APPLE SILICON · DIE FLOORPLANS · N3B → N3E → N3P

Apple M3 to M5 Floorplans

Three generations on the same 3 nm family, seen from above. The interesting story is not core counts. It is where the matrix arithmetic moved, and then, at M5 Pro, the die splitting in two.

M3 → M4 · width, not layout

N3B to N3E, 25 billion to 28 billion transistors. Two extra efficiency cores, wider decode and better branch prediction on the performance cores, second-generation ray tracing, and a Neural Engine that roughly doubles to 38 TOPS. The floorplan is recognisably the same chip, grown.

M4 → M5 · AI moves into the GPU

N3P, and identical core counts on paper. That is the point. The gain comes from ten new matrix engines inside the GPU cores, second-generation Dynamic Caching, and 153 GB/s of bandwidth to feed them. Metal 4's Tensor API is the software half of the same change.

M5 → M5 Pro/Max · the die splits

Monolithic becomes two hybrid-bonded dies. One CPU die serves both Pro and Max, while the GPU die carries 20 or 40 cores plus all four memory controllers. This is the reticle-and-yield escape hatch that lets Apple keep scaling the GPU without growing one enormous die.

The whole family, three generations

CHIP	NODE	CPU	GPU	NEURAL ENGINE	MEMORY BW	MAX RAM	TRANSISTORS
M3	TSMC N3B	8-core · 4P + 4E	10-core	16-core · 18 TOPS	100 GB/s	24 GB	25 B
M3 Pro	TSMC N3B	12-core · 6P + 6E	18-core	16-core · 18 TOPS	150 GB/s	36 GB	37 B
M3 Max	TSMC N3B	16-core · 12P + 4E	40-core	16-core · 18 TOPS	400 GB/s	128 GB	92 B
M4	TSMC N3E	10-core · 4P + 6E	10-core	16-core · 38 TOPS	120 GB/s	32 GB	28 B
M4 Pro	TSMC N3E	14-core · 10P + 4E	20-core	16-core · 38 TOPS	273 GB/s	64 GB	n/d
M4 Max	TSMC N3E	16-core · 12P + 4E	40-core	16-core · 38 TOPS	546 GB/s	128 GB	92 B
M5	TSMC N3P	10-core · 4P + 6E	10-core + 10 NA	16-core · faster	153 GB/s	32 GB	n/d
M5 Pro	TSMC N3P × 2 dies	18-core · 6 Super + 12 Perf	20-core + 20 NA	16-core	307 GB/s	64 GB	n/d
M5 Max	TSMC N3P × 2 dies	18-core · 6 Super + 12 Perf	40-core + 40 NA	16-core	614 GB/s	128 GB	n/d

