# Developing Soft and Parallel Programming Skills Using Project-Based Learning

Project Report

Fall-2018

Submitted By PAJ_VR

Prashant Vemulapalli, Anthony Davis, Jaiana Butler, Vincent Lee, Ryan Barrett

## Planning and Scheduling(Task 1)

| Name: | Email: | Task(s): | Duration | Due Date | Note: |
|---|---|---|---|---|---|
| **Prashant Vemulapalli** | Pvemulapalli1@student.gsu.edu | **Creating the report and working on questions** | **3 hours** | **12/03/2018** | |
| **Ryan Barret** | **Rbarrett7@student.gsu.edu** | **Planning and scheduling and working on the pi** | **6 hours** | **12/03/2018** | |
| **Anthony Davis** | **Adavis183@student.gsu.edu** | **Collaborating with the report and working on questions** | **6 hours** | **12/03/2018** | |
| **Jaiana Butler** | **Jbutler36@student.gsu.edu** | **Recording the video and collaborating with the report** | **2 hours** | **12/03/2018** | |
| **Vincent lee** | Vlee18@student.gsu.edu | **Working on the questions and working on the pi** | **2 hours** | **12/15/2018** | |

## Parallel Programming Skills(Task 3)

### a) Foundation:

1) Read this paper " Introduction to Parallel Programming and MapReduce " and answer the following questions:

- What are the basic steps (show all steps) in building a parallel program? Show at least one example.

Ans: 1. Determine which parts of the program, if any, can be processed concurrently.

2. Determine what sort of implementation is to be used for the task that is to be processed in parallel.

3. Determine how to the task will be parallelized with the implementation that has been chosen.

Example: summing an array

1. Can this task be parallelized? Yes, the task of summing the entire array can be broken into separate tasks, that can be processed concurrently.

2. The task of parallelizing the summing of an array can be implemented using master-worker implementation.

3. The task will be parallelized by having the master initialize the array and send each worker the subarray that is to be summed. Each worker will then sum their assigned subarrays, and then each return their result to the master, who will then return the final sum.

- What is MapReduce?

Ans: MapReduce is a programming model that uses a distributed system of machines which do parallel programming with the map and reduce functions to abstract the processing of large amounts of data.

- What is map and what is reduce?

Ans: Map is a function that take a pair of values in the form of a key and value and returns a set of pairs each consisting of another key and value. The reduce function takes input as a key and a set of values for the key and merges all the inputs for a particular key together into a single set.

- Why MapReduce?

Ans: MapReduce allows for calculations to be done on large amounts of data, without needing to take into consideration the exact details such as parallelization, data distribution, load balancing and fault tolerance.

- Show an example for MapReduce?

Ans: A task that could be completed using MapReduce is sorting emails by recipient. Email archives could be passed to the map function, and for each email sent to a target email, the map function would return a set of key-values pairs with the key being the target email address, and the value being the email address the email was sent from. Then the reduce function would output a pair with the target email being the key, and the values being a list of the sender email addresses.

- Explain in your own words How MapReduce model is executed?

Ans: After splitting up the input data into different shards and starting up multiple copies of the program on other machines, all but one of these copies are assigned the role of workers, while one is assigned the role of master. The master then assigns a certain number of map tasks “M” and a certain number of reduce tasks “R” to the workers. Then the map workers parse the input from their respective input shards and passes any resulting key-value pairs to the map function. Once the map function produces the intermediate key-value pairs, they are temporarily stored in memory, and eventually are stored in a local disk. The location of the pairs on the local disk are passed to the master, who then passes the location to the reduce workers. When the reduce workers receive the location of the pairs, it reads the pairs using remote calls, and then groups all the data together by sorting the data using specific occurrences of keys. The sorted data set is then iterated over by the reduce workers who then passes each unique key, and its corresponding values over to the reduce function. The results of the reduce function is included in the output file that is returned to the user program by its MapReduce call, when all map/reduce tasks are finished, and the master notifies the user program.

- List and describe three examples that are expressed as MapReduce Computations?

Ans: Count of URL Access Frequency:

The map function parses webpage request logs and returns a set of pairs for each URL consisting of the URL for the key, and 1 for the value. Every unique URL and its associated values is passed to the reduction function which sums all values belonging to a URL to produce a final pair listing the URL and total count.

Reverse Web-Link Graph:

The map function parses webpages using a target URL for the key, and for each webpage that links to the target URL, outputs a pair with the target URL, and the URL of the source webpage.

The reduce function then outputs a pair of the target URL, and a list of all source URLs that link to the target URL.

Inverted Index:

The map function parses documents using a given word for the key, and outputs pairs with the given word as the key, and the document id as the value. The reduce function then outputs a pair consisting of the word for the key, and a list of all corresponding document ids for the value.

2)When do we use OpenMP, MPI and, MapReduce(Hadoop) and why?

Ans: OpenMp is used when you want to implement parts of your code in parallel with the different processors sharing the same memory and is beneficial as in many situations it may be more efficient than writing manual code. MPI is used when you want parallel code to run on multiple separate machines each with their own memory, due to it using a distributed memory implementation in which processors have separate memory. Hadoop MapReduce is used when there are very large amounts of data to be processed, and when some processing errors might be expected when handling this data, due to the its ability to handle fault tolerance.

