http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (21): 6086-6089, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.6086.6089

Workload Prediction in Load Balancing and Resource Management System

Qiao Zhang and Chuanchang Liu
State Key Lab of Networking and Switching, Beijing University of Posts and Telecommunications, Beijing, 100876, China

Abstract: Cloud computing is becoming the primary source of computing power, which has the advantage of high availability, high flexibility, low cost, dynamic resource sharing. Workload balancing is necessary in cloud computing, aiming at using resources at the most balanced. However, virtualization technology used in cloud computing will generate a lot of virtual resources, which will easily cause workload imbalance in cloud environment, making some computing nodes overburdened while some are unoccupied. So we propose a workload balancing and resource management system, by workload prediction and resource adjustments, workload in each virtual machine can be better balanced. In this study, we focus on the workload prediction module. Experimental results demonstrate that our prediction module can get relatively accurate prediction results, which can make big contribution to the workload balancing and resource management system.

Key words: Cloud computing, workload balancing, workload prediction

INTRODUCTION

With the rapid development of Internet, our work and life has undergone tremendous changes. The more we rely on network, the higher requirement we call for computing resources, software and hardware resources and service resources. Cloud computing came into being under this circumstance. Cloud computing can handle and store massive data and has the advantage of high availability, high flexibility, low cost, dynamic resource sharing, etc. It deals with computation, software, data access and storage service, which don't require end-user to know the physical al location and configuration of the system.

Cloud Computing can be divided into IaaS (Infrastructure as a Service), PaaS (Platform as a Service), SaaS (Software as a Service). In this study, we use IaaS to provide the required infrastructure as a service. IaaS provides more flexibility to maintain the proper size of resources according to different demand that may be changed dynamically, this flexibility can eliminates the cost of purchasing and maintaining infrastructure for cloud users. By using virtualization technology, cloud computing can divide one physical host into multiple virtual hosts, each virtual host can provide services just like independent host. Virtualization technology has greatly improved hardware utilization, IaaS users can acquire hardware resources according to their real needs and dynamically scale up and down the resources. But at the same time, virtualization generates a lot of virtual resources, which will easily cause workload imbalance in cloud environment, making some computing nodes overburdened while some are unoccupied (Jadeja and Modi, 2012).

Therefore, workload balancing plays an important role in IaaS cloud computing platform, whose goal is to distribute cloud users' requests properly to make sure the quality of service and at the same time, reduce the waste of resources.

In this study, we propose a workload balancing and resource management system in IaaS computing platform. This system is mainly composed of 3 parts: Workload Prediction Module, Workload Balancer, Resource Manager. By communicating between workload Balance Module and Resource Manager, the workload imbalance problem can be better eased, cloud users can also get high-quality service and resource will be used at the best. This study will also focus on a key part of this system--workload prediction. Accurate prediction is the premise of resource management and workload balancing.

RELATED WORK

There have been many workload balancing algorithm in cloud environment, such as round robin, weighted round robin, respond time balancing, least connection balancing and so on. Different algorithms need to be chosen considering of different conditions, such as the type of request, server's capacity, etc. This study will design a more synthesized model to adapt to most of the situations.

Corresponding Author: Qiao Zhang, State Key Lab of Networking and Switching, Beijing University of Posts and Telecommunications, Beijing, 100876, China

We choose OpenStack to be our IaaS experiment platform. OpenStack is a collection of open source technology that provides massively scalable open source cloud computing software to provide scalable, elastic cloud computing for both public and private clouds, large and small. OpenStack is made up of five important parts: Nova (Compute), Swift (Object Storate), Glance (Image service), Keystone (Identity), Horizon (Dashboard). We have deployed some businesses onto OpenStack to do experiments. Mostly, these businesses are social networking tools. Because social networking tool is a trend in the future and this research will help these tools to extend to cloud.

In this study, we are going to describe the model which can handle the problem of workload imbalance. In this model, we want to match different type requests onto different VMs with corresponding capability. This study also presents a key part of this model-workload prediction.

SYSTEM ARCHITECTURE

This section will introduce a workload balancing and resource management system and explain how it works. The architecture of our system is showed as Fig. 1. We have built a controller node and several cloud compute nodes and each node is a physical machine. The controller node runs OpenStack-network, Openstack-scheduler, Openstack-API and some other controlling module, while the cloud compute node runs OpenStack-compute which can support VMs running.

Present system is built in controller node, it includes 2 key parts: Workload balancing module and resource manager module, they are responsible for the request distribution and resource adjustment. Workload Balancing includes two important parts: workload prediction and workload balancer.

