ANALYSIS ON PERFORMANCE WITH COMBINATION OF SCHEDULING AND REPLICATION IN DATA GRIDS

Lakshmi C S*

PG Student
Department of Computer Science and Engineering
Karunya University
Coimbatore, Tamil Nadu, 641114, India
sathishlakshmi18@gmail.com

Arul Xavier V. M

Assistant Professor
Department of Computer Science and Engineering
Karunya University
Coimbatore, Tamil Nadu, 641114, India
arulvmax@gmail.com

Abstract

Data grid is the storage component of a grid environment. Amount of data transferred among nodes can be reduced by submitting the jobs to the nodes that having maximum requested files by scheduling and reducing the access latency by initiating data replication strategies. All techniques of scheduling and replication address some issues with performance metrics. In this paper, different combination of scheduling and replication is studied to find out which one is acquiring better evaluation with attributes like network bandwidth, dynamic behavior of user, latency and cost of replication.

Keywords: *Grid computing, Data grid, Scheduling, Replication*

1. Introduction

Current trends in technology of scientific disciplines are going through a bulk amount of data collections. The huge amount of data and computation involved problems regarding data access, processing and distribution. Managing the data is a great challenge to meet. Data grid is a solution for this. According to Kesselman and Foster Grid Computing defines as “a growing technology that facilitates the executions of large-scale resource intensive applications on geographically distributed computing resources”. It facilitates flexible, secure, coordinated large scale resource sharing among individuals, institutions and resources. A classification of grid as follows – “A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities” and “A data grid deals with data – the controlled sharing and management of large amounts of distributed data.”

While dealing with the data grid since it is distributed the performance of networks plays an important role in scheduling and replication. The Fig 1 shows the data grid architecture supported for the combination of replication and scheduling [5] represents an organisation unit which group of sites that are geographically close to each other comprises the computer which are connected by a high bandwidth, the performance of system is underlying available network bandwidth and data access latency, especially the hierarchy of bandwidth appears.

2. Motivation

In distributed environment of grid, availability of data, response time, access cost, bandwidth consumption, reliability, scalability are some very important performance metrics to be evaluated. The motivation of this survey is to explore the issues while providing with combination of replication and scheduling strategies in data grid environment and the compare their performances

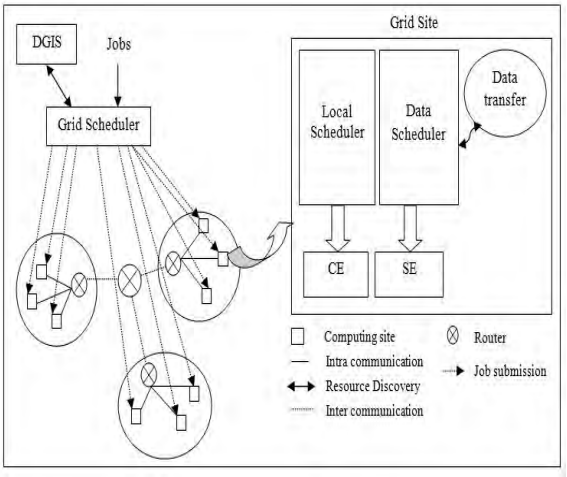

Fig 1.Data Grid Architecture

3. History

The need for data grids was first (categorized) recognized when the scientific community concerning climate as modeling, where terabyte and petabyte sized data sets were data transferring is between sites. More recent research requirements for data grids have been driven by the Large Hadron Collider (LHC) at CERN. Other uses for data grids involve governments, hospitals, schools and businesses where efforts are taking place to improve services and reduce costs by providing access to dispersed and separate data systems through the use of data grids.

The data grid has also been defined more recently in terms of usability; what must a data grid be able to do in order for it to be useful to the scientific community. Several criteria's are, users should be able to search and discover applicable resources within the data grid from datasets, then users should be able to locate datasets within the data grid that are most suitable for their requirement from amongst numerous replicas, also users should be able to transfer and move large datasets between points in a short amount of time, then the data grid should provide a means to manage multiple copies of datasets within the data grid. And finally, the data grid should provide security with user access controls within the data grid, i.e. which users are allowed to access which data. The data grid is an evolving technology that continues to change and grow to meet the needs of an expanding community.

