NVIDIA Nemotron 3 Ultra Hardware Requirements: VRAM and Cost

Published July 15, 2026 36 min read

Executive Summary

NVIDIA's **Nemotron 3 Ultra** (formally **NVIDIA-Nemotron-3-Ultra-550B-A55B**) is a 550-billion-parameter Mixture-of-Expe

At the weight level, community benchmarking site llmrun.dev measures the BF16 checkpoint (560.5B parameters counting embe

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 H2

This report walks through every documented deployment tier for Nemotron 3 Ultra: local and workstation-class hardware, si

Introduction and Background

Nemotron 3 Ultra sits at the top of NVIDIA's third-generation Nemotron family of open-weight language models, released al

The practical consequence of this architecture is that Nemotron 3 Ultra behaves very differently from a same-sized dense

NVIDIA's own sources describe pretraining scale in two slightly different ways worth noting honestly rather than glossing

Because the weights, datasets, and fine-tuning recipes are all open, Nemotron 3 Ultra is being evaluated by a wide spectr
```Those two audiences face almost entirely different hardware and cost questions, and this report addresses both, workin

# 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 cons

**DGX Spark**, NVIDIA's smallest AI desktop appliance, pairs a GB10 Grace Blackwell Superchip with 128 GB of coherent uni

**DGX Station** takes a different approach: a single deskside unit built around one GB300 Grace Blackwell Ultra Desktop S

Below the DGX tier, the model becomes genuinely impractical. Community quantizer Unsloth published dynamic GGUF quantizat

Despite the hardware barrier, community interest in local quantizations has been substantial. llmrun.dev's tracker shows

_Table 1_ below breaks out the full community GGUF quantization ladder for the BF16 checkpoint as measured by llmrun.dev,

**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
| **Q4\_K\_M** (recommended default) | 369.9 GB | 336.31 GB | "Offers the best balance of quality and VRAM usage" (Source
| **Q5\_K\_S** (higher quality) | 423.9 GB | not separately published | "Provides better quality if you have the VRAM" (S
| **BF16** (full precision) | 1,233.2 GB | 1,121.05 GB | Full-precision baseline; "the full-precision BF16 version is 112

No tier in _Table 1_ fits inside a single consumer or most professional GPUs. llmrun.dev's own compatibility checker stat

# 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 GP

For the BF16 checkpoint, NVIDIA's Hugging Face model card lists a minimum of **8x GB200, B200, GB300, or B300; 16x H100;
```

For the NVFP4 checkpoint, the requirement drops substantially. NVIDIA's build.nvidia.com model card, mirrored on the equi

NVIDIA's NIM (NVIDIA Inference Microservices) support matrix goes a level deeper, publishing the exact tensor-parallel (T

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 nati
H200	141 GB HBM3e	8 (TP8, PP1, single node)	4 or 8 (TP4 or TP8, single node)	Higher memory than H100 may al
B200	180 to 192 GB HBM3e	8 (TP8, PP1, single node)	4 (recommended minimum; 2 or 8 also validated)	NVIDIA's
B300	288 GB HBM3e	4 or 8 (single node)	4 (recommended minimum; 2 or 8 also validated)	Matches B200 topology
GB200 (Grace Blackwell)	192 GB per GPU	8 (TP4, PP2, 2 NVL trays)	2 or 4 (single node)	Superchip pairs Grace
GB300 (Grace Blackwell Ultra)	288 GB per GPU	4 (single node)	2 or 4 (single node)	Same silicon family as DG

Table 2 makes the core planning tradeoff visible at a glance: choosing NVFP4 over BF16 roughly halves GPU count at ever

Independent community measurement broadly confirms NVIDIA's figures from the weights side. Model-hardware tracker llmrun.

