Research on the Reliable Replica Creation Technique Based on Virtual Ring Strategy

Liu Dong, Li Kang and Cui Yunfei, *Member, IAENG*

***Abstract*—In big data times, the replica management of mass data meets great challenges with the increasing of data capacity and network complexity. The paper researched on the replica creation strategy of mass data from two aspects. Firstly, it improved the traditional consistent hash strategy by rearranging the nodes in virtual ring. The new method made two neighbouring nodes in virtual ring belong to two different racks. Then, the paper put forward a replica creation algorithm which had rack-aware ability, and simulated the processes of replica creation using the new algorithm. The presented algorithm is applicable to mass data management to improve the data reliability and availability.**

***Index Terms*—replica, consistent hash, virtual ring, rack-aware, mass data**

I. Introduction

In 2011, the Digital Universe Study, published by International Data Center (IDC), said that the global information gross would double every two years, and the gross data, which were created and copied in 2011, was 1.8ZB [1]. With the coming of Big Data times, the storage and the management of mass data have meet many new challenges, one of which is the redundant data management, or replica management. The introducing of replica is to improve data reliability and access efficiency. However, too much replica will occupy restricted storage resources and cause many additional problems such as more resources waste and data maintenance expenses. Therefore, it has become one of the key research points to explore better replica management strategies in order to optimize data access performances and improve data availability meanwhile [2][3].

At present, there have been some researches on replica management technology in data-intensive application. Megastore, Google's distributed memory system, and Spanner, the system suggested in recent years, treated replica management as a key technology, where Chubby lock service and classical Paxos algorithm were used to coordinate and deploy replica. Amazon's Dynamo adopted an improved distributed hashing algorithm to deal with data distribution and replication problem [4]. HDFS, the file systems used in Hadoop, uses a rack-aware method to improve data management performance, and the detail strategy is being developed [5]. Besides, China Southeast University's Computer Science and Engineering College pointed out that more attention should be paid to the replica management strategy corresponding to decentralized cooperated data centers [6].

This paper focuses on the replica creation strategy based on consistent hash method. It aims to present an improved strategy so as to provide both data reliability and data availability.

The second section introduces some basic principles about replica creation strategy based on consistent hash. Section 3 presents a node arranging strategy based on virtual ring. Section 4, the primary content of the paper, puts forward a replica creationg algorithm based on rack-aware strategy. The conclusion is made in section 5.

II. Replica Creation Strategy Based on Consistent Hash

Consistent hash algorithm, put forward by MIT in 1997, is one of the most popular distributed hash table protocols at present [7]. It includes three steps. Firstly, the storage space is virtualized to a ring, on which the hash values of all the storage nodes are mapped, and each node in the ring has a hash value. Then, the target data is hashed. At last, the data is stored in the node whose mapped hash value is the nearest value to the data hash result.

In consistent hash algorithm, the nodes are unevenly distributed on the virtual ring if there are not enough nodes. Besides, the algorithm neglects the performance diversity among data nodes. Dynamo improved consistent hash algorithm by introducing virtual nodes. In the improved algorithm, each virtual node belongs to a physical node, and each physical node, depending on its processing capability, can own one or more virtual nodes. All virtual nodes have equal processing capability and distribute randomly on virtual ring.

Accept the default of 3 replicas. After the first replica is stored in the corresponding node, the second replica is stored in the node following to the node which the first replica exists.

The strategy guarantees the data redundancy, and provides sufficient data availability. However, it cannot guarantee that two neighboring nodes in virtual ring belong to two different physical nodes or two racks, which brings latent reliability problem.

For example, the node *A*, *B*, *C* in Fig.1 store 3 replicas of data object *K*. If *A*, *B* and *C* belong to the same physical node,

Manuscript received January 21, 2013. This work was supported in part by the National Natural Science Foundation of China under Grant 60904082.

Liu Dong is with the Academy of Equipment, Beijing, 101416 China (phone: 86-10-66364613; e-mail: ld5m@163.com).

Li Kang is with the Academy of Equipment, Beijing, 101416 China (e-mail: likang123@yahoo.com.cn).

Cui Yunfei is with the Academy of Equipment, Beijing, 101416 China (e-mail: sducyf123@yahoo.com.cn).

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

data *K* would be unavailable for user after the physical node drops out or fails.

