A COMPARATIVE ANALYSIS OF LOCALIZED COMMAND LINE EXECUTION VERSUS OF THE REMOTE EXECUTION THROUGH COMMAND LINE AND TORQUE SUBMISSION OF MATLAB® SCRIPTS FOR THE CHARTING OF CRESIS FLIGHT PATH DATA

JerNettie Burney & Robyn Evans
Mentor: Je'aime Powell Principal Investigator: Dr. Linda B. Hayden
1704 Weeksville Road, Box 672
Elizabeth City, North Carolina 27909

Abstract- The Polar Grid team was able to install the newest cluster system, "Madogo", at Elizabeth City State University. With this cluster, the team focused on testing for the rate at which Madogo will be able to process and run different jobs. The team pulled information of flight paths from Greenland into MATLAB so that they could be converted from text files into actual script files. With these m-files, the team was able to create a code within MATLAB® that could plot the flight path data into a graph with the axes of the graph being labeled latitude for the x-axis and longitude for the y-axis.

The team took the master code of the graph and ran jobs through the command line of MATLAB® to Madogo and then ran the code through the command line of MATLAB® to the cluster system located at Indiana University. The team was then able to compare the speed of the completion of jobs from a local area versus one that is at a far distance and had the jobs submitted from offline. A comparison was done with TORQUE submission versus MATLAB® submission to see of which program would be able to complete the jobs faster.

The paper focuses upon what the procedure was in order to complete the research along with the conclusion that was reached.

Keywords – ANOVA, binary, cluster, jobs, LINUX, MATLAB®, node, Perl Script, ssh key, TORQUE

I. NATURE AND BACKGROUND OF THE STUDY

A. Introduction

The Polar Grid team focused on finding the most efficient and inexpensive way to run jobs. MATLAB was utilized in the conversion of datasets gathered from the 2007 Greenland field deployment from texts file into MEX files. The team then did research on plotting one of the datasets into a single graph. This paved the way for all of the data sets to be plotted onto the same graph. The focus was then directed towards finding ways to make the data into one stand-alone application.

B. Statement and Background of the Problem

Formed by the National Science Foundation (NSF) in 2005, the Center for Remote Sensing of Ice Sheets (CReSIS) strives for the development of technologies that allows them to be able to measure and predict any change of sea level caused by a mass balance of ice sheets, located in both Greenland and Antarctica.

CReSIS offers research in a selection of majors for students and faculty; to work together in unison with various scientists and engineers no matter what the nationality; and aid to the to the constant study of what different factors may play a part in the ongoing change of the climate.

Within CReSIS, there are six universities that are also aiding in their efforts to conduct research upon this topic. These universities include the University of Kansas where the CReSIS headquarters is stationed, Elizabeth City State University, Haskell Indian Nations University, Ohio State University, Pennsylvania State University, and the University of Maine.

During the 2008 Greenland Research Expedition in Ilulissat, Greenland there were a few selected representatives involved to assist in conducting research. These included representatives from Elizabeth City State University (ECSU), the University of Kansas (KU), and Indiana University (IU). Dr. Eric Akers—who is a professor at ECSU—and Mr. Je'aime Powell, a graduate student who is also at ECSU, joined in their research efforts in order to fulfill their project requirements at the Center of Excellence in Remote Sensing Education and Research (CERSER).

There were six antennas that were installed onto a Twin Otter aircraft were able to collect data every 100 milliseconds. Because of the short amount of time that was given for the antennas to collect the data and also the large amount of distance that aircraft traveled, gathered datasets required hundreds of gigabytes of storage. Before anything could take place, this data need to be stored in a redundant manner. If a hard disk was corrupted or unsuccessful without redundancy, that season's information was lost.

To solve this issue, a 30Tb (14 TB usable) RAID 10 storage center, was use so that any data that was stored was copied as soon as it was sent to the cluster by using "rsync". Rsync was a method that used error checking while copying. Backing of all data was accomplished by using 2Tb external drives that reflected each day independently.

The MEX files used by the Center for Remote Sensing of Ice Sheets (CReSIS) are solely dependent on MATLAB®. In order to compile their data, a MATLAB® compiler was needed. However, in order for this compiler to function properly, the MATLAB® Distributing Tool Kit must be installed. The toolkit has a cost $50,000 for seven nodes. The purposes of the team's research became to offer CReSIS with a more productive, versatile, and inexpensive method for processing their data.

