# Tutorial 4: WRES1201 Computer System Architecture

1. What is the general relationship among access time, memory cost and capacity?

Faster access time, greater cost/bit, smaller capacity;

Slower access time, lower cost/bit, greater capacity;

1. What are the differences among *direct mapping*, *associative mapping* and *set associative mapping*?

In a cache system,

**direct mapping** maps each block of main memory into only one possible cache line. **Associative mapping** permits each main memory block to be loaded into any line of the cache.

**Set-associative mapping**, the cache is divided into a number of sets of cache lines; each main memory block can be mapped into any line in a particular set.

1. Consider a machine with a byte addressable main memory of 216 bytes and block size of 8 bytes. Assume that a direct mapped cache consisting of 32 lines is used with this machine.
   1. How is a 16-bit memory address divided into tag, line number, and byte number?

8bit tag; 5bit lines; 3bit byte number

* 1. Into what line would bytes with each of the following addresses be stored?
     1. 0001 0001 0001 1011
     2. 1100 0011 0011 0100
     3. 1101 0000 0001 1101
     4. 1010 1010 1010 1010
     5. 0001 0001 0001 1011 slot 3
     6. 1100 0011 0011 0100 slot 6
     7. 1101 0000 0001 1101 slot 3
     8. 1010 1010 1010 1010 slot 21
  2. Suppose the byte with address 0001 1010 0001 1010 is stored in the cache. What are the addresses of the bytes stored along with it?

0001 1010 0001 1000 till 0001 1010 0001 1111

* 1. How many total bytes of memory can be stored in the cache?

256 bytes

* 1. Why is the tag also stored in the cache?

Because two items with two different memory addresses can be stored in the same place in the cache. The tag is used to distinguish between them.

1. A cache has 64KB capacity, 128byte per lines and is 4-way set-associative. The system containing the cache uses 32-bit addresses. How many lines and sets does the cache have?

64KB/128byte=512 lines;

4-way set = 4 lines per set

512/4=128 sets

1. Describes three method used by cache memory to do replacement policy?

**Least recently used (LRU):** replace block which in the cache longest with no reference to it. A list to put the most recent used cache at the top.

First in first out (FIFO): replace block which in the cache longest. Arrange the cache with the first in data at the bottom of the list.

Least frequently used (LFU): replace block which experienced the fewest references. A counter is used to count the number of times the data referenced.

1. What are the differences between write-through and write-back?

Write through: the data in the main memory is updated together with the data inside the cache.

Write back: the data is only updated in the cache. A bit (use bit or dirty bit) used to show that an update occurred in the cache. When, the data is replaced and the dirty or use bit is 1, then the main memory will be updated.

1. Consider a single-level cache with an access time of 2.5ns, a line size of 64 bytes, and a hit ration of H= 0.95. Main memory uses a block transfer capability that has first-word (4 bytes) access time of 50ns and an access time of 5 ns for each word thereafter.
   1. What is the access time when there is a cache miss? Assume that the cache waits until the line has been fetched from main memory and then re-executes for a hit.

Tmiss = 2.5 + 50 + (15)(5) + 2.5 = 130 ns

* 1. Suppose that increasing the line size to 128 bytes increases the H to 0.97. Does this reduce the average memory access time?

Tmiss = 2.5 + 50 + (31)(5) + 2.5 = 210 ns

and *Ts* = *H* × *T*1 + (1 – *H*) × (*T*1 + *T*2) = (0.97)(2.5) + (0.03)(210) = 8.725 ns

1. For a system with two levels of cache, define Tc1 = first level cache access time; Tc2 = second level cache access time; Tm = main memory access time; H1 = first level hit ratio; H2 = combined first/second level hit ratio. Provide an equation for Ta for a read operation.

Ta = [H1Tc1 + (1 – H1)Tc2] + (1 – H2)Tm