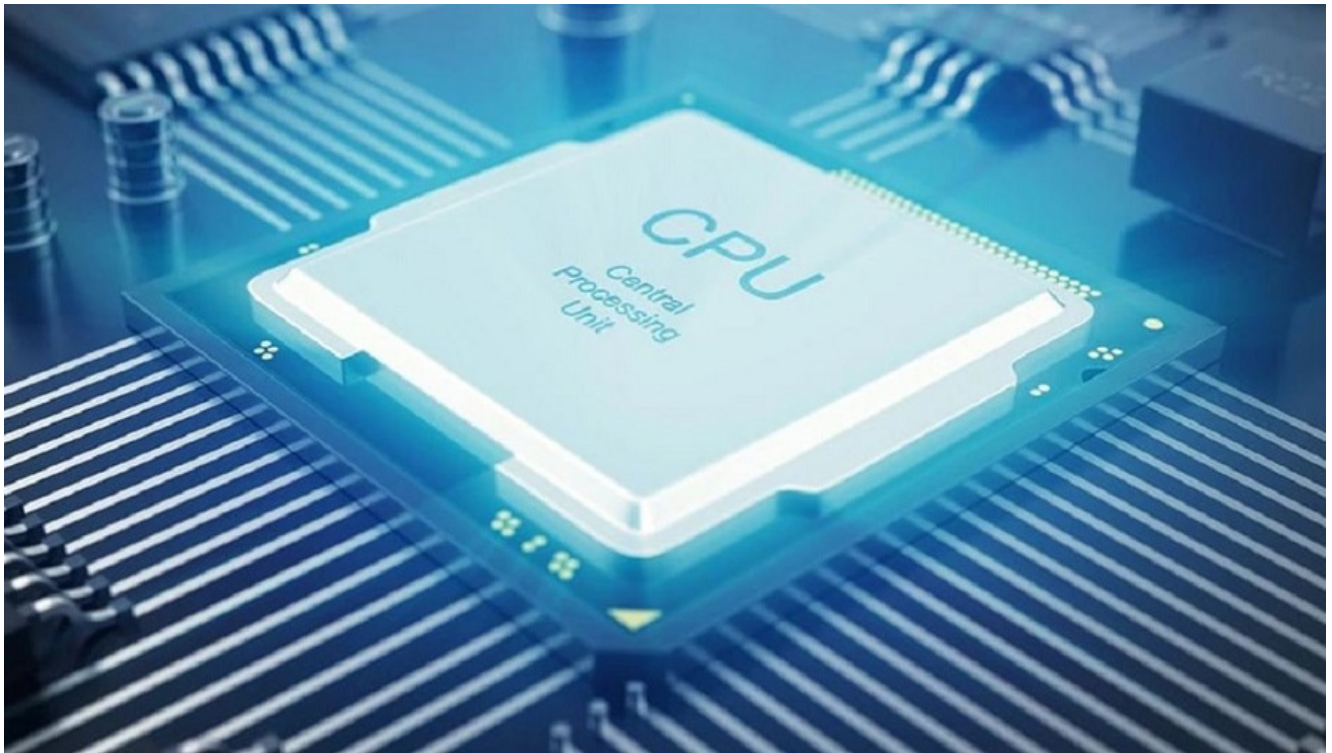


Memoria caché: qué es y qué diferencias hay entre los tipos L1, L2, L3 y L4



Hace ya algún tiempo publicamos un artículo dedicado a repasar a fondo los elementos más importantes que dan vida a un procesador. Entre ellos se encuentra la memoria caché, un componente que ha vivido una importante evolución durante las últimas décadas pero que todavía hoy sigue siendo **un gran desconocido**.

