International Journal of Grid Distribution Computing
Vol.7, No.6 (2014), pp.175-190
http://dx.doi.org/10.14257/ijgdc.2014.7.6.14

Optimization Model of Reliable Data Storage in Cloud Environment Using Genetic Algorithm

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

[1]*National Time Service Center, Chinese Academy of Sciences, 3 East Shuyuan Road, Xi'an 710600, China*

[2]*University of Chinese Academy of Sciences, 19A Yuquan Road, Beijing 100049, China*

[3]*Key Laboratory of Precision Navigation and Timing Technology, Chinese Academy of Sciences, 3 East Shuyuan Road, Xi'an 710600, China*

[4]*Xidian University, 2 South Taibai Road, Xi'an 710071, 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 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

ISSN: 2005-4262 IJGDC

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) Storage reliability: With the development of cloud computing, in recent years, more and more individuals and enterprises choose to migrate their businesses to large DC to reduce their own operating cost. The value of data embodies in its context as well as potential benefit brought for business development, hence data is becoming more important day by day, and thus higher requirements of data storage reliability are demanded by users. Massive data storage presents a great deal of challenges. In view of DC level, massive data storage management system is highly complex. As for storage equipment level, numerous cheap storage devices are used by DC to control cost; therefore data loss caused by failure of storage device tends to occur easily. For example, Data service interruption accidents occurring to Google and Amazon began concerning users about DC storage reliability. Therefore, how to improve storage reliability of large-scale data in DC has become a hot topic for research in recent years.

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_j in server s_i
- Intrinsic reliability $R(s_i)$ owned by the server s_i 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:

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

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:

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

3.5. Total Cost for Storage

Total storage cost of data file F is the sum of storage fee, migration cost and communication cost of each data block.

$$\begin{aligned} C(F,S) &= C_c(F) + \sum_{k=1}^{M}\sum_{j} C_{mig}(f_k, s_i, s_j) + C_{com}(F) \\ &= \sum_{j=1}^{M}\sum_{i=1}^{N} C(s_i, L(f_j))X(f_j, s_i) + \sum_{k=1}^{M}\sum_{j} C_m X(f_k, s_i) d(s_i, s_j) \\ &+ \sum_{k=1}^{M}\sum_{i=1, i\neq j}^{N}\sum_{j=1}^{N} [X(f_k, s_i)W(s_i, s_j)d(s_i, s_j)] \end{aligned} \tag{8}$$

In the time of migration, all servers S are required to be traversed, after which minimum migration cost is obtained.

3.6. Reliability of Data Storage

$R(s_i)$ is intrinsic reliability owned by server s_i, $TR(s_i, s_j)$ is reliability of data transmission among servers, where, $TR(s_i, s_i) = 1, TR(s_i, s_j) \in [0,1)$, and $TR(s_i, s_j)$ is inversely proportional to $d(s_i, s_j)$. When $TR(s_i, s_j) = 0$, data is unable to be migrated between severs s_i and s_j.

The p type of hard disk has intrinsic reliability, denoted as $P(h), h = 1, 2, \cdots, K$.

Data storage reliability, which is the product of transmission reliability, hardware reliability and server reliability during data transmission, has the following formula:

$$S_r(f_i, R(s_j), X(f_i, s_j), TR(s_j, s_B)) = \sum_{j=1}^{N} R(s_j) X(f_i, s_j) TR(s_j, s_B) P(Y_{ij}) \tag{9}$$

4. Mathematical Model

4.1. Objective Analysis

1) Cost calculation: when data file F is submitted by user from certain server s_B, the total cost of data storage, which is the sum of storage cost, communication cost and migration cost, is calculated by DC. The objective is acquisition of optimal storage solution to minimize total storage cost.

2) Storage reliability: After data is submitted by user, DC needs to reach maximum reliability, which mainly includes transmission reliability and device dependability during storage.

4.2. Constraints

1) Storage constraint: Total data size in each server is no more than available capacity of hard disk;

2) Off-side storage of backup data: Each data file and its backup should not be stored in the server in identical location;

3) Reliability constraint: Intrinsic reliability owned by hard disk or server is no less than safety factor owned by data block file.

