



UNIVERSIDAD DEL BÍO-BÍO
FACULTAD DE CIENCIAS EMPRESARIALES

Memory Hierarchy

Heterogeneous Computing

Professor: Dr. Joel Fuentes - jfuentes@ubiobio.cl

Assistants:

- Daniel López - daniel.lopez1701@alumnos.ubiobio.cl
- Sebastián González - sebastian.gonzalez1801@alumnos.ubiobio.cl

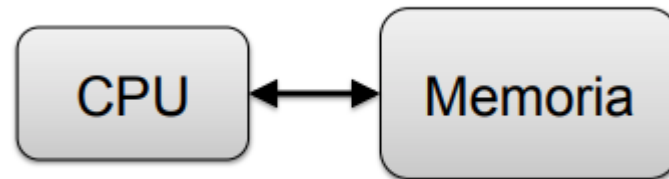
Course Webpage: <http://www.face.ubiobio.cl/~jfuentes/classes/ch>

Contents

- Basic concepts
- Memory hierarchy
- Caching

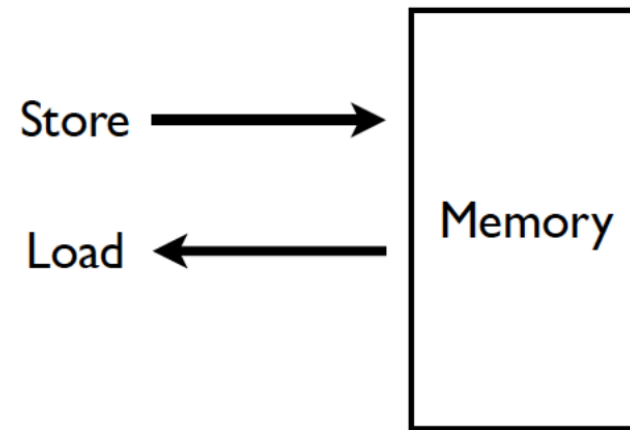
Basic concepts

- Limitaciones de eficiencia entre CPU y memoria suelen ser latencia y ancho de banda
- Efficiency limitations between CPU and memory are usually latency and bandwidth
- **Latency:** The time required for a single access
 - Memory access time \gg Processor's cycle time
- **Bandwidth:** Number accesses per unit of time

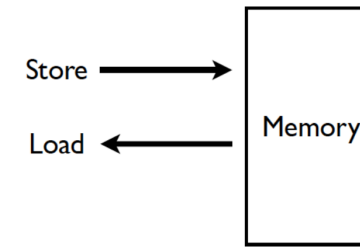


Basic concepts

- Programmer's view:

