國立高雄大學資訊工程學系 計算機結構期末考考卷

姓名： 學號：

1. (10%) Please compare branch prediction buffer and branch-target buffer, e.g., access time, contents.
   - A branch-*predication* buffer is accessed during the ID cycle, so that at the end of ID we know the branch – target address.
   - For a branch-*target* buffer, we access the buffer during the IF stage using instruction address of the fetched instruction.

2. (20%) 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".

Victim caches

+ Decreases conflict misses (and some capacity misses)

- May increase miss penalty

- Extra hardware

Software controlled prefetching

+ May decrease stalls due to demand misses

- Introduces explicit instruction overhead which may be higher than the benefit

- Requires software support

Increasing block size

+ Decreases compulsory and capacity misses by "prefetching" data

+ Increases capacity misses by decreasing number of unique blocks in cache

- Increases memory traffic

- May increase miss penalty

Higher associativity

- Decreases conflict misses by hashing data

- More complex replacement protocol

Non-blocking caches

+ Allows multiple outstanding cache misses, allowing misses to overlap
+ Increases memory system bandwidth
- Requires dynamically scheduled processor to fully exploit
- Complex hardware

3. (12%) Please explain the three types of cache misses.
   A. Compulsory
   B. Capacity
   C. Conflict (collision)

   What is "cache trash"?

- Compulsory

  The very first access to a block cannot be in the cache, so the block must be brought into the cache.

- Capacity

  If the cache cannot contain all the blocks needed during execution of a program, capacity misses will occur because of blocks being discarded and later retrieved.

- Conflict (collision)

  If the block placement strategy is set associative or direct mapped conflict will occur because a block may be discarded and later retrieved if too many blocks map to its set.

- Trash
  - If the upper-level memory is much smaller than what is needed for a program, and a significant percentage of the time is spent moving data between two levels in the hierarchy.

4. (8%) Please describe the advantages and disadvantages of conditional instruction. Give a simple code example to illustrate how they work.

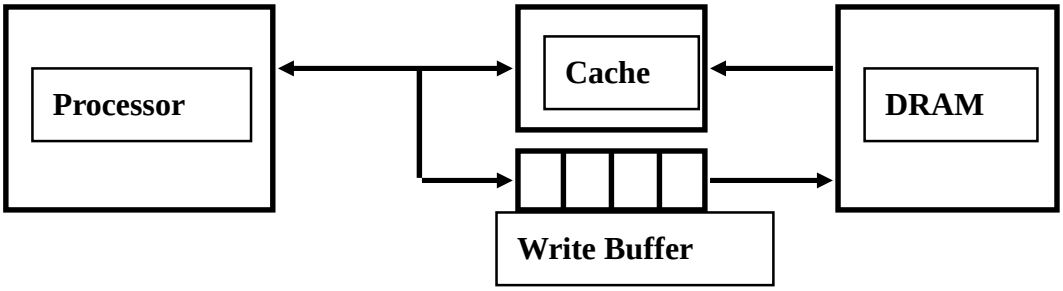
- Advantages of conditional instructions
  - Can eliminate simple branches
  - Can reduce code size
- Drawbacks to conditional instructions
  - Still takes a clock even if condition is false
  - Need the condition to be evaluated early to be useful
  - Could result in higher CPI or lower clock rate

– Could make control and datapath design complex

if (R1 == 0) R2 = R3
BNEZ R1, L
ADD R2, R3, R0

- Instead implement this using a conditional move instruction
  CMOVZ R2, R3, R1

5. (6%) Please explain how the write back strategy uses a write buffer to reduce write stall.



- A Write Buffer is needed between the Cache and Memory
  – Processor: writes data into the cache and the write buffer
  – Memory controller: write contents of the buffer to memory

6. (16%) Consider the following description of a computer and its cache performance:
   Block size = 1 word
   Memory bus width = 1 word
   Miss rate = 3%
   Memory accesses per instruction = 1.2
   Cache miss penalty = 64 cycles
   Average cycles per instruction (ignoring cache miss) = 2
   (a) Please compute the CPI for the original computer.
   (b) If we change the block size to 2 words, the miss rate falls to 2%. Please compute the CPI.
   (c) If we change the block size to 2 words, the miss rate falls to 2%. What is the improvement in performance of interleaving two ways and four ways versus doubling the bandwidth the width of memory and the bus?

   Assume that the performance of the basic memory organization is
   - 4 clock cycles to send the address
   - 56 clock cycles for the access time per word
   - 4 clock cycles to send a word of data

(a) 2 + (1.2*3%*64) =
(b) 2-word block
32-bits bus and memory, no interleaving = 2 + (1.2*2%*2*64) =
(c) 2-word block
Two-way 32-bits bus and memory, interleaving = 2 + (1.2*2%*2*(4+56+8))
Four-way and doubling 64-bits bus and memory, no interleaving = 2 + (1.2*2%*1*64)

7. (5%) Please explain what is 2:1 cache rule of thumb.

A. Miss Rate of direct mapped cache size N = Miss Rate 2-way cache size N/2
B. Hold for cache sizes less than 128 KB.

8. (12%) Please describe how Merging Arrays, Loop Interchange, Loop Fusion, and Blocking reduce cache misses, respectively.

Merging Arrays: improve spatial locality by single array of compound elements vs. 2 arrays
Loop Interchange: change nesting of loops to access data in order stored in memory
Loop Fusion: Combine 2 independent loops that have same looping and some variables overlap
Blocking: Improve temporal locality by accessing "blocks" of data repeatedly vs. going down whole columns or rows

9. (10%) What are Write Invalidate Protocol and Write Broadcast Protocol? Please compare the two protocols.

- Write Invalidate Protocol:
  - Multiple readers, single writer
  - Write to shared data: an invalidate is sent to all caches which snoop and *invalidate* any copies
  - Read Miss:
    - Write-through: memory is always up-to-date
    - Write-back: snoop in caches to find most recent copy
- Write Broadcast Protocol (typically write through):
  - Write to shared data: broadcast on bus, processors snoop, and *update* any copies
  - Read miss: memory is always up-to-date

Invalidate:

+ Multiple writes by the same processor to the cache block only require one invalidation
+ No need to send the new value of the data (less bandwidth)
– Caches must be able to provide up-to-date data upon request
– Must write-back data to memory when evicting a modified block

Usually used with write-back caches (more popular)

Update:

+ New value can be re-used without the need to ask for it again
+ Data can always be read from memory
+ Modified blocks can be evicted from caches silently
– Possible multiple useless updates (more bandwidth)

Usually used with write-through caches (less popular)

10. (14%) What is cache coherence? Please complete the cache coherence diagram of an example snooping protocol and the state transition diagram of an example directory protocol.