POSSIBILITIES OF OPTIMIZATION IN SCHEDULING AT GRID ENVIRONMENTS: A VIEW FROM DISTRIBUTED DATA MINING

P. Vishvapathi[1], Dr. S.Ramachandram[2], Dr.A.Govardhan[3]
CMR Engineering College, Hyderabad, India[1]
College of Engg. Osmania University, Hyderabad, India[2]
JNT University, Hyderabad, India[3]
vpujala@gmail.com, schandram@gmail.com, govardhan_cse@yahoo.co.in

Abstract: *In this information era grid computing has emerged as an important new branch of distributed data mining. It is majorly focused on large-scale data sets, high-performance and utilization. The distributed data mining datasets require resources to be heterogeneous and distributed. In many distributed environment oriented applications it is necessary to perform the analysis of very large data sets. Generally, large-scale data sets are geographically distributed and structurally complex. In this paper we are discussing about the complexity involved in data transferring from grid to nodes and it's scheduling. We are focusing on scheduling in a grid environments at architecture level which provides effective computational support for distributed applications and environments in knowledge discovery domain. Further this paper is focusing careful attention to computing and communication resources within existing infrastructure*.

Keywords: **Distributed data mining; Grid environments; Communication resources; Architectures; Topologies.**

1. Introduction

"Grid is a type of parallel and distributed system that enables the sharing, selection, and aggregation of geographically distributed "autonomous" resources dynamically at runtime depending on their availability, capability, performance, cost, and users' quality-of-service requirements." –(Beker et.al, 2002)

In these days, the distributed data mining (DDM) has been notified as one of the major and important technology in knowledge discovery industries with the rapid development of information access technologies. In the knowledge industries such as science, engineering and medicine, distributed data mining has been utilized as tool for automating the process of analyzing and interpreting large datasets (Holtman et al., 2001)**.** The knowledge discovery process is extraction of new data patterns, previously unknown information from large datasets (Stockinger et.al 2001). In order to extract data patterns in knowledge discovery process the data mining and machine learning methods have to apply. The data mining process is interactive and controllable by the user, so that the process required interactive environment to set certain parameters to access and analyze the large datasets (Holtman et al., 2001). Information overloading, organizing, communicating is still an open ended research problems in knowledge discovery and distributed environments, and working on very large datasets using conventional computing machines are time consuming and huge cost process to get effective results.

Generally, a distributed environment is use to store, process large-sets of data and connected to communicate using internet, intranets, local and wide area networks, adhoc wireless and sensor networks (Mckinley, et.al, 1996). One of the major focuses of grid environments are learning patterns from these distributed data sources and performance for machine learning tasks for mining and knowledge discovery.

The distributed data mining is one of the important topics to solve many of these problems. Data has geographically distributed and get updated frequently by the users of several organizations and individuals in various locations and the data belongs to different domains (Krauter, et.al, 2002). The data stored in different high-end servers and it is represented in various formats and storage methods. Grids are taken care of data resource access to distributed computing, management and analysis also allowing data intensive applications to improve significantly data access in part of quality of service (QoS) (Papazoglou, M. P. 2003). As discussed the grids which are handling large-sets of data spatially distributed and affected to temporal updatation are categorized as generic data management systems and particularly called as data mining grids, it involve a great number of challenges (Papazoglou, M. P. 2003). This paper is discussing state-of-the-art in models organizational data grids, elements of distributed data and discuss data transferring from grid to nodes and at scheduling, where we have a clear scope to apply mining techniques to optimize the communication and data resources.