C. Hypotheses

The team hypothesized that running jobs through TORQUE on Madogo would be the most efficient since it can access MATLAB without having to bring up the actual program. It was also hypothesized that running the jobs through the Quarry cluster of Indiana University command line of MATLAB would be the less efficient to submit, because it must be accessed through remote computing using the SSH terminal function.

D. Limitations and Delimitations

A conflict arose while making the MEX files into binary standalone file pertaining to shared libraries. The MATLAB Computing Complier (MCC) could not recognize the G++ program installed onto the system as a compiler. Therefore, the scripts could not be compiled as a C or C++ program. We used the MATLAB® Deployment Tool to be able to create a distribution. Thus, the G++ program could not be used in compiling the scripts.

Also the Elizabeth City cluster—Madogo was not used to make a direct TORQUE comparison. For TORQUE, there was no shared home directory along with no shared *SSH* keys which prevented the use of the ECSU cluster.

II. Review of Literature

A. Prior Research

At Indiana University, a TORQUE Perl script was created as a result of their research in finding a way for people to submit MATLAB batch jobs on the Quarry cluster. This was essential to the team's research in that TORQUE was needed to use as a means to submit the MATLAB jobs to Quarry, and returning the team's job scheduler run times.

A research team composed of researchers from Elizabeth City State University, Indiana University, and the University of Kansas, obtained the datasets used for this research project. To do this, they utilized a twin otter plane with antennas attached below the wings. The aircraft contained GPS which had the ability to record the flight path. This research was done during the 2007 Greenland Research Expedition in Ilulissat, Greenland.

Mr. Je'aime Powell conducted further research while at ECSU. In his research, the ANOVA statistical process was introduced and revealed the different factors needed in order for the use of it to be used properly.

III. Methodology

A. Definition of the Population

There were two types of data collected for this research project: datasets and the job scheduler run times.

We were able to access the datasets through the CReSIS website on the Greenland Data page. The data was available in three different formats: MEX, .txt, and as a PDF file formats. The team chose to use the .txt files. Txt files were chosen due to the incapability of the MEX files being readable. The files were converted into a MATLAB® format, or a .m file, for editing. There were a total of twenty-six datasets, each file containing between 3,000 to 8,000 sets of longitude/latitude points.

This data was uploaded into MATLAB® from the command line. We timed how long it took for the job to run and repeated the process twenty times. The batch was run again on both the Indiana University's Quarry cluster via the command line and through TORQUE. The times it took to run on these systems became the team's run time data and were used for a comparative analysis.

B. Procedure

The group started off by defining each dataset by saving the files in a common directory. The files were then converted, changing them from .txt files to .m files. By doing this, the group was able to edit the files, deleting the column headings; this step would later allow us to use the load command.

The next step was to load the files into MATLAB®. To do this the 'load' command was utilized. The plot command was then used to plot the dataset on a scatter-plot graph. Next, a method had to be identified to upload multiple datasets into MATLAB®. The 'hold on' and 'hold off' commands were used to chart the flight path taken by the Twin Otter plane on September 23, 2007.

After plotting multiple datasets into a single graph, a file, named "code.m," was created. This file was capable of running from the command line using MATLAB®. A Perl script was then created to run the jobs twenty times automatically while logging the times; from these times an average run time generated.

The next step in the project, dealt with the use of TORQUE and turning the code into a stand-alone binary. It was attempted to run the jobs on TORQUE through Madogo but problems were encountered: there was no shared home directory or shared ssh keys.

After trying to correct this issue, it was decided to utilize the resources available through Indiana University (IU). Access was granted to IU's Quarry cluster. MATLAB® had to be softadded to Quarry, so the created codes could be run.

The Perl script was utilized at this point, enabling a second dataset to be formed. To run the code, a TORQUE script was created. Another Perl script was created to submit the TORQUE jobs twenty times. To do this, the 'qsub' command had to be used. The wall times were then emailed in a list format.

Problems with trying to convert the code into a standalone binary also occurred. At first there was an 'ifconfig' error, which was fixed by adding '/sbin' to the path of all users. Once completing this step accessing the deployment tool on MATLAB® was utilized to create a binary. Though the code should have been capable running, errors were still received. The 'mcc' command was utilized to fix this problem, but the 'mcc' command could not find the libraries with G++. To resolve this issue, Mathworks, the company who created MATLAB® was contacted. MATLAB® worked with the team's code and looked for a solution to resolve the internal problems.

