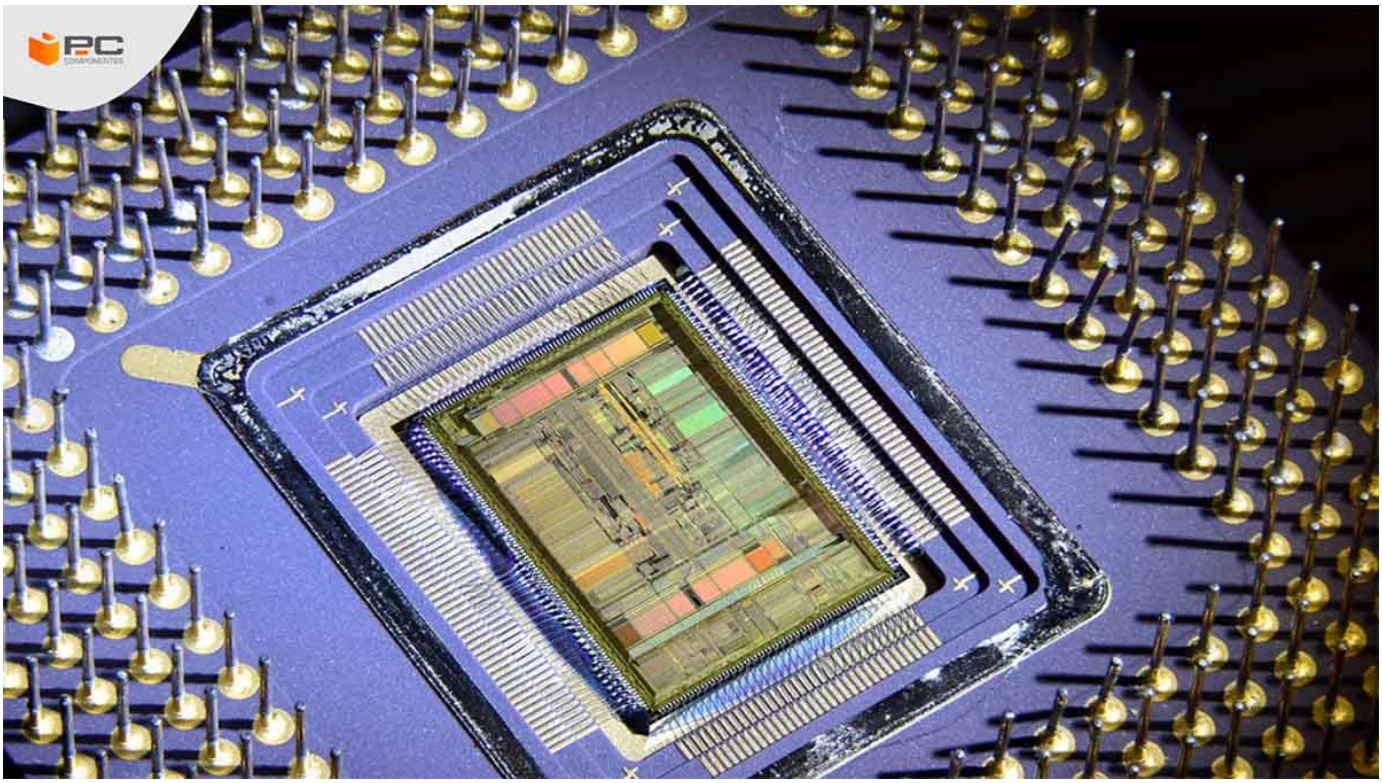
¿Qué es la memoria caché? Conoce su utilidad en CPU y GPU

Componentes Ángel Aller.6 de septiembre de 2023

La memoria caché está dentro del chip y es importante saber por qué es interesante tener más MB en el chip. Explicamos este concepto básico en informática a fondo.



ÁNGEL ALLER

— Geek inconformista.