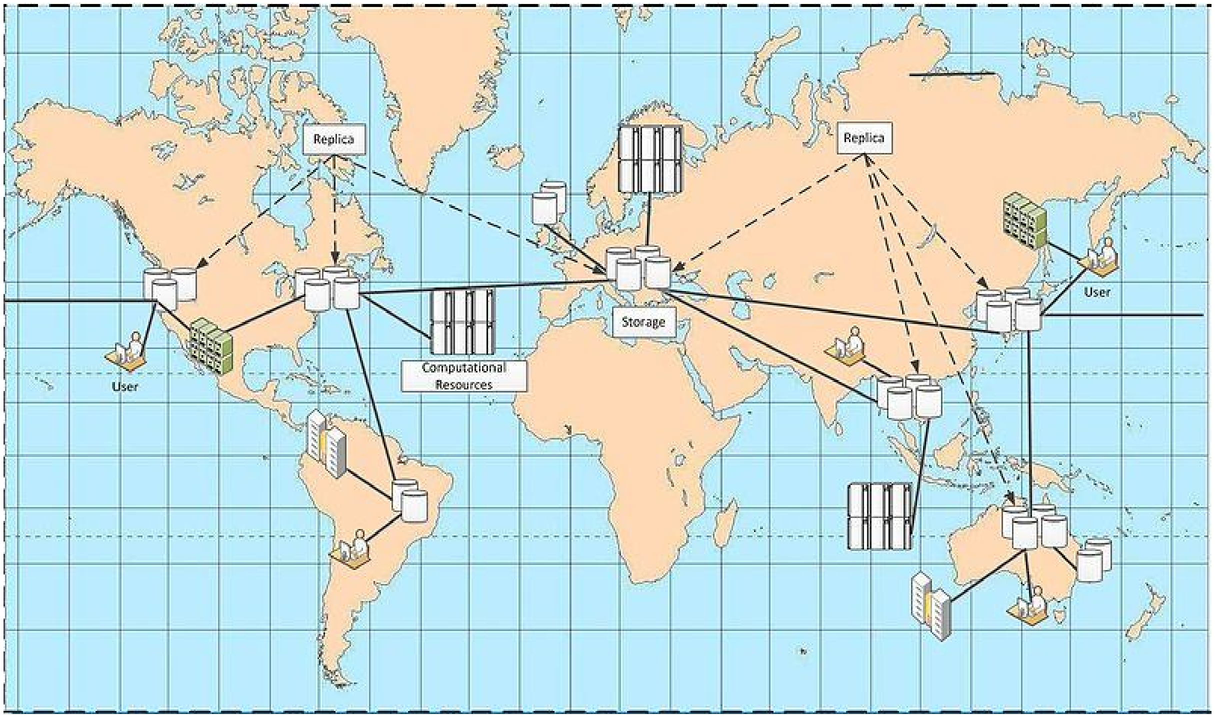


Figure 1: Overview of a High level Data Grid (Allcock, B; Chervenak, A; Foster, I; et al. 2005; Venugopal, S; Rajkumar; Ramamohanarao, K. 2006)

2. History

In mid 90s grid environments have established to link supercomputing sites to provide resources to analyze several high-performance applications (Berman et.al 2003). Before that i.e. early 80s the researchers from different disciplines have proposed and started to work on large-scale computational infrastructure is to play a key role in solving computational problems to achieve significant results. The researches have introduced two fundamental concepts of grid computing is that “coordination and distribution” to solve such problems are inherent in multidisciplinary areas which are having geographically dispersed collaborations (Berman et.al 2003). In part collaborations researchers have developed FAFNER (Factoring via Network-Enabled Recursion) was developed for factoring large prime numbers (De Roure et.al 2003) and the contributors supported by providing computational power to the web servers and the related calculations were done by a CGI script at the each server.

These are all to enrich the computational performance on a range of high-performance applications (DeFanti et.al, 1996).In the history of grid evolution generally the I-WAY is to be considered as the first modern grid. It has started as a project in mid 90s and was experimented on high-end computers which have advanced visualization environments numerous linking with high-performance network (DeFanti et.al, 1996).

At present the grid infrastructures have been allowing to couple more specialized supercomputing centres. The grid environments are now more ubiquitous because increasing of utilization computing techniques by definitions and standards (Venugopal et.al, 2006). For instance, in present scenario the grid architectures are designing based on standards of geographical policies and it supposed to meet various requirements from business, legal and social, security, here the Web Services technologies are also one of the major trait in the design to show the performance and application aspects. Also the grid architectures have to justify the local needs from the aspects of global changes. Such that grids are designing and building on the accessibility of the internet to access geographically distributed resources for effective utilization (Sumithra, R and Paul, S, 2010). For that the grid computing technologies are providing a platform for the next generation of Internet-enabled HPC solutions, such as workflow specification, enactment and execution are essential supports for business process management, QoS negotiation and application enabling etc (Ranganathan et.al 2002; Venugopal et.al, 2005).

