15-213

"The course that gives CMU its Zip!"

Cache Memories
October 6, 2006

Topics

- Generic cache memory organization
- Direct mapped caches
- Set associative caches
- Impact of caches on performance
- The memory mountain

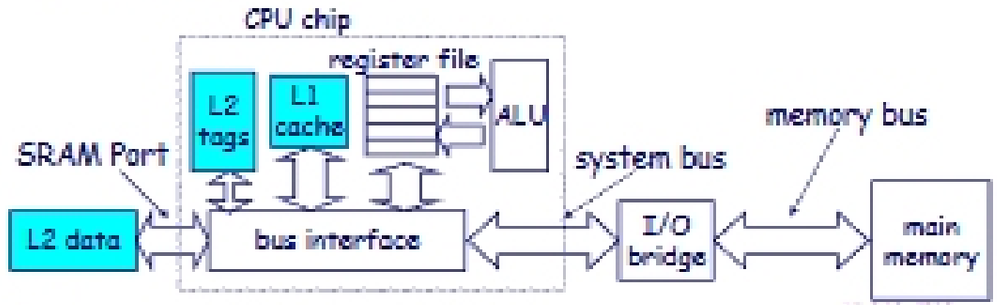
Cache Memories

Cache memories are small, fast SRAM-based memories managed automatically in hardware.

- Hold frequently accessed blocks of main memory

CPU looks first for data in L1, then in L2, then in main memory.

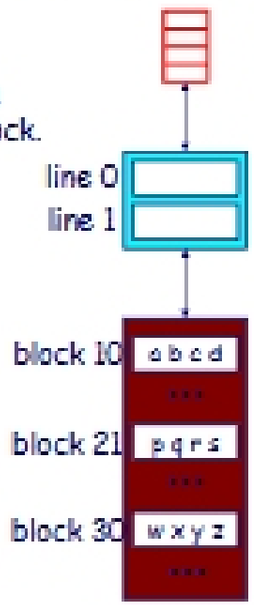
Typical system structure:



Inserting an L1 Cache Between the CPU and Main Memory

The transfer unit between the CPU register file and the cache is a 4-byte block.

The transfer unit between the cache and main memory is a 4-word block (16 bytes).



The tiny, very fast CPU register file has room for four 4-byte words.

The small fast L1 cache has room for two 4-word blocks.

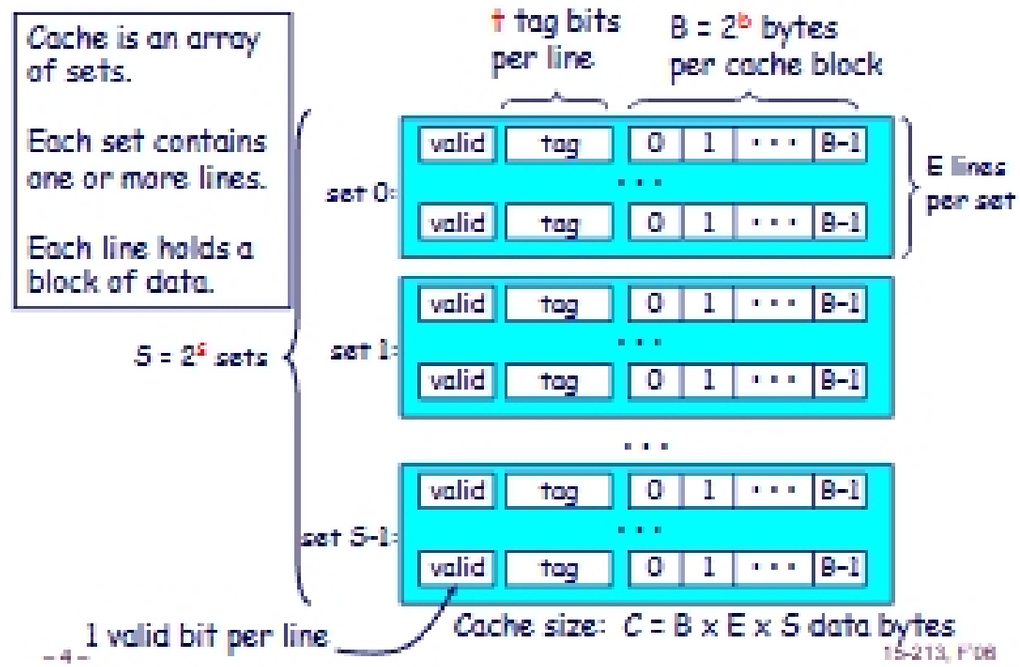
The big slow main memory has room for many 4-word blocks.

