6.004 Recitation Problems
L15 – Memory Hierarchy

Keep the most often-used data in a small, fast SRAM (often local to CPU chip). The reason this strategy works: LOCALITY.
– Temporal locality: If a location has been accessed recently, it is likely to be accessed (reused) soon
– Spatial locality: If a location has been accessed recently, it is likely that nearby locations will be accessed soon

AMAT(Average Memory Access Time) = HitTime + MissRatio * MissPenalty
Problem 1. ★
Belly Eyelash is designing a processor and is analyzing the performance of different numbers of cache levels. Without any caches, Belly’s main memory has an access time of 140 cycles.

(A) As a test, Belly adds a 32 KB Level 1 (L1) cache that has single-cycle reads/writes, and runs a long computation, during which she observes a new AMAT of 10. What is the hit ratio for the Level 1 cache during this test?

\[ \text{AMAT} = 10 = 1 + (1 - \text{HitRatio}) \times 140 \]
\[ \text{HitRatio} = \frac{131}{140} \]

(B) Next, Belly adds a 256 KB Level 2 (L2) cache between the L1 cache and main memory. The L2 cache takes 4 cycles to decide if the memory access is a hit or a miss. After running the same computation with this new memory hierarchy, Belly observes an improved AMAT of 1.45. Assume that the hit rate for the L1 cache is the same as in (A). What is the hit ratio for the Level 2 cache during this test?

\[ \text{AMAT} = 1.45 = 1 + (\frac{9}{140}) \times (4 + (1 - \text{HitRatio}) \times 140) \]
\[ 0.45 \times (140/9) - 4 = 7 - 4 = (1 - \text{HitRatio}) \times 140 \]
\[ \text{HitRatio} = \frac{137}{140} \]

(C) Finally, Belly adds a 10 MB Level 3 (L3) cache between the L2 cache and main memory. After running the same test as before, she observes that the main memory was accessed once for every 140 accesses to the L3 cache. If Belly wants to achieve an AMAT of 1.3 for this computation, what is the maximum number of clock cycles that the Level 3 cache can take to decide if a memory access is a hit or a miss? Assume that the L1 and L2 caches can not be changed and they have the same hit rates observed in (A) and (B).

\[ \text{AMAT} = 1.3 = 1 + (\frac{9}{140}) \times (4 + 3 \times \text{HitTime} + (1/140) \times 140) \]
\[ 0.3 \times 140 = 9 \times 140 \times (4 + 3 \times \text{HitTime} + 1) \]
\[ \text{HitTime} = \frac{0.3 \times 140^2 - 36 \times 140}{27} - 1 \]
To achieve AMAT <= 1.3, HitTime <= 30
Basic Cache Algorithm (Reads)

On reference to Mem[X], look for X among cache tags

HIT: X = Tag(i) for some cache line i

MISS: X not found in Tag of any cache line

Return Data(i)

Read Mem[X]

Select a line k to hold Mem[X]

Write Tag(k) = X, Data(k) = Mem[X]

Q: How do we "search" the cache?

Direct-Mapped Caches

- Each word in memory maps into a single cache line
- Access (for cache with $2^w$ lines):
  - Index into cache with W address bits (the index bits)
  - Read out valid bit, tag, and data
  - If valid bit == 1 and tag matches upper address bits, HIT
- Example 8-line direct-mapped cache:

```
32-bit BYTE address
00000000000000000000000000000011 01000
```

Valid bit  Tag (27 bits)  Data (32 bits)

0  1  2  3  4  5  6  7

Tag bits  Index Offset bits
Example: Direct-Mapped Caches

64-line direct-mapped cache → 64 indexes → 6 index bits

Read Mem[0x400C]

0100 0000 0000 1100
TAG: 0x40
INDEX: 0x3
OFFSET: 0x0

HIT, DATA 0x42424242

Would 0x4008 hit?
INDEX: 0x2 → tag mismatch
→ MISS

Part of the address (index bits) is encoded in the location
Tag + Index bits unambiguously identify the data’s address

Block Size

- Take advantage of spatial locality: Store multiple words per data block
  - Another advantage: Reduces size of tag memory!
  - Potential disadvantage: Fewer blocks in the cache
- Example: 4-block, 16-word direct-mapped cache

Valid bit Tag (26 bits) Data (4 words, 16 bytes)

32-bit BYTE address
Tag bits: 26 (=32-4-2) Index bits: 2
(16 bytes/block)

Block offset bits: 4
(4 indexes)
Problem 2.