3. Models of Organizational Data Grids and Services

A grid is to be showed as an integrated computational and collaborative environment and it is a platform to perform high-level activities like data analysis and data transfer for various application domains around the world (Sumithra, R and Paul, S 2010). Generally the users are interacting with a grid resource to solve computational problems where the interaction is extended to middleware development, advanced networking and storage management of resources and it turns in to action of resource discovery, scheduling, and data processing, performance of application jobs on the distributed resources which grid resources (Ranganathan, et.al, 2002; Venugopal, et.al 2006).From the end-user point of view, Grids can be used to provide the following types of services.

Computational services: In grid world some of the major computational grids are NASA IPG , the World Wide Grid, and NSF TeraGrid. Computational grids often provide computational services such as data privacy and security for executing web applications and application jobs on distributed resources. Those resources can be computationally individual and collective(Berman, et.al 2003)**.**

Data services: The computational grid services carry out the processing of data sets. The data services will taken care of the secure access of distributed data sets, sharing, processing and also it manages the data scalable storage i.e. it may replicate, catalogue and creates an illusion of mass storage from different data sets in different spaces (Foster, et.al. 2002; Foster, et.al., 2003)**.**

Application services: Application services are influenced by emerging web technologies. The application service is the combination of data and computational services. It manages the applications and provides secured access to the remote software applications to access the libraries without any disturbance to other services i.e. transparently for example NetSolve (Foster, et.al. 2002; Foster, et. al., 2003).

Information services: it is a major part in Web to depart the data for many scientific experiments. It handles low-level details like representation, storage, access, sharing and maintenance of the information by the extraction and presentation of data with meaning with the combined services of computational, data, or application services (Foster, et.al. 2002;

Foster, et.al., 2003).

Knowledge services: "*Knowledge is an understood as information applied to achieve a goal, solve a problem, or execute a decision*" (Baker, M and Rajkumar, B and Laforenza, D., 2002)**.** The end-users expect the services of the information within a prescribed format i.e. easy usable way. This is concerned the knowledge services are acquire, use, retrieve, publish, assist to the users to reach the goal and objectives like data mining techniques for building a new knowledge-bases automatically **(**Baker et.al 2003).

The researchers have kept in many of these aspects and elements of a data grid while development of following modern grid architectures especially, the focus on those topologies which are involved in setting up large computational infrastructure.

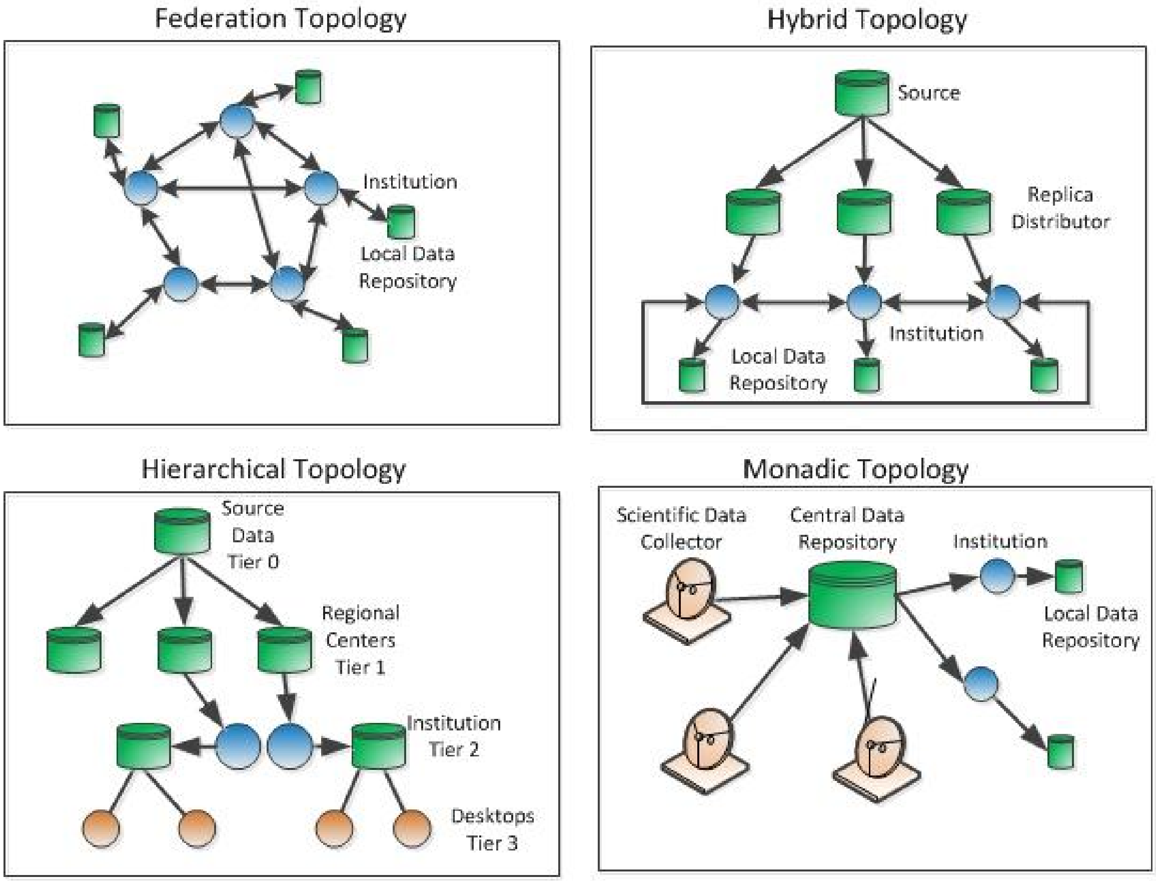


Figure 2: Models of Organizational Data Grids (Venugopal, S. and Rajkumar, B and Ramamohanarao, K, 2006).

Federation: This model is used in BioInformatics research Network (BIRN) 2005 in United States(Biogrid, 2005). This model is developed by institutions who are working on existing databases to share their data (Rajasekar et al., 2004)**.** In this data grid each institute having an authentic control to access the data. After receiving an authentic request from a federation institute. The data grid transfers the data to the particular request according to their degree of control and access. The autonomy of control and degree of authentication of each access point is based on the constraints, replication and synchronisation of the registered users i.e. institutions.

Hybrid: This model is combination of the above models and production usage i.e. this model has been developed in the need of collaborations and analysis sharing (Venugopal, S. and Rajkumar, B and Ramamohanarao, K, 2006).

Hierarchical: In this architecture the data has distributed across worldwide collaborations from a single source. This model have tiered infrastructure and distribution of content, i.e. transfer of data from the CERN to various locations established as regional centres and from there to national and institutional centres or groups of researchers (Aderholz , M. et.al. 2000). This tiered model has been proposed in CERN, the main source of storage and computing placed in CERN and tier-1, tier-2 centre has maintain the sufficient bandwidth, storage and computing requirements to conduct robust experiments on massive distributed data. This data identifies based on the metadata and since the data is consistent and single source it is simpler to maintain (Aderholz , M. et.al. 2000; Ahamed, B.B., and Hariharan, S, 2011).

Monadic: This grid form having central repository and a single point of access provides the data to the user queries and answers through a centralised interface like a web portal which also required the user authentication (Venugopal, S. and Rajkumar, B 2005). The data has collected from various spatially distributed sources and the interface has been connected through sensor networks to the access point. The NEESgrid (Network for Earthquake Engineering Simulation) model is an example of Moandic architecture (Venugopal, S. and Rajkumar, B and Ramamohanarao, K, 2006). In this model the central repository does not improves the local data instead of in fault tolerance cases it replicated itself. Thus, this model serves better in overhead of replications (Krauter, K and Rajkumar,B and Maheswaran, M 2002).

3.1 Elements of a Data Grid

***3.1.a.* Organisation:** the data sources are organized in a system is based on various organizational characteristics and requirements of Data Grids like source of data, single or distributed, size of data and sharing mode. These characteristics are manifest in different methods and it is central to every Data Grid (Venugopal, et.al, 2006)**.**

***3.1.b.* Data transport:** Data transport is the one a fundamental concept of a Data Grid. It deals with security, access controls and management of data transfer or communication between nodes across resources (Allcock, et.al, 2005).

***3.1.c.* Replication:** Data replication is bounded to bandwidth and storage size of data available at different sites while accessing by the authentic members of geographically distributed and collaborations of the grid. In part of its' responsibilities it ensures scalability of collaboration, reliability of data access and preservation of the bandwidth where an underling requirements of transfer and storage (Casanova, et.al 2000; Congiusta, et.al, 2003).

