ECE 411 Spring 2012
Midterm 1 practice homework problems

MP1

1) General Questions - LC3b design, Assembly

   a) Assume all memory accesses take 5 cycles. How many cycles will this program take to execute?

   LDR r1, r0, off1
   LDR r2, r0, off2
   ADD r1, r1, r1
   AND r3, r1, r2
   STR r3, r0, off3

   b) Write a routine to count the number of set bits in a word using LC3b assembly.

   For example, with input 0xFFFF, the output would be 0x0010 (16). With input 0x1000, the output would be 0x0001 (1). Hint: Come up with a C algorithm first!

2) Consider your MP1 datapath and control unit.

   Your cruel TA has decided that you should implement AddM to replace the add immediate instruction. Instead of acting as an immediate value, the 5 low bits of the instruction are an offset (from PC) to a memory location, at which is stored the second operand. For example:

   ADD DR, SR, offset5

   a) Give the RTL for this instruction.

   b) What changes should be made to the datapath?

   c) What changes should be made to the control unit (what new states must be added)?
MP2.1

1) Suppose that you are given a memory with only one memory write signal, which writes an entire 16 bit word (in MP2.1, the memory uses two signals - MWRITEH_L and MWRITEL_L).

   a) What instruction would see the largest impact (e.g., what instruction would need to change the most from the original design) from this change?

   b) What changes need to be made to the datapath so that this instruction will work with the new memory?

   c) …to the control unit (state machine)…?

   d) Would this affect the performance of your design? Explain.

2) From the MP1 design, you are going to add LDB instruction. What changes should be made to the design? List any addition of the components or signals.

   LDB: Opcode(15:12), DR (11:9), BaseR(8:6), Offset6(5:0)
   DR = ZEXT(mem[BaseR+SEXT(offset6)]);
   setcc;
ISA Trade-offs

1) General
   a) What does RISC stand for? CISC?

   b) Give an example of each architecture.

   c) Give 2 advantages of RISC over CISC.

   d) What is an advantage of using variable-length instructions? A disadvantage?

   e) What is an advantage to having a large number of architectural registers? A disadvantage?

   f) What are the four classifications of ISAs as discussed in class?

   g) Write the following in each of the 4 ISA classes.
      \[ C = A + B - 3 \]
2) Write the RTL code for the following Assembly instructions.

<Assembly instruction>

Loop: 1) LD R3, 0(R1)
    2) ADD R5, R3, R0
    3) ADDI R1, R1, #4
    4) SUBI R4, R4, #1
    5) BRnp R4, Loop
Performance

1) Various Performance Questions

    a) Does it make sense to compare the MIPS of two different processors? Why or why not?

    b) Briefly explain the difference between latency and throughput.

    c) A program is inherently 4% serial. What is the maximum speedup that can be obtained by parallelizing the program?

    d) For a given ISA, how might performance be increased?

2) A compiler designer is trying to evaluate two code sequences for a particular machine, which runs with 500MHz. Based on the hardware implementation, there are three different classes of instructions: Class A, Class B, Class C, and they require 3, 1, and 2 cycles respectively.

The first code sequence has these instructions in total: 1 of A, 2 of B, and 2 of C
The second code sequence has these instructions in total: 5 of A, 1 of B, 3 of C

    a) What is the CPI for each sequence?

    b) What is the execution time for each sequence?

    c) What is the MIPS according to each sequence?
2) Amdahl’s law
   a) Assume a new execution mode provides a 1.5x speedup to the sections of the programs where it applies. What percentage of a program must run with new execution mode for an overall speed up of 10%?

   b) Two enhancements are proposed: one that can enhance 40% of the execution time with speedup of 1.5, and another that can enhance different 25% of execution time with some (greater) speedup value X. How much should X be for better enhancement than first?

3) Consider the following two versions of matrix multiplication: Which version runs faster? Why? (Hint: consider cache locality)

   (1)
   ```java
   for (i=1, i <= n, i++) {
     for (j=1, j <= n, j++) {
       for (k=1, k <= n, k++) {
         c[i,j]=a[i,k]*b[k,j]+c[i,j]
       }
     }
   }
   ```

   (2)
   ```java
   for (i=1, i <= n, i++) {
     for (k=1, k <= n, k++) {
       for (j=1, j <= n, j++) {
         c[i,j]=a[i,k]*b[k,j]+c[i,j]
       }
     }
   }
   ```
Memory hierarchy

1) General
   a) What is spatial locality?
   b) What is temporal locality?

c) Suppose the following access times and hit rates.
   L1 cache: 5ns / 95%
   L2 cache: 30ns / 98%
   Main Memory: 1000ns / 99%
   Disk: 20ms / 100%
   What is the average memory access time?

2) Suppose a processor with virtual memory has pages of size 1MB, 48 bit virtual addresses and 32 bit physical addresses. Answer the following questions.
   a) How many bits long is the page offset in a virtual address?
   b) How many bits long is a virtual page number?
   c) Is the page number in the most significant bits of a virtual address, or the least significant bits?
   d) How many bits long is a physical page number?
   e) How many bits long is the page offset in a physical address?
f) How much virtual address space is covered by a 8 entry TLB?

g) Rounding (up) the page table entry up to the nearest byte if necessary, how much space would be required to hold the page table for this system?

3) A computer has a cache memory and a main memory with the following features:
   - Memory cache access time: 4 ns
   - Main memory access time: 80 ns
   - The time needed to load a line in cache is 120 ns.
   - Write-through policy.

If the hit ratio in this computer is 95% and the 90% of the memory accesses are read operations, what is the average access time?

