Create a document that has your answers to the 4 “book” questions. Be sure to show your work. Correct answers without a show of work will receive partial credit only. You may use a calculator, spreadsheets, or you may write a program to help you arrive at an answer. However, answers must be yours and yours only. Although you may discuss the homework with your fellow students, do no discuss or share solutions. A violation of this policy will result in you failing the course. Please ask if something is unclear. This homework also includes a programming task. All submissions are via Canvas.

**Question 1**: (5 pts) On your first day of the job at *Concurrency Computer R Us*, the VP calls you into her office and tells you the following: “We've designed a pipeline that includes 8 distinct steps. Assume that the pipeline does not stall, that it is fully loaded, and each stage of the pipeline takes an equal amount of time (the pipeline is balanced). Under these conditions, how many independent instructions must be executed so that the speedup of the pipelined versus a general-purpose, non-pipelined architecture is 8?”

**Question 2**: (5 pts) Assume that you have a single (non superscalar) 6-stage pipeline, that is made up of stages A-F. The pipeline is originally empty. The A stage consumes 1ns, the B stage 3ns, the C stage 1ns, the D stage 4ns, the E stage 5ns, and the F stage 1ns. For a program with 8,000 independent instructions, what is the average number of nanoseconds that an instruction is stalled in the pipeline? Do NOT count the time WAITING to enter the pipeline as a stall. Do NOT balance the stages. As soon as a stage is available and an instruction needs that stage, then the stage MUST be used (ie, do not modify the order in which the instructions are processed and do not introduce artificial delays).

**Question 3**: (5 points) Assume you are an employee of *We-Tolerate-Errors*, a software company that specializes in writing programs intended to be run concurrently. You are working on a program for which you have the executable but not the source code (you cannot modify the program). All that you've been told is that the program has 3 threads with 3, 4, and 2 instructions, respectively:

- **Thread A**: a1 < a2 < a3
- **Thread B**: b1 < b2 < b3 < b4
- **Thread C**: c1 < c2

where < specifies the ordering that we’ve seen in lecture. The three threads are run concurrently, so although instructions in a thread are executed sequentially, the orderings among instructions from different threads is unpredictable. The following 2 are example instruction histories that might be realized:

- **History 1**: a1 < a2 < b1 < a3 < c1 < c2 < b2 < b3 < b4
- **History 2**: a1 < b1 < b2 < c1 < c2 < a2 < b3 < a3 < b4

Also assume that dependencies among variables in use by the threads necessitate that the following constraints (instruction histories) be maintained:

- **Constraint 1**: a1 < b3
- **Constraint 2**: c2 < a2

*We-Tolerate-Errors* guarantees their software to be correct to a certain fault tolerant level. The question you must answer: **Assuming each possible history is realized only once, what percentage of the invocations of the software (when run concurrently) will give an incorrect result?** An Incorrect result is achieved when an execution history violates one of the constraints.
**Question 4**: (5 pts) Assume that you have a computer whose cache access time is 14ns and main memory access time is 140ns. Further assume that you have a computationally intensive program, *quadratics v2.0*, that is crucial to your employer’s business. *Quadratics 2.0* has a cache reuse ratio of 0.65. A new version of that program, *quadratics v3.0*, has just been released. Unfortunately, the cache reuse ratio of the trial version of *quadratics v3.0* when run on your computer is 0.55. Your boss asks you to determine which hardware upgrades to the cache or main memory are needed (the size of main memory and size of cache must stay the same) so that *quadratics v3.0* runs just as fast as does *quadratics v2.0*. Give a quantitative (numerical) answer, and not a qualitative one. In other words, do not just say, “Improve the cache access time.”

**Programming Task** (20 points)

This programming task builds on what you learned in the threading labs. You will write a program to calculate the product of two square matrices. You’ll perform the calculation with and without threads, and compare the run-times.

Matrix multiplication is a binary operation that takes as input a pair of matrices, and produces an output, the product matrix. It is an often-used procedure in many algebra-intensive applications and high-performance programs. For the square matrices A and B shown above, their product is the following:

\[
A = \begin{pmatrix}
A_{11} & A_{12} & \cdots & A_{1n} \\
A_{21} & A_{22} & \cdots & A_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
A_{n1} & A_{n2} & \cdots & A_{nn}
\end{pmatrix}, \quad B = \begin{pmatrix}
A_{11} & A_{12} & \cdots & A_{1n} \\
A_{21} & A_{22} & \cdots & A_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
A_{n1} & A_{n2} & \cdots & A_{nn}
\end{pmatrix}
\]

\[
AB = \begin{pmatrix}
(AB)_{11} & (AB)_{12} & \cdots & (AB)_{1n} \\
(AB)_{21} & (AB)_{22} & \cdots & (AB)_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
(AB)_{n1} & (AB)_{n2} & \cdots & (AB)_{nn}
\end{pmatrix}, \quad (AB)_{ij} = \sum_{k=1}^{m} A_{ik}B_{kj}
\]

Refer to the lecture slides for a review of matrix multiplication.

