Research on online cloud storage technology

ZOU Shan-hua[1, 2],FANG Ning-sheng[2, 3],GAO Wei-jie[3]

（1. Wuxi Taihu University，Wuxi 214064；2. Jiangsu Key Construction Laboratory of IoT Application Technology Wuxi 214064;3. Southeast University Nanjing 210018）

Abstract: Cloud computing is a popular concept nowadays is rapidly plays an important role in all walks of life to seize. As a typical application of SaaS, online cloud storage provides cloud data storage and handling, the major Internet companies are facing the public with their cloud network disk, the user can be your own file upload to the cloud, and then will be able to access these files from a variety of devices and locations. Cloud network disk in the technology still need to consider quite a lot of problems, such as user authentication, physical data storage, space compression. At the same time, users are increasingly concerned about the privacy issues, data reliability issues and how to profit is also a need to consider the object. Through understanding to personal online cloud storage implementation may be met, by reading the relevant information and documents, and combined with the knowledge learned in daily study, from a certain extent, put forward the corresponding solutions.

Key words: Cloud computing; Cloud storage; SkyDrive cloud; Solutions

Once the concept of cloud storage is put forward, it has been supported and concerned by many manufacturers, which can be seen from various kinds of cloud and network disk in the current market. At present, the more famous and powerful personal online cloud storage providers [1,2] include apple icloud, Dropbox, Google drive, Microsoft onedrive, mega, etc. abroad, and Baidu cloud, 360 cloud disk, Tencent micro cloud, Huawei dbank, etc. in China. The main selling point of foreign personal online cloud storage lies in its confidentiality and reliability, while domestic major providers are committed to promoting the features of user experience, such as providing huge storage space, data "second transmission". The technical implementation of these functions will be mentioned below.

1. Technical problems and solutions of online cloud storage

1.1 basic storage services

Traditionally, all the storage devices are inside the host / server and cannot be shared with other hosts. With the development of network, hosts can use network to transmit data. This architecture is called server centered storage architecture, in which each server has its own storage device. The maintenance of a server or the failure of a server will lead to the inaccessibility of information, resulting in the problems of information difficult to protect, difficult to manage, isolated information island and high maintenance costs.

In order to solve these problems,a new architecture, called information centric architecture, has emerged. In this architecture, the storage device is managed centrally and no longer attached to the server. Storage devices can be shared between multiple servers. When you deploy a new server, allocate storage for it from the shared storage device. The capacity of shared storage can be increased dynamically by adding new devices without affecting the availability of information. This architecture makes information management simpler and more cost-effective.

In this architecture, the most typical is NAS,as shown in Figure 1.NAS[3] is a dedicated high-performance file sharing and storage device. It is a solution of enterprise file server, that is, the part of information storage in information centric architecture. NAS devices use their own operating system, integrated software and hardware components to meet specific file service requirements. NAS optimizes file I/O and excels at all kinds of general file servers in transmission speed. At the same time, today's NAS even supports file level virtualization, which eliminates the dependency between file data and

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00025

physical storage. Even when files are moved in physical media, they can be accessed continuously.

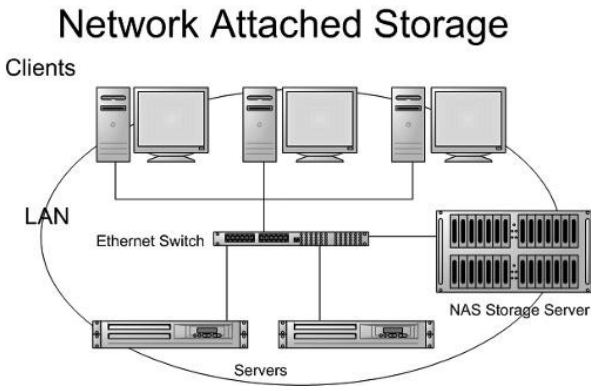

Figure 1 network attached storage (NAS)

The information in the online cloud storage system is of great magnitude, and it will be accessed by multiple users at the same time, and there is no need to protect data reliability all the time. Therefore, this information centric storage architecture fully meets the needs of online cloud storage. On the other hand, the birth of cloud computing and virtualization also makes the centralized management of shared resources possible.

1.2 independent redundant disk array

Nowadays, the price of physical storage is very low. People who are keen on technology and downloading can easily have six pieces of storage (2TB each). In the data center, there are tens of thousands or even hundreds of thousands of pieces of physical storage. Although the failure probability of a single physical memory is very small, when a large number of physical memories are operating at the same time, it can be known by using simple probability theory knowledge, and it is almost impossible that no failure occurs. For example, for a hard disk with an annual failure rate of 0.01%, the probability of normal operation for one year is 99.99%. When there are 10000 pieces in a data center, the probability of no failure in one year is:

$$(1-0.01\%)^{10000} = 36.79\%$$

Mechanical wear and mechanical damage are the causes of hard disk drive failure, which can not be avoided. And with the increase of the number of disks, the probability of overall failure will be more and more large. So we need some methods to prevent data loss caused by hard disk drive failure.

In 1987, Patterson, Gibson and Katz of the University of California, Berkeley, published a paper called "an example of redundant disk array (RAID)". They first proposed the concept of raid. RAID technology [4,5] forms multiple disks into a whole, so that it can provide data protection technology in case of hard disk failure. At the same time, RAID technology can also improve the performance of the storage system, because multiple hard drives can provide I / O services at the same time. At present, the more common RAID levels are shown in Table 1. Raid-3 and RAID-4 are very similar to RAID-5 in technology, so they are not very common now.

Table 1 Introduction to RAID levels

level	Brief description
RAID 0	Striped array without fault tolerance
RAID 1	Disk image
RAID 1+0 /RAID0+1	The application of combining RAID 1 and RAID 0
RAID 3	Parallel access striped array with special check disk
RAID 4	Distributed array with independent disk access and special verification disk
RAID 4	Striped array with independent disk access and distributed verification
RAID 6	Striped array with independent disk access and dual distributed verification

RAID technology has a good application in online cloud storage, which ensures the stability of data in a certain sense. Through the use of striping technology, different RAID levels have different application scenarios, and their overhead, read-write performance are also different. A lot of relevant information can be found on the Internet, so we will not go into details here. The only thing that needs to be mentioned is the emergence of RAID technology, which enables us to accept the failure of a small amount of hard disks in a short period of time, and at the same time, we can automatically repair them through the hot spare disks, which also realizes the

beautiful vision that managers only need to sit in chairs and look at the screen, where the red dot lights up, they will replace a disk drive in the past.

The information in the online cloud storage system is of great magnitude, and it will be accessed by multiple users at the same time, and there is no need to protect data reliability all the time. Therefore, this information centric storage architecture fully meets the needs of online cloud storage. On the other hand, the birth of cloud computing and virtualization also makes the centralized management of shared resources possible.

1.3 load balancing and content distribution network (CDN)

There are many files and many users on Baidu online disk [1], so Baidu can't have only one server, otherwise many users visiting at the same time will cause Baidu online disk to crash. Therefore, in the process of implementing online cloud storage, multiple servers must be deployed, so we must consider load balancing to allocate requests to each server.

The earliest cloud storage load balancing [6] is implemented by using local DNS, which allocates several mappings for the same host name, and uses basic scheduling algorithms such as polling and random allocation to allocate requests. At present, this method is still used in many small and medium-sized websites, such as BiliBili video website (because it indicates which DNS server is directed to which server in the crash information), but it also has a big disadvantage, that is, it is unable to realize the dynamic monitoring of each server. If one of the servers goes down, the DNS server can't be found in time, which leads to the access failure of the users assigned to the server. At the same time, load balancing through DNS can not judge the load of each server immediately. If the scheduling algorithm is not perfect, one server may be almost idle while the other server is under high load pressure.

At present, the most popular load balancing is to use reverse agent. Reverse proxy is similar to forward proxy. We use proxy to visit "some websites that can't be accessed directly", so the firewall can't identify which websites we are visiting, so as to achieve the purpose of indirect access. And the reverse proxy is similar. When we visit a website where the reverse proxy is deployed, we do not visit the current server, but the reverse proxy server. When the request arrives at the reverse proxy server, the reverse proxy then forwards the request to the server. At present, the common reverse proxy servers are built with nginx and other servers, because they have many allocation strategies to ensure the average allocation of access requests. In fact, reverse proxy is similar to dynamic DNS service, but it can achieve the dynamic monitoring function that DNS cannot.

At the same time, personal online cloud storage is often used by users to store various unstructured large-scale data, such as movies, videos, photos, etc., and the current mainstream "online disk" has realized various online preview functions. If a few years ago, online video has been buffering and previewing photos have been showing red forks, we will not care much about it, but now, we will definitely feel that this network disk is made of a lot of slag, so we give up using it. In order to solve this problem, there is a content distribution network (CDN).

Simply speaking, CDN is one or more servers that store some static files, and save the files through copying, caching and other ways. Because the files in online cloud storage can usually be classified as static data, the use of CDN can be very appropriate. In the era of no CDN, all data is obtained from the main server. If the server is in Beijing and we visit in Guangzhou, the speed of access will slow down due to factors such as transmission distance, operator and loan what's the relationship between this and payment for goods? Please put it another way. After using the CDN service, the CDN server will be deployed in different geographical locations, and the CDN server will cache files after the first request of users, or actively request data from the primary server and cache. Thus, when the user sends the request to the server, the server judges the user's geographical

location through IP, operator and other information, and allocates the nearest CND server to accelerate the loading of static data.

1.4 disaster recovery

If RAID technology can guarantee the security of data to a certain extent, then when the data center has an earthquake, flood or large-scale power failure, RAID technology can't help. In order to deal with this kind of catastrophic destruction, we must take corresponding countermeasures to protect the data security. Backup is the primary requirement for disaster recovery. When the primary location fails to work due to a disaster, the backup copy will be used to recover the data in the second location. There are different backup schemes for different information availability requirements. In the early days, people used the way of tape backup, and the backup tape media was transported to different places for storage. However, this method has a long recovery point, which will lead to data loss and "backlogging" in a period of time before the disaster. Among all kinds of video games, the game "backgammon" is the thing that players feel most bitter about, as well as in online cloud storage. Users may store the very important data in the "network disk" after modification, so they cannot accept the loss of data. At present, the remote replication technology is often used to copy the data to the disaster recovery location in real time, so that the production system can be recovered in a relatively short time in case of a disaster.

Remote replication[7] is divided into synchronous mode and asynchronous mode, and there are host based replication, array based remote replication, network-based remote replication and other technologies. Readers can find relevant materials for details. At the same time, because online cloud storage is a service based on cloud computing, when a disaster occurs, because there are multiple CDN sites and multiple backup servers, users can't even perceive the occurrence of the disaster, only feel that the speed of downloading or linking becomes slower.

reference:

[1] Li Xinyu. Introduction to network cloud disk - Take 360 cloud disk and Baidu cloud as examples [J]. Wireless Internet technology, 2014,01:38

[2] China's personal cloud storage applications [J]. Communication world, 2012, 08:46-47

[3] Zhang Yang. Design and implementation of a personal cloud storage service system [D]. University of Electronic Science and technology, 2012

[4] Chen Huaying. Raid reliability analysis of disk array [J]. Journal of University of Electronic Science and technology, 2006, 03:403-405

[5] Cao Yang. Implementation and development of RAID technology [J]. Computer learning, 2006, 04:43-44 + 60

[6] Extension conservation. Research on cloud computing and cloud data storage technology [J]. Computer development and application, 2010, 09:1-3 + 9

[7] Li Ling. Research on data security in cloud computing services [D]. University of science and technology of China, 2013