

Paul Pearce  
Statnews Undergraduate Researcher

My contribution to the StatNews Statistically News Analysis project involved working with Professor El Ghaoui primary in three areas: 1) Parallelization of the word imager across multiple threads and eventually multiple machines. 2) The building, configuration, and maintaining of a small cluster of machines for use on the project. This also involved setting up network access for the machines. 3) Design, creation, and population of a SQL database to house the current and future news articles and headlines associated with the word imager, based on the existing flat text files.

For the first task, I was given a functional piece of python code written by graduate student Brian Gawalt that ran linearly on a single machine in a single thread. My goal was to parallelize as much of this algorithm as possible, in order to spread the load across the future cluster. After profiling the code to identify where exactly the algorithm spent most of its execution time, as well as what portions would benefit from parallelization, I broke the algorithm up into separate components that could be executed concurrently in multiple threads. Initially this concurrent execution was within a single machine, and then eventually across the cluster.

The main area of the algorithm that was parallelized involved work that was proportional to the size of the input data set. One of the main benefits of this form of parallelization is as the data set size increases more machines can be brought online to keep the runtime roughly constant. With enough parallel computing power, the runtime of the algorithm is now a function of the number of words in the English language, instead of the number of words in our data sets.

In the course of working on the parallelization of the word imager I created a simple manager for the Parallel Python library that allows future parallelization of Statnews python applications to our cluster with reduced effort. The manager library abstracts away the details of working with the underlying library and the details of the cluster, in an effort to make parallelization across the cluster more accessible to other team members.

My work involving the cluster consisted on advising Professor El Ghaoui on machines too purchase, the setup of the machines from their boxes, the installing the operating systems, and the configuration and selection of their software. I was also responsible for building the network the cluster ran on, and getting network access to the cluster's router. The network I built consisted of a router and switch I selected specifically for the purposes of our cluster and its LAN.

For the database aspect of my work, I designed and implemented a series of

tables that would store the entire data set used for the word imager. I based my designs on prior proposals from graduate student Yanpei Chen. The goal of the database design was to maximize the speed of read queries, as each thread would be making heavy use of such queries in the course of its operation, and several different designs were implemented and tested. An additional goal was to make the tables as extensible as possible for future use for other tasks. Once I felt the tables were optimized, I populated the databases from the text files the project had been using, and setup the imager to use the database. For optimization purposes, each machine in the cluster had its own database server running a mirrored copy of the data.

Once the cluster was functional I deployed the algorithm, and as a result of the parallelization and database improvements, the runtime was reduced from 4 minutes on a simulated data set to approximately 30 seconds. From here undergraduate Hisham Zarka began to take over where I left off, with the migrating the imager to the Statnews.org website.