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 mu

NVIDIA's NIM container also documents per-GPU runtime tuning as a practical node configuration reference. For a 2-GPU B20

For serving software, NVIDIA validates three inference engines for Nemotron 3 Ultra: vLLM, SGLang (recommended container

Enterprises without in-house GPU infrastructure can reach the same topologies through managed cloud services, including A

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 OpenRo

It is possible to sanity-check self-hosted compute economics against these API rates using NVIDIA's own published through

Table 3 below summarizes these calculations alongside the OpenRouter API rates. All self-hosted figures are illustrativ

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 (Source: [lamb
Self-hosted, chat workload	4x B200, NVFP4, MTP	201.4 tok/s/GPU (805.6 tok/s total)	\$27.16/hr (Source: [lambda.ai
Self-hosted, agentic workload	8x H200, NVFP4, MTP	27.4 tok/s/GPU (219.2 tok/s total)	\$50.44/hr (Source: [corewea
Managed API, list price	DeepInfra via OpenRouter	not applicable, per-token billing	not applicable	\$2.20 (list)
Managed API, effective average	OpenRouter, all providers, 7-day weighted	not applicable, per-token billing	not ap

The pattern in _Table 3_ is counterintuitive to a common assumption that self-hosting is automatically cheaper than API a

Comparative Context and Market Positioning

Nemotron 3 Ultra's hardware demands should be read against its own performance claims and against sibling and rival model

One recurring theme across community deployment reports is that NVFP4, not BF16, is the practical default. The same check

Data Analysis and Evidence

Aggregating the cloud GPU rental market provides useful bounds for budgeting a Nemotron 3 Ultra deployment. Tracking serv

Pricing trends over the trailing twelve months add helpful context for budgeting multi-year deployments. getdeploying.com

Applying these ranges to the GPU counts in _Table 2_ produces the following approximate hourly cost bands for a single no

Community-reported hardware audits add a valuable reality check to vendor specifications, since they surface configuratio

Multi-Token Prediction itself accounts for a meaningful share of Blackwell's throughput advantage, and its benefit scales

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

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

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 tradition

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 GB30

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

Implications and Future Directions

The clearest implication of Nemotron 3 Ultra's hardware profile is that NVFP4 quantization has effectively become the def

A second implication concerns the growing bifurcation between datacenter-scale and prosumer-scale hardware paths. NVIDIA'

Finally, the cost calculations in this report's Table 3 suggest that self-hosted infrastructure investment is best justif

Frequently Asked Questions (FAQs)

What is Nemotron 3 Ultra's total parameter count? Nemotron 3 Ultra has 550 billion total parameters, with 55 billion

What is the minimum GPU memory required to run Nemotron 3 Ultra? There is no single-GPU answer; the model requires mu

```
**How many GPUs does Nemotron 3 Ultra need?** As few as four B200, B300, GB200, or GB300 GPUs for the recommended NVFP4 s

**Can Nemotron 3 Ultra run on a single GPU or a consumer graphics card?** No. Even the most memory-rich single consumer G

**Can Nemotron 3 Ultra run on a DGX Spark or other unified-memory device?** A single DGX Spark (128 GB) cannot; NVIDIA's

**Does the 1M-token context window affect hardware requirements?** Yes. NVIDIA's NIM documentation notes the model's nati

**Which inference engines support Nemotron 3 Ultra?** NVIDIA validates vLLM, SGLang, and TensorRT-LLM for Nemotron 3 Ultr

**How much does it cost to run Nemotron 3 Ultra via API?** As of July 2026, list pricing on OpenRouter starts at \$0.50 p

**What does it cost to self-host Nemotron 3 Ultra?** This report's illustrative calculations, combining NVIDIA's publishe

**What is the difference in hardware requirements between the BF16 and NVFP4 checkpoints?** NVFP4 roughly halves the mini

**How much local storage does deploying Nemotron 3 Ultra require?** NVIDIA's NIM documentation states the model cache alo

# Conclusion

Nemotron 3 Ultra's hardware requirements scale directly with the choices an organization makes about precision and deploy
```

Tags: 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