Advanced Science and Technology Letters
Vol.50 (CST 2014), pp.74-79
http://dx.doi.org/10.14257/astl.2014.50.12

Genetic Algorithm Based Optimization Model for Reliable Data Storage in Cloud Environment

Feng Liu[1,2,3], Haitao Wu[1,3], Xiaochun Lu[1,3], Xiyang Liu[4], Lei Fan[4]

[1] National Time Service Center, Chinese Academy of Sciences, 3 East Shuyuan Road, 710600 Xi'an, China
[2] University of Chinese Academy of Sciences, 19A Yuquan Road, 100049 Beijing, China
[3] Key Laboratory of Precision Navigation and Timing Technology, Chinese Academy of Sciences, 3 East Shuyuan Road, 710600 Xi'an, China
[4] Xidian University, 2 South Taibai Road, 710071 Xi'an, China
elkhood@163.com

Abstract. Massive data storage is one of the great challenges for cloud computing service, and reliable storage of sensitive data directly affects quality of storage service. In this paper, based on analysis of data storage process in cloud environment, the cost of massive data storage is considered to be comprised of data storage price, data migration and communication; and the storage reliability consists of data transmission reliability and hardware dependability. A multi-objective optimization model for reliable massive storage is proposed, in which storage cost and reliability are the objectives. Then, a genetic algorithm for solving the model is designed. Finally, experimental results indicate that the proposed model is positive and effective.

Keywords: Cloud storage, multi-objective optimization model, genetic algorithm, reliable storage

1 Introduction

The development of information-based society entails that more and more resources are being digitized, causing endless growth of data resource storage capacity, and thus resulting in substantial increase of storage costs. Moreover, different applications require different storage capacity. However, storage space assigned to these applications is often not fully utilized. The service provider is facing the tradeoff between rapid growth of information resources and control of the costs. On one hand, service providers not only produce enormous important data by themselves, but also require massive information resources. On the other hand, large amount of storage equipment and manpower are needed to store the information resources. Therefore, new data storage devices required by service provider should include features such as storage virtualization [1], dynamically extensible storage capacity [2][3] and reliable data storage.

Cloud storage is a new concept derived from cloud computing, referring to a system which assembles plentiful different small and large storage devices of same

ISSN: 2287-1233 ASTL

type in the grid by cooperating application software to jointly and externally provide functions of data storage and business access using cluster application, grid technology or Distributed File System (DFS). Compared with traditional storage devices, cloud storage is not only hardware but also a complex system comprised of network device, storage device, server, application software, public access interface, Access Network (AN), client program, etc. Each part provides data storage and business access through application software with the storage device as the core. Strictly speaking, cloud storage is a service rather than a storage device. Cloud storage provider may provide personalized storage services according to clients' demands, such as storage space, network bandwidth, data safety, disaster recovery performance, etc.

Currently, cloud storage is a field in which hi-tech enterprises compete; many hi-tech enterprises have launched their own cloud data storage products including Cloud Drive by Amazon, Live Mesh by Microsoft, Does by Google, etc. Cloud storage has become a tendency for future storage development. With development of cloud storage technology, applications that combine techniques of all kinds of searches and applications pertaining to cloud storage should be improved from the view of safety, reliability, data access, etc.

2 Problem Analysis

2.1 Risk Analysis of Data Resources in Cloud Storage Environment

According to literature [4], Gartner indicates that seven security risks are faced by cloud computing: privileged user access, regulatory compliance, data location, data segregation, recovery, investigative support, and long-term viability. Among different safety risks, data disaster recovery is the most important issue which needs to be considered by cloud storage providers and clients. Enterprises hand over their sensitive and important commercial data to cloud service provider. Therefore, loss of the data will not only bring enterprises fatal disaster but also cause legal dispute between enterprises and their clients. For example, in banking system, client data is vital for both the bank and its clients; Unrecoverable data loss will result in incalculable loss for the bank and its clients. Therefore, for cloud storage providers, it is their own business objective to guarantee the safety, reliability and recoverability of data. Nevertheless, when providing some important clients with storage service, legal clauses are signed with the clients, so that if data loss occurs, clients will be compensated. Therefore, guarantee of data disaster recoverability is of significant importance for cloud storage providers.

2.2 DC (Data Center)

Wikipedia defines a data center as "a complex set of facility which includes not only computer systems and associated components such as telecommunications and

storage systems, but also redundant data communications connections, environmental controls and various security devices." In literature [5], Google illustrates datacenters as "buildings with multi functions, where multiple servers and communication gear are collocated because of their common environmental requirements and physical security needs, and ease of maintenance", other than just "a collection of co-located servers".

1) Server: Compared with PC, server has more reliable continuous operating capability, more powerful storage and network communication capability, faster failure recovery capability and has easier extension space. Data backup functions are also required by applications which are sensitive to data. In reality, servers are located in geographically different locations. Therefore the influence caused by geographical distance between servers should be considered when storing the data. In practical applications, servers are usually connected by internet or routers. Considering each server as node, the connections as sides, connection of servers can be shown as a graph, more specifically a complete graph.

2) Storage capability: Hardware is vital for data storage, which directly affects DC's data storage capability and storage safety. Nowadays, various types of hard disks can be found in the market, each having its own characteristics. Hard disk interface is the connecting component between hardware and host system, and functions as the data transmission device between HDD (Hard Disk Drive) caching and host memory. Different hard disk interfaces determine speed of connection between hard disk and computer. In the whole system, quality of hard disk interface directly affects program running speed and system performance level. In DC, the future of enterprise-based disk array lies in mixed use of SSD (Solid State Disk) and HDD. For the sake of cost control, both should be taken into consideration. HDD can provide massive storage space, while SSD can provide high performance. Generally, for hard disks used for server configuration circulated in the market, its price is directly proportional to its stability and data read/write speed;

3 Model Establishment

Since cloud storage is a burgeoning service, the users are highly concerned about data storage reliability and safety, in addition to the price of the service. To some extent, the data submitted by users may involve their privacy or interest. Therefore, more suitable storage service which meets their requirements may be chosen by users regarding storage price vs. reliability. Hence, when building the mathematic model, storage service cost and storage reliability are both considered as objectives, and multi-objective optimization model of reliable storage of cloud storage is proposed.

3.1 Assumptions

Assuming server cluster owned by certain DC is $S = \{s_1, s_2, \ldots, s_N\}$, the servers are allocated at different places with geographical difference among them. The matrix $(d(s_i, s_j))_{N \times N}$ represents spatial distance between two servers,

where $d(s_i, s_i) = 0, i = 1 \sim N$ and $\forall i, j = \{1, 2, \ldots, N\}, i \neq j$, then $d(s_i, s_j) \geq 0$. Assuming data set submitted to DC by users is $F = \{f_1, f_2, \ldots, f_M\}$, where f_i is indivisible data block, each data block has its own safety level $L(f_i), i = 1 \sim M$.

Different prices are charged for each server storing data with different safety level mainly because data files with different safety level have difference backup file numbers. The number and location of backups are decided according to the safety level of data file specified. Therefore, when providing storage service, data files submitted by users may be migrated among servers in DC according to requirement. Thus, during migration, an overhead cost arises from communication and migration, along with some problems related to reliability. Reliability in migration is mainly related to connection stability and distance among servers.

For convenient model description, the following parameters are introduced:

- Data file storage identification: if the data file f_i is stored in the k type hard disk of server s_j, then mark the matrix $Y_{ij} = k$, or $Y_{ij} = 0$
- Sign function: $sign(\cdot)$ is defined as
  $sign(y) = \begin{cases} 1 & if\ y > 0 \\ 0 & else \end{cases}$
- Function $X(f_j, s_i)$ represents the storage space of data file f_i in server s_i
- Intrinsic reliability $R(s_i)$ owned by the server s_j itself
- Reliability $TR(s_i, s_j)$ of data transmission among servers,
  where $TR(s_i, s_j) \in [0,1]$, and $TR(s_i, s_j)$ is inversely proportional to $d(s_i, s_j)$
- There are K types of hard disk in DC, current available capacity of various hard disks in server s_i can be denoted as $\{A_1(s_i), A_2(s_i), \cdots, A_K(s_i)\}$, and the h type of hard disk has reliability of its own: $P(h), h = 1, 2, \cdots, K$
- Data storage reliability, which is the product of transmission reliability, hardware reliability and server reliability during data transmission; the formula is:

$$SR(f_i, R(s_j), X(f_i, s_j), TR(s_j, s_B)) = \sum_{j=1}^{N} R(s_j) X(f_i, s_j) TR(s_j, s_B) P(Y_{ij}) \quad (1)$$

3.2 Storage Cost

Different prices are charged for each server providing storage service to data at different safety levels, denoted as $C_s(s_i, L(f_j))$. Plus, data files with different safety levels have different backup file numbers; higher safety levels correspond to more backups. Here, safety level is made equivalent to the amount of data file backups, and these backups are stored in servers at different geographic locations which complies with our practice of data storage in cloud computing. Hence our assumption is reasonable. The storage price of file f_i is:

$$C_c(f_i) = \sum_{i=1}^{N} C(s_i, L(f_j)) X(f_j, s_i) \tag{2}$$

The storage price of data file F submitted by user is:

$$C_c(F) = \sum_{j=1}^{M} \sum_{i=1}^{N} C(s_i, L(f_j)) X(f_j, s_i) \tag{3}$$

3.3 Migration Cost

Since servers are located in different geographic locations, different migration distance occurs during migration of data block from one server to another. Hence, migration cost is related to migration distance and size of data file. Therefore, the migration cost of data block f_k migrated from the server s_i to the server s_j is expressed as:

$$C_{mig}(f_k, s_i, s_j) = C_m X(f_k, s_i) d(s_i, s_j) \tag{4}$$

Where, C_m is parameter of migration cost.

