一、 select the bast one.　(2 * 35 = 70)

1. Which of the following hazards is NOT introduced by the pipeline technique?

A. Structure Hazard　　B. Exception Hazard

C. Data Hazard　　D. Control Hazard

（B）

2. For a RISC-V processor that implements precise exceptions and runs in M-mode (machine mode), which one of the following statements is wrong?

A. Instructions before the faulting instruction can complete

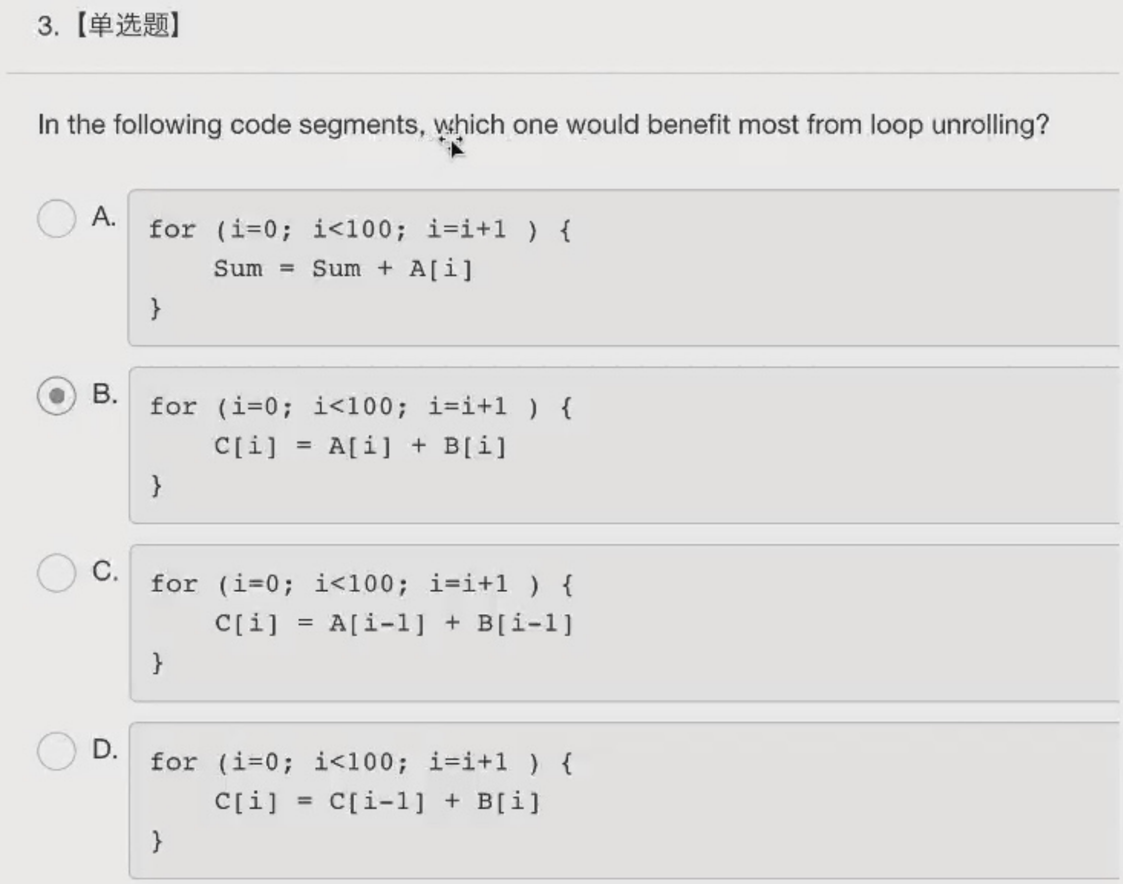
B. Instructions after the faulting instruction might be restarted later

C. Address of the instruction that caused the exception will be written to the mtvec register

D. The mtval register can be used to store exception-specific information like the faulting virtual address

(C)

3.



3. 【单选题】

In the following code segments, which one would benefit most from loop unrolling?

○ A.

```
for (i=0; i<100; i=i+1 ) {
    Sum = Sum + A[i]
}
```

● B.

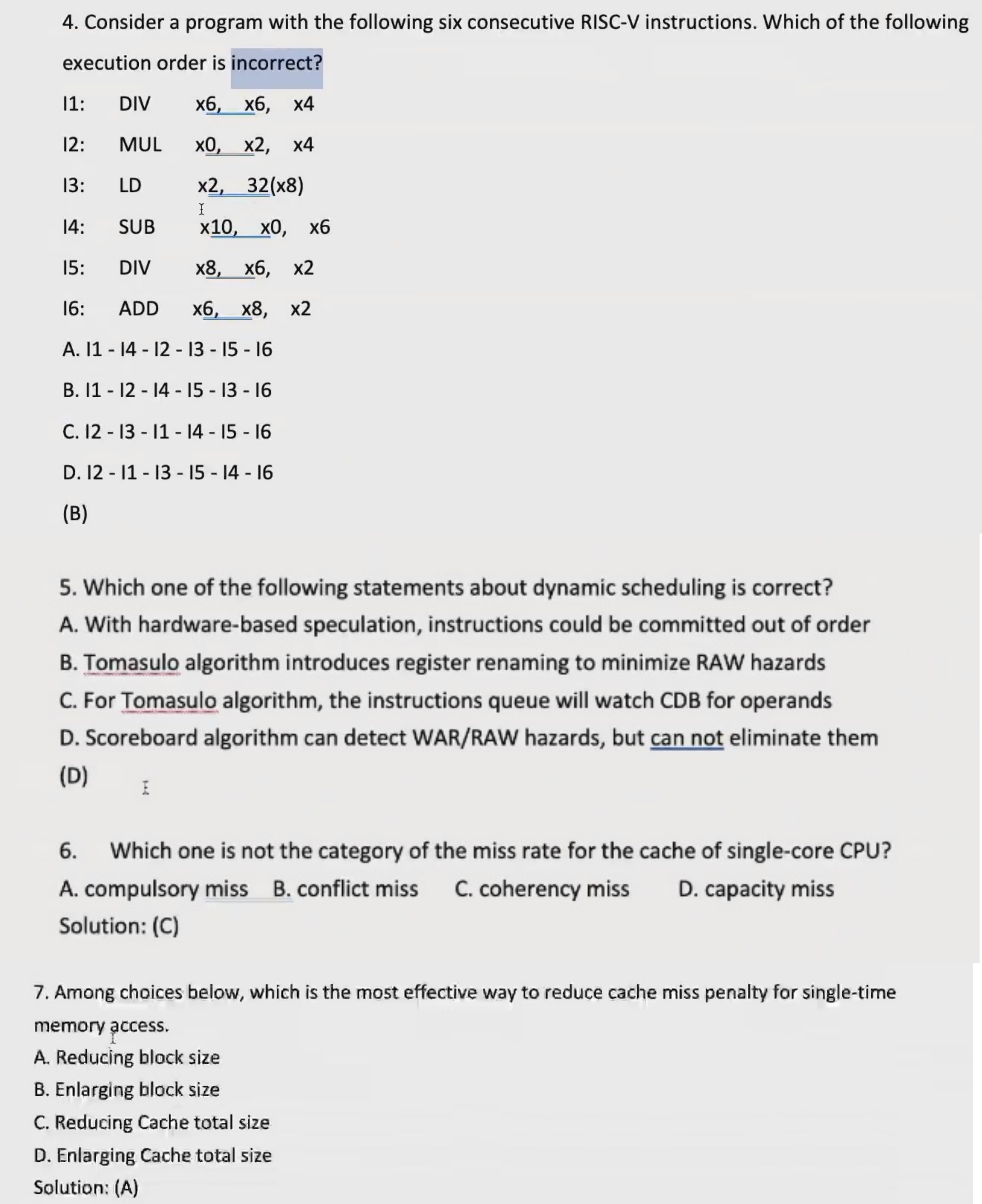
```
for (i=0; i<100; i=i+1 ) {
    C[i] = A[i] + B[i]
}
```

○ C.

```
for (i=0; i<100; i=i+1 ) {
    C[i] = A[i-1] + B[i-1]
}
```

○ D.

```
for (i=0; i<100; i=i+1 ) {
    C[i] = C[i-1] + B[i]
}
```



4. Consider a program with the following six consecutive RISC-V instructions. Which of the following execution order is incorrect?

```
I1:    DIV     x6,  x6,  x4
I2:    MUL     x0,  x2,  x4
I3:    LD      x2,  32(x8)
I4:    SUB     x10, x0,  x6
I5:    DIV     x8,  x6,  x2
I6:    ADD     x6,  x8,  x2
```

A. I1 - I4 - I2 - I3 - I5 - I6

B. I1 - I2 - I4 - I5 - I3 - I6

C. I2 - I3 - I1 - I4 - I5 - I6

D. I2 - I1 - I3 - I5 - I4 - I6

(B)

5. Which one of the following statements about dynamic scheduling is correct?

A. With hardware-based speculation, instructions could be committed out of order

B. Tomasulo algorithm introduces register renaming to minimize RAW hazards

C. For Tomasulo algorithm, the instructions queue will watch CDB for operands

D. Scoreboard algorithm can detect WAR/RAW hazards, but can not eliminate them

(D)

6. Which one is not the category of the miss rate for the cache of single-core CPU?

A. compulsory miss B. conflict miss C. coherency miss D. capacity miss

Solution: (C)

