Assignment 7

Monday, December 4, 2023 2:10 PM

CS 219 – Assignment #7

Purpose: Become familiar with MIPS cache implementation

Points: 100

Reading/References: Chapter 5 (5.1, 5.2, 5.3, 5.4)

Assignment:
Answer the following questions:

1) Define cache? Discuss whether the cache size should be small or large to overcome misses like compulsory, capacity, conflict? [10 pts]

2) Explain the Temporal and Spatial locality principle and give examples for each locality principle. [10 pts]

3) Explain memory hierarchy level and talk about their speeds. Give example for each level-memory type and where they are placed in a typical computer architecture. [10 pts]
[Hint: Refer to A Typical Memory Hierarchy on Ch5 slides on Canvas]

4) In general, cache access time is proportional to capacity. Assume that main memory accesses take 70 ns. The following table shows data for L1 caches attached to each of two processors, P1 and P2. [10 pts]

	L1 Size	L1 Miss rate	L1 Hit time
P1	1KB	11.4%	0.62ns
P2	2KB	8.0%	0.66ns

a) If the L1 hit time determines the cycle times for P1 and P2, what are their respective clock rates? [Hint: clock rate= 1/L1 hit time]
b) What is the AMAT for each of P1 and P2?
AMAT= Hit time + Miss rate x Miss penalty

1. A cache is a small, fast, and expensive memory used to reduce the speed gap between the CPU and the main memory. The cache cannot be too small since that would increase the number of capacity misses. Both a large cache and small cache have a compulsory miss. Large caches have less capacity misses and collision misses, but have slower overall search times.

2. The temporal principle is the idea that the same location will be referenced again in the near future such as a counting variable in a for loop. The spatial principle is the idea that the nearby memory locations will be referenced in the near future such as the content inside a for loop.

3. In the Hierarchy of Memory there are 5 layers. The layers are the processor at the top, then L1$ cache sits a bit slower on-chip. The L2$ is next and is off-chip. The main memory is off-chip and is slower than the L2 memory. Finally the slowest memory is the secondary memory is at the bottom. An example of processor memory are the registers in the register file blocks. The example of L1 memory is instructional or data cache. L2$ memory is typically a second level cache memory using SRAM. Then the main memory typically uses DRAM. Finally the secondary memory uses a disk or SSD.

5) Here is a series of address references given as words addresses: [15 pts]

5, 18, 1, 1, 2, 3, 11, 10, 21, 18, 17, 10, 12, 3, 11, 7, 10, 22, 4, 22.

a. Assuming a direct-mapped cache with 8 one-word blocks that is initially empty, label each reference in the list as a hit or a miss and show the contents of the cache (including previous, over-written values). You do not need to show the tag field. When done, include the hit ratio. If you have entry in cache, then valid bit is 1 else zero.

	Cache Set	valid	Address
0	000	0	
1	001	1	1, 17
2	010	1	18, 2, 10, 18, 10,
3	011	1	3, 11, 3, 11
4	100	1	12, 4
5	101	1	5, 21
6	110	1	22
7	111	1	7

4a. P1 clk rate = $\frac{1}{L1} = \frac{1}{0.62 \times 10^{-9}} = 1.613 \times 10^{8} = \boxed{1.6 \times 10^{8}\ Hz}$

P2 clk rate = $\frac{1}{L1} = \frac{1}{0.66 \times 10^{-9}} = 1.515 \times 10^{8} = \boxed{1.5 \times 10^{8}\ Hz}$

4b. P1 AMAT = Hit Time + Miss Rate x Miss Penalty =

$= 0.62ns + (0.114 \times 70ns)$

$= \boxed{8.6ns}$

P2 AMAT = $0.66 + (0.08 \times 70ns)$

$= \boxed{6.26ns}$

Table 1: Reference address to cache mapping

Address Decimal	Address Binary	Line ID/ Cache Set	Hit/Miss	
5	0000 0101	101	Miss	1
18	0001 0010	010	Miss	2
1	0000 0001	001	Miss	3
1	0000 0001	001	Hit	4
2	0000 0010	010	Miss	5
3	0000 0011	011	Miss	6
11	0000 1011	011	Miss	7
10	0000 1010	010	Miss	8
21	0001 0101	101	Miss	9
18	0001 0010	010	Miss	10
17	0001 0001	001	Miss	11
10	0000 1010	010	Miss	12

Note: For Q5 and Q6, fill/use Table 1 for each part (a, b, c, d) in addition to the provided tables.

