Assignment # 4

1. A computer has a cache of 256KB with a block size of 64 Bytes. Assume that the main memory is 4 GB. What is the address format used to access this cache, if the cache is organized as
   - directed mapped?
   - 4-way set associative?

2. Consider a processor with 3-level cache memory hierarchy sitting in between the CPU and Memory followed by a an SSD. Assume a base CPI of 1.6 and an execution profile with
   - an L1-cache miss rate of 15% where the penalty is 10 cycles,
   - an L2-cache miss-rate of 5% where the penalty is 40 cycles,
   - an L3-cache miss-rate of 1% where the penalty is 120 cycles, and
   - a memory page fault with frequency 0.01% with a SSD access time of 5000 cycles.

Assume in the target execution profile 20% of the instructions reference the memory. What is the CPI for the given execution profile?

3. This problem evaluates the cache performances for different loop orderings. You are asked to consider the following two loops, written in C, which calculate the sum of the entries in a 128 by 64 matrix of 32-bit integers:

   **Loop A:**
   ```c
   sum = 0;
   for (i=0, i < 128; i = i + 1)
       for (j=0, j < 64; j = j + 1)
           sum = sum + A[i][j]
   
   **Loop B:**
   ```
   ```c
   sum = 0;
   for (j=0, j < 64; j = j + 1)
       for (i=0, i < 128; i = i + 1)
           sum = sum + A[i][j]
   ```

The matrix A is stored contiguously in memory in row-major order. Row major order means that elements in the same row of the matrix are adjacent in memory as shown in the following memory layout: A[i][j] resides in memory location [4*(64*i + j)]
Memory Location:

<table>
<thead>
<tr>
<th>0</th>
<th>4</th>
<th>252</th>
<th>256</th>
<th>4*(64*127+63)</th>
</tr>
</thead>
</table>

Assume that the caches are initially empty. Also, assume that only accesses to matrix A cause memory references and all other necessary variables are stored in registers. Instructions are in a separate instruction cache.

(a) Consider a 4KB direct-mapped data cache with 8-word (32-byte) cache lines. Calculate the number of cache misses that will occur when running Loop A and Loop B.

(b) Consider a direct-mapped data cache with 8-word (32-byte) cache lines. Calculate the minimum number of cache lines required for the data cache if Loop A / Loop B is to run without any cache misses other than compulsory misses.

4. You are designing a write buffer between a write-through L1 cache and a write-back L2 cache. The L2 cache write data bus is 16 bytes wide and can perform a write to an independent cache address every 4 processor cycles.

(a) How many bytes wide should each write buffer entry be?

(b) What speedup could be expected in the steady state by using a merging write buffer instead of a non-merging buffer. Assume 64-bit stores are executed to zero memory. Assume further that all necessary instructions other than stores could be issued in parallel with the stores and the blocks are present in L2 cache?