Experto en ordenadores y gaming

Qué es la memoria caché y cómo funciona

La memoria caché es un tipo de memoria volátil (SRAM) con poca capacidad que se encuentra dentro de la CPU o GPU, cuyo fin es ofrecer el acceso a datos de uso frecuente. Se trata de una memoria temporal y tiene distintos niveles, siendo lo más común L1, L2 y L3. La caché es más rápida que la memoria RAM (DRAM), pudiendo ejecutar instrucciones y leer o escribir datos a mayor velocidad. También existe la memoria caché en los navegadores de Internet para 'almacenar' versiones de las páginas web que visitas y cargarlas más rápido.

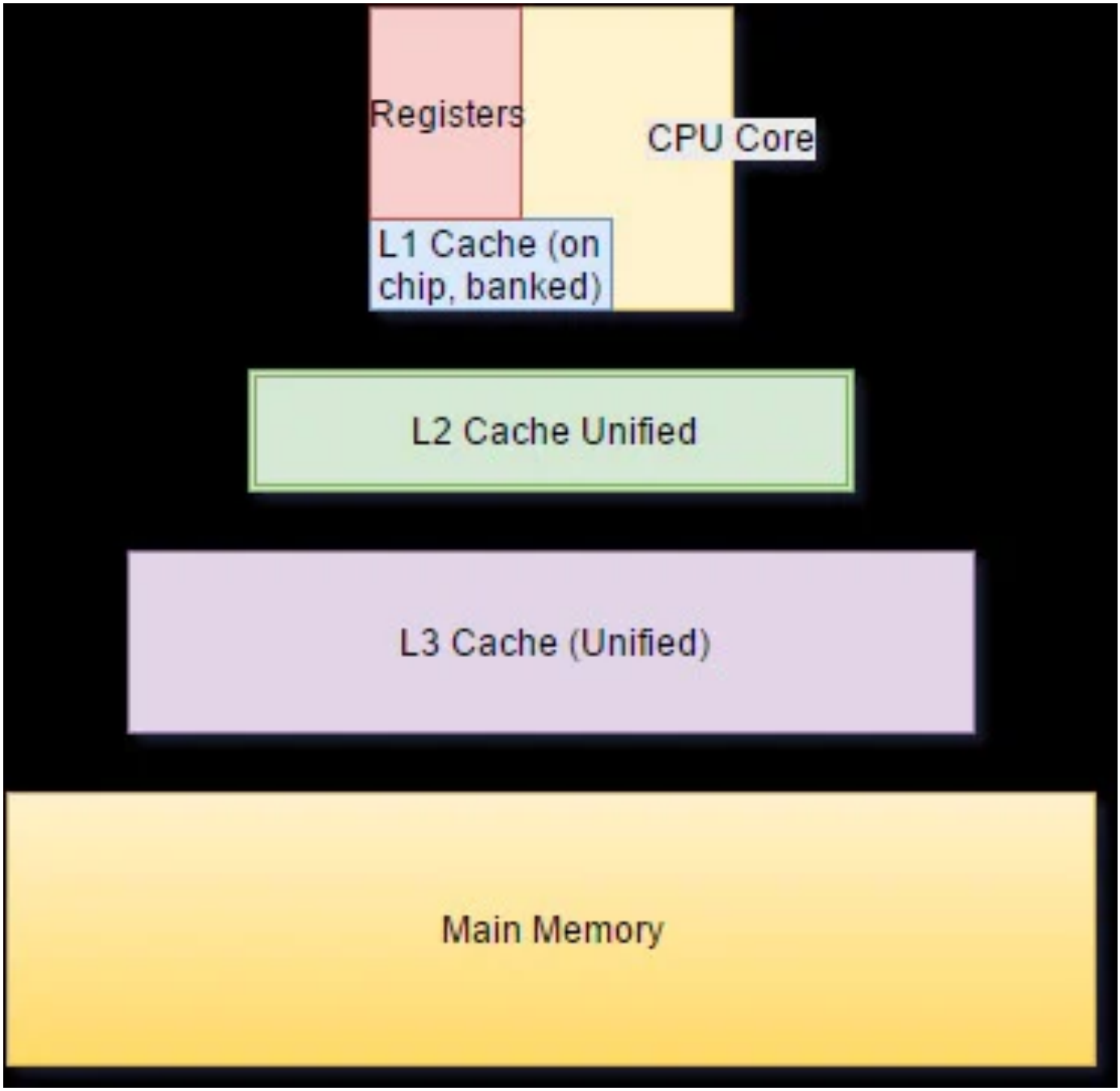
Decimos que es una memoria temporal, y es que aprovechan “la localidad temporal”. Hacemos referencia a los **datos que no han cambiado y que se utilizan varias veces**, de ahí que se use para almacenar datos o instrucciones cuyo uso futuro es previsible.

