Question 1

Design a 8-way set associative cache that has 16 blocks and 32 bytes per block. Assume a 32 bit address. Calculate the following:

How many bits for the byte offset? [5]

How many bits for the set (index) field? [1]

How many bits for the tag? [26]

#sets = 16/8 = 2, so set(index) bits = 1. 32 bytes per block, so offset = 5. Tag = 32-5-1 = 26

Question 2

T/F High associativity in a cache reduces compulsory misses.

It reduces conflict misses.

Question 3

Assume a 64KB cache with four-word block size (a word is 4 bytes) and a 32-bit address. If a block has 28 tag bits, what is the type of this cache?

2-way set associative / Direct mapped / Fully associative / 4-way set associative

address = 32 bits, tag = 28 bits, as a block is 4 words i.e. 4x4 bytes, offset = 4 bits, that means, index = 0 bits. Thus, its fully associative

Question 4

TLBs are typically built to be fully-associative or highly set-associative. In contrast, first-level data caches are more likely to be direct-mapped or 2 or 4-way set associative. Give a reason why this is so.

*TLBs are typically built to be fully associative to prevent lots of conflict misses because different processes have their own virtual address space and it's unfair to concern only one process's memory.*

Question 5

T/F For a given capacity and block size, a set-associative cache implementation will typically have a lower hit time than a direct-mapped implementation.

Set-associative cache will have a lower miss rate (leading to lower AMAT) but given a set has multiple blocks the hit time for a particular block is higher.

Question 6

T/F In a Write-Through cache, a read miss always causes a write to the lower memory level.

Write through writes back to the lower level memories when there is a write miss.

Question 7

The speed of the memory system affects the designer’s decision on the size of the cache block. Which of the following cache designer guidelines is generally valid?

The shorter the memory latency, the smaller the cache block

The shorter the memory latency, the larger the cache block

The higher the memory bandwidth, the smaller the cache block

The higher the memory bandwidth, the larger the cache block

Question 8

T/F RAID systems can have catastrophic failures.

Question 9

Consider a system with a two-level cache having the following characteristics: L1 Cache has an access time of 1 clock cycle with an average hit rate of 95%; L2 Cache has an access time of 5 clock cycles (after L1 miss) with an average miss rate of 10%. If L2 misses take 40 clock cycles, what would be the average memory access time in clocks?

AMAT = 1+0.05\*(5 + 0.1\*(40)) = 1.45 cycles

Question 10

For a data cache with a 92% hit rate and a 2-cycle hit latency, calculate the average memory access latency in cycles. Assume that latency to memory and the cache miss penalty together is 124 cycles. Note: The cache must be accessed after memory returns the data.

AMAT = hit time + miss rate\*(miss penalty) = 2 + 0.08\*124 = 11.92