CS430 Computer Architecture

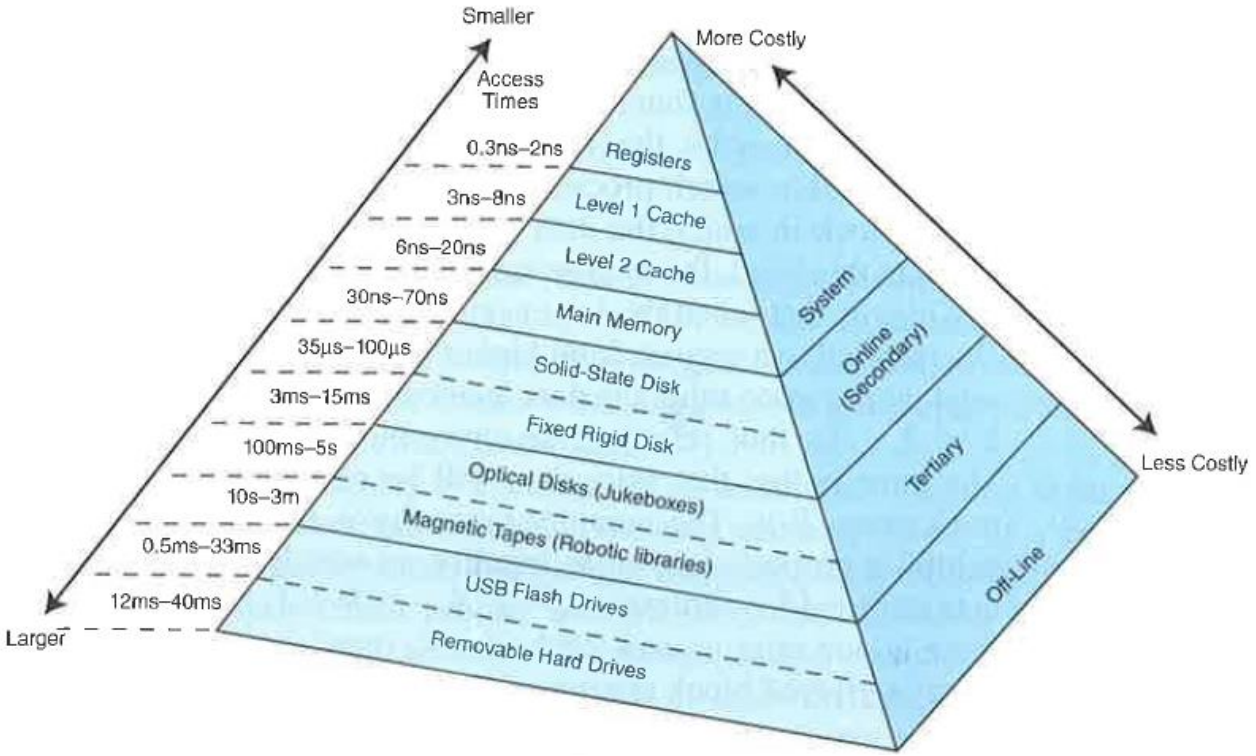
Spring 2015

Chapter 4
Cache Memory

Table 4.1 Key Characteristics of Computer Memory Systems

Location	**Performance**
Internal (e.g. processor registers, cache, main memory)	Access time
External (e.g. optical disks, magnetic disks, tapes)	Cycle time
Capacity	Transfer rate
Number of words	**Physical Type**
Number of bytes	Semiconductor
Unit of Transfer	Magnetic
Word	Optical
Block	Magneto-optical
Access Method	**Physical Characteristics**
Sequential	Volatile/nonvolatile
Direct	Erasable/nonerasable
Random	**Organization**
Associative	Memory modules

Memory Hierarchy



The Essentials of Computer Organization and Architecture 4th

Memory Design

- Goal
  1. Provide for large-capacity memory in computer systems because the capability is needed
     a. The capacity is needed
     b. The cost per bit is low
  2. Use lower-capacity memories
     a. Short access times are needed for better performance
  3. Decrease the frequency of access to slower memory

Locality of Reference

- During program execution, memory and data references tend to cluster due to the fact that programs contain
  a. functions
  b. loops
- Locality of reference allows the designer to take advantage of small high speed memory

What is Memory?