General Organization of a Cache

Cache is an array of sets.

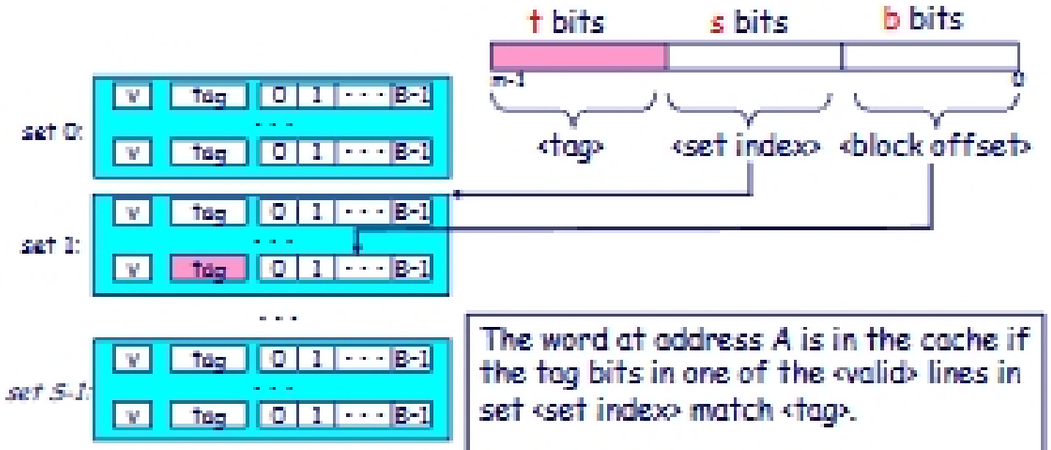
Each set contains one or more lines.

Each line holds a block of data.



Addressing Caches

Address A:

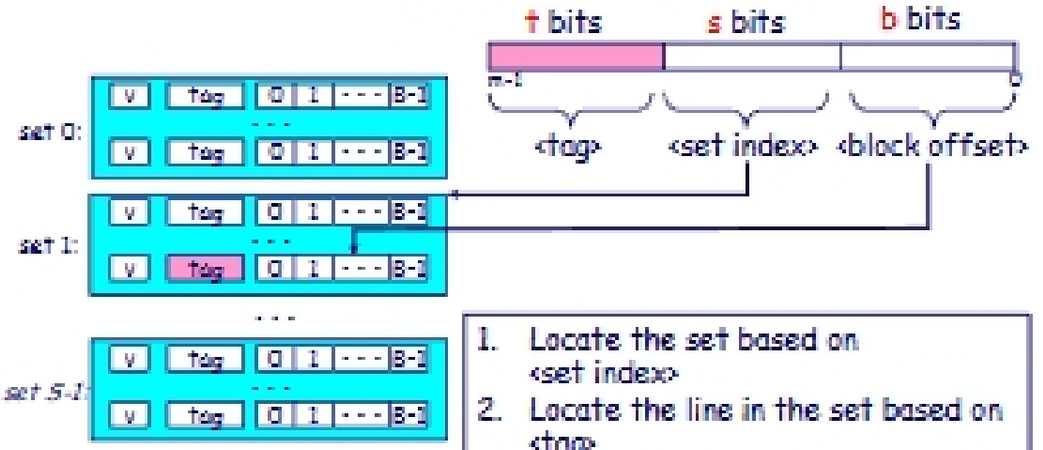


The word at address A is in the cache if the tag bits in one of the <valid> lines in set <set index> match <tag>.