4.3. Multi-objective Optimization Model

By aggregating each index analyzed above, multi-objective optimization model of reliable storage in cloud computing environment is obtained:

$$\begin{cases} \min \ C(F,S)=\sum_{j=1}^{M}\sum_{i=1}^{N}C(s_i,L(f_j))X(f_j,s_i)+\sum_{k=1}^{M}\sum_{j}C_m X(f_k,s_i)d(s_i,s_j) \\ \qquad\qquad +\sum_{k=1}^{M}\sum_{i=1,i\neq j}^{N}\sum_{j=1}^{N}[X(f_k,s_i)W(s_i,s_j)d(s_i,s_j)] \\ \max \ SR(F,S)=\sum_{i=1}^{N}\sum_{j=1}^{M}S_r(f_i,R(s_j),X(f_i,s_i),TR(s_j,s_B)) \\ \qquad\qquad =\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{j=1}^{N}R(s_j)X(f_i,s_j)TR(s_j,s_B)P(Y_{ij}) \\ s.t. \quad L(f_i)\le R(Y_{ij}), i=1,2,\ldots,M \\ \qquad f_i \le AY_{ij}(s_j), i=1,2,\ldots,M \\ \qquad \sum_{j=1}^{N}sign(Y_{ij})\ge L(f_i), i=1,2,\ldots,M \\ \qquad R(s_i)\in[0,1), i=1,2,\ldots,M \\ \qquad P(h)\in[0,1), h=1,2,\ldots,K \end{cases} \tag{10}$$

Since objective functions in the above mathematical model contain both maximized objective and minimized objective, it is not convenient to optimize both the objectives using the algorithm. For a more convenient solution of above multi-objective optimization problem, we equivalently transform them to multi-objective minimization problem:

$$\begin{cases} \min \ C(F,S) \\ \min \ SR'(F,S)=1-SR(F,S) \\ s.t. \quad L(f_i)\le R(Y_{ij}), i=1,2,\ldots\ldots,M \\ \qquad f_i \le A_{Y_{ij}}(s_j), i=1,2,\ldots\ldots,M \\ \qquad \sum_{j=1}^{N}sign(Y_{ij})\ge L(f_i), i=1,2,\ldots\ldots,M \\ \qquad R(s_i)\in[0,1), i=1,2,\ldots\ldots,M \\ \qquad P(h)\in[0,1), h=1,2,\ldots\ldots,K \end{cases} \tag{11}$$

After equivalent transformation, the less $SR'(F,S)$ is, the higher reliability is. Hence, in the following numerical experiment, the problem (11) is validated and solved.

5. GA for Solving Multi-Objective Optimization

Currently, there are three types of methods for solving multi-objective optimization: aggregate multi-objective into single objective, non-Pareto and Pareto. Although the method of aggregate multi-objective into single objective is simple in design and efficient in operation, at the end only one efficient solution can be obtained. Multiple efficient solutions can be obtained by Non-Pareto, however these solutions are all concentrated at endpoints of effective interfaces with too many non-inferior solutions being lost. Pareto-based optimization method usually maps multi-objective values directly to fitness function, and effective solution set is searched by comparing dominant relation of function values. Thus a series of non-inferior solutions can be obtained, making it an effective method. Due to connotative concurrency, randomness and high robustness of GA, it is mostly applied to Pareto-based optimization method. A series of classic algorithms are proposed and successfully applied.

5.1. Flow of GA

According to GA concept, simple GA flow can be shown as Figure 1.

5.2. Solution Algorithm

In this paper, the noted multi-objective genetic algorithm NSGA-II is used for solving multi-objective optimization of reliable storage. According to features of the model in this paper, parameters of NSGA-II are set as:

- Population size: $N = 100$
- Crossover: using simplex crossover operator for individual crossover, where, crossover probability is $p_c = 0.95$, high crossover probability is used to raise updating speed of algorithm population, therefore increasing the convergence rate of the algorithm