3) In your own words, explain what a Drug Design and DNA problem is in no more than 150 words?

Ans: When pharmaceutical companies design medicine, they find ligands, that can be used to change the shape of a protein. The shape of a protein in the body determines what function it serves, and is first determined by the DNA, which serves as a blueprint for proteins. To find these ligands, drug design software is created to test generated ligands for the best fit to a protein, scores them based on how well they fit and produce the desired change, with the highest scoring ones being created and tested. One software implementation is representing the protein as a string of characters, generating ligands as character strings of random length and characters, scoring each ligand using parallel processing, and sorting based on scores. Command line arguments could be used to determine the number of threads to compute ligand scores, the maximum length for the ligand, and how many ligands to score.

## 4) Parallel Programming Basics: Drug Design and DNA in Parallel:

**Measure Run-Time:**

| Implementation | Time(s) |
|---|---|

| | |
|---|---|
| dd Serial | 122.35 |
| dd omp | .02 |
| dd threads | .02 |

| Implementation | Time(s) 2 Threads | Time(s) 3 Threads | Time(s) 4 Threads |
|---|---|---|---|
| dd_omp | .02 | .04 | .23 |
| dd_threads | .02 | .05 | .16 |

**Discussion Questions:**

1) Which approach is the fastest?

Ans: The c++11 and openmp approaches are very similar, but c++11 seems to be very slightly faster. Sequential is by far the slowest.

2) Determine the number of lines in each file (use wc -l). How does the C++11 implementation compare to the OpenMP implementation?

Ans: The c++11 implementation has 208 lines, and the open mp has 194 lines. They seem fairly similar as far as the methods they contain.

3) Increase the number of threads to 5 threads. What is the run time for each?

Ans: 1) Time -p ./dd_threads 5 = .645

2) time -p ./dd_omp 5 = ..99

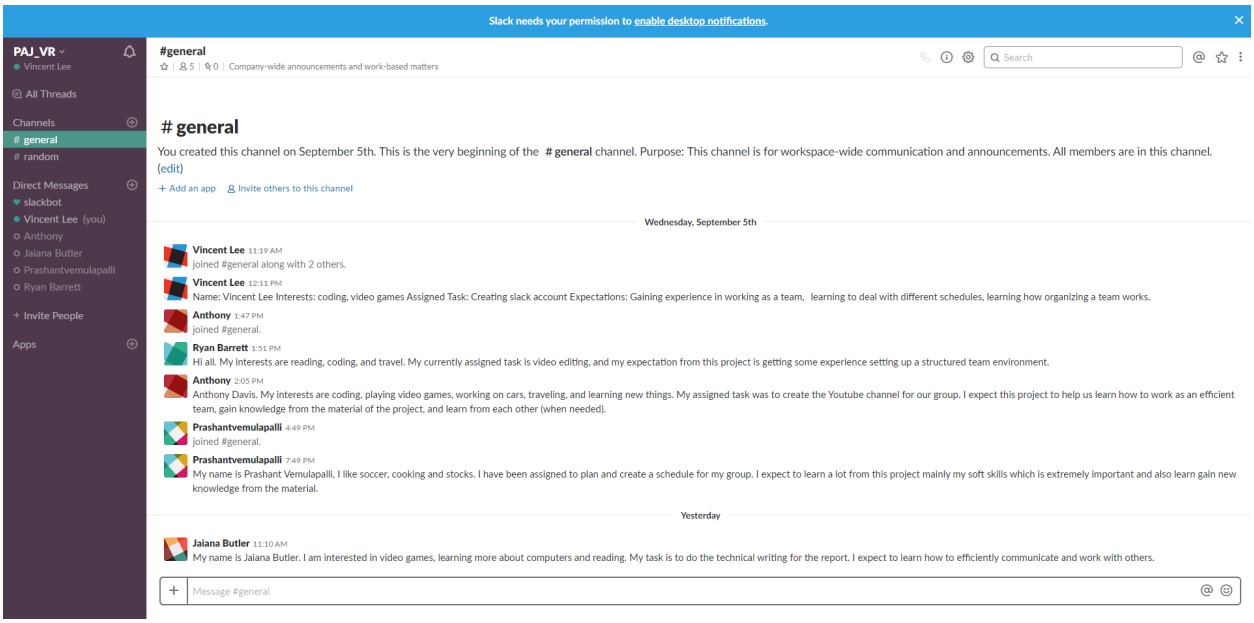
4) Increase the maximum ligand length to 7 and re-run each program. What is the run time for each?

