IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 7, 2018, accepted July 22, 2018, date of publication August 13, 2018, date of current version September 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2862468

Date Hierarchical Storage Strategy for Data Disaster Recovery

TIAN JUN-FENG, ZHANG JIA-YAO, AND DU RUI-ZHONG
College of Computer Science and Technology, Hebei University, Baoding 071000, China
Key Lab on High Trusted Information System in Hebei Province, Baoding 071000, China

Corresponding author: Zhang Jiayao (2218624065@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61170254, Grant 60873203, and Grant 61379116, and in part by the Hebei Natural Science Foundation under Grant F2016201244.

ABSTRACT Cloud storage has become a widely accepted data storage model in recent years. With the development of cloud storage applications in various fields, cloud storage security issues have aroused people's special attention. In this paper, taking aim at ensuring data reliability and satisfying QoS requirements for different users, a method named hierarchical storage-based data disaster recovery strategy was proposed. Cloud computing fault tolerance model is constructed by grading the storage resources of cloud service providers according to the service level to meet the different QoS requirements of the cloud users. On this basis, some storage levels are classified to improve resource utilization. The correctness of the classification strategy is analyzed theoretically, finally, the effectiveness of the proposed strategy is validated by emulation experiment. This paper has the theoretical and practical value to improve the overall experience of users.

INDEX TERMS Cloud computing, data disaster recovery, warehouse storage system, storage resources classification.

I. INTRODUCTION

As a new Internet-based service model, cloud storage has attracted more and more attention [1]. However, while providing good storage services to users, cloud storage inevitably has the security risks posed by natural disasters or man-made damage. Once a cloud service provider's storage service has a security issue, the user's stored data is at risk of data corruption, data loss, and the like [2].

In 2012, the Amazon data center in North Carolina was struck by a storm over the eastern United States, shutting down the Amazon Web Services (AWS) cloud service and affecting customers such as Netflix, Pinterest, and Instagram [3]. The unprecedented prosperity of information technology and industry also makes the events that endanger the information security constantly happening. The destruction of hostile forces, attacks by hackers, malware intrusions, etc. can all cause data damage [4]. According to IDC's report, 55% of U.S. companies that experienced a system disaster in the 1990s immediately went bankrupt because they could not continue to operate, 29% went bankrupt in two years and only about 16% can continue to operate. Research shows that if an organization shut down its data and applications for 1 hour, the company lost 150,000 to 6,450,000 U.S. dollars [5]. Therefore, disaster recovery backup storage resources is particularly important.

In recent years, it has drawn the attention of industry and academia through the special data disaster recovery technology to ensure the high availability of data [6]. Data disaster recovery refers to the establishment of one or more data backups in different places. Data redundancy and geographical dispersion are used to recover data after a disaster. Distributed storage systems such as Amazon S3, Google and other file systems default to 3-Replicas data backup mechanism [7].Cloud service providers can build their own disaster recovery center by purchasing and maintaining a large amount of hardware and software resources to ensure that different data backups can be stored in geographically isolated data centers. However, when system tasks are small, a large amount of resources are free. Wood *et al.* [8] has confirmed that the cost of renting other cloud resources is far less than the cost of establishing a data disaster recovery center. A cloud service provider rents the resources of other cloud platforms to store its own data backups in the form of pay-as-you-go according to its own need [9]. However, the QoS (eg. response bandwidth, security, reliability, etc.) requirements of different users or different tasks of the same user are different. The storage capacity and throughput capacity of the cloud service provider are limited, how to satisfy the QoS requirements of users and increase the resource utilization of the cloud service providers is a question worth studying.

In this paper, we propose a disaster recovery storage strategy—Hierarchical Storage Based Data Disaster Recovery Strategy (HSBDDRS), which takes into account both the QoS requirements of user tasks and the storage rates of cloud service providers. The main work includes: 1) Proposing a cloud storage resource ranking strategy to meet user QoS requirements; 2) Proposing a storage resource hierarchical classification strategy to improve the storage resource utilization of cloud service providers; 3) Considering the user's QoS requirements and resource utilization, a hierarchical and classified hybrid storage model is proposed; 4)Based on customer satisfaction and cloud Service provider resource utilization as the goal, the storage resource hierarchical classification strategy is analyzed and compared with several other storage strategies through simulation experiments.

II. RELATED WORK

The traditional disaster recovery service is to establish their own remote data center to ensure data reliability, but it takes a huge cost [10]. Wood *et al.* proposed a cloud service model named disaster recovery as a service, that is, providing data disaster recovery service to users through a virtual cloud platform. In addition, the data disaster recovery system RUBIS is established for the application service of the website, and the cost of the public cloud disaster recovery service is evaluated. The example shows that the disaster recovery cloud model greatly reduces the overhead. however, there is no in-depth study on the impact of cloud storage disaster recovery time and communication delay [8];Data access overhead is directly related to the storage location of the data backup. In order to reduce the data access overhead, JI *et al.* use the existing computing resources of each university to back up each other's data. On the education cloud platform Ren, a data backup strategy based on node location is designed, according to the geographical location to select the backup node [11]. He *et al.* [12] also considered the location distance of nodes, designed a file backup method based on minimizing the access overhead for multi-user data access, and calculated the location distance relationship between multiple users and data storage nodes to minimize the data access overhead. The network connection about distributed system may fail, which requires the path backup to ensure reliability. In order to reduce the overhead of path backup, Hou *et al.* [13] proposed a selective protection scheme of balancing path backup and cost, which selectively backed up some unreliable links and reduced some overhead.

The concern of Bermbach *et al.* [14] is the issue of communication delays for multiple cloud service provider data backups. They set up the access queue according to the response time of each node, and select appropriate nodes in order and back up the data so as to alleviate the communication delay problem. Wood *et al.* proposed a cloud-based backup method based on pipe synchronization. When a user submits a service request, the cloud storage system allows the remote asynchronous backup of data to be performed simultaneously with the data storage of the foreground server. The user can perform other webpage operations without waiting for response from the remote server [15]. Xiang Fei proposed a new 'clouds-based' data disaster recovery strategy-RCDDRS. Cloud providers can rent other cloud platform resources for their own data according to the pay-as-you-go method. This data redundancy method based on multiple cloud platforms is called "rich clouds" mode. Cloud providers can selectively back up data to their own disaster recovery center or storage resources of other cloud platforms according to their resource conditions and task types. This strategy minimizes data disaster recovery costs and reduces RTO, and dynamically adjusts policies according to the data storage situation at different times. However, this strategy only considers the average disaster recovery time of the entire user, ignoring that different users or different tasks of the same user have different QoS requirements, it is difficult to meet the high level of user experience [16]. Fan Guisheng's utility-based data fault tolerance strategy, using formal and game theory, builds a cloud computing fault tolerance model to effectively characterize cloud computing and fault tolerance. Based on the analysis of cloud computing fault tolerance requirements, the problem of cloud computing fault tolerance utility optimization is transformed into a multi-application game, but the QoS is not fully considered [17]. Lin Guoyong, Huang Fan first established disaster data distribution hierarchy structure model. They use multiple users QoS overhead fitness operation mechanism and particle clustering algorithm to weigh the data resources in cloud computing task allocation, reduce the operation time, improve the ability of data disaster backup, but ignores that the hierarchical structure can reduce the storage resource utilization [18].