The word contents begin at offset <block offset> bytes from the beginning of the block.

Addressing Caches

Address A:

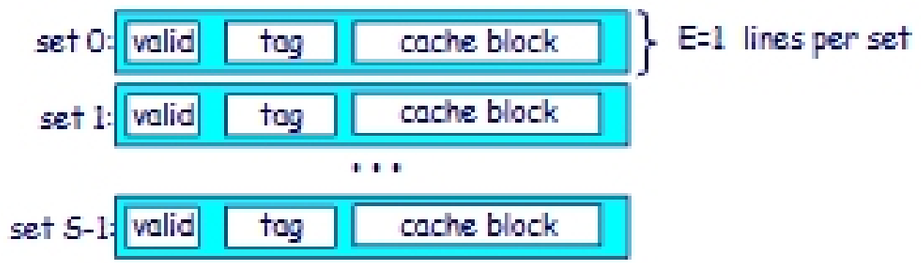


1. Locate the set based on <set index>
2. Locate the line in the set based on <tag>
3. Check that the line is valid
4. Locate the data in the line based on <block offset>

Direct-Mapped Cache

Simplest kind of cache, easy to build (only 1 tag compare required per access)

Characterized by exactly one line per set.

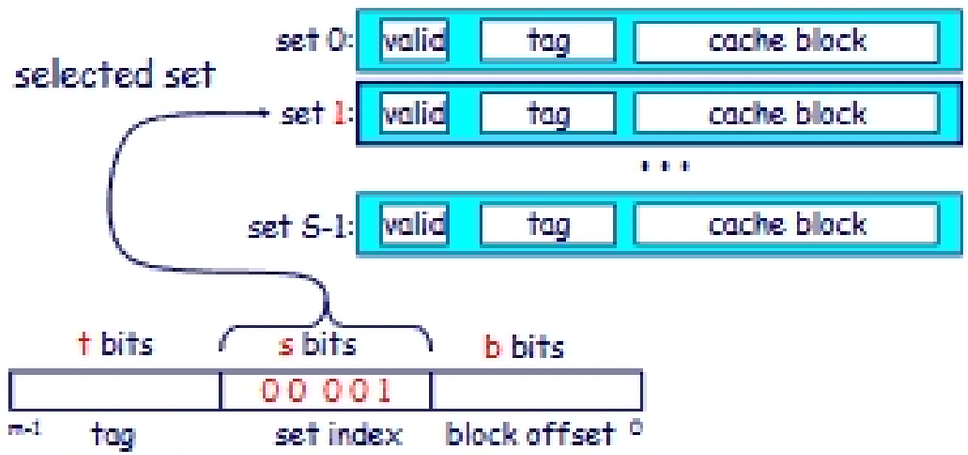


Cache size: C = B x S data bytes

Accessing Direct-Mapped Caches

Set selection

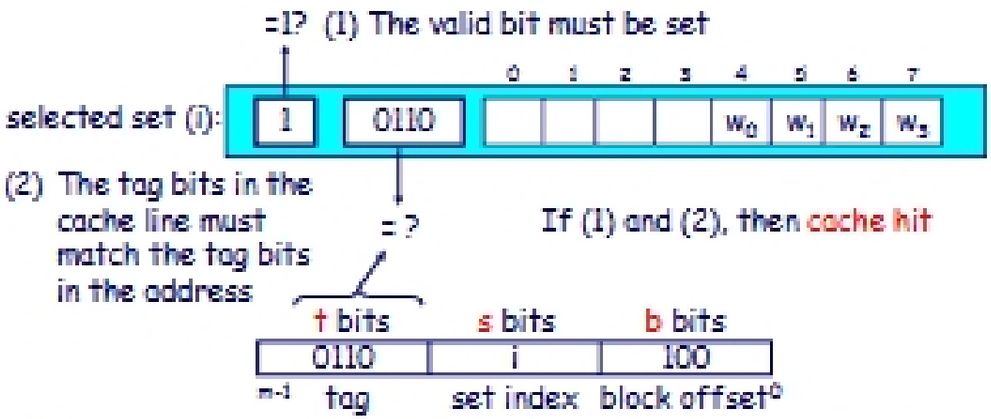
- Use the set index bits to determine the set of interest.



