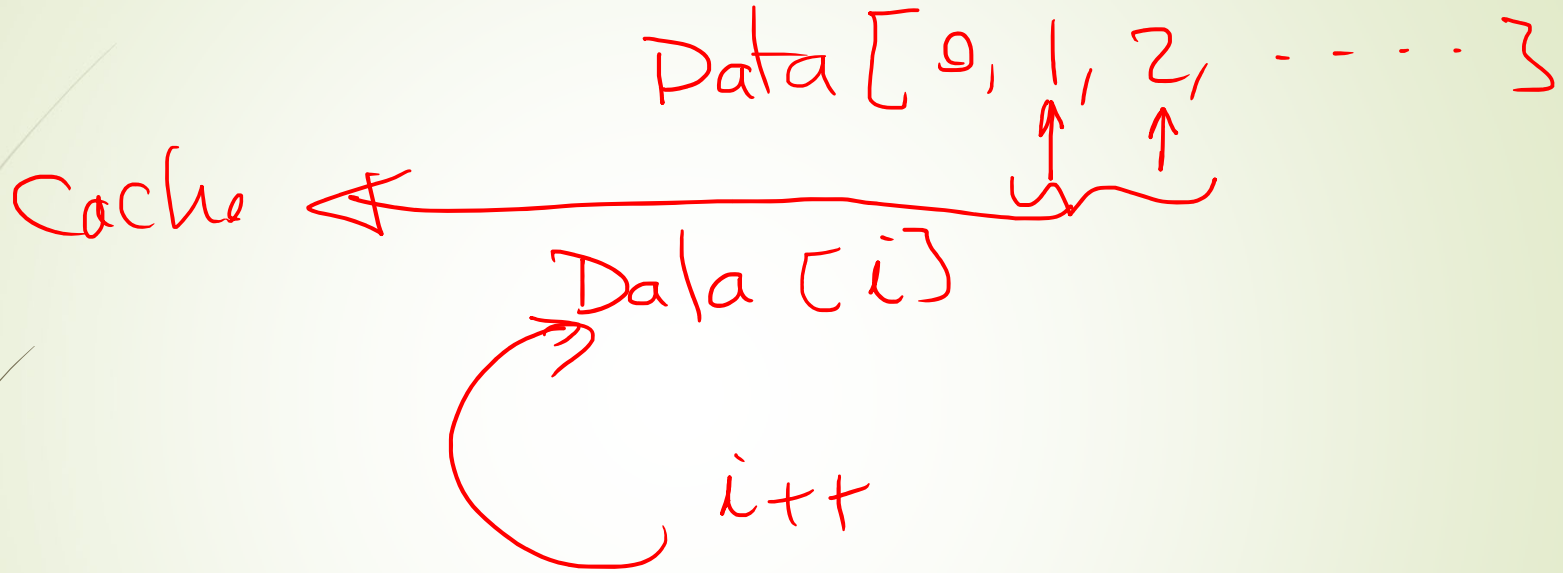
# EGC442
# Class Notes
# 4/25/2023

**Baback Izadi**
Division of Engineering Programs
bai@engr.newpaltz.edu


Data [0, 1, 2, ----3
Cache
Data [i]
i++

Assume a direct-mapped cache with the following parameters:
Address size: 32 bits
Cache data size: 2 KiB
Cache block: 2 words

1) The cache contains _____ words.

512

Check Show answer

**Correct**

512

The number of words is obtained by dividing the cache data size in bits by the number of bits per word.

$$= \frac{2\text{KiB}}{32 \text{ bits per word}}$$
$$= \frac{2 \cdot (1024 \text{ bytes} \cdot 8 \text{ bits per byte})}{32 \text{ bits per word}}$$
$$= 512 \text{ words}$$

2) The cache contains _____ blocks.

256

Check Show answer

**Correct**

256

If each block contains 2 words, then the number of cache blocks is found by dividing the total number of words by 2.

$$= \frac{512 \text{ words}}{2 \text{ words per block}}$$
$$= 256 \text{ blocks}$$

3) The size of the tag field is _____ bits.

21

Check Show answer

**Correct**

21

8 bits are used to index the cache ($\log_2 256 = 8$). The least significant 2 bits of the address specify a byte and are ignored. One bit is used to specify the word within the cache block. Thus, 32 - 8 - 2 - 1 = 21, bits are used in the tag.

