Project Layout

Group Members: Amruta Abhyankar (aa1589), Michael Wu (mw811), Christopher Salandra (cms647)

Project #1: Map Reduce

For the first project of the semester, the group was assigned to work on a Map Reduce for a given input text file. The text file is a normal file you would see in any document. The team needed to write a program that would take the input file what was given, and output another text file that would print out all the words that was used in the text file. We must solve these problems in two different ways. For the first way, we must create the file output using multithreading. Having multiple threads running at once, retrieving their own section of data that they received. And then finally sort the words within each thread, and then give a output file with all the words that were sorted through all the treads combined. With the second part of the project, we need to find the same output file, but only this time we'll be using multiple processes. With this one, we call and use multiple processes. Each process will contain a specific amount if words that was gathered from the input file.

In this project, the team created a program that would take in the input file and either make several threads or processes. From there the words in the input file would be distributed as evenly as possible to the several threads or processes. From there each threads or processes' will go into a wordsort, where the works are sorted out in alphabetical order. This way all the similar words are next to each other. From there we then did a number sort with ints. This counted the amount of times the same exact word in the thread or process appeared. Finally, once all the threads or processes are individually sorted, we did a merge sort that would combine all the words from the threads or processes together. This will also add the number that represented the amount of times a certain word was represented in each thread or process. Then we created a function that gave the output on a output text file.

From the testing and evaluation of each framework, we had noticed a few key concepts. The first thing to begin with was we were able to tell that the threads ran and edited the output data faster than the processes framework. This may be true due to it being quicker to switch between threads than it is to switch between processes. Though the threads took less time to work on, it also had the most problems. It took the group longer to get an good understanding of

thread than it did process. It also seemed that the thread part of the project was causing more problems for the group than the processes. So though we thought that the threading would be the easy and most useable decision, we soon realize the difficulties behind the program and making the code function into the commands and tasks we want it to do.

During the project, the major issues we ran into was our knowledge based on the what the project demanded. Some of us never worked with threads and processes that much before, so it was a big learning curve to get the hang of implementing all the code to get our project to run. One of us never even used the iLab machines before, so the who working environment was a change and difficult to get use to. During the project, we were learning how the code worked as we were implementing it. It something the group had to adapt to. And it was harder for some to adapt to more than others. So we were teaching each other at many times the new ideas we had to use in order to get our mapreduce to function properly. When it comes to the actual code, there were many difficulties the group faced. One was trying to control the C++ objects in the shared memory, which we later found out was not possible. Another idea that was strange to us was using the C style struct. Then when it came to the extra credit, it was hard to use the multithreading to read from the input files. Sitting all the lines and then putting the strings or numbers into a vector of multiple maps. Even with all these difficulties, the team was able to come together and figure out how to create a project that completed the functions we were tasked with.