Computer Architecture Recitation 4

29.05.2025

Question 1

In a computer system, the address bus and data bus are 20 bits and 8 bits wide, respectively. The system includes a cache memory that can hold 4 KB data.

Suppose you are running a program with the following data access pattern. The hexadecimal numbers represent the addresses of data accessed consecutively. The pattern is executed only once. In the beginning, the cache memory is empty.

\$00050, \$0005C, \$01052, \$00057, \$01054

Question 1

a) Data transfer between main memory and cache is in blocks of 16 bytes. The cache control unit uses direct mapping.

   i) Into what fields does the cache control unit divide the physical address? How many bits are there in each field? Explain.

   ii) For each address in the given data access pattern, write which event (a hit or a miss) occurs in cache memory.

b) Data transfer between main memory and cache is in blocks of 8 bytes. The cache control unit uses 2-way set associative mapping.

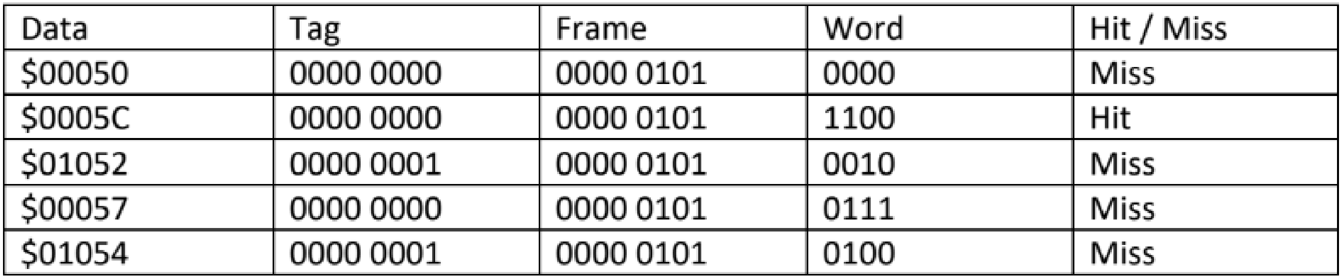
   i) Into what fields does the cache control unit divide the physical address? How many bits are there in each field? Explain.

   ii) For each address in the given data access pattern, write which event (a hit or a miss) occurs in cache memory.

Solution 1a

i) Data is transferred in blocks of 16 bytes: $16{=}2^4 \Rightarrow$ w=4 bits
4K cache memory $\Rightarrow 2^2 * 2^{10} = 2^{12}$ bytes
number of frames in cache $\Rightarrow 2^{12}/2^4{=}2^8 \Rightarrow$ f=8 bits
Tag bits $\Rightarrow$ t = a - (w+f) $\Rightarrow$ 20 - (4+8) = 8 $\Rightarrow$ t=8 bits

ii)



Data	Tag	Frame	Word	Hit / Miss
$00050	0000 0000	0000 0101	0000	Miss
$0005C	0000 0000	0000 0101	1100	Hit
$01052	0000 0001	0000 0101	0010	Miss
$00057	0000 0000	0000 0101	0111	Miss
$01054	0000 0001	0000 0101	0100	Miss

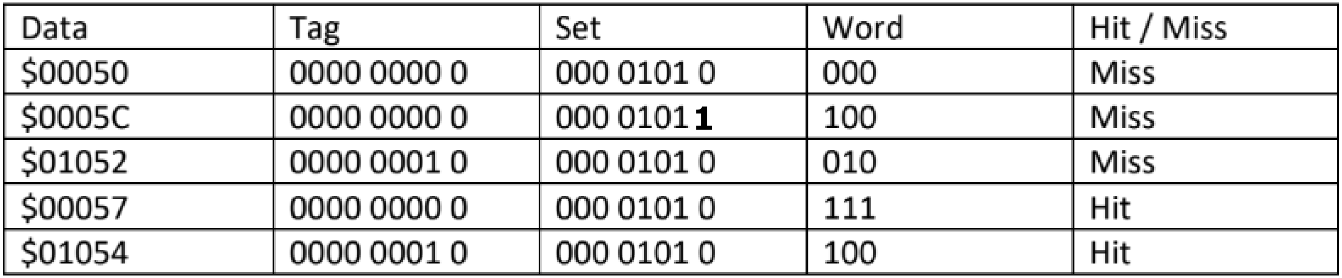
Solution 1b

i) Data is transferred in blocks of 8 bytes: $8=2^3 \Rightarrow w=3$ bits
4K cache memory $\Rightarrow 2^2 * 2^{10} = 2^{12}$ bytes
number of frames in cache $\Rightarrow 2^{12}/2^3=2^9 \Rightarrow f=9$ bits
2^9 frames, each set contains 2 frames $\Rightarrow 2^9/2 = 2^8$ sets in cache memory $\Rightarrow$ set num length = 8 bits
Tag length $\Rightarrow t = a - (w+s) \Rightarrow 20 - (3+8) = 9 \Rightarrow t=9$ bits

ii)

!!!SET OF SECOND ADDRESS WAS GIVEN WRONG DURING RECITATION!!!

