

Ejercicios Tema 4

Centros de Datos UZ, 2025-26

Lacueva Sacristán, Héctor

21/10/2025

Índice

1	T4.3 IBM Telum II chip (2h, mayor parte del tiempo, encontrar información)	1
1.1	Introducción	1
1.2	Caché L2	1
1.3	Caché L3 Virtual	1
1.4	Caché L4 Virtual	1
1.5	Conclusión	2
2	Referencias	2

1 T4.3 IBM Telum II chip (2h, mayor parte del tiempo, encontrar información)

IA: Si, Copilot, para que hiciera un resumen de la página (*Telum II at Hot Chips 2024* 2024) después de haberla leido, este luego ha sido comprobado y adaptado a mi manera.

Lee el artículo (*IBM Boosts Mainframes with 50% More AI Performance* 2024) y profundiza en una aspecto que te haya llamado la atención.

1.1 Introducción

A diferencia de la mayoría de los procesadores comerciales, IBM adopta un enfoque **no convencional y jerárquico** para el manejo de la memoria caché. En lugar de utilizar niveles físicos tradicionales (L1, L2 y L3), **Telum II emplea una arquitectura de caché virtual**, donde las memorias L2 cooperan para simular niveles superiores (L3 y L4) de manera dinámica y eficiente.

1.2 Caché L2

Según (*Telum II at Hot Chips 2024* 2024), Telum II tiene 10 cachés L2 de 36MB “on-chip” (8 asociados a los cores, otra al DPU (accelerador de entrada/salida) y la otra no está asociada a nada). Cuenta por lo tanto, con 360MB en un chip con una latencia de apenas 3,6 ns, superando a muchas arquitecturas tradicionales.

1.3 Caché L3 Virtual

En lugar de una caché L3 física, Telum II implementa una caché L3 virtual que aprovecha la capacidad agregada de las L2. Cuando una L2 necesita liberar espacio, los datos expulsados no se envían directamente a la memoria principal, sino que se reubican en otra L2 menos saturada, según un indicador interno denominado Saturation Metric.

La L2 no asociada a ningún núcleo actúa como un buffer adicional, sirviendo como destino preferente para estas líneas expulsadas. Además, el sistema evita duplicaciones: si una línea ya existe en otra L2, se transfiere la propiedad sin replicar el contenido.

1.4 Caché L4 Virtual

La idea de las **cachés virtuales** no se limita a un solo chip. En sistemas mainframe, varios procesadores Telum II pueden trabajar juntos dentro de un mismo **módulo o “drawer”**. IBM extiende su estrategia de caché para crear una **caché L4 virtual de 2.8 GB**, distribuida entre distintos procesadores.

Cuando una línea de la L3 virtual necesita ser reemplazada, puede almacenarse en la memoria caché disponible de otro chip del sistema. Así, IBM logra **aprovechar toda la capacidad combinada de las L2** de múltiples procesadores, creando un sistema de memoria en varios niveles con una **latencia sorprendentemente baja (48.5 ns)**, incluso al cruzar los límites de los chips.

1.5 Conclusión

El sistema de caché del **IBM Telum II** representa una solución altamente sofisticada y eficiente para los problemas clásicos de latencia y ancho de banda de la memoria. Mediante la creación de **niveles de caché virtuales (L3 y L4)** a partir de la cooperación entre las L2, IBM consigue reducir drásticamente los accesos a memoria externa, mejorar el rendimiento por hilo y mantener una excelente eficiencia energética.

En resumen, **Telum II transforma su enorme caché L2 en una red inteligente de memoria compartida**, capaz de adaptarse dinámicamente a las necesidades del sistema, y extendiendo el concepto de caché más allá del propio chip hacia el conjunto completo del mainframe.

2 Referencias

IBM Boosts Mainframes with 50% More AI Performance: Z17 Features Telum II Chip with AI Accelerators. 2024. <https://www.tomshardware.com/tech-industry/ibm-boots-mainframes-with-50-percent-more-ai-performance-z17-features-telum-ii-chip-with-ai-accelerators>.

Telum II at Hot Chips 2024: Mainframe with a Unique Caching Strategy. 2024. <https://chipsandcheese.com/p/telum-ii-at-hot-chips-2024-mainframe-with-a-unique-caching-strategy>.