

Apple Mac Studio (M5 Max / M5 Ultra)

Also known as: Mac Studio 2026 · Mac Studio M5 Ultra · Mac Studio M5 Max · M5 Ultra 512GB

The 2026 Mac Studio is a 19.7 cm square desktop built on Apple's M5 Max or M5 Ultra SoC, with CPU, GPU, Neural Engine and up to 512 GB of unified memory sharing one memory pool at up to 1.2 TB/s. For inference that combination is the point: a single quiet box can hold model weights that would otherwise need several datacenter GPUs, and its bandwidth is several times that of the 128 GB LPDDR5x desk boxes. The trade is the software stack. There is no CUDA; models run through MLX, LM Studio and other Metal backends, and multi-box scaling goes over Thunderbolt 5 with RDMA, which Apple has supported only since macOS 26.2. Buy it for large-model, low-concurrency local inference and development; do not plan a CUDA serving stack or a rack deployment around it.

Up to 512 GB Unified memory · M5 Ultra; 512 GB option listed as coming late October 2026, 256 GB orderable now	1.2 TB/s Memory bandwidth · M5 Ultra; M5 Max up to 614 GB/s	Up to 80-core GPU · M5 Ultra, Neural Accelerators in each GPU core
32-core Neural Engine · M5 Ultra; 16-core on M5 Max	Thunderbolt 5 with RDMA Clustering · 6x TB5 on M5 Ultra, up to 120 Gb/s each; macOS 26.2 or later	385 W max, 9 W idle Power · Apple wall measurement, M5 Ultra 80-core GPU, 512 GB, 16 TB

COMPUTE		01/05
M5 Ultra CPU	30-core or 36-core · base 10 super + 20 performance cores	
M5 Ultra GPU	64-core or 80-core · hardware ray tracing, Neural Accelerators per core	
M5 Max CPU	18-core · 6 super + 12 performance cores	
M5 Max GPU	32-core or 40-core · Neural Accelerators per core	
Neural Engine	32-core (Ultra), 16-core (Max)	
Tensor throughput (TFLOPS/TOPS)	Not published · Apple gives only relative claims, e.g. up to 4.3x the peak AI compute of M3 Ultra (vendor claim)	

MEMORY & STORAGE		02/05
M5 Ultra memory	96 GB, 256 GB, 512 GB · 512 GB listed as coming late October 2026	
M5 Ultra bandwidth	1.2 TB/s · Apple: 50 percent higher than the prior generation (vendor claim)	
M5 Max memory	36, 48, 64 or 128 GB	
M5 Max bandwidth	460 GB/s (32-core GPU), 614 GB/s (40-core GPU) · 40-core GPU requires at least 48 GB	
Memory model	Unified, shared by CPU, GPU and Neural Engine · no separate VRAM; usable model size is capacity minus OS and app footprint	
SSD	Up to 16 TB (Ultra), up to 8 TB (Max)	
CONNECTIVITY & CLUSTERING		03/05
Thunderbolt 5	6 ports (Ultra), 4 ports (Max) · up to 120 Gb/s; Max has 2 front USB-C at 10 Gb/s instead	
RDMA over Thunderbolt 5	Supported from macOS 26.2 · enabled per machine; MLX docs say from macOS recovery with rdma_ctl	
Cluster library	JACCL (Apple, open source) · MLX backend; MLX docs: fully connected topologies only	
Apple-tested cluster	4 nodes, full mesh · up to 3x faster distributed inference than one system (vendor claim); max node count not published	
Ethernet	1x 10 GbE RJ-45 · 1, 2.5, 5, 10 Gb	
Other I/O	2x USB-A, HDMI 2.1, SDXC, 3.5 mm · Wi-Fi 7, Bluetooth 6 (Apple N1)	
POWER & PHYSICAL		04/05
Max continuous power rating	480 W · 100 to 240 V AC internal supply	
Measured max / idle (Ultra)	385 W / 9 W · wall, 1,314 / 31 BTU/hr, 36-core CPU, 80-core GPU, 512 GB, 16 TB	
Measured max / idle (Max)	200 W / 7 W · wall, 682 / 24 BTU/hr, base 32-core GPU, 36 GB	
Dimensions	9.5 x 19.7 x 19.7 cm · H x W x D	
Weight	3.6 kg (Ultra), 2.7 kg (Max)	
Operating temperature	10C to 35C	

