iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 7 **Issue:** III **Month of publication:** March 2019

**DOI:** http://doi.org/10.22214/ijraset.2019.3372

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887*
*Volume 7 Issue III, Mar 2019- Available at www.ijraset.com*

# TBHR: A Dynamic Data Management in Grid Computing

Ms Prema. R[1], Dr Antony Selvadoss Thanamani[2]
[1]*Head and Assistant Professor, Department of Computer Applications, Indo Asian Women's Degree College, Bangalore-560043*
[2]*Head & Associate Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi – 642 001*

***Abstract:** A data grid connects huge numbers of computers which is located various places all over the world for sharing data and resources. Day by day the size of the data in data grid is increasing more than Terabytes. Accessing such huge sized data is a serious challenge to network and Grid designers. Data replication is an efficient technique which is used to create multiple copies of file and store them on more than one place, so that user can easily access with in a nearest location. In this paper, a dynamic data replication strategy, called Transmogrified Bandwidth Hierarchy Replication Algorithm is proposed to overcome the drawbacks of BHR, Modified BHR, DMDR, RSCA, PHFS algorithm.*
***Keyword:** Data Replication, Data Grid, Data Management*

## I. INTRODUCTION

In today's world almost all the disciplines like nuclear and particle physics, bioinformatics, high energy etc., requires huge storage size and computing resources. The dynamic data management in grid is fully based on the efficiency of data replication and data placement. In a data grid, periodically new data is generated. They have to be stored in proper sites for future processing. Currently, such data placement decisions are made either randomly or manually per user requests. However, without a careful control on the data placement, a site could get overloaded easily if it stores too many popular data sets. Effective data placement is complicated due to a number of factors.

The characteristics of data set are varying on size, popularity and are processed by different number of jobs. As a result, storage and computing resources required by each data set are different. An intelligent data placement algorithm must take data heterogeneity into account. Clustering the sites of data grid requires different storage capacity [9]. So it is difficult to guess the required size to arrange the data properly. Therefore, we required efficient data placement algorithm to place the data. These are all the challenges which we have considered as a drawback and we gave solution in this paper. Particularly, we proposed data replication algorithm for load balancing in data grids.

## II. RELATED WORKS

Dynamic replication efficient technique which reduces the average job execution time in data grid. Sang-Min Park, Jai-Hoon Kim et al., proposed [7] a novel dynamic replication strategy, called BHR, which reduces data access time by avoiding network congestions in a data grid network. BHR strategy, provides benefits from 'network-level locality' which represents that required file is located in the site which has broad bandwidth to the site of job execution. They evaluated the BHR strategy by implementing it in an OptorSim, a data grid simulator initially developed by European Data Grid Projects. The simulation results show that BHR strategy can outperform other optimization techniques in terms of data access time when hierarchy of bandwidth appears in Internet. BHR extends current site-level replica optimization study to the network-level. Grid computing is emerging as a key part of the infrastructure for a wide range of disciplines in science and engineering, including astronomy, high energy physics, molecular biology and earth sciences. These applications handle large data sets that need to be transferred and replicated among different grid sites. A data grid deals with data intensive applications in scientific and enterprise computing. Data grid technology is developed to permit data sharing across many organizations in geographically disperse locations. Replication of data to different sites will help researchers around the world analyse and initiate future experiments. The general idea of replication is to store copies of data in different locations so that data can be easily recovered if a copy at one location is lost or unavailable. In a large-scale data grid, replication provides a suitable solution for managing data files, which enhances data reliability and availability. Sasi and Thanamani proposed [8], a Modified BHR algorithm to overcome the limitations of the standard BHR algorithm. The algorithm is simulated using a data grid simulator, OptorSim, developed by European Data Grid projects. The performance of the proposed algorithm is improved by minimizing the data access time and avoiding unnecessary replication.