4) The total block size is _____ bits.

64

Check Show answer

5) The total cache size is _____ bits.

22016

Check Sl

**Correct**

64

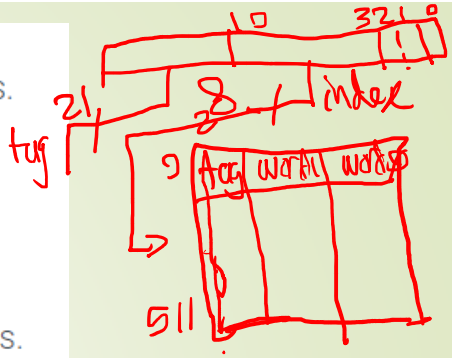
Each block contains 2 words, so 2 × 32 = 64 bits.

**Correct**

22016

number of blocks × (block size + tag size + valid field size)
= 256 × (64 bits + 21 bits + 1 bit)
= 22016 bits

In KiB, the total cache size is 22016 / (1024 × 8) = 2.7 KiB. The total number of bits in the cache is about 2.7 KiB/ 2 KiB = 1.35 times as many needed just for the storage of data.

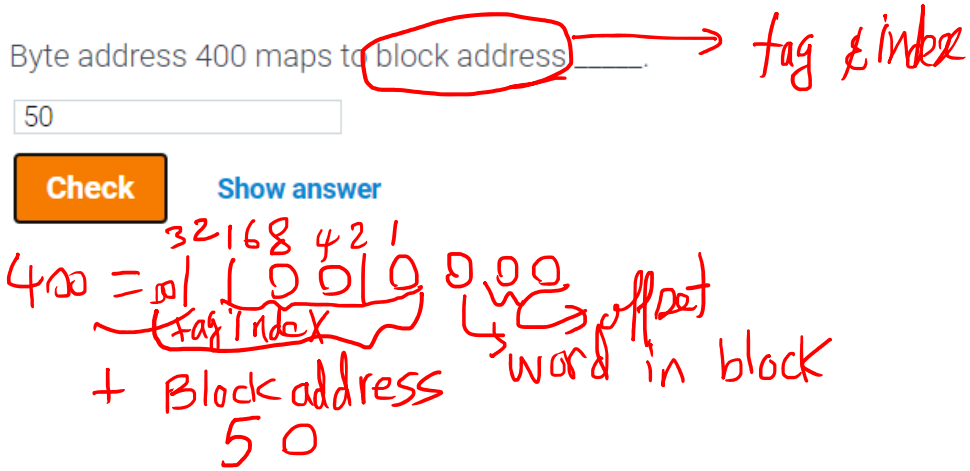


1. Assume a direct-mapped cache with 32 blocks and a block size of 8 bytes.

I) Byte address 400 maps to block address ____.

50

Check Show answer




**Correct**

50

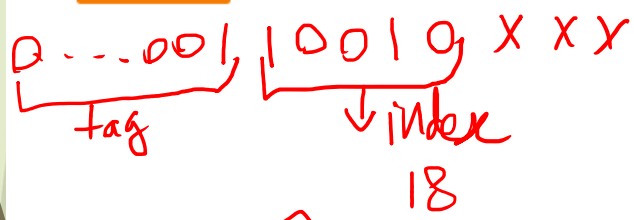
The block address is given by:

$$= \frac{\text{Byte address}}{\text{Bytes per block}}$$
$$= \frac{400}{8}$$
$$= 50$$

II) Byte address 400 maps to block number ____.

18

Check Show answer




**Correct**

18

The block addressed is used to map to the block number:

= (Block address) modulo (number of blocks in the cache)
= (50) modulo (32)
= 18

III) Byte address 360 maps to block number ____.

13

Check Show answer

2) The miss rate may increase if the block size becomes a significant fraction of the cache size.

- (●) True
- ( ) False

**Correct**

Fewer blocks can be held within the cache as block size increases, which results in replacing a block from the cache before many of that block's words are accessed.

3) Which of the following items does NOT contribute to the cost of a miss penalty?