4. Issues involved in grid environment

Certain issue arises in grid environment due the following reasons:

- *Dynamic behavior:* The users can enter and leave the grid at any time. So the number of participants also may vary at any time .In order to provide better results it should adaptive in kind.
- *Grid architecture:* A data grid can be supported by much architecture like multi-tier, tree like structure, graph topology, peer to peer topology or it can be hybrid in nature. According to this the scheduling and replication are performed.
- *Available storage space:* As the grid environment is handling with huge amount of data, availability of storage space is not sufficient to store replicas.
- *Cost of replication:* There is a price to be paid when data is replicated. The files in grid can change by users in any time, so to maintain consistency is a problem. To specifically evaluate this criteria the strategy must check cost of replication.

5. Replication and scheduling techniques

We have categorised the combination of replication and scheduling into several groups according to their increment in performances and evaluated the metrics. The Table 1 and Table 2 chosen for bandwidth configuration and job configuration have been used for evaluation in the following methods discussed below. There are three regions to analyse the configuration. Table 1 shows values for bandwidth configuration (assumes Mbps is Megabit per second) and Table 2 shows job configuration. The analysis is chosen for 500 jobs in which 6 types of job types are taken with number of file access types as 10.The work done is discussed individually in this section.

Table 1: Bandwidth Configuration

Parameter	Value
Inter-LAN	1000 Mbps
Intra-LAN	100 Mbps
Intra-Region	10 Mbps

Table 2: Job Configuration

Parameter	Value
Number of job types	6
Number of file access per jobs	10
Size of single file	1 GB
Total size of files	100 GB

5.1 Hierarchical Replication Scheduling & Hierarchical Cluster Scheduling

In this job scheduling policy called Hierarchical Cluster based Scheduling (HCS) and a dynamic replication Strategy called Hierarchical Replication Strategy (HRS) to improve data access in data grid. They considered the bottleneck to support fast data access in grids. To address these problem two aspects [6] of inter-communication: Job Scheduling and replication mechanism is introduced. HCS uses hierarchical scheduling to reduce search time for a suitable node with data. Scheduling depends on the criteria's – the location of required data, access cost, and job queue length of the computing node. To perform replication, HRS increases the chance of accessing the most nearest node.

5.2 Hierarchical scheduling and Replication strategy

While analyzing through replication algorithm for a 3-level hierarchical structure and scheduling algorithm. Replication [1] checks for replica feasibility and Scheduling determines the most appropriate region, LAN & site respectively .So most requested files available then significantly reduce total transfer time & bandwidth i.e. network traffic. If there is no enough space for replica to accommodate then it performs the deletion of file that have low cost of transfer i.e. the file available in local LAN. While scheduling is performed, the region having maximum request of files has been chosen as per it minimize the data transfer time hence reduces the job execution time.

5.3 Enhancement of Replication Strategy & Enhancement of Scheduling Strategy

In this a new combination of data replication and job scheduling algorithm for 2-layer hierarchical structure by reducing job data access time correspondingly with job execution time. Scheduling evaluates the required data and job queue length of the node called as Enhanced of Scheduling Strategy (ESS) and Enhancement of Replication Strategy (ERS) which takes bandwidth as a factor to evaluate replica selection and placement. It also increases the chances to access data at nearby nodes and prevent full storage. It achieves good network utilization, reduce data access time through the scheduling that not only consider computation capability, job type, data location also the cluster information in job placement decision.ESS[2] always scheduled the needed data and ERS for replica management.

5.4 Combined Scheduling Strategy & Dynamic Hierarchical Replication Algorithm

In this proposed two algorithms: first, a novel job scheduling algorithm called Combined Scheduling Strategy (CSS) that considers the number of jobs waiting in queue, the location of required data for the job, and computational capability; second, a dynamic data replication strategy called Dynamic Hierarchical Replication Algorithm (DHRA) that improves file access time.CSS [3] reduce job execution time by reducing data access time by finding the best region i.e. region with most requested files hence reduces the total transfer time corresponding with network traffic. DHRA stores each replica in best site, also selects replica location for executing jobs by evaluating the number of requests that are waiting in the queue, data transfer time and access latency.CSS and DHRA have less job execution time.

