Dynamic Replication Strategies for Object Storage Systems*

Tan Zhipeng and Feng Dan
Key Laboratory of Data Storage System, Ministry of Education
School of Computer, Huazhong University of Science and Technology,
Wuhan 430074, Hubei, China
Corresponding authors: zhipengtan@163.com

Abstract: Object replication is an important and sophisticated question, it can reduce access latency and bandwidth consumption, replication can also help in load balancing and can improve reliability by creating multiple replicas of the same storage object. But there are many difficulties to realize object replication in efficiently. In this paper, we discuss the architecture of object storage system with object replication, and analyze key question of dynamic replication such as replica granularity、When to replicate storage object and Where to place the replicas etc.. In the theory, we give the formulas of these question, the analysis prove these formulas are reasonable.

Key words: Object Storage System; Dynamic Replication; Management; Architecture

1. Introduction

In the past few years, network storage technology have great research findings, such as direct-attached storage(DAS)、storage area networks(SAN)、and network-attached storage(NAS)[1]. DAS connects block-based storage devices directly to I/O bus of a hostmachine (e.g.,via SCSI or ATA/IDE). There are some advantages as follows as are: first, the server is it's central, storage security is the server's responsibility; second DAS offers high performance to do I/O operation by the high performance server. But there are several shortcomings, the high performance, scalability, and large capacity are limited; and bandwidth consumption is huge. NAS is a device

* It was supported by the National Basic Research Program of China (973 Program) under Grant No. 2004CB318201.

TAN Zhi-Peng is a Ph.D candidate in Department of Computer Science, Huazhong University of Science & Technology. He is interested in computer architecture、information storage and database technology etc.；**Feng dan** is a professor of Huazhong University of Science & Technology. She is interested in computer architecture、mass information storage and disk array etc..

which can give file service. It integrates hardware and software, now it's protocol is NFS or CIFS. There are several advantages too, first, it is easy to realize file sharing on heterogeneous platform; Second, it can utilize LAN to protect our investment; Third, it can adapt sophisticated network environment. But it's disadvantages are as follows: if application need high speed of access, we can't use NAS, and NAS need high bandwidth too. SAN is network storage system which can transfer data between server and storage system directly by Fibre Channel. It has high performance、high speed of access、high scalable、high availability and concentrated management. But there exist bad interoperability and data file sharing[1]. In order to overcoming these defects, researchers advance object storage system. Now object storage system has become a research hotspot. Unlike traditional storage schema, Storage objects are a new approach to storing data on a magnetic disk drive. Objects are units of disk space managed by the storage device itself. These objects may be of any length and are used to contain operating system and user data. The storage device determines where they are allocated on the media. It modifies traditional block-based interface, by moving low-level storage functions into the storage device itself and accessing the device through a standard object interface[2]. The object storage architecture, combines the two key advantages of today's storage systems, performance and file sharing, and eliminates the drawbacks that have made them unsuitable for Linux cluster deployments. First, the object storage architecture provides a method for allowing computer nodes to access the storage device directly and in parallel providing very high performance. Second, it distributes the system metadata allowing shared file access without a central bottleneck [3,4].

In object storage system, object is base unit of storage、access、and management, it distributes the burden of data storage, computation, communications and administration among thousands of individual client workstations. When a user generates a request for an object, large amounts of bandwidth could be consumed to transfer the object from the server to the client. Furthermore the latency involved could be significant considering the size of the object involved. Our study investigates the usefulness of creating replicas to distribute storage object sets among in the network object storage system. The main aims of using replication are to reduce access latency and bandwidth consumption. Replication can also help in load balancing and can improve reliability by creating multiple copies of the same object. Static replication can be used to achieve some of the above-mentioned gains but has the drawback that it cannot adapt to changes in user behavior. In network object storage system, where the data amounts to petabytes, and the user community is in the order of thousands around the world, static replication does not sound feasible. Such a system needs dynamic replication strategies, where replica creation, deletion and management are done automatically and strategies have the ability to adapt to changes in user behavior.