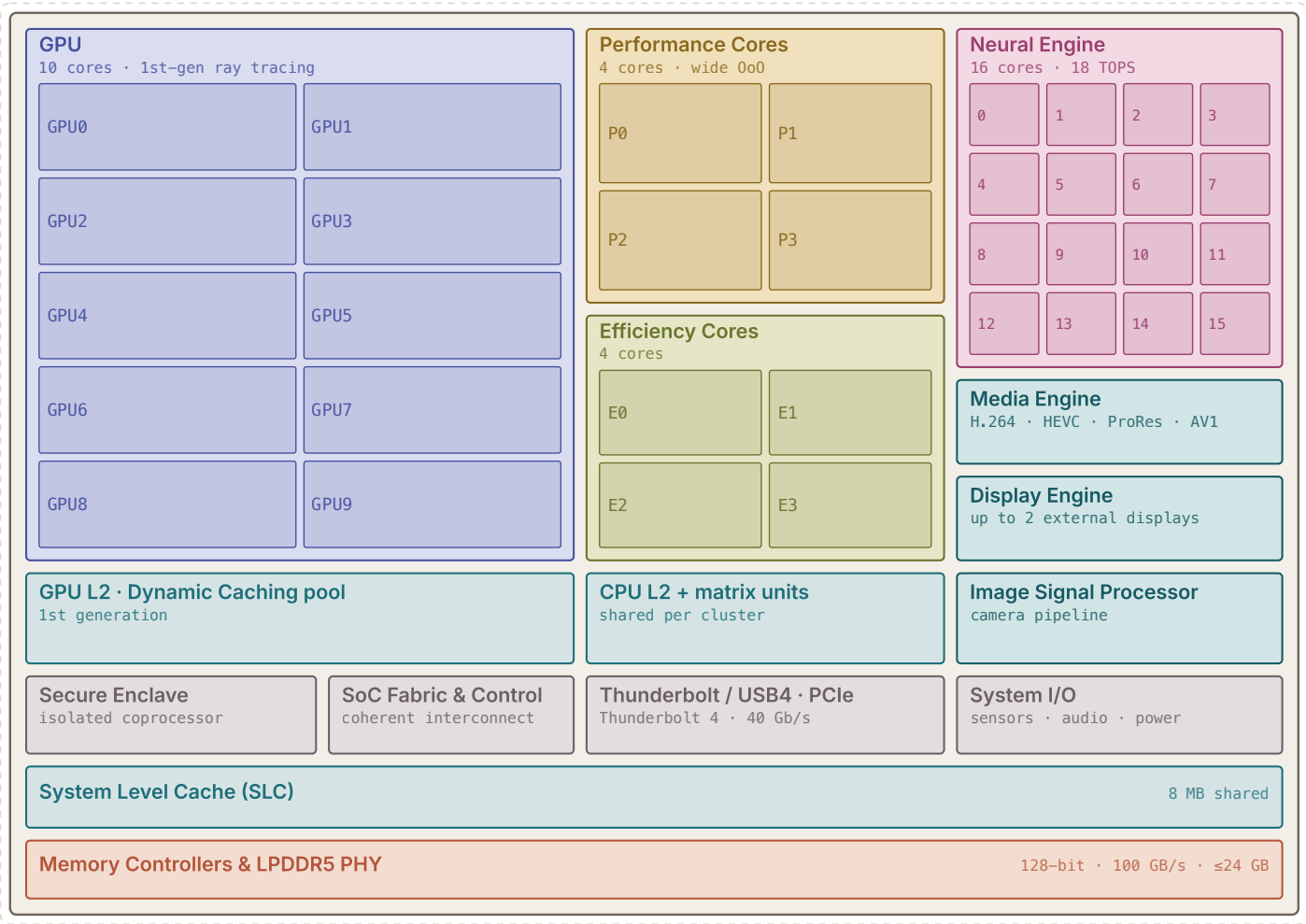
> **What this document is.** Block identities, core counts, process nodes and bandwidths come from Apple's own specifications and the Counterpoint teardown of the M5 Pro. Block positions and relative sizes in the floorplans that follow are schematic, arranged the way published die-shot annotations of M3 and M4 show Apple organising these SoCs. Nothing is traced from a photograph: read the topology as accurate and the geometry as illustrative.

M3

Released: Oct 2023 · Process: TSMC N3B · Transistors: 25 B · Die area: ~146 mm² est. · CPU: 8-core (4P + 4E) · GPU: 10-core · Bandwidth: 100 GB/s · Max memory: 24 GB

CPU · performance / super cores
CPU · efficiency / performance cores
GPU cores
Neural Engine & Neural Accelerators
Cache (L2 / SLC / Dynamic Caching)
Memory controllers & PHY
Media, display & image signal
I/O, security & fabric
Inter-die bond interface

M3 · monolithic die

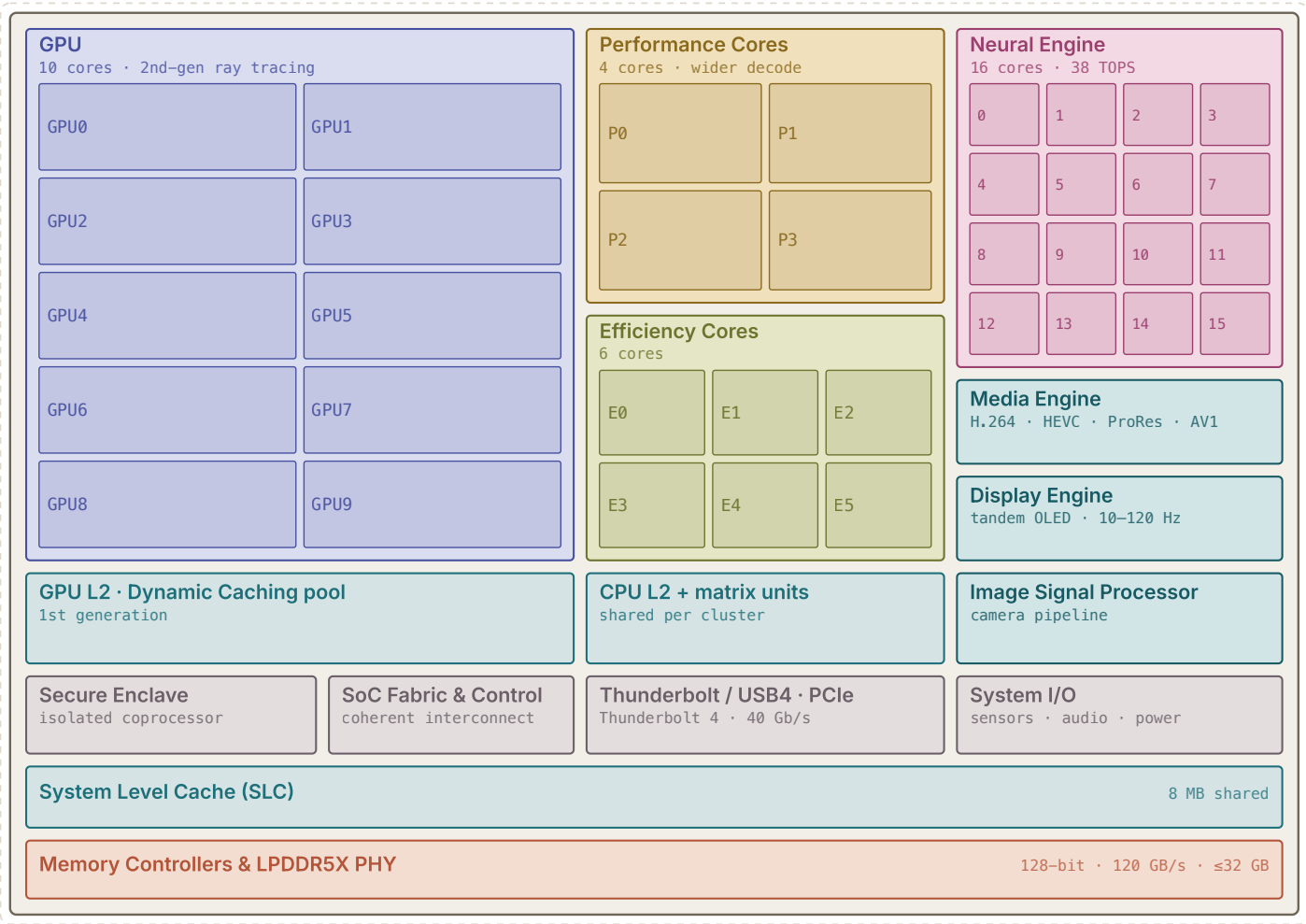


M4

Released: May 2024 · Process: TSMC N3E · Transistors: 28 B · Die area: ~165 mm² est. · CPU: 10-core (4P + 6E) · GPU: 10-core · Bandwidth: 120 GB/s · Max memory: 32 GB

CPU · performance / super cores
CPU · efficiency / performance cores
GPU cores
Neural Engine & Neural Accelerators
Cache (L2 / SLC / Dynamic Caching)
Memory controllers & PHY
Media, display & image signal
I/O, security & fabric
Inter-die bond interface

M4 · monolithic die

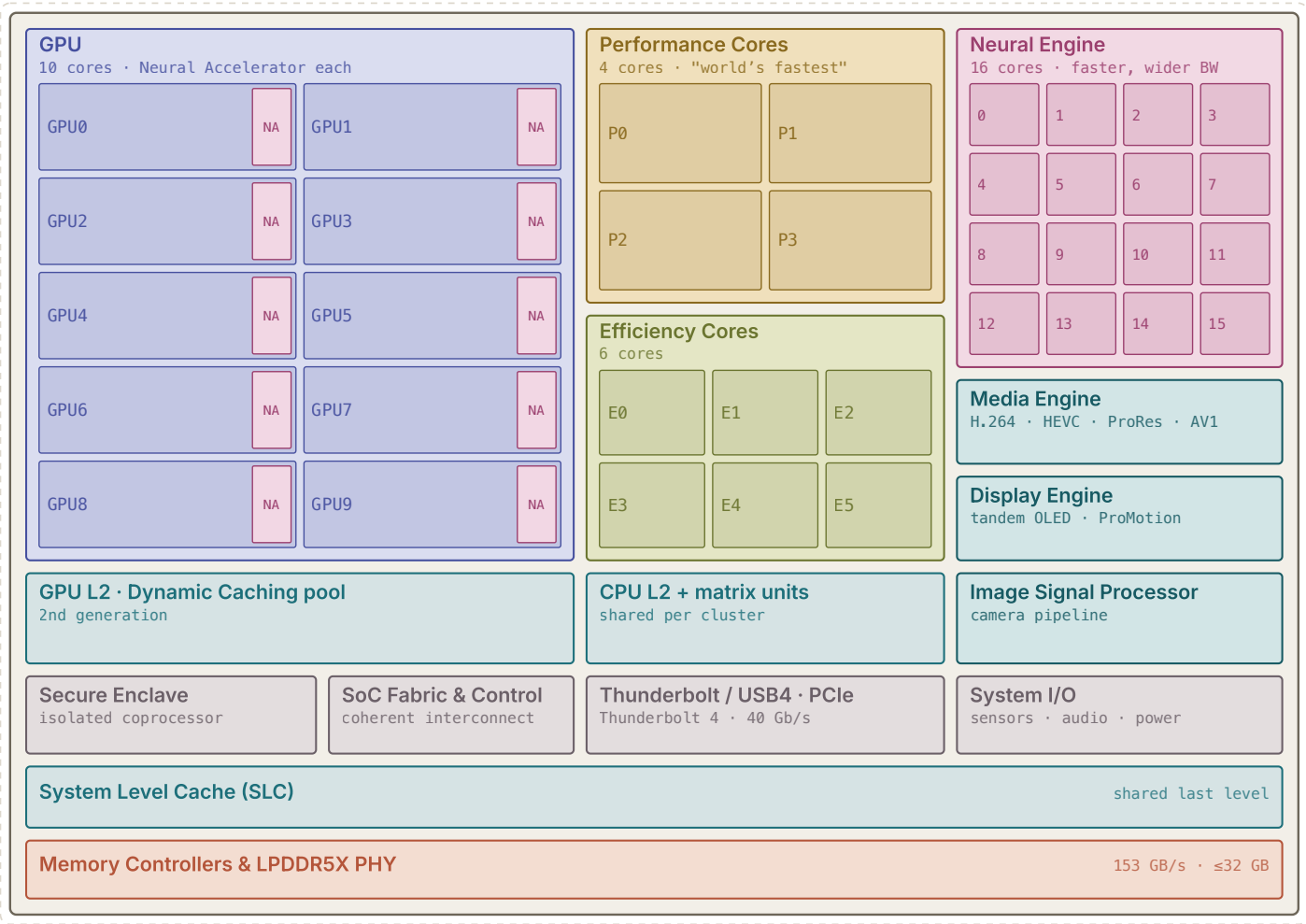


M5

Released: Oct 2025 · Process: TSMC N3P · Transistors: not disclosed · Die area: not disclosed · CPU: 10-core (4P + 6E) · GPU: 10-core + 10 NA · Bandwidth: 153 GB/s · Max memory: 32 GB

CPU · performance / super cores
CPU · efficiency / performance cores
GPU cores
Neural Engine & Neural Accelerators
Cache (L2 / SLC / Dynamic Caching)
Memory controllers & PHY
Media, display & image signal
I/O, security & fabric
Inter-die bond interface

