5.1) Unroll the following loop four times and optimize the resulting instruction sequence. Assume the loop count is a multiple of four. Further assume the 5-stage MIPS pipeline with ID-stage branch resolution and make sure that no bubbles need to be generated by the CPU.

**lbl: sw $zero, 0($t1)**

**addi $t1, $t1, 4**

**bne $t1, $t2, lbl**

sw $zero, 0($t1)

addi $t2, $t2, 4

addi $t3, $t3, 8

addi $t4, $t4, 12sw $zero, 0($t2)sw $zero, 0($t3)sw $zero, 0($t4)

bne $t1, $t2, lbl

**5.2) What is the best CPI a superscalar CPU with *k* pipelines can reach?**

The best CPI for a superscalar CPU with K pipelines is K. One cycle per instruction

**5.3) Why do modern superscalar cores typically not have more than eight parallel pipelines?**

Because beyond 8 parallel pipelines it doesn’t really make a difference on performance.

**5.4) In the dual issue MIPS VLIW, the data memory has only one read and one write port, just like in the single-pipeline MIPS implementation. Why does the number of ports not double?**

**5.5) In a system with 48-bit addresses and 64-bit words, how many bits of an effective address are used for the byte offset, word (block) offset, the index, and the tag in a 32kB direct-mapped cache with 64 bytes per block?**

Cache size = 32KB = 32 \* 2^10 bytes

Cache line size = 64 bit words = 2^6 bytes

Number of cache lines = 32\*2^10 bytes / 64 bytes = 512 = 2^9

Index bits = 9

Word Offset = 6

Tag bits = 48 – 9 – 6 = 33

**5.6) In a system with 32-bit words and addresses, how many bits of total storage, including data, valid, dirty, and tag bits, does a direct-mapped cache with 16 words per block and 512 lines require?**

**5.7) Assume an ideal load CPI of 1, an L1 miss penalty of 5 cycles, and an L2 miss penalty of 60 cycles. Assume further than 4% of the load accesses to the L1 data cache miss and that 20% of the load accesses to the L2 miss. Compute the average memory access time for load instructions.**

1 + 4% \* (5 + 20% \* 60) = 1.68 cycles

**5.8) For performance reasons, the L1 cache is usually split into an instruction cache and a data cache. However, all other cache levels are unified. Explain why it is not necessary to split, for example, the L2 cache.**

It increases bandwidth – processors can read data from the instruction cache and the data cache simultaneously in order to satisfy the demands of the pipeline without stalling.

5.9) In the following C/C++ code segment, which performs a matrix-matrix multiplication, specify, for the matrixes *a*, *b*, and *c*, whether their accesses primarily have spatial, temporal, or no locality. Accesses fewer than *n* words apart are considered to have spatial locality.

**for (k = 0; k < n; k++)**

**for (i = 0; i < n; i++)**

**for (j = 0; j < n; j++)**

**c[i][j] += a[i][k] \* b[k][j];**

5.10) In a direct-mapped cache with one byte per line and four lines, which of the following byte addresses hit and miss, assuming the cache is initially empty?

**0, 1, 2, 3, 4, 5, 4, 3, 2, 1, 0, 1, 2, 3, 4, 0**

0 - miss

1 - miss

2 - miss

3 - miss

4 - miss

5 - miss

3 - hit

2 - hit

1 – miss

0 - miss

1 - hit

2 - hit

3 - hit

4 – miss

0 - miss