~~000 0101 0~~ - -> 000 0101 1



Data	Tag	Set	Word	Hit / Miss
$00050	0000 0000 0	000 0101 0	000	Miss
$0005C	0000 0000 0	000 0101 **1**	100	Miss
$01052	0000 0001 0	000 0101 0	010	Miss
$00057	0000 0000 0	000 0101 0	111	Hit
$01054	0000 0001 0	000 0101 0	100	Hit

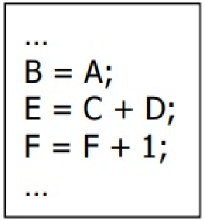
Question 2

A single-CPU computer system has 16Mi bytes of main memory and cache memory that is used only for data, not for instructions. Prefix Mi represents Megabinary as explained in Chapter 1 of lecture notes.

- Data transfer between main memory and cache in blocks/frames of 64 bytes
- Cache memory can hold 256 frames.
- The cache control unit uses 2-way set associative mapping.
- For write operations, Simple Write Back (SWB) with Write Allocate (WA) is used.
- When necessary, LRU is used as the replacement algorithm. If a set is empty, assume that frame 0 is older than frame 1.

Question 2

The CPU runs the piece of pseudocode given on the below.



```
...
B = A;
E = C + D;
F = F + 1;
...
```

- Variable A is read and written to variable B.
- Variables C and D are read, and their sum is written to the variable E.
- The variable F is read, incremented, and the result is written to the same variable.
- Each variable is one byte.
- The addresses of the variables are given below (Symbol $ denotes hexadecimal): A:$000122, B:$002100, C:$00013C, D:$00013D, E:$00013E, F:$006100
- Initially, none of these variables are in the cache memory.

Question 2

a) Into what fields does the cache control unit divide the main memory address? Give the lengths of the fields.

b) Consider the given code segments. For each line,

   i) Which set/frames of cache memory does the cache control unit place variables into? Give your answers in decimal. (Example: “set number: 73, frame: 0” or “set number 85, frame: 1”)

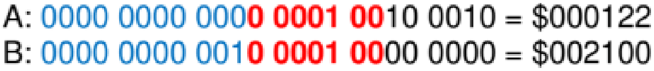
   ii) How many read misses, read hits, write misses, write hits, and block transfers occur during the run of these statements?

Solution 2a

Main Memory: 16 MiB = 2^4 x 2^{20} = 2^{24} bytes
Data transfer in 64 bytes = 2^6 bytes
Number of frames in cache = 256 frames = 2^8 frames
256 frames, each set 2 frames = 256/2 = 128 = 2^7 sets
Word number length = 6 bits
Set number = 7 bits
Tag length = 24 - (6+7) = 11 bits

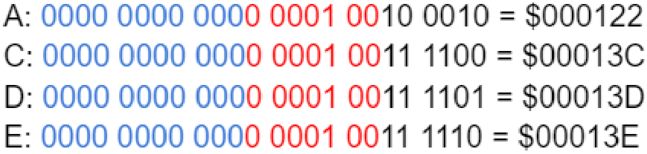
Solution 2a

i) Starting with line B=A. Initially, neither A or B is in cache memory.



A: 0000 0000 0000 0001 0010 0010 = $000122
B: 0000 0000 0010 0001 0000 0000 = $002100

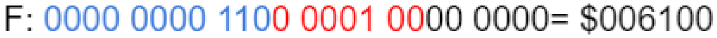
Read A: A is placed into set number: 4, frame: 0 because frame 0 is older. (LRU) Write to B: B is placed into set number 4, frame: 1 because of Write Allocate (WA) method.
For line E=C+D, A, C, D and E all map to the same set and they all have the same tag value. They are all in the same block of main memory. Since this block has been copied into the cache for A, variables C, D and E are already in the cache. C, D and E are in set number: 4, frame: 0.



A: 0000 0000 0000 0001 0010 0010 = $000122
C: 0000 0000 0000 0001 0011 1100 = $00013C
D: 0000 0000 0000 0001 0011 1101 = $00013D
E: 0000 0000 0000 0001 0011 1110 = $00013E

Solution 2b

For line F=F+1, due to the previous line E=C+D, frame 0 is newer than frame 1. B is in set number: 4, frame: 1 so B is replaced. F is placed into set number: 4, frame: 1.



