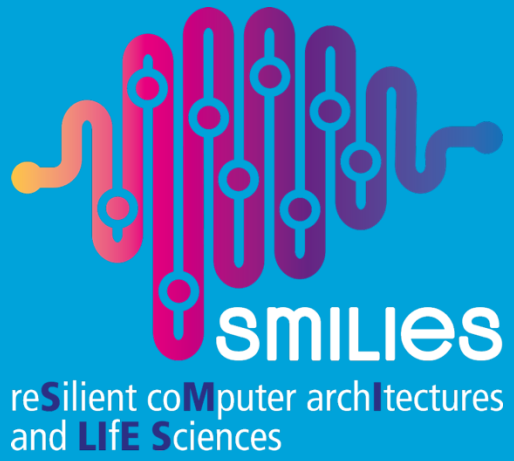


Politecnico di Torino

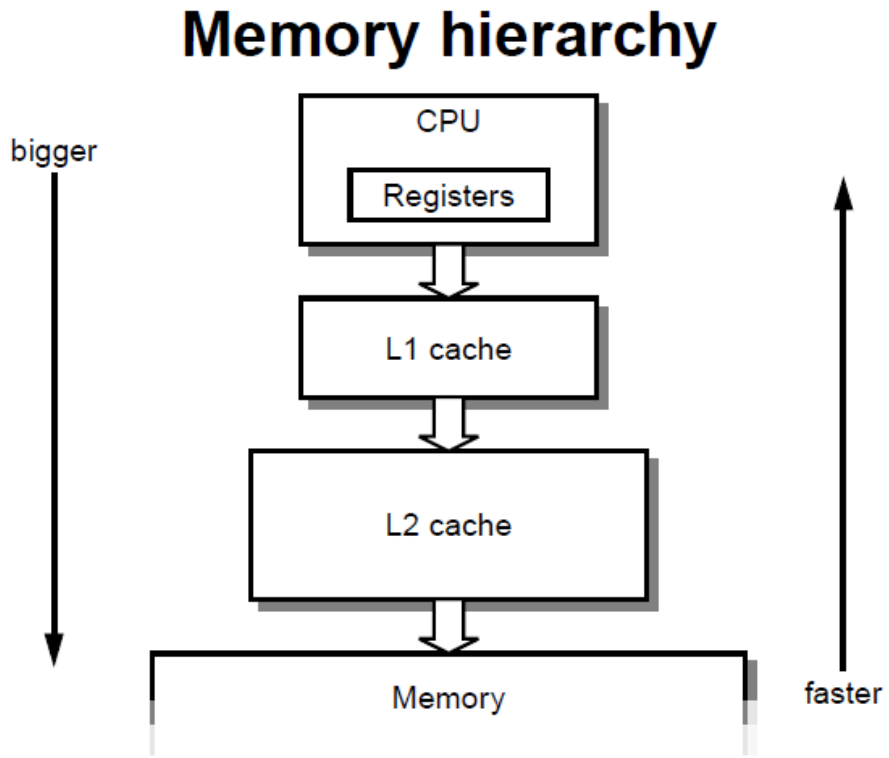
Department of Control and Computer Engineering

CACHE MEMORY

STEFANO DI CARLO

MEMORY HIERARCHY

- A typical system has several different memory subsystems.