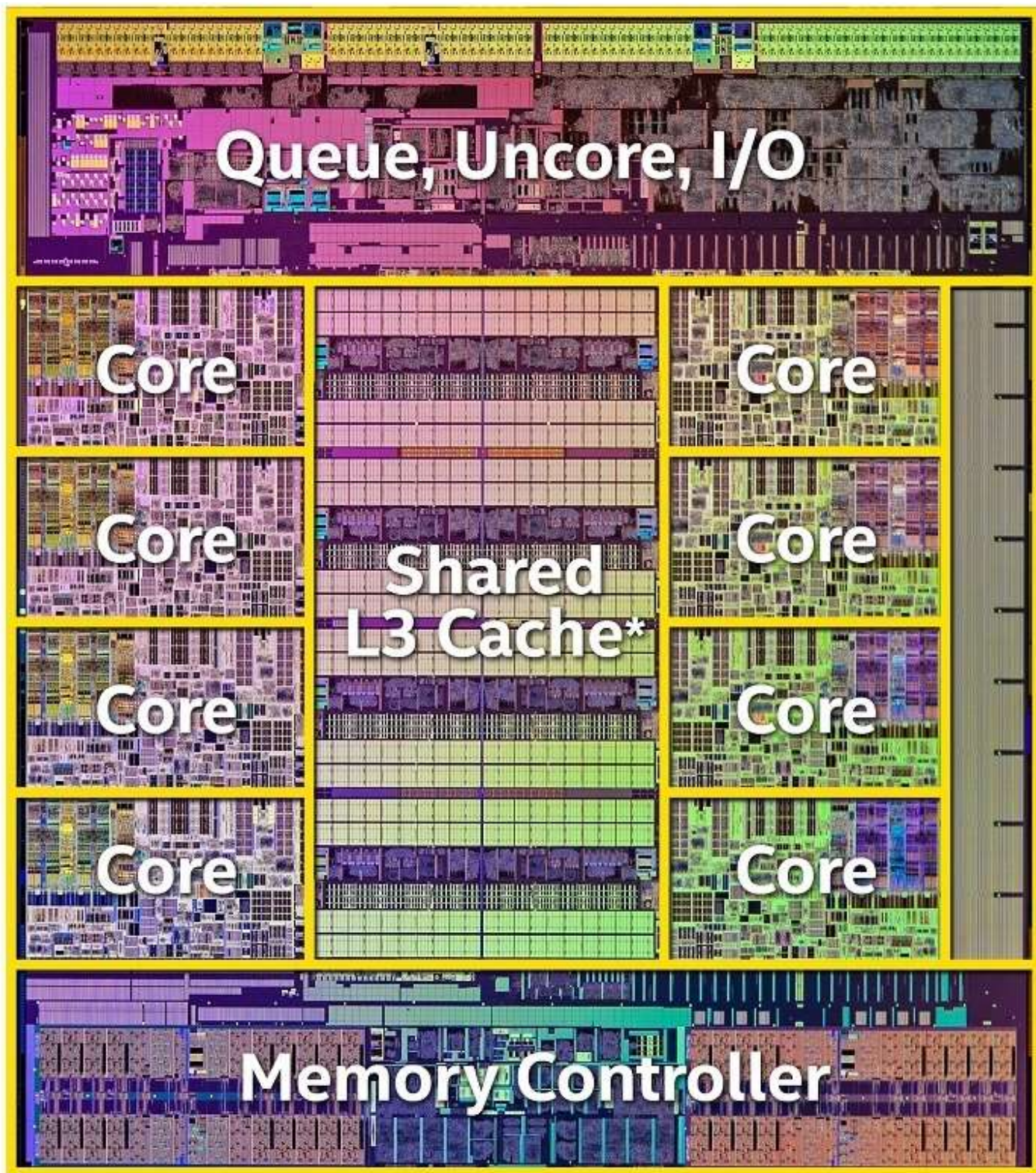
Sabemos que muchos de nuestros lectores tienen claro el papel que juegan en un procesador los núcleos, el IPC y el proceso de fabricación, pero la memoria caché sigue **generando algunas dudas**, y por ello hemos decidido hacer este artículo especial, en el que os vamos a contar todo lo que debéis saber sobre la memoria caché en general, y también sobre todos los tipos que existen.

Actualmente podemos diferenciar la memoria caché en **L1, L2 y**

L3, nomenclaturas con las que se identifican los distintos niveles en los que se encuadran. También se utiliza, aunque con poca frecuencia, un tipo de memoria caché conocida como **L4**, sobre todo en procesadores que vienen con GPUs integradas.

Sobre todo esto vamos a profundizar en las siguientes líneas, así que poneos cómodos, que hay muchas cosas interesantes que leer. Como siempre os invitamos a que nos dejéis **cualquier duda en los comentarios**.

Memoria caché: ¿qué es y cómo funciona?



La memoria caché es, como su propio nombre indica, un tipo de memoria específica. En un ordenador existen diferentes tipos de memoria que se agrupan en dos grandes variantes: **la memoria volátil y la memoria no volátil**, siendo la primera aquella que pierde los datos cuando apagamos el equipo (memoria RAM, por ejemplo) y la segunda aquella que los mantiene (un disco duro o un SSD, por ejemplo).

Pues bien, la memoria caché es un tipo de memoria específica que está preparada para **servir de apoyo al procesador**, y que es capaz de trabajar a velocidades muy elevadas. Esta memoria se divide en diferentes niveles como dijimos, pero su función base es la misma: servir de buffer de apoyo a la CPU para mejorar su rendimiento.

Para ello actúa como un sistema de **almacenamiento de instrucciones y de datos** capaz de comunicarse con el procesador a gran velocidad. Normalmente en la memoria caché se almacenan datos e instrucciones que el procesador va a necesitar para completar diferentes tareas a corto plazo, y a los que podrá acceder a mayor velocidad que si los buscase en la memoria RAM.