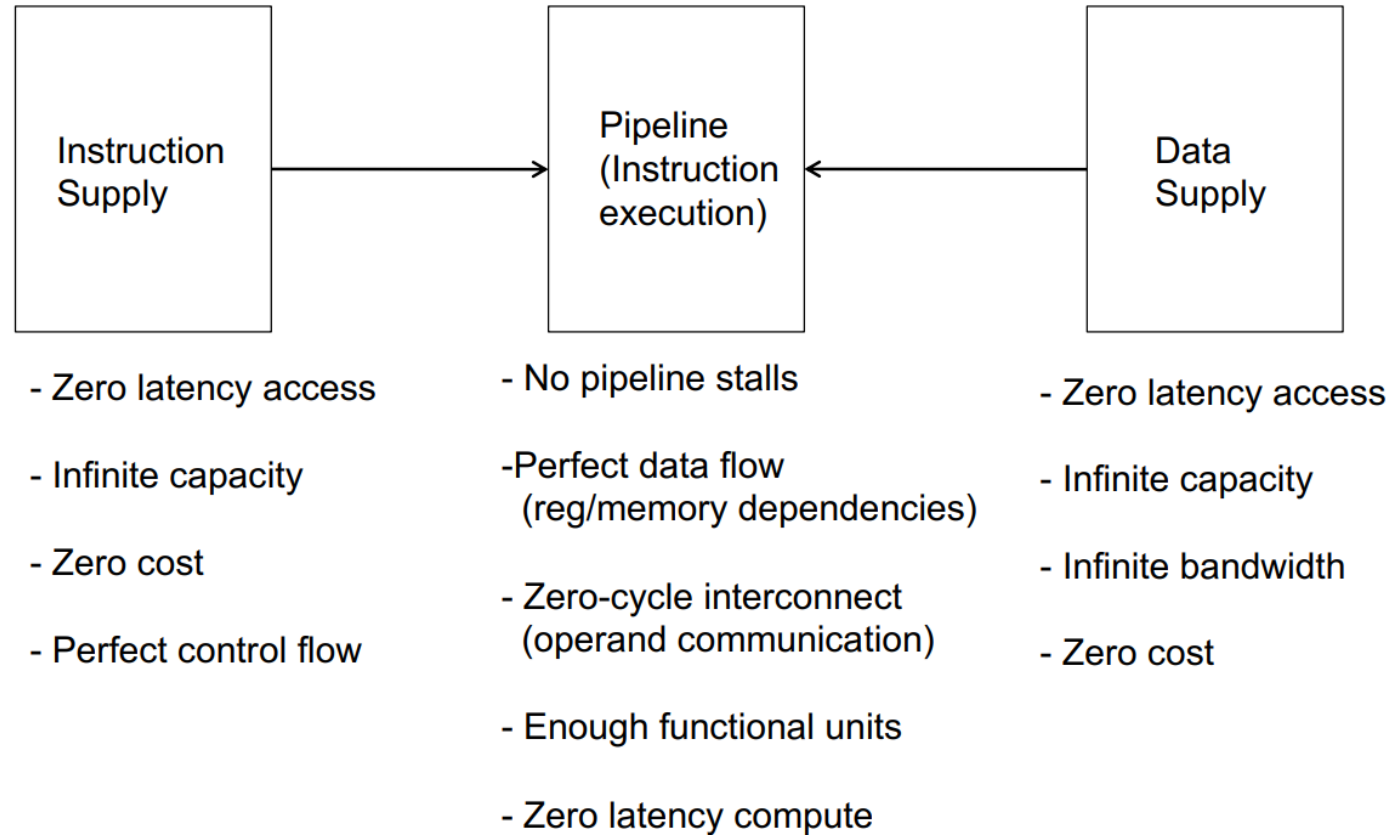


Virtual memory vs Physical memory

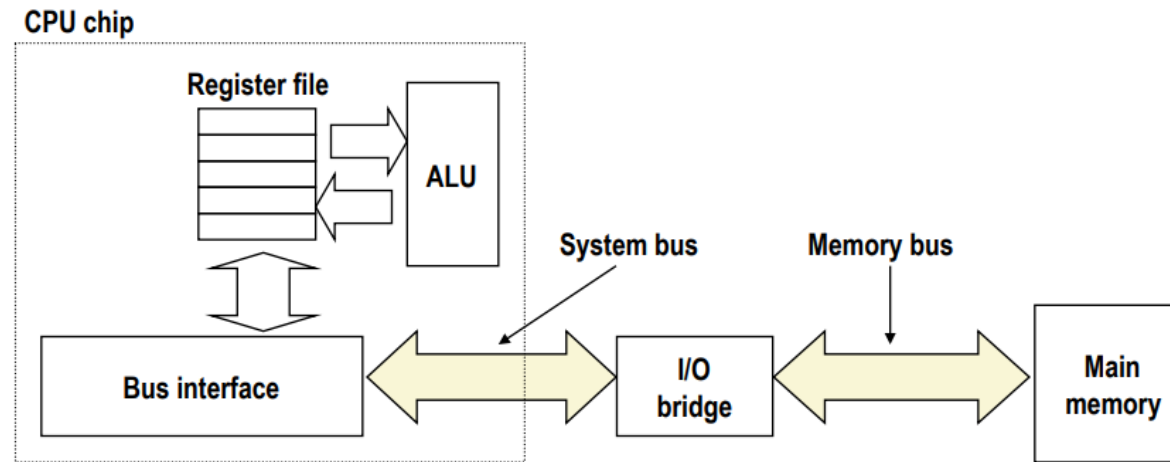


- A programmer sees **virtual memory**
 - And assumes an infinite amount of memory
- Reality: The size of the **physical memory** is much smaller than what the programmer assumes
- The system (software and hardware) maps virtual memory addresses to physical memory
 - This mapping is completely transparent to the programmer

In an ideal world...



