Unified Memory Systems with 128GB+ — Late-2026 Buying Comparison

Local-LLM desktop & laptop boxes · prices and availability as of November 6, 2026 · the 2026 DRAM shortage has pushed prices 30–75% above launch

System	Memory / Bandwidth	Price (Nov 2026)	Availability	Best for
Mac Studio — M5 Max	128GB max · 614 GB/s	~$4,000–4,100 configured	In stock, ships now	Best $/GB in the top tier; quiet daily-driver computer
Mac Studio — M5 Ultra	96GB std; 256/512GB · 1.2 TB/s	$5,499 base · 256GB ≈ $9,599 · 512GB $22,299	In stock; 512GB ships late Oct	Biggest coherent memory pool on a desk; fastest bandwidth
Strix Halo mini PCs GMKtec EVO-X2 · Minisforum MS-S1 Max · Framework Desktop · HP Z2 Mini G1a	128GB · 256 GB/s	$3,449–4,349	GMK/Minisforum in stock; HP ~4–6 wks; Framework 128GB ships Nov	Cheapest path to 128GB; general-purpose Windows/Linux PC
AMD Ryzen AI Halo Ryzen AI Max+ PRO 495 "Gorgon Halo"	**192GB** · LPDDR5X-8533	$6,799–9,249 (AMD box: $7,299)	Linux version in-store at Micro Center	Only >128GB system under $10K besides the Mac
NVIDIA DGX Spark (GB10)	128GB · 273 GB/s	$6,950 launch price was $3,999	Out of stock at NVIDIA Marketplace	Full CUDA stack; 2-box clustering for 405B-class models
NVIDIA DGX Spark 64GB	64GB · 273 GB/s	$4,999	Shipping since Oct 23 (Acer, ASUS, Dell, Gigabyte, HP, MSI)	Entry-level CUDA box with ConnectX-7
Microsoft Surface Laptop Ultra RTX Spark / N1X	up to 128GB · 273 GB/s	$5,899 (128GB) · $2,599 base (24GB)	Shipping since Oct 16	Only laptop in this class; Blackwell RTX, 6,144 CUDA cores

Pros & cons

Mac Studio M5 Max (128GB)
+ Lowest cost per GB in the top tier; 614 GB/s ≈ 2.3× the AMD/NVIDIA boxes; silent; it's a real computer.
− No CUDA (MLX only); 128GB is the ceiling; weak for serving many concurrent users.

Mac Studio M5 Ultra
+ 1.2 TB/s bandwidth; Neural Accelerators in every GPU core (≈4× faster prompt processing); up to 512GB coherent memory.
− Memory upgrades are brutal (+$4,000 for 256GB); base chip is binned; Apple removed the 512GB option from the previous gen mid-cycle.

Strix Halo mini PCs (128GB)
+ Cheapest 128GB anywhere (~$3,500); good Vulkan / llama.cpp / LM Studio experience; normal Windows or Linux PC.
− ROCm is rough (vLLM / FlashAttention-2 need manual builds); ~75GB max GPU-assignable memory; dense 70B decode ≈ 2.6 tok/s.

Ryzen AI Halo (192GB)
+ Only sub-$10K option with more than 128GB besides the Mac; headroom for 100B+ dense or 200B+ MoE models.
− Expensive; same ROCm caveats; brand-new platform with a thin OEM ecosystem.

NVIDIA DGX Spark (128GB)
+ Full CUDA / TensorRT-LLM / NIM stack on day one; dev–prod parity with cloud GPUs; two units cluster over 200GbE for up to 405B models.
− Price up 75% since launch and out of stock; 273 GB/s caps dense decode; locked appliance, not a general PC.

DGX Spark 64GB
+ Cheapest entry to NVIDIA's software stack with built-in ConnectX-7 clustering.
− Half the memory — can't hold 100B-class models; weak value at $4,999.

Surface Laptop Ultra
+ Portability; Blackwell RTX with 6,144 CUDA cores — more raw compute than the Spark or Halo iGPU.
− Laptop thermals; same ~273 GB/s bandwidth ceiling as the rest; specs sourced from third-party reports.

Three takeaways

1. **Prices won't improve by waiting.** The 2026 DRAM shortage has pushed every system 30–75% above launch price and analysts expect it to persist into 2027.
2. **Bandwidth, not capacity, is the bottleneck.** Every AMD/NVIDIA box runs 256–273 GB/s, so dense 70B decode is slow on all of them (~2.6 tok/s). MoE models are the sweet spot; the Mac's 546–1,200 GB/s is the only meaningful fix.
3. **Quick picks.** Price-per-GB: Strix Halo mini (~$3.5K). CUDA/dev–prod parity: DGX Spark. Speed + general computer: Mac Studio M5 Max 128GB.

Compiled November 6, 2026 · Sources: NVIDIA Marketplace, Apple, MacRumors, The Register, ServeTheHome, Videocardz, TechPowerUp, MacWorld / dev.to pricing analyses, openclawdc.com benchmark comparisons. Surface Laptop Ultra details come from third-party sources — verify on the Microsoft Store before purchase. Prices move weekly in the current memory market.