Apple’s M5 Ultra: 512GB of Unified Memory Is the Number That Matters

2026-08-26T10:37:15.548032Z

Apple’s M5 Ultra: 512GB of Unified Memory Is the Number That Matters

body{font-family:-apple-system,BlinkMacSystemFont,’Segoe
UI’,sans-serif;max-width:680px;margin:0 auto;padding:24px;color:#111;line-height:1.7}
h1{font-size:24px;font-weight:700;margin-bottom:8px;line-height:1.3}
.hero{width:100%;border-radius:8px;margin:16px 0 28px;display:block}
p{margin:0 0 18px}
a{color:#0057ff}
.source{font-size:13px;color:#888;margin-top:32px;border-top:1px solid
#eee;padding-top:16px}

Apple’s M5 Ultra: 512GB of Unified Memory Is the Number That Matters

Apple announced two chips on August 25, 2026 - the M6 and the M5 Ultra - and the figures that matter for local inference are memory, not FLOPS.

The M5 Ultra is Apple’s first quad-die design, fusing two dual-die M5 Max chips. It scales to up to 512GB of unified memory and 1.2TB/s of memory bandwidth, and Apple claims up to 4.5x the peak GPU compute for AI compared to M3 Ultra. Source: https://www.apple.com/newsroom/2026/08/apple-introduces-m6-and-m5-ultra-for-a-big-leap-in-performance-and-ai-compute/

The M6 is Apple’s first 2-nanometer chip. In the Mac mini it pairs a 12-core CPU and 12-core GPU with Neural Accelerators in every GPU core, up to 32GB of unified memory and up to 170GB/s of bandwidth. Apple states a nearly 30 percent increase in peak GPU compute for AI compared to M5.

Why memory bandwidth is the number to watch: for a quantized model that fits in unified memory, token generation is bandwidth-bound, not compute-bound. A 70B model at 4-bit needs roughly 35GB of weights read per token. At 170GB/s that ceiling is a few tokens per second; at 1.2TB/s it is an order of magnitude higher. The 512GB ceiling on M5 Ultra is what puts models in the hundreds of billions of parameters in reach of a desktop at all.

Two corrections to the numbers circulating since the announcement. There is no "M5 Pro" in this launch - the Mac Studio ships with M5 Max (up to 128GB unified memory, up to 614GB/s) and M5 Ultra. And Apple’s own comparison for the M6 is against M5, at nearly

30 percent more peak GPU compute for AI; the widely repeated "4.8x faster than M4" figure does not appear in Apple’s press release.

Starting prices were not stated in Apple’s newsroom post.

Source:
https://www.apple.com/newsroom/2026/08/apple-introduces-m6-and-m5-ultra-for-a-big-leap-in-performance-and-ai-compute/
(https://www.apple.com/newsroom/2026/08/apple-introduces-m6-and-m5-ultra-for-a-big-leap-in-performance-and-ai-compute/)

https://myspacexai.site/issues/apple-s-m5-ultra-512gb-of-unified-memory-is-the-number-that-matters