- ( ) Latency to access the first word
- ( ) Transfer time of the block
- (●) Latency to determine the cache block

**Correct**

The cache block calculation is independent of a miss penalty. The cache block is calculated first, then the processor can determine if the access request results in a cache hit or miss.

4) The processing of a cache miss creates a _____.

- (●) pipeline stall
- ( ) interrupt

**Correct**

Multiple clock cycles are needed to access the memory, so the contents of the temporary registers and programmer-visible registers are frozen until the data is available.

5) If an instruction access results in a miss, then the address of the instruction at _____ is fetched from the memory.

- ( ) PC
- (●) PC - 4
- ( ) PC + 4

**Correct**

The program counter is incremented in the first clock cycle of execution, so the address of the instruction that generated an instruction cache miss is equal to PC - 4.

| | | |
|---|---|---|
| **Write-through scheme** | A value is read from the cache and modified. The modified value is written to the cache and the corresponding memory location.<br>While the write-through scheme is simple to implement, processor performance is slowed due to the large number of clock cycles required for each write to memory. | Correct |
| **Write buffer** | A value is read from the cache and modified. The modified value is written to the cache and to a queue that stores the value while waiting to be written to the corresponding memory location.<br>The processor can continue execution after writing the data into the write buffer, which helps to improve the processor's performance. Processor stalls may still occur if the rate at which the memory can complete writes is less than the rate at which the processor is generating writes. | Correct |
| **Write-back scheme** | A value is read from the cache and modified. The modified value is written to the cache. The modified value is only written from the cache to memory when the cache block is replaced.<br>Write-back schemes can improve performance, especially when processors can generate writes as fast or faster than the writes can be handled by main memory. | Correct |

7 Assume the miss rate of an instruction cache is 2% and the miss rate of the data cache is 4%. If a processor has a CPI of 2 without any memory stalls, and the miss penalty is 100 cycles for all misses, determine how much faster a processor would run with a perfect cache that never missed. Assume the frequency of all loads and stores is 36%.

**Answer**

The number of memory miss cycles for instructions in terms of the Instruction count (I) is

$$\text{Instruction miss cycles} = \text{I} \times 2\% \times 100 = 2.00 \times \text{I}$$

As the frequency of all loads and stores is 36%, we can find the number of memory miss cycles for data references:

$$\text{Data miss cycles} = \text{I} \times 36\% \times 4\% \times 100 = 1.44 \times \text{I}$$

The total number of memory-stall cycles is 2.00 I + 1.44 I = 3.44 I. This is more than three cycles of memory stall per instruction. Accordingly, the total CPI including memory stalls is 2 + 3.44 = 5.44. Since there is no change in instruction count or clock rate, the ratio of the CPU execution times is

$$\frac{\text{CPU time with stalls}}{\text{CPU time with perfect cache}} = \frac{\text{I} \times \text{CPI}_{\text{stall}} \times \text{Clock cycle}}{\text{I} \times \text{CPI}_{\text{perfect}} \times \text{Clock cycle}} = \frac{\text{CPI}_{\text{stall}}}{\text{CPI}_{\text{perfect}}} = \frac{5.44}{2}$$

The performance with the perfect cache is better by $\frac{5.44}{2} = 2.72$.

1) The instruction cache miss rate is ____.
   - (•) 2%
   - ( ) 4%
   - ( ) 36%

**Correct**

The example assumes the miss rate of the instruction cache is 2%, where a miss results in a 100 cycle penalty. For I instructions, the number of memory-stall cycles due to instruction misses is I × 2% × 100 = 2.00 I.

2) The number of memory-stall cycles for data misses in terms of the instruction count (I) is ____.
   - ( ) I × 4% × 100
   - ( ) I × 36% × 4%
   - (•) I × 36% × 4% × 100

**Correct**

36% of instructions access the data cache, of which 4% result in a miss. For I instructions, the number of memory-stall cycles due to data misses is I × 36% × 4% × 100 = 1.44 I.

3) The total CPI is ____.
   - ( ) 1.44
   - ( ) 3.44
   - (•) 5.44

**Correct**

The total CPI includes the number of cycles per instruction without any memory stalls, the instruction miss cycles, and the data miss cycles. Ex: 2 + 2.00 + 1.44 = 5.44 cycles per instruction.

