Disclaimer:

1. The solution is just for your reference. They may contain some mistakes. DO TRY to solve the problems by yourself. Please also pay attentions to the course website for the updates.

Selected Question: 5.1.1~5.1.3, 5.2.1~5.2.3, 5.3.1~5.3.5, 5.5.1~5.5.4, 5.7.2&5.7.3

5.1.1. 4

5.1.2 I, J, and B[I][0]

A[0][0] = B[0][0] + A[0][0]

A[0][1] = B[0][0] + A[1][0]

A[0][2] = B[0][0] + A[2][0]

….

A[0][7999]= B[0][0] + A[7999][0]

A[1][0]=B[1][0]+ B[0][1]

A[1][1]=B[1][0]+ B[1][1]

A[1][2]=B[1][0]+ B[2][1]

….

Therefore, I, J, and B[I][0] exhibit temporal locality

5.1.3

A[I][J] only. Note that A[J][I] doesn't exhibit spatial locality because only data within the same row are stored contiguously. Access pattern for A[J][I] are A[0][0], A[1][0], A[2][0]

Access pattern for A[I][J] = A[0][0], A[0][1], A[0][2] ….

Access pattern for B[I][0] = B[0][0], B[0][0]….

Access pattern for A[J][I] = A[0][0], A[1][0], A[2][0]

Therefore, only A[I][J] exhibits spatial locality

5.2.1

Word Address	Word address in Binary	Tag	Index	Hit/Miss
3	0000 0011	0000	0011	M
180	1011 0100	1011	0100	M
43	0010 1011	0010	1011	M
2	0000 0010	0000	0010	M
191	1011 1111	1011	1111	M
88	0101 1000	0101	1000	M

190	1011 1110	1011	1110	M
14	0000 1110	0000	1110	M
181	1011 0101	1011	0101	M
44	0010 1100	0010	1100	M
186	1011 1010	1011	1010	M
253	1111 1101	1111	1101	M

5.2.2

Each block contains two words, and there are 8 blocks in the cache. Therefore, index has 3 bits (bit 3:1), and tag has 4 bits (bit 7:4)

Word Address	Word address in Binary	Tag	Index	Hit/Miss
3	0000 0011	0	1	M
180	1011 0100	11	2	M
43	0010 1011	2	5	M
2	0000 0010	0	1	H
191	1011 1111	11	7	M
88	0101 1000	5	4	M
190	1011 1110	11	7	H
14	0000 1110	0	7	M
181	1011 0101	11	2	H
44	0010 1100	2	6	M
186	1011 1010	11	5	M
253	1111 1101	15	6	M

5.2.3

Word Address	Word address in Binary	Tag	Cache1		Cache2		Cache3	
			index	hit/miss	index	hit/miss	index	hit/miss
3	0000 0011	0	3	M	1	M	0	M
180	1011 0100	22	4	M	2	M	1	M
43	0010 1011	5	3	M	1	M	0	M
2	0000 0010	0	2	M	1	M	0	M
191	1011 1111	23	7	M	3	M	1	M
88	0101 1000	11	0	M	0	M	0	M
190	1011 1110	23	6	M	3	H	1	H
14	0000 1110	1	6	M	3	M	1	M
181	1011 0101	22	5	M	2	H	1	M
44	0010 1100	5	4	M	2	M	1	M
186	1011 1010	23	2	M	1	M	0	M

253	1111 1101	31	5	M	2	M	1	M

Cache 1 miss rate = 100%

Cache 1 total cycles = 12 × 25 + 12 × 2 = 324

Cache 2 miss rate = 10/12 = 83%

Cache 2 total cycles =10 × 25 + 12 × 3 = 286

Cache 3 miss rate= 11/12 = 92%

Cache 3 total cycles = 11 × 25 + 12 × 5 = 335

Cache 2 provides the best performance.

5.3.1. Offset has 5 bits. 2^5=32 bytes = 8 words

5.3.2 Index has 5 bits. 2^5=32 entries

5.3.3

Data size = 32 * 32*8 = 8192 bits

Total storage bits if no valid bit = 32* (22+ 32*8)= 8896

Ratio = 8896/8192 = 1.086 (without valid bits)

Note that is the valid bits are include:

Total storage bits if there are valid bits = 32 * (22+ 32*8+1) = 8928 bits

Ratio = 8928/8192 = 1.089 (with valid bits)

5.3.4 3 blocks are replaced.

