

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

- (1) DRAM and Disk access times are rapidly converging.
- (1) Measuring performance on multiprocessors using linear speedup instead of execution time is a good idea.
- (1) Amdahl's law applies to parallel computers.
- (1) You can predict cache performance of Program A by analyzing Program B.
- (1) Computer components fail suddenly, with little warning.
- (1) Linear speedups are not 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 hierarchy.
- (1) A program's locality behavior is constant over the run of an entire program.
- (1) Communication is 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 impacts the implementability of a virtual machine monitor.

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

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

OLTP

VMM

SRAM

SMP

TPC

COMA

NUMA

MPP

DSM

4. (4) What are the four classifications of cache misses?

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

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

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

Circle the correct answer:

8. (1) Relaxing the requirement that Writes complete before Reads yields a model known as
Total Store Ordering Partial Store Ordering Weak Ordering

9. (1) Relaxing the requirement that Writes complete before Writes yields a model known as
Total Store Ordering Partial Store Ordering Weak Ordering

10. (1) Which type of operation is necessary in order to support synchronization?
Nuclear Atomic Radioactive

11. (1) Which benchmarks are most affected by the window size?
Integer Float Equally affected

12. (1) Which benchmarks are most affected by the accuracy of the branch predictor?
Integer Float Equally affected

13. (1) Alias analysis has the most impact on which benchmark?
Integer Float Equally affected

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

- | | |
|-------------------------|--|
| Small and simple caches | Larger block size |
| Bigger caches | Way prediction |
| Trace caches | Higher associativity |
| Multilevel caches | Pipelined caches |
| Non-blocking caches | Multibanked caches |
| Compiler optimizations | Victim cache |
| Priority to Read Misses | Critical Word First/Early Restart |
| Merging Write Buffer | Avoiding Address Translation when Indexing Cache |
| Hardware Prefetching | Software prefetching |

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

- | | |
|-------------------------|--|
| Small and simple caches | Larger block size |
| Bigger caches | Way prediction |
| Trace caches | Higher associativity |
| Multilevel caches | Pipelined caches |
| Non-blocking caches | Multibanked caches |
| Compiler optimizations | Victim cache |
| Priority to Read Misses | Critical Word First/Early Restart |
| Merging Write Buffer | Avoiding Address Translation when Indexing Cache |
| Hardware Prefetching | Software prefetching |

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

- | | |
|-------------------------|--|
| Small and simple caches | Larger block size |
| Bigger caches | Way prediction |
| Trace caches | Higher associativity |
| Multilevel caches | Pipelined caches |
| Non-blocking caches | Multibanked caches |
| Compiler optimizations | Victim cache |
| Priority to Read Misses | Critical Word First/Early Restart |
| Merging Write Buffer | Avoiding Address Translation when Indexing Cache |
| Hardware Prefetching | Software prefetching |

Short Answers:

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

18. (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.

19. (6) What is an instruction window? How does it impact ILP?

20. (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.

21. (7) What is simultaneous multithreading? What characteristics of multi-issue processors is this trying to take advantage of? (Be relatively detailed in your answer)

22. (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?

23. (4) Suppose you want to achieve a speedup of 80 with 100 processors. What fraction of the original computation can be sequential? (Full credit if you show your work, half-credit if you just write the answer.)

24. (4) 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.