**CSE 415: Introduction to Parallel Computing**

**Spring 2021 Midterm**

Monday, March 8, 2021

**Important information:**

* There are 100 points plus 7 bonus points.
* This is an open book, open notes exam, conducted via Zoom. You are expected to turn on your webcam & share your screen during the entire duration of the exam. If you do not have a working webcam, please use your smartphone as a webcam. Zoom chat will be used for Q&A.
* No calculators, cell phones or any electronic devices other than your computer (and smartphone as a webcam) are allowed during the exam. You can leave an answer in fractional form, if necessary.
* Time limit: 90 minutes (10:15 am – 11:45 am). You must return your completed exam in electronic form through your gitlab repo by 11:50 am.
* THE CONTENTS OF THIS EXAM ARE INTELLECTUAL PROPERTIES OF THE INSTRUCTOR AND MSU. YOU ARE NOT ALLOWED TO SHARE THIS EXAM OR ANY PORTION OF IT WITH ANYONE ELSE IN HARDCOPY OR ELECTRONIC FORM.

**Additional information:**

* There are two additional scrap pages at the end. Please feel free to use more paper if needed. Make sure to put your name on any additional sheets that you use.
* Use the space underneath a question to write your answers. If the space is not enough, you may continue on the scrap pages (or the additional sheets). When you do so, please put a mark to indicate that your answer is continued on the scrap pages.
* The ordering of questions is no indication of their relative difficulties, work on them in any order you feel comfortable.
* Don’t just state an answer, show your work so that you can get partial credit.

|  |  |
| --- | --- |
| NAME: | Jered Brophy |

|  |  |
| --- | --- |
| **Question #** | **Points** |
| **1** | /35 |
| **2** | /12 |
| **3** | /12 |
| **4** | /18 |
| **5** | /18 |
| **6** | /12 |
| **Total:** | **/107** |

1. **Short Answer Questions (7x5 = 35 points total)**

Provide brief (2-3 sentences) answers to each question.

1. What is simulation based engineering, and how can it help a company’s bottom line?

Simulation based engineering is a tool that is used for the design and evaluation of a complex system. Through simulation based engineering a company can measure out many things about the system they may not have recognized otherwise. For example: good estimate on price, good estimate on materials needed

1. [*True or False? Briefly explain.*] Data locality is critical only for distributed memory architectures and not for shared memory computers because shared memory computers have hardware that makes the entire memory globally accessible.

False. Data locality will be critical regardless if memory is shared or not.

1. **"Rules of Thumb":** Give your best guesses:
   1. How many clock cycles to access L1 cache?

5 cycles

* 1. How many clock cycles to access main memory?

100+

* 1. How many clock cycles to access a hard drive?

200

1. [*True or False? Briefly explain.*] Pipelining is a hardware technology that reduces the execution time of a sequential program by reducing the latency of each individual instruction’s execution.

True, through pipelining multiple instructions can be made in one line of code reducing the need for a larger amount of instructions.

1. Consider a vector processor and a loop for which the number of iterations is not a multiple of the maximum vector length. Can this loop be still vectorized? If yes, how; if no, why?

Yes, a vector can have empty slots

**2. Architecture (12 points)**

Consider a shared memory system with non-uniform memory accesses (NUMA). Suppose that it takes 10 ns to access the local cache (which is a single level cache), 100 ns to access local memory, and 400 ns to access remote memory. When you execute your program on this machine, you observe a 20% cache miss rate. Of those misses, 50% are served from the local memory, and 50% are served from the remote memory. Calculate the average memory access time (AMAT) for this execution.

**3. Latency and Bandwidth (12 pts)**

If it takes 33 microseconds to send a 5KB message and 75 microseconds to send a 12KB message between two given nodes on a distributed memory system, estimate the latency (⍺ in microseconds) and bandwidth (β in MB/s) between these two nodes.

**Note:** 1 microsecond is 10-6 seconds.

**4. Caching (18 points)**

Suppose that you are writing a board game for a simple handheld device with a single level cache. Cache block (or cache line) size is 2 bytes. To start things off you need to mark the cells as white (W) or black (B) on a board with 64 cells in an alternating sequence as follows:

BWBWBWBWBW . . . BWBWBWBW

**Version1: Version2:**

char board[64]; char board[64];

void init\_board() { void init\_board() {

int i; int i;

char c = ‘W’;

for( i=0; i<64; i+=2)

board[i] = ‘W’; for( i=0; i<64; ++i ) {

for( i=1; i<64; i+=2) board[i] = c;

board[i] = ‘B’; c = (c==’W’) ? ‘B’ : ‘W’;

} }

You are trying to choose between the two versions of the board initialization function given above. Since the platform is a low-cost handheld device, performance is critical. Assuming cold start, which version of the initialization function would you choose (or would it not matter), if the device has a cache size of

a) 32 bytes?

Version1

b) 64 bytes?

Version 2

Justify your answer by analyzing the number of cache misses under both scenarios.

**Note:** In the C language, a char is 1 byte long. For simplicity, assume that variables i, c and board (which is a pointer) are stored in registers during the entire computation and do not occupy any cache space. ‘?’ operator returns the value before colon, if the expression inside () is true, otherwise it returns the value after the colon.

**5. Roofline Model (18 points)**

Consider matrix-vector multiplication which can be implemented in single-precision (4 bytes per element) as follows:

/\* matrix-vector product loop \*/

float a[dim][dim];

float b[dim], c[dim];

for (i = 0; i < dim; i++)

for (j = 0; i < dim; j++)

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

* 1. What is the arithmetic intensity (flops/byte) of each iteration of the loop? Think of how many bytes need to be transferred over the memory bus (both reads and writes) and how many floating point operations are performed in total.
  2. Assuming a computer with a peak CPU performance of 60 Gflops, and 60 GB/s memory bandwidth, apply the roofline model to determine the peak execution speed (in Gflops). Is this computation compute-bound or memory-bound?

**6. Vectorization (12 pts)**

Which of these kernels are vectorizable, which are not? Briefly explain.

1. **for** (i=0; i < N; i=i+1)

a[i] = 2.3\*b[i];

1. **for** (i=1; i < N; i=i+1)

b[i+1] = b[i]+b[i-1];

1. **for** (i=0; i < N; i=i+1)

**if** (i%2)

a[i] = b[i]\*b[i]+b[i];

**B is vectorizable. Having b always equal a whole number is necessary.**

**Scrap page**