Research on Cloud Computing Based on Deep Analysis to Typical Platforms

Tianze Xia, Zheng Li, and Nenghai Yu

MoE-Microsoft Key Laboratory of Multimedia Computing and Commucation
University of Science and Technology of China
{Tianze Xia,Zheng Li,Nenghai Yu}Zenith@mail.ustc.edu.cn

Abstract. Cloud Computing, as a long-term dream of turning the computation to a public utility, has the potential to make IT industry great changed: making software more charming as a service and changing the way hardware designed and purchased. Along with the rapid development of Cloud Computing, many organizations have developed different Cloud Computing platforms, expressing their different understandings of the Cloud. Based on these facts, this paper has analyzed these understandings, introduced and tested several typical kinds of Cloud Computing platforms, and contrasted among them. The purpose of the study is to give a deep insight to the trend of Cloud Computing technology and to provide reference on choosing Cloud Computing platforms according to different needs.

Keywords: Cloud computing, Hadoop, Enomaly, Eucalyptus, trends.

1 Introduction

With Cloud, developers worked for new Internet services no longer need the large capital outlays in hardware to deploy their service. They do not have to worry about over-provisioning for a service whose popularity does not meet predictions, or under-provisioning for one that then becomes popular. Since the concept of Cloud Computing has been made, it gets rapid development. Not only Google, Microsoft, Amazon, such commercial companies launched their own Cloud Computing products, but there are also many open-source software groups published their attractive platforms.

The first part of this paper has studied the principle and structure of Hadoop then tested. The second part studied and tested two Iaas Cloud platforms—Eucalyptus and Enomaly. Finally some issues about the tendency of the Cloud Computing will be raised in this paper.

2 Studies on Cloud Computing Platforms

2.1 Hadoop

Hadoop is a distributed computing framework provided by Apache[4]. It is a Cloud service similar to PaaS. The core of Hadoop is MapReduce[8] and Hadoop Distributed

M.G. Jaatun, G. Zhao, and C. Rong (Eds.): CloudCom 2009, LNCS 5931, pp. 601–608, 2009.

File System (HDFS).[5][9] The idea of MapReduce is from a Google's paper and is now widely circulated[8]. HDFS provided underying support for distributed storage.

Hadoop is most suitable for mass data analysis. Mass data is divided and sent to a number of nodes. And then by each node's parallel computing the output result will be integrated.

2.2 Eucalyptus

Eucalyptus is developed by University of California (Santa Barbara) for Cloud Computing research[2]. It is compatible with the EC2's API system. Although it supports for Amazon in the syntax of interfaces and achieve the same functionality (with some exceptions), but is almost completely different inside. The design goal for Eucalyptus is easy to expand, and easy to install and maintain. Eucalyptus will become an integrated part of Ubuntu Linux. Ubuntu users will be able to easily use Eucalyptus building private clouds.

2.3 Enomaly's Elastic Computing Platform (ECP)

Enomaly's Elastic Computing Platform (ECP) is an open source web-based virtual infrastructure platform.[3] It is also an IaaS platform. Its design goal is to manage the Distributed Virtual Server environment that is complicated.

2.4 Conclusion

From the introduction of each platform, we can see that Hadoop might be the most proper for processing data; Eucalyptus might be good for Linux users to build their private IaaS Cloud; And ECP may be suitable for IT managers to supervise their Clouds. However, all these are just the official introduction given by each of them, the real features and performances of them need to be dug and evaluated by experiments and tests. In the following parts, we will perform tests to give a deep analysis for each of these platforms.

3 Research on Hadoop

3.1 Architecture of Hadoop

There are many elements in Hadoop. At the bottom there is Hadoop Distributed File System (HDFS)[5].It stores all the data in Hadoop system. On top of HDFS is MapReduce engine that consists of JobTrackers and TaskTrackers.

HDFS using master /slave architecture. HDFS is a cluster of one Namenode and numbers of Datanodes[7]. Namenode is a central server that is responsible for managing the file system namespace, and controls the access of files. From the internal perspective, a file is divided into one or more data blocks, these blocks are stored in a group of Datanodes. Namenode executes the operation of namespace,such as open, close, rename files or directories. It is also responsible for map a certain data block to a certain Datanode. Datanode is responsible for handling file system requests.

Namenode usually runs independently in a machine. It is responsible for the management of file system name space and controls the access to external clients. NameNode decide how to map files to data blocks in Datenodes. The actual I / O stream does not go through it. When an external client sent a request to access a file, Namenode responds a Datanode's IP address that contains the file's copy. Then the Datanode responds for the client's request. This feature of hadoop is very important. It does not move data to a certain location to process, but move processing to data. So use hadoop to process data is very efficient.

3.2 Test on Hadoop

During the test of hadoop, we run WordCount[10] for a pdf text file on different numbers machines. The pdf's size is 300MB. Table 1 shows the results. The advantage of hadoop mentioned above can be clearly seen.

Table 1. The result of Hadoop test

Number of Datanodes	1	3	7	15
Time cost(seconds)	931	511	274	153

Also we do a file write test. The test system has 16 machines (15 machines are configured as datanode) and connected with 1000M Ethernet. The results are shown in Table 2.

Table 2. The result of HDFS test

fileSize (byte)	118147302	12340000	708002	82545	50193
Time cost(ms)	28432	13[illegible]	63	15	16

3.3 Current Deployment of Hadoop and Some Issues

The feature of Hadoop's map/reduce and the HDFS make it very easy to handle vast amounts of data[12]. Because of this and other features like easy to extend, reliable, Yahoo! has chosen Hadoop as its cloud computing platform[6], built the world's largest Hadoop platform—Yahoo! Search Webmap[11]. In addition, Yahoo! And Carnegie - Mellon University launched the Open Academic Clusters-M45 that has more than 500 machines today and has completed many valuable projects. [13]The index of Amazon's search portal—a9.com is also accomplished by Hadoop. The Facebook.com use Hadoop to build the entire site's database,which currently has more than 320 machines for log analysis and data mining.

During the use and the test of hadoop, we found some issues.

a) The performance of hadoop is not stable. Some application might cost different times. This problem makes hadoop OK to process offline data but unsafe to handle real-time tasks.
b) Hadoop is based on Java. This makes it compatible on different systems but limit its performance. In a test from Open Cloud Consortium[10], Sector[18]

which is written in C++ is about twice as fast as Hadoop. To build a Hadoop C++ version is a hopeful way.

4 Research on Eucalyptus and ECP

4.1 Eucalyptus

Eucalyptus will be soon integrated into Ubuntu. Ubuntu users can easily use Eucalyptus building private clouds, just like Amazon Web Services LLC (AWS).And more this private cloud can work together with AWS to create a "composite cloud". [14]
Eucalyptus has three components [17]:

a) Cloud Manager (CM) : The CM is responsible for processing incoming user-initiated or administrative requests, making high-level VM instance scheduling decisions, processing service-level agreements(SLAs) and maintaining persistent system and user metadata.
b) Instance Manager (IM): The IM executes on the physical resources that host VM instances and is responsible for instance start up, inspection, shutdown and cleanup.
c) Group Manager (GM): The GM is responsible for gathering state information from its collection of IMs, scheduling incoming VM instance execution requests to individual IMs, and managing the configuration of public and private instance networks.

Communications between these three components is "SOAP with WS-security". In every cluster there will be only one node operating as Cluster Controller. Every node has a node controller.

4.2 Enomaly's Elastic Computing Platform (ECP)

Compared with Eucalyptus, ECP has these features:

a) A number of server entities can be managed as a virtual cluster
b) Support a wide range of virtualization environments。 It has a long history of complete KVM support.
c) Provide an easy-to-use web management. After deployment, almost all operations can be completed by the web interface.
d) Support python language. That makes the expansion and maintenance are simple.
e) Valet feature is quite handy when building clusters of VM's.
f) Use KVM as the virtualization hypervisor, by VM Creator in ECP, virtual machines that OS is pre-installed can be produced in less than 1 minute.

The specific performance of ECP will be described below.

4.3 Test of Eucalyptus and ECP

Eucalyptus and Enomaly's Elastic Computing Platform (ECP) are also IaaS platforms and both provide the use of virtual machines. So we made a Comparison test between them.

In the test, Eucalyptus adopts Xen and takes ubuntu 8.04LTS as the VM's OS. ECP adopt KVM and Debain 4.0. We also made an original OS that runs on real machine to be the comparison platform. It is fedora Core 6.

Eucalyptus and ECP's VM both have 512MB RAM, single virtual CPU, 10GB system image + 20GB supplementary storage. The comparison platform is also build as this.

Gzip Compression tests the efficiency of virtual CPUs. The shorter time costs the more effective CPU is virtualized.

Lame Compilation tests the cloud VMs' overall performance. The RAMspeed batch copy and batch run tests the cloud VMs' memory speed.

Table 3 to Table 6 gives the results.[16]

Table 3. Gzip Compression Test Result

Platforms	Fedora Core 6	Eucalyptus	ECP
Time (seconds)	67.73	103.30	97.72

This test runs "time gzip –c test.tar > test.tar.gz" The size of test.tar is 801MB.

Table 4. Lame Compilation Test Restlt

Platforms	Fedora Core 6	Eucalyptus	ECP
Time (seconds)	9.24	35	43.27

This test runs "time make –j 5".

Table 5. RAMspeed Batch Copy Test Result

Platforms	Fedora Core 6	Eucalyptus	ECP
Speed(MB/s)	1201.45	1124.36	622.20

This test runs as INTEGER BatchRun Copy.

Table 6. RAMspeed Batch Run Test Result

Platforms	Fedora Core 6	Eucalyptus	ECP
Speed(MB/s)	1345.50	1298.84	783.39

This test runs as INTEGER BatchRun Add.

Test Conclusion:

We can see from the test that the performance of VM that the two IaaS platform provided has a gap between the actual system there. This should be the cost of network communication and VM scheduling. But for these two types of cloud computing platform, the computing power of the VM is similar. In the VM memory speed test, Eucalyptus leads significantly. The reasons for this result may be caused by different virtualization technologies. And it is also possible that because of different mechanisms of these two platforms, such as scheduling. We shall study this problem in future work.

4.4 Comparison of EC2 Eucalyptus and ECP and Some Issues

Finally we compared some key features of Eucalyptus ECP and Amazon EC2. List as Table 7.

Table 7. The comparison of EC2 Eucalyptus & ECP

	EC2	Eucalyptus	ECP
Flow control in data transport	O	O	O
Billing Mechanism	O	X	X
Storage Mechanism	O(s3[15])	O(Walrus)	O
Block Storage Mechanism	O(EBS)	O(Walrus)	O
Load Balance & Live Migration	O(SQS)	X	X
Target Customers	Users	Administrators&Users	Administrators

a) Both Eucalyptus and ECP do not support virtualization technology from commercial company such like VMware or Hyper-V.
b) Eucalyptus will be contained in Ubuntu 9.10. This makes it easy to be employed. However it lacks a method like appzero[19] that can inosculate itself to other public clouds. This can help people in this scene: employ Eucalyptus as a development platform to build applications and then run these applications on AWS seamlessly.
c) The Web management of ECP is a special feature. It is better if the Web system can provide management on not only VMs but virtual networks, virtual applications and storage just by drag and clicks.

5 Predictions and Suggestions

At the basis of research for the developing states on Cloud Computing in different companies, we have made predictions for the possible developing directions of Cloud Computing. What is more, from the experiments and tests results described above, we made an analysis for the performances of the typical existing Cloud Computing platforms, as well as given some suggestions for choosing proper platforms based on different needs.

5.1 Predictions for Trend of Cloud Computing

1. Most clouds today are designed to work just within one data center. An interesting research direction is to develop appropriate network protocols, architectures and middleware for wide area clouds that span multiple data centers.
2. To investigate how different clouds can interoperate. That is, how two different clouds, perhaps managed by two different organizations, can share information. And how the “composite cloud” can be build.
3. A practical question is to develop standards and standard based architectures for cloud services. And thus develop a way to benchmark clouds.

4. Neither of the two IaaS platforms discussed in this paper support VM's Live Migration. When and how to migration VMs from one server to another shall be a key problem.
5. There are lots of open-source Cloud Computing platforms in the level of IaaS and PaaS but few in SaaS. More platforms like 10GEN[1] is needed.
6. Almost all open-source Cloud Computing platforms are based on Linux. A useful research direction is to make open-source platforms to support multiple operating systems such like Microsoft Widows. This can make the Cloud Computing more popular in everyone's life.
7. Nowadays the ability of Cloud Computing cannot fully meet the requirements of entertainment. Image this Scene: a user adopted Amazon EC2 and wants to play 3D games in the VM. Of course the experience he gets will not be good today. The virtualization technologies have not covered GPU yet. This is another charming direction. GPU has more Floating-point computing power than CPU. If GPUs can be virtualized, it is also benefit for Scientific Computing.

5.2 Suggestions for Choosing the Proper Platforms

As shown above, Hadoop suits when the ability of intensive data processing is required, like data mining, Target Recognition on remote-sensing image and so on. Eucalyptus is a good choice when you are using Ubuntu. It makes a simple way to build up private Cloud and can work with Amazon EC2 smoothly. So it also suits to the companies or groups that deployed Amazon EC2 but want to process private data in their own Cloud. Also the Appscale[20] can run on Eucalyptus. This means people who want to deploy Google App Engine can choose Eucalyptus. And for Enomaly's Elastic Computing Platform, Python support and management of servers as virtual clusters are its strengths.

References

1. 10gen, http://www.10gen.com/
2. Rich, W., Chris, G., Dan, N.: Eucalyptus: An Open-source Infrastructure for Cloud Computing, http://open.eucalyptus.com/documents/eucalyptus-slides-wolski-cloud_expo_apr08.pdf
3. Enomaly, http://www.enomaly.com/Product-Overview.419.0.html
4. Hadoop, http://hadoop.apache.org/core/
5. Hadoop DFS User Guide, http://hadoop.apache.org/core/docs/r0.17.2/hdfs_user_guide.html
6. Yahoo! Developer Network Blog (2008), http://developer.yahoo.net/blogs/hadoop/2008/02/yahoo-worlds-largest-production-hadoop.html
7. Doug, Cutting. Hadoop: Funny Name, Powerful Software (2008), http://www.tuxme.com/index2.php?option=com_content&do_pdf=1&id=27470
8. Hadoop Wiki, http://wiki.apache.org/hadoop/PoweredBy

9. Michael, N.: Running Hadoop on Ubuntu Linux (Multi-Node Cluster), http://wiki.apache.org/hadoop/Running_Hadoop_On_Ubuntu_Linux_(Single-Node_Cluster)
10. Open Cloud Testbed, http://www.opencloudconsortium.org/testbed.html
11. Scott, D.: Yahoo's Doug Cutting on MapReduce and the Future of Hadoop (2007), http://www.infoq.com/articles/hadoop-interview
12. Robert, L., Grossman, Yunhong, G.: On the Varieties of Clouds for Data Intensive Computing. IEEE Computer Society Technical Committee on Data Engineering (2009)
13. Open Cloud Consortium (2008), http://www.opencloudconsortium.org
14. Michael, A., Armando, F., et al.: Above the Clouds: A Berkeley View of Cloud Computing, http://nma.berkeley.edu/ark:/28722/bk000471b6t
15. Amazon Simple Storage Service, https://s3.amazonaws.com
16. Ramspeed, http://www.alasir.com/software/ramspeed/
17. Daniel, N., Rich, W.: The Eucalyptus Open-source Cloud-computing System, http://open.eucalyptus.com/documents/nurmi_et_al-eucalyptus_open_source_cloud_computing_system-cca_2008.pdf
18. Sector-Sphere, http://sector.sourceforge.net/
19. Appzero, http://www.trigence.com/
20. Appscale, http://code.google.com/p/appscale/