Caching

Caching is a general technique for exploiting the locality of data reference to make it appear as if there is a large amount of fast memory.

Memory Technology	Typical Access Time	$/MB ('97)
SRAM	5 - 25ns	$100 - $250
DRAM	60 - 120ns	$5 - $10
Magnetic Disk	10^7 - 2x10^7ns	$0.10 - $0.20

© Larry Snyder, 2000 All rights reserved

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

Spatial 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
```

© Larry Snyder, 2000 All rights reserved

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

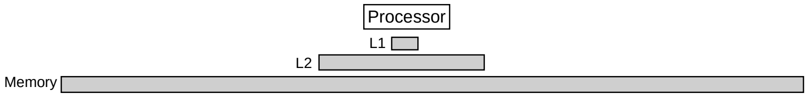
Block -- unit of memory transfer between two levels in a memory hierarchy. Also called a *cache line*

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



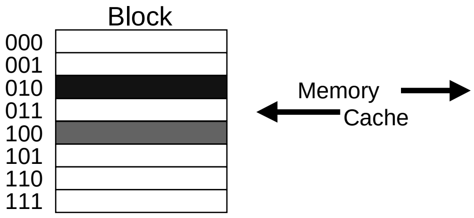
© Larry Snyder, 2000 All rights reserved

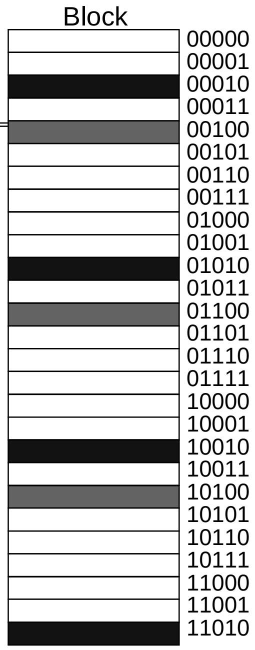
Direct-mapped Cache

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.

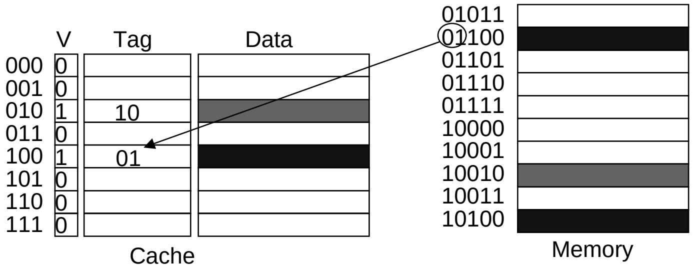




© Larry Snyder, 2000 All rights reserved

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.

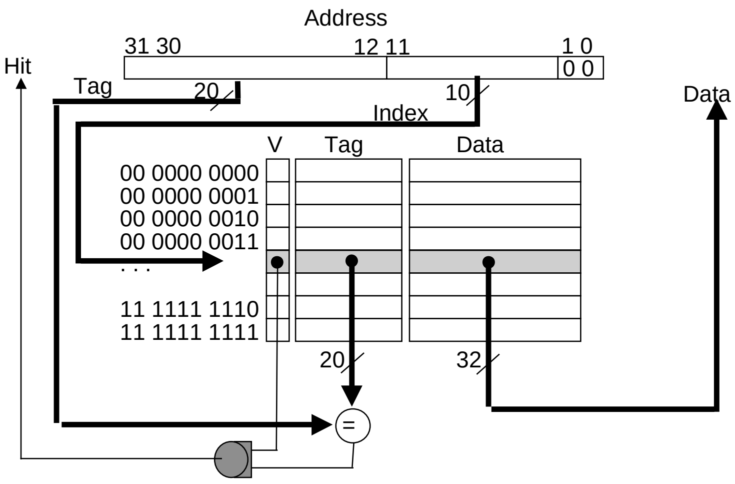


	V	Tag	Data
000	0		
001	0		
010	1	10	▨
011	0		
100	1	01	■
101	0		
110	0		
111	0		

01011	
01100	■
01101	
01110	
01111	
10000	
10001	
10010	▨
10011	
10100	■

© Larry Snyder, 2000 All rights reserved

Direct-Mapped Cache Operation



© Larry Snyder, 2000 All rights reserved

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.

© Larry Snyder, 2000 All rights reserved

Block size = 4 bytes
Address = 7 bits
T = 3, CA = 2, BA = 2

Reference Sequence

Memory Refs: 16, 24, 29, 25, 9, 26, 10, 16, 17, 18, 19

		16	24	29
00	---	001 F	001	001
01	---	---	---	---
10	---	---	001 F	001
11	---	---	---	001 F
	Initial	0010000 miss	0011000 miss	0011101 miss

	25	9	26	10
00	001	001	001	001
01	---	---	---	---
10	001	000 F	001 F	000 F
11	001	001	001	001
	0011001 hit	0001001 miss	0011010 miss	0001010 miss

	16	17	18	19
00	001	001	001	001
01	---	---	---	---
10	000	000	000	000
11	001	001	001	001
	0010000 hit	0010001 hit	0010010 hit	0010011 hit

© Larry Snyder, 2000 All rights reserved