Lecture #21 Worksheet, Answer Master

Fill in blanks to answer questions below. Then email this sheet to your TA.

1. **Early in the history of computers, memory speed was not a problem. Why is data access speed from memory such a large issue today?**

   Today, computer CPU processing cycles are generally quite a bit faster than memory (that is, DRAM) access times. This creates a problem in that if memory access is not speeded up, the CPU is slowed down awaiting data.

2. **Another problem with memories is that working memory must be very fast, but there must also be a large amount of archival storage. State the dichotomy between these two requirements.**

   Fast memory is expensive (increasing the cost of the computer), but storage (archival) memory is very slow.

3. **Computer designers balance these needs by using very fast ("cache") memory to speed up processing. Summarize the three memory requirements for a modern computer.**
   a. **There must be enough fast memory (cache) to avoid slowing down the CPU.**
   b. **There must be sufficient DRAM to avoid having to access bulk memory (disk or SSM)**
   c. **There must be enough bulk storage to archive all necessary working databases and files.**

4. **Thus the modern computer uses multiple types of memory to assure that the computer (a) can process as rapidly as possible, and (b) has plenty of archival memory for storing large files. State the types of normal memory types used, versus the speed given, using the hints below.**

   a. **Registers (in the ALU, superfast)**
   b. **L1/L2 cache (very near the ALU, very, very fast).**
   c. **L3 cache (still on the ALU die, near the ALU, very fast).**
   d. **DRAM (not on the CPU chip, nearby, fast).**
   e. **Disk or solid state memories (bulk storage, relatively slow).**
   f. **Flash drives or CD's/DVD's (truly archival storage).**

5. **L1, L2, and L3 caches are the "fast" memories used to help with CPU speed, augmenting DRAM speed). In general, concerning speed, L1 > L2 > L3. What one characteristic gives each type of cache its relative speed advantage?**

   Distance to the ALU governs speed, L1 closest, L2 next, L3 farthest away.

6. **Registers are the fastest memory elements, adjacent to the ALU. What is the basic component of a register.**

   Registers are made up of D flip-flops.

7. **Random-access memory is fast electronic memory. There are two kinds, static ram (SRAM) and dynamic ram (DRAM). What is the basic component of SRAM (also referred to as "cache").**

   Cache or SRAM is, like the register block, composed of D flip-flops.

8. **State the relative sizes (in Kbytes or Mbytes) of L1, L2, and L3 cache.**

   L1 cache is typically small, 64 Kbytes or so. L2 cache is larger in most CPU's, typically 0.5-1.- Mbyte. L3 cache is 8-25 Mbytes, typically shared by the 4-20+ CPU's on a chip.

9. **On slides 15 and 16, note the relative location of L1, L2, and L3 caches to the processing elements. Based on that estimate, estimate the speed ratios of the cache elements.**

   Exact numbers are hard to estimate, but a cursory examination indicates that for the three cache types, L1 speed = 2× L2 speed = 4× L3 speed, based on proximity to the ALU elements.

10. **Cache is very fast and helps the CPU run faster. What are the two reasons that more cache is not used?**

    Cache memory uses lots of power, so it runs hot and would require a bigger power supply and CPU cooling. It is also very expensive compared to DRAM.

11. **DRAM memory, while quite fast in terms of access, is far slower than cache. Other than Rambus Corporation, which makes very-high-performance DRAM systems, most DRAM providers have memory that is several times**

slower than cache. This is because DRAM is NOT made of flip-flops, but of single-transistor memory cells that consume a very small amount of power and are very efficient storage devices. They have one disadvantage, which is spelled out in their name. What is it?

DRAM memory, based on capacitors, leaks the charge stored in the capacitor over time. It thus "loses its memory," and must be periodically refreshed.

12. **Based on slides 21-25, describe <u>briefly</u> how a DRAM cell works.**

    The transistor is turned on by activating the word line, voltage is applied to the bit line, and the transistor is charged to a one. If a zero is desired, the transistor is turned on, and the charge is drained off the capacitor.

13. **How is data read from the DRAM?**

    The transistor is turned on and the output sent to another circuit. If current then flows, a 1 is present. If no current flows, a 0 is present.