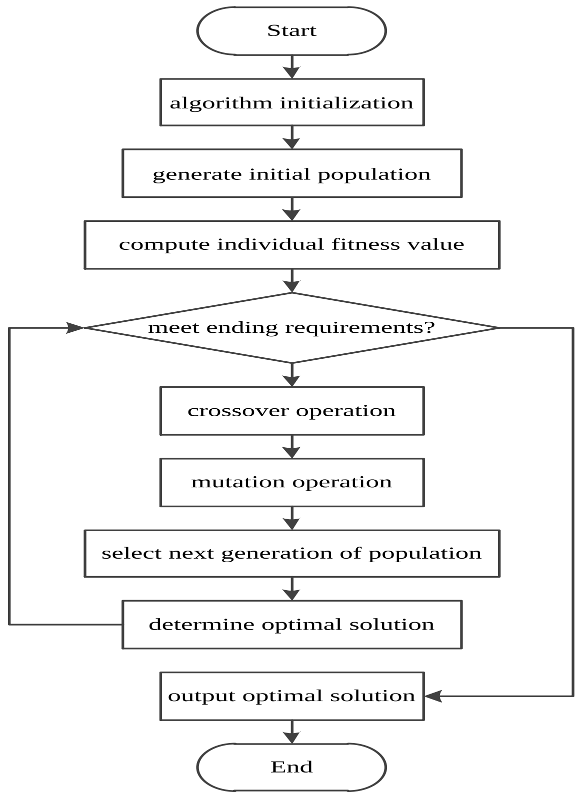

Figure 1. Flow Chart of GA

- Mutation: using Gauss mutation operator for individual mutation, mutation probability is $p_m = 0.9$. Since constrained multi-objective optimization is a problem, feasible solution is difficult to be obtained. Hence, mutation probability is set as 0.9 to improve search performance of the algorithm.
- Strategy selection: using roulette selection strategy to select individuals for crossover or mutation, elitism strategy is used to preserve next generation of population.
- Stop condition: when algorithm iterations reach 400 times, or solution found by algorithm does not change in 30 iterations, the algorithm will stop running and the found solution will be output.

The literature [6] has theoretically proved that utilization of elitism strategy and Gauss mutation can guarantee GA's convergence to optimal solution by probability 1. Therefore, we adopt the two strategies to guarantee convergence of the algorithm.

6. Simulation Experiment

6.1. Experiment Setup

For the experiments, we adopt two cloud computing service environments, namely EC2 (from Amazon) and GoGrid, to validate the correctness and availability of the proposed model. Storage type configuration and storage price used are obtained from data openly published by the two companies, in which cloud storage prices of EC2 from Amazon are classified as US-east price, US-west price and EU-west price. Since storage of data file of higher safety level requires specific storage device and encryption algorithm with corresponding safety level for processing, storage of this kind of data file demands more CPU computing time, resulting in higher storage price. The detailed data is shown in Table 1-5.

Table 1. Amazon EC2 Storage Instance and Price

Name	CPUs	Memory(GB)	Price($/hour)		
			US-east	US-west	EU-west
m1.small	1	1.7	0.085	0.095	0.095
m1.large	4	7.5	0.34	0.38	0.38
m1.xlarge	8	15	0.68	0.76	0.76
c1.medium	5	1.7	0.17	0.19	0.19
c1.xlarge	20	7	0.68	0.76	0.76
m2.xlarge	6.5	7	0.50	0.57	0.57
m2.2xlarge	13	34.2	1.00	1.14	1.14
m2.4xlarge	26	68.4	2.00	2.28	2.28

Table 2. GoGrid Storage Type and Price

Name	CPUs	Memory(GB)	Price($/hour)
			EU-west
X-Small	0.5	0.5	0.095
Small	1	1	0.19
Medium	2	2	0.38
Large	4	4	0.76
X-Large	8	8	1.52
XX-Large	16	16	3.04

Table 3. Price of EC2 and Gogrid Network Data Transmission

Cloud Service	Price($/GB)	
	Data Input	Data Output
EC2	0.10	0.15
GoGrid	0.00	0.29

Table 4. Available Bandwidth from UA to Cloud Service Provider

Cloud Service	Average Bandwidth (MB/s)	Transmission Time/GB(s)	Stdev(%)
EC2 us-east-1	1.54	665	40.87
EC2 us-west-1	1.19	864	42.43
EC2 eu-west-1	24.16	42	23.58
GOGrid EU West	5.91	173	32.06

Table 5. Private Storage Types

Name	CPUs	Memory(GB)
Small	4	2
Large	8	8
Xlarge	16	16

6.2. Analysis of Experimental Results

The experimental data file is stored as one integral data file and also divided into a group of sub data files. When the data file is treated as one integral data file, the storage solution is to store all data files in an identical server. When the data file is treated as a group of sub data files, they can be stored in different servers. This setting helps to observe the effect of solutions of different types of data file storage in cloud storage environment on storage cost and reliability. The experimental result is shown in Figure 3-5. Where, each point in the figures represents corresponding storage cost and reliability concluded by optimal storage solution, using the multi-objective optimization model for storage reliability proposed in this paper.

