

All homework in this course should be done individually. Violation of this rule may result in a zero grade being assigned for the homework, or if flagrant or repeated an F for the course.

- 1). What is meant by “hitting a wall” as it is described in [McKee95]. Do you think that this technological wall has been overcome? Why or why not? Do you see any imminent “walls” limiting the performance of microprocessor systems?
- 2). The memory hierarchy of a modern workstation appears to be a single unified address space to the program. Explain why the memory hierarchy of a modern workstation - with many levels and complexities - is constructed in a complex pyramid encompassing multiple memory technologies.
- 3). Provide (3) goals of virtual memory.
- 4). Suppose that you have a two-level cache, L1, and L2. The hit rate for the L1 cache is 95% and the hit rate for the L2 cache is 80%. A memory reference which hits in the L1 cache takes 1 cycle, a memory reference which misses in L1 but hits in L2 takes 8 cycles, and a reference which misses in both and accesses DRAM takes 120 cycles. What is the average memory access time, assume that accesses to the three memory levels are strictly serial?
- 5). Locality
  - a) Describe in words, or with pseudo-code, a program that would exhibit a high degree of temporal locality, but relatively little spatial locality, with respect to its data references.
  - b) Describe in words, or with pseudo-code, a program that would exhibit a high degree of spatial locality, but very little temporal locality, with respect to its data references.
- 6). The following is a string of accesses to a direct mapped L1 cache with a 16 byte linesize and 32 lines. Classify each access as a Hit or one of the four following types of misses: Compulsory, Capacity, Conflict or Coherence.
  - 0x000
  - 0x008
  - 0x080
  - 0x200
  - 0x008

0x208

0x090

- 7). Some cache enhancements are more appropriate to either I or D side accesses. For each of the following state whether they are effective for I, D or Both and justify your answer.
- Non-Blocking
  - Stream Buffer
  - Trace Cache
- 8). The case has been made that cache inclusion is automatically maintained in a two-level cache hierarchy, where L1 is closer to the processor than L2, iff  $a_2 \geq a_1$  and  $b_2 \geq b_1$ , where  $a_n$  is the associativity of cache level  $n$ , and  $b_n$  is the block size of cache level. For this statement to hold true, what else must be true of the controller and replacement policy of these caches.
- 9). Explain how it is possible for memory accesses to appear at the DRAM controller “out-of-order” when compared to the sequential program ordering. What functional units are necessary in the processor for this to occur, and what role do they play?
- 10). Web search
- Describe only the on-chip memory hierarchy (levels of cache, associativity, size, linesize [if you can find it] for each cache) for the following processors:
    - Alpha 21264
    - Pentium 4(Northwood)
    - Athlon (Thoroughbred)
    - Itanium
  - Describe the distinguishing features of the memory hierarchy for the following:
    - Cray C90
    - Unisys ES7000
- Please provide citations/URLs

Provide a Definition for the following Terms

- Memory Hierarchy
- Temporal Locality
- Spatial Locality
- Non-Volatile Memory

- 15). Split Cache
- 16). Basic Block
- 17). Victim Cache
- 18). MSHR
- 19). Trace Cache
- 20). Register Update Unit
- 21). RAW Hazard