Accessing Direct-Mapped Caches

Line matching and word selection

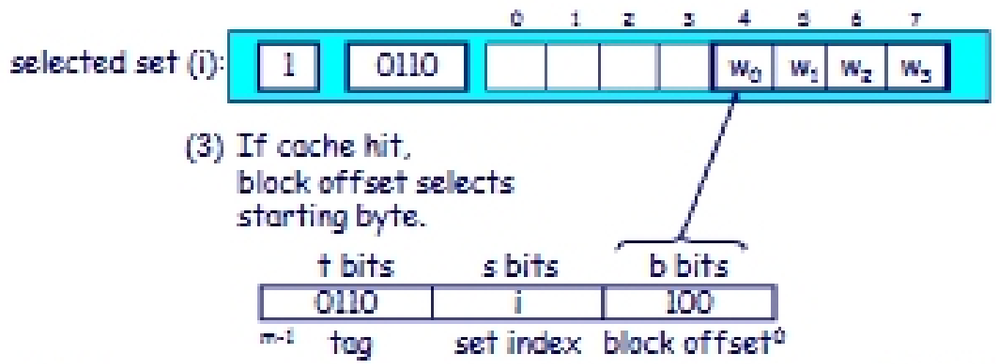
- Line matching: Find a valid line in the selected set with a matching tag
- Word selection: Then extract the word



Accessing Direct-Mapped Caches

Line matching and word selection

- Line matching: Find a valid line in the selected set with a matching tag
- Word selection: Then extract the word



Direct-Mapped Cache Simulation

t=1 s=2 b=1

x	xx	x

M=16 byte addresses, B=2 bytes/block,
S=4 sets, E=1 entry/set

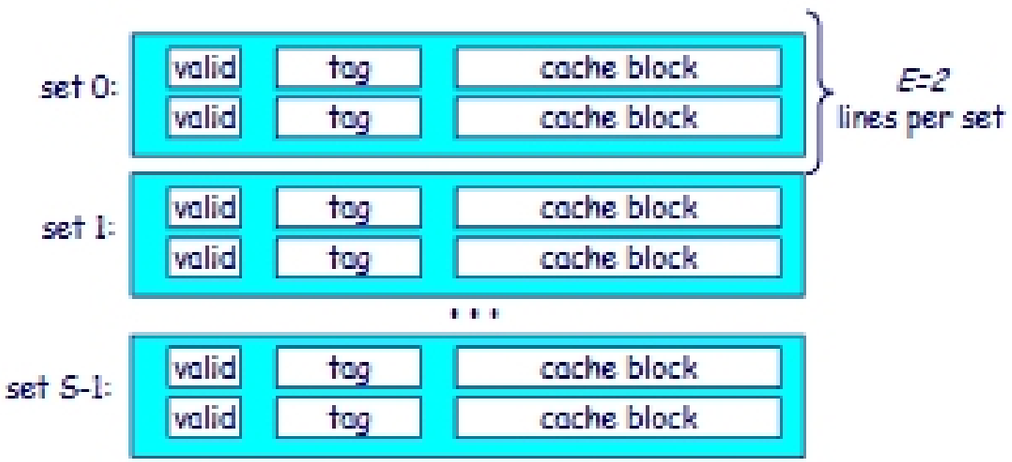
Address trace (reads):

0	$[0000_2]$,	miss
1	$[0001_2]$,	hit
7	$[0111_2]$,	miss
8	$[1000_2]$,	miss
0	$[0000_2]$	miss

v	tag	data
1	0	M[0-1]
1	0	M[6-7]

Set Associative Caches

Characterized by more than one line per set



E-way associative cache