5.5 Enhanced Dynamic Hierarchical Replication & Weighted Scheduling Strategy

In this Weighted Scheduling Strategy (WSS) that uses hierarchical scheduling to reduce the search time for an appropriate computing node and then a dynamic data replication strategy, called Enhanced Dynamic Hierarchical Replication (EDHR) that improves file access time. The EDHR [4] strategy improves by using an economic model for file deletion when there is not enough space for replicas. The economic model is based on the future value of a data file. EDHR considers the frequency of requests of the replica and the last time the replica was requested for replica placement and also minimizes access latency by selecting the best replica when various sites hold replicas. The replica selection strategy selects the best replica location for the users' running jobs by considering the replica requests that are waiting in the storage and data transfer time.

6. Features comparison and its tabular representation

Now to summarize all the algorithms discussed in section 5 in such a way to identify the different parameters consider whether it meets the performance metrics by evaluating with the advantages and disadvantages of the combination in Table 1 accordingly with configuration mentioned above in section 5.

6.1 Benefits of scheduling and replication algorithms

- Availability: Generally in distributed database environment replication is the only way to improve data availability.
- Reliability: When replication increases the availability, the reliability is improved. As number of replicas is more as chance of user request is served properly hence system is reliable.
- Adaptability: This is a parameter provided by replication since the nature of grid is dynamic as nodes keep entering and leaving the grid frequently so algorithm must be adaptive to provide support to all nodes in data grid.
- Performance: AS availability of data increases the performance of the data grid environment also increases.

6.2 Different parameters and their importance

To gain the benefits mentioned above, a set of parameters is discussed. All replication strategies use any subset of these parameters.

- Reduced bandwidth consumption
- Less maintenance cost
- Reduced access latency
- Job execution time
- Balanced workload
- Quality assurance
- Strategic replica placement

Only few strategies consider providing fault tolerance and quality assurance. Idea to place replica closer to the user, minimize response time and job execution time. This will increase system throughput. Almost all replication algorithms try to reduce the access latency thus reducing the job response time and hence increase the performance of the data grids.

Similarly almost all the replication strategies try to reduce the bandwidth consumption to improve the availability of data and performance of the system. The target is to keep the data as close to user as possible, so that data can be accessed efficiently. Through the graph while analyzing performance, it is cleared that combination of Enhanced Hierarchical Replication Strategy and Weighted Scheduling Strategy in data grid environment evaluates better performance as the number of jobs increased.

