MANAGING REPLICATION AND TRANSACTIONS USING NEIGHBOUR REPLICATION ON DATA GRID

NORAZIAH BINTI AHMAD

Thesis Submitted in Fulfillment of the Requirement for the Doctor of Philosophy in the Faculty of Science and Technology Universiti Malaysia Terengganu

July 2007

Abstract of thesis presented to the Senate of Universiti Malaysia Terengganu (UMT) in fulfilment of the requirement for the degree of Doctor of Philosophy

MANAGING REPLICATION AND TRANSACTIONS USING NEIGHBOUR REPLICATION ON DATA GRID

NORAZIAH BINTI AHMAD

July 2007

Chairperson : Professor Md. Yazid Mohd. Saman, Ph.D

Member : Associate Professor Muhammad Suzuri Hitam, Ph.D
Professor Mustafa Mat Deris, Ph.D

Faculty : Science and Technology

Replication is a useful technique for distributed database systems. Through this technique, a data object will be accessed (i.e., read and written) from multiple locations. Thus, it increases the data availability and accessibility to users despite site and communication failures. The all-data-to-all sites replication schemes such as Read-One-Write-All (ROWA) and Tree Quorum (TQ) are the popular techniques being used for replication and management of data in this domain. However, these techniques have its weaknesses in terms of data storage capacity and also data access times due to some number of sites must agree in common to execute certain transactions. In this study, the all-data-to-some sites scheme called Neighbour Replication on Grid (NRG) technique is proposed by considering only neighbours that have the replicated data. It is based on the logical structure of sites/servers in order to form a read or a write quorum in distributed database systems. The proposed technique considers only neighbours obtain a data copy. For simplicity, the neighbours are assigned with vote one and zero otherwise. The assignment provides a minimum communication cost with high system availability, due to the minimum

number of quorum size required. In addition, it minimizes the storage capacity as well as data access time.

A series of experiment was carried out by using three servers. Neighbour Replication on Grid (NRG) daemon is developed under Linux platform in the local area network (LAN) environment. It was carried out in Shell and Perl programming integrated with File Transfer Protocol (FTP) for the communications agent. The experimental results showed that the proposed model work successfully in managing replication and transaction when no failures occurred. Besides, the reconciliation and resolving conflict during system recovery are also supported when primary and neighbour replicas have failure.

ACKNOWLEDGEMENTS

I would like to express my most sincere gratitude to the supervisory committee Prof. Dr. Mustafa Mat Deris for his continuing support, professional guidance and for giving me an opportunity to learn what research is all about. Special gratitude also to Prof. Dr. Md. Yazid Mohd. Saman and Assoc. Prof. Dr. Muhammad Suzuri Hitam for their contributions, guidance and time towards this research. Not forget to Ms. Noraida Haji Ali, for her ideas during an early of this study.

Sincerely thanks should be forwarded to Vice Chancellor Universiti Malaysia Pahang (UMP), YH Prof. Dato' Dr. Mohamed Said bin Mat Lela for the KUKTEM Scholarship and also for the sponsorships to the International Symposium on DCABES 2006 and IIGSS Workshop 2007. Not forget to Dean of Faculty Computer System and Software Engineering (FSKPP), Prof. Dr. Abdullah Embong; Assoc. Prof. Ruzaini Abdullah Arshah; staffs in HR Department and all friends in UMP.

Special gratitude also to my family, especially to my father, Ahmad Zakaria; in memory of my mother, Esah Ibrahim; my beloved husband, Zakaria Mamat; my sister, Mazni Ahmad; my brothers Mohd Rashidi and Abdul Aziz; my mother in-law, Wan Sanah Wan Ibrahim; and my father in-law, Mamat Endut; for their patience and morale support.

Finally, I thank to all my friends especially to Rabiei, Norhayati, Nathrah, Suryani, Shuhadah, Ahmed, Che Norhaslida and others who have contribute in this research.

I certify that an Examination Committee has met on 8th of July 2007 to conduct the final examination of Noraziah Ahmad on her Doctor of Philosophy thesis entitled "Managing Replication and Transactions Using Neighbour Replication On Data Grid" in accordance with the regulations approved by the Senate of Universiti Malaysia Terengganu. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

Mohd. Pouzi Hamzah, Ph.D.
Faculty of Science and Technology
Universiti Malaysia Terengganu
(Chairperson)

Md. Yazid Mohd. Saman, Ph.D.
Professor
Faculty of Science and Technology
Universiti Malaysia Terengganu
(Member)

Muhammad Suzuri Hitam, Ph.D.
Associate Professor
Faculty of Science and Technology
Universiti Malaysia Terengganu
(Member)

Mustafa Mat Deris, Ph.D.
Professor
Faculty of Information Technology and Multimedia
Universiti Tun Hussien Onn Malaysia
(Member)

Hamidah Ibrahim, Ph.D.
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Independent Examiner)

Mashkuri Yaacob, Ph.D.
Professor
Universiti Tenaga Nasional
(Independent Examiner)

MOHD AZMI AMBAK, Ph.D.
Professor / Dean of Graduate School
Universiti Malaysia Terengganu

Date:

This thesis submitted to the Senate of Universiti Malaysia Terengganu and has been accepted as fulfilment of the requirements for the degree of Doctor of Philosophy.

