ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 133

2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)

An Improved Scheme of Privacy Preserving Based on Lagrange Interpolation in Cloud Storage

Yu Jin[1,2] and Yadan Wang[1,2]

[1]College of Computer Science and Technology, Wuhan University of Science and Technology

[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan, China

[*] Corresponding author

***Abstract*—The increasing maturity of cloud computing has led many companies to store critical information in the cloud. The privacy preserving in the cloud storage still remains major concern because the management of the data might not be fully trustworthy. Privacy preserving brings in concern for data confidentiality. In this paper, the work pays more attention to the issues involved in the data confidentiality and service availability. It proposes an improved scheme of privacy preserving based on Lagrange interpolation which uses multi-clouds instead of single cloud service provider. Compared with its previous data hiding scheme based on Lagrange interpolation algorithm and Multi-clouds, this improved scheme only uses Lagrange interpolation to ensure both data confidentiality and service availability . The experiment was done through deploying this system in four clouds. It shows that the improved scheme outperforms the existing multi-clouds researches in terms of the cost of storage space and performance in the situation that data confidentiality and service availability are ensured.**

Keywords-multi-clouds; cloud storage; privacy preserving; data confidentiality; Lagrange interpolation

I. INTRODUCTION

Current trends show that the increasing maturity of cloud computing technology will lead many organizations to migrate their IT infrastructure to cloud service providers [1]. Even governments and companies maintain critical data in the cloud. The fast access to applications or the decreasing of the infrastructure costs are provided by cloud computing [2].

Although the cloud storage service brings many benefits, the privacy preserving in cloud storage services is considered to be one of the critical issues due to the valuable stored information for users in the cloud. Cloud service providers should address privacy preserving issues as a matter of high and urgent priority [3]. This paper pays more attention to the issues related to the data confidentiality and service availability. It proposes an improved scheme of privacy preserving based on Lagrange interpolation which uses multi-clouds instead of single cloud service provider, such as in Amazon cloud service [4]. It is found that the research into the use of multi-clouds providers to maintain security has received less attention from the research community than has the use of single clouds [5]. So the researches about multi-clouds are relatively less.

The previous data hiding scheme based on Lagrange interpolation algorithm and Multi-clouds [6] uses Reed Solomon coding to ensure service availability. On the other hand data confidentiality is ensured through Lagrange interpolation. In this paper, the proposed improved scheme only uses Lagrange interpolation to ensure data confidentiality and service availability simultaneously. So the response times of data uploaded and downloaded will have a more significant improvement than the previous data hiding scheme.

The performance of the previous data hiding scheme is superior to the DEPSKY-CA [6]. So, it can be concluded that the improved scheme is superior to the previous data hiding scheme and DEPSKY-CA [8] in term of the performance in the situation that data confidentiality and service availability are ensured. And compared with MCDB [9], the previous data hiding scheme and the improved scheme cost the less storage spaces.

The remainder of this paper is organized as follows. Section 2 reviews the previous data hiding scheme, then the improved scheme of privacy preserving based on Lagrange interpolation is proposed. Section 3 describes the realization of the improved scheme, with a thorough data flow explanation. Section 4 analyses and evaluates the improved scheme by experiment and theory. Section 5 concludes the paper with the suggestion of future work.

II. RELATED WORK

The security researches in single cloud service provider have some limitations, such as the failure of service availability, the malicious insiders in single cloud and vender lock-in issue. So the term "multi-clouds" was introduced by Vukolic[7]. Existing researches have focused on the multi-clouds environment [8~12] which control several clouds and avoid dependency on any one individual cloud.

A. Overview of the Lagrange Interpolation Algorithm

In the previous data hiding scheme [6], the basic principle of the Lagrange interpolation algorithm has been introduced. It can be used to hide the data value.

The k data points $(x_1,d_1),(x_2,d_2)\cdots(x_k,d_k)$ can construct the interpolation polynomial $L(x)$ with degree k-1. Other n (n>=k) independent variables $\alpha_1,\alpha_2,\alpha_3...\alpha_n$ are substituted into the polynomial $L(x)$. So the $d_1,d_2....d_k$ will be hidden into $L(\alpha_1),L(\alpha_2)....L(\alpha_n)$ .If arbitrary k data points $(\alpha_k,L(\alpha_k))$ and $x_1,x_2...x_k$ are obtained, the original

$d_1, d_2 ... d_k$ can be recovered. Formula one describes the process of constructing the polynomial $L(x)$.

$$\begin{cases} d_1 = ax_1^{k-1} + bx_1^{k-2} + \cdots v \\ d_2 = ax_2^{k-1} + bx_2^{k-2} + \cdots v \\ \vdots \\ d_k = ax_k^{k-1} + bx_k^{k-2} + \cdots v \end{cases} \rightarrow L(x) \quad (1)$$

B. The Proposing of The Improved Scheme

In the previous data hiding scheme, the limitations of existing multi-clouds researches have been discussed. DEPSKY-CA has a relatively low performance. MCDB consumes more storage spaces.

