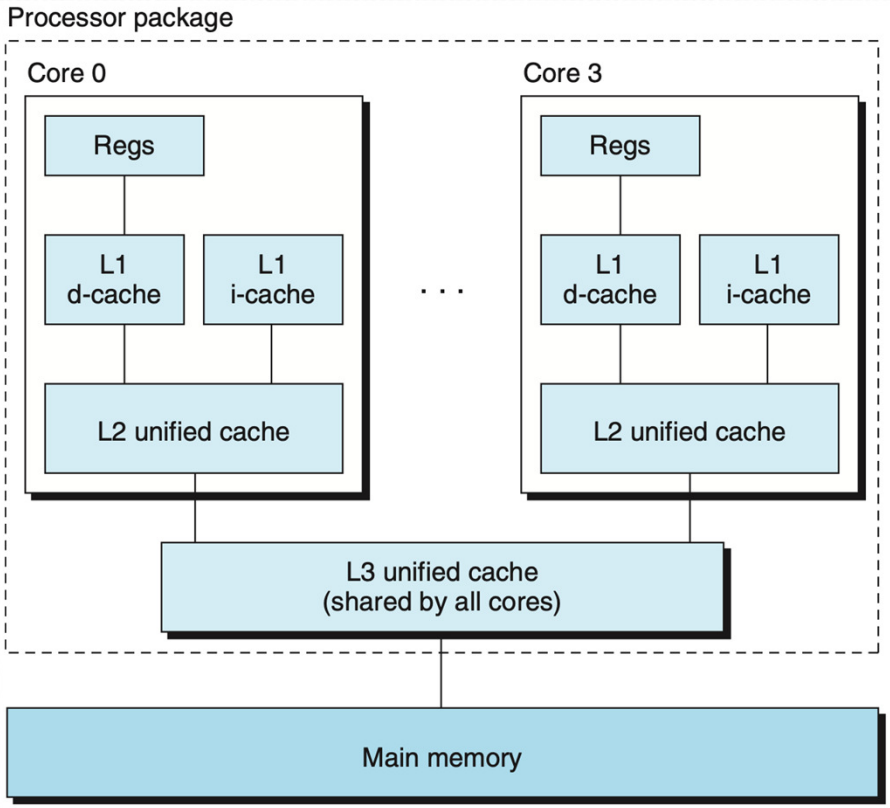
Cachés multinivel

Cachés multinivel

- Tener uno o mas niveles de cachés.
- Los sistemas de dos niveles son comunes.
- El caché de segundo nivel (L2) se accesa cuando hay una falla en el caché primario (L1).
- Si el caché L2 contiene el dato, el castigo por falla de L1 es el tiempo de acceso de L2 que es mucho menor que el tiempo de acceso a la memoria
- Si el dato no está en L1 ni en L2, se accesa la memoria y el castigo por falla es mayor.

Intel Core i7



Fuente: CS-APP, p. 668

Intel Core i7

Cache type	Access time (cycles)	Cache size (C)	Assoc. (E)	Block size (B)	Sets (S)
L1 i-cache	4	32 KB	8	64 B	64
L1 d-cache	4	32 KB	8	64 B	64
L2 unified cache	10	256 KB	8	64 B	512
L3 unified cache	40–75	8 MB	16	64 B	8,192

Fuente: CS-APP, p. 668

Ejemplo

- Se tiene una CPU con:
  - CPI = 1 con caché perfecto (sin fallas).
  - Velocidad de reloj = 5 GHz.
- Tiempo de acceso de la memoria = 100 ns.
- Tasa de fallas del caché primario = 2%.
- ¿Qué tan rápido es el sistema si se agrega otro nivel de caché con tiempo de acceso de 5 ns y capaz de reducir la tasa de fallas a la memoria a 0.5%?

Ejemplo

- El castigo por falla es:

$$\text{castigo por falla} = \frac{\text{tiempo de acceso}}{\text{periodo de reloj}}$$

- Para la memoria principal el castigo es:

100 ns / 0.2 ns = 500 ciclos.

- El CPI efectivo con un nivel de caché es:

$$\text{CPI total} = \text{CPI base} + \text{ciclos de detención por instrucción}$$

- Para la CPU con un nivel de caché:

CPI = 1 + 0.02 x 500 = 11.0

Ejemplo

- Con dos niveles de caché, una falla en el caché L1 se puede resolver por L2 o por la memoria.
- Si se resuelve en L2, el tiempo de acceso a L2 es el castigo por falla.
- En otro caso, el castigo por falla es la suma de los tiempos de acceso a L2 y a la memoria.
- El castigo por falla por un acceso a L2 es:

5 ns / 0.2 ns = 25 ciclos

Ejemplo

- El CPI efectivo para un caché de dos niveles:

  CPI total = CPI base + detenciones primarias por instrucción + detenciones secundarias por instrucción

- Para la CPU con dos niveles de caché:

  $1.0 + 0.02 \times 25 + 0.005 \times 500 = 1 + 0.5 + 2.5 = 4$

- Por lo tanto, la CPU con dos niveles de caché es más rápida que la CPU con un nivel en:

  $11 / 4 = 2.75$

Cachés multinivel

- Se introducen dos nuevos conceptos:
- Tasa de fallas global (global miss rate) es la fracción de referencias que fallan en todos los niveles.
- Tasa de fallas local (local miss rate) es la fracción de referencias que fallan en un nivel.
- El caché primario es mas pequeño y con menor tiempo de acceso que los secundarios.

Ejemplo

- Suponer que de 1000 referencias a la memoria, hay 40 fallas en L1 y 20 fallas en L2.
- Calcular las tasa de fallas.
- Para L1: tasa de fallas = 40 / 1000 = 4%.
- Para L2:
  - Tasa de fallas local = 20 / 40 = 50%.
  - Tasa de fallas global = 20 / 1000 = 2%.

Ejemplo

- Con los datos anteriores, suponer lo siguiente:
  - El castigo por falla de L2 a la memoria es de 200 ciclos.
  - El tiempo de acceso (hit time) de L2 es de 10 ciclos.
  - El tiempo de acceso de L1 es de 1 ciclo.
  - Hay 1.5 referencias a la memoria por instrucción.
- Calcular el AMAT.

Ejemplo

- $AMAT = Hit\ time_{L1} + Miss\ rate_{L1} \times (Hit\ time_{L2} + Miss\ rate_{L2} \times Miss\ penalty_{L2})$
- $AMAT = 1 + 0.04 \times (10 + 0.5 \times 200) =$
- $1 + 0.04 \times 110 = 5.4$ ciclos