7. Among choices below, which is the most effective way to reduce cache miss penalty for single-time memory access.

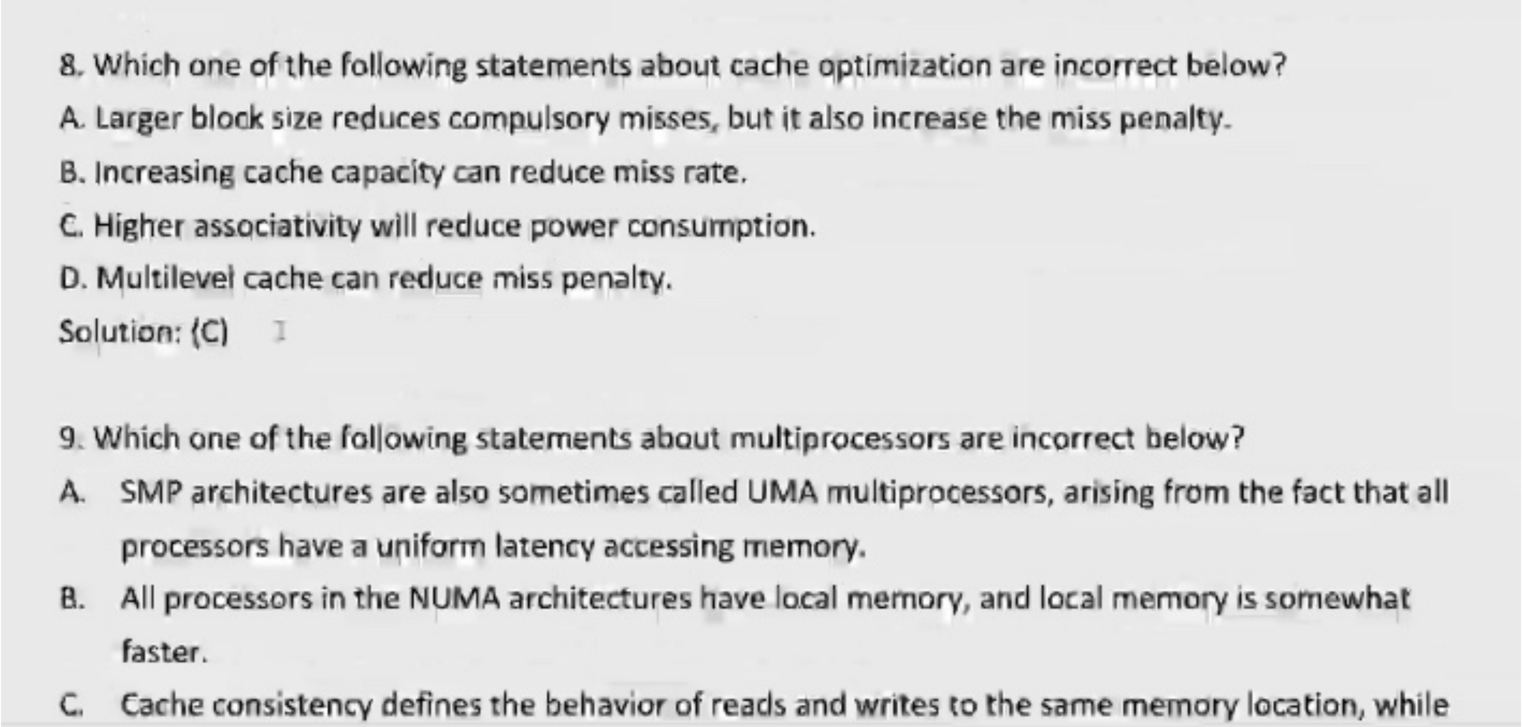
A. Reducing block size

B. Enlarging block size

C. Reducing Cache total size

D. Enlarging Cache total size

Solution: (A)



8. Which one of the following statements about cache optimization are incorrect below?

A. Larger block size reduces compulsory misses, but it also increase the miss penalty.

B. Increasing cache capacity can reduce miss rate.

C. Higher associativity will reduce power consumption.

D. Multilevel cache can reduce miss penalty.

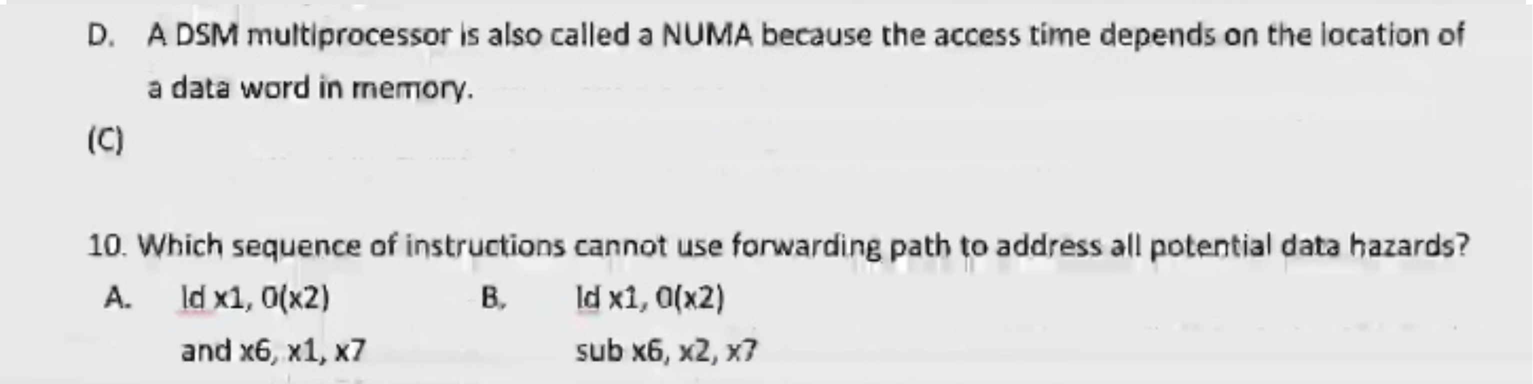
Solution: (C)

9. Which one of the following statements about multiprocessors are incorrect below?

A. SMP architectures are also sometimes called UMA multiprocessors, arising from the fact that all processors have a uniform latency accessing memory.

B. All processors in the NUMA architectures have local memory, and local memory is somewhat faster.

C. Cache consistency defines the behavior of reads and writes to the same memory location, while



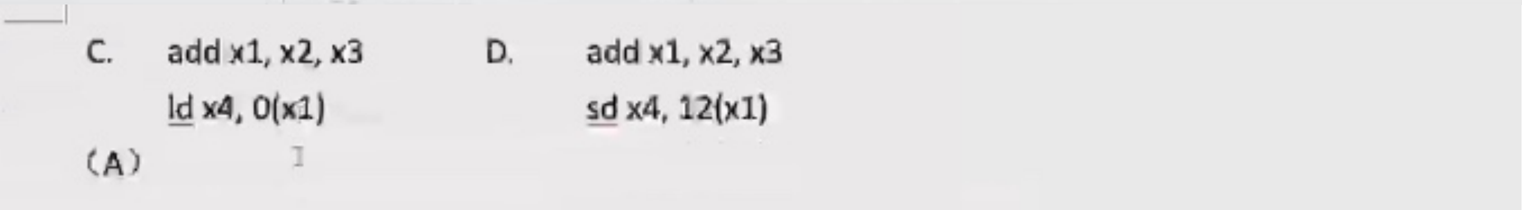
D. A DSM multiprocessor is also called a NUMA because the access time depends on the location of a data word in memory.

(C)

10. Which sequence of instructions cannot use forwarding path to address all potential data hazards?

A. ld x1, 0(x2)
and x6, x1, x7

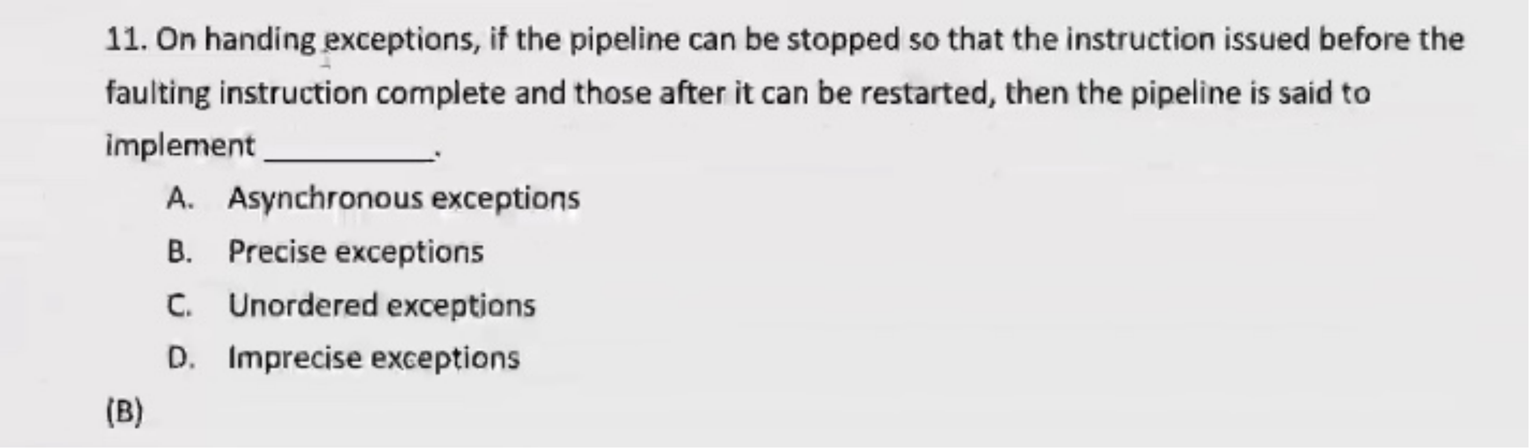
B. ld x1, 0(x2)
sub x6, x2, x7



C. add x1, x2, x3
ld x4, 0(x1)

D. add x1, x2, x3
sd x4, 12(x1)

（A）



11. On handing exceptions, if the pipeline can be stopped so that the instruction issued before the faulting instruction complete and those after it can be restarted, then the pipeline is said to implement __________.

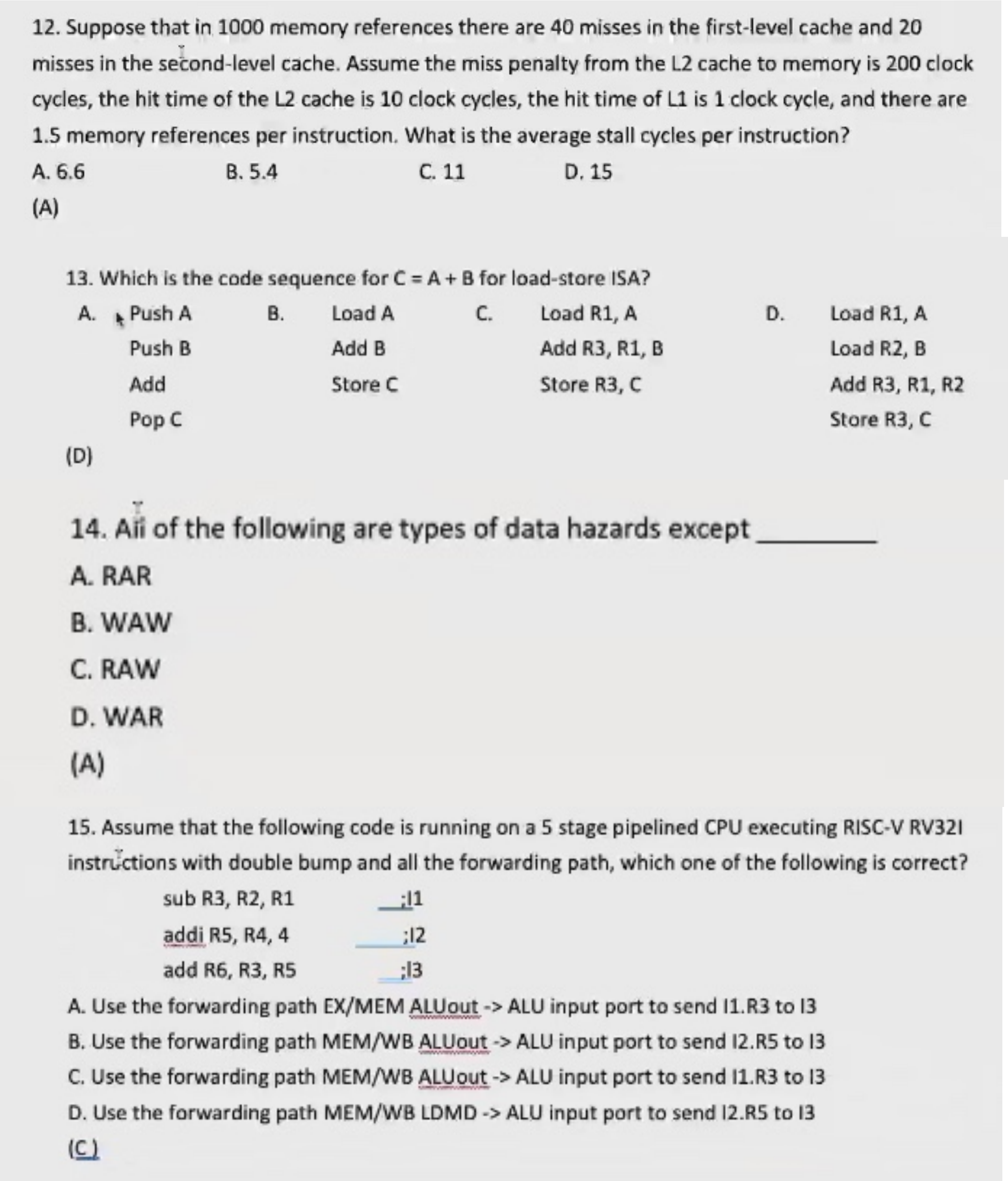
A. Asynchronous exceptions

B. Precise exceptions

C. Unordered exceptions

D. Imprecise exceptions

(B)



12. Suppose that in 1000 memory references there are 40 misses in the first-level cache and 20 misses in the second-level cache. Assume the miss penalty from the L2 cache to memory is 200 clock cycles, the hit time of the L2 cache is 10 clock cycles, the hit time of L1 is 1 clock cycle, and there are 1.5 memory references per instruction. What is the average stall cycles per instruction?

