

CPE 412/512

Fall Semester 2016

Exam 2 (Take-Home)

Name \_Kyle Ray\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**INSTRUCTIONS: CPE 512 students must work all five (5) problems. CPE 412 students should work any four (4) problems clearly indicating which problem they will omit. In addition to electronic submissions on the UAH Canvas course administration system student are required to submit a complete hardcopy of this exam by its due date at the beginning of class on Monday November 13, 2017.**

Statement of Compliance with UAH Academic Misconduct Policies

I \_Kyle Ray\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ certify that I have worked independently of others on this test and the work that I am presenting is my own. I am familiar with the UAH academic misconduct policy as outlined in the UAH student handbook and have agreed to abide by the policies that are stated in this document.

Alternative Statement

I \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ am unwilling to sign the above statement of compliance because I cheated on the problems listed below:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Fully answer the following questions. Justify your answers by providing the page/section number reference from your text which supports your answer or by providing a complete citation of any external sources that you have used.

a) What OpenMP directive is used to provide mutual exclusion synchronization in sections of code? Give an example of its use and explain why it is needed.

**The critical directive is used for mutual exclusion.**

**#pragma omp critical (optional critical\_section\_name)**

**Usage:**

**// Some code**

**#pragma omp critical my\_critical\_section**

**{**

**// Critical Section Code**

**}**

b) Give the output of the following OpenMP code fragment if possible or discuss why this is not allowed. What is the effective scope of the sum variable? How could the answer be affected if the reduction clause is not used? Assume that there are 4 threads all together:

int i; double sum = 0.0;

#pragma omp parallel for reduction(+:sum)

for (i=1; i <= 4; i++)

sum = sum + i;

 cout << "The sum is " << sum << endl;

**Output: The sum is 10**

**The effective scope of the sum variable is shared. The answer will not be affected by removing the reduction clause. The reduction clause will create a private variable of sum for each thread which in turn will add the thread value of (i) to the private sum. Then at the end of the loop the reduction clause will add the private sums to the shared sum variable thus the reduction has just added another step. The for loop without the reduction clause will just add the value of (i) to the current shared value of sum, and because addition is commutative it will not matter which order this happens as we will get the same answer.**

c) For the following pthread code what is (are) the possible value(s) of num if there are no assumptions regarding the targeted system or the manner in which the pthread scheduler will operate? Justify your answer.

**TODO: Answer this question whenever we get the pthread code**

d) What is the difference between the ordered directive and the critical section directives in OpenMP? Give separate examples of the use of both constructs. Which constructs might be useful to support local synchronization? Which constructs are useful for or global synchronization? Explain.

**Ordered Directive - The ordered construct will make it so the code for the structured block will execute sequentially.**

**Example.**

**#pragma omp parallel for schedule(static, 1)**

**{**

**For (int I = 0; I < n; i++)**

**{**

**// Do something here first**

**#pragma omp ordered**

**{**

**// Do these things in order**

**}**

**// Do some more stuff here**

**}**

**// Do something here first**

**}**

**Critical Directive – The critical directive will lock the structured block of code after the call to the directive until the current thread is finished, at which point it will release and allow the next thread in line to execute. This allows threads to execute code that might affect shared memory variables without entering in race conditions.**

**Example**

**// Some code**

**#pragma omp critical my\_critical\_section**

**{**

**// Critical Section Code**

**}**

**TODO: Check on the implicit barrier for both of these directives. Maybe an example of global synchronization would be the parallel directive as it has an implicit barrier to wait for everyone.**

**The ordered directive is not really used for synchronization, but to make sure that a block of statements are executed in sequential order.**

**The critical section however is a good example of local synchronization as it will block a portion of code from the rest of the threads on the team until the locking thread is finished with the task at which point the previously locking thread can move on to the next set of instructions.**

e) What is the difference between the single and master section directives in OpenMP? Give separate examples of the use of both constructs.

**Master Directive – The master directive specifies that a piece of code should be executed by the master thread.**

**Example:**

**#pragma omp parallel num\_threads(thread\_count)**

**{**

**// Do some stuff**

**#pragma omp master**

**{**

**// Allow master to do something specific**

**// Maybe output something to std out**

**}**

**}**

**Single Directive – The single directive specifies that a piece of code should be executed by a single thread, but the thread is not required to be the master thread.**

**Example:**

**#pragma omp parallel num\_threads(thread\_count)**

**{**

**// Do some stuff**

**#pragma omp single**

**{**

**// Allow a single thread to do something specific**

**// Note the thread doesn’t have to be the master thread**

**}**

**}**

f) Explain why static mapping of data blocks to processors may be bad for the Mandelbrot program that was discussed in the class lecture.

**TODO: Figure out what the Mandelbrot program is and answer the question**

g) What is the basic idea of a Monte Carlo simulation. Describe how this technique can be used to solve the numerical integration problem? What makes this technique so attractive for parallel processing? What are the major issues that could affect its accuracy when it is parallelized?

**TODO: Answer this question fully**

**Basic idea of the Monte Carlo simulation is to exercise all test cases for a given system.**

**TODO: How can this be used to solve the numerical integration problem?**

**This technique is attractive to parallel programming because most use cases of the Monte Carlo technique are very time consuming. The use or parallel processing can divide the tasks up and hopefully complete the suite of simulations in a timelier manner.**

**TODO: What are the major issues that could affect its accuracy when it’s parallelized?**

2. You have been provided with a sequential exhaustive search traveling salesman program that you are to parallelize using OpenMP, or pThreads. This sequential program which is named tsp\_serial.cpp, can be found on the CPE 512/412 CanvasTM site (or can be copied from /cpe412/exam2/tsp\_serial.cpp on the UAH Jetson system).

a) Examine this program and modify it so that you can measure the execution time. Using the queuing system on the Jetson system, as discussed in class, record the execution times for a 3,4,5,6,7,8,9,10,11,12, and 13 city tour. What is the time complexity of this algorithm? Develop an equation that can be used to estimate the execution time of the base sequential algorithm. Why do you think this problem is a challenging problem to solve, and inexact heuristics are utilized instead of exhaustively searching through all solutions as is done in this program?

b) Develop a general multi-threaded version of the sequential program in either pThreads or OpenMP that will effectively divide up the amount of work that is performed. Using a single node of the Jetson queuing system, measure the execution times for an 8,9,10,11,12, and 13 city tour on a 1, 2, 3, and 4 thread implementation. For each multi-threaded implementation show the speedup, efficiency, and cost as a function of the number of cities in the tour.