8) Find the AMAT (average memory access time) for a processor with a 2 ns clock cycle time, a miss penalty of 40 clock cycles, a miss rate of 0.05 misses per instruction, and a cache access time (including hit detection) of 1 clock cycle. Assume that the read and write miss penalties are the same and ignore other write stalls.

$$\text{AMAT} = \text{Time for a hit} + \text{Miss rate} \times \text{Miss penalty}$$

$$\text{AMAT} = 2\text{ns} + 0.05 \times 40 \times 2\text{ns}$$

$$= 6\text{ns}$$

9) If the clock rate is increased without changing the memory system, the fraction of execution time due to cache misses ____ relative to total execution time.

- (●) increases
- ( ) decreases

**Correct**

The amount of time spent on memory stalls takes an increasing fraction of the execution time. In the above example, if the CPI is reduced from 2 to 1, while retaining the same memory system, the fraction of time taken on memory stalls increases from 63% to 77%.

10) AMAT considers the average time to access data for ____.

- ( ) misses
- (●) both hits and misses

**Correct**

The hit time, as well as the miss penalty, affects performance. AMAT provides a simplified way to examine alternative cache designs considering both hits and misses and the frequency of different accesses.

The speed of the memory system affects the designer's decision on the size of the cache block. Complete the following cache designer guidelines.

1) The shorter the memory latency, the ____ the cache block.

- (•) smaller
- ( ) larger

**Correct**

A lower miss penalty can enable smaller blocks because a shorter amount of memory latency is available to amortize.

2) The higher the memory bandwidth, the ____ the cache block.

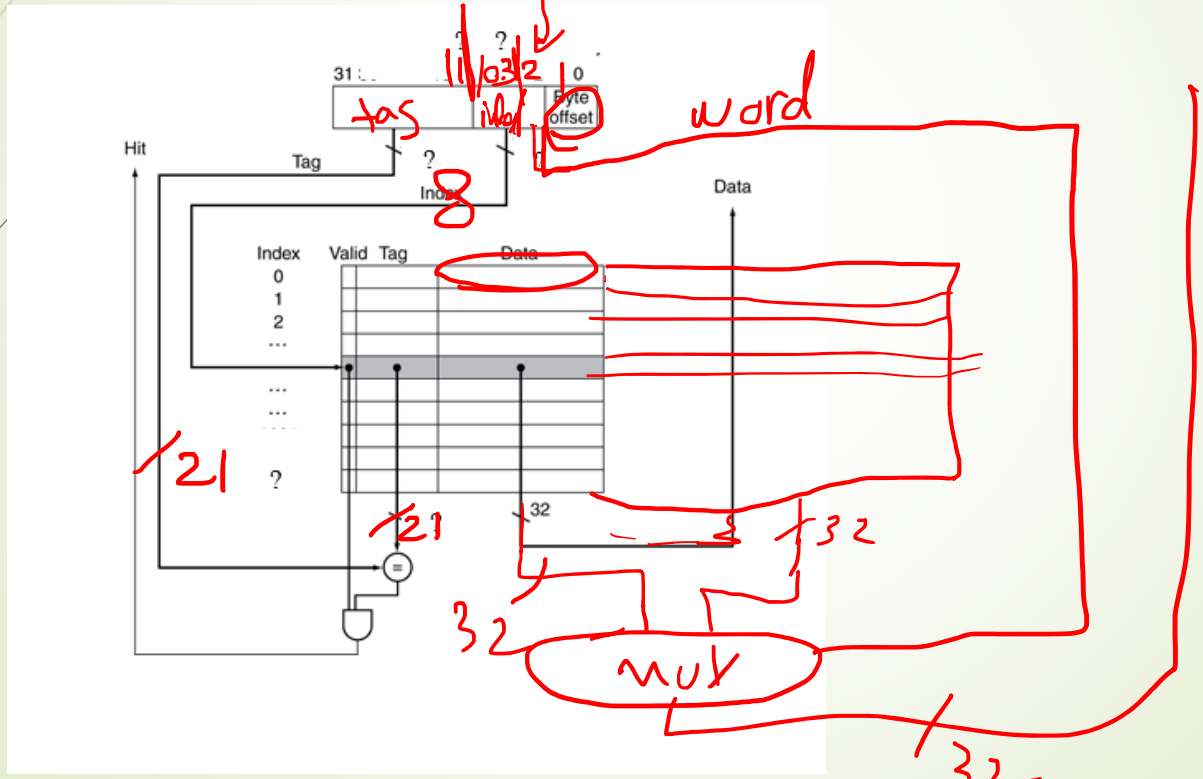
- ( ) smaller
- (•) larger

**Correct**

A higher memory bandwidth usually leads to larger blocks because the miss penalty is only slightly larger.

