
Arquitectura
Apple
Silicon
(M1/M2/M3/M4)
1
About
2
Education
3
Experience
4
Projects
5
Testimonial
6
Contact

Equipo

- Daniel Eduardo Alvarez Terrazas
- Karol Burrola Torres
- Anette Cazares Suarez
- Melina Gonzalez Méndez

¿QUÉ ES APPLE SILICON?

Apple Silicon es la familia de procesadores

diseñados por Apple basados en arquitectura ARM para:

MacBook

iMac

iPad

Mac Studio

Características:

Arquitectura ARM64

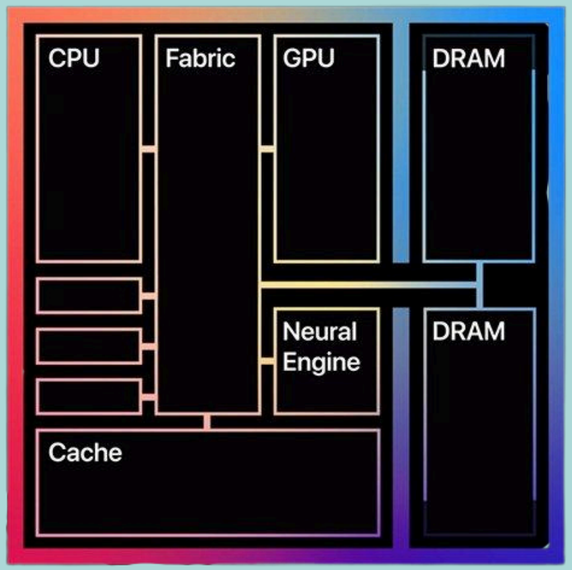
Diseño SoC (System on Chip)

CPU + GPU + IA integrados

Alto rendimiento por watt

El cambio de Intel (x86) a Apple Silicon (ARM) fue un cambio de paradigma de arquitectura CISC a RISC.

UNIFIED MEMORY ARCHITECTURE (UMA)

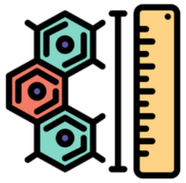


- EL PROBLEMA DEL MODELO TRADICIONAL: para que la GPU procese datos que preparó la CPU, esos datos deben copiarse físicamente de la RAM a la VRAM a través del bus PCIe.
- LA SOLUCIÓN: CPU, GPU y Neural Engine comparten el mismo banco físico de memoria. Lo que la CPU escribe, la GPU lo lee desde la misma dirección de memoria - sin copia, sin latencia de transferencia.

CHIP M1

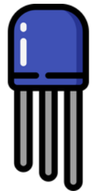
5 nm

Nanómetros correspondientes



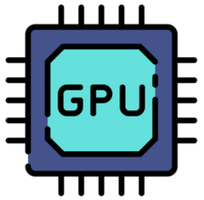
16 mill

Transistores en el chip



8 Núc.

GPU integrada



68 GB/s

Ancho de banda

11 TOPS

Neural Engine (Operaciones/segundo)

Diagrama del SoC M1 — 5nm TSMC

4 × Núcleos P (Firestorm) — Alto Rendimiento — 3.2 GHz	4 × Núcleos E (Icestorm) — Alta Eficiencia — 2.0 GHz
GPU de 8 núcleos — Metal — 128 GFLOPS	
Motor Neuronal — 16 núcleos — 11 TOPS	ISP · Codificador de Medios · Secure Enclave

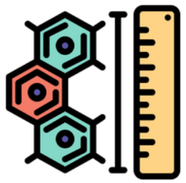
¿QUÉ LOGRÓ EL M1 FRENTE A INTEL?

Instrucciones ARM de 32 bits fijos (RISC):	Las etapas IF y ID son deterministas: el hardware sabe exactamente dónde empieza cada instrucción sin búsqueda previa. Menor Ciclo Por Instrucción (CPI) frente a Intel.
Decodificador de 8 vías (superescalar):	Procesa 8 instrucciones por ciclo de reloj simultáneamente. Intel decodifica 4-6 micro-operaciones, con una etapa extra de traducción de instrucciones CISC.
Memoria Unificada de 68 GB/s:	CPU y GPU comparten el mismo banco físico de memoria. Elimina la copia de datos CPU→GPU que en Intel consumía hasta el 25% del tiempo en tareas gráficas.