III. HIERARCHICAL CLOUD STORAGE STRATEGY FOR DISASTER RECOVERY

Due to different hardware performance and network conditions, different data nodes of cloud service providers provide different service levels. The upper application's preference for QoS presents a personalized trend, Traditional copy selection strategies that use a single criterion can greatly affect the user experience. For this reason, we put forward a Hierarchical Storage Based Data Disaster Recovery Strategy-HSBDDRS for such applications that are sensitive to quality of service. The cloud service provider's storage resources composed of data nodes are ranked according to the service level they can provide from high to low. The storage resources for the node at the i-th position in the list is called the i-th level storage resources. When users back up data, they will choose the appropriate hierarchical storage resources for data storage according to their QoS requirements and service cost.

A. HIERARCHICAL STORAGE STRATEGY DESCRIPTION

Because the weights of different level of data on different attributes are also different, the order of the ranked list is also different. Therefore, we consider the preference of different data for each attribute, and divide the storage levels for different types of data. Xiong [19] conducted fine-grained analysis from three aspects: reliability attributes, temporal attributes, and security attributes, and constructed a three-dimensional QoS model. When the system selects a location for storage, it must not only pay attention to the user's QoS requirements, but also take into account the acceptance scope of different users for service costs. Although the storage resources with high service levels can well meet the QoS requirements of various users, their storage costs are relatively high. This paper fully considers the service levels of various grades of storage resources, user QoS requirements, and the service costs paid by users. The replica storage model is designed as shown in Figure 1.

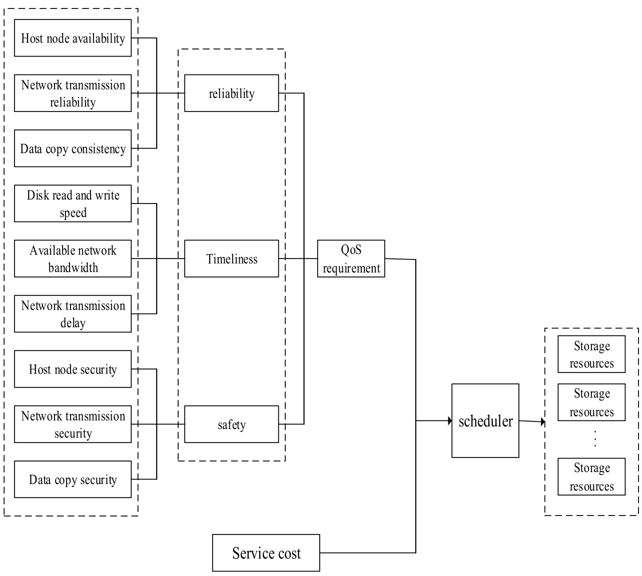

FIGURE 1. Copy storage model.

The user's satisfaction with various attributes of the storage resource is:

$$S^{(i)} = \begin{cases} 1 & if\ Us^{(i)} \geq Ur^{(i)} \\ \dfrac{Us^{(i)}}{Ur^{(i)}}, & if\ Us^{(i)} < Ur^{(i)} \end{cases} \quad (i =, 2, \ldots, 9) \tag{1}$$

Ur(i) represents a user's demand value for the i-th attribute, and Us(i) represents a service level provided by the storage resource of the storage node for the i-th attribute. A user's satisfaction degree for the i-th level for a specific user Q_i is:

$$Q_i = \sum_{i=1}^{9} k_i x S^{(i)} \quad (i = 1, 2, \ldots 9) \tag{2}$$

Consider the user-constrained service overhead, based on the hardware performance of each storage class, and then make the following evaluations:

$$\begin{aligned} &\max(Q_i) \\ &\text{s.t } cost_{user} \geq \cos t_i \end{aligned} \tag{3}$$

Among them, $cost_i$ is the unit cost of the i-th leve storage resource, and $cost_{user}$ is the maximum value of the unit cost that the user accepts. The storage level resource with the highest satisfaction level is selected for data storage within the cost constraint range.

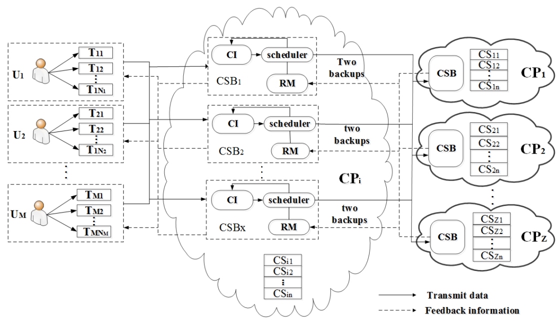

FIGURE 2. General frame of HSBDDRS.

The overall framework of HSBDDRS is shown in Figure 2. U_i represents the i-th ($i = 1, 2, \ldots, M$) cloud users. Cloud users refer to the tenants who complete the relevant work using the cloud platform, which may be individuals, enterprises or other cloud service providers, and each user has several tasks. T_{ij} denotes the j-th ($j = 1,2,\ldots, N_i$) tasks of the i-th user; CP_k ($k = 1,2,\ldots, Z$) denotes the k-th cloud service provider. CSB_i ($i = 1, 2, \ldots, x$) represents the cloud service provider's cloud service proxy. CS_{ij} represents the j-th level of storage resource of the i-th cloud service provider. Cloud Service Broker (CSB) includes Cloud Interface (CI), Resource Monitor (RM) and Scheduler. In Figure 2, it is assumed that CP_i is a cloud service provider that accepts user task requests locally. CP_j ($j = 1, 2, \ldots, Z, j \neq i$) is another cloud service provider that provides data disaster recovery services. Each cloud service provider divides storage resources from high to low according to the level of service they can provide. Upon receiving the user request, the CP_i selects a matched data backup storage location according to the QoS requirements of the user and rents different cloud service provider resources to store the data backup. In order to ensure data reliability, HSBDDRS uses the current 3-Replicas policy to back up data, that is, save 3 copies for each user task to implement disaster recovery storage for user tasks. One of the backups is stored at the local data center and the other two are backed up geographically. Users interact with the cloud platform through the Cloud Interface (CI), such as providing applications to the cloud platform and querying the current status of tasks. RM monitors the local cloud resources in real time and monitors the scalability of cloud resources of other cloud service providers and forwards the information to the Scheduler. Based on the task request information of the users and the current cloud service providers, the user choose cloud storage resources that cost relatively little under the condition of satisfying user QoS requirement, and feeds back the scheduling results to the CI.