F: 0000 0000 1100 0001 0000 0000= $006100

ii) **Again, starting with B=A.**
Read of A will be a miss. So, 1 read miss for A.
The frame 0 of set 4 of cache will be replaced by A. The replaced frame will be written back to the main memory because of SWB. (block transfer). If we assume that the frame 0 in cache is empty (not valid), it will not be written back to main memory.
The block that contains A will be brought in from main memory to cache. (block transfer)

Solution 2b

There will be a write miss for B.
The replaced frame will be written back to the main memory because of SWB. (block transfer). If we assume that the frame 1 in cache is empty (not valid), it will not be written back to main memory.
The block that contains B will be brought in from main memory to cache because of WA. (block transfer)

For line E=C+D,
Read C and read D will be hits since they are already in the cache. For E, there will be 1 write hit. Since WB method is used, data is not written to the main memory.

Solution 2b

For line F=F+1,

Read F will be a miss, so 1 read miss occurs. and it will be brought to the cache.

As the writing operation is done after reading, the block will already be in the cache, so there will be 1 write hit.

The replaced frame is always written back to the main memory. (SWB) It is not checked whether the frame is changed or not. 1 block transfer for F and 1 block transfer for B.

Then,

Number of read misses: 2 Number of read hits: 2

Number of write misses: 1 Number of write hits: 2

Number of block transfers: 6

Question 3

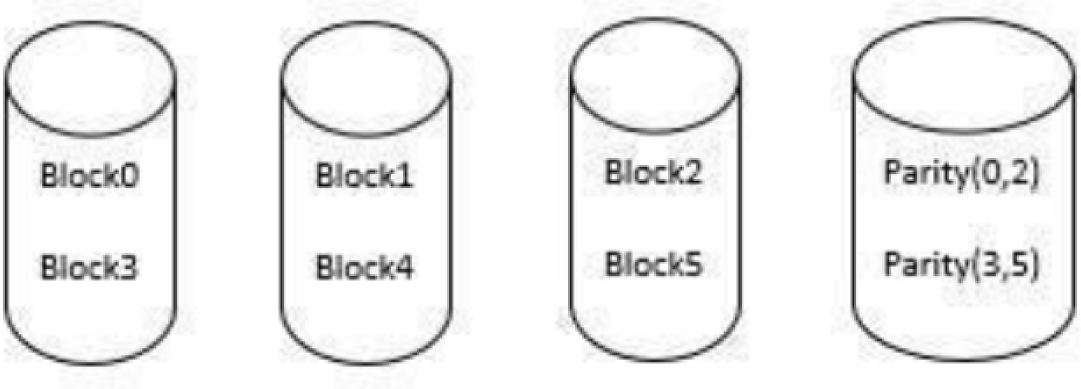
a) Draw a RAID 4 system with a total of 4 disks (data + parity), and distribute 6 blocks (block 0 – block 5) over the disks. Assume that the access time for each disk is t_a.

i) How long does it take to read words from two blocks (e.g., block 0 and block 4) in two different disks?

ii) How long does it take to update (write) words of two blocks (e.g., block 0 and block 4) in two different disks? Explain.

b) Answer the questions in part (a) (i and ii) for the RAID 5 system.

Solution 3a


Block0
Block3
Block1
Block4
Block2
Block5
Parity(0,2)
Parity(3,5)

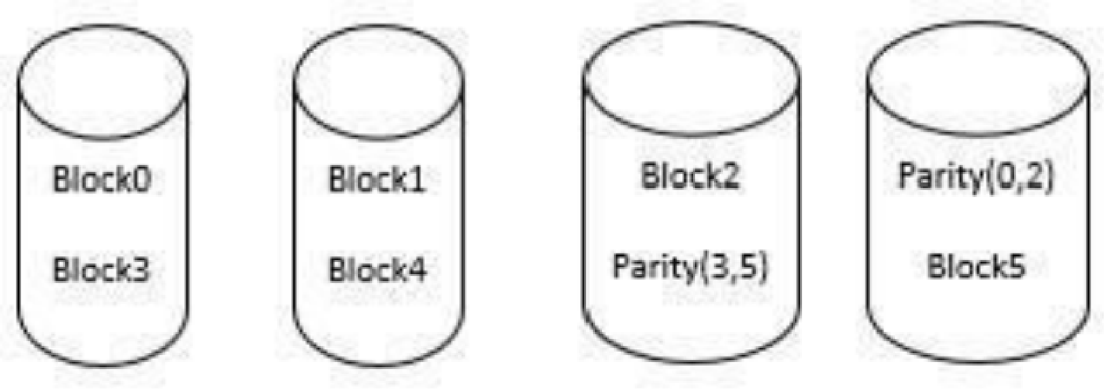
Solution 3a

i) **Two different blocks from two different disks can be read in $\mathbf{t_a}$.**

ii) Two read and two write operations (**$\mathbf{2t_a}$**) should be performed for an update operation in RAID 4 (see the lecture notes). Since parity update operations cannot be performed independently (in parallel) (there is only one parity disk), it takes **$\mathbf{4t_a}$** to update words of two blocks in two different disks.