Cuando el procesador encuentra lo que buscaba en la caché **se produce un acierto**. Dependiendo de la caché en la que se produzca el acierto tendremos un mayor o un menor rendimiento. La caché L1 es más rápida que la L2 pero tiene menor capacidad, y lo mismo ocurre con la caché L3, que es más grande pero más lenta que la L2.

Al conseguir un acierto el procesador no tiene que recurrir a la memoria RAM, que es más lenta y además tiene una latencia mayor, lo que se traduce en un tiempo de acceso mucho más largo. Si no encuentra lo que necesitaba en la caché **se producirá un fallo**, y tendrá que realizar otra búsqueda en la memoria RAM.

Contar con una mayor cantidad de memoria caché, especialmente de tipo L2 y L3, que son las que se pueden implementar en mayores capacidades, se aumenta la cantidad de aciertos del procesador, y con esto se evitan dos cosas muy importantes que podrían reducir el rendimiento:

1. Que el procesador tenga que **volver a completar ciclos de trabajo** para conseguir los datos o instrucciones que necesita y que no están ni en la caché ni en la RAM.

2. Que tenga que **acceder a la memoria RAM** (más lenta que la caché) para obtener dichos datos o instrucciones.

La velocidad de la memoria caché importa, y mucho, pero como he dicho anteriormente también es clave **la capacidad** de la misma, aunque sobre este tema hablaremos en el siguiente punto. Es importante recordar que en las generaciones más antiguas de procesadores la memoria caché L2 y L3 estuvo integrada durante mucho tiempo en la placa base, y no en el encapsulado del procesador. Esto suponía que la información tenía que recorrer una distancia importante en cada acceso lo que acababa teniendo un impacto considerable en el rendimiento.

Actualmente todos los tipos de memoria caché se encuentran **integrados en el encapsulado del procesador**, con la única excepción de la L4 que, por cuestiones de tamaño y de espacio, suele ir en un encapsulado propio ubicado justo a continuación del encapsulado principal. Esto tiene consecuencias importantes, ya que este diseño permite incorporar una mayor cantidad de memoria, pero al estar colocada en un encapsulado separado está más lejos del procesador y se aumenta la latencia, lo que reduce el rendimiento y aumenta los tiempos de acceso.

