GRID PORTAL FOR BIOINFORMATICS SEQUENCES ALIGNMENT APPLICATIONS

AZLAN ARIFIN

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2008

GRID PORTAL FOR BIOINFORMATICS SEQUENCES ALIGNMENT APPLICATIONS

A thesis submitted to

the Faculty of Computer Science & Information Technology.

University of Malaya

in Partial Fulfillment of the Requirements for

the Degree of Master of Information Technology.

By

AZLAN ARIFIN

(WGD040006)

JUNE, 2008

Supervisor: Mr. Liew Chee Sun

ABSTRACT

The aim of present study is to develop a grid portal for bioinformatic's sequences alignment applications. Further enhancement has been done to the campus grid called GeRaNIUM. A web-based interface called GeRaNIUM Grid Portal (GGP) was successfully developed and implemented on GeRaNIUM to provide a simple user friendly interface for grid users with varying IT experience. This Portal allows grid users to submit jobs in a secure, reliable and scalable manner. However, most of bioinformatic's users are hesitant to run a parallel job in a grid environment. In order to convince users to use parallel computing in grid environment, a comparison study has been done to compare the performance of process runtime among workstation and cluster computing in the GeRaNIUM grid environment. The main comparison was based on process runtime and consistency of results produced. The produced results in this study have shown that output consistency is achieved while the computing speed is increased in the grid environment.

Findings of this study indicated that parallel computing could speed up the runtime of Bioinformatics applications by parallelizing the sequences alignment process into collective resources. Besides, parallel computing not only accelerated the process but also produced reliable outputs which might convince users to use parallel computing. The runtime on different problem sizes showed that parallel computing was more effective on running problems that had shorter length of sequences rather than processing a longer length of sequences. The main contribution in this project was the development of Grid portal which would make it easier for a GeRaNIUM user to exploit grid applications anytime and anywhere.

ACKNOWLEDGEMENT

Thanks God, finally I am here, writing the acknowledgement.

I never thought how much it could take in preparing this piece of work. By the way, I always thought that things never come easy to me. Anyway, with this opportunity, I would like to extend my thanks to my supervisor of whom this piece of work would never really look like a piece of work at all without his dedication and belief. I am most indebted to Mr Liew Chee Sun who has really taken great efforts to ensure that this work is true and accurate. I would like to thank Assoc. Prof. Dr. Amir Feisal Merican for his suggestion and advice about proposing this field of research to me and his effort to provide facilities on doing this work. I would also like to thank with all staff at Bioinformatics and Bio-Computing Division, Institute of Biological Sciences, University of Malaya for their support and help during the time I used facilities in the respective departments.

I am most grateful to have very understanding parents, brothers, especially my mother whom I regard as my mentor in everything. Lastly, I would never have done all this without the support of my beautiful and lovely wife, Aini Suraya Ahmad Ghazali.

Kuala Lumpur, Friday 16 June 2008

Azlan Bin Arifin

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATION

MPICH-G2	Message Passing Interface Framework- Globus 2
MPI	Message Passing Interface
HGP	Human Genome Project
SCE	Scalable Cluster Environment
GUI	Graphical User Interface
VRML	Virtual Reality Modeling Language
BLAST	Basic Local Alignment Search Tool
TMHMM	Transmembrane Helics Markov Model
EMBOSS	European Molecular Biology Open Software Suite
GeRaNIUM	Grid-Enabled Research Network and Info-structure of the University of Malaya
GGP	GeRaNIUM Grid Portal
eth0	Ethernet 0
eth1	Ethernet 1
MyREN	Malaysia Research and Education Network
DNS	Domain Name Server
PBE	pre-boot execution
DHCP	Dynamic Host Configuration Protocol
IP	Internet Protocol
OS	Operating System
TFTP	Trivial File Transfer Protocol

CA	Certificate Authority
GRAM	Grid Resource Allocation and Management
GPT	Grid Packaging Tool
GSI	Grid Security Infrastucture
HPC	High Performance Computing
CoG	Comodity Grid
JDK	Java Development Kit
FTP	File Transfer Protocol
GASS	Global Access Secondary Storage
IT	Information Technology
NCBI	National Centre for Biotechnology Information
VPN	Virtual Private Network
MDS	Monitoring and Discovery System
CPU	Central Processing Unit

CHAPTER 1

INTRODUCTION

1.1 Introduction to Workstation, Cluster and Grid Computing.

The increasing interest in high performance computing has heightened the need of computational resources to solve large scale computational problems. The bizarre technological improvements over the past few years in areas such as microprocessors, memory, networks, and software, have made it possible to assemble groups of economical personal computers and/or workstations into a cost effective system with high processing power. Several studies have observed that parallel applications were successful in solving computational problems which were too large to be solved with previous workstations (Bodlaender 1994; Cheetham et al. 2003; Hsun-Chang et al. 2005).

Unlike past workstations, cluster systems can be used as a multi-purpose computing platform to run high-performance computing applications. According to Foster (1999) “Clusters are groups of computers that are relatively close proximity and that are managed as a tightly coupled unit on dedicated network”. Thus, cluster computing is a potential way of doing parallel computing which provides a high performance platform for a parallel and distributed application. The development of cluster computing environment has offered tools that are currently available in workstations and also has increased the ratio power/price of commodity hardware (Foster et al., 1999). Nowadays, a new technology called grid computing has been developed. Grid computing was developed by enabling linked and coordinated use of geographically distributed resources or clusters for purposes such as large-scale computation and distribution of data analysis. Grid computing technology is not a new initiative. The concept of using multiple distributed resources to work cooperatively on a single application has been around for several decades. Together with the development of Globus (Foster et. all, 1997) and MPICH-G2 (Karonis et. all, 2003), re-structuring and executing of these parallel applications have already been developed for cluster platforms. Grid environment enables organizations to share computing power and information resources across departmental and organizational boundaries in a secure, reliable and highly efficient manner.

1.2 Bioinformatics Sequence Alignment Applications

There are a lot of bioinformatic softwares or tools available for sequence alignment studies for example, ClustalW, T-Coffee, HMMER, Geneious and others (List of Sequence alignment software-Wikipedia, 2007). Those sequence alignment tools are able to identify alignments within multiple DNA or protein sequences. In addition, the programs were developed to be a fully automatic program for global multiple alignment of DNA and protein sequences. The alignment process is progressive and is able to consider the sequence redundancy. Consequent to the sequence alignment process is the development of phylogeny trees which gives a picture of evolutionary relationships among organisms. Currently, software developers are trying to transform sequence alignment tools from single processing to parallel processing since the alignment process seems to have the potential to be parallelized. A new version of ClustalW called ClustalW-MPI enables ClustalW to be implemented in a parallel and distributed systems environment. ClusalW-MPI uses a message-passing library called MPI (Message Passing Interface) and runs on distributed workstation clusters as well as on traditional parallel computers.

1.3 Motivation

The variety and complexity of data generated in biological fields make computational processes on biological data to become computationally intensive. Often the calculations require high performance computing to undertake the task in reasonable time limits. As a solution, parallel computing is the best method to overcome this problem because parallel computing can provide a high performance computing power in a scalable and affordable manner.

Grid computing has become popular in the last decade. This is mainly due to the demand in the use of distributed and parallel computing resources as a metacomputer. Grid provides a computing power available for use by anyone, anywhere. Grid enables organizations to share computing and information resources across departmental and organizational boundaries in a secure and highly efficient manner. Users could run their computational jobs in a collective machine attached to the grid environment. University of Malaya (UM) has established a campus-wide grid environment called GeRaNIUM**.** **GeRaNIUM**, an acronym for **G**rid-**E**nabled **R**ese**a**rch **N**etwork and **I**nfo-structure of **U**niversity **M**alaya which has some initial resources located at different locations around the campus (Por et. al, 2006). However, GeRaNIUM needs an easier administration system or architecture to enable users to collaborate easily and gain access to more computing power. For this purpose, GeRaNIUM needs to be enhanced with a more systematic architecture.

Most of grid computing systems need users to be familiar with UNIX command line. Researchers who are not familiar with the UNIX command line will need a user friendly system and interface to make it easier to exploit grid computing technology. Therefore, web-based grid portal need to be implemented on GeRaNIUM to provide a uniform working environment. Portal can be accessed through web browser with any operating system at user's local machine. Grid Portal also enable users with any level of IT experience, to use grid services with ease. The use of grid portal allows users to have a centrally hosted and hence centrally administered, user interface.

A study has been done on the performance evaluation of ClustalW-MPI in distributed cluster and grid computing (Hsun-Chang et al., 2005). However, it seems that the consistency of output produced has been overlooked. Users are still hesitating to use a parallel computing system regardless of the fact that the output produced is the same as in a single processing. We need to look at the consistency of output produced to convince researchers to switch to parallel applications. There is also the question from users about how big is the impact given by parallel computing to accelerate computational tasks compared to the single processing used previously.

1.4 Objectives

The aim of the study is to develop grid portal for campus grid environment which runs a bioinformatics sequence alignment application. This study also compares the runtime performance between a single processing in workstation and parallel computing in cluster computing. The process runtime and output produced by the application will be examined to evaluate the reliability of parallel computing products and performances. The objectives of the study can be summarized as follows:

- To provide a user-friendly interface that allows users using bioinformatics sequence alignment application on grid.
- To extend the campus grid architecture by providing a server to manage user's cert, thus making grid environment more centralized and manageable.
- To conduct a research and study of single processing and parallel computing performance using bioinformatics sequence alignment application.
- To examine output produced by bioinformatics sequence alignment application running in single processing and parallel processing.

1.5 Scope

In conjunction with the objectives of the thesis, the scope of the thesis is defined in order to provide a basic guideline that enables the study to be conducted within a certain range and depth. The following statements summarize the scope of the thesis in accordance with the stated objectives.

- Review related works done on the development of grid portal.
- Review related works done on the comparison of single processing and parallel computing.
- Extend campus grid environment by placing a Portal Server, Combi Cluster and Bigjam workstation to run a bioinformatic sequence alignment application in single and parallel processing.
- Setup Portal server for user's cert manager and for grid portal platform.
- Develop a web-based grid portal to manage campus grid resources and users' activities.
- Test the grid portal using bioinformatic sequence alignment application.
- Study on bioinformatic sequence alignment tools and the implementation on single machine, cluster computer and grid environment.
- Run sequence alignment process and collect output and process runtime taken by each task.
- Preparing methods to check the consistency of output produced by single processing and parallel processing.

1.6 Thesis Organization

This report contains a total of 5 chapters. The organizations of these chapters are as follows:

- Chapter 1

This chapter is the introduction of the project that briefs on workstation, cluster and grid computing technology. It also introduces bioinformatics sequence alignment applications the motivation of doing this project, the objectives and the scope of the project.

- Chapter 2

This chapter is the literature review. In this chapter, the concept and several studies about different computing platforms namely workstation, cluster computing and grid computing will be introduced. Then, grid portal on previous studies were reviewed to understand the grid portal concept. Lastly, studies about the application used as a benchmark application to run a sequence alignment application.

- Chapter 3

In this chapter, the development process of GeRaNIUM grid environment will be described. Basically, this chapter will be focusing on the development phase of grid portal. This chapter will then explain on the methodology used in doing a comparison on the runtime performance and how the output was analyzed. Lastly, evaluation and discussion about the methodology and results will be explained.

- Chapter 4

This chapter presents the steps for using grid environment through grid portal. The later section of this chapter will explain the method to establish a cross-ca-trust within grid environment. Next, the job submission workflow through grid portal will be described briefly. The final section presents the implementation of single processing and parallel processing through the grid portal.

- Chapter 5

This chapter addressed the overall research target to compare the runtime performance of computational biology applications through a single processing (workstation) and parallel processing (cluster). A discussion about issues on implementing GeRaNIUM grid environment and grid portal will also be discussed briefly in this chapter.

- Chapter 6

This chapter summarizes the efforts of the study and provides recommendations for possible future work in the related field.

CHAPTER 2

LITERATURE REVIEW

2.1 Studies on Workstation, Cluster and Grid Computing

2.1.1 Workstation for Single Processing.

Workstation is a high-end desktop or desk side microcomputer designed for technical applications (Workstation – Wikipedia, 2007). Workstation usually offers higher performance of memory capacity, processing power and multitasking ability as well. The high capability offered is usually optimized for displaying and manipulating complex data such as 3D mechanical design, engineering simulation results, and mathematical plots. The Pilot Ocean Data System (PODS) at Jet Propulsion Laboratory developed computational workstations to support the analysis of remotely sensed data by oceanographers (Kuykendall et al, 1984). This system stored and process images from satellite data for about 100 megabytes per day. After a certain time they realized that storage capacity and the performance of workstation need to be upgraded which probably required an allocation of more funds.

The first implementation of human genome sequence mapping in the Human Genome Project (HGP) used a single robotic workstation (Brignac, 1997). This project emphasized on establishing high-resolution genetic and physical maps to organize large-scale sequencing process. Thus, they developed a very high-throughput and autonomous robotic workstation to quickly and efficiently complete the sequencing of the 3 billion nucleotide base pairs that make up the human genome. This robotic workstation is capable of operating without any human intervention in a 24-hour-a-day, continuous-run mode which maximize throughput and effectively reduced labor costs associated with sequencing. However, this automatic system can only process about 21, 500 samples a day. Referring to Genebank statistic (Figure 2.1), genome data increases rapidly throughout the year which in 2002 saw about 40 millions sequences of genome data banked (GeneBank Data Statistic, 2007). From the GeneBank statistic it seems that, we need a more powerful system than the autonomous robotic workstation to complete the sequencing process within a reasonable time. Instead of changing to a new system, it is suggested that we need a system that has the ability to either handle growing amounts of work in a graceful manner, or to be readily enlarged. For example, the system has the capability to increase total throughput under an increased load when resources (typically hardware) are added. For that purpose cluster computing is the best candidate to solve this problem.

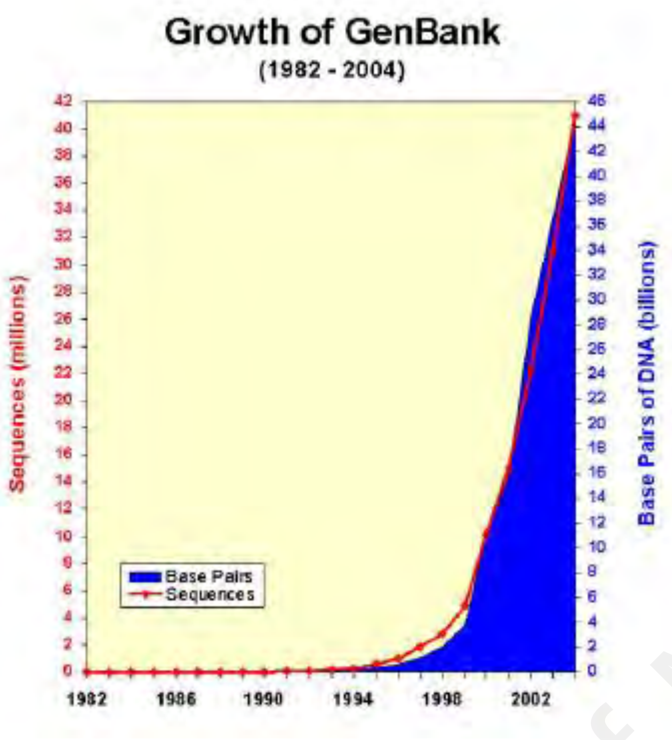


Figure 2.1. *Historical Growth of GenBank Databases Represented in Gene Sequences and DNA Base Pairs (GeneBank Data Statistic, 2007).*

In this thesis, a personal computer with high specification will be used to get a benchmark performance of single processing jobs. The result from this study would suggest that overloading in workstation can be solved by using parallel computing or can be called as cluster computing.

2.1.2 Cluster Computing for Parallel Processing.

Clusters have become the high performance compute (HPC) engine of choice for many industries seeking raw number crunching power with greater flexibility, reliability, scalability and price/performance over traditional workstation or supercomputers. Cluster, which is developed for parallel computing system, has one master node and one or more compute nodes, or cluster nodes (Chao-Tung Yang et al., 2004). Cluster processing performance could be increased by adding more nodes attached to the master node.

Several studies have developed a very large scale cluster systems. For example, the computational plant project (Reisen et. al, 1999) at Sandia National Laboratory, giga-plant system (Halstead et. al, 1999) at Ames Laboratary, and the planned Chiba City System (Evard, 1999) at Argonne National Laboratory. Those studies developed cluster system that consists of nodes starting from 64 to several hundreds. However, as the cluster system size increases from dozens, to hundreds, and even to thousands of processors, management becomes exponentially complex, and can be a daunting challenge for them. Keeping software up to date, monitoring hardware and software status, and even performing routine maintenance requires significant effort. Those issues were addressed by Scalable Cluster Environment (SCE) project (Putchong et al., 2000) and Linux NetworX project (Joshua Harr et. all, 2002). Those projects have developed a software suite that includes tools to install compute node software, manage and monitor compute nodes, and a batch scheduler to address the difficulties in deploying and maintaining clusters. The encouraging development of cluster computing system has convinced several bioinformatics researchers choosing cluster computing to run bioinformatics complex tasks (Gracanin, 2005; Vazquez-Poletti et. al, 2007). From their results, they found that cluster computing has accelerated their computing performance over that provided by a single workstation which is much more cost-effective of comparable speed and availability. However, the study of bioinformatics has led to a huge abundance of computer applications and statistical techniques to manage biological information and to facilitate biological research. Those biological information and Bioinformatics applications are sometimes located in heterogeneous environment and geographically dispersed which need to be integrated so as to expedite a given study. For that purpose, grid computing environment has become a potential solution to this problem.

2.1.3 Grid Computing

The term "Grid" was first used in the mid-1990s to denote a distributed computing infrastructure for advanced science and engineering (Hey, 2002). Grid computing reaches the meaning of grid itself which is used before this in the electricity power grids, providing a computing power available for anyone anywhere to use. In fact, grid computing is also another way to apply parallel computing but in a lager environment than cluster. Grid is a heterogeneous environment which allows collective resources to have a different operating system and hardware, while cluster is a homogenous environment. Grid technology is an opportunity to normalize the access for an integrated exploitation. Grid should be allowed to present software, servers and information systems with homogenous means. In fact, grid is a system that coordinates resources that are not subject to centralized control, using an open standard, general-purpose protocols and interfaces to deliver non-trivial qualities of service (Foster, 2002). The environment of Grid computing allow test computing infrastructure capable of providing shared data and computing resources. The uses of grid allow users to handle the exponentially growing database and to speed up their calculation in data processing by using existing sources. Besides, grid enable users to share inexpensive access to computing power, storage systems, data sources, applications, visualization devices, scientific instruments, sensors and human resources across a distance department and organization in a secure and highly efficient manner. Therefore, researchers will be able to collaborate more easily and will also gain more access to more computing power, enabling more studies to be run and larger problems to be considered.

Data in Bioinformatics field is growing steadily. Researchers from Bioinformatics European Institute have done a research on multiple sequences alignment using ClustalW which is a bioinformatics sequence alignment applications (Thompson et. al, 1994). ClustalW were used to find diagnostic patterns to characterize protein families; to detect or demonstrate homology between new sequences and existing families of sequences; to help predict the secondary and tertiary structures of new sequences; to suggest oligonucleotide primers for PCR; as an essential prelude to molecular evolutionary analysis. However, they found that the rate of appearance of new sequence data is steadily increasing and the development of efficient and accurate automatic methods for multiple alignments is, therefore, of major importance. They need programs that can cope with the large volumes of data to produce an accurate sequence alignment process.

Currently, there are a lot of biological data available in the public domain like Swissprot (Apweiler et. al, 2004), EMBL (Kanz et. al, 2005). The enormous number of biological data makes biological data to be located in different resources across various sites. This issue creates the need for bioinformatics researchers to access the diverse applications and data sources by visiting many web servers which might increase the overall time for the execution of the experiment. GeneGrid which is a UK e-Science industrial project had addressed this issue and had successfully developed a system that integrates numerous bioinformatics programs and various databases (Kelly et al., 2005). GeneGrid provides a platform for bioinformatics researchers to access their collective skills, experiences and results through the creation of 'Virtual Bioinformatics Laboratory'. GeneGrid creates a simple user friendly interface that enables the seamless integration of a myriad of heterogeneous applications and datasets. As a result, researchers can run bioinformatics application such as the multiple sequence alignment with no worry of time and volumes of data located in a distributed location. However, this project was overlooked to be presented as the performance of bioinformatics programs on grid environment which is important for convincing researchers to use grid computing technology.

