Cloud computing and its key techniques

Songjie
Software school
Xiamen University
Xiamen,China
309597136@qq.com

Junfeng Yao
Software school
Xiamen University
Xiamen,China
yao0010@xmu.edu.cn

Chengpeng Wu
Software school
Xiamen University
Xiamen,China
munihei1986@yahoo.com.cn

***Abstract*—Cloud computing is a new computing model; it is developed based on grid computing. The authors introduced the development history of cloud computing; took cloud computing of Google techniques as an example, summed up key techniques, such as data storage technology (Google File System) , data management technology (Big Table) , as well as programming model and task scheduling model (Map-Reduce) , used in cloud computing. Finally the paper analyses the challenge of cloud computing and pointed out the broad development prospects of cloud computing.**

Keywords-cloud computing; data storage; data management; programming model

I INTRODUCTION

Cloud Computing is the latest developments of computing models after Distributed Computing, Parallel Processing, Grid Computing and so on. Cloud computing achieve multi-level virtualization and abstraction through effective integration of variety of computing, storage, data, applications and other resources, users can be easy to use powerful computing and storage capacity of cloud computing only need to connect to the network.

There is no doubt that cloud computing is the most popular topic in IT industry in 2009, Google, Amazon, Yahoo and other Internet service providers, IBM, Microsoft and other IT vendors have put forward their own cloud computing strategy, various telecom operators are also have put a great deal of attention on cloud computing, the very low cost of cloud computing platform becomes the focus of the industry. Google claims that because of the use of cloud computing technology, its computing cost is only competitors 1/100, the storage cost is only competitors 1/30. If this is really the case, then how did Google do it?

In order to meet the needs of operational management, telecom operators built many large-scale IT systems, such as China Mobile has built business support systems, network management systems and management information systems etc, these IT systems are generally built on the basis of high-performance UNIX server cluster, compared with Google cloud computing which is built on the basis of a large number of low-cost x86 server clusters, there exist obvious differences in technical architecture between them. This article attempts to find out the root cause of Google cloud computing platform have a very low computational cost and storage cost by the comparative study of cloud computing platform and traditional IT systems in the basis of in-depth analysis of cloud computing platform key technologies.

II THE KEY TECHNOLOGY OF CLOUD COMPUTING

Cloud computing is a new type of super-computing way, take the date as the center, is a data-intensive super-computing. In data storage, data management, programming models, and many others have their own unique technology which also involves a number of other technologies at the same time. This chapter mainly introduces cloud computing unique technologies, including data storage technology, data management technology, programming model, etc.

At present, Google is the biggest users of cloud computing, have its own cloud computing platform. Cloud

computing infrastructure model that used by Google mainly includes three both independent and closely system, also known as the three magic weapons of cloud computing including: Google File System built on cluster of Google; Big Table which is a model-simplified large-scale distributed database developed by Google; and the Map/Reduce programming model which put forward based on the characteristics of the application of Google.

A. Data Storage Technology

To ensure high availability, high reliability and economy, cloud computing use a distributed storage methods to store data, use a redundant storage methods to ensure the reliability of data storage, namely, store multiple copies for a data. In addition, the cloud computing system needs to meet the needs of large numbers of users and to provide the service to a large number of users in parallel. Therefore, the data storage technology of cloud computing must have characteristics of high throughput and high transfer rate.

The data storage technologies of cloud computing mainly have non-open source GFS (Google File System) of Google and open source HDFS (Hadoop Distributed File System) of the implementation of GFS of the GFS Hadoop development team developed. Most IT vendors, including Yahoo, Intel, their "cloud" plans are using the HDFS data storage technology.

The Google File System that designed and implemented by Google have many of the same goals such as performance, scalability, reliability and availability with the distributed file system before but also influenced by application workloads and technological environment of Google. Mainly embodied in the following four aspects: a cluster node failure is the norm, not an exception; the concept of the file size of Google file system is not the same as the file size of the usually file system, file size is usually to G bytes ; files' read and write model of Google file system is different with the traditional file system; file system is no longer transparent in some specific operations, but also need the assistance of the application, the co-design of applications and file system API have improve the overall system flexibility.

Given the system architecture of Google File System, a GFS cluster contains a single master and multiple chunkservers and is accessed by multiple clients. Files are divided into fixed-size chunks. Each chunk is identified by an immutable and globally unique 64 bit chunk handle assigned by the master at the time of chunk creation. Chunkservers store chunks on local disks as Linux files and read or write chunk data specified by a chunk handle and byte range. To ensure reliability, each chunk is replicated on multiple chunkservers, by default we store three replicas. The master maintains all file system metadata, including namespace, access control information, the mapping information from files to chunks, and the current locations of chunks. GFS client code is embedded in each program, which implements the Google File System API, to help the application communicate with the master and chunkservers to read or write data. Clients interact with the master for metadata operations, but all data-bearing communication goes directly to the chunkservers. Through the co-design of server and client, Google File System can get the maximum performance and availability results to its own applications.

B. Data Management Technology

Cloud computing system process and analysis large data sets to provide efficient services to users. Therefore, the data management technology must be able to efficiently manage large data sets. Then, how to find specific data in the large-scale data has also become a problem of cloud computing data management technology which must be addressed.

The characteristics of cloud computing is the mass of data storage and reading and then a lot of analysis, the data read operation frequency far more than the data update frequency; the data management in cloud is a read optimized data management. Therefore, the data management of cloud systems often uses the data management model of column stored in the field of database management. That is to take the table divided by column and storage.

The most famous data management technology of cloud computing is the Google's Big Table data management technology.

As many applications of Google (including Search History, Maps, Orkut and RSS readers, etc.) need to manage a large number of formatting or semi-formatted data, a common feature of these applications is the need to support the massive data storage, reading and then a lot of analysis, data read operation frequency far more than the data update frequency, etc., for which Google has developed a weak-consistency large-scale database system -- Big Table.

BigTable are optimized for data read operations, using of distributed data storage management model which based on column storage to improve data access efficiency. The basic elements of BigTable are row, column, record tablet, and timestamp. Among them, the record tablet is a line of collection of row.

The data items are ordered according to lexicographic order of the row keywords, each row is divided into the records dynamically each sever node Tablet Server responsible for the management board of about 100 records. Timestamp is a 64 bit integer that represents the different versions of the data. Column cluster is a collection of several columns, access limits in BigTable control in the granularity of column cluster.

BigTable system relies on the underlying structure of the cluster system, one is distributed cluster task scheduler, one is the GFS file system, and one more is distributed lock service Chubby. Chubby is a very robust coarse-grained locks, BigTable use Chubby to save the Root Tablet pointer, and use one server as the master to store and operate metadata. When a client reads data, the user firstly obtain the location information of Root Tablet from ChubbyServer, and read the corresponding location information of the metadata table Metadata Tablet, and then read the location information of User Table which contains location information of the target data from the Metadata Tablet, and then read location information items in the target data from the User Table.

The master server of BigTable not only manage metadata, but also responsible for remote management and load allocation for the Tablet Server. Client control communications through the programming interface and the master server to obtain metadata, and it communicate data with the Tablet Server. But specific read and write requests handled by the Tablet Server.

Similar with the system mentioned above, BigTable is also the co-design of client and server, making the performance to the greatest extent possible consistent with application requirements.

C. Programming Model

In today of the high-performance computing has not completely universal, and even the most professional development of parallel programs have lagged far behind the hardware development, a common programming model is lacked to achieve parallelism of existing procedures. The Google and Hadoop project is representative today. Google has developed a programming tool Map/Reduce of java, Python, C++, it is a simplified distributed programming model and an efficient task scheduling model for parallel computing of large-scale data sets. The idea of Map/Reduce model is to take the problem be decomposed into Map (mapping) and Reduce (simplified) way, the first data partitioning by Map program, scheduling to the number of computer to processing, to achieve the effect of distributed computing, and then the summary of the results output through Reduce program. Map and Reduce Operating was proposed in 1956 by the McCarthy and was absorbed and expanded by the genius of Google. The Hadoop from open source community Apache imitate the Map/Reduce, but it is still very unstable. To make it easier for users to enjoy the service come from cloud computing and users can use the programming model to write simple programs to accomplish a specific purpose, the programming model of Cloud computing must to be very simple. The complex parallel executing and task scheduling of background must be transparent to users and programmers.

Cloud computing mostly adopted Map/Reduce programming model. Today, most IT companies proposed the programming model of the "cloud" planned, are programming tools based on the ideas of Map/Reduce.

Map/Reduce is not just a programming model, but also an efficient task scheduling model. Map/Reduce programming model is not just applicable to cloud computing, but also have the same good performance on

multi-core and multi-processor, cell processor and heterogeneous cluster.

While the programming model is only applicable to write programs of the task within loosely coupled and high paralleled. How to improve the programming model to allow programmers to write tightly coupled program easily and can be efficiently scheduling and executing tasks in run-time, is the future developing direction of Map/Reduce programming model.

Map-Reduce is a programming model which can process and produce large-scale data set, programmers specify each block data processing in the Map function and specify how to reduce to the middle results of block data processing in the Reduce function. Users only need to specify the map and reduce functions to write the distributed parallel program. When running the Map/Reduce program on the cluster, the programmer don't need worry about how to partitioning, distribution and scheduling to the data of inputting, and the system will also handle the failure of the cluster nodes and the management of communication between nodes in the cluster. Figure1 below gives a specific implementation procedure of the Map-Reduce.

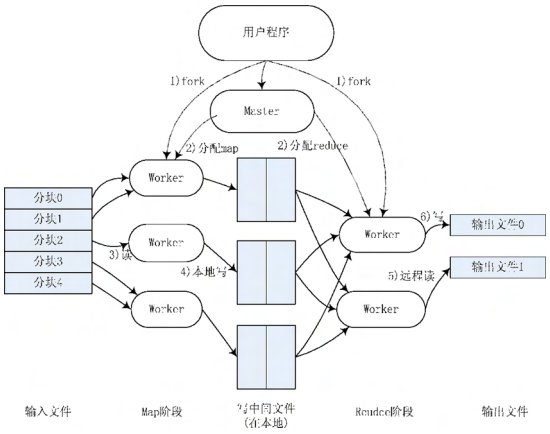

Figure 1

III THE CHALLENGE OF THE CLOUD COMPUTING AND CONCLUSION

This article discusses the concept of cloud computing, implementation mechanism, architecture and several forms and characteristics of cloud computing. And take Google as an example to introduce the basic architecture of cloud computing: Google File System, BigTable and Map/Reduce.

Of course, cloud computing platforms are also facing many challenges, including the following:

Data security and privacy protection: the security of user data is considered to be the security problem of computing platforms, security problem of computing platform is an important issue of cloud computing. Cloud computing infrastructure with a multi-tenant properties, manufacturers generally cannot guarantee that the data of two different users to achieve physical separation. In addition, considering the massive expansion (scalability) requirements, the physical location of the data may not be guaranteed.

Data access and storage model: Now that the storage models provided by big vendors need to adapt to many different usage scenarios. Thus, they may favor to simple memory model or a simple hierarchical model which based on binary object. Although it has brought significant flexibility, it also increase the burden to the application logic explains the relationship between different data elements.

Lack of standards and vendor locking: Most vendors have defined standards-based mechanisms (HTTP, REST, and SOAP, etc.) to access and use its services. However, the standard of development services in cloud computing is just rising, and now the lack the function of write once and run everywhere.

Services Interoperability: Currently, cloud computing doesn't have enough support for the interoperability of services, this have a lots problems for the service of cross-platform or the services between different service.

These are all the challenges the cloud computing model faced. Although there are many problems of cloud computing need to be resolved, but cloud computing will be greater development.

Cloud computing has broad prospects for development, is also the rapid development of related key technologies.

REFERENCES

[1] Wangpeng .Cloud computing key technology and application examples [M].Beijing:Publish house of People's Posts and Telecommunications, 2010.

[2] China cloud computing. http://www.chinacloud.cn/show.aspx?. Id = 515 & cid = 17.

[3] Caijian, Shumei Wang. Cloud computing example analysis based on Google [J].journal of computer knowledge and technology, 2009, (5): 7093-7095.

[4] Jinhua Han. The review of Cloud computing [J].journal of enterprise technology development, 2010, (29) : 13-14.

[5] Chenquan, Qianni Deng. Computing clouds and the key technology [J], 2009, (29) : 2562-2567.

[6] Sunjian, Xiaojing Jia. The technical framework and the influence of cost of Google cloud computing platform [J].journal of research and development, 2010, (1) : 38-44.

[7] Chenkang. The exploring of Cloud computing back-stage large-scale data processing technology [J], telecom engineering technology and standardization, 2009, (11) : 12-16.

[8] Liaoli, Zhangtao. The reviews of Cloud Computing and the prospects [J], information technology, 2009, (29) : 21

[9] Chenkang, Weimin Zheng. Cloud Computing: system examples and researching status [J]. Journal of software, 2009 (5): 1337-1348.

[10] Shenlin, Zhongwen Xia. Education practice research based on Google cloud computing [J], teachers' growth, 2009 (9): 42aa -42dd.

[11] Borthakur D .The Hadoop Distributed File System: Architecture and Design.2007

[12] Dean J.Ghemawat S. Map/Reduce: Simplied Data Processing on Large Clusters.2004

[13] W. E. Howell, "The Growth of Cloud Drops in Uniformly Cooled Air," J. Meteor., vol. 6, Apr. 1949, pp. 134-149.

[14] M. P. Meyers, R. L. Walko, J. Y. Harrington, and W. R. Cotton, New RAMS Cloud Microphysics Parameterization. Part II. The Two-moment Scheme," Atmos. Res., vol. 45, Aug. 1997, pp. 3-39.