CACHES

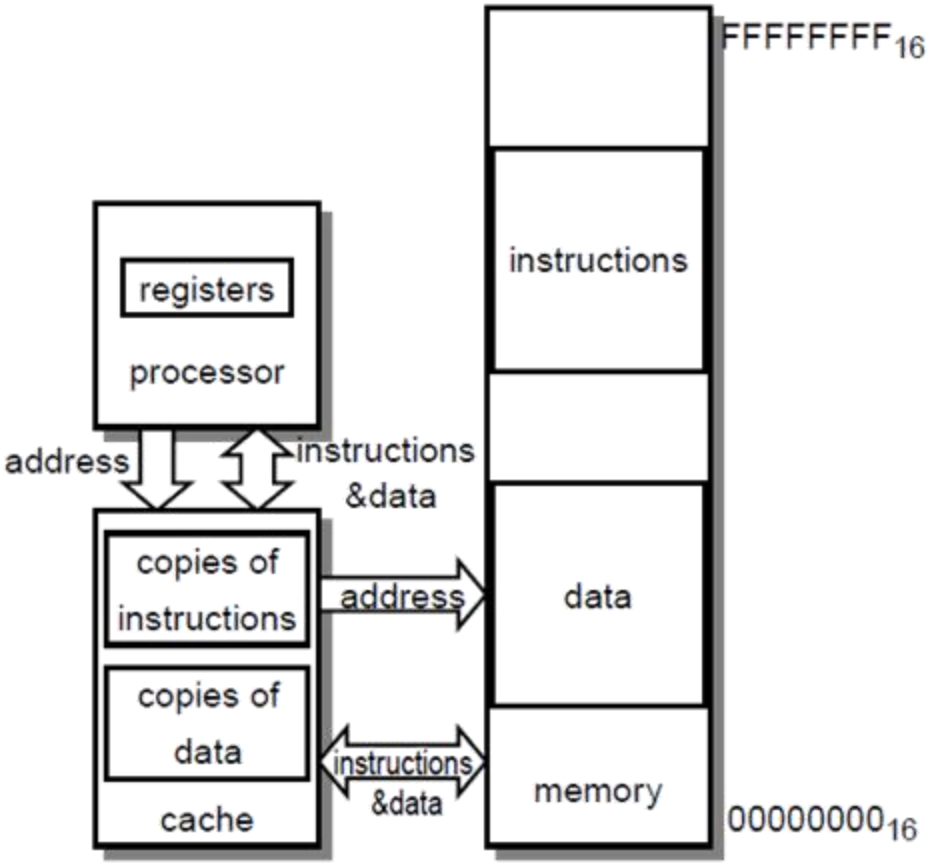
- A cache is a small on-chip memory which automatically:
  - keeps copies of recently used memory values
  - supplies these to the processor when it asks for them again thereby avoiding an off-chip memory access
  - Decides which values to over-write when it is full

CACHES - CLASSIFICATION

- Based up on the storage of Instruction and data, caches can be classified into
  - Unified Cache
  - Modified Harvard

