THE SUPPORTED GRID TOOL FOR MEASURING AN APPLICABILITY OF MEDICAL DATASET VISUALIZATION

IBRAHIM SALEM NAJAM

A project report submitted in partly fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science & Information
Universiti Teknologi Malaysia

NOVEMBER 2008

ABSTRACT

Grid computing is the technology that allow scientist to share their personal computer to store, process and visualize a large amount of data economically and efficiently. Medical dataset such as Computerized Tomography (CT) Scan and Magnetic Resonance Imaging (MRI) may contain a huge size of data that requires a powerful computer to visualize it. Existing grid tools and middleware are used to monitor and manage grid resources to achieve high computing processor. However, validating the medical dataset and measuring appropriate number of resources is an important task that may reflect the overall performance of grid computing. For example any damages in a dataset or missing in any sequence slices may result an error in processing. Additionally the number of resources for a specific size of dataset is essential to be identified. This project is aimed to study existing grid performance tools and develop the supported grid tool to measure an applicability of medical dataset in visualization process. Problem formulation and scope identification is the first step taken. Some analysis on grid performance tools namely Ganglia, Hawkeye and GridIce is performed to identify the performance criteria and supported grid tool development and evaluation is finally obtained. The developed supported tools is using dataset scanning technique and be able to identify the suitable and non-suitable medical dataset for visualization. The tool shows that the requirement of medical dataset with the size of 60 MB is six standard grid resources.

ABSTRAK

Pengkomputeran grid adalah teknologi yang membenarkan para saintis berkongsi komputer peribadi mereka untuk menyimpan, memproses and seterusnya menggambarkan (visualize) data yang besar secara ekonomik dan berkesan. Set data perubatan seperti Computerized Tomography (CT) Scan dan Magnetic Resonance Imaging (MRI) biasanya mengandungi sais data besar yang memerlukan komputer yang berkuasa tinggi untuk menggambarkannya. Peralatan grid dan perisian-perantara sedia ada digunakan untuk memantau dan menguruskan sumber-sumber grid untuk menghasilkan pemproses komputer yang berkuasa tinggi. Walau bagaimanapun, menentusahkan set data perubatan dan menentukan bilangan sumber yang bersesuaian adalah tugas yang penting dalam menentukan prestasi pengkomputeran grid secara keseluruhan. Contohnya sebarang kerosakan pada set data atau kehilangan jujukan lapisan data akan menyebabkan kesalahan semasa pemprosesan. Tambahan lagi bilangan sumber yang diperlukan bagi sesuatu sais set data tertentu adalah penting untuk dipastikan. Projek ini bertujuan untuk mengkaji peralatan prestasi grid sedia ada dan membina peralatan grid sokongan untuk mengukur kebolehgunaan sesuatu set data perubatan untuk proses visualisasi. Pembinaan masalah dan mempastikan skop projek adalah langkah pertama yang dilakukan. Beberapa analisa terhadap peralatan prestasi grid seperti Ganglia, Hawkeye dan GridIce dilakukan untuk mengenalpasti kritiria prestasi dan akhirnya pembinaan dan penilaian terhadap peralatan sokongan grid dihasilkan. Peralatan sokongan grid yang telah dibina menggunakan teknik mengimbas mampu mengenalpasti set data perubatan yang sesuai dan tidak sesuai bagi proses visualisasi. Peralatan ini juga menunjukkan bahawa keperluan untuk set data perubatan bersais 60 MB adalah enam sumber grid yang piawai.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

The idea of Grid computing has emerged over the past few years to allow different users to share resources that could be exhausted for them if they tried to use standalone computing platforms. It is a viable technique aimed to enable large-scale resource sharing among geographically distributed collaborations. The advances in network technologies that significantly increased over wide area connections helped pave the way to the Grid vision. Many communities have shown interest in leveraging Grid technologies to cope with their ever-increasing requirements in their fields of expertise. Examples include governmental organizations, biotechnological and health organizations, physicists, and economists. The envisaged patterns of usage range from distributed supercomputing for high-throughput and data intensive computing to on-demand and collaborative computing. Distributed networked resources such as desktops, servers, storage, databases, even scientific instruments combined create Grid computing. Grid Computing is used to achieve higher throughput and to deploy massive computing power wherever and whenever it is needed the most. Users of the Grid can find resources quickly, use them efficiently,

and scale them seamlessly (Mc Cormick et al. 1987; Bethel et al. 2000; Zhang et al. 2001; Foster et al. 2002).

The existence of basic Grid middleware such as Globus Legion or UNICORE (Ulrich, 2001) allowed the construction of some of the first Grid environments. The Global Grid Forum (GGF) aims to promote and support the development, deployment, and implementation of Grid technologies and applications via creation and documentation of best practices technical specifications, user experiences, and implementation guidelines.