Hints for direct-mapped cache (one-word blocks)
(**Hint 1:** Read section 5.3 and map the cache locations similar to Figure 5.8,
Hint2: Convert each address that is given in decimal to binary, then based on last 3 digits, map into one of the 8 cache entries.
Hint 3: For the first time, any address is considered as miss and if the address is repeated from the given list then it is considered as hit. To consider hit the previous entry should be the same in that cache. For example: 1, 9, 1, 2, 1 are the addresses, in binary 0001, 1001, 0001, 0010, 0001 the addresses (1, 9, 1, 1) will be in 001 cache entry and address 2 in 010 cache entry. In this (1, 9, 1, 2, 1) for the first time 1-miss, then 9-miss, 1-miss (1 is miss here because previous entry is 9), 2-miss, 1-hit(1 is hit here because previous entry in that cache list is 1)
Hint 4: To find hit ratio: number of hits divided by total entries, in this example it is 1/5 or 20%.)

10	0000 1010	010	Miss	8
21	0001 0101	101	Miss	9
18	0001 0010	010	Miss	10
17	0001 0001	001	Miss	11
10	0000 1010	010	Miss	12
12	0000 1100	100	Miss	13
3	0000 0011	011	Miss	14
11	0000 1011	011	Miss	15
7	0000 0111	111	Miss	16
10	0000 1010	010	Hit 1	17
22	0001 0110	110	Miss	18
4	0000 0100	100	Miss	19
22	0001 0110	110	Hit 1	20
Hit Ratio	3/20 = 15% Hit Ratio			
Comments				

Hint 4: To find hit ratio: number of hits divided by total entries, in this example it is 1/5 or 20%.)

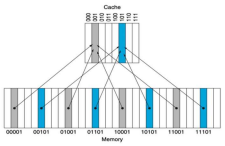

FIGURE 5.8 A direct-mapped cache with eight entries showing the addresses of memory words between 0 and 31 that map to the same cache locations. Because there are eight words in the cache, an address X maps to the direct-mapped cache word X modulo 8. That is, the low-order $\log_2(8) = 3$ bits are used as the cache index. Thus, addresses 00001_, 01001_, 10001_, and 11001_ all map to entry 001_ of the cache, while addresses 00101_, 01101_, 10101_, and 11101_ all map to entry 101_ of the cache.

b) Assuming a direct-mapped cache with 16 one-word blocks that is initially empty, label each reference in the list as a hit or a miss and show the contents of the cache (including previous, over-written values). You do not need to show the tag field. When done, include the hit ratio. [15 pts]

	Cache Set	Valid	Address
0	0000		
1	0001	1	1, 17
2	0010	1	18, 2, 18
3	0011	1	3
4	0100	1	4
5	0101	1	5, 21
6	0110	1	22
7	0111	1	7
8	1000		
9	1001		
10	1010	1	10,
11	1011	1	11
12	1100	1	12
13	1101		
14	1110		
15	1111		

Table 1: Reference address to cache mapping

	Address Decimal	Address Binary	Line ID/ Cache Set	Hit/Miss
1	5	0000 0101	0101	Miss
2	18	0001 0010	0010	Miss
3	1	0000 0001	0001	Miss
4	1	0000 0001	0001	Hit ✓
5	2	0000 0010	0010	Miss
6	3	0000 0011	0011	Miss
7	11	0000 1011	1011	Miss
8	10	0000 1010	1010	Miss
9	21	0001 0101	0101	Miss
10	18	0001 0010	0010	Miss
11	17	0001 0001	0001	Miss
12	10	0000 1010	1010	Hit ✓
13	12	0000 1100	1100	Miss
14	3	0000 0011	0011	Hit ✓
15	11	0000 1011	1011	Hit ✓
16	7	0000 0111	0111	Miss
17	10	0000 1010	1010	Hit ✓
18	22	0001 0110	0110	Miss
19	4	0000 0100	0100	Miss
20	22	0001 0110	0110	Hit ✓
Hit Ratio	6/20 = 0.3 = 30% Hit Ratio			
Comments				

Hit Ratio	6/20 = 0.3 = 30% Hit Ratio
Comments	

c) Show the hits and misses and cache contents (including previous, overwritten values) for a direct-mapped cache with four-word blocks and a total size of 8 words. You do not need to show the tag field. When done, include the hit ratio. [10 pts]

