Organización. Jerarquía de Memoria

Motivación:

- ✓ ¿Cómo clasificamos las técnicas basadas en organización del hardware?
- ✓ ¿Cuáles son las principales técnicas relativas al sistema de memoria?
- ✓ ¿Cuál es el principio de funcionamiento de la jerarquía de memoria?
- ✓ ¿Cómo se organiza la memoria principal del sistema?

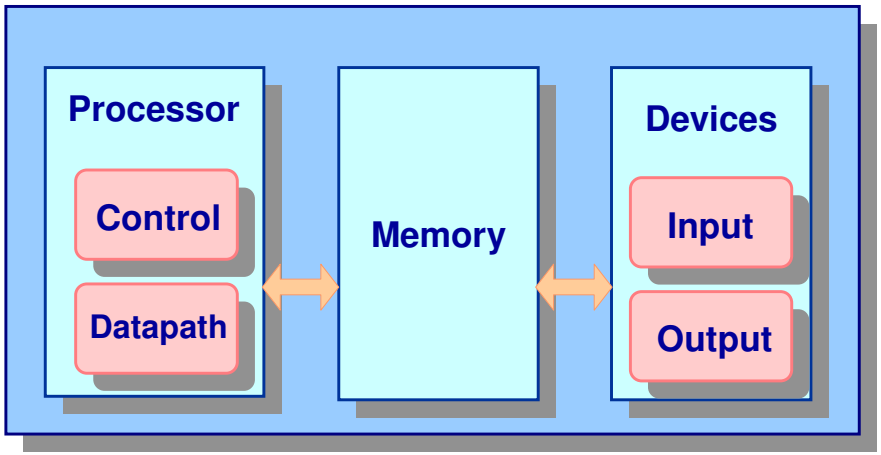
Organización. Jerarquía de Memoria

- Técnicas basadas en Organización del Hardware
- Sistema de Memoria. Jerarquía de Memoria
- Organización de la Memoria Principal

Organización del Hardware

- MEMORIA
  - ✓ Jerarquía de Memoria
- PROCESADOR
  - ✓ Segmentación
    - ejecución solapada de instrucciones
  - ✓ Paralelismo (➔ realización de actividades en paralelo)
    - Múltiples Unidades Funcionales (diferentes/replicadas)
    - Superescalabilidad:
    - Múltiples núcleos:
- E/S
  - ✓ Arrays de Discos (Sistemas RAID)
    (Redundant Array of Independent Disks)

Sistema de Memoria



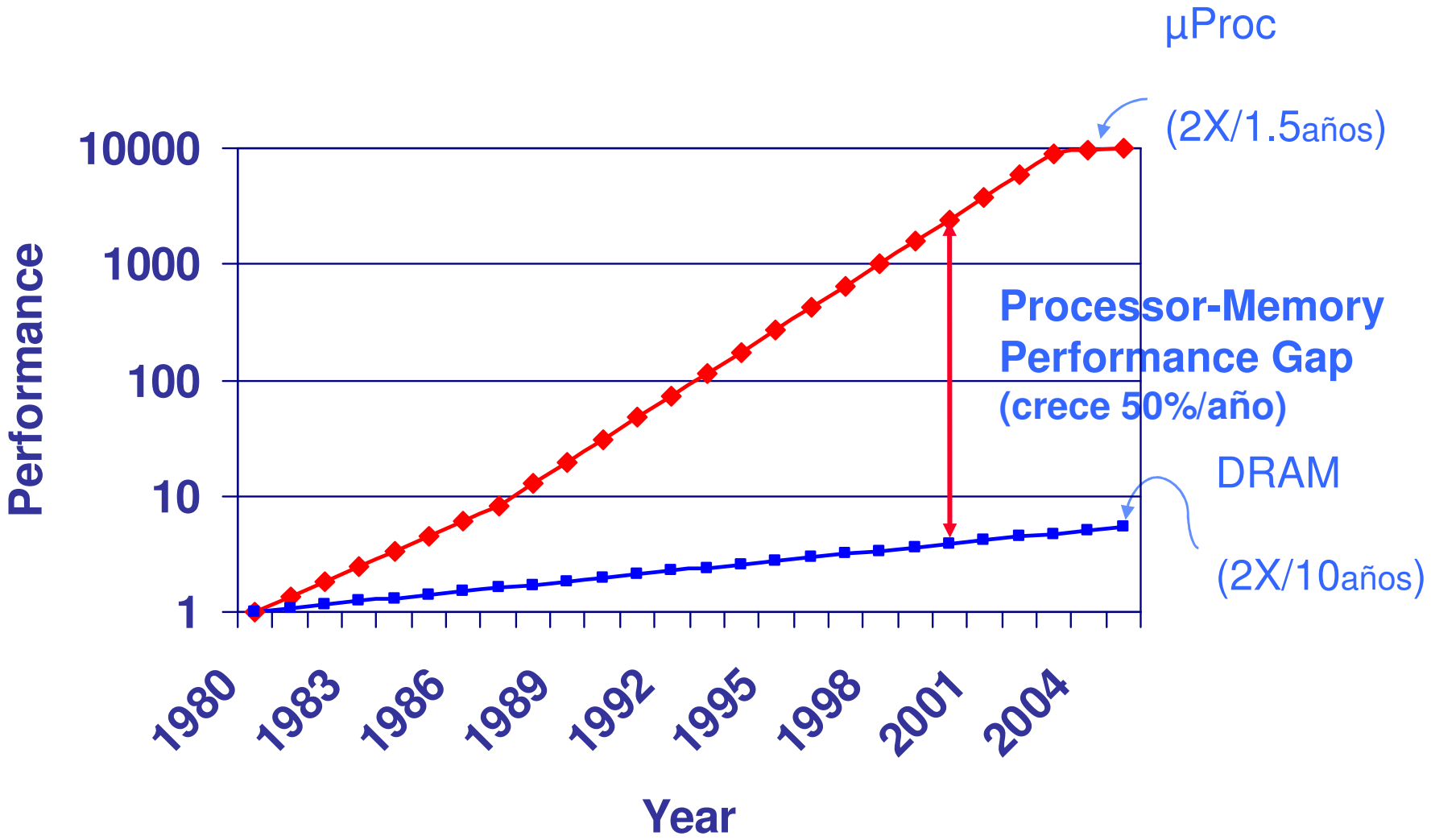
- **El acceso a memoria condiciona en gran medida el rendimiento del computador**