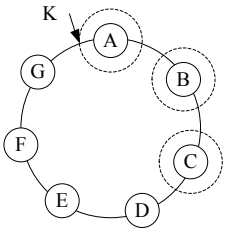

Fig. 1 The layout of replica based on consistant hash algorithm

In order to deal with the problem, we should focus on the nodes arranging in virtual ring. A better method is to make two neighboring nodes in virtual ring belong to two different racks. Then, data reliability can be improved.

III. Node Arranging Strategy Based on Virtual Ring

If two replicas of the same data object are stored in two neighboring virtual nodes, the corresponding two physical nodes should not be placed in the same rack. In order to satisfy the requirement, the virtual ring should be constructed by a new method, by which virtual nodes on the ring are arranged in a suitable way. We put out a new strategy for virtual ring, named as Virtual Ring Creation Strategy (VRCS).

In VRCS, all the physical nodes in a data center are numbered according to the following rules:

(1) All racks are numbered in sequence;

(2) Number all the physical nodes in the first rack from number one;

(3) According to the previous method, number all the physical nodes in all the racks.

Let the total number of racks be N, n be rack number, m be the physical number in a rack, and M_n be the total number of physical nodes in the nth rack.

The creation strategy of VRCS is described as follows:

(1) Virtualize a ring space for nodes;

(2) Evaluate the ID value for each physical node using the following method:

Let H be the mapping function, and $H(n,m)$ the ID of the mth physical node in the nth rack (n=1,2,…,N). $H(n,m)$ is evaluated by:

1) Set $x = 1$;

2) If the 1st node in the nth rack exist, set $x = x + 1$, and $H(n,1) = x$;

3) If the 2nd node in the nth rack exist, set $x = x + 1$, and $H(n,2) = x$;

…

M) If the Mth node in the nth rack exist, set $x = x + 1$, and $H(n,M) = x$.

It is obviously that the value range of virtual ring is [1, $\sum_1^n M_n$].

(3) Arrange all the nodes on the virtual ring according to their ID values.

Using VRCS, any two neighboring virtual nodes represent two physical nodes in two different racks.

IV. Replica Creationg Algorithm Based on Rack-Aware Strategy

VRCS is the first step for replica creation, since it just arranges physical nodes on virtual ring. Next, we will discuss how to store replica in data nodes, namely the replica creation problem.

In order to provide high data reliability and availability without losing data access performance, rack-aware idea should be taken into account.

The basic principles of rack-aware strategy can be summarized as below:

- Store the replicas of the same data object in different racks so as to improve data reliability and system fault-tolerant ability.
- If a data is frequently accessed, its two replicas should be stored in the same rack so as to decrease the network transmission across racks.
- Replicas should be arranged among the nodes in cluster as evenly as possible.

Suppose that the default redundancy degree be 3, rack-aware replica creation strategy would arrange 3 replicas into two racks. Namely, the first replica is stored in the local rack, the second replica in a different rack to the first replica, and the third replica in a different node of the same rack to the first or the second replica.

If all the replicas of the same data object are stored in the same rack, it will bring in great one-point risk. Since the rack failure will lead to the unavailability of all replicas. Rack-aware strategy can avoid the data losses caused by rack failure.

Normally, the network bandwidth in a rack is much higher than the bandwidth between two racks. Therefore, it will decrease consistent maintenance cost and improve data access performance if two replicas are stored in two different nodes of the same rack.

A. Algorithm Description

The Rack-Aware Replica Creation Algorithm, RRCA, is detailed below:

Function:

Get the storage position of r replicas according to the known storage position of z replicas, where r is data redundancy degree.

Input:

H_j, the storage positions of previous z replicas, where $0<j\leq z$; r, redundancy degree.

Output:

H_i, the storage positions of the ith replica, where $z<i\leq r$.

1. Begin
2. i = z+1;
3. while (i ≤r)
*4. h=MAX{mod(H_i,N*M)}, (0<j<i+1);*
5. //the position of the ith replica
6. H_i=h+1;
7. if(i == r)
8. break;
9. // n is the most frequently accessed rack
10. DA(n) = MAX{DA(n_j)}, (0<j≤i);

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

11. // m is a random node in rack n
12. m = random node from n, where space(n, m) >size(D)
13. // the position of the (i+1)th replica
14. H_{i+1} = H(n,m);
15. i+=2;
16. end

It should be mentioned that they are the physical nodes that are numbered in virtual ring. In fact, we can still divide a high-performance physical node into several virtual nodes in order to take full advantage of the computing and storage capacity of the physical node. If we still use the above mentioned virtual ring and replica creation algorithm, some replicas maybe are stored in different virtual nodes, which are exist in the same physical node actually. If so, data reliability cannot be guaranteed. Therefore, it is a tradeoff between utilizing hardware performance and improving data reliability. In order to balance these two aspects, we can arrange the second replica to a different physical node that exists in the same rack.

B. Simulation

In this section, we will simulate VRRC and RRCA in different data scale and redundancy degree.

Let data redundancy degree *r* be 5. There are *N*=5 racks in a data center. The number of nodes in each rack is a random integer between 10 and 15. And there is only one data object. Fig.2 shows the replicas distribution of one data object using RRCA.

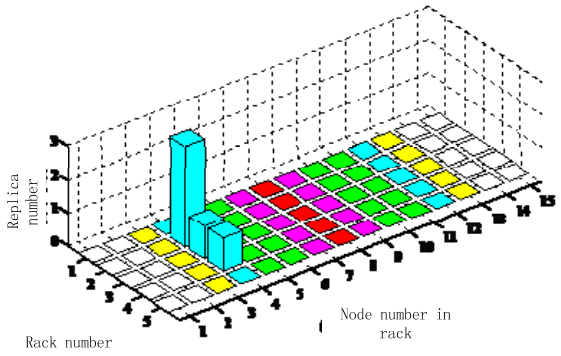

Fig. 2 Distribution of five replicas of one data object

Let data redundancy degree *r* be 3. There are *N*=10 racks in a data center. The number of nodes in each rack is a random integer between 10 and 30. And there are 100,000 data objects. Fig.3 shows the replicas distribution of one data object using RRCA.

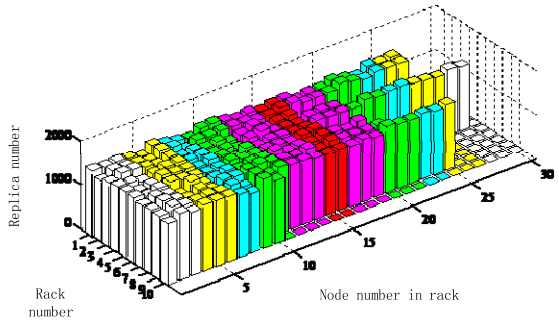

Fig. 3 Distribution of replicas (1)

In the above circumstance, let data redundancy degree *r* be 5. The distribution of replicas is shown in Fig.4.

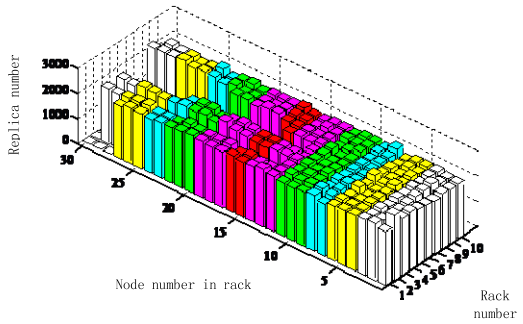

Fig. 4 Distribution of five replicas (2)

RRCA has linear computing complexity, so it takes less computing resources. Besides, the algorithm uses explicit formula to compute replica position, which brings great facilities for data reviewing and management.

V. Conclusions

The paper analyzed the consistent hash based replica arrangement method, and put forward a new virtual ring creation method in order to improve data storage reliability. Based on the virtual ring strategy, a rack-aware replica creation algorithm is presented, which guarantees not only data reliability but also data access capacity.

The paper is an exploration to replica management in big data field. More researches are still under proceeding by us.

References

[1] EMC. "IDC Digital Universe Study- Extracting Value from Chaos". 2011,6
[2] C. Lynch. How do your data grow? Nature, Vol.455, (2008), pp.28-29
[3] Dealing with Data. Science. Vol.331, 2011.02
[4] G. DeCandia, D Hastorun, etc. "*Dynamo: Amazons Highly Available Key-value Store*". In Proceedings of Twenty-First ACM SIGOPS Symposium on Operating Systems Principles, ACM Press New York, NY, USA, 2007, pp.205-220
[5] L. Peng. "*Cloud Computing (Second Edition)*". Beijing: electronics industry press, 2011, pp.43-54
[6] L. Junzhou, J. Jiahui, S. Aibo. "*Cloud computing: architecture and key technologies*". Journal on Communications. Vol.32, Issue 7, 2011, pp.3-21
[7] David Karger, Eric Lehman, Tom Leighton, etc. Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web. In: Proc. of Twenty-Ninth Annual ACM Symposium on theory of Computing. ACM Press,1997, pp. 654-663

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)