

128GB+ Unified Memory Systems for Local AI — Comparison

Prices/availability checked October 2026 · Whole category inflated ~30–75% in 2026 on the DRAM/HBM shortage ("RAMpocalypse")

At a Glance

System	Price	Memory	Bandwidth	\$ / GB	Platform	Availability
Bosgame M5 / GMKtec EVO-X2 (Ryzen AI Max+ 395)	\$3,600–3,650	128GB	256 GB/s	~\$28	x86 · Win/Linux/ Proxmox	Bosgame low stock; EVO-X2 in stock (Amazon/Micro Center)
AMD Ryzen AI Halo (first-party dev kit)	\$3,999	128GB	256 GB/s	~\$31	x86 · Win 11 Pro	Retail (Micro Center), 2TB SSD
Mac Studio M5 Max, 40-core GPU	~\$5,100	128GB	614 GB/s	~\$40	macOS/ARM · MLX/Metal	Apple direct only (not Amazon)
NVIDIA DGX Spark (GB10)	\$6,950	128GB	273 GB/s	~\$54	DGX OS/ARM · CUDA	NVIDIA direct; raised from \$3,999 in Feb
ASUS Ascent GX10 (same GB10 chip)	~\$6,200	128GB	273 GB/s	~\$48	CUDA · ConnectX-7	Amazon 3rd-party, thin stock, check warranty
Framework Desktop (Ryzen AI Max+ PRO 495)	\$6,799 (192GB)	32–192GB	256 GB/s	~\$35 (192GB)	x86 · Linux-friendly	Batch pre-order; batch 1 sold out day 1
Mac Studio M5 Ultra	\$5,499 / \$9,499	96GB / 256GB	1.2 TB/s	\$57 / \$37	macOS/ARM · MLX/Metal	Apple; 512GB config late Oct (80-core GPU)

Pros & Cons

Strix Halo mini PCs (~\$3,600–3,999)

Best value — cheapest 128GB behind a GPU

- + Cheapest \$/GB; gpt-oss-120b decode ~54 tok/s, only 8–20% behind DGX Spark at half the price
- + Normal x86 PC: Windows, Ubuntu, Fedora, Proxmox; llama.cpp via Vulkan/ROCm
- + ~\$3,999 AMD Halo = polished first-party version (Win 11 Pro, 2TB)
- Weak prompt processing (~459 tok/s) — long-context/RAG pre-fill takes 40+s
- GPU memory cap: 96GB on Windows BIOS; Linux kernel param needed for ~115–120GB
- Cheap units have thin networking (Bosgame: 1×2.5GbE); poor batched-serving scaling

NVIDIA DGX Spark / ASUS GX10 (\$6,200–6,950)

Buy for the software, not the tokens

- + Full CUDA stack: PyTorch/vLLM/TensorRT-LLM NGC containers, NVFP4; fine-tune code ports easily
- + ~5× faster prompt processing vs Strix Halo; crushes batched serving (2,760 vs 314 tok/s vLLM)
- + New 64GB model \$4,999 (Oct 23, Acer/ASUS/Dell/Gigabyte/HP/MSI) — same chip/bandwidth
- 128GB model = \$54/GB; ~120GB usable; ARM limits x86-only wheels/containers
- Barely beats \$3,600 Strix box on single-user decode — bad buy for chatting alone

Mac Studio M5 Max, 40-core (~\$5,100)

Fastest single-user token generation

- + 614 GB/s: ~88 tok/s gpt-oss-120b, ~120 tok/s on 8B Q4 — 1.5× the Spark's decode
- + Best bandwidth-per-dollar in the 128GB class
- No CUDA; MLX/Metal only; thinner fine-tuning tooling
- GPU memory capped ~75% (~96GB); sysctl bump resets on reboot; RAM soldered at purchase
- Must buy from Apple; 32-core M5 Max tops at 36GB — don't pick it for this

Framework Desktop, PRO 495 (\$6,799)

Newest entrant (Sept 30) — only x86 box past 128GB

- + Up to 192GB; repairable; strong Linux support; long-context-friendly chip (495)
- Same 256 GB/s bandwidth wall as \$3,600 boxes — ~2× price for capacity + form factor
- Batch pre-order model = availability roulette

Mac Studio M5 Ultra (\$5,499–\$9,499+)

The >128GB escape hatch

- + 1.2 TB/s; 256GB config (~\$9,499) runs several large models at once, ~218GB usable
- + 512GB arriving late Oct — biggest single-box unified memory pool
- Base SKU is only 96GB; 128GB needs the 80-core GPU → ~\$12K real-world configs

Bottom Line

Cheapest 128GB lab box: GMKtec EVO-X2 / Bosgame M5 (~\$3,600), Linux w/ GPU-memory kernel param. **Fastest solo inference:** Mac Studio M5 Max 40-core 128GB (~\$5,100). **CUDA, fine-tuning, multi-user serving:** DGX Spark / GX10 (64GB at \$4,999 is the new value pick). **>128GB in one box:** M5 Ultra 256GB (Mac) or Framework 192GB (x86). **Note vs a discrete 128GB GPU rig (RTX PRO 6000, ~1.8 TB/s):** these boxes win on idle power, big-MoE single-stream decode, and dev-env capacity — not raw throughput/\$.

Sources: [computingforgeeks.com](https://www.computingforgeeks.com) DGX Spark vs Strix Halo vs Mac Studio (Oct 13, 2026, prices/API stock checked Oct 13) · shattered.io DGX Spark 64/128GB pricing (Oct 2) · [The Register](https://www.theregister.com) (memory-crunch repricing) · [Micro Center](https://www.microcenter.com) & AMD (Ryzen AI Halo \$3,999) · [Framework/Videocardz](https://www.frameworkvideo.com) (192GB desktop, Sept 30) · [Apple newsroom](https://www.apple.com/newsroom) & [Forbes](https://www.forbes.com) (M5 Ultra pricing) · [llama.cpp/Ollama/StorageReview/Hardware](https://llama.cpp) Corner benchmarks. LLM tok/s figures from published runs — runtimes differ between rows.