AB_{MEM} = Ancho de banda de memoria
AB_{CPU} = Ancho de Banda de CPU
$AB_{E/S}$ = Ancho de banda de E/S

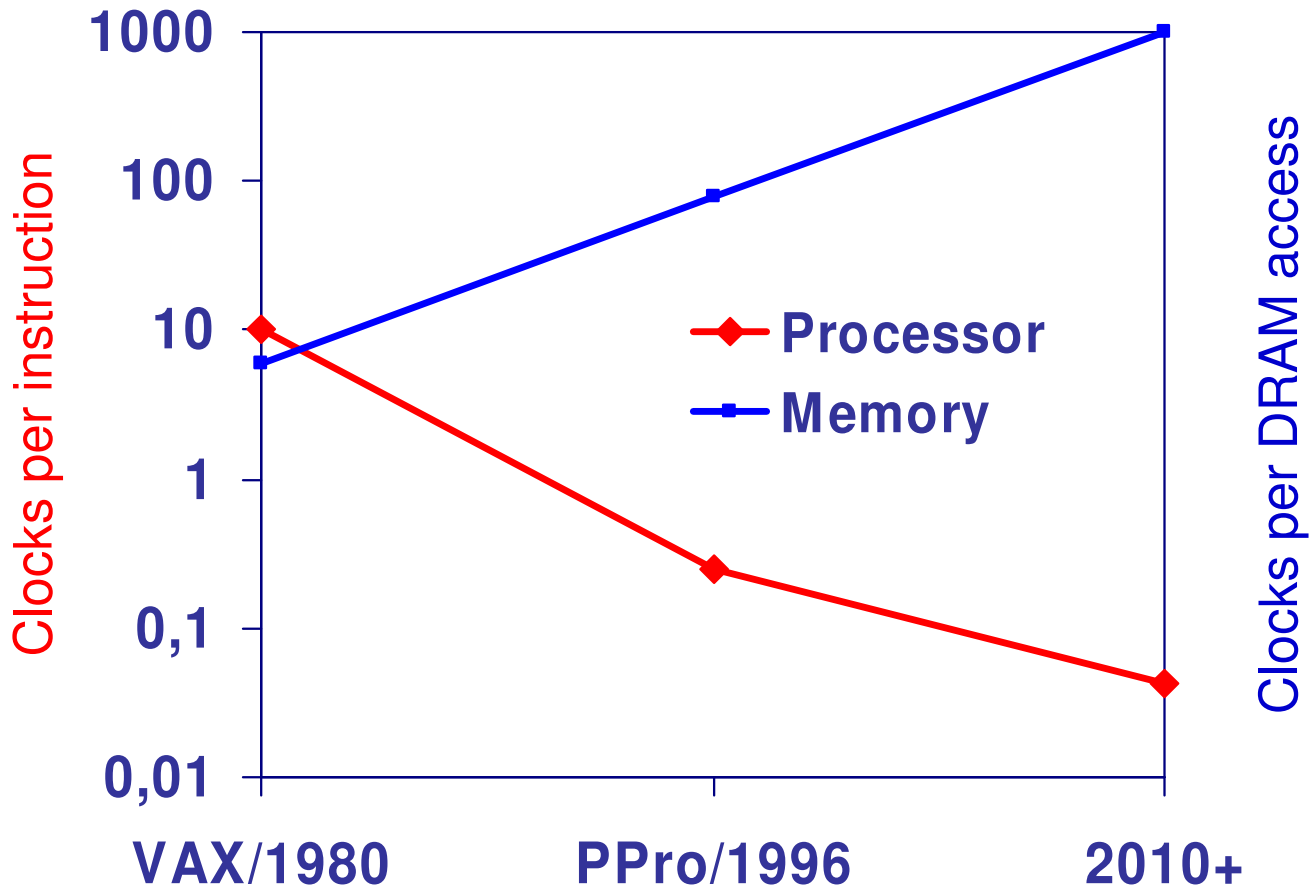
Relación ideal a cumplir en un computador de altas prestaciones:

