Caching

Memory is a commodity that every computer user would like more of. Moreover almost every program would run faster if the memory it used could be accessed faster. Unfortunately, memory has the property that the faster it can be accessed, the more expensive it is. Caching is a general technique for exploiting the locality of reference of data to make it appear as if there is a large amount of fast memory.

Memory Technology	Typical Access Time	$/MB (93)
SRAM	8 - 35ns	$100 - $400
DRAM	90 - 120ns	$25 - $50
Magnetic Disk	10^7 - 2x10^7ns	$1 - $2

Terminology

Locality -- the property of memory references to cluster.

Temporal locality -- the tendency of the time intervals between references to a given address to be small.

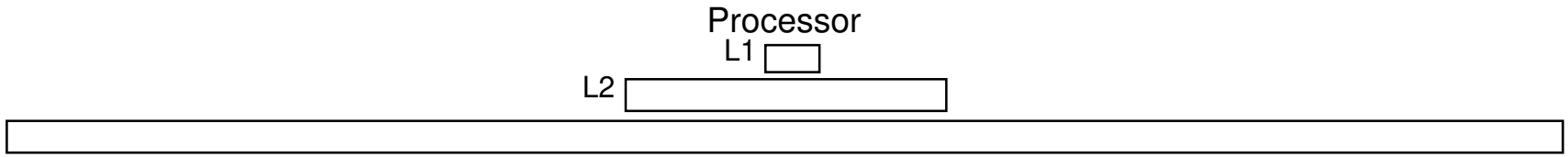
Spacial locality -- the tendency of the distances between consecutive memory references to be small.

Memory hierarchy -- a characteristic of computer design in which a series of storage technologies are used such that the access time is faster as the memory is closer to the processor and the capacity is larger as the memory is further from the processor.

```
        . . .
add     $4,$5,$6
lcw1    $f2,0($4)
lcw1    $f3,4($4)
sw      $7,0($29)
sw      $8,4($29)
bne     $9,$0,loop
        . . .
```

Terminology (Continued)

Cache -- memory closest to the processor in a memory hierarchy.
Caching -- any storage management technique exploiting locality.
Upper/lower level -- memory closer/further from the processor
Block -- unit of memory transfer between two levels in a memory hierarchy
Hit/Miss -- accessing data present/not present in a hierarchy level.
Hit rate -- ratio of hits to total references. Miss rate = 1 - Hit rate.
Hit time -- time to hit in the cache.
Miss penalty -- time to move a block from a lower level in the hierarchy and satisfy the processor's request.

