The Use Of Parallelization Support To Speed Up PlotViz3

Khaliq Satchell, Mentor(s): Dr. Geoffrey Fox, Yang Ruan, Gregor von Laszewski :: Indiana University :: Elizabeth City State University

Abstract

Keywords: Bioinformatics, Genomic Sequences, PlotViz3, Phylogenetic Trees, Multithreading, Parallelization, C++

In biology there is a scientific field that develops methods and software tools for organizing and analyzing biological data. That field is bioinformatics and it combines computer science with other fields in order to study biological data and processes which in turn can provide meaningful information on genomic sequences. Currently, there is a software called PlotViz3, a three-dimensional data point browser, which can be helpful for scientists in the field of bioinformatics.

PlotViz3 can be used to interactively discover intrinsic structures efficiently of which are high-dimensional and contain large volumes of data. This means that scientists will be able to find the correlations between the DNA sequence clusters that they have data for more effectively than their previous methods such as phylogenetic trees. This software should be accessible to every scientist working in bioinformatics but has yet to be put out there for them because the process is not easily done. Once it is basic enough for simple execution then scientist will have a new and more efficient tool for analyzing organism's genomic sequences.

The purpose of this project is to add parallelization support to the code for multithreading PlotViz3. The code in the software uses the C++ programming language which is what we shall be using to make improvements. In the end, adding this support will speed up the virtualization process in the software and make it less time consuming when looking for results quickly and efficiently.

Design

Adding in parallelization support for multithreading within PlotViz3 using multi-core processing will undoubtedly assist in speeding the process the software goes through when rendering data. Multi-core processing is essentially multiple processes running all at once rather than one process at a time which is more time consuming when data is needed quickly. In order to find out how to add the parallelization support into the code, the environment had to be built onto the system that is being used. Through an Integrated Development Environment (IDE) the code has to be thoroughly read through and understood. Only then can the code be altered and updated. Then a testing a phase is required to make sure that any improvement to the speed of the software has been made.

Methods

Building the environment for PlotViz3's code was the first step in the research process. The steps to build the environment in the Windows 7 Ultimate 64 Bit operating system are as follows:

1. Download and install Microsoft Visual Studio 2010, CMAKE, and NSIS.
2. Download and extract pviz3 and pviz3dev files to the root directory.
3. Add a new folder in pviz3dev called "workspace".
4. Open the Control Panel, go to System Security and click on System.
5. In System, click on Advanced System Settings.
6. In the Advanced tab click on Environment Variables.
7. Click New located under User Variables for (Username).
8. Add "PVIZ3DEV_HOME" for variable name and "C:\pviz3dev" for variable value and click OK.
9. Add "PVIZ3_SRC" for variable name and "C:\pviz3" for variable value and click OK.
10. Type in "PVIZ3DEV_WORKSPACE" for variable name and "C:\pviz3dev\workspace" for variable value and click OK.
11. Click OK to exit out of Environment Variables and exit out of Advanced System Settings.
12. Logout of computer and log back in to set changes.
13. Open Visual Studio x64 Win64 Command Prompt (2010) by opening the start menu and going to All Programs.
14. In All Programs, click on Microsoft Visual Studio 2010.
15. In Microsoft Visual Studio 2010, click on Visual Studio Tools and then click on Visual Studio x64 Win64 Command Prompt (2010).
16. In the command prompt, type "cd %PVIZ3_SRC%\build" and press enter.
17. Type in "%PVIZ3_SRC%\config.bat" and press enter.
18. After the operation is complete, type "nmake" and press enter.
19. A solution file called PVIZ3(PVIZ3.sln) will be created for Visual Studio 2010 in this location: C:\pviz3\build-vs.
20. Open the solution file and build PACKAGE and INSTALL.
21. After they both succeed you are finished.

Once those steps were completed, we had to start reading the code to get a general idea of what was used and what each piece of it was doing. Afterwards, we had to find out what was needed to be done in order to multithread PlotViz3. Research showed that three C++ header files, an object file library, and an application extension should be added to Microsoft Visual Studio 2010. Those files are a necessity since they allow us to add parallelization support to the code. Once the files were added it was just a matter of looking over the code and finding what were the best possible parts to change.

PlotViz3 and It's Code

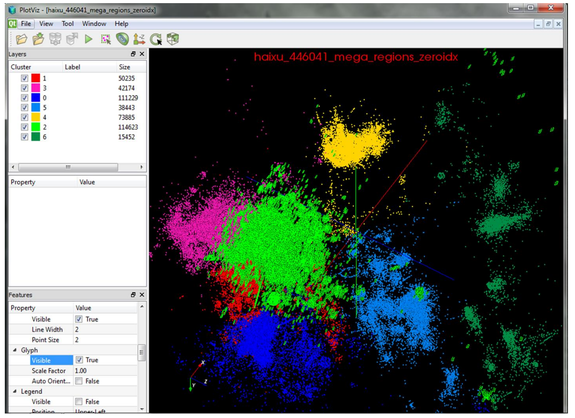

Fig 1: The visualization of genomic sequences in PlotViz3

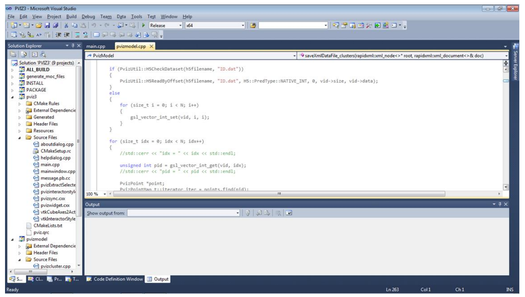

Fig 2: A portion of PlotViz3's code

Conclusion

The research was only half completed but has helped in making sure that work can be continued off of it. We were able to find the necessary components which can help in applying parallelization support to PlotViz3. With these components in place, future researchers tasked to this project will be able to only focus on finding out which pieces of the code should be changed and testing to see if it works.

Acknowledgments

Special thanks to:
Dr. Linda Hayden; principal investigator of CERSER
Dr. Lamara Warren; director of IU-SROC
Dr. Geoffrey Fox; mentor
The National Science Foundation (NSF)

References

1. CHOI, J. Y., RUAN, Y., BAE, S.-H., QIU, J. and FOX, G. 2010. Introduction. PlotViz – A tool for visualizing large and high dimensional data. http://salsahpc.Indiana.edu/pviz3/#intro.

2. CHOI, J. Y., RUAN, Y., BAE, S.-H., QIU, J. and FOX, G. 2010. Instructions for Developers. PlotViz – A tool for visualizing large and high dimensional data. http://salsahpc.Indiana.edu/pviz3/#intro.

3. MOHTASHIM, M. 2014. C++ Multithreading. Tutorialspoint. http://www.tutorialspoint.com/cplusplus/cpp_multithreading.htm.

4. RUAN, Y., HOUSE, G. L., EKANAYAKE, S., SCHUTTE, U., BEVER, J. D., TANG, H. and FOX, G. 2014. Integration of Clustering and Multidimensional Scaling to Determine Phylogenetic Trees as Spherical Phylograms Visualized in 3 Dimensions. http://grids.ucs.indiana.edu/ptliupages/publications/PhylogeneticTreeDisplayWithClustering.pdf.

Primary Contacts

Dr. Geoffrey Charles Fox, Indiana University, gcfexchange@gmail.com

Gregor von Laszewski, Indiana University, laszewski@gmail.com