Very Short Answer:

(1) (1) True or False: You can predict cache performance of Program A by analyzing Program B.

(2) (2) What is the goal of the memory heirarchy?

(3) (2) The memory heirarch works because of what two principles?

(4) (4) List 4 of the 5 Miss Penalty Reduction Techniques.

(5) (4) What are the 4 types of parallel processors, according to Flynn?

(6) (1) What does COMA stand for?

(7) (2) What does UMA and NUMA stand for?

Short Answer:

(1) (2) Briefly describe Interleaved Memory and how it works.

(2) (2) What is a victim cache, and how does it work?

(3) (2) What is the goal of the Virtual Memory system?

(4) (2) What is Cache Coherence, and why is it necessary?

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

(6) (12) Dopey is a computer which has a CPI of 1.0 when all memory accesses hit in the cache. The only data accesses are loads and stores, which total 50% of the instructions. If the miss penalty is 25 clock cycles and the miss rate is 2%, how much faster would Dopey be if all instructions were cache hits?

(7) (12) Some memory systems handle TLB misses in software (as an exception), while others use hardware for TLB misses.
What are the tradeoffs between these two methods for handling TLB misses?

Will TLB miss handling in software always be slower than TLB miss handling in hardware? Why or why not?

Are there page table structures that would be difficult to handle in hardware, but possible in software? Are there any such structures that would be difficult for software to handle but easy for hardware to manage?

(8) (12) 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]
WriteMem[100]
ReadMem[200]
WriteMem[200]
WriteMem[100]

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

(9) (12) Suppose you want to achieve a speedup of 80 with 100 processors. What fraction of the original computation can be sequential?

(10) (12) Assume that words A and B are in two different locations in the same cache block, which is in the shared state in the caches P1 and P2. In the following sequence of events, identify each miss as either a true sharing miss, a false sharing miss, or a hit. (Any miss that would occur if the block size were one word is referred to as a true sharing miss.)

Time	P1	P2
1	Write A	
2		Read B
3	Write A	
4		Write B
5	Read B	

For example, the event at time 1 is a true sharing miss, since A was read by P2 and needs to be invalidated from

(11) (12) Suppose there are 10 processors on a bus that each try to lock a variable simultaneously. Assume that each bus transaction (read miss or write miss) is 100 clock cycles long. You can ignore the time of the actual read or write of a lock held in the cache, as well as the time the lock is held. Determine the number of bus transactions required for all 10 processors to aquire the lock, assuming they are all spinning when the lock is released at time 0. About how long will it take to process the 10 requests? Assume that the bus is totally fair so that every pending request is serviced before a new request and that the processors are equally fast.