Memoria caché: tipos y jerarquía

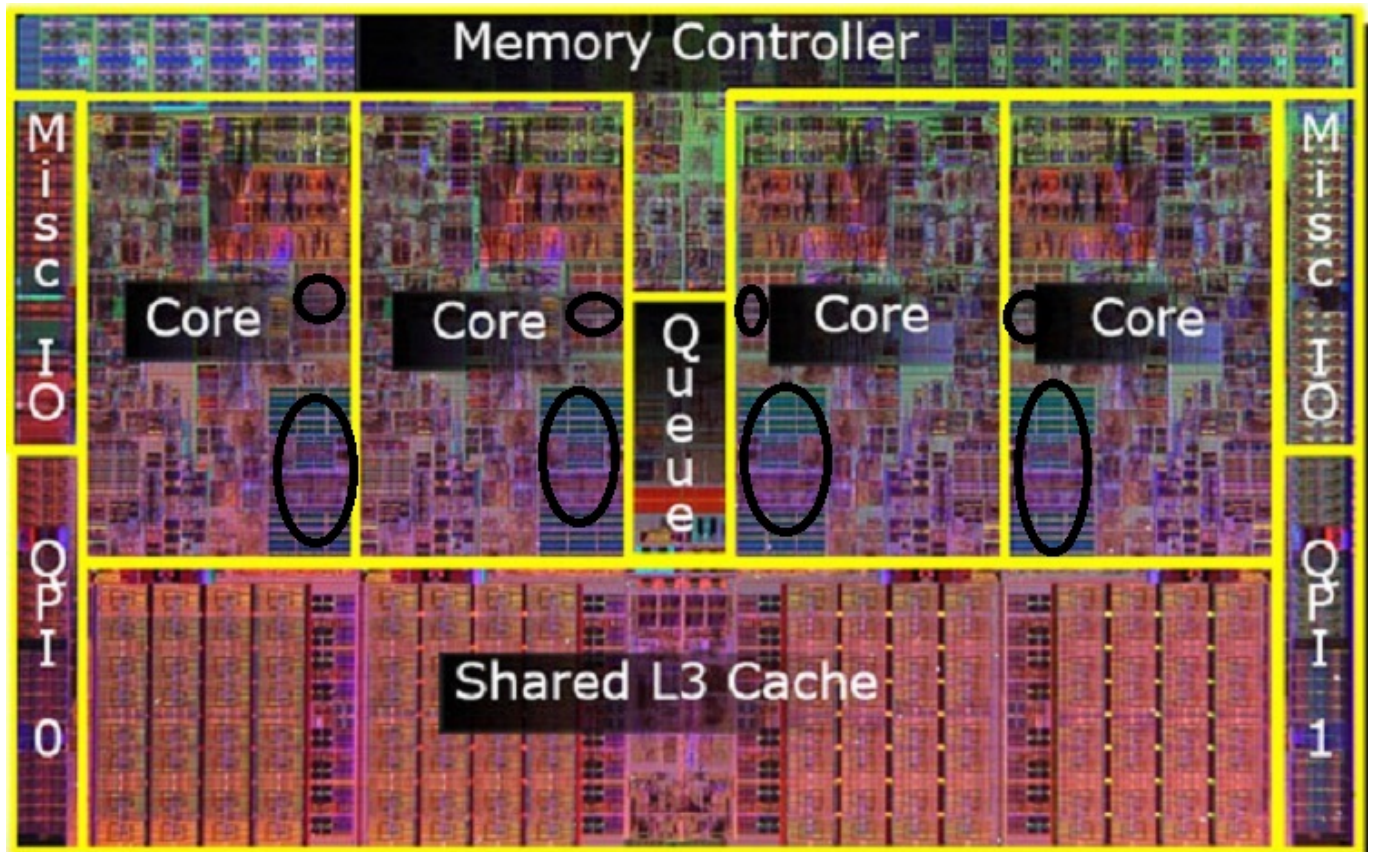


Imagen de un procesador Intel Nehalem. Como vemos los cuatro núcleos tienen acceso a la caché L3. Los círculos grandes identifican la caché L2, y los pequeños la caché L1.

Ya hemos visto qué es y qué trabajo realiza la memoria caché. En general es un apoyo importante para el procesador que, como dijimos, se divide en un total de **tres niveles** generales al que podemos sumar un cuarto nivel formado por la caché L4, que no resulta nada común y que juega un papel particular, como vamos a explicar más adelante,

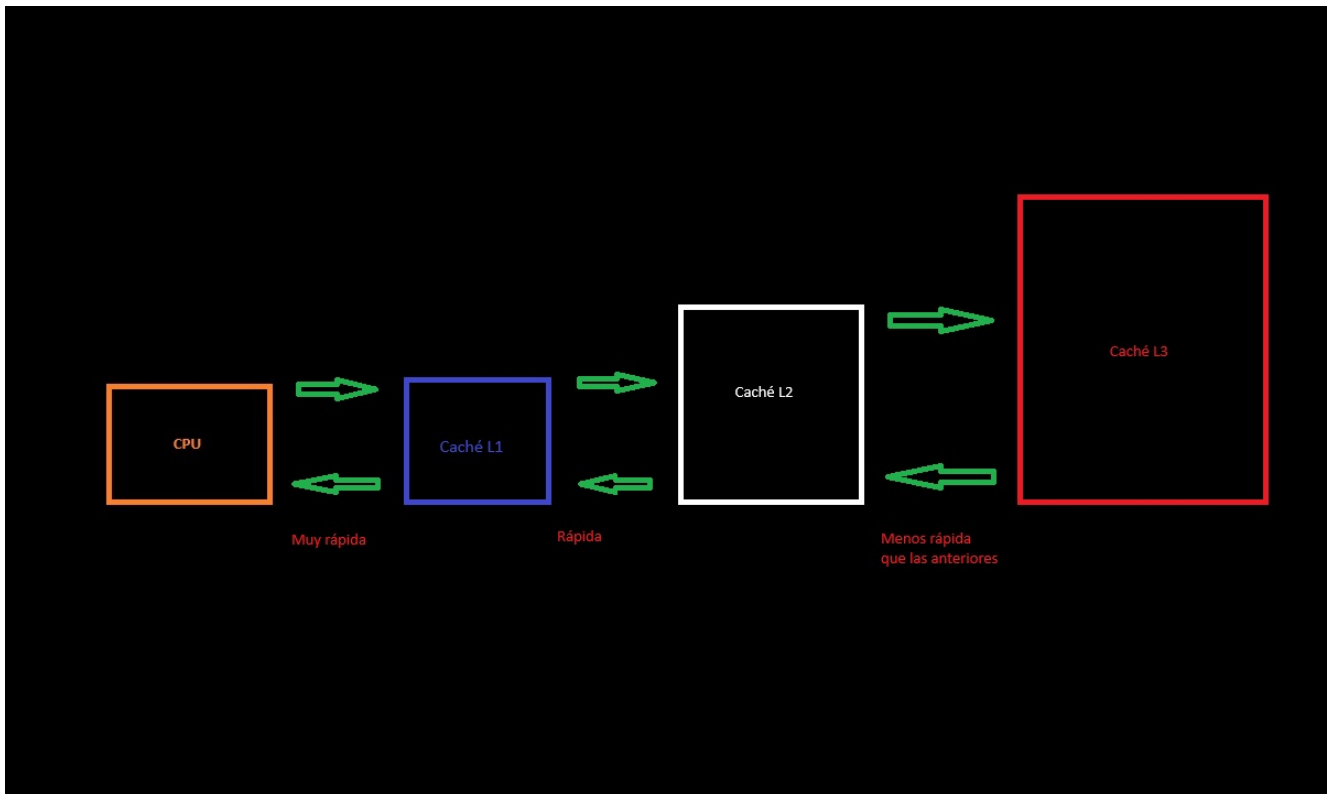
La diferenciación entre memoria caché L1, L2 y L3 obedece a un orden de jerarquía establecido por cercanía al procesador, velocidad y capacidad. Utilizaremos dos ejemplos extremos para que veáis las grandes diferencias que pueden haber en términos de cantidad en función de cada CPU, aunque **la función de la memoria caché será la misma** en todos los casos.

