Speeding Up via OpenMP, TPU, and GPU

Date assigned: Week 9 (March 22, 2024)
Date due: Wednesday, **April 24**, 2024 at 11 a.m. sharp. (No late submission allowed.)
Submission at: Your section's GCR
Max points: 2500 points
Collaboration: Not allowed. It is a single-person assignment. You can discuss only the requirements if there is any confusion in those.

Note: Please carefully read the full assignment first before starting.

Disclaimer: Read Carefully!

This assignment uses services from Google Cloud. To complete this assignment, you will rely on a $300 new user free credit provided to each new user, and partly on small personal funds such as ~ $2 per hour for TPU and $1 per hour for a GPU. You will need to provide your credit/debit card information to activate your GCP account. It is your sole responsibility that whatever services you use at GCP must remain within the free tier credits. Any overages incurred will be your sole responsibility. Course staff (instructor, TA), FAST School of Computing, CS department, and the FAST university bear no responsibility (implied or explicit) in this regard.

You should conserve your free credit as much as possible because you will need it again for the third assignment.

As a precautionary step, always turn off / terminate all services when not actively using them. Forgetting to shut off resources, or ignoring what resources are in use are the common reasons for taking thousands of dollars in overage bills. While doing this assignment, always time-bound your work (for example, you must finish before the hour ends so that you can turn off a VM or other resources before the hour completes.)

This assignment aims to allow you to speed up a serial code using OpenMP, TPUs, and GPUs.

What to submit:

(1) You will need to provide us with a report that will journal your journey on how you solved the problems posed in this assignment, along with the specific answers to the questions.
(2) Your code with a README that will tell how to execute your code

We will continue with our matrix multiplication example from the first assignment. You need to follow the input and output rules as we laid them in the first assignment.

Part 0: [100 points] Create a new Google Cloud Platform (GCP) account to get a $300 free tier credit. Paste appropriate proof that you have created your personal account at GCP with $300 free credit. You will need to teach yourself how to do so using online resources.

Part 1: [500 points] Parallelize your c-based matrix multiplication code (from assignment 1) using openMP. You can keep your SIMD instructions enabled. You will use the same target machine you used in assignment 1. Please provide the execution time and add the graph to the graphs you already have from assignment 1.

Part 2: [500 points] Repeat part 1 with an instance of a c3-standard-22 machine in GCP. At the time of this writing, this machine has 22 virtual cores and costs $1.416712 per hour in one of the regions. Find the region which can give you the cheapest such instance. Again provide the execution time and add the graph to the previous ones.

Part 3: [500 points] GCP provides special accelerators primarily for ML workloads called Tensor Processing Units (TPU). Get a VM instance with just one TPU, from the region providing it at the most economical price. You must research how to program TPU to do matrix multiplication using its systolic array. Report the execution time and add it to the graphs. Also, provide some form of proof (such as a picture of the result or screencast) to show us that you used a TPU from your account. You might not need to select the latest generation of TPU. You might use an earlier version of TPU if your free credits are not supporting TPUs.

You should first understand and write the code before getting a TPU. Once you get a TPU, you should try to complete your work as soon as possible. Remember some services are billed on a per-second basis while others might be on a per-hour basis. Carefully understand the pricing model associated with the resources you are using.

Part 4: [500 points] Research what types of GPUs are available at GCP and which one will be suitable for matrix multiplication of the 4096 X 4096 matrix. Learn to program your selected GPU and do matrix multiplication. Provide execution time, and add to the graphs. Tell us which

GPU you used and in which region. Also, provide screencast proof that you used a GPU from your account.

Part 5: [200 points] Comparing the speed up you achieved via OpenMP (on a 22-core machine), one TPU, and one GPU which gave you the best results? Provide reasons why one gave a better performance than the others.

Part 6: [200 points] We found raw performance in the previous part. Now, let's find out performance per watt by finding out how much energy was consumed for the work we did. Now based on performance per watt, which is better among the three (22-core general-purpose machine, GPU, or TPU)? Draw a graph and also tell us how you calculated energy/power?