Coloquialmente, se dice que es una versión más pequeña y más ágil de la memoria RAM habitual. Y es que **almacena la información que es usada por la CPU o GPU habitualmente**; una vez que la CPU o GPU necesite esos datos, tira de la caché si es posible.

Cuando la CPU o GPU encuentra esos datos se le denomina “**cache hit**” (acierto de caché), lo que se traduce en que el **la CPU/GPU ha podido acudir a los datos rápidamente** sin perjudicar el rendimiento. Al final, su capacidad es limitada, por lo que ahí solo encontraremos los datos que necesita la CPU.

Por el contrario, **si la CPU o GPU acude a la memoria caché y no encuentra los datos**, se llama “**cache miss**” (fallo de caché). Así que, el chip tendrá que ir a la RAM o a los discos duros para encontrar los datos que busca.

La jerarquía de la memoria caché: niveles L1, L2 y L3



Los datos primero se almacenan en la L1, luego en la L2 y finalmente en la L3; cuando no queda espacio, se recurre a la memoria RAM. Incluso si la RAM se queda sin capacidad, el sistema hace uso de la memoria de los HDD o SSD.

El nivel más rápido es L1, seguido de L2 y el más lento es L3. Dependiendo de la arquitectura del procesador, veremos cómo los núcleos acceden a la memoria cache: hay arquitecturas que distribuyen determinados MB por varios núcleos, así como otras asignan MB específicos a cada núcleo.

Para evitar que el rendimiento caiga, lo ideal es que “todo quede en casa”; es decir, que la CPU o GPU solo recurran a su memoria caché. La **memoria caché está en el**

interior del chip, normalmente en un encapsulado o complejo distinto, pero está en el mismo die. De ahí que no decaiga el rendimiento: la interconexión es total y no hay prácticamente latencia.

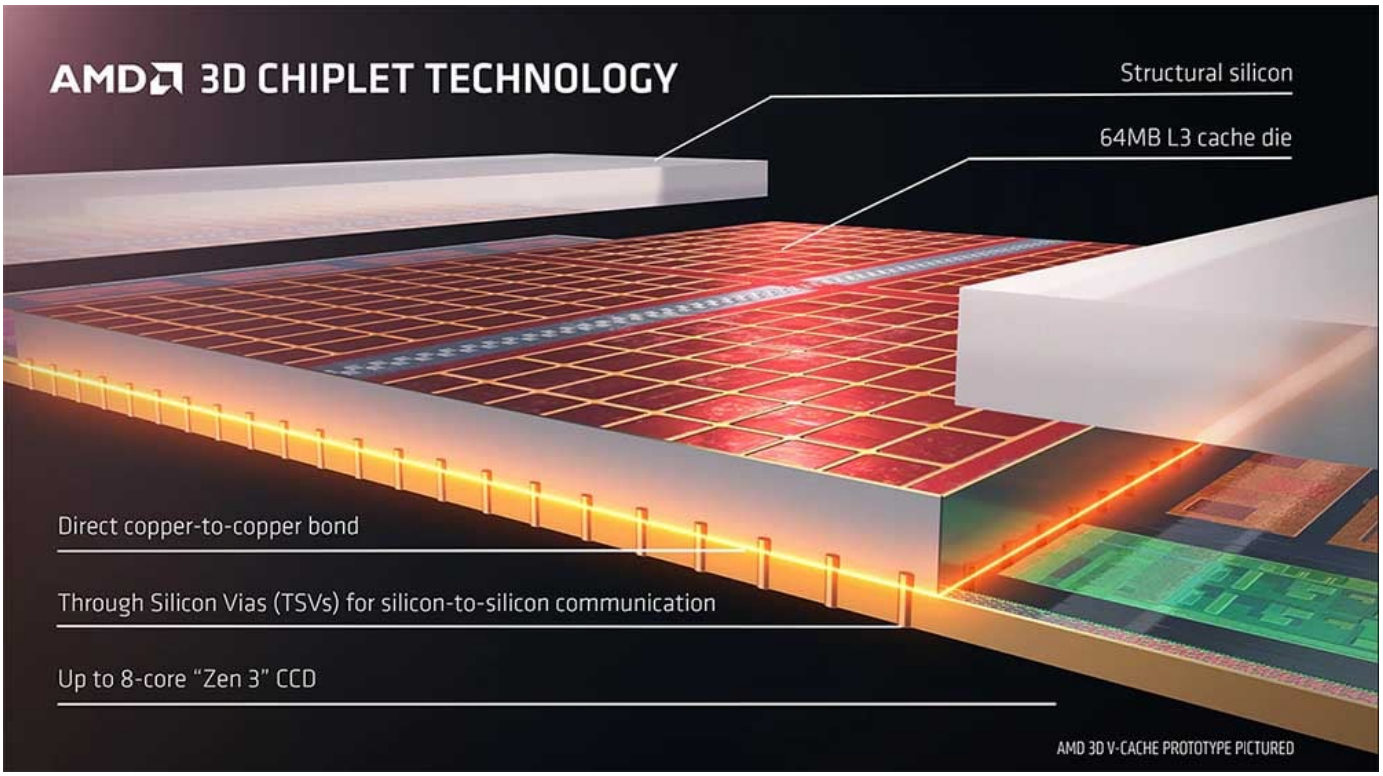
El problema es dar con un cache miss y tener que acudir a la RAM o a los discos duros: se introduce latencia en las interconexiones.

