

Analysis of Jobs in a Multi-Organizational Grid Test-bed

Bu-Sung Lee¹, Ming Tang¹, Junwei Zhang¹, Ong Yew Soon¹,
Cindy Zheng², Peter Arzberger², David Abramson³

¹*School of Computer Engineering, Nanyang Technological University, Singapore*

²*San Diego Supercomputer Center, USA*

³*Faculty of Information Technology, Monash University, Australia*

{ebslee, tangm, jwzhang, asysong}@ntu.edu.sg, {zhengc, parzberg}@sdsc.edu,
david.abramson@infotech.monash.edu.au

Abstract

The inevitable move from a single large scale server to a distributed Grid environment is beginning to be realized across international Grid test-bed like Pacific Rim Applications and Grid Middleware Assembly (PRAGMA). Although jobs submitted to a single large server have been widely analyzed, job characteristics in a Grid environment are different as we found in our analysis of jobs submitted in the PRAGMA Grid test-bed. This paper reports on the analysis of jobs submitted across the PRAGMA Grid test-bed. The job types are categorized and the runtime of jobs is captured, using the Multi-Organization Grid Accounting System (MOGAS), and analyzed. The number of jobs submitted across organizations, indicating the level of resource sharing among participants, is also captured by the system.

1. Introduction

The Grid technology has been in the headline for the past decade with government been a major driver through research programs, e.g. TeraGrid [2] in the USA, China Grid [3], and e-Science program in UK [4]. These government funding has reach \$100s million and is usually a mix of application focus and/or building up an infrastructure, e.g. TeraGrid.

Major multinational companies, e.g. IBM, SUN Microsystems and Oracle, have also labeled the Grid computing as one of the key technologies in their companies. This has resulted in great advancement of the Grid technology into the industry. In the area of standardization Global Grid Forum (GGF) [17] has certainly played a critical role in establishing best practices for Grid implementers and practitioners.

An important part of the development was the setting up of international Grid test-bed. There are certainly a number of international test-bed, e.g. CERN LHC EGEE

[5] and PRAGMA [1]. We have chosen PRAGMA for our study due to the heterogeneous nature of the Grid resources, software and the international participation in the test-bed.

In this paper we will first introduce the PRAGMA community. In Section 3, the multi-organization Grid accounting tool, MOGAS, which is deployed across PRAGMA will be briefly described. Section 4 will report on the analysis of the job data captured across the PRAGMA test-bed and our major finding. The paper will conclude with Section 5.

2. PRAGMA

Pacific Rim Applications and Grid Middleware Assembly (PRAGMA) was formed in March 2002, led by SDSC and has grown in leap and bounce during the last few years. Currently it has established an international Grid test-bed as shown in Figure 1. The test-bed sites are connected to each other via international and national research and education networks, e.g. Abilene [18], SingAREN [19], AARnet [20], NICT [21], and APII [22] network links.

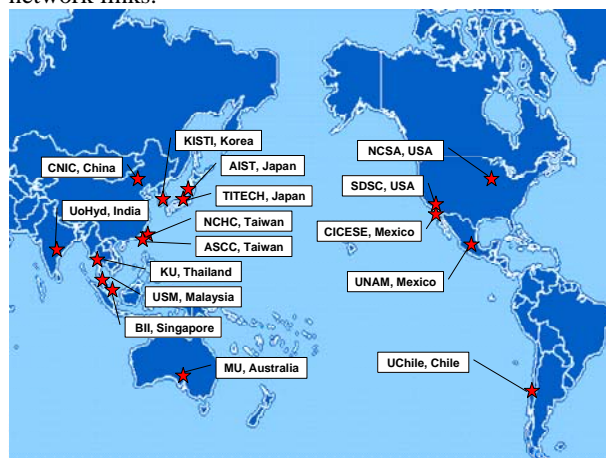


Figure 1. PRAGMA Grid Test-bed.

Table 1. PRAGMA Grid resources with MOGAS installed.

