Q1. Assume a 2-level cache system with the following specifications. L1 Hit Time = 1 cycle, L1 Miss Rate = 2.5%, L2 Hit Time = 6 cycles, L2 Miss Rate = 17% (% L1 misses that miss), L2 Miss Penalty = 120 cycles. Compute the average memory access time.

Q1. Assume a 2-level cache system with the following specifications. L1 Hit Time = 1 cycle, L1 Miss Rate = 2.5%, L2 Hit Time = 6 cycles, L2 Miss Rate = 17% (% L1 misses that miss), L2 Miss Penalty = 120 cycles. Compute the average memory access time.

AMAT for first level cache

AMAT = ht1 + mr1 × MP

Similarly, AMAT for second level cache

AMAT= ht1 + mr1 × (ht2 + mr2 × MP)

Substituting the values in above formula

AMAT = 1 + 0.025 × (6 + 0.17 × 120) = 1.66 CC

Q2. A cache has access time (hit latency) of 10 ns and miss rate of 5%.An optimization was made to reduce the miss rate to 3% but the hit latency was increased to 15 ns. Under what condition this change will result in better performance (Lower AMAT)?

Q2. A cache has access time (hit latency) of 10 ns and miss rate of 5%.An optimization was made to reduce the miss rate to 3% but the hit latency was increased to 15 ns. Under what condition this change will result in better performance (Lower AMAT)?

- HT1=10 ns, MR1=0.05 (original)
- HT2=15 ns, MR2=0.03(after optimization)
- Miss penalty (MP) remains the same in both.
- the new AMAT to be less than the original:

$$AMAT2 < AMAT1$$

$$15 + 0.03 \times MP < 10 + 0.05 \times MP$$

- Solve for Miss Penalty (MP):

$$15 + 0.03\ MP < 10 + 0.05\ MP$$

$$5 < 0.02MP$$

$$MP > 250ns$$

This shows that the optimization results in better performance only if the miss penalty is greater than 250 ns

Q3. Hit rate of 95% (128B blocks, cache hit latency of 5ns Main memory takes 50 ns to return first word (32 bits)) of a block and 10 ns for each subsequent word.
(a) What is the miss latency of the cache?
(b) If doubling the cache block size reduces the miss rate to 3%, does it reduces AMAT?

(a)

- Number of words per block: 128B/4B =32 words per block
- Miss Penalty Calculation:
  - It takes 50 ns to transfer the first word, then 10 ns for each of the remaining 31 words.
  - MP=50+(31×10)=360*MP*=50+(31×10)=360 ns
- AMAT Formula:
  AMAT=Hit time+(Miss rate×Miss penalty)*AMAT*=Hit time+(Miss rate×Miss penalty)
  AMAT1=5+0.05×360=5+18=23ns

(b)

- New block size: 256 Byte blocks
- New number of words per block: 256B/4B=64 words per block
- New Miss Penalty:
  - MP=50+((64-1)×10)=680MP=50+(63×10)=680 ns
- Reduced miss rate: 3% = 0.03
- New AMAT:
  AMAT2=5+0.03×680=5+20.4=25.4 ns

Q4. A 16KB direct mapped 256B block unified cache is attached to a 16MB main memory system. The word length as well as instruction length of the processor is 16 bits. Consider a program that consists of a main routine M which in turn calls a subroutine S. M consists of 12 instruction words which are loaded in the main memory from the address 0x4230FA onwards. The last five instructions of M is a loop that is iterated 10 times. The second instruction in the loop is a call to subroutine S. S consists of 4 instruction words loaded in the main memory from the address 0x70F168. The last instruction of S is a subroutine return back to M. The only two data words that are used by M and S are at addresses 0x748074 and 0x846064. Assume the caches are initially empty. Ignore OS level interruption and subsequent cache impact on context switching.

Tag	Index	Offset
6	6	8

1	0x 4230FA
2	0x 4230FC
3	0x 4230FE
4	0x 423100
5	0x 423106
6	0x 423108
7	0x 42310A
8	0x 42310A
9	0x 42310C
10	0x 42310E
11	0x 423110

Tag	Index	Offset
6	6	8

1	0x 4230FA
2	0x 4230FC
3	0x 4230FE
4	0x 423100
5	0x 423106
6	0x 423108
7	0x 42310A
8	0x 42310A
9	0x 42310C
10	0x 42310E
11	0x 423110

Tag	Index	Offset
6	6	8

1	0x 707168
2	0x 70716A
3	0x 70716C
4	0x 70716E

M

1	0x 4230FA (48)
2	0x 4230FC (48)
3	0x 4230FE (48)
4	0x 423100 (49)
5	0x 423102 (49)
6	0x 423104 (49)
7	0x 423106 (49)
8	0x 423108 (49)
9	0x 42310A (49)
10	0x 42310C (49)
11	0x 42310E (49)
12	0x 423110 (49)

10

Tag	Index	Offset
10	6	8

1	0x 748074 (0)

1	0x 846064 (16)

S

1	0x 707168 (49)
2	0x 70716A (49)
3	0x 70716C (49)
4	0x 70716E (49)

Find the number of cache misses occurred during the execution of the program

Find the number of cache misses occurred during the execution of the program

M1

M4

S1, M10, S1, M10,.....(10 TIMES)= 22 MISSES

How many cache block evictions happened during the execution of the program?

How many cache block evictions happened during the execution of the program?

20 evictions

List out the block numbers(in decimal) in the cache that are non-empty after the execution

List out the block numbers(in decimal) in the cache that are non-empty after the execution

All blocks Except 0, 16, 48, 49

Q.5 A CPU has a cache with block size 64 bytes. The main memory has k banks, each bank being c bytes wide. Consecutive c-byte chunks are mapped on consecutive banks with wrap-around. All the k banks can be accessed in parallel, but two accesses to the same bank must be serialized. A cache block access may involve multiple iterations of parallel bank accesses depending on the amount of data obtained by accessing all the k banks in parallel. Each iteration requires decoding the bank numbers to be accessed in parallel and this takes k/2 ns. The latency of one bank access is 80ns. If c=2 and k=24, the latency of a cache block starting at address zero from main memory is:

- In each parallel bank access, each of the 24 banks supplies 2 bytes, for a total per iteration:

$$\text{Bytes per iteration} = k \times c = 24 \times 2 = 48 \text{ bytes}$$

- The cache block size is 64 bytes.
- The first iteration gives 48 bytes.
- The remaining bytes after one iteration:

$$\text{Remaining} = 64 - 48 = 16 \text{ bytes}$$

In the second iteration, all banks can be used, but only 8 banks need to be accessed (since 16/2=8), supplying 16 bytes.

- Therefore, it takes 2 iterations to fetch the full cache block.
- Bank access latency per iteration: Bank access latency=80 ns
- Total time per iteration: Time per iteration=12+80=92 ns
- Two iterations required: Total latency=92 ns×2=184 ns

Final calculation table:

Iteration	Bytes Fetched	Decoding Time (ns)	Access Latency (ns)	Total Iteration Time (ns)
1	48	12	80	92
2	16	12	80	92