HGASA: An Efficient Hybrid Technique for Optimizing Data Access in Dynamic Data Grid

R. Kingsy Grace[1(✉)] and R. Manimegalai[2]

[1] Sri Ramakrishna Engineering College, Coimbatore, India
kingsygrace.r@srec.ac.in
[2] Park College of Technology, Coimbatore, India
mmegalai@yahoo.com

Abstract. Grid computing uses computers that are distributed across various geographical locations in order to provide enormous computing power and massive storage. Scientific applications produce large quantity of sharable data which requires efficient handling and management. Replica selection is one of the data management techniques in grid computing and is used for selecting data from large volumes of distributed data. Replica selection is an interesting data access problem in data grid. Genetic Algorithms (GA) and Simulated Annealing (SA) are two popularly used evolutionary algorithms which are different in nature. In this paper, a hybrid approach which combines Genetic Algorithm with Simulated Annealing, namely, HGASA, is proposed to solve replica selection problem in data grid. The proposed algorithm, HGASA, considers security, availability of file, load balance and response time to improve the performance of the grid. GridSim simulator is used for evaluating the performance of the proposed algorithm. The results show that the proposed algorithm, HGASA, outperforms Genetic Algorithms (GA) by 9 % and Simulated Annealing (SA) by 21 % and Ant Colony Optimization (ACO) by 50 %.

Keywords: Replica selection · Data grid · Computational grid · Genetic algorithm · Simulated annealing

1 Introduction

The two major categories of grid computing [1] are: (i) Computational grid and (ii) Data grid. Computational grid is mainly used for compute intensive applications and data grid is an infrastructure for storing and sharing large volumes of data, for data intensive applications Data replication in data grid reduces the access latency in distributed systems by keeping multiple copies of the data file in geographically distributed sites [2]. In a data grid system, there are hundreds of clients across the globe submitting job requests. Usually, a grid job accesses multiple files for its job execution. In data-intensive applications, when a job accesses large file, the unavailability of that file can cause the whole job to hang up. Any node or network failure causes file unavailability. As a result, there has been an increasing research interest focusing on how to maximize the file availability. Data replication reduces the access latency in distributed systems by

N. Bjørner et al. (Eds.): ICDCIT 2016, LNCS 9581, pp. 132–136, 2016.
DOI: 10.1007/978-3-319-28034-9_17

keeping multiple copies of the data file in geographically distributed sites [2]. Each grid site has its own capabilities and characteristics; therefore, selecting one particular site which has the required data among many such sites, is an important and significant decision [3]. The replica selection problem has been investigated by many researchers and only response time is considered as a criterion for the selection process. In this work, the replica selection problem is addressed as an important decision to guarantee efficiency and to ensure the satisfaction of the grid users by providing replicas with reduced latency and improved security. The main contribution of this work is to produce an alternative solution to the replica selection problem based on response time, availability, security and load balancing. This work extends the replica selection using genetic algorithm [4] by employing hybrid approach. The proposed replica selection algorithm is based on both genetic algorithm and simulated annealing and is called as Hybrid of Genetic Algorithm and Simulated Annealing (HGASA). HGASA is implemented using GridSim 5.1 simulation toolkit [5]. Performance of the proposed HGASA approach is compared with Genetic Algorithm (GA), Ant Colony Optimization (ACO) algorithm, Simulated Annealing (SA) in terms of TASL (Response Time, Availability, Security and Load Balancing) value [4].

2 Related Work

Replica selection is one of the important tasks of data management in data intensive application. It decides which replica location is the best place to access the data for users. If several replicas are available for a file, the optimization algorithm determines which replica should be selected to execute the job. Lin et al. have proposed a Network Coordinate (NC) based nearest replica selection service called Rigel in [6]. Tim and Ambramson have proposed GriddLeS Data Replication Service (GRS) which provides limited support for automatic replica selection in [3]. Naseera and Murthy have proposed predictive replica selection using neural networks [7] based on [8]. Ishii and Mello have proposed a solution for Data Access Problem (DAP) in [9]. It is a prediction based optimization approach. Sun et al. have proposed an Ant Colony Optimization (ACO) algorithm for replica selection in [10]. It reduces data access latency, decreases bandwidth consumption and distributes the load evenly. Jadaan et al. have proposed a rank based elitist clustering genetic algorithm for replica selection in data grid [4].

3 HGASA: Hybrid of Genetic Algorithm and Simulated Annealing for Replica Selection

Genetic Algorithm (GA) was introduced by J. Holland in 1975 [11] and had been used for solving searching, learning and optimization problems. GA is a global search technique which is based on the mechanism of biological evolution inspired by Darwin's theory of evolution [12]. GA consists of two types of operations, namely, mutation and crossover. These operations are repeatedly applied to a population of chromosomes for obtaining a possible solution for the given search space. Simulated Annealing (SA) is a heuristic optimization algorithm [13] and is analogous to annealing in metals and solids.

SA was first introduced by Metropolis et al. in [14]. The idea in Metropolis et al. is used by Kirkpatrick et al. in [13] to search for an optimal solution in optimization problems. Combining GA which is a global search technique with SA which is a local search technique gives the benefit of both, at the same time avoids problems such as premature convergence and local optimum [15]. In the proposed replica selection architecture, if a user requests for a replica, the replica selection algorithm gets all information regarding the replica from the Replica Location Service (RLS) [16]. The best replica location site is selected based on four parameters: response time, availability, security and load balancing. The network related information such as bandwidth is gathered with the help of Network Weather Service (NWS). A hybrid evolutionary algorithm, HGASA, which employs both Genetic Algorithm (GA) and Simulated Annealing (SA), is proposed in this work. TASL values are used to compare the performance of the proposed algorithm with the existing algorithms. During GA implementation, a Model Replica (MR) [4] is set with maximum (100 %) of response Time T, Availability A, Security S and Load balancing L. i.e. MR (T, A, S, L) = (100, 100, 100, 100). The distance between MR and the available replica is computed using the Eq. (1). T_1, A_1, S_1 and L_1 are the TASL values of MR and T_0, A_0, S_0 and L_0 are the TASL values of available replica.

$$TASL = \sqrt{((T_1 - T_0) + (S_1 - S_0)^2 + (A_1 - A_0)^2 + (L_1 - L_0))} \tag{1}$$

The replicas that are closer to MR are grouped to form a cluster. The replica with the shortest distance from the MR is selected as the best replica. The cluster metric, M, is calculated as in [4]. The implementation parameters for the proposed algorithm, HGASA are initial population is 50, mutation probability is 0.9, crossover probability is 0.1, initial temperature is 10000 and cooling rate is 0.9. The implementation of HGASA algorithm for replica selection problem is shown in Algorithm 1.

4 Experimental Results

The ACO, GA, SA and HGASA algorithms are implemented using Intel CORE i5 processor and simulated in GridSim toolkit [5] for selecting best replica location. The number of sites in the grid network will be defined by the user and with varying performance in time, availability, security and load balancing. The number of grid sites is twenty in the simulation of existing and proposed algorithms. The performance of the ACO, GA, SA and HGASA are calculated for two different scenarios such as 10 user requests and 25 user requests. The efficiency is calculated using the Formula in [4]. GA is 44 % more efficient than ACO algorithm for selecting replica in data grid. HGASA shows 21 %, 9 % and 50 % more improvement in efficiency when compared to SA, GA and ACO respectively. When two or more sites have the best possible performance in terms of response time, security, availability and load balancing which are equal in proportional value but vary in the order, then randomly one among them is selected for creating the replica. All the factors are equally considered and one factor is not preferred over the other during replication. The response time, security, availability and load balancing for all the twenty sites are generated randomly from 75 to 95.

```
Algorithm HGASA ()
{
1. Input all necessary Parameters
2.  Generate Initial Population
3. Generate Initial Temperature to all the individuals
4. Calculate the fitness of all the chromosomes
5. Select a random node as parent node
6. Perform crossover with some other node satisfying fitness function
7. Check the fitness probability of the child chromosome [17]
8. Apply mutation operation
9. Calculate the fitness function of the new population [4]
10. Set T[i]=T[i]*Cooling Rate
11. If the child's fitness is better than the parent, the child replaces the
    parent
12. Repeat steps 5-11 until termination condition is met
}
```

Algorithm 1: HGASA

5 Conclusion

HGASA based Replica Selection in Data Grid improves the efficiency of selecting the best replica site for user requests during job execution. The efficiency is improved by increasing the number of parameters such as response time, availability of the file, security and load balancing. The efficiency of the HGASA algorithm is compared with GA, SA and ACO. ACO algorithm does not deal with availability, security and load balancing, and therefore not efficient when compared to genetic algorithm, simulated annealing and HGASA. The efficiency of genetic algorithm and simulated annealing is 44 % and 36 % greater than ACO algorithm. The proposed algorithm, HGASA, performs better than all the three algorithms, namely, GA, SA and ACO by 9 %, 21 % and 50 % respectively. The efficiency can be improved further by considering other parameters such as bandwidth, scheduling strategies, access pattern that are important for job execution.

Acknowledgements. The authors would like to thank the Management & Principal of Sri Ramakrishna Engineering College, and the Head of the Department of Computer Science and Engineering, for their support in completing this work.

References

1. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
2. Khanli, L.M., Isazadeh, A., Shishavan, T.N.: PHFS: a dynamic replication method, to decrease access latency in the multi-tier data grid. Future Gener. Comput. Syst. **27**(3), 233–244 (2011)

3. Tim, H., Abramson, D.: The griddles data replication service. In: Proceedings of the 1st International Conference on E-Science and Grid Computing, pp. 271–278 (2005)
4. Jadaan, O.A., Abdulal, W., Hameed, M.A.: Enhancing data selection using genetic algorithm. In: Proceedings of IEEE International Conference on Computational Intelligence and Communication Networks, pp. 434–439 (2010)
5. Buyya, R., Murshed, M.: GridSim: A toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing. J. Concurrency Comput. Pract. Experience **14**, 1175–1220 (2002)
6. Lin, Y., Chen, Y., Wang, G., Deng, B.: Rigel: a scalable and lightweight replica selection service for replicated distributed file system. In: 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, CCGC, pp. 581–582 (2010)
7. Naseera, S., Murthy, K.V.M.: Performance evaluation of predictive replica selection using neural network approaches. In: Proceedings of International Conference on Intelligent Agent and Multi-Agent Systems, IAMA 2009, p. 1 (2009)
8. Rahman, R.M., Baker, K., Alhajj, E.: A predictive technique for replica selection in grid environment. In: Seventh IEEE International Symposium on Cluster Computing and the Grid, pp. 163–170 (2007)
9. Ishii, R.P., De Mello, R.F.: An online data access prediction and optimization approach for distributed systems. IEEE Trans. Parallel Distrib. Syst. **23**(6), 1017–1029 (2012)
10. Sun, M., Sun, J., Lu, E., Yu, C.: Ant algorithm for file replica selection in data grid. In: Proceedings of First International Conference on Semantics, Knowledge and Grid, p. 64 (2005)
11. Holland, J.: Adaptation in Natural Artificial Systems. University of Michigan Press, Ann Arbor (1992)
12. Olivas, E.S., Guerrero, J.D., Martinez-Sober, M., Magdalena-Benedito, J.R., Serrano Lopez, A.J.: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI Global, Hershey (2010). doi:10.4018/978-1-60566-766-9
13. Kirkpatrick, S., Gelatt, C., Vecchi, M.: Optimization by simulated annealing. Science **220**, 671–680 (1983)
14. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculation by fast computing machines. J. Chem. Phys. **21**(1087), 1087–1091 (1953)
15. Yoshikawa, M., Yamauchi, H., Terai, H.: Hybrid architecture of genetic algorithm and simulated annealing. Eng. Lett. **16**(3), EL_16_3_11 (2012)
16. Chervenak, A., Schuler, R., Ripeanu, M., Amer, M.A., Bharathi, S., Foster, I., Kesselman, C.: The globus replica location service: design and experience. IEEE Trans. Parallel Distrib. Syst. **20**(9), 1260–1272 (2009)
17. Gandomkar, M., Vakilian, M., Ehsan, M.: A combination of genetic algorithm and simulated annealing for optimal DG allocation in distribution networks. In: Proceedings of Canadian Conference on Electrical and Computer Engineering, pp. 645–648 (2005)