128GB+ Unified Memory Systems

Six boxes you can actually buy as of mid-October 2026 — priced, ranked, and with the tradeoffs nobody puts in the marketing copy.

> **The context that matters first:** this entire category got squeezed by the 2026 DRAM shortage (the press calls it the “RAMpocalypse”). Every price below is well above launch pricing, and analysts expect continued upward drift into early 2027 rather than relief. Unified memory capacity is now the headline spec that gates what you can run — but for single-user inference, **memory bandwidth is the speed limit**, and these boxes differ by ~3x on that number.

AT A GLANCE

System	Silicon	Memory & bandwidth	Price (Oct 2026)	OS	Availability
Strix Halo mini PCs GMKtec / Minisforum / Beelink	AMD Ryzen AI Max+ 395 (Radeon 8060S iGPU)	128GB LPDDR5X ~215 GB/s real	**$3,450–4,350** GMKtec $3,499.99 in stock	Win / Linux	In stock
AMD Ryzen AI Halo	AMD Ryzen AI Max+ 395	128GB @ 256 GB/s	**$3,999 list** street $4,700–4,900	Win / Linux	Shipping since Jul
Surface RTX Spark Dev Box	Nvidia RTX Spark (20-core Grace + 6,144 CUDA)	128GB, ~1 petaflop	**$5,999**	Win 11 Pro	Pre-order, ships Nov
NVIDIA DGX Spark 128GB	Nvidia GB10 Grace Blackwell	128GB @ 273 GB/s	**$6,950** was $4,699 ↑ $2,251	DGX OS (Linux)	Nvidia + OEM partners
Mac Studio M5 Max 40-core GPU config	Apple M5 Max	128GB @ **614 GB/s**	**~$5,099**	macOS	Shipping since Sep 22
Mac Studio M5 Ultra	Apple M5 Ultra	96GB base, to 512GB @ 1.2 TB/s	**$5,499+**	macOS	Shipping; 512GB late Oct

PROS & CONS

NVIDIA DGX Spark 128GB

The serving / CUDA pick — if you can stomach the hike

Nvidia split the line on Oct 2: a 64GB model landed at $4,999 while 128GB jumped to $6,950, same chip and same bandwidth either way.

- + Best serving throughput here: 2–13x the AMD Halo in vLLM tests, with the biggest gaps in FP4 where Blackwell has dedicated hardware.
- + Real 200Gb ConnectX fabric — two units cluster into a genuine 256GB tensor-parallel machine.
- + ~13,400 MB/s SSD; the mature CUDA ecosystem everyone else is chasing.
- − A $2,251 price increase with no published explanation from Nvidia.
- − 273 GB/s still means dense 70B single-user decode crawls. It’s a batch/serving monster, not a fast chat box.
- − Locked to Nvidia’s Linux-based DGX OS.

Surface RTX Spark Dev Box

The Windows-native CUDA option — launched Oct 7

Microsoft’s answer: 20-core Grace CPU, 6,144 CUDA-core Blackwell GPU, 128GB, rated at one petaflop.

- + The only CUDA-capable unified-memory box with a first-party Windows 11 Pro dev image plus Surface warranty. Matters if your team lives in WSL and Visual Studio.
- + Day-one tooling support (CPU-Z, HWMonitor) already landed, so less early-adopter tax than usual.
- − A $1,000+ premium over DGX Spark for the convenience layer.
- − Storage capacity still unpublished, no third-party thermal or sustained-load reviews, US-exclusive via Microsoft.com.

AMD Ryzen AI Halo

The only box in the group that’s a normal PC

AMD’s own 128GB Strix Halo dev box in a 150mm coaster-sized chassis, 120W over USB-C.

- + Lowest official list price ($700 under DGX Spark on paper), 2TB SSD, 10GbE, and identical hardware in both Windows and Linux SKUs.
- + Doubles as a real desktop — the Radeon 8060S games, it runs your existing software, and it isn’t a paperweight once the AI phase ends.
- − Newegg marketplace pricing ($4,700–4,900) erases the headline discount entirely.
- − Half the Spark’s SSD speed; single 10GbE port ends any clustering conversation.
- − ROCm remains second-class to CUDA. 192GB PRO 495 version is still “coming soon,” no price.

