Cairo University
Electronics and Communications Department
Dr. Hossam A. H. Fahmy

Computer architecture
MSc Final , June 2012
Duration: 120 minutes

This exam has **(10)** pages. Please check that it is complete.

Please write clearly any assumptions you make.

1. Consider the following progression of on-chip caches for three generations of chips.

   (a) Assume the first generation chip had a direct-mapped 8KB single-cycle first-level data cache. Assume T_{hit} for this cache is one cycle, T_{miss} is 100 cycles (to access off-chip memory), and the miss rate is 20%. What is the average memory access latency? (3 points)

   (b) The second generation chip used the same direct-mapped 8KB single-cycle first-level data cache, but added a 96KB second-level cache on the chip. Assume the second-level cache has a 10-cycle hit latency. If an access misses in the second-level cache, it takes an additional 100 cycles to fetch the block from the off-chip memory. The second-level cache has a global miss rate of 4% of memory operations (which corresponds to a local miss rate of 20%). What is the average memory access latency? (3 points)

   (c) The third generation chip replaced the two-levels of on-chip caches with a single-level of cache: a set-associative 64KB cache. Assume this cache has a 5% miss rate and the same 100 cycle miss latency as above. Under what conditions does this new cache configuration have an average memory latency lower than the second generation configuration? (4 points)

2. The following code implements the Newton-Raphson iteration $x_{i+1} = x_i(2 - bx_i)$ to find a better estimate of the reciprocal of b starting from an initial estimate x_0.

(At the start of the code, R1 holds the value of $-b$, R2 holds the value 2, R3 holds the value of x_0, and R7 holds the requested number of iterations.)

```
Loop:  Mul  R5, R1, R3  ;  R5 $= -bx_i$
       Add  R6, R2, R5  ;  R6 $= 2 - bx_i$
       Mul  R3, R3, R6  ;  $x_{i+1} = x_i(2 - bx_i)$
       Dec  R7          ;  decrement R7
       BNZ  Loop        ;  branch if not zero to restart the loop
```

Indicate clearly whether the use of the following compiler optimization techniques enhances the code performance: loop unrolling, SW pipelining, and trace scheduling. (10 points)

3. Please specify whether each of the following statements is true or false and indicate your reasons. Note: *A wrong answer subtracts half a point.* (12 points)

   (a) The CPI of a processor can never be less than one.

   (b) Vector instructions improve the code density and reduce the instruction bandwidth.

   (c) Cache coherence is not required if there is no sharing between multiple processes.

   (d) On the average, a directory-based (networked) cache coherency protocol reduces the invalidate and update traffic.

   (e) In case of a deadlock in the interconnection network of a parallel processor, the packets are dropped to break the deadlock.

   (f) To calculate the component of the seek time in accessing a disk, we use the average of the maximum and minimum seek times provided by the manufacturer.

   (g) It is impossible to access both the TLB and the cache in parallel. The TLB access must always be first.

   (h) Vector instructions may replace some loop constructs.

(i) Vector processors provide a good speedup on small problems while multiple-issue machines provide a good speedup on large scientific loads.

(j) High performance buses are usually wide buses with multiple masters.

(k) Out-of-order execution machines save their results out of the original program's order.

(l) The compulsory miss rate in the cache is not affected by increasing the size of a cache line (number of bytes per line).

4. Please choose the correct answer for each of the following questions and indicate your reasons. Note: *A wrong answer subtracts one point.* (22 points)

   (a) What characteristic of RAM memory makes it not suitable for permanent storage?

      i. too slow
      ii. unreliable
      iii. it is volatile
      iv. too bulky

   (b) Computers use addressing mode techniques for

      i. giving programming versatility to the user by providing facilities as pointers to memory counters for loop control
      ii. to reduce no. of bits in the field of instruction
      iii. specifying rules for modifying or interpreting address field of the instruction
      iv. All the above

   (c) The average time required to reach a storage location in memory and obtain its contents is called the

      i. seek time
      ii. turnaround time
      iii. access time
      iv. transfer time

   (d) The idea of cache memory is based

      i. on the property of locality of reference
      ii. on the heuristic 90-10 rule
      iii. on the fact that references generally tend to cluster
      iv. all of the above

(e) Cache memory acts between

i. CPU and RAM
ii. RAM and ROM
iii. CPU and Hard Disk
iv. None of these

(f) Write Through technique is used in which memory for updating the data

i. Virtual memory
ii. Main memory
iii. Auxiliary memory
iv. Cache memory

(g) Cache memory consists of

i. Static RAM
ii. Dynamic RAM
iii. Magnetic memory
iv. None of these

(h) The instructions which copy information from one location to another either in the processors internal register set or in the external main memory are

i. Data transfer instructions.
ii. Program control instructions.
iii. Input-output instructions.
iv. Logical instructions.

(i) In which addressing mode the operand is given explicitly in the instruction

i. Absolute.

ii. Immediate.
iii. Indirect.
iv. Direct.

(j) A stack organized computer has

i. Three-address Instruction.
ii. Two-address Instruction.
iii. One-address Instruction.
iv. Zero-address Instruction.

(k) A page fault

i. Occurs when there is an error in a specific page.
ii. Occurs when a program accesses a page of main memory.
iii. Occurs when a program accesses a page not currently in main memory.
iv. Occurs when a program accesses a page belonging to another program.

5. An instruction requires four stages to execute: stage 1 (instruction fetch) requires 3 ns, stage 2 (instruction decode) = 0.9 ns, stage 3 (instruction execute) = 2 ns and stage 4 (store results) = 1 ns. An instruction must proceed through the stages in sequence. (2 points)

   (a) What is the minimum asynchronous time for any single instruction to complete?

   (b) We want to set this up as a pipelined operation. Assume that none of the stages may be subdivided into smaller stages. However, you may allow a single stage to take multiple clock cycles if you want and wait for its result to complete before allowing another instruction to enter such a stage. How many stages should we have and at what rate should we clock the pipeline? Depending on your answer, how frequently can we initiate the execution of a new instruction, and what is the latency? (8 points)

Left blank, use as you like

Left blank, use as you like

End of exam.