Data replication in data grids is an efficient technique that aims to improve response time, reduce the bandwidth consumption and maintain reliability. In this context, a lot of work is done and many strategies have been proposed. Unfortunately, most of existing replication techniques are based on single file granularity and neglect correlation among different data files. Indeed, file correlations become an increasingly important consideration for performance enhancement in data grids. In fact, the analysis of real data intensive grid applications reveals that job requests for groups of correlated files and suggests that these correlations can be exploited for improving the effectiveness of replication strategies. Lakshmi and Thanamani proposed [4], a new dynamic data replication strategy, called DMDR, which consider a set of files as granularity. Their strategy gathers files according to a relationship of simultaneous accesses between files by jobs and stores correlated files at the same site. In order to find out these correlations data mining field is introduced. They choose the all confidence as correlation measure. In data grid, it is an important research filed to complete interoperability of data. In the mean time, share of data also becomes the crucial problem. Data replication, as a solved solution of data share, goes into more and more vital. Gui Liu, HaiLaing et al., proposed [2] a strategy called replication strategy based on clustering analysis (RSCA), which confirms the correlation among the data files accessed according to the access history of users. And then, through clustering analysis operation obtains the correlative files sets related to the access habits of users. At the same time, it produces the data files replica on the basis of those sets, which achieves the aim of pre-fetching and buffering data. The experimental results show that RSCA is effective and available. Contrast to other dynamic replication strategies, it has reduced not only the average response time of client nodes, but also those of the bandwidth consumption. Data replication is a method to improve the performance of data access in distributed systems. Dynamic replication is a kind of replication that adapts replication configuration with the change of users' behaviour during the time to ensure the benefits of replication. Leyli Mohammad Khanli, Ayaz Isazadeh et al., proposed [5] a new dynamic replication method in a multi-tier data grid called predictive hierarchical fast spread (PHFS) which is an extended version of fast spread (a dynamic replication method in the data grid). Considering spatial locality, PHFS tries to predict future needs and pre-replicates them in hierarchal manner to increase locality in accesses and consequently improves performance. They also compared PHFS and CFS (common fast spread) with an example from the perspective of access latency. The results shows that PHFS causes lower latency and better performance in comparison with CFS. Replica value determination plays an important role in the problem of replication in data grid, since only replicating high value replicas can improve the replication efficiency [10]. However, replica value determination methods applied now do not perform well, because they cannot well adapt to the changes of file access patterns [1]. Tian Tian, Junzhou Luo et al., addresses this problem by introducing a new replica value determination method through data mining, and based on this method they proposed a pre-fetching based replication algorithm in a virtual organization (VO) environment in order to carry out a better replication optimization [6]. The major idea of this algorithm is to make use of the characteristic that members in a VO have similar interests of file. Experiments show that compared with traditional replication algorithms their new replication algorithm shows better performance and efficiency in data grid. Creating many replicas in the processing of data-intensive jobs in data grid is an efficient strategy. Replica replacement is the crucial step to this strategy. Economic model, popularity model and hybrid model etc. have been proposed to solve this issue of replica replacement with analysis and prediction based on each data file, however, these models neglect association relationships among different data files. To find out these association relationships hidden in data-intensive jobs, Apriori algorithm in data mining field is adopted to analyze behaviours of each data-intensive job. Jianhua Jiang, Huifang et al., proposed [3] an associated replica replacement algorithm based on Apriori approach in data grid. This algorithm has two major steps: 1) associated behaviour analysis and classification of data files in each node; 2) generation and application of replica replacement rules. This algorithm is simulated in Optorsim to be compared with LFU algorithm. The experiment shows that there is a relative advantage compared with LFU in mean job times of all jobs, number of remote file access and effective network usage perspectives.

## III. REPLICATION STRATEGIES

In general, replication algorithms can be broadly classified into two categories such as static and dynamic [4] a shown in Fig.1.

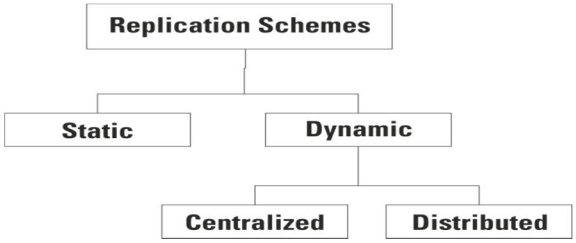



Figure 1: Types of Replication Schemes

In static method the created replica will exist until user deletes replica manually or up to the end of its duration. 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.
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.
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. Reliability, Availability, Adaptability, Scalability and high Performance are the advantages of replication strategies.

## IV. TBHR ALGORITHM

In data grid finding interoperability between data and sharing data are an important research areas. Data replication, as a solved solution of data share, goes into more and more vital. In data grid, data replication is a suitable technique which reduces more bandwidth consumption and speeds up the response time. So to solve these problems due to the lack of most of the existing replication techniques, in this research paper a lot of work is done as well as new strategy has been proposed. In data grid, file correlation become an important consideration for performance enhancement. The analysis of real data intensive grid applications proves that groups of file correlations can be exploited for improving the effectiveness of replication strategies. In order to solve such problems a new dynamic algorithm is proposed in this paper which works in two phases. Initially, clustering analysis is conducted on user file access history and correlated file sets are related to the access habits based on user are extracted. Then replication is done within the correlated file sets and stored in the place where file has been accessed frequently. So it minimizes the job execution time, reduced the storage space and reduced bandwidth consumption.

