Unified Memory Systems — 128 GB+ RAM Comparison

Recent releases — Pricing & availability — Pros & cons (2026)

Executive Summary

Memory bandwidth dominates decode speed — not raw TOPS. tokens/sec ≈ bandwidth / model_size_in_GB. In 2026's DRAM shortage, **capacity and bandwidth-per-dollar** decide the buy. Four systems with 128 GB+ unified memory: Mac Studio M4/M3 Ultra (Apple), DGX Spark (NVIDIA), and MacBook Pro M5 Max. Each trades speed for flexibility or CUDA for capacity.

Side-by-Side Comparison

System	Max Unified RAM	Bandwidth	~Price (128GB+)	Best For
Mac Studio M4 Max	128 GB	546 GB/s	~$3,999–$4,499	Budget + macOS versatility
Mac Studio M3 Ultra	512 GB	819 GB/s	~$3,999 (96 GB) / ~$5,999+ (512 GB)	Largest model capacity; fastest decode
NVIDIA DGX Spark	128 GB (fixed)	273 GB/s	~$4,699 street	CUDA parity; cloud deployment
MacBook Pro M5 Max	128 GB	614 GB/s	~$3,899+	Portable on-device AI; privacy

Apple Mac Studio M4 Max

Pros:

- Lowest entry into unified-memory AI at $1,999 base; 128 GB undercuts competitors
- Full macOS workstation — Final Cut, Xcode, Logic Pro alongside AI
- Near-silent, very power-efficient (~$28/yr electricity)
- 128 GB matches DGX Spark capacity at lower absolute price

Cons:

- No CUDA support — blocks TensorRT-LLM, vLLM, cloud-GPU-native tooling
- Weaker long-context prefill throughput than NVIDIA (compute-bound step)
- 128 GB is the max on this chip — no higher tier available

Apple Mac Studio M3 Ultra

Pros:

- Highest memory ceiling — 512 GB loads 200B-class models locally (4x DGX Spark)
- Fastest token generation — 819 GB/s bandwidth (~16–22 tok/s on dense 70B)
- Best price-per-GB at ~$12/GB for 512 GB config
- Full macOS workstation versatility

Cons:

- Expensive at max capacity; 512 GB tiers intermittently unavailable (DRAM shortage)
- Higher power draw (~270 W peak) and heat under sustained loads
- Same no-CUDA limitation as all Apple Silicon

NVIDIA DGX Spark

Pros:

- Full CUDA parity — TensorRT-LLM, PyTorch run identically on desk and cloud GPUs
- Strong prefill throughput on long prompts (Blackwell Tensor Cores)
- ConnectX-7 networking (200 Gb/s) pools two units to 256 GB combined
- Compact desk appliance; quiet for sustained workloads

Cons:

- Fixed 128 GB soldered — no upgrade path; can't match Mac Studio capacity

- Slowest token generation (~5–15 tok/s on 70B with common tools) due to 273 GB/s
- Street pricing inflated to ~$4,699 by 2026 DRAM shortage (MSRP $3,999)
- Single-purpose AI appliance — not a general daily computer

Apple MacBook Pro M5 Max

Pros:

- Only laptop with 128 GB unified memory — portable on-device AI
- Efficient (~$28/yr electricity), near-silent operation
- Runs 70B passably at estimated 12–18 tok/s
- Privacy-first — all inference stays on-device

Cons:

- Portable = higher thermal/power limits than desktop equivalents
- Same no-CUDA limitation
- 128 GB cap on M5 Max chip

Key Takeaways

1. **1. Bandwidth > TOPS** for decode — tokens/sec ≈ bandwidth / model_size_in_GB
2. **2. Capacity vs. Speed tradeoff** — Mac Studio M3 Ultra offers 512 GB but at high cost; DGX Spark has fixed 128 GB with better prefill via Blackwell Tensor Cores
3. **3. 2026 DRAM/GDDR7 shortage** inflated prices across all categories — DGX Spark street ~$4,699, NVIDIA PRO 6000 street $12k–$14.5k
4. **4. Choose by your stack:** CUDA/research/fine-tuning → **DGX Spark**; max local models + macOS → **Mac Studio M3 Ultra**; budget + portable → **MacBook Pro M5 Max** or **Mac Studio M4 Max**

Sources: Digital Applied — 'Best Hardware to Run Local AI Models in 2026' (Jun 28, 2026); Tech Insider — 'DGX Spark vs Mac Studio 2026' (Jul 30, 2026). Prices reflect late-2026 street figures adjusted for DRAM shortage.