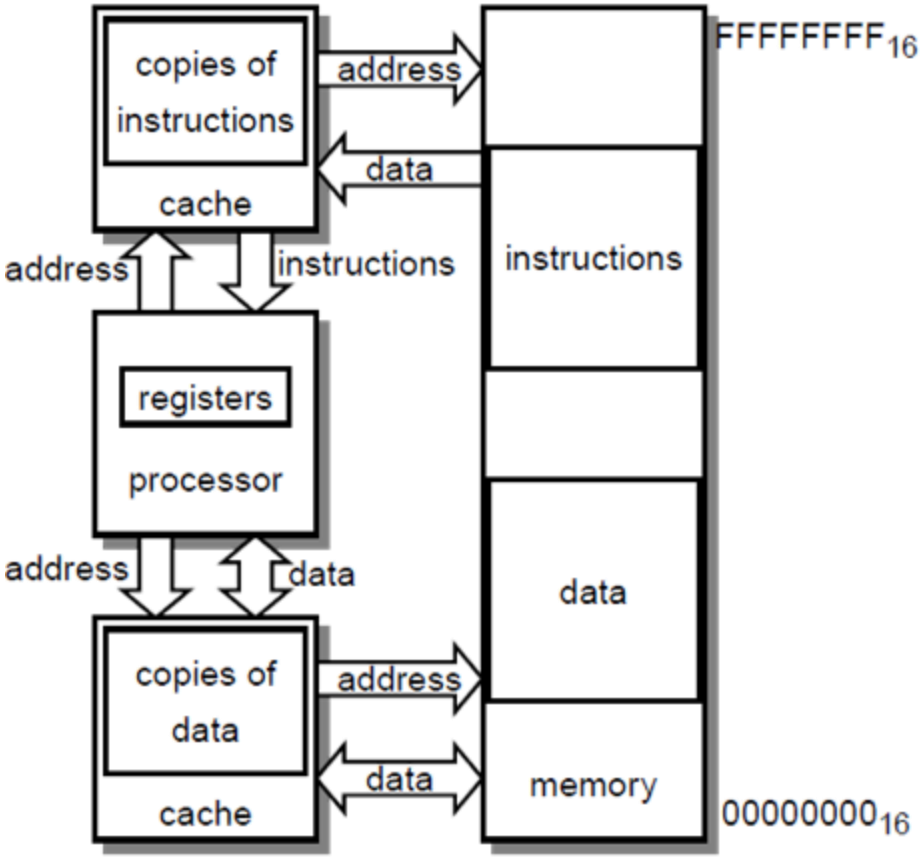
UNIFIED CACHE

- This is a single cache for both instructions and data



MODIFIED HARVARD

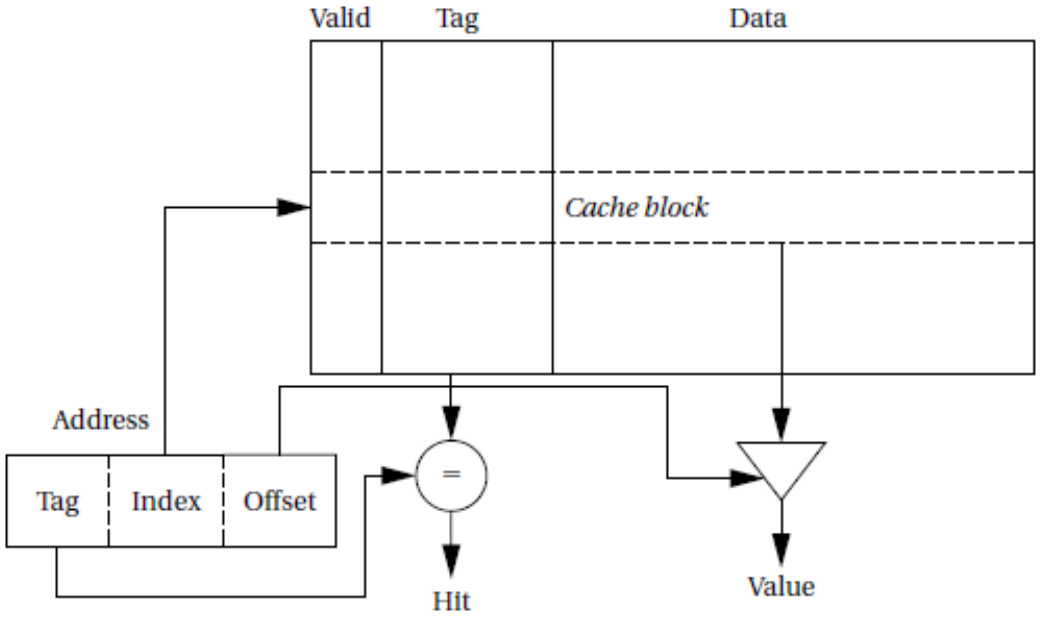
- Separate cache for both instructions and data



MEMORY HIERARCHY

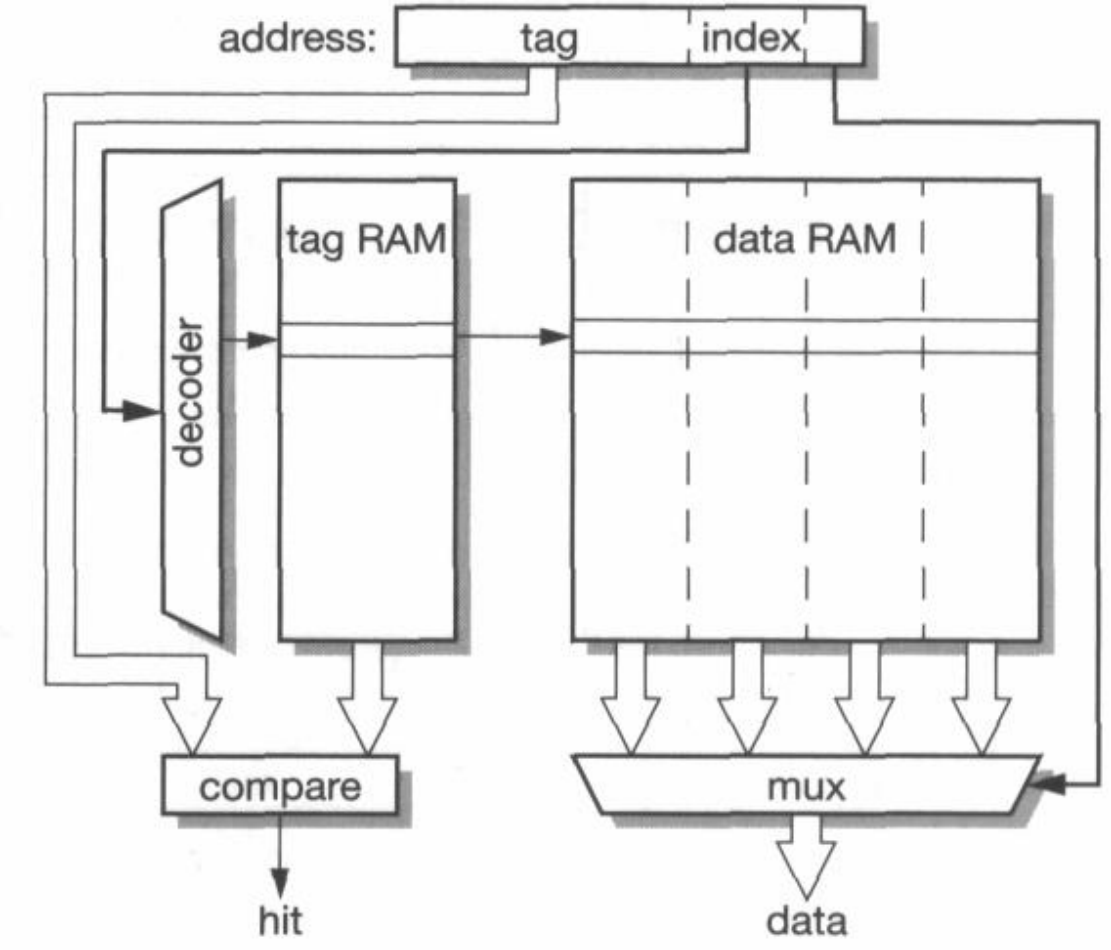
- An access to an item which is in the cache is called a hit
- An access to an item which is not in the cache is a miss.
- The proportion of all the memory accesses that are satisfied by the cache is the hit rate, usually expressed as a percentage, and the proportion that are not is the miss rate.