Problema: Velocidad (CPUs) >> Velocidad (MEMs)

Rendimiento de Procesadores y Memorias



Limitaciones de la Memoria



Impacto de la Memoria sobre el Rendimiento

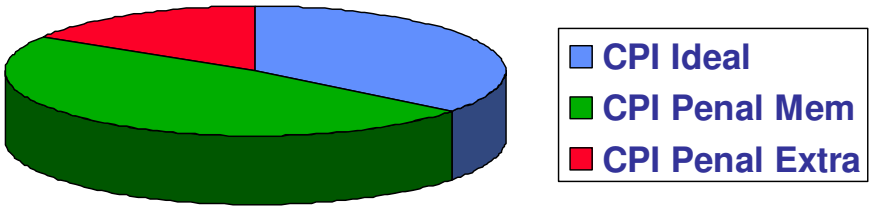
- Suponer que el procesador ejecuta
  - ✓ 50% arit/log, 30% ld/st, 20% control
  - ✓ CPI ideal = 1.1

  y que el 10% de los accesos a memoria penalizan 50 ciclos
- CPI = CPI ideal + ciclos perdidos por instrucción

  = 1.1 ciclos + [0.3(mem/inst) x 0.1(penal/mem) x 50(ciclos/penal)]

  = 1.1 ciclos + 1.5 ciclos = 2.6 ciclos

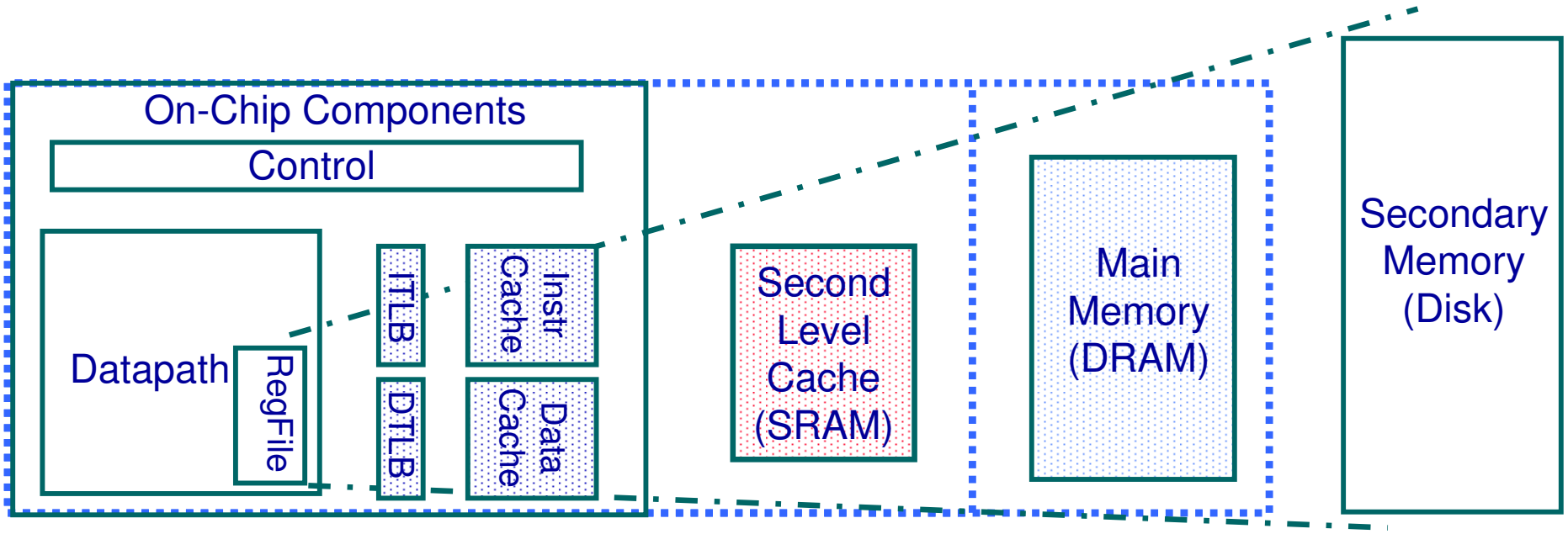
  ➔
- Cada 1% de penalizaciones extra implica un aumento de 0.5 en el CPI



