DECISION MAKING SCHEME TO ENSURE POWER AND RESOURCE USAGE FOR SERVICE PROVIDERS

Pradheeba.P[1,], Jeba.N[2], Dr. Rajendran. T[3]

[1.]*PG Scholar, Angel College of Engineering & Technology, Tamilnadu..* prathi.deepa11@gmail.com

[2.]*Assistant Professor, Angel College of Engineering & Technology, Tamilnadu..* jeba.acet@gmail.com

[3.]*Dean, Angel College of Engineering & Technology, Tamilnadu.* rajendran_tm@yahoo.com

Abstract: **Many cloud service providers face problems due to VM migration in the clustered server. Due to over migration of VM there is a loss of power, resource, costs etc. Sometimes clusters may be overloaded, demand of the CPU performance exceeds the available capacity, so a violation of the SLAs established between the resource providers may occur. To reduce these facts decision making schema is proposed in this paper. This method will be used to decide whether to migrate the VM and to switch the cluster in sleep mode or to switch the state of cluster from sleep mode to active mode. This decision is made by analyzing the current load and future load that is predicted and displayed. In this paper all the loads are being monitored and updated in the database and displayed to the service providers for decision making.**

Keywords: Clusters, Load Prediction, Virtual Machine Migration, Decision Making.

1.INTRODUCTION

Despite of all the hype surrounding the cloud service providers are still reluctant to manage the cloud. Migration is one of the major issues which increase the cost, resource, power etc., for cloud service providers and complications with VM migration within cluster and to another cluster continue to plague for service providers. Cloud service users need not to be vigilant in understanding the risks of migration. Here we study a different problem: how can a cloud service provider optimize the power, cost and resource? It is up to the cloud provider to make sure the underlying physical machines (PMs) have sufficient resources to meet their needs. VM live migration technology makes it possible to change the mapping between VMs and PMs, while applications are running [3], [8]. However, a policy issue remains as how to decide the mapping adaptively so that the resource demands of VMs are met while the number of PMs used is minimized. This is challenging when the resource needs of VMs are heterogeneous due to the diverse set of applications they run and vary with time as the workloads grow and shrink.

Energy efficiency can be achieved at different levels - computation, data processing, power distribution at the rack level and server level, power generation and transmission etc. By adaptively predicting future loading on the cloud and dynamically enabling the precise number of machines to turn on, we target higher 80-90% average loading across the entire set of active cloud computing nodes. This allows us to place many unloaded processors in low power sleep mode. The load prediction algorithms determine the joint allocation of high and low power processors.

To realize these promises, cloud providers need to be able to quickly plan and provision computing resources, so that the capacity of the supporting infrastructure can closely match the needs of new applications or computing tasks. This prediction can be applied separately to the different SaaS and PaaS session types as well as to a combined total session count, which appears as the number of Web session requests to a cloud platform. Other resource measures such as the number of VMs on demand at any time, amount of RAM, disk space, CPU or network bandwidth may also be modeled using this approach, for load prediction on clouds.

Live migration (Clark et al., 2005) allows a virtual machine to be migrated from one physical server to another, without interrupting the application running in that virtual machine [10]. Ideally, layout of virtual machines on physical servers can be dynamically adjusted with live migration to a state that Service Level Agreement (SLA) is always satisfied as long as there are idle resources available in any physical server in the system. When application load may get high enough that no matter how the hypervisor allocates the resources to virtual machines, there are always some applications that cannot get enough resource to achieve acceptable performance regarding to SLA.

Live migration, however, incurs network overhead because it involves transferring the memory image along with other states of a virtual machine from one server to another. When network resource is busy, using live migration may make the situation worse. In addition, migration may last for an uncertain period of time depending on network traffic; therefore it is unwise to incorporate migration for a transient overload. One of the main drawbacks of migration where cloud service providers face problems due to in the clustered server is that due to over migration of VM there is a loss of power, resource, costs etc. Sometimes clusters may be overloaded, demand of the CPU performance exceeds the available capacity, so a violation of the SLAs established between the resource providers may occur. To reduce these facts decision making schema is proposed in this paper.

2.RELEATED WORKS

In this section, we first review related works addressing the power management, VM Live migration and prediction methods in the cloud.

Recently, Andreolini et al (2009) presented [9] an overview of dynamic management of virtualized app environments, for cloud applications. This work focused on supporting VM migration decisions in a cloud environment, to answer questions such as which VMs should be migrated and when, using an algorithm which does not depend on the instantaneous behavior or average trends but uses load trend behavior, a better method to avoid false (inaccurate) prediction or alarms. For example, if the CPU utilization more than 85% for only 2-3 seconds, no replacement measures are triggered. This work is not directly relevant to the present cloud workload characterization, which is a prerequisite input needed for such VM migration decisions.

Kumar et al. [5] have proposed an approach for dynamic VM consolidation based on an estimation of “stability” – the probability that a proposed VM reallocation will remain effective for some time in the future. Predictions of future resource demands of applications are done using a time-varying probability density function. The problem is that the authors assume that the parameters of the distribution, such as the mean and standard deviation, are known prior. They assume that these values can be obtained using offline profiling of applications and online calibration. However, offline profiling is unrealistic for IaaS environments. Moreover, the author assume that the resource utilization follows a normal distribution, whereas numerous studies [1], [4], [7] have shown that resource usage by applications is more complex and cannot be modeled using simple probability distributions.

Kusic et al. [6] have defined the problem of power management in virtualized heterogeneous environments as a sequential optimization and addressed it using Limited Look-ahead Control (LLC). The objective is to maximize the resource provider’s profit by minimizing both power consumption and SLA violation.

Cardosa et al. [2] have proposed an approach for the problem of power-efficient allocation of VMs in virtualized heterogeneous computing environments. They have leveraged the min, max and shares parameters of Xen’s VMM, which represent minimum, maximum and proportion of the CPU allocated to VMs sharing the same resource. However, the approach suits only enterprise environments as it does not support strict SLAs and requires the knowledge of application priorities to define the shares parameter.

From the literature review, it is clear that algorithms for load prediction, critical to the success of cloud computing, are not simple, deterministic functions of raw resource measures over time.

3.DESIGN AND IMPLEMENTATION

A. System Architecture

According to Figure 1 our system architecture mainly composed of Cloud Service Providers, Users, Clusters, Management System and Database. In the cloud the Virtual Machines are grouped based on their categories inside various servers to form n number of clusters.

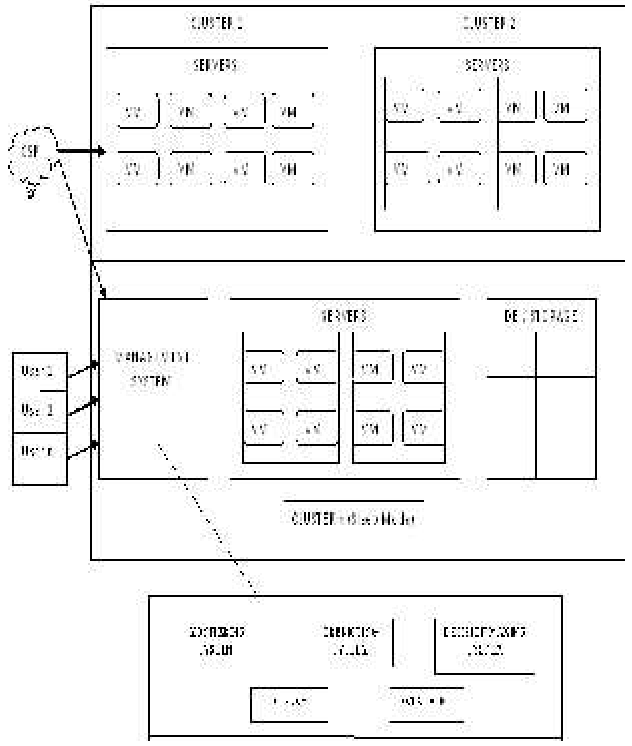

Figure 1 System Architecture

Each user’s applications will be running in the VM of clusters. Some clusters are in sleep mode which are activated when the load and utilizations for users increases. VM migration takes place when the load is minimum. For this the total utilization of entire cloud, individual users and clusters should be monitored and tabulated. All those information’s are stored in the database. Decision of activation of new cluster from sleep mode to active mode or to migrate the VM of minimum running load clusters to another cluster and put the entire cluster in sleep mode. In this paper the management system is the focusing component in which the decision making schema is proposed.

B. Proposed Model

In this paper our proposed schema is the Management system which consists of Monitoring System, Predicting System, Decision Making System, Display System and Database System. This schema is used for the benefit of Cloud Service Providers in cost reduction, power efficiency, minimized VM migration and efficient resource usage. The steps involved in these activities are given in data flow follows:

Step 1: Initially we have to monitor all the parameters and utilizations such as Processor utilization, RAM utilization, Disk/Storage utilization, and Network utilization etc., for the whole Physical Machine & Virtual Machine of individual users in cloud and the Total utilizations for individual clusters.
Step 2: The monitoring information’s that is collected, should be tabulated and stored in the database.
Step 3: Then the current usage of the individual cluster and users should be calculated for all the parameters and stored in the database.
Step 4: After that, the future load is predicted, in which the periodic load for next 8 hours is being predicted and stored in the database.
Step 5: A threshold level is fixed for every cluster in the cloud. Now compare the load of current usage with future usage.
Step 6: If the future load is high i.e. overloaded, then the decision of activating the new cluster which is in sleep mode is taken by the admin in Cloud Service Providers. For that the specific cluster should selected for activation as per the utilizations requires for future or claim to another Service Provider.
Step 7: If current load is below threshold then the decision for migration of VM, of low running application in a cluster to another and make the entire in sleep mode. For migration the target cluster is been identified.
Step 8: These decisions are taken by the Admin in the Cloud Service Providers from the information’s that is displayed. Frequently collected data’s are stored in the database about the current and future load

These steps are clearly explained in the figure 2.

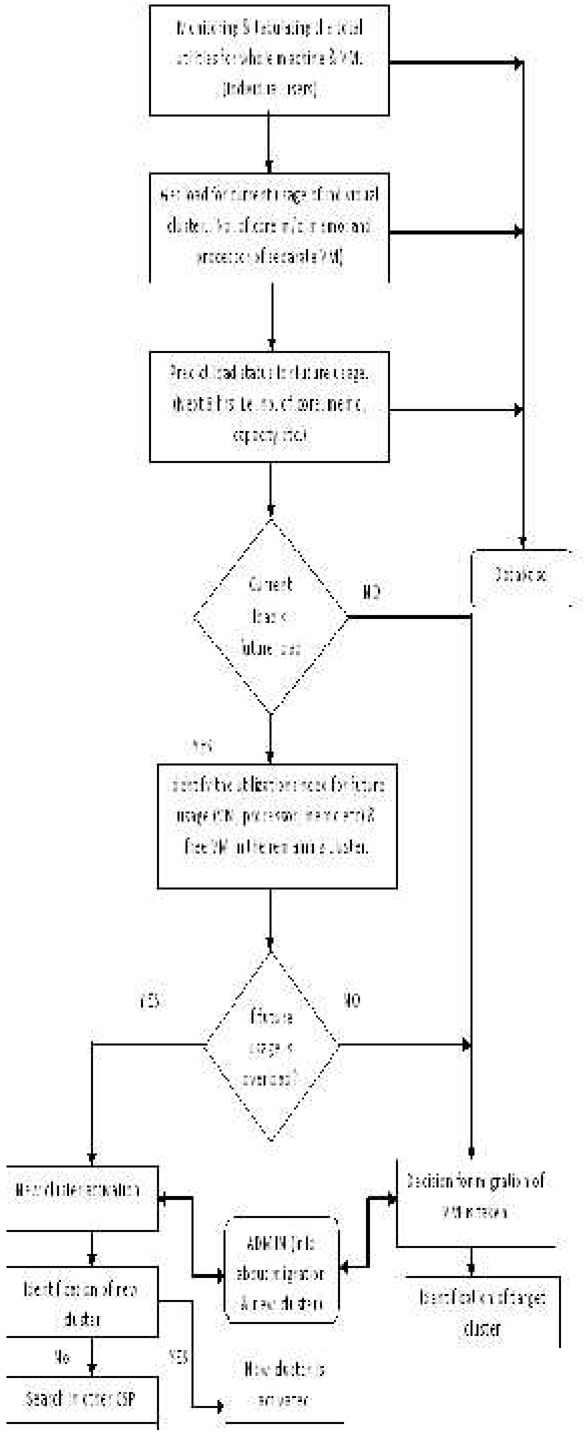
Monitoring & Tabulating the total utilities for whole machine & VM. (Individual users)
Net load for current usage of individual cluster. (No. of cores utilization and processor of separate VM)
Predict load status for future usage. (Next 8 hrs i.e. no. of cores, memory, capacity etc.)
Database
Current load < future load
NO
YES
Identify the utilization needs for future usage (CPU, processor, memory etc) & Free VM in the remaining cluster.
Future usage is overloaded?
YES
NO
New cluster activation
ADMIN (Info about migration & new cluster)
Decision for migration of VM is taken
Identification of new cluster
Identification of target cluster
No
YES
Search in other CSP
New cluster is activated

Figure 2 Data flow

A.. *Prediction Method*

In this project, the important section is predicting the future load for each parameter in clusters. We found that two categories of load prediction algorithm. One category composed of variations of the Exponentially Weighted Moving Average (EWMA) algorithm [12]. It is designed based on the assumption that the future value of a random variable has strong relation to its recent history. Algorithms of the other category adopt the auto-regressive (AR) model [11]. It requires more computation than EMWA based algorithms. But it can incorporate periodicity, which is hard to be utilized in EWMA alternatives, for better precision. The following table shows the monitoring information's which is used to predict future load.

	Cluster 1	Cluster 2	Cluster 3
Number of servers	16	12	18
CPU	8 core	12 core	32 core
Memory	16 GB	32 GB	64 GB
Network	1 GB	1GB	1GB
Threshold	80%	70%	90%

Table 1 Monitoring Cluster Utilities

EWMA

With the original EWMA (Exponentially Weighted Moving Average), load at time t is calculated by:

$$E(t) = \alpha_{*}\, O(t)+(1-\alpha)_{*}\, E(t-1),\ 0 \leq {}_{*} \leq 1$$

where $E(t)$ and $O(t)$are the estimated and the observed load at time t, respectively. The parameter alpha reflects a trade off between stability and responsiveness. We measure the load every minute and predict the load in the next few hours.

For example, when we see a sequence of $O(t) = 10;$ 20; 30; and 40, it is reasonable to predict the next value to be 50. Unfortunately, when α is between 0 and 1, the predicted value is always between the historic value and the observed one. To reflect the "acceleration", we take an innovative approach by setting α to a negative value. When $-1 \leq \alpha < 0$, the above formula can be transformed into the following:

$$E(t) = -|\alpha|_{*}\, E(t - 1) + (1 + |\alpha|)_{*}O(t) = O(t) + |\alpha|_{*}\,(O(t) - E(t - 1))$$

Hence, we use two parameters, ↓α and ↑α to control how quickly $E(t)$ adapts to changes when $O(t)$ is increasing or decreasing, respectively.

AR Model

In some works, future load is modeled as a linear function of several other factors such as the load history, time, or resource allocation. The parameters can be calculated by training with data in the past. Then the model can predict the future load. This methodology is called Auto-Regression (AR), represented as AR (p), where p is the number of factors considered in this model. AR model works well for periodical load.

In this model, the current value of the time series process y (t) is expressed as a linear combination of its previous values [y (t−1), y(t−2),……] and a random noise a(t). The order of this series depends on the oldest previous value at which y (t) is regressed on.

An auto-regressive process of order , AR (p), can be written as:

$$y(t)= {}_{1}y(t-1)+\varnothing_{2}y(t-2)+..+\varnothing_{p}y(t-p)+a(t) \qquad (1)$$

By introducing the backshift operator β the defines $y(t-1)=\beta y(t)$, and consequently $y(t-m)=\beta^{m}y(t)$, Equation (1) can be expressed as:

$(\beta)y(t)=a(t)$ (2) where, $(\beta)=1-\varnothing_{1}\beta-\emptyset_{2}\beta^{2}\cdots-\varnothing_{p}\beta^{p}$

It models the load at time t as a linear function of the average of n latest observations. The load is modeled as a linear function of six independent variables, two of the most recent observations and four of the observations at the same time in last four weeks.

4. CONCLUSION

In this research, we propose a decision making schema which is an efficient work for the benefit of the Cloud Service Providers. The monitoring function views the current load trend and prediction algorithm for predicting the future load under real time constraint. With that result comparison is made and the decision is displayed to the admin of the service provider. This approach is suitable to support different decision systems on cloud platforms, even for highly variable workloads, and is characterized by a computational complexity that is compatible to run-time decisions.

REFERENCES

[1] Barford. P, Crovella. M., "Generating representative web workloads for network and server performance evaluation", *ACM SIG METRICS Performance Evaluation Review 1998; 26(1):151 160.*

[2] Cardosa. M, Korupolu. M, Singh. A." Shares and utilities based power consolidation in virtualized server environments", *Proceedings of the 11th IFIP/IEEE Integrated Network Management (IM 2009), Long Island, NY, USA, 2009.*

[3] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, “Live migration of virtual machines,” in *Proc. of the Symposium on Networked Systems Design and Implementation (NSDI’05)*, May 2005.

[4] Feitelson. D. G.”Workload modeling for performance evaluation”, *Lecture notes in computer science 2002; 2459:114– 141.*

[5] Kumar S, Talwar V, Kumar V, Ranganathan P, Schwan K, “ Loosely coupled platform and virtualization management in data centers”, *Proceedings of the 6th international conference on Autonomic computing (ICAC 2009), Barcelona, Spain, 2009; 127–136.*

[6] Kusic D, Kephart JO, Hanson JE, Kandasamy N, Jiang G. “Power and performance management of virtualized computing environments via lookahead control*”, Cluster Computing 2009; 12(1):1–15.*

[7] Li H. “Workload dynamics on clusters and grids”, *The Journal of Supercomputing 2009; 47(1):1–20.*

[8] M. Nelson, B.-H. Lim, and G. Hutchins, “Fast transparent migration for virtual machines,” in *Proc. of the USENIX Annual Technical Conference*, 2005.

[9] M. Andreolini, S. Casolari, M. Colajanni, M. Messori, "Dynamic load management of virtual machines in a cloud architecture", *Proc of First Int. Conference on Cloud Computing (ICST CLOUDCOMP2009), Munich, Germany, Oct. 2009.*

[10] Sanjay Kumar; Vanish Talwar; Vibhore Kumar; Parthasarathy Ranganathan; Karsten Schwan, “Vmanage: loosely coupled platform and virtualization management in data centers”, *In ICAC '09: Proceedings of the 6th international conference on Autonomic computing, ACM, New York, NY, USA, 2009; pp. 127–136.*

[11] Weijia Song and Zhen Xiao, “An Infrastructure -as-a-Service Cloud: On-Demand Resource Provisioning”, *2013 IGI Global copyright.*

[12] Zhen Xiao, Weijia Song, and Qi Chen (2012), ” Dynamic Resource Allocation using Virtual Machines for Cloud Computing Environment”, *IEEE Transaction On Parallel And Distributed Systems,* vol. n, no. n, month year.

BIOGRAPHY

Ms. PRADHEEBA.P completed her Bachelor of Engineering from Nehru Institute of Engineering And Technology affiliated to Anna University, Coimbatore. She is currently pursuing her Master of Engineering in Angel College of Engineering And Technology affiliated to Anna University, Chennai. Her area of interest is Networking and Cloud Computing. She has presented papers in various National conferences and International conferences on emerging technologies. She has participated in many seminars, workshops, national conference and international conferences on various topics. She is a active member in IAENG. She has published paper in IEEE Explore and in International Journal of Internet Computing.

Ms.JEBA. N completed her Bachelor of Engineering and worked as a Project Trainee for several months. She completed her Master of Engineering and currently working as an Assistant Professor in Angel College of Engineering And Technology. She has presented two papers in International conference and three papers in National conference. Her area of interest is Mobile Ad-Hoc Network.

Dr. RAJENDRAN. T completed his PhD degree in 2012 at Anna University, Chennai in the Department of Information and Communication Engineering. Now he is working as a Dean for Department of CSE & IT at Angel College of Engineering and Technology, Tirupur, Tamilnadu, India. His research interest includes Distributed Systems, Web Services, Network Security, SOA and Web Technology. He is a life member of ISTE & CSI. He has published more than 51 articles in International/ National Journals/Conferences. He has visited Dhurakij Pundit University in Thailand for presenting his research paper in International conference. He was honored with Best Professor Award 2012 by ASDF Global Awards 2012, Pondicherry.