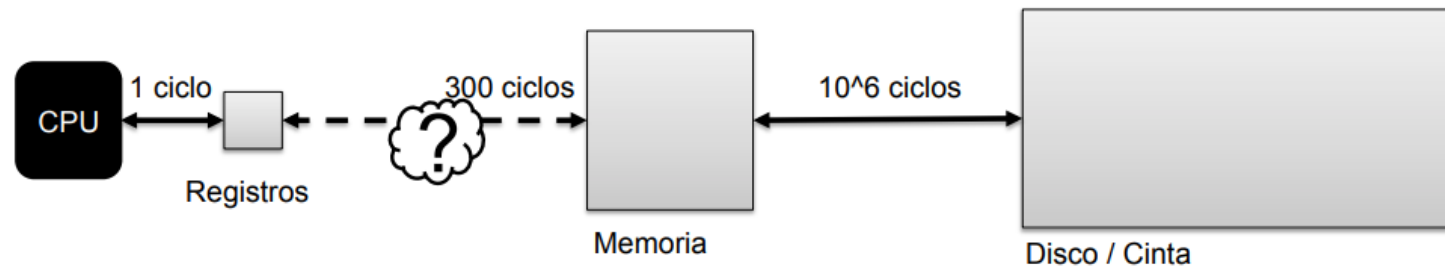
Memory hierarchy is born



- Idea: **Have multiple storage tiers**
- **Progressively larger and slower** as they get further away from the processor, **but** ensuring that the largest amount of data that the processor needs is at the fastest tiers.

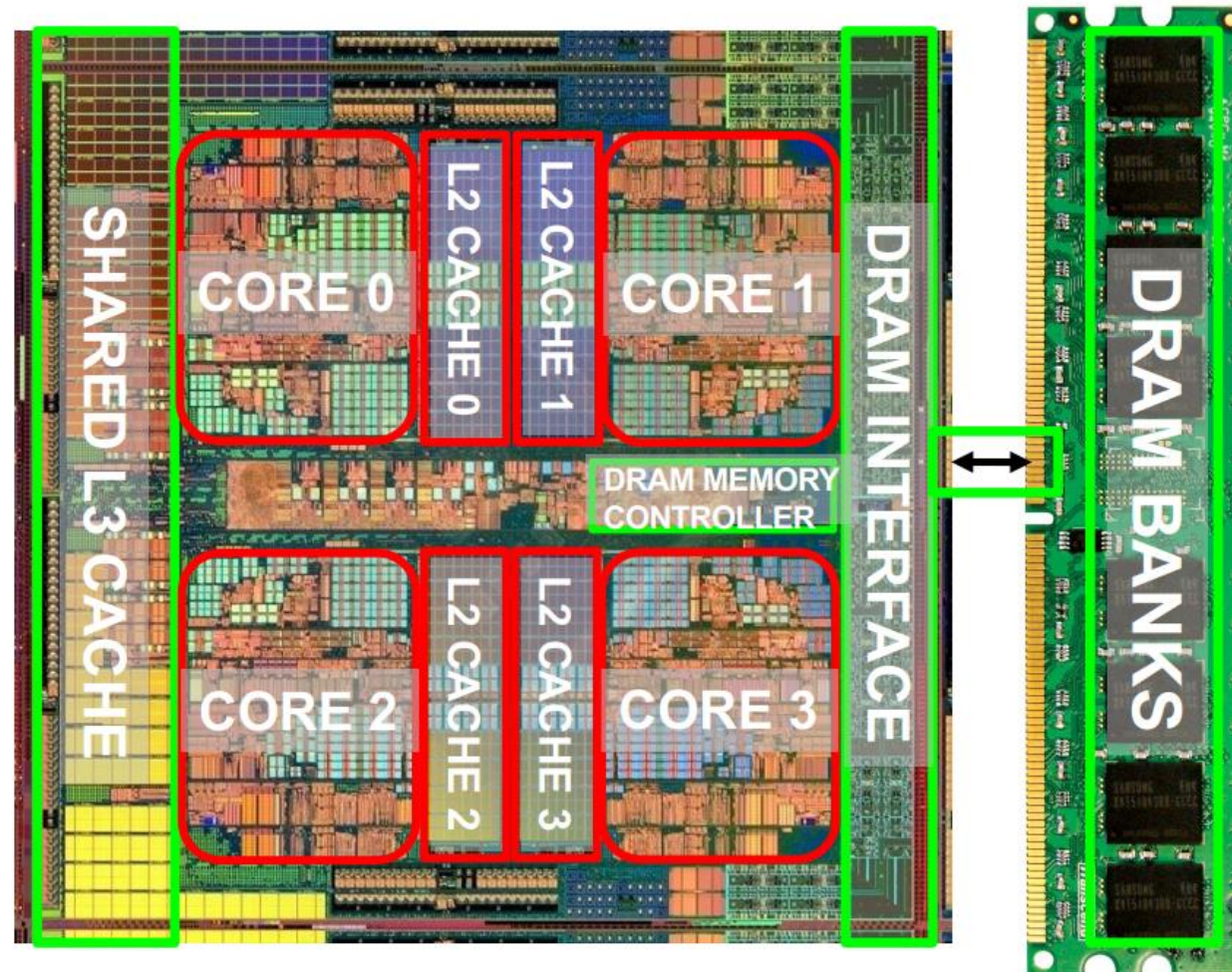
Memory hierarchy is born

- There is a long way (number of cycles) to read/write data between different memories.

