Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 33 (2012) 1791 – 1797

Physics Procedia

2012 International Conference on Medical Physics and Biomedical Engineering

Research on Key Technologies of Cloud Computing

Shufen Zhang,Hongcan Yan ,Xuebin Chen

Hebei United University
NO.46 Xinhua West Street, Tangshan 063009, Hebei Province, China

Abstract

With the development of multi-core processors, virtualization, distributed storage, broadband Internet and automatic management, a new type of computing mode named cloud computing is produced. It distributes computation task on the resource pool which consists of massive computers, so the application systems can obtain the computing power, the storage space and software service according to its demand. It can concentrate all the computing resources and manage them automatically by the software without intervene. This makes application offers not to annoy for tedious details and more absorbed in his business. It will be advantageous to innovation and reduce cost. It's the ultimate goal of cloud computing to provide calculation, services and applications as a public facility for the public, So that people can use the computer resources just like using water, electricity, gas and telephone. Currently, the understanding of cloud computing is developing and changing constantly, cloud computing still has no unanimous definition. This paper describes three main service forms of cloud computing: SAAS, PAAS, IAAS, compared the definition of cloud computing which is given by Google, Amazon, IBM and other companies, summarized the basic characteristics of cloud computing, and emphasized on the key technologies such as data storage, data management, virtualization and programming model.

Keywords-cloud computing; virtualization; MapReduce; BigTable

1. Introduction

Under the traditional model, the enterprises need to purchase not only infrastructure such as hardware but also software licenses to establish an IT system, and need specialized personnel to maintain. It needs to upgrade various facilities including hardware and software to meet the demands when the scale of business extends. For the enterprises, what they really need is only the tool which can complete the work and improve efficiency but the hardware and software themselves. For the individuals, we need to install much software to use the computer normally, but much software isn't free, so it is very uneconomical for

 doi:10.1016/j.phpro.2012.05.286

the users who don't often use the software. Could have a kind of service that provides all the software we need for us to rent? Thus we only need to pay a small amount of rent while we using it, so we can save much money. We use electricity every day, which is supplied by the power station instead of our own generators; we use water every day, which is supplied by the water plants instead of our own wells. This mode saves resources greatly and facilitates our lives. Facing the problems brought by the computer, can we use the computer resources like using water and electricity? These ideas eventually lead to the emergence of cloud computing.

Dell put forward the concept of Cloud Computing form the enterprise level firstly, however, but IBM-Google parallel computing project and Amazon EC2 had deep impact to the concept of cloud computing. Then more and more media, companies, technical staff began to chase the cloud, even put a lot of IT innovations into the concept of cloud computing. This promoted and developed the concept of cloud computing and industry of cloud computing, and formed a complete industrial chain including IAAS, PAAS, XAAS and many hardware manufacturers and infrastructure operators participated in it.

2. Summariness of Cloud Computing

Could computing is a new type of computing mode, It distributes computation task on the resource pool which consists of massive computers, so the application systems can obtain the computation power, the storage space and software service according to its demand. This kind of resource pool is called "cloud". The Clouds are some virtual computation resources which can be maintained and managed by themselves, usually they are some large-scale server clusters, including calculating server, storage server, the broadband resources and so on. Cloud computing can concentrate all the computing resources and manage them automatically through the software without intervene. This makes application offers not to annoy for tedious details and more absorbed in his business. It will be advantageous to innovation and reduce cost. This is favorable to innovation and cost reduction. It's the ultimate goal of cloud computing to provide calculation, services and applications as a public facility for the public, So that people can use the computer resources like using water, electricity, gas and telephone.

2.1 Characteristics of Cloud Computing

Currently, the understanding of cloud computing is developing and changing constantly, cloud computing still has no unanimous definition. But, there are four key elements in the definition of cloud computing.

1) The hardware and the software are resources which are provided to the users in form of service through the Internet. For example, Amazon EC2 provides processing ability which is baled into resources for the users; Google App Engine compresses application software and hardware which isnecessary for Web application from designing and development to deployment and implementation, then provides to the users. The resource of cloud computing is expanded to the software categories, including software platform, the Web service and the application program rather than limiting in physical categories, such as processor's period, network bandwidth.
2) These resources can be dynamically extended and configured as needed. For example, Amazon EC2 can initialize resources of 200 virtual servers for the Washington Post society, and recycles these resources in 9 hours while the mission completed.
3) These resources are existed physically in form of distributing and sharing, but logicaly it is presented in form of single and whole.For example, IBM owns 8 institutes in the world, IBM RC2 connected the data centers of these the institutes via the intranet to provide service for the researchers around the world As an end user, These researchers don't know and care that a certain

science operation is completed by which server. Because distributed resources of cloud computing hide the implementation details, and finally it is presented to the users in form of single and whole.