3.1.d. **Scheduling.** This element deals with bandwidth availability and latency of data transfer between computational node to storage resources where the requested job has submitted and retrieved as per job requirements. Scheduling of data-intensive jobs of large data sets which are geographically-distributed with multiple replicas is different from regular computational jobs (Casanova, et.al 2000; Congiusta, et.al, 2003).

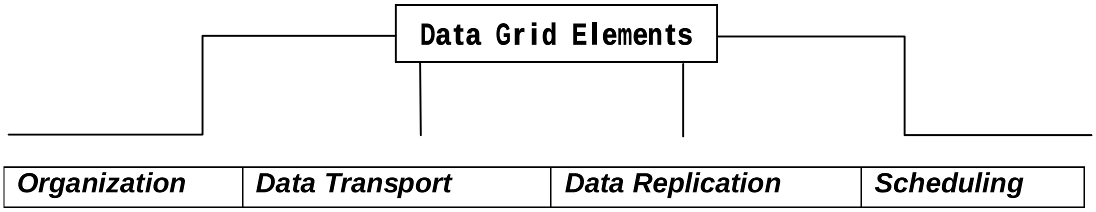
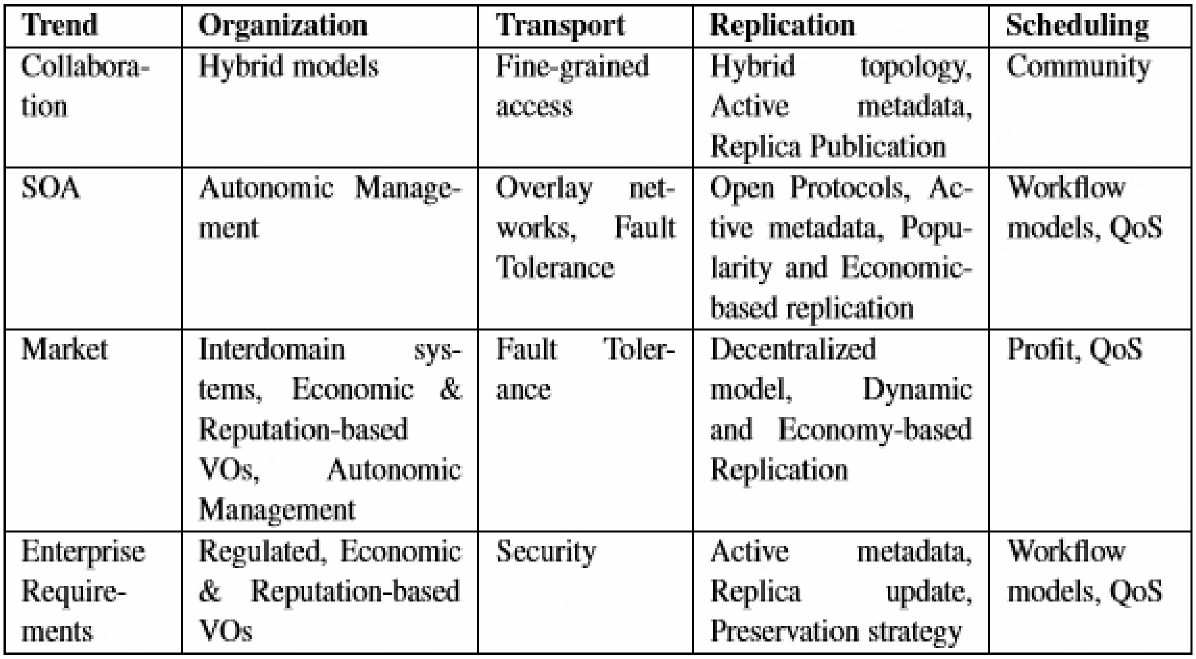


Figure 3: General Elements in Data Grids.

4. Trends in Services and Scheduling in Grid Technologies

Data grid researchers have investigated on replication and scheduling mechanisms in data grid technologies and presented as “*the Data Grid technologies are only beginning to be employed in production environments and are still evolving to meet present and future requirements”* (Yu and Rajkumar, 2004; Venugopal, S. and Rajkumar, B and Ramamohanarao, K, 2006). Also kept challenges for future researchers in the space of emerging trends of data grids technologies with respect to needs of information market noted as key properties are:



Trend	Organization	Transport	Replication	Scheduling
Collaboration	Hybrid models	Fine-grained access	Hybrid topology, Active metadata, Replica Publication	Community
SOA	Autonomic Management	Overlay networks, Fault Tolerance	Open Protocols, Active metadata, Popularity and Economic-based replication	Workflow models, QoS
Market	Interdomain systems, Economic & Reputation-based VOs, Autonomic Management	Fault Tolerance	Decentralized model, Dynamic and Economy-based Replication	Profit, QoS
Enterprise Requirements	Regulated, Economic & Reputation-based VOs	Security	Active metadata, Replica update, Preservation strategy	Workflow models, QoS

Figure 4: Key Properties of Data Grid technologies (Venugopal, S. and Rajkumar, B and Ramamohanarao, K, 2006)

Increasing collaboration is all about the hybrid topologies are functioning on the basis of sharing resources by participants of various communities. It should ensure that the resource allocation must fair shares to everyone. This requirement brought the community-based schedulers have to assign quotas to participants based on priorities and resource availability (Papazoglou, M. P., and Georgakopoulos, D., 2003).

Service Oriented Architectures (SOAs) is another key aspect of the trend and web services. The main difference between SOA and the client and server architecture is SOA having the "*ability for web services to be composed of other services by building on standard protocols and invocation mechanisms"* (Venugopal, S. and Rajkumar, B and Ramamohanarao, K, 2006).
To build a high level of transparency in SOA required selection of right services and constituent technologies with required QoS parameters. It will create an impact on both replication and resource allocation (Papazoglou, M. P., 2003; Papazoglou, M. P., and Georgakopoulos, D., 2003).

Market mechanisms are the trends of socio dynamic of nature of consumers and it get controlled by the user-defined QoS parameters like a dynamic system by Lin (2005) which drives based on cost of data movement. In this, it functions on basis of demand and supply patterns to decide the value of the resources which can be computational either data-intensive.

In this, the domain based content providers will get incentives for their resource consumption outside of their domain also provides ways to open new applications and it will be based on SOAs (Papazoglou, M. P., 2003). These types of mechanisms are having broad scope and consumers from specific domains to various. The utility functions will be guided by the policies of resource allocation and replication due to the user-defined QoS parameters (Venugopal, S. and Rajkumar, B., 2005; Venugopal, S. and Rajkumar, B and Ramamohanarao, K, 2006).

Enterprise requirements deals with massive volumes of distributed data that can be more than terra bytes for business functions, production systems. The major challenge is now to organize the massive data in a time-bound extraction format i.e. more precision required in the extracted data from the storage devices. In this type of problems also need to consider reliability, security and privacy for that required consistent computational model for enterprise computing in data grids (Venugopal, S. and Rajkumar, B and Ramamohanarao, K, 2006).

Above table shows the trend and key features of the data grids based on the elements of organisation, data transport, data replication and scheduling.

4.1 Scheduling in Grid Mechanisms

According to Magowan (2003), in existing grid mechanisms some of major challenges are replication, data transfer and scheduling to work with new distributed data sources such as distributed databases (Magowan, 2003; Venugopal and Rajkumar, 2005; Venugopal, S. and

Rajkumar, B and Ramamohanarao, K, 2006). The large datasets are scattered at geographically and distributed. In the scheduling process it handles the data intensive jobs which are different from the computational jobs. Also it considers the factor of bandwidth for communicating between computational nodes to storage place of resources as discussed above i.e. in trends and architectures. As mentioned above the end-users deals with web applications and they create huge data, in such case the major focus of SOAs in grids and have to satisfy the requirements of the Web Service Resource Framework (WSRF) (Foster et al., 2005) specified by the grid standards community.

According to the current literature, the scheduling in grid mechanism have been frames on several factors such as application areas, scope, utility and locality. Here we are looking this scheduling from large-scale communication of data between nodes to users. This process placed on readily available documents from the resources i.e. data sources to the web pages i.e. applications and other sources i.e. users of the grid enviroments.

In scheduling, according to application models the strategies have composed by the target requirements like the processes at global grids, independent tasks with certain conditions such as deadline for an application and workflow process i.e. a sequence of tasks which the job is dependent on the result of preceded (Venugopal, S. and Rajkumar, B and Ramamohanarao, K, 2006). And then the scope is dependent on the application model and relates to the user’s perspective if it is individual. Otherwise it is independent in utilization of resources. While scheduling based on the application model and its service we should be aware of the fluctuations like resource availability and load of the jobs (Wasson and Humphrey, 2003; Dumitrescu and Foster, 2004; Dumitrescu, C L and Wilde, Michael and Foster, I, 2005;, Dumitrescu, C and Raicu, I and Foster, I, 2005).