C. Statistical Methods and Tests Used to analyze the Data

In both hypotheses, the question was posed dealing with productivity. To test both theorems, a function in Microsoft Excel, the analysis of variance, also known as ANOVA, was utilized.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
ECSU MATLAB ®	20	84.75	4.2375	0.0445355 26
IU MATLAB ®	20	392.6	19.63	0.2161157 89
Torque	20	223	11.15	36.134210 53

ANOVA

Sthe team'sce of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2377.480 75	2	1188.7403 75	97.986939 49	3.58737E-19	3.15884 2719
Within Groups	691.5023 75	57	12.131620 61			
Total	3068.983 125	59				

ANOVA is a statistical method for making simultaneous comparisons between two or more mean. The function attempts to prove a significant relationship between variables—that there was or was not a change. To use ANOVA, a null-hypothesis is needed. The null hypothesis states:

$$\mathbf{H_o = \mu_1 = \mu_2 = \mu_3 \quad (No\ Change)}$$

$$\mathbf{H_1 \neq \mu_1 \neq \mu_2 \neq \mu_3 \quad (Change)}$$

In the group's study, ANOVA was performed to determine if there was statistically enough variance between the means. This was tested within a 5% significance to note a difference in submission times. If the p-value of the ANOVA table was higher than the level of significance, then the hypotheses would be rejected. However, if the p-value was lower, then the hypotheses could be accepted.

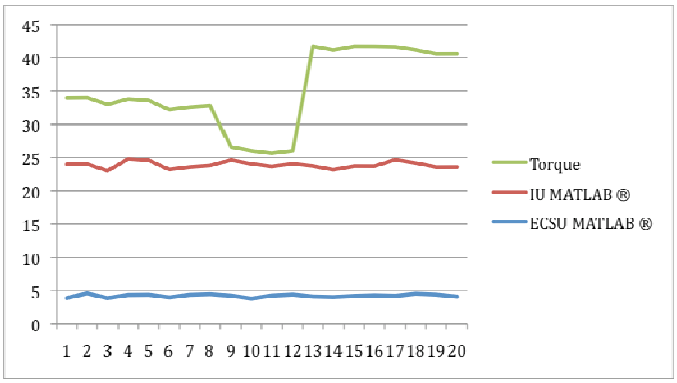
The results of the ANOVA chart were placed in a Tukey or chart. This allowed the direct comparison of all three variances to be compared to speed, making it statistically possible to determine which method for processing the CReSIS data was the most efficient.

IV. Analysis of Data

A. Results of the Statistical Analysis of Data

After conducting the necessary research, it was discovered that TORQUE not only had a very high variance, but it took the longest to process data.

B. Tables, Figures, etc. used for Data Analysis



C. Decision about the Hypothesis

The team hypothesized that the TORQUE submissions will take the shortest amount of time to submit and it was also predicted the submissions from the command line of MATLAB through Quarry would take the longest. Both of these hypotheses were proved wrong in that the TORQUE submissions not only took the longest, but it had a very high variance.

V. Summary, Conclusions, and Recommendations

A. Conclusions Resulting from Statistical Analysis of the Data

The team looked at the variance (the difference between points) that the ANOVA produced. When the graph was examined, the team concluded that TORQUE was the least stable of the three and took the longest to submit its jobs.

With the ANOVA table, it was proved that ECSU was not only the most efficient, but had the least variance with only 0.02. This showed that the team's original hypothesis

was incorrect. It was surprising, because it was thought that TORQUE would be the second most efficient.

Indiana University's MATLAB submissions were the second most efficient and had less variant which was not expect for the reason that it had to be imported to a different location then the results had to be sent back to the team for analysis.

B. Shortcomings

The MATLAB® complier did not recognize the G++ complier for the reason that the MATLAB® complier was not able to locate the libraries; therefore the creation of standalone binaries could not be fulfilled.

The team attempted to use the MATLAB® deployment tool. Though it could create distributions, it could not create the libraries. For this reason, the jobs couldn't be complied nor turned into a stand alone

There were also internal conflicts that occurred when TORQUE was used. The problem was that TORQUE did not have a shared home directory so that all of the users would be able to shared their files along with no shared *ssh* keys. Because of this situation, ECSU's TORQUE could not be used, and there could not be a direct cluster comparison

C. Future Works

As a continuation of the team's research, it would valuable if TORQUE was studied in depth to analyze why there was a great spike in the run times. Eventually, the TORQUE cluster at ECSU, named Madogo, will need to have a common home directory and ssh keys added to it; this will allow the team's script to run on the system.

