Virtualization Technology to Allocate Data Centre Resources Dynamically Based on Application Demands in Cloud Computing

Namita R. Jain, PG Student, *Alard College of Engg & Mgmt.*, Rakesh Rajani, Asst. Professor, *Alard College of Engg & Mgmt.*

Abstract— Cloud computing is on demand service as it offers dynamic, flexible and efficient resource allocation for reliable and guaranteed services in pay-as-you-use manner to the customers. In Cloud computing multiple cloud users can request number of cloud services simultaneously, so there must be a facility provided such that all resources are obtainable to requesting user in efficient, well organised and proper manner to satisfy their need without compromising on the performance of the resources. Cloud computing has its era and become a new age technology that has got huge importance and potentials in enterprises and markets. Clouds can make it possible to access applications and associated data from anywhere, anytime. One of the major challenges in cloud computing is related to optimizing the resources being allocated. The other challenges of resource allocation are meeting customer demands, data center management and application requirements.

Here the design, implementation, and evaluation of a resource management system for cloud computing services are presented. The system multiplexes virtual to physical resources adaptively based on the changing demand. The skewness metric is used to combine Virtual Machines (VMs) with different resource characteristics appropriately so that the capacities of servers are well utilized. The algorithm helps to achieve both overload avoidance and green computing for systems with multi resource constraints.

Index Terms— Cloud Computing, Green Computing, Overload Avoidance, Resource Management, Virtual Machine Monitor, Virtualization

*****

I. INTRODUCTION

CLOUD computing is widely used as a vital tool for utility computing services. Cloud computing is defined as the use of computer technology along with internet. This computing is dynamically scalable with virtualized resources provided as a service over the internet. In it user must not have information of the tools or infrastructure that supports them. Various providers already has Cloud Computing solutions available, where a group of virtualized resources and dynamically scalable computing authorities (such as power, storage, platforms, and services) are delivered on demand to requested clients over the Internet in a pay-as-you-use manner. For that the large Data Centers are created and in it 'n' numbers of servers are located. Clients have to decide whether to choose private cloud (where cloud infrastructure is worked exclusively for specific business and managed by the third party) or public cloud (i.e. cloud infrastructure is available to the general public and is owned by an organization selling cloud services). The cloud computing deployment models include three main services that are, Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The modern data centers are accommodating a variety of applications. The series of application can be changed from the smallest to the highest. The data center resources are allocated to applications, based on peak load characteristics, in order to maintain isolation and provide performance securities. But one of the major challenges in cloud computing is resource optimization [1].

In cloud computing dynamic resource allocation is widely used concept since last few years. They indulged in it with innovative concepts, new ways and practices to face this type of experiments. Since data centers host multiple applications on a common server platform; they can dynamically reallocate resources among different applications.

Namita Jain is a PG student of ALARD College of Engineering, affiliated to Pune University. Her first review paper was published in International Journal of Computer Science and Management Research eETECME, ISSN 2278-733X dated October 2013. . Her second paper was published in International journal of software & hardware research in Engg., ISSN No: 2347-4890 Vol. 2 Issue 1, dated Jan. 2014.
Rakesh Rajani is Asst. Professor with ME Department of Computer Science, ALARD College of Engineering, affiliated to Pune University. (e-mail:rakeshom123@rediffmail.com).

II. LITERATURE SURVEY

Zhen Xiao, Weijia Song, and Qi Chen [1] develop a resource allocation system, which is used to prevent the overloading in the system. The concept of skewness metric is also introduced in it. By decreasing the skewness, they improve the overall utilization of the servers.

Liang-Teh Lee, Kang-Yuan Liu, Hui-Yang Huang and Chia-Ying Tseng [2] & Anton Beloglazov and Rajkumar Buyya [3] introduces various mechanisms for energy saving. In it, many efforts have been made to reduce the energy consumption in the data centers & to fulfill the concept of green computing. The dynamic Voltage Frequency Scaling (DVFS) is used to adjust the CPU power according to its load. But [1] does not use the DVFS to support green computing.

In [3] different allocation policies are introduced. They propose three stages of VM placement that are reallocation according to current utilization of multiple system resources. Second, optimization of virtual network topologies and third is VM reallocation using thermal state of resources.