Jerarquía de Memoria

- **Hecho**: las memorias baratas y de elevada capacidad (DRAM) son lentas y las rápidas (SRAM) son caras y de baja capacidad
- ¿Como podemos diseñar una memoria que de la ilusión de ser de elevada capacidad, barata y al mismo tiempo rápida la mayor parte del tiempo?
  - ✓
  - ✓
- **Jerarquía de memoria**
  - ✓ organizando el almacenamiento en varios niveles jerárquicos …
  - ✓ y aprovechando los **principios de localidad temporal y espacial** …
  - ✓ se puede presentar al usuario tanta memoria como permite la tecnología barata …
  - ✓ pero con la velocidad que proporciona la tecnología más rápida

Jerarquía de Memoria típica

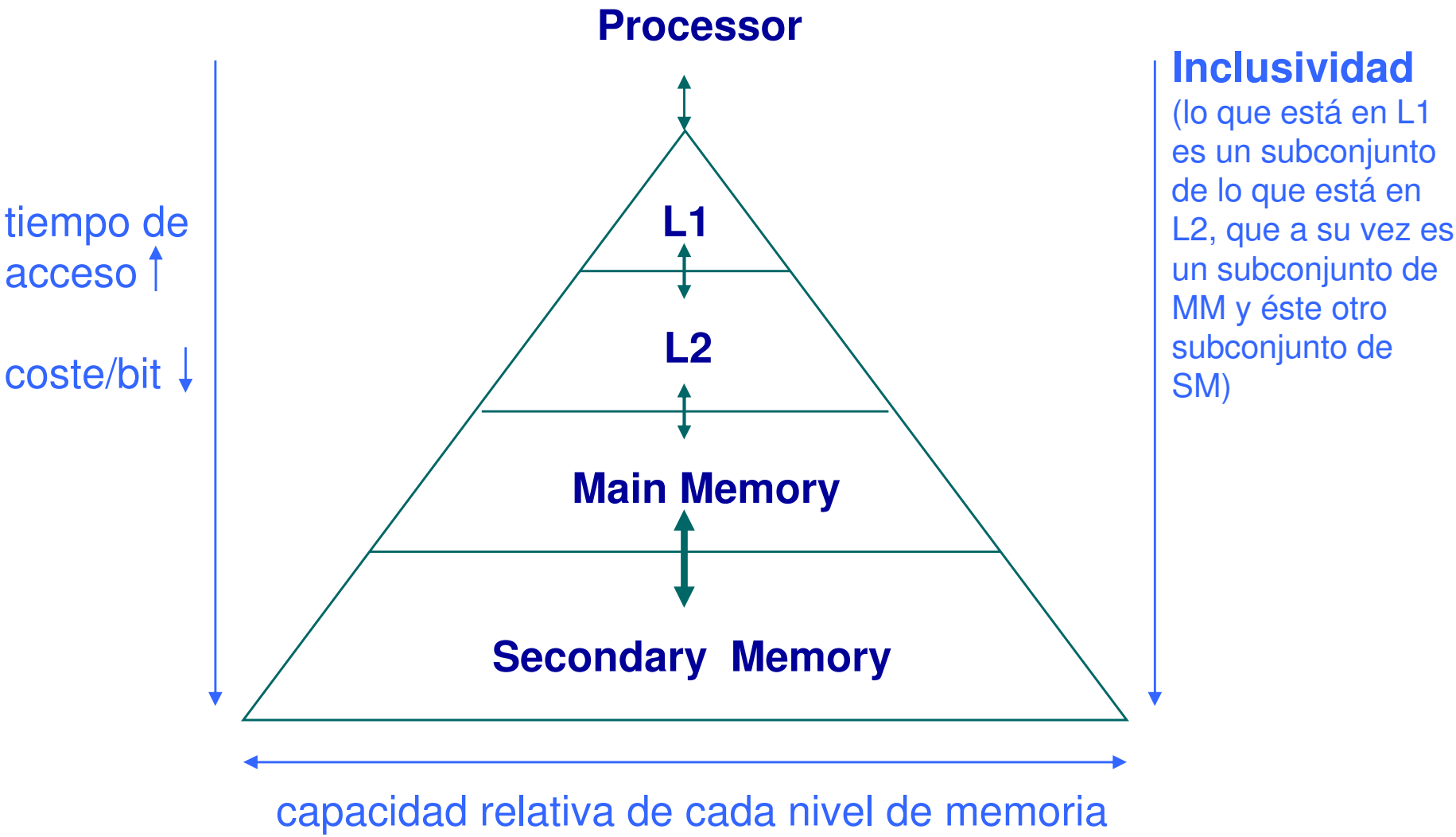


Velocidad ***(ciclos):*** ½'s — 1,000's

Tamaño ***(bytes):*** 100's — G's - T's

Coste ***(€/bit):*** el mayor — el menor

Características de la Jerarquía de Memoria



Tecnologías de Memoria implicadas

- Las Caches utilizan memoria *SRAM* por velocidad
  - ✓ Baja densidad (6 transistores/celda), alto consumo, muy rápidas pero caras
  - ✓ Estática: el contenido almacenado permanece mientras haya alimentación
- La Memoria Principal utiliza *DRAM* por tamaño
  - ✓ Alta densidad (1 transistor por celda), bajo consumo, baratas pero lentas