A. 6.6 B. 5.4 C. 11 D. 15

(A)

13. Which is the code sequence for C = A + B for load-store ISA?

A.	Push A	B.	Load A	C.	Load R1, A	D.	Load R1, A
	Push B		Add B		Add R3, R1, B		Load R2, B
	Add		Store C		Store R3, C		Add R3, R1, R2
	Pop C						Store R3, C

(D)

14. All of the following are types of data hazards except ________

A. RAR

B. WAW

C. RAW

D. WAR

(A)

15. Assume that the following code is running on a 5 stage pipelined CPU executing RISC-V RV32I instructions with double bump and all the forwarding path, which one of the following is correct?

```
sub R3, R2, R1          ;I1
addi R5, R4, 4          ;I2
add R6, R3, R5          ;I3
```

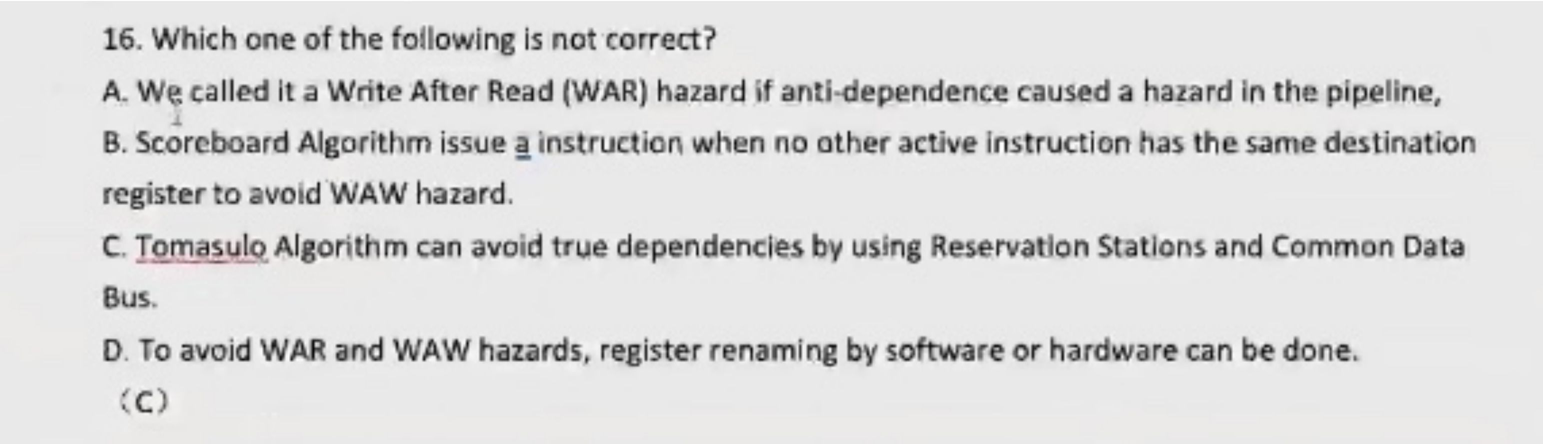
A. Use the forwarding path EX/MEM ALUout -> ALU input port to send I1.R3 to I3

B. Use the forwarding path MEM/WB ALUout -> ALU input port to send I2.R5 to I3

C. Use the forwarding path MEM/WB ALUout -> ALU input port to send I1.R3 to I3

D. Use the forwarding path MEM/WB LDMD -> ALU input port to send I2.R5 to I3

(C)



16. Which one of the following is not correct?

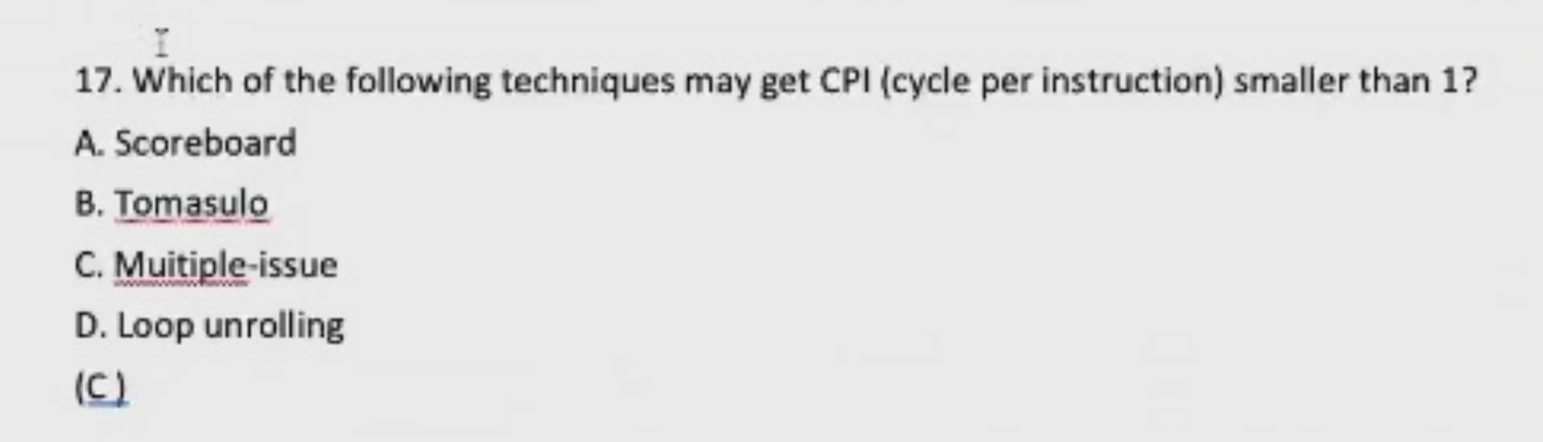
A. We called it a Write After Read (WAR) hazard if anti-dependence caused a hazard in the pipeline,

B. Scoreboard Algorithm issue a instruction when no other active instruction has the same destination register to avoid WAW hazard.

C. Tomasulo Algorithm can avoid true dependencies by using Reservation Stations and Common Data Bus.

D. To avoid WAR and WAW hazards, register renaming by software or hardware can be done.

（C）



17. Which of the following techniques may get CPI (cycle per instruction) smaller than 1?

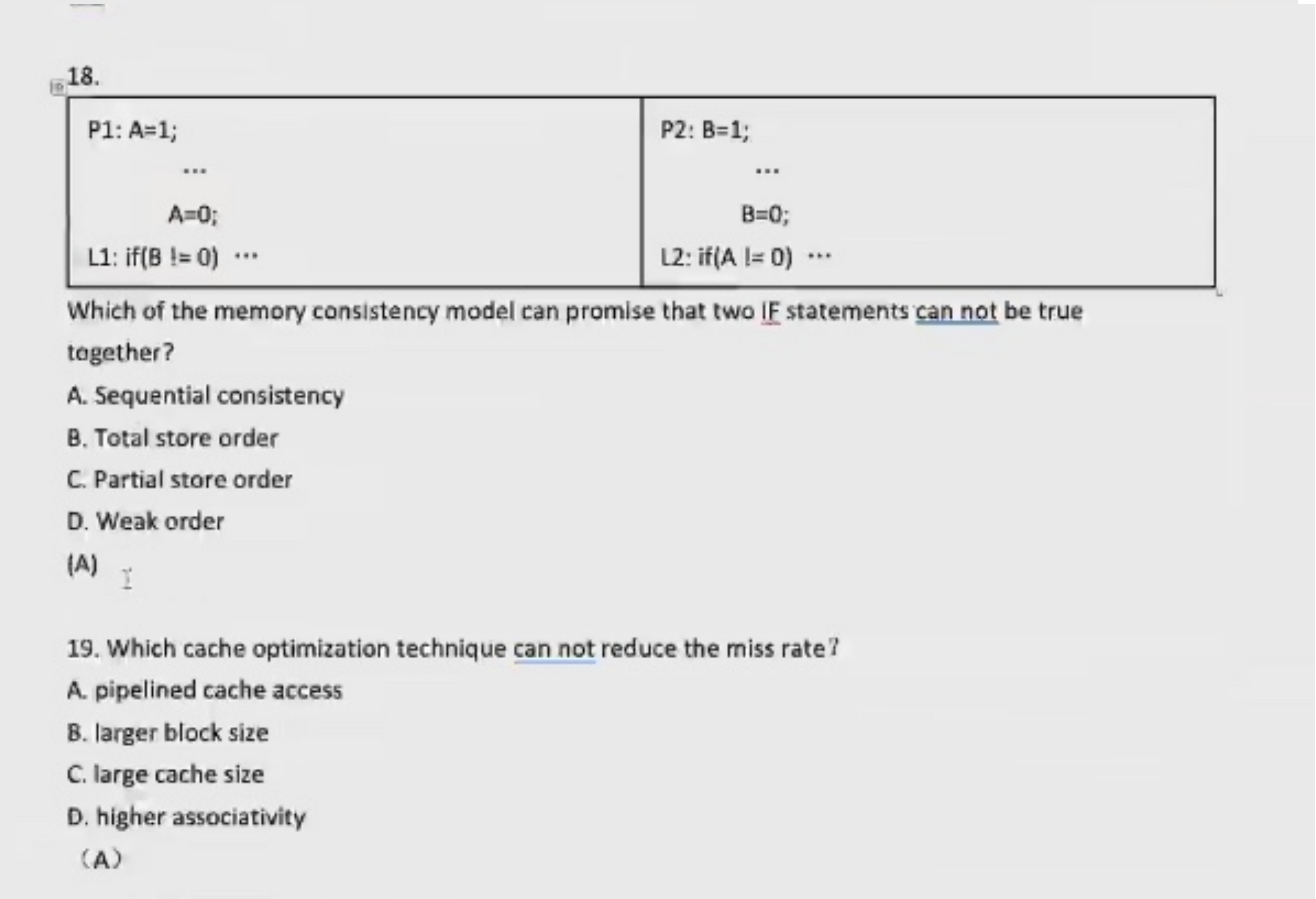
A. Scoreboard

B. Tomasulo

C. Muitiple-issue

D. Loop unrolling

(C)



18.

