COMP25212

Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

System Architecture

Date: Tuesday 31st May 2016

Time: 09:45 - 11:45

Please answer any THREE Questions from the FOUR questions provided

Use a SEPARATE answerbook for each SECTION

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text

[PTO]

Section A

1. **Caches**

a) What are the three categories of cache misses? Explain the cause in each case. (3 marks)

b) Explain the effect of either increasing or decreasing each of the following cache parameters on each of the three previous cache miss categories. You should only consider the effect of modifying one parameter at a time (NOT combined effects). (5 marks)

- cache size
- line size
- associativity

c) Describe the two forms of locality exploited by caches and give at least one example for each type of locality occurring in programs. (4 marks)

d) Consider a CPU with a single level of cache such that the cache hit ratio is **h**, the cache access time (hit time) is $\mathbf{t_h}$ and the main memory access time (miss time) is $\mathbf{t_m}$.

1) What is the perceived average memory access time **T** for this CPU ? *[Note: You are expected to provide a symbolic expression and explain your reasoning.]* (3 marks)

2) What is the ***maximum speedup*** S_{max} of the memory system that could be achieved with a single level of cache if the hit ratio **h** is fixed but the hit time $\mathbf{t_h}$ could be reduced to 0? (2 marks)

3) Evaluate this maximum speedup S_{max} when **h** is 90%, 95% and 98%. *[Note: you are expected to provide numerical answers.]* (1 mark)

4) Consider that the main memory access time is $\mathbf{t_m}$ = 200 CPU clock cycles. What is the average perceived memory access time T_{min} in the three hit ratio scenarios in d3) if $\mathbf{t_h}$ can be reduced to 0 ? (1 marks)

5) What could be done to improve the performance of this memory system? (1 mark)

2. **Storage**

a) A server hard disk drive (HDD) is specified as having a mean "seek time" of 4 milliseconds, a "rotation speed" of 10,000 revolutions per minute and a "transfer rate" of 200 megabytes per second.

1) Explain what each term in this specification means for the performance of disk operation. (3 marks)

2) How long on average would a transfer of 8 kilobytes take from a random position on the disk? And for 8 megabytes? Assume the data to transfer is contiguous in the device. (4 marks)

b) A solid state drive (SSD) is specified as having a streaming read transfer rate of 500 megabytes per second, a streaming write transfer rate of 250 megabytes per second, a random 4 kilobytes read latency of 12 microseconds and a random 4 kilobytes write latency of 24 microseconds.

1) Considering the HDD from the previous question, approximately how much faster would this SSD be to read 8 kilobytes from a random position? And 8 megabytes? (2 marks)

2) Same question for writing 8 kilobytes and 8 megabytes. (2 marks)

3) Could multiple HDDs be used to bridge this performance gap with the SSD? Explain for each of the two scenarios above how this can be achieved or why it cannot. (3 marks)

c) Explain two main reasons, aside from increasing storage capacity, why more than one disk drive may be used in a single system and how these goals may be achieved. (6 marks)

[PTO]

Section B

3. **Pipelining**

a) What is pipelining in the context of processor design? What benefits does it provide? What issues arise from this architecture? (6 marks)

b) Draw the data-dependency graph for the following application. (4 marks)

```
MUL R2, R0, R1
MUL R3, R1, R1
MUL R4, R0, R0
SUB R5, R1, R0
CMP R5, #0
ADD R6, R4, R3
ADD R7, R6, R2
DIV R8, R7, R5
```

c) Based on that dependency graph, discuss how suitable that code is for being executed in a superscalar processor (assume all arithmetic operations can be done in a single clock cycle). (2 marks)

d) Explain the benefits and limitations of reordering instructions in the compiler or in hardware. (4 marks)

e) What needs to be changed in an in-order processor to transform it into an out-of-order processor? (4 marks)

4. **Multithreading / Multicore**

a) Explain cache thrashing and false sharing and give examples of when they can occur. (4 Marks)

b) Define what is meant by superscalar, multithreading and multicore, three techniques we can use to exploit parallelism in hardware. (6 Marks)

c) We are involved in the purchase of a high-performance server for a SME that works on data analytics. We have obtained the following specifications for 3 different configurations:

	Config 1	**Config 2**	**Config 3**
# of cores	8	6	8
Multithreading	No	4-way fine-grain	2-way SMT
Superscalar	4-way	4-way	2-way
Clock freq.	1.5 GHz	2.5 GHz	3 GHz

Assuming that the 3 are priced at a similar range and that memory and cache systems are similar, which one you will choose (and why – *please give numeric answers*) if the main concern of the development team was increasing: (8 marks)

1) single-thread peak performance

2) peak IPC of the system (instructions per cycle)

3) peak computing throughput of the system (instructions per second)

4) the total number of hardware threads supported by the system

d) What is 'cache coherence' in the context of multicore processing systems? Why is it important? (2 marks)

END OF EXAMINATION