DIRECT MAPPED CACHE

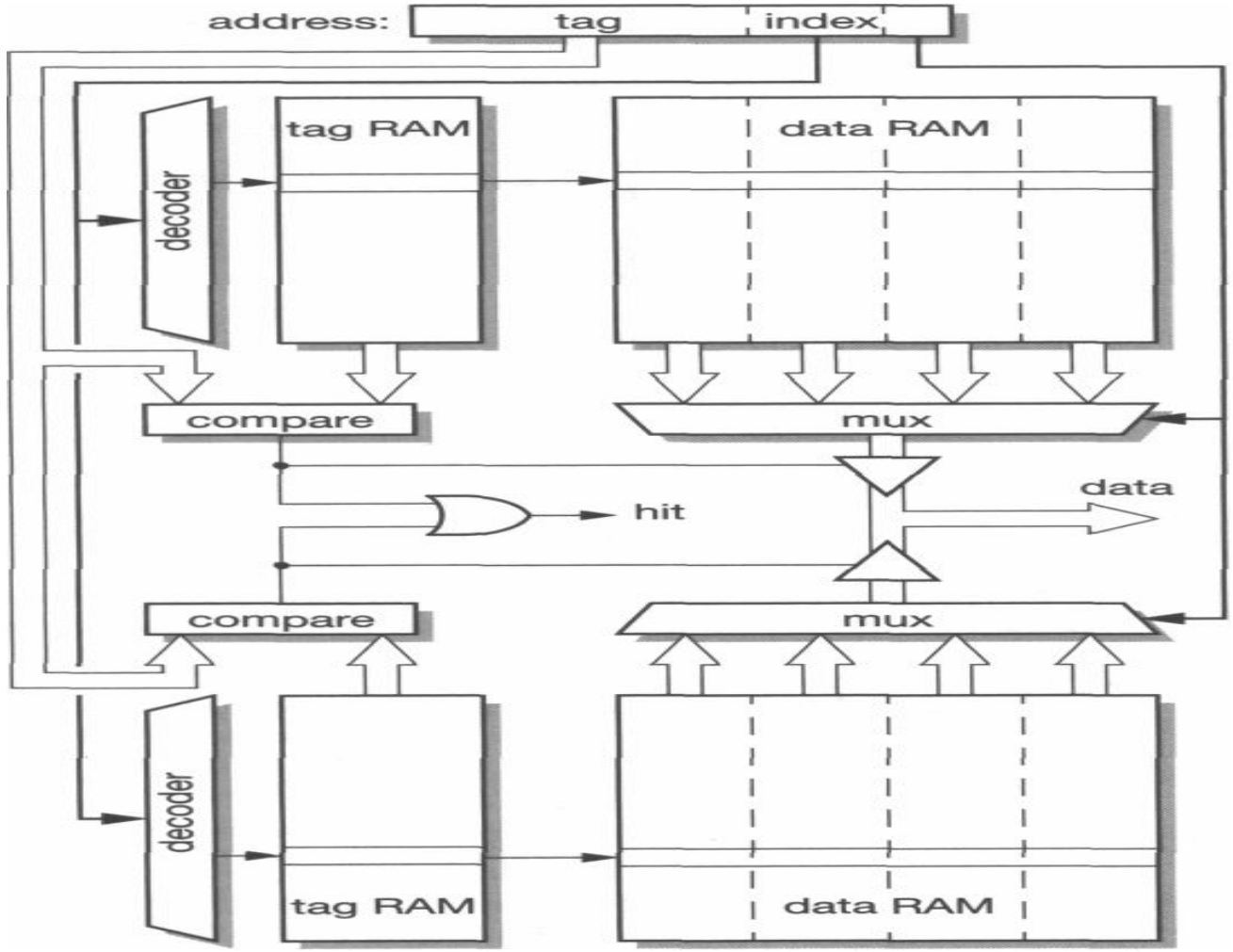


- Writing is more complicated than reading
  - Write-through
  - Write-back

DIRECT MAPPED CACHE



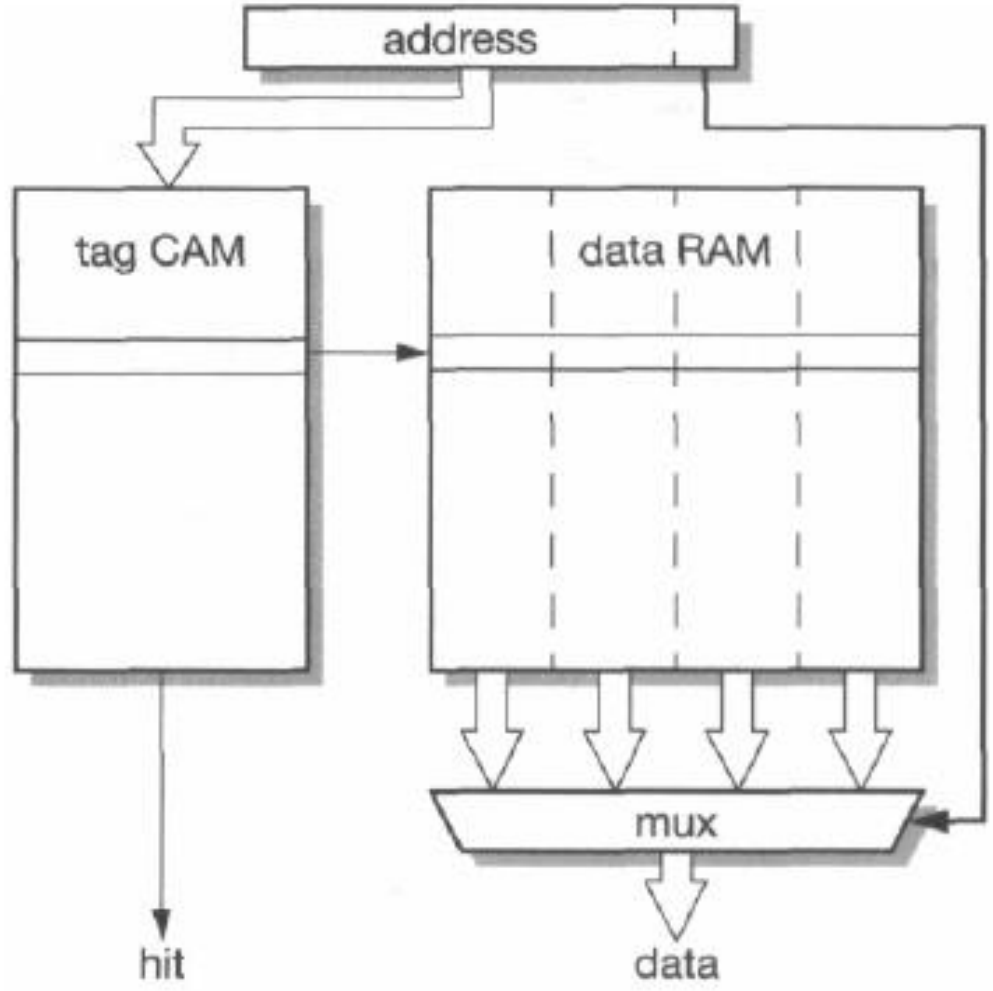
TWO WAY SET ASSOCIATIVE CACHE



TWO WAY SET ASSOCIATIVE CACHE

- two (smaller) cache blocks
- two chances to store any line
- better hit rate
- more expensive
- can extend to 4-way, etc.

