Memoria

Introducción

Locality

- **Locality.** Principio que establece que en un instante dado un programa accesa un conjunto pequeño de ítems (instrucciones o datos).
- **Locality en tiempo.** Si un ítem es accesado, es posible que vuelva a ser accesado pronto.
- **Locality en espacio.** Si se accesa un ítem, es posible que los ítems con direcciones cercanas van a ser accesados pronto.
- Para tomar ventaja del principio de locality, se establece una **jerarquía de memoria.**

Jerarquía de memoria

- Una jerarquía de memoria consiste de múltiples niveles de memoria con diferentes velocidades y tamaños.
- Las memorias rápidas son más caras que las memorias lentas y por lo tanto son más pequeñas.
- La memoria rápida está cerca del procesador y la memoria lenta está debajo.

Jerarquía de memoria

Speed	Processor	Size	Cost (\$/bit)	Current technology
Fastest	Memory	Smallest	Highest	SRAM
	Memory			DRAM
Slowest	Memory	Biggest	Lowest	Magnetic disk or Flash memory

Fuente: COD-RISC2, p. 389

Explicación

- La jerarquía de memoria consta de varios niveles pero la información se copia solo entre niveles adyacentes a la vez.
- Entre el caché y la memoria principal, la unidad mínima de información que se copia es un **bloque** o **línea**.
- Entre la memoria principal y el disco duro, la unidad se llama **página**.
- Un dato no puede estar en un nivel a menos que esté también en el nivel inferior.

Definiciones

- **Bloque o línea.** Unidad mínima de información que almacena el caché. Tamaño típico: 64 bytes.
- **Página.** Unidad mínima de información que almacena la memoria principal. Tamaño típico: 4 / 8 KB.
- **Éxito (hit).** Cuando el dato solicitado por el procesador está en el nivel superior.
- **Falla (miss).** Cuando el dato solicitado por el procesador *no* está en el nivel superior y se tiene que buscar en el nivel inferior.

Definiciones

- **Tasa de éxitos (hit rate)**. Fracción de los accesos a la memoria encontrados en el nivel superior.
- **Tasa de fallas (miss rate)**. Fracción de los accesos a la memoria no encontrados en el nivel superior. Equivale a $(1 - \text{tasa de éxitos})$.
- **Tiempo de éxito (hit time)**. Es el tiempo para determinar si el acceso es un éxito o una falla.
- **Castigo por falla (miss penalty)**. Tiempo que se pierde cuándo se produce una falla.

Tecnologías de memoria

Memory technology	Typical access time	\$ per GB in 2004
SRAM	0.5–5 ns	\$4000–\$10,000
DRAM	50–70 ns	\$100–\$200
Magnetic disk	5,000,000–20,000,000 ns	\$0.50–\$2

Memory technology	Typical access time	\$ per GiB in 2020
SRAM semiconductor memory	0.5–2.5 ns	\$500–\$1000
DRAM semiconductor memory	50–70 ns	\$3–\$6
Flash semiconductor memory	5,000–50,000 ns	\$0.06–\$0.12
Magnetic disk	5,000,000–20,000,000 ns	\$0.01–\$0.02

Tecnologías de memoria

Year introduced	Chip size	\$ per GiB	Total access time to a new row/column	Average column access time to existing row
1980	64 Kibibit	\$6,480,000	250 ns	150 ns
1983	256 Kibibit	\$1,980,000	185 ns	100 ns
1985	1 Mebibit	\$720,000	135 ns	40 ns
1989	4 Mebibit	\$128,000	110 ns	40 ns
1992	16 Mebibit	\$30,000	90 ns	30 ns
1996	64 Mebibit	\$9,000	60 ns	12 ns
1998	128 Mebibit	\$900	60 ns	10 ns
2000	256 Mebibit	\$840	55 ns	7 ns
2004	512 Mebibit	\$150	50 ns	5 ns
2007	1 Gibibit	\$40	45 ns	1.25 ns
2010	2 Gibibit	\$13	40 ns	1 ns
2012	4 Gibibit	\$5	35 ns	0.8 ns
2015	8 Gibibit	\$7	30 ns	0.6 ns
2018	16 Gibibit	\$6	25 ns	0.4 ns

Fuente: COD-RISC2, p. 394