資訊工程學系 計算機結構

姓名： 學號：

[Limits to ILP] (24%)

1. (6%) Please compare Hardware Speculation and Software Speculation.
   A. Hardware-based speculation maintains a completely precise exception model even for speculated instructions.
   B. Hardware-based speculation does not require compensation or bookkeeping code, which is needed by ambitious software speculation mechanisms.
   C. Hardware-based speculation with dynamic scheduling does not require different code sequences to achieve good performance for different implementations of architecture.
   D. The major disadvantage of supporting speculation in hardware is the complexity and additional hardware resources required.
2. (9%) Please describe Fine-Grained Multithreading and its corresponding advantages and disadvantages.

It switches between threads on each instruction, causing the execution of multiples threads to be interleaved. It is usually done in a round-robin fashion, skipping any stalled threads. Besides, CPU must be able to switch threads every clock.

- **Advantage is it can hide throughput losses that arise from both short stalls, since instructions from other threads executed when one thread stalls**
- **Disadvantage is it slows down execution of individual threads, since a thread ready to execute without stalls will be delayed by instructions from other threads**

2. (9%) Please describe Coarse-Grained Multithreading and its corresponding advantages and disadvantages.

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

- **Advantages**
  - **Not need to have very fast thread-switching**
  - **Doesn't slow down thread, since instructions from other threads issued only when the thread encounters a costly stall**
- **Disadvantage is hard to overcome throughput losses from shorter stalls, due to pipeline start-up costs**

- **Since CPU issues instructions from 1 thread, when a stall occurs, the pipeline must be emptied or frozen**
- **New thread must fill pipeline before instructions can complete**

[Multiprocessors and Thread-Level Parallelism] (46%)

1. (6%) In the directory protocol, there are three kinds of nodes, i.e., Local node, Home node, and Remote node. What roles do the nodes play in the protocol?

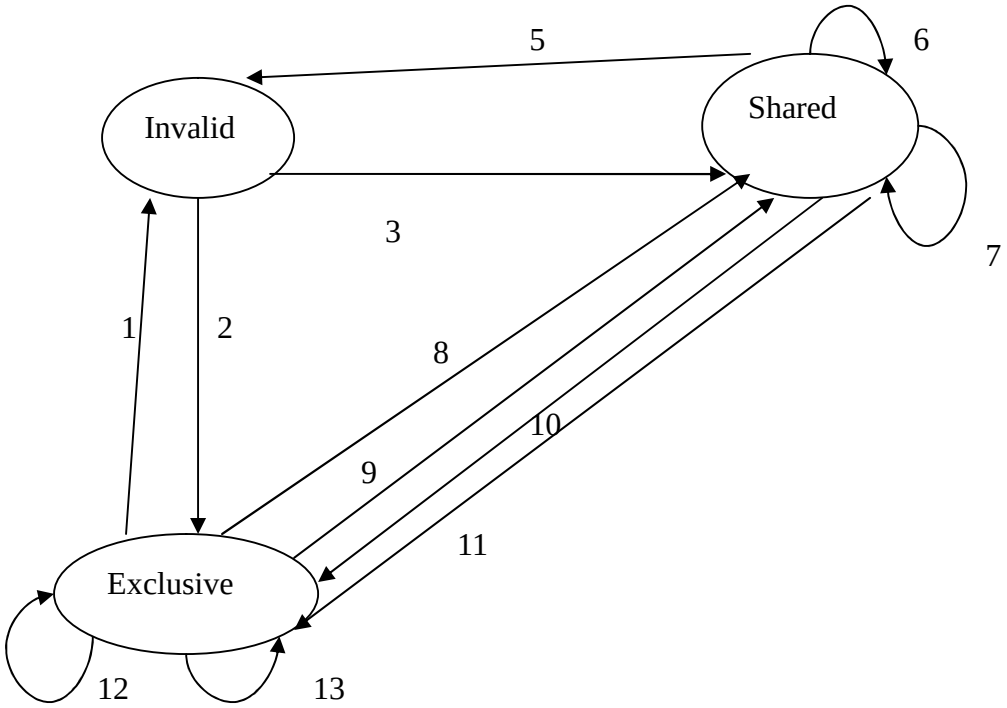
   - **Local node where a request originates**
   - **Home node where the memory location and the directory entry of an address reside**
   - **Remote node has a copy of a cache block, whether exclusive or shared**

2. (6%) What is multiprocessor Cache Coherence Problem?
   When the memory data item is held in the individual cache of two processor, two different values are saw.

3. (7%) 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

4. (11%) A write-back cache use the write-invalidate cache coherence protocol. In the **directory 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 type (Please note CPU request)	Function
1	Fetch/invalidate	Data write back
2	CPU write miss	Not listed
3	CPU read miss	Not listed
5	Invalidate	Not listed
6	CPU read hit	Not listed
7	CPU read miss	Not listed
8	CPU read miss	Data write back, place read miss on bus
9	Fetch	Data write back
10	CPU write hit	Send invalidate message
11	CPU write miss	Not listed
12	CPU write hit/ CPU read hit	 X
13	CPU write miss	Data write back

5. (10%) Assume there are two processors in a system which uses the **snooping protocol** to solve conference problems. A1 and A2 map to the same cache block and the initial values are 0. Please describe the related procedures of different parts in each step.

	P1			P2			Bus				Memory	
step	State	Addr	Value	State	Addr	Valu	Actior	Proc	Addr	Value	Add	Valu
P1 Write 10 to A1	Excl.	A1	10				WrMs	P1	A1			
P1: Read A1	Excl.	A1	10									
P2: Read A1				Shar.	A1		RdMs	P2	A1			
	Shar.	A1	10				WrBk	P1	A1	10	A1	10
				Shar.	A1	10	RdDa	P2	A1	10	A1	10
P2: Write 20 to A1	Inv.			Excl.	A1	20	Inv	P2	A1		A1	10
P2: Write 40 to A2							WrBk	P2	A1	20	A1	20
				Excl.	A2	40	WrMs	P2	A2		A1	20

6. (6%) Please compare Write invalidate protocol and Write update protocol.

A. Write invalidate protocol

i. To ensure that a processor has exclusive access to a data item before it writes that item.

ii. It invalidates other copies on a write.

B. Write update or write broadcast protocol:

iii. To updates all the cached copies of a data item when that items is written.

iv. Because a write update protocol must broadcast all writes to shared cache lines, it consumes considerably more bandwidth.

[Memory Hierarchy Design] (30%)

1. (12%) Please discuss the following compiler optimization techniques that can improve the cache performance. You can give an example code and explain it

- ***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**

2. (9%) 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.

甲、*Non-blocking cache*

1. allows data cache to continue to supply cache hits during a miss

乙、Merging Write Buffer

2. 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.
3. If so, the new data are combined with that entry.

丙、*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

3. (3%) 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. (6%) 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.

甲、 **Higher associativity**

i. Reduce conflict misses
ii. Increase hit time

乙、 **Larger block size**

i. Reduce compulsory misses
ii. Increase miss penalty

丙、 **Bigger caches**

i. Reduce capacity misses
ii. Potentially longer hit time