J.S. Chase, D. C. Anderson, P.N. Thakar, A. M. Vahdat, and R. P. Doyle [5], investigates the policies for allocating resources in a hosting center, with a principal focus on energy management. This present the design and implementation of a flexible resource management architecture called Muse. The server distribution is controlled and the requests are routed properly to particular servers through a reconfigurable interchanging framework. A simple adaptive strategy is demonstrated to vigorously allocate adequate resources for each service. It is used to help its existing contents i.e. loads to be avoided from over provisioning.

P. Braham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R.Neugebauer [6], concentrated on virtual machine monitor like Xen. VMM like x86 allows multiple operating

systems to share conventional hardware in a harmless and resource managed approach, but without losing either performance or functionality. This is possible by virtual machine abstraction to which operating systems such as Linux, Windows XP etc. can be ported effortlessly.

S. Genaud and J. Gossa, "Cost-wait trade-offs in client-side resource provisioning with elastic clouds," finds that recent Infrastructure-as-a-Service offers, such as Amazon's EC2 cloud, provide virtualized on-demand computing resources on a pay-per-use model. From the user point of view, the cloud provides an inexhaustible supply of resources, which can be dynamically claimed and released. This drastically changes the problem of resource provisioning and job scheduling. This article presents how billing models can be exploited by provisioning strategies to find a trade-off between fast/expensive computations and slow/cheap ones for indepedent sequential jobs. We study a dozen strategies based on classic heuristics for online scheduling and bin-packing problems, with the double objective of minimizing the wait time (and hence the completion time) of jobs and the monetary cost of the rented resources.

M. Mao, J. Li, and M. Humphrey, "Cloud auto-scaling with deadline and budget constraints", said clouds have become an attractive computing platform which offers on-demand computing power and storage capacity. Its dynamic scalability enables users to quickly scale up and scale down underlying infrastructure in response to business volume, performance desire and other dynamic behaviors. However, challenges arise when considering computing instance non-deterministic acquisition time, multiple VM instance types, unique cloud billing models and user budget constraints. Planning enough computing resources for user desired performance with less cost, which can also automatically adapt to workload changes, is not a trivial problem. In this paper, we present a cloud auto-scaling mechanism to automatically scale computing instances based on workload information and performance desire. Our mechanism schedules VM instance startup and shut-down activities.

Y. Zhang, G. Huang, X. Liu, and H. Mei, "Integrating resource consumption and allocation for infrastructure resources on-demand," investigates infrastructure resources on-demand requires resource provision (e.g., CPU and memory) to be both sufficient and necessary, which is the most important issue and a challenge in Cloud Computing. Platform as a service (PaaS) encapsulates a layer of software that includes middleware, and even development environment, and provides them as a service for building and deploying cloud applications. In PaaS, the issue of on-demand infrastructure resource management becomes more challenging due to the thousands of cloud applications that share and compete for resources simultaneously. The fundamental solution is to integrate and coordinate the resource consumption and allocation management of a cloud application. The difficulties of such a solution in PaaS are essentially how to maximize the resource utilization of an application, and how to allocate resources to guarantee adequate resource provision for the system. In this paper, we propose an approach to managing infrastructure resources in PaaS by leveraging two adaptive control loops: the resource consumption optimization loop and the resource allocation loop. The optimization loop improves the resource utilization of a cloud application via management functions provided by the corresponding middleware layers of PaaS.

III. PROPOSED SYSTEM

A. System Architecture

In this paper, we present the design and implementation of an automatic, computerized resource management system. Fig. 1 shows the system architecture of dynamic resource allocation using virtual machine in cloud computing. The Resource Allocator is used to allocate the resources (CPU, Memory, storage, I/O etc.). The Load Predictor is used to predict the upcoming resource demands of virtual machines. Based on the past statistics (Result, information) the future load of physical machines is also predicted by load predictor. Load prediction algorithm will be explained in the next section. The skew analyzer is used to compute the inequality or unevenness in the deployment of several resources on a server. In other terms skew analyzer is used to improve the complete utilization of server resources. The use of Hot-Spot monitor is to check whether the overall resource utilization of any physical machine is exceeding hot threshold or not. If it exceeds, then some virtual machines running on it will be transferred properly, to reduce the load of physical machine. The Virtual Machine Monitor such as Xen is used to provide a mechanism for mapping virtual machines to physical resources. But the main issue arises in the system is that how it will come to know whether the mapping is done properly or not. And the question also arises that how to resolve the mapping properly such that not only the resource demands of virtual machines are fulfilled but also the number of physical machines use is minimized. The VM Handler & Manager is used when the resource requirements of VMs are heterogeneous and unrelated due to the different sets of applications they run and differ with time as the loads grow and shrink.

