Key Technologies Research of Distributed Storage Based on Cloud Computing

Shaolong Shan^{1, a}, Dongling Wu^{1, b}

¹ Tangshan Vocational & Technical College

^a 29463989@qq.com, ^b 176288793@qq.com

Keywords: Key technologies; distributed storage; cloud computing

Abstract. Considered as the next generation computing model, cloud computing plays an important role in scientific and commercial computing area and draws great attention from both academia and industry fields. In the paper, cloud computing has been described in detail, while the key technologies of distributed storage based on cloud computing is in a detailed description. And elaborates that with the development of computers, the technology has been used in various fields, and plays a positive role in many fields of development.

Introduction

Cloud computing as the next generation model, has widely used in many areas of scientific computing and business computing and others. Data Center as the basis of cloud computing, needs to address the distribution of mass data storage under environmental problems in scalability, fault tolerance and cost aspects. This requires in-depth study of data center physical network topology construction techniques, increases data fault tolerance techniques to reduce energy consumption as well as a variety of energy-saving technologies. Build a scalable, fault-tolerant and low-cost distribution of storage data center, involving a variety of technical methods in a cloud computing environment. These technologies are currently distributed in the storage areas of research, but there are still a great distance with the actual deployment of the application .

This paper studies the challenges facing the construction of distributed storage in the cloud computing environment, as well as a variety of key technologies to solve these challenges. The latest research progress of these key technologies are classified depending on the technical standards. And based on the classification, analyze and compare a number of related technical methods, point out the problems of these technologies. Finally, propose some directions for future research.

Cloud computing

Considered as the next generation computing model, cloud computing plays an important role in scientific and commercial computing area and draws great attention from both academia and industry fields. Cloud computing is hailed as revolutionary computing model.

Cloud computing^[1] is computing model to provide users with a configurable, shared infrastructure computing resources. It enables the user to facilitate real-time access to the network, storage, and other computing resources in the case of a cloud service provider rarely involved. Cloud computing providers build one or several large-scale data centers, data center by connecting a large number of nodes and network equipment together and then provide the basis for a variety of levels of service to users, such as infrastructure services, platform services, storage services and software services. Cloud computing^[2] has a very large scale, high scalability, high reliability, virtualization, low-demand services and prices and other characteristics, which can well meet the massive data storage requirements.

In the cloud computing environment, massive data is stored on different nodes in the same data center, or even a node on different data centers, but the location and organization of data is transparent to the user. An easy to use interface to data center users only need to provide access to data by service providers. Data storage, organization, management, and reliability, availability,

guaranteed by the cloud provider is responsible. Cloud computing allows users not to have to build their own data centers, reducing the cost to the user. They only need to pay a fee based on their needs and can easily store data onto the data center to submit the task when, at last, get results.

The key technologies of distributed storage based on cloud computing

The background of the key technologies of distributed storage based on cloud computing. Today, with the a variety of information processing technology relying on the rapid development of computer technology emergence, the communication technology class is rapid development. Such large amounts of data is only stored on the computer has been unable to meet the needs of large amounts of data. Based on this, the key technology is based on distributed storage cloud appears.

Under cloud computing environment, data center consist of a large amount of computers, usually up to millions, and stores petabyte even exabyte of data, which may easily lead to the failure of the computers or data. The large amount of computers composition not only leads to great challenges to the scalability of the data center and its storage system, but also results in high hardware infrastructure cost and power cost. Therefore, fault-tolerance, scalability, and power consumption of the distributed storage for a data center becomes key part in the technology of cloud computing, in order to ensure the data availability and reliability.

The key technologies. The traditional way to improve the scalability is usually by way of redundancy of the disk reserved. This method can guarantee enough storage space a certain extent. However, cloud computing node scale of modern data center environment is tens or even hundreds of thousands. In addition, the data stores in the data center or even EB or PB calculated. The size of the data center and the size of the data storage will expand rapidly with the applications increasing. Therefore, any cloud service provider's data center by way of the disk reserved in the beginning of the establishment can not be fully planned. Large size and rapid growth of distributed storage scalability put forward higher requirements, not only requires the data center network have good scalability, and organizational structure of the data must also have good scalability to adapt to expand the application requirements.

Conventional methods to improve fault tolerance, such as high performance servers, dedicated storage device or RAID technology is costly, which is difficult to meet the profit of cloud computing providers. And the huge scale of nodes and size of the data greatly enhance the probability of failure. In the cloud computing environment, the failure behavior becomes the norm. Frequent failure behavior and enormous losses, makes fault tolerance have become an important challenge to cloud computing environments distributed storage facing an urgent problem. In the cloud computing environment, improving the distribution of fault-tolerant storage, not only study the relationship between interconnected nodes, to improve fault tolerance physical topology, but also research organization and management of data stored on the node to improve fault tolerance data.

Traditional distributed storage nodes because of its small size and data, and less consideration of energy consumption, enterprises are generally willing to exchange for cost efficiency and reliability. In the cloud computing environment, storage of huge size distribution, consumption spending is also great. In order to enable the device to normal operation, should include energy consumption of refrigeration equipment. As service providers rely on corporate profits, reduce energy consumption and thus reduce costs is a necessary goal of cloud computing. You can save energy and reduce energy consumption and promote environmental protection. Energy-saving technology has become one of the key technologies in distributed storage design and efficiency, fault-tolerant parallel.

The architecture of the key technologies of distributed storage based on cloud computing. The traditional enterprise data centers powers by the data center network to switch to the center of the building. This architecture uses a switch to connect the server. Forwarding packets borne entirely by the switch, server is only responsible for storing and processing data. Most of the traditional structure use a tree structure. Switch to the center of the structure is generally connected to a three-tree structure, namely the edge layer, aggregation and core layers. A rack of servers are

generally connected to a chassis switch, constitute edge layer. Edge layer switch certain structure and aggregation layer switches to connect, complete and balanced through bandwidth aggregation. Finally, the aggregation layer switches and core layer routing devices are connected, to provide users with access from outside the data center routing.

Simple and intuitive tree structure, easy to operate, easy to connect and realized by increasing the frame and corresponding switches can be more convenient to expand, but there are a lot of questions: Link bandwidth limited capacity; poor flexibility, lower server utilization is low; switch serious waste of resources; unable to meet the communications needs.

Server-centric architecture^[3] for each server is by installing multiple NICs for server interconnection, no switches and routers forwarding device data. Data forwarding all the work are done by the server. CamCube is a server-centric architecture, and its topology is shown as Figure 1^[4].

Figure 1 topology structure of CamCube

CamCube is a pure interconnected by service data center network structure, no switches and routers. Between servers directly connected via multiple network cards, routing packets forwarding work entirely borne by the server. One server nodes are directly connected with the other two servers in each direction, three-dimensional, three-dimensional form a cyclic structure.

CamCube structure and wiring connections are relatively simple, but have a high link redundancy. Network does not switch the device and no single point of tree structure bottlenecks, which allows servers to directly interact with the underlying network. The upper application can develop more efficient routing algorithm based on the demand. However, all of the data forwarded by the server to complete the work, you need to take up the computing resources of the server. The server load pressure, causes a decline in the computational efficiency of the server. And the number of cards in each server installation are limited, which are currently the largest installation of two cards for each machine. Each card has six ports, so the size of a limited number of nodes.

Hybrid architecture server nodes connected through the switch while installing multiple NICs for each server so while participating in routing and forwarding data packets. You can design a more unique and apply to specific application scenarios network structure. It is designed to address the shortcomings in server-centric and switch-center of the structure. Hybrid structure can be constructed out of more freedom and flexibility to adapt to specific application scenarios and network structure.

Conclusion

Considered as the next generation computing model, cloud computing plays an important role in scientific and commercial computing area and draws great attention from both academia and industry fields. Under cloud computing environment, data center consist of a large amount of computers, usually up to millions, and stores petabyte even exabyte of data, which may easily lead to the failure of the computers or data. The large amount of computers composition not only leads to great challenges to the scalability of the data center and its storage system, but also results in high hardware infrastructure cost and power cost. Therefore, fault-tolerance, scalability, and power consumption of the distributed storage for a data center becomes key part in the technology of cloud computing, in order to ensure the data availability and reliability. In this paper, a survey is made on the state of art of the key technologies in cloud computing in the following aspects: Design of data center network, organization and arrangement of data, strategies to improve fault-tolerance, methods to save storage space, and energy. These technologies are currently distributed in the storage areas of research, but there are still a great distance with the actual deployment of the application.

This paper studies the challenges facing the construction of distributed storage in the cloud computing environment, as well as a variety of key technologies to solve these challenges. The latest research progress of these key technologies are classified depending on the technical standards. And based on the classification, analyze and compare a number of related technical methods, point out the problems of these technologies. Finally, propose some directions for future research.

References

- [1] U.S.Department of Commerce.The NIST definition of cloud computing.National Institute of Standards and Technology, 2011.
- [2] Chen K,Zheng WM. Cloud computing:System instances and current research.Journal of Software,2009,20(5):1337-1348(in Chinese with English abstract).
<http://www.Jos.Org.cn/1000-9825/3493.htm> [do:10.3724/SP.J.1001.2009.03493]
- [3] Abu-Libdeh H,Costa P,Rowstron A,O'Shea G, Donnelly A. Symbiotic routing in future data centers.In::Proc. of the SIGCOMM 2010.New Delhi: ACM Press, 2010.51–62.[doi: 10.1145/1851182.1851191]
- [4] Ratnasamy S,Francis P,Handley M, Karp R, Shenker S.A scalable content-addressable network. In: Proc.of the SIGCOMM 2001.San Diego: ACM Press, 2001.161–172. [doi:10.1145/383059.383072]