M4 Max and M3 Ultra for Local LLMs: Apple Silicon in 2026

February 21, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** The M4 Max Mac Studio (from $1,999) runs 70B models in its 128GB configuration at ~15-18 tok/s via MLX. But there's no M4 Ultra — Apple skipped it. The Mac Studio's top-end option is the M3 Ultra ($3,999+), a 2024 chip with 800 GB/s bandwidth and up to 192GB unified memory. For most local AI builders, the M4 Max with 128GB is the sweet spot: current-gen architecture, enough memory for 70B models, 50-80W power draw, near-silent operation. The M3 Ultra is for people who need 192GB — running 120B+ models or loading multiple 70B models simultaneously. Use MLX, not llama.cpp. It's 30-50% faster on Apple Silicon.

📚 **More on this topic:** Best Local LLMs for Mac · Mac vs PC for Local AI · Running LLMs on Mac M-Series · VRAM Requirements

If you follow Apple Silicon and local AI, you were expecting 2025 to bring the M4 Ultra — a 256GB+ chip that would make the Mac Studio the definitive local AI workstation. It didn't happen. The M4 Max chip lacks UltraFusion support, the die-to-die interconnect that combines two Max chips into one Ultra. Apple hasn't said whether this is permanent or just delayed.

What they shipped instead: a Mac Studio with the M4 Max (current gen, up to 128GB) starting at $1,999, and the M3 Ultra (previous gen, up to 192GB) starting at $3,999. A newer chip with less memory, or an older chip with more. For local AI, that tradeoff matters more than any benchmark.

The Specs That Actually Matter

Spec	M4 Max (40-core GPU)	M3 Ultra (60-core GPU)
Memory Bandwidth	**546 GB/s**	**800 GB/s**
Max Unified Memory	**128GB**	**192GB**
CPU Cores	16 (12P + 4E)	28 (20P + 8E)
GPU Cores	40	60

Spec	M4 Max (40-core GPU)	M3 Ultra (60-core GPU)
Neural Engine	16-core	32-core
Power (AI load)	50-80W	100-200W
Mac Studio Price	**From $1,999**	**From $3,999**

Memory bandwidth determines LLM inference speed. Every token requires reading through the model's weights in memory. More bandwidth, more tokens per second. More GPU cores barely help for this workload.

The M3 Ultra has 47% more bandwidth than the M4 Max. That translates directly to ~40-50% faster token generation on the same model. But the M4 Max has a newer architecture with better per-core IPC and Thunderbolt 5. For prompt processing — the initial computation before tokens start flowing — the M4 Max's architectural improvements partially close the gap.

This is the same reason an M3 Max generates tokens faster than an M4 Pro despite being older. Bandwidth trumps architecture generation for this workload.

M4 Max: Current Gen, 70B Capable

The M4 Max Mac Studio comes in two GPU configurations:

- **$1,999:** 14-core CPU, 32-core GPU, 36GB, 512GB SSD
- **$2,499:** 16-core CPU, 40-core GPU, 48GB, 1TB SSD
- **BTO:** Up to 128GB unified memory, up to 8TB SSD

The 128GB configuration is what matters for serious local AI. It's the first sub-$4,000 Apple machine that comfortably runs 70B models.

M4 Max Benchmarks (MLX, Q4 Quantization)

Model	VRAM Needed	128GB Config	64GB Config	Notes
Llama 3.2 8B	~5 GB	~90-100 tok/s	~90-100 tok/s	Way beyond reading speed
Qwen3 14B	~9 GB	~55-65 tok/s	~55-65 tok/s	Sweet spot quality at great speed

Model	VRAM Needed	128GB Config	64GB Config	Notes
Qwen3 32B	~20 GB	~30-40 tok/s	~30-40 tok/s	Expert-level, still smooth
Llama 3.3 70B	~40 GB	~15-18 tok/s	Tight fit	Needs 128GB; fast enough for chat
DeepSeek V3.2 MoE (Q3)	~90 GB	Fits (slow)	Does not fit	Experimental — ~5-8 tok/s

These numbers use MLX, Apple's native ML framework. Ollama (which uses llama.cpp) would be 30-50% slower on the same hardware. More on that below.

At 15-18 tok/s, a 70B model is genuinely usable for interactive chat. Not blazing, but faster than you read. For 8B-32B models, the M4 Max is overkill on speed — you're buying it for the 70B capability and the headroom.

M3 Ultra: More Memory, More Bandwidth, Last Gen

The M3 Ultra Mac Studio starts at $3,999 with a 28-core CPU, 60-core GPU, and configurable up to 192GB unified memory. The chip itself is from 2024 — the M3 generation — but its raw bandwidth advantage over the M4 Max is significant.