Many scientific and engineering applications require access to large amounts of distributed data. These applications require widely distributed access to data by many people in many places. A data grid is a grid computing system that deals with data controlling the sharing and management of large amounts of distributed data. The data grid creates virtual collaborative environments that support distributed but coordinated scientific and engineering research. These techniques aim to minimize costs incurred by accessing data in a Grid environment and thereby increase the performance of single applications and the throughput of the entire system (Marian et al. 2002; Bonnassieux et al. 2002; Bowman, 2004; Allcock et al. 2003).

Data Grid allows users' data, once stored on local disk, to be located on remote network stores. Doctors and physicists now have instruments and simulations producing digital images representing medical datasets larger than local storage capacity. The digital imaging modalities include: Computer Tomography (CT), Digital X-ray, Magnetic Resonance Imaging (MRI), Nuclear Medicine, Ultrasound and Ophthalmology. A very effective method of assessing the value of dataset is visualization. Using web and Java-based technology allows user to render the large data at remote server without downloading the full dataset.

1.2 Problem Statement

There are many performance tools available for measuring the grid performance. Some of the examples are Ganglia, Hawkeye and GridIce. The criteria presented in different interface targeted to specific communities such as end user, application developers, middleware developers and system administrators. However, most of the interface give a very general information and not specific for a certain application, and they only focused for the suitable data which leads towards visualization and if some data or slices are damage then application cannot process for that or some unexpected result occur, as it is mentioned that Medical dataset visualization demand some specific presentation and indication.

1.3 Research Objectives

The objectives of this research are stated as below:

- To study and compare grid tools for performance measurement.
- To develop the grid tool for measuring an applicability of medical dataset.
- To evaluate the applicability of visualization for different types of medical dataset.

1.4 Scope of Study

To achieve the objectives of this study, the following scopes are considered:

- Write software-using Java programming language to generate XML scripts describing structures of large datasets.
- Developing suitable user interface for medical data set visualization.
- Ganglia, Hawkeye and GridIce are used for analyze the user interface.
- Testing the user interface to know whether it works accordingly.
- Testing will be carried out with the help of developed user interface to assess the performance measurement tool.

1.5 Importance of Research

One of the central challenges of Grid computing today is that Grid applications are prone to frequent failures and performance bottlenecks. Intervening layers of application, middleware, and operating systems often hide the real causes of failure. For example, assume a simple Grid workflow has been submitted to a resource broker, which uses a reliable file transfer service to copy several files and then runs the job. Normally, this process takes 15 minutes to complete, but two hours have passed and the job has not yet completed. In today's Grid, it is difficult to determine what, if anything went wrong. Is the job still running or did one of the software components crash? Is the network particularly congested? Is the CPU particularly loaded? Is there a disk problem? Is a software library containing a bug installed somewhere? In the simple case where the resources and middleware are only servicing one workflow, current Grid monitoring systems can answer these questions by correlating the workflow performance with the timestamps on the associated monitoring data. But the whole point of a Grid is that resources and

middleware are shared by multiple workflows. In this case, workflows will interleave their usage of middleware, hosts and networks. At the highest level of middleware, e.g., the resource broker in the examples above, there may be an identifier that can track the workflow. But once the workflow leaves that layer, there is very little beyond rough time correlation to help identify which monitoring data is associated with which workflow. With enough monitoring data, and enough time spent in analysis, troubleshooting is still possible. Also we can get responses in a visualized form.

REFERENCE

Allen G., Benger W, Goodale T., Hege H.C., Lanfermann G., Merzky A., Radke T., Seidel E. and Shalf J. (2000). The Cactus Code: A Problem Solving Environment for the Grid. In Proceedings of the Ninth International Symposium on High Perform*ance Distributed Computing (HPDC'00),* IEEE August 2000, p. 253–262.

Allcock W., Bester J., Bresnahan J., Chervenak A., Foster I., Kesselman C., Meder S., Nefedova V., Quesnel D., and Tuecke S. (2001). Secure, efficient data transport and replica management for high performance data-intensive computing. International conference on High Performance data. July 2001. Kaufman.

Bethel W., Tierney B., Lee J., Gunter D. and Lau S. (2000). Visapult Using High-Speed WANs and Network Data Caches to Enable Remote and Distributed Visualization, IEEE May 2000.

Brett B., Mark D. and David (2005). Serverside Visualization of Massive Datasets Thompson3. Proceedings of the First International Conference on e- Science and Grid Computing (e-*Science'05)*

Bonnassieux F., Harakaly R., and Primet P. (2002). MapCenter: an Open GRID Status Visualization Tool. In ISCA 15th International Conference on Parallel and Distributed Computing Systems, Louisville, Kentucky, USA, and September 2002.

Cooke A., Shrinkman Z., Manulkin C. and French F. (2003). R-GMA: An Information Integration System for Grid Monitoring. In Proceedings of the Tenth International Conference on Cooperative Information Systems, August 2003.

Daniel K., Gunter, Keith R., Jackson, David E., Konerding, Jason R., Lee J., Brian L. and Tierney. (2005). Essential Grid Workflow Monitoring Elements. The 2005 International Conference on Grid Computing and Applications (GCA'05), LBNL-57428.

David E., Gunter, Keith R., Jackson, Daniel K., Konerding, Jason R., Lee J., Brian L. and Tierney (2005) Grid Workflow Data collecting Elements, The 11th International Conference on Grid Workflow and Applications (GWA'05) , KBML-58427.

Engel K., (2000). Remote 3D Visualization using Image- Streaming Techniques. P. 47-54. Meeting, volume 2231 of Lecture Notes in Computer Science in Medical, Linz, Austria, p 77-98.March 2002.

Engel K. Ertl T., Sommer O., Sevior K. and Ernist C. (2000): Combining Local and Remote Visualization Techniques for Interactive. Volume of Rendering in Medical Applications.

Foster C., Kesselman M., Nick K. and Tuecke S. (2002). The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Technical report on Grid Services. Globus, February 2002.

Haber R.B. and Mc Nabb D.A. (1990). Visualization Idioms: A Conceptual Model for Scientific Visualization Systems. Ins: Visualization in Scientific Computing, for Stereoscopic Visualization, Proceedings of UK e-Science Second All Hands Meeting.

Ian Bowman (2004). Performance Modeling for 3D Visualization in a Heterogeneous Computing Environment. Retrieved on 17 July 2007 from http://vis.lbl.gov/Publications/2004/Bowman-PGVLBNL-56977.pdf

Lorensen, William and Harvey E. Cline Marching Cubes: A High Resolution 3D Surface Construction Algorithm. Computer Graphics (SIGGRAPH 87 Proceedings) 21(4) July 1987, p. 163-170). Retreived on 15 December 2008 from http://www.cs.duke.edu/education/courses/fall01/cps124/resources/p163-orensen.pdf

Mc Cormick B.H., DeFanti T.A. and Brown M.D. (1987) “Visualization in Scientific Computing”, Computer Graphics 211-214.

Mahovsky J. and Benedicenti L. (2003). Architecture for JAVA-Based Real-Time Distributed Visualization. IEEE Transactions on Visualization and Computer Graphics, 9(4): 570 – 579, December 2003.

Marian B., Wlodzimierz F. and Roland W. (2002). The crossgrid performance analysis tool for interactive grid applications. In D. Kranzlmller, P. Kacsuk, J. Dongarra, and J. Volkert, editors, Recent Advances in Parallel Virtual Machine and Message Passing Interface, 9th Europea*n PVM/MPI Users' Group Meeting,* volume 2474 of Lecture Notes in Computer Science, pages 50-60, Linz, Austria, September 2002.

Osborne J. and Wright H. (2003). SuperVise: Using Grid Tools to Support Visualization. In Proceedings of the Fifth International Conference on Parallel Processing and Applied Mathematics (PPAM 2003).

Primet P., Harakaly R., Bonnassieux F., and Allien F. (2004). Maps Center: an Open GRID Status Visualization Tool. In ISCA 17th International Conference on Series and Distributed Computing Grid Systems, Louisville, Kentucky, USA, October 2004.

Rich Wolski (2003). Experiences with Predicting Resource Performance. On-line in Computational Grid Settings (ACM SIGMETRICS Performance. Evaluation Review: Volume 30, Number 4, pp 41-49, March, 2003.

Roland W., Funika W., Marian B., Wlodzimierz M. and Wismller M. (2003). The crossgrid performance analysis tool for interactive grid applications. Recent Advances in Parallel Virtual Machine and Message Passing Interface, ninth *European PVM/MPI Users' Group*

Solveig A., Dirk D. and Paul M. (1999). Metadata Monitoring Requirements, pages 12, 14 and 16.

Thomas Sandholm and Jarek Gawor. (2003). Globus Toolkit 3 Core: A Grid Service Container Framework. Globus Toolkit 3 Core White Paper, July 2003.

Ulrich, K. (2001). European Patent No. FAP1162184. Retrieved on May 19, 2006, from http://www.unicode.org/grid/WhatIsgridservise.html

William J., Schroeder K., Jonathan M., Zarge A., William E., Lorensen. (1992). Decimation of triangle meshes, ACM 572 574 SIGGRAPH Computer Graphics, V.26, n.2, p. 65-70, July 1992.

Xiaoyu Z., Chandrajit B. and William B. (2001). Scalable Isosurface Visualization of Massive Datasets on COTS Clusters: Proceedings of the IEEE 2001 symposium on parallel and large-data visualization and graphics.