1) *Algorithm:* TBHRA Region Based Algorithm
2) *Inputs:* Grid Topology, job and bandwidth details
3) *Outputs:* Mean job Execution Time, Average Storage Used, Network Usage, Number of replications, Shortest path, Characteristic Path Length.

### A. Methods

1) if (Requested File not in Local Site)
fetch from the nearby site within the region
2) Create Cluster on file accessing history in the grid over a period of time
3) Proceed to replicate among the correlated file sets, which is related to the access habits of users
4) if (free space available in SE)

```
Store it;
Else {
if (duplication if replica in other sited within region)
Terminate optimizer;
Else {
Sort files in SE using LFU
For (each file in SE)
{
if file is duplicated in other sites within region
Delete it;
if (of the enough space to store new Replica)
Break;
}}
```

5) if (not enough free space)

```
{
Sort files in SE using LFU
 For(each file in sorted list)
 {
```

if (access frequency of new Replica > access frequently of the file
Delete file
if (enough free space)
break
}}
if (enough free space)
Store new replica
}

## V. WORKING PRINCIPLE OF TBHR ALGORITHM

Initially, the user submits a job to the grid. The data are produced in a master site, then master site distributes data to each other region header. The jobs are assigned to computing elements, the places where the jobs are executed. When the job needs the data, and it is not present in the local storage, replication takes place. The replicated files are not stored in all the requested sites.

Instead, the file is stored in the site where the file is accessed for the maximum time, with the assumption that files recently accessed by a client are likely to be accessed by nearby clients and the files accessed recently are likely to be accessed again. By storing it in the maximally accessed site and in the region header, the storage cost and also the mean job execution time can be reduced. If there is no space and the replica is duplicated in other sites within the region other than the region reader, the optimizer can be terminated. If there are no duplications, the least frequently accessed file is deleted and the new replica is stored.

Whenever an optimizer is considering a file request, it performs the following tasks in order to optimize replication.

1) *Replica Decision:* If a requested file is not present on a site's storage element, this process decides whether local replication of this file take place. If the optimizer decides not to replicate a file, the job must access the file remotely.
2) *Replica Selection:* When considering which replica to read or replicate locally, this process selects the best of those available.
3) *File Replacement:* When a remote replica has been selected for replication to the site's storage element, the storage element might not have sufficient spare capacity. In this case, one or more replicas must be deleted. In this algorithm, the least frequently accessed file is deleted.

## VI. EXPERIMENTAL RESULTS

The efficiency is calculated based on Performance Metrics, such as Saved Storage Space, Shortest Path, Characteristic Path Length, and Reduced Bandwidth Consumption. TBHRA is compared with one other dynamic replication strategy on the MATLAB.

Table 1: Job Execution Scenarios

| Job Execution Scenarios | Values |
|---|---|
| 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 |
| Reduced Bandwidth Consumption | Network Utilization |
| Shortest Path | Specifies shortest path of given job |
| Storage Space | Storage space reduced by avoiding storing files in unnecessary locations |

1) *Saved Storage Space*: The saved storage space is shown below;

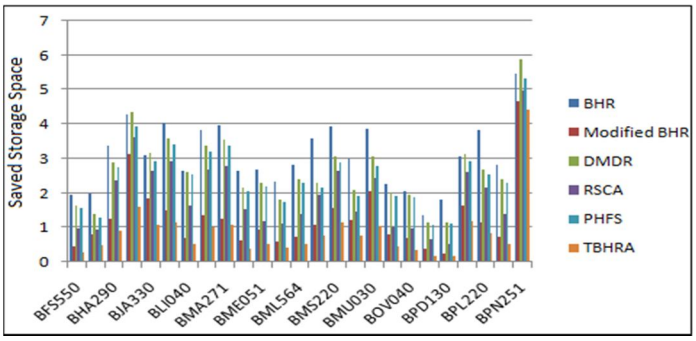



Figure 2: Saved Storage Space

2) *Reduced Bandwidth Consumption*: The bandwidth consumption for random access pattern is shown below;