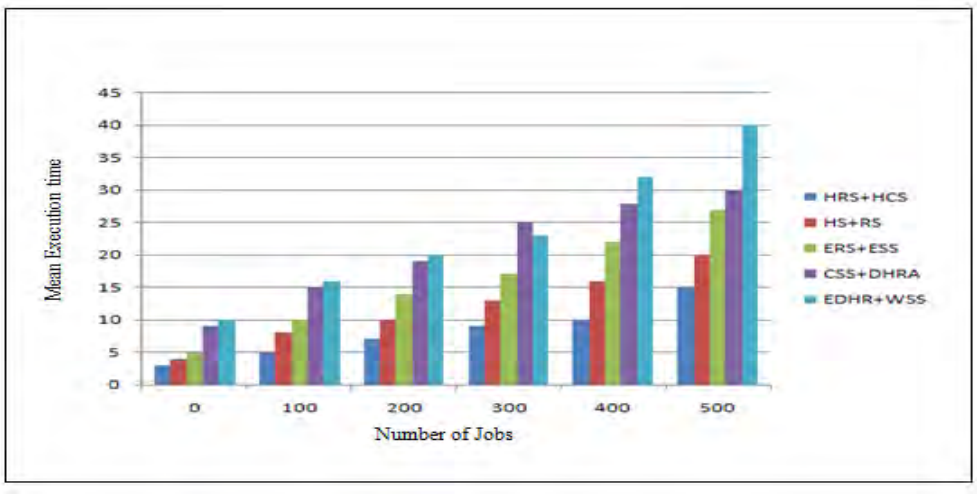

Fig 6.1 Scheduling performance with various replication strategies

Table 3: Features and its comparison

Sl. No	Method	Factors influences	Criteria	Advantages	Disadvantages	Performance
1	HRS& HCS	-Storage space -Bandwidth	Cluster grid topology	-reduce job execution time -increase robustness	Arise load balancing issues	Less performance
2	Hierarchical scheduling and Replication strategy	Bandwidth	3 - level hierarchical architecture	-reduce network traffic -reduce file access time	Not much scalable	Overall performance is better with 30%
3	ERS& ESS	-CPU workload -Computational capability -Network load	2 – layered hierarchical structure	-prevent full storage of data	Fault tolerance occurs	Total execution time is 10% & 30% faster
4	CSS& DHRA	-Data locality -CPU workload -Computational capability -Network load -Response time	3 - level hierarchical architecture	-reduce job execution time	Cost estimation is not effectively done	Increases as both file size & number of jobs increases
5	EDHR& W SS	-Bandwidth -Network locality -Network workload	3 - level hierarchical architecture	-Minimize data access time -avoid unnecessary replication	Replica consistency & management to be maintained	Performance improved by 40% in execution time

Conclusion

We have analyzed five appropriate techniques on the combination of scheduling and replication strategy algorithms for a data grid environment. It can be evaluated that different strategies have presented their own terms for the evaluation of proposed methods. The issues in respective papers have been resolved and discussed. From this analysis in data grid environment, it can be seen that there is still a lot of work done in the field of data replication in data grid environment. It has been observed that there exists no standard architecture for a data grid environment, mostly referred with hierarchical architecture but generally is more realistic. While plotting the graph the performance is better with EDHR and WSS. Data replication is a frequently used technique that can enhance the data availability and access latency. Since a grid environment is dynamic, network latency and user behavior may change. To address these issues a well-designed replication and scheduling strategies should be combined.

References

[1] A.Horri, R. Sepahvand, et al,(2008): A hierarchical scheduling and replication strategy, International Journal of Computer Science and Network Security 8.

[2] F. Jolfaei and A. Haghighat (2012): Improvement of Job Scheduling And Tow Level Data Replication Strategies In Data Grid International Journal of Mobile Network Communications & Telematics (IJMNCT) Vol.2, No.3.

[3] N. Mansouri,G.Dastghaibyfard (2013): Combination of data replication and scheduling algorithm for improving data availability in Data Grids, Springer Science, Pages 711-722.

[4] N.Mansouri, G.Dastghaibyfard (2013): Enhanced Dynamic Hierarchical Replication and Weighted Scheduling Strategy in Data Grid, J. Parallel Distrib. Comput. 73 Pages 534-543.

[5] Park S-M, Kim J-H, Go Y-B, Yoon W-S (2003) Dynamic grid replication strategy based on Internet hierarchy. In: International workshop on grid and cooperative computing. Lecture notes on computer science, pp 1324–1331.

[6] R.Chang, Jih-Sheng Chang,et al. (2007): Job scheduling and data replication on data grids, Future Generation Computer Systems, Volume 23, Pages 846-860.

[7] V. Andronikou, K. Mamouras, et al. (2012), Dynamic QoS-aware data replication in grid environments based on data importance, Future Generation Computer Systems 28 (3) 544–553.

Lakshmi C. S completed B.Tech in Information Technology from The Rajaas Engineering College, Anna University, Chennai. Currently pursing M.Tech in Computer Science and Engineering in Karunya University, Coimbatore. Areas of research are in Cloud Computing, Grid Computing.

Arul Xavier V. M completed B.Tech in Information Technology from Velammal Engineering College, Madras University, Chennai. He obtained his M.E. in Computer Science and Engineering from Noorul Islam College of Engineering, Anna University. He is currently working as Assistant Professor - Department of Computer Science and Engineering in Karunya University, Coimbatore. Presently, he is doing Ph.D under Anna University, Coimbatore, and his areas of interest in research are Wireless Networks, Distributed Computing and Grid Computing.