14. **If a DRAM cell "1" is read, what must then occur?**

    The "1" must be rewritten, similar to a refresh phase.

15. **What happens during a refresh cycle?**

    Each bit is "read." If it was a 1, it is rewritten. If a 0, no action is required.

16. **Answer the questions on slide 28, then check your answers on slide 29. Note that this exercise will help you answer the 35 questions which are the basis of your Test 2 bonus problem.**

17. **Although hard disk drives are slowly phasing out, their relative inexpensiveness and long life make them still viable for many computers. What is a hard disk made of, and what is it usually coated with?**

    Hard disks are made of aluminum and typically coated with nickel.

18. **From slide 31, data is written on the disk by a stationary (and very tiny) magnetic coil suspended over the spinning surface. What is the circular line of data called?**

It is called a track.

19. **From slides 32-35, what is the metal device called that holds the recording device, or "head?"**

    It is called the positioning arm.

20. **How is the head positioned?**

    The positioning arm pivots and moves in an arc, positioning the head.

21. **Are multiple disks common in a disk drive?**

    Yes.

22. **The disk case is rigid to protect the drive and heads. How does the unit protect the disk or heads from sudden jarring, which might cause problems?**

    The unit is suspended on springs or soft, flexible material to protect it.

23. **Data retrieval from a disk is slow, since it is a mechanical device. Why is the so-called "seek time" so long, and why is "rotational time" second in length?**

    The seek time is the time to move the positioning arm to the correct track, which can take milliseconds. The rotational time can be anything from a very short time (head got to track just as correct data appeared beneath it) or quite long (head just missed start of data, so disk must rotate a full rotation for the data to be available).

24. **There are many other storage media, mostly magnetic storage. What are the two slowest?**

    Magnetic tapes, which must be retrieved, mounted in a tape player, and then read, and floppy disks, which also must be retrieved and put into a reader.

25. **Solid state drives, sometimes called flash drives, are chips that are much slower than DRAM, but much faster than magnetic or optical drives. Study the information about SSD's on slides 39-43. What are the (1) largest and (2) cheapest SSD's available at present?**

a. **Largest: Nimbus ExaDrive—100 Tbytes.**
b. **Cheapest: The 240 Gbytes Toshiba TR200--$89.**

26. The memory hierarchy in a modern computer mixes types and speeds to give the best performance. It makes use of the principles of temporal locality and spatial locality. State these principles:

a. **Principle of temporal locality—Recently-used instructions and data in the computer will probably be reused.**
b. **Principle of spatial locality—Recently-accessed instructions and data in an area of memory will probably be accessed again.**

27. The hierarchy principle: The closer to the ALU the memory is placed, the faster it must be. Study the diagram on slide 46 and note the relative speed of the memories.

28. Using cache memory is the key to speeding up the CPU. Use slides 44-48 for the following questions.

29. Explain the "shuffling" technique used with cache to speed up the CPU.

Required information from DRAM is shuffled through the cache units, moving closer to the CPU, as it is required.

30. How do the two principles discussed above aid in placing the instructions/data from DRAM in the much smaller space of cache, since cache is so limited?

Since code/data recently used is often reused, and as code/data in one area of memory is often used again (typically in the same program), these two indicators help the cache units to move likely-needed instructions and data from DRAM to cache.

31. As a small amount of cache, say 1 Kbyte, might represent all the available storage for 1 Mbyte of DRAM, how does the CPU know when it accesses that Kbyte of cache that the desired 1 Kbyte from DRAM is correct?

The cache unit uses "validity indicators" to assure the CPU that the correct data is resident in that cache area.

32. Slide 52 shows the cache management unit. What is it sometimes referred to?

It is referred to as the "translation look-aside buffer."

33. **Do exercise 2 on slide 54, then check your answers on slide 55.**

34. **Note: slides 56-99 are provided to give students an idea of "what is going on" in the computing world at present, and what to expect in the near future. Please peruse the slides at your leisure, taking the time to understand where our "world of computing" is headed. These slides may tell you a bit about what your job may be like in the future! You will not be graded on this material, but you will most surely want to understand, as electrical or computer engineers, how these developments will affect all of us in the coming years.**