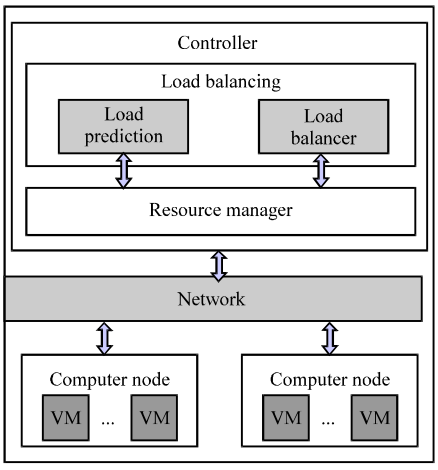

Fig. 1: Architecture of workload balancing and resource management system

Workload prediction: Workload prediction is the foundation of the whole system. It is used to predict the number of requests of different businesses deployed on current cloud platform in one scheduling period. By some study, we find that social networking users' behavior has certain periodicity (Fan *et al*., 2012). This will help to predict the workload in next period. Then workload prediction module can send the prediction results to resource manager.

Resource manager: According to the workload prediction next cycle, resource manager will do some adjustments, scale up or down Virtual Machine (VM) resources. If the workload in next cycle is heavier, resource manager will increase the number of instance, to balance the workload in each instance; While if the workload in next cycle is less than current cycle, resource manager will decrease the number of instance to reduce the waste of VM resources. And it will send a list to workload balancer, which contains information about how businesses are deployed in each instance. So workload balancer can make a decision to better distribute these requests. Whenever the VM list changes, resource manager will update the list to workload balancer.

Workload balancer: When request is coming to workload balancing module, workload balancer will check the VM list which is received from resource manager to see which instance can handle this type of request, then workload balancer will examine these instances' status, including CPU utilization, I/O utilization, Internet condition and so on. Then it will distribute each request to the most suitable virtual machine.

IMPLEMENTATION OF WORKLOAD PREDICTION

Workload prediction plays a very important part in the whole system, which undertakes the responsibilities of predicting the workload in each period, so the resource manger can know how to adjust current VM quantity. In this section, we will introduce a method to predict workload based on historical data.

There are many predicting methods, such as statistical methods, neural networks and prediction models. Our system is to do a research on social network which has its own character, such as time regular pattern

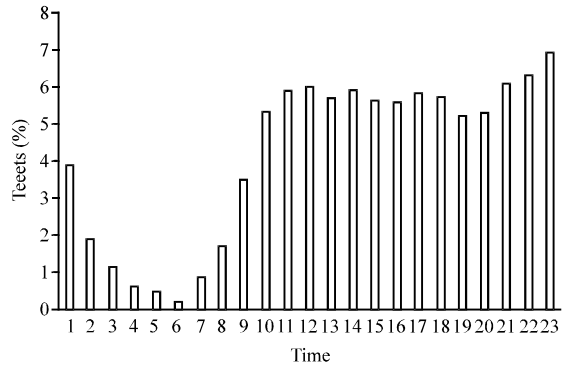

Fig. 2: Daily pattern of tweets in social network

and low-high-stable life cycle. In consideration of all these factors, we choose statistical methods based on historical data to do workload prediction.

Time period distribution: We have collect the time slots of users' behavior in social network (Fan *et al*., 2012). Figure 2 has shown that from 00: 00 a.m. to 10: 00a.m. Less users are using social network, while from 10: 00 a.m. to 21: 00 p.m. more users are using social network and from 21: 00 p.m. to 24: 00 p.m. most users are active.

Besides, there are no obvious weekend effects (Fan *et al*., 2012). So in the workload prediction, we divide time period to several sections as shown in Table 1.

So we are going to predict workload according to this time slot.

Build historical information table: Similar businesses will have the similar workload in the same condition (Zhao *et al*., 2013), so we can predict the workload from prior historical information. These prior historical information is from the research datasets in this field in a long period and is reliable. Then we will update this historical information by updating fresh data constantly. We define the prior historical information as follows: T (A1, A2, A3……An). T is the name of historical information table, A1 to An is the different business names. Each business has some specific properties to define it: A (a1, a2……an). an is used to define each property of one business. Time slot is the time cycle we defined in part A. Workload is to record in current period, the number of requests that visit this type of task. Record frequency is to calculate the number of records that corresponding to this workload, total frequency records the total records of the current task in the same time slot (Li *et al*., 2008).

Using information theory, the probable of task U is P (U), the uncertainty of task U is I (U), then:

Table 1: Time slot in social network

Period	Time slot
Period 1	00: 00 a.m. ~ 01: 00 a.m.
Period 2	01: 00 a.m. ~ 04: 00 a.m.
Period 3	04: 00 a.m. ~ 06: 00 a.m.
Period 4	06: 00 a.m. ~ 08: 00 a.m.
Period 5	08: 00 a.m. ~ 09: 00 a.m.
Period 6	09: 00 a.m. ~ 10: 00 a.m.
Period 7	10: 00 a.m. ~19: 00 p.m.
Period 8	19: 00 p.m. ~ 21: 00 p.m.
Period 9	21: 00 p.m. ~24: 00 p.m.

$$P(U) = \frac{\text{Record times}}{\text{Total times}} \quad (1)$$

$$I(U) = \frac{\log}{1} P(U) \quad (2)$$

We can use uncertainty to describe the reliability of a task (Li *et al*., 2008). So when we do workload prediction, we will choose the most reliable record.

Time prediction strategy: For existing business, we can find clue from the existing records in historical information and use the most reliable records (Zhang and Meng, 2012). For new business, we can find the most similar business in historical information by comparing the specific property value to init the first period's prediction result.

Update historical information: At each end of the task period, new period prediction should be sent to resource manager for further VM adjustment. At the same time, current workload result (which will become the history) should be returned to workload prediction center to update the historical information. There are three conditions:

- This type of task doesn't have records in historical information, then add a new record to the historical information, set probable to 1, set uncertainty to 0
- There exists record of the type of task but the workload deviation is too much between the history record and new record, then add a new record to historical information, recalculate this type of task's workload, total records, probable and uncertainty
- If the new record is consistent with the existing record, that is to say, the workload deviation can be acceptable, then just update the record times to add 1 and update the total records, probable and uncertainty of the relative records. When a record's uncertainty is less than a threshold, this record can be regarded as reliable enough and the other record of this type of task can be ignored

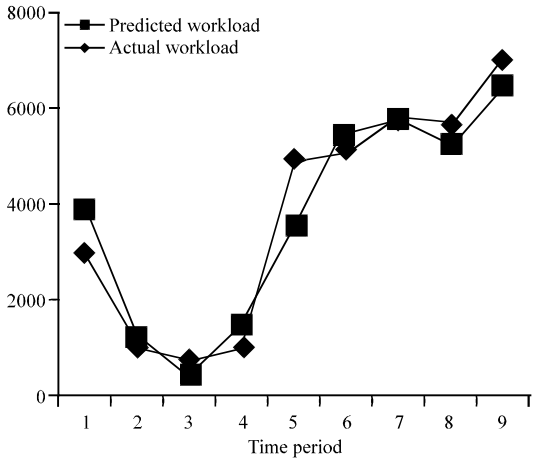

Fig. 3: Comparison of prediction and actual workload of business1

In this way, we can keep a historical information that give clue to the workload next period to do workload prediction

EXPERIMENTS

In this section, we experimentally evaluate the accuracy of the prediction model. We set up one controller node and five compute nodes in a cloud cluster on OpenStack. The hardware configurations of all the compute nodes are almost the same. And we deploy ten different businesses which are social network tools onto these VMs.

In order to simulate real use case, we collect some real datasets of the similar type of businesses. Using these datasets, we simulate 1, 000 users visiting these businesses. Figure 3 shows the comparison of predicted and real scene results of two businesses. The absolute average of prediction deviation is 2.23%, which means this prediction model can get relatively accurate result and well predict business requests of social network businesses.

CONCLUSION

In this section, we experimentally evaluate the accuracy of the prediction model. We set up one controller node and five compute nodes in a cloud cluster on OpenStack. The hardware configurations of all the compute nodes are almost the same. In this experiment, we care about whether the prediction result is near to the real scene condition, so in some extent, we can ignore the performance of the VM. And we deploy ten different businesses which are social network tools onto these VMs.

The key contribution in this study is that it proposes an architecture of workload balancing and resource management system and figure out how to do workload prediction-foundation of this system. The workload prediction can offer an accurate prediction result to resource management part, then this part can do corresponding adjustment to the VMs, scale up or down the resources. Using this system, workload can be predicted, resources will be adjusted according to the pediction result, then workload balancer can distribute requests to appropriate VMs. In this way, the performance of cloud can be improved and become more efficient. In the future, we will do research on resource management and request distribution.

ACKNOWLEDGMENT

This study is supported by the National Grand Fundamental Research 973 Program of China under Grant No. 2011CB302506 and the Technology Development and Experiment of Innovative Network Architecture under Grant No. CNGI-12-03-007.

REFERENCES

Fan, P.Y., H. Wang, Z.H. Jiang and P. Li, 2012. Measurement of microblogging network. J. Comput. Res. Dev., 49: 691-699.

Jadeja, Y. and K. Modi, 2012. Cloud computing-concepts, architecture and challenges. Proceedings of the International Conference on Computing, Electronics and Electrical Technologies, March 21-22, 2012, Kumaracoil, India, pp: 877-880.

Li, Y., L. Meng and L.Y. Xu, 2008. Evaluation and prediction of workload based on historical statistic. Comput. Appl. Software, 25: 25-26.

Zhang, J.Z. and X.F. Meng, 2012. Research on mobile web search. J. Software, 23: 46-64.

Zhao, G.F., B. Li, C. Xu and H. Tang, 2013. Research on lifecycle of mobile social network. Chin. J. Comput., 36: 727-737.