Chapter6. Memory

2019年11月26日 14:07

NOTE TAKING AREA

Types of memory

Silicon memory: read-only memory (ROM), read-write memory (RAM).

- Programmable ROM (PROM): programmed once only.
- Erasable PROM (EPROM): contents can be erased by exposure to ultraviolet light and reprogrammed.
- Electrically erasable PROM (EEPROM): erase by an electrical signal.
- Flash erasable PROM (FEPROM, flash memory): similar to EEPROM, but erase over a significant part of the integrated circuit.

ROM is non-volatile, while RAM is volatile.

RAM: random access memory, read or store time doesn't depend upon the order in which the data is accessed.

- Dynamic RAM (DRAM): contents must be refreshed every few milliseconds.
- Static RAM (SRAM): uses D type latches (gated SR latch), doesn't need to be refreshed. Faster than DRAM.

Memory cache

Save frequently used data into memory cache in faster memory.

Cache relies on 2 features: temporal and spatial locality.

Cache operation / behavior: hit ratio h, miss ratio $m = 1 - h$.

Average access time: $h \times t_c + (1 - h) \times t_m + a$, t_c is cache access time, while t_m is main memory access time, a is extra time delay.

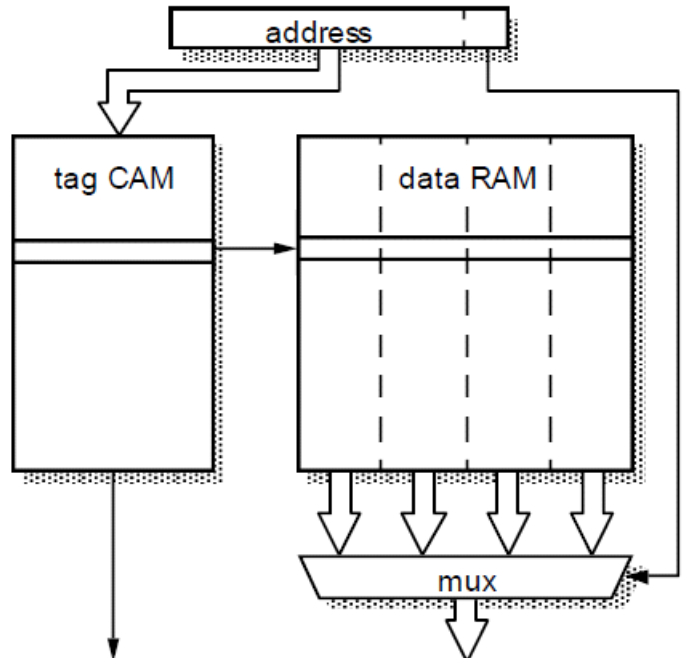
Cache control systems: associative cache, direct mapped cache.

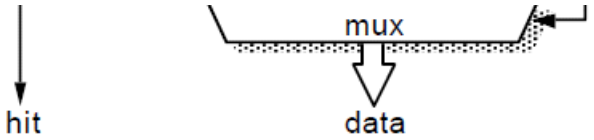
Associative cache

Contents of associative cache:

- Number of bytes in a block $X = 2^a$, a is a fixed integer, typically 4.
- Block number: $Y - a\ bits\ with\ Y\ bits\ main\ memory\ address$.
- Line length: a large word length, longer than memory word length.
- Valid bit (tag): if the line holds a valid copy of main memory words.

Fully associative cache organization relies on content addressed memory (CAM) to measure hit or miss (in parallel):





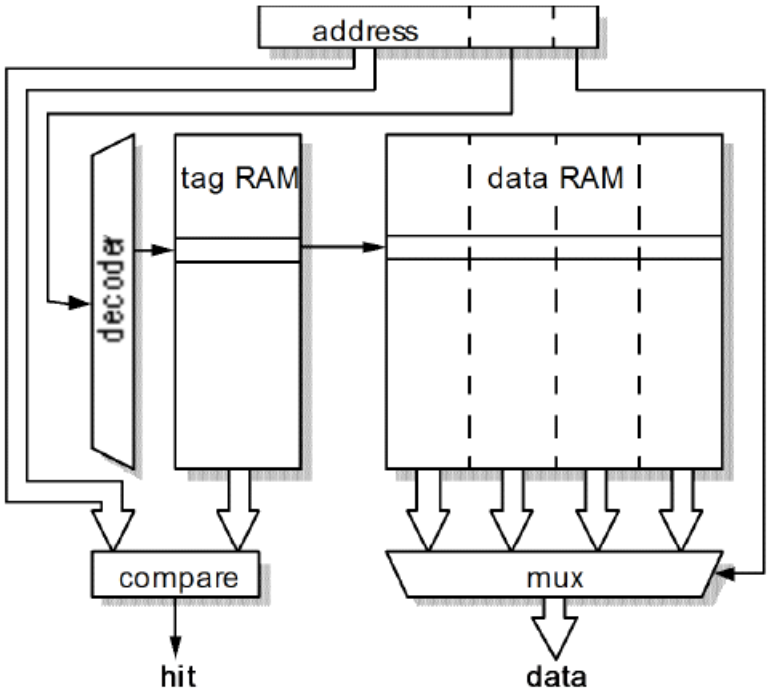
Action of associative cache control: compare hit, verify valid bit, use.

Direct-mapped cache

Contents in direct-mapped cache:

- Number of lines: 2^M $with\ M\ bits\ used\ to\ identify\ each\ cache\ line$.
- Least significant a bits are discarded as each line contains 2^a bytes.
- Following M bits are used as the cache line number.
- The most significant $Y - M - a$ bits are stored in tag field.

Direct-mapped cache controller: access cache line, verify valid bit, use.



Problems in cache systems

Remove useless entries when cache is full: only for associative cache.

- Entry has not been accessed for the longest time.
- Entry which has been in longest.
- Replace at random.

Write access problem: cache contents changed but main memory unchanged.

Write through and copy back.

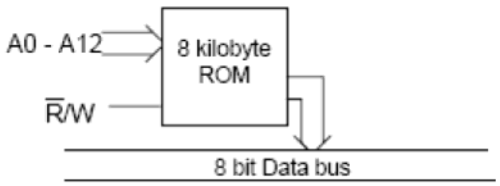
Memory addressing

Tri-state gates: enable, input, and high impedance.

enable	input	output
0	0	high-impedance
0	1	high-impedance
1	0	0
1	1	1

Memory and address lines: 2^N $byte\ with\ N\ address\ lines$.

8 kilobyte 13
16 14
etc.



R*/W is the corresponding signal means whether memory can access.
Split memory into little pieces: most significant bits defines which memory block, while least significant bits determines memory address (see example).

Memory map: defines location of each memory chip.

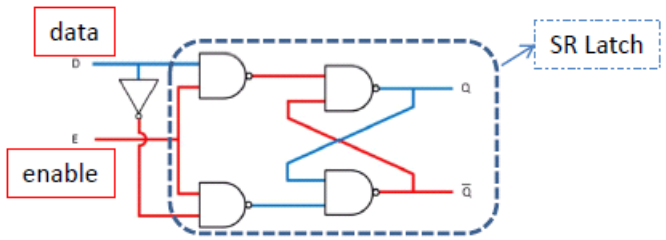
Input and output

Memory mapped I/O: assign input / output ports to memory address.
Use memory loads and memory writes to execute I/O.

Example: GPIO in ARM7.

CUE COLUMN

SR latch



Multiple levels of cache example

	Typical Size	Access time
CPU registers	64 to 256 bytes	1 cycle
Level 1 cache	16k to 64k bytes	1 to 2 cycles
Level 2 cache	128k to 1M bytes	8 cycles
Main memory	256M to Gbytes	16 cycles
Back-up	Very large	Block transfer

Associative cache example

Cache line number ↓	Valid bit	Block number	Block contents
0			
1			
2			
3	1	10110010	01001010010010001111101....
last			
No. of bits→	1	$Y - a$	8×2^a

8 bits per byte

Cache line length is = $(1 + Y - a + 8 \times 2^a)$ bits

number of bytes

Example: number of bytes stored in a cache line is 16.

data byte at main memory address 0x37B6A038 is 0xD4 – highlighted.

Tag field	Block contents
0x9AB6301	0xC2A1096C 6FE3D674 A956410B D4FE3D67
0x54B09FF	0xD4FE3D67 A956410B 6FE3D674 C2A1096C
0x37B6A03	0x6FE3D674 C2A1096C D4FE3D67 A956410B
0xF31A004	0xC2A1096C D4FE3D67 6FE3D674 A956410B
0xFF4E502	0x6FE3D674 C2A1096C A956410B D4FE3D67
0x07D3A00	0xA956410B C2A1096C 6FE3D674 D4FE3D67
…	… … … …
0x96301AB	0x6FE3D674 C2A1096C D4FE3D67 A956410B

0 8 15

Direct-mapped cache example

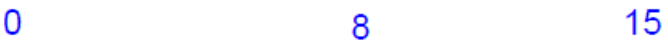
Cache line number ↓	Valid bit	Tag field	Block contents
0			
1			
2			
3	1	01011010	01001010010010001111010...
2^M-1			
No. of bits→	1	$Y - M - a$	8×2^a

So if M = 4 and a = 4 the contents of main memory address 01011010 0011 0000 and the following $2^4 - 1$ addresses is 01001010 01001000 1111010.....

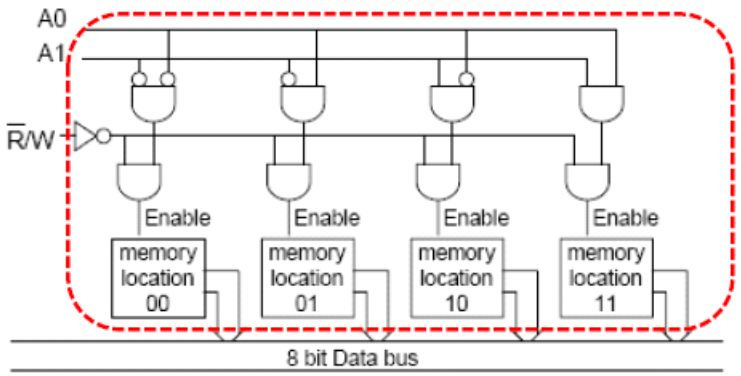
For cache line length 16 bytes and 256 lines:

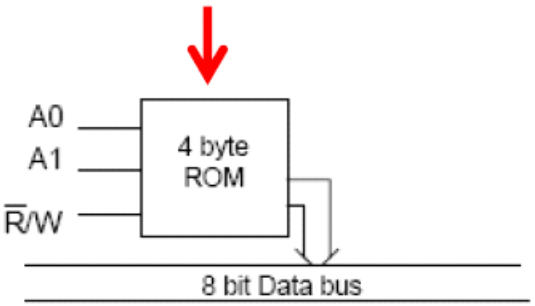
The data byte at main memory address 0x37B6A038 is 0xD4 – highlighted.

Cache line no.	Tag field	Block contents
0x00	0x07D3A	0xC2A1096C 6FE3D674 A956410B D4FE3D67
0x01	0x9AB63	0x6FE3D674 C2A1096C D4FE3D67 A956410B
0x02	0xFF4E5	0xC2A1096C 6FE3D674 A956410B D4FE3D67
0x03	0x37B6A	0x6FE3D674 C2A1096C D4FE3D67 A956410B
0x04	0xF31A0	0xC2A1096C 6FE3D674 A956410B D4FE3D67
…	…	… … … …
0xFF	0x54B09	0x6FE3D674 C2A1096C D4FE3D67 A956410B



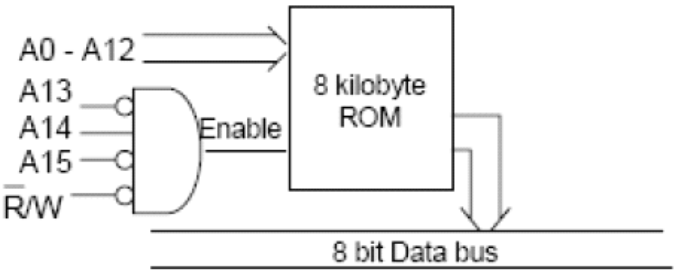
Simple ROM addressing example

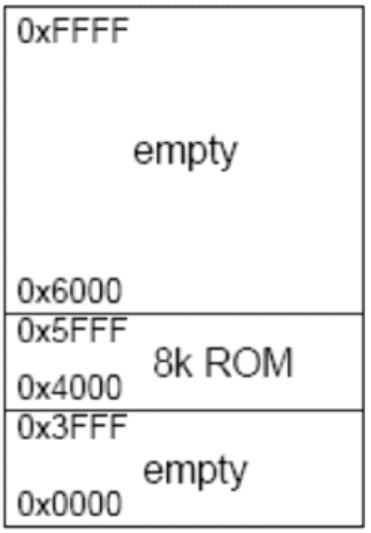




In this case, if A0=0 and A1=1, with R*/W=0, then location 00 is enabled.
For 8 kilobyte ROM addressing example:

The memory is only enabled when A13 = 0, A14 = 1, A15 = 0 and R*/W = 0.





If the address bus is 32 bits wide then all the address lines not connected to the memory chip, A13 to A31, are decoded.

SUMMARIES

1. Types of memory: ROM and RAM;
2. Memory cache: average accessing time, associative cache, and direct-mapped cache;
3. Problems in cache systems: replace entries, write access;
4. Memory addressing: tri-gates, memory map;
5. Memory mapped I/O.