EE/CSCI 451
Spring 2015
Homework 1

Assigned: January 21, 2015
Due: January 29, 2015, before 4:30 pm, Locker #??, EEB basement
Total Points: 100

1 [20 points]

Explain the following terms:

1. Superscalar processor
2. Row major layout
3. Cache pollution
4. Instruction-level parallelism
5. Cache line
6. Data dependencies
7. Very Long Instruction Word (VLIW) Processor
8. Spatial locality
9. Single instruction multiple data (SIMD)
10. Implicit parallelism

2 [15 points]

1. Consider a symmetric multiprocessing with a distributed shared-address space. Consider a simple cost model in which it takes 10 ns to access local cache, 100 ns to access local memory, and 400 ns to access remote memory. A parallel program is running on this machine. The program is perfectly load balanced with 80% of all accesses going to local cache, 10% to local memory, and 10% to remote memory. What is the effective memory access time for this computation? If the computation is memory bound, what is the peak computation rate?

2. Now consider the same computation running on one processor. Here, the processor hits the cache 70% of the time and local memory 30% of the time. What is the effective peak computation rate for one processor? What is the fractional computation rate of a processor in a parallel configuration as compared to the serial configuration?

Notice that the cache hit for multiple processors is higher than that for one processor. This is typically because the aggregate cache available on multiprocessors is larger than on single processor systems.

3 [15 points]

Consider a memory system with a single cycle cache and 100 cycle latency DRAM with the processor operating at 1 GHz. Assume that the processor has two multiply-add units and is capable of executing four instructions in each cycle of 1 ns. Consider the problem of computing the dot product of two vectors on such a platform. A dot-product computation performs one multiply-add (2 FLOPs) on a single pair of vector elements, i.e., each floating point operation requires one data fetch.

1. In each memory cycle, the processor fetches one word (one vector element). What is the peak achievable performance (in FLOPS) of a dot product of two vectors?

2. In each memory cycle, the processor fetches four words. What is the peak achievable performance of a dot product of two vectors?

```
/* dot product loop */
for (i = 0; i < dim; i++)
    dot_product += a[i] * b[i];
```

4 [15 points]

Consider a memory system with a single cycle cache and 100 cycle latency DRAM with the processor operating at 2 GHz. The processor has two multiply-add units and is capable of executing four instructions in each cycle. In each memory cycle, the processor fetches four words. Now consider the problem of multiplying a dense matrix with a vector using a two-loop dot-product formulation. The matrix is of dimension $4K \times 4K$. (Each row of the matrix takes 16 KB of storage.) Assume the vector is cached, what is the peak achievable performance of this technique using a two-loop dot-product based matrix-vector product? (State your assumptions)

```
/* matrix-vector product loop */
for (i = 0; i < dim; i++)
    for (j = 0; j < dim; j++)
        c[i] += a[i][j] * b[j];
```

5 [10 points]

For the same memory system and processor in Problem 3, consider the problem of multiplying two dense matrices of dimension $4K \times 4K$. What is the peak achievable performance using a three-loop dot-product based formulation? (Assume that matrices are laid out in a row-major fashion.)

```
/* matrix-matrix product loop */
for (i = 0; i < dim; i++)
    for (j = 0; j < dim; j++)
        for (k = 0; k < dim; k++)
            c[i][j] += a[i][k] * b[k][j];
```

6 [15 points]

For the same memory system and processor in Problem 3, consider a program which needs to read w words from DRAM and reuse them for k times. The size of the cache line is equal to one word. We assume that the cache is large enough to store the w words and has an ideal cache placement strategy (none of the data items is overwritten by others). The computation itself of the program takes n cycles. What is the total execution time of the program, including the memory read time and computation time? (State your assumptions)

7 [10 points]

What are the major differences between message-passing and shared-address-space computers? Also outline the advantages and disadvantages of the two.