Unified-Memory Boxes for Local LLMs — 128GB+ Buyer's Guide

Research summary · October 18, 2026 · Focus: local inference capacity, price & availability (US)

Context: The 2026 DRAM shortage ("RAMpocalypse") has roughly **doubled** prices on unified-memory machines since mid-2026 — 128GB Strix Halo boxes went from ~$2,000 (June) to $3,449–4,349 now. No near-term relief expected; prices are still drifting up.

Comparison Table

System	Chip	Memory / Bandwidth	Price (128GB)	Availability	Best for
GMKtec EVO-X2	Ryzen AI Max+ 395	128GB LPDDR5X-8000 · ~256 GB/s · ~96GB GPU-usable	$3,499 (1TB) / $3,649 (2TB)	In stock (Amazon + GMKtec)	Cheapest 128GB in stock; value
Framework Desktop	Ryzen AI Max+ 395	Same silicon as above	$3,449 DIY (no SSD)	Out of stock — pre-order	Linux, repairability, PCIe slot
Minisforum MS-S1 Max	Ryzen AI Max+ 395	Same silicon · dual 10GbE, USB4 v2, PCIe x4, 2U	$3,879 (2TB)	Taking orders	Homelab serving node
Beelink GTR9 Pro	Ryzen AI Max+ 395	Same silicon · dual 10GbE (E610), vapor chamber	$4,349 (2TB)	35-day pre-sale	Serving node (10GbE), polished I/O
Bosgame M5	Ryzen AI Max+ 395	Same silicon · 2.5GbE only	$3,599 (2TB)	Low stock on Amazon (~13)	Cheapest 128GB on Amazon
NVIDIA DGX Spark (+ Gigabyte/Lenovo/MSI/ASUS/Dell/HP/Acer GB10 clones)	GB10 Grace Blackwell (Arm)	128GB unified · **273 GB/s** · ConnectX-7 200Gb/s (2 units → ~405B models)	$6,950 (NVIDIA); OEMs from ~€5,038 ex-VAT	Founders Ed. discontinued, limited; OEMs shipping	CUDA stack, fine-tuning, multi-unit linking
Mac Studio M5 Max	Apple M5 Max (40-core GPU)	128GB max · **614 GB/s**	~$5,099 (128GB/512GB)	Shipping since Sep 22, 2026	Speed per GB; dense models
Mac Studio M5 Ultra	Apple M5 Ultra (80-core GPU)	96GB base → **512GB max** · **1.2 TB/s**	$5,499 (96GB base); 128GB ≈ $6.5–7k est.	Shipping; 512GB config late Oct	200B+ dense models at interactive speed
192GB boxes (GMKtec EVO-X5, Framework, ACEMAGIC F9A, MS-S1-P495)	Ryzen AI Max+ PRO 495 ("Gorgon Halo", new)	192GB LPDDR5X-8533 · ~273 GB/s · ~160GB GPU-assignable	$6,499–7,399	Pre-order / late-Oct ship; batches selling out	MoE models needing 96–160GB VRAM

Pros & Cons by Platform

Strix Halo mini PCs (Ryzen AI Max+ 395)

+ Best $/GB of model space: ~96GB usable VRAM > any single consumer GPU
+ Real-world: gpt-oss-120B MoE ~52–56 tok/s; full desktop x86 Linux/Windows
+ EVO-X2 loud fans/2.5GbE; MS-S1 Max + GTR9 Pro get dual 10GbE
- ~256 GB/s ceiling: dense 70B only 4–6 tok/s (capacity machine, not speed)
- RAM soldered; prices up 75–100% since June

NVIDIA DGX Spark (GB10)

+ Full CUDA/NVIDIA AI stack; best fine-tuning tooling of the group
+ ~1 PFLOP FP4; ConnectX-7 lets 2 units run ~405B-param models
- 273 GB/s hurts: Llama 3.3 70B ≈ 3 tok/s
- 128GB SKU repriced $4,999 → $6,950 on Oct 2; Arm/Linux only; worst $/GB here

Mac Studio (M5 Max / M5 Ultra)

+ M5 Max: 614 GB/s ≈ 2.4× Strix Halo — fastest dense-model tokens/sec per $ at 128GB
+ M5 Ultra: 1.2 TB/s, up to 512GB — only box here for 200B+ dense locally
+ Silent, compact, ships now; strong MLX/LM Studio ecosystem
- Memory/storage non-upgradeable — configure for life; 128GB is Max's hard ceiling
- Prices up 25–37.5% YoY; less CUDA flexibility

192GB Ryzen AI Max+ PRO 495 (new)

+ Only way to run Qwen3-235B Q4 (~142GB), MiniMax-M2 (~138GB) on one box
- Bandwidth barely moves (~273 GB/s) — models that fit a 395 run nearly the same speed at ~half the price

Recommendations

Best value: GMKtec EVO-X2 ($3,499) — cheapest 128GB in stock; gpt-oss-120B runs great. | **Best per-box performance:** Mac Studio M5 Max 128GB (~$5,099) — ~2.4× the bandwidth for dense-model speed. | **Homelab node:** Minisforum MS-S1 Max — dual 10GbE + PCIe for $3,879. | **Skip:** $6,950 DGX Spark 128GB (unless you need CUDA/linking) and 192GB PRO 495 boxes (unless chasing 235B-class MoE).

Sources: RedShark (Mac Studio M5 pricing/specs) · pi3g DGX Spark Buyer's Guide, Oct 2026 · Shattered.io, DGX Spark 64/128GB repricing, Oct 2, 2026 · ComputingForGeeks Strix Halo 5-way comparison, prices re-checked Oct 10, 2026 · TechRepublic M5 Max vs Ultra, Sep 22, 2026 · Apple Newsroom. Caveat: M5 Ultra 128GB BTO price is extrapolated from B&H option grid — treat as estimate. Amazon/store prices re-check before buying; memory-market drift is upward.