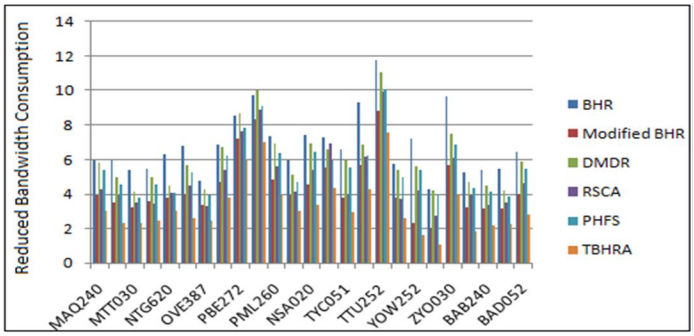



Figure 3: Reduced Bandwidth Consumption

3) *Characteristic Path Length*: The characteristic path length for random access pattern is shown below;

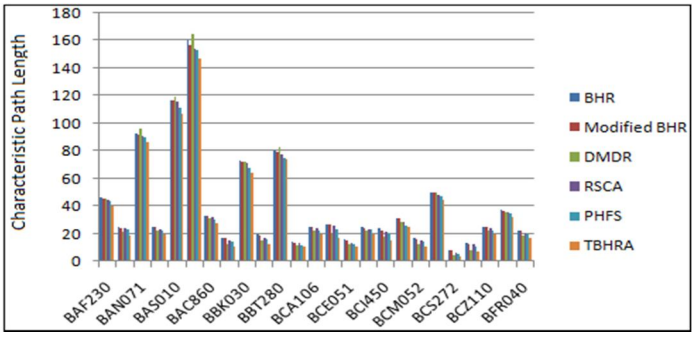



Figure 4: Characteristic Path Length

4) *Job Execution Time*: The job execution time for random access pattern is shown below;

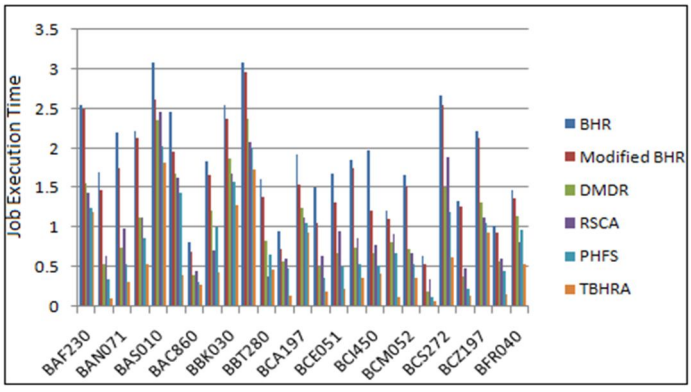



Figure 5: Job Execution Time

## VII. CONCLUSION

In this paper, we have presented a TBHRA (Transmogrified Bandwidth Hierarchy Replication Algorithm) for data grids. In this paper, the relationship between files and groups of related files are considers as a granularity for replication. Indeed, considering correlated files for replication improves significantly performance evaluation metrics of replication strategies, including characteristic path length, neighbourhood connectivity, job execution time and shortest path compared with other algorithm. The experimental result shows that TBHRA gives better performance than other algorithms. As a part of our future work, we plan to deploy our replication strategy in a real grid environment.