Organization Name	Country	Host Name	# Nodes	# CPUs	Memory (GB)	Job Manager	CPU Model	CPU Speed (MHz)
AIST	Japan	ume	33	66	34.460	SGE	i686	1396
KISTI	Korea	jupiter	17	17	15.103	PBS	i686	1694
NCSA	USA	tgc	13	52	26.169	PBS	i686	2400
SDSC	USA	rocks-52	16	62	32.493	SGE	i686	2388
SDSC	USA	rocks-47	4	4	8.140	SGE	ia64	900
TITECH	Japan	gsic-presto	9	9	7.566	PBS	i686	1195
UNAM	Mexico	malicia	6	6	3.510	PBS	i686	1894
USM	Malaysia	hawk	17	17	8.700	SGE	i686	2800

The goals of the community are as follows:

- Establish sustained collaboration.
- Advance the use of Grid technologies for applications among the community of investigators working with leading institutes around the Pacific Rim.
- Working closely with established activities that promote grid activities or the underlying infrastructure, both in the Pacific Rim and globally.

The resources that are currently available on the test-bed are varied. Table 1 sums up some of the resources that are currently shared by the community and running MOGAS.

There are number of middleware technologies and application that are deployed and tested across the PRAGMA test-bed [9]. Some of the major application deployments are as follows:

- mpiBLAST [14]. The Basic Logical Alignment Search Tool (BLAST) [15] is a DNA and protein sequence alignment tool, and it is developed by National Center for Biotechnology Information (NCBI). mpiBLAST is a parallelization of BLAST, and it segments the BLAST database and distributes it across cluster nodes to enable BLAST queries to be processed on many nodes simultaneously.
- QMMD [16]. QMMD enables large scale simulation with quantum accuracy. It combines classical molecular dynamics (MD) simulation with quantum mechanical (QM) simulation. QMMD is implemented with Ninf-G2 [12].
- Savannah Forest Fire analysis application from Monash University. This application studies

Savannah fire impact on northern Australian climate. It is based on Nimrod/G [13].

- MM5 Climate Simulator aims to generate regional climate change scenarios over southern South America to improve information on the climate conditions under which nature and human systems will be exposed. It studies the sensitivity of climate to surface conditions, mainly soil moisture and sea surface temperatures, in particular, their impact on precipitation anomalies over southern South America.

3. MOGAS

In the Grid environment where resources are been shared among collaborating organization, it is essential that we are able to log and capture the level of usage across multi-organization. This forms the impetus for the development of a monitoring and accounting tool, Multi-Organization Grid Accounting System (MOGAS) [11].

MOGAS was build on top of the Globus core: Grid File Transfer Protocol (GridFTP), Grid Resource Allocation Manager (GRAM) and Grid Security Infrastructure (GSI). At each of the grid computing resource, Globus core calls are used where possible and wrappers were developed where necessary.

MOGAS tasks are broken up into a number of functions:

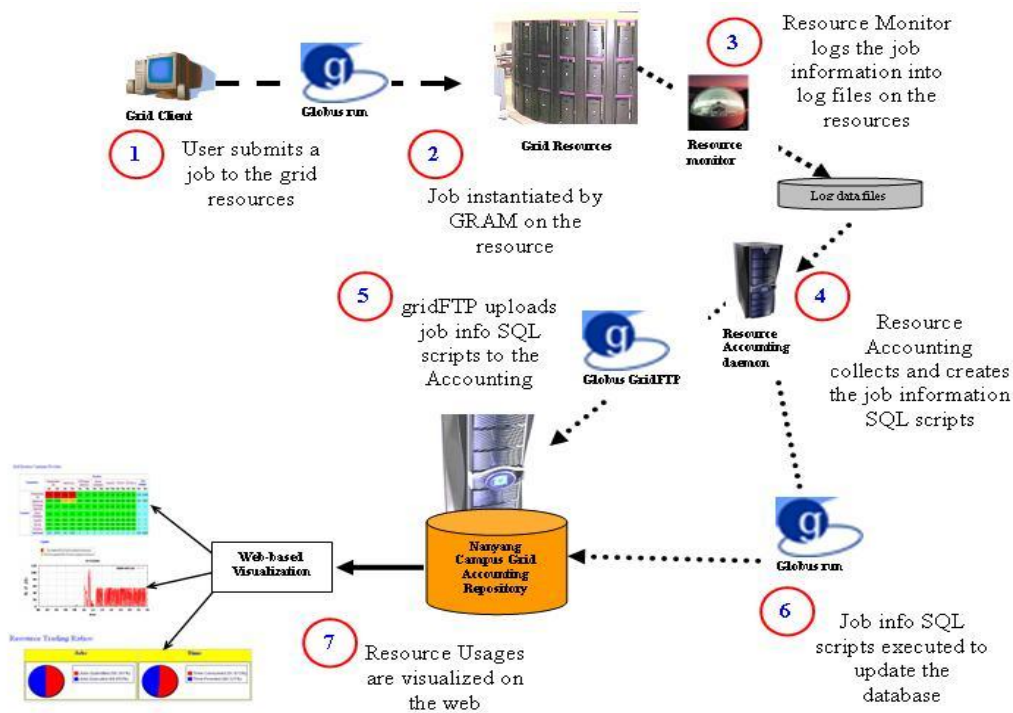


Figure 2. Workflow overview of the MOGAS

- Identification of user, organization, virtual organization. One of the key issues is the identification of the organization. There has been deployment where each user is given a specific account. However, in a large Grid environment it would not be practical as a large number of resources that need to be update whenever there is a change in the user database. In MOGAS we use the security certificate, part of GSI, to determine the consumer organizations, and the account name to determine the virtual organization. The consumer organization is responsible for giving out the certificate and thus the users can access the PRAGMA resources.
- Monitor and update resource usage. MOGAS is required to record the resources used by each job. Globus core does not provide any facility to monitor the resource usage. Two modules were developed on top of Globus services
 - RMon to monitor and log the resource usage. RMon talks to the scheduler. Presently, three schedulers are supported by MOGAS: LSF [7] from Platform computing, Sun Grid Engine (SGE) [23] from SUN Microsystem Pte Ltd and PBSPro [8].
 - MOGAS daemon to update the database with job usage information.
- Organization of resource usage database. An SQL database is used to store the information. Tables were created to ease access.
- User Web access and visualization. A portal was developed for user access to the job information. A simple web-base Java Servlet was used for the graphic display of job information.

Figure 2 shows the workflow of MOGAS and is explained as follows:

- 1) User submits a Globus job to the Grid resources.
- 2) The GRAM gatekeeper at the Grid resources, will instantiate the job.
- 3) The Resource Monitor (RMon) which is deployed on the gatekeeper will then log the job information, including the Grid user id, the job name and the start and end time onto the job log files.
- 4) The MOGAS daemon deployed on the Grid resources will periodically scan the accounting log file directories and create the SQL scripts for updating the accounting database. When the log files are updated into the DBMS, the log files will be renamed and updated.
- 5) After creating the SQL script, the daemon will utilizes GridFTP to upload the script onto the DBMS server.

- 6) Next, the daemon will initiate a Globus job run to execute the SQL scripts and update the repository.
- 7) Upon user request, the results are presented on the web using the visualization tools. Users can view a number of information about the resource usage and job workload in PRAGMA.

MOGAS provides users with the following information:

- **Overall Consumer/Provider Usage.** The consumer/provider matrix view presents the usage of the Grid resources based on one-to-one relationship between a providing organization and a consuming organization.
- **Specific Grid Organization Consumer/Provider Usage View.** This view enables users to look at a specific organization contribution and consumption of resources.
- **Resource Workload View.** This view shows a plot of the number of concurrent jobs per day, to provide an idea of the workload of the resources in the organization.
- **Accounting Report View.** Some of the available reports are: the transaction history for specific organization or specific project; most frequently executed jobs; and the inter-organizational transaction summary.

4. Analysis of Grid Jobs

The PRAGMA test-bed provides an ideal Grid environment for us to analyze the job characteristics as well as distribution across multiple organizations. MOGAS was deployed across a subset of the test-bed covering 7 organizations. The Grid job characteristics are captured by MOGAS and visualized in [10]. The following is an analysis of the data captured over a period of 3 months, from 1 Sept'2005 till 25 Nov'2005.