- Before getting into more specifics of cache design, we need to get a feel for what cache memory is.
- Main memory is made up of dynamic RAM (DRAM)
- Cache memory is made up of static RAM (SRAM)

SRAM vs DRAM

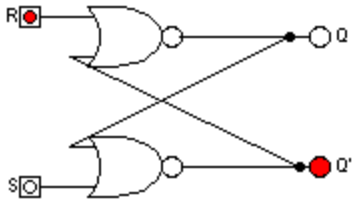
- SRAM
  1. requires no refresh
  2. has a shorter cycle time (time between the start of one mem access to the time when the next access can start) than DRAM
  3. relatively insensitive to disturbances such as electrical noise
- DRAM
  1. requires refresh every 10-100 ms
  2. sensitive to disturbances

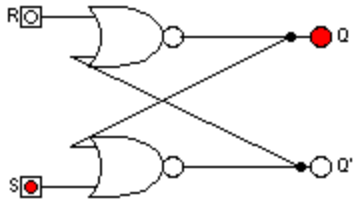
Note: Main memory is DRAM, On-chip cache is SRAM, Off-chip cache depends

SRAM

- SRAM is made from flip-flop logic-gate technology
- S-R flip-flop (or S-R latch)

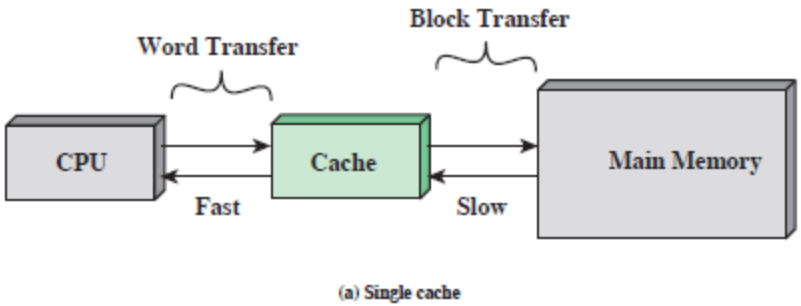
Characteristic Table		
Current Inputs	Current State	Next State
SR	Q_n	Q_{n+1}
00	0	0
00	1	1
01	0	0
01	1	0
10	0	1
10	1	1
11	0	undefined
11	1	undefined



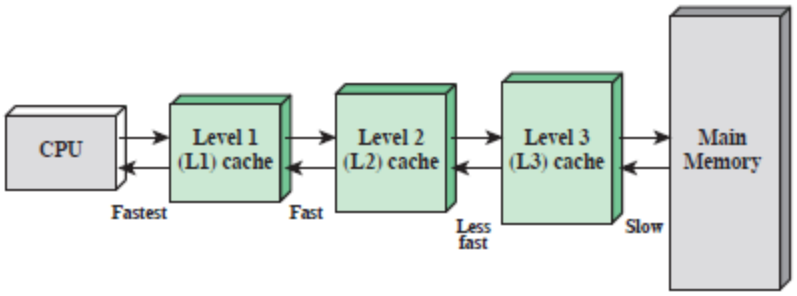


Cache Memory

- The goal of cache memory is to yield a memory speed of the fastest memories available while at the same time providing a much larger memory at the cheapest possible price.



(a) Single cache



(b) Three-level cache organization

Cache/Memory Structure

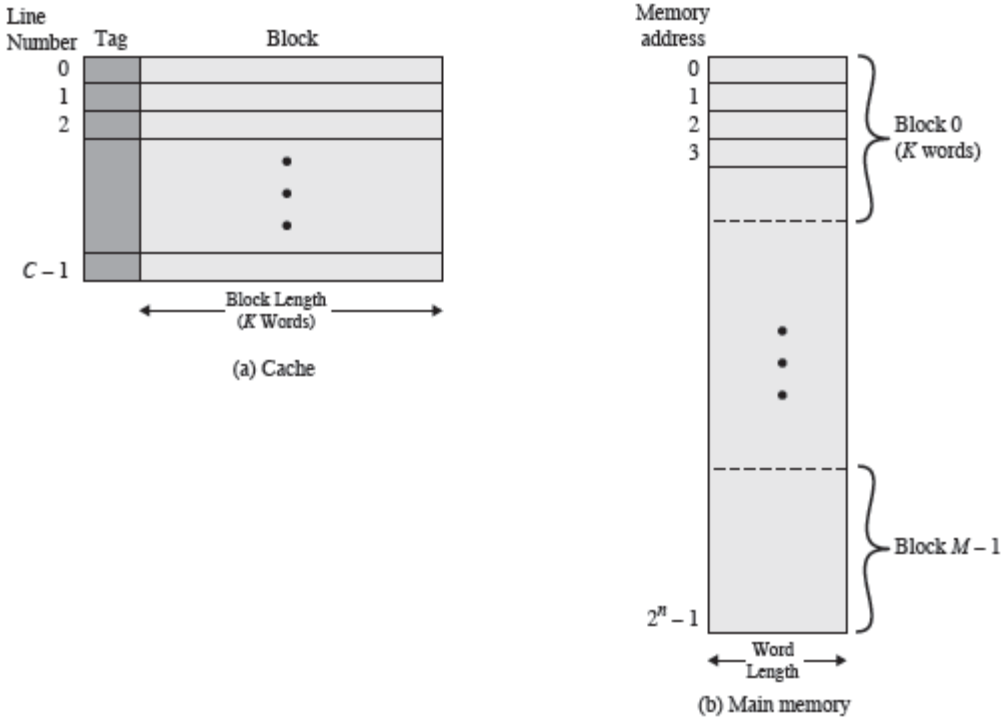


Figure 4.4 Cache/Main-Memory Structure

Cache/Memory Structure

- Things to note:
  1. A main memory has 2^n addressable memory locations
  2. Each word has a unique $n - bit$ address
  3. A block of $K - words$ is mapped to a line in the cache
  4. The number of words in a cache line is called the line size

Questions

Q1: How many blocks of information exist in main memory?

Q2: What is the line size in the above picture?

Q3: If M is the number of main memory blocks and C is the number of cache lines, would you say that C is a little smaller than M or a lot smaller than M? Why?