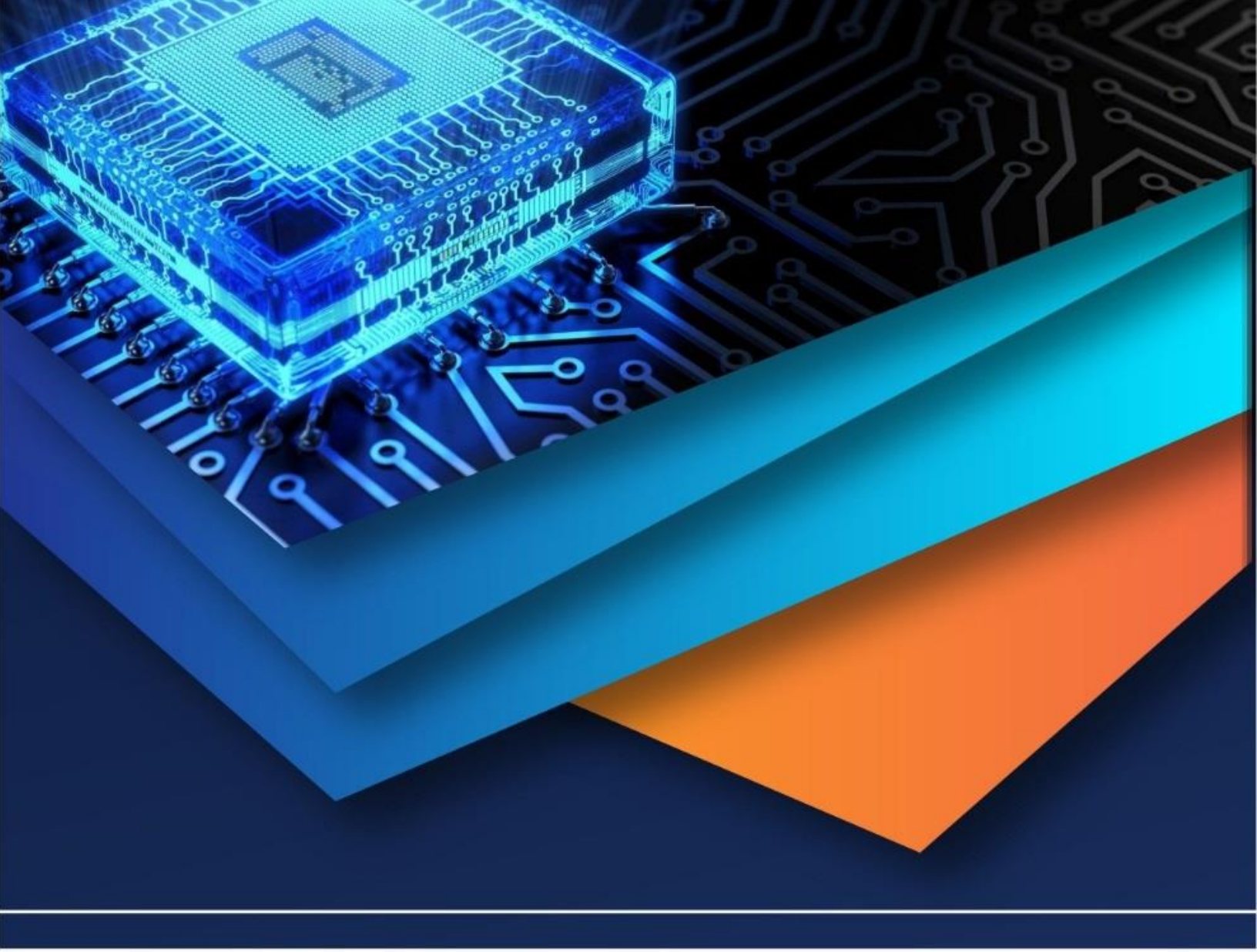
## REFERENCES

[1] Chamseddine Hamdeni, Tarek Hamrouni, Faouzi Ben Charrada, "DISQUEV: Looking for Distribution Quality Evolution as a New Metric for Evaluating Replication Strategies", Computer Systems and Applications (AICCSA) 2017 IEEE/ACS 14th International Conference on, pp. 295-302, 2017.
[2] Gui Liu, HaiLiang Wei et al., "Research on Data Interoperability Based on Clustering Analysis in Data Grid", International Conference on Interoperability for Enterprise Software and Applications China, IEEE, ISBN:978-0-7695-3652-1.
[3] J. H. Jiang et al., "ARRA: An Associated Replica Replacement Algorithm Based on Apriori Approach for Data Intensive Jobs in Data Grid", Key Engineering Materials, Vols. 439-440, pp. 1409-1414, 2010.
[4] Lakshmi, 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, 2016, pp. 389-394.
[5] Leyli Mohammad Khanli, Ayaz Isazadeh et al., "PHFS: A dynamic replication method, to decrease access latentcy in the multi-tier data grid", Future Generation Computer Systems, Elsevier, 27(2011), pp.233-244.
[6] Luo, Junzhou & Wang, Xiaopeng & Song, Aibo. (2005). "A semantic access control model for grid services", 350 - 355 Vol. 1. 10.1109/CSCWD.2005.194196.
[7] Sang-Min Park, Jai-Hoon Kim et al., "Dynamic Data Grid Replication Strategy Based on Internet Hierarchy", International Joint Research Project) by Ministry of Information & Communication in South Korea.
[8] Sashi, K & Selvadoss Thanamani, Antony, "A new dynamic replication algorithm for European data grid", COMPUTE 2010 - The 3rd Annual ACM Bangalore Conference, 10.1145/1754288.1754305, 2010.
[9] Subhendu Sekhar Paik, Rajat Subhra Goswami et al., "Intelligent Data Placement in Heterogeneous Hadoop Cluster", International Conference on Next Generation Computing Technologies, NGCT 2017, Communications in Computer and Information Science, vol 827, Springer, Singapore.
[10] Yi, Kan & Ding, Feng & Wang, Heng. (2010). "Integration of Task Scheduling with Replica Placement in Data Grid for Limited Disk Space of Resources", 2010 Fifth Annual ChinaGrid Conference. 37-42. 10.1109/ChinaGrid.2010.29.

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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY