NVIDIA L40S

Also known as: NVIDIA L40S PCIe · PNY Quadro L40S

The L40S is NVIDIA's Ada Lovelace generation, air cooled PCIe accelerator: 48 GB of GDDR6 with ECC, 864 GB/s of bandwidth, no NVLink, and a 350 W ceiling in a standard dual-slot server card. It predates the Blackwell RTX PRO 6000 Server Edition by two years and remains in active production and wide OEM and channel distribution. It is the default air cooled PCIe workhorse GPU Smith specs into single node and small cluster inference builds under about eight GPUs when a model comfortably fits in 48 GB and budget or lead time rules out the newer, pricier Blackwell part.

48 GB GDDR6 ECC Memory	**864 GB/s** Memory bandwidth	**733 TFLOPS** FP8 tensor · dense, 1,466 TFLOPS with sparsity
350 W TDP	**Dual-slot, passive** Form factor · 4.4 in x 10.5 in	**PCIe Gen4 x16** Interconnect · no NVLink

COMPUTE	01/05
CUDA cores	18,176
Tensor cores	568 · 4th generation
RT cores	142 · 3rd generation
FP32	91.6 TFLOPS
TF32 tensor	183 TFLOPS · 366 TFLOPS with sparsity
FP16 / BF16 tensor	362 TFLOPS · 733 TFLOPS with sparsity
FP8 tensor	733 TFLOPS · 1,466 TFLOPS with sparsity
RT core throughput	209 TFLOPS

MEMORY	02/05
Capacity	48 GB GDDR6 · ECC
Bandwidth	864 GB/s
MIG	Not supported

INTERCONNECT & FORM FACTOR 03/05

PCIe	Gen4 x16 · 64 GB/s bidirectional
NVLink	Not supported
Dimensions	4.4 in x 10.5 in · dual-slot
Power connector	16-pin
Video engines	3x NVENC, 3x NVDEC · includes AV1 encode and decode
Display outputs	4x DisplayPort 1.4a

POWER & THERMAL 04/05

Max power	350 W
Cooling	Passive · requires server chassis airflow
NEBS	Level 3 ready · telecom/edge deployment qualified

PLATFORM & SOFTWARE 05/05

vGPU	Supported
Secure Boot	With Root of Trust
MIG	Not supported
Server density	Up to 8-10 GPUs per node · per reference OEM configs, e.g. Thinkmate QH16-24E9-10GPU

PROCUREMENT

Typical lead time

Typically 2 to 4 weeks through direct-purchase channels for standalone cards (reported).

Pricing

Reported street pricing roughly $7,500 to $10,000 per card as single units; one reseller listed $7,569, marked down from a $9,450 baseline. Bulk and OEM pricing typically comes in below single-unit list (reported).

- Mature, high-volume Ada Lovelace part with wide OEM and channel availability (PNY board partner, CDW, Insight, Thinkmate, ASA Computers), unlike allocation-gated SXM/HBM parts.
- No MIG support, unlike the newer RTX PRO 6000 Server Edition; multi-tenant isolation has to happen at the software or scheduler layer instead.
- NEBS Level 3 ready, relevant for telecom and edge inference deployments.
- Reference OEM chassis ship up to 8 to 10 cards per node; check chassis PSU headroom at 350 W times the card count before ordering.

FIELD NOTES

- No NVLink, same constraint as the RTX PRO 6000 Server Edition: any multi-GPU tensor-parallel split rides PCIe and the host, so plan L40S nodes as pipeline-parallel or one-model-per-GPU, not as a tensor-parallel training pod.
- 48 GB is the real ceiling: a 70B model in FP8 needs at least two cards. The L40S is cost-effective for whatever fits on one card, roughly up to 30B-class in FP8 depending on context length and KV cache.
- VRAM per dollar is worse than the RTX PRO 6000 Server Edition on paper (48 GB vs 96 GB, at a narrower price gap than 2x), but the L40S wins on maturity: broader driver validation, wider OEM support, and no GDDR7 shortage premium.
- No MIG means you cannot hard-partition one card into isolated tenant slices the way the newer Blackwell part allows; use time-slicing or software-level multiplexing (Triton, vLLM multi-model serving) instead.
- Passive cooling, same as every other data center Ada or Blackwell PCIe card: it lives in a qualified server chassis only, never a generic tower.

QUESTIONS WE GET ON THIS PART

How much does an L40S cost?

Reported street pricing runs about $7,500 to $10,000 per card, with at least one reseller listing at $7,569 marked down from a $9,450 baseline. Bulk or OEM pricing typically comes in below single-unit list.

L40S vs RTX PRO 6000 Server Edition for inference?

The RTX PRO 6000 Server Edition roughly doubles memory (96 GB vs 48 GB) and adds FP4 support, and NVIDIA claims up to 5x higher LLM inference throughput over the L40S in agentic AI workloads. The L40S remains the cheaper, more mature, higher-availability choice for models that comfortably fit in 48 GB.

Does the L40S support NVLink or multi-GPU scaling?

No. The L40S has no NVLink; multi-GPU setups communicate over PCIe Gen4 x16 and the host system, which is adequate for running independent model replicas across cards but not for low-latency tensor-parallel splits of a single large model.

Is the L40S being discontinued now that the RTX PRO 6000 Server Edition has shipped?

NVIDIA has not published an end-of-life notice for the L40S; it remains listed as an active, shipping product alongside the newer Blackwell Server Edition, which NVIDIA positions as a performance upgrade rather than a replacement that forces L40S retirement.

REFERENCES

[1] https://www.nvidia.com/en-us/data-center/l40s/

[2] https://acecloud.ai/wp-content/uploads/2025/06/l40s-datasheet.pdf

[3] https://www.fluence.network/blog/nvidia-l40s/

[4] https://www.thundercompute.com/blog/nvidia-l40-pricing

[5] https://www.asacomputers.com/nvidia-l40s-48gb-graphics-card.html

[6] https://www.thinkmate.com/systems/servers/gpx/l40s?b=

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