**Memorias caché**

**Memoria Caché L1**

**Tecnología usada**

* Implementada en SRAM de alta velocidad.
* Protocolos de coherencia (MESI, MOESI, etc.) para mantener consistencia.  
  [*https://dl.acm.org/doi/pdf/10.1145/3410463.3414623*](https://dl.acm.org/doi/pdf/10.1145/3410463.3414623)

**Capacidades**

* Típicamente 32–64 KB por núcleo (separadas en datos e instrucciones).  
  [*https://pmc.ncbi.nlm.nih.gov/articles/PMC8537857/*](https://pmc.ncbi.nlm.nih.gov/articles/PMC8537857/)

**Velocidad de acceso**

* 1.2–1.8 ns (3–5 ciclos de CPU).
* Ancho de banda: 30–80 GB/s según arquitectura.  
  [*https://dl.acm.org/doi/pdf/10.1145/3410463.3414623*](https://dl.acm.org/doi/pdf/10.1145/3410463.3414623)

**Ventajas**

* Latencia ultrabaja, indispensable para el pipeline.
* Minimiza fallos tempranos y acelera ejecución.  
  [*https://pmc.ncbi.nlm.nih.gov/articles/PMC8537857/*](https://pmc.ncbi.nlm.nih.gov/articles/PMC8537857/)

**Desventajas**

* Capacidad muy limitada.
* Alto costo por área/energía.
* Vulnerable a conflictos de mapeo y ataques por canales laterales.  
  [*https://dl.acm.org/doi/pdf/10.1145/3410463.3414623*](https://dl.acm.org/doi/pdf/10.1145/3410463.3414623)

**Memoria Caché L2**

**Tecnología usada**

* SRAM con mayor latencia que L1.
* Algoritmos de reemplazo más sofisticados (ej. Multi-step LRU, bypass selectivo).
* Estrategias dinámicas de particionado o clustering para minimizar interferencias entre hilos y equilibrar acceso.<https://arxiv.org/pdf/2112.09981>

<https://arxiv.org/pdf/2402.07578>

**Capacidades**

* Típicamente en el rango de 256 KB a 2 MB por núcleo o clúster.

<https://arxiv.org/pdf/2402.07578>

**Velocidad de acceso**

* 2.7–9 ns (7–20 ciclos).
* Latencias en el rango aproximado de 7–20 ciclos, dependiendo del método de particionado y del camino de acceso.  
  <https://arxiv.org/pdf/2112.09981>

<https://arxiv.org/pdf/2402.07578>

**Ventajas**

* Mejora el balance de rendimiento y la justicia entre aplicaciones concurrentes.
* Reduce interferencias y saturaciones en el uso del cache compartido.

<https://arxiv.org/pdf/2402.07578>

**Desventajas**

* Consumo energético mayor que L1.
* Puede duplicar datos si es inclusiva (overhead).
* Mayor complejidad de control.  
  <https://arxiv.org/pdf/2112.09981>

**Memoria Caché L3 (LLC)**

**Tecnología usada**

* SRAM más densa, interconectada mediante anillos o mallas on-chip.
* Algoritmos avanzados: RRIP, SHiP, Hawkeye, ML-based (ej. PAIC, HAPC).  
  <https://jcst.ict.ac.cn/fileup/1000-9000/PDF/JCST-2023-2-12-1573-391.pdf>

**Capacidades**

* 8–64 MB por socket (algunos servidores más de 100 MB).

<https://www.mdpi.com/2072-666X/13/11/2014>

**Velocidad de acceso**

* 24–38 ns (≈37–100 ciclos).
* Latencia depende de la distancia del núcleo (efecto NUMA).  
  <https://jcst.ict.ac.cn/fileup/1000-9000/PDF/JCST-2023-2-12-1573-391.pdf>

**Ventajas**

* Amortigua accesos a DRAM.
* Soporta conjuntos de trabajo grandes.

Mejora la coherencia y el rendimiento multicore.  
<https://www.mdpi.com/2072-666X/13/11/2014>

**Desventajas**

* Mayor latencia.
* Congestión de interconexión.
* Sensible a patrones irregulares (HPC, big data).
* Elevado consumo energético por MB.  
  <https://jcst.ict.ac.cn/fileup/1000-9000/PDF/JCST-2023-2-12-1573-391.pdf>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Memorias Caché** | | | | | |
| **Memoria caché** | **Tecnología / Interfaz** | **Velocidad de lectura / escritura** | **Capacidad** | **Ventajas** | **Desventajas** |
| Memoria Caché L1 | Implementada en SRAM de alta velocidad [1].  Protocolos de coherencia (MESI, MOESI, etc.) para mantener consistencia [1]. | 1.2–1.8 ns (3–5 ciclos de CPU) .  Ancho de banda: 30–80 GB/s según arquitectura [1]. | Típicamente 32–64 KB por núcleo (separadas en datos e instrucciones) [2]. | * Latencia ultra baja, indispensable para el pipeline [2]. * Minimiza fallos tempranos y acelera ejecución [2]. | * Capacidad muy limitada [1]. * Alto costo por área/energía [1]. * Vulnerable a conflictos de mapeo y ataques por canales laterales [1]. |
| Memoria Caché L2 | SRAM con mayor latencia que L1 [3].  Algoritmos de reemplazo más sofisticados (ej. Multi-step LRU, bypass selectivo) [4].  Estrategias dinámicas de particionado o clustering para minimizar interferencias entre hilos y equilibrar acceso [4]. | 2.7–9 ns (7–20 ciclos) [3].  Latencias en el rango aproximado de 7–20 ciclos, dependiendo del método de particionado y del camino de acceso [4]. | Típicamente en el rango de 256 KB a 2 MB por núcleo o clúster [4]. | * Mejora el balance de rendimiento y la justicia entre aplicaciones concurrentes [4]. * Reduce interferencias y saturaciones en el uso del cache compartido [4]. | * Consumo energético mayor que L1 [3]. * Puede duplicar datos si es inclusiva (overhead) [3]. * Mayor complejidad de control [3]. |
| Memoria Caché L3 | SRAM más densa, interconectada mediante anillos o mallas on-chip.  Algoritmos avanzados: RRIP, SHiP, Hawkeye, ML-based (ej. PAIC, HAPC) [5]. | 24–38 ns (≈37–100 ciclos).  Latencia depende de la distancia del núcleo (efecto NUMA) [5]. | 8–64 MB por socket (algunos servidores más de 100 MB) [6]. | * Amortigua accesos a DRAM [6]. * Soporta conjuntos de trabajo grandes [6]. * Mejora la coherencia y el rendimiento multicore [6]. | * Mayor latencia [5]. * Congestión de interconexión [5]. * Sensible a patrones irregulares (HPC, big data) [5]. * Elevado consumo energético por MB.aja capacidad en comparación con medios posteriores [5]. |