B. WAREHOUSE STORAGE STRATEGY AND CLOUD STORAGE

Although the above strategy can satisfy each user's QoS requirement, it does not consider the cloud service providers' storage resource utilization. Because the size and storage time of each task are different, and the data access volume also dynamically changes with time, so the data storage volume of each level of storage resources fluctuates: at some times, some levels storage resources are tight, and some level storage resources are largely idle. If the access volume of a certain level storage resource is too large, its service level will also be affected. Using data migration technology can solve the above problems. When access to a certain level of storage resources becomes large and resources are scarce, migrating backup data to other levels of storage resources with similar service levels can effectively alleviate this situation, but a large amount of data migration will inevitably consume additional costs and take up a lot of bandwidth. In addition, many operations cannot be migrated. In order to reduce costs and increase resource utilization, HSBDDRS classifies hierarchical storage resources. As shown in Figure 3, the level storage resources with similar service levels are be grouped together. The k_m level storage resources are divided into m classes, and the i class has $(k_i - k_{i-1})$ level storage resources. Based on the above-mentioned hierarchical storage strategy, the system considers the user's QoS requirements and storage costs to find suitable storage classes for them.

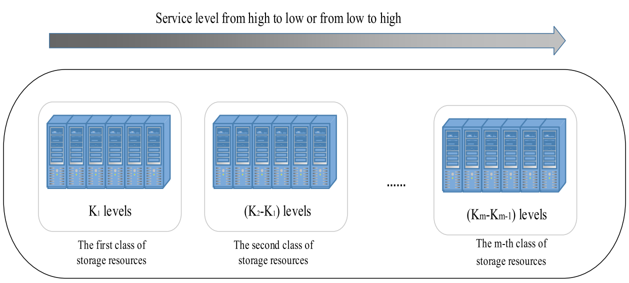

FIGURE 3. **Classification storage.**

In order to improve the utilization of storage resources, HSBDDRS applied Guo Xiaolong's limited-based warehouse storage strategy to the cloud platform to classify hierarchical storage resources [20]. Classification storage is widely used in warehouse management, which is to classify various items according to the turnover rate, and arrange them according to the turnover rate of each group of items. Items in the same class are stored randomly. Guo Xiaolong's warehouse storage model overthrows the previous theory of "the more types of items in the warehouse, the shorter the average travel time of stored items". The function of the number of items in the area is as follows:

$$A_i(N'_k) = 0.5 \times (1 + N'^{-\varepsilon}_k) \times Q(i) \tag{4}$$

Where N'_k is the number of items in class k; Q (i) is the upper limit of the storage space required for item i; $A_i(N'_k)$ is the storage space for the classified data of rank i; ε is the sharing coefficient of the storage space between different items.

From (4), we can see that as the number of types of warehouse items increases, the required storage area for item i decreases. The relationship between the average travel time and the number of categories is the bowl-shaped model. When the items are divided into a few categories, the average travel time is the shortest. The theory is also validated in reality: In warehouse operations, most warehouses are classified into smaller classes (categories 3 to 5 are more common).

This section will improve the warehouse storage classification strategy and apply it to the backup storage strategy of cloud service provider resources to analyze and verify whether it is applicable to the cloud platform.

Related symbols and their definitions:

N: the number of storage levels;

Q (i): The maximum amount of data stored at the i-th level over a period of time;

SQ(i): The maximum amount of storage of class i over a period of time;

q_i: service level provided by i-th level storage resource;

QoS_i: QoS requirement of user i;

A_i (N'_k) : The storage space required for the i-th level of storage resources after classification;

h: service level difference between adjacent level storage resources;

C: The maximum difference between the lowest level and the highest level of storage resources in each storage class, recorded as the maximum level difference of the level resource in the storage class;

N_i: The number of levels of the first i classes of storage resources;

N'_i : The number of levels of the i-th class of storage resources;

S_{ij}: Satisfaction of the i-th user task stored in the j-th storage level resource.

Warehouse storage strategies classify items into categories. In each category, items are stored randomly. If on a cloud platform, the level storage resources in each class are randomly stored, the load will be unbalanced and the resource utilization will be reduced. In order to ensure the utilization of storage resources, in each category, the data is stored according to Zheng's comprehensive load reference comparison method [21]. Taking a working server (assumed to be a server M) as a benchmark, the information of other servers (assumed to be server i) and the information of the benchmark servers are weighted into the following formulas. The ratio is

$$axN_{1i}x\frac{C_i}{N_{1m}xC_m} + bxN_{2i}x\frac{M_i}{N_{2m}xM_m} + cxN_{3i}x\frac{D_i}{N_{3m}xD_m} + dxN_{1i}x\frac{Net_i}{Net_m} \tag{5}$$

i is the i-th server; m is the benchmark server; N1 is the CPU processing capacity; N2 is the memory parameter; N3 is the hard disk parameter; C is the CPU usage rate; M is the memory usage rate; D is the hard disk throughput: Net is the network traffic; a is the CPU comparison weight; b is the memory comparison weight; c is the hard disk comparison weight; d is the network comparison weight. The initial value of a, b, c, d is set to 1. Depending on the actual situation of the data center, the system can increase or decrease one of the weights to emphasize or weaken the load performance of a certain area. After getting the information of all the running servers and the benchmark server in the server information table, the scheduler chooses the server with the smallest ratio as the server with the lightest load, and the users who enter the cluster will access the server with the lightest load.

It is assumed that the storage space of each level storage resources is equal, and the average data storage amount and the maximum data storage amount in each level storage resources are randomly generated. The access frequency of each level storage resources is Poisson distribution. After being classified, the storage space required by the i-th level storage resource is:

$$A_i(N_k') = \frac{SQ(k)}{\sum_{j=(N_k-N_{(k-1)}+1)}^{N_k} Q(j)} xQ(i) \tag{6}$$

After a large number of data access experiments, the relationship between the storage space required for the i-th level storage resource and the number of levels in this class is as follows:

$$A_i(N_k') = 0.5 \times (1 + N_k'^{-\varepsilon}) \times Q(i) \tag{7}$$

In Eq. (7), ε is the sharing coefficient among the resource data of each level in each class, and the size of ε is related to the average data storage capacity in each level of storage resource, the arrival time of the task, the task update, etc., The average amount of data stored in the storage resource has a significant impact. Under normal circumstances, the greater the amount of data storage, the smaller the value of ε.

Below we take a special case to give the theoretical derivation of equation (7). It is assumed that the average amount of data stored in each level of storage resources is Q (i) / 2 and varies periodically from 0-Q (i). Further coordinate the arrival times of all the stored data so that the backup data of the first level of storage resources arrives at the time T + 0, the backup data of the second level of storage resources arrives at the time T+T/N_k', the backup data of the i-th level of storage resources arrives at the time T + T/N_k' ∗ (i-1). In this case, get the total storage needed

$$\sum_{i=1}^{N_k'} i \times Q(i)/N_k' = (1 + N_k) \times Q(i)/2 \tag{8}$$

The required storage space for level i is

$$A_i(N_k') = 0.5 \times (1 + N_k'^{-1}) \times Q(i) \tag{9}$$

In this case, the storage coefficient between all levels of storage resources is 1. However, the amount of data, the data arrival time, and the data storage time of each storage resource are different, the value of ε will be far less than 1. After a large number of storage experiments, the value of ε is between 0.10-0.25.

In order to analyze the classified resource utilization, HSBDDRS proposed the relative storage rate of cloud service providers, denoted as R, and the relative storage rate R_i in each category is defined as follows:

$$R_i = \frac{Q(i)}{A_i(N_k')} \tag{10}$$

This section considers both user satisfaction and storage rates, and analyzes whether it is possible to find a storage resource classification scheme that makes these two aspects relatively optimal.

The most important part of the QoS requirements for replica storage is the response bandwidth, so this section uses the response bandwidth as an evaluation indicator for a simple analysis. For the sake of calculation, the following analysis of this section is restricted as follows:

(1) the storage space of each level storage resource is equal;
(2) the number of levels in each class is equal;
(3) the average amount of data stored in each level storage resources is substantially the same in a period of time.

Assume that the data nodes of the cloud service provider's storage resource are divided into N levels according to the response bandwidth. From the first level to the N-th level, the response bandwidth of data nodes is getting higher and higher. Assume that the bandwidth of the first-level storage resource is unit 1, the bandwidth of the i-th level storage resource is $1 + (i-1)*$h, and h is recorded as a service level difference. With the increase of response bandwidth, the unit cost of data storage increases. These storage resources are divided into m classes according to the storage level, that is, $N_i' = N/m$. Then, the user's satisfaction and the relative storage rate of the cloud service provider are calculated. bw_j and BW_i are the response bandwidth of the j-th level storage resource and the user i's required bandwidth, respectively. The expression of user satisfaction S_{ij} is as follows:

$$S_{ij} = \begin{cases} 1, & if\ bw_j \geq BW_i \\ \frac{bw_j}{BW_i}, & if\ bw_j < BW_i \end{cases} \tag{11}$$

For ease of presentation, let $N_i' = n$. The relationship between rank number n in the first class and average user satisfaction S′ is as follows:

$$S' = \frac{1}{4}(3 - \frac{1}{n} + \frac{1}{n^2}) + \frac{1}{2}\sum_{i=2}^{n} \frac{1}{1 + (i\text{-}1) \times h} \tag{12}$$

Taking h as 5%, 10%, 20% and 30%, we get the relationship between the number n of storage levels in the first class and the average customer satisfaction S′, as shown in Figure 4:

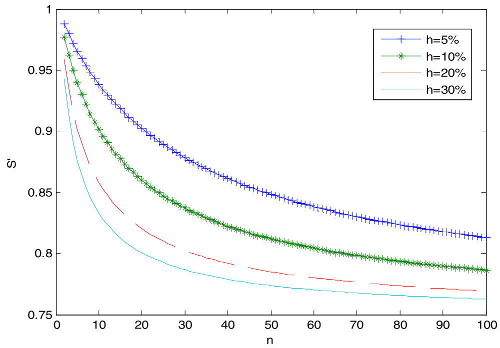

FIGURE 4. The relationship between n and S.

As can be seen from Figure 4, the average user satisfaction in each class is related not only to the number of storage levels but also to adjacent storage service level difference. Through the analysis of the relationship curve, we can see that within a certain range, the average user satisfaction of the public storage area is basically close when the maximum level difference C ($C = h * N'_k$) in the storage class is a fixed value.

Taking the value of as 0.10,0.2,0.25, respectively, get the relationship curve between the level number n and the relative storage rate R in each category, as shown in the following figure:

It is concluded that in the first class, when the user's satisfaction requirement is greater than 0.9, the maximum level difference C in the storage class ranges from 0.9 to 1, and the relative storage rate is between 1.03 and 1.36. The storage rate is Better. When the average user satisfaction S is greater than 0.95, the maximum level difference C in the storage class ranges from 0.3 to 0.4, and the relative storage rate ranges from 1 to 1.25, and the storage rate is good. However, the average user satisfaction S less than 0.98, it cannot satisfy the users with high satisfaction requirements.

C. HIERARCHICAL CLASSIFICATION HYBRID STORAGE STRATEGY

HSBDDRS improves the above classified storage strategy and proposes a hierarchical and classified hybrid storage strategy to solve the problem of low user satisfaction caused by the classification of storage resources.

Related symbols and meanings:

A (i): storage space for the i-th level dedicated storage area;

B (i): storage space of the i-th level public storage area;

S: The minimum average value of user satisfaction required by the data stored in the public storage area in each storage class;

S′: The total average user satisfaction,

P: The proportion of dedicated storage area;

R′: total relative storage rate;

OH: unit overhead, referring to the storage cost of task backup in unit time and unit storage space;

H: unit cost difference of adjacent storage level resources;

MDS (i): Maximum data storage for i-th storage level;

MS: The minimum data storage rate of storage resources for each storage level.

Pt (i): The amount of data storage for the i-th storage level, which is Poisson distribution.

Hierarchical classification Hybrid storage idea is that each level of storage resources is divided into a part of storage resources called public storage resources, and the remaining storage resources are used as dedicated storage resources. Data is stored preferentially in dedicated storage areas, as shown in Figure 9.

The backup data storage process is as follows:

(1) According to the user's QoS requirements and their storage overhead constraints, the system will preferentially search for a suitable storage level resource in the dedicated storage area according to the hierarchical storage strategy.
(2) When the storage space of the corresponding level of storage resources is insufficient, the system will give up the storage of data in the dedicated storage area of this level. In the public storage area, the system finds suitable storage classes according to the hierarchical storage strategy.
(3) After the storage class is selected in the public storage area, in order to ensure load balancing and improve resource utilization, the system selects the lightest load class for data storage in the storage class of the public storage area according to the comprehensive load reference comparison method.

In order to guarantee the QoS requirement of the user task, it is necessary to limit the maximum level difference C of the level resources in each storage class. The larger the C, the less secure the data stored in the public storage area. When $Q(i) > A(i)$, through a large number of storage simulation experiments, the functions of the storage space required by the i-level storage resource and the number of storage levels for each storage class are obtained as follows:

$$A_i(N'_k) = A(i) + 0.5 \times (1 + N'^{-\varepsilon}_k) \times (Q(i)\text{-}A(i)) \tag{13}$$

The principle of HSBDDRS is that data is preferentially stored in the dedicated storage space A(i), so the data volume of each storage level resource will inevitably affect the service level. Assuming that the data storage rate is Poisson distribution, the function relationship between the total average user satisfaction S′ and the user satisfaction S in the public storage area is:

$$S' = \begin{cases} \dfrac{A(i) + (P_t(i) - A(i))xS}{P_t(i)}, & if\ Pt(i) \geq A(i) \\ 1, & if\ Pt(i) < A(i) \end{cases} \tag{14}$$

Where

$$p = \frac{A(i)}{A(i) + B(i)} \tag{15}$$

TABLE 1. Attribute weights for different types of data.

different types of data	Response time weight	error rate weight
time-based data	0.7	0.3
reliability-based data	0.3	0.7

The relative storage rate of data is:

$$R=\frac{\sum_{i=1}^{N} Q(i)}{\sum_{i=1}^{N} A_i(N_k')} \tag{16}$$

The larger P is, the more data is stored in the dedicated storage area, and the higher the user satisfaction, the lower the storage rate. In the simulation experiment, we will find the P value so that the user satisfaction and relative storage rate are good.

IV. EXPERIMENTAL DESIGN AND RESULTS ANALYSIS

A. EXPERIMENT DESCRIPTION

This article implements simulation under CloudAnalyst, a cloud computing simulation analyzer. The basic configuration of the system running platform is as follows: Intel(R) Core(TM) i5-4570 CPU @ 3.2GHz, 8G memory. Because the security dimension is difficult to analyze and belongs to a higher level of attribute requirements, the experiment in this paper only considers reliability and timeliness.

On the simulation platform, each storage node is generated by a virtual machine. The service level of each storage node is evaluated by CPU speed, CPU utilization, available bandwidth, and bit error rate. Configure detailed configuration in the data center, including CPU speed, network bandwidth, and bit error rate. The configuration information in the user group includes: the number of requests, request size, peak start and end times, the number of peak users, and the number of users in general.

The processor speed, CPU utilization, and available bandwidth determine the response time, so the service level is evaluated in terms of response time and bit error rate. The system divides the data into two types from reliability and time. The weights of these two types of data for each attribute are as follows:

Different data have different acceptance ranges for maximum response time and error rate. The system sets response time and error rate acceptance ranges for different types of users. Within this range, this type of user satisfaction is 1.

B. STORAGE RATE EXPERIMENT

The experiment analyzed several cases where the minimum storage rate of the data was 0.3 and 0.5 respectively. In the initial case, the CPU speed, network bandwidth, and bit error rate of the first level storage resource are set to 100 GHz, 1000 M/s, and 3%, respectively. Set the number of storage levels N to be 100. The CPU speed, network bandwidth, and bit error rate of the i-th level storage resource are $(100-(i-1)xh)$GHz, $(1000-(i-1)xh)$M/s, and $(3+0.5hx(i\text{-}1))\%$ respectively. Where h is equal to 0.2. The minimum response delay and the maximum response delay are the range of response delays required for different types of data. Assuming that there are 100 kinds of data, the response time requirements are distributed between the minimum response time and the maximum response time obtained before, and the error rate is 3%-12.9%. The 100 data accesses were randomly distributed. Among them, the first 50 are time-based data, and reliability-based data. Data access frequency is Poisson distribution. As can be seen from Figure 5, when the number of storage levels in each storage class is 20, rising trend of the relative storage rate becomes slow, so C is set to 6. After a large number of storage experiments, the curves for P and S are:

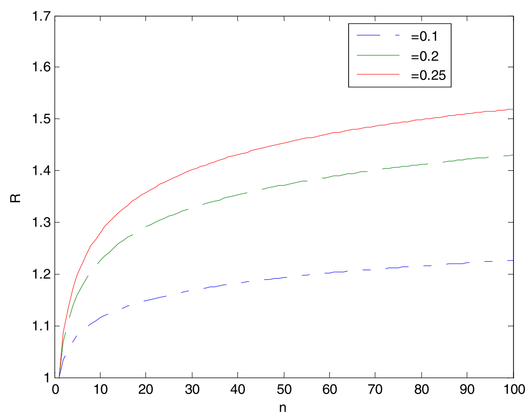

FIGURE 5. The relationship between n and R.

From Figure 7 and Figure 8, it can be seen that when the total average user satisfaction is constant, as the occupation ratio P of the dedicated storage area increases, S decreases. Because as P increases, the proportion of data in the dedicated storage area increases, the amount of data stored in the public storage area decreases, and the effect of the public storage area on user satisfaction will become smaller. Therefore, the public storage area requires the lowest user satisfaction.

When S' is 0.98, 0.95, and 0.90, we discuss how to store data so that the total relative storage rate is maximized.

As shown in Figure 5 and Figure 6, when the total average user satisfaction S' is low, for example, less than 0.90, as long as the dedicated storage area ratio P is greater than the minimum storage rate MS, the S value is almost less than 0.8. Because S is still greater than 0.8 when the number of levels of each storage class reaches the maximum value, when the proportion of the dedicated storage area P is greater than the minimum storage rate MS, and the number of storage levels

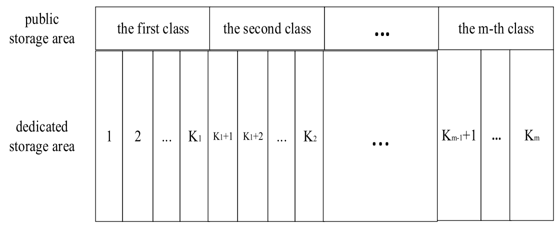

FIGURE 6. Hierarchical Classification Hybrid Storage Strategy.

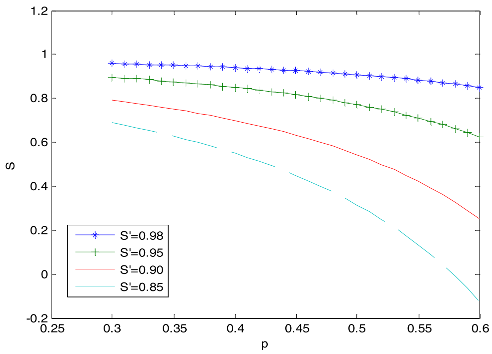

FIGURE 7. When MS = 0.3,the relationship between P and S.

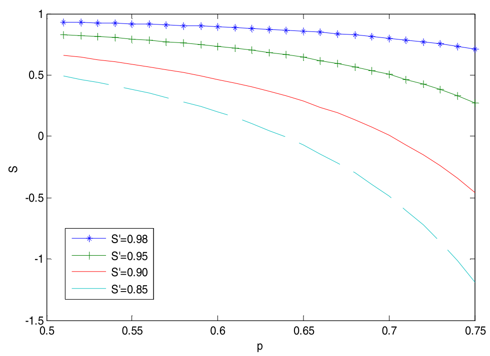

FIGURE 8. When MS = 0.5,the relationship between P and S.

in each storage class reaches the maximum value under the constraint condition, the relative storage rate is the largest.

The following experiments were performed when S' was 0.98 and 0.95. Through a large number of storage experiments, the relationship between P and N_i' is obtained when the total average user satisfaction is 0.98, 0.95, 0.90, and 0.85, respectively. The system records the maximum amount of data storage Q(i) for each level storage resource and the maximum amount of data storage BQ(j) for each class of storage in the public storage resources. Calculate total relative storage rate by formula

$$\mathrm{R}' = \frac{\sum_{i=1}^{N} Q(i)}{\sum_{j=1}^{m} BQ(j) + \sum_{i=1}^{N} [\min(Q(i), A(i)]} \tag{17}$$

When R' is 0.98, the curve of P and the total relative memory rate R' is shown in the following figure:

When 0.95, the curve of P and the total relative memory rate R' is shown in the following figure:

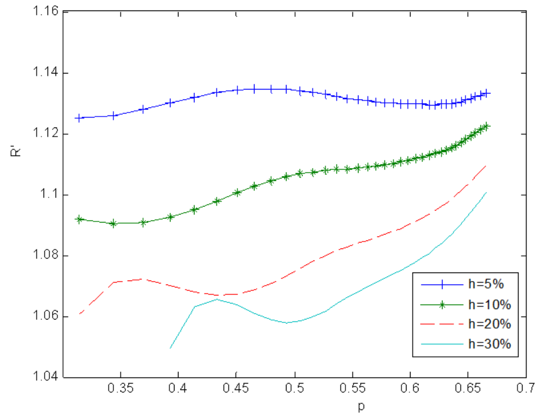

FIGURE 9. When MS = 0.3,the relationship between P and R′.

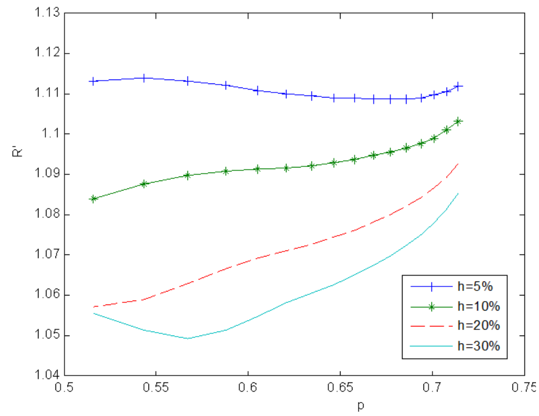

FIGURE 10. When |*rmMS* = 0.5, the relationship between P and R′.

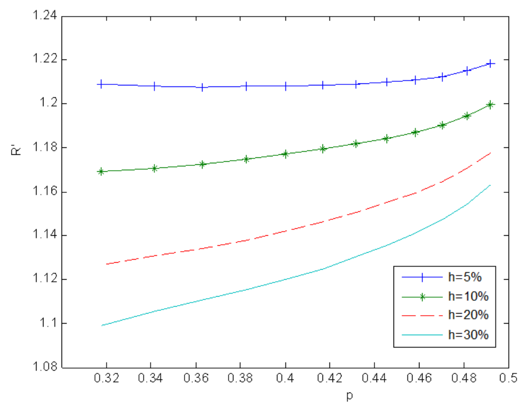

FIGURE 11. When *MS* = 0.3,the relationship between P and R′.

As can be seen from Figures 9, 10, 11 and 12, when the total average customer satisfaction rate is constant, as P increases, the dedicated storage area increases, and the number of storage levels in each class increases. The increase of the

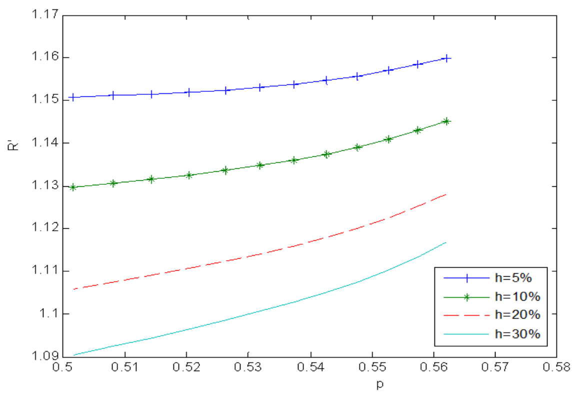

FIGURE 12. When $MS = 0.5$, the relationship between P and R$'$.

dedicated storage area will make the total relative storage rate decrease, and the increase of the storage levels in each class will increase the total relative storage rate. As P changes, these two factors affect each other so that the total relative storage rate fluctuates. As P increases, as the number of storage levels in each class reaches its maximum, the total relative storage rate begins to decline at this moment as public storage areas decrease. As shown in the figure above, when P is respectively 0.67, 0.72, 0.5 and 0.56, the number of storage levels in the public area reaches the maximum. At this time, if P increases, the relative storage rate keep falling until 1. This concludes that when the required average user satisfaction is low ($S' < 0.9$), P is only greater than the minimum storage rate. When the average user satisfaction required is normal (eg, $S' = 0.95$), P can be between 0.5 and 0.6. When the required total average customer satisfaction is high (eg, $S' = 0.98$), P can be between 0.65 and 0.7.

C. COMPARISON OF STORAGE COSTS

In the following, the storage cost is compared. Experiments have selected 100 storage nodes. These nodes can be accessed by multiple data. The storage resource unit overhead OH with the lowest service level is set to unit 1, and the unit overhead of the i-th storage resource is 1+0.1x(i-1). When the total customer satisfaction is 0.98, 0.95, and 0.90 respectively, the storage costs of FCSS (Federated Cloud Storage System) storage strategy, RCDDRS storage strategy, and HSBDDRS storage strategy are compared, as shown in Figure 13:

As can be seen from Figure 13, the data disaster recovery costs of this strategy are basically lower than those of the other two strategies. Because data are stored on the appropriate data nodes according to the user's QoS requirements for the HSBDDRS storage strategy in this paper which has a lower unit cost. When the requirements of S' are different, only P varies, and the unit cost of the public storage area is close to the average unit cost in a class. Therefore, with the change of S', the unit cost will not change too much.

D. COMPARISON OF USER EXPERIENCES

Figure 14 and Figure 15 respectively show the user satisfaction of the four strategies. The experiment selected

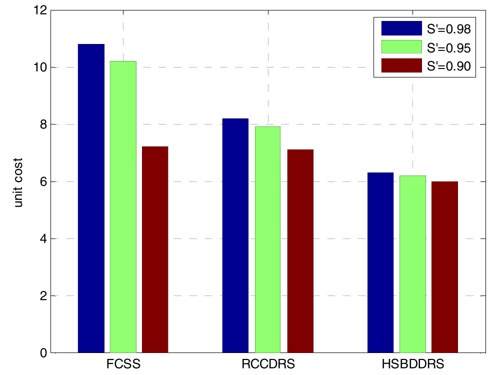

FIGURE 13. The comparison of a task unit cost.

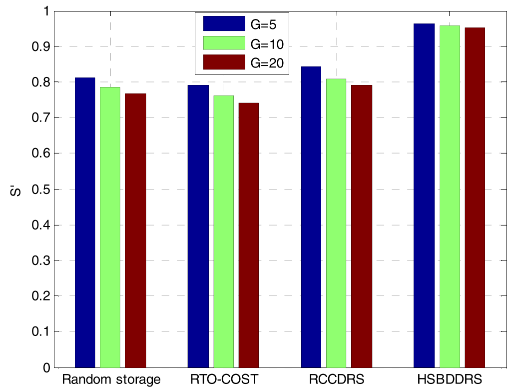

FIGURE 14. When $AS = 0.3$,user satisfaction for several strategies.

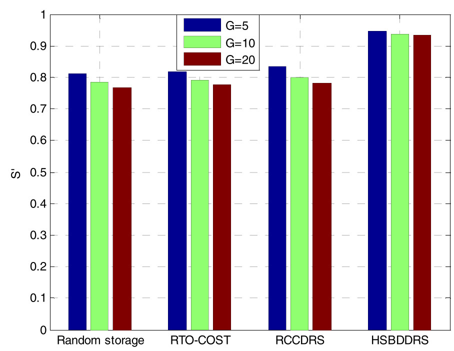

FIGURE 15. When $AS = 0.5$,user satisfaction for several strategies.

100 storage nodes. The difference in the storage capacity and service level of each data node will affect the user experience. The gaps G between the highest and lowest service levels of the data nodes are 5, 10, and 20 respectively, and the average data storage rate AS is chosen to be 0.3 and 0.5, and then user satisfaction analysis is performed. From Figure 14 and Figure 15, it can be seen that the total user satisfaction of HSBDDRS is significantly higher than the three strategies, which are close to 95%. Because data is preferentially stored in a dedicated storage area to ensure user QoS requirements, only when the storage space of the dedicated storage area is insufficient, some tasks are stored in the public storage

area, so the user satisfaction is high. From the comparison of the two figures, as the amount of data increases, the total user satisfaction of HSBDDRS decreases, because with the increase of data, the proportion of public storage area data increases, which will reduce user satisfaction.

E. LOAD BALANCE COMPARISON

Classifying storage resources is to increase resource utilization. Under normal circumstances, the more balanced use of resources, the higher the utilization of resources. So we use unbalanced degrees to analyze the resource utilization under the HSBDDRS strategy. When P is different, resource utilization is also different. Assumes that p was 0.3,0.5,0.7. We compare the strategy with the comprehensive utilization product method, the comprehensive load reference comparison method, and the random storage strategy. The comprehensive utilization product method is to use the CPU, memory, network bandwidth utilization to measure the physical server and virtual machine load, the formula is as follows:

$$V = \frac{1}{(1 - CPU_{util})} x \frac{1}{(1 - Mem_{util})} x \frac{1}{(1 - Net_{util})} \tag{18}$$

Among them, $\mathrm{CPU_{util}}$, $\mathrm{Mem_{util}}$, $\mathrm{Net_{util}}$, respectively are the server's CPU, memory and network utilization. In the case of different numbers of storage nodes, the imbalance degree in several strategies is shown in Figure 16:

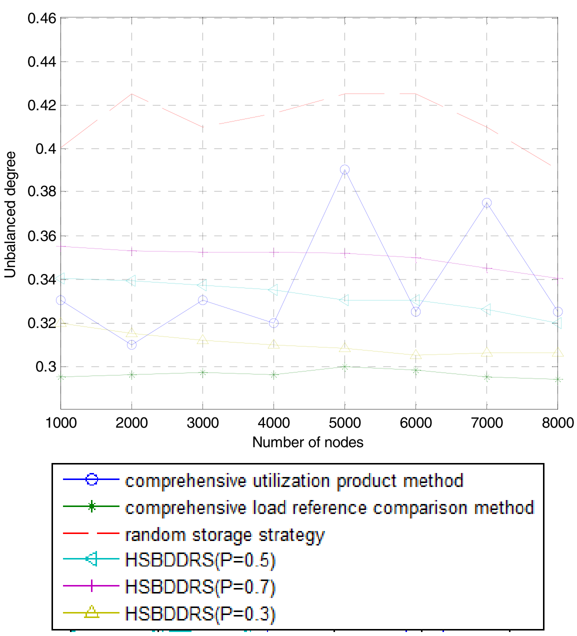

FIGURE 16. Unbalanced degree at different physical scales.

The comprehensive utilization product method only considers resource utilization, and does not consider the physical hardware's own hardware processing capability. Over time, there will be unbalanced loads across data centers. In this paper, the storage strategy uses the comprehensive load reference comparison method to store the data stored in the public storage area and selects the lightest data node for data storage.

It can be seen from the figure above that with the increase of p, the imbalance increases. Because p increases, the public storage area decreases, the overall imbalance increases. The unbalance of the HSBDDRS strategy is maintained at 0.3-0.36, which is generally lower.

In summary, the strategy of this paper can reasonably choose the level storage resources of cloud service providers for data disaster recovery according to the QoS requirements of users. Compared with the existing strategies, the strategy of this paper can better meet the QoS requirements of different users while taking into account the storage rate and resource utilization.

V. CONCLUSION

This paper proposes a hierarchical disaster recovery storage strategy. This strategy uses a new optimization method that can provide different services according to different QoS requirements of user tasks and also consider cloud services Provider's resource utilization.

REFERENCES

[1] C. Hong, M. Zhang, and D.-G. Feng, "Achieving efficient dynamic cryptographic access control in cloud storage," *J. Commun.*, vol. 32, no. 7, pp. 125–132, Jul. 2011.

[2] W. Wan, W. Suo, Y. Chen, and T. Wang, "A data distribution strategy based on the X-RDP array code," *J. Commun.*, vol. 34, no. 1, pp. 67–75, 2013.

[3] Xiamen Lingzhong Technology. (Jul. 2012). *Amazon Cloud Service has Brought Disaster to Instagram and Other Lingzhong Editions Due to Lightning Strikes*. [Online]. Available: http://www.xmlzkj.com/info.asp?id=1189

[4] H. Zhang and W. B. Han, "Overview of network security," *Sci. China, Inf. Sci.*, pp. 125–164, Jan. 2016.

[5] R. M. Zhong, "Rich cloud: Research on implementation mechanism for mutual backup based reliable cloud storage system across heterogeneous cloud platforms," Beijing Univ. Posts Telecommun., Beijing, China, Tech. Rep., 2014.

[6] *A Disaster Recovery very Solution Selection Methodology*, IBM Corp., New York, NY, USA, Feb. 2004.

[7] W. H. Li, Y. Yang, and D. Yuan, "A novel cost-effective dynamic data replication strategy for reliability in cloud data centres," in *Proc. IEEE 9th Int. Conf. Dependable, Auto. Secure Comput.*, Sydney, NSW, Australia, Aug. 2011, pp. 496–502.

[8] T. Wood, E. Cecchet, K. K. Ramakrishnan, P. J. Shenoy, J. E. van der Merwe, and A. Venkataramani, "Disaster recovery as a cloud service: Economic benefits & deployment challenges," in *Proc. 2nd USENIX Conf. Hot Topics Cloud Comput.*, 2010, pp. 8–15.

[9] H. Wang, Q. Jing, R. Chen, B. He, Z. Qian, and L. Zhou, "Distributed systems meet economics: Pricing in the cloud," in *Proc. 2nd USENIX Conf. Hot Topics Cloud Comput.*, Boston, MA, USA, 2010, pp. 1–7.

[10] R. W. Watson, "High performance storage system scalability: Architecture, implementation and experience," in *Proc. 22nd IEEE/13th NASA Goddard Conf. Mass Storage Syst. Technol.*, 2010, pp. 145–159.

[11] J. I. Khan and O. Y. Tahboub, "Peer-to-peer enterprise data backup over a ren cloud," in *Proc. 8th Int. Conf. Inf. Technol., New Gener.*, Apr. 2011, pp. 959–964.

[12] D. He, Y. Liang, and Z. Hang, "Replicate distribution method of minimum cost in cloud storage for Internet of things," in *Proc. Int. Conf. Comput. Inf. Secur.*, May 2011, pp. 89–92.

[13] M. J. Hou, D. Wang, M. Xu, and J. Yang, "Selective protection: A cost-efficient backup scheme for link state routing," in *Proc. 29th IEEE Int. Conf. Distrib. Comput. Syst.*, Jun. 2009, pp. 68–75.

[14] D. Bermbach, M. Illems, and S. Tai, "MetaStorage: A federated cloud storage system to manage consistency-latency tradeoffs," in *Proc. IEEE 4th Int. Conf. Cloud Comput.*, Melbourne, VIC, Australia, Jul. 2011, pp. 452–459.

[15] T. Wood, H. A. Lagar-Cavilla, and K. K. Ramakrishnan, "PipeCloud: Using causality to overcome speed-of-light delays in cloud-based disaster recovery," in *Proc. 2nd ACM Symp. Cloud Comput.*, 2011, p. 17.
[16] F. Xiang, C. Y. Liu, B. X. Fang, C. L. Wang, and R. M. Zhong, "Novel 'rich cloud' based data disaster recovery strategy," *J. Commun.*, vol. 6, pp. 92–101, Jun. 2013.
[17] G. Fan, H. Yu, L. Chen, and D. Liu, "Utility-based fault tolerant strategy and model for cloud computing," *Sci. China Inf. Sci.*, vol. 41, pp. 158–176, Jan. 2014.
[18] G. Y. Lin and F. Huang, "An improved data disaster allocation algorithm for cloud computing," *Sci. Technol. Eng.*, pp. 260–264, Jan. 2017.
[19] R. Q. Xiong, "Research on replica placement and selection strategies in heterogeneous cluster storage system for big data," Southeast Univ., Dhaka, Bangladesh, Tech. Rep., 2015.
[20] Y. G. Yu, B. M. R. de Koster, and X. L. Guo, "Class-based storage with a finite number of items: Using more classes is not always better," *Prod. Oper. Manage. Soc.*, vol. 24, no. 8, pp. 1235–1247, Aug. 2015.
[21] H. Zheng, L. Zhou, and J. Wu, "Design and implementation of load balancing in Web server cluster system," *J. Nanjing Univ. Aeronaut. Astronaut.*, vol. 38, no. 3, 2006.

TIAN JUN-FENG was born in 1965. He received the Ph.D. degree from the China University of Technology. He is currently a Professor with Hebei University. His research interests include information security and trusted computing.

ZHANG JIA-YAO was born in 1991. He received the bachelor's degree from Hebei University, where he is currently pursuing the degree. His research interests include information security and trusted computing.

DU RUI-ZHONG was born in 1975. He received the Ph.D. degree. He is currently a Professor. His main research interests include network security, trusted computing, and network technology.