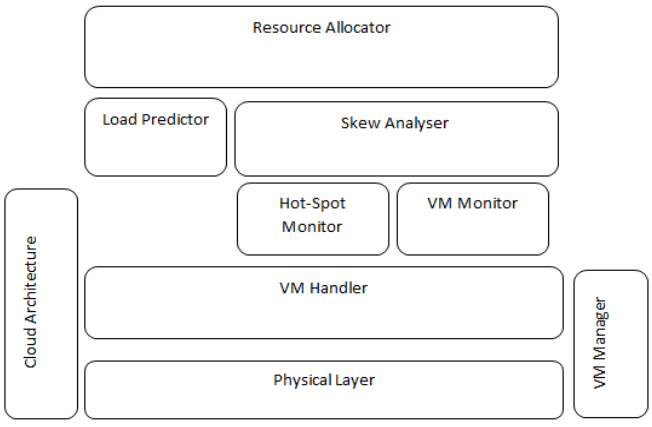


FIG.1. SYSTEM ARCHITECTURE

 ISSN: 2321-8169
1929 – 1933

B. Algorithms Used

Here the aim is to use the skewness algorithm, load prediction algorithm and the concept of green computing, illustrated as follows-

I. Skewness Algorithm

Skewness is termed as an extent of the asymmetry or unevenness of the probability distribution where positively skewed or negatively skewed distribution may be used. The concept of skewness is presented to figure out the inequality in the utilization of numerous resources on a server.

Skew Analyser is used to detect the skewness [1] by following formula:-

$$skewness\ (p) = \sqrt{\sum_{i=1}^{n}(\frac{ri}{\bar{r}} - 1)^2} \quad (1)$$

Where, r is the average utilization of all resources for server p

Utilization of VM Ri= (utilized Memory + utilized CPU)/2
Total Utilization R = ∑ ri

Figure 2 illustrates the effect of skewness algorithm.

II. Hot Spot Algorithm

The algorithm is periodically executed to check the status of resource allocation by evaluating and predicting the future resource demands of virtual machines. Here if the consumption of any resources is beyond the hot threshold, then some virtual machines which are running on it are migrated away.

Following formula is used to check the hot threshold [1]-

$$temparature\ (p) = \sum_{r \in R}(r - rt)^2 \quad (2)$$

Where, R is the set of overloaded resources in server p and rt is the hot threshold for resource r.
If (temperature > hot_ threshold) then
Send request to VM handler to migrate VM

III. Load Prediction Algorithm

As its name indicates this algorithm is used to evaluate the upcoming resource usages of requests (applications) accurately without actually observing inside the VMs. In simple terms it is used to predict the future resource load and resource needs of VMs. It is possible by observing past logs generated and estimating the future load.

Load prediction has noteworthy impact on resource allocation technique. Because if the system has an over-estimated load, then virtual machines may allocate additional resources than essential, due to that some of the resources are wasted. On the other hand, if the system has an under-estimated load, then resource allocation may be inadequate. So, it is important that the system neither should be overloaded nor under loaded otherwise this situation can reduce the performance of its VMs.

Load Prediction Algorithm:

1. For All PMs in Cloud calculate
2. For Each VM in PM calculate, predicted Load as
   $E(t) = \alpha * E(t-1) + (1-\alpha) * O(t),\ 0 \leq \alpha \leq 1,$
   $E(t) = -|\alpha| * E(t-1) + (1+|\alpha|) * O(t)$
   $E(t) = O(t) + |\alpha| * (O(t) - E(t-1))$

Where E (t-1) is expected Load and O (t-1) is observed load at time t-1. We use no. of Jobs at time t-1 as Observed load O (t-1) and estimate load at time t where $0<\alpha<1$

IV. Green Computing

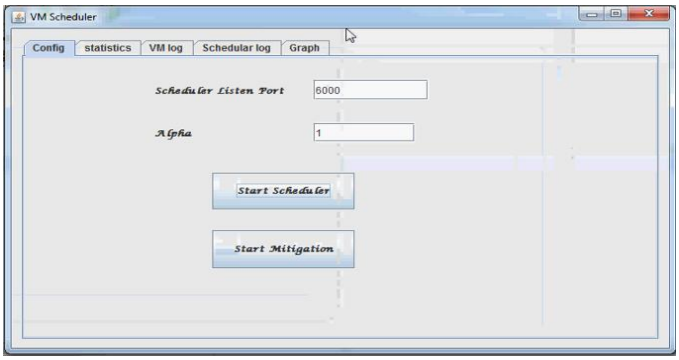
When the resource utilization of active servers is too low, some of them can be turned off to save energy. This is handled in our green computing algorithm. The challenge here is to reduce the number of active servers during low load without sacrificing performance either now or in the future. We need to avoid oscillation in the system.

Our green computing algorithm is invoked when the average utilizations of all resources on active servers are below the green computing threshold. We sort the list of cold spots in the system based on the ascending order of their memory size. Since we need to migrate away all its VMs before we can shut down an under-utilized server, we define the memory size of a cold spot as the aggregate memory size of all VMs running on it. Recall that our model assumes all VMs connect to share back-end storage. Hence, the cost of a VM live migration is determined mostly by its memory footprint. The complementary file explains why the memory is a good measure in depth. We try to eliminate the cold spot with the lowest cost first.

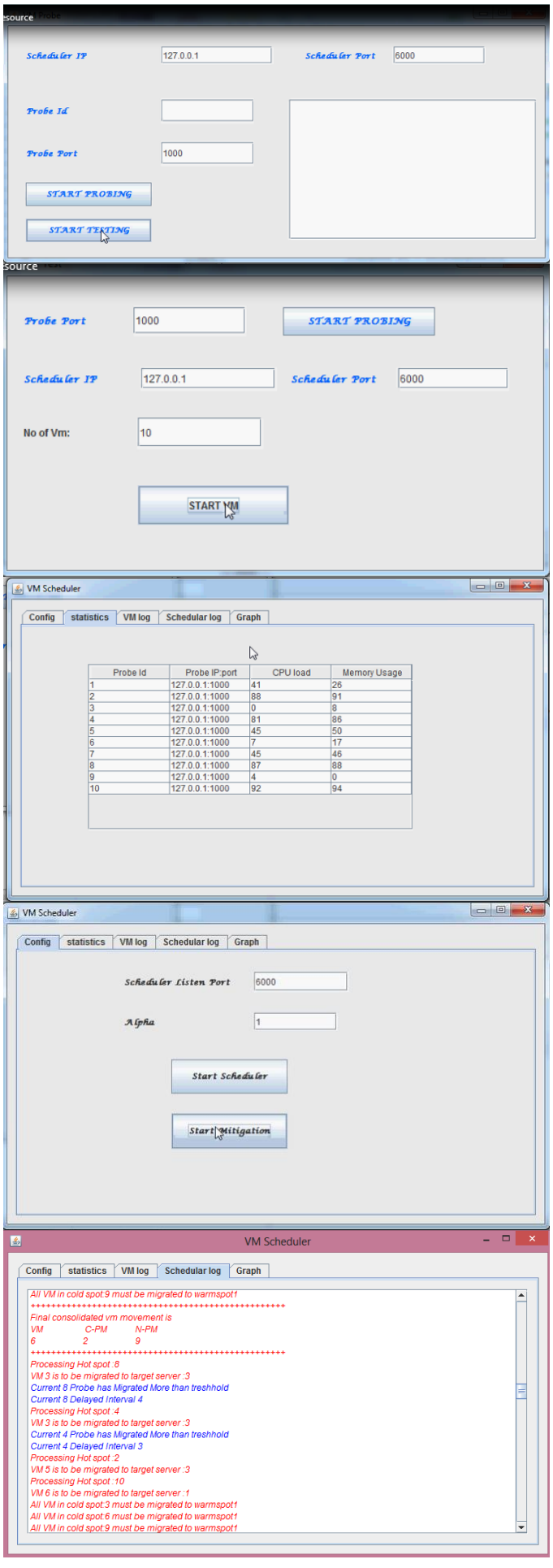
V. Consolidated movements

The movements generated in each step above are not executed until all steps have finished. The lists of movements are then consolidated so that each VM is moved at most once to its final destination. For example, hot spot mitigation may dictate a VM to move from PM A to PM B, while green computing dictates it to move from PM B to PM C. In the actual execution, the VM is moved from A to C directly.

IV. Results



ISSN: 2321-8169
1929 – 1933



V. CONCLUSION & FUTURE SCOPE

Cloud Computing is a grace of computing where, dynamically accessible and often virtualized resources are provided as a service over the internet. Dynamic resource allocation is emergent and growing need of cloud providers. It is useful in cloud environment for more number of users and with the less response time. Recent computers are sufficiently powerful to use virtualization to present the deception of many smaller VMs, each running a separate OS instance. Here a system is successfully presented that uses virtualization technology to allocate data center resources dynamically based on application demands. The concept of "skewness" is introduced and achieved to measure the unevenness in the multidimensional resource utilization of a server. The concept of green computing is introduced by optimizing the number of servers in use. In future this concept can be extended.

ACKNOWLEDGMENT

I would like to thank Mr. Rakesh Rajani, my project guide for their helpful comments and suggestions. I express my sincere and profound thanks to Ms. Vani Hiremani & Ms. Kalpana Saharan, which always stood as the helping and guiding support for me. I would like to thank my family, GOD and all the people who gave me an unending support right from stage the idea were conceived.

REFERENCES

[1] Zhen Xiao, Weijia Song, and Qi Chen, "Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment," IEEE Transactions on Parallel and Distributed Systems, Vol. 24, No. 6, June 2013

[2] Liang-Teh Lee, Kang-Yuan Liu, Hui-Yang Huang and Chia- Ying Tseng, "A Dynamic Resource Management with Energy Saving Mechanism for

Supporting Cloud Computing," in International Journal of Grid and Distributed Computing Vol. 6, No.1, Feb, 2013.

[3] Anton Beloglazov and Rajkumar Buyya, "Energy Efficient Resource Management in Virtualized Cloud Data Centers," 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010.

[4] "Amazon elastic compute cloud (Amazon EC2)," http://aws.amazon.com/ec2/, 2012.

[5] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat, and R. P. Doyle, "Managing energy and server resources in hosting centers," in Proc Of the ACM Symposium on Operating System Principles (SOSP'01), Oct.2001.

[6] P. Braham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R.Neugebauer, I. Pratt, and A. Warfield, "Xen and the art of virtualization," in Proc. of the ACM Symposium on Operating Systems Principles (SOSP'03), Oct. 2003.

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

[8] T.Wood,P. Shenoy, A. Venkataramani, and M.Yousif, "Black-box and gray-box strategies for virtual machine migration," in Proc. Of the Symposium on Networked Systems Design and Implementation (NSDI'07), Apr. 2007.

[9] G. Chen, H. Wenbo, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao, "Energy-aware server provisioning and load dispatching for connection-intensive internet services," in Proc. of the USENIX Symposium on Networked Systems Design and Implementation (NSDI'08), Apr. 2008.

[10] M. Armbrust, A. Fox, Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, Stoica, Zaharia., "Above the clouds: A Berkeley view of cloud computing," University of California, Berkeley, Tech. Rep., Feb 2009.

[11] C.A. Waldspurger, "Memory Resource Management in VMware ESX Server," Proc. Symp. OS Design and Implementation (OSDI '02), Aug. 2002.

[12] Narander Kumar, Shalini Agarwal, Vipin Saxena, " Overload Avoidance Model using Optimal Placement of Virtual Machines in Cloud Data Cetres", International Journal of Computer Applications (0975 – 8887) Volume 73– No.11, July 2013

[13] Namita R. Jain, Rakesh Rajani, "A review of Virtualization Technology to Allocate Data Centre Resources Dynamically Based on Application Demands in Cloud Computing", International Journal of Computer Science and Management Research eETECME, October 2013, ISSN 2278-733X.

[14] Namita R. Jain, Rakesh Rajani, "Virtualization Technology to Allocate Data Centre Resources Dynamically Based on Application Demands in Cloud Computing", International journal of software & hardware research in Engg., ISSN No: 2347-4890 Vol. 2 Issue 1, Jan. 2014

AUTHORS

1. Namita R. Jain currently a PG student at Alard College of Engg, Pune University. Her current research focuses on Cloud Computing & virtualization.
2. R. Rajani, Assistant Professor with ME Department of Computer Science, ALARD College of Engineering, Pune University. His current research focuses on Cloud Computing & Networking.