Hence, the migration cost of data file F migrated from server s_i to another server is:

$$\sum_{k=1}^{M} \sum_{j} C_{mig}(f_k, s_i, s_j) = \sum_{k=1}^{M} \sum_{j} C_m X(f_k, s_i) d(s_i, s_j) \tag{5}$$

3.4 Communication Cost

Data transmission among servers mainly embodies in communication flow due to different geographic locations of servers, hence, distance between source server (s_i) and host server (s_j) is another factor influencing communication cost, expressed as:

$$C_{com}(f_k) = \sum_{i=1, i\neq j}^{N} \sum_{j=1}^{N} [X(f_k, s_i) W(s_i, s_j) d(s_i, s_j)] \tag{6}$$

Where, $W(s_i, s_j)$ represents communication cost of servers.

The whole communication cost of data file F after storage is:

$$C_{com}(F) = \sum_{k=1}^{M} \sum_{i=1, 1\neq j}^{N} \sum_{j=1}^{N} [X(f_k, s_i) W(s_i, s_j) d(s_i, s_j)] \tag{7}$$

4 Conclusion

By analyzing data storage information in cloud environment, a multi-objective optimization model for reliable storage is built. In view of data files with safety requirement stored by users, in the model, both data storage cost including storage price, migration cost and communication cost and data reliability including

transmission reliability in storage process and storage device reliability after data storage are considered. In order to validate correctness and availability of the model, a multi-objective GA is designed under the framework of algorithm NSGA-II for model solution. In the experiments, computations are carried out through constructing several storage situations by using parameters published by existing commercial cloud storage services. The experimental results validate correctness and availability of the model, and show that the model can provide multiple storage solutions for users so that storage resources in DC can be effectively utilized.

References

1. Aameek, S., Madhukar, K., Dushmanta, M.: Server-storage Virtualization: Integration and Load Balancing in Data Centers[C]. In: Proceeding of the 2008 ACM/IEEE Conference on Supercomputing (2008) 1-12
2. Gregory, C., Idit, K,, Rachid, G.: Reliable distributed storage[J]. IEEE Computer Society, Vol. 42, No. 4 (2009) 60-67
3. Deng, Y.H., Wang, F., Helian, N.: Dynamic and Scalable Storage Management Architecture for Grid Oriented Storage Devices [J]. Parallel Computing, Vol. 34, No. 6 (2008) 17-31
4. Heiser, J., Nicolett, M.: Assessing the Security Risks of Cloud Computing. http://www.gartner.com (2008)
5. Luiz, B., Urs, H.: The Datacenter as a Computer. Morgan & Claypool Publishers, California (2009)
6. Fan, L.: Research on Efficient and Intelligent Algorithms for Two Classes of Complex Optimization Problem. Xidian University (2012)
7. Ghemawat, S., Gobioff, H., Leung, S.T.: The Google File System[C]. In: Proceedings of the 19th ACM Symposium on Operating Systems Principles, New York (2003) 32-47
8. Wang, C., Wang, Q., Kui, R., Lou, W.J.: Ensuring Data Storage Security in Cloud Computing[C]. In: 17th IEEE International Workshop on Quality of Service (2009) 1-9
9. Alves, A., Viegas, C., Nejdl, P.: A Distributed Tabling Algorithm for Rule Based Policy Systems. IEEE Computer Society, Vol. 15, No. 5 (2004) 123-132
10. Thomas, C.: The Basics of Reliable Distributed Storage Networks [J]. IEEE Computer Society, Vol. 6, No. 3 (2004) 16-24
11. Yang, J.F., Chen, Z.B.: Analysis and Research of Cloud Computing System Instance[C]. In: Proceedings of the 2010 Second International Conference on Future Networks (2010) 88-92
12. Hayes, B.: Cloud Computing [J]. Communications of the ACM, Vol. 51, No. 7 (2008) 30-34
13. Gurumurthi, S.: Architecting Storage for the Cloud Computing Era [J]. IEEE Computer Society, Vol. 34, No. 6 (2009) 124-135
14. Wu, J.Y., Ping, L.D., Ge, X.P.: Cloud Storage as the Infrastructure of Cloud Computing[C]. In: Proceedings of the 2010 International Conference on Intelligent Computing and Cognitive Informatics (2010) 380-383
15. Buxmann, P., Hess, T., Lehmann, S.: Software as a Service. Business and Economics, Vol. 50, No. 6 (2008) 500-503