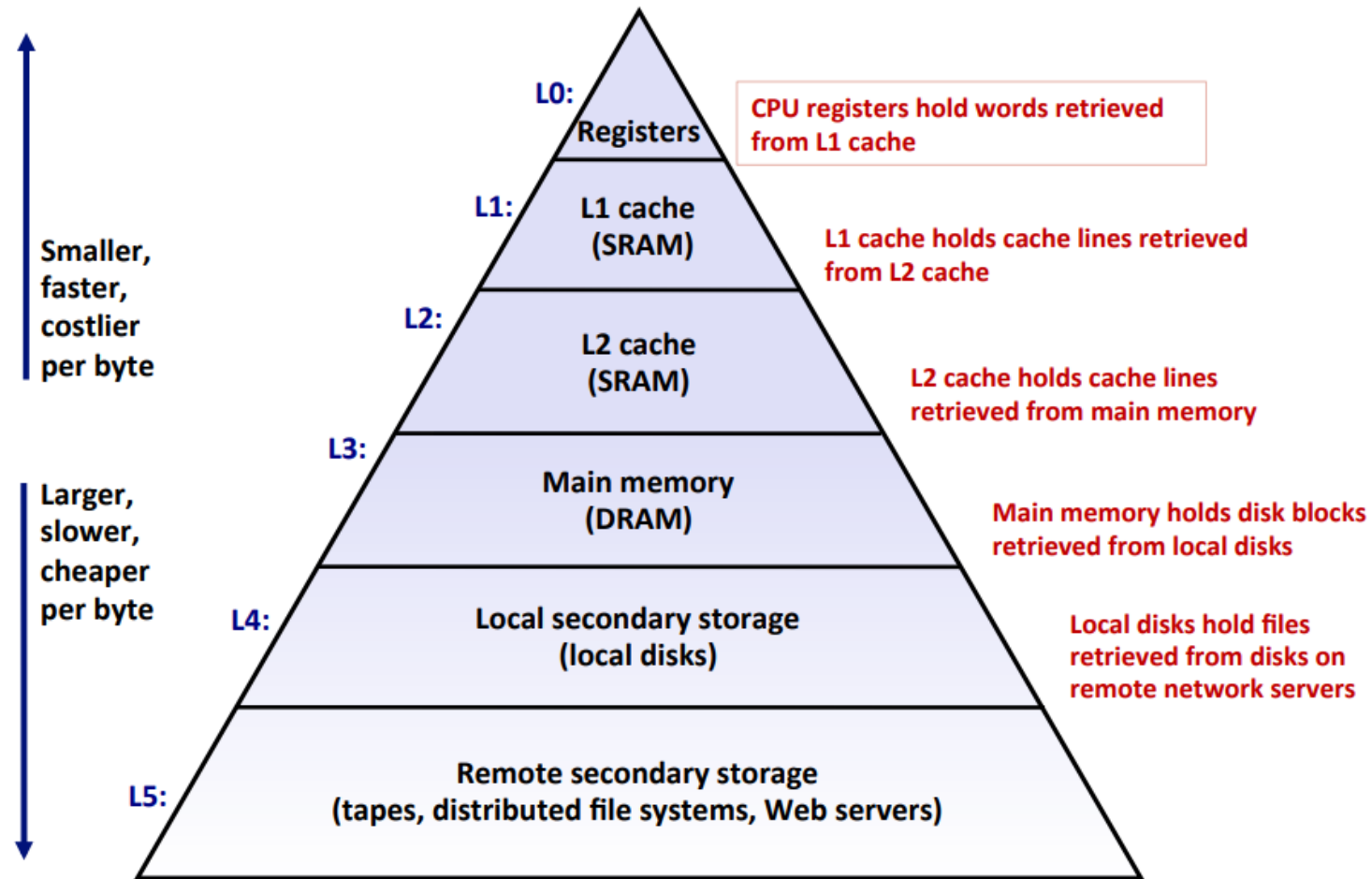


- What would improve this process?

Memory hierarchy



Memory hierarchy

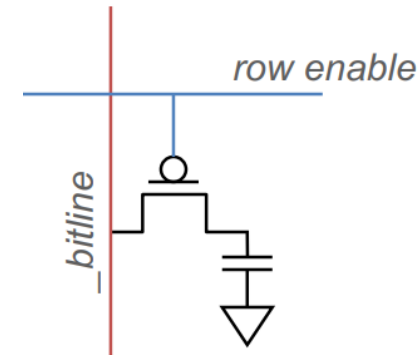


Main problems

- The larger the memory, the slower it is
 - As it grows larger, it is more complex to determine the location (physical memory address)
- Faster memories are more expensive
 - Technologies: SRAM vs DRAM vs Disk
- Higher bandwidth is more expensive
 - Need more memory banks, ports, higher frequency, faster technology, etc

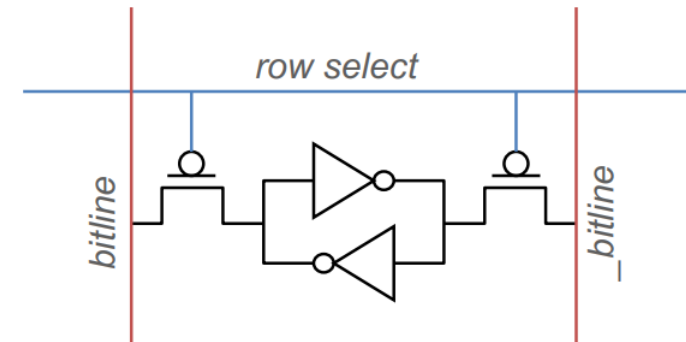
DRAM

- Dynamic Random Access Memory
- The capacitor's state of charge indicates the stored value
 - Whether the capacitor is charged or discharged indicates a stored value of 1 or 0.
 - 1 capacitor
 - 1 access transistor
- Capacitor leaks through the RC path
 - DRAM cells lose charge over time
 - DRAM cells need to be refreshed