The researches shown in scheduling when a job is coupled to data replication while execution of a compute node fetches the data from the remote storage and creates a copy of data at computation to get a quick access in future requests.

These copies were managed by the storage management schemes such as LRU (Least Recently Used) and FIFO (First In First Out). This kind of management can be prejudiced by the requirements of compute nodes. Also there is high probability reduction of storage space, the process of creation replica, registering into catalogue creates burden and low performance in job execution. In decoupled scheduler, the storage space is required during execution of job and the space requirement is transient, so that it shows the performance comparatively increases and reduce the complexity in designing of algorithms (Toporkov, Victor V and Tselishchev, Alexey, 2010; Foster, Ian and V"ockler, Jens and Wilde, Michael and Zhao, Yong, 2003; Foster, Ian and Kesselman, Carl and Nick, Jeffrey M and Tuecke, Steven, 2002).

However, to bring the QoS in the grid mechanism we need to optimize the communication resources and load on storage space, for that we are proposing a domain based data organizing manifest. In this manifest the scattered geographically distributed data’s meta information will be organized as schema as per the locality and geographical utility. This will reduce the load and creates balance job execution and communication. Once the user submitted a job to schedule it verifies the locality features as described in the organizing manifest in schedule taxonomy then executes at compute node such that it is reducing the load and reaches the objectives with minimum resources.

5. Future Work

In future we would conduct experiments on the proposed model and present the results, the future scope of this work is to integrate in existing health grid environment to produce optimized results. Therefore, this can provide efficient results and reduce the cost of computing, communication by decreasing data processing time at scheduling and replication, and optimizing resources and distributing workloads, thereby we can achieve much faster results on large operations and at lower costs.

The Distributed data mining DDM) systems uses high-end multiple processors and data sources to speed up job execution of data mining algorithms and enables efficient data distribution on within grid environments. Also we would extend the scope of this work is to automatic knowledge extraction with implementation of distributed data mining tasks on grid environments as the main aim of grid computing i.e. provide a platform to organizations and developers to create distributed computing environments that can utilize computing resources as per their requirements.

References

1. Aderholz , M. et.al. (2000). Monarc project phase2 report. Tech. rep., CERN.Mar.
2. Ahamed, Bagrudeen B and Hariharan, S., (2011), A Survey On Distributed Data Mining Process Via Grid, *International Journal of Database Theory and Application*, 4(3) pp77-90
3. Allcock, B and Chervenak, A and Foster, I and Kesselman, C and Livny, M, (2005), Data Grid tools: enabling science on big distributed data, *Journal of Physics: Conference Series*, 16(1), pp.571, IOP Publishing
4. Baker, M and Rajkumar, B and Laforenza, D, (2002) Grids and Grid technologies for wide-area distributed computing, *Software: Practice and Experience*, 32(15), pp.1437—1466, Wiley Online Library
5. Berman, Fran and Fox, Geoffrey C and Hey, Anthony JG, (2003) *The Grid: past, present, future,*
6. Wiley and Sons
7. Biogrid, (2005). http://www.biogrid.jp/. BIOMEDICAL INFORMATICS RESEARCH NETWORK (BIRN). 2005. http://www.nbirn.net.
8. Casanova, H., Ledgrand, A., Zagorodnov, D., and Berman, F. (2000). *Heuristics for Scheduling Parameter Sweep Applications in Grid environments.* In Proceedings of the 9th Heterogeneous Computing Systems Workshop (HCW 2000), IEEE.
9. Congiusta, Antonio and Pugliese, Andrea and Talia, Domenico and Trunfio, Paolo, (2003), Designing grid services for distributed knowledge discovery, *Web Intelligence*

and Agent Systems, v-1 (2), pp.91-104, IOS Press.