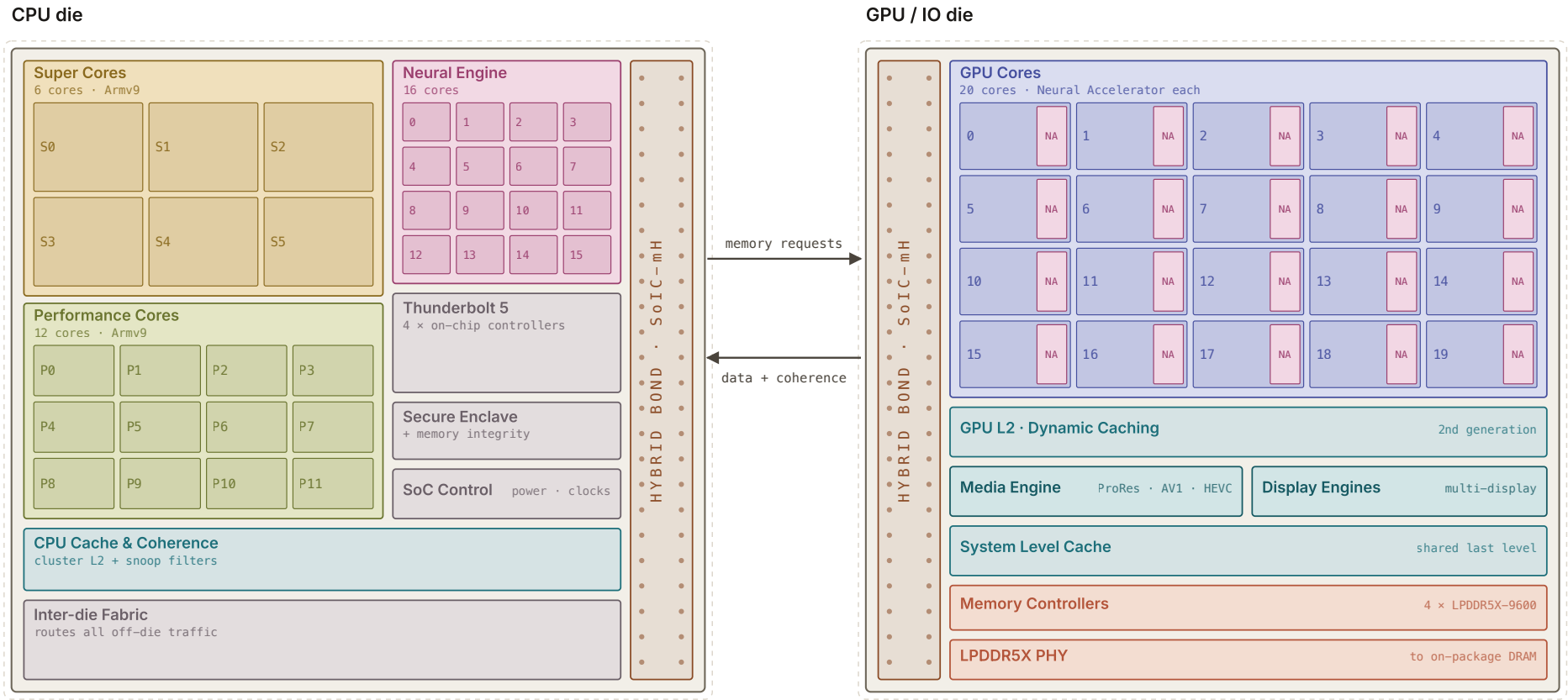
M5 · monolithic die



M5 Pro

Released: Mar 2026 · Process: TSMC N3P × 2 dies · Packaging: TSMC SoIC-mH · CPU: 18-core (6 Super + 12 Perf) · GPU: 20-core Pro / 40-core Max · Bandwidth: 307 / 614 GB/s · Max memory: 64 / 128 GB · Interconnect: Fusion Architecture

CPU · performance / super cores
CPU · efficiency / performance cores
GPU cores
Neural Engine & Neural Accelerators
Cache (L2 / SLC / Dynamic Caching)
Memory controllers & PHY
Media, display & image signal
I/O, security & fabric
Inter-die bond interface