**Coding Details**

- Write a function for calculating the product of square matrices, using the naïve (above) method, which has a runtime of \(O(n^3)\), where \(n\) is the dimension of the square matrix.

- When you generate your matrices, they should be composed of doubles or integers. It is up to you whether the matrix values are randomly generated, or hard-coded (for example, matrix A can be composed of all 1s, and matrix B composed of all 2s).

- Use any programming language of your choice. I recommend using C++, for which you’ve been shown (in lecture and in lab), how to implement threads. If you select another language it is up to you to make sure that the language enables threading. Include a README file with compile and invocation instructions.

- Write a program with a main routine that invokes the matrix multiplication method. Your program must take as input (as command-line arguments) the following 3 arguments: a) designate how many times the matrix method should be invoked, b) whether threading should be used (a thread should consist of invoking the matrix multiplication method), and c) how big the matrix should be (see next section). If threading is specified, your program should create a separate thread for each calculation. If threading is NOT specified, then your program should perform the calculations in serial.
Running your Program (see Figure 1 for pseudocode schematic)

- Run your program on the Linux account that you have access to.
- Run the program on matrices that are 200 x 200, 400 x 400, 600 x 600, and 800 x 800
- For each of the 4 sizes of matrices mentioned above, have your program perform matrix multiplication 1, 2, 4, 8, 16 and 32 times, each with (threaded) and without threading (serial)
- Time each run of your program, using the linux/unix `time` command, as was shown in lab or using the wall time routine.

Figure 1: Pseudocode for four separate invocations of the program. Note that the `matrixProduct()` invocations for the non-threaded versions could have been invoked using for loops instead of sequentially calling those methods one-by-one. Numerical values of outputs are for demonstration only.

Analyze the run-time

- Record the run-times of your program, and tally the results in a table. For example, the following table has the user run-time values (in seconds) for the threaded invocation of the program. The values in the table are NOT real; they are shown only for demonstration purposes.

<table>
<thead>
<tr>
<th>Threaded</th>
<th>Matrix Size</th>
<th>1x</th>
<th>2x</th>
<th>4x</th>
<th>8x</th>
<th>16x</th>
<th>32x</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200 x 200</td>
<td>0.114</td>
<td>0.114</td>
<td>0.114</td>
<td>0.014</td>
<td>0.171</td>
<td>2.272</td>
</tr>
<tr>
<td></td>
<td>400 x 400</td>
<td>0.930</td>
<td>0.830</td>
<td>0.956</td>
<td>0.963</td>
<td>1.102</td>
<td>1.462</td>
</tr>
<tr>
<td></td>
<td>600 x 600</td>
<td>2.163</td>
<td>2.128</td>
<td>2.294</td>
<td>2.195</td>
<td>2.790</td>
<td>3.339</td>
</tr>
<tr>
<td></td>
<td>800 x 800</td>
<td>7.018</td>
<td>6.715</td>
<td>6.762</td>
<td>6.696</td>
<td>8.605</td>
<td>45.456</td>
</tr>
</tbody>
</table>
• Plot the run-times of your algorithm, for the four matrix sizes, when threading was used, and when it wasn’t. Two sample plots are shown in Figure 2. Be sure to label your axes.

Figure 2: Execution times for threaded (left) and serial (right), for different matrix sizes, for different invocation counts (1x dark blue, 2x orange, 4x yellow, 8x green, 16x purple, 32x light blue).

Writeup

Compose a brief document (1 page should suffice!) that includes:

• Tabulated run-times
• Correctly labeled (axes, titles, units, etc.) plots of run-times
• High-level architecture specs about your computer (number of cores and cache/memory sizes). See Lab 1 for how to gain access to that information.

The document should also include at MOST 2 paragraphs discussing/explaining the following:

Discussion point 1: Does threading improve the performance of your program? If yes, is it reasonable to assume that such a performance gain should always exist, or only under special circumstances? Explain.

Discussion point 2: Is there a thread “limit”, at which point adding more threads does not significantly improve the performance of your program? Regardless of whether you answer yes or no, explain this in the context of the system architecture for the linux account where you ran your program.

Submission and Rubric

Submit to canvas two documents (answers to 4 book questions and code write-up), along with your code.

<table>
<thead>
<tr>
<th>Question/Component</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 1</td>
<td>5</td>
</tr>
<tr>
<td>Question 2</td>
<td>5</td>
</tr>
<tr>
<td>Question 3</td>
<td>5</td>
</tr>
<tr>
<td>Question 4</td>
<td>5</td>
</tr>
<tr>
<td>Programming task – program written, commented, and README included</td>
<td>3</td>
</tr>
<tr>
<td>Programming task – program compiles and runs as intended</td>
<td>5</td>
</tr>
<tr>
<td>Programming task writeup – table of run times</td>
<td>2</td>
</tr>
<tr>
<td>Programming task writeup – two plots</td>
<td>4</td>
</tr>
<tr>
<td>Programming task writeup – answers to 2 discussion points</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
</tr>
</tbody>
</table>