FULLY ASSOCIATIVE CACHE



FULLY ASSOCIATIVE CACHE

- More places to store given line
- Even better hit rate
- Even more expensive
- (Potentially) slower
- Requires CAM (Content Addressable Memory)

COMPARISON

Direct mapped	Set associative	Fully associative
If each block has only one place that it can appear in the cache, it is said to be direct mapped	If a block can be placed in a restricted set of places in the cache, the cache is said to be set associative	If a block can be placed anywhere in the cache, the cache is said to be fully associative.
simple, cheap, fast	compromise	slow, expensive
subject to ‘thrashing’	may be 2-, 4-, 8-, etc. way	best hit rate
choice for large caches	often preferred	choice for small caches

COMPARISON

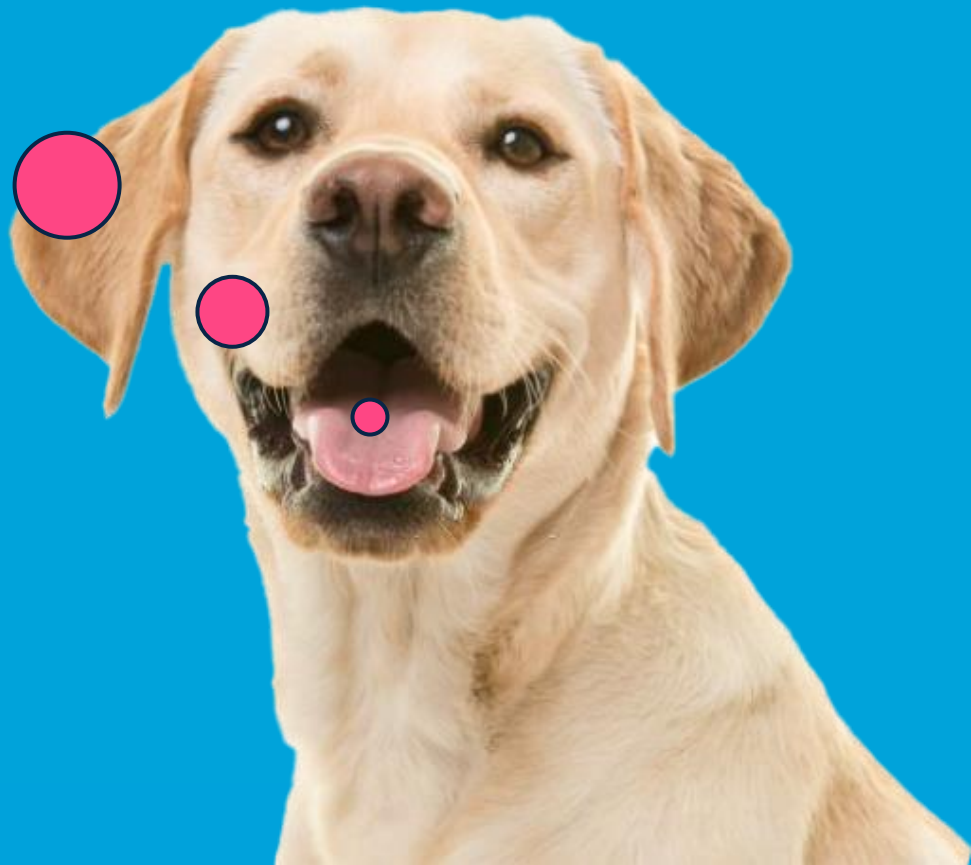
- Write-through
  - All write operations are passed to main memory; if the addressed location is currently held in the cache, the cache is updated to hold the new value. The processor must slow down to main memory speed while the write takes place.
- Write-through with buffered write
  - Here all write operations are still passed to main memory and the cache updated as appropriate, but instead of slowing the processor down to main memory speed the write address and data are stored in a write buffer which can accept the write information at high speed. The write buffer then transfers the data to main memory, at main memory speed, while the processor continues with its next task.
- Copy-back (also known as write-back)
  - A copy-back cache is not kept coherent with main memory. Write operations update only the cache, so cache lines must remember when they have been modified (usually using a dirty bit on each line or block). If a dirty cache line is allocated to new data it must be copied back to memory before the line is reused.

SMILIES
reSilient coMputer archItectures and LIfE Sciences

Politecnico di Torino
Department of Control and Computer Engineering

DAUIN

QUESTIONS?



THANK YOU!