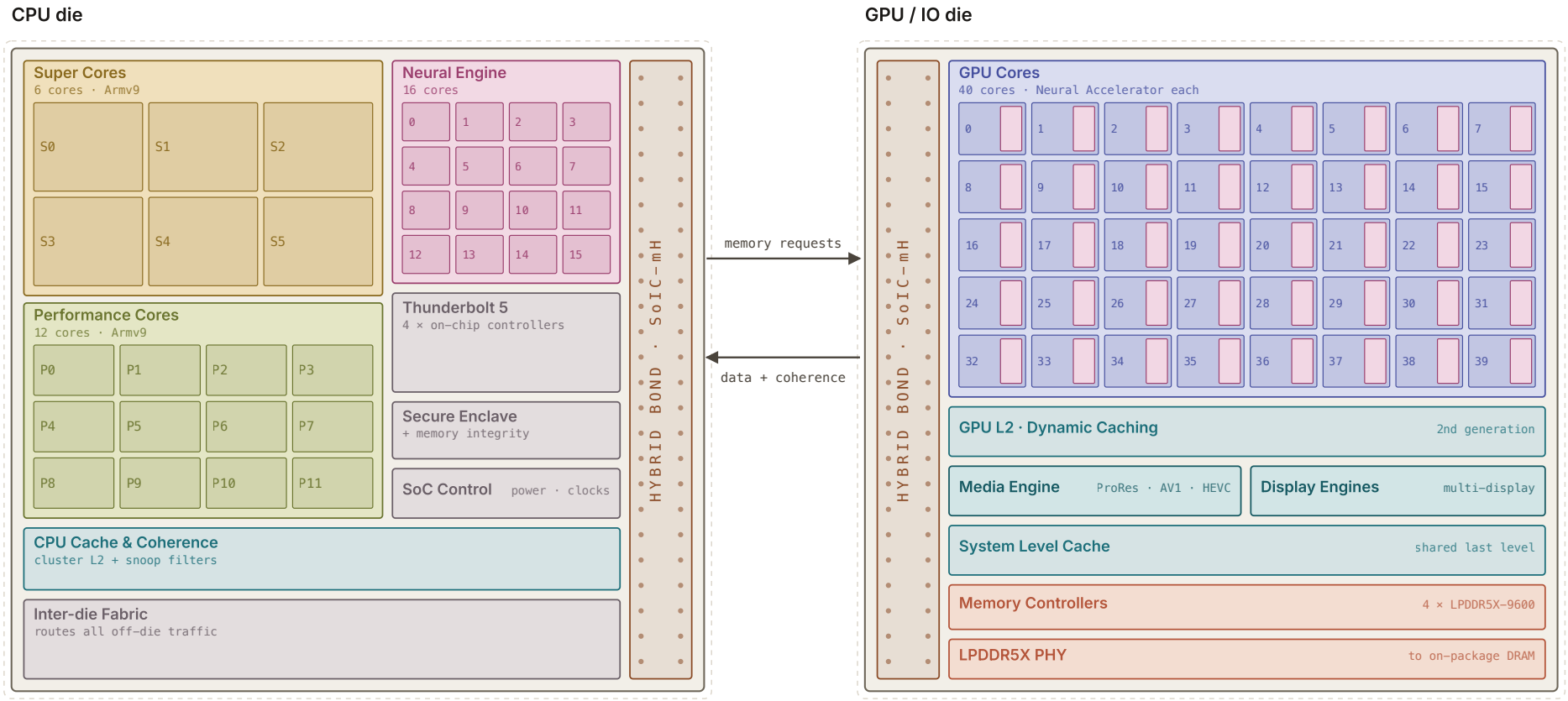


unfolded view — the dies are bonded face-to-face, not side by side

M5 Max

Released: Mar 2026 · Process: TSMC N3P × 2 dies · Packaging: TSMC SoIC-mH · CPU: 18-core (6 Super + 12 Perf) · GPU: 20-core Pro / 40-core Max · Bandwidth: 307 / 614 GB/s · Max memory: 64 / 128 GB · Interconnect: Fusion Architecture

CPU · performance / super cores
CPU · efficiency / performance cores
GPU cores
Neural Engine & Neural Accelerators
Cache (L2 / SLC / Dynamic Caching)
Memory controllers & PHY
Media, display & image signal
I/O, security & fabric
