

Tenstorrent TT-QuietBox 2 (Blackhole)

Also known as: TT-QuietBox 2 · QuietBox 2 · QB2 · TT-QuietBox · QuietBox Blackhole

The TT-QuietBox 2 is Tenstorrent's liquid-cooled desk-side tower: two p300c cards carrying four Blackhole ASICs (480 Tensix cores, 128 GB GDDR6 in total) behind an AMD Ryzen 7 9700X host with 256 GB of DDR5, on a 1,600 W supply that plugs into a standard 120 V outlet. At \$9,999 it is the cheapest packaged way to get 128 GB of accelerator memory with a vendor-supported open-source stack, and Tenstorrent says it runs open-weight models up to about 120B parameters on the box. The trade-offs are a non-CUDA software stack, four separate 32 GB devices rather than one memory pool, and a peak draw near 1,500 W, which is a lot for a desk circuit. It replaces the first Blackhole QuietBox (four p150c cards, \$11,999, now sold out).

4x Blackhole ASIC Accelerators · on 2x p300c cards, 480 Tensix cores	128 GB GDDR6 Accelerator memory · 32 GB per chip; 1,024 GB/s per card	2,654 TFLOPS Peak compute · BLOCKFP8, vendor claim
AMD Ryzen 7 9700X Host · 8 cores, 256 GB DDR5-5600	~1,500 W peak Power · 1,600 W PSU, 120 V outlet	\$9,999 Price · list, Tenstorrent store

ACCELERATORS		01/05
Cards	2x Blackhole p300c · dual-ASIC, liquid-cooled; p300c is not sold separately	
ASICs	4x Blackhole · software sees four separate devices	
Tensix cores	480 · 120 per chip	
On-chip SRAM	720 MB · 180 MB per chip	
Peak compute	2,654 TFLOPS · BLOCKFP8, vendor claim; other precisions not published	
Card power	600 W TBP per p300c · per the p300c spec page	

MEMORY		02/05
Accelerator memory	128 GB GDDR6 · 32 GB per chip; chips do not share memory	
Memory bandwidth	1,024 GB/s per card · 512 GB/s per chip	
System memory	256 GB DDR5-5600	
Storage	4 TB NVMe PCIe Gen5	

HOST & CONNECTIVITY		03/05
CPU	AMD Ryzen 7 9700X · 8 cores, 3.8 GHz base	
Accelerator host link	PCIe Gen4 · per the QB2 guide	
Card-to-card link	Samtec ARP6 cable · connects the two p300c cards internally	
Network	Gigabit Ethernet · plus Wi-Fi 6 dual-band and Bluetooth 5.3	
POWER & PHYSICAL		04/05
Power supply	1,600 W	
Peak draw	~1,500 W	
Outlet	Standard 120 V wall outlet · no rack required	
Cooling	Liquid · 38 dBA max	
Dimensions	9.1 x 17.8 x 15.6 in · tower	
Weight	20 kg (44 lb)	
SOFTWARE		05/05
OS	Ubuntu 24.04 LTS · preinstalled	
Stack	TT-Metalium, tt-inference-server, TT-Studio · open source; containers preinstalled; not CUDA	
Model ceiling	Up to ~120B parameters · vendor claim, e.g. GPT-OSS 120B on-device	
Throughput claim	476.5 tokens/s on Llama 3.1 70B · vendor claim; batch size and precision not stated	

PROCUREMENT

Typical lead time

Listed in stock on Tenstorrent's store, shipping in about 2 weeks plus 2 to 5 weeks transit, so plan on 4 to 7 weeks from a direct order. Tenstorrent announced the QB2 on March 13, 2026 with worldwide shipments from Q2 2026.

Pricing

\$9,999 list on Tenstorrent's store. The earlier QuietBox models are listed sold out: TT-QuietBox (Blackhole, 4x p150c) at \$11,999 and TT-QuietBox (Wormhole, 4x n300) at \$15,000.

- Sold direct from Tenstorrent at a published price; the p300c cards inside are not sold separately, so this box is the only way to buy them.
- The QB2 ships with fewer, denser cards (two dual-chip p300c) than the first Blackhole QuietBox (four p150c), at a lower price for the same 128 GB of GDDR6.
- Preinstalled software can be out of date on arrival; Tenstorrent's own guide tells you to upgrade the stack before running tutorials. Budget setup time.
- No enterprise support terms or warranty length are stated on the product page; confirm them in writing for a fleet purchase.

FIELD NOTES

- 128 GB is four 32 GB islands, not one pool. A 120B model has to be split across the four chips with TT-NN multi-device code; it is not the same as a 128 GB unified-memory box where any process sees all of it.
- Peak draw of about 1,500 W on a 120 V outlet is roughly 12.5 A, above the 12 A (80%) continuous load usually allowed on a 15 A household circuit (our arithmetic). Put it on a dedicated 20 A circuit for sustained inference.
- Each p300c is rated 600 W, so the two cards alone account for up to 1,200 W of the budget; the 8-core Ryzen host is sized to feed the accelerators, not to do heavy CPU work.
- The 2,654 TFLOPS figure is BLOCKFP8 and the 476.5 tokens/s Llama 3.1 70B figure comes without a batch size; neither compares directly with NVIDIA FP8 or FP4 numbers. Benchmark your own model on the shipped software release.
- Buy it to develop on Tenstorrent's open stack or to run a supported open-weight model privately at a desk. For CUDA tooling or the widest day-one model support, a GB10 or RTX-class box is the lower-risk choice.

QUESTIONS WE GET ON THIS PART

What is inside the TT-QuietBox 2?

Two liquid-cooled Blackhole p300c cards with four Blackhole chips in total (480 Tensix cores, 720 MB SRAM, 128 GB GDDR6), an AMD Ryzen 7 9700X host, 256 GB DDR5-5600, a 4 TB PCIe Gen5 NVMe drive and a 1,600 W power supply, in a tower that weighs 20 kg.

How much does the QuietBox cost?

The TT-QuietBox 2 lists at \$9,999 on Tenstorrent's store. The earlier Blackhole QuietBox (four p150c cards) listed at \$11,999 and the Wormhole QuietBox at \$15,000; both are shown as sold out.

What models can it run?

Tenstorrent says open-weight models up to about 120B parameters, citing GPT-OSS 120B on-device and 476.5 tokens/s on Llama 3.1 70B (vendor figures, batch size not stated). Each model needs support in Tenstorrent's TT-Metalium/TT-NN stack, so check the supported-model list first.

Can it run from a normal wall socket?

Yes, Tenstorrent says it plugs into a standard 120 V outlet and needs no rack. Peak draw is about 1,500 W, though, so a dedicated circuit is the safer setup for sustained load.

REFERENCES

- [1] <https://tenstorrent.com/hardware/tt-quietbox>
- [2] <https://docs.tenstorrent.com/tt-quietbox2-guide>
- [3] <https://docs.tenstorrent.com/tt-quietbox2-guide/first-timer/01-what-just-arrived/>
- [4] <https://docs.tenstorrent.com/aibs/blackhole/p300.html>
- [5] <https://kyodonewsprwire.jp/release/202603135607>