MOHD AZMI AMBAK, Ph.D.
Professor / Dean of Graduate School
Universiti Malaysia Terengganu

Date:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1SR	One-copy Serializable
CERN	European Organization for Nuclear Research
DDS	Distributed Database Systems
DDBMS	Distributed Database Management Systems
FTP	File Transfer Protocol
GS	Grid Structure
HEP	High Energy Physics
HPN	Highest Priority Neighbour
LAN	Local Area Network
LHC	Large Hadron Collider
NRG	Neighbour Replication on Grid
NRG_NTD	NRG *Need-To-Do* table
NTM	NRG Transaction Manager
P2P	Peer-to-peer
PNGC	Primary-Neighbours Grid Coordination
POSIX	Portable Operating Systems Interface
ROWA	Read-One-Write-All
SMTP	Simple Mail Transfer Protocol
TQ	Tree Quorum
VT	Voting
WAN	Wide Area Network

CHAPTER 1

INTRODUCTION

Several research articles have been published regarding distributed databases. Among them were those by Agrawal et al. [8, 9, 10, 22], Berstein et al. [46, 47], Chung [58], Garcia-Molina and Barbara [16], Maekawa [31], Mustafa et al. [33, 34, 35, 36, 66], Nicola [39], Stockinger [17] and Zhou et al. [68]. Those articles revealed that replicated data management is one of the current issues that still unsolved in distributed databases. Therefore, the study on this basis is initiated.

A database (DB) can be defined as a shared collection of logically related data that has been designed to meet the information needs of an organization and to be used by multiple users [15, 61]. The emergence of the network and additional communication facilities to a database system, can take it from centralize to a decentralize concept [11]. Distributed database system (DDS) is one of the major developments in the database area, where it moves from centralization that resulted in monolithic gigantic databases towards more decentralization [33, 35]. DDS is defined as a collection of multiple independent databases that operate on two or more computers that are connected and share data over the network [15, 61]. Meanwhile, a Distributed Database Management System (DDBMS) can be defined as the software that permits the management of the distributed database and make the data distribution transparent to users [61]. Nowadays, commercial database system such as *Oracle Database 10g* provides the required support for data distribution and inter-database communication

[53]. As a result of remarkable advance communication technologies, wireless and mobile computing concepts become reality. These concepts allow for even higher degrees of distributedness and flexibility in distributed databases [6, 35].

With the advances in distributed processing and distributed computing, the database research communities have carried out considerable works to address issues of data distribution, distributed design, distributed query processing, distributed transactions management and etc [61]. One of the major issues in data distribution is replicated data management. Typical replicated data management parameters are data availability and communication costs. These parameters share an inverse dynamic relationship: the higher the data availability with a low communication cost, the better the system is [31, 33].

1.1 Data Replication

Replication is an act of reproducing. It also addresses the management of the complete copying process [28]. Any type of data processing object can be implemented. These include the data files [13, 20, 23, 25, 42]; entire databases, specific tables, data within a specific tablespace [53]; service types such as Telnet, File Transfer Protocol (FTP), Simple Mail Transfer Protocol (SMTP) [38, 40] and etc. In a distributed database that relies on replication, the DDBMS may maintain a copy of fragment at several different sites [61]. Besides replicating, it also encompasses the administration to guarantee those data consistency across multiple sites.

Recently, scientific research and commercial application generate large amount of data that are required by users around the world. To illustrate this, the High Energy Physics (HEP) area deploys a new particle accelerator, the Large Hadron Collider (LHC) [17, 42]. It starts working at European Organization for Nuclear Research (CERN) in the year 2007. Several HEP experiments will produce Petabytes of data per year for decades. Those data will need to be managed and stored at more than hundreds of participating institutions. Thus, the data replication is a very useful technique to manage the large scale data across widely distributed communities.

Data replication plays an increasingly important role in this evolving world of distributed databases. Through this technique, an object will be accessed (i.e., read and written) from multiple locations such as a local area network or geographically distributed network world wide. For example, a student's results in a college, will be read and updated by lecturers of various departments. The financial instruments' prices will be read and updated from all over the world [36, 45]. Therefore, this technique provides high availability, fault tolerance and enhance the performance of the system [27, 33, 34, 36, 42].

Two approaches commonly used for replication, namely synchronous and asynchronous. Synchronous means move or operate together at the same time with each other while asynchronous is otherwise. Thus, synchronous replication provides what is called '*tight consistency*' between data stores. This means that the latency between data consistency is zero. If any copy is updated, the update immediately applied to all other copies within the same transaction. Data at all sites is always the same and exactly consistent, no matter from which replica the updated originated.

Conversely, asynchronous replication provides what is called *'loose consistency'* between data stores. This means that the latency between data consistency is always greater than zero. If one copy is updated, the change will be propagated and applied to the other copies within separate transactions. This copy changes may occur over seconds, minutes, hours, or even days later. Thus, some degree of lag always exists between the originating transaction that committed, and the effects of the transaction available at other replica(s) [28].

1.2 Problem Statements

Single centralized DB has low availability and reliability because if the DB site goes down the whole system fails. DDS has high availability and reliability. However, DDS introduces high redundancy as more than one site is used. This also creates low data consistency and data coherency as more than one replicated data need to be updated. Thus, some of the research problems that arise in DDS can be stated as follows:

- How does high availability of data achieved?
- How does redundancy reduced while there is an increase in the data storage capacity? How does replicated data minimized?
- How are synchronization mechanisms done in order to maintain the consistency of data when changes are made by transactions?

By storing multiple copies of data at several sites in the system, there is an increased data availability and accessibility to users despite site and communication failures. It is an important mechanism because it enables organizations to provide users with

access to current data where and when they need it. However, the storage capacity becomes an issue as multiple copies of data are replicated on different sites. Of course this way of data organization increases the data storage capacity. At the same time, expensive synchronization mechanisms are needed to maintain the consistency and integrity of data when changes are made by the transactions. This suggests that proper strategies are needed in managing replicated data in distributed database systems.

One of the simplest techniques is all-data-to-all sites replication scheme, namely Read-One-Write-All (ROWA) technique [55]. Read operations on a data object are allowed to read any copy, and write operations are required to write all copies of the data object. This technique has been proposed for managing data in mobile and peer-to-peer environment [6, 61]. It provides read operations with high degree of availability at low communication cost. However, it severely restricts the availability of write operations since they cannot be executed at the failure of any copy. This technique results in the imbalance of availability as well as the communication cost of read and write operations. The read operations have a high availability and low communication cost. Meanwhile, the write operations have a low availability with higher communication cost. Voting (VT) techniques [12] became popular because they are flexible and are easily implemented. This technique has been applied to the primary cluster for managing replicated data [20]. One weakness of these techniques is that writing an object is fairly expensive: a write quorum of copies must be larger than the majority of votes. To optimize the communication cost and the availability of both read and write operations, the quorum technique generalizes the ROWA technique. It imposes the intersection requirement between read and write operations [55]. Write operations can be made fault-tolerant since they do not need to access all

copies of data objects. Dynamic quorum techniques have also been proposed to further increase availability in replicated databases [21, 57]. However, these approaches do not address the issue of communication cost of read operations. Another technique, called Tree Quorum (TQ) technique [8, 9, 58] uses quorums that are obtained from a logical tree structure imposed on data copies. This technique has been proposed for persistent consistent distributed database commit in peer-to-peer network [3]. Nonetheless, this technique also has some drawbacks. If more than a majority of the copies in any level of the tree become unavailable, write operation cannot be executed. Several researchers have proposed logical structures on the set of copies in the database. To create intersecting quorums, the logical information has been deployed. The technique that uses a logical structure such as Grid Structure (GS) technique, executes operation with low communication costs while providing fault tolerance for both read and write operations [10]. However, this technique still requires that a bigger number of copies be made available to construct a quorum.

1.3 Objectives and Scopes of Research

The objectives of the research are as follows:

i - To propose & to develop a new data replication model & to design its transaction management for the grid structure.

ii – To analyze the model mathematically and to implement the model.

iii – To test and to evaluate the performances of the proposed model.

This thesis only concentrates on the synchronous replication. Therefore, this research will focus on replication technique, in order to obtain high system availability with

low communication costs, in managing data replication by means of the synchronous replication.

In this research, the Neighbour Replication on Grid (NRG) technique is proposed based on the logical structure, in order to improve the read and write operations. Consequently, the implementation of this technique combines the replication and the transaction techniques. The motivations of this implementation are to show the clarity of the algorithms and also NRG technique can be used in the practical applications. A simplify approach to data distribution has been implemented by using the replication servers. A replication server handles the replication of data to remote sites. Through this mere simplification, it allows clearer presentation of an algorithm when less locking information of data is deployed.

1.4 Organization of Thesis

This thesis is organized as follows: Chapter 2 reviews the five major replica techniques, namely Read-One-Write-All, Lazy Replication, Voting, Tree Quorum and Grid Structure. In Chapter 3, the concepts of transaction, serializability, quorum intersection property, communication cost and availability are described. The performance analysis of the existing replication techniques are given in Chapter 4. Chapter 5 proposes the Neighbour Replication on Grid (NRG) technique. The performance comparisons with other replication techniques are then carried out in Chapter 6. Next, Chapter 7 presents the combination of replication and transaction techniques. In Chapter 8, the implementation of those techniques is described. Finally, the conclusion and the future works are presented in Chapter 9.

1.5 Conclusion

This chapter introduces the database systems and distributed database systems concepts. Data replication which is the main focus attention this research has been extensively elaborated. Together with advantages of data replication brings specific problems. It occurs since multiple copies of data are replicated at different sites. The storage capacity and expensive synchronization mechanisms to maintain the consistency and integrity of data become the issues. Thus suggests that proper strategies are required to solve the problems, which are the significance of this research.