Recently Released Unified-Memory Systems (128 GB+)

Comparison of price, availability, and pros/cons | Last updated: 26 October 2026

Executive Summary

- In 2026, AI hardware is dominated by on-chip/unified memory — not discrete RAM. Four systems shipping with 128 GB+ unified memory were compared:
- • **AMD Gorgon Halo (Ryzen AI Max+ PRO 495)** leads on memory ceiling (192 GB, 160 GB GPU-addressable) and enables 300-billion-parameter model runs, but OEM pricing is volatile ($6,399–$9,249) and it's Windows-on-ARM.
- • **Apple M5 Max (16" MacBook Pro)** has the highest memory bandwidth (614 GB/s) and best-in-class macOS ecosystem, but 128 GB is an upgrade on top of the 36 GB base and macOS walls off CUDA.
- • **NVIDIA DGX Spark 128GB** is the most feature-complete AI workstation (128 GB, 273 GB/s, Linux, up to 70B fine-tuning), but its 128 GB SKU jumped ~75% in price to $6,950.
- • **NVIDIA RTX Spark (Windows partners)** offers the lowest entry price ($2,599+) and CUDA access, but suffers thermal throttling and limited large-model fit under sustained load.
- Key takeaway: memory prices are at an all-time high through 2026–2027 due to DRAM/HBM supply crunch. If you need the most memory, the Gorgon Halo is best — but watch OEM pricing volatility. Apple M5 Max offers the best bandwidth and ecosystem; DGX Spark is best for Linux-native local AI work.

At a Glance — Quick Reference

System	Memory	Bandwidth	Price Range	OS	Availability
AMD Gorgon Halo (Ryzen AI Max+ PRO 495)	192 GB unified (160 GB GPU)	~273 GB/s	$6,399–$9,249	Windows (on ARM)	Sept 2026; GMKtec, Minisforum, TOPC + AMD
Apple M5 Max (16" MacBook Pro)	128 GB unified (upgrade from 36 GB)	614 GB/s	~$3,499–$4,999	macOS	Apple retail + major retailers
NVIDIA DGX Spark 128 GB	128 GB unified	273 GB/s	$6,950	Linux (DGX OS)	Oct 2026; Acer, ASUS, Dell, Gigabyte, HP, MSI
NVIDIA RTX Spark (Windows partners)	up to 128 GB unified	~300 GB/s	$2,599–$6,229+	Windows (CUDA)	Oct 2026; ASUS, Dell, HP, Lenovo, Microsoft, MSI

1. AMD Gorgon Halo — Ryzen AI Max+ PRO 495

Architecture: Ryzen AI Max+ PRO 495 (16 Zen 5 cores / 32 threads, Radeon 8065S GPU RDNA 3.5, XDNA 2 NPU 55 TOPS, 131 TOPS total). 192 GB LPDDR5X-8533 unified memory, up to 160 GB dynamically assignable to GPU, ~273 GB/s. Announced Sept 28, 2026.

Pros:

- Highest memory ceiling in the comparison: 192 GB total, 160 GB GPU-addressable — enables 300B-parameter model runs at 4-bit quantization
- Windows-based (Ryzen AI Halo platform); ROCm available (less mature than CUDA but accessible)
- 55 TOPS NPU / 131 TOPS total platform AI performance
- Lower floor price at ~$6,399 — cheaper than DGX Spark 128GB at $6,950
- 50% larger memory pool and 67% larger GPU-addressable ceiling vs predecessor Strix Halo

Cons:

- Wide OEM pricing spread ($6,399–$9,249 on identical silicon) — pricing volatile, not AMD-standard MSRP
- Windows on ARM — no native CUDA; ROCm maturity lags CUDA
- Only three OEMs confirmed early in launch (GMKtec, Minisforum, TOPC)
- No published detailed specs for 490/485 variants in the family

2. Apple M5 Max — 16" MacBook Pro

Architecture: M5 Max chip, 18-core CPU (6 super + 12 performance cores), up to 40-core GPU, 128 GB unified memory (M5 Max base is 36 GB, upgradable to 128 GB), up to 614 GB/s bandwidth, 2 TB SSD base. macOS Tahoe.

Pros:

- Highest memory bandwidth of the set: 614 GB/s unified memory
- Best-in-class integrated GPU (up to 40-core M5 Max) for local AI
- macOS ecosystem with Apple Intelligence / on-device AI integration

- Premium build quality, 16.2" XDR display, 22-hour battery life
- Significant promotions bring 128GB config down to ~$3,499 from $4,999 MSRP

Cons:

- macOS walls off CUDA/ROCm — no native Windows or CUDA ecosystem
- 128 GB is an upgrade on top of 36 GB base — you pay for a smaller starting point then upgrade
- Absolute price remains higher than DGX Spark 64 GB sibling and AMD floor tier
- Unified memory architecture ties CPU and GPU together — no independent discrete GPU to leverage

3. NVIDIA DGX Spark 128 GB

Architecture: GB10 Grace Blackwell Superchip (Arm Grace CPU + Blackwell GPU, shared unified memory pool), 128 GB unified memory, 273 GB/s bandwidth, 4 TB SSD, DGX OS / Linux. Announced Oct 2025 at $3,999; 64 GB sibling introduced Oct 2, 2026 at $4,999.

Pros:

- Most feature-complete AI workstation: Linux-native, DGX OS, NVIDIA AI stack
- 273 GB/s bandwidth with 128 GB — fine-tunes models up to ~70B parameters
- Up to 8-unit cluster support, 1 TB total unified memory combined
- Strong thermal/architecture design suited for sustained AI workloads

Cons:

- Price surge is severe (+~75% from original launch) — no official reason given, tied to industry memory supply crunch
- Linux-only — no native Windows/CUDA access
- Available only through OEM partners, no direct NVIDIA retail
- 128 GB ceiling is still below AMD's 192 GB and above Apple's 128 GB upgrade tier

4. NVIDIA RTX Spark — Windows Partner Systems

Architecture: N1X SoC (MediaTek-derived CPU + Blackwell GPU, up to 20 CPU cores, 6,144 CUDA cores in top SKU), up to 128 GB unified memory, ~300 GB/s, Windows, CUDA ecosystem. Partner systems announced Oct 16, 2026.

Pros:

- Lowest entry price ($2,599) — most accessible in this category
- ~300 GB/s unified bandwidth, slightly above DGX Spark's 273 GB/s
- Windows + CUDA — most mature AI toolchain among the set
- Serious local AI benchmarks: HP OmniBook Ultra outperforms DGX Spark on token throughput
- Lightweight, premium-built notebook form factor for prosumer buyers

Cons:

- Thermal throttling under sustained agentic AI load: GPU temps exceeded 90°C, causing lockups and graphical artifacts
- Windows memory allocation is less flexible than Linux — no dynamic GPU memory allocation, limiting large-model fits
- Limited model availability for large local models (e.g., 200B+)
- Pricing not competitive at current memory prices — base $3,200, tested $6,229
- Only available through partner OEMs, not direct NVIDIA retail

---

Sources: NVIDIA newsroom & The Register, AMD Ryzen AI Max PRO 400 Series, Apple press releases & product pages, Tech Insider (HUMAIN/Qualcomm), Shattered.io, MacPrices.net, and The Register benchmarking (Oct 2026). All prices as of 26 October 2026 unless noted. Unified memory = on-chip/SoC-level memory, not discrete RAM. AMD Gorgon Halo OEM prices range due to OEM-dependent pricing; AMD does not publish a single MSRP for the chip.