Inter-die bond interface

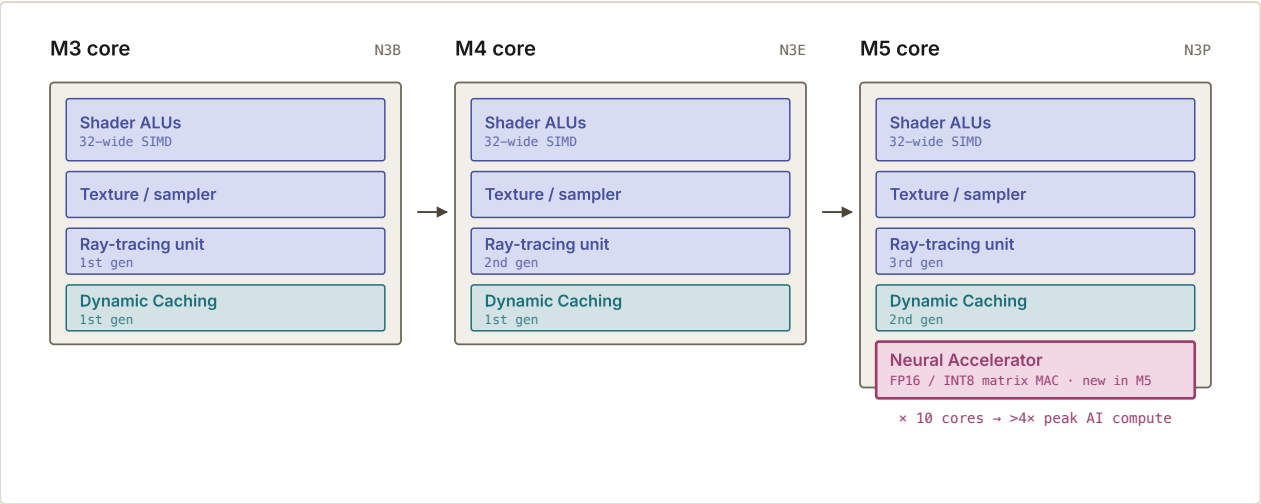


unfolded view — the dies are bonded face-to-face, not side by side

Zoom · one GPU core

where the matrix unit landed

M3 gave every GPU core a ray-tracing unit and a flexible on-chip memory pool. M4 sharpened both. M5 added something categorically different: a Neural Accelerator, a small FP16 and INT8 matrix engine, inside each of the ten cores. That is the whole "over four times the peak GPU compute for AI" claim, and it is why the M5's AI story lives in the GPU rather than in the Neural Engine.