P1: A=1;	P2: B=1;
…	…
A=0;	B=0;
L1: if(B != 0) …	L2: if(A != 0) …

Which of the memory consistency model can promise that two IF statements can not be true together?

A. Sequential consistency

B. Total store order

C. Partial store order

D. Weak order

(A)

19. Which cache optimization technique can not reduce the miss rate?

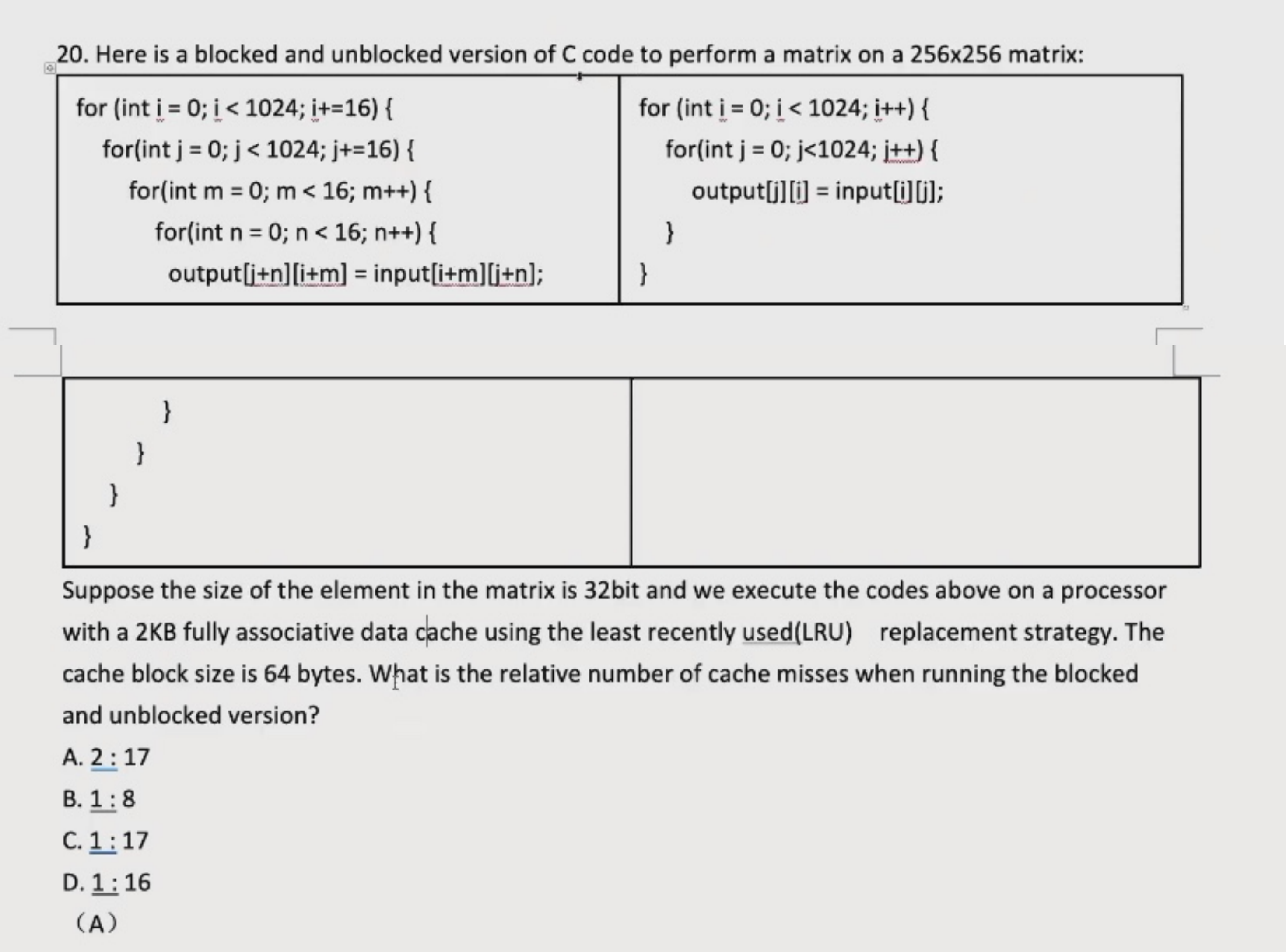
A. pipelined cache access

B. larger block size

C. large cache size

D. higher associativity

（A）



20. Here is a blocked and unblocked version of C code to perform a matrix on a 256x256 matrix:

```
for (int i = 0; i < 1024; i+=16) {
   for(int j = 0; j < 1024; j+=16) {
      for(int m = 0; m < 16; m++) {
         for(int n = 0; n < 16; n++) {
           output[j+n][i+m] = input[i+m][j+n];
         }
      }
   }
}
```

```
for (int i = 0; i < 1024; i++) {
   for(int j = 0; j<1024; j++) {
      output[j][i] = input[i][j];
   }
}
```

Suppose the size of the element in the matrix is 32bit and we execute the codes above on a processor with a 2KB fully associative data cache using the least recently used(LRU)   replacement strategy. The cache block size is 64 bytes. What is the relative number of cache misses when running the blocked and unblocked version?

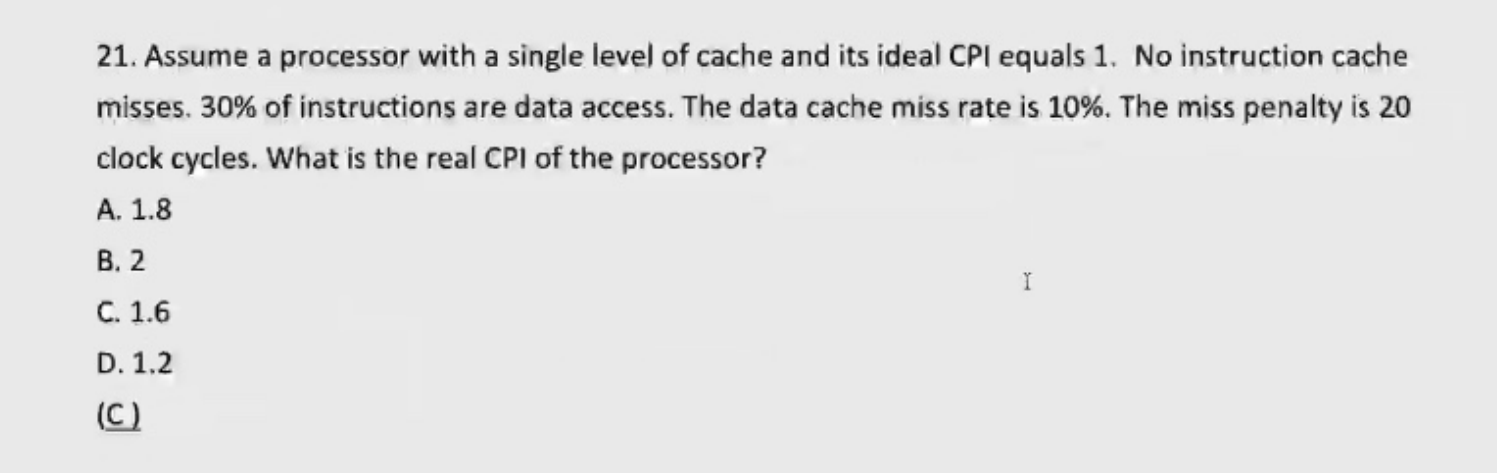
A. 2 : 17

B. 1 : 8

C. 1 : 17

D. 1 : 16

（A）



21. Assume a processor with a single level of cache and its ideal CPI equals 1. No instruction cache misses. 30% of instructions are data access. The data cache miss rate is 10%. The miss penalty is 20 clock cycles. What is the real CPI of the processor?

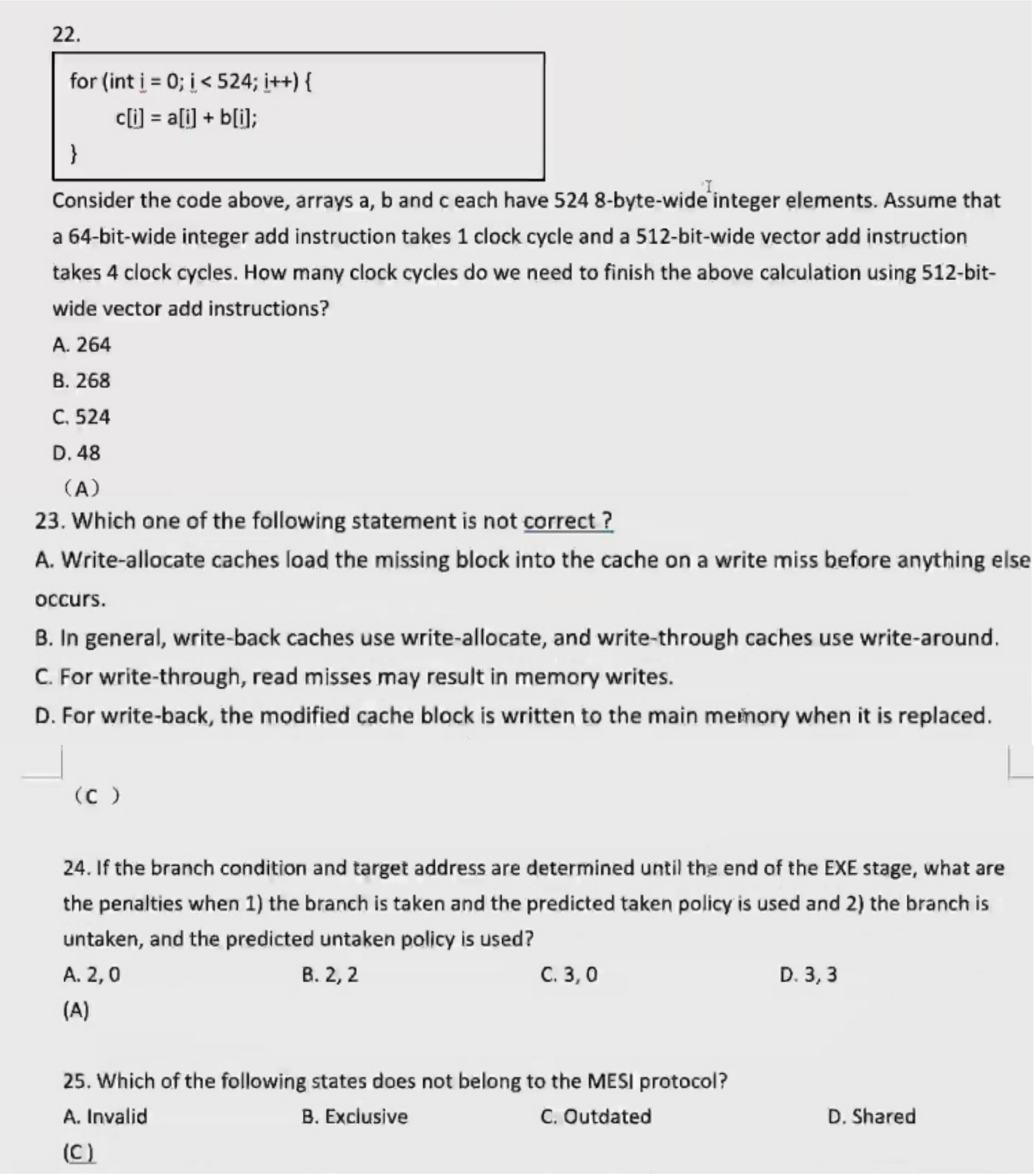
A. 1.8

B. 2

C. 1.6

D. 1.2

(C)



22.

```
for (int i = 0; i < 524; i++) {
        c[i] = a[i] + b[i];
}
```

Consider the code above, arrays a, b and c each have 524 8-byte-wide integer elements. Assume that a 64-bit-wide integer add instruction takes 1 clock cycle and a 512-bit-wide vector add instruction takes 4 clock cycles. How many clock cycles do we need to finish the above calculation using 512-bit-wide vector add instructions?

