EECS 470 Fall 2025
Homework 5

Due December 3rd at 11:59pm.

This is an individual assignment, all of the work should be your own. **Remember you may drop one homework assignment.**

1) Consider the following access pattern: A, B, C, A. Assume that A, B, and C are memory addresses each of which are in a different block of memory. Further, assume A, B and C are generated in a uniformly random way and that a "true" LRU replacement algorithm is used. Further, assume that any given block has an equal chance of being placed in either "way". To receive credit you must show your work. What is the probability that the second instance of "A" will be a hit if:
   a) The cache has 4 lines and is direct-mapped **[1 point].**
   b) The cache has 4 lines and is fully-associative **[1 point].**
   c) The cache has 8 lines and is direct-mapped **[1 point].**
   d) The cache has 4 lines and is two-way associative **[2 points].**

2) Consider a 4KB direct-mapped cache backed-up by a 1MB 4-way associative cache. The L2 is inclusive of the L1. Say the L1 has a T_{hit} of 2 cycles, the L2 has a T_{hit} of 10 cycles, and the L2 has a T_{miss} of 200 cycles. Assume the T_{miss} time does not include the time to figure out if we have a hit or not. Answer the following questions:
   a) What would be the average memory access time if the L1 has a hit rate of 95% and the L2 has a hit rate of 50%? Assume accesses are only sent to the L2 if they miss in the L1 **[1 point].**
   b) There is a proposal to add an L0 cache of 1KB. The L0 would have a 1 cycle access time and memory requests would only be passed on to the L1 if the L0 missed. The L1 would be inclusive (See wikipedia for a definition if needed) of the L0 and accesses that hit in the L1 and/or L2 before would continue to do so. What hit-rate would be needed to get a tie with the original configuration? (Use the hit-rate numbers from part "a" as needed.) **[2 points]**
   c) In part b we said that the L1 is only accessed if the L0 misses. Let's assume we instead accessed both the L1 and L0 in parallel. What would be the disadvantages of doing this as compared to the proposal in part b? **[2 points]**

3) Consider two L1 cache designs: Design A is a 64KB direct-mapped cache with 256-byte blocks, and Design B is a 64KB sub-blocked cache where the block size is 256 bytes and the sub-block size is 32 bytes. Assume memory addresses are 32 bits wide.
   a) Draw a picture of the tag array for Design A. Indicate the number of tags, and width of each tag. Calculate the total storage (in bits) for the entire tag array. **[1 point]**
   b) Repeat question 3a for Design B. Don't forget to include the necessary sub-block valid bits**! [1 point]**
   c) Suppose there is a 1MB array of 4-byte integers in main memory, aligned to a 256-byte boundary. If the processor walks the array from the first element to the last.
      i) How many misses does each cache design incur? **[0.5 points]**
      ii) How much total data is transferred from main memory by each design? **[0.5 points]**
   d) Now suppose the processor walks through the array starting at index zero and accesses only every 64^{th} element.
      i) How many misses does each cache design incur**? [0.5 points]**
      ii) How much total data is transferred from main memory by each design? **[0.5 points]**

4) Chips'n'Dip Processors is trying to improve memory system performance in their next-generation processor design. Their current design has a 64KB 2-way L1 and a 512MB 4-way L2. The L1 has a hit time of 2 cycles, the L2 access latency is 12 cycles, and main memory access latency is 250 cycles (L1, L2, and main memory accesses are performed in series). For the suite of benchmarks they use to evaluate their design, they achieve a 5% L1 miss rate and a 40% (local) L2 miss rate.
   a) What is the effective access time of their current design? **[2 points]**
   b) They are considering increasing the L2 to a 2MB 8-way design. The larger L2 has a 14 cycle access latency. What must the (local) L2 miss rate of the new design be to achieve a 15% reduction in overall average memory access time? **[2 points]**

5) For each of the scenarios described below, match the workload to the most appropriate processor design. You must **explain your choice to receive credit.** Your choices are:

A. normal superscalar (single-threaded, single core)
B. vector machine
C. 2-way simultaneous multithreaded superscalar
D. 4-core chip multiprocessor (single-threaded cores)
E. 10-thread fine-grain multithreading

a) A scientific kernel that primarily performs matrix multiplication **[1 point]**

b) A network packet processing application where many independent packets must be checked against virus signatures. The data structure for virus signatures is too large to fit in an L1 cache, but can fit in an L2. **[1 point]**

c) Compiling a single, large C source file with gcc. **[1 point]**

d) A parallel computational fluid dynamic solver, where each thread consists of long sequences of floating point operations. Each thread's performance is bounded by floating point execution bandwidth. **[1 point]**

e) A multiprogrammed workload consisting of ray tracing software, which heavily utilizes the floating point unit, and a text parser, which contains no floating point code. **[1 point]**

6) Given a virtually indexed, physically-tagged cache that is four-way associative and has 64-byte blocks,
a) what is the largest **total size** the cache could have if pages were 4KB in size? **[1 point]**
b) What is the largest **number of sets** it could have in that case? **[1 point]**
c) How would these two values change if the cache were direct-mapped? **[1 point]**

Now, answer these questions about virtual memory based systems:
d) Describe one hardware and one software solution to mitigate the synonym problem in the D-cache. **[1 point]**
e) In one or two sentences, explain why it is difficult to support multiple page sizes in a hardware TLB. **[1 point]**

7) Draw the state transition diagram for a three-state MSI (Modified-Shared-Invalid) coherence protocol for a snoopy bus-based symmetric multiprocessor. Draw only the stable states (i.e., your diagram should have three states, "Modified", "Shared", and "Invalid"). Include transition arrows for all events that can occur in a state. For each transition, label it with "Event => Reaction". **[2 points]**