# Introduction to Cache Memories

(Notes based on: *Computer Architecture: A Quantitative Approach, 5th Edition*, John L. Hennessy and David A. Patterson, Morgan Kaufmann, 2012)

The unit of information transfer between the cache and main memory is called a block. A block consists of one or more words.

Cache questions:

1. Where can a block be placed in the cache? (placement policy)
2. How is a block found if it is in the cache? (block identification)
3. Which block should be replaced on a miss? (replacement policy)
4. What happens on a write to the cache? (write strategy)

Answer to Question 1: direct mapped, fully associative, set associative

Direct mapped and fully associative can be considered as the two extreme cases of a set-associative cache:

- Direct mapped is essentially one-way set associative
- Fully associative cache with m blocks is essentially m-way set associative

Most caches in practical use are direct mapped, 2-way set associative or 4-way set associative.

Answer to Question 2: Each block frame in the cache has a tag that identifies the address of the block that is stored there.

All possible tags are searched in parallel to have good performance.

A valid bit along with each tag to indicate if the data in that block frame is valid or not.

Answer to Question 3: Random, LRU, FIFO

Random: May use pseudo-random (LFSR-based) to get reproducible and verifiable results.

LRU: Uses corollary of temporal locality – if a block has not been used for a long time, then it is unlikely to be needed in the near future

FIFO: Can be considered as an approximation to LRU that replaces the oldest block in the cache, even if it has been used recently. May be simpler to implement.

Answer to Question 4:

Most cache access are reads, because:

- all instruction access are reads
- most instructions don't write to memory

Two options:

- Write-through: New info written to both the cache and to main memory.
- Write-back: New info only written to the cache. Main memory is only updated when a modified cache block is replaced.

For write-back, can use a dirty bit to indicate if a cache block has been modified:

- "dirty" means it has been modified
- "clean" means that it has not been modified

So, if the block is clean, there is no need to update main memory when it is replaced.

Advantages of write-through:

- Easier to implement
- Cache and main memory are always coherent (i.e., hold the same information), which is more convenient for multiprocessors.

Advantages of write-back:

- Writes occur at the speed of the cache
- Multiple writes within the same block only require one write to main memory (i.e., at the time when the block is replaced)
- Uses less memory bandwidth, so attractive for multiprocessors.
- More power efficient

Split Caches: Separate caches for data and instructions.

- A processor can simultaneously access an instruction and a data word.
- Most recent processors use separate I and D caches.

Multilevel caches: Create more levels of the memory hierarchy (L1, L2, etc.).