

Qualcomm Cloud AI 100 Ultra

Also known as: Cloud AI 100 Ultra · AI 100 Ultra · QAIC Ultra · AI 80 Ultra (cut-down sibling)

The Cloud AI 100 Ultra is Qualcomm's inference-only PCIe accelerator, not a GPU: four Cloud AI 100 SoCs (64 AI cores in total) sit behind an on-card PCIe switch, with 128 GB of LPDDR4x at 548 GB/s and 576 MB of on-die SRAM, all inside a 150 W, single-width, full-height three-quarter-length card. It is the big brother of the 75 W, single-SoC Cloud AI 100 Standard and Pro cards. The attraction is memory capacity per watt: Qualcomm rates it for 100B-parameter models on one 150 W card. The trade is low memory bandwidth for its capacity, a compiler-based toolchain (Cloud AI SDK) instead of CUDA, and no training. GPU Smith treats it as a candidate for power-capped, inference-only boxes running models the Qualcomm toolchain already supports, not as a general-purpose GPU replacement.

Up to 870 TOPS INT8 · vendor peak figure	Up to 288 TFLOPS FP16 · vendor peak; SDK docs list 290	128 GB LPDDR4x ECC Memory · 548 GB/s on-card
576 MB On-die SRAM · 9 MB per AI core	150 W TDP · per card	PCIe FH3/4L, single width Form factor · PCIe Gen4 x16 host link

COMPUTE01/05	
SoCs per card	4x Cloud AI 100 · behind an on-card PCIe switch
AI cores	64 · 16 per SoC
Peak INT8	Up to 870 TOPS
Peak FP16	Up to 288 TFLOPS · product page; the SDK hardware table lists 290 TFLOPS
Data types (SoC)	FP16, INT16, INT8, FP32 · as listed for the Cloud AI 100 SoC
Process node (SoC)	7 nm

MEMORY02/05	
DRAM	128 GB LPDDR4x · with ECC
DRAM bandwidth	548 GB/s · aggregate across the four SoCs
On-die SRAM	576 MB · 9 MB per AI core, software managed
Model size	100B parameters on one card · vendor claim; 175B on two cards

PROCUREMENT

Typical lead time

Not published. Qualcomm's product page carries a Buy Now link but no lead time; the card mostly reaches buyers inside OEM servers or integrated systems, so lead time follows the integrator's quote.

Pricing

Card price not published; quote-only. The one public price point is rental: Cirrascale lists an 8x AI 100 Ultra bare-metal server (128 vCPUs, 512 GB RAM, 2x 3.84 TB NVMe) at \$4,699 per month on a monthly term or \$3,759 per month on an annual term (reported, October 2026).

- HPE named the ProLiant DL380a Gen11 as a supported server at launch; ask any other OEM for a written qualification before ordering cards for its chassis.
- Innodisk sells the APEX-X100-Q, an Intel Core i7-13700E system with one AI 100 Ultra pre-installed, 192 GB DDR5 and a 2 TB SSD, for edge deployments.
- Cirrascale runs both bare-metal Ultra servers and a hosted Inference Cloud on the Qualcomm AI Inference Suite: the cheap way to test your model on the toolchain before buying hardware.
- The AI 80 Ultra shares the card, memory and 150 W TDP but has half the AI cores (32); check which SKU a quote is for.

FIELD NOTES

- Bandwidth is the constraint, not capacity. 548 GB/s across 128 GB is well below what current GDDR7 and HBM accelerators deliver, so token generation on a dense model that fills the card will be slow per stream; the card makes sense where you need many models or big contexts resident at low watts, not maximum tokens per second per user.
- It is four SoCs behind a PCIe switch, not one big chip. Dividing the card totals by four gives roughly 32 GB and 137 GB/s per SoC (inferred, Qualcomm does not state the split), and a model larger than one SoC is partitioned across them by the compiler. Expect the per-SoC layout to matter for batch size and model placement in the same way multi-GPU layout does.
- The software path is compile-ahead: models are compiled with the Cloud AI SDK, with Hugging Face models deployed through Qualcomm's Efficient Transformers library. Check that your exact architecture and quantization are on Qualcomm's supported list before you commit; there is no CUDA fallback when they are not.
- Qualcomm's own documents disagree on some figures: the product page lists the Pro and Standard at 400 and 350 TOPS INT8 (200 and 175 TFLOPS FP16), while the SDK hardware table lists 375 and 325 TOPS (125 and 110 TFLOPS), and the Ultra at 288 versus 290 TFLOPS FP16. Size against measured throughput on your model, not either peak.
- Qualcomm's data center roadmap has moved to rack-scale parts (AI200 with 768 GB of LPDDR per card, AI250 with near-memory compute, both liquid-cooled at 160 kW per rack). Qualcomm describes them as accelerator cards and racks, but has published no air-cooled PCIe card for a single-box build; plan a multi-year deployment on the AI 100 Ultra knowing the roadmap is aimed at racks.

QUESTIONS WE GET ON THIS PART

What is the difference between Cloud AI 100 Ultra, Pro and Standard?

Ultra is four Cloud AI 100 SoCs on one 150 W full-height card with 64 AI cores, 128 GB LPDDR4x and 548 GB/s. Pro and Standard are single-SoC, 75 W half-height half-length cards with 16 and 14 AI cores, 32 GB and 16 GB of memory, and 137 GB/s, on a PCIe Gen4 x8 link.

How big a model fits on one Cloud AI 100 Ultra?

Qualcomm claims 100 billion parameter models on a single card and 175 billion on two cards, with larger models across more cards. That is a capacity claim; at 548 GB/s per card, decode speed on models that large will be well below what a GPU with HBM or GDDR7 delivers.

Can it run CUDA or train models?

No. It is an inference accelerator programmed through Qualcomm's Cloud AI SDK and AI Inference Suite. Qualcomm's positioning is train elsewhere, then compile and deploy the trained model onto the card.

How much does a Cloud AI 100 Ultra cost?

Qualcomm does not publish a card price; it is quote-only through Qualcomm and server integrators. The only public price is rental: Cirrascale lists an 8-card Ultra server at \$4,699 per month, or \$3,759 per month on an annual term (reported, October 2026).

REFERENCES

[1] <https://www.qualcomm.com/data-center/products/cloud-ai-100-ultra>

[2] <https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/Prod-Brief-QCOM-Cloud-AI-100-Ultra.pdf>

[3] https://docs.qualcomm.com/bundle/publicresource/80-99100-3/topics/index_Architecture.md

[4] <https://www.qualcomm.com/news/onq/2023/11/introducing-qualcomm-cloud-ai-100-ultra>

[5] <https://www.qualcomm.com/news/releases/2025/10/qualcomm-unveils-ai200-and-ai250-redefining-rack-scale-data-cent>

[6] <https://www.qualcomm.com/data-center>

[7] <https://www.cirrascale.com/pricing>

[8] <https://cirrascale.com/ai-innovation-cloud/qualcomm-cloud-ai>

[9] https://www.innodisk.com/upload/file/APEX-X100-Q_Datasheet.pdf