Firstly, in order to further analyze the effect on migration frequency of data files during storage to storage reliability, relative experiments for analysis are performed without considering the storage cost. In the experiments, sub data files with higher safety level are of higher proportion. In order to improve precision of analysis, multiple experiments are performed to record average migration frequency of all sub data files and reliability of corresponding solution, the result of which is shown in Figure 2. According to Figure 2, storage reliability of data files decreases sharply with increasing average migration frequency. During storage, reliability of data to be stored is mainly influenced by stability of network transmission. When providing storage service, if all data is stored in one server without any data file transferring, most reliable storage can be achieved for all data files. However, the cost for storage service is highest in this case. In order to reduce storage cost, data files may be transferred, but that will obviously reduce storage reliability of the data files, which is evident in Figure 2. In cloud environment, higher length and higher frequency of data files transfer result in lower reliability for the whole data storage service.

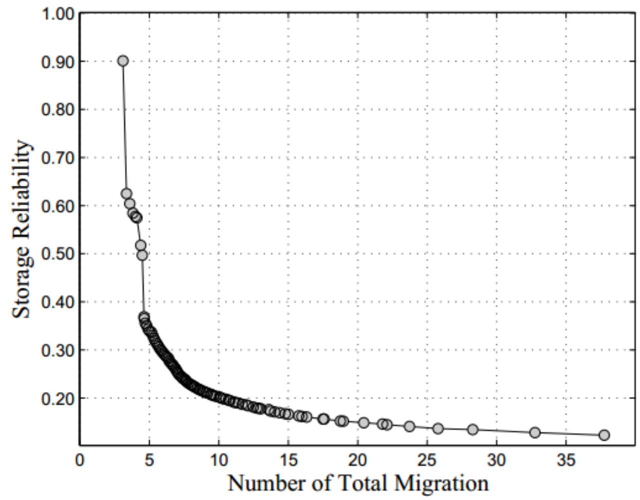

Figure 2. Effect of Total Migration Frequency of Data Files on Storage Reliability

In the next phase of the experiment, all sub files of experimental data files are stored as one integral file in order to observe effect of sub data files on storage solution. Two types of conditions are tested: equal proportion distribution of sub data files at various safety levels, and higher proportion of files at higher safety levels compared to those at lower safety levels. The results are shown in Figure 3 and Figure 4. According to both figures, it is obvious that for increased proportion of sub data files at higher safety levels, storage cost is also increased, and storage solution produced by the model is evidently less than when sub data files at various safety levels are equally distributed. This complies with practical application, thereby validating the correctness of the model.

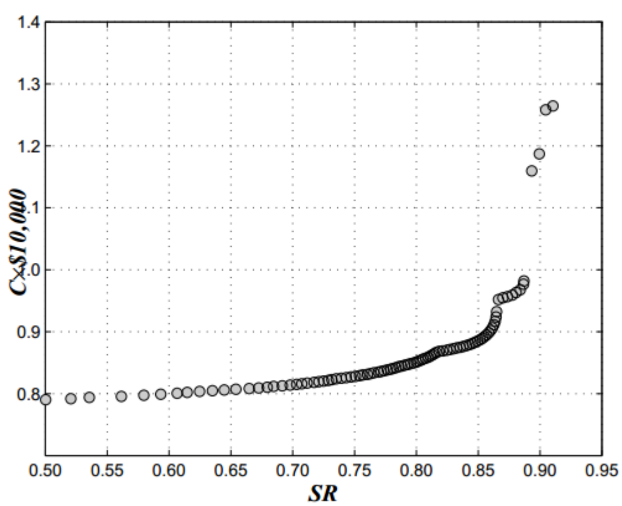

Figure 3. Integral Storage of Data Files
(data files at various safety levels are distributed equally in proportion)

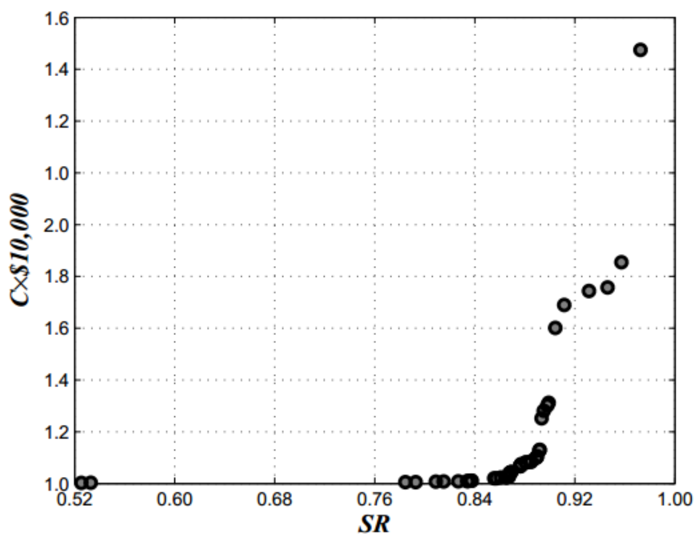

Figure 4. Integral Storage of Data Files
(proportion of data files at higher safety levels is higher than those at lower safety levels)