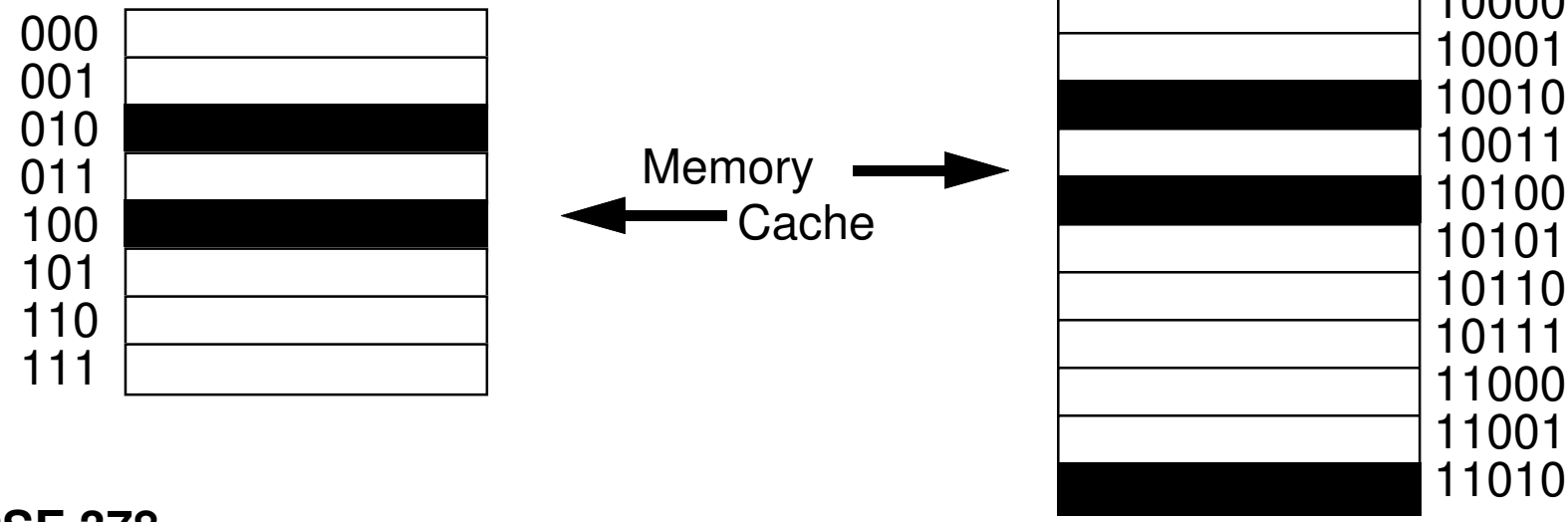


A Direct-Mapped Cache

Questions asked of a caching technique:

- Where is a block stored?
- How is a block found?
- What block is stored at a location?

A direct mapped cache of size 2^k uses the k lsb's of the (block) address.