M3 Ultra Benchmarks (MLX, Q4 Quantization)

Model	VRAM Needed	192GB Config	Notes
Llama 3.3 70B	~40 GB	~25-30 tok/s	Comfortably fast, room for large context
Qwen3 72B	~43 GB	~22-28 tok/s	Excellent reasoning speed
Llama 4 Maverick MoE (Q3)	~80 GB	~12-16 tok/s	Fits; usable for coding and analysis
120B+ models (Q4)	~70-80 GB	~10-15 tok/s	M3 Ultra territory — won't fit on M4 Max
Multiple models loaded	—	Yes	Run 70B + 32B simultaneously

The M3 Ultra's advantage is raw capacity and speed. 192GB at 800 GB/s means you can run models that won't fit on an M4 Max, and run everything else 40-50% faster. Running a 70B model at 25-30 tok/s versus 15-18 tok/s is a noticeable difference in interactive chat.

The question is whether that's worth $2,000+ more and a previous-generation architecture.

MLX: The Software That Makes It Work

If you're running LLMs on Apple Silicon using Ollama's default llama.cpp backend, you're leaving 30-50% of your speed on the table.

MLX is Apple's machine learning framework, purpose-built for the Metal GPU API and unified memory architecture. The performance gap is consistent and well-documented:

- **30-50% faster token generation** than llama.cpp across model sizes
- **Up to 2.5x faster** on specific models (Qwen3 32B on some configs)
- **LM Studio** now defaults to MLX on Mac — the easiest way to get the boost
- **MLX-LM** (command line) gives direct access for power users

The reason is focus. llama.cpp supports dozens of hardware backends — NVIDIA, AMD, Intel, CPU, Vulkan, Metal. MLX only supports Apple Silicon. That specialization pays off in every benchmark.

For a $2,000-$4,000 Mac Studio, the difference between 15 tok/s (llama.cpp) and 22 tok/s (MLX) on a 70B model is the difference between "usable" and "comfortable." Use MLX.

Mac Studio vs PC for Local AI

Factor	Mac Studio M4 Max (128GB)	PC + RTX 4090 (24GB)	PC + 2× RTX 3090 NVLink (48GB)
Memory	128GB unified	24GB VRAM	48GB VRAM (pooled)
Bandwidth	546 GB/s	1,008 GB/s	~1,872 GB/s
Largest model (Q4)	70B comfortable	32B max	70B via NVLink
70B Q4 speed	~15-18 tok/s	Won't fit	~25-35 tok/s
Power under AI load	50-80W	350-450W system	700W+ system
Noise	Near silent	Fan noise	Significant fan noise
Price	~$3,000-$3,500	~$2,200 GPU + $800 PC	~$1,800 GPUs + $800 PC

The Mac advantage: memory capacity. 128GB of unified memory runs 70B models natively. On PC, that requires dual GPUs with NVLink — which limits you to the used RTX 3090, the only consumer card that supports it — or $10,000+ datacenter GPUs. The Mac also runs near-silent at a fraction of the power.

The PC advantage: bandwidth. An RTX 4090's 1,008 GB/s is nearly 2x the M4 Max. For models that fit in 24GB VRAM, discrete GPUs are substantially faster. A dual 3090 NVLink setup pushes ~1,872 GB/s combined — 3.4x the M4 Max.

The comparison breaks down at model size boundaries. The Mac doesn't win on speed. It wins on what fits. See our Mac vs PC guide for the full breakdown.

→ Check what fits your hardware with our Planning Tool.

Power and Silence

This is where Apple Silicon genuinely has no competition.

Machine	AI Workload	Idle	Annual Cost (8hr/day)
Mac Studio M4 Max	50-80W	5-8W	~$42-63
Mac Studio M3 Ultra	100-200W	10-15W	~$79-158
PC + RTX 4090	350-450W	60-80W	~$276-355
PC + 2× RTX 3090	700-800W	80-100W	~$552-631

At US average electricity rates ($0.18/kWh). A Mac Studio M4 Max running a 70B model draws about 60W. A dual-3090 PC doing the same work draws 700W+. The annual electricity difference is $400-$500. Over three years, that's $1,200-$1,500 — enough to significantly offset the Mac's higher hardware cost.

And the Mac Studio at 60W under AI load is near-silent. You can run it as an always-on inference server in a bedroom or office without hearing it. Try that with dual RTX 3090s.

The Verdict

Buy the M4 Max Mac Studio (128GB) if:

- You want to run 70B models on Apple Silicon — get the 128GB config

- You want current-gen with Thunderbolt 5
- Budget is $2,000-$3,500
- You value silence and low power for always-on inference
- 128GB covers your needs (it does for 95% of people)

Buy the M3 Ultra Mac Studio if:

- You need 192GB for 120B+ models or multiple simultaneous models
- Raw inference speed on 70B matters (40-50% faster than M4 Max)
- Budget is $4,000+
- You're building a workstation that needs to handle anything

Skip both if:

- Your workload is 8B-32B models — the M4 Pro Mac Mini is half the price
- Speed per token matters more than capacity — discrete GPUs have 2-3x more bandwidth
- You primarily do image generation or fine-tuning — compute-bound workloads favor NVIDIA

The M4 Max Mac Studio at $1,999 is the most accessible 70B-capable Apple machine ever shipped. The M3 Ultra at $3,999 is for people who need headroom beyond 128GB. The missing M4 Ultra? It might come with M5. Apple isn't saying.

For most local AI builders on Mac, the M4 Max 128GB is the play. Run MLX. Load a 70B model. Enjoy the silence.