In this paper we study dynamic replication strategies for network object storage system. There are many related questions of resource discovery so that the request goes to the nearest replica and furthermore how to distribute these requests among replicas to archive best results etc.. The replica placement issue is key question of dynamic replication. The three fundamental questions any replica placement strategy has to answer are: When replicas should be created? Which files should be replicated? Where replicas should be placed?

The rest of the paper is organized as follows. Relate work of dynamic replication will be introduced in section 2, Section 3 describes the dynamic replica management and architecture of object storage system. Section 4 discusses the analysis of replication in object storage system. We end the paper with our conclusions and future directions in Section 5.

2. Relate work

Replication will give several advantages for distributed system、mobile database and network storage system etc. Replication offers more performance and fault tolerance advantages than is available with simple distribution of objects. Replication allows data access to be spread over multiple data servers, it reduces network traffic and reduces access latency; In the network, bandwidth is a valuable resource, replication can reduce bandwidth consumption too; There are several replicas of the same data, so replication improves data availability; on the other hand, if there is overloaded at one site, replication can copy some data to the other site, to realize load balancing. So replication is a important technology, there some research about replication. In the grid, file replication is a research hotpot, and there are lots of achievements [4,5,6]. For example, dynamic replication adapts to changes in user behavior and impossible to manually monitor and mange very huge data sets with thousands of user, difference of caching to replication are: first, caching is decided by client but replication is decided by server; second, the file has to be requested first but the file need not be explicitly asked for. The performance evaluation of dynamic replication in the grid has proved dynamic replication can advance performance of system.

Lots of research has been conducted on using replication to improve data access efficiency and fault tolerance. The general idea of replication is to store copies of data in different locations so that data can be easily recovered if one copy of data at one location is lost. Also, if replication of data can be placed at where the users access them, data access performance can improve dramatically. Replication is a solution to many Grid-based applications such as Climate data analysis and Grid Physics Network, which require responsive navigation and manipulation of large-scale datasets [5].

Dynamic replication Strategies is the key issue, there are six replication Strategies and replica replacement Strategies have been discussed[6,7]. The articles have introduced replica granularity, there exists two kinds of replica granularity as follows as are: one is block-based replication the other is file-based replication. There has no study of object-based replication in really. In peer-to-peer storage, there exists replication research too[9], it discussed the questions like above the questions, and it has got some achievements. But in object storage system, there is little research about replication, the article studies dynamic replication of object storage in systematically.

3. Dynamic Replication Management and Architecture

In the solution we propose, each site can obtain an understanding of the state of the system and takes object replication and migration decisions. The system works as follows. Each node in the object storage system is authorized to select and create replicas for the objects it stores. A node decides where to replicate an object using a performance model that compares the costs and the benefits of creating replicas of a particular object in certain locations. Which replicas are deleted according to the local policy of the host node. That is that dynamic replication management should answer three questions: when the object should be replicated? Which the object should be replicated and deleted? Where the object should be placed?

As shown in Figure 1, in object storage system, Synchronization server can keep replica consistency between storage site and between the object storage systems; the replica manager decide which object need to be replicated? when storage devices should replicate the objects? where the replicas will be placed? The metadata manager servers are the central of object storage system, they manage object-based reading and writing、 security and others management. There are replicas of object at every storage site or object-based storage devices; the whole object storage system can make up of one or several storage networks.

Dynamic replication management can be very expensive. For example, an application may require real-time data that changes frequently, so replicas of the data needs to be updated frequently. In such case, replication is virtually useless. The main goals of dynamic replication management as follows are:

(1)Dynamic: replica can be created and deleted dynamically when the need arises.

(2)Efficient: replica should be created with reasonable amount of time and resources.

(3)Adaptable: the replication process must handle various network speed，heterogeneous storage environment, various processing speeds, and failures.

(4) Flexible: replica should be able to join and leave the grid when it is needed.

(5) Replica Consistency: in an environment where updates to a replica are needed, different degrees of consistency and update frequencies should be provided.

(6)Scalable: the replication system should be able to handle a large number of replicas and simultaneous replica creation.

In our management, object replication is a multi-step process:

(1)On the source sites, there are agent engine that can get information of storage devices, when the storage device is overload or the accessing of object in abundance, an object copier tool is used to copy the objects that need to be replicated into a new object. At the same time the replication manager and Synchronization manager will do some control for replica consistency、efficient、adaptable、scalable and so on.

(2)The new object is moved to the destination storage device using a best algorithm.

(3)When the replica has arrived in destination storage device, the replication manager decided the replica to be deleted at the source site according the situations of the object storage system besides replica consistency.

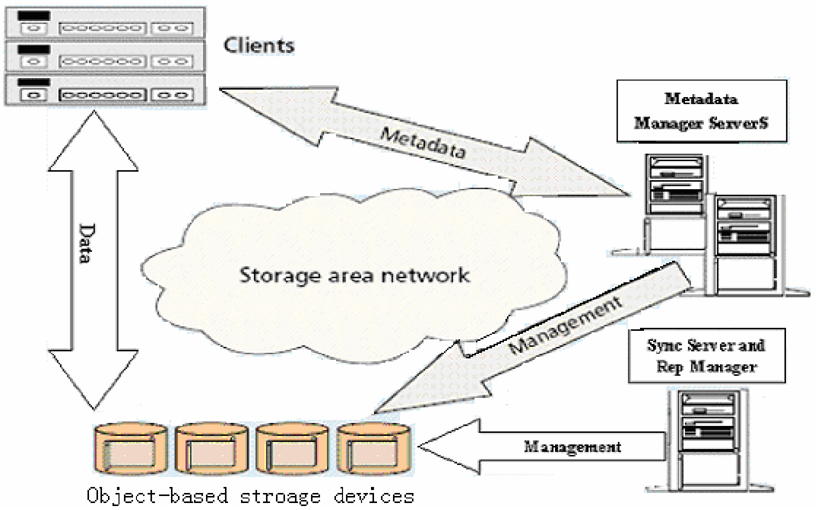

Fig.1 Architecture of the replication object storage system

4. Analysis of dynamic Replication

In the object storage system, the object is a base unit of the system, and object has attributes and operations, in order to help implement of dynamic replication, we expand the data structure of object, as shown Table 1.

Table 1**.** The data structure of object

Object id
Object name
Object size
Object cache size
Object pointer
Object timestamp
Object status
Object access frequency
Object access regulation
Object network bandwidth requirement
Object transfer time

4.1 Replica Granularity

Objects are replicated among many storage devices in the storage system based upon which storage devices download those objects. Whole object replication is simple to implement and has a low state cost it must only maintain state proportional to the number of replicas. However, sometimes one operation is concerned a huge object or lots of objects, if the system replicate these objects, the cost of replicating can be cumbersome in both space and time.

The agent engine can decide number of replicas and granularity of replication object according the application's reading/writing the objects. The replication divides the object into an ordered sequence of fixed size small object. The divide can guarantee large objects to be spread across many peers even if the huge object is larger than what any single peer is able to store, and not replicate part of huge object or entire object that the application does not write/read. However, downloading an object requires that enough hosts storing objects replicas are available. If any one replicated object is unavailable, and the applications do not access the replicated objects, the replication has little significance.

In the network object storage system where all data are accessed at the object level, a natural strategy of data replication would be object-level replication. While the fact that each site uses object-oriented software is a big motivation to use object-level replication as the replication strategy, there is another characteristic particular to the object storage system that makes object-level replication efficient. The replica granularity is important for object replication. If the replication object is huge and the replication object would be copied time after time, the storage space and transfer time will be a new question. Perhaps the object replication does not improve performance of storage system. At the same time, if the replication object is small, building new object with object replicas is difficult.

In the paper the replica granularity will be confirmed by the information got from the object data structure, the information such as object size(O_{size})、 object cache size(OC_{size})、 object access frequency($OA_{frequency}$)、 Object transfer time (OT_{time}) and object network bandwidth requirement($OR_{bandwidth}$). Then there exist the functions of replica granularity and numbers of replicas as follows:

replica granularity = function(OC_{size}, $OA_{frequency}$, $OR_{bandwidth}$)

numbers of replicas = function(O_{size},OC_{size}, $OA_{frequency}$, OT_{time},$OR_{bandwidth}$)

Through we know there exist functions of of replica granularity and numbers of replicas, in this we can't give the exact calculating formula, we are studying them, and we will introduce them in the other paper in the future.

4.2 Analysis of Dynamic Replication

Dynamic replication can realize the reorganization of system and the load balancing of the system, especially in the internet object-based network storage system, in which case that we can't exactly know which OSD (Object Storage Device) to be storied will be reasonable. But replication time and replication location are not favourable, it will not improve data availability. We will analyze replication time and replication location on the under. First, we give a dynamic feedback adjustment model of the variable threshold and span, then we calculate the time and location by the dynamic feedback adjustment model of the variable threshold and span.

4.2.1 When to replicate storage object

Definition 1: In an object-based storage system, each the storage object can be accessed by all users, but when the users exceed a threshold, the performance will decrease in obviously, we describe the maximum connection number with one positive integer C, and describe optimal connection number with one positive integer T(T<C), and the span can be described with one positive integer a, the actual connection number with positive integer L, then L is the function of T and a, and the relationship as follows:

$$F(L,T,a)=\begin{cases} 0 \text{ when } L<T-a\text{, the number of connections is few} \\ 1 \text{ when } L\in[T-a,\ T+a]\text{, the number of connects is moderate} \\ 2 \text{ when } L>T+a\text{, the number of connects is greate} \end{cases}$$

As showed in Figure 2, the number of current connections of storage object exists in one of the three regions described in the figure.

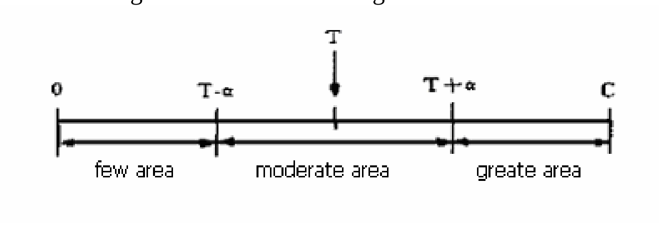

Fig. 2 the rationale of threshold and span

When L<T-a, the current connection of the object is few, it is not the time of object replication. On the other hand, the object storage device is the best location where the object replicas can be placed; when the current connection number in the moderate region, then it means it is not the time of object replication. but the object storage device is not the best location where the object replicas can be placed; when the current connection number is overload, it means that the object must be replicated immediately, and the object storage device is not the location where the object replicas can be placed.

Definition 2: One object-based storage system which is composed of n pieces of object storage devices, we can definite the whole load balancing target of the system as V, then V as follow is:

$$V=\sum_{i\in N}(r_i-\ r_0)^2 \qquad (1)$$

$$\text{thereinto } r_i=\frac{L_i}{C_i};\ r_0=\frac{\sum_{i\in N}L_i}{\sum_{i\in N}C_i}$$

Among the upper formula, we define ri as the rate of connection of one storage object at one storage device and r0 as the whole rate of connection of the storage device.

Obviously, r0 means the whole situation of connection of the storage device, if r0 is more bigger the storage object must be replicated at once, and transfer the replicas to the storage devices whose r0 is small. So we can select the best time to replicate the storage object through the calculation of the V-value.

4.2.2 Where to place the replicas

Storage object replication is really the copy of the storage object. If the replicas can not be placed the best position, the performance of the storage system will not be improved. On the contrary, replication will decrease the performance. Suppose the current connections of the number i object storage device as Li, because the replicas will be transferred to the storage device which will generate new connections, and define the connections as Pk, So whether the storage device can accept the replicas or not is predicated by this:

$$L_i + P_k > T_i + a_i \qquad (2)$$

Once the storage device can accept the replicas of the storage object, the storage device is the best position where the replicas should de placed. According to the formula 2, we should choose the best object storage device which can bring V-value to be least. Then we put forward the reference of algorithm that we select the storage device to place the replicas as follows:

$$\begin{aligned} & Min \sum \left(\frac{WS_i * Q + L_i}{C_i} - r_0 \right)^2 \qquad (3) \\ & s.t. \quad \sum WS_i = 1 \\ & WS_i * Q + L_i \leq C_i \text{ thereinto } Q = P_k, \ WS_i = 0, 1 \end{aligned}$$

Obviously, the replication of the storage object causes the redistribution of the loading of storage devices. Through the formula 3, we can ascertain the position which is the best of storage device to place the replicas, after we research the impact which caused by placing replicas. This is a representative problem of planning 0 and 1, there are some algorithms like either greed method or branch and bound method that can solve the problem.

5. Conclusions and future work

Object replication is an important and sophisticated question, because it can reduce access latency and bandwidth consumption, replication can also help in load balancing and can improve reliability by creating multiple replicas of the same storage object. In the general, Object replication can give great help to improve whole performance of object storage system. On the other hand, there are many difficulties to realize object replication in efficiently. In this paper, we discuss the architecture of object storage system with object replication, and analyze key question of dynamic replication such as replica granularity、When to replicate storage object and Where to place the replicas etc.. In the theory, we give the formulas of these question, the analysis prove these formulas are reasonable. In our future work, we will keep going to study these key questions of dynamic replication, and give practical algorithms of dynamic replication, we will program a prototype system also.

References:

[1] Mike Mesnier, Gregory R.Ganger, Erik Riedel, Object-Based Storage, IEEE Communications Magazine • August 2003, 84-90.

[2]Holtman K.,van der Stok P., Willers I. “Towards Mass Storage Systems with Object Granularity.” In Proc. of the Eighth NASA Goddard Conference on Mass Storage Systems and Technologies, Maryland, USA, March 27-30, 2000, p. 135-149.

[3]K.Holtman, H. Stockinger. Building a Large Location Table to Find Replicas of Physics Objects. Computing in High Energy Physics (CHEP 2000), Italy, Feb 2000.

[4]K.Holtman. Object Level Physics Data Replication in the Grid. VII International Workshop on Advanced Computing and Analysis Techniques in Physics Research ACAT’2000, Chicago, USA, October 16-20, 2000.

[5]A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, and S. Tuecke. The Data Grid: Towards and Architecture for the Distributed Management and Analysis of Large Scientific Data Sets. Journal of Network and Computer Applications, 23(3):187-200, 2000.

[6]Kavitha Ranganathan and Ian Foster, Design and Evaluation of Dynamic Replication Strategies for a High-Performance Data Grid,

[7] Holtman K., Stockinger H. “Building a Large Location Table to Find Replicas of Physics Objects.” In Proc. of CHEP 2000, Padova, Italy, February 2000.

[8] A. L. Chervenak, Naveen Palavalli, Shishir Bharathi, Carl Kesselman, Robert Schwartzkopf, "Performance and Scalability of a Replica Location Service," presented at High Performance Distributed Computing Conference (HPDC-13),Honolulu, HI, June 2004.

[9]Ranjita Bhagwan, David Moore, Stefan Savage, and Geoffrey M. Voelker, Replication Strategies for Highly Available Peer-to-Peer Storage,

[10] F. Dabek, M. Kaashoek, D. Karger, R. Morris, and I. Stoica. Wide-area cooperative storage with cfs. In proceedings of the 18th ACM Symposium on Operating System Principles (SOSP), 2001.

[11] C. C. Aggarwal and P. S. Yu. On disk caching of web objects in proxy servers. In Proc. Int’l. Conf. Info and Knowledge Management, CIKM’97, pages 238 – 245, Las Vegas, Nevada, 1997.