For other future works, the team would like to see the team's scripts tested using CSARP-Lite, a more complex program. The .m file code used should be complied and converted to a standalone binary. A comparison of Condor job scheduler verses TORQUE should also be compared to see how this affects the run times (there should not be a difference since Condor performs the same function as TORQUE); and to compare running MATLAB® using TORQUE verses MDCE.

REFERENCES

[1] How use ANOVA: unknown, "ANOVA." *ANOVA*. 07Aug 2003. Elon University. 28 Jul 2009 <http://org.elon.edu/econ/sac/anova.htm>.

[2] How to install ANOVA: unknown, "Activating the Anova Data Analysis Function in Excel 2007." *Technology Training at SUNY Cortland*. 27 Nov 2007. Wordpress. 28 Jul 2009 <http://cortlandttc.wordpress.com/2007/11/27/activating-the-anova-data-analysis-function-in-excel-2007/>.

[3] How to submit a compiled Matlab job to Torque: Lepora, N.. "How to submit compiled MATLAB(R) code to a cluster." 9 Feb 2009 4. Web.28 Jul 2009. <http://www.lepora.com/publications/lepora2009.pdf>.

[4] How to submit a job through Quarry: unknown, "At IU, on Quarry, how do I submit a Matlab batch job?." *University Information Technology Services* . 10 July 2009. Indiana University. 28 Jul 2009 <http://kb.iu.edu/data/ayfu.html>.

[5] What is Quarry: unknown, "University Information Technology Service." At IU, what is Quarry?. 30 June 2009. Indiana University. 28 Jul 2009 <http://kb.iu.edu/data/avju.html>.

[6] MATLAB operator for Mac OS X: Lord, Steven. ""unexpected MATLAB operator" when run from command line." [Weblog unknown] 17 Dec 2008. Derkiler. Web.28 Jul 2009. <http://newsgroups.derkeiler.com/Archive/Comp/comp.soft-sys.matlab/2008-12/msg03512.html>.

[7] Ways to use the mcc complier: unknown, "Using the mcc command." MATLAB complier. MathWorks. 28 Jul 2009 <http://www.mathworks.com/access/helpdesk/help/toolbox/compiler/index.html?/access/helpdesk/help/toolbox/compiler/br2jggs-10.html&http://www.google.com/search?hl=en&client=firefox-a&rls=org.mozilla:en-US:official&hs=cax&q=how%20

[8] What is TORQUE and how to use it: unknown, "Managed Services and Consulting." How to use TORQUE & usage. 18 May 2009. University of Colorado at Boulder. 28 Jul 2009 <http://www.colorado.edu/its/managedservices/support/torque.html>.

[9] How to create an object: unknown, "MATLAB programming." Object-Orientated Programming Example. 10 Dec 1997. unknown. 28 Jul 2009 <http://www.engin.umd.umich.edu/CIS/cthe team'sse.des/cis400/matlab/oop.html>.

[10] How to load data into MATLAB: unknown, "MATLAB HyperText Reference." Loading Data into MATLAB. unknown. unknown. 28 Jul 2009 <http://web.cecs.pdx.edu/~gerry/MATLAB/plotting/loadingPlotData.html>.

[11] How to do Matrix Multiplication in MATLAB: unknown, "Matrix Multiplication Example." MATLAB Programming. unknown. unknown. 28 Jul 2009 <http://www.engin.umd.umich.edu/CIS/cthe team'sse.des/cis400/matlab/mamu.html>.

[12] Knight, Andrew. Basics of MATLAB and Beyond. Boca Raton, Florida: Chapman & Hall/CRC Press LLC, 2000.

[13] MATLAB Hello World Program: unknown, "Hello World! Example." MATLAB Programming. unknown. unknown. 28 Jul 2009 <http://www.engin.umd.umich.edu/CIS/cthe team'sse.des/cis400/matlab/hello wor.html>.

[14] Phillips, Fred Young. Thinkwork. West Prot, Conneticut: Greenwood Publishing Group, Inc., 1992.