國立高雄大學資訊工程學系 計算機結構期末考考卷

1. (16%) Please discuss the following compiler optimization techniques that can improve the cache performance. You can give an example code and explain it
   - **A. *Merging Arrays*: improve spatial locality by single array of compound elements vs. 2 arrays**
   - **B. *Loop Interchange*: change nesting of loops to access data in order stored in memory**
   - **C. *Loop Fusion*: Combine 2 independent loops that have same looping and some variables overlap**
   - **D. *Blocking*: Improve temporal locality by accessing "blocks" of data repeatedly vs. going down whole columns or rows**

2. (12%) There are numerous techniques for improving the performance of caches. Some reduce the frequency of misses, some reduce the miss penalty, some reduce hit times, and so on. Many optimizations involve trade-offs, making one performance factor worse in return for reducing another. In some cases, the so-called optimization can actually hurt performance. For the following optimizations, briefly summarize the trade-offs involved and how it helps or can hurt performance. Be as specific as possible, for example, if an optimization reduces conflict misses, say so, rather than just say "it reduces misses".

   A. Higher associativity
   - i. Reduce conflict misses
   - ii. Increase hit time

   B. **Larger block size**
   - i. Take advantage of spatial locality
   - ii. Reduce compulsory misses
   - iii. Increase the miss penalty

   C. **Bigger caches**
   - i. Reduce capacity misses
   - ii. Potentially longer hit time of the larger cache memory
   - iii. Higher cost and power

3. (4%) Please write the equation of the average memory access time.

$$\text{Average Memory Access Time} = \text{Hit Time} + \text{Miss Rate} \times \text{Miss Penalty}$$

4. (20%) Assume there are two processors in a system which uses the directory to solve conference problems. A1 and A2 map to the same cache block and their initial values are equal to 0. Please describe the related procedures of different parts in each step.

Processor 1 Processor 2 Interconnect Directory Memory

	P1			P2			Bus				Directory			Memory
step	State	Addr	Value	State	Addr	Value	Action	Proc.	Addr	Value	Addr	State	{Procs}	Value
P1: Write 10 to A1														
P1: Read A1														
P2: Read A1														
P2: Write 20 to A1														
P2: Write 40 to A2														

5. (10%) Please explain what are "Fine-Grained Multithreading", and "Course-Grained Multithreading", respectively.

Fine-Grained Multithreading

i. Switches between threads on each instruction, causing the execution of multiples threads to be interleaved

ii. Usually done in a round-robin fashion, skipping any stalled threads

iii. CPU must be able to switch threads every clock

Course-Grained Multithreading

- Switches threads only on costly stalls, such as L2 cache misses

6. (8%) In parallel processing, there are two challenges. What are they?

A. Limited parallelism available in programs

B. long latency to remote memory

7. (5%) What is the cache coherence problem?

Two different processors can have two different values for the same location.

8. (5%) Please explain why snooping and directory schemes adopt “write invalidate protocol” instead of “Write update protocol”.

   Write update protocol consumes considerably more bandwidth

9. (8%) Conference miss can be divided into two categories, i.e., true sharing misses and false sharing misses. What is “false sharing miss”? Please provide a method to eliminate all false sharing misses.

   False sharing misses when a block is invalidated because some word in the block, other than the one being read, is written into

   i. Invalidation does not cause a new value to be communicated, but only causes an extra cache miss

   Block is shared, but no word in block is actually shared

   ⇒ miss would not occur if block size were 1 word

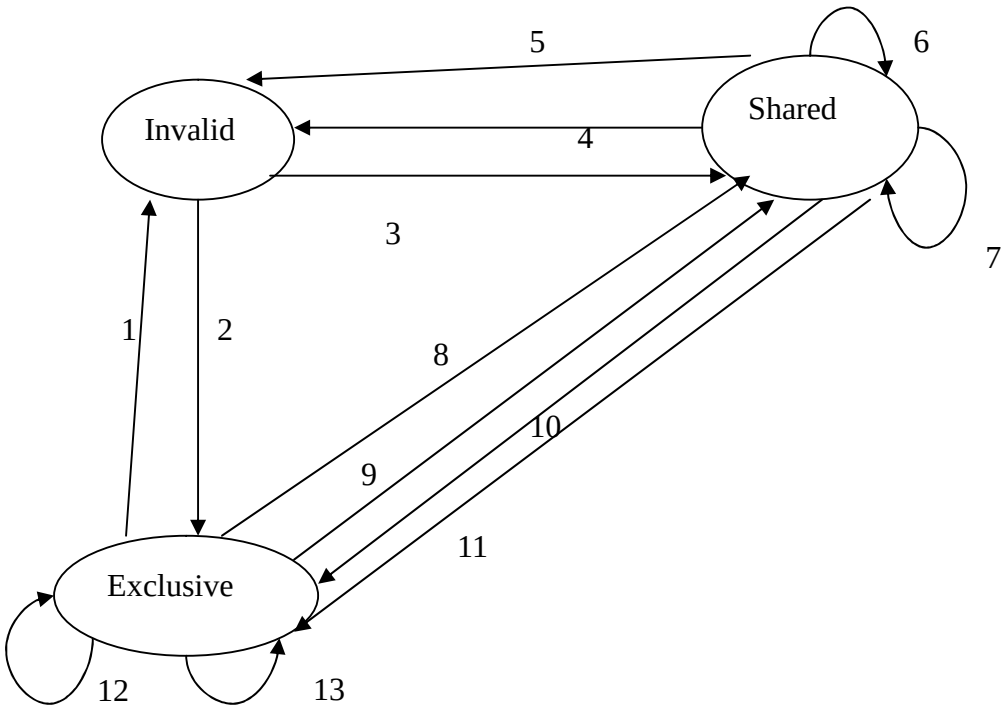
10. (5%) How does snooping scheme enforce serialization of write accesses (write serialization)?

    By broadcast medium. If two processors attempt to write shared blocks at the same time, their attempts to broadcast an invalidate operation will be serialized when they arbitrate for the bus.

11. (12%) There are numerous techniques for improving the performance of caches. Please describe the main idea behind the following techniques, i.e., nonblocking caches, **Merging Write Buffer**, critical word first.

    A. *Non-blocking cache*
       i. allows data cache to continue to supply cache hits during a miss
    B. Merging Write Buffer
       i. If the buffer contains other modified blocks, the addresses can be checked to see if the address of this new data matches the address of a valid write buffer entry.
       ii. If so, the new data are combined with that entry.
    C. *Critical Word First*—Request the missed word first from memory and send it to the CPU as soon as it arrives; let the CPU continue execution while filling the rest of the words in the block

12. (10%) A write-back cache use the write-invalidate cache coherence protocol. In the snooping protocol, the cache controller must maintain the state transitions for each cache block. Please complete the state transition diagram for processor requests and bus requests. Please write the request and the corresponding function for each edge in the following diagram. (one point for each row)



Index	Request	Function
1	Bus write miss	Write back block, abort memory access
2	CPU write miss	Place write miss on bus
3	CPU read miss	Place read miss on bus
4	Bus Write miss	X
5	Bus invalidate	X
6	CPU read hit	Read data in cache
7	CPU read miss	Place read miss on bus
8	CPU read miss	Write back block, Place read miss on bus
9	Bus read miss	Write back block
10	CPU write hit	Place invalidate on bus
11	CPU write miss	Place write miss on bus
12	CPU write hit/ CPU read hit	X
13	CPU write miss	Write back block, Place write miss on bus