10. DeFanti, Thomas A and Foster, Ian and Papka, Michael E and Stevens, Rick and Kuhfuss, Tim,(1996), Overview of the I-WAY: Wide-area visual supercomputing, *International Journal of High Performance Computing Applications,*10(2,3), pp.123—131, SAGE Publications

11. De Roure, David and Baker, Mark A and Jennings, Nicholas R and Shadbolt, Nigel R, (2003), The evolution of the grid, Grid computing: making the global infrastructure a reality, 13 pp.14—15, John Wiley & Sons

12. Dumitrescu, C L and Wilde, Michael and Foster, I, (2005), A model for usage policy-based resource allocation in grids,. *Sixth IEEE International Workshop on Policies for Distributed Systems and Networks, 2005*, pp.191—200, IEEE

13. Dumitrescu, C and Raicu, I and Foster, I, (2005), Di-gruber: A distributed approach to grid resource brokering, *Proceedings of the 2005 ACM/IEEE conference on Supercomputing*, pp.38, IEEE

14. Foster, I., Kesselman, C., Nick, J. M., and Tuecke, S. (2002). Grid services for distributed system integration. Computer 35, 6 pp.37–46.

15. Foster, I and Jens, V and Michael, Wand Yong, Z., (2003), The virtual data grid: A new model and architecture for data-intensive collaboration,Conference on Innovative Data Systems Research, Citeseer

16. Foster, I and Kesselman, C and Nick, J. M and Tuecke, S., (2002), Grid services for distributed system integration, *Computer,* 35(6), pp.37—46, IEEE

17. Holtman, K. et.al . (2001). *CMS Requirements for the Grid.* In Proceedings of 2001 Conference on Computing in High Energy Physics(CHEP 2001), Science.

18. Krauter, K and Rajkumar,B and Maheswaran, M (2002), A taxonomy and survey of grid resource management systems for distributed computing, *Software: Practice and Experience*, 32(2), pp.135—164, Wiley Online Library

19. Lin, H. (2005). Economy-Based Data Replication Broker Policies in Data Grids. Tech.rep., Univer- sity of Melbourne, Australia. Jan. BSc Honours Thesis.

20. Magowan, J.(2003). *A view on relational data on the Grid.* In Proceedings of the 17th International Symposium on Parallel and Distributed Processing (IPDPS ’03), IEEE.

21. Mckinley, K. S., Carr, S., and Tseng, C.-W. (1996). Improving data locality with loop trans- formations. *In ACM Trans. Program. Lang. Syst.* Vol. 18. pp.424–453. ACM Press,

22. Papazoglou, M. P. (2003), *Service-oriented computing: Concepts, characteristics and directions, WISE 2003*. Proceedings of the Fourth International Conference on Web Information Systems Engineering, pp.3—12, IEEE

23. Papazoglou, M. P. And Georgakopoulos, D. (2003). *Service-oriented computing. Commun.* 46, 10.

24. ACM

25. Ranganathan, K., and Foster,, I., (2002), Decoupling computation and data scheduling in distributed data-intensive applications, *HPDC-11 Proceedings.* 11th IEEE International Symposium on High Performance Distributed Computing, pp.352—358, IEEE

26. Ranganathan, K., Iamnitchi, A., and Foster, I. (2002). *Improving data availability through dynamic model-driven replication in large peer-to-peer communities.* In Proceedings of the 2nd IEEE/ACMInternational Symposium on Cluster Computing and the Grid (CCGRID'02). IEEE.

27. Stockinger, H., Samar, A., Allcock, B., Foster, I., Holtman, K., and Tierney, B. (2001). *File and object replication in data grids*. In Proceedings of the 10th IEEE Symposium on High Performance and Distributed Computing (HPDC-10). IEEE.

28. Sumithra, R and Paul, S (2010)*,Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery, International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp.1-5, IEEE

29. Toporkov, Victor V and Tselishchev, Alexey, (2010) Safety scheduling strategies in distributed computing, *International Journal of Critical Computer-Based Systems*, 1(1) pp.41—58, Inderscience

30. Venugopal, S. and Rajkumar, B and Ramamohanarao, K, (2006) A taxonomy of data grids for distributed data sharing, management, and processing, *ACM Computing Surveys (CSUR)*, 38(1) pp3, ACM

31. Venugopal, S. and Rajkumar, B (2005). A Deadline and Budget Constrained Scheduling Algorithm for e-Science Applications on Data Grids. In Proceedings of the 6th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP-2005), *Lecture Notes in Computer Science*, vol. 3719. Springer-Verlag.

32. Yu, J., and Rajkumar, B. (2004). *A novel architecture for realizing grid workflow using tuple spaces.* In Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing (GRID'04). IEEE.