CHIP M2

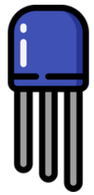
5 nm+

Nanómetros correspondientes



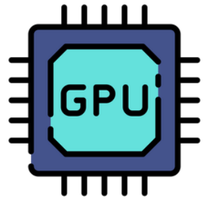
20 mill

Transitores (+25% VS M1)



10 Núc.

GPU integrada (+35% VS M1)



100 GB/s

Ancho de banda (+47% VS M1)

CARACTERÍSTICAS

- **Memoria unificada:** La memoria integrada en el chip permite que la CPU y la GPU compartan información al instante, evitando cuellos de botella y mejorando el rendimiento.
- **CPU nuevo:** Incorpora núcleos más rápidos, mayor memoria caché y mejoras en los núcleos de eficiencia para ofrecer un mejor rendimiento.
- **Neural Engine:** Para tareas de IA (como reconocimiento de voz, procesamiento de imágenes o modelos matemáticos predictivos), el M2 cuenta con un NE de 16 núcleos capaz de realizar 15.8 billones de operaciones por segundo
- **Motor multimedia:** Incluye hardware dedicado especialmente para procesar videos pesados

CARACTERÍSTICAS

- **Rendimiento superior por Watt:** Permite que los sistemas tengan una duración larga de batería y puedan ejecutarse de manera silenciosa, manteniendo la temperatura, incluso en juegos con gráficos avanzados
- **Hardware de video:** El potente motor de video ProRes de Apple permite reproducir múltiples instancias de videos 4K y 8K.
- **Secure Enclave:** Ofrece seguridad encargada de mantener contraseñas y datos biométricos seguros a nivel hardware
- **La familia de los M2:** El chip M2 se usa en dispositivos como la MacBook Air y el iPad Pro, y su arquitectura permite crear versiones más potentes como el M2 Pro, M2 Max y M2 Ultra

CHIP M2: RELACIÓN

EXPLICACIÓN

Cuello de botella de Von Neumann: En la organización de computadoras clásica, la CPU y la Memoria están separadas físicamente por un bus de sistema, (Cuello de botella de Von Neumann). El M2 soluciona esto al al implementar una Arquitectura de Memoria Unificada (UMA). La RAM LPDDR5 está encapsulada en la misma placa que el procesador.

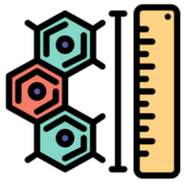
Paralelismo a nivel de datos (Neural Engine): El M2 usa paralelismo SIMD para procesar muchos datos al mismo tiempo, lo que acelera enormemente las tareas de inteligencia artificial y multimedia.

Arquitectura superescalar del M2: El M2 deja atrás el procesamiento secuencial tradicional usando una arquitectura superescalar con pipeline avanzado, capaz de ejecutar varias instrucciones al mismo tiempo y fuera de orden para aprovechar al máximo el rendimiento del chip.

CHIP M3

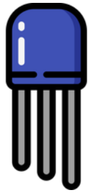
3 nm+

Nanómetros correspondientes



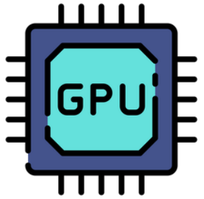
25 mill

Transitores (+25% VS M2)



10 Núc.

GPU integrada (+65% vs M1)



100 GB/s

Consumo energético a igual rendimiento

CARACTERISTICAS

- **Memoria unificada:** CPU y GPU comparten memoria para mejorar velocidad y comunicación.
- **Mayor rendimiento gráfico:** Mejor desempeño en videojuegos, renderizado y aplicaciones profesionales.
- **Neural Engine:** Optimiza tareas de inteligencia artificial y aprendizaje.
- **CPU más rápido:** El chip M3 es hasta un 60% más rápido que el chip M1.
- **Procesamiento multimedia:** Optimizado para edición de video, fotografía y contenido 3D.
- **Eficiencia térmica:** Mantiene alto rendimeinto con menor generación de calor.

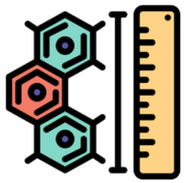
CARACTERISTICAS

- **Dynamic Caching:** Administra dinámicamente la memoria local de la GPU en tiempo real. La memoria se asigna según las necesidades de cada tarea.
- **Ray Tracing por hardware:** Simula el comportamiento real de la luz dentro de escenas 3D. Genera reflejos, sombras e iluminación más precisos.
- **Mesh Shading:** Moderniza el pipeline gráfico tradicional, reduce trabajo innecesario, mejora el rendimiento y permite escenas 3D más complejas y detalladas.
- **CPU más rápido:** Los núcleos de rendimiento y eficiencia han sido mejorados. El chip M3 es hasta un 60% más rápido que el chip M1.

CHIP M4

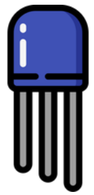
3 nm+

Nanómetros correspondientes



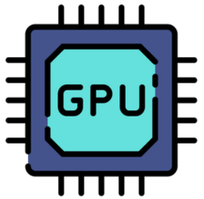
28 mill

Transistores (+12% vs M3)



10 Núc.

GPU integrada



120 GB/s

Ancho de banda

CARACTERÍSTICAS

- **Innovación en el ISA (ARMv9 y SME):** Adopta formalmente la arquitectura ARMv9, integrando extensiones matriciales escalables (SME) que llevan el paralelismo de datos (SIMD) al siguiente nivel para acelerar Inteligencia Artificial local.
- **CPU:** Arquitectura base de 10 núcleos configurada en 4 núcleos de Rendimiento (P-Cores) y 6 núcleos de Eficiencia (E-Cores).