A. 264

B. 268

C. 524

D. 48

（A）

23. Which one of the following statement is not correct ?

A. Write-allocate caches load the missing block into the cache on a write miss before anything else occurs.

B. In general, write-back caches use write-allocate, and write-through caches use write-around.

C. For write-through, read misses may result in memory writes.

D. For write-back, the modified cache block is written to the main memory when it is replaced.

（C ）

24. If the branch condition and target address are determined until the end of the EXE stage, what are the penalties when 1) the branch is taken and the predicted taken policy is used and 2) the branch is untaken, and the predicted untaken policy is used?

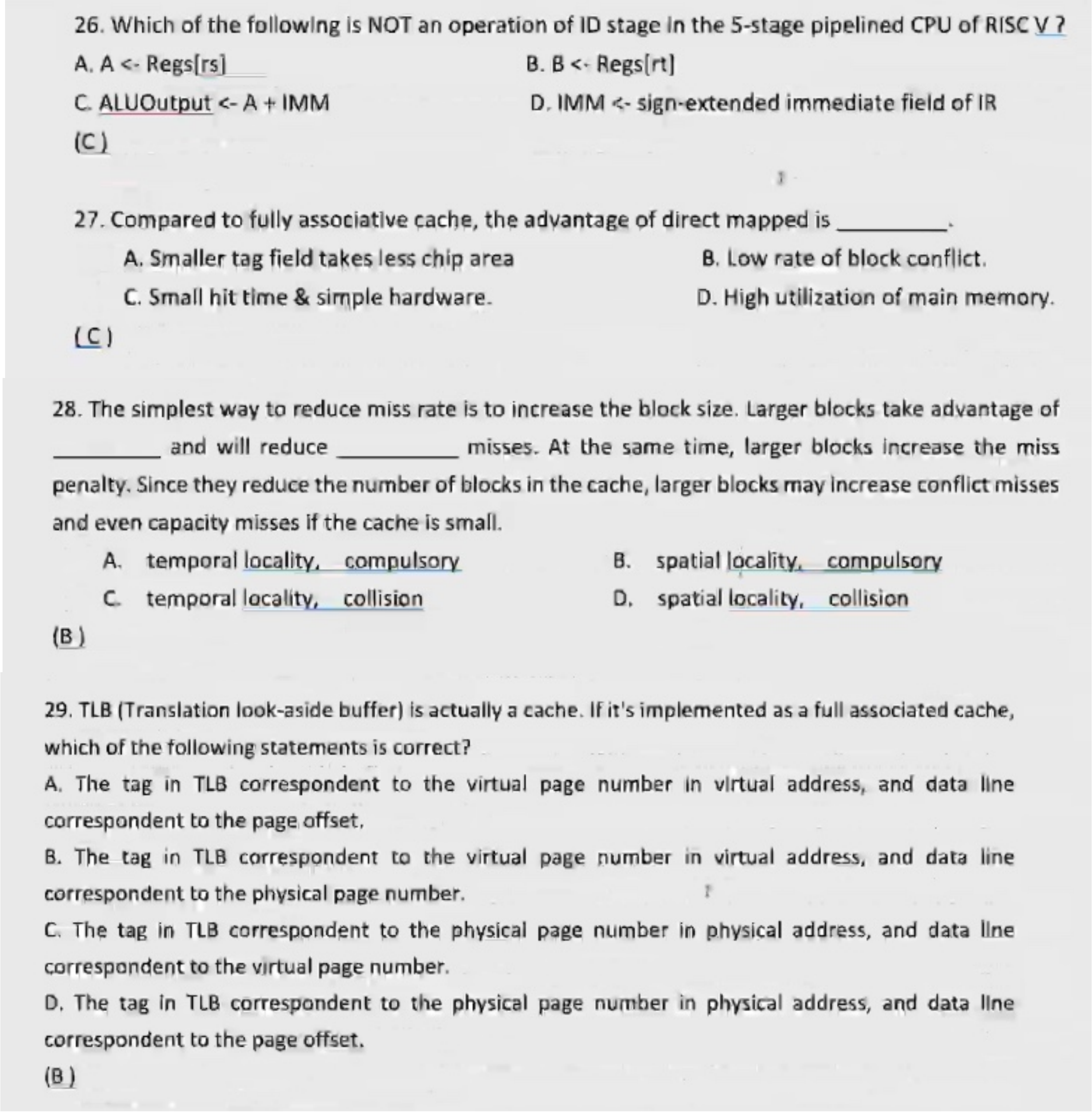
A. 2, 0　　B. 2, 2　　C. 3, 0　　D. 3, 3

(A)

25. Which of the following states does not belong to the MESI protocol?

A. Invalid　　B. Exclusive　　C. Outdated　　D. Shared

(C)



26. Which of the following is NOT an operation of ID stage in the 5-stage pipelined CPU of RISC V ?

A. A <- Regs[rs] B. B <- Regs[rt]

C. ALUOutput <- A + IMM D. IMM <- sign-extended immediate field of IR

(C)

27. Compared to fully associative cache, the advantage of direct mapped is _________.

A. Smaller tag field takes less chip area B. Low rate of block conflict.

C. Small hit time & simple hardware. D. High utilization of main memory.

(C)

28. The simplest way to reduce miss rate is to increase the block size. Larger blocks take advantage of _________ and will reduce __________ misses. At the same time, larger blocks increase the miss penalty. Since they reduce the number of blocks in the cache, larger blocks may increase conflict misses and even capacity misses if the cache is small.

A. temporal locality, compulsory B. spatial locality, compulsory

C. temporal locality, collision D. spatial locality, collision

(B)

29. TLB (Translation look-aside buffer) is actually a cache. If it's implemented as a full associated cache, which of the following statements is correct?

A. The tag in TLB correspondent to the virtual page number in virtual address, and data line correspondent to the page offset.

B. The tag in TLB correspondent to the virtual page number in virtual address, and data line correspondent to the physical page number.

C. The tag in TLB correspondent to the physical page number in physical address, and data line correspondent to the virtual page number.

D. The tag in TLB correspondent to the physical page number in physical address, and data line correspondent to the page offset.

(B)

30. Consider the following description of a memory hierarchy.

Virtual address wide = 46 bits, Memory physical address wide = 36 bits, Page size = 8KB. Cache capacity =16KB.

If it's a direct mapped cache, then tag in the 16KB cache is ______ bits.

A. 25 B.24 C.23 D. 22

(D)

31. Consider the following description of a memory hierarchy.

Virtual address wide = 46 bits, Memory physical address wide = 36 bits, Page size = 8KB. Cache capacity =16KB.

If it's a 8-way physical addressed cache, then tag in the 16KB cache is ______ bits.

A. 25 B.26 C.27 D. 28

(A)

32. Multiple-issue processors have two basic types: ______ and ______, the purpose of which is to pursue CPI<1. The difference is that ______ is the hardware-intensive technique using hardware to do hazard detection, while ______ use compiler-based approaches using primarily software to detect hazard.

A. Multiprocessor, superscalar; Superscalar, multiprocessor

B, VLIW, superscalar; VLIW, superscalar

C. Superscalar, VLIW; Superscalar, VLIW

D, superscalar, VLIW; VLIW, superscalar

(C)

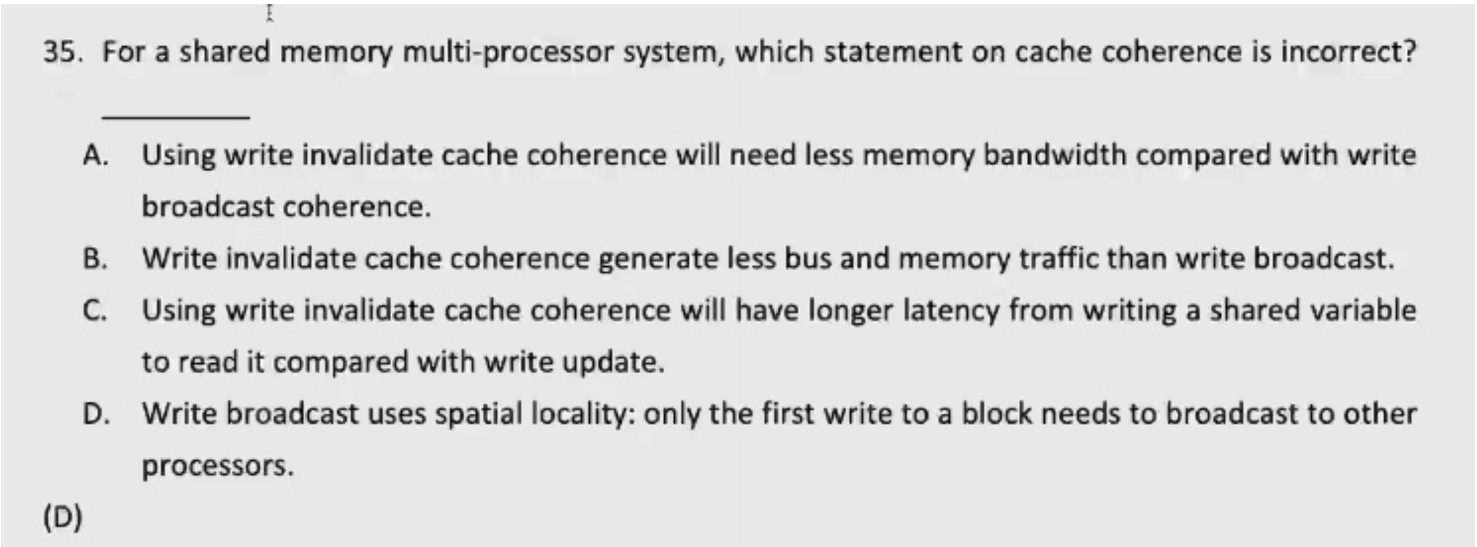
33. Consider the following description of a memory hierarchy.