Solo tenéis que comprobar por vosotros mismos las especificaciones de los procesadores y tarjetas gráficas para ver qué caché llevan. Comprobaréis que hay ciertas diferencias, pero patrones en común:

- En los procesadores encontramos la memoria caché L1, L2 y L3.
- En las tarjetas gráficas podemos encontrar los 3 niveles, pero solo en AMD.
- La memoria L1 suele tener KB, sin llegar a 1024 KB (1 MB), siendo la más pequeña.
- Con la L2 ya empezamos a ver más 1 MB o más por núcleo.
- La L3 es la que más capacidad tiene de las 3, llegando a ver especificaciones por encima de los 100 MB.

Sin embargo, lo más interesante de cara al rendimiento son la L1 y la L2 porque son una memoria caché más rápidas, que es la clave.

La evolución de la memoria caché en GPU y CPU



La memoria caché lleva con nosotros décadas, pero se ha reivindicado frente a la DRAM y la VRAM como protagonista en los últimos tiempos. **AMD fue quien impulsó el enfoque de equipar más memoria caché en sus CPUs y GPUs** para ganar FPS.

El equipo rojo detectó un problema: la superficie del die del chip es limitada, por lo que no podían instalar mucha memoria caché. Su solución fue **3D V-Cache**, que no es más que **memoria caché L3 apilada en vertical e interconectada** por un interposer. El primer procesador con 3D V-Cache fue el Ryzen 7 5800X3D, lanzado en 2022.