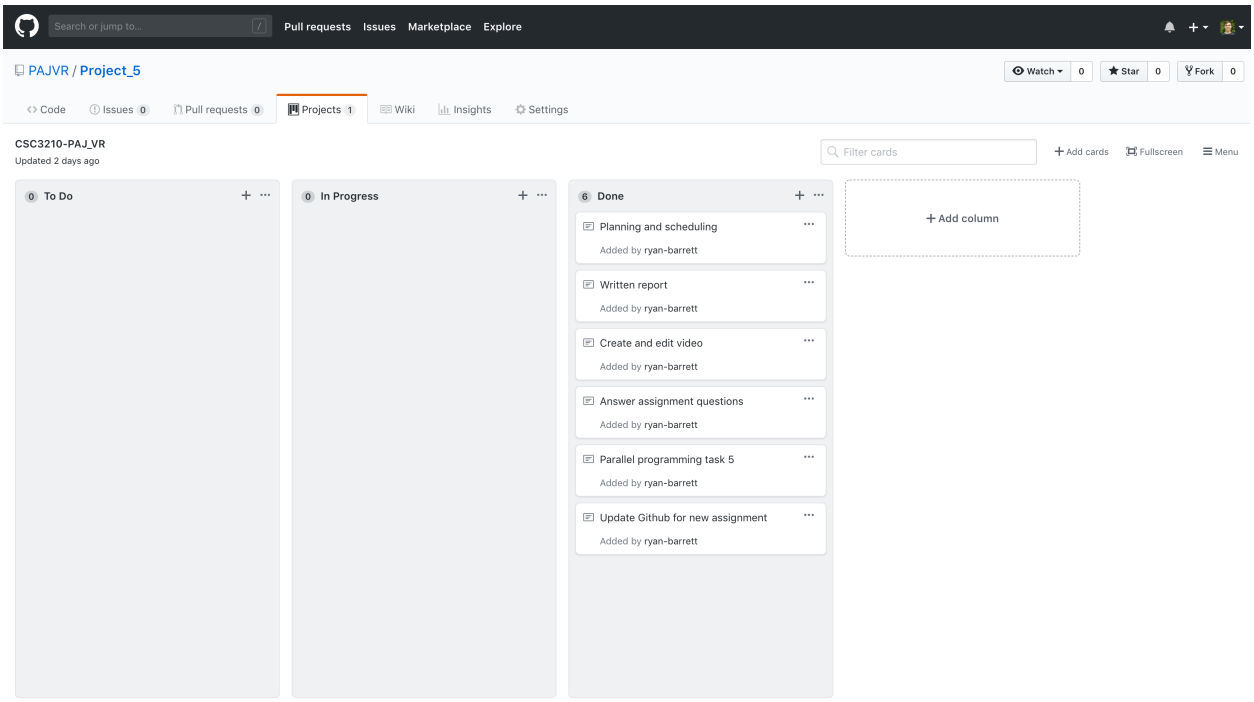
Ans: dd_omp = 0.02 & dd_threads = 0.02

## Appendix

https://youtu.be/PMkSzjRplFk

pajvr.slack.com


Slack needs your permission to enable desktop notifications.
PAJ_VR
Vincent Lee
All Threads
Channels
# general
# random
Direct Messages
slackbot
Vincent Lee (you)
Anthony
Jaiana Butler
Prashantvemulapalli
Ryan Barrett
Invite People
Apps
#general
Company-wide announcements and work-based matters
# general
You created this channel on September 5th. This is the very beginning of the # general channel. Purpose: This channel is for workspace-wide communication and announcements. All members are in this channel. (edit)
Add an app
Invite others to this channel
Wednesday, September 5th
Vincent Lee 11:19 AM
joined #general along with 2 others.
Vincent Lee 12:11 PM
Name: Vincent Lee Interests: coding, video games Assigned Task: Creating slack account Expectations: Gaining experience in working as a team, learning to deal with different schedules, learning how organizing a team works.
Anthony 1:47 PM
joined #general.
Ryan Barrett 1:51 PM
Hi all. My interests are reading, coding, and travel. My currently assigned task is video editing, and my expectation from this project is getting some experience setting up a structured team environment.
Anthony 2:05 PM
Anthony Davis. My interests are coding, playing video games, working on cars, traveling, and learning new things. My assigned task was to create the Youtube channel for our group. I expect this project to help us learn how to work as an efficient team, gain knowledge from the material of the project, and learn from each other (when needed).
Prashantvemulapalli 4:49 PM
joined #general.
Prashantvemulapalli 7:49 PM
My name is Prashant Vemulapalli, I like soccer, cooking and stocks. I have been assigned to plan and create a schedule for my group. I expect to learn a lot from this project mainly my soft skills which is extremely important and also learn gain new knowledge from the material.
Yesterday
Jaiana Butler 11:10 AM
My name is Jaiana Butler. I am interested in video games, learning more about computers and reading. My task is to do the technical writing for the report. I expect to learn how to efficiently communicate and work with others.
Message #general


Search or jump to...
Pull requests
Issues
Marketplace
Explore
PAJVR / Project_5
Watch 0
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Projects 1
Wiki
Insights
Settings
CSC3210-PAJ_VR
Updated 2 days ago
Filter cards
Add cards
Fullscreen
Menu
0 To Do
0 In Progress
6 Done
Planning and scheduling
Added by ryan-barrett
Written report
Added by ryan-barrett
Create and edit video
Added by ryan-barrett
Answer assignment questions
Added by ryan-barrett
Parallel programming task 5
Added by ryan-barrett
Update Github for new assignment
Added by ryan-barrett
Add column