The RISC-V Engineering Team is working on the design of a cache. They’ve decided that the cache will have a total of $2^{10} = 1024$ data words, but are still thinking about the other aspects of the cache architecture.

First assume the team chooses to build a direct-mapped cache with a block size of 4 words.

(A) Please answer the following questions:

- **Number of lines in the cache:** 256
- **Tag:** 20bit, **Index:** 8bit, **Offset:** 4bit
- **Number of bits in the tag field for each cache entry:** 20

(B) This cache takes 2 clock cycles to determine if a memory access is a hit or a miss and, if it’s a hit, return data to the processor. If the access is a miss, the cache takes 20 additional clock cycles to fill the cache line and return the requested word to the processor. If the hit rate is 90%, what is the processor’s average memory access time in clock cycles?

**Average memory access time assuming 90% hit rate (clock cycles): 4**

\[
AMAT = \text{HitTime} + (1 - \text{HitRatio}) \times \text{MissPenalty} = 2 + (1 - 0.90) \times 20 = 4
\]
Problem 3. ★

(A) The timing for a particular cache is as follows: checking the cache takes 1 cycle. If there's a hit the data is returned to the CPU at the end of the first cycle. If there's a miss, it takes 10 additional cycles to retrieve the word from main memory, store it in the cache, and return it to the CPU. If we want an average memory access time of 1.4 cycles, what is the minimum possible value for the cache's hit ratio?

Minimum possible value of hit ratio: \(0.96\)

\[1.4=1+(1-HR)\times10 \Rightarrow HR=0.96\]

(B) If the cache block size, i.e., words/cache line, is doubled but the total number of data words in the cache is unchanged, how will the following cache parameters change? Please circle the best answer.

- \# of offset bits: UNCHANGED ... +1 ... -1 ... 2x ... 0.5x ... CAN’T TELL
- \# of tag bits: UNCHANGED ... +1 ... -1 ... 2x ... 0.5x ... CAN’T TELL
- \# of cache lines: UNCHANGED ... +1 ... -1 ... 2x ... 0.5x ... CAN’T TELL

Consider a direct-mapped cache with 64 total data words with 1 word/cache line. This cache architecture is used for parts (C) through (F).

(C) If cache line number 5 is valid and its tag field has the value 0x1234, what is the address in main memory of the data word currently residing in cache line 5?

Main memory address of data word in cache line 5: \(0x123414\)

Tag: 24bits, Index: 6bits (000101), block offset: 2bits (00)

The program shown on the right repeatedly executes an inner loop that sums the 16 elements of an array that is stored starting in location 0x310.

The program is executed for many iterations, then a measurement of the cache statistics is made during one iteration through all the code, i.e., starting with the execution of the instruction labeled outer_loop: until just before the next time that instruction is executed.

\[
\begin{align*}
. &= \emptyset & & // \text{tell assembler to start at} \\
\text{outer_loop:} \\
& \text{addi} \ x4, \ x0, \ 16 & & // \text{initialize loop index J} \\
& \text{mv} \ x1, \ x0 & & // \text{x1 holds sum, initially } 0 \\
\text{loop:} & & & // \text{add up elements in array} \\
& \text{subi} \ x4, \ x4, \ 1 & & // \text{decrement index} \\
& \text{slli} \ x2, \ x4, \ 2 & & // \text{convert to byte offset} \\
& \text{lw} \ x3, \ 0x310(x2) & & // \text{load value from A[J]} \\
& \text{add} \ x1, \ x1, \ x3 & & // \text{add to sum} \\
& \text{bne} \ x4, \ x0, \ \text{loop} & & // \text{loop until all words are summed} \\
& \text{j} \ \text{outer_loop} & & // \text{perform test again!}
\end{align*}
\]
(D) In total, how many instruction fetches occur during one complete iteration of the outer loop? How many data reads?

Number of instruction fetches: **83**

Instruction fetch = 2+5*16+1=83

Number of data reads: **16**

(E) How many instruction fetch misses occur during one complete iteration of the outer loop? How many data read misses? Hint: remember that the array starts at address 0x310.

Number of instruction fetch misses: **4**

Number of data read misses: **4**

Data: 0x310, 0x314, 0x318, 0x31C, 0x320, 0x324, 0x328 ……
Instruction: 0x0 addi, 0x4 mv, 0x8 subi, 0xC slli, 0x10 lw, 0x14 add, 0x18 bne, 0x1C j ……

Bold Blue addresses are cache misses / conflicts.

(F) What is the hit ratio measured after one complete iteration of the outer loop?

Hit ratio: **91/99**, where **99=83+16**  **91=(83-4)+(16-4)**