Real Time Analysis of Data Grid Processing for Future Technology

Thurein Kyaw Lwin
Saint-Petersburg State University
198504 St.Petersburg, Russia,
Peterhof, Universitetsky pr., 35
e-mail: trkl.mm@mail.ru

Alexander Bogdanov
Russian Economic University
198504 St.Petersburg, Russia,
Peterhof, Universitetsky pr., 35
e-mail: bogdanov@csa.ru

ABSTRACT
The purpose of this paper is to examine the distributed databases and some problems of Grid dealing with the data replication from different points of view. This paper is referred to find out the most efficient commonalities between Data Grid and managed data stored in object-oriented databases. We discuss the features of Distributed Database System (DDBS) such as design, architecture, performance and concurrency control and submit some research that has been carried out in this specific area of Distributed Database System (DDBS). Query optimization, distribution optimization, fragmentation optimization, and joint option to the optimization of Internet are included in our research. Our project design, advantages of its results and examples concerning this topic are presented in this paper.

Keywords
Database System, Cosolidation Technology, Hybrid Cloud, Distributed System, centralized databases, federated databases, Virtual server, Virtual Processing.

1. INTRODUCTION
Some of the related works are launched in the database community as well as in the Grid community. At first, we discuss about the aspects of Grid computing coming from the high performance computing and cluster computing where several processors or workstations are connected high speed interconnect in order to the process of a mutual program. Distributed database (DDB) is a collection of multiple database and logically interrelated databases of distributed data over a computer network [2]. A distributed database management system (DDBMS) can be managed by the software of distributed database and provides an access mechanism that makes this distribution transparent to the users. We spread of the data among the sites of the distributed network, which has its own computer and data storage facilities. All of this distributed data are considered to be a single logical database [3]. When a process of anywhere in the distributed network queries, it is not necessary to know where the data location in network.

2. ANALYSIS OF THE DISTRIBUTED DATA PROCESSING
We tested about the distributed system at our Saint Petersburg State university research center and checked how to improve the performance and extend the range of applications of scientific methods and algorithms for parallel and distributed data processing. The purpose is to provide an operating environment for the database and its consolidation in distributed computing. This network can be used in research institutions and commercial enterprises, whose resources may be located in local area network and wide area network in remote locations [4,5]. Some problems of prototype system architecture, algorithm development and adapting existing software products could be solved by the usage of this network. Such as the system is implemented in the form of blocks and can make the distributed virtual computer system and it can be named virtual TESTBED [9].

3. FUNDAMENTAL STRUCTURE OF OUR PROJECT'S ARCHITECTURE
Our project is carried out at the Saint Petersburg State University research center and it can be divided into 4 modules, they are Module Network, Random Dynamic Process, Data Grids Replacement, Data Secure.

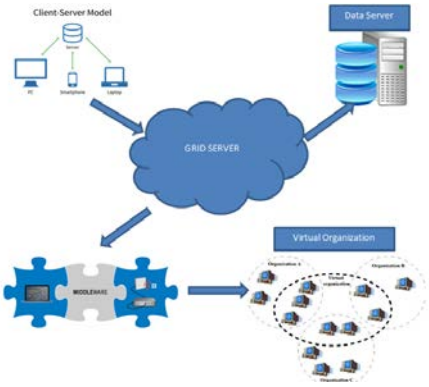

Fig.1. Our Grid Architecture in SPBSU

The Clients server and Main servers are operated over a computer network on hardware separated from our project. A server machine is running one or more server programs with a high performance and sharing resources for clients. A client also shares any data from resources, clients need to initiate communication sessions with servers, which await incoming requests [10]. The delivery packet is sent to the destination at a node in order to minimize the probability, that packets are eavesdropped over a specific link of a randomization process for packet deliveries. In this process, the previous next hop for the source node is identified in the first step of the process [7,10]. That process can randomly pick up a neighbouring node and the current packet transmission with the next hop. The exclusion for the next hop is a selection that avoids transmitting two packets of consecutive data as shown in figure 1. We investigated the problem of optimal allocation of sensitive data objects and the partitioned by using secret sharing scheme or delete coding replicated. We considered to achieve secure, and high-performance processing of the data stored in data grid storage. We decomposed the allocation problem into two problems of intra cluster and inter cluster that shared the allocation problems separately and independently [7,11]. Data grid is a distributed computing architecture, integrated as a large number of data sets and computing resources into a single virtual data management system (SVDMS). It can do sharing and coordination of the data from various resources and provides various services to fit the needs of

high-performance distributed computing. Replication data can be widely distributed to achieve better access performance and load sharing in data grids [8].

4. REAL TIME DISTRIBUTED DATABASE

We installed Hadoop and checked the basics of working with file system in a real time application in our server. We used one server for the standalone operation of Hadoop and the others for the fully-distributed operation. We separated the name node and the data node in the fully-distributed operation of Hadoop. The Fully-distributed operation is good for processing large data. Processing small data with the fully-distributed operation is undesirable because the time it takes to collect distributed data in the same nodes during a Reduce operation outweighs the advantages of distributing data to each node using Map. In this test, the fully-distributed operation of 10 million rows of data outperformed other test conditions. The test result varies depending on the amount of data and the system specification for the future of Big data and data distributed system as shown in figure 2.