Ver detalles

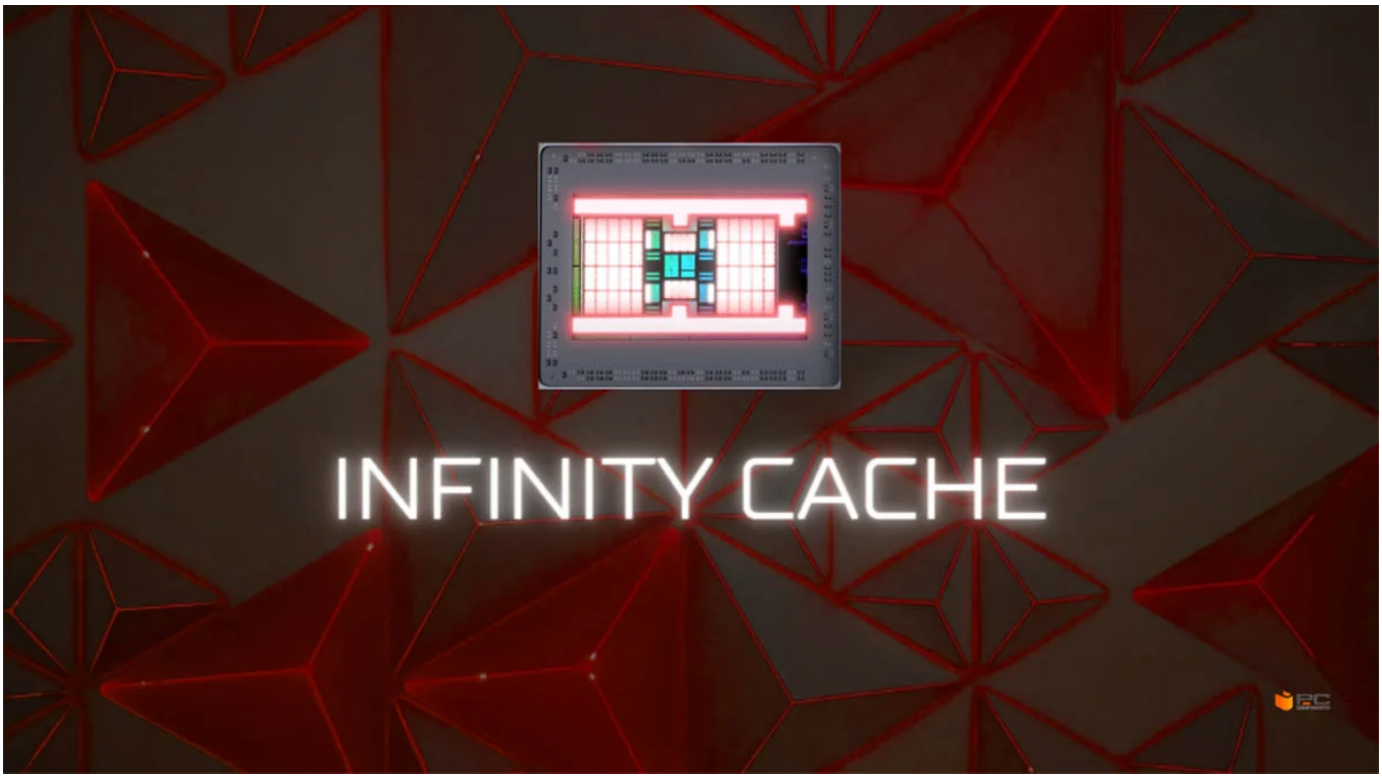
Si no puedes instalarla una al lado de otra, ¿por qué no amontonarla? Eso sí, el coste de producción es elevado y solo se puede vender dicha tecnología a entusiastas.

Este fabuloso AMD Ryzen 7 5800X3D venía con una arquitectura **Zen 3**, lo que significa que la caché L1 y L2 tenían una asignación individual por núcleo (64KB y 512 KB respectivamente), mientras que la L3 era compartida por todos los núcleos (8 en total). **Esta CPU venía con 96 MB L3**, lo que era una locura en ordenadores personales.

Con la llegada de los AMD Ryzen 7000, vimos que la marca aumentó la caché en todos sus niveles, ¡y parece qué seguirán haciéndolo así! De hecho, han lanzado más procesadores con la tecnología 3D V-Cache.

En la división Radeon también vieron que **la caché podía impulsar a que las GPUs sacaran más FPS**. En este caso, AMD decidió apostar por una **caché L3 más grande**

dentro del chip para que la GPU no tuviese que acudir a la VRAM para acceder a un dato.



De este modo, las AMD Radeon RX 6000 equiparon esta novedad, lo que se tradujo con un aumento considerable de la caché L3. Esto no lo tenía en cuenta NVIDIA antes, vieron el impacto de **Infinity Cache** en los FPS y decidieron aumentar memoria caché en sus chips.

Concretamente, vimos un aumento de memoria caché L2 en las NVIDIA RTX 4000, algo que sin duda les ha ayudado.