4) A computer has a cache memory and a main memory with the following features:
   - Size: 16KB with lines of 32 bytes (8 words)
   - Access time: 10ns

The cache memory is connected to a memory using a bus. The main memory transfers a block of 8 words in 120 ns. What is the hit ratio needed to obtain an average access time of 20 ns?
5) In each part of this problem, you will be given a graph showing the average memory access time that a theoretical computer sees when executing a program, as a function of the size of the data array the program references. Each graph will also be marked with a set of letters that represent values on the graph. You know about the computer’s memory system consists of a single-level cache, a main memory, and a virtual memory. The computer also has a magic instruction memory, such that the only memory references the program makes are to its data array. You know nothing about the program, except that it accesses a single array of data whose size can be varied in different runs of the program.

For each graph, you need to do two things. First, you need to give a text description of how the program accesses its memory array, given the average access times shown on the graph. This description should explain why you believe that the description is correct. Second, for each of the marked points on the graph, state which parameter of the computer system the value represents. Do not write pseudo-code or attempt to provide an example of a program that would generate the graph shown. Describe the program’s behavior using a few sentences of text.

Example (albeit nonsensical) Answer: “This program executes a loop. In each iteration of the loop, a random-number generator selects an item in the array to be read. This is evident because the access time is mostly independent of the size of the array, with slight variations because the number of cache hits the program sees is a function of the sequence generated by the random number generator, and varies from run to run. In the graph, the value A indicates the clock speed of the computer, B indicates the number of entries in the branch predictor, and C indicates Professor Kumar’s estimate of the average number of hours of sleep his students have gotten in the last week.

a) How is the program accessing data?

A =
B =
C =
D =
E =

b) For this part, consider only the cache and the main memory. (i.e., assume that the program never page faults). How is the program accessing data?

A =
B =
C =
6) A computer's memory system has three levels: a (unified) cache, a main memory, and a virtual memory.

Assume that the hit rate of the cache is 98%, that memory accesses that hit in the cache take 2 cycles, that the main memory has a hit rate of 95%, that memory accesses that hit in the main memory take 100 cycles, and that the virtual memory handles a memory operation in 100,000 cycles. All of the latencies given are the total time to complete a memory reference that reaches a given level of the hierarchy (i.e., a request that is completed by the main memory takes 100 cycles, not 2 cycles to find that the data is not in the cache + 100 cycles to handle the request in the main memory.)

a) What is the average time required to complete a memory operation in the system?

b) You are considering three different optimizations to the memory system. One would increase the hit rate of the main memory to 97%, one would increase the hit rate of the cache to 99%, and one would decrease the access time of the memory to 30 cycles. If each of these optimizations cost the same amount of memory, which would you recommend that a customer purchase first? Second, assuming that the customer has already purchased the one you recommended first, what would you recommend the next?
Caches

1) Cache basics
   a) What are the three types of cache misses? Explain each (1 sentence).
   
   b) What is a fully associative cache? Direct mapped?
   
   c) Give one advantage and disadvantage each of both fully associative and direct mapped caches.
   
   d) What are two possible write policies for a cache? What do they mean?

2) Consider a 16KB cache. Each line contains 64 bytes. The system uses 48 bit addresses.
   
   a) Suppose the cache is direct mapped and word (2 bytes) addressable. How many tag, set, and offset bits are needed? How many sets are there?
   
   b) The cache is fully associative and byte addressable. How many tag, set, and offset bits are there? How many sets?
   
   c) The cache is 8-way set associative and word addressable. How many tag, set, and offset bits are there? How many sets are there?
3) A system’s data cache has a capacity of 8KB, and 32-byte cache lines. It is byte addressable. Addresses in the system are 32 bits long. Note that the comparators are the size of tag for each cache line.

   a) How large are the tag, set, and offset fields within the address if the cache is direct-mapped? How many comparators are needed for the hit/miss evaluation?

   b) How large are the tag, set, and offset fields within the address if the cache is four-way set-associative? How many comparators are needed for the hit/miss evaluation?

   c) How large are the tag, set, and offset fields within the address if the cache is fully-associative? How many comparators are needed for the hit/miss evaluation and why?

4) On a machine with 32bit byte-addressable space, the cache has a capacity of 64KB and 512-byte lines. Fill the following tables for different cache designs.

   a) Direct mapped

<table>
<thead>
<tr>
<th>Tag</th>
<th>Index</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

   b) 8-way set-associative

<table>
<thead>
<tr>
<th>Tag</th>
<th>Index</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

   c) Fully associative

<table>
<thead>
<tr>
<th>Tag</th>
<th>Index</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

d) Where will you find the word at address 0x2012abcd when using 8-way associative scheme?
5) (LRU execution) For a 4-way set associative cache with initial status given below, write down the state of a cache set for each memory access of sequence B, D, A, C, B, A, C, B using LRU and FIFO scheme. Show the cache miss rates.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

B: 
D: 
A: 
C: 
B: 
A: 
C: 
B: 

6) A system’s data cache has a capacity of 16KB, and 128-byte cache lines. Addresses in the system are 32 bits long. Assume byte addressability.

a) How large are the tag, set, and offset fields within the address if the cache is direct-mapped?

b) How large are the tag, set, and offset fields within the address if the cache is four-way set-associative?

c) How large are the tag, set, and offset fields within the address if the cache is fully-associative?
Single / Multicycle

1) What are the 5 stages in the instruction execution cycle? Give a brief explanation of each.

2) Describe the pros and cons of single cycle and multiple cycle datapath design.