SRAM

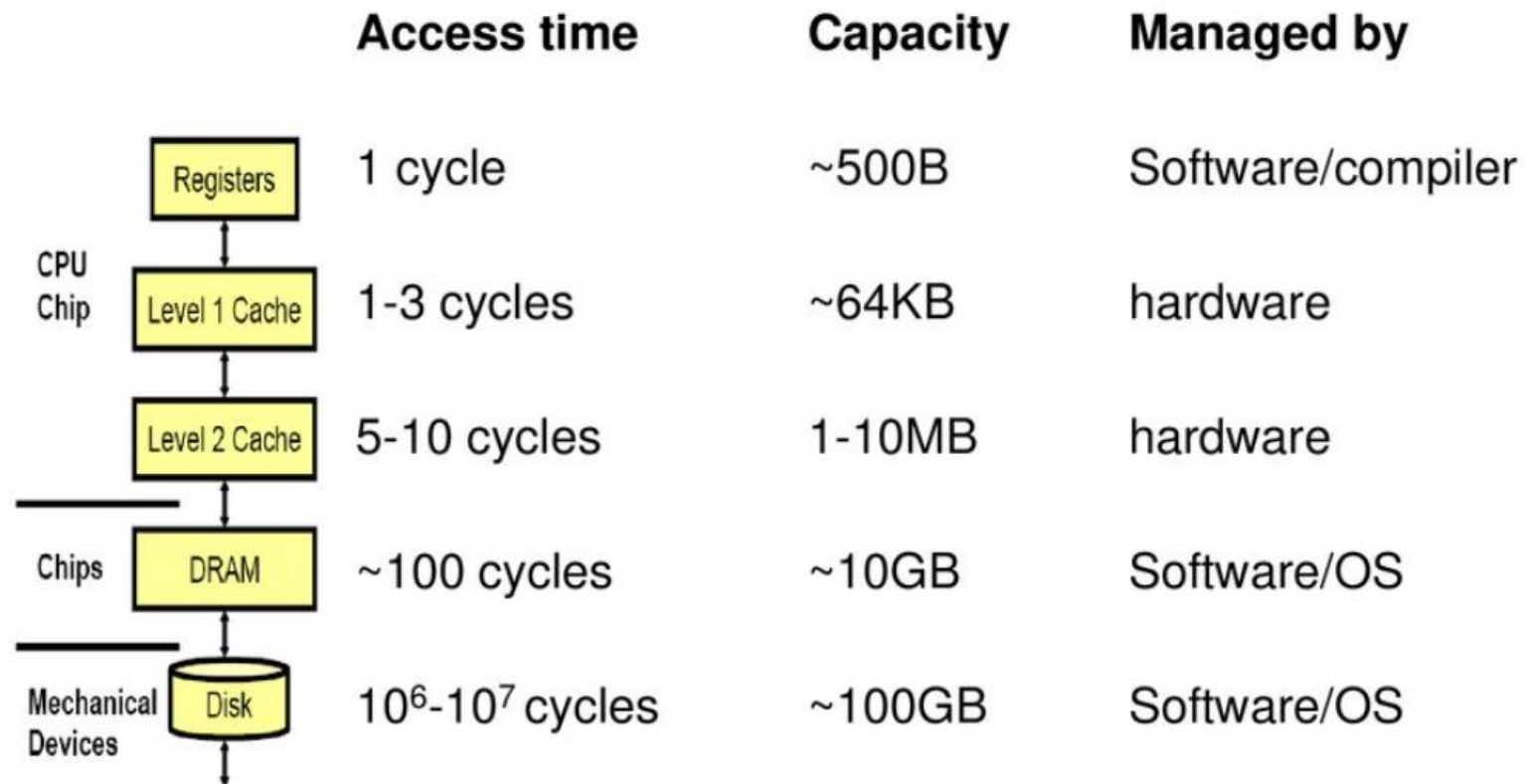
- Static Random Access Memory
- Typically known as Cache Memory
- Two cross-coupled inverters store a single bit
 - Feedback path enables the stored value to persist in the cell
 - 4 transistors for storage
 - 2 transistors for access



DRAM vs SRAM

- DRAM
 - Slower access (capacitor)
 - Higher density (1 Transistor 1 capacitor per cell)
 - Lower cost
 - Requires memory refreshing (energy, performance, circuitry)
 - Manufacturing requires putting capacitor and logic together
- SRAM
 - Faster access (no capacitor)
 - Lower density (6 transistors per cell)
 - Higher cost
 - Does not require memory refreshing
 - Manufacturing compatible with logic process (no capacitor)

