The Wotan Beowulf Cluster Project

ABSTRACT

With the emergence of Big Data, the use of cluster computing to analyze large quantities of data has become an important area of commercial and scientific interest. To investigate cluster computing, we, a group of high school students, constructed a Beowulf cluster last year in liaison with a university. The Beowulf cluster architecture was used because this used low-cost, consumer-range computers. The cluster was built by connecting obsolete computers to a network and using a program to distribute tasks across this network. It pursued the purpose of contributing to scientific research and to create a local computing resource. This year, we continued our research by installing a new framework, Hadoop, to the cluster. This framework is currently dominant in Big Data analysis, preferred for its novel approach to cluster computing. Hadoop creates a server to deploy Java programs. The data that is handled is mapped and split apart (called the "MapReduce" model) into chunks, so the data is more manageable for individual nodes in the cluster. We met numerous challenges in installing Hadoop, mainly Java language difficulties and script troubles. We had to find an efficient way to create long, custom scripts on every one of the nodes. Eventually we were able to use a Network File System to standardize the script process. Java programming with Hadoop's libraries was also unfamiliar territory. We found that Hadoop was a very complex framework, focusing on small computations. It split the cluster into individual environments and clients, each requiring individual attention. After the installation of Hadoop, we were able to begin to apply our cluster to various small projects. One program we ran was TeraSort, which benchmarked our cluster by generating random, generic datasets and sorting them. We also wrote our own program to calculate the digits of pi. In the future, we plan to investigate algorithms for gene assembly, in which raw biological data is prepared for research. In preparation, we downloaded a large set of raw strawberry DNA and ran a third-party program to assemble it. Overall, this year we gained great insight on the challenges and processes of current cluster computing practices.