The previous data hiding scheme [6] uses Reed Solomon coding to encode the data blocks to generate a redundant block. Then Lagrange interpolation is used to hide each data block.

The basic principle is shown in Fig I, suppose that the size of D is 9 bytes, where $D = \{d_1, d_2, d_3 \cdots d_9\}$.

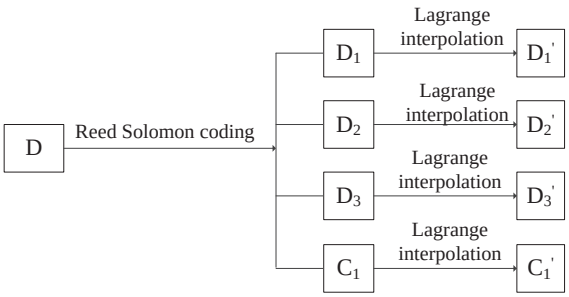

FIGURE I. THE PRINCIPLE OF THE PREVIOUS DATA HIDING SCHEME

First the D is divided into three data blocks, where $D_1 = \{d_1, d_2, d_3\}$ $D_2 = \{d_4, d_5, d_6\}$ $D_3 = \{d_7, d_8, d_9\}$. After Reed Solomon coding the redundant block $C_1 = \{d_{10}, d_{11}, d_{12}\}$ is generated. One of the four data blocks can fail.

Then Lagrange interpolation hides each data block. For each data block, the three bytes d_i, d_{i+1}, d_{i+2} and three x values x_1, x_2, x_3 are used to construct the interpolation polynomial with degree 2. Another three independent variables $\alpha_1, \alpha_2, \alpha_3$ that are different from x_1, x_2, x_3 are put into the polynomial, d_i, d_{i+1}, d_{i+2} will be hidden by $\beta_i, \beta_{i+1}, \beta_{i+2}$.

So the previous data hiding scheme uses Reed Solomon coding to ensure service availability, and uses Lagrange interpolation to ensure data confidentiality.

While as mentioned in section 2.1, If $n>k$, the Lagrange interpolation algorithm also can generate the redundancy. So the previous scheme can be improved. Fig II presents the basic idea to improve the previous scheme. For the data D, each three bytes d_i, d_{i+1}, d_{i+2} and three x values x_1, x_2, x_3 are used to construct the interpolation polynomial with degree 2. Then another four independent variables $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are put into the polynomial. Thus, d_1, d_2, d_3 will be hidden by $\beta_{11}, \beta_{21}, \beta_{31}, \beta_{41}$, d_4, d_5, d_6 will be hidden by $\beta_{12}, \beta_{22}, \beta_{32}, \beta_{42}$, and d_7, d_8, d_9 will be hidden by $\beta_{13}, \beta_{23}, \beta_{33}, \beta_{43}$. After Lagrange interpolation, the original D will be hidden into four data blocks, where $D_1 = \{\beta_{11}, \beta_{12}, \beta_{13}\}$ $D_2 = \{\beta_{21}, \beta_{22}, \beta_{23}\}$ $D_3 = \{\beta_{31}, \beta_{32}, \beta_{33}\}$ $D_4 = \{\beta_{41}, \beta_{42}, \beta_{43}\}$. One of data blocks can fail; the original data can be recovered by three data blocks.

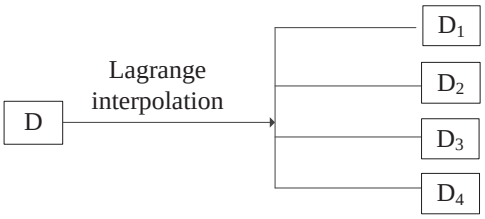

FIGURE II. THE IDEA OF AN IMPROVED SCHEME

Based on the above discussion, it can be observed that Lagrange interpolation can ensure data confidentiality and service availability at the same time. So an improved scheme of privacy preserving based on Lagrange interpolation is proposed.

And the response times of data uploaded and downloaded will have a more significant improvement than the previous data hiding scheme.

III. THE IMPROVED SCHEME OF PRIVACY PRESERVING

This section presents the architecture of the improved scheme, and then describes the specific achievement of the improved scheme.

The achievement of the improved scheme is based on two assumptions: data integrity is ensured; different cloud service providers will not collude with each other.

A. The Architecture of the Improved Scheme

Fig III presents the architecture of the improved scheme. As shown in Fig III, it deploys the system on four cloud service providers. The client performs Lagrange interpolation algorithm. The data is handled into four blocks to store in the four clouds.

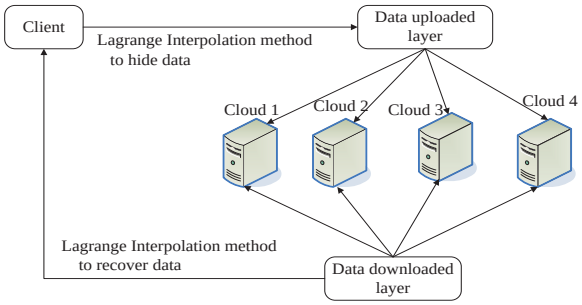

FIGURE III. ARCHITECTURE OF THE IMPROVED SCHEME

B. The Specific Process of the Improved Scheme

As the previous data hiding scheme, this improved scheme also contains two parts, data uploaded and data downloaded.

1) The Steps of Data Uploaded

For the data $D=\{d_1,d_2\cdots d_s\}$, D must be a multiple of 3. If D is not a multiple of 3, D will be filled into a multiple of 3.

• Step 1: The client uses Lagrange Interpolation algorithm to handle the D into four blocks.

• Step 2: The four blocks and related information will be uploaded into four cloud service providers.

The details of the Lagrange interpolation how to handle the data can be described in the algorithm 1. In algorithm 1, D_1,D_2,D_3,D_4 represent the data blocks that will be stored in cloud service providers. The three data points $(x_1,d_i),(x_2d_{i+1}),(x_3d_{i+2})$ are used to construct the interpolation polynomial with degree 2. The four independent variables $\alpha_1,\alpha_2,\alpha_3,\alpha_4$ will be substituted into the polynomial. Finally, the original d_i,d_{i+1},d_{i+2} will be replaced by $\beta_{1t},\beta_{2t},\beta_{3t},\beta_{4t}$.

Algorithm 1: Lagrange interpolation algorithm to hide data and generate the redundancy.

Input: $D=(d_1,d_2\cdots d_s)$, $x[3]=\{x_1,x_2,x_3\}$ and $\alpha[4]=\{\alpha_1,\alpha_2,\alpha_3,\alpha_4\}$.

Output: D_1,D_2,D_3,D_4.

1. Begin
2. *i*=1,*t*=1
3. for *i*<=*S*-2 do
4. $L(x) \longleftarrow (x_1,d_i),(x_2,d_{i+1}),(x_3,d_{i+2})$
5. $\beta_{1t},\beta_{2t},\beta_{3t},\beta_{4t} \longleftarrow L(x) \longleftarrow \alpha_1,\alpha_2,\alpha_3,\alpha_4$
6. *i*=*i*+3, *t*++
7. End for
8. End

After algorithm 1, $t=S/3$, the data D is split into D_1,D_2,D_3,D_4 where:

$$D_1=\{\beta_{11},\beta_{12}\cdots\beta_{1t}\} \quad D_2=\{\beta_{21},\beta_{22}\cdots\beta_{2t}\} \quad D_3=\{\beta_{31},\beta_{32}\cdots\beta_{3t}\}$$
$$D_4=\{\beta_{41},\beta_{42}\cdots\beta_{4t}\}$$

After that, the four parts $(\alpha_1,D_1),(\alpha_2,D_2),(\alpha_3,D_3),(\alpha_4,D_4)$ will be stored in four cloud service providers randomly, as shown in Fig IV. The original three independent variables x_1,x_2,x_3 are kept secret by the data owner.

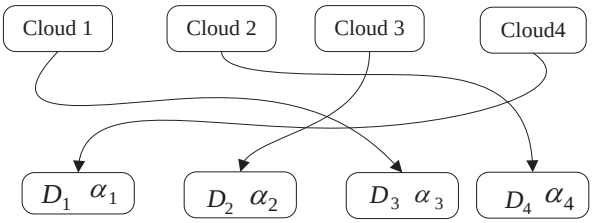

FIGURE IV. RANDOM DISTRIBUTION

2) The Steps of Data Downloaded

Due to the Lagrange interpolation generates a redundancy. So user can download data blocks from three cloud service providers. Suppose that the user downloads data information from cloud 1, cloud 3 and cloud 4.

• Step1: D_1,D_2,D_3 and $\alpha_1,\alpha_2,\alpha_3$ are downloaded from three cloud service providers.

• Step 2: The client gets x_1,x_2,x_3 from the user, and then adopts Lagrange Interpolation to handle the downloaded data blocks.

The following algorithm 2 describes the details of Lagrange Interpolation algorithm to handle data blocks.

Algorithm 2: Lagrange interpolation algorithm to recover the original data

Input: D_1,D_2,D_3, $\alpha_1,\alpha_2,\alpha_3$ and x_1,x_2,x_3.

Output: The original data *D*.

1 Begin
2. *i*=1,*m*=1
3. for *i*<=*t* do
4. $L(x) \longleftarrow (\alpha_1,\beta_{1i}),(\alpha_2,\beta_{2i}),(\alpha_3,\beta_{3i})$
5. $d_m,d_{m+1},d_{m+2} \longleftarrow L(x) \longleftarrow x_1,x_2,x_3$
6. *m*=*m*+3, *i*++
7. End for
8. End

After algorithm 2, the data $D=\{d_1,d_2\cdots d_s\}$ is recovered.

IV. IMPLEMENTATION AND EVALUATION

A. The Implementation

The experimentation that is used to examine the improved scheme is written in Java. The multi-clouds system is deployed by four Cassandra databases. The design of the interface uses SWT.

B. Evaluation

1) Service availability & data confidentiality analysis

In the previous data hiding scheme, the service availability has been proved. Similarly, this improved scheme uses

Lagrange interpolation algorithm produces the redundancy block. So it can tolerate one cloud fails. Thus the service availability is ensured.

The previous data hiding scheme and DEPSKY-CA ensure the data confidentiality at the same level. Now the data confidentiality analysis of the improved scheme will be done in three situations.

1) The data information in one cloud is revealed

If the data information (α_1, D_1) is revealed, the malicious attacker gets no information about the original data D. So data confidentiality is ensured in this situation.

2) The data information in two clouds is revealed

For example, the data information (α_2, D_2) and (α_3, D_3) is attacked by the malicious attacker.

According to the $D_3 = \{\beta_{31}, \beta_{32} \cdots \beta_{3t}\}$, $D_2 = \{\beta_{21}, \beta_{22} \cdots \beta_{2t}\}$, α_3 and α_2, t interpolation polynomials with degree 2 can be defined in the following.

$$\begin{aligned} L_1(x) &= a_1 x^2 + b_1 x + c_1 \\ L_2(x) &= a_2 x^2 + b_2 x + c_2 \\ &\vdots \\ L_t(x) &= a_t x^2 + b_t x + c_t \end{aligned} \tag{2}$$

In (2), a_t, b_t and c_t are unknown coefficients. Two points (α_3, β_{31}) and (α_2, β_{21}) are put into $L_1(x)$, (α_3, β_{32}) and (α_2, β_{22}) are put into $L_2(x) \ldots (\alpha_3, \beta_{3t})$ and (α_2, β_{2t}) are put into $L_t(x)$. After that, *t* equation groups are given by the following.

$$\begin{aligned} &\begin{cases} \beta_{31} = a_1 \alpha_3^2 + b_1 \alpha_3 + c_1 \\ \beta_{21} = a_1 \alpha_2^2 + b_1 \alpha_2 + c_1 \end{cases} \\ &\begin{cases} \beta_{32} = a_2 \alpha_3^2 + b_2 \alpha_3 + c_2 \\ \beta_{22} = a_2 \alpha_2^2 + b_2 \alpha_2 + c_2 \end{cases} \\ &\vdots \\ &\begin{cases} \beta_{3t} = a_t \alpha_3^2 + b_t \alpha_3 + c_t \\ \beta_{2t} = a_t \alpha_2^2 + b_t \alpha_2 + c_t \end{cases} \end{aligned} \tag{3}$$

Each interpolation polynomial with degree 2 needs three equations to rebuilt, while each equation group only has two equations. So (3) cannot reconstruct the $L_1(x)$ to $L_t(x)$. Thus in this situation, data confidentiality is also ensured.