  - ✓ Dinámicas: necesitan ser "refrescadas" periódicamente (~ cada 8 ms)

Métricas de rendimiento de Memoria

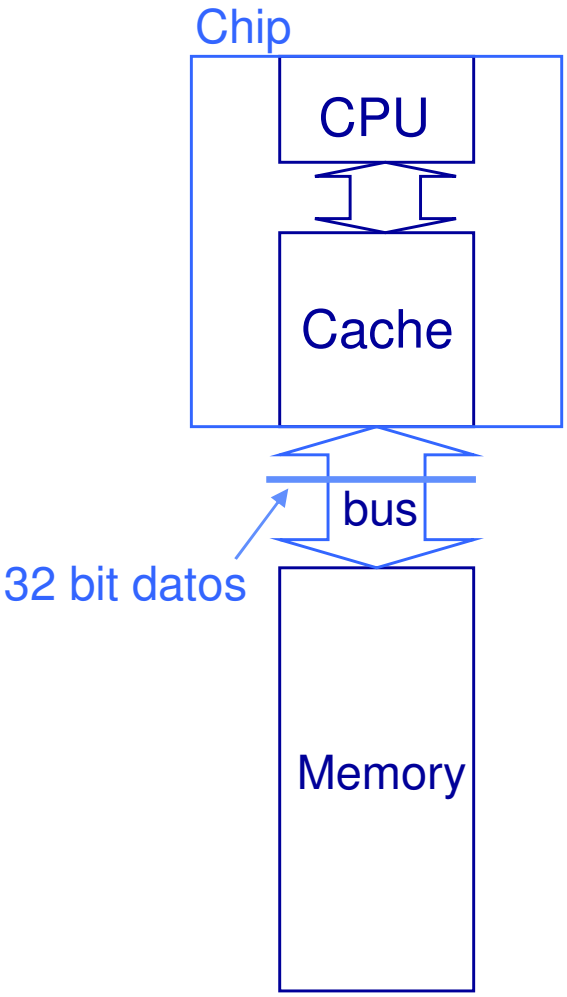
- Latencia: tiempo para acceder a una palabra
  - ✓ ***Tiempo de acceso***: *tiempo entre solicitud y recepción/ escritura del dato*
  - ✓ ***Tiempo de ciclo***: *tiempo entre solicitudes*
  - ✓ DRAMs: *Tiempo de ciclo > Tiempo de acceso*
  - ✓ Tiempos de ciclo típicos:
- Ancho de Banda: *capacidad de proporcionar datos por unidad de tiempo*
  *= ancho del canal de datos * ratio al que puede ser utilizado*
- Tamaño: *DRAM / SRAM =*
- Tiempo de ciclo: *DRAM / SRAM =*
- Coste: *SRAM / DRAM =*

Latencia y Ancho de Banda para DRAM

	DRAM	*Page DRAM*	*FastPage DRAM*	*FastPage DRAM*	*Synch DRAM*	*DDR SDRAM*
Module Width	16b	16b	32b	64b	64b	64b
Year	1980	1983	1986	1993	1997	2000
Mb/chip	0.06	0.25	1	16	64	256
Die size (mm^2)	35	45	70	130	170	204
Pins/chip	16	16	18	20	54	66
BandWidth (MB/s)	**13**	**40**	**160**	**267**	**640**	**1600**
Latency (nsec)	**225**	**170**	**125**	**75**	**62**	**52**

- En el período en que se duplica el ancho de banda de memoria, la latencia tan solo mejora en un factor entre 1.2 y 1.4
- Para poder proporcionar esos niveles de ancho de banda, la DRAM tiene que ser organizada internamente de forma adecuada