CARACTERÍSTICAS

- **El cambio más importante TOPS = Billones de Operaciones Por Segundo (Tera Operations Per Second).** El Neural Engine del M4 duplicó su rendimiento respecto al M3 (18 → 38 TOPS) en esta generación. Ejecuta modelos de lenguaje de 3 mil millones de parámetros completamente en el dispositivo, sin conexión a internet, a ~30 palabras por segundo.
- **Los P-Cores (núcleos de alto rendimiento)** de tercera generación logran un 40% más de rendimiento en tareas de un solo hilo frente al M2. Esto es directo a la Ley de Amdahl: al hacer más rápida la parte secuencial, el speedup total del sistema sube aunque la fracción paralela no cambie.

CHIP M4

Ejemplo 1: El impacto de los P-Cores del M4 en la Ley de Amdahl

Supongamos un escenario con una fracción secuencial $f_{sec} = 0.25$ y una fracción paralela $f_{par} = 0.75$

Limite teorico:

$$S_{max} = \frac{1}{f_{sec}} = \frac{1}{0.25} = 4\times$$

un incremento de cuatro veces

Efecto M4 (P-Cores un 40% más rápidos):

El nuevo tiempo efectivo de la fracción secuencial se reduce:

$$f_{sec\_nueva} = \frac{0.25}{1.4} \approx 0.178$$

Calculando el nuevo límite de Speedup con el M4:

$$S_{max\_M4} = \frac{1}{0.178} \approx 5.61\times$$ un incremento de 5.61 veces

La optimización del rendimiento secuencial mediante P-Cores eleva el límite teórico de escalabilidad del sistema. Esta estrategia permite mitigar las restricciones de speedup absoluto impuestas por la Ley de Amdahl.

CHIP M4

Ejemplo 2: Ecuación de Rendimiento de la CPU (RISC vs CISC)

Tiempo de ejecución = $IC \times CPI \times T_c$

Sistema CISC (Intel i9):

$IC = 1 \times 10^9$ (Menos instrucciones, pero complejas)

$CPI = 1.5$ (Penalización por decodificación variable)

$T_c = 0.18ns$ (Reloj de 5.5 GHz)

$$Tiempo_{CISC} = (1 \times 10^9) \times 1.5 \times (0.18 \times 10^{-9}\ \text{s})$$

$$Tiempo_{CISC} = 0.27 \text{ segundos}$$

Sistema RISC (Apple M4):

$IC = 1.2 \times 10^9$ (Más instrucciones reducidas)

$CPI = 0.7$ (Decodificador superescalar de 8 vías)

$T_c = 0.22ns$ (Reloj de 4.4 GHz)

$$Tiempo_{M4} = (1.2 \times 10^9) \times 0.7 \times (0.22 \times 10^{-9}\ \text{s})$$

$$Tiempo_{M4} = 0.18 \text{ segundos}$$

El diseño RISC superescalar del M4 demuestra que la reducción del CPI es más efectiva que el incremento de la frecuencia de reloj. La eficiencia arquitectónica prevalece sobre la velocidad de ciclo en la optimización del tiempo de ejecución.

TABLA COMPARATIVA — M1 al M4

Característica	M1	M2	M3	M4
Proceso de fabricación	5nm (N5)	5nm (N5P)	3nm (N3B)	3nm (N3E)
Transistores	16 Milllones	20 Millones	25 Millones	28 Millones
Núcleos CPU (base)	4 R + 4 E	4 R + 4 E	4 R + 4 E	4 R + 6 E
Ancho decodificador	8 instrucciones	8 instrucciones	8 instrucciones	8 instrucciones
Neural Engine	11 TOPS	15.8 TOPS	18 TOPS	38 TOPS
Ancho de Banda	68 GB/s	100 GB/s	100 GB/s	120 GB/s

CONCLUSIONES

- **El triunfo de la eficiencia (RISC vs CISC):** El rendimiento ya no depende de subir los GHz. Al usar RISC y un pipeline superescalar ancho, lograron reducir drásticamente el CPI (Ciclos por Instrucción).

- **Superando el muro de la memoria:** Con la Memoria Unificada (UMA), la CPU, GPU y NPU comparten el mismo espacio físico. Se evitan las copias innecesarias de datos y se maximiza la localidad espacial y temporal de los datos.

- **Dominio de la Ley de Amdahl:** Su diseño híbrido es el mejor ejemplo de esta ley. Los P-Cores aceleran la fracción secuencial del código y los E-Cores manejan la fracción paralelizable ahorrando energía.

- **El futuro es el paralelismo (SIMD e IA):** El Neural Engine saltó de 11 a 38 TOPS (del M1 al M4). El rendimiento moderno exige hardware dedicado y el uso de extensiones matriciales para ejecutar IA localmente.


RACIAS!