3) The data information in three clouds is revealed

Suppose that (α_1, D_1) (α_2, D_2) and (α_3, D_3) the three data blocks are obtained by the malicious attacker.

For $D_3 = \{\beta_{31}, \beta_{32} \cdots \beta_{3t}\}$, $D_2 = \{\beta_{21}, \beta_{22} \cdots \beta_{2t}\}$, $D_1 = \{\beta_{11}, \beta_{12} \cdots \beta_{1t}\}$, α_3, α_2 and α_1, t interpolation polynomials with degree 2 can be constructed. Three points (α_3, β_{31}) (α_2, β_{21}) and (α_1, β_{11}) can construct $L_1(x)$, (α_3, β_{32}) (α_2, β_{22}) and (α_1, β_{12}) can construct $L_2(x) \ldots . (\alpha_3, \beta_{3t}), (\alpha_2, \beta_{2t})$ and (α_1, β_{1t}) can construct $L_t(x)$.

Although the *t* interpolation polynomials with degree 2 are rebuilt, the three values x_1, x_2, x_3 are kept by the user. Of course the malicious attacker can try to use a large amount of computation to crack the three *x* values. So in this situation, the original data might be recovered.

As analyzed in the previous research [6], if the three data blocks are obtained by the malicious attacker the original data might be recovered in the previous data hiding scheme and DEPSKY-CA.

So it can be concluded that the previous data hiding scheme, the improved scheme and the DEPSKY-CA ensure data confidentiality at the same level.

2) Performance analysis

In the previous data hiding scheme, the cost of time between Lagrange interpolation algorithm and AES when the data with different size is handled has been compared as shown in fig V.

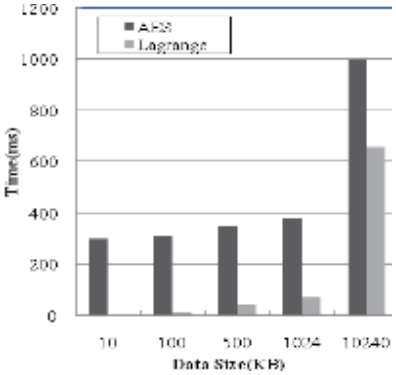

FIGURE V. TIME COMPARISON BETWEEN LAGRANGE INTERPOLATION AND AES

DEPSKY-CA uses AES for symmetric cryptography to achieve confidentiality. So it will need much time in the processes of data upload and data download.

Now the response time of the data uploaded is compared between DEPSKY-CA, the previous data hiding scheme and the improved scheme considering five data unit sizes: 10KB, 100KB, 500KB, 1MB and 10MB.

Fig. VI shows the time cost in the response time of data uploaded with data size. This shows that DEPSKY-CA system needs the most time, the improved scheme needs the least time in the procedure of data uploaded.

Then the response time of the data downloaded is compared between three schemes considering five data unit sizes: 10KB, 100KB, 500KB, 1MB and 10MB.

Fig. VII shows the time cost in the response time of data downloaded with data size. This implies that DEPSKY-CA system needs the most time, the improved scheme needs the least time in the procedure of data downloaded.

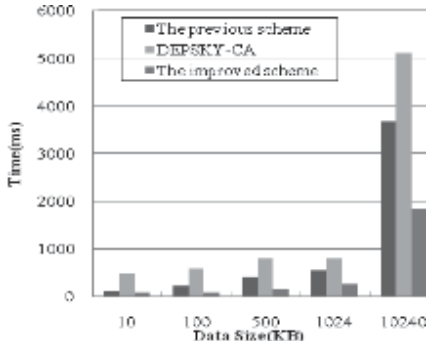

FIGURE VI. UPLOADED TIME COMPARISON, VARYING DATA SIZE

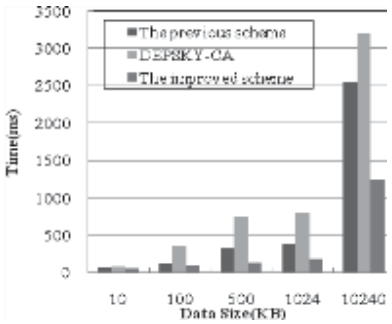

FIGURE VII. DOWNLOADED TIME COMPARISON, VARYING DATA SIZE

Through performance analysis, it can be proved that the improved scheme outperforms the DEPSKY-CA system and the previous data hiding scheme for both data uploaded and downloaded operations.

3) Storage space analysis

In MCDB, the data is handled by the secret sharing algorithm [13]. So in each cloud, the data will be stored in whole. While, the previous data hiding scheme and the improved scheme store one third data in each cloud.

Table I describes the whole cost of storage space in MCDB, the previous data hiding scheme and the improved scheme when the data size is *S* bytes.

TABLE I. THE COST OF STORAGE CAPACITY

Scheme	Storage space
The previous scheme	*S*/3+*S*
The improved scheme	*S*/3 +*S*
MCDB	*S**3

Through the above storage space analysis, the improved scheme and our previous data hiding scheme cost the same storage space. MCDB needs more storage space.

After service availability analysis, data confidentiality analysis, performance analysis and storage space analysis, It can be observed that the improved scheme outperforms the DEPSKY-CA and the previous data hiding scheme in term of the response times of data uploaded and downloaded, outperforms the MCDB system in term of the cost of storage space in the situation that data confidentiality and service availability are ensured.

V. CONCLUSION AND FUTURE WORK

On the basis of the previous data hiding scheme, this paper proposes an improved scheme of privacy preserving based on Lagrange interpolation. The improved scheme mainly improves the performance of the previous data hiding scheme by using Lagrange interpolation to ensure both service availability and data confidentiality.

In the previous data hiding scheme and the improved scheme, if three data blocks are attacked by the malicious attacker, the interpolation polynomial will be reconstructed. The original data might be recovered by large computation. So in the future, the work can continue to research on this aspect.

ACKNOWLEDGMENT

This work was supported by the National Nature Science Foundation of China under Grant No.61303117. The corresponding author is Yu Jin.

REFERENCES

[1] H. Stevens and C. Pettey. Gartner Says Cloud Computing Will Be As Influential As E-business. In Gartner Newsroom Online Ed., June 26 2008. http: //www.gartner.com/it/page.jsp?id=707508.

[2] S. Subashini and V. Kavitha, A survey on security issues in service delivery models of cloud computing, Journal of Network and Computer Applications (2011), pp. 1-11.

[3] P. BNA. Privacy & security law report, 03/09/2009. Copyright 2009 by The Bureau of National Affairs, Inc. (800-372-1033), 2009.

[4] Amazon, Amazon Web Services. Web services licensing agreement, (2010).

[5] AlZain, Mohammed A; Pardede, Eric; Soh, Ben; Thom, James A, Cloud Computing Security: From Single to Multi-clouds, Hawaii International Conference on System Sciences (HICSS 2012) 45th, IEEE, Jan. 2012, pp. 5490-5499.

[6] Yu Jin, Yadan Wang, Wei Xia, Li Deng, Heng He. A Data Hiding Scheme Based on Lagrange Interpolation Algorithm and Multi-Clouds[C]//Parallel Architectures, Algorithms and Programming (PAAP), 2015 Seventh International Symposium on. IEEE, 2015: 210-216.

[7] M. Vukolic, The Byzantine empire in the intercloud, ACM SIGACT News, 41, 2010, pp.105-111.

[8] A. Bessani, M. Correia, B. Quaresma, F. André and P. Sousa, DepSky: dependable and secure storage in a cloud-of-clouds, EuroSys'11:Proc. 6thConf. on Computer systems, 2011, pp. 31-46.

[9] AlZain M, Soh B, Pardede E. MCDB: Using Multi-clouds to Ensure Security in Cloud Computing[C]//Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on. IEEE, 2011: 784-791.

[10] H. Abu-Libdeh, L. Princehouse and H. Weatherspoon, RACS: a case for cloud storage diversity, SoCC'10:Proc. 1st ACM symposium on Cloud computing, 2010, pp. 229-240.

[11] K.D. Bowers, A. Juels and A. Oprea, HAIL: A high-availability and integrity layer for cloud storage, CCS'09: Proc. 16th ACM Conf. on Computer and communications security, 2009, pp.187-198.

[12] C. Cachin, R. Haas and M. Vukolic, Dependable storage in the Intercloud, Research Report RZ, 3783, 2010.

[13] Shamir A. How to share a secret [J]. Communications of the ACM, 1979, 22(11): 612-613.