A Dynamic Data Replication in Grid System using Rule Based Classifier Algorithm

Sonali Warhade, Prashant Dahiwale, M. M. Raghuwanshi

***Abstract*— A data grid is a structural design or cluster of services that provides persons or assortments of users' ability to access modify and transfer very great amount of geographically distributed data. As a result of this needed massive storage resources for store massive data files. For take away that drawback we have a tendency to use dynamic data replication is applied to scale back data time interval and to utilize network and storage resources expeditiously. Dynamic data replication through making several replicas in numerous websites. Here, through up the modified BHR (MBHR) methodology, we have a tendency to project a dynamic algorithmic program for data replication in data grid system. This algorithmic program uses the rule based classifier algorithm uses three parameter for locating appropriate web site wherever the file is also needed in future with high likelihood. The algorithmic program predicts future wants of replicated appropriate grid web site square measure supported file access history**

***Index Terms*— Data Grid; Data Replication; MBHR, replicated site.**

I. INTRODUCTION

Large scale geographically distributed systems are getting additional and additional widespread. It's additionally referred to as grid system. The term grid system is same because the distributed system. it's the contain network structure and rising computation that square measure accustomed easy accessibility information and resources over network [1]. Grids is employed to sharing, selection, and aggregation of an outsized varied resources as well as supercomputers, storage systems, data sources, and specialized devices that square measure geographically distributed and owned by completely different organizations for finding large-scale machine and data intensive issues in science, engineering, and commerce.

The necessary data is present in much website victimization the grid system the user will straightforward to access, while not the duplication of replicate regionally. These services square measure provided by associate integrated grid services platform in order that that user will access the resources transparently and effectively. Managing this data during a centralized location will increase the data. Thus to scale back the data time interval, "replication is used" [2].

Manuscript received April 12, 2016

Sonali Warhade, Scholar Student, Dept of Comp. Sci. Engg., RGCER, Nagpur, India

Prashant Dahiwale, Assistant Professor, Dept of Comp. Sci. Engg., RGCER, Nagpur, India

M. M. Raghuwanshi, Professor, Dept of Comp. Tech., YCCE, Nagpur, India

One in all the necessary functions in data reproduction techniques is replica placement. Replica placement is that the main drawback of putting duplicate copies of data within the best suited node within the data grid. The replica placement may be due by victimization the four topology in data grid system: multi-tier, graph topology, hierarchical and peer to peer [3].

There square measure two sorts of data replication techniques, name as, static and dynamic. In static data replication, the set of replicas to be created, and also the node wherever the replicas ought to place is decide statically at the time of grid system setup. The static replication ways square measure straightforward to implement however not principally used as a result of it doesn't support data duplication throughout job execution. Static replication ways have the benefits of reduced overhead and fast job programming [3].

Dynamic replication [3] can adapt changes supported user requests, storage capability and information measure. Dynamic replication techniques square measure capable of creating intelligent choices to position data within the grid supported storage capability and node handiness. Replicas square measure created in distributed fashion. During a distributed theme, replicas square measure created during a few selective nodes additionally to the top node. As a result of intelligent higher cognitive process and putting the replica in step with surroundings conditions makes dynamic replication ways higher than the static replication ways. Further, the dynamic replication ways don't produce the replica of all the files, solely widespread files square measure replicated. So, to induce the high handiness, fault tolerance and potency, the necessary data files ought to be dynamically adjusted.

In data replication has several benefits however there square measure some drawbacks additionally. an excessive amount of replicas might not forever will increase the info handiness as a result of typically brings reserve disbursal and it's once more a challenge to position the new replicas on completely different nodes in step with the present environments conditions within the distributed systems like cloud systems [4]. Keeping on top of drawbacks in mind to attain higher dynamic replication systems, 3 necessary issues should be solved. 1) Its necessary to search out the data that ought to be replicated in distributed system and once, thus on meet the user needs like increase in data access speed and reduction in waiting time. 2) To attain system handiness demand, it's necessary to search out the quantity of recent replicas that ought to be created within the system 3) to scale back the information measure needs and obtain the utmost job execution rate, it's necessary to see wherever we should always place the replica? Of these square measure necessary issues that require to be addressed

II. LITERATURE SURVEY

There square measure some recent works that address the matter of planning and or replication in data grid likewise because the combination between them.

In [5] there square measure 3 kinds of replication algorithms were introduced, least frequently Used (LFU) and 2 economic methods. In each algorithm is want time replication continuously occurs. Once a needed file isn't obtainable regionally and native storage isn't enough area, then LFU and economic technique work differently. In LFU replication continuously replicate data and if the actual replica placements haven't needed area then delete the last often access enter recent time. In Economic algorithmic program continuously replication occur and if replica placement haven't enough area then calculate worth of every enter data storage and also the file have less value that file square measure deleted. They act supported the distribution. the smallest amount Recently Used (LRU) algorithmic program is additionally continuously occur the replication and if replica placement has not enough area then the files that are used less time in recently is deleted. The LRU algorithmic program is provide high performance as compared to the LFU.

In [4] information measure Hierarchy Replication (BHR) algorithmic program is introduced. During this algorithmic program produce the set of region by victimization the closely set sites are going to be organized as cluster. That organized cluster is understood as network region. In BHR technique, if the needed data is present within the network region then it straightforward to access and it's required less time and value for access the data. as a result of the information measure between the actual region sites is high. In requested website haven't enough area for store the new replica, then the files that square measure duplicated in region these file deleted and store the new replica. BHR is employed to scale back the entomb region transmission and increasing performance. The access history of each website, that contains details of often accessed files, is maintained within the region header (shown in Fig 1).

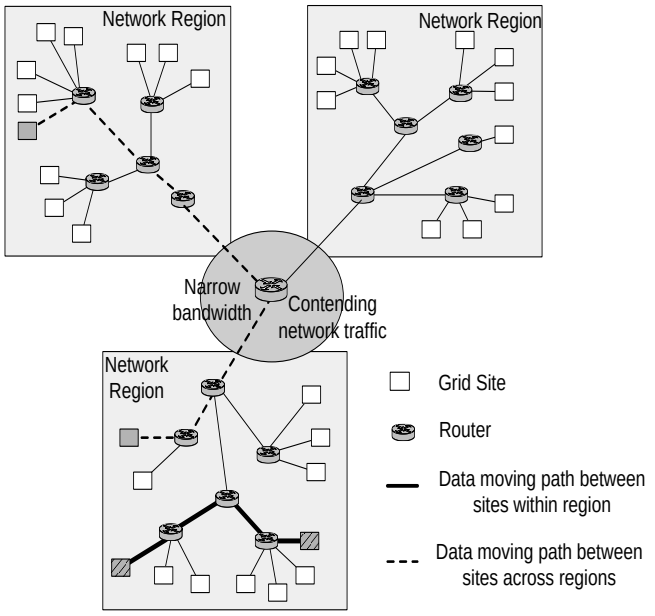

Fig1: Replication Strategy based on Bandwidth Hierarchy

In [8] one new algorithmic program is purposed that is named modified BHR, it's a modified version of BHR algorithmic program. During which attempt to replicate the file inside the region. During which used the one parameter for replica placement. That parameter is variety of access file. During which place the replicas within the web site that have most variety of request for access the file. During which victimization the history of replication set the parameter. During this paper OptorSim simulator is employed for evolution the purposed algorithmic program. It's accustomed avoid the inessential replication.

In [8] one new algorithmic program is purposed that is modified version of BHR algorithmic program. During which attempt to replicate the file inside the region. During which used three parameters for replica placement. In which used the Fuzzy based logic for finding the appropriate website for replication using three parameters. Those parameters are no. of access to file, last access time interval, sum of bandwidth of the node to other node. During which place the replicas within the web site that have most variety of result. During which victimization the history of replication set the parameter. During this paper OptorSim simulator is employed for evolution the purposed algorithmic program. It's accustomed avoid the inessential replication.

III. PROPOSED WORK

A. Proposed Approach

Data replication could be a technique that has been wide studied and applied in grid system. Dynamic replication creates and deletes replicas per the changes of user behavior. Dynamic replication ways have the power to adapt changes in user behavior, therefore dynamic replication is additional appropriate for grid system. In our approach, we have a tendency to use the graph topology to form the grid system. Here, through raising the modified BHR (MBHR) methodology, we have a tendency to apply dynamic replication technique victimization the no of parameters to calculate the replication node. The no of parameter is simply like as range of access file, information measure of the node to alternative nodes; last time interval etc. victimization that parameter calculates the replication node so transfers the data to it explicit node. Those parameters are passed to Rule Based Classifier Algorithm, and using that algorithm calculate the most replication node so transfer the data to it explicit node. The algorithmic program predicts future wants of replicated appropriate grid web site square measure supported file access history. Victimization that approach enhance the system and straightforward to share data at intervals the time. It's used avoid the duplication of replication.

B. Proposed Algorithm

In the grid system no of jobs square measure present, that needed great amount of data for execution. In grid system, employment needed data then it first access from its original website if needed data doesn't present then it access from native region. Then if a job doesn't present data at native region then it has access from remote region. Thus that needed most price and time for access data. Therefore, to reduce the access price and time we tend to use the dynamic data replication technique. Data replication is employed to reduce the step to access great amount of data from remote website. In which replicating the data in geographically

distributed data stores for simply access data. thus replication could be a time overwhelming and costly method, the node chosen for replication ought to have the very best requesting likelihood within the future, and though the file is requested by another node, the price of transferring the file to the replication node ought to be as low as potential.

The planned algorithmic program, just like the modified BHR algorithmic program [8], could be a dynamic algorithmic program. Within the algorithmic program replicate files among a region, and store replica in appropriate website having most likelihood for the placement of future access. In modified BHR algorithmic program select right node for replication solely with one parameter, it select best node with higher access frequency because the appropriate place for replication. Within the planned algorithmic program, select right replication node with the 3 parameter as range of access file, information measure of the node to different nodes and last time interval for select the most effective replication website. The detailed algorithm is given below. Here, we use a Rule Based Classifier algorithm in our proposed algorithm to find the best location or website for replication.

For find the best location or website for replication used the Rule Based Classifier algorithm, we have considered 24 identical rules, some of which have been explained in Table I. The replication node in the present algorithm is selected on the basis of the value allocated to each node; this value is calculated, on the basis of a Rule Based Classifier algorithm, Fig. 2. Shows a schematic diagram of our Rule Based Classifier algorithm, the output of which is the value of the node. Through taking into three parameters including: i) the number of past accesses to the file in the node. A node with a great number of accesses to the file in past, is more likely to be requesting that file by the node again in the near future. Therefore, the value of node will be greater and will be a suitable candidate place for replication of the file. ii) Bandwidth, which is another important parameter in the appropriateness of a node for being the replication site. The more the bandwidth of a node is, the less the transferring cost from that node to other nodes and vice versa will be; therefore, the value of that node for being the replication site will be higher; this input parameter for Rule Based Classifier algorithm considered as the sum of the bandwidths connecting the node to the other nodes within a region, which is calculated for node i in region c as:

$$WB_c = \sum_{1}^{n} (bw_{ij}) \qquad (1)$$

Where n is the number of the nodes within region c, and bw_{ij} is the connecting bandwidth of node i and j; and iii) the time of the last access to the file in a node: The nodes that have recently accessed to the file are more likely to be requester node for that file in the future again. So the smaller time difference between the current time and the last access time of that file on the node makes value of node larger.

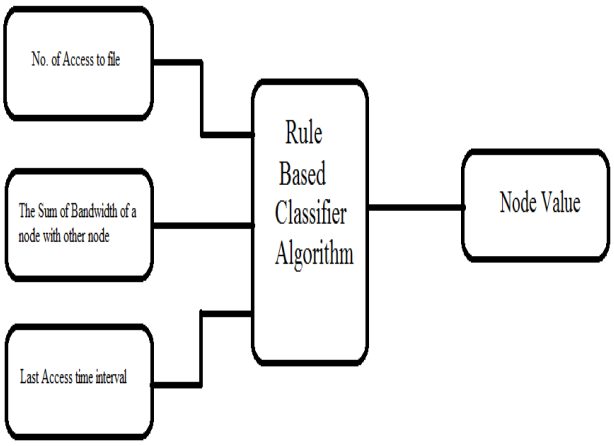

Fig. 2. A schematic drawing of the Rule based classifier algorithm for determining the value of a node.

1. Rule definition of rule based classifier algorithm

Number	Description
1	If (number_of_accesses is high) and (bandwidth is average) and (last_access_time_interval is low) then (node_value is high)
2	If (number_of_accesses is average) and (bandwidth is high) and (last_access_time_interval is low) then (node_value is high)
3	If (number_of_accesses is high) and (bandwidth is high) and (last_access_time_interval is high) then (node_value is average)

In our algorithmic program, once the work needs the data that doesn't exist within the native storage, replication takes place. The replicated files square measure keep within the best sites wherever the file can most likely be accessed within the future within the different hand. To get the most effective replication website, every node having a storage component was given a amount calculated through a perform. The algorithmic program calculates its amount by line of work a function having 3 input parameters explained above. When the number is calculated for all the sites existing within the region wherever the request for the file was received from, the algorithmic program chooses the node with the very best price of amount because the replication website. If the chosen node has enough area offered, the file is keep within the selected website and within the region header. On the opposite hand, if the chosen website doesn't have enough area and another copy of the file exists in another site among the region instead of at intervals the region header, the optimizer may be terminated. If duplication of the file doesn't exist, the smallest amount frequently accessed file is deleted, and therefore the new replica is keep. Required info just like the history of access to files in every node, and therefore the price of bandwidths in every region square measure keep within the region header.

IV. RESULT

The result shows the output of the particular grid system. The implementation is in java with simulation. The implementations are as follows:

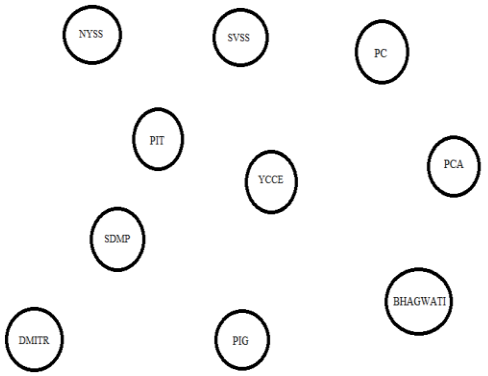

Fig 3: Creation of Network Structure

Fig 3 describes the "Network Structure". In which shows set of the nodes which are connect to each other by plot button. It is in the form of graph topology.

2. Set the region.

Region Point	Name of the node
E1	YCCE
E2	SVSS
E3	NYSS
User node	YCCE

2. describes the "Set the Region". After the network structure creation, set region by the E1, E2 and E3 region points i.e. YCCE, SVSS and NYSS node and user node is about by YCCE site. The user node is employed to point out native node. And region is employed to point out the set of the node.

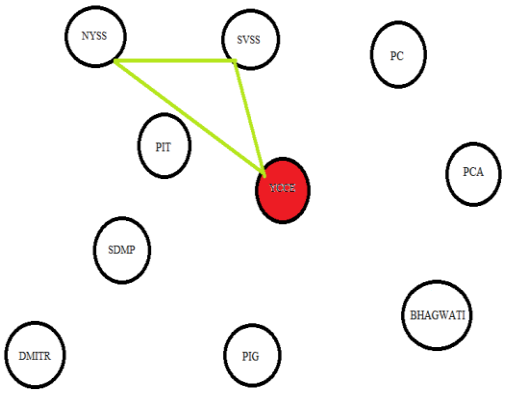

Fig 4: Plot the Network Structure

Fig 4 describes the "Plot the Network Structure". Plot the graph by victimization the PLOT button and show the distance between those nodes in a very region. In which the red color is employed to point out the user node. That user node is show the bottom node. And green color is employed to point out the sides between the nodes in particular region.

3. Show the result of Search content

Search File name	Result
Sonali.txt	No result found at root node. No result found at child node SVSS. Found 1 Result. Found- E:\Project:\2201New\Searchfolder\nyss\sonali.txt

3 describes the "Shows the Result". In which shows the particular search file and the result about the particular file.

4. Show the result of calculate parameter

Name of the node	No of Access to File	Sum of Bandwidth of Node with other node	Last Access time
YCCE	7	1100	2016-02-15 12:45:27.0
NYSS	8	600	2016-02-15 12:45:30.0
SVSS	7	500	2016-02-15 12:45:28.0

4. describes the "Shows the result of Calculate parameter". Shows the result of calculate three parameter i.e. no of access to file in particular node, sum of Bandwidth of node with other node and the last access time by victimization the Calculate "Calculate Parameter" button.

CONCLUSION

In this paper, a new dynamic replication algorithm is proposed. This algorithm is same as the modified BHR algorithm. In which used the set of parameter for placement of replica i.e. no of access to file, last access time interval and the sum of bandwidth. Hence, in comparison with the Modified BHR method which only takes the number of accesses into account to select the right replication site, our proposed algorithm has better results, reflecting the significance of taking the highest number of parameters into account in decision-making process. It should be use to minimize the access cost and access time. It should give better result as compare to other replication algorithm.

ACKNOWLEDGMENT

I have this opportunity to express our deep sense of gratitude and whole hearted thanks to my Guide Prof. Prashant Dahiwale, Computer Science and Engineering Department for this invaluable guidance, inspiration and encouragement .It is because of his that I could synchronies my efforts in this Project greatly indebted to his for piloting me whenever I faced difficulties in my Project work. I take this opportunity to express my profound sense of gratitude and respect to all those who helped me through the duration of this work. This acknowledgment would be incomplete without rendering

impartial gratitude to all those who directly or indirectly helped me in my Project work.

REFERENCES

[1] Houda Lamehamedi, Boleslaw Szymanski, and Zujun Shentu, "Data Replication Strategies in Grid Environments," in IEEE Computer Society Press, Los Alamitos, CA, 2002, pp. 378-383.

[2] Sheida Dayyani and Mohammad Reza Khayyambashi, "A Comparatives study of Replication Technique in Grid Computing Study," in *(IJCSIS) International Journal of Computer Science and Information Security, Vol. 11, No. 9, September 2013*.

[3] Alireza Souri and Amir Masoud Rahmani, "A survey for Replication Placement Technology in Data Grid Environment," in *I.J. Modern Education and Computer Science*, 2014, 5, 46-51.

[4] K. Ranganathan, I. Foster, "Design and Evaluation of Dynamic Replication Strategies for a High-Performance Data Grid," International Conference on Computing in High Energy and Nuclear Physics, 2001

[5] M. Tang, B.S. Lee, X. Tang, and C.K. Yeo, "Combining data Replication Algorithm and Job Scheduling Heuristics in Data Grid", LNCS 3648, pp.381 -390, 2005.

[6] R.S. Chang, H.P. Chang, "A Dynamic Data Replication Strategy Using Access-Weights in Data Grids," Supercomputing, Vol. 45, No. 3, pp. 277-295, 2008.

[7] A.R. Abdurrab, T. Xie, "FIRE: A File Reunion Based Data Replication Strategy for Data Grids," International Conference on Clustering Computing and the Grid, pp. 215-223, 2010.

[8] T. Amjad, M. Sher, A. Daud, "A Survey of Dynamic Replication Strategies for Improving Data Availability in Data Grids", Future Generation Computer Systems, Vol. 28, No. 2, pp. 337-349, 2012.

[9] Y. Yuan, et al., "Dynamic Data Replication based on Local Optimization Principle in Data Grid", International Conference on Grid and Cooperative Computing, pp. 815-822, 2007.

[10] F. Jolfaei, A.T. Haghighat, "Improvement of Job Scheduling and Tow Level Data Replication Strategies in Data Grid", Mobile Network Communications & Telematics, Vol. 2, No. 3, 2012.

[11] S. M. Park, J.H. Kim, Y.B. Ko, and W.S.Yoon, "Dynamic Data Grid Replication Strategy Based on Internet Hierarchy" in Grid and cooperative computing, vol-3033, M. Li, X. H. Sun, Q. Deng, and J. Ni. Eds. Ed: Springer Berlin Heidelberg, 2004, pp. 838-846.

[12] Foster, and K. Ranganathan, "Design and evaluation of dynamic replication strategies for a high performance data grid", in proceedings of international conference on computive in High Energy and nuclear physical, Beijing, China, September 2001.

[13] MahsaBeigrezaei, Hamidreza Rashidy Kanan and AbolfazlToroghi Haghighat, "A New Fuzzy Based Dynamic Data Replication Algorithm in Data Grids", in 2013 13th Iranian Conference on Fuzzy Systems (IFSC) 978-1-4799-1228-5/13/$31.00 ©2013 IEEE

[14] Wenhao LI, Yun Yang, and Dong Yuan, "A Novel Cost-effective Dynamic Data Replication Strategy for Reliability in Cloud Data Centres", in 2011 Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, 978-0-7695-4612-4/11 $26.00 © 2011 IEEE.

[15] S. K. Singh, A. Prasad P. K. Singhan and R. K. Singh, "A Replica Placement and Replacement Algorithm for Data-Grid in DRTDBS", in International Conference on Grid and Cooperative Computing, pp. 815-822, 2007