Computer Architecture
Comprehensive Examination
Fall 1995-96

SOLUTIONS: COMPUTER ARCHITECTURE

This exam is closed book. There are 55 points in total, and 60 minutes for the exam. Write your answers in the spaces provided. Where possible show all intermediate steps in computing your answers; that will allow us to give partial credit for incorrect answers.

1	15	
2	5	
3	5	
4	5	
5	5	
6	20	
TOTAL	**55**	

28

Question 1 - Memory System Design (15 points)

Imagine you have a cache-coherent multiprocessor system in which a cache-miss causes the hardware to place not only the memory block that you missed into the cache, but also the following memory block. Furthermore, whenever a cache line is accessed by the processor for the first time, the following cache line is also brought in.

(a) (8 points) What are the advantages and disadvantages of the above approach as compared to simply doubling the cache line size in the base system?

Answer:
Advantages: (1) Reduced false sharing in MPs (i.e., reduced bouncing of lines between one processor and another even though they are writing to distinct portions of the line); (2) Can totally eliminate cache misses in case of unit-stride data accesses; (iii) Lower miss penalty as cache-line-size is smaller.

Disadvantages: (1) More complex hardware; (2) More tag overhead with smaller lines; (3) Makes less efficient use of the bus, as several smaller transfers rather than a single large transfer is being done (can reduce number of processors that can be supported on a given bus); (4) may fetch more unnecessary data than the scheme with double cache line size.

(b) (4 points) What information would you need to quantitatively evaluate whether to implement the suggested approach on a particular machine?

Answer:
- effective miss rate of the two schemes (impacts stall time)
- average miss penalty under the two schemes (impacts stall time)
- bus bandwidth required per processor under the two schemes (determines when contention will kick in and limit number of processors that can be supported)

(c) (3 points) Would you expect the above approach to be more effective for instruction cache or data cache?

Answer: The prefetching approach should be quite effective for both instruction and data caches. The utility for data caches is a function of the spatial locality in the application (e.g., it may not work too well for programs that make heavy use of pointers, or where the arrays are being accessed in a non-unit-stride manner). The instruction reference stream of programs usually has more spatial locality, and therefore the benefits should spread across a larger segment of programs.

29

Question 2 - Processor Pipelines (5 points)

The standard DLX pipeline is quite short; only 5 stages deep. In contrast, the recently announced intel P6 processor has a 13-deep pipeline. Discuss, in general, the advantages and disadvantages of long versus short pipelines.

Answer:

Long Pipeline Advantages: It allows for higher clock rate (as work done in each stage is smaller), potentially offering higher performance.

Long Pipeline Disadvantages: (1) Requires extra concurrency, which may or may not be available in the program; said another way, there will be more load-delay slots and branch-delay slots that will need to be filled, which may not always be possible; (2) if branch-prediction is used, there will be higher penalty for mis-predicted branches; (3) tougher to balance the work between the various pipe stages; (4) more extensive and complex bypassing logic is needed, implying more complex hardware.

Question 3 - Precise vs. Imprecise Exceptions (5 points)

The DLX architecture presents a precise exception model.

(a) (2 points) Explain briefly what the term "precise exception model" means.

Answer: If the pipeline can be stopped so that the instructions just before the faulting instruction are completed, and those after it can be restarted from scratch, the pipeline is said to have precise exceptions.

(b) (3 points) Give a example of how the precise exception model could be violated in the DLX pipeline if the exceptions were signalled in the order in which they occurred rather than by waiting until the excepting instruction reaches a particular point in the pipeline. State any assumptions that you make.

Answer:

```
LW   R4, 256(R2)        ==> this instr has a protection fault

ADD  R7, R4, R5         ==> this instr lies on a different page
                            than the previous one, and the IFETCH gets
                            a page fault.
```

Question 4 - Virtual Memory Systems (5 points)

Assume a machine has a virtual address size of 32 bits, a page size of 4KB, and a 4-way associative TLB with 64 total entries. Indicate how a 32-bit virtual address is broken down into page offset, TLB index, and TLB tag fields. Show the width (in bits) and position of each of these fields.

`Answer:`

16-bit TLB tag	4-bit	12-bit offset
	TLB index	

Question 5 - Cache Organization (5 points)

Consider 2 caches of the same total size but different organizations.

Cache-Organization-1:
Total Size: 8 bytes
Block Size: 1 byte
Associativity: direct mapped
Replacement policy: LRU

Cache-Organization-2:
Total Size: 8 bytes
Block Size: 1 byte
Associativity: 2-way set associative
Replacement policy: LRU

Assume both caches are initially empty (i.e., contain no valid data).

Provide a reference stream of no more than 4 references that exhibits a "higher" miss rate on cache-2 than it does on cache-1. Each element of the reference stream should be a byte address. Indicate next to each reference whether it misses or hots in each cache.

`Answer:`

	Org-1	Org-2
LD 0	miss	miss
LD 4	miss	miss
LD 12	miss	miss
LD 0	hit	miss

31

Question 6 - System Performance (20 points)

Consider the following 2 systems.

System-1:
200 MHz processor
Two load delay slots (first filled 75% and second filled 20% of the time)
256 KB data cache (miss rate 5%, miss penalty 25 cycles)

System-2:
150 MHz processor
One load delay slot (filled 75% of the time)
8KB first-level data cache (miss rate 10%; miss penalty 7 cycles if data in level-2 cache and 32 cycles if data must be fetched from main memory)
1 MB second-level data cache (global miss rate 2%)

Now consider the following workload running on both systems:
Loads: 25%
Stores: 10%
Other: 65%

(a) (14 points) Ignoring instruction-cache effects, determine which system is faster for the given workload. Assume that neither system has any write buffers. State any other assumptions that you make.

Answer:

```
System-1 CPI  = 0.65 + 0.25 (1 + .05*25 + .25 + .80) + 0.1 (1 +
                  .05 * 25)
              = 1.7
              This implies performance = 200/1.7 = 118 MIPS

System-2 CPI  = 0.65 + 0.25 (1 + 0.1 (.98*7 + .02*32) + 0.25) +
                  0.1 (1 + 0.1(.98*7 + .02*32))
              = 1.325
              This implies performance = 150/1.325 = 113 MIPS
```

Thus system-1 is faster in this case.

(b) (3 points) Assume system-2's clock rate is increased to 200 MHz, but all other parameters remain the same. Now which system is faster?

Answer: System-2's performance is now 200/1.325 = 151 MIPS. Thus system-2 will be faster if the same clock rate as system-1 can be achieved.

(c) (3 points) Assume an infinitely-deep write buffer is added to system-2 inbetween the processor and the first-level cache, but all other parameters remain the same (system-2's clock rate is 150 MHz). Now which system is faster?

Answer: The implication is that all write-stalls will now be hidden by the write buffer. Thus:

System-2 CPI = 0.65 + 0.25 (1 + 0.1 (.98*7 + .02*32) + 0.25)
= 1.15
This implies performance = 150/1.15 = 130 MIPS, which is faster than system-1 at 200 MHz.

33