Tutorial
Cache Memory

1. A set-associative cache consists of 64 lines divided into sets of 4 lines each. Main memory has a 32-bit address and the block size is 128 bytes.
   a) Show the format of the main address.
   b) How many sets are there?

2. A computer has a 32-bit address. It has an 8 Kbyte cache with a 2-way set-associative organization. The block size is 8 bytes.
   a) Show the format of the main address.
   b) How many blocks are mapped to each set?
   c) How many lines are there in each set?
   d) How many sets are there?

3. A computer uses a direct mapped cache that is 64 Kbytes. The main memory is 1 Gbyte. The block size is 8 bytes.
   a) Show the format of the main address.
   b) How many lines are there in the cache?
   c) How many blocks are mapped to the same line?

4. What is the replacement algorithm used with direct caches?
5. For a direct-mapped cache design with a 32-bit address and byte-addressable memory, the following bits of the address are used to access the cache:

   a)

Tag	Index	Offset
31-10	9-5	4-0

   i. What is the cache line size (in words)?
   ii. How many entries (cache lines) does the cache have? Cache lines are also called blocks.

   b)

Tag	Index	Offset
31-12	11-6	5-0

   i. What is the cache line size (in words)?

ii. How many entries (cache lines) does the cache have?

6. You have an L1 data cache, L2 cache, and main memory. The hit rates and hit times for each are:

50% hit rate, 2 cycle hit time to L1.
70% hit rate, 15 cycle hit time to L2.
100% hit rate, 200 cycle hit time to main memory

i. What fractions of accesses are serviced from L2? From main memory?

ii. What is the miss rate and miss time for the L2 cache?

iii. What is the miss rate and miss time for the L1 cache? (Hint: depends on previous answer)

iv. If main memory is improved by 10%, what is the improvement in L1 miss time?

Additional Problems

1. You have a 2-way set associative L1 cache that is 8KB, with 4-word cache lines. Writing data to L2 takes 10 cycles. You get the following sequence of writes to the cache -- each is a 32-bit address in hexadecimal (along with their decimal values):

0x1000 4096
0x1004 4100
0x1010 4112
0x11c0 4554
0x2000 8192
0x21c0 8640
0x3400 13312
0x3404 13316
0x3f00 16128
0x2004 8196
0x1004 4100

a. How many cache misses occur with an LRU policy?
b. How many cache misses occur with a most-recently used policy?
c. Would the miss-rate increase or decrease if the cache was the same size, but direct-mapped? Explain.
d. How long does a read-miss eviction take if the cache is write-back, write-allocate? What about a write-miss? (Assume the cache line is dirty and Assume that writing and reading data to/from L2 takes 10 cycles.)
e. How long does a read-miss eviction take if the cache is write-through, write-allocate? What about a write-miss?

2. You have 3 cache designs for a 16-bit address machine.

C1:

Direct-mapped cache.
Each cache line is 1 byte.
10-bit index, 6-bit tag.
1 cycle hit time.

C2:

2-way set associative cache.
Each cache line is 1 word (4 bytes).
7-bit index, 7-bit tag.
2 cycle hit time.

C3:

fully associative cache with 256 cache lines.
Each cache line is 1 word.
14-bit tag.
5 cycle hit time.

a. What is the size of each cache?
b. How much space does each cache need to store tags?
c. Which cache design has the most conflict misses? Which has the least?
d. If someone told you the hit rate for the 3 caches is 50%, 70% and 90% but did not tell you which hit rate corresponds to which cache, which cache would you guess corresponded to which hit rate? Why?
e. Assuming the miss time for each is 20 cycles, what is the average service time for each? (Service Time = (hit rate)*(hit time) + (miss rate)*(miss time)).

3. In 1000 memory references there are 40 misses in L1 and 20 misses in L2. Assume all are reads with no bypass.

   L1: Hit time = 1 cycle
   L2: Hit time = 10 cycles
   Miss penalty = 100 cycles
   What is the average memory access time?

4. Consider a cache of 4 lines of 16 bytes each. Main memory has blocks of 16 byte each. Thus block 0 has bytes with addresses 0 through 15, and block 1 has addresses 16 through 31 etc. Consider a program that accesses memory in following sequence of addresses:

Once: 63 through 70
Loop 10 times: 15 through 32; 80 through 95

a) Suppose the cache is organized as direct mapped. Memory blocks 0, 4, and so on are assigned to line 1; blocks 1, 5, and so on to line 2; and so on. Compute the hit ratio.
b) Suppose the cache is organized as two-way set associative, with two sets of two lines each. Even-numbered blocks are assigned to set 0 and odd-numbered blocks are assigned to set 1. Compute the hit ratio for the two-way set-associative cache using the least recently used replacement scheme.

5. Consider a memory system that consists of only 8 blocks and there is a 4 line cache. The cache is fully associativity uses LRU. If a program requires the sequence of blocks shown below, what is in the cache when the program ends?

the sequence of lines in the cache as the program progresses.
1 3 7 0 3 5 1 3 4 7 7 5 2