Cache Read

- RA is read address

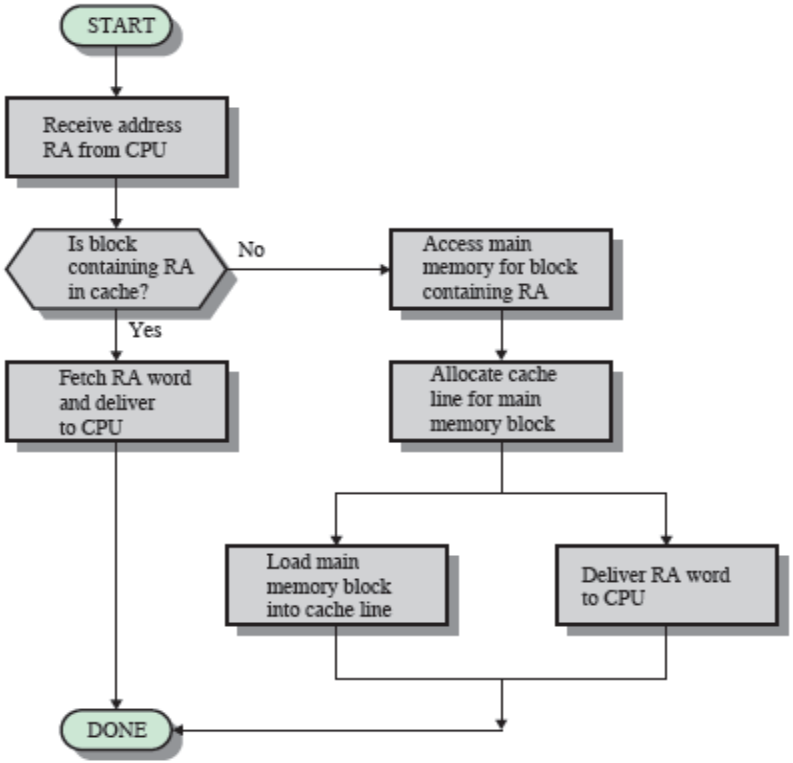


Figure 4.5 Cache Read Operation

Cache Organization

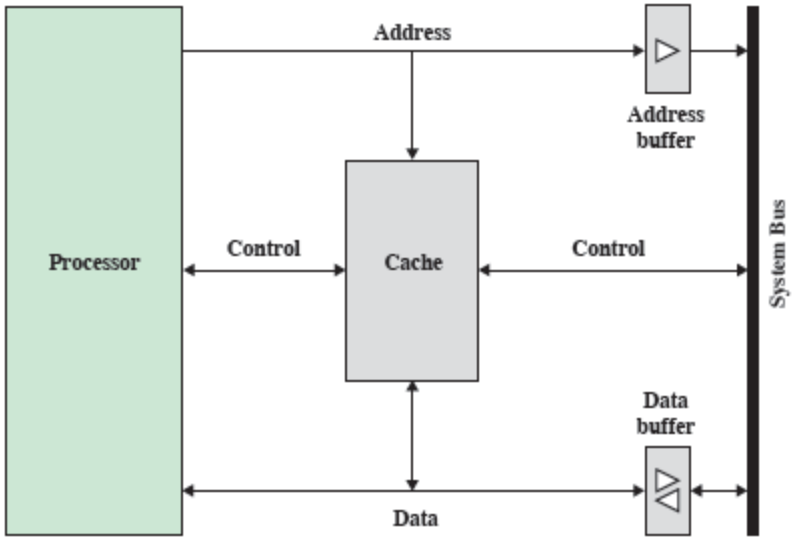


Figure 4.6 Typical Cache Organization

Elements of Cache Design

Table 4.2 Elements of Cache Design

Cache Addresses	**Write Policy**
Logical	Write through
Physical	Write back
Cache Size	**Line Size**
Mapping Function	**Number of caches**
Direct	Single or two level
Associative	Unified or split
Set Associative	
Replacement Algorithm	
Least recently used (LRU)	
First in first out (FIFO)	
Least frequently used (LFU)	
Random	

Cache Sizes

Table 4.3 Cache Sizes of Some Processors

Processor	Type	Year of Introduction	L1 Cache[a]	L2 cache	L3 Cache
IBM 360/85	Mainframe	1968	16 to 32 kB	—	—
PDP-11/70	Minicomputer	1975	1 kB	—	—
VAX 11/780	Minicomputer	1978	16 kB	—	—
IBM 3033	Mainframe	1978	64 kB	—	—
IBM 3090	Mainframe	1985	128 to 256 kB	—	—
Intel 80486	PC	1989	8 kB	—	—
Pentium	PC	1993	8 kB/8 kB	256 to 512 KB	—
PowerPC 601	PC	1993	32 kB	—	—
PowerPC 620	PC	1996	32 kB/32 kB	—	—
PowerPC G4	PC/server	1999	32 kB/32 kB	256 KB to 1 MB	2 MB
IBM S/390 G6	Mainframe	1999	256 kB	8 MB	—
Pentium 4	PC/server	2000	8 kB/8 kB	256 KB	—
IBM SP	High-end server/ supercomputer	2000	64 kB/32 kB	8 MB	—
CRAY MTA[b]	Supercomputer	2000	8 kB	2 MB	—
Itanium	PC/server	2001	16 kB/16 kB	96 KB	4 MB
Itanium 2	PC/server	2002	32 kB	256 KB	6 MB
IBM POWER5	High-end server	2003	64 kB	1.9 MB	36 MB
CRAY XD-1	Supercomputer	2004	64 kB/64 kB	1MB	—
IBM POWER6	PC/server	2007	64 kB/64 kB	4 MB	32 MB
IBM z10	Mainframe	2008	64 kB/128 kB	3 MB	24-48 MB
Intel Core i7 EE 990	Workstaton/ server	2011	6 × 32 kB/32 kB	1.5 MB	12 MB
IBM zEnterprise 196	Mainframe/ Server	2011	24 × 64 kB/ 128 kB	24 × 1.5 MB	24 MB L3 192 MB L4

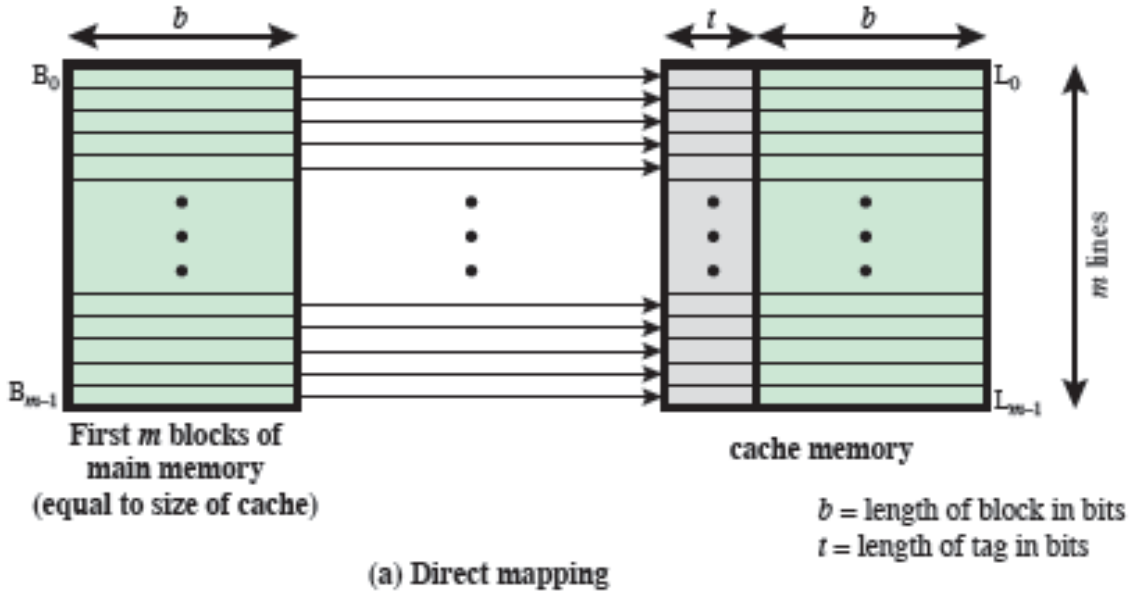
[a] Two values separated by a slash refer to instruction and data caches.
[b] Both caches are instruction only; no data caches.