In the next phase, by storing each sub file of experimental data individually, the effect of proportion of sub files at different safety levels in data file on storage solution is observed. The conditions mentioned in the previous paragraph are also considered here. The results are shown in Figure 5 and Figure 6. According to the results, when sub files at various safety levels in experimental data are distributed with same proportion, more uniformly distributed storage solutions can be obtained using the model proposed in this paper, In the output storage solutions, storage reliability distributes moderately uniformly between 0.63~0.95, which is beneficial for providing clients with multi representative solutions. When proportion of sub files of experimental data at higher safety levels is higher, number of storage solutions, uniformity and universality of distribution clearly decrease. According to comparison between Figure 5 and Figure 6, the change of storage cost for storage solution shown in Figure 6 is obviously greater than that shown in Figure 5. This is because different storage prices are provided for storing files at higher safety levels by different servers or storage devices, larger data file at higher safety level will result in dramatic change of storage cost. Figure 6 shows when reliability is between 0.83~0.95, storage solutions without great change of storage cost are observed. The quick increase of migration frequency of data files in these solutions result in decrease of reliability during migration.

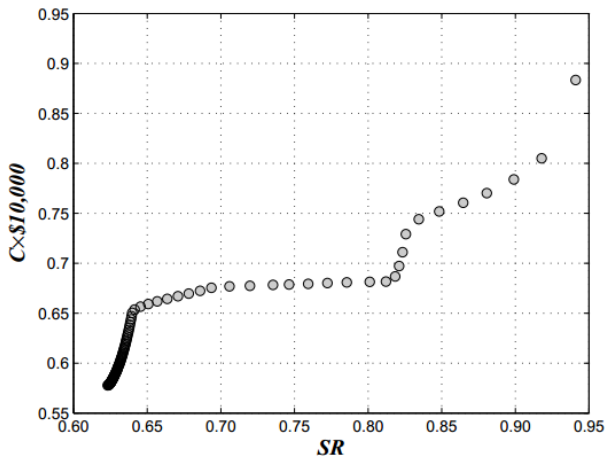

Figure 5. Respective Storage of Sub Data Files
(data files at various safety levels are distributed equally in proportion)

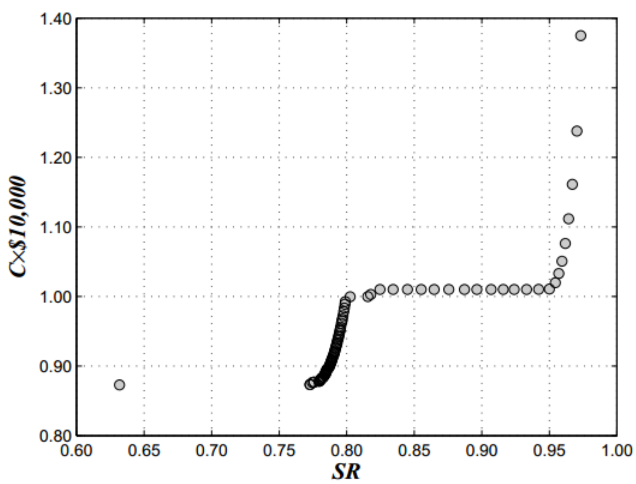

Figure 6. Respective Storage of Sub Data Files
(the proportion of data files at higher safety levels is higher than that at lower safety levels)