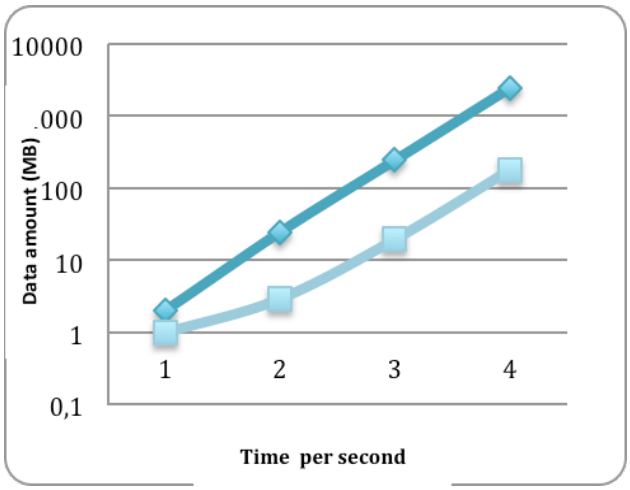

Fig.2. Real time Distributed Data in Our Database Server

5. CONCLUSIONS AND PROJECT REALIZATION

We combined the data partition schemes with dynamic replication to achieve data security, and access performance in data grids processing. The partition data need to be properly allocated to achieve the actual performance benefits in replication process. Our project design can provide the following advantages:

• Data can be secured

• Enables the sharing of coordinated data from various resources and provides various services with distributed and data intensive computing

• Replication techniques are frequently used to improve the data and reducing of client response times and communication

• Single point access can't be accepted in this system

We implied our project of distributed database to the problem of data replication in the Grid middleware with large amounts of data Network. We developed the algorithms to allocate shared data in Data grid and distributed computing architecture that integrates a large number of data sets and computing resources into a single virtual data management system. In our project, data can be often stored in database management systems and it is reasonable to try to understand the key features of both communities with distributed databases and Grid, which can combine of common ideas for an efficient Data Grid and analyze of the different data systems for future technology [12].

6. ACKNOWLEDGEMENT

This Research is supporting by Faculty of Applied Mathematics and Control Processes, Department of Computer Modeling and Multi Processing system of Saint-Petersburg State University. We thank our colleagues from our Department who provided insight and expertise that greatly assisted the research.

REFERENCES

[1] The European DataGrid Project: http://www.cern.ch/grid

[2] Wolfgang Hoschek, Javier Jaen-Martinez, Asad Samar, Heinz Stockinger, Kurt Stockinger "Data Management in International Data Grid Project", *1st 1EEE, ACM International Workshop on Grid Computing (Grid'2000)*, Bangalore, India, 17-20 pp 18-21, 2010.

[3] The Globus Project: http://www.globus.org

[4] The Apache Software Foundation "The Apache Cassandra Project, 2014." http://www.cassandra.apache.org

[5] Dr. H. Hakimzadeh "Advanced Database – B561. Spring 2005", Department of Computer and Information Sciences Indiana University South Bend.

[6] Heinz Stockinger "Distributed Database Management Systems and the Data Grid", CERN, European Organization for Nuclear Research, Geneva, Switzerland Institute for Computer Science and Business Informatics, University of Vienna, Austria.

[7] A.S.Syed Navaz, C.Prabhadevi, V.Sangeetha " Data Grid Concepts for Data Security in Distributed Computing", Department of Computer Applicatons, Muthayammal College Of Arts & Science, Namakkal, India, *International Journal of Computer Applications* (0975 – 8887) Volume 61– No.13 pp 63-66, January 2013.

[8] K. Ranganathan, I. Foster "Identifying Dynamic Replication Strategies for a High Performance Data Grid", Second Int'l Workshop Grid Computing, white paper 2001.

[9] Bogdanov A. V., Thurein Kyaw Lwin, Elena Stankova "Storage Database System in the Cloud Data Processing on the Base of Consolidation Technology", *Computational Science and Its Applications - ICCSA 2015, 15th International Conference*, Banff, Canada, June 22-25, 2015, Proceedings, Part IV. Springer Volume 9158 of the series Lecture Notes in Computer Science pp 311-320. DOI: 10.1007/978-3-319-21410-8_24

[10] Y. Deswarte, L. Blain, and J.C. Fabre, "Intrusion Tolerance in Distributed Computing Systems", *Proc. IEEE Symp. Research in Security and Privacy*, 2015.

[11] Foster and A. Lamnitche "On Death, Taxes, and Convergence of Peer-to-Peer and Grid Computing", *Second Int'l Workshop Peer-to-Peer Systems (IPTPS)*, 2003.

[12] L. Xiao, I. Yen, Y. Zhang, F. Bastani, "Evaluating Dependable Distributed Storage Systems," Int'l *Conferene Parallel and Distributed Processing Techniques and Applications (PDPTA)*, pp 32-35, 2017.