Byte Address	Binary Address	Tag	Index	Hit/Miss
0	0000 0000 0000	0	00000	M
4	0000 0000 0100	0	00000	H
16	0000 0001 0000	0	00000	H
132	0000 1000 0100	0	00100	M
232	0000 1110 1000	0	00111	M
160	0000 1010 0000	0	00101	M
1024	0100 0000 0000	1	00000	M
30	0000 0001 1110	0	00000	M
140	0000 1000 1100	0	00100	H
3100	1100 0001 1100	3	00000	M
180	0000 1011 0100	0	00101	H
2180	1000 1000 0100	2	00111	M

5.3.5 4 hits in 12 accesses. Hit ratio = 4/12 = 0.33

5.5 Media applications that play audio or video ales are part of a class of workloads called "streaming" workloads; i.e., they bring in large amounts of data but do not reuse much of it. Consider a video streaming workload that accesses a 512 KiB working set sequentially with the following address stream (assuming the addresses are given as byte address):

0,2,4,6,8,10,12,14,16,…

5.5.1[5] <§§5.4, 5.8>Assume a 64 KiB direct- mapped cache with a 32-byte block. What is the miss rate for the address stream above? How is this miss rate sensitive to the size of the cache or the working set? How would you categorize the misses this workload is experiencing, based on the 3C model?

5.5.1 Assuming the addresses given as byte addresses, each group of 16 accesses will map to the same 32-byte block so the cache will have a miss rate of 1/16. For example, accesses 0, 2, 4, 6, …….30 will be mapped into the same 32-byte. All misses are compulsory misses. The miss rate is not sensitive to the size of the cache or the size of the working set. It is, however, sensitive to the access pattern and block size.

5.5.2The miss rates are 1/8, 1/32, and 1/64, respectively. The workload is exploiting spatial locality.

5.5.3 In this case the miss rate is 0.

5.5.4

Average Memory Access Time (AMAT) = (Time for a Hit) + (Miss Rate) x (Miss Latency). Since CPI is given as one, the time for a hit in this case is one cycle.

Block size (B)	Miss Rate	Latency	
8 bytes	4%	8*20=160 cycles	1+4%*160=7.4 cycles
16 bytes	3%	320 cycles	1+3%*320=10.6 cycles
32 bytes	2%	640 cycles	1+2%*640=13.8
64 bytes	1.5%	1280 cycles	1+1.5%*1280 = 20.2
128 bytes	1%	2560 cycles	1+1%*2560=26.6

Therefore, block size=16 bytes is optimal.

5.7 This exercise examines the impact of different cache designs, specifically comparing associative caches to the direct-mapped caches from Section 5.4. For these exercises, refer to the address stream shown in Exercise 5.2. (Note: the address stream is 3, 180, 43, 2, 191, 88, 190, 14, 181, 44, 186, 253. Each one is word address)

5.7.2

Each block has 1 word, and the cache would have 8/ 1 = 8 blocks.

Since this cache is fully associative and has one-word blocks, the word address is equivalent to the tag. The only possible way for there to be a hit is a repeated reference to the same word, which does not occur for this sequence.

Tag	Hit/Miss	Contents (with LRU)
3	M	3
180	M	3,180
43	M	3,180,43
2	M	3,180,43,2
191	M	3,180,43,2,191
88	M	3,180,43,2,191,88
190	M	3,180,43,2,191,88,190
14	M	3,180,43,2,191,88,190,14
181	M	181,180,43,2,191,88,190,14
44	M	181,44,43,2,191,88,190,14
186	M	181,44,186,2,191,88,190,14
253	M	181,44,186,253,191,88,190,14

5.7.3

Each block has 2 word, and the cache would have 8/ 2 = 4 blocks.

Tag has 7 bits.

Address	Binary Address	Tag	Hit/Miss	Contents (with LRU)
3	0000 0011	1	M	1
180	1011 0100	90	M	1,90
43	0010 1011	21	M	1,90,21
2	0000 0010	1	H	1,90,21
191	1011 1111	95	M	1,90,21,95
88	0101 1000	44	M	1,90,21,95,44
190	1011 1110	95	H	1,90,21,95,44
14	0000 1110	7	M	1,90,21,95,44,7
181	1011 0101	90	H	1,90,21,95,44,7
44	0010 1100	22	M	1,90,21,95,44,7,22
186	1011 1010	93	M	1,90,21,95,44,7,22,93
253	1111 1101	126	M	1,90,126,95,44,7,22,93

The final reference replaces tag 21 in the cache, since tags 1 and 90 had been reused at time=3 and time=8 while 21 hadn't been used since time=2.

Miss rate = 9/12 = 75%

Miss rate if MRU is used = 75%

This is the best possible miss rate, since there were no misses on any block that had been previously evicted from the cache. In fact, the only eviction was for tag 21, which is only referenced once.