On Construction of Cluster and Grid Computing Platforms for Parallel Bioinformatics Applications

Chao-Tung Yang, Tunghai University, Taiwan

Wen-Chung Shih, Asia University, Taiwan

ABSTRACT

Biology databases are diverse and massive. As a result, researchers must compare each sequence with vast numbers of other sequences. Comparison, whether of structural features or protein sequences, is vital in bioinformatics. These activities require high-speed, high-performance computing power to search through and analyze large amounts of data and industrial-strength databases to perform a range of data-intensive computing functions. Grid computing and Cluster computing meet these requirements. Biological data exist in various web services that help biologists search for and extract useful information. The data formats produced are heterogeneous and powerful tools are needed to handle the complex and difficult task of integrating the data. This paper presents a review of the technologies and an approach to solve this problem using cluster and grid computing technologies. The authors implement an experimental distributed computing application for bioinformatics, consisting of basic high-performance computing environments (Grid and PC Cluster systems), multiple interfaces at user portals that provide useful graphical interfaces to enable biologists to benefit directly from the use of high-performance technology, and a translation tool for converting biology data into XML format.

Keywords: *Bioinformatics, Biology Databases, Cluster Computing, Grid Computing, High-Performance Technology*

1. INTRODUCTION

Bioinformatics is a combination of biology and information technology and includes any computational tools and methods for managing, analyzing and manipulating large sets of biology data. Thus, computing technologies are vital for bioinformatics applications (Konishi et al., 2002; Trelles et al., 1998). For example, biology problems often require repeating the same task millions of times such as when searching for sequence similarities in existing databases or comparing groups of sequences to determine evolutionary relationships. In such cases, the high-performance computers to process this information are indispensable. Biologi-

DOI: 10.4018/jghpc.2011010104

cal information is stored on many computers around the world. The easiest way to access this information is to join these computers together through networking. Such activities require high-performance computing infrastructures (Prodan & Fahringer, 2002) with access to huge databases of information.

The major advances in computer technology and computer science over the past 30 years have dramatically changed much of our society. Currently, many parallel versions of bioinformatics applications can be used to conduct computing tasks on Linux PC Cluster or Grid systems, including, HMMer (http://hmmer.wustl.edu/), FASTA (ftp://ftp.virginia.edu/pub/fasta/), mpiBLAST (http://mpiblast.lanl.gov/index.html), ClustalW-MPI (Li, 2003), FastDNAml (Stewart et al., 2001), and TREE-PUZZLE (Schmidt et al., 2992). Using these parallel versions of bioinformatics software for sequence alignment or analysis can always save enormous amounts of time and cost. The use of parallel software versions and cluster system is cost-effective and it will become more and more popular in the near feature.

Computing technologies today represent promising future possibilities. Currently, it is still very difficult for researchers who are not specialized in Information Technology (IT) to fully utilize these high-performance computing technologies. IT engineers are therefore playing an important role in improving the research environment. The mission imposed on us is to provide user-friendly interfaces for researchers who are not specialists in IT to be able to benefit directly from the use of high-performance technology. The user portal enables interactions between application users and applications obtaining parametric inputs for problems and reporting results upon execution completion (Pierce, 2002; Stocker, 2004; Sturn, 2003; Suzumura, 2004).

Large quantities of biological data have been made accessible to the scientific community through numerous genome websites such as the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) and the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/index.html), and many of these genome websites distribute large datasets as flat files (e.g., tab-delimited files). Flat files are text files lacking any form of markup language. The tasks of automating the processes of information retrieval and integration of heterogeneous biological data are difficult due to their unstructured formats. The data involved may range from nucleic acid and protein sequences, to three-dimensional protein structures, and relationships among various metabolic pathways. Furthermore, different approaches are used for data modeling, storage, analysis, and querying purposes. Therefore, Molecular Biology databases have only a few widely accepted schemas.

As a consequence, integration and interoperability of Molecular Biology databases are issues of considerable importance. The eXtensible Markup Language (XML) (http://www.w3.org/XML/) proposed by the World Wide Web Consortium (W3C) (http://www.w3.org) has emerged as a popular format for representing and exchanging information over the Web. XML was originally designed to overcome the limitations of HTML and flat files. In this paper, we present an approach to converting data from various databanks into XML format for storage in XML database management systems. Our system uses an open-source project called BioJava (http://www.biojava.org/) to translate biological data into XML format, simplifying biological data translation. The details are described below.

In the present study, THUBioGrid, an experimental distributed computing application for bioinformatics (BioGrid) is proposed (Yang, Hsiung, & Kan, 2005a, 2005b; Yang, Kuo, & Lai, 2004, 2005; Yang, Kuo, Li, & Gaudiot, 2004). THUBioGrid incorporates directory services (data and software), grid computing methods (security, authentication, data transport and remote jobs), and gene sequence/genomic data processing methods. It uses Java CoG Kit plus bioinformatics Java packages to perform various computational tasks. The performance of THUBioGrid has been tested by executing the FASTA and mpiBLAST programs for pro-