Sistemas de Memoria con soporte para CACHE



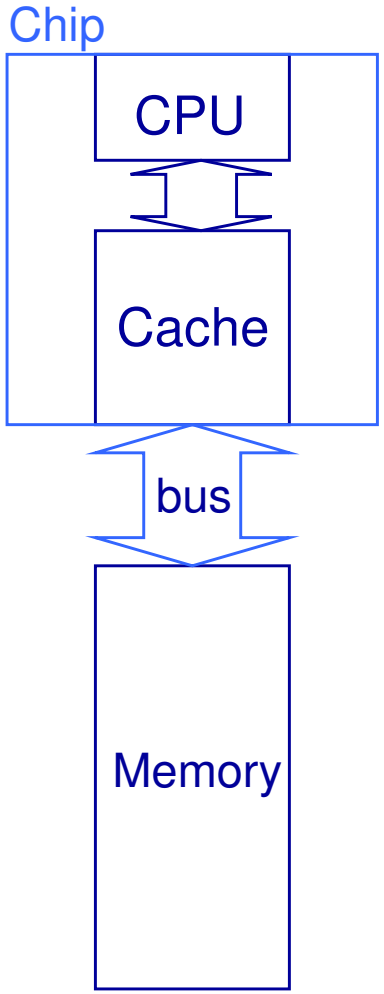
- **La conexión externa del chip y la arquitectura de la memoria condicionan fuertemente el rendimiento del sistema**

Organización de la memoria a nivel de palabra

➔

- Asume:
  - ✓ 1 ciclo de reloj para enviar la dirección
  - ✓ 8 ciclos de tiempo de acceso
  - ✓ 1 ciclo para devolver la palabra de datos
- Ancho de banda *Bus de Memoria-Cache*
  - ✓ Número de *bytes* accedidos desde memoria y transferidos hacia la *cache* por ciclo

Organización de Memoria a nivel de palabra



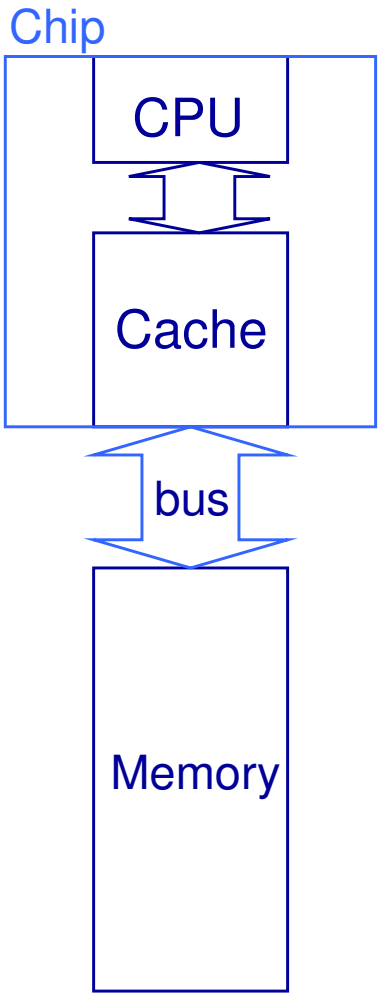
- Si el tamaño de bloque de la *cache* es de una palabra, en caso de fallo de la *cache* tendrá lugar una parada durante el tiempo necesario para cargar la palabra desde memoria principal

  ciclo para enviar la dirección
  ciclos para leer la DRAM
  ciclos para retorno de la palabra
  total de ciclos para el fallo

- Número de *bytes* transferidos por unidad de tiempo (ancho de banda) para fallo de *cache:*