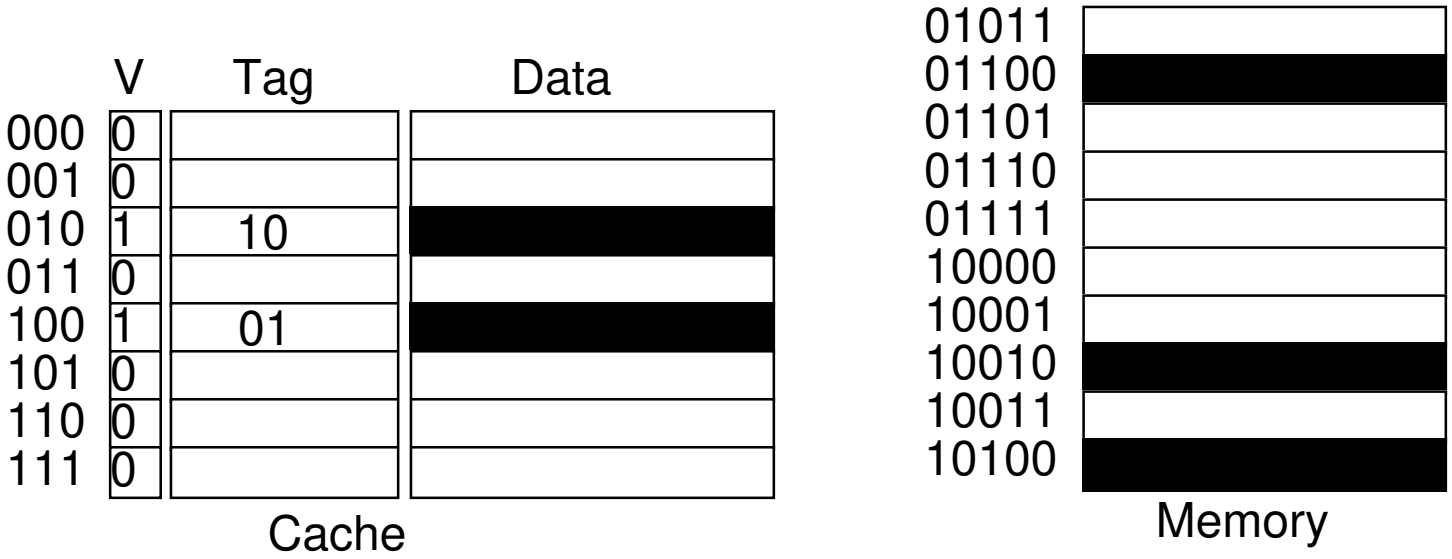


Direct-Mapped Cache Fields

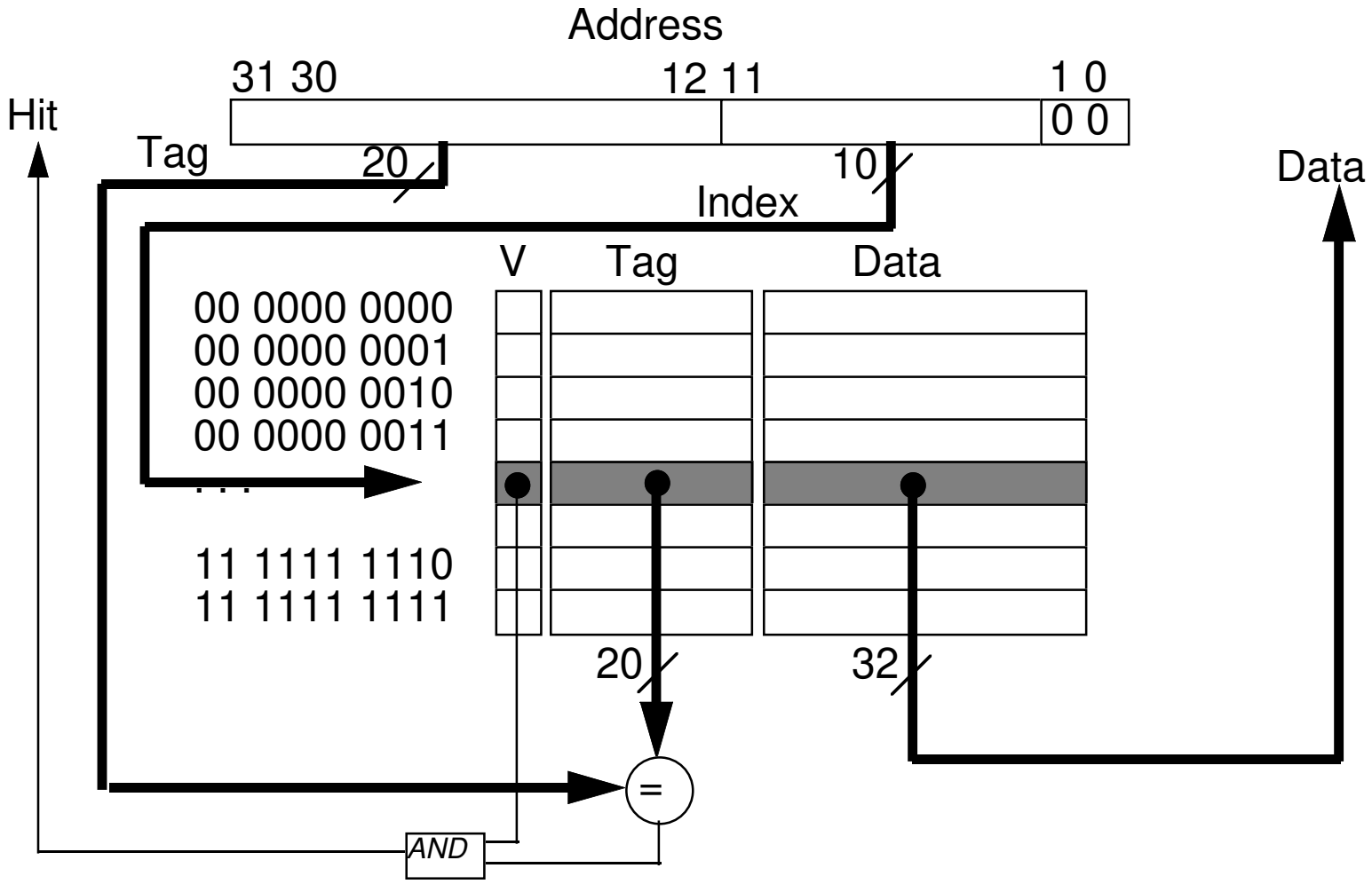
The Tag field stores the msbs of the address.

The Valid Bit indicates whether the data in the cache block is correct and available.

The Data field stores the contents of the block.



Direct-Mapped Cache Operation



Handling Misses

The processor has to stall the instruction that missed; instruction misses stall the pipeline at IF while data misses stall in MEM.

Operations by controller on a miss:

1. Compute PC-4
2. Access address in main memory and wait for completion.
3. Move data to cache, write tag bits, set Valid.
4. Restart execution pipeline at the fetch for instruction misses, or MEM for data misses.