4) The users use the resources when they need and pay according to the actual dosage, but needn't to manage them.

2.2 Service Forms of Cloud Computing

1) Software as a Service(SaaS)

SaaS provider dispose the applied software unified on their server, the user can subscribe applied software service from the manufacturer through Internet .The Provider supply software pattern through Browser, and charge according to the quantity of software and using time.

The advantage of this kind of service pattern is that the provider maintains and manages software, supplies the hardware facilities, the users can use software everywhere when they own the terminal which can log in Internet. Under this pattern, the users can use the corresponding hardware, the software and the maintenance service via the Internet, by paying some rents rather than liking traditional pattern which made users to spend much funds on them. This is the most benefit business pattern of the network application. For small business, SaaS is the best way to use advanced technology. At present, Salesforce.com is famous company for providing these services, so as Google Doc, Google Apps and Zoho Office.

2) Platform as a Service(PaaS)

PaaS takes develop environment as a service to supply .It is a kind of distribution platform server, the manufacturers supply service to the users, such as develop environment, server platform and hardware resources, and the users customize and develop their own application and transfer to other customers through their server and Internet. PaaS can provide the middleware platform, application development, database, application server and experiment for the enterprise and the individual. Google App Engine is the representative product, as well as fore.com and 800 APP.

3) Infrastructure as a Service(IaaS)

IaaS takes infrastructure which is made of many servers as a measurement service to the customers. It integrates memory and I/O devices, storage and computing ability into a virtual resources pool, and provides storage resources and virtualization service for the whole industry. This is a way of hosted hardware, and the customer pays when they use the hardware. For example, Amazon Web Service and IBM Blue Cloud all rent the infrastructure as a service. The advantage of IaaS is that the user only need low cost hardware and rent computing ability and storage ability according to his need, greatly reduced cost of the hardware. Currently, Google cloud application has most representatives, such as Google Docs Google Apps, Google Sites and Google App Engine.

3. Key Technologies of Cloud Computing

Cloud computing systems use many technologies, of which the programming model, data management, data storage, virtualization are the key technologies.

3.1 Virtualization

Virtualization is a method of deploying computing resources. It separates the different levels of the application system including hardware, software, data, networking, storage and so on, breaks the division among the data center, servers, storage, networking, data and the physical devices, realize dynamic architecture, and achieves the goals of managing centralized and use dynamically the physical resources

and virtual resources, improving the flexibility of the system, reducing the cost, improving the service and reducing the risk of management.

In the cloud computing environment, all virtualization solutions are system integration solutions including servers, storage systems, network devices, software and service. They include multiple layers of virtualization technologies such as hardware virtualization, network infrastructure virtualization, application virtualization and desktop virtualization, and combine several layers flexibly to realize the different models of virtualization solutions according to the application environment.

In the whole cloud computing virtualization strategy, We can make use of various mechanisms which is provided by virtualization technique, quickly imitate different environment and experiment without important hardware and physical resources ,and achieve the purpose of building operate system and application, raising the safety and realizing management environment, for later in a more simplified and effective way to put them into the production environment. Thus provide greater flexibility and quickly identify potential conflicts. In the meantime, We can make use of server virtualization technique to integrate a large number of scattered and underutilized physical servers to less independent and aggregate physical servers, even make a large network virtual machine to replace thousands of server and make it run under the high utilization in long time, thus bitterly manage IT cost, maximize energy efficiency and advance using rate of resource. We can also make use of storage virtualization technique to support the varied disk storage system in network environment, through integrating the storage capacity to a storage resources pool, help IT system to simplify storage foundation structure, manage the life cycle of information system and maintain business continuity. We also make use of application and desk virtualization technique to provide application infrastructure virtualization function, lower the cost of establish, management and run the application, and achieve the purposes of improving flexibility and agility, ensuring business process integrity, raising application function and bitterly manage running status of the application. In addition, virtualized system management and supervision service can help us detect, monitor and manage all the virtual and physical resources including system and software through a common access point, and provide complete cross-enterprise service management, decrease the amount of managing tool which is used to support various type servers .

3.2 Mass Distributed Storage

In order to ensure high credibility and economy, cloud computing adopts distributed storage to save data, using redundancy storage to ensure the reliability of stored data， and using high credible software to make up the incredibility of the hardware, therefore providing the cheap and credible mass distributed storage and computing system. The data storage system of cloud computing are Google File System (GFS) and Hadoop Distributed File System (HDFS) which is developed Hadoop team.

1) GFS

GFS is a distensible distributed file system. It is used in large and distributed applications which need to access mass data. The designing ideology of GFS is different from the traditional file system, which is designed for dealing large-scale data and the application property of Google. It runs on the cheap and common hardware, but it can provide fault tolerance function. It can provide high-performance service to a great deal of users.

Figure 1 shows the system structure of GFS. A GFS clusters includes master server and many chunk servers, and it can be visited by several clients. The file is spilt into fixed size blocks. When creating a block, the server distributes an unchanged and globally unique 64 handles to identify it. The block server treats the blocks as Linux files and saves them at the local hard disk, read or write block data according to specified handles and byte range. To ensure the credibility, each block will be copied to several block severs, and default saving three copy. The master server manages all metadata of the file system,

including namespace, access control information, the reflecting information from file to blocks and position of the blocks. The code of GFS Client is embedded in each program, realizes Google file system API, which helps the application communicate with master server and block server, and then read or write data. Exchange between client and server is only the operations of metadata, all the data communication directly contact with block server, which consumedly raised the efficiency of system, and keeping master server from overloading.

2) HDFS

HDFS is a distributed file system which is applicable to running on commodity hardware. It is very similar to the existing distributed file system, but also with a significant difference, for example, HDFS is highly fault- tolerant and it can run on the cheap hardware; HDFS can provide data access with high throughput, so it applies to the application of large-scale dataset.

HDFS adopt Master/Slave architecture, a HDFS cluster makes up of a Namenode and several Data nodes. Namenode is a center server which is responsible for managing the file system namespace and the client access to files. Usually a node has a datanode which is responsible for managing storage of the node.

Seeing from the inner part, a file is split into one or more block, which are saved on a set of Datanodes. Namenode implements the namespace operations of file system, for example, open, close, rename file or directory, in the meantime, it is also responsible for determining the mapping from the data block to datanode, an establishing， deletion and replication data piece in namenode.

Unitary of Namenode consumedly simplified the structure of system. Namenode is responsible for preserving and managing all HDFS metadata, reading or writing of user's data is on Datanode not by Namenode.

3.3 Parallel programming model

To enable users efficiently to use cloud computing resources and more easily enjoy services that cloud computing brings about; cloud computing programming model must make task scheduling and parallel execution transparent to users and programmers. Cloud computing adopts MapReduce programming model, which decomposes the task into multiple subtasks, and through two steps (Map and Reduce) to realize scheduling and allocation in the large-scale node.

MapReduce is a parallel programming system developed by Google. It puts parallelism and fault tolerance, data distribution, and load balance in a database, and all the operations of data are summarized in two steps: Map and Reduce. When the Programmer submitted his parallel processing procedures to MapReduce, he just need to definite two functions: Map and Reduce. According to the size of input data and configure information, MapReduce can automatically initialize them to several same Map tasks and Reduce tasks, and then process them using different data blocks by calling Map function and Reduce function.

MapReduce system mainly consists of three modules: client, master and worker. The client is responsible for submitting parallel processing assignments composed by the users to master node; master node will automatically decompose user's task into Map missions and Reduce missions, and delivered to worker nodes; worker nodes request to the master node for the work tasks, at the same time, the distributed file system consisting of many worker nodes will be used for storing input and output data of MapReduce.

MapReduce is mainly used in mass data processing. One of the features of the task scheduling strategy is scheduling priority the task the node which the data belong. This kind of scheduling scheme which is based on data position enables Map tasks to read and process data locally when the worker node which request task saves the data needing to process, thus reduces the network overhead and improve the performance of the system.

3.2 Data management

Cloud computing needs to process and analyze mass and distributed data, therefore, data management technology must be able to efficiently manage large data sets. There are two kinds of data management technology in cloud computing system: BigTable of Google and HBase developed by Hadoop team.

BigTable is based on GFS, Scheduler, Lock Service and MapReduce. Each Table is a multi-dimensional sparse map. Row, column, Tablet and timestamp are the basic elements of BigTable. Tablet is a collection of rows.

Data items in BigTable are ordered according to the sequence of keyword in the dictionary, with each row dynamically delivered to Tablets. Each node manages about 100 Tablets. Timestamp is a 64-bit integer, representing the different versions of data. Column family is an aggregation of several columns, whose granularity decides access authority.

BigTable needs three components to execute: one database which is linked to each client, a master server, and several Tablet servers. The master server is responsible for arranging Tablets to Tablet servers, load balance and garbage collection, etc. Tablet servers are responsible for managing a group of Tablets, processing the requests of read or write, etc.

To ensure the high scalability of data structure, BigTable adopts three-level hierarchical way to store location information, as shown in figure 2.

The first level is chubby file which contains the position information of root Tablet. There is only one root Tablet, which contains all the location information of MetaData tablets; a MetaData contains the position information of many user tables.

When reading data, the client firstly gets the location of root Tablet from chubby file, and it gets the position information of MetaData tablet from root Tablet, then it gets the location of user table which contains the location information of the object data, and then it read the location information of the object data from user table, according to the information, read data from a special location in the server.

4. Conclusions

Cloud computing is a new kind of commercial computing model developed on the basis of grid computing, public computing and SaaS. It can distribute computing tasks to the resources pool consisting of massive computers, enabling different application systems to acquire computing power, storage space and various software services according to needs. The ultimate goal of cloud computing is to provide calculation, services and applications as a public facility for the public, So that people can use the computer resources just like using water, electricity, gas and telephone. Therefore, the enterprises can save many costs purchasing hardware and software. This paper introduces the definition of could computing and its main service patterns, summarizes the characteristics, and focused on the key technologies such as the data storage, data management and programming model.

References

[1] Amazon. Amazon elastic compute cloud (Amazon EC2). 2009. http://aws.amazon.com/ec2/

[2] SANJAY GHEMAWAT; HOWARD GOBIOFF; PSHUN-TAK LEUNG. The Google file system. Proceedings of the nineteenth ACM symposium on Operating systems principles. Oct. 2003

[3] Jinesh Varia. Cloud architectures- Amazon web services [EB/OL]. ACM Monthly Tech Talk , http://acmbangalore.org/events/monthly-talk/may-2008--cloud-architectures---amazon-web-services.html, May, 2008

[4] Sims K. IBM introduces ready-to-use cloud computing collaboration services get clients started with cloud computing. 2007. http://www-03.ibm.com/press/us/en/pressrelease/22613.wss.
[5] Google Docs [URL]. http://docs.google.com/, access on Sep. 2008
[6] IBM Blue Cloud project [URL].http://www-03.ibm.com/press/us/en/pressrelease/22613.wss/,access on June 2008
[7] S. Ghemawat, H. Gobioff, and S. Leung. The Google file system. In *Proceedings of the 19th ACM Symposium on Operating Systems Principles*, pages 29–43, 2003
[8] Aymerich, F.M. Fenu, G. Surcis, S. An approach to a Cloud Computing network. Applications of Digital Information and Web Technologies, 2008
[9] Zhang Weimin, Tang Jianfeng. Cloud Computing. Science Press, 2009

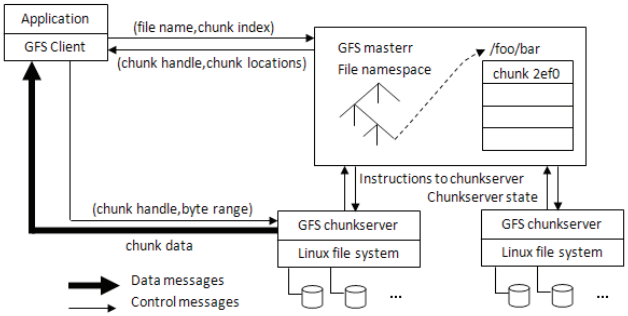

Figure 1. Google File System Architecture

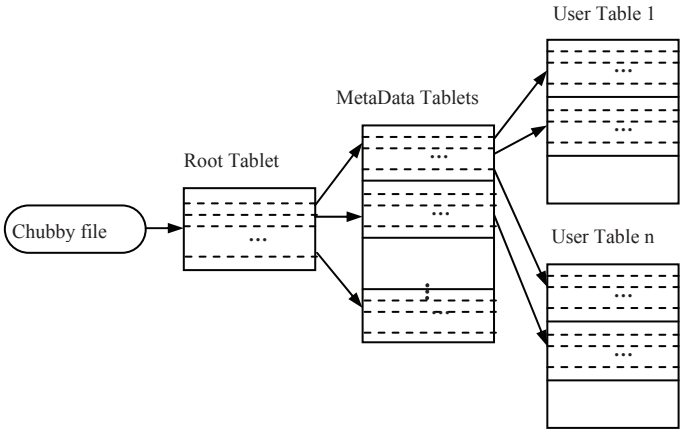

Figure 2.Tthe Structure of Sstoring TabletsLlocation Information