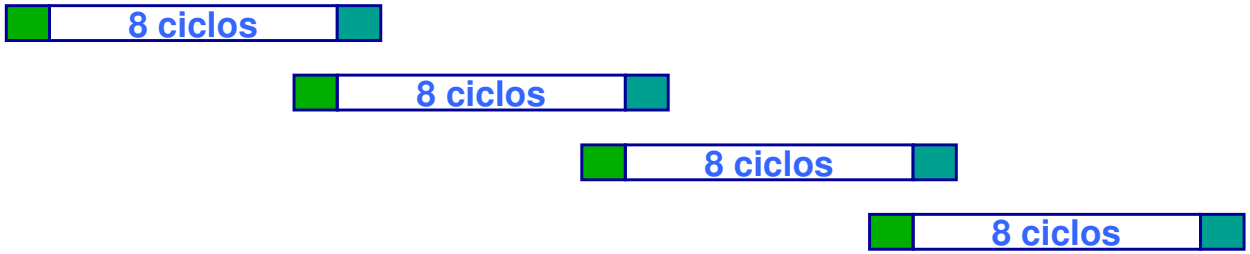
  bytes por ciclo

Organización de Memoria a nivel de palabra



- **¿Que ocurre si el tamaño de bloque es de 4 palabras?**

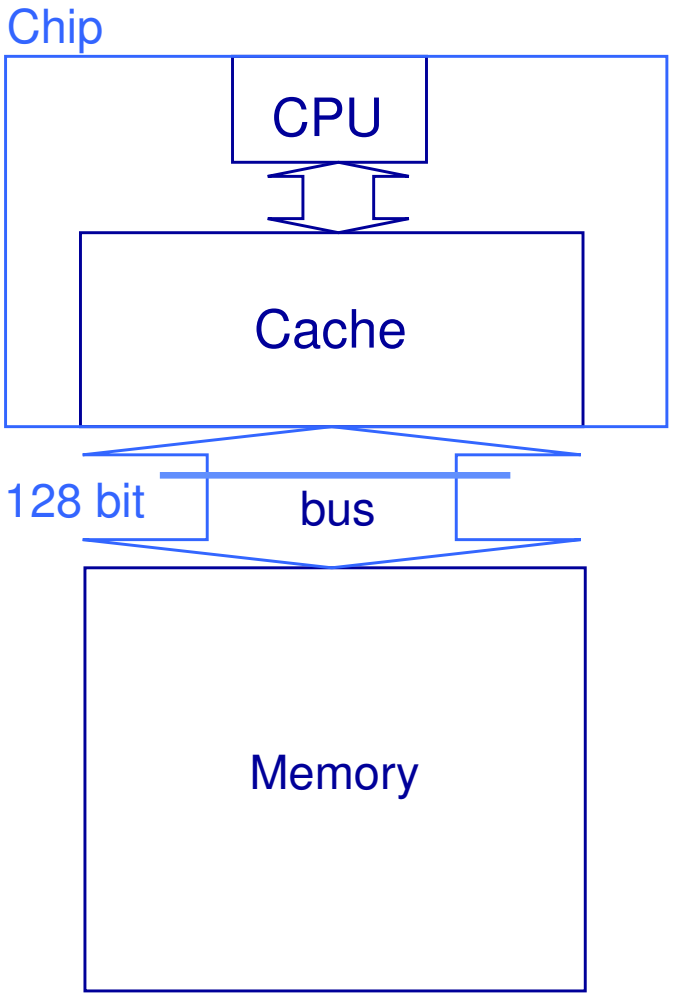
1 ciclo para enviar la dirección
ciclos para leer la DRAM
ciclos para retorno de la última palabra
total de ciclos para el fallo



- Número de *bytes* transferidos por unidad de tiempo (ancho de banda) para fallo de *cache:*

bytes por ciclo

Organización de Memoria más ancha



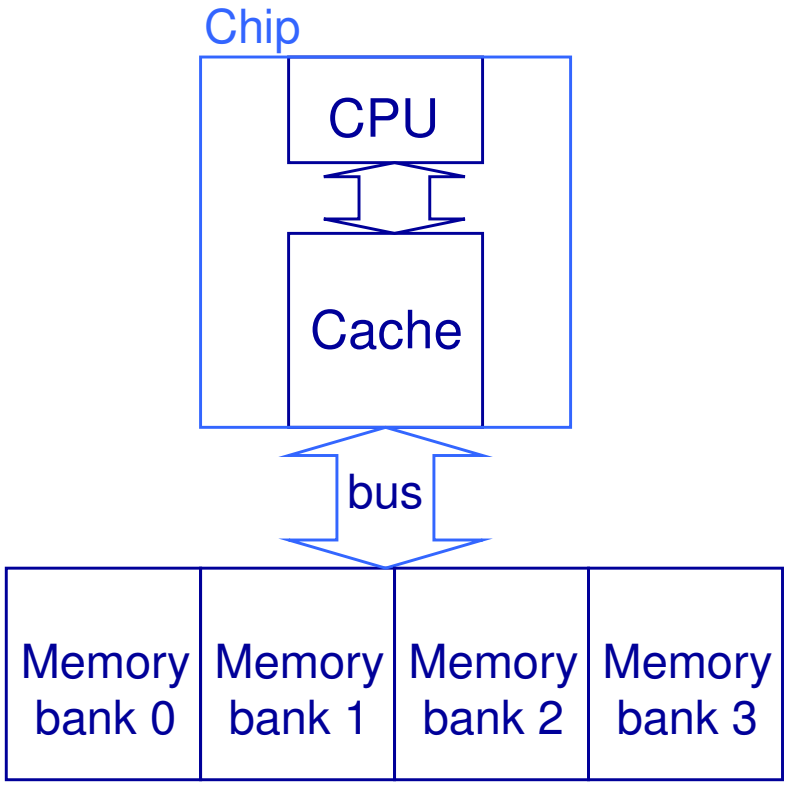
- **¿Que ocurre si la memoria es 4 veces más ancha?**

  ciclo para enviar la dirección
  ciclos para leer la DRAM
  ciclos para retorno de las 4 palabras
  total de ciclos para el fallo