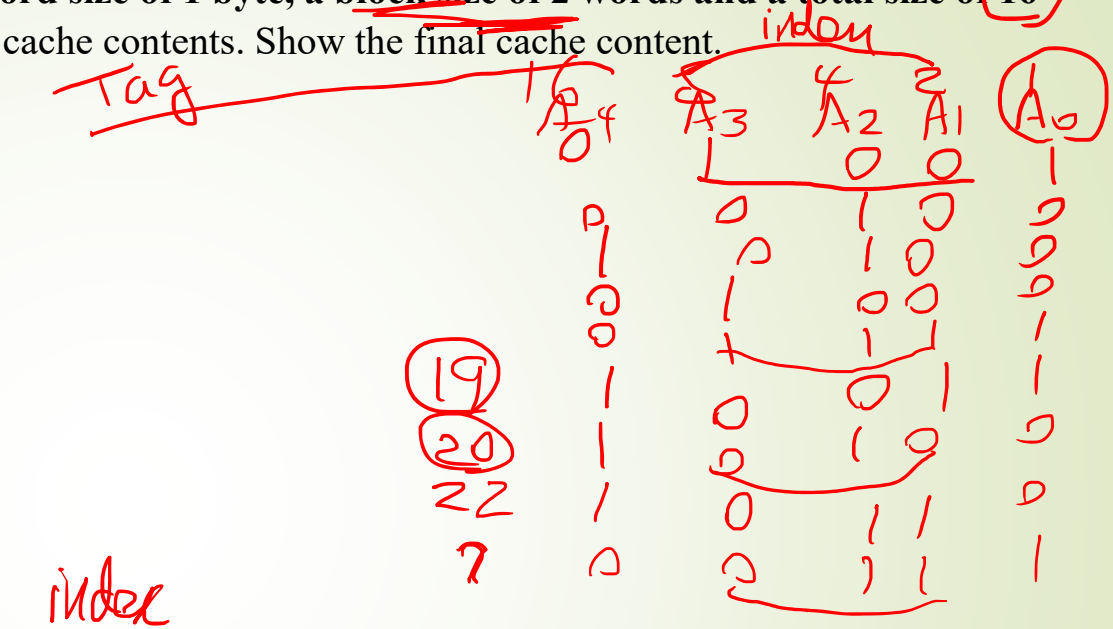
11) Design a direct-mapped cache with the following parameters:

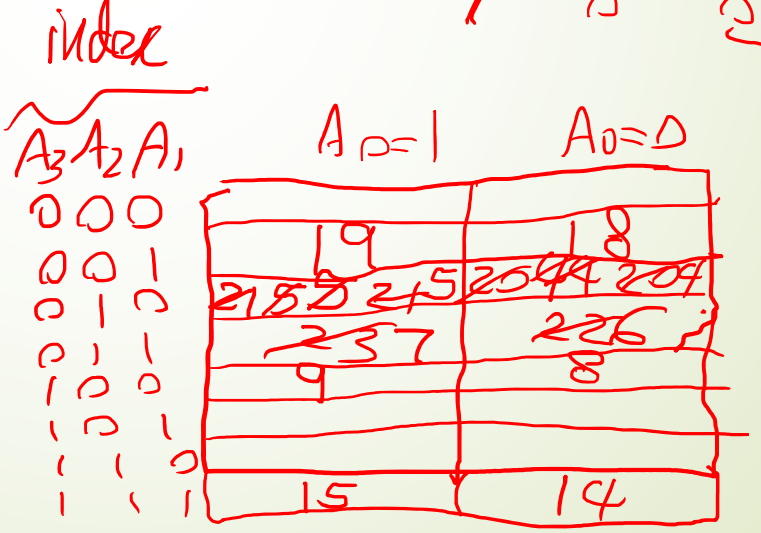
- Address size: 32 bits
- Cache data size: 2 KB
- Cache block: 2 word



