Recently Released Unified Memory Systems (128GB+) — Price / Availability / Pros & Cons Comparison

Source data as of September 2026 · Best local LLM hardware choices

Comparison Overview

System	Unified Mem	Price (U.S.)	Availability	Bandwidth	Best For
Mac Studio M5 Max	128 GB	$2,499	Sep 22, 2026	614 GB/s	On-device LLMs, privacy, quiet desktop
Mac Studio M5 Ultra	512 GB	$5,499	Late Oct 2026+	1.2 TB/s	Massive frontier models, extreme pro + AI
RTX Spark (Win Laptops/SFF)	128 GB	~$2K–$4.5K	Fall 2026	273 GB/s	CUDA ecosystem, agents, capacity over speed
DGX Spark Desktop	128 GB LPDDR5x	~$4,699	Now	273 GB/s	~200B models, CUDA parity, agent stacks

Pros & Cons per System

Mac Studio M5 Max	■ **Pros:** 128GB unified → large local LLMs on-device; near-silent (~80W), battery-optional; macOS Core AI + MLX framework; Thunderbolt 5 RDMA clustering for 3x distributed inference	■ **Cons:** No CUDA; thinner advanced-ML tooling vs NVIDIA; education pricing varies by region
Mac Studio M5 Ultra	■ **Pros:** Up to 512GB unified memory — frontier models on a single machine; 4.3x AI compute vs M3 Ultra; 33× 8K ProRes streams; Thunderbolt 5 multi-system clustering	■ **Cons:** Very expensive ($5,499+); 512GB tier delayed until late Oct; Apple Upgrade lease at $110/mo
RTX Spark (Windows)	■ **Pros:** 128GB unified memory on consumer PCs; full CUDA ecosystem (TensorRT, PyTorch, Ollama); Windows ML + NVIDIA OpenShell for local agents; multiple OEM options	■ **Cons:** 273 GB/s bandwidth → slow dense 70B decode (~5 tok/s); Arm compatibility still maturing (Prism emulation); Fall 2026 launch
DGX Spark Desktop	■ **Pros:** Load ~200B models; full CUDA datacenter parity; tiny desktop form factor; low power (~140W), cheap electricity (~$49/yr)	■ **Cons:** Slowest decode (273 GB/s caps dense 70B at ~5 tok/s); requires NVIDIA-optimized stack; not ideal for interactive chat speed

Key Takeaways

- **Memory bandwidth > raw TOPS** for local LLM decode — tokens/sec ≈ bandwidth ÷ model size [3]
- **Capacity is the scarce resource** in 2026 due to DRAM/GDDR7 shortage; 128GB unified systems let you run models 32GB cards can't hold [3][6]
- **Best single-box 70B speed**: RTX PRO 6000 (~32 tok/s) or Mac Studio M3 Ultra (~16–22 tok/s); M5 Max estimated ~12–18 tok/s [3][5]
- **Best for CUDA/homelab tooling**: RTX Spark or DGX Spark — capacity + full NVIDIA stack, but accept slower dense decode [3][6]
- **Most efficient on-device privacy**: Mac Studio M5 Max — 128GB, near-silent, ~$28/yr electricity [3][5]

Sources: [3] Digital Applied — Best Hardware to Run Local AI in 2026 (Jun 28, 2026) · [5] Apple — Mac Studio M5 Max/Ultra Press Release (Aug 25, 2026) · [6] NVIDIA/Microsoft — RTX Spark Announcement (May 31, 2026)