Finally, comprehensive analysis is made by comparing Figures 3, 4, 5 and 6. Among sub files at various safety levels of experimental data files distributed with various proportions, the storage cost of storing sub files of same reliability separately is noticeably less than storing the sub files as one integral file. According to Figure 3, if the experimental data files are stored as one integral file, the solution of storage cost with reliability greater than 0.5 obtained by the model in this paper is between 10~16 thousand dollars. Storage reliability and storage cost corresponding to different storage solutions vary significantly. Storage cost increases sharply with increased storage reliability. This is due to the fact that when integrally storing 120 GB-sized data files, one first needs to find a server with available storage space of over 120 GB. Moreover, storage prices of servers in different location are different. Migration from the server where data was initially submitted, to a different target server will not only produce higher migration cost and communication cost, but also reduce reliability of data storage. According to Figure 5, if the sub data files of experimental data files are stored individually, the optimal solution to storage cost obtained by the model in this paper is between 5500~9500 dollars, the storage reliability is between 0.60~0.95, the distribution is relatively uniform, and storage cost and reliability is relatively stable. Compared with Figure 3, respective storage of sub files of data files can obviously reduce storage cost, and on the premise that storage reliability is guaranteed, more storage solutions can be found by the proposed model. Hence, individual storage of data files at various safety levels can not only realize cost reduction for data storage with prerequisite of storage reliability guaranteed, but also achieve more storage solutions, which is helpful for service providers to provide more diversified storage solutions so as to improve their attractiveness in storage service.

7. 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.

Acknowledgement

This paper is a revised and expanded version of a paper entitled Genetic Algorithm Based Optimization Model for Reliable Data Storage in Cloud Environment presented at CST 2014, Indonesia, June 19-22, 2014.

References

[1] S. Aameek, K. Madhukar and M. Dushmanta, “Server-storage Virtualization: Integration and Load Balancing in Data Centers”, Proceeding of the 2008 ACM/IEEE Conference on Supercomputing, **(2008)** November 15-21, Texas, USA.

[2] C. Gregory, K, Idit and G. Rachid, “Reliable distributed storage”, IEEE Computer Society, vol. 42, no. 4, **(2009)**.

[3] Y. H. Deng, F. Wang and L. N. He, “Dynamic and Scalable Storage Management Architecture for Grid Oriented Storage Devices”, Parallel Computing, vol. 34, no. 6, **(2008).**

[4] J. Heiser and M. Nicolett, “Assessing the Security Risks of Cloud Computing”, http://www.gartner.com, **(2008).**

[5] B. Luiz and H. Urs, “The Datacenter as a Computer”, Morgan & Claypool Publishers, California **(2009).**

[6] L. Fan, “Research on Efficient and Intelligent Algorithms for Two Classes of Complex Optimization Problem”, Xidian University, **(2012).**

[7] S. Ghemawat, H. Gobioff and S. T. Leung, “The Google File System”, Proceedings of the 19th ACM Symposium on Operating Systems Principles, **(2003)** October 19-22, New York, USA.

[8] A. Alves, C. Viegas and P. Nejdl, “A Distributed Tabling Algorithm for Rule Based Policy Systems”, IEEE Computer Society, vol. 15, no. 5, **(2004).**

[9] C. Thomas, "The Basics of Reliable Distributed Storage Networks", IEEE Computer Society, vol. 6, no. 3, **(2004).**
[10] S. Ahmad, B. Ahmad, S. M. Saqib and R. M. Khattak, "Trust Model: Cloud's Provider and Cloud's User", International Journal of Advanced Science and Technology, vol. 44, **(2012).**
[11] S. Gurumurthi, "Architecting Storage for the Cloud Computing Era", IEEE Computer Society, vol. 34, no. 6, **(2009).**
[12] P. Buxmann, T. Hess and S. Lehmann, "Software as a Service", Business and Economics, vol. 50, no. 6, **(2008).**

Authors

Feng Liu, he received his master's degree in Measuring and Testing Technologies and Instruments from Graduate School of Chinese Academy of Sciences, China, in 2008. Now he is studying the PhD of Astrometry and Celestial Mechanics in University of Chinese Academy of Sciences, China. His current research interests include the applications of cloud computing and parallel computing in GNSS.

Haitao Wu, he received his PhD in Astrometry and Celestial Mechanics from Graduate School of Chinese Academy of Sciences, China, in 2002. Now he is a research professor in National Time Service Center (NTSC), Chinese Academy of Sciences. His current research interests focus on overall technical and applications of satellite navigation system.

Xiaochun Lu, she received his PhD in Astrometry and Celestial Mechanics from Graduate School of Chinese Academy of Sciences, China, in 2004. Now she is a research professor in National Time Service Center (NTSC), Chinese Academy of Sciences. Her current research interests focus on satellite navigation signal design and assessment.

Xiyang Liu, he received his PhD in Circuits and Systems from Xidian University, China, in 2007. Now he is a professor in School of Software of Xidian University. His current research interests focus on distributed computing and software testing.