- **Caché L1:** es el nivel más básico, la más cercana al procesador y la más rápida. También es la que menos capacidad tiene, por ejemplo un procesador de dos

núcleos como el Pentium G4560 tiene 64 KB en total (32 KB por núcleo), aunque los procesadores más potentes y avanzados pueden contar con varios megabytes de este tipo de caché. Por ejemplo, el AMD EPYC 9654 tiene 6 MB de caché L1.

- **Caché L2:** es un nivel intermedio que presenta un buen equilibrio entre capacidad, cercanía y velocidad. Siguiendo el ejemplo anterior el Pentium G4560 tiene 512 KB (256 KB por núcleo) de caché L2, pero un AMD EPYC 9754 llega a los 96 MB en total, ya que tiene 1 MB por núcleo.
- **Caché L3:** posiciona en un nivel inferior a la anterior tanto en cercanía como en velocidad, pero tiene una capacidad mucho mayor. El Pentium G4560 tiene 3 MB de caché L3 compartida en sus dos núcleos, mientras que el AMD viene con un total de 384 MB de caché L3, compartida a razón de 32 MB por cada chiplet de 8 núcleos.
- **Caché L4:** es un tipo de memoria caché poco habitual que se utiliza normalmente como apoyo para mejorar el rendimiento de GPUs integradas. Por ejemplo, el Core i5 5775C venía con 6 MB de caché L3 y 128 MB de eDRAM como caché L4, que se utilizaba como buffer para la gráfica Intel Iris Pro 6200 que integraba. Así se mejoraba el ancho de banda y se reducía el impacto de tener que recurrir a la RAM como memoria gráfica.

Como dije anteriormente, cuando un procesador busca instrucciones y datos que necesita primero recurre a la memoria caché, y lo hace **siguiendo el orden de jerarquía establecido**. Esto quiere decir que primer busca en la caché L1, si no encuentra nada recurre a la caché L2 y finalmente a la caché L3.

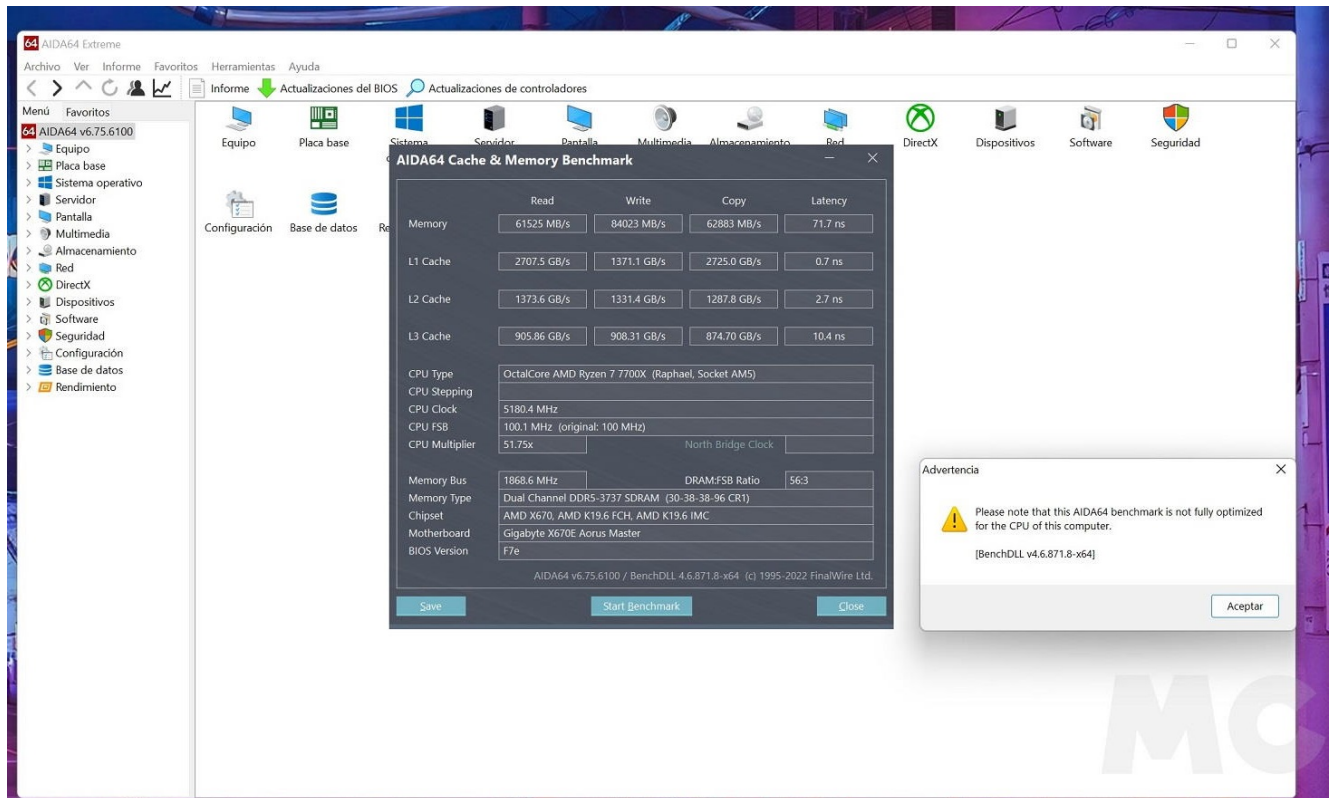


En caso de que ninguna de las cachés contenga lo que está buscando no tiene más opción que **recurrir a la memoria RAM**, y si tampoco está en ella tiene que realizar un ciclo de trabajo completo, tras el cual podrá guardar todo o parte del resultado en la memoria caché o en la memoria RAM, dependiendo del tamaño y de la cantidad de caché que esté disponible.

Para acceder a cada nivel de memoria caché y encontrar, o no, lo que se está buscando tiene que pasar un tiempo determinado. Esto produce **un retraso que se conoce como latencia**, puede tener un gran impacto en el rendimiento y se ve afectado por la distancia a la que se encuentre la caché. La latencia de acceso a la caché L1 es la más baja, y se incrementa de forma considerable en los niveles L2 y L3, sobre todo si esta última está externalizada en un encapsulado propio.

Al acceder a la memoria RAM **la latencia aumenta significativamente**, y es lógico porque esta ya no está integrada en el procesador, sino que utiliza sus propias ranuras en la placa base y se comunica con la CPU a través del bus de memoria. Para que entendáis mejor este tema solo tenéis que ver los datos de la imagen adjunta, una prueba de

rendimiento realizada con AIDA64.

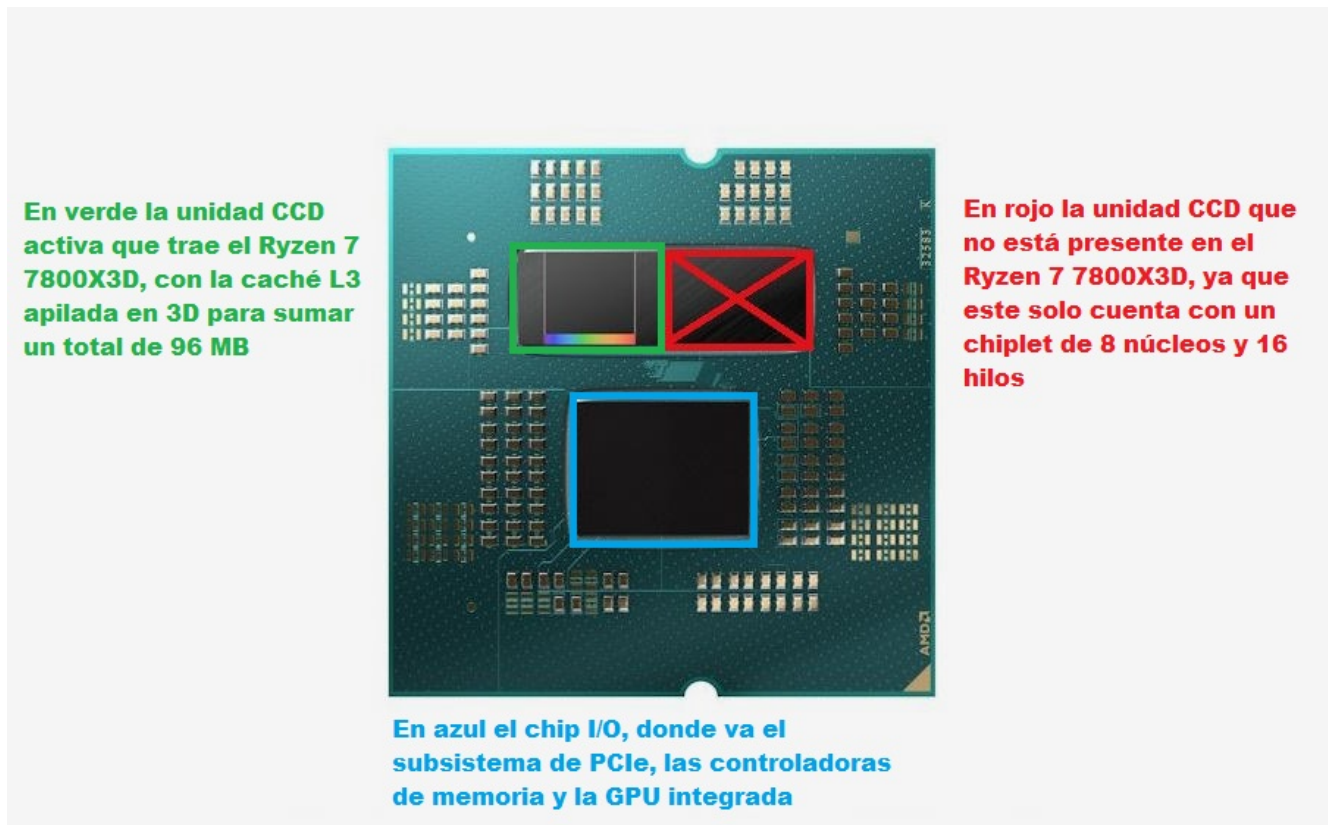


Con un Ryzen 7 7700X la latencia media de la memoria DDR5 utilizada, que funcionaba en doble canal a 6.000 MT/s y CL30, es de 71,7 nanosegundos. La latencia de la caché L1 de ese procesador es de **solo 0,7 nanosegundos**, y tiene un pico de velocidad de lectura de 2.707,5 GB/s. Es muy rápida, y ofrece un rendimiento claramente superior al de la RAM, pero su tamaño es de solo **512 KB** en este procesador.

Cuando saltamos a la caché L2 la capacidad total sube a **8 MB**, pero la latencia **aumenta a los 2,7 nanosegundos** y la velocidad general cae, puesto que tenemos un pico en lectura de 1.373,6 GB/s, pero los valores siguen siendo muy buenos. La caché L3 tiene una capacidad de **32 MB** en este procesador, y como podemos ver su latencia es de 10,4 nanosegundos. La velocidad de lectura cae hasta los **905,8 GB/s**.

Esta explicación nos ayuda a entender también la referencia que hemos hecho anteriormente a los diseños que integraban la caché L2 y L3 en la placa base. **La distancia entre componentes** aumentaba enormemente la latencia y reducía en gran medida el

rendimiento, algo que se pudo superar al montar los tres tipos principales de caché en el encapsulado de la CPU.



La caché L3 es la que mayor capacidad tiene, pero al aumentar su tamaño se incrementa el espacio que esta ocupa a nivel de silicio, algo que actúa como un techo de cristal que AMD ha sido capaz de superar gracias al apilado en 3D. Sus procesadores Ryzen 5000X3D y 7000X3D tienen 32 MB de caché L3 por cada unidad CCD o chiplet, y traen además **un chiplet con 64 MB de caché L3 adicional apilado** sobre una de esas unidades CCD.

En la imagen podéis ver el diseño que ha utilizado AMD en el Ryzen 7 7800X3D. Gracias a ese chip adicional de caché apilada en 3D este procesador cuenta con un total de **96 MB de caché L3**. Este tipo de memoria caché **influye mucho en el rendimiento en juegos**, porque permite almacenar más datos e instrucciones que son fundamentales para este tipo de aplicaciones, y **aumenta en gran medida la tasa de aciertos del procesador al buscar en la caché**, reduciendo con ello su dependencia de la memoria RAM y la cantidad de llamadas a dicho componente.

Notas finales

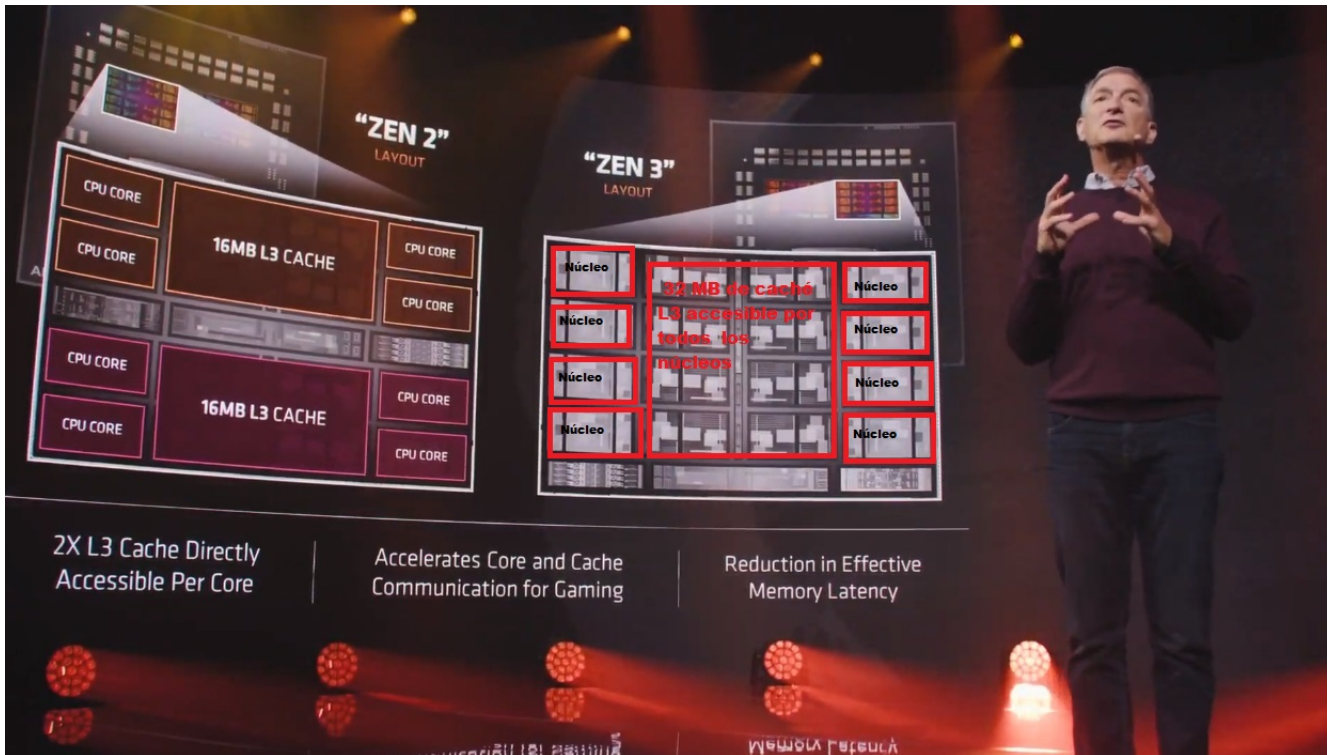


Imagen de un procesador Ryzen 5000 basado en Zen 3. Como vemos la caché L3 se comparte en bloques de ocho núcleos, un diseño que se ha mantenido con Zen 4.

A día de hoy la caché ha dejado de ser un problema, ya que **incluso los procesadores más básicos** vienen equipados con una cantidad razonable de memoria caché L3, algo que, sin embargo, no ocurría hace unos años, cuando lo normal era que este tipo de memoria quedase reservada exclusivamente a los procesadores más potentes.

No debemos obsesionarnos con este tema. Es cierto que el impacto de la memoria caché en el rendimiento general del sistema puede ser grande, pero este **no tiene porque ser uniforme**, ya que depende de otros aspectos como la arquitectura del procesador y su potencia bruta, y también, de forma indirecta, de la velocidad de trabajo y la latencia de la memoria RAM.

Por ejemplo, la arquitectura de núcleo monolítico que utilizan los procesadores Intel permite trabajar con memorias caché L3

más rápidas que tienen una latencia inferior, mientras que en el caso de los procesadores AMD Ryzen de primera generación ocurría todo lo contrario, debido a la arquitectura MCM que separa la caché L3 **en bloques de núcleos**, conocidos como unidades CCX en Zen y Zen+.

Los núcleos de un procesador Intel con ese diseño monolítico también pueden acceder a toda la memoria caché L3 disponible, cosa que no ocurría en los procesadores Zen 2 y anteriores. Los Ryzen 3000 solo podían acceder a 16 MB de caché L3 por cada grupo de cuatro núcleos, a pesar de que habían 32 MB disponibles. **Por suerte esto cambió con Zen 3 y Zen 4**, ya que ambas arquitecturas hicieron posible que los 8 núcleos de cada unidad CCD pudieran acceder a los 32 MB de caché L3 disponibles.

Si contamos con una memoria RAM más rápida **es posible reducir también el impacto** que tiene la ausencia de datos necesarios en la caché, que como vimos obliga al procesador a buscar en la RAM. Precisamente en este sentido, para aumentar la tasa de aciertos y reducir la dependencia de la memoria RAM, AMD optó por la caché L3 apilada en 3D, un movimiento que le ha salido realmente bien, y que le ha permitido tener en su catálogo el procesador más potente para juegos, el Ryzen 7 7800X3D, todo gracias a esos 96 MB de L3 que incorpora.

Fuente: muycomputer.