SOFTWARE		05/05
OS	macOS · no Linux or Windows option from Apple	
CUDA	Not supported · Apple GPU, Metal only	
ML frameworks	MLX, MLX LM, Core AI · MLX is Apple's open-source framework; Core AI is new in 2026	
Runtimes named by Apple	LM Studio, exo · exo for clustering multiple machines	
PROCUREMENT		
Typical lead time M5 Max and M5 Ultra standard configurations ship from Apple's store and retail (customer arrivals began September 22, 2026). The 512 GB M5 Ultra option is not yet orderable: Apple's store lists it as coming late October 2026, and Macworld reports orders slipping to late January, reported. Pricing Apple US store, 2026-10-14: M5 Max from \$2,499; M5 Ultra (30-core CPU, 64-core GPU, 96 GB, 1 TB) \$5,499. Options: 36-core CPU / 80-core GPU +\$1,300, 256 GB +\$4,000, 16 TB SSD +\$7,500. Top orderable configuration (80-core GPU, 256 GB, 16 TB) is \$18,299; the inference-relevant 80-core GPU, 256 GB, 1 TB build is \$10,799. 512 GB price: not published.		
- Storage is the expensive line, not memory: 16 TB adds \$7,500. For inference, keep weights on external or networked storage and buy the 1 TB or 2 TB SSD. - The 512 GB option is announced but unpriced; do not budget a 512 GB cluster until Apple lists it. - Memory and SSD are configured at order time; plan the final capacity before purchase. - Education pricing (\$5,099 for the M5 Ultra base) and Apple Upgrade leasing (from \$110.10 per month for M5 Ultra) exist; enterprise volume terms are quote-only.		
FIELD NOTES		
- No CUDA is the real constraint. CUDA-only serving stacks, custom CUDA kernels and NCCL-based multi-GPU code do not run here; the working paths are MLX and MLX LM, LM Studio and other Metal backends. Check that your model and quantization format are supported in MLX before you buy. - Memory bandwidth decides decode speed. At 1.2 TB/s the M5 Ultra has several times the bandwidth of 273 GB/s GB10 boxes, and reviewers report about 55 tokens/s on a 4-bit 27B model and about 80 tokens/s on a 122B-A10B mixture-of-experts model in LM Studio, reported. It is a single-user or small-team box, not a high-concurrency serving node. - RDMA over Thunderbolt 5 is new and still rough. MLX's JACCL backend needs a full mesh with a cable between every pair, and RDMA must be enabled locally from recovery, not remotely. Users have reported about 3.5 to 3.8 GB/s point-to-point on M3 Ultra, roughly 37 percent of theoretical single-port bandwidth, plus a GPU-stream timeout and resource leaks that needed a reboot, reported. - Apple's 3x for four boxes is a vendor claim measured on one 72B dense model with a 32K prompt. A cluster lets you hold a model that one box cannot (Apple demonstrated a 1-trillion-parameter model on four M3 Ultras), but tensor parallelism over Thunderbolt does not scale like NVLink. - Efficient and quiet, but a desktop. Apple measured 385 W at the wall under maximum load and 9 W idle for the top M5 Ultra configuration, so it fits an office circuit. It is macOS only, with no rack-mount version and no out-of-band management port listed in its specs, so fleet management is up to you.		

QUESTIONS WE GET ON THIS PART

How much does a Mac Studio for AI cost?

On Apple's US store on 2026-10-14, the M5 Ultra starts at \$5,499 (96 GB). With the 80-core GPU and 256 GB of memory it is \$10,799, and the top orderable configuration with a 16 TB SSD is \$18,299. The 512 GB memory option is listed as coming late October 2026 and has no published price.

Does the Mac Studio support CUDA?

No. The GPU is Apple silicon, programmed through Metal. Local inference and fine-tuning run on MLX, MLX LM, LM Studio and similar Metal-backed tools; CUDA-only code has to be ported or run elsewhere.

Can I cluster several Mac Studios?

Yes. From macOS 26.2, Apple supports RDMA over Thunderbolt 5, and MLX uses it through the JACCL backend. Apple tested a four-machine full mesh and claims up to 3x faster distributed inference than one machine (vendor claim). JACCL needs a direct cable between every pair of machines, so the cable count grows quickly.

M5 Max or M5 Ultra for local LLMs?

The M5 Max tops out at 128 GB and 614 GB/s (40-core GPU), which suits models that fit in roughly 100 GB. The M5 Ultra offers 256 GB now, 512 GB announced, and 1.2 TB/s, so it is the choice for the largest open-weight models and for faster decode on large dense models.

REFERENCES

[1] <https://www.apple.com/mac-studio/specs/>
[2] <https://www.apple.com/mac-studio/>
[3] <https://www.apple.com/newsroom/2026/08/apple-introduces-new-mac-studio-with-m5-max-and-m5-ultra/>
[4] <https://www.apple.com/shop/buy-mac/mac-studio>
[5] <https://www.apple.com/shop/buy-mac/mac-studio/m5-ultra-chip-30-core-cpu-64-core-gpu-96gb-memory-1tb-storage>
[6] <https://support.apple.com/en-us/102027>
[7] <https://developer.apple.com/videos/play/wwdc2026/233/>
[8] <https://ml-explore.github.io/mlx/build/html/usage/distributed.html>
[9] <https://github.com/ml-explore/mlx/issues/3207>
[10] <https://www.bgr.com/2264114/mac-studio-m5-ultra-review/>
[11] <https://www.macworld.com/article/2973459/2026-mac-studio-m5-release-date-specs-price-rumors.html>