Virtual address wide = 46 bits, Memory physical address wide = 36 bits, Page size = 8KB. Cache capacity =16KB.

If it's a virtual indexed and physical tagged cache, then tag in a 2-way associative cache is ______ bits.

A. 26 B.25 C.24 D. 23

(D)



35. For a shared memory multi-processor system, which statement on cache coherence is incorrect? ________

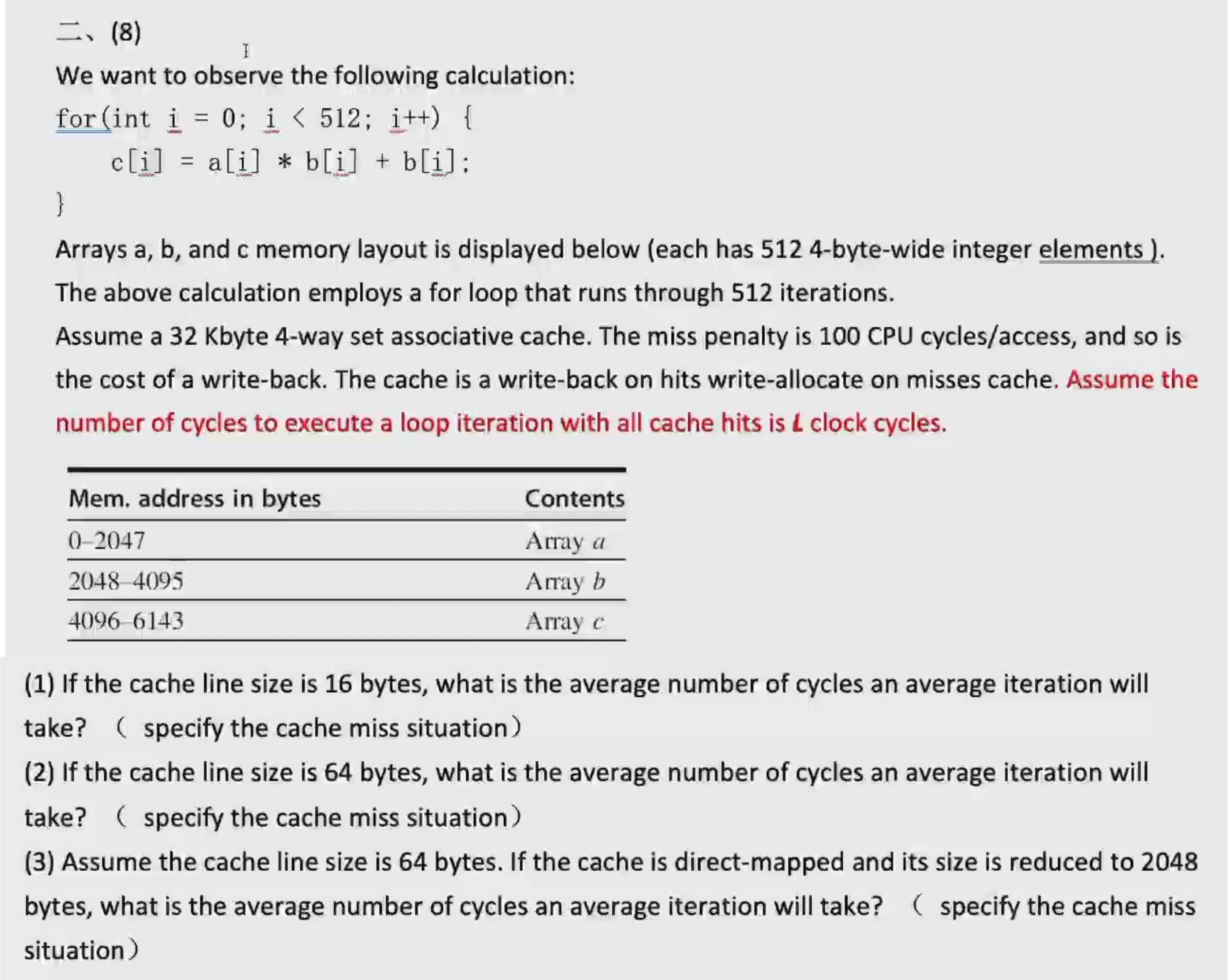
A. Using write invalidate cache coherence will need less memory bandwidth compared with write broadcast coherence.

B. Write invalidate cache coherence generate less bus and memory traffic than write broadcast.

C. Using write invalidate cache coherence will have longer latency from writing a shared variable to read it compared with write update.

D. Write broadcast uses spatial locality: only the first write to a block needs to broadcast to other processors.

(D)



二、(8)

We want to observe the following calculation:

```
for(int i = 0; i < 512; i++) {
    c[i] = a[i] * b[i] + b[i];
}
```

Arrays a, b, and c memory layout is displayed below (each has 512 4-byte-wide integer elements). The above calculation employs a for loop that runs through 512 iterations.

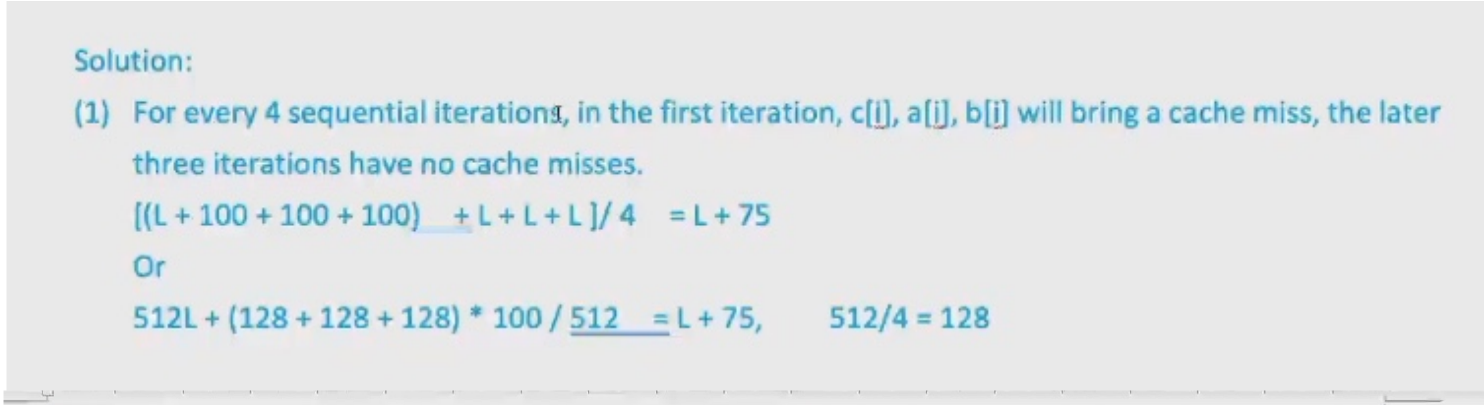
Assume a 32 Kbyte 4-way set associative cache. The miss penalty is 100 CPU cycles/access, and so is the cost of a write-back. The cache is a write-back on hits write-allocate on misses cache. Assume the number of cycles to execute a loop iteration with all cache hits is ***L*** clock cycles.

Mem. address in bytes	Contents
0–2047	Array *a*
2048–4095	Array *b*
4096–6143	Array *c*

(1) If the cache line size is 16 bytes, what is the average number of cycles an average iteration will take? （ specify the cache miss situation）

(2) If the cache line size is 64 bytes, what is the average number of cycles an average iteration will take? （ specify the cache miss situation）

(3) Assume the cache line size is 64 bytes. If the cache is direct-mapped and its size is reduced to 2048 bytes, what is the average number of cycles an average iteration will take? （ specify the cache miss situation）



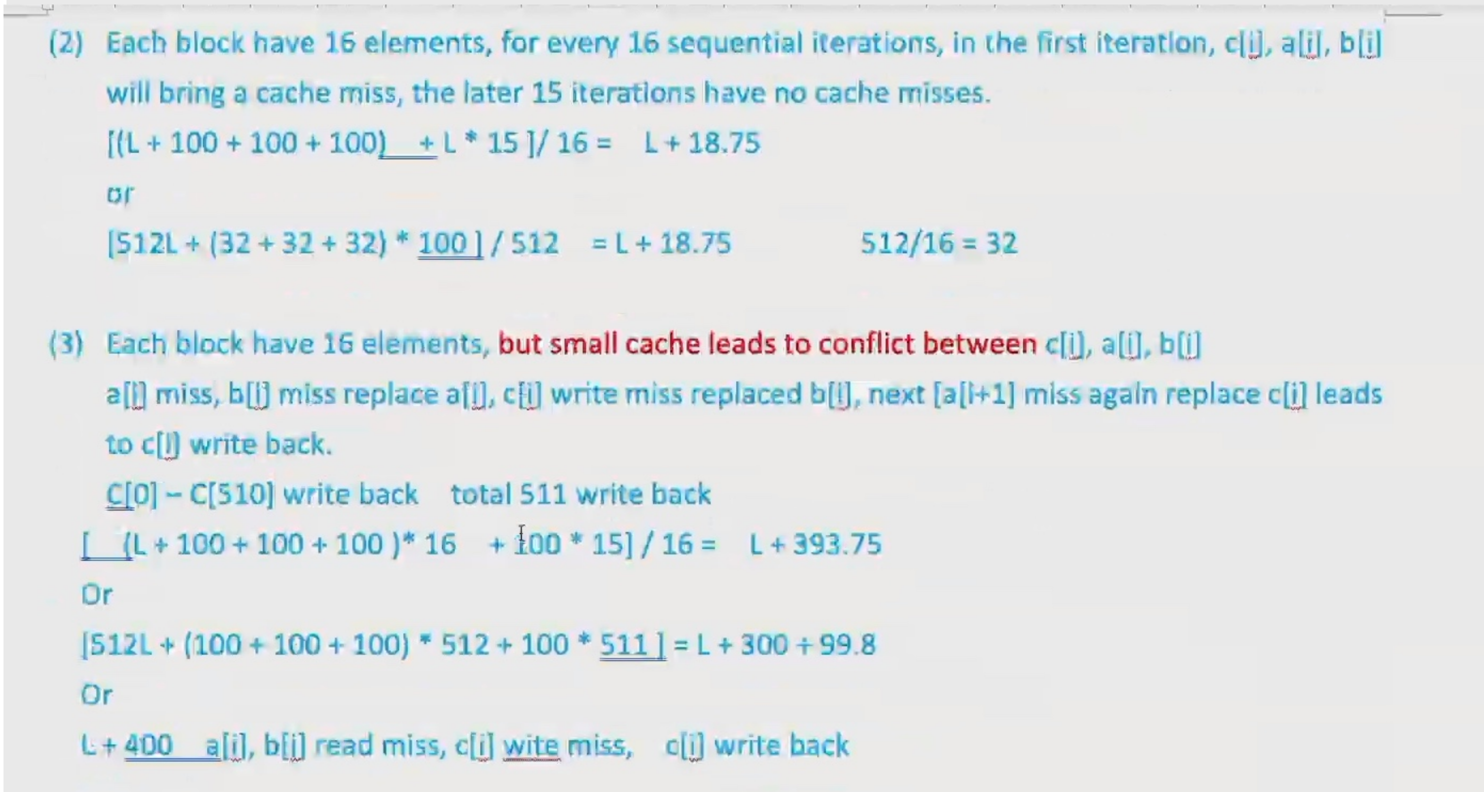
Solution:

(1) For every 4 sequential iterations, in the first iteration, c[i], a[i], b[i] will bring a cache miss, the later three iterations have no cache misses.