Several studies have been done on the ClustalW performance at different platforms of computing technology (Kuo-Bin Li, 2002; Hsun-Chang et. al, 2005). Those researches used ClustalW-MPI which is a parallel version of ClustalW to present the efficiency and the performance of ClustalW on single computing and parallel computing environment. In their research they found that the parallelization of ClustalW process using 2 to10 processors has speed up lengthy multiple alignments with relatively inexpensive PC clusters. Nevertheless, in my observation during the literature review, there is no study that proposes the resulting consistency of parallel sequences alignments (ClustalW-MPI) compared to the single sequence alignments (ClustalW). In this thesis the consistency of output produced from different computing platforms is proposed to proof the validity and reliability of parallel computing technology.

In the University of Malaya, **GeRaNIUM**, an acronym of **G**rid-**E**nabled **R**ese**a**rch **N**etwork and **I**nfo-structure of **U**niversity of **M**alaya was proposed as a project to establish a campus-wide computational grid working environment to utilize clusters located at different department such as Combi Cluster at Bioinformatics, Perdana Cluster at Center of Information Technology, FSKTM Cluster at Faculty of Science Computer and Information Technology, Cadcam Cluster at Faculty of Engineering and Biotech cluster which all have applications related to certain fields of study (Por et. al, 2006). An experimental grid test-bed was successfully implemented during a workshop on Grid computing at the University on 23rd to 25th August 2005 to attach the machine between Combi Cluster and Perdana Cluster.

In this thesis, GeRaNIUM grid project will be planned to provide more services in GeRaNIUM that will benefit all researchers and scientists in the campus who are doing their computational experiments. Portal Server was set up to manage resources and also worked as a broker for user's jobs.

2.2 Grid Portal

Most of the researchers are not familiar with command line which is mostly used in grid computing. They were confronted with UNIX command for submitting, altering, deleting and scheduling their jobs running on grid. In order to provide a graphical user interface which is easier for the researcher than typing a UNIX command, a Grid Portal is the ideal solution. Grid Portal web interface provides a uniform working environment on all clusters connected to grid. According to APAC Australian grid subproject, Chemistry Grid Portal was developed which aims to allow user utilizing grid resources without knowing specifications of each computer system environment (Zhongwu Zhou et al., 2005). Therefore, scientists will be able to focus on their research with improved accessibility and productivity. In addition, it provides an easy-to-use user interface for accessing input or output, running various applications jobs on a variety of group computer resources without logging onto those platforms and it also has the ability to transfer data between various resources, Chemistry Grid Portal has allowed user to complete their tasks in the shortest time and in an efficient way. As a result, it has provided a learning curve for scientists to exploit grid technology.

There are some other related works with respect to grid portal. The Australian Biogrid Portal (Buyya et al., 2005) provides the biotechnology sector in Australia a web interface that enables researchers to perform drug-lead exploration on national and international computing Grids. The GENIUS grid portal (Andronico et al., 2003) is a problem solving environment that allows scientists to access, execute and monitor distributed applications that make use of grid resources by only using a conventional web browser. However, those grid portals were developed by large team and each has high programming knowledge and skill. This is impossible for smaller team with lack of programming skill to develop the grid portal.

Curently, there are many portal toolkits available providing a simple way for developers to create grid portal. As an example, the GridPort Toolkit (GridPort). GridPort enables a rapid development of highly functional grid portals that simplify the use of underlying grid services for the end-user (Thomas et. al, 2001). GridPort comprises a set of portlet interfaces and services in the portal layer that provide access to a wide range of backend grid and information services. The services available in portlet are provided by lower-level grid technologies including the Globus Toolkit, the Grid Portal Information Repository (GPIR), and Condor (Foster et al., 1997). Portlets expose the backend services via customizable web interfaces in order to enable personalization of grid portal user interfaces. Portal services support the portlets inside the portal layer by augmenting their capabilities in an extensible and reusable way while tying the portlets together in order to make them more cohesive. GridPort is intended for use by developers of grid-enabled portals, portlets, and applications. Nevertheless, GridPort toolkit is not a Java based toolkit

Another portal toolkit available for grid portal development is Grid Portal Development Kit (GPDK). GPDK provides grid functionality to web sites using JAVA Beans which encapsulate grid functionality (Charles, 2003). By using beans, the GPDK functionality is accessible using the JSP (Java Server Pages) thus, allow user to take a relatively straightforward static web site and quickly add grid functionality. The disadvantage to this approach is that beans must be developed to support each capability.

Several developers involved in grid portal development used GridSphere as portal toolkit to provide a Web portal interface for their grid environment (Lambert et al., 2006; Akram et al., 2005; Zhongwu Zhou et al., 2005). GridSphere is an open-source and widely used tool for portal development (Lambert et al., 2006). GridSphere enables developers to quickly develop and package third-party portlet web applications that can be run and administered within the GridSphere portlet container. GridSphere is used to develop the components that make up the portal, namely the presentation view, presentation logic, and the application logic. The view is implemented with JSP and the logic with portlets that control the presentation flow. The emphasis is on the application logic, developed as a portlet service, which interfaces with the Grid environment.

Because of the widely used of GridSphere as a portal toolkit, GridSphere has been used to develop the GeRaNIUM Grid Portal. The aim in developing the grid portal is to provide a GeRaNIUM grid interface for users to submit jobs, view their results and view resources in GeRaNIUM grid environment as well.

2.3 Studies on Bioinformatics Multiple Sequence Alignment Applications

Multiple sequence alignment of many nucleotides or amino acids is an important application in bioinformatics. The multiple sequence alignment technique identifies diagnostic patterns or motif to characterize protein families. This technique can also detect or demonstrate homology between new sequences and existing families of sequences. Thus this technique helps to predict the secondary and tertiary structures of the new sequence. The prediction process is an essential prelude to molecular evolutionary analysis.

Many multiple sequence alignment tools have been proposed to reduce the high computation time of fully performing alignment of all sequences. Implementations of various multiple sequence alignment heuristics include MSA (Lipman et. al, 1989), PRALINE (Simmossis et. al, 2005), T-Coffee (Notredame et. al, 2000) and DIALIGN P (Schmollinger et. al, 2004). However, ClustalW is the most popular tool for aligning multiple protein or nucleotide sequences (Thompson et al, 1994). The alignment is achieved via three steps: pair-wise alignment, guide-tree generation and progressive alignment. ClustalW-MPI is a distributed and parallel implementation of ClustalW. According to Kuo-Bin Li (2003) suggestion, multiple sequences alignment in ClustalW could be easily parallelized with MPI (a popular message passing programming standard) since most of alignments are time independent on each other. This assumption brings to the development of ClustalW-MPI during Kuo-Bin studies in distributed system and parallel computing. All three steps have been parallelized to reduce the execution time.

The software uses a message-passing library called MPI (Message Passing Interface) and able to run on parallel computing system.

For the conclusion, overall in this thesis ClustalW will be used as single computing tasks and ClustalW-MPI for parallel computing tasks. GeRaNIUM Grid Portal will be developed to run ClustalW and ClustalW-MPI applications. The process runtime and output produced will be recorded to present both computing technology performances and the resulting consistency.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This research was planned to develop a portal for GeRaNIUM grid environment concentrating on bioinformatics sequences alignment application namely ClustalW. Throughout the development process some works has been done on GeRaNIUM grid environment especially in the implementation of Portal Server. In the system testing phase, comparison of single and parallel computing performance was prepared and results consistency was presented.

3.2 Purpose of Research

This research was conducted to provide a grid portal for bioinformatics sequences alignment application. In order to convince researcher to use grid portal and parallel computing technology, output obtained from the job submissions through grid portal were examined and presented. Comparison of single and parallel computing performance was also presented to show that parallel computing is potentially to accelerate researcher's tasks. This research was also conducted to locate Portal Server in GeRaNIUM grid environment. The development of Portal Server is to provide a platform for grid portal and managing user's cert. This is to make GeRaNIUM grid environment more secured and easy to manage.

3.3 Research Procedure

3.3.1 Preparation and Planning

Firstly, GeRaNIUM grid environment was prepared for the grid portal development project. GeRaNIUM is a project establishing a campus-wide computational grid working environment in the University of Malaya (Por et. al, 2006). GeRaNIUM utilizes open source software and applications. GeRaNIUM was officially launched on 14th July 2005. An experimental grid testbed was successfully implemented during a workshop on Grid computing that I had attended from 23rd to 25th August 2005 at the university. In the GeRaNIUM testbed, Perdana Cluster was successfully attached with Combi Cluster (Figure 3.1).

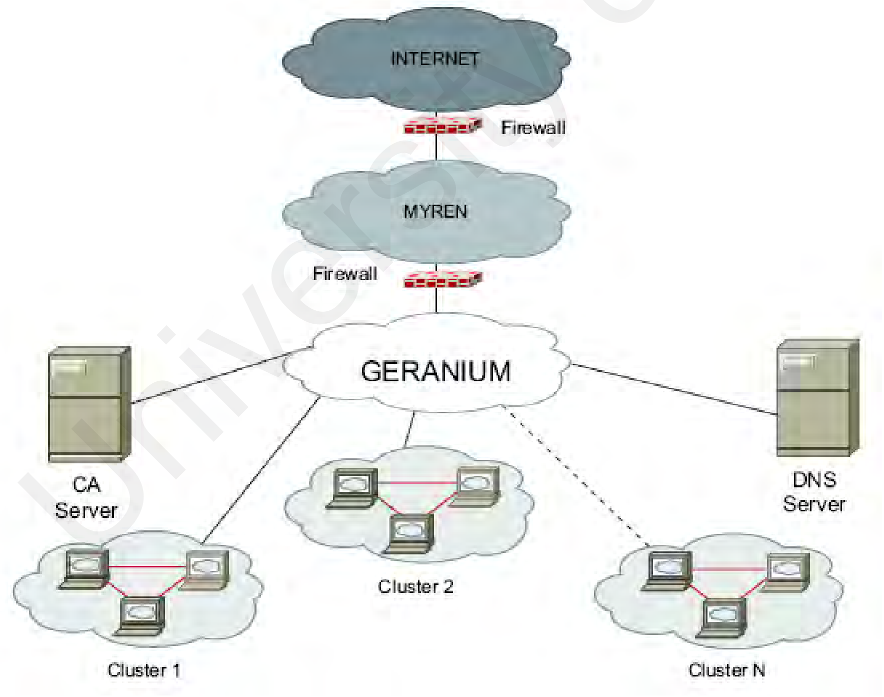


***Figure 3.1** GeRaNIUM current architechture (Retrieved from: Por et. al, 2006)*

However, CA Server might not be fully utilized because it is only stores certificates in GeRaNIUM. In addition, GeRaNIUM doesn't provide user interface that users have to face with command line in order to exploit GeRaNIUM.

Next, GeRaNIUM Grid Portal was developed in Portal Server. GridSphere (Novothy et. al, 2004) has been used as a grid portal toolkit for the GeRaNIUM Grid Portal development. GridSphere has gained wide usage in the Grid community. The UK E-Science Program (UK National E-Science Centre, 2007), D-Grid (D-Grid Initiative, 2007), K*Grid (Korean National Grid, 2007) and many other projects around the world have adopted GridSphere as their Grid portal development platform. According to GeneGrid project, GridSphere has made GeneGrid Portal capable to provide a secure central access point for all users to GeneGrid environment (Sachin et. al, 2005). GeneGrid Portal has also concealed the complexity of interacting with many different grid resources types and applications from the end users' perspective and providing a web-based user friendly interface which users already familiar with. The successful result from GeneGrid that drastically reduced learning curve for the scientists in order to exploit grid technology has influenced this thesis to choose GridSpehere product in the grid portal development. According to Novothy's project (Novothy et. al, 2004), GridSphere portal framework is base on Apache web server, the Jakarta Tomcat Servlet container. Tomcat is a Java based application which has tremendous popularity as a language that provides greater support for the development component based architectures. However, Tomcat not provides a compiler for Java Server Page (JSP) web language and not allows web application reloading in their container. This would bring some problem for

GridSphere to quickly develop and package third-party portlet web application namely GridPortlet. Grid Portlet provides an interface for user to register clusters, view clusters specification, browse file in clusters, retrieve user's credential and monitor job submitted to clusters (Russell et. al, 2006).

Several studies have used ClustalW as an application to compare sequences alignment performances at different computing platform (Kuo-Bin Li, 2003; Hsun-Chang et al., 2005). Those studies have presented that the alignment process runtime on lengthy sequences can be reduced by parallel computing or cluster. However, research on ClustalW results was not presented which is very important to consider at the results consistency when running on different computing technology.

From previous study, sequences alignment between bird, rodent and fish was studied using MrBayes sequence alignment application (Huelsenbeck et. al, 2001). This application was successfully presented the relationship between bird, rodent and fish by producing aligned sequences and phylogeny tree. The aligned sequences were presented to show the relationship between those organisms according to genetic relationship. Phylogeny tree was developed to show evolutionary relationship among those organisms. However, MrBayes application needs a reasonably fast computer that has a lot of memory to ensure the efficiency of alignment process when dealing with lengthy sequences. This problem could be solved if we run sequences alignment onto the parallel application

3.3.2 Research Phase

In this project, those challenges or problems arise in previous study would be resolved. For the GeRaNIUM, CA Server has been replaced with Portal Server that works as certificate repository and also as a platform for grid portal. By doing this, Portal Server would be fully utilized.

In order to make Portal Server as a platform for grid portal development, Tomcat 5.5 (The Apache Tomcat 5.5 Servlet/JSP Container, 2007) which is the latest version of Tomcat has been installed. Compared to previous version of Tomcat, Tomcat 5.5 was chosen because it uses Eclipse JDT Java compiler for compiling JSP pages which is the main web language used in GridSphere. Besides, Tomcat 5.5 allows web application reloading which is suite with GridSphere version 2.1, the latest version of GridSphere. GridSphere-2.1 (GridSphere Portal Framework, 2007) provides a quickly develop and package third-party portlet web applications that can be implemented and administered within the GridSphere portlet container. The latest version, GridPortlet-1.3 has been used and some configuration was done at certain applications to make it appropriate with GeRaNIUM grid environment.

Lastly, ClustalW was used in this study as a benchmark program to present performance of the alignment process in single machine and ClustalW-MPI to present performance in parallel computing. Results obtained from those applications were presented to compare performances of different computing platform and to show the consistency of output as well.

3.3.3 Development Phase

The development phase was carried out with three phases. The first phase is the development of Portal Server in GeRaNIUM. The second phase is the development of GeRaNIUM Grid Portal and the last phase is system testing using Bioinformatics sequences alignment application.

1- Development of Portal Server

In this project, all resources in GeRaNIUM can only be accessed through Portal Server (Figure 3.2). By using this architecture, computing resources would be more secured and jobs would be more efficiently managed by Portal Server. Besides managing users' job, Portal Server also manages resources in GeRaNIUM through a user-friendly web interface. Furthermore, Portal Server which has two network connections was connected to the campus networks as the first network (eth0), and soon will be connected to MyREN (Malaysia Research and Education Network) as the second network (eth1). This machine was also registered to campus Domain Name Server (DNS) as *http://portal.geranium.um.edu.my*. Domain name for all Clusters and resources in GeRaNIUM were also registered to DNS as well but, Portal Server is accessible internally and externally and clusters are only accessible internally. This protocol was planed to make users outside campus only can access GeRaNIUM resources indirectly but through Portal Server.

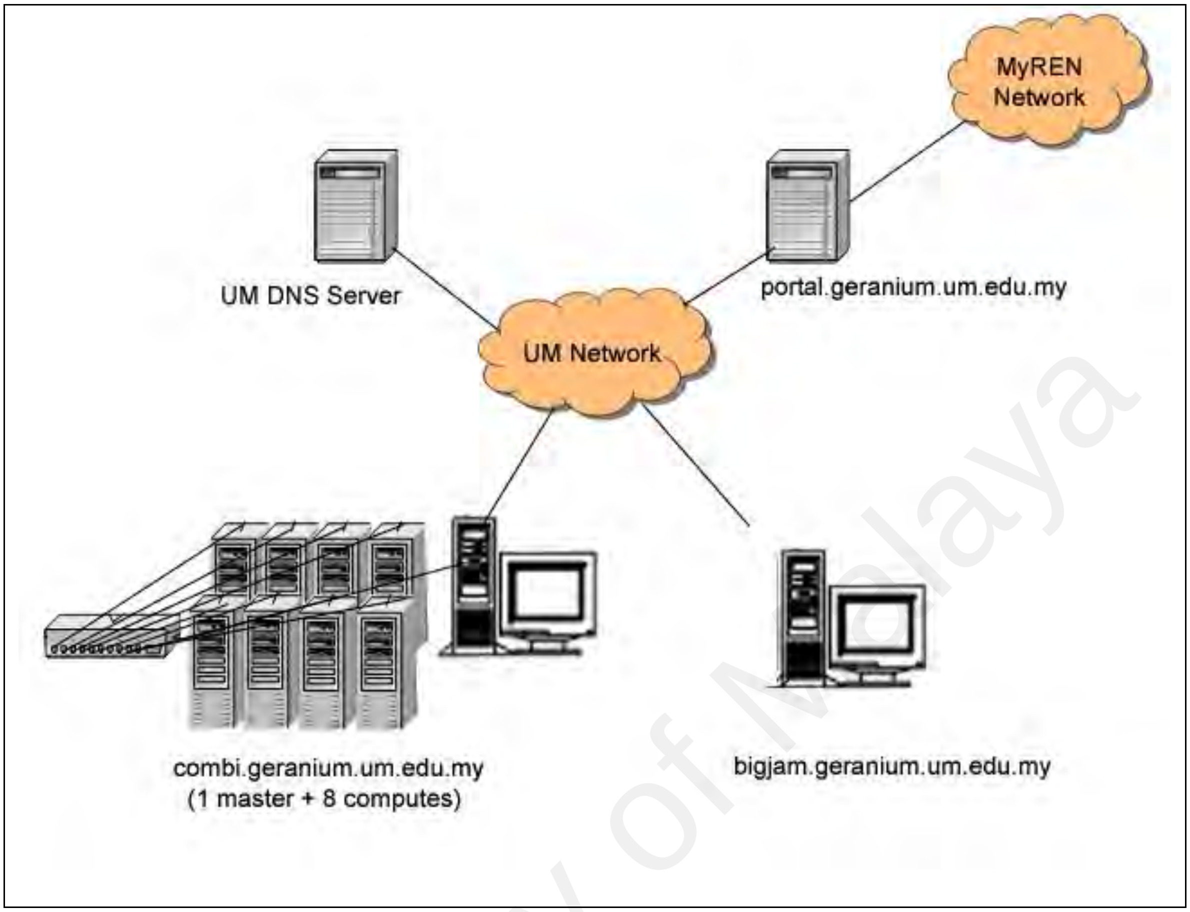


Figure 3.2 *GeRaNIUM's proposed architecture.*

Referring to Figure 3.2, Portal Server, Combi Cluster and Bigjam workstation were constructed using Rocks version 4.1 (Rocks, 2004). Every cluster in GeRaNIUM consists of a front-end node and several compute nodes. As in Combi Cluster which is located at Bioinformatics Department in University of Malaya, has 80 GB disk capacity, 1 GB of memory capacity and 2 physical network ports for master node and each compute node consists of 40 GB of disk capacity, 512 MB of memory capacity and 1 physical network port. Within a cluster, a switch connects 8 compute nodes to the master node. However, Portal Server was differently assembled because this machine just only has a specification like master node which has 2 physical network ports but without compute node. This machine was setup to make it works as a GeRaNIUM's manager, proxy server and web server as well.

Once Rocks was successfully installed on master node, compute nodes were automatically installed by master node using pre-boot execution (PBE) environment. During the installation process, each compute node used DHCP to request for an IP address from the master node. Compute nodes then automatically downloaded the operating system (OS) from master node via TFTP. Installation of compute nodes using PBE is diskless and less time consuming compared to the installation of master node.

Before each cluster can distribute jobs internally (intra-cluster) or externally (inter-cluster) in GeRaNIUM, generation and signing of certificates need to be done. GeRaNIUM utilizes a single Certificate Authority (CA) server which was placed in Portal Server. Portal Server provides trusted CA to every cluster in GeRaNIUM. The rational of having only one CA Server is to enable centralized control and monitoring of certificates signing.

Certificate generation within intra-cluster was done firstly in order to establish a cross-ca-trust. In a cluster, a common user was created to apply a certificate from root. Root is an administrator in Linux operating system. Certificate applied by common user will be signed by root. Once certificate was signed, user needs to have a temporary short-lived credential which allows user to submit an intra-cluster’s job.

After completing an intra-cluster certificate signing, an inter-cluster cross-ca-trust was established between Portal Server and a cluster. As been mentioned previously, Portal

Server works as a manager for clusters certificate. Any cluster in GeRaNIUM environment need to apply certificate from Portal Server. In order to make a cluster trust Portal Server, certificate setup package from Portal was installed in the cluster. Once the certificate signing and exchange process between clusters and Portal server completed, user could submit job from Portal Server to clusters.

2- Grid Portal Development.

In the grid portal development, GridSphere-2.1 has been used as a portlet container and GridPortlet-1.3 to provide services in grid environment. Firstly, GridSphere-2.1 was installed then ready for administrator to login. In the administrator layout, users creation and some customization on portal layout was made such as configuration of users' security level, password and users' layout. Next, GridPortlet-1.3 was installed in gridsphere folder and can be accessed through gridsphere container. User can start or stop gridportlet service under portlet application manager. Once gridportlet started, a Grid tab appears for user to access services available. In the gridportlet, there is a Registry application accessible only by administrator to register resources in grid environment. But before gridportlet can be exploited, users need to apply for credential. Application of credential is to make proxy for users that allow them to act on behalf of grid portal and also to minimize exposure of user's private key. Gridportlet provides an online credential application. Nevertheless, for a secure certificate application; MyProxy need to be installed in Portal Server. MyProxy combines an online credential repository with an online certificate authority to allow users to securely obtain credentials when and where needed (MyProxy, 2007). Once MyProxy was installed,, user can apply new credential and renew their expired credential through the credential application form in gridportlet online.

During the grid portal development phase, some customization at gridsphere's main page was done to add some information about GeRaNIUM's project, resources and applications in GeRaNIUM. The information was put in html based and accessible at the main page in different tabs.

3- System Testing

In the system testing, jobs submission through grid portal was done using bioinformatics sequence alignment tasks. Firstly, sequences were prepared before it can be run in ClustalW. The sequences used in this study focus on the alignment of cytb gene sequences in fish, rodent and bird. The sequences were taken from National Centre for Biotechnology Information (NCBI). The sequences were selected and grouped according to the number of organisms and the length of sequence in base pair (bp) (Table 1). The purpose of doing this is to check the efficiency of parallel processing which is influenced by the sequences length and the number of organisms. Besides, it is also to check the grid portal performances and limitations. All sequences were kept in FASTA format and saved into a text file.

TABLE 1 Groups of data.

Group	**Detail**
1.	~1000bp for each 30 organisms
2.	~1000bp for each 60 organisms
3.	~1000bp for each 90 organisms
4.	~1000bp for each 120 organisms
5.	~2000bp for each 30 organisms
6.	~2000bp for each 60 organisms
7.	~2000bp for each 90 organisms
8.	~2000bp for each 120 organisms
9.	~3000bp for each 30 organisms
10.	~3000bp for each 60 organisms
11.	~3000bp for each 90 organisms
12.	~3000bp for each 120 organisms
13.	~4000bp for each 30 organisms
14.	~4000bp for each 60 organisms
15.	~4000bp for each 90 organisms
16.	~4000bp for each 120 organisms

Method used to compare between single and parallel processing are:

- Firstly, every data was executed through Grid Portal using ClustalW on workstation (single computing).
- Elapsed times taken were recorded and average to 6 executions time.
- Output produced from single processing was downloaded and saved to be used as a standard output.
- Then every data was executed in parallels on cluster with 2 to 9 processors respectively using ClustalW-MPI.
- Runtime taken were recorded and average to 6 executions times
- Outputs produced from parallel processing were compared with the standard output (single processing).
- The comparison done was focused on consistency of sequences alignment produced in parallel processing refers to single processing output.
- Finally, the analysis also compared phylogeny trees produced to check whether the tree built by parallel processing is acceptable or similar with the standard output.

3.3.4 Results Evaluation

Overall, the development of GeRaNIUM Grid Portal is successful even it has some limitations to be considered. At the file upload application in gridportlet, port for gridFTP need to be opened. This problem will not allow users to upload and download files to their resources through grid portal. So, during the test phase, the input and output files need to be copied to the resources using command line. However, once input file located in the selected machine, job still can be run through grid portal. Through Job Submission Portlet, user can select ClustalW application, set job scheduler, and input file location and number of processors to be used using a web based interface. Users can view job details, process runtime, job status and results from the Job Submission Portlet.

From the testing phase, it was found that grid portal can submit job at different computing platform. By using single machine, the process runtime showed at gridportlet was longer than using parallel machine. Process runtime at parallel machine can be accelerated when the number of processors was added in the Job Submission Portlet.

Results produced from single and parallel computing were examined to look at the consistency. Referring to previous study (Huelsenbeck et. al, 2001), first hypothesis of result produced at both computing platform is presented at Figure 3.3. The phylogeny trees produced by both computing platform in this study is merely same with the phylogeny tree produced in previous study.

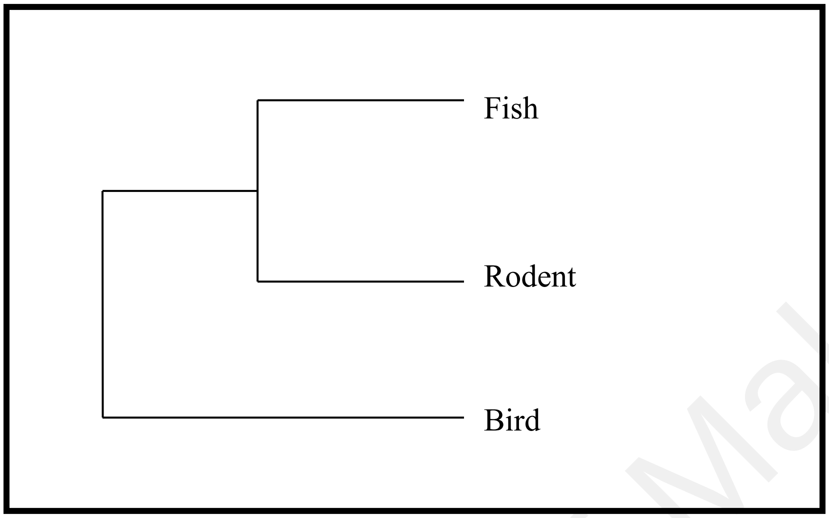


Figure 3.3 First hypothesis of phylogeny tree obtained from sequences alignment.

3.3.5 Discussion

GeRaNIUM Grid Portal is successfully provides an initial interface for GeRaNIUM grid environment. From the grid portal, user can easily exploit GeRaNIUM to submit job especially for Bioinformatics sequences alignment application. Administrator can also ease to manage resources and users' cert in GeRaNIUM at anytime and anywhere with this online system. Results produced by single and parallel computing technology look consistent and reliable and can be used to convince researchers to use GeRaNIUM grid portal.

However, some enhancement and future works need to be done at this initial grid portal. GridFTP application need to be considered in order to make File Manager Portlet in gridportlet can be used for file transportation and management. For the sequences alignment application, performance of longer sequences and more variety of data need to be considered running in a larger grid environment. Besides, load balancing system need to be added in GeRaNIUM in order to balance overwhelming between grid resources to handle larger data and complex tasks. Lastly, it is suggested that GeRaNIUM Grid Portal need to be tested using other applications from other fields.

CHAPTER 4

SYSTEM IMPLEMENTATION

4.1 Establishing Cross-CA Trust

In this section, the process of certificate generation and exchange will be described within intra-cluster and then within inter-cluster.

4.1.1 Intra-cluster certificate signing

Firstly, common user namely *geranium-test* was created in this thesis as a user for Portal Server and every cluster in GeRaNIUM as well. Certificate was requested for *geranium-test* from *root or* administrator as in the following command:

```
[geranium-test@combi ~]# grid-cert-request
```

The command created a directory named ".globus" in a user's directory which contained usercert.pem, usercert_request.pem and userkey.pem files. After that, user's certificate was signed by *root* using this command:

```
[root@combi ~]# local-ca-sign
```

The command signed the usercert_request file which was stored in user's directory and simultaneously added line to grid-mapfile (located at /etc/grid-security/ directory). As an example, once *root* in Combi signed *geranium-test*'s certificate, line in grid-mapfile was automatically added as can be seen in the following:

```
[root@combi ~]# local-ca-sign

# Modifying /etc/grid-security/grid-mapfile ...
New entry:

"/O=Grid/OU=University                                        of
Malaya/OU=combi.geranium.um.edu.my/OU=geranium.um.edu.my/CN=Grid
Test" geranium-test

 (1) entry added

Modifying /etc/grid-security/grid-mapfile ...
```

Then to have a temporary short-lived credential, a command (`grid-proxy-init`) was issued by *geranium-test*, thus allowing *geranium-test* to submit an intra-cluster job. An intra-cluster job submission was proved successful when "GRAM Authentication Successful" message appeared after "`globusrun -a -r localhost`" command was invoked by *geranium-test* (as in Figure 4.1).

```
[geranium-test@combi ~]$ grid-proxy-init
Your          identity:         /O=Grid/OU=University         of
Malaya/OU=combi.geranium.um.edu.my/OU=geranium.um.edu.my/CN=Grid
Test
Enter GRID pass phrase for this identity:********
Creating                                                   proxy
............................................... Done
Your proxy is valid until: Tue Mar 27 01:19:16 2007


[geranium-test@combi ~]$ globusrun -a -r localhost

GRAM Authentication test successful
```

Figure 4.1 *`grid-proxy-init` commands to retrieve proxy certificate*

4.1.2 Inter-cluster certificate signing and exchange.

After *geranium-test* was successful in running job internally, job was also tested to run an inter-cluster job submission. In order to do this, a cross-ca trust was established between Portal Server and clusters in GeRaNIUM. For example to make Combi Cluster trust Portal Server's CA, a CA setup package from Portal was copied to Combi. After that, as common user namely *globus* in cluster, `gpt-build` and `gpt-install` command were invoked in order to install the copied package (as in Figure 4.2).Then as a *root*, `setup-gsi` command was invoked in order to complete the installation of Portal CA setup package, thus made clusters trust Portal Server's CA (as in Figure 4.3).

```
 [globus@combi~]              $GLOBUS_LOCATION/sbin/gpt-build
          /tmp/globus_simple_ca_b44f6af3_setup-018.tar.gz

gpt-build ====> CHECKING BUILD DEPENDENCIES FOR
globus_simple_ca_ b44f6af3_setup
gpt-build ====> Changing to
/home/globus/BUILD/globus_simple_ca_ b44f6af3_setup-0.18/
gpt-build ====> BUILDING globus_simple_ca_ b44f6af3_setup
gpt-build ====> Changing to /home/globus/BUILD
gpt-build ====> REMOVING empty package globus_simple_ca_
b44f6af3_setup-noflavor-data
gpt-build ====> REMOVING empty package globus_simple_ca_
b44f6af3_setup-noflavor-dev
gpt-build ====> REMOVING empty package globus_simple_ca_
b44f6af3_setup-noflavor-doc
gpt-build ====> REMOVING empty package globus_simple_ca_
b44f6af3_setup-noflavor-pgm_static
gpt-build ====> REMOVING empty package globus_simple_ca_
b44f6af3_setup-noflavor-rtl
[globus@combi globus]$ $GLOBUS_LOCATION/sbin/gpt-
postinstall
running /opt/globus/setup/./setup-ssl-utils. b44f6af3..[
Changing to /opt/globus/setup/globus/. ]
setup-ssl-utils: Configuring ssl-utils package
Running setup-ssl-utils-sh-scripts...

***********************************************************

Note: To complete setup of the GSI software you need to run
the
following script as root to configure your security
configuration
directory:

/opt/globus/setup/globus_simple_ca_b44f6af3_setup/setup-gsi

For further information on using the setup-gsi script, use
the -help
option. The -default option sets this security
configuration to be
the default, and -nonroot can be used on systems where root
access is
not available.
***********************************************************
setup-ssl-utils: Complete
```

Figure 4.2 *Progress during installation of Portal's CA on Combi Cluster by user globus.*

```
[root@combi ~]$ $GLOBUS_LOCATION/setup/globus_simple_ca_
b44f6af3_setup/setup-gsi
setup-gsi: Configuring GSI security
Installing /etc/grid-security/certificates//grid
security.conf. b44f6af3...
Running grid-security-config...
Installing Globus CA certificate into trusted CA
certificate directory...
Installing Globus CA signing policy into trusted CA
certificate directory...

WARNING: Can't match the previously installed GSI
configuration files to a CA certificate. For the
configuration files ending in "00000000" located in
/etc/grid-security/certificates/, change the "00000000"
extension to the hash of the correct CA certificate.

setup-gsi: Complete
```

Figure 4.3 *Installation of Portal's GSI package in Combi.*

The next step is to ensure that *geranium-test* from Portal can submit job to any clusters attached with GeRaNIUM. As a *root* in Portal, all certificates and *certificate signing policy* files in certificates directory (located at /etc/grid-security/certificates) in each clusters were copied to certificate directory in Portal and the same thing was done to other clusters as well (as in Figure 4.4). Then the line in grid-mapfile for *geranium-test* at the Portal was copied and added to cluster's grid-mapfile and vice versa. After that IP address and hostname of the Portal were added to each hosts file at clusters (located at /etc/) and vice versa. Lastly, geranium-*test* from Portal can be proved successful to submit job to other cluster by using `globusrun` command pointed to the selected cluster. The "Gram Authentication Successful" message appeared after invoking globusrun command (Figure 4.5).

```
[root@portal certificates]# scp combi:/etc/grid-
security/certificates/\*.0 .
root@combi's password:********
b9495a68.0                 100% 1436     1.4KB/s   00:00

[root@portal certificates]# scp combi:/etc/grid-
security/certificates/\*.signing_policy .
root@combi's password:********
b9495a68.signing_policy    100% 2114     2.1KB/s   00:00

[root@portal certificates]# scp portal:/etc/grid-
security/certificates/\*.0 combi:/etc/grid-
security/certificates/
root@portal's password:********
b44f6af3.0                 100% 1436     1.4KB/s   00:00

[root@portal certificates]# scp portal:/etc/grid-
security/certificates/\*.signing_policy combi:/etc/grid-
security/certificates/

root@portal's password:********
b44f6af3.signing_policy    100% 2114     2.1KB/s   00:00
```

Figure 4.4 *Copying certificate files in Portal and Combi.*

```
 [geranium-test@portal ~]$ globusrun -a -r localhost

GRAM Authentication test successful

[geranium-test@portal ~]$ globusrun -a -r
combi.geranium.um.edu.my

GRAM Authentication test successful
```

Figure 4.5 *Message to show job was successfully submitted.*

4.2 Establishing of GeRaNIUM Grid Portal.

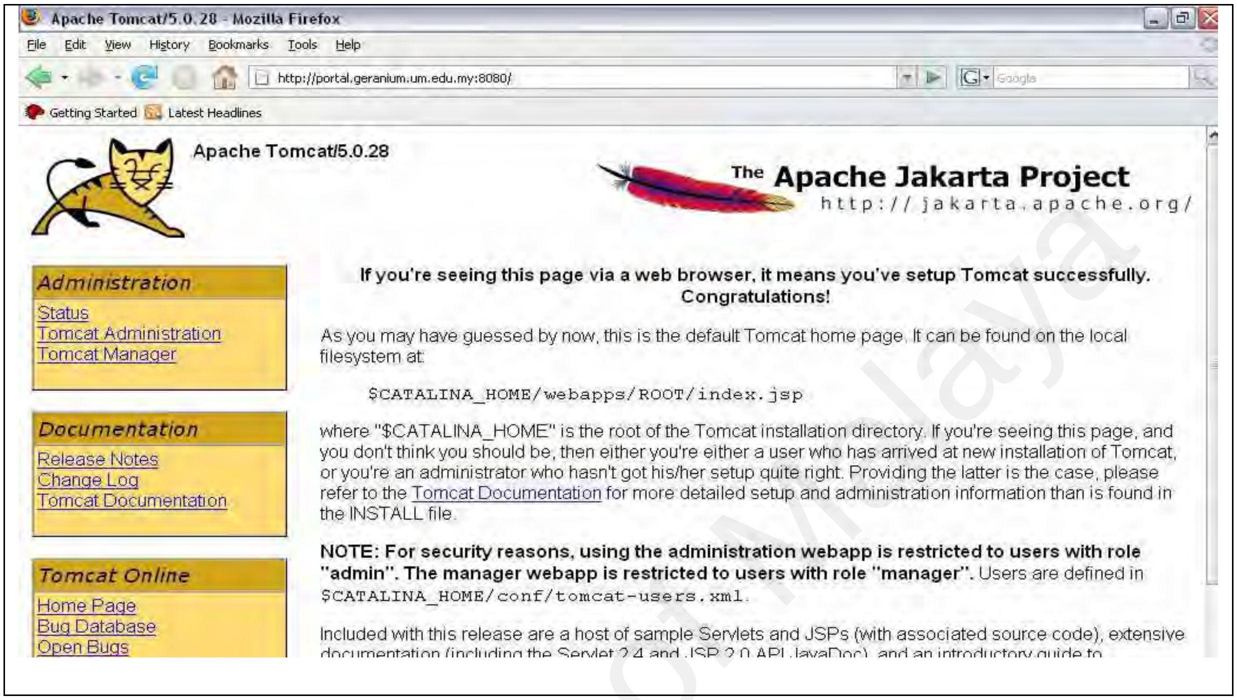
In this section, configuration on Portal Server will be explained until the implementation of GeRaNIUM Grid Portal. In addition, credential retrieval flow and job submission workflow through Grid Portal will be described briefly.

4.2.1 Server configuration: Pre-requisite tools and applications.

GeRaNIUM Grid Portal (GGP) was implemented on Portal Server which uses Rocks version 4.1 embedded with Centos-4.4 operating system. Rocks-4.1 has a complete package of grid contains Globus Toolkit version 4.0.2 that is compatible with GridSphere-2.1, Apache-Ant version 1.3 and Java Development Kit version 1.5. Those tools were required for the installation of GGP.

GGP was developed using GridSphere-2.1 Portal Framework. GridSphere version 2.1 was the most stable software when it was released during GGP development in 2005. GridSphere-2.1 also compatibles with GridPortlets-1.3 which is a third-party portlet provides a grid application for GeRaNIUM. GridPortlets-1.3 provides a collection of simple, easy-to-use and well integrated portlets. The configured GridPortlet made use of the Java Comodity Grid (CoG) Kit version 1.2 due to the “ogsa-3.2.1” that was set in grid portlets properties during the installation. The CoG Kit tool has been used to perform many tasks on the Portal Server, including retrieving credentials from Myproxy, submitting jobs to Globus Gatekeepers, transferring files with Grid Ftp and setting up GASS server to collect job output. Besides, Tomcat version 5 was also installed on

Portal Server as a hosting environment for GGP which also makes Portal Server accessible through the Portal address, at *http://portal.geranium.um.edu.my* (Figure 4.6).



If you're seeing this page via a web browser, it means you've setup Tomcat successfully. Congratulations!

As you may have guessed by now, this is the default Tomcat home page. It can be found on the local filesystem at:

```
$CATALINA_HOME/webapps/ROOT/index.jsp
```

where "$CATALINA_HOME" is the root of the Tomcat installation directory. If you're seeing this page, and you don't think you should be, then either you're either a user who has arrived at new installation of Tomcat, or you're an administrator who hasn't got his/her setup quite right. Providing the latter is the case, please refer to the Tomcat Documentation for more detailed setup and administration information than is found in the INSTALL file.

NOTE: For security reasons, using the administration webapp is restricted to users with role "admin". The manager webapp is restricted to users with role "manager". Users are defined in `$CATALINA_HOME/conf/tomcat-users.xml`.

Included with this release are a host of sample Servlets and JSPs (with associated source code), extensive documentation (including the Servlet 2.4 and JSP 2.0 API JavaDoc), and an introductory guide to

Figure 4.6 *Apache Tomcat 5.0.28 installed on Portal Server*

4.2.2 Configuration of GeRaNIUM Grid Portal

GeRaNIUM Grid Portal (GGP) is capable in providing a secure central access point for all users and researchers in GeRaNIUM. GGP also serves to conceal the complexity of interacting with many different resources and applications from the end users' perspective and providing a user-friendly interface due to various users' IT experiences.

GGP enables researchers to utilize GeRaNIUM services and resources through a web-base interface. Researchers can view details (status, load, job queues) on resources in GeRaNIUM. Researchers can also view network bandwidth and latency of a cluster, the aggregate capabilities of all nodes, other applications installed on clusters, submit job and manage data and files in a collective clusters.

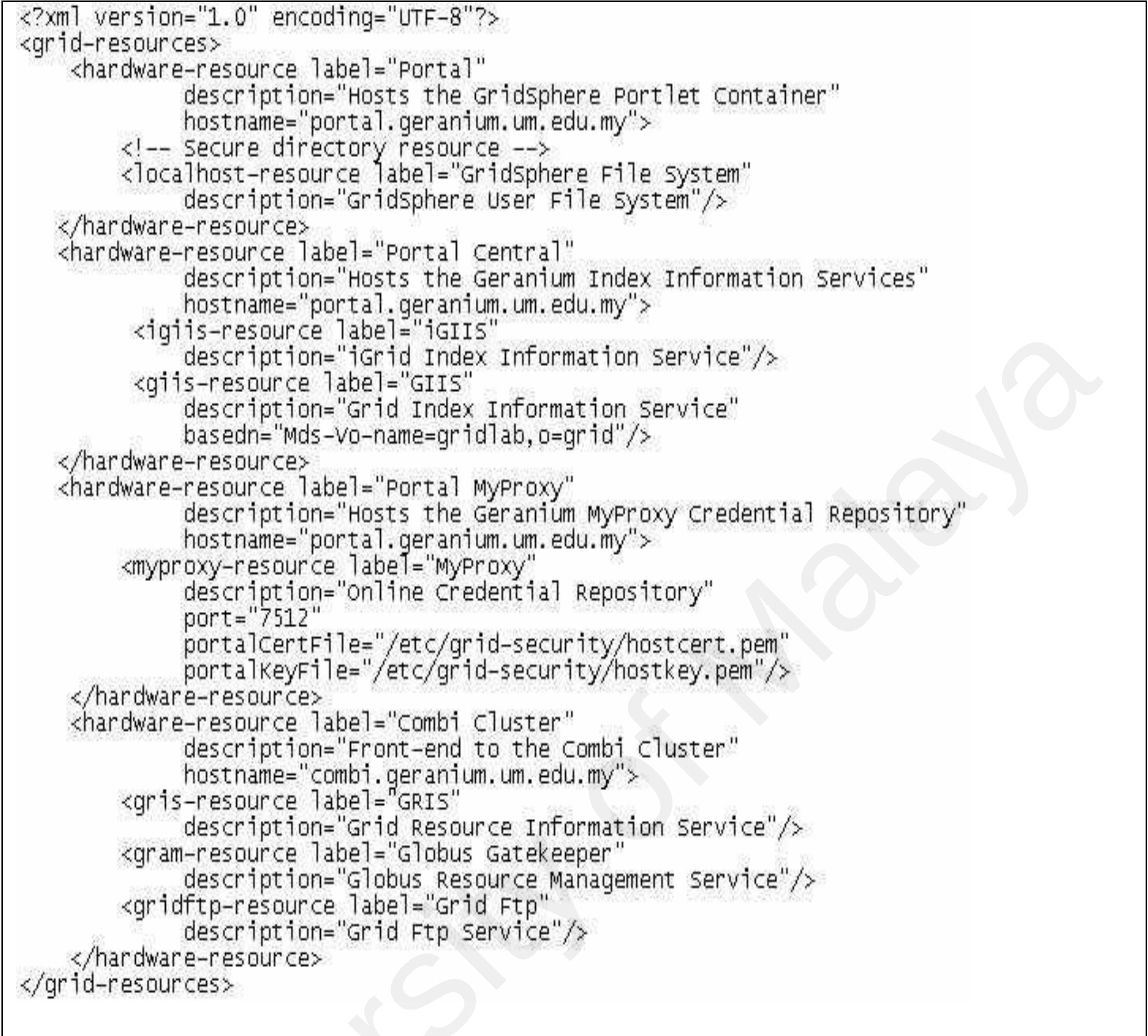
GridSphere-2.1 was installed in Portal Server relied on apache-ant for the compilation and deployment. Firstly, the JAR file (junit-3.8.1.jar) from GridSphere's lib folder (located in GridSphere installer folder) was copied to lib folder in ANT_HOME (apache-ant directory). Then "`ant install`" command was invoked in GridSphere installer folder. The command compiled, deployed and built GridSphere source onto tomcat servlet container. "BUILD SUCCESSFUL" message appeared when the installation complete. Lastly, tomcat server was started by typing command `$CATALINA_HOME/bin/startup.sh` in command windows thus, made GridSphere ready to be accessed at *http://portal.geranium.um.edu.my:8080/gridsphere.*

The next part is the installation of GridPortlet version 1.3. This portlet run relies on GridSphere which provides resource registry portlet, credential manager portlet, resource browser portlet, file browser portlet and job submission portlet. GridPortlet-1.3 installer has been copied into GridSphere's projects folder for the installation process. GridPortlet also relied on apache-ant so "`ant install`" command was issued in GridPortlet's directory to build, compile and deploy the application. After the "BUILD SUCCESSFUL" appears, GridPortlet can be accessed from GridSphere portal framework. Lastly, some configurations were done such as, banner designation, mainpage configuration and more to make GridSphere appropriate with GeRaNIUM environment.

4.2.3 Grid Portlet services.

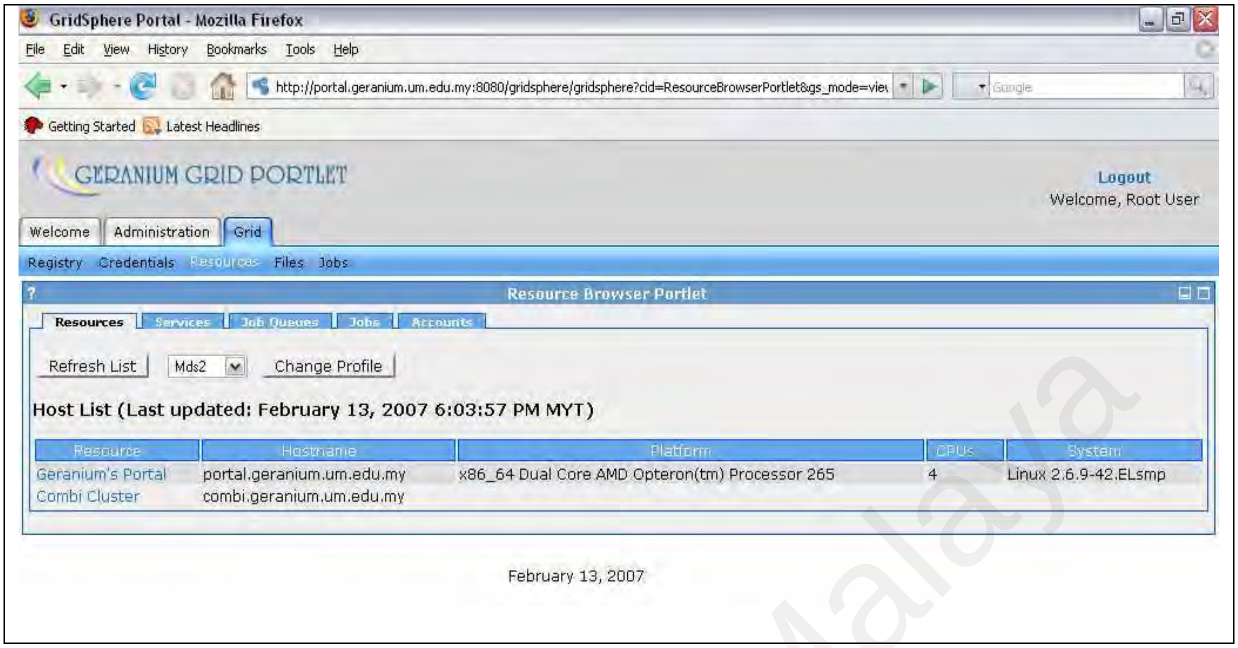
Installation of gridportlet made GGP to have interfaces for registering clusters, viewing clusters specification, browsing files, retrieving user's credential and monitoring user's jobs. All resources were registered, administered and setup through the **Registry Portlet** which allows the alteration of Resources.xml (Figure 4.7) file (located at /opt/tomcat5/webapps/gridportlets/WEB-INF/ directory) remotely. All registered resources can be viewed in the **Resources Portlet** which allows users to view hardware descriptions, services, softwares and user accounts in each cluster (Figure 4.8).

From the gridportlet, users can view, browse and download files in the **Files Portlet**. In this portlet, users are allowed to view, download or upload files at the collective resources. The core part of this portlet is a **Jobs Portlet** which allows user to submit job with an on-click-wizard interface. But before they are allowed to submit job, they need to apply for credentials that will be explained in Section 4.2.4.



```
<?xml version="1.0" encoding="UTF-8"?>
<grid-resources>
    <hardware-resource label="Portal"
             description="Hosts the GridSphere Portlet Container"
             hostname="portal.geranium.um.edu.my">
        <!-- Secure directory resource -->
        <localhost-resource label="GridSphere File System"
             description="GridSphere User File System"/>
   </hardware-resource>
   <hardware-resource label="Portal Central"
             description="Hosts the Geranium Index Information Services"
             hostname="portal.geranium.um.edu.my">
         <igiis-resource label="iGIIS"
             description="iGrid Index Information Service"/>
         <giis-resource label="GIIS"
             description="Grid Index Information Service"
             basedn="Mds-Vo-name=gridlab,o=grid"/>
   </hardware-resource>
   <hardware-resource label="Portal MyProxy"
             description="Hosts the Geranium MyProxy Credential Repository"
             hostname="portal.geranium.um.edu.my">
        <myproxy-resource label="MyProxy"
             description="Online Credential Repository"
             port="7512"
             portalCertFile="/etc/grid-security/hostcert.pem"
             portalKeyFile="/etc/grid-security/hostkey.pem"/>
    </hardware-resource>
    <hardware-resource label="Combi Cluster"
             description="Front-end to the Combi Cluster"
             hostname="combi.geranium.um.edu.my">
        <gris-resource label="GRIS"
             description="Grid Resource Information Service"/>
        <gram-resource label="Globus Gatekeeper"
             description="Globus Resource Management Service"/>
        <gridftp-resource label="Grid Ftp"
             description="Grid Ftp Service"/>
    </hardware-resource>
</grid-resources>
```

Figure 4.7 *Registration of resources in Resource Portlet*


GridSphere Portal - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://portal.geranium.um.edu.my:8080/gridsphere/gridsphere?cid=ResourceBrowserPortlet&gs_mode=vie
Getting Started Latest Headlines
GERANIUM GRID PORTLET
Logout
Welcome, Root User
Welcome Administration Grid
Registry Credentials Resources Files Jobs
Resource Browser Portlet
Resources Services Job Queues Jobs Accounts
Refresh List Mds2 Change Profile

Host List (Last updated: February 13, 2007 6:03:57 PM MYT)

Resource	Hostname	Platform	CPUs	System
Geranium's Portal	portal.geranium.um.edu.my	x86_64 Dual Core AMD Opteron(tm) Processor 265	4	Linux 2.6.9-42.ELsmp
Combi Cluster	combi.geranium.um.edu.my			

February 13, 2007

Figure 4.8 *Resources Portlet*

4.2.4 Application of user's credential

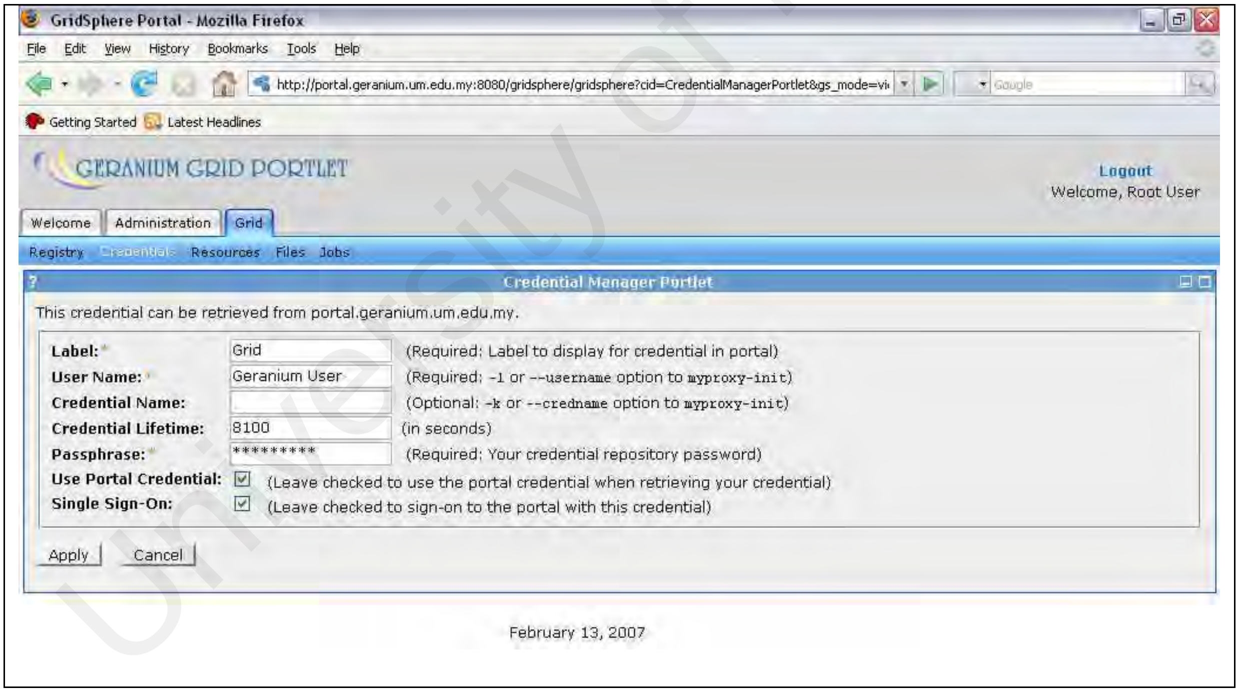
In order to submit job through Jobs portlet, a credential for users was created and delegated by administrator using Globus tool and MyProxy. This credential provides proxies for users to make them act on behalf and also to minimize exposure of user's private key. In the **Credential** portlet, users are allowed to retrieve credentials by specifying their credential details in the New Credential form. This form also allowed users to renew credential when it expired.

As an example, *geranium-test* credential was created using globus and delegated to Portal Server using MyProxy Client Programme. The delegation made *geranium-test* has permission to remote computing resources in GeRaNIUM. To get the credential `myproxy-init` command was invoked with some option for username (`-l`), hostname (`-s`) and credential lifetime (`-c`) (as in Figure 4.9).

```
[geranium-test@portal   ~]#   myproxy-init   -l   "Geranium
Account" -c 0 -s portal.geranium.um.edu.my
```

Figure 4.9 *Command to retrieve credential with some option.*

Once user's credential was successfully delegated by administrators, users can specify their credential details in the New Credential form. This form will delegate their credential with GGP (Figure 4.10) thus; allowing them to submit jobs through the Jobs Portlet. The job workflow will be explained in the next section.


GridSphere Portal - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://portal.geranium.um.edu.my:8080/gridsphere/gridsphere?cid=CredentialManagerPortlet&gs_mode=vi
Google
Getting Started
Latest Headlines
GERANIUM GRID PORTLET
Logout
Welcome, Root User
Welcome
Administration
Grid
Registry
Credentials
Resources
Files
Jobs
Credential Manager Portlet
Apply
Cancel
February 13, 2007

This credential can be retrieved from portal.geranium.um.edu.my.

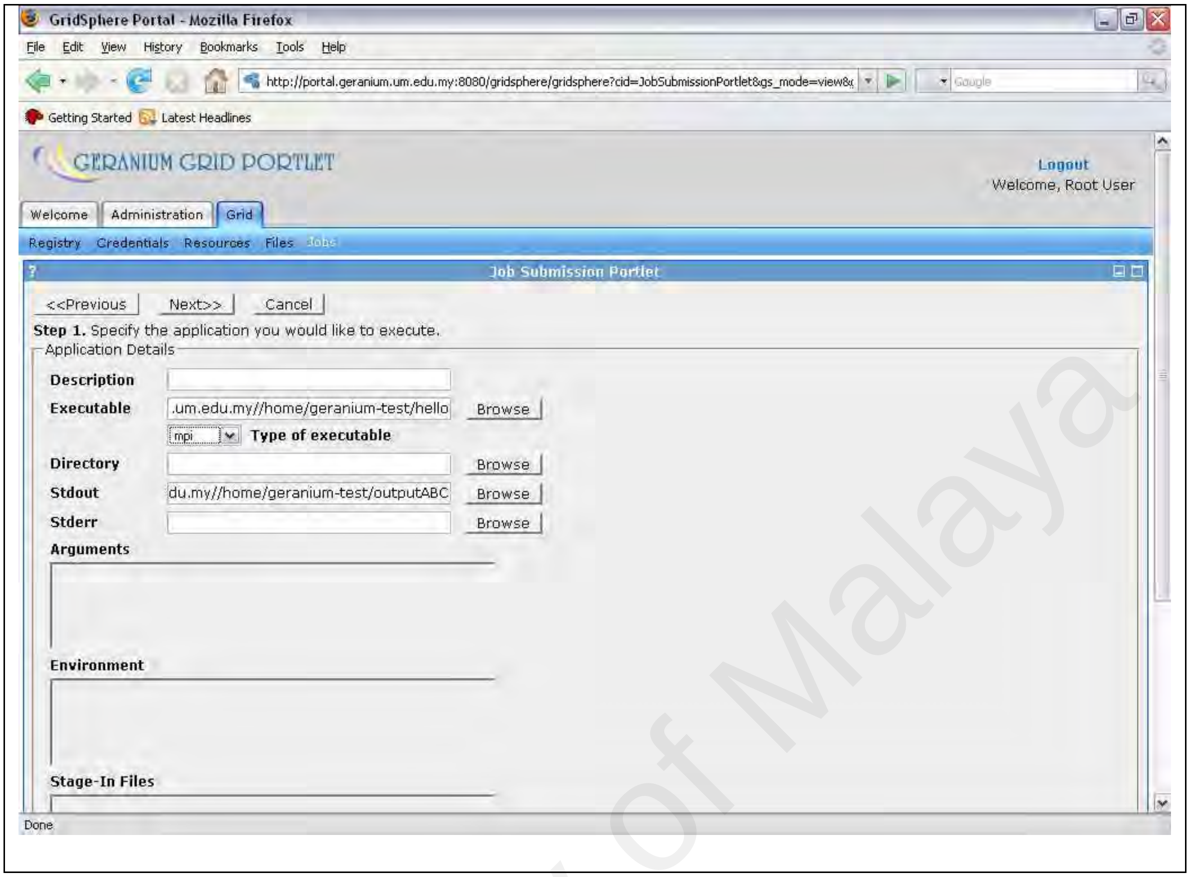
Label: *	Grid	(Required: Label to display for credential in portal)
User Name: *	Geranium User	(Required: -l or --username option to myproxy-init)
Credential Name:		(Optional: -k or --credname option to myproxy-init)
Credential Lifetime:	8100	(in seconds)
Passphrase: *	*********	(Required: Your credential repository password)
Use Portal Credential:	☑	(Leave checked to use the portal credential when retrieving your credential)
Single Sign-On:	☑	(Leave checked to sign-on to the portal with this credential)

Figure 4.10 *An example of complete forms for new credential application.*

4.3 Job submission workflow through GeRaNIUM Grid Portal

GeRaNIUM Grid Portal (GGP) provides an on-click-wizard interface for users to submit job to GeRaNIUM. Jobs Portlet in GGP could reduce the learning curve for users to use grid services anytime and anywhere since GGP is an online web interface.

To submit job through portal, the first wizard in Jobs Portlet display a form for users to specify their job details. In this form users need to specify their job details in the description textbox then specifying the executable file in the Executable browsing box. The executable file in grid resources can be browsed and chosen in the File Browser Portlet after clicking the browse button (Figure 4.11). Next, users can specify the location of a file to be processed in the arguments area.



http://portal.geranium.um.edu.my:8080/gridsphere/gridsphere?cid=JobSubmissionPortlet&gs_mode=view&

Step 1. Specify the application you would like to execute.

Application Details

Description

Executable .um.edu.my//home/geranium-test/hello

mpi **Type of executable**

Directory

Stdout du.my//home/geranium-test/outputABC

Stderr

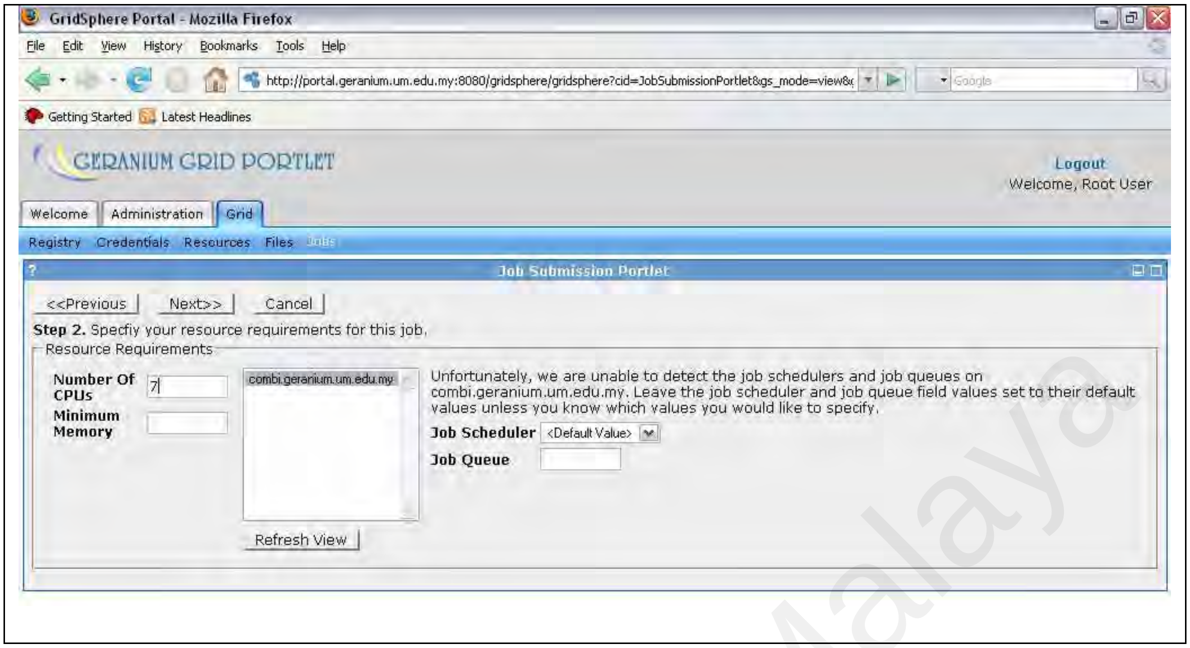
Arguments

Environment

Stage-In Files

Figure 4.11 *The first wizard of Jobs Portlet.*

In the second wizard, users need to specify the requirement for their job. Users can specify how many processor, memory, scheduler and clusters they want to use (Figure 4.12). The next wizard display user's specified information and requires users to confirm for submission. The last wizard will display user's job status. If the job status is successful, users can display job's output or download it to display on their machine (Figure 4.13).


GridSphere Portal - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://portal.geranium.um.edu.my:8080/gridsphere/gridsphere?cid=JobSubmissionPortlet&gs_mode=view&
Google
Getting Started Latest Headlines
GERANIUM GRID PORTLET
Logout
Welcome, Root User
Welcome Administration Grid
Registry Credentials Resources Files Jobs
Job Submission Portlet
<<Previous Next>> Cancel
Refresh View

Step 2. Specfiy your resource requirements for this job.

Resource Requirements

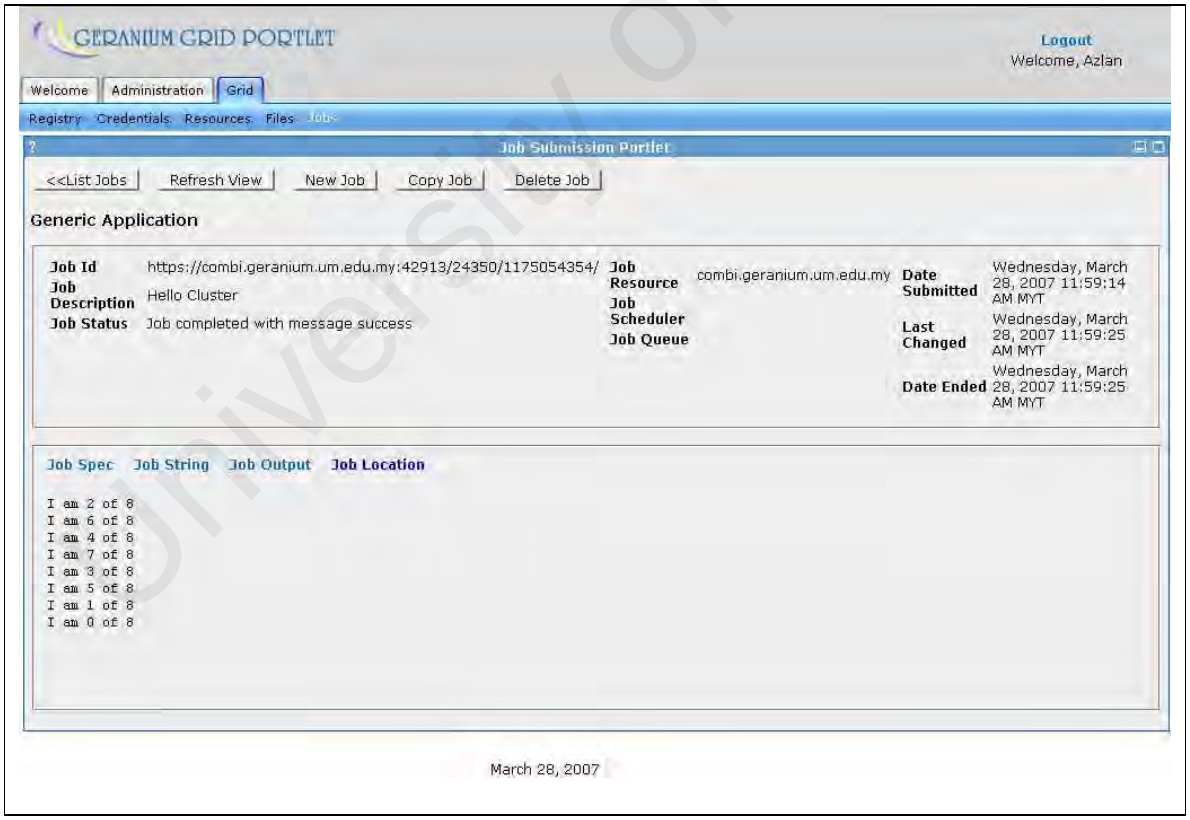
Number Of CPUs	7
Minimum Memory	

combi.geranium.um.edu.my

Unfortunately, we are unable to detect the job schedulers and job queues on combi.geranium.um.edu.my. Leave the job scheduler and job queue field values set to their default values unless you know which values you would like to specify.

Job Scheduler	<Default Value>
Job Queue	

Figure 4.12 The second wizard of Jobs Portlet.


GERANIUM GRID PORTLET
Logout
Welcome, Azlan
Welcome Administration Grid
Registry Credentials Resources Files Jobs
Job Submission Portlet
<<List Jobs Refresh View New Job Copy Job Delete Job
Job Spec Job String Job Output Job Location
March 28, 2007

Generic Application

Job Id	https://combi.geranium.um.edu.my:42913/24350/1175054354/
Job Description	Hello Cluster
Job Status	Job completed with message success
Job Resource	combi.geranium.um.edu.my
Job Scheduler	
Job Queue	
Date Submitted	Wednesday, March 28, 2007 11:59:14 AM MYT
Last Changed	Wednesday, March 28, 2007 11:59:25 AM MYT
Date Ended	Wednesday, March 28, 2007 11:59:25 AM MYT

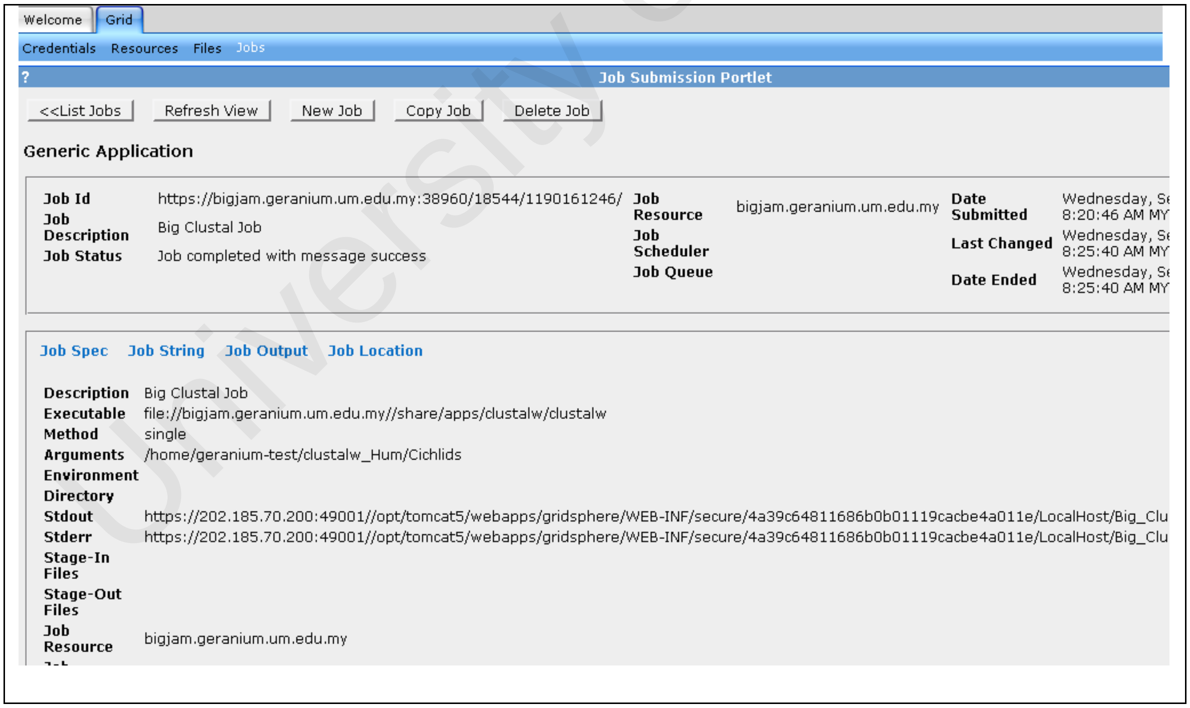
```
I am 2 of 8
I am 6 of 8
I am 4 of 8
I am 7 of 8
I am 3 of 8
I am 5 of 8
I am 1 of 8
I am 0 of 8
```

Figure 4.13 The last wizard of Jobs Portlet displays job status and job output.

4.3.1 Job submission in single and parallel processing

a) Single Processing.

In this study, various input files were used according to the sequence length and the number of organisms. Each data was aligned with a single processing using ClustalW in a workstation using GGP (Figure 4.14). During this process, each data was constructed in pairwise alignment, and then the guide-tree was generated followed by multiple sequences alignment. ClustalW produced the aligned sequences in .aln file and phylogeny tree in .dnd file. The process runtime were recorded and averaged to 6 execution times respectively. The phylogeny trees taken to be used as a control result.



Generic Application

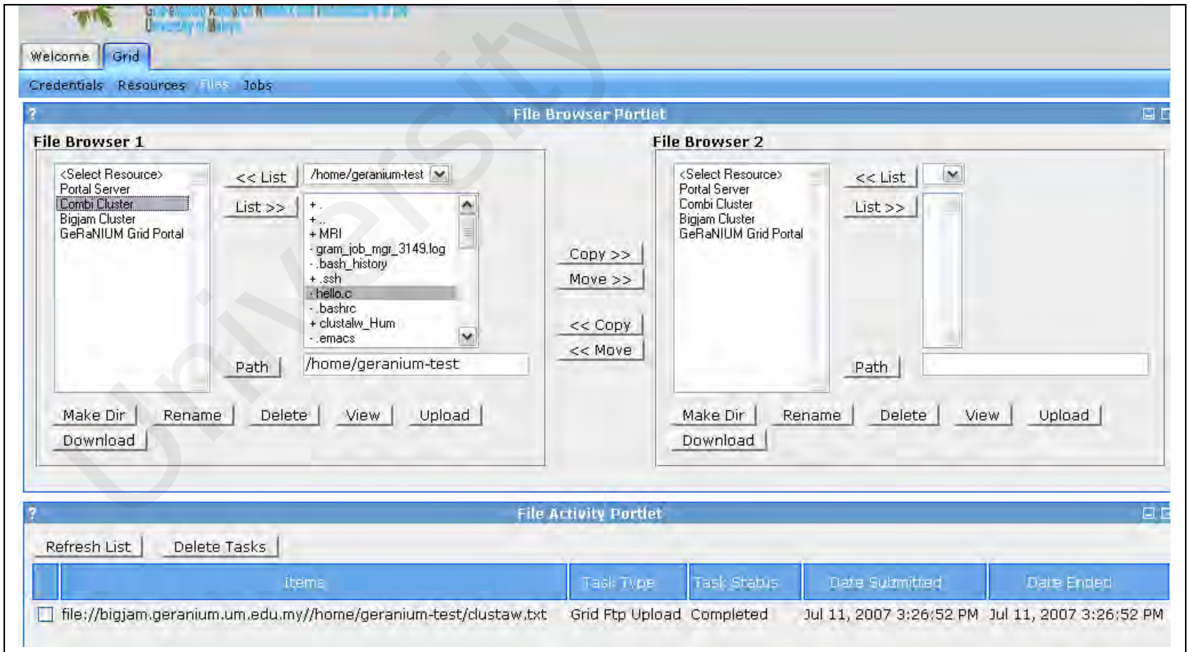
Job Id	https://bigjam.geranium.um.edu.my:38960/18544/1190161246/	**Job Resource**	bigjam.geranium.um.edu.my	**Date Submitted**	Wednesday, Se 8:20:46 AM MY
Job Description	Big Clustal Job	**Job Scheduler**		**Last Changed**	Wednesday, Se 8:25:40 AM MY
Job Status	Job completed with message success	**Job Queue**		**Date Ended**	Wednesday, Se 8:25:40 AM MY

Description	Big Clustal Job
Executable	file://bigjam.geranium.um.edu.my//share/apps/clustalw/clustalw
Method	single
Arguments	/home/geranium-test/clustalw_Hum/Cichlids
Environment	
Directory	
Stdout	https://202.185.70.200:49001//opt/tomcat5/webapps/gridsphere/WEB-INF/secure/4a39c64811686b0b01119cacbe4a011e/LocalHost/Big_Clu
Stderr	https://202.185.70.200:49001//opt/tomcat5/webapps/gridsphere/WEB-INF/secure/4a39c64811686b0b01119cacbe4a011e/LocalHost/Big_Clu
Stage-In Files	
Stage-Out Files	
Job Resource	bigjam.geranium.um.edu.my

Figure 4.14 *Using GGP to submit job to Bigjam workstation*

b) Parallel Processing

In order to look at cluster computing performance, Combi Cluster was used as a cluster computing platform to run a sequence alignment process in parallel. As stated in the previous chapter, ClustalW-MPI was installed in Combi Cluster to align previous problems in parallel. Firstly those problems were aligned using 2 processing power. The processing power was increased from 2 to 9 processing power respectively. The alignment processes were done similarly as in single ClustalW but in ClustalW-MPI the alignment processes ran in parallel. The process runtime displayed in GGP were recorded and averaged to 6 execution times respectively. Phylogeny trees produced were downloaded and then were compared with control result from the single processing (Figure 4.15).



File Browser 1

<Select Resource>
Portal Server
Combi Cluster
Bigjam Cluster
GeRaNIUM Grid Portal

/home/geranium-test

+ .
+ ..
+ MRI
- gram_job_mgr_3149.log
- .bash_history
+ .ssh
- hello.c
- .bashrc
+ clustalw_Hum
- .emacs

Path: /home/geranium-test

File Browser 2

<Select Resource>
Portal Server
Combi Cluster
Bigjam Cluster
GeRaNIUM Grid Portal

	Items	Task Type	Task Status	Date Submitted	Date Ended
□	file://bigjam.geranium.um.edu.my//home/geranium-test/clustaw.txt	Grid Ftp Upload	Completed	Jul 11, 2007 3:26:52 PM	Jul 11, 2007 3:26:52 PM

Figure 4.15 *File Browser Portlet to download and upload file.*

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Performance Results

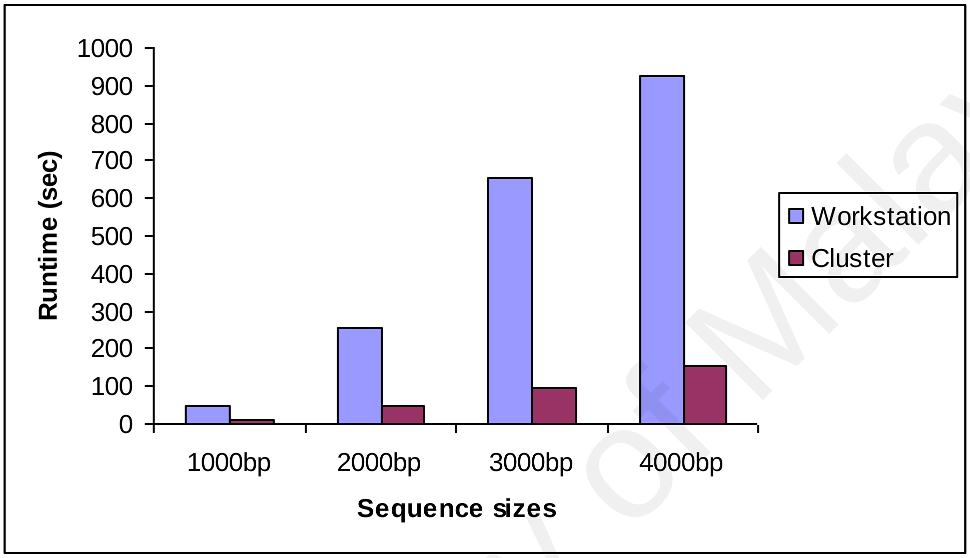


Figure 5.1 *Comparison of runtime performance between workstation and cluster.*

From Figure 5.1, the bar chart illustrates the results of runtime taken between workstation and cluster (refer to Appendix A). 30 sequences of organisms with different length were used to run sequence alignment process. The process runtime taken were compared to look at the performance of single computing and cluster computing. From the bar chart, the process runtime fall drastically when the alignment processes run in cluster computing. Although the length of organisms was expanded, the process runtime were still shorter in cluster than in workstation. From the graph, the speedup performance also becomes bigger when running longer sequences length. As an example, the runtime at 1000bp length run in workstation and clusters has speed up about 76 percents. This speed up percentage increased to 83 percents when processing the 4000bp length on both machines. From this trend, it is anticipated that cluster computing would gives a big impact and has a potential to accelerate bioinformatics complex tasks.

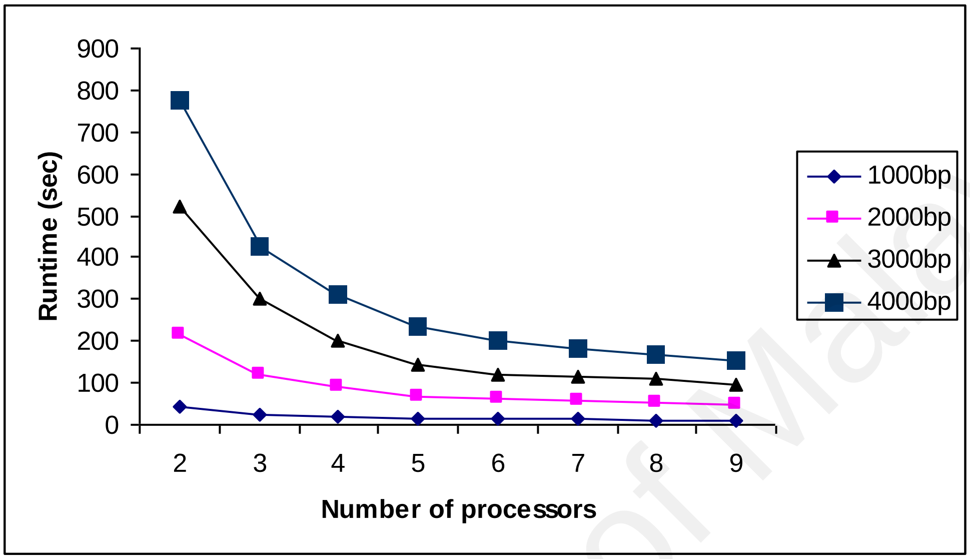


Figure 5.2 *Runtime performace of ClustalW-MPI on different number of processors.*

The graph in Figure 5.2 shows the accelerating of sequence alignment process through the increment of processors. This result was produced by running 30 samples of sequences with different length. As shown from the graph, the runtime taken by all processes gradually decrease when the number of processors was added from 2 to 6 processors but then started to stabilize from 7 to 9 processors at any sequences length. The stabilization might occur because of parallel computing slowdown. Parallel slowdown is a phenomenon in parallel computing where parallelization of a parallel computer program beyond a certain point causes the program to run slower (takes more time to run to completion).

Parallel slowdown is typically the result of a communications bottleneck. From the graph, when processing nodes were added from 6 to 9 processors, possibly each processing node spends progressively more time doing communication than useful processing. Probably start from 6 processors, the communications overhead created by adding another processing node surpasses the increased processing power that node provides, resulting in parallel slowdown.

5.2 Discussion

5.2.1 Single and Parallel Processing Performance

From this study, results found that the execution time on a single processing tends to drag with the amount of data (Figure 5.1). Meaning, the longer sequence length to be aligned, the more time program needs to complete the alignment process. This is due to the ClustalW algorithm principle in pairwise aligning, where all sequences were being compared one-to-one to generate a similarity matrix. This matrix was used in building a guided tree for the multiple sequence alignment that aligns sequences many-to-many. As a proposed solution, all the processes are much better running on a parallel system rather then using a stand alone workstation or a single processor.

According to Figure 5.1 and Figure 5.2, the performance of computing power could be increased by parallelizing the computing process and by adding more processors on cluster. Parallel processing will make longer sequences processed in shorter execution time by distributing big job into smaller task to the collective processors. This technique would not overwhelm machines capability. However, in workstation, we need to have more budgets to buy a new system in order to have a higher capability performance with higher specification of processors, memory and disk space. This is not only a waste on valuable resources but would also increase the cost of research. By clustering existing workstations, we would not only have a high performance computing power which is affordable but also could increase the performance by simply adding more processors in an existing cluster.

5.2.2 Result Consistency.

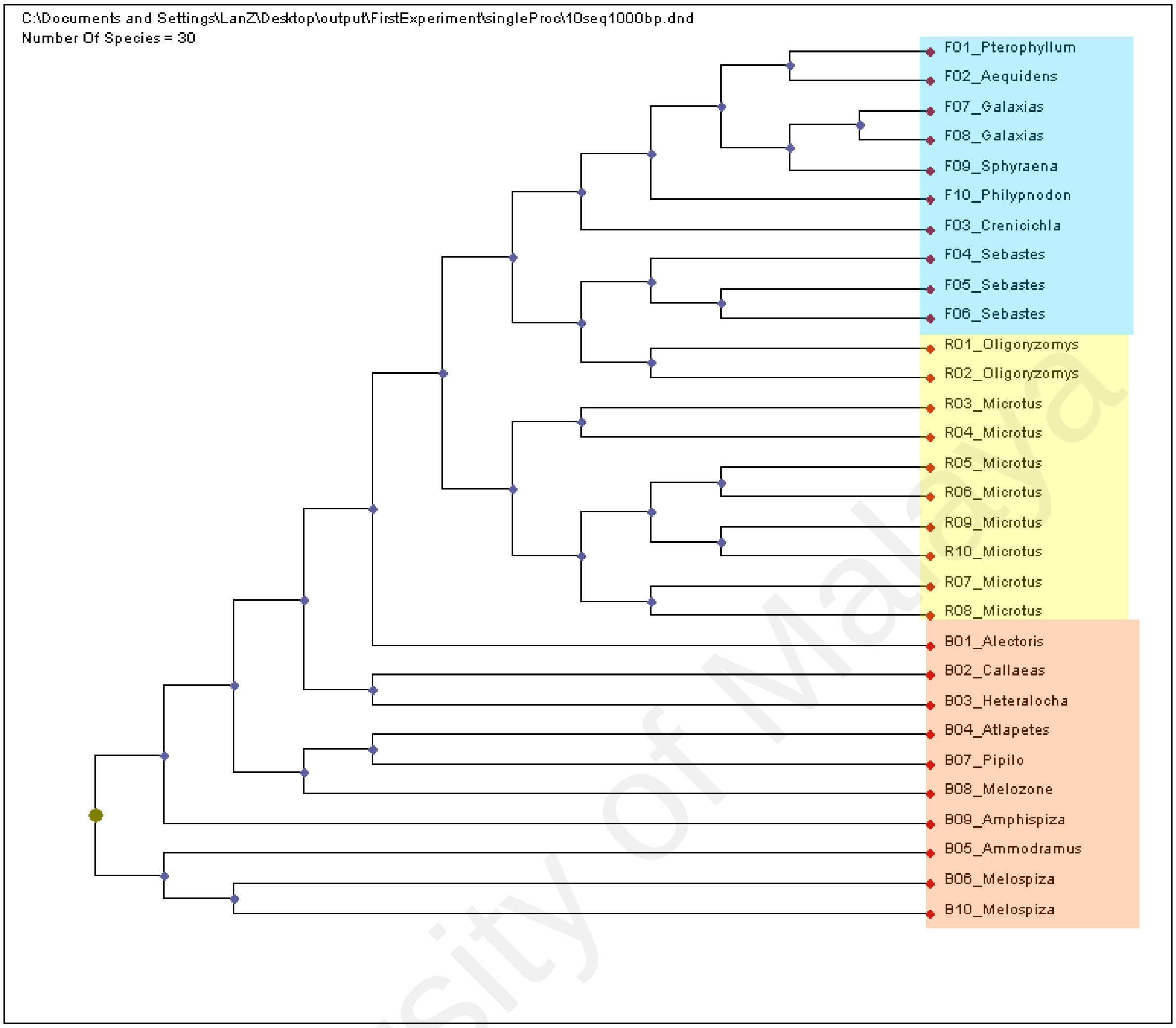
Output produced from single processing and parallel computing were analyzed and studied. The aligned sequences (.aln files) and phylogeny trees (.dnd files) produced were compared to look at the result consistency. For the aligned output (.aln), the sequences aligned by both processes are almost the same. Despite the fact that organisms arrangement in cluster was sometimes slightly different compared to workstation, the gap added and deleted in the alignment was totally the same and consistent in every organism (Figure 5.3 and Figure 5.4). This might be caused by the method used in single processing align samples by order until the last sample. In contrast, parallel processing divided organisms to compute nodes first, then; every sequence of organism was aligned separately. After that, the aligned sequences were resent to master node randomly. The resending process would make organisms' arrangement in cluster dissimilar compared to the single processing arrangement. However, the alignment of sequences at each organism is totally the same. The effect of this issue will produce a different phylogeny tree and make some species to be put in a different group. However, when we look at the phylogeny tree carefully, it seems that the main branch is consistently maintained as in a single processing (Figure 5.6).

```
F01_Pterophyllum      ATCTACCTTCACATCGGAC------GAGGACTTTACTACGGTTCATACCTCTATAAAGAA
F02_Aequidens         ATCTATCTTCACATCGGCC------GAGGACTTTATTACGGCTCATACCTCTACAAAGAA
F07_Galaxias          ATTTATATGCACATTGGAC------GAGGACTTTATTATGGATCTTACCTCTATAAGGAG
F08_Galaxias          ATTTATATGCACATTGGAC------GAGGACTTTATTACGGGTCTTACCTCTATAAGGAG
F09_Sphyraena         ATTTACTTCCACATTGGCC------GAGGACTTTACTACGGCTCTTACTTGAATAAAGCA
F10_Philypnodon       CTCTACTTACACATCGGAC------GAGGCCTATATTACGGATCCTACCTATATAAAGAA
F03_Crenicichla       ATCTACCTCCATATTGGCC------GCGGACTCTACTATGGCTCCTATCTCTACAAAGAG
B06_Melospiza         ATCTATCTACATATCGGCC------GAGGCATCTACTACGGCTCATACCTAAACAAAGAA
B10_Melospiza         ATCTACCTACACATCGGTC------GAGGCATCTACTACGGCTCATACCTAAACAAAGAA
B05_Ammodramus        ATCTACCTACACATCGGCC------GAGGCATCTACTACGGCTCATACCTAAACAAAGAA
B09_Amphispiza        ATCTACCTACATATCGGCC------GAGGAATCTATTACGGCTCATACCTAAACAAAGAG
B04_Atlapetes         ATCTACCTACACATCGGCC------GAGGAATCTACTACGGCTCATATCTCTACAAAGAA
B07_Pipilo            ATCTACCTACACATCGGCC------GAGGAATTTATTACGGCTCATACCTAAACAAAGAA
B08_Melozone          ATCTACCTACACATCGGCC------GAGGAATTTATTACGGCTCATACCTGAACAAAGAA
B02_Callaeas          ATCTACCTACATATCGGCC------GAGGCCTCTACTACGGCTCATACATAAACAAAGAG
B03_Heteralocha       ATCTACCTACATATCGGCC------GAGGACTCTACTACGGCTCATACCTGAACAAAGAG
B01_Alectoris         ATTTTCCTCCACATCGGAC------GCGGCCTATACTATGGCTCCTATCTCTACAAAGAA
R03_Microtus          CTATTTCTACATGTAGGGC------GAGGTGTTTACTACGGCTCCTACAACATAATCGAA
R04_Microtus          CTATTTCTACATGTAGGGC------GAGGTGTTTACTACGGCTCCTACAACATAATCGAA
R05_Microtus          CTATTCCTGCACGTAGGAC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R06_Microtus          CTATTCCTGCACGTAGGAC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R09_Microtus          CTATTCCTGCACGTAGGAC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R10_Microtus          CTATTCCTGCACGTAGGAC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R07_Microtus          CTATTCCTGCACGTAGGGC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R08_Microtus          CTATTCCTGCACGTAGGAC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R01_Oligoryzomys      AACCACTTTCATATAAGC--------AAGAATTAAACATTAAT-GTATTAAAACATTATA
R02_Oligoryzomys      AACCACTTTCATATAAGC--------AAGAATTAAACATTAAT-GTATCAAGACACTATA
F05_Sebastes          CGTTACCTACGTAGGGTACTGCAGAGAGTAGGTTGGTGATGACGGTGGCACCTCAAAAGG
F06_Sebastes          CGTTACCTACGTAGGGTACTGCAGAGAGTAGATTGGTGATAACGGTGGCACCTCAAAAGG
F04_Sebastes          CGTTACCCACGTAGGGTACTGCAGAGAGTAGGTTGGTGATAACGGTGGCACCTCAAAAGG
```

Figure 5.3 *The output from single processing alignment.*

```
F01_Pterophyllum      ATCTACCTTCACATCGGAC------GAGGACTTTACTACGGTTCATACCTCTATAAAGAA
F02_Aequidens         ATCTATCTTCACATCGGCC------GAGGACTTTATTACGGCTCATACCTCTACAAAGAA
F07_Galaxias          ATTTATATGCACATTGGAC------GAGGACTTTATTATGGATCTTACCTCTATAAGGAG
F08_Galaxias          ATTTATATGCACATTGGAC------GAGGACTTTATTACGGGTCTTACCTCTATAAGGAG
F09_Sphyraena         ATTTACTTCCACATTGGCC------GAGGACTTTACTACGGCTCTTACTTGAATAAAGCA
F10_Philypnodon       CTCTACTTACACATCGGAC------GAGGCCTATATTACGGATCCTACCTATATAAAGAA
F03_Crenicichla       ATCTACCTCCATATTGGCC------GCGGACTCTACTATGGCTCCTATCTCTACAAAGAG
B06_Melospiza         ATCTATCTACATATCGGCC------GAGGCATCTACTACGGCTCATACCTAAACAAAGAA
B10_Melospiza         ATCTACCTACACATCGGTC------GAGGCATCTACTACGGCTCATACCTAAACAAAGAA
B05_Ammodramus        ATCTACCTACACATCGGCC------GAGGCATCTACTACGGCTCATACCTAAACAAAGAA
B09_Amphispiza        ATCTACCTACATATCGGCC------GAGGAATCTATTACGGCTCATACCTAAACAAAGAG
B04_Atlapetes         ATCTACCTACACATCGGCC------GAGGAATCTACTACGGCTCATATCTCTACAAAGAA
B07_Pipilo            ATCTACCTACACATCGGCC------GAGGAATTTATTACGGCTCATACCTAAACAAAGAA
B08_Melozone          ATCTACCTACACATCGGCC------GAGGAATTTATTACGGCTCATACCTGAACAAAGAA
B02_Callaeas          ATCTACCTACATATCGGCC------GAGGCCTCTACTACGGCTCATACATAAACAAAGAG
B03_Heteralocha       ATCTACCTACATATCGGCC------GAGGACTCTACTACGGCTCATACCTGAACAAAGAG
B01_Alectoris         ATTTTCCTCCACATCGGAC------GCGGCCTATACTATGGCTCCTATCTCTACAAAGAA
R03_Microtus          CTATTTCTACATGTAGGGC------GAGGTGTTTACTACGGCTCCTACAACATAATCGAA
R04_Microtus          CTATTTCTACATGTAGGGC------GAGGTGTTTACTACGGCTCCTACAACATAATCGAA
R05_Microtus          CTATTCCTGCACGTAGGAC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R06_Microtus          CTATTCCTGCACGTAGGAC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R09_Microtus          CTATTCCTGCACGTAGGAC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R10_Microtus          CTATTCCTGCACGTAGGAC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R07_Microtus          CTATTCCTGCACGTAGGGC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R08_Microtus          CTATTCCTGCACGTAGGAC------GAGGAATTTACTACGGCTCCTACAACATAATCGAA
R01_Oligoryzomys      AACCACTTTCATATAAGC--------AAGAATTAAACATTAAT-GTATTAAAACATTATA
R02_Oligoryzomys      AACCACTTTCATATAAGC--------AAGAATTAAACATTAAT-GTATCAAGACACTATA
F05_Sebastes          CGTTACCTACGTAGGGTACTGCAGAGAGTAGGTTGGTGATGACGGTGGCACCTCAAAAGG
F06_Sebastes          CGTTACCTACGTAGGGTACTGCAGAGAGTAGATTGGTGATAACGGTGGCACCTCAAAAGG
F04_Sebastes          CGTTACCCACGTAGGGTACTGCAGAGAGTAGGTTGGTGATAACGGTGGCACCTCAAAAGG
```

Figure 5.4 *The output from parallel processing alignment.*



```
C:\Documents and Settings\LanZ\Desktop\output\FirstExperiment\singleProc\10seq1000bp.dnd
Number Of Species = 30
```

Figure 5.5 *The phylogeny tree produced by single processing.*

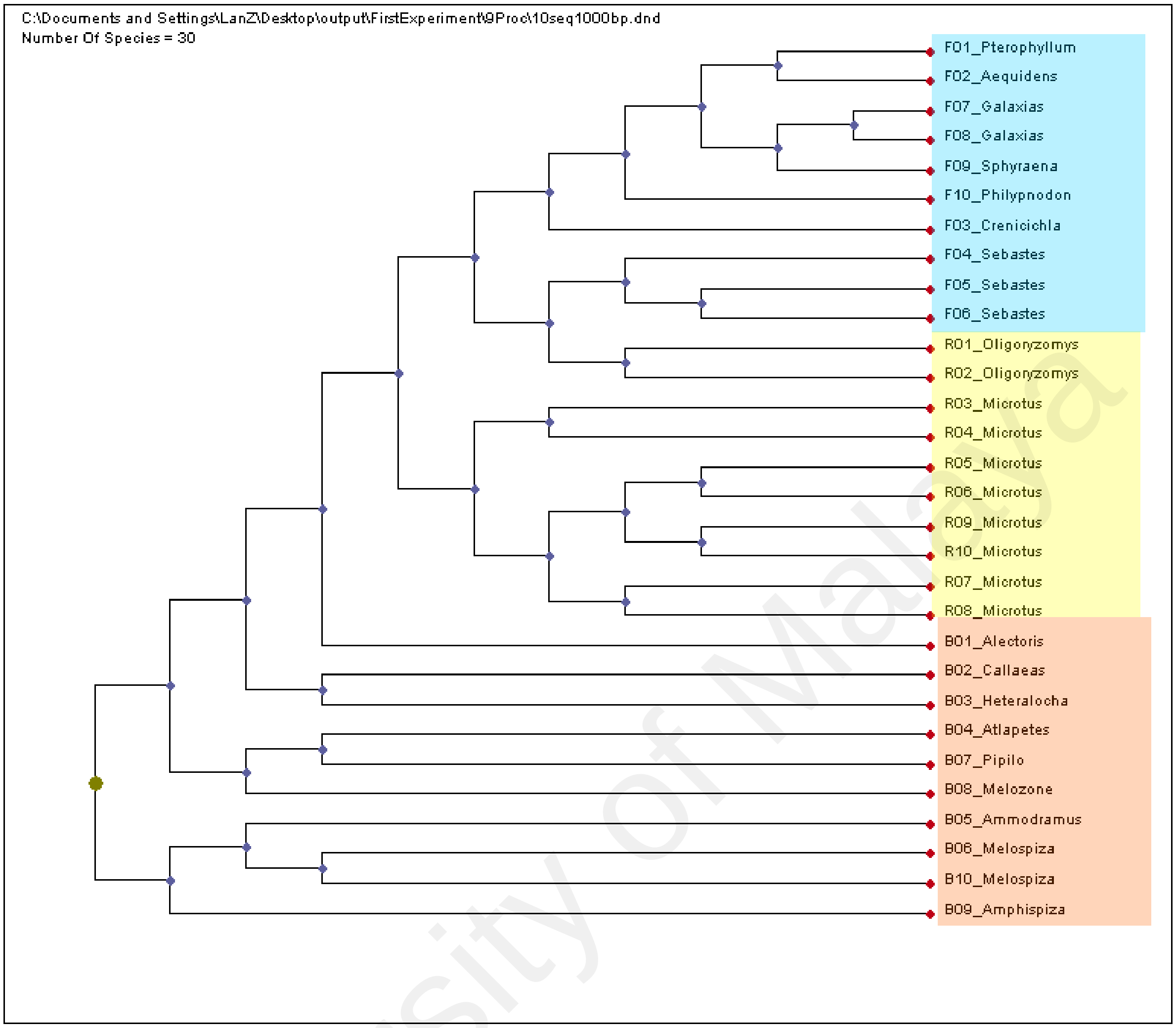


Figure 5.6 *The phylogeny tree produced by parallel processing.*

Results in Figure 5.3 and Figure 5.4 indicate that sequences aligned in parallel processing produced a similar output as in the single processing. The deletion and gap added in parallel processing was handled consistently. Referring to Figure 5.5 and 5.6, phylogeny trees produced from both processing were quite similar. Organisms with same species were grouped in a similar branch. The different between Figure 5.5 and 5.6 is at species B09_Amphispiza. In single processing, this species was grouped in the first branch but in parallel processing it was put in the second branch. This probably happened due to the organism arrangement in the previous parallel processing. The algorithm created in ClustalW-MPI is another possible cause to this issue but, the output from ClustalW-MPI is still acceptable. As shown from the tree, the main branch in ClustalW-MPI's tree is still maintained as in the first hypothesis of clustalW's tree (Figure 3.3). For instance, Fish species were put in the same branch with Rodent species and Bird species were located in a separate branch at both trees. As shown at the highlighted region in Figure 5.4 and Figure 5.5, organisms that have the same genus were consistently grouped in the same branch.

5.3 Issues on GeRaNIUM and GeRaNIUM Grid Portal.

5.3.1 GeRaNIUM Grid Environment Issues

In this work, GeRaNIUM grid environment was successfully extended involving Portal Server, Combi Cluster and Bigjam workstation. GeRaNIUM was extended by locating Portal Server as a broker for campus grid. Portal Server was assembled to manage activities occurred on GeRaNIUM including offering users credential, acting as MyProxy server, operating as a gateway for valid users in GeRaNIUM and also providing users interface for monitoring resources, submitting job and browsing resource in permitted location. Any machine or cluster in campus that was allowed to share or attached with GeRaNIUM needs to trust Portal Server Certificates Authority's (CA) first before providing services to GeRaNIUMs' users. Besides doing this, GeRaNIUM's credential has been centralized and easier to observe and managed by GeRaNIUM's administrator.

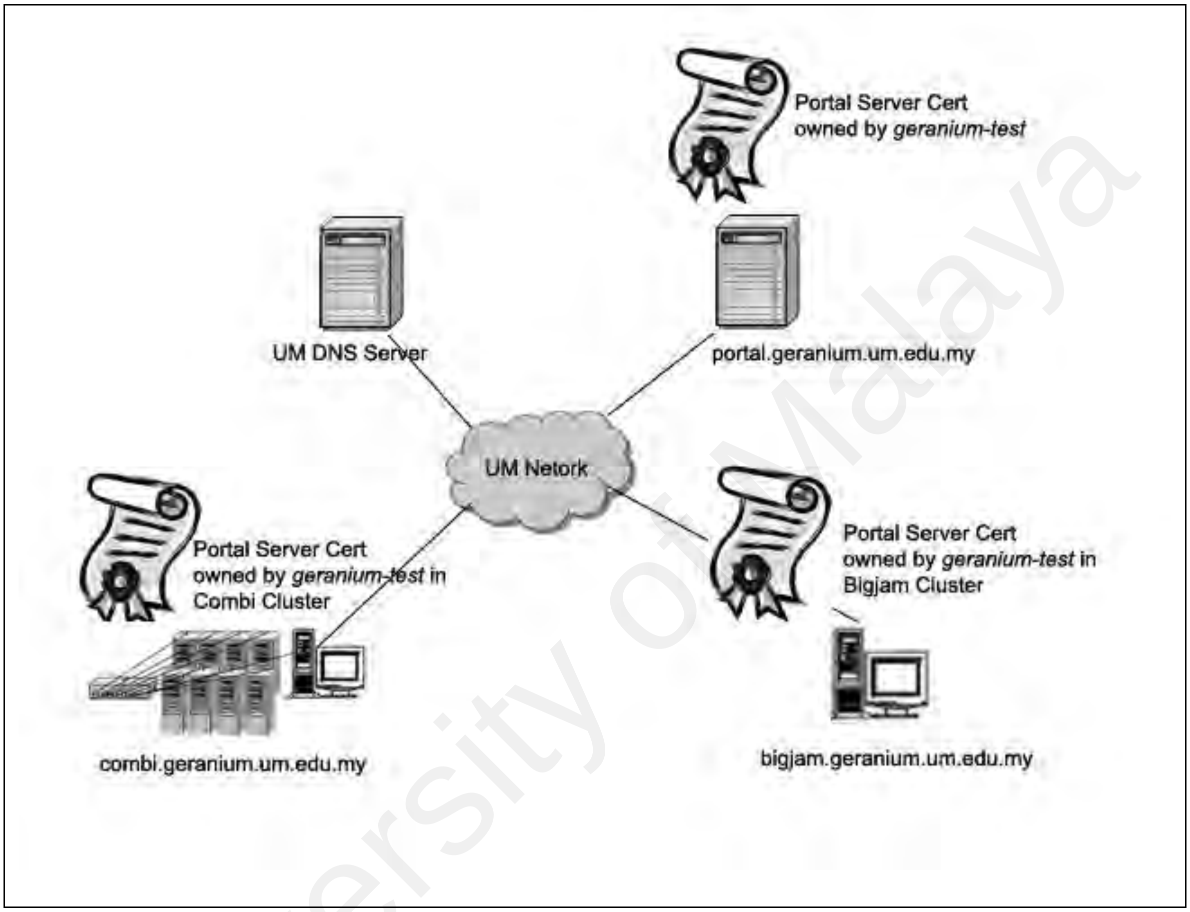


Figure 5.8 *Portal Server certificate installed to every cluster.*

As shown from Figure 5.8, Portal Server's certificate was installed and trusted in Combi cluster and Bigjam workstation. Any machines or clusters trust Portal Server's CA could be considered as GeRaNIUM's member. Any users created and validated in Portal also validated to other clusters or machines attached in GeRaNIUM. Those users also could start using available services and submitting job to any GeRaNIUM's resources. During this thesis user *geranium-test* was successfully created and validated in GeRaNIUM and also has been used as a grant to submit job in Combi Cluster or Bigjam workstation through Portal or directly to those clusters itself.

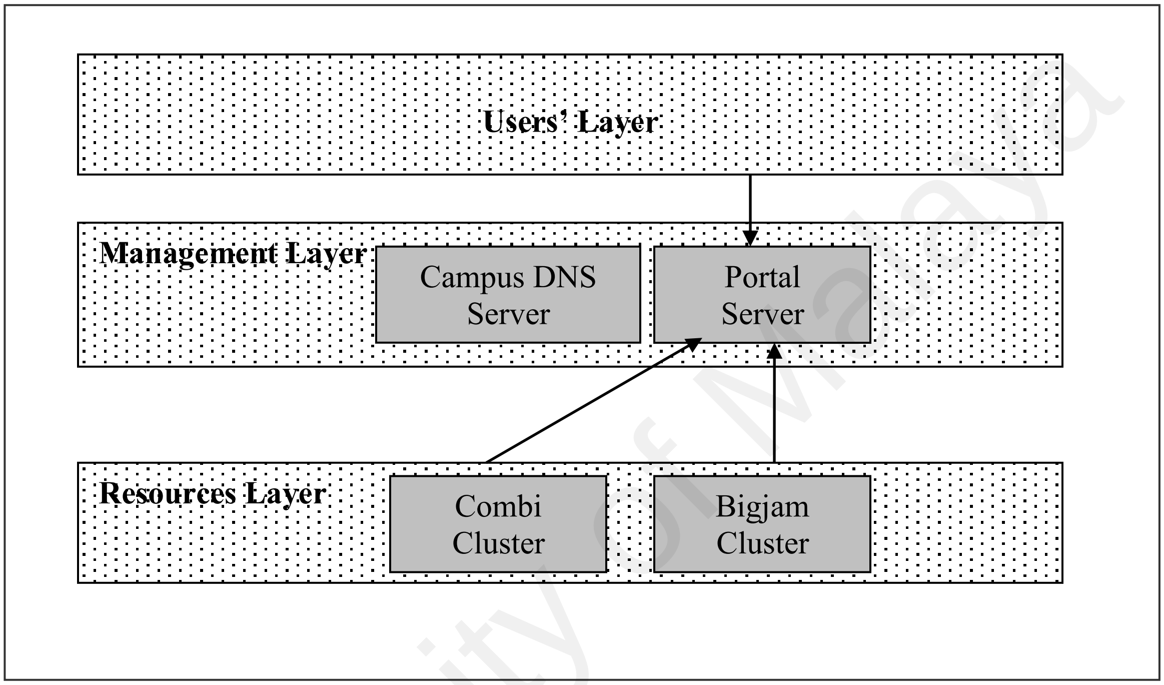


Figure 5.9 *GeRaNIUM environment layers.*

The development of GeRaNIUM grid environment during this thesis has divided GeRaNIUM into 3 layers as shown in Figure 5.9. The first layer is resources layer which locates clusters and machines that provide services to users through Portal Server. Although all clusters available are located in a distributed environment and different places, they were grouped in the same layer in GeRaNIUM. The more resources added at this layer will make more services being offered in GeRaNIUM. The next layer is the management layer which locates campus DNS Server and Portal Server. Campus DNS Server part in this layer is to managed resources' domain name in GeRaNIUM. The hostname of resources which was highly used in GeRaNIUM during grid computing implementation was referred to campus DNS Server to identify the collective resources. Portal Server that was put in this layer has a role as a GeRaNIUM's broker. As a broker it works as a hub between users and GeRaNIUMs' resources and also provides uniform working environment for users. To make the first layer and second layer work efficiently, those layers need a Globus Toolkit which was installed to every resources. Globus Toolkit enables GeRaNIUM to share computing power, databases, and other tools securely online across distributed locations. This toolkit completes with tools for security, information infrastructure, resource management, data management, communication, fault detection, and portability (Globus, 2007). Furthermore, Globus toolkit provides services and libraries for resource monitoring, discovery, and management, plus security and file management within GeRaNIUM. As soon as first and second layer can work efficiently, they are ready to be utilized by users who have been validated in the management layer by Portal Server.

To conclude from the development of GeRaNIUM, it was found that this initial campus grid architecture has the potential for further development which will give a lot of benefits to University of Malaya researchers. In order to make GeRaNIUM become useful for researchers from anywhere and anytime, it is suggested that more resources need to be added in GeRaNIUM. Consequently, this could make GeRaNIUM able to fulfill researchers' needs from any field. GeRaNIUM grid environment needs an intensive management and also good collaboration from researchers within the campus and the whole of Malaysia as well.

5.3.2 GeRaNIUM Grid Portal Issues

GeRaNIUM Grid Portal (GGP) project is a big challenge that has to be faced from the development process all the way to the implementation stage. In the development process it has been proven that software compatibility is the most important thing to note due to the utilization of open source software's. Only certain version is compatible with certain version of software. In this thesis, it was found that Globus toolkit version 4 is compatible with Tomcat version 5. This version of tomcat provides an efficient host environment for Gridsphere-2.1 which worked as a grid portal framework. This portal framework is a web portal which enabled it to develop and package third-party portlet web applications that have been run and administered within the gridsphere container. Gridsphere-2.1 is compatible with Gridportlet-1.3 which provides a user friendly portlet for user to submit jobs and also provides high level of grid resources management as well.

In order to have an efficient and good management of grid environment some policies were made and agreed on. In this thesis a public credential was created as a credential for all users who want to use collective resources in GeRaNIUM. This policy was created to avoid generating many credentials, which would give problems to administrator who manage many credentials at a time. Moreover, for each resources attached to GeRaNIUM it was restricted only to three applications. By doing this, the administrators would not be overwhelmed in resources and also would make it easier for them to manage and maintain. In this way, they would become experts in certain related applications.

In addition, grid portal that was located in Portal Server has made GeRaNIUM capable to share applications within resources in a high level management. It provides an interface for GeRaNIUM environment that enabled researchers to share computing and information resources across departmental and organizational boundaries in a secure and highly efficient manner. The interface has a potential to decrease the learning curve for researchers to use grid computing and provides a uniform working environment

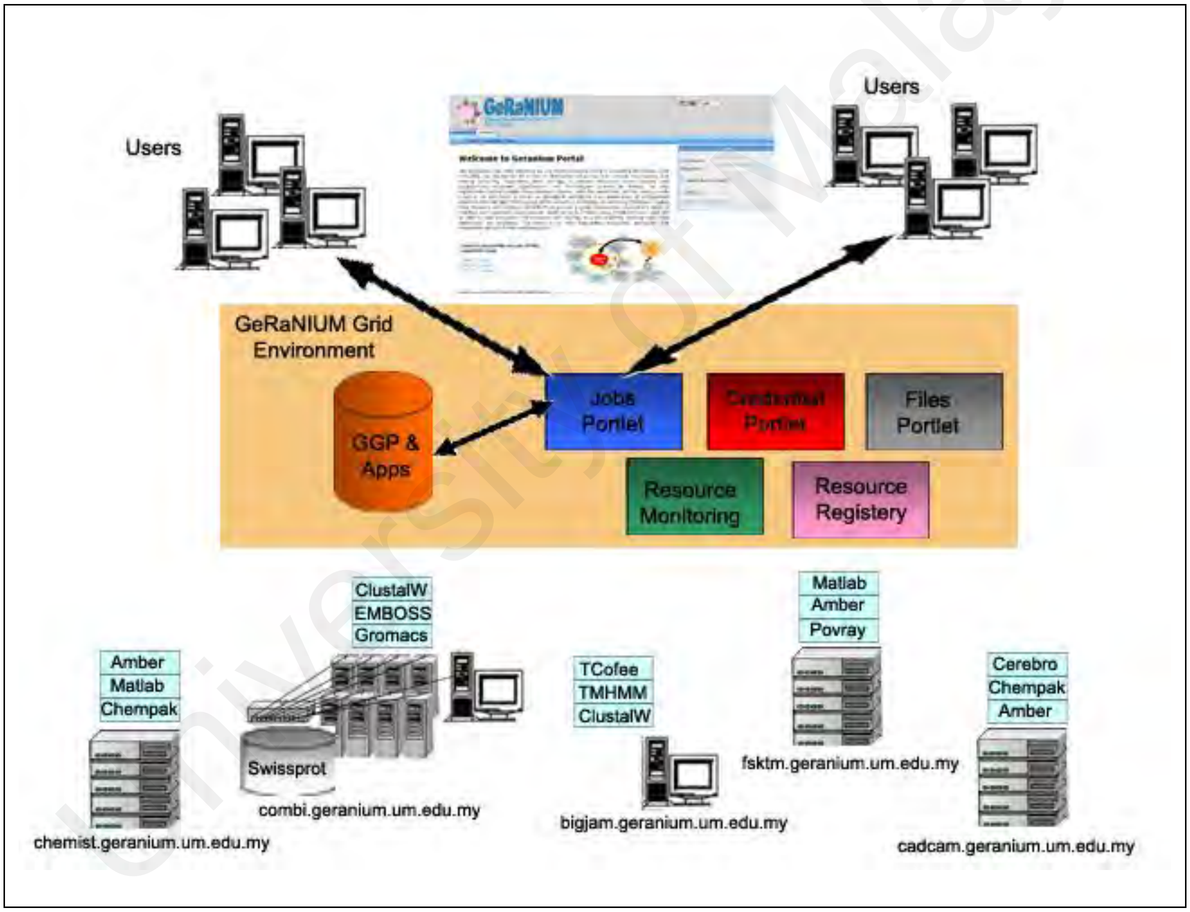
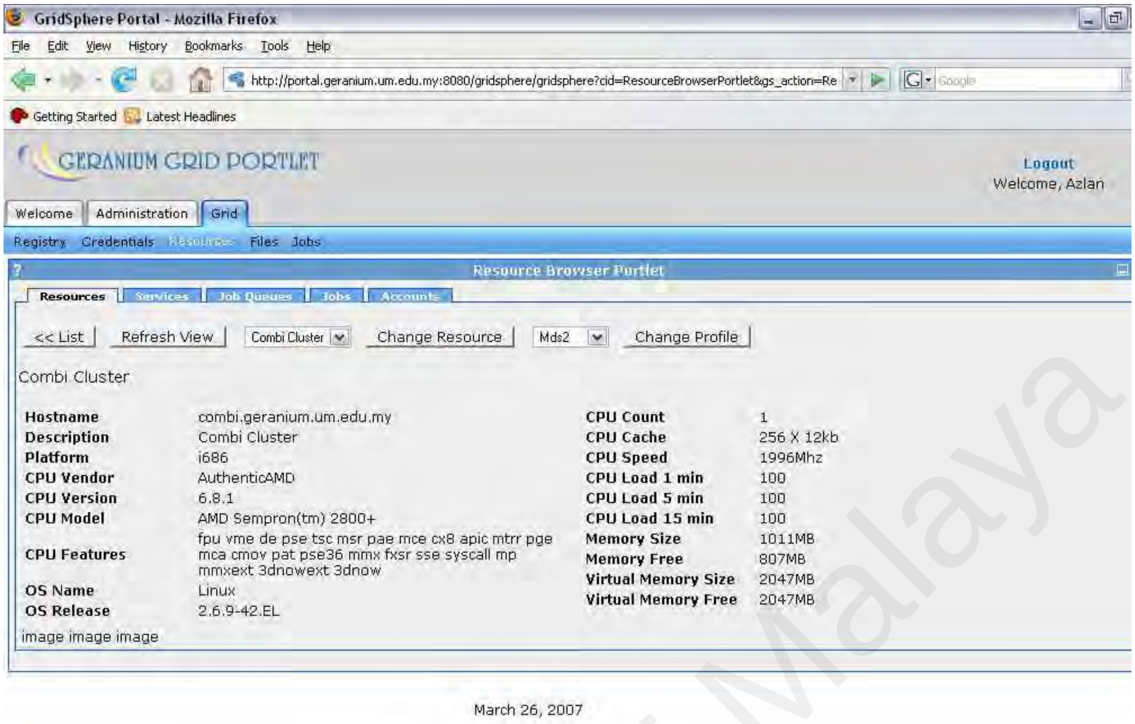


Figure 5.10 *Applications sharing through GeRaNIUM Grid Portal.*

As can be seen in Figure 5.10, collective resources were recommended to install maximum with 3 applications related to researchers' field. Those applications are accessible through grid portal. Users who have a credential could login to grid portal portlet which would allow them to access Jobs Portlet for using a job submission service. Validated user also could access to Credential Portlet to retrieve and renew a credential, Files Portlet to upload and download their files and Resource Browser Portlet to access collective resources. Resource Registry Portlet in GGP is an interface accessible by administrator to add resources available and add information of resources that will appear in Resource Browser Portlet for users. Besides, Resource Browser Portlet also retrieved resources' information from MDS (Figure 5.11). The Monitoring and Discovery System (MDS) is the information services component of the Globus Toolkit that provides information about the available resources on the Grid and their status (Globus, 2007). Through the FrontPage of grid portal guest may get the information of GeRaNIUM including services available, clusters specification, GeRaNIUM researchers' contacts and more but they are not allowed to access GeRaNIUM services (Figure 5.12).

As a result, GeRaNIUM Grid Portlet probably has a potential to be an initial manager to give a momentum to the growth of UM campus grid environment. The centralized management pursue in grid portal makes GeRaNIUM easy to manage and access by users without having to learn technical commands and process to use grid. By keeping a web based approach to GeRaNIUM interface, this might increase the interest from several users who are already familiar in using web pages to exploit applications and databases.



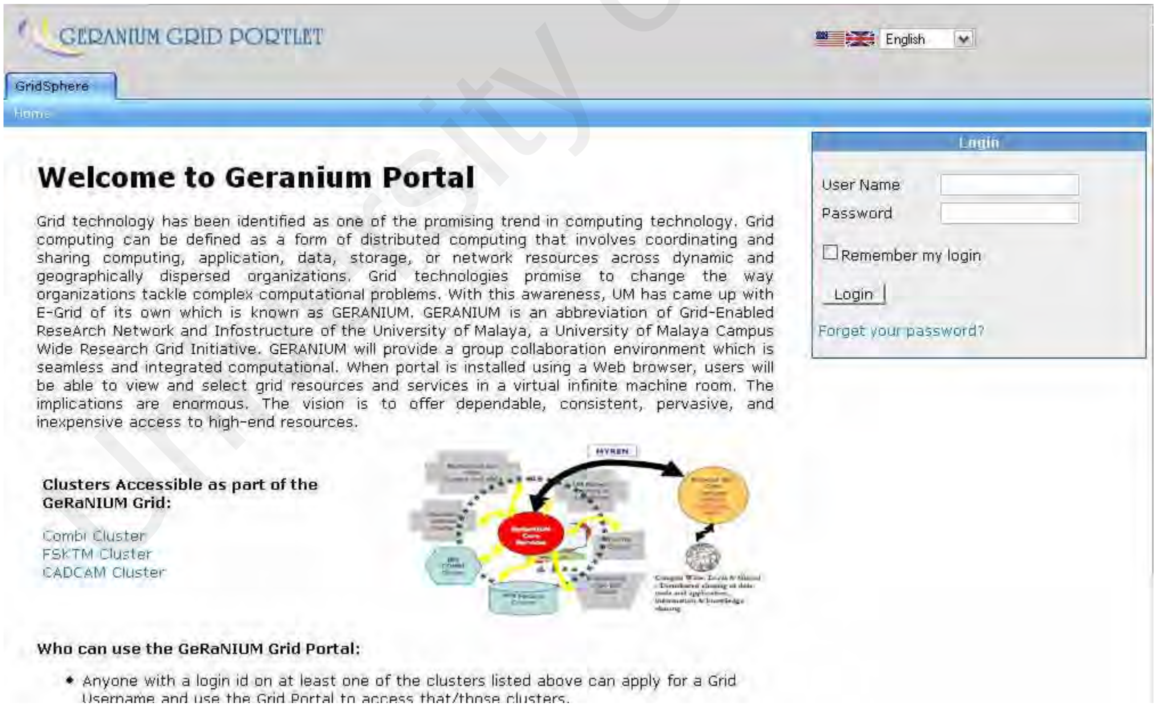
Combi Cluster

Hostname	combi.geranium.um.edu.my	**CPU Count**	1
Description	Combi Cluster	**CPU Cache**	256 X 12kb
Platform	i686	**CPU Speed**	1996Mhz
CPU Vendor	AuthenticAMD	**CPU Load 1 min**	100
CPU Version	6.8.1	**CPU Load 5 min**	100
CPU Model	AMD Sempron(tm) 2800+	**CPU Load 15 min**	100
CPU Features	fpu vme de pse tsc msr pae mce cx8 apic mtrr pge mca cmov pat pse36 mmx fxsr sse syscall mp mmxext 3dnowext 3dnow	**Memory Size**	1011MB
		Memory Free	807MB
		Virtual Memory Size	2047MB
OS Name	Linux	**Virtual Memory Free**	2047MB
OS Release	2.6.9-42.EL		

image image image

March 26, 2007

Figure 5.11 *Resource Browser Portlet views cluster specification.*



Welcome to Geranium Portal

Grid technology has been identified as one of the promising trend in computing technology. Grid computing can be defined as a form of distributed computing that involves coordinating and sharing computing, application, data, storage, or network resources across dynamic and geographically dispersed organizations. Grid technologies promise to change the way organizations tackle complex computational problems. With this awareness, UM has came up with E-Grid of its own which is known as GERANIUM. GERANIUM is an abbreviation of Grid-Enabled ReseArch Network and Infostructure of the University of Malaya, a University of Malaya Campus Wide Research Grid Initiative. GERANIUM will provide a group collaboration environment which is seamless and integrated computational. When portal is installed using a Web browser, users will be able to view and select grid resources and services in a virtual infinite machine room. The implications are enormous. The vision is to offer dependable, consistent, pervasive, and inexpensive access to high-end resources.

User Name

Password

☐ Remember my login

Login

Forget your password?

Clusters Accessible as part of the GeRaNIUM Grid:

Combi Cluster
FSKTM Cluster
CADCAM Cluster

Who can use the GeRaNIUM Grid Portal:

- Anyone with a login id on at least one of the clusters listed above can apply for a Grid Username and use the Grid Portal to access that/those clusters.

Figure 5.12 *The main page of GeRaNIUM Grid Portal.*

CHAPTER 6

CONCLUSION

6.1 Dissertation summary

The aim of the present study was to develop a grid portal for bioinformatics sequences alignment applications. In order to convince users to use parallel computing in grid environment, a comparison study was done to compare the performance of runtime process among workstation and cluster computing located in the grid environment. This study is useful to propose that parallel computing have a potential in providing a high-performance computing power which is better than workstation in several aspects.

During this study, GeRaNIUM's architecture was extended with locating Portal Server as a manager for GeRaNIUM and for hosting a grid portal. A first web-based interface called GeRaNIUM Grid Portal (GGP) for GeRaNIUM was successfully implemented on GeRaNIUM which provides user an interface to submit jobs and manage clusters easily.

In contrast from certain studies done in the performance evaluation of a sequence alignment processes, the present study considered a qualitative analysis on output produced by a single processing and parallel processing. The phylogeny tree and aligned sequence produced from both technologies were studied to determine the consistency of result. The result is to encourage researchers previously working on the sequence alignment in single processing to switch to a parallel processing which could give better performance with the same results as in a single processing.

6.2 Contribution and Findings

One of the contributions from this study is the enhancement of GeRaNIUM. This study has successfully located Portal Server to become the central of the campus grid certificates. Portal Server makes every machine attached to the campus grid to trust Portal Server certificate first before they are allowed to provide services to user. Portal Server also has been located as a platform for a grid portal development. Grid portal which was developed during this project could be utilized as the first user credential system in campus grid. In addition, Portal Server could monitor resources and user's job by using a web-based interface. Before users want to utilize resources in grid, they can apply for a credential from grid portal easily before they can access and submit job to resources.

Another finding in this project is the performance comparison between workstation and parallel computing in grid environment. The study investigated on the performance of running biological application namely ClustalW. The present study successfully showed that the development of grid environment will lower expenses of having a higher performance computing power which are previously available in workstation. The development indicated that grid environment has a potential to be expanded in University of Malaya that would encourage sharing of inexpensive computing power, storage systems, data sources, applications, visualization devices, scientific instruments, sensors and human resources over the network and distributed location. By using parallel computing with grid environment, users may not have to buy new resources when doing bigger tasks. They can just add new CPUs to existing resources. This will not only dissipate resources available but also decreases expenses to upgrade computing performance rather than expense for higher-performance workstation. Moreover, parallel computing would decrease the process runtimes by distributing a big job to a number of processors into a smaller task. It is to be hoped that the initiative of this new technology would also attract the involvement of researchers from the fields of humanities, arts and social sciences in the university.

Lastly, a significant finding is related with the performance evaluation on single and parallel processing doing a sequences alignment process. During this study, computing performance can be accelerated by parallelizing job into a small task in a cluster. This study also found that results produced by parallel processing using ClustalW-MPI were acceptable and reliable, thus; hopefully will throw away taxonomic worries about using parallel computing among researchers. This will also influence them to use cluster and grid computing technology with confidence. In addition, this study also illustrated that the alignment process in ClustalW was handled consistently whether the process was in a single processing or in a parallel processing.

6.3 Limitations

Several limitations of the present study should be noted. Firstly, in the development of Portal Server, there is a limitation with CentOS operating system. CentOS do not allow Portal Server located in two environments which are campus grid environment and Malaysian Research & Education Network (MyREN). CentOS operating system only allows one gateway in one time although Portal Server has two network connections. In the proposed architecture, Portal Server was planned to be connected to MyREN which could provide a better network bandwidth and possibly will be connected to National Grid environment. The other problem faced in the GeRaNIUM's architecture enhancement was, some researchers were not willing to share their resources because they were worried that their resources might be overloaded after attaching with grid. Consequently, this study had limited resources to test.

Furthermore, the other problem faced during this study is on searching of software compatibility. During the implementation of GeRaNIUM Grid Portal (GGP), software compatibility needs to be considered due to the utilization of open source software. In the installation of Gridsphere and Gridportlet it was found that Gridpshere version 2.1 which is compatible with Globus version 4.1 is also compatible with Gridportlet version 1.3. The limitation had caused this study to drag longer because a lot of time was spent to test on software compatibility.

In conclusion, despite limitations described above, this study contributes to the on-going literature on individual's creativity in organization and provides support in interactive approach.

6.4 Potential future enhancement and research.

Grid portal development is still on going and needs more improvements and additional functions with more intuitive and user-friendly graphical interfaces especially within File Browser Portlet and Job Submission Portlet. It is to be hoped that Grid portal will allow users to view graphical data such as 3D model or animation model using Java Applet. Users might not need to install viewer application to view their data. Furthermore, it is also suggested that GeRaNIUM needs to be placed on a dedicated network to avoid network traffic when running a job and view 3D data as well.

This study was focused only on the bioinformatics sequence alignment application, ClustalW to compare the performance of single processing (workstation) and parallel processing (cluster computing). It would be interesting to test Grid Portal and GeRaNIUM too with different application from other field. Additional portlets provide more services and applications can be added if necessary.

The performances studied in this thesis were focused on several aspects which are the performance between workstation and cluster computing, and the performance of cluster with the increment of processors. During the making of this thesis, I had also tried to study the performance of cluster computing influenced by different problem sizes. The problem sizes were considered at the length of sequences and the number of organisms to be aligned. From the result, it seems that ClustalW worked more efficiently on the samples with shorter length of sequences without considering the number of organisms to be aligned. However, due to the limited data this study needs to be further studied using more samples of data with different proportions to prove that the efficiency of sequence alignment applications in cluster computing is influenced by the length of sequences or by the number of organisms.

As for the future, it would be exciting to find out the performance of grid computing in a larger place such as within a faculty or in a different campus and looking at the issues on implementing grid in a larger environment.

REFERENCES

(Akram et al., 2005) A. Akram, D. Chohan, X.D. Wang, X. Yang and R. Allan, “A Service Oriented Architecture for Portals Using Portlets”, *UK e-Science AHM2005*, Nottingham, UK, 2005, accepted.

(Andronico et al., 2003) G. Andronico, R. Barbera, A. Falzone, G. Lo Re, A. Pulvirenti, A. Rodolico, The GENIUS web portal: grid computing made easy, 2003, Proceedings of the International Conference on Information Technology: Computers and Communications (ITCC.03)

(Apweiler et. al, 2004) R. Apweiler, *et al*, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Res.,* 32, D115-9, 2004.

(Bodlaender, 1994) Bodlaender, H.L., Downey, R.G., Fellows, M.R. 1994. *Application Parameterized Complexity to Problems of Parallel and Distributed Computing*. Unpubl. Abstract.

(Brignac, 1997) Stafford Brignac, 1997, “Sequencing Support System: A Robotic System for Processing DNA Samples”, IEEE Engineering in Medicine and Biology.

(Buyya et al., 2005) Buyya R, Gibbins HA, Nadiminti K, Chhabra R, Beeson B & Smith B. 2005. The Australian BioGrid Portal: Empowering the Molecular Docking Research Community. In R Chhabra & J Young (eds), *Proceedings of the APAC Conference and Exhibition on Advanced Computing, Grid Applications and eResearch*. 1-18. Canberra, ACT, Australia: Australian Partnership for Advanced Computing.

(Cesati, 1998) Cesati, M., Ianni, M.D. 1998. *Parameterized Parallel Complexity*. In Proceedings of the 4th International Euro-Par Conference, 892-896.

(Chao-Tung Yang et al., 2004) Chao-Tung Yang, Chuan-Lin Lai. 2004, *Apply Cluster and Grid Computing on Parallel 3D Rendering*, IEEE International Conference on Multimedia and Expo (ICME).

(Charles, 2003) Charles, S., 2003, *Integrating Grid Capabilities into the CHEF Collaborative Portal Framework,* Technical Report NEESgrid-2003-01.

(Cheetham et al., 2003) Cheetham, J., Dehne, F., Rau-Chaplin,R., Stege, U., Taillon, P.J. 2003. *A Parallel FPT Application For Clusters*. Proceedings of the 3rd IEEE/ACM International Sysposium on Cluster Computing and the Grid.

(Digipede, 2003) *Grid and Cluster Computing: Options for Improving Windows Application Performance. www.digipede.net*. 2003. Digipede Technologies.

(D-Grid Initiative, 2007) D-Grid Program, Retrieved: 1/10/2007. From: *http://www.d-grid.de/*

(Downey, 1999) Downey, R.G., Fellows, M.R., Stege, U. 1999. *Parameterized Complexity: A Framework for Systematically Confronting Computational Intractability*. AMS-DIMACS Processdings, Vol.49, AMS, 49-99.

(Evard, 1999) R. Evard. Chiba city. In *Extreme Linux Workshop, USENIX Annual Technical Conference,* Monterey, California, June 8-11 1999.

(Foster et al., 1999) Foster, I. and Kesselman, C. 1999. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufman Publishers.

(Foster et al., 1997) Foster, I. and Kesselman, C. 1997. *Globus: A metacomputing infrastructure toolkit. Supercomputer Applications*. 11(2): 115-128, 1997.

(Foster, 2002) Foster, I. 2002. *What Is The Grid? A Three Point Checklist*, Daily News and Information For The Global Grid Community, Vol. 1, No. 6, pp. 8.

(GeneBank Data Statistic,2007) GeneBank Data Statistic, Retrieved: 2/3/2007. From: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html.

(Geranium, 2006) Geranium. 2006. *Grid infrastructure for scientific computing: Towards development of a campus wide research grid for the University of Malaya*. Unpubl. report GeRaNIUM First Year Report

(Globus, 2007) About the Globus Toolkit, available from: *http://www.globus.org/toolkit/about.html* [Accessed 3 March 2007].

(Gracanin, 2005) Denis Gracanin: *An Approach to Distributed Interactive Simulation and Visualization of Complex Systems using Cluster Computing.* 2005. CSB Workshops 2005: 292-298

(GridSphere Portal Framework, 2007) GridSphere Portal Framework, Retrieved: 5/10/2007.From: http://www.gridsphere.org

(Halstead et. al, 1999) D. M. Halstead, B. Bode D. Turner, and V. Lewis. Gigaplantscalale cluster. In *Extreme Linux Workshop, USENIX Annual Technical Conference,* Monterey, California, June 8- 11 1999.

(Hey, 2002) Hey, T., 2002, *Unlocking the Power of The Grid*, IEE Review.

(Hsun-Chang et al., 2005) Hsun-Chang Chang, Kuan-Ching Li, Yaw-Ling Lin, Chao-Tung Yang, Hsiao-His Wang, Liang-The Lee. 2005. *Performance Issues of Grid Computing Based on Different Architecture Cluster Computing Platforms*. Proceedings of the 9th International Conference on Advanced Information Networking and Applications.

(Huelsenbeck et. al, 2001) John P. Huelsenbeck, Fredrik Ronquist,Barry Hall, 2001, *MrBayes: A program for the Bayesian inference of phylogeny,*Department of Biology, University of Rochester, Rochester, NY 14627, U.S.A.

(Joshua Harr et. all, 2002) Joshua Harr, Greg Denault, 2002. *Issues Concerning Linux Clustering: Cluster Management and Application Porting.* Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS.02)

(Kanz et. al, 2005) C. Kanz, P. Aldebert, N. Althorpe *et al*, "The EMBL Nucleotide Sequence Database," *Nucleic Acids Res.,* vol. 33 Database Issue, pp. D29-33, Jan 1. 2005.

(Karonis et al., 2003) Karonis, N., Toonen, B., and Foster, I. 2003. *Mpich-g2: A grid-enabled implementation of message parsing interface. Journal of Parallel and Distributed Computing*, 63(5): 551-563.

(Kelly et al., 2005) Kelly, N., McCurley, M., McKee, S. 2005. *The Genegrid Portal: A User Interface for a Virtual Bioinformatics Laboratory.* Proceeedings of UK e-Science All Hands Meeting.

(Kocher, 1989) Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Paabo, A., Villablanca, F. X. & Wilson, A. C. 1989. *Dynamics of mitochondrial DNA evolution in animals: ampli¢cation and sequencing with conserved primers*. Proc. Natl Acad. Sci. USA 86, 6196-6200.

(Korean National Grid,2007) Korean National Grid Program, Retrieved: 1/10/2007. From: *http://gridcenter.or.kr/*

(Kuo-Bin Li, 2003) Kuo-Bin Li. 2003. *ClustalW-MPI: ClustalW Analysis Using Distributed and Parallel Computing* , Bioinformatics, 19(12), 1585--1586.

(Kuykendall et. al.,1984) Frank Kuykendall, Philip M. Zion, 1984, "The Pilot Ocean Data System Science Workstation", IEEE.

(Lambert et al., 2006) Lambert, Tan, Turner, Gayle, Prandy, Sinnott (2006) "Developing a Grid Enabled Occupational Data Environment", Paper presented to the 2nd International Conference on eSocial Science, Manchester, 28-30 June

(Lipman et. al, 1989) D. J. Lipman, S. F. Altschul, and J. D. Kececioglu, 1989 *A tool for multiple sequence alignment*, Proc. Nail. Acad. Sci. USA, vol. 86, pp. 4412–4415.

(List of Sequence alignment software- Wikipedia, 2007) List of Sequence Aligment Software- Wikipedia the free encyclopedia. Retrieved: 30/9/2007 From: *http://en.wikipedia.org/wiki/Sequence_alignment_software*

(MyProxy, 2007) MyProxy Credential Management, Retrieved: 1/4/2007. From: *http://grid.ncsa.uiuc.edu/myproxy/*

(Notredame et. al, 2000) C. Notredame, D. G. Higgins, and J. Heringa, 2000, *Tcoffee:A novel method for fast and accurate multiple sequence alignment*, IDEAL J. Mol. Biol. (2000), vol. 302, pp. 205–217.

(Novotny et. al, 2004) J. Novotny, M. Russell, O. Wehrens, 2004, *GridSphere: An Advanced Portal Framework*, Proceedings of EuroMicro Conference (2004), 412-419

(Por, 2006) L.Y. Por, M.T. Su, T.C. Ling, C.S. Liew, T.F. Ang, K.K. Phang. Issues of Establishing a Campus-wide Computational Grid Infrastructure in the GERANIUM Project. *Proceedings of the International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, 2006.*

(Putchong, 2000) P. Uthayopas, T. Angsakul, and J. Maneesilp. System management framework and tools for beowulf cluster. In *Proceedings of HPCAsia2000*, Beijing, May 2000.

(Reisen et. al, 1999) R. Reisen, R. Brightwell, L. A. Fisk, T. Hudson, and J. Otto. Cplant*. In *Extreme Linux Workshop, USENIXAnnual Technical Conference,* Monterey, California, June 8-1 1 1999.

(Rocks, 2004) Rocks Cluster Distribution: Award Winning Open Source High Performance Linux Cluster Solution, http://rocks.npaci.edu/rocks, November 30, 2004

(Sachin et. al, 2005) Sachin Wasnik, Mark Prentice, Noel Kelly, P.V. Jithesh, Paul Donachy, Terence Harmer, Ron Perrott, Mark McCurley, Michael Townsley, Jim Johnston, Shane McKee, 2005, *Resource Monitoring and Service Discovery in GeneGrid,* UK e-Science All Hands Meeting (AHM 2005), Nottingham, Sep 19-22, 2005.

(Savvas et. al, 2004) Savvas, I.K., Kechadi, M-T. 2004. *Dynamic Task Scheduling in Computing Cluster Environments*, Proceedings of the ISPDC/HeteroPar'04 IEE.

(Schmollinger et. al, 2004) M. Schmollinger, K. Nieselt, M. Kaufmann, and B. Morgenstern, 2004, *Dialign p: Fast pair-wise and multiple sequence alignment using parallel processors*, BMC Bioinformatics 2004, vol. 5, no. 128.

(Simmossis et. al, 2005) V. A. Simossis and J. Heringa, 2005 ,*Praline: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information*, Nucleic Acids Research, 2005, vol. 33, pp. 289–294.

(The Apache Tomcat 5.5 Servlet/JSP Container, 2007) The Apache Tomcat 5.5 Servlet/JSP Container, Retrieved: 5/10/2007.From: *http://tomcat.apache.org/tomcat-5.5-doc/index.html*

(Thomas et. al, 2001) M. Thomas, J. Boisseau, S. Mock, M. Dahan, K. Mueller, D. Sutton, 2001, *The GridPort Toolkit Architecture for Building Grid Portals,* Proceedings of the 10th IEEE Intl. Symp. on High Perf. Dist. Comp.

(Thompson et. al, 1994) J.D. Thompson, D.G. Higgins, and T.J. Gibson, 1994, *Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice*, Nucleic Acids Research 22, vol. 22, no. 22, pp. 4673–4680.

(UK National E-Science Centre,2007) UK National E-Science Centre, Retrieved: 1/10/2007. From: *http://www.nesc.ac.uk/*

(Vazquez-Poletti et. al, 2007) Vázquez-Poletti, J.L., Huedo, E., Montero, R.S., Llorente, I.M.: *Workflow Management in a Protein Clustering Application*.2007. In: Proc. 5th Intl. Work. Biomedical Computations on the Grid (BioGrid 2007). 7th IEEE Intl. Symp. Cluster Computing and the Grid (CCGrid 2007), pp. 679–684. IEEE Computer Society Press, Los Alamitos

(Workstation – Wikipedia, the free encyclopedia, 2007) Workstation – Wikipedia, the free encyclopedia. *http://en.wikipedia.org/wiki/workstation*. Last modified: Feb 20 2007

(Zhongwu Zhou et al., 2005) Zhongwu Zhou, Feng Wang, Billy D. Todd, 2005, *Development of Chemistry Portal for Grid-enabled Molecular Science*, IEEE Proceedings of the First International Conference on e-Science and Grid Computing (e-Science'05).