Updating Block0:	Read Block0 and Parity(0,2):	$\mathbf{t_a}$
	Update Block0 and Parity(0,2):	$\mathbf{t_a}$
	Total: $\mathbf{2t_a}$	
Updating Block4:	Read Block4 and Parity(3,5):	$\mathbf{t_a}$
	Update Block4 and Parity(3,5):	$\mathbf{t_a}$
	Total: $\mathbf{2t_a}$	
	Total: $\mathbf{4t_a}$	

Solution 3b


Block0
Block3
Block1
Block4
Block2
Parity(3,5)
Parity(0,2)
Block5

Solution 3b)

i) **Same as in RAID 4: $\mathbf{t_a}$.**

ii) For each data update, two read and two write operations are necessary. Different from RAID 4, now parity update operations can be performed in parallel, because parity strips are distributed to different disks: $\mathbf{2t_a}$.

Updating Block0:	Read Block0 and Parity(0,2):	$\mathbf{t_a}$
	Update Block0 and Parity(0,2):	$\mathbf{t_a}$
	Total:	$\mathbf{2t_a}$
Updating Block4:	Read Block4 and Parity(3,5):	$\mathbf{t_a}$
	Update Block4 and Parity(3,5):	$\mathbf{t_a}$
	Total:	$\mathbf{2t_a}$

For two update operations, different disks are accessed, so these operations can be performed in parallel.

Total: $\mathbf{2t_a}$

Question 3

In a symmetric multiprocessor (SMP) system using a shared bus:

- There are two CPUs (CPU1 and CPU2) that have local cache memories.
- The system does not have a shared L2 cache.
- Main memory size is 64K words.
- Each local cache memory can store 4K words of data.
- Data transfer between main memory and cache memories is in blocks of 16 words.
- The cache control unit uses direct mapping.
- Cache memory is used only for data, not for instructions.
- For write operations, write back (WB) method is used.
- To provide cache coherence, the snoopy MESI protocol is used.
- There is a shared array A in the system.
- The starting address of array A in main memory is \$000C and its size is 10 words.

Question 4

a) Assume that the valid copy of array A is residing in main memory and in cache memory of CPU2. The cache memory of CPU1 is empty. A program running in CPU1 clears the array by assigning zero to all elements using a loop.

   i) Which control messages are sent by the MESI cache controllers during the run of the loop? Write the messages in the order they are sent.

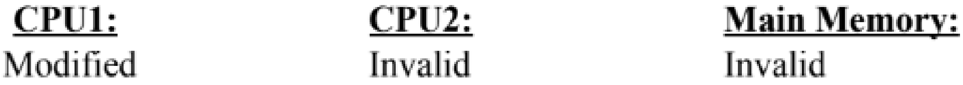
   ii) What are the MESI states of the corresponding frames storing array A in the caches of the CPUs? Is the value in main memory valid after this clear operation?

b) After the clear/write operation of CPU1, CPU2 uses a loop to read all elements of array A. Which control messages are sent by the MESI cache controllers during the run of the loop? Write the messages in the order they are sent.

Solution 4a

i) Array A (A[0], ... , A[9]) address range: $000C-$001F
Array A occupies 2 frames (Frame 0 and Frame 1) in each cache memory.
A[0] ← 0; write miss
1. Control message from CPU1: ”read-with-intent-to-modify”
A[1] ← 0, ..., A[3] ← 0; write hit; No message
A[4] ← 0; write miss
2. Control message from CPU1: ”read-with-intent-to-modify”
A[5] ← 0, ..., A[9] ← 0; write hit; No message

ii)



CPU1:	CPU2:	Main Memory:
Modified	Invalid	Invalid

Solution 4b

Read A[0]; read miss

1. Control message from CPU2: ”read”

CPU1 blocks the read operation, writes Frame 0 back to main memory.

2. Control message from CPU1: ”Shared”

Read A[1]-A[3]; read hit; no message

Read A[4]; read miss;

3. Control message from CPU2: ”read”

CPU1 blocks the read operation, writes Frame 1 back to main memory.

4. Control message from CPU1: ”Shared”

Read A[5]-A[9]; read hit; no message