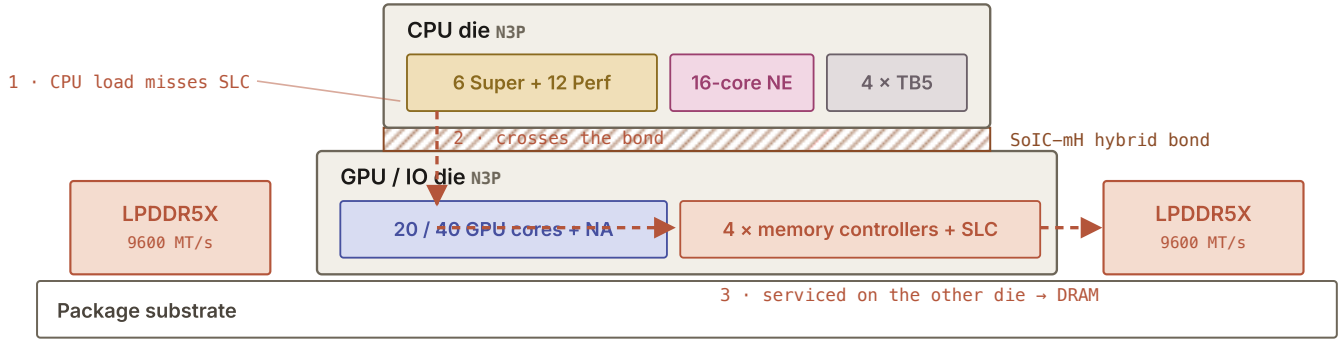


One GPU core, three generations. The shader ALUs, the shared Dynamic Caching pool and the ray-tracing unit persist. M5 inserts a dedicated matrix-multiply block next to the ALUs, sharing the same core-local memory rather than sitting off on the Neural Engine across the fabric.

M5 Pro and M5 Max · one SoC, two dies

TSMC SoIC-mH · Fusion Architecture

The M5 Pro and M5 Max are not monolithic. Apple splits the SoC into a CPU die and a GPU and IO die, bonded with TSMC's SoIC-mH hybrid bonding. Both dies are N3P, and Pro and Max ship the same CPU die. The consequence worth staring at: all four LPDDR5X-9600 memory controllers live on the GPU die, so every CPU access to unified memory crosses the bond interface. Hybrid bonding is what makes that affordable, with direct copper-to-copper pads at a pitch fine enough that the inter-die hop costs a fraction of the energy an organic-substrate link would.



Package cross-section. The CPU die is bonded face-to-face onto the GPU and IO die; only the lower die talks to the package substrate and therefore to the LPDDR5X stacks. The traced path is a CPU load missing in cache.

> **On the M5 Pro/Max core names.** Apple and the teardown both describe the 18-core CPU as six Super Cores plus twelve Performance Cores on Armv9, not the familiar performance and efficiency split. Read literally that is a renaming of the tiers. Apple has not said this outright, so treat the mapping as inference rather than fact. The base M5 of October 2025 still used the old wording: four performance and six efficiency cores.

Figures as published by Apple, except die areas (third-party measurement, marked est.) and the M5 transistor count, which Apple has not disclosed. Bandwidth is peak unified-memory bandwidth. · Sources: apple.com/newsroom — M3 family 2023-10, M4 2024-05, M5 2025-10, M5 Pro and M5 Max 2026-03; Counterpoint Research, "Apple M5 Pro chip teardown analysis". · theisegoria.github.io/explore/apple-silicon-floorplans.html