- Número de *bytes* transferidos por unidad de tiempo (ancho de banda) para fallo de *cache:*

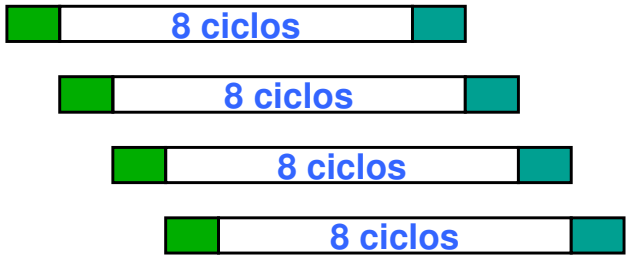
  bytes por ciclo

Organización de Memoria entrelazada *(interleaved)*



- **¿Que ocurre si el tamaño de bloque es de 4 palabras?**

  1 ciclo para enviar la dirección
  8 ciclos para leer la DRAM
  8 ciclos para retorno de la última palabra
  17 total de ciclos para el fallo



- Número de *bytes* transferidos por unidad de tiempo (ancho de banda) para fallo *de cache:*

  4 *bytes* por ciclo

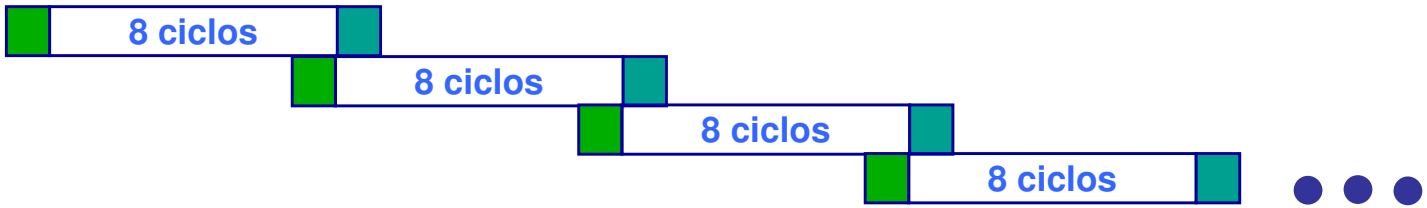
Sistemas de Memoria DRAM (Resumen)

- Es importante casar las características de la *cache* ...
  (accede a un bloque de una vez, normalmente de más de una palabra)
  - ✓ con las características de la DRAM:
    - utilizar DRAMs que soporten acceso rápido a un conjunto de palabras, preferiblemente cuando coincida con el tamaño de bloque de la *cache*
  - ✓ con las características del Bus de Memoria:
    - el *bus* puede soportar los ritmos y patrones de acceso de la DRAM si se quiere incrementar el ancho de banda hacia la *cache*

Ancho de Banda Memoria máximo ➔

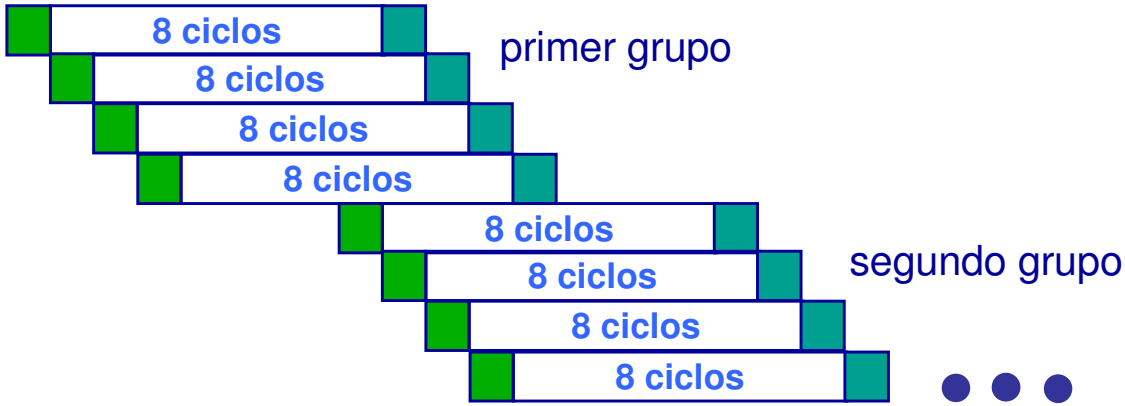
Ejercicio 4.1

Sin entrelazado:



Tiempo sin entrelazado =

Con entrelazado ➔ acceso secuencial a 4 grupos de 4 palabras, cada uno de ellos en paralelo:



Tiempo con entrelazado =

G =