**TODO: Finish this problem**

3. Expand the mm\_mult\_serial.cpp program on the CPE 512/412 CanvasTM site (or copy this program from /cpe412/exam2/mm\_mult\_serial.cpp on the UAH Jetson system) to create a hybrid multi-threaded/message passing implementation of a matrix/matrix multiplication program where the first matrix is of size lxm and the second matrix is of size mxn. Assume that the quantities being multiplied are of type float. Write the program in a general manner to allows the number of message passing processes and the number of threads per message passing process to be independently set by the user at run time. The program should be designed using the same row-wise decomposition method that was used in the two homework assignments. It should be written in a manner that will result in the total amount of computation to be divided as evenly as possible among the message-passing processes with the computation that is assigned to each message-passing processes in turn being further divided as evenly as possible between the associated threads. You are to use a combination of MPI and either OpenMP or pThreads to complete this problem. The number of threads should be a command-line parameter while the number of message-passing processes should be set by the Jetson Queuing system when you specify the number of nodes that are to be employed. The l, m, and n dimensions should also be command line parameters that can be set at run time. In other words, if the executable is named mm\_mult\_hybrid then the syntax needed to execute the code should take on the general form shown below:

srun mm\_mult\_hybrid [No. Threads per Process] [dim\_l] [dim\_m] [dim\_n])

where the number of MPI processes to be employed which is specified when the job is submitted to the Jetson queuing system.

a) Illustrate the correctness of your program for all possible combinations of thread numbers and number of MPI processes whose product is equal to 8 (i.e. 1 MPI process 8 threads per process, 2 MPI processes and 4 threads per process, 4 MPI process and 2 threads per process, and 1 MPI process and 8 threads per process) for cases where dim\_l, dim\_m, and dim\_n take on distinct values.

b) Then perform a set of timing experiments under the same set of MPI process/thread combinations of part a (but with the output suppressed) for the cases where the dim\_l=dim\_m=dim\_n=5,000. Record the execution time associated with each multi-threaded case and compare these times with that of the original serial program. What is the relative speedup and efficiency for each case. Is there a significant difference between the various parallel implementations? If so give a possible explanation as to why the execution time was not the same for each case given that the total number of threads is the same in all cases and your code was designed to evenly distribute the workload among the threads.

**TODO: Update the command line read to read in the thread count for the openmp process.**

**TODO: Make the test run and do all of the analysis, doesn’t look like adding the openmp to my implementation is actually speeding anything up.**

**TODO: Finish this problem**

4. Answer the following question for the code segment shown below assuming that we have a computing node that has 4 active cores that are dedicated to processing our problem and that were are utilizing OpenMP to parallelize a for-loop that initializes the upper triangle portion of a 100 ∞ 100 matrix to the values returned by the function, init\_element(x,y), which itself always executes in constant amount of time regardless of the values associated with its two arguments.

#pragma openmp parallel for schedule( ... )

for (i = 0; i < 99; i++) {

for (int j = i+1; j < 100; j++) {

a[i][j] = init\_element(i,j);

}

}

Notice that the arguments to the schedule() clause have been left undefined. Below are six example schedule clauses that could be used. Rank these clauses from slowest to fasted by closely examining the characteristics of this problem. To do this note that each iteration of the inner loop above does just one assignment and we can estimate the execution time by counting how many assignments each thread does. (Also note the total number of assignments the problem performs is exactly 4,950 assignments). For each schedule clause, estimate how long the parallelized loop will run. Explain how you arrived at your estimates.

**TODO: Finish this problem**

5. Expand upon the bounded\_buffer.cpp that is provided on Canvas (or copy this program from /cpe412/exam2/bounded\_buffer.cpp on the Jetson system) to create two separate OpenMP implementations of the producer/consumer bounded buffer-problem discussed in class that utilizes a common shared memory (between threads) and counting semaphores to ensure proper synchronization. [Note you will also need to link this program with the util.o file that is also provided on Canvas or on the Jetson system at /cpe412/exam2/util.o] One implementation should use OpenMP's standard parallel and work sharing data parallel constructs to start up the separate consumer and producer threads. The other implementation should utilize the OpenMP tasking model. Verify that both models function correctly and answer the following questions:

a) Carefully explain the characteristics of the producer and consumer bounded buffer problem? What are the major synchronization challenges that are involved.

b) Describe the function of the main, consumer, and producer module? What are the output files diary and con\_? and prod\_? reporting?

c) Execute both versions of the producer/consumer program with two producers, two consumers, number of messages of 100 and with a buffer size of 5. How many actual threads are generated when you run each version? Are they the same? How do they relate to the number of producers/consumers that are specified at run time?

d) How is it possible for the buffer size to be smaller than the number of data that is to be passed between producer and consumer? Explain how this is actually implemented in the program.

e) Experiment with different size buffers, messages, and consumer processes and producer processes. From the diary file and other files can you determine if the semaphore releasing strategy is fair? Explain your answer.

**TODO: Try to finish this problem and hope for some credit.**