1. (11) **True or False:**

(1) Measuring performance on multiprocessors using linear speedup instead of execution time is a bad idea.

(1) Amdahl’s law applies to parallel computers.

(1) You can predict cache performance of Program A by analyzing Program B.

(1) Linear speedups are needed to make multiprocessors cost-effective.

(1) It is not necessary to simulate very many instructions in order to get an accurate performance measure of the memory heirarchy.

(1) Scalability is almost free.

(1) A program’s locality behavior is constant over the run of an entire program.

(1) Communication is not a significant problem for parallel processor systems.

(1) Memory Bandwidth is the most important thing when designing a memory system.

(1) The instruction set architecture does not impact the implementability of a virtual machine monitor.

(1) Professor Farrens never wore shorts to class.

2. (3) What is the goal of the memory heirarchy? What two principles make it work?

3. (12) What do the following acronyms stand for:

SMP TPC COMA

OLTP VMM SRAM

NUMA MPP DSM

MIMD SIMD SISD

For the next 6 problems, circle the correct answer.

4. (1) Which type of cache miss can be changed by altering the mapping scheme?

Capacity Coherence Compulsory Conflict

5. (1) Which type of cache miss can be reduced by using longer lines?

Capacity Coherence Compulsory Conflict

6. (1) Which type of operation is necessary in order to support synchronization?

Nuclear Atomic Radioactive

7. (1) Relaxing the requirement that Writes complete before Reads yields a model known as

Total Store Ordering Partial Store Ordering Weak Ordering

8. (1) Relaxing the requirement that Writes complete before Writes yields a model known as

Total Store Ordering Partial Store Ordering Weak Ordering

9. (1) Which type of operation is necessary in order to support synchronization?

Nuclear Atomic Radioactive

10. (7) Which of the following does the book list as techniques for reducing the Miss Rate? (Circle the correct answers)

Small and simple caches
Bigger caches
Trace caches
Multilevel caches
Non-blocking caches
Compiler optimizations
Priority to Read Misses
Merging Write Buffer
Hardware Prefetching

Larger block size
Way prediction
Higher associativity
Pipelined caches
Multibanked caches
Victim cache
Critical Word First/Early Restart
Avoiding Address Translation when Indexing Cache
Software prefetching

11. (4) Which of the following does the book list as techniques for reducing the Hit Time? (Circle the correct answers)

Small and simple caches
Bigger caches
Trace caches
Multilevel caches
Non-blocking caches
Compiler optimizations
Priority to Read Misses
Merging Write Buffer
Hardware Prefetching

Larger block size
Way prediction
Higher associativity
Pipelined caches
Multibanked caches
Victim cache
Critical Word First/Early Restart
Avoiding Address Translation when Indexing Cache
Software prefetching

12. (7) Which of the following does the book list as techniques for reducing the Miss Penalty? (Circle the correct answers)

Small and simple caches
Bigger caches
Trace caches
Multilevel caches
Non-blocking caches
Compiler optimizations
Priority to Read Misses
Merging Write Buffer
Hardware Prefetching

Larger block size
Way prediction
Higher associativity
Pipelined caches
Multibanked caches
Victim cache
Critical Word First/Early Restart
Avoiding Address Translation when Indexing Cache
Software prefetching

Short Answers:

13. (5) What is a victim cache, and how does it work?

14. (8) Describe the difference between shared memory and message passing machines. Include the impact on design, cost, and programming model.

12. (8) What is Cache Coherence, and why is it necessary? Snooping is one main approach to providing coherence - state what the other main approach is, and briefly outline how each of them work.

15. (4) Flynn broke parallel machines into 4 different classifications. Which of the four is most flexible? Why?

16. (5) What pair of instructions are used to implement a lock in RISC systems (as described in the text)? Describe how this pair works together in order to accomplish the goal.

17. (4) Assume a relatively large fully associative write-back cache that contains no valid data. Given the following sequence of 5 memory operations (the address of the operation is in the square brackets):

WriteMem[100]
ReadMem[100]
WriteMem[100]
WriteMem[200]
WriteMem[100]

What are the number of hits and misses when using write allocate versus no-write allocate?

18. (8) Suppose that in 1000 memory references there are 40 misses in the first-level cache and 20 misses in the second-level cache. What are the various miss rates? In addition, assume the miss penalty for the L2 cache is 200 clock cycles, the hit time of the L2 cache is 10 clock cycles, the hit time of the L1 cache is 1 clock cycle, and there are 1.5 memory references per instruction. What is the average memory access time in cycles? What is the average stall cycles per instruction? (Ignore writes.)

19. (8) Assume that L2 has a block size four times that of L1. Show how a miss for an address that causes a replacement in L1 and L2 can lead to a violation of the inclusion property.