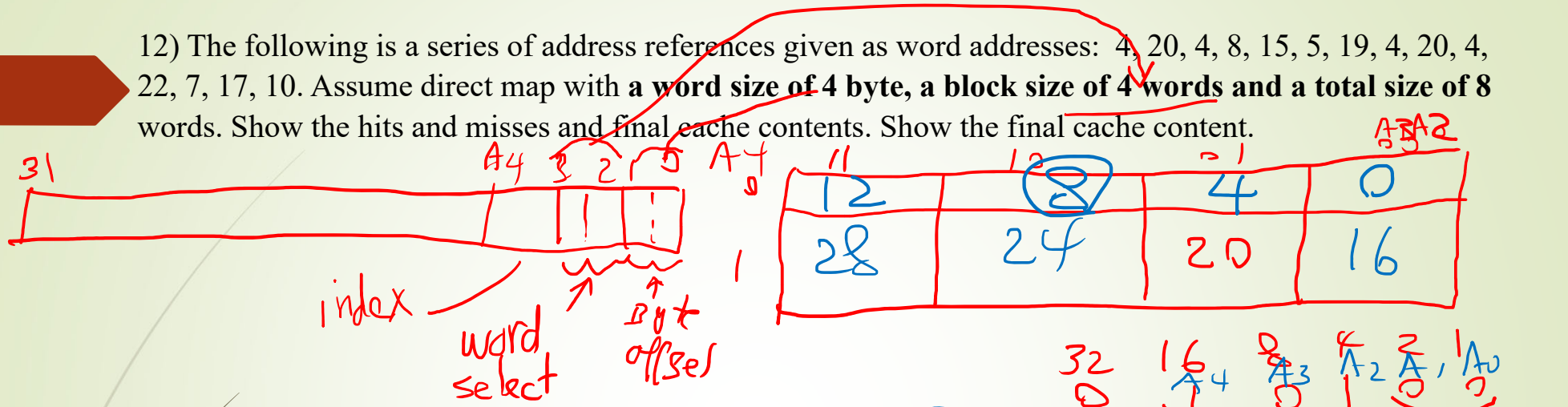
12) The following is a series of address references given as word addresses: 9, 4, 20, 4, 8, 15, 5, 19, 4, 20, 4, 22, 7, 17, 10. Assume direct map with **a word size of 1 byte, a block size of 2 words and a total size of 16** words. Show the hits and misses and final cache contents. Show the final cache content.

| Location | Hit/Miss? |
|---|---|
| 9 | M |
| 4 | M |
| 20 | M |
| 4 | H |
| 8 | M |
| 15 | M |
| 5 | H |
| 19 | M |
| 4 | H |
| 20 | M |
| 4 | H |
| 22 | M |
| 7 | M |



12) The following is a series of address references given as word addresses: 4, 20, 4, 8, 15, 5, 19, 4, 20, 4, 22, 7, 17, 10. Assume direct map with **a word size of 4 byte, a block size of 4 words and a total size of 8 words**. Show the hits and misses and final cache contents. Show the final cache content.




| Location | Hit/Miss? |
|---|---|
| 9 | |
| 4 | |
| 20 | |
| 4 | |
| 8 | |
| 15 | |
| 5 | |
| 19 | |
| 4 | |
| 20 | |
| 4 | |
| 22 | |
| 7 | |

13) Assume an instruction cache miss rate for an application is 2% and the data cache miss rate of 4%. Assume further that our CPU has a CPI of 2 without any memory stalls and the miss penalty is 40 cycles for all misses.

a. Determine the overall CPI with the indicated misses, provided the frequency of all loads and stores in the application is 20%.

b. Suppose we increase the performance of the machine in the above example by reducing its CPI from 2 to 1 via pipelining. Determine the new overall CPI.

.32

.8 .2 * .04 * 40

base

.02 x 40

Effective CPI = 2 + 2% * 40 + 20% * 4% * 40

Inst.

clk = 5 ns

= 3.12

3.12 * 5 ns = 15.60 ns to exec. an inst.

b. $E_{CPI}$ = 1 + .8 + .32

= 2.12