# Computer Architecture 2 Midterm Exam

ELE3020, Hanyang University, 2019 Fall,

Department: ID: Name:

1.(20pt) Fill the blanks.
(a) ____ locality : Tendency to re-use recently accessed data items
(b) ____ locality : Tendency to reference data items that are close to other recently accessed items
(c) _____ : Minimum unit of information transfer between cache and main memory
(d) _____ : Minimum unit of information transfer between main memory and storage in virtual memory system

2. (15pt) Compute the sizes of the tag bits and the index bits.
(a)
- direct mapped
- 8 KB of data with 2-word blocks
- 32-bit address

(b)
- 4 way set associative
- 32 KB of data with 4-word blocks
- 32-bit address

(c)
- fully associative
- 64 KB of data with 16-word blocks
- 32-bit address

3. (10pt) Consider a 8 Mb DRAM device. In the DRAM, assume that memory array is composed of 2048 rows and 4096 columns.
Assume that it requires 50 ns to load a single row line into a row buffer, and 3 ns to load a bit from the row buffer. Using 32 DRAM devices in parallel, we construct a 32 MB DRAM (8Mbit x 32) in which a word can be transferred at once.

(a) Suppose that word load addresses are given as following: 0x0, 0x4, 0x8, 0xc, 0x100000, 0x100004, 0x100008. 0x10000c, Computer the time to load the eight words.
(b) Suppose that load addresses are given as following: 0x0, 0x100000, 0x4, 0x100004, 0x8, 0x100008, 0xc, 0x10000c. Computer the time to load the eight words.

4. (15pt) Assume that instruction cache miss rate is 2%, data cache miss rate is 10%, CPI (clock cycle per instruction) is 2 without any memory stall, and miss penalty is 200 cycles. In addition, assume that the frequency of loads/stores is 30%. Suppose that the CPU clock frequency is 2GHz. Assume that a given program A requires 10 seconds to complete on this machine without considering memory stall.

(a) Compute CPI with memory stall and compute the execution time of program A.
(b) When CPI without any memory stall becomes 1, compute CPI with memory stall and compute the execution time of program A.
(c) If the CPU clock rate is doubled with the same memory (so the clock frequency becomes 4GHz) when CPI without memory stall is 2, compute CPI with memory stall and compute the execution time of program A.

5. (10pt) Compare write-through and write-back in terms of performance and complexity. For write-back cache, explain the dirty bit.

See the back

6. (10pt) Cache misses can be classified into compulsory, capacity, and conflict misses. We have a fully associative cache of which size is 32KB. When an application is running on the cache, cache miss rate is 5%. When the cache size increases to infinite, the cache miss rate reduces to 4% for the same application. When the 32KB fully associative cache changes to 32KB 4-way set associative cache, the miss rate becomes 7%. For the fully associative cache of which size is 32KB, what are compute compulsory, capacity, and conflict cache miss rates?

7. (10pt) Suppose that a cache has 128 blocks that are 32 bytes each. Show how to break the following address into the tag, set index, block offset, and byte offset if the cache is direct-mapped, 2-way set associative, 4-way set associative, 8-way set associative and fully associative?

0000 1000 0101 1100 0001 0001 0111 1001

8. (20pt) Caches are important to providing a high-performance memory hierarchy to processors. Below is a list of 32-bit memory address references, given as word addresses (so you need to multiply 4 to change them to the byte addresses).
3, 180, 43, 2, 191, 88, 190, 14, 181, 44, 186, 253

You are asked to optimize a cache design for the given references. There are three direct-mapped cache designs possible, all with a total of 8 words of data: C1 has 1-word blocks, C2 has 2-word blocks, and C3 has 4-word blocks. In terms of miss rate, which cache design is the best? If the miss penalty is 20 cycles, and C1 has an hit time of 1 cycles, C2 takes 2 cycles, and C3 takes 3 cycles, which is the best cache design? Note that you need to show the reasons.

9. (30pt) Consider the following C code.

```
int a[1024], b[1024], c[1024], d[1024];

void main(){
        int s=0;
        for(int i=0; i<1024; ++i) {
                s += a[i]*b[i]+c[i]*d[i];
        }
}
```

Assume that the starting addresses of a, b, c, d are 0x0000, 0x1000, 0x2000, and 0x3000, respectively.

(a) (10pt) Consider a direct mapped cache which has 8 blocks and the block size is 16-word. Compute the data cache miss rate for the above program. When the block size increases to 32-word with 8 blocks, compute the data cache miss rate.

For the above program, the cache miss rate is too high. To improve the cache hit rate, we either increases the cache associativity or optimize the program.

(b) (5pt) For the above program, what is the minimal associativity to minimize the conflict cache misses. And for the revised cache, what is the data cache miss rate?

There are two ways to optimize the program.

(c) (10pt) Change the program to improve the cache hit rate and write the revised program. In addition, for the same cache (16word block size) in (a), compute the data cache miss rate.

(d) (5pt) Instead of changing the program, the change of the starting addresses of arrays improves the cache hit rate since the above program has high conflict misses. Change the starting addresses of arrays a, b, c, and d to minimize the conflict misses and compute the data cache miss rate.

10. (15pt) Assume that TLB hit time is 0.5 cycles, cache index time (time to load a cache line) is 0.5 cycles, and cache tag comparison time is 0.5 cycles.
Compute hit time (TLB hit+cache hit) for
(a) physically indexed and physically tagged cache
(b) virtually indexed and physically tagged cache, and
(c) virtually indexed and virtually tagged cache

11. (10 pt) Describe how OS handles the page fault.

12. Assume 4 KB pages, a four-entry fully associative TLB, and LRU replacement. If pages must be brought in from disk, increment the next largest page number.

| 4095, 31272, 15789,15000,7193,4096,8912 |
|---|

TLB

| Valid | Tag | Physical Page Number |
|---|---|---|
| 1 | 11 | 12 |
| 1 | 7 | 4 |
| 1 | 3 | 6 |
| 0 | 4 | 9 |

Page table

| Valid | Physical page or in disk |
|---|---|
| 1 | 5 |
| 0 | Disk |
| 0 | Disk |
| 1 | 6 |
| 1 | 9 |
| 1 | 11 |
| 0 | Disk |
| 1 | 4 |
| 0 | Disk |
| 0 | Disk |
| 1 | 3 |
| 1 | 12 |

9.1 [10pt] Given the address stream in the table, and the shown initial state of the TLB and page table, show the final state of the system. Also list for each reference if it is a hit in the TLB, a hit in the page table, or a page fault.

9.2 [10pt] Repeat 9.1, but this time use 16 KB pages instead of 4 KB pages. What would be some of the advantages of having a larger page size? What are some of the disadvantages?

There are several parameters that impact the overall size of the page table. Listed below are several key page table parameters.

| | Virtual address size | Page size | Page table entry size |
|---|---|---|---|
| a. | 32 bits | 4 KB | 4 bytes |
| b. | 64 bits | 16 KB | 8 bytes |

9.3 [10pt] Given the parameters in the table above, calculate the total page table size for a system running five applications that utilize half of the memory available.