4.1. Consumer/Provider Statistics

Figure 3 shows the distribution of the number of jobs submitted by each organization. Thus, based on the records there are 10 consumer organizations. As seen in the figure, most of the jobs are submitted by Monash University, for the Savannah Forest Fire application. The second highest number of jobs is from SDSC, and this is expected as they host the Grid Operation Center for PRAGMA. Most of the jobs from SDSC is for monitoring and ensuring that the Grid test-bed is up and running.

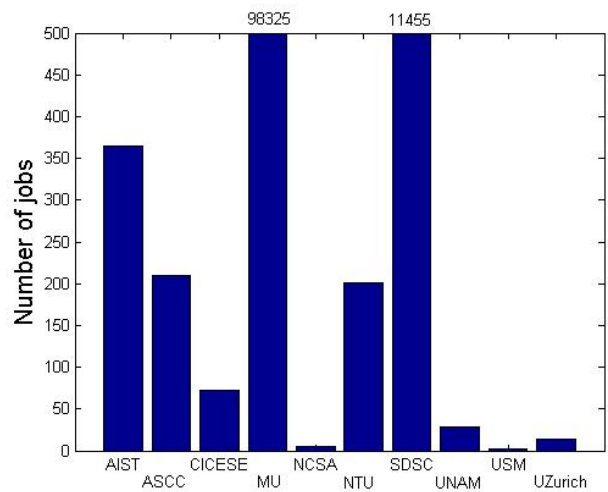


Figure 3. The number of jobs submitted by each organization.

Figure 4 shows the distribution of the number of jobs that have run on each organization. In total there are 7 resource provider organizations. Most of the jobs are run on AIST, NCSA and SDSC. The clusters at these organizations have been the forerunners in the installing of middleware and applications running on the test-bed.

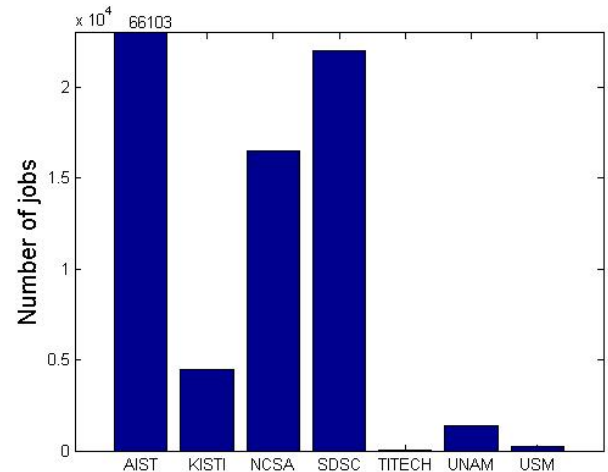


Figure 4. The number of jobs that have run on each organization.

To understand how organizations on the test-bed are sharing/using resources across the test-bed we proposed to use a new index called *Provider Diversity* (PD) index, which is defined as:

$$PD\ index = \frac{1}{M} \sum_{i=1}^M \frac{P_i}{N}$$

where P_i is the number of providers for consumer organization i , N is the total number of provider organizations, and M is the total number of consumer organizations.

A high PD index indicated a test-bed where users in the Grid environment are using multiple resource providers. Taking the information obtained from MOGAS, there are 7 provider organizations and 10 consumer organizations. Note that an organization can be both a consumer and a provider of job service. Table 2 shows the data of how many provider organizations have been used by the consumers.

Table 2. Consumer organization and the number of provider organizations.

Consumer organization (i)	Number of Provider organizations (P_i)
AIST	4
ASCC	3
CICESE	1
MU	5
NCSA	1
NTU	7
SDSC	7
UNAM	1
USM	1
UZurich	1

According to the data shown in Table 2, we can get the PD index:

$$PD\ index = (31/7)/10 = 0.44$$

Thus, on average an organization use about half of the number of providers in the Grid environment of PRAGMA.

4.2. Job Characteristics

It is important for us to understand the characteristics of the jobs submitted across the Grid test-bed.

The distribution of job runtime for the PRAGMA test-bed is shown in Figure 5. As seen in Figure 5, the duration of majority of the jobs are short, less than 15 seconds. However, it is noted that some jobs has a runtime of more than 50 hours.

As noted there are three peaks in the distribution for jobs less than 15 seconds. An analysis of the peaks shows the following:

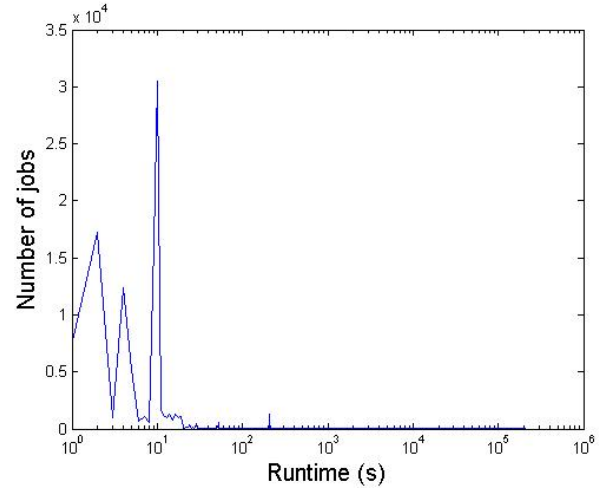


Figure 5. Job vs. runtime distribution

1. The peak at runtime 10 seconds. Most of these jobs are globus-url-copy, and they are mainly submitted from Monash University (MU) for the Savannah application. These GridFTP jobs copy data file from the data servers at MU to local disks and vice versa across the Grid test-bed. This is because the simulation is distributed across multiple organizations and the computation is time-step, at each time-step the simulation results are copied to the MU and data for the next time-step is copied to the respective sites/organization. As the size of the data file is small, as only a portion of data are transferred at a time, the latency of network connection setup and GSI authentication become the major contributors to the data access time. Therefore, the jobs' runtime is very consistent. A snapshot of GridFTP and Nimrod jobs of Savannah submitted by MU in a randomly selected hour is shown in Figure 6. From it we can see that the data transfer jobs and Savannah jobs were submitted alternately at intervals.
2. Peak at runtime 4 seconds. This peak is mainly caused by the "ls" jobs and "/bin/sh /etc/nimrodresourcecost.conf" jobs from MU.
3. Peak at runtime 2 seconds. This peak is mainly caused by the "mkdir" jobs submitted by MU. The "mkdir" jobs from MU create the job running directory.

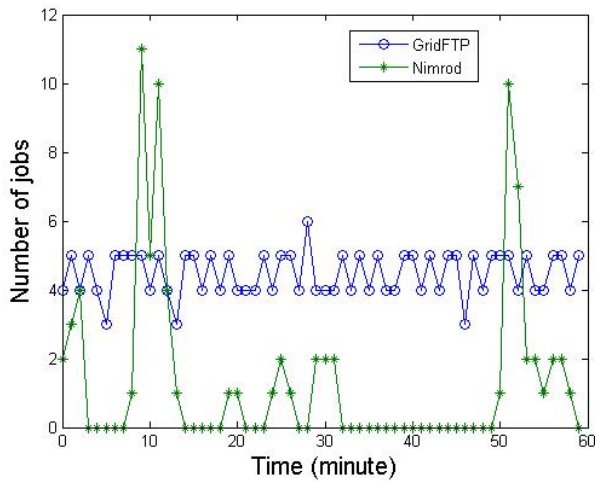


Figure 6. A snapshot of GridFTP and Nimrod jobs submitted by MU in an hour.

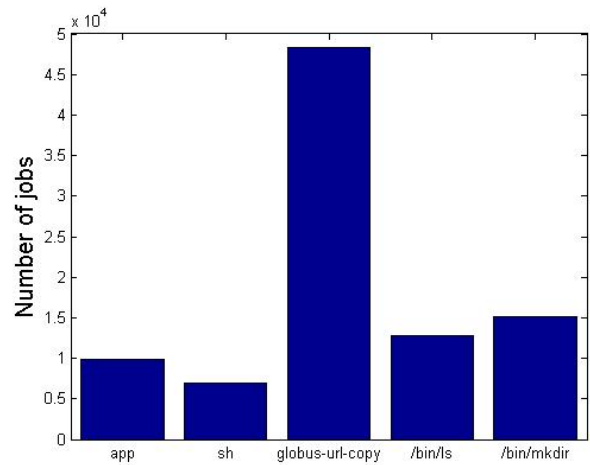


Figure 7. The types of jobs submitted by MU.

Further analysis is done for the jobs submitted from MU. The jobs from MU include the computing application and environment setup. The number of jobs for different types is shown in Figure 7. The job types are:

- “app” is the Savannah Forest Fire application job;
- “sh” stands for the command of “/bin/sh /etc/nimrodresourcecost.conf” supposed to do system configuration;
- “globus-url-copy” is for remote data copies, and “/bin/lis” and “/bin/mkdir” are used for the application environment setup or test.

The application jobs only account for 11% of the total submitted jobs, while most of the jobs are for computing environment setup and test. As expected the largest number of job types is the GridFTP as the application has large amount of data to be transferred to the resource location.

However, using the number of jobs can be deceiving when trying to optimize the environment. The accumulated runtime for the jobs submitted from MU is shown in the Figure 8. The figure certainly differs greatly from Figure 7. In Figure 8, the accumulated runtime of the application jobs, where some jobs runs for more than 50 hours, is significantly higher than the accumulated runtime of the other jobs.

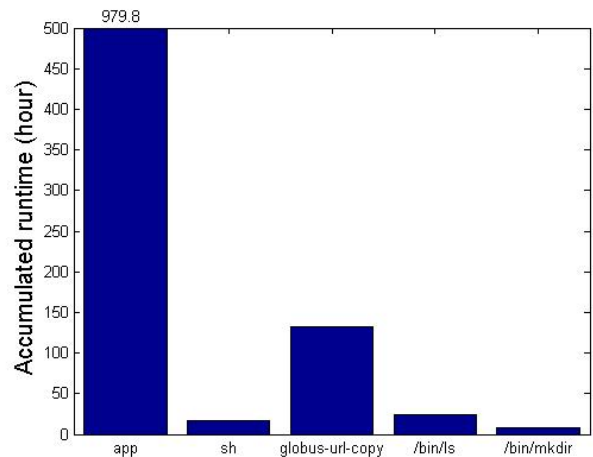


Figure 8. The accumulated runtime for each type of jobs submitted by MU.

Similar results were also obtained from the other application, i.e. QMMD and mpiBLAST, where majority of jobs are for setting up the environment and transfer of data to the respective computer cluster.

The characteristics of jobs submitted by SDSC are different from other organizations. About 84% jobs from SDSC run the commands of “/bin/echo”, “/bin/hostname”, and “/usr/bin/cal”. They are basically jobs with commands to test and check that the clusters are up and running. This is because the Grid Operation Center for PRAGMA is located at SDSC.

The distribution of the number of running jobs in the PRAGMA test-bed at different time of one day is shown in Figure 9. From this figure, we can find that the peak time is from 13PM to 18PM in one day, while the lull time is from 3AM to 8AM. This is because most of the application jobs are submitted from MU, which is in GMT+10 time zone.

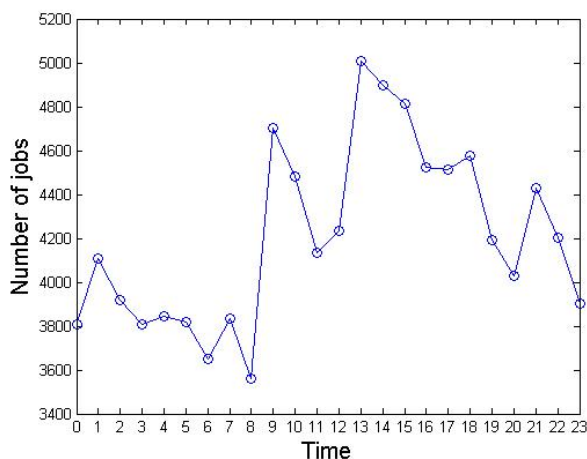


Figure 9. The distribution of the running jobs at different time (GMT+8).

5. Conclusion

The data obtained from the job log of PRAGMA using MOGAS has given us a better understanding on how the resources are used across the test-bed. It has highlighted the nature of the jobs and the level of resource sharing across the test-bed.

For Grid environment there is a necessity to monitor status of the different clusters. In this case it was done by SDSC which hosts the PRAGMA test-bed Grid Operation Center. An additional observation is that there are a lot of environment settings and uploading of data across multiple organization before an application is executed, e.g. the Savannah program. This differs from large scale single servers where the data is already on the server or need to be loaded only once. AIST, SDSC and NCSA clusters are widely used by members of PRAGMA. This is because of the software and infrastructure support that is readily available at the site.

The PRAGMA Grid test-bed has certainly provided us with a rich set of job information. It has highlighted the differences between single large server submission and job submission in a Grid environment. Furthermore, it has also highlight the great need to develop tools to make the job submission more transparent to the application user.

Acknowledgement

The authors would like to acknowledge the support given by all PRAGMA members who have been forthcoming in the deployment of MOGAS. This project is funded by National Grid Office of Singapore.

Reference

- [1] Pacific Rim Applications and Grid Middleware Assembly (PRAGMA), <http://www.pragma-grid.net/>.
- [2] The TeraGrid Project, <http://www.teragrid.org/>.
- [3] China National Grid Project (CNGrid), <http://www.cngrid.org/>
- [4] e-Science, <http://www.rcuk.ac.uk/escience/>
- [5] The Enabling Grids for E-sciencE (EGEE) project, <http://public.eu-egee.org/>.
- [6] The Global Grid Forum (GGF), <http://www.gridforum.org/>.
- [7] Load Sharing Facility (LSF), <http://accl.grc.nasa.gov/lsf/>.
- [8] PBSPro, <http://www.altair.com/software/pbspro.htm>.
- [9] Cindy Zheng, David Abramson, Peter Arzberger, et al., "The PRAGMA Testbed: Building a Multi-Application International Grid", International Workshop on Test-bed held in conjunction with International Symposium on Cluster Computing and Grid 2006 (CCGrid 06), 16-19 May 2006.
- [10] MOGAS for PRAGMA, <http://www.ntu-cg.ntu.edu.sg/pragma/>
- [11] Hee-Khiang Ng, Quoc Thuan Ho, Junwei Zhang, Dudy Lim, Yew Soon Ong, "Nanyang Campus Inter-organisation Grid Monitoring System", Proceedings of GridAsia Workshop on Grid Computing & Applications 2005, ISBN 981-05-3055-2, pp. 118-127.
- [12] Ninf, <http://ninf.apgrid.org/>.
- [13] Nimrod: Tools for Distributed Parametric Modelling, <http://www.csse.monash.edu.au/~davida/nimrod/>.
- [14] mpiBLAST, <http://mpiblast.lanl.gov/>.
- [15] The Basic Local Alignment Search Tool (BLAST), <http://www.ncbi.nlm.nih.gov/blast/>.
- [16] QM-MD simulation, <http://pragma-goc.rocksclusters.org/applications/qmmd/overview.html>.
- [17] Global Grid Forum (GGF), <http://www.gridforum.org/>.
- [18] Advanced Networking for Leading-edge Research and Education (Abilene), <http://abilene.internet2.edu/>.
- [19] Singapore Advanced Research Education Network (SingAREN), <http://www.singaren.net.sg/>.

- [20] Australia's Research and Education Network (AARNet),
<http://www.aarnet.edu.au/>.
- [21] National Institute of Information and Communications
Technology (NICT), <http://www.nict.go.jp/>.
- [22] Asia Pacific Information Infrastructure (APII),
<http://www.apiicc.org/>.
- [23] Sun Grid Engine (SGE),
<http://www.sun.com/software/gridware/index.xml>.