Strix Halo mini PCs

Cheapest way into 128GB — GMKtec, Minisforum, Framework, Beelink

Same silicon as the Halo, thinner wallets. GMKtec EVO-X2 at $3,499.99 is the cheapest box here that can hold a 120B-class model in memory.

- + On Linux with Vulkan, MoE models fly: ~100 tok/s on Qwen3-30B-A3B, 31–55 tok/s on gpt-oss 120B.
- + The only desktop route to a 120B-class model in unified memory under $4,500.
- − Prices are up 73%+ from launch — the Framework Desktop that was $1,999 now lists at $3,449 and is out of stock.
- − Prompt processing is the weak spot: ~500 tok/s best case, an order of magnitude below a discrete GPU.
- − ROCm needs `HSA_OVERRIDE_GFX_VERSION`; Windows costs 20–30% performance and 14GB of GPU-addressable memory vs Linux.
- − Dense 70B is a non-starter (~5 tok/s). **This hardware is for MoE models.**

Mac Studio M5 Max / M5 Ultra

The bandwidth king — refreshed Sept 22

- + 614 GB/s on the 128GB M5 Max, roughly 3x Strix Halo, so dense models that crawl on AMD are genuinely usable. Best interactive speed per dollar.
- + ~4x faster LLM prompt processing than the M4 Max generation, near-silent, and MLX works out of the box.
- + M5 Ultra reaches 512GB at 1.2 TB/s — the only box here that goes beyond 128GB.
- − No CUDA. You live inside Apple’s MLX / Metal ecosystem.
- − No clustering option at all.
- − Memory is non-upgradeable and soldered — size it right at checkout. 128GB forces the 40-core GPU bump, so price the entry M5 Ultra too.
- − This generation added $500 (Max) to $1,500 (Ultra) over the prior launch prices.

NVIDIA DGX Spark 64GB

Honourable mention — the new entry door

Announced Oct 2 at $4,999, sold through Acer, ASUS, Dell. Same GB10 chip and same 273 GB/s as the 128GB model — Nvidia cut capacity, not compute. Comfortable up to 30B-class models; the wall shows up with bigger fine-tunes or several models loaded at once.

WHICH TO BUY

- **CUDA, serving, or a growth path to 256GB:** DGX Spark — if $6,950 is acceptable to you.
- **Cheapest 128GB today:** GMKtec EVO-X2 or Minisforum — but only if you’ll run Linux and stick to MoE models.
- **One machine for work, games, and local AI:** Ryzen AI Halo — just check its live street price against the Spark before assuming the discount is real.
- **Windows-native CUDA, fleet-manageable:** Surface RTX Spark Dev Box, if the $5,999 and the November wait are fine.
- **Fastest single-user inference, including dense models:** Mac Studio M5 Max 128GB. Bandwidth bottlenecks interactive use, and Apple has 2.3–3x of it.
- **Need more than 128GB:** Mac Studio M5 Ultra is effectively your only option today.

Meta-caveat: all of these are priced by a memory market that nobody expects to loosen before ~2027. Buy for the models you actually run now, not the one you’re imagining six months out.

Sources: [1] Nvidia newsroom / Tom’s Hardware / Hardware Busters (shattered.io, 2 Oct 26) — Spark 64GB $4,999; 128GB repriced to $6,950. [2] Microsoft Windows Devices blog (tech-insider.org, 7 Oct 26) — Dev Box $5,999, ships Nov 26. [3] AMD / Tom’s Hardware / StorageReview (thedeskbrief.com, 25 Sep 26) — Halo pricing; vLLM results vs Spark. [4] Apple Newsroom (25 Aug 26) — M5 Max / Ultra specs and pricing. [5] openclawdc.com (3–5 Oct 26) — Strix Halo vendor pricing, bandwidth and model benchmarks.