 per million output tokens depending on GPU generation and workload type, generally higher than managed API pricing unless utilization is sustained and high (see *Table 3*).

What is the difference in hardware requirements between the BF16 and NVFP4 checkpoints? NVFP4 roughly halves the minimum GPU count at every tier compared to BF16, because it uses 4-bit rather than 16-bit weights; NVIDIA's TensorRT-LLM guide notes the NVFP4 checkpoint on Hopper GPUs still requires double the GPU count of Blackwell because Hopper lacks native 4-bit tensor cores and must dequantize on the fly (nvidia.github.io).

How much local storage does deploying Nemotron 3 Ultra require? NVIDIA's NIM documentation states the model cache alone ranges from approximately 330 GB (NVFP4) to 1.1 to 1.7 TB (BF16), on top of a roughly 38 GB container image (docs.nvidia.com).

Conclusion

Nemotron 3 Ultra's hardware requirements scale directly with the choices an organization makes about precision and deployment tier, and this report's central finding is that those choices carry roughly a two-times swing in minimum GPU count and an order-of-magnitude swing in implied per-token compute cost. The

NVFP4 checkpoint on modern Blackwell-class GPUs, four B200, B300, GB200, or GB300 units in a single node, represents NVIDIA's own recommended baseline and the most GPU-efficient path to running the full model; the BF16 checkpoint, older Hopper-generation hardware, or higher-GPU-count topologies each add substantial cost and complexity without necessarily improving output quality for most production workloads. For teams without dedicated infrastructure, managed API access through OpenRouter's aggregated providers, AWS SageMaker JumpStart, or the model's other cloud partners remains, based on this report's cost comparison, generally more economical per token than self-hosting unless request volume is both large and sustained. For teams that do need local or on-premises deployment, whether for data residency, fine-tuning, or air-gapped requirements, NVIDIA has published two concrete, tested hardware paths below full datacenter scale: a four-node DGX Spark cluster at roughly \$18,800 in hardware, or a single DGX Station at roughly \$94,000 and up, each suited to slightly different operational tradeoffs between networking complexity and single-box simplicity. Given how quickly cloud GPU pricing, quantization tooling, and NVIDIA's own reference architectures are moving in 2026, teams evaluating Nemotron 3 Ultra should treat every dollar figure in this report as a snapshot to be re-verified against current vendor pricing pages at the time of an actual deployment decision, rather than a permanent benchmark.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16/blob/main/README.md>
2. <https://research.nvidia.com/labs/nemotron/Nemotron-3-Ultra/>
3. <https://build.nvidia.com/nvidia/nemotron-3-ultra-550b-a55b/modelcard>
4. <https://gpusmith.com/articles/llm-inference-hardware-sizing-guide>
5. <https://llmrun.dev/model/nvidia-nvidia-nemotron-3-ultra-550b-a55b-bf16>
6. https://www.reddit.com/r/LocalLLaMA/comments/1twwycb/nemotron_3_ultra_reality_check_no_onebox_128gb/
7. <https://gpusmith.com/articles/gpu-cloud-rental-prices-h100-h200>
8. <https://gpusmith.com/articles/nvidia-ai-gpu-price-index-trends>
9. <https://getdeploying.com/gpus/nvidia-h100>
10. <https://getdeploying.com/gpus/nvidia-b200>
11. <https://docs.nvidia.com/dynamo/dev/recipes/nemotron-3-ultra>
12. <https://gpusmith.com/articles/own-vs-rent-gpus-tco-comparison>
13. <https://gpusmith.com/articles/llm-cost-per-million-tokens-benchmark>
14. <https://openrouter.ai/nvidia/nemotron-3-ultra-550b-a55b>
15. <https://gpusmith.com/articles/nvidia-dgx-spark-review-specs-performance>
16. <https://marketplace.nvidia.com/en-us/enterprise/personal-ai-supercomputers/dgx-spark/>
17. <https://www.exxactcorp.com/Exxact-VWS-158270643-E158270643>
18. <https://docs.nvidia.com/nemotron/nightly/usage-cookbook/Nemotron-3-Ultra-Base/README.html>
19. <https://nvidia.github.io/TensorRT-LLM/latest/deployment-guide/deployment-guide-for-nemotron-3-on-trtllm.html>
20. <https://developer.nvidia.com/blog/nvidia-nemotron-3-ultra-powers-faster-more-efficient-reasoning-for-long-running-agents/>

21. <https://www.nvidia.com/en-us/products/workstations/dgx-station/>
22. <https://www.servethehome.com/nvidia-dgx-station-systems-available-at-last-gb300-gb200-workstations-for-your-desktop/>
23. https://www.reddit.com/r/unsloth/comments/1twsd68/nvidia_nemotron_3_ultra_is_out_now/
24. <https://docs.api.nvidia.com/nim/reference/nvidia-nemotron-3-ultra-550b-a55b>
25. <https://docs.nvidia.com/nim/large-language-models/2.0.6/day-0/support-matrix-day-0.html>
26. <https://docs.nvidia.com/nim/large-language-models/2.0.6/day-0/get-started-nemotron-3-ultra.html>
27. <https://aws.amazon.com/blogs/machine-learning/nvidia-nemotron-3-ultra-now-available-on-amazon-sagemaker-jumpstart/>
28. <https://www.coreweave.com/pricing>
29. <https://lambda.ai/pricing>
30. <https://build.nvidia.com/nvidia/nemotron-3-super-120b-a12b/modelcard>
31. <https://getdeploying.com/gpus/nvidia-h200>
32. <https://getdeploying.com/gpus/nvidia-b300>

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the [hardware library](#), [GPU references](#), [networking references](#) and [vendor research](#). [Contact GPU Smith](#) with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: <https://gpusmith.com>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.