CSC506 Homework due Friday, 5/28/99 - Cache questions & VM

Question 1. What advantage does a Harvard cache have over a unified cache?

Question 2. Why would you want the system I/O to go directly to main memory rather than through the processor cache?

Question 3. How long does it take to access any random word from Synchronous DRAM that is rated at 10 ns?

Question 4. If an 8-way set-associative cache is made up of 32 bit words, 4 words per line and 4096 sets, how big is the cache in bytes?

Question 5. What is the shortest time it would take to load a complete line in the above cache using fast page mode DRAM that has a RAS access time of 50 ns, a CAS access time of 13 ns, a cycle time of 95 ns and a fast page mode cycle time of 35 ns?

Question 6. What is the shortest time it would take to load a complete line in the above cache using EDO DRAM that has a RAS access time of 50 ns, a CAS access time of 13 ns, a cycle time of 84 ns and an EDO cycle time of 20 ns?

Question 7. What is the shortest time it would take to load a complete line in the above cache using synchronous DRAM that requires 5 clock cycles from RAS to the first data out and is clocked at 100 MHz?

Question 8. If the memory bus speed is 50 MHz, which of the above three DRAMs would you use and why?

Question 9. If a memory system consists of a single external cache with an access time of 20 ns and a hit rate of 0.92, and a main memory with an access time of 60 ns, what is the effective memory access time of this system?

Question 10. We now add virtual memory to the system described in question 9. The TLB is implemented internal to the processor chip and takes 2 ns to do a translation on a TLB hit. The TLB hit ratio is 98%, the segment table hit ratio is 100% and the page table hit ratio is 50%. What is the effective memory access time of the system with virtual memory?

Question 11. LRU is almost universally used as the cache replacement policy. Why?

Question 12. In an *n*-way set-associative cache, it is preferable to start a read to all lines in a set in parallel, even though only one line, at most, will be used. Why is it reasonable to do this?

Question 13. In contrast to starting access to all ***n*** lines of an ***n***-way set associative cache even though we know we won't use them all, an access to main memory is not started until we know we have a cache miss. Why don't we start the main memory access in parallel with searching the cache so we can overlap part of the access time?

Question 14. Stone says that a simple rule of thumb is that doubling the cache size reduces the miss rate by roughly 30%. Given that the cache in question 9 is 256K bytes, what is the expected percentage improvement in the effective access time if we double the cache size to 512K bytes?

Question 15. What is the expected percentage improvement in the effective access time over that in question 14 if we double the cache size again to 1024K bytes?

Question 16. What is the advantage of using a write-back cache instead of a write-through cache?

Question 17. What is a disadvantage of using a write-allocate policy with a write-through cache?

Question 18. What is "bus-snooping" used for?

Question 19. You find that it would be very inexpensive to implement small, direct-mapped cache of 32K bytes with an access time of 30 ns. However, the hit rate would be only about 50%. If the main memory access time is 60 ns, does it make sense to implement the cache?

Question 20. Would it make a difference on question 19 if we doubled the cache size to 64K bytes?