NVIDIA RTX PRO 6000 Blackwell Server Edition

Also known as: RTX 6000 Blackwell Server Edition · RTX PRO 6000 BSE · RTX PRO 6000 Blackwell SE

The RTX PRO 6000 Blackwell Server Edition is NVIDIA's passively cooled, PCIe form factor Blackwell part built for standard server chassis rather than SXM baseboards. It carries 96 GB of GDDR7 on a 512 bit bus, no NVLink, and a 600 W ceiling, so it drops into off the shelf 2U to 6U rackmount servers from Dell, HPE, Cisco, Lenovo, and Supermicro at up to eight cards per node instead of requiring a liquid cooled SXM baseboard and NVLink fabric. This is the class of part GPU Smith specs for single node and small cluster private inference builds under about eight GPUs, where the memory capacity and FP4 throughput matter more than multi-node interconnect.

96 GB GDDR7 ECC Memory · 512-bit bus	**1,597 GB/s** Memory bandwidth · Server Edition rating	**4 PFLOPS** FP4 tensor · dense, 5th-gen Tensor Cores
2 PFLOPS FP8 tensor · dense	**600 W** TDP · configurable down by OEM	**Dual-slot FHFL, passive** Form factor · PCIe 5.0 x16, no NVLink

COMPUTE	01/05
CUDA cores	24,064
Tensor cores	752 · 5th generation
RT cores	188 · 4th generation
FP32	120 TFLOPS
TF32 tensor	234 TFLOPS · dense
FP16 / BF16 tensor	1 PFLOPS · dense
FP8 tensor	2 PFLOPS · dense
FP4 tensor	4 PFLOPS · dense, Transformer Engine
RT core throughput	355 TFLOPS

MEMORY	02/05
Capacity	96 GB GDDR7 · ECC
Bus width	512-bit
Bandwidth	1,597 GB/s
MIG	Up to 4 instances · 24 GB each, isolated compute and QoS per instance

INTERCONNECT & FORM FACTOR	03/05
PCIe	Gen 5 x16
NVLink	Not supported
Dimensions	4.4 in x 10.5 in · dual-slot, full-height full-length
Alternate SKU	Single-slot FHXL, liquid-cooled · for higher rack density
Video engines	4x NVENC (9th gen), 4x NVDEC (6th gen), 4x JPEG

POWER & THERMAL	04/05
Max board power	600 W
Cooling	Passive · requires qualified server chassis airflow
Rack density	Up to 8 GPUs per node · 6U reference design, e.g. Aivres KR6268-X3

PLATFORM & SOFTWARE	05/05
Software stack	NVIDIA AI Enterprise · LLM inference, agents, generative AI
OEM qualification	Dell, HPE, Cisco, Lenovo, Supermicro · and other NVIDIA partner-ecosystem server vendors
Availability start	May 2025 · announced GTC March 2025
Inference positioning	Up to 5x LLM inference throughput vs L40S · NVIDIA claim, agentic AI workloads

PROCUREMENT

Typical lead time

7 days to 3 weeks from channel stock at resellers such as Central Computer and ServerSupply for single cards; multi-GPU OEM server builds run longer depending on chassis backlog (reported).

Pricing

Reported street pricing roughly $11,000 to $15,000 per card as of July 2026 (NVIDIA Marketplace approximately $13,250, PNY channel approximately $11,360, Newegg $13,400 to $15,000, refurbished units $9,500 to $11,000). Prices have risen since the March 2025 launch, attributed to a GDDR7 memory shortage (reported).

- Sold through normal server OEM and distribution channels (Dell, HPE, Cisco, Lenovo, Supermicro, PNY, CDW), not allocation-gated the way SXM/HGX parts are.
- Passive Server Edition cards need a chassis NVIDIA or the OEM has actually qualified for the airflow curve; they will not run correctly in an arbitrary unqualified box.
- A single-slot, liquid-cooled FHXL SKU exists for teams that want higher GPU density per rack unit than the air-cooled dual-slot card allows.
- GDDR7 supply constraints are the reported driver behind rising 2026 street prices; budget for volatility on multi-unit orders.

FIELD NOTES

- No NVLink on this card: cross-GPU traffic for tensor-parallel serving rides PCIe 5.0 and the host, not a GPU-to-GPU fabric. Fine for pipeline-parallel or one-model-per-GPU serving, a real bottleneck if you try to tensor-parallel-split a large model across many cards in one node.
- 96 GB on a single card is the actual advantage over the L40S, not raw FLOPS: a lot of 70B-class models fit in FP8 on one GPU with zero parallelism, which simplifies the whole serving stack.
- MIG partitioning (up to 4 x 24 GB, isolated QoS) is a genuinely useful multi-tenant knob the L40S does not have: run several isolated small-model endpoints on one card instead of a single underutilized process.
- VRAM per dollar beats HBM/SXM parts (H100, H200, B200) by a wide margin, at the cost of no NVLink and lower per-GPU compute. Right part for sub-8-GPU inference-only nodes, wrong part for multi-node pretraining.
- Passive cooling means this card only belongs in a chassis with a validated airflow curve for it; do not assume it drops into an arbitrary 2U box the way an actively cooled workstation card would.

QUESTIONS WE GET ON THIS PART

RTX PRO 6000 Server Edition vs H100 for inference?

The Server Edition has more memory (96 GB vs 80 GB) and lower street price than an H100, and NVIDIA's own benchmarks show it ahead of H100 SXM on single-GPU LLM throughput in some configurations. It gives up NVLink and HBM bandwidth, so H100/H200 SXM nodes still win on multi-GPU tensor-parallel training and very large batch serving.

RTX PRO 6000 vs L40S for inference?

NVIDIA claims up to 5x higher LLM inference throughput on the Server Edition versus the L40S, driven by double the memory (96 GB vs 48 GB), higher bandwidth GDDR7, and FP4 support the L40S lacks. For teams already running L40S fleets, the Server Edition is the straightforward generational upgrade path, not a different product category.

How much does the RTX PRO 6000 Blackwell Server Edition cost?

Reported street pricing in mid-2026 runs roughly $11,000 to $15,000 per card depending on channel (PNY, CDW, NVIDIA Marketplace, Newegg), up from the low-$9,000s range shortly after launch, attributed to a GDDR7 memory shortage. Refurbished units trade around $9,500 to $11,000.

Does it need special server hardware?

Yes. The Server Edition ships as a passively cooled card that depends entirely on host chassis airflow, so it only runs correctly in a server platform NVIDIA or the OEM has qualified for it, such as Dell, HPE, Cisco, Lenovo, or Supermicro systems. It is not a drop-in for an arbitrary desktop or an unqualified 2U chassis.

REFERENCES

[1] https://www.nvidia.com/en-us/data-center/rtx-pro-6000-blackwell-server-edition/

[2] https://lenovopress.lenovo.com/lp2263-thinksystem-nvidia-rtx-pro-6000-blackwell-server-edition-pcie-gen5-gpu

[3] https://blogs.nvidia.com/blog/rtx-pro-6000-blackwell-server-edition/

[4] https://gpupoet.com/gpu/learn/card/nvidia-rtx-pro-6000-blackwell-server

[5] https://www.thundercompute.com/blog/nvidia-rtx-pro-6000-pricing

[6] https://www.newegg.com/nvidia-900-2g153-0000-000-rtx-pro-6000-96gb-graphics-card/p/N82E16814132104

[7] https://resources.arccompute.io/hubfs/Data%20Sheets/Datasheet%20-%20NVIDIA%20RTX%20PRO%206000%20Server%20Edition%20(Aivres).pdf