$[(L + 100 + 100 + 100) + L + L + L]/ 4 = L + 75$

Or

$512L + (128 + 128 + 128) * 100 / 512 = L + 75,$ $512/4 = 128$



(2) Each block have 16 elements, for every 16 sequential iterations, in the first iteration, c[i], a[i], b[i] will bring a cache miss, the later 15 iterations have no cache misses.

$[(L + 100 + 100 + 100) + L * 15]/ 16 = L + 18.75$

or

$[512L + (32 + 32 + 32) * 100] / 512 = L + 18.75$ $512/16 = 32$

(3) Each block have 16 elements, but small cache leads to conflict between c[i], a[i], b[i]

a[i] miss, b[i] miss replace a[i], c[i] write miss replaced b[i], next [a[i+1] miss again replace c[i] leads to c[i] write back.

C[0] – C[510] write back total 511 write back

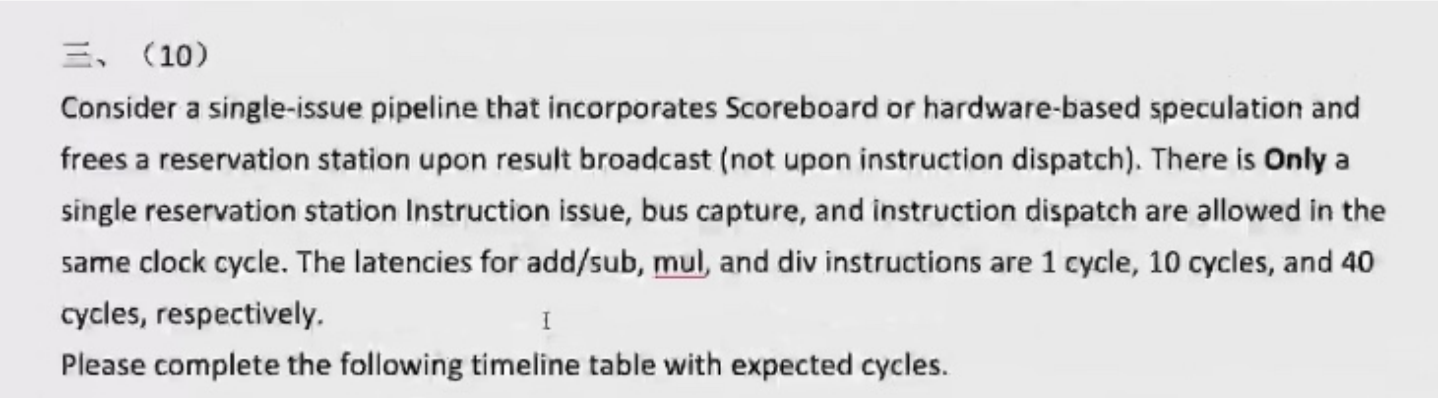
$[(L + 100 + 100 + 100)* 16 + 100 * 15] / 16 = L + 393.75$

Or

$[512L + (100 + 100 + 100) * 512 + 100 * 511] = L + 300 + 99.8$

Or

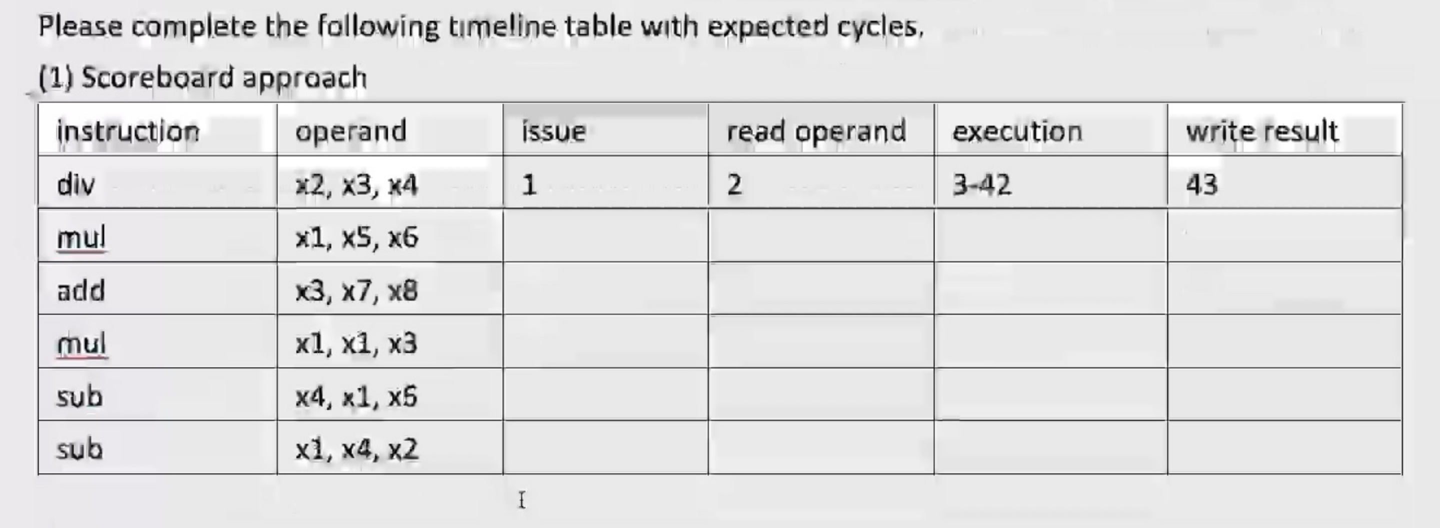
L + 400 a[i], b[i] read miss, c[i] wite miss, c[i] write back



三、（10）

Consider a single-issue pipeline that incorporates Scoreboard or hardware-based speculation and frees a reservation station upon result broadcast (not upon instruction dispatch). There is **Only** a single reservation station Instruction issue, bus capture, and instruction dispatch are allowed in the same clock cycle. The latencies for add/sub, mul, and div instructions are 1 cycle, 10 cycles, and 40 cycles, respectively.

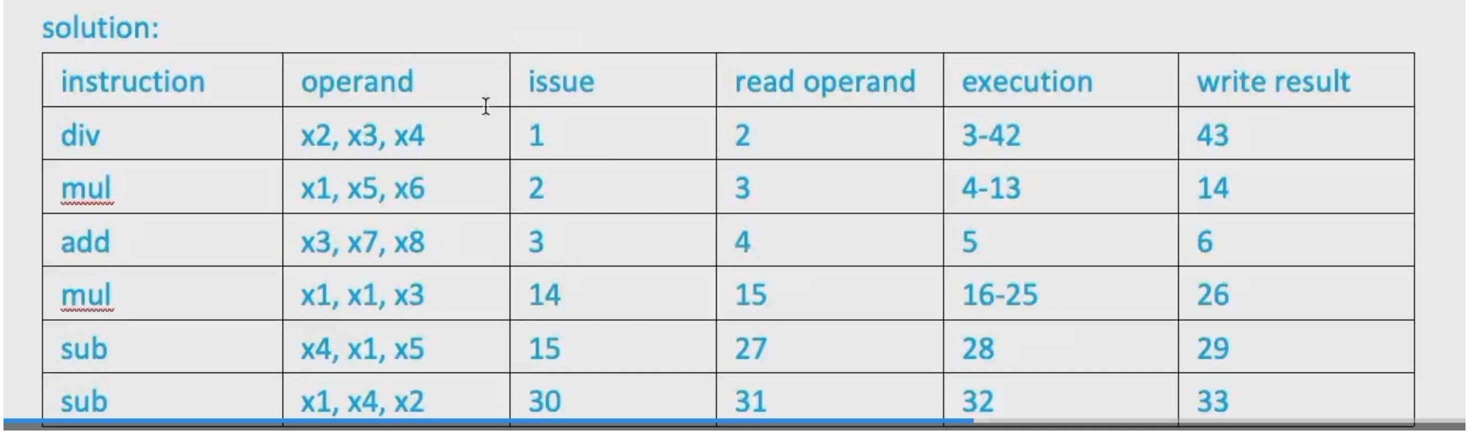
Please complete the following timeline table with expected cycles.



Please complete the following timeline table with expected cycles,

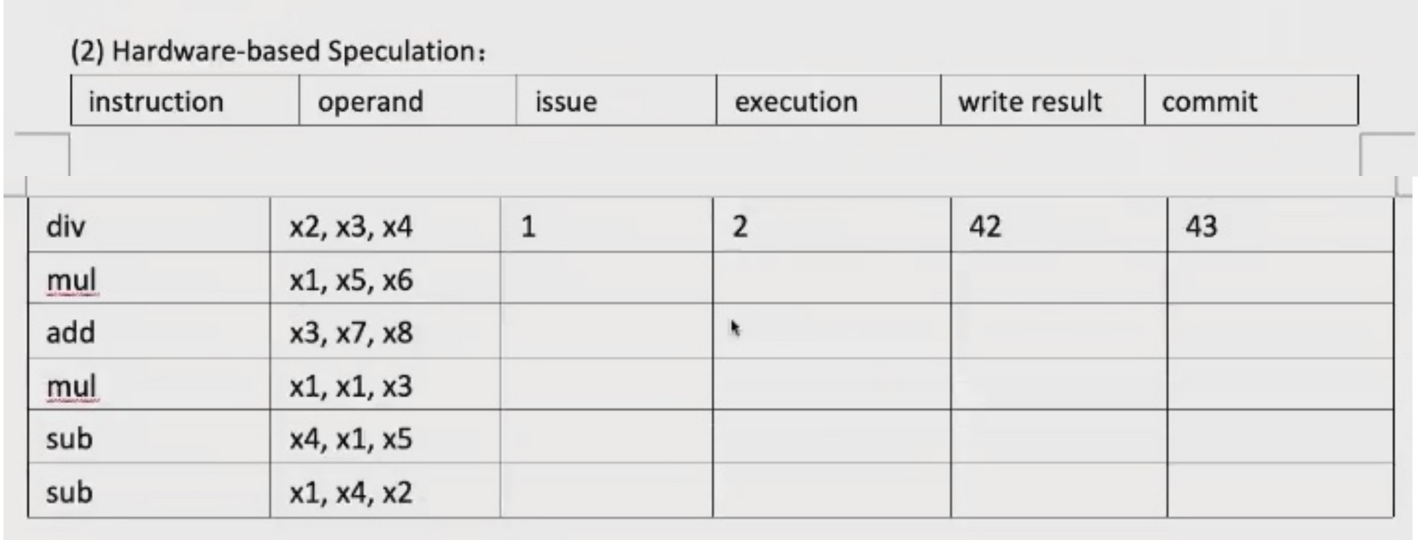
(1) Scoreboard approach

instruction	operand	issue	read operand	execution	write result
div	x2, x3, x4	1	2	3-42	43
mul	x1, x5, x6				
add	x3, x7, x8				
mul	x1, x1, x3				
sub	x4, x1, x5				
sub	x1, x4, x2				