Cache Mapping Functions

- Direct Mapping - simplest of the cache mapping schemes

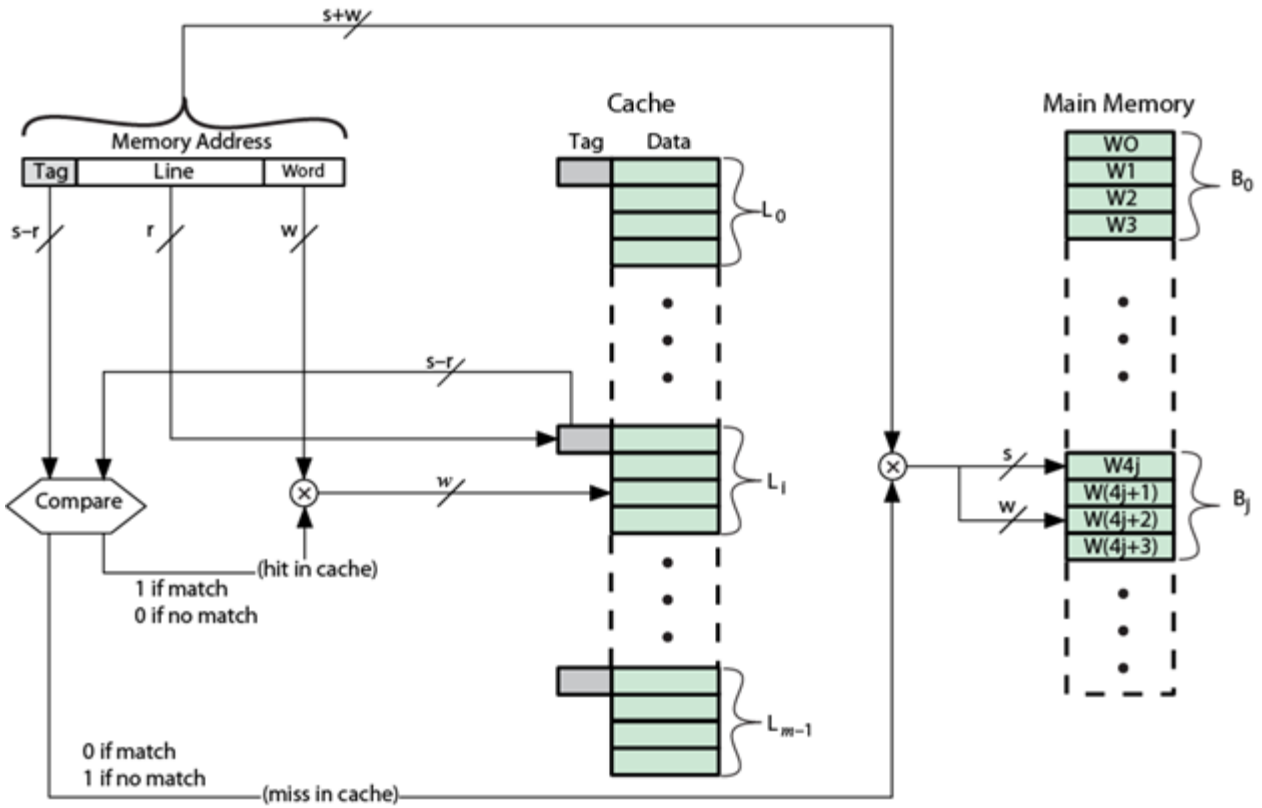
$$i = j \; modulo \; m \; where$$
$$i = cache \; line \; number$$
$$j = main \; memory \; block \; number$$
$$m = number \; of \; lines \; in \; the \; cache$$

Direct Mapping



(a) Direct mapping

Direct Mapping



Direct Mapping

- To summarize:
  - Address Length is (s + w) bits
  - Number of addressable units is 2^{s+w} bytes
  - Block size = line size = 2^{w} bytes
  - Number of blocks in main memory is = 2^{s}
  - Number of cache lines is 2^{r}
  - Tag size is (s-r) bits