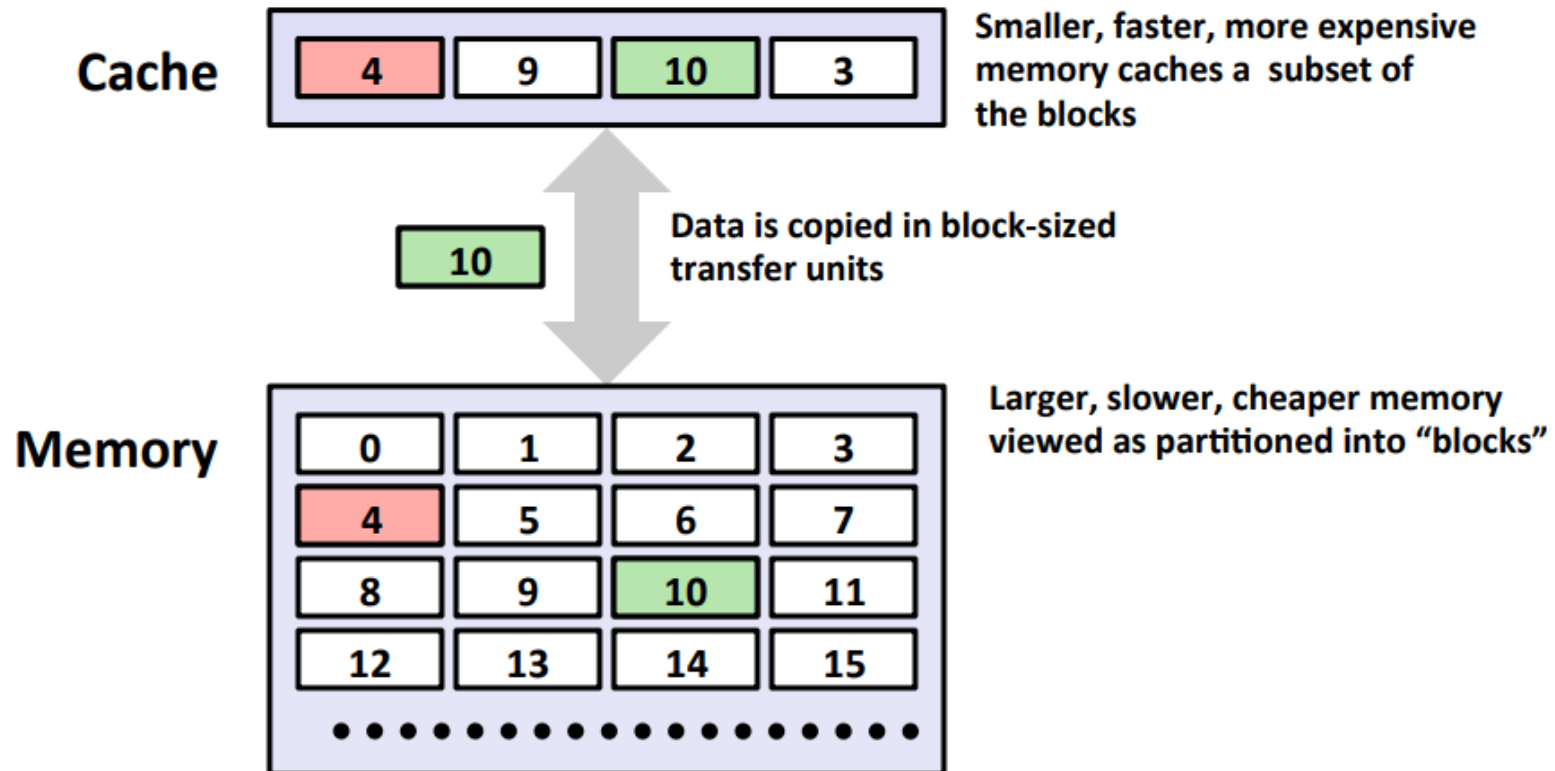
Memory hierarchy



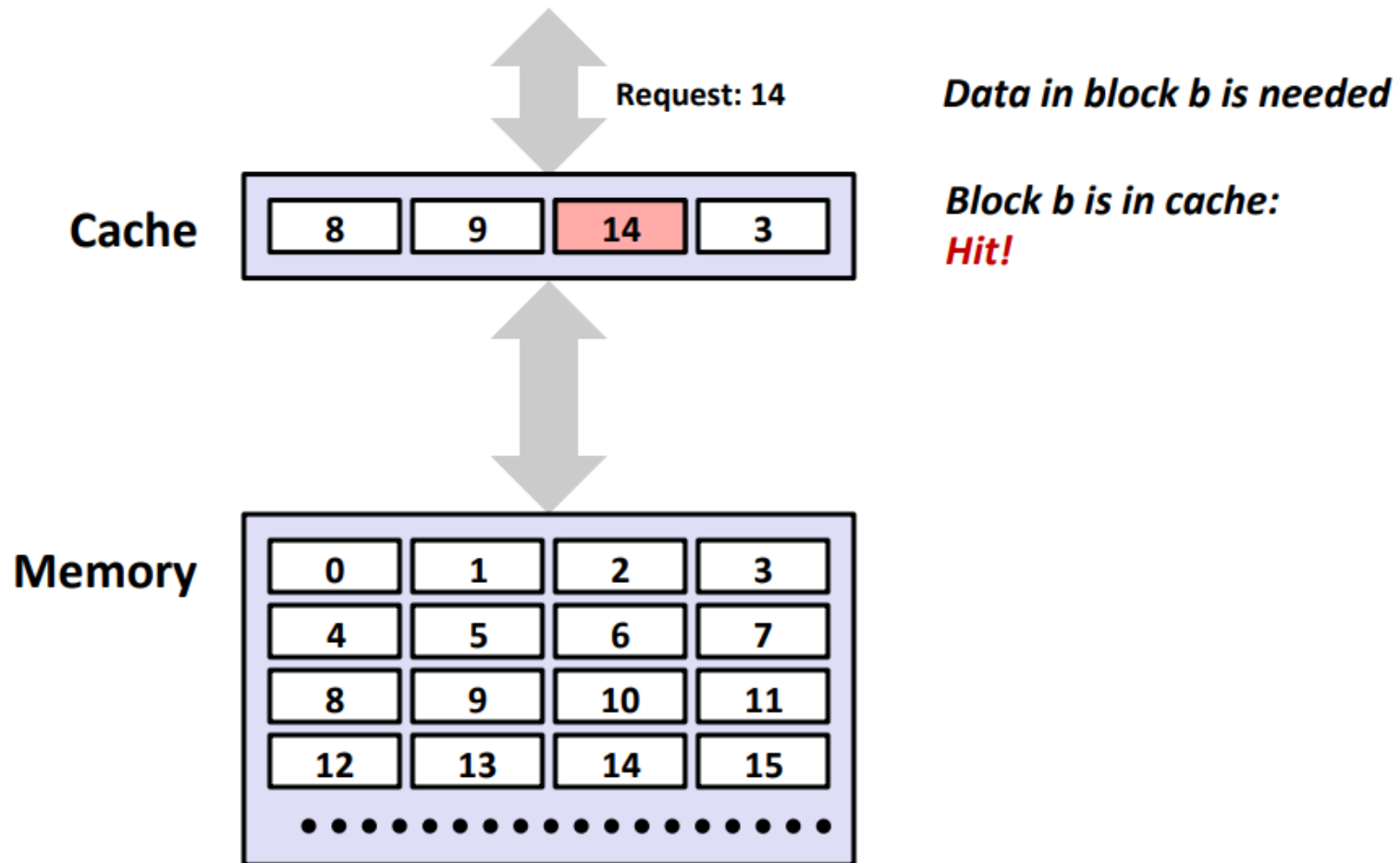
Conceptos sobre Caching

- **Block line:** Unit of storage in the cache
 - Memory is logically divided into cache blocks that map to locations in the cache
- **Hit:** If the data is in cache, use cached data instead of accessing memory
- **Miss:** If the data is not in cache, bring block into cache
- **Placement:** Where and how to find a block in cache
- **Replacement:** What data to remove to make room in cache

Caching



Caching: Hit



Caching: Miss