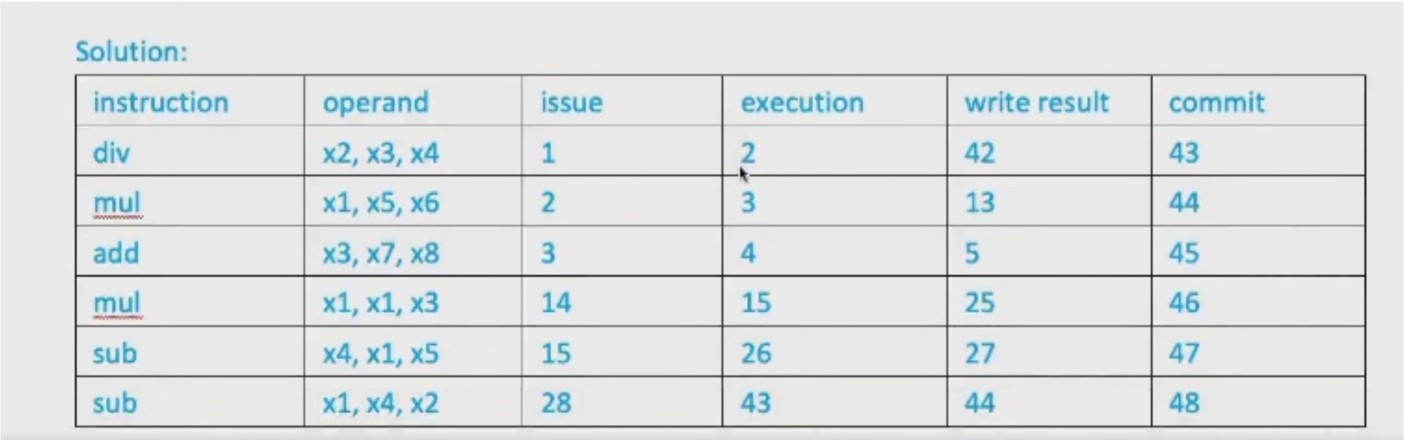
solution:

instruction	operand	issue	read operand	execution	write result
div	x2, x3, x4	1	2	3-42	43
mul	x1, x5, x6	2	3	4-13	14
add	x3, x7, x8	3	4	5	6
mul	x1, x1, x3	14	15	16-25	26
sub	x4, x1, x5	15	27	28	29
sub	x1, x4, x2	30	31	32	33



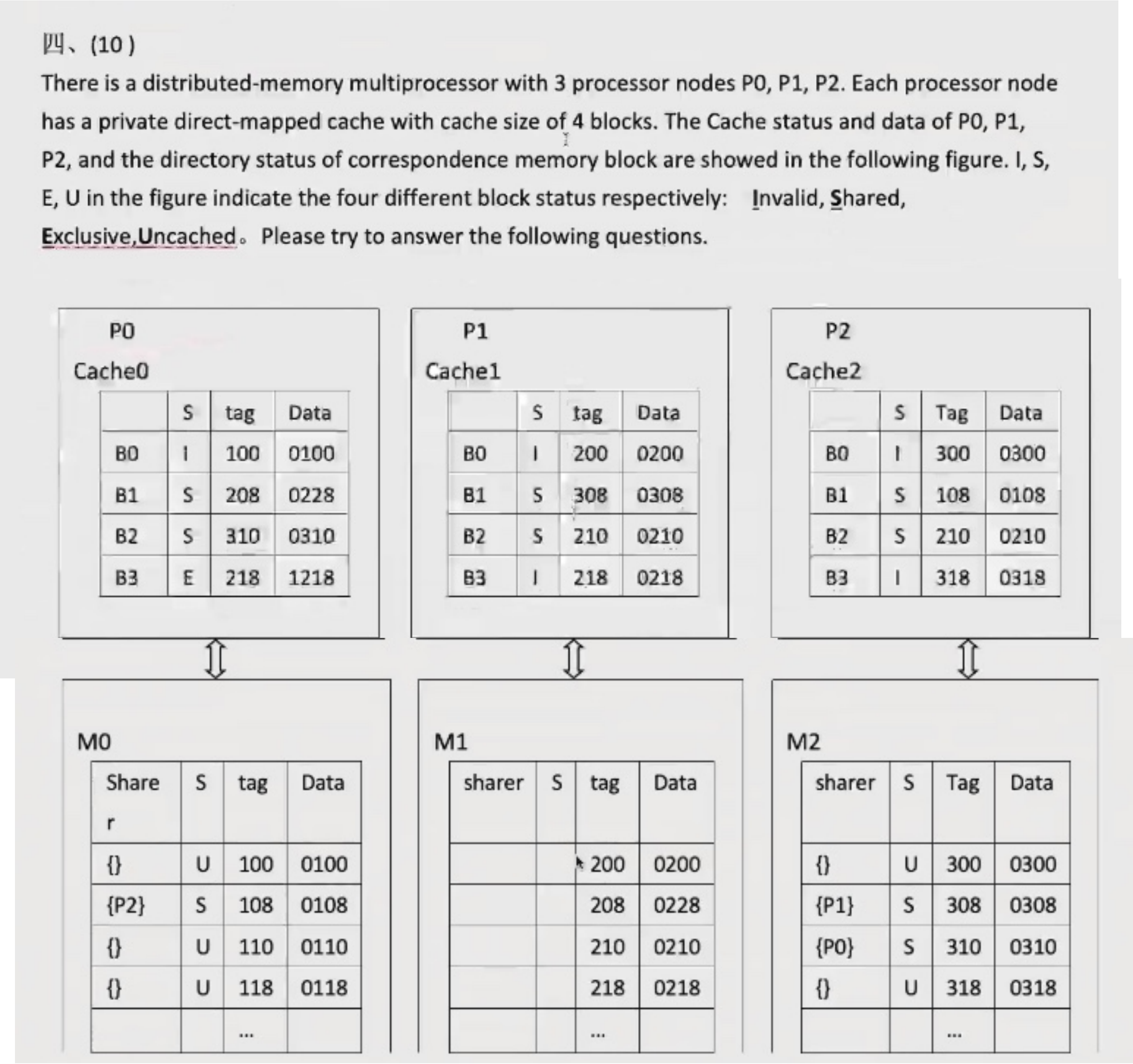
(2) Hardware-based Speculation:

instruction	operand	issue	execution	write result	commit
div	x2, x3, x4	1	2	42	43
mul	x1, x5, x6				
add	x3, x7, x8				
mul	x1, x1, x3				
sub	x4, x1, x5				
sub	x1, x4, x2				



Solution:

instruction	operand	issue	execution	write result	commit
div	x2, x3, x4	1	2	42	43
mul	x1, x5, x6	2	3	13	44
add	x3, x7, x8	3	4	5	45
mul	x1, x1, x3	14	15	25	46
sub	x4, x1, x5	15	26	27	47
sub	x1, x4, x2	28	43	44	48



四、(10)

There is a distributed-memory multiprocessor with 3 processor nodes P0, P1, P2. Each processor node has a private direct-mapped cache with cache size of 4 blocks. The Cache status and data of P0, P1, P2, and the directory status of correspondence memory block are showed in the following figure. I, S, E, U in the figure indicate the four different block status respectively: **I**nvalid, **S**hared, **E**xclusive,**U**ncached。 Please try to answer the following questions.

P0

Cache0

	S	tag	Data
B0	I	100	0100
B1	S	208	0228
B2	S	310	0310
B3	E	218	1218

P1

Cache1

	S	tag	Data
B0	I	200	0200
B1	S	308	0308
B2	S	210	0210
B3	I	218	0218

P2

Cache2

	S	Tag	Data
B0	I	300	0300
B1	S	108	0108
B2	S	210	0210
B3	I	318	0318

M0

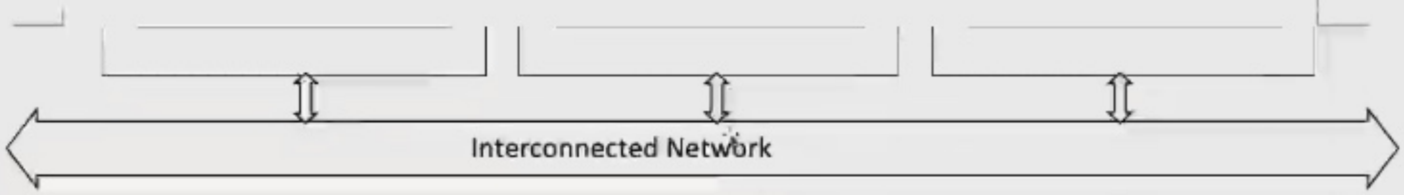
Sharer	S	tag	Data
{}	U	100	0100
{P2}	S	108	0108
{}	U	110	0110
{}	U	118	0118
		...	

M1

sharer	S	tag	Data
		200	0200
		208	0228
		210	0210
		218	0218
		...	

M2

sharer	S	Tag	Data
{}	U	300	0300
{P1}	S	308	0308
{P0}	S	310	0310
{}	U	318	0318
		...	



1) Complete the directory information in M1. (2 points)

2) According to the example "P0 read 300" in the following figure, please draw the graph to show the procedures to complete the event of "P2 read 218" with sending messages between processor nodes, updating contents of directory items and cache status and values. Assuming that the event "P2 read 218" starts from the initial status in above figure. (4 points)

3) Similarly, please draw the graph to show the procedures to complete the event of "P0 Write 210 with value 8888." （4 points）Assuming that the event "P0 Write 210 with value 8888" starts from the initial status in above figure.

Example: P0 read 300:

P0 send ReadMiss for Tag(300) to P2;
P2 modify the directory " M2, 300, {}, U" to " M2， 300， {P0}， S"
P2 send datareply back to P0 with the value 0300

Solution:

(1) 200: {}, U;
208: {P0}, S
210: {P1, P2} , S
218: {P0}, E

(2) P2 read 218
P2 send ReadMiss to P1
P1 send Fetch 218 to P0
P0 write back 218, 1218 to P1, modify cache "B3, E, 218, 1218" to "B3, S, 218, 1218"
P1 modify directory "M1, 218, {P0}, E, " to "M1, 218, {P0, P2}, S"
P1 datareply back to P2, 218, 1218
P2 modify cache: "B3, I, 318, 0318" to "B3, S, 218, 1218.

(3) P0 Write 210 with value 8888.
P0 send writemiss 210 to P1
P1 send invalidate to P1, P2,
P1 modify cache "B2, S, 210, 0210" to "B2, I, 210, 0210", sendback ACK to P1