INTEROPERABILITY OF DATA MANAGEMENT

IN GRID COMPUTING

[1] Ms Prema.R , [2] Dr Antony Selvadoss Thanamani

Head and Assistant Professor, Department of Computer Applications,Indo Asian Women's Degree College, Bangalore-560043.

Head & Associate Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi – 642 001

Abstract

Replication is a widely used method in data grid which aims to reduce bandwidth consumption, improve response time and maintain reliability. In data grid, for performance enhancement file correlations become an important consideration. The scrutiny of actual data intensive grid applications suggests that correlations can be broken for improving the efficiency of replication strategies and acknowledge that job requests for group of correlated files. In this paper an algorithm namely, Data mining based dynamic replication algorithm (DMDR) has been implemented, which consider a set of files as granularity. This work collects files based on a relationship of concurrent accesses between files by jobs as well as stores related files at the same time. So DM (Data Mining) is introduced to find out these correlations. All confidence measure is chosen for correlation measure.

Index Terms : DMDR Algorithm, grid Architecture

I INTRODUCTION

Data grid is a large scale sharing and computational resources that are not restricted within a particular location. It is now gaining much important interest in important areas such as high energy physics, bioinformatics, earth observations, global climate changes, image processing. At present data grid facing challenges to improve the data management since the techniques must revamp for heterogeneity, dynamicity and autonomy of data sources. Replication is the only technique which is used to improve the data management [1]. It creates multiple copies of the same file in more than one storage locations for load balancing, to improve response time, to reduce bandwidth consumption, and to maintain the reliability.

Most of the existing replication techniques, neglects correlation among different data files as well as those are all based on the single file granularity. In fact, in most of the applications, data files may be correlated for accessing and to consider together to reduce access cost [2]. By analysing Coadd, which is a actual data intensive grid application, found that there is strong correlation between jobs tend to demand groups of correlated files and requested files. The knowledge of correlation between data is extracted from historical data using the techniques of data mining field.

Data mining tools are used to extract meaningful information from vast amount of data. Even though data mining has been applied for sectors, the applications of data mining in the context of replication is limited. In this regard, data mining techniques are used by several works to explore file correlations in data grids [2]. The main approaches to mining file correlations can be classified into two categories: access sequence mining and association rule mining [3].

On the other hand, various studies have shown the limits of association rule mining based on the support and confidence approach and confirm that it results on an extremely large number of rules, with a most of them do not reflect the true correlation relationship and redundant among data. This paper introduced frequent correlated pattern mining to overcome this drawback. Replacing the support measure with a more expressive measure provides better captures whether the items in a set are associated. Correlated patterns mining was shown to be more complex and more informative than frequent patterns mining and association rules [4].

II RELATED WORK

In this section we discuss various strategies followed for data management by different authors in their paper. In [1] Lakshmi.R, Antony Selvadoss Thanamani proposed various strategies. Existing strategies are based on the single file granularity so they proposed new strategy called DMDR. They chose all the confidence as correlation measure.

In [2] TarekHamrouni, Sarra Slimani, et al, proposed new dynamic periodic decentralized data replication strategy as well as they introduced new pattern mining algorithm to find out frequent correlated pattern. Using OptorSim simulator they proved the performance with compared to other strategies. In [3] Sean Carlistode Alvarenga, SylvioBarbonJr, proposed a new approach to facilitate the investigation of intrusion alert. Process mining approach is used to extract the information. Large models are clustered into smaller through hierarchical clustering.

In [4] HyunwooKim, SuckwonHong, et al., proposed an systematic approach and identified potential area for concentric diversification. Text mining is used to construct an integrated patent-product database. Then association rule-mining is used to construct a product ecology network. Next link prediction analysis of conducted to identify potential areas for concentric diversification. At last quantitative indicators are developed to assess the characteristics of the areas identified.

In [5] Sheida Dayyani, Mohammad Reza Khayyambashi, presents a survery on new replication techniques which is proposed by other researchers. Then they made study on those replication strategies and finally they summarize the results about replication techniques. In [6] Sheida Dayyani, Mohammad Reza Khayyambashi, proposed dynamic hierarchical replication with threshold. It is an enhanced version of dynamic hierarchical replication strategy used for characterizing the number of appropriate sites for replication. They used OptorSim for simulation and proved, it provides better performance compared with other algorithms.

In [7] Ming Tang, Bu-Sung LeeX, et al., evaluated various replication algorithms performance with various simulations and proved it reducing the job turnaround time remarkably. In [8] TaoWang, Shihong Yao et al., prosed a well designed dynamic replication strategy which consists of replica replacement algorithm, replica layout algorithm and replica selection algorithm. OptorSim is used for Simulation and proves it provides better performance.

In [9] N. Mansouri, Gh. Dastghaibyfard, proposed modified dynamic hierarchical replication strategy. OptorSim is used for simulation which shows Modified Dynamic Hierarchical Replication achieves better performance compare to other strategies in terms of network usage, storage usage and job execution time. In [10] Babu R., Rao S., proposed check point based optimal replication which checkpoint for replication and calculates the number of copies by setting weight for each data access record. OptorSim is used for simulation and proves it provides better result. In [11] A.Vergnol, J.Sprooten, et al., proposes to enhance congestion management by used real-time supervisor. This method reduced the redispatching costs as well as increased the network reliability. EUROSTAG software is used for simulation.

Data grid architecture

Grid computing is a broad area distributed computing environment which enables selection, aggregation, sharing of geographically distributed resources. As well as it is an significant mechanism for utilizing distributed computing resources. These resources are spread in different locations, but finally organized to give an integrated service. Grid computing is a special networking infrastructure which is designed to provide reliable access to computational resources and data over network, across various domains. Nowadays storing, retrieving and managing different experimental data is a tendency from many projects.

These data play a fundamental role in all kinds of several scientific applications such as particle physics, high energy physics, data mining, climate modelling, earthquake engineering and astronomy [5, 6]. Storing such amount of data in the same location is difficult, even impossible. Moreover, an application may need data produced by another geographically remote application. For this reason, a grid is large scale resource sharing and problem solving mechanism in virtual organizations and is suitable for the above situation.

Data Grids provide geographically distributed storage resources to the large computational problems which requires evaluating and mining huge amounts of data. The Grid resources, including network bandwidth, data storage and computing facility are consumed by jobs [7]. Based on the system status and job requirements, the grid scheduler decides whether to run the incoming job or not. In data-intensive applications, a data location required by job impacts the performance and Grid scheduling decision greatly. Creating data replicas offers higher accessing speed than a single server as well as it provides wide decision space for grid scheduler to attain best performance from the job perspective.

Managing this data from centralized location will increases the data access time as well as it takes more time to execute the job. So replication is used to reduce data accessing time. The Data Replication is the process of creating and placing the copies of data. Creating replica contains the status of the replicated data and reproducing the structure, while the placement phase consists in selecting the suitable site of this new replica, based on the replication. The following figure 1 shows the Grid architecture;

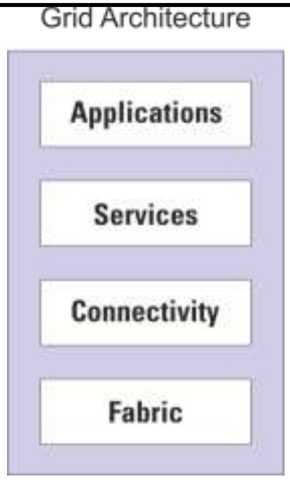

FIGURE 1 : GRID ARCHITECTURE

Replication in Data Management

Replication strategies can be classified into two categories such as static and dynamic. And further dynamic can be classified in to centralized and distributed as shown in following figure 2.

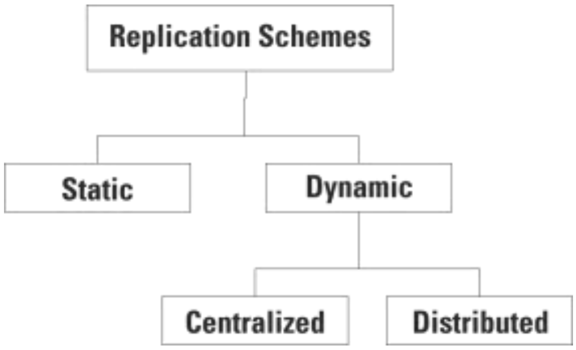

FIGURE 2 TYPES OF REPLICATION SCHEMES

In static method the created replica will exist until user deletes replica manually or up to the end of its duration [8]. The Static replication strategy is not suitable for large amount of data and it cannot adopt the frequent changes in user behaviour. But in other part the advantages of static replication strategies are: quick job scheduling and absence of overhead of dynamic algorithms [9].

But, dynamic strategies create as well as delete replicas based on the users file access pattern. The main advantages of dynamic strategies are the more appropriate dynamic replication and variable user requirements for these systems. On other part, significance of dynamic algorithm to transfer huge amount of data is the disadvantage because it leads to strain on the network's resources [10].

A dynamic replication scheme might be implemented either in a distributed or centralized approach. Drawback of this approach is the overload of central decision. In case of the decentralized manner, further synchronization is involved making the task hard [11]. Reliability, Availability, Adaptability, Scalability and high Performance are the advantages of replication strategies.

DMDR Algorithm (Data mining based Dynamic Replication Algorithm)

The proposed DMDR also based on network level locality. The modified algorithm replicates the files from neighbouring region and stores it in where files can be frequently accessed in future.

DMDR Region based algorithm

Inputs: Bandwidth, Grid Topology, Data and Storage Space

Output: Find best candidate for replication, job execution time, Load balancing, No of replications, RFA (remote file access), LFA (local file access), Network usage, Creation of new replica, Network utilization.

Method:

```
If (needed replica is not in a site)
{
        Search replica within the region
```

```
            If (replica is within the region)
                        Create list of candidate replica within the region
}
Else
{
            Search replica in other region
             Create list of candidate replica on other region
}

Fetch the most availability replica in candidate
replica list
If(there is sufficient space to store the new replica)
{
            Store it;
            Exit
}
 If(replica is in the same region)
{
            Exit; // avoid duplication in the same region
}
End process
```

IV WORKING PRINCIPLE OF DMDR ALGORITHM

Dynamic replication is optimization technique which reduces total access time, increase network bandwidth, availability of data by taking into account of different issues. The proposed algorithm called DMDR addressed these issues by using agents which needs to be addressed before replicating.

i) Replica creation: Data replication decides when and number of copies required to create. Replica is created in DMDRA when the file is not available on node.

ii) Replica placement: Once the replica is created, it decides to place the replica in suitable place where we can get fast access and less access latency. In DMDR algorithm replica is placed on the basis of the storage availability and file popularity.

iii) Replica selection: Next step is selecting best replicas from the group of available replicas. Best replica is chosen based on criteria such as availability status, workload of node, computing capacity of node and available bandwidth.

iv) Storage space: Once replica is chosen, we need to check whether required storage space is available or not. If it is not available then some special strategies need to be followed to place the replica.

v) Adaptability: In order to provide good results, replica strategy must be adaptive to the dynamic nature. While executions suppose the file is not available, immediately in DMDRA, the replica of file created, so that grid can adaptive to dynamic nature.

Experimental Results

The efficiency is calculated based on Performance Metrics, such as, shortest path, characteristic path length and neighbourhood connectivity. DMDRA is compared with one other dynamic replication strategy on the MATLAB. At the time of simulation in DMDRA, there exists only one replica in a region based on the popularity of the file.

Table 1: Job Execution Scenarios

Job Execution Scenarios	Values
Number of Post	200,400,600,800,1000
Number of post types	10
Algorithm Scheduling	Random
Size	2 GB

Table 2: Performance Metrics

Performance Metrics	Description
Execution Time	Job execution and Waiting Time
Network Usage	Network Utilization
Shortest Path	Specifies shortest path of given job

Shortest Path: The job execution time for random access pattern is shown below;

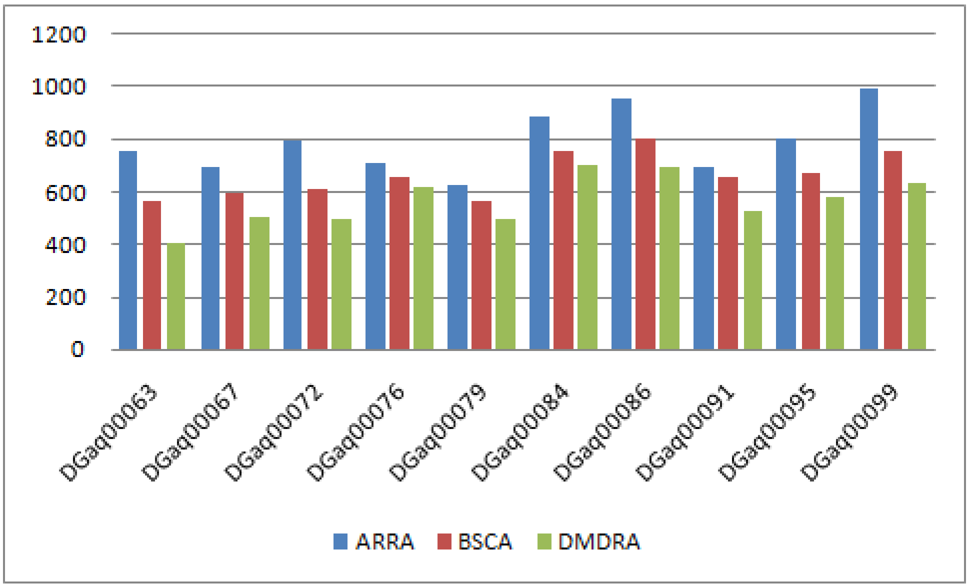

FIGURE 3 SHORTEST PATH: THE JOB EXECUTION TIME FOR RANDOM ACCESS PATTERN

V CONCLUSION

This paper implements the DMDR algorithm to address the problem in a data grid environment for optimization of replication. To attain good network, bandwidth consumption has been reduced. Our work concentrates on correlated between files and considers groups of correlated files as granularity for replication. Our work also compared with two other algorithms called Associated Replica Replacement Algorithm and Based on Support and Confidence Dynamic Replication algorithm. The result proves that Data mining based Dynamic Replication Algorithm gives better performance than other two algorithms.

REFERENCES

[1] Lakshmi.R, Antony Selvadoss Thanamani, "Performance Evolution of Dynamic Replication in a Data Grid using DMDR Algorithm, "International Journal of Engineering Research & Technology, ISSN:2278-0181, Vol.5 Issue 10, October 2016.

[2] TarekHamrouni, Sarra Slimani, et al, "A data mining correlated patterns-based periodic decentralized replication strategy for data grids", Elsevier, Volume 110, Pages 10-27, December 2015.

[3] Sean Carlistode Alvarenga, SylvioBarbonJr, et al., "Process mining and hierarchical clustering to help intrusion alert visualization", Elsevier, Volume 73, Pages 474-491, March 2018.

[4] HyunwooKim, SuckwonHong, et al., "Concentric diversification based on technological capabilities: Link analysis of products and technologies", Elsevier, Volume 118, Pages 246-257, May 2017.

[5] Sheida Dayyani, Mohammad Reza Khayyambashi, "A Comparative Study of Replication Techniques in Grid Computing Systems", International Journal of Computer Science and Information Security, Vol. 11, No. 9, September 2013.

[6] Sheida Dayyani, Mohammad Reza Khayyambashi, "A Novel Replication Strategy in Data Grid Environment with a Dynamic Threshold", International Journal of Computer Science Engineering (IJCSE), ISSN:2319-7323, Pages: 244-252, Vol.3 No.05, Sep 2014.

[7] Ming Tang, Bu-Sung LeeX, et al., "Combining Data Replication Algorithms and Job Scheduling Heuristics in the Data Grid", Springer, pp 381-390, 2005.

[8] TaoWang, Shihong Yao et al., "Dynamic replication to reduce access latency based on fuzzy logic system", Elsevier, Volume 60, Pages 48-57, May 2017.

[9] N. Mansouri, Gh. Dastghaibyfard, "Improving Data Grids Performance by using Modified Dynamic Hierarchical Replication Strategy", Iranian Journal of Electrical & Electronic Engineering, Vol. 10, No. 1, March 2014 .

[10] Babu R., Rao S., "Dynamic Checkpoint Data Replication Strategy in Computational Grid", Springer, New Delhi, Vol 266, pp 95-105, March 2014.

[11] A.Vergnol, J.Sprooten, et al., "Line overload alleviation through corrective control in presence of wind energy", Elsevier, Volume 81, Issue 7, Pages 1583-1591, July 2011.

About the Authors

Ms. Prema .R is presently working as Head, Dept of Computer Applications, Indo Asian Women's Degree College, India (affiliated to Bangalore University,Bangalore . She is pursuing her Ph.D in Bahrathiar University, Coimbatore. Her areas of interest include Algorithm Analysis and design, Data Mining, Data Sturctures.Operation Research. She has 10 years of teaching experience.

Dr. Antony Selvadoss Thanamani is presently working as Professor and Head, Dept of Computer Science, NGM College, Coimbatore, India (affiliated to Bharathiar University, Coimbatore). He has published more than 100 papers in international/

national journals and conferences. He has authored many books on recent trends in Information Technology. His areas of interest include ELearning, Knowledge Management, Data Mining, Networking, Parallel and Distributed Computing. He has to his credit 32 years of teaching and research experience. He is a senior member of International Association of Computer Science and Information Technology, Singapore and Active.