Virtual Machine Provisioning in Cloud Computing Environment – A Performance Evaluation Study

Anand Motwani[1], Sukanya Rajak[2]

[1]*Associate Professor & Head,* [2]*PG Scholar*

Department of CSE, NRIIRT, Bhopal

Abstract- Cloud Computing systems fundamentally provide on demand access to large shared pool of computational and data resources through a variety of means similar to existing Grid and high performance computing (HPC) resource management and programming systems. The Virtualization is key technology to cloud computing. As efficient resource provisioning is a challenging problem in cloud based environments due to its dynamic nature. The dynamic and on-demand attribute is required to support heterogeneous applications with diverse Quality of Service (QoS) requirements. Virