Scientific Data Sharing Using Clustered-based Data Sharing (CDS) in Grid Environment

Rohaya Latip[1], Hamidah Ibrahim[2], Feras Ahmad Al-Hanandeh[3]

University of Putra Malaysia, Serdang, Selangor, MALAYSIA
{rohaya, hamidah}@fsktm.upm.edu.my

Al Balqa Applied University, JORDAN
feras@hu.edu.jo[3]

ABSTRACT

In a large dynamic network, data can be shared and accessed in an easy manner among researchers from various locations and accessed provided an efficient protocol to manage the data. A protocol is needed to ensure the data is consistent and accessible. In this paper, we introduced a new protocol, named Clustered-based Data Sharing (CDS) for data sharing in a large dynamic network such as grid computing by using Clustered-based techniques to improve the accessibility. This is due to Clustered-based techniques can ensure that the data is accessible among the members of the grid. To evaluate our model, we developed a simulation model in Java. Our results show that CDS achieves high data accessibility even when the number of nodes or sites are increasing.

Keywords
Data sharing, cluster grid, data accessibility.

1. INTRODUCTION

A grid is known as a large scale geographically distributed hardware and software infrastructure composed of heterogeneous network resources owned and shared by multiple administrative organizations which are coordinated to provide transparent, dependable, pervasive and consistent computing support to a wide range of applications. These applications can perform distributed computing, high throughput computing, on demand computing, data intensive computing, collaborative computing or multimedia computing (Bote-Lorenzo et al., 2004).

By having large scale infrastructure of network the researchers can store their scientific data in the distributed database and share their data with other researchers and students. But if the data contains sensitive, personal information about human subjects, it may violate the Data Protection Act and ethics codes. Therefore sharing data is not advisable.

Most of the time data sharing among researchers are to preserve their data for their own future use, they will benefit it by being able to identify, retrieve, and understand the data. Other reasons why researchers should share their data is for teaching purposes. The collected data may be ideal for students to learn how to collect and analyse similar types of data by themselves.

Other than having grid environment for data sharing, sharing data can be obtained by informal and formal methods. Informal methods are accustomed to sharing data within their research group or extended virtual organization or an international research partnership. In these situations, risk is perceived as low and trust is high, particularly with co-investigators or co-authors. Data sharing within a group can be accomplished in a number of ways, but most of them require centrally administered, authenticated access. Examples shared network file, intranets, research project, online collaboration tools or groupware, proprietary enterprise and social networking sites for researchers. Researchers can directly communicate by asking questions about the intent of use and the user's experience. In these situations, direct or peer to peer sharing of data files may take place via email attachments, posting files on a website or ftp (file transfer protocol) server, with the link sent by email laptop to flashdrive, passing through Compact Disc (CD) or Digital Video Disc (DVD), social networking sites, file sharing sites, and mobile devices such as Bluetooth networking (Viki, 2009).

Formal methods of data sharing among researchers normally involve additional effort such as preparing a public use version of the dataset, cleaning the data of routine errors, documenting or annotating the data to improve its understandability, and possibly reformatting the data, so that it can be reused in another research context. Once this work is complete the data may be disseminated or published through posting data on a website or upload the data to a distributed, dynamic database.
The rest of this paper is organised as follows. Section 2 presents briefly the Clustered-based Data Sharing (CDS) model in grid environment. Next, Section 3 provides the

experimental setup of CDS. Section 4 then analyse the CDS result of accessibility in grid environment. Finally, a summary is provided and the paper is concluded.

2. CLUSTER-BASED DATA SHARING (CDS) MODEL

CDS model is developed for large dynamic network such as grid environment. The number of sites can be in any number where CDS can manage the data for sharing. Sites are clustered and connected with other clustered site which is known as member of the grid and one site is selected to have the master data file. We assume that replica copies of data are in data files.

The selected clustered site in grid is selected based on the nearest location with most frequently requested nodes. After the site is defined, the primary copy of the data is placed on the site while the replicas of the data are distributed to the other site in different cluster of the grid.

The framework of the CDS is as in Figure 1 where a middleware of Glite is used to distribute the data in each cluster grid.

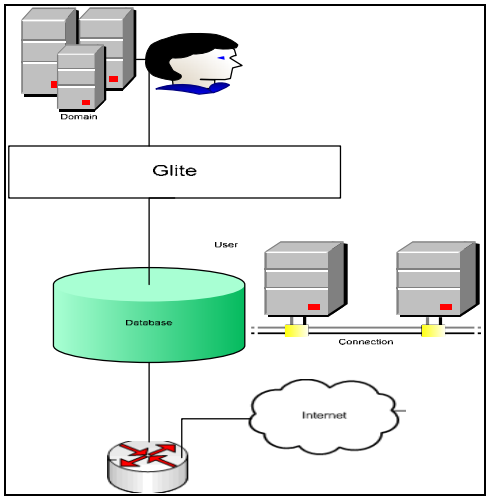

Figure 1: Glite as the Middleware

3. EXPERIMENTATION SETUP

The simulator is a modified version of (Rohaya, 2008) running in Java and the simulator flow is shown in Figure 2.

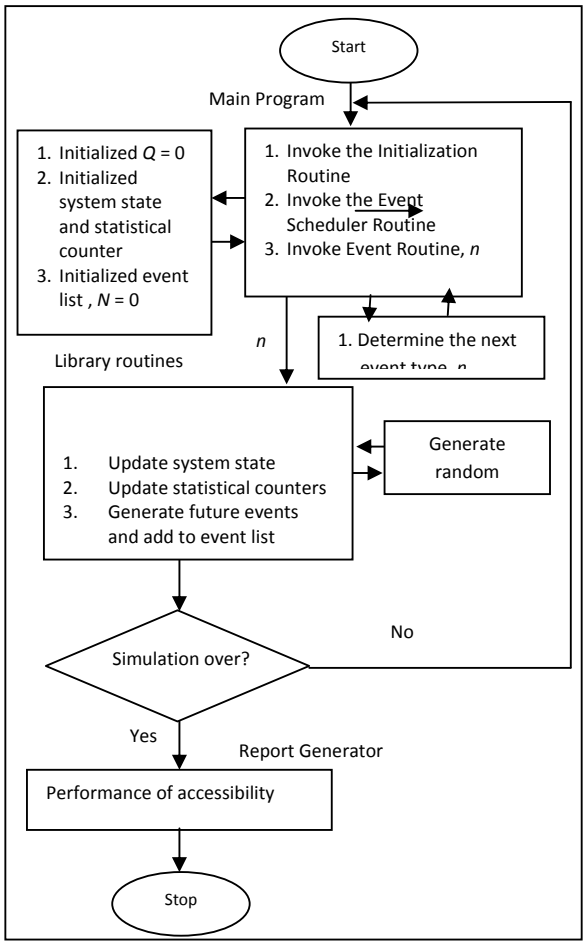

Figure 2: Flow of the Simulator

Figure 3 shows the network design for the simulator to run the CDS model. One clustered location has at least five nodes and a database to store the scientific data. The hardware requirements are presented in Table 1. Most of the parameters are taken form Lamehamedi and Szymanski (2007) and Ranganathan and Foster (2004) works.

Table 1: Hardware requirements

Number of sites/nodes in one cluster	5 – 2401
Size of data packet	2G
Storage capacity	1 Terabyte
Inter-arrival rate	5s

Scientific data means the data stored at each cluster are data used for research in different field. The data are stored in data files.

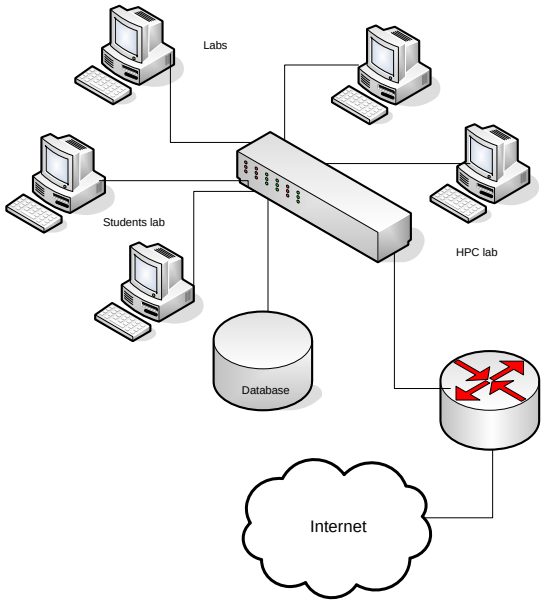

Figure 3: Network design

4. PERFORMANCE ANALYSIS AND RESULT

We used the discrete event simulation to evaluate the performance of the proposed data sharing model, Cluster-based Data Sharing (CDS) protocol. We used the data accessibility to measure the effectiveness of the CDS protocol.

Data accessibility is improved by storing multiple copies of data at different sites. This is because, data remaining available to users despite site and communication failures. Accessibility, $A(t)$, refers to the probability that the system is operational according to its specification at a given point in time. The failure of accessing is following the Poisson distribution. The failure rate, λ, and the repair time is exponential with a mean repair time, $1/\mu$, where μ is the repair rate. The accessibility $A(t)$, can be written as in Eq. (1).

$$A(t) = \frac{\lambda}{\lambda + \mu} \quad (1)$$

In particular, the data is accessible when there is availability of a read or write operation of the system. To estimate the operation availability, all copies are assumed to have the same availability of a data item. The notations below are defined:

- $A_{X,R}$ represents the accessibility of X model in Read operation.
- $A_{X,W}$ represents the accessibility of X model in Write operation.
- $AC(X)$ is the system accessibility of X model.

A_{CDS} represents the accessibility of CDS. If the probability that an arriving operation of read and write for the data file are f and $(1 - f)$ respectively, then the accessibility of X model is expressed as in Eq. (2) taken from Maekawa (1992).

$$AC(X) = (f\, A_{X,R} + (1 - f)\, A_{X,W}) * 100\% \quad (2)$$

From the equations discussed above, the percentage of accessibility for 49, 81 and 121 nodes are shown in Figure 4. The probability, p is the probability of the data being accessed. The probability is in the range of 0 to 1.

Figure 4 shows that the percentage of accessibility increased even when the the number of nodes is increasing. This is because each cluster has its own data copies to access. The member of the cluster grid will first access the data copies at the local cluster and if the database is busy or unavailable, the requester will request the data at another cluster. Therefore accessibility of the scientific data is always available.

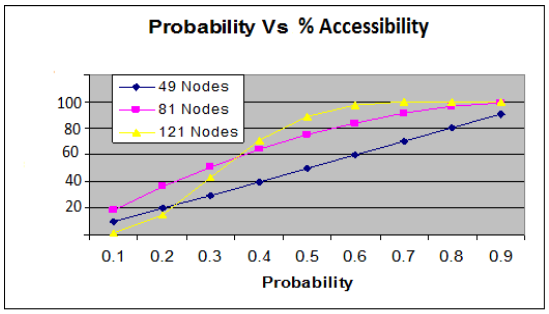

Figure 4: Percentage of Accessibility Result

Figure 4 shows that the increment of percentage of accessibility is 16.67% for probability 0.7 and at probability 0.8 the percentage of accessibility increases to 23.08%. Therefore the CDS model is suitable for dynamic and large network such as grid environment.

5. CONCLUSION AND FUTURE WORK

In this paper, CDS model selects a site from the large number of nodes to act as a database and replicate the data copies. Selection is based on the nearest node with the frequently requested nodes. Meaning all clusters have their own database to manage and share their data. By having database locally at each cluster will make data sharing more accessible and efficient.

The simulator is developed to study the performance of the CDS model. The results demonstrate that the CDS model increases the percentage of accessibility performance even when the number of nodes is increasing in size.

In this paper, consistency was not discussed. Therefore in the future, we will investigate the consistency of data. These investigation will be used to evaluate the performance of our CDS model for updating data. Currently we have improved the model to a large network and allocate a certain number of nodes into

cluster. We manage to find the optimized number of nodes in each cluster to implement this protocol.

ACKNOWLEDGEMENT

Thank you to Malaysian Ministry of Science, Technology and Innovation (MOSTI) for supporting our project under the Fundamental Grant No: 02-01-07-269FR.

REFERENCES

Bote-Lorenzo, M.L., Dimitriadis, Y.A., and Gómez - Sάnchez, E. (2004). Grid Character and Users: A grid Definitions. *Lecture Notes of Computer Science Springer Verlag Berlin Heidelberg. 2970*, 291-298.

Lamehamedi, H. and Szymanski, B.K. (2007). Decentralized Data Management Framework for Data Grids, *Future Generation Computer Systems.* 23(1), 109-115.

Maekawa, M. (1992). A $\sqrt{n}$ Algorithm for Mutual Exclusion in Decentralized Systems. *ACM Transactions Computer System. 3*(2), 145-159.

Ranganathan, K. and Foster, I. (2004). Simulation Studies of Computation and Data Scheduling Algorithms for Data Grids. *Journal of Grid Computing.1*(1), 53-62.

Rohaya, L., Hamidah, I., Mohamed, O., Md Nasir, S. and Azizol, A. (2008). Quorum Based Data Replication in Grid Environment. International Conference on Rough Set and Knowledge Technology 2008, 379-386.

Viki, G. (2009). Article from http://www.ed.ac.uk/schools-departments/information-services/services/research support/data-library/research-data-mgmt.