Cache Set	valid	Address
00/ 10	1	[16,17,18,19], [0,1,2,3], [8,9,10,11], [16,17,18,19], [8,9,10,11], [12,13,14,15], [0,1,2,3], [8,9,10,11],
01/11	1	[4,5,6,7], [20,21,22,23], [4,5,6,7], [20,21,22,23], [4,5,6,7], [20,21,22,23]

Table 1: Reference address to cache mapping

	Address Decimal	Address Binary	Line ID/ Cache Set	Hit/Miss
1	5	0000 0101	1	Miss
2	18	0001 0010	0	Miss
3	1	0000 0001	1	Miss
4	1	0000 0001	1	Hit ✓
5	2	0000 0010	0	Hit ✓
6	3	0000 0011	1	Hit ✓
7	11	0000 1011	1	Miss
8	10	0000 1010	0	Hit ✓
9	21	0001 0101	1	Miss
10	18	0001 0010	0	Miss
11	17	0001 0001	1	Hit ✓
12	10	0000 1010	10	Miss
13	12	0000 1100	00	Miss
14	3	0000 0011	1	Miss
15	11	0000 1011	1	Miss
16	7	0000 0111	1	Miss
17	10	0000 1010	0	Hit ✓
18	22	0001 0110	0	Miss
19	4	0000 0100	0	Miss
20	22	0001 0110	0	Miss
	Hit Ratio	6/20 = 30% Hit Rate		
	Comments			

d) Show the hits and misses and cache contents (including previous, overwritten values) for a direct-mapped cache with four-word blocks and a total size of 16 words. You do not need to show the tag field. When done, include the hit ratio. [10 pts]

Cache Set	valid	Address
00	1	[0,1,2,3], [12,13,14,15],
01	1	[4,5,6,7], [20,21,22,23],
10	1	[16,17,18,19], [8,9,10,11]
11	1	[8,9,10,11], [0,1,2,3], [8,9,10,11], [4,5,6,7]

Table 1: Reference address to cache mapping

	Address Decimal	Address Binary	Line ID/ Cache Set	Hit/Miss
1	5	0000 0101	0101	MISS
2	18	0001 0010	0010	MISS
3	1	0000 0001	0001	MISS
4	1	0000 0001	0001	Hit ✓
5	2	0000 0010	0010	Hit ✓
6	3	0000 0011	0011	Hit ✓
7	11	0000 1011	1011	MISS
8	10	0000 1010	1010	Hit ✓
9	21	0001 0101	0101	MISS
10	18	0001 0010	0010	Hit ✓
11	17	0001 0001	0001	Hit ✓
12	10	0000 1010	1010	Hit ✓
13	12	0000 1100	1100	MISS
14	3	0000 0011	0011	MISS
15	11	0000 1011	1011	MISS
16	7	0000 0111	0111	MISS
17	10	0000 1010	1010	MISS
18	22	0001 0110	0110	Hit ✓
19	4	0000 0100	0100	Hit ✓
20	22	0001 0110	0110	Hit ✓
Hit Ratio		10/20 = 50% Hit Rate		
Comments				

e) Compare the mapping techniques used above (a, b, c, and d) and write your observations that includes hit ratio and the reasons why the hit ratio is more compared to others.
[Hint: Improvement due to more one-word blocks; Improvement due to multiple words in each block, etc.] [10 pts]

It was shown in the previous parts that the ideal situation is to have an increase in multiple word blocks for the highest hit rate. The collision misses were common in the one-word block & multi-word block setups when there were too few blocks. With 1-word blocks with a large capacity, there were many compulsory misses. The situation with the best hit rate was with a large number of multi-word blocks.

e) Compare the mapping techniques used above (a, b, c, and d) and write your observations that includes hit ratio and the reasons why the hit ratio is more compared to others.
[Hint: Improvement due to more one-word blocks; Improvement due to multiple words in each block, etc.] [10 pts]

It was shown in the previous parts that the ideal situation is to have an increase in multiple word blocks for the highest hit rate. The collision misses were common in the one-word block & multi-word block setups when there were too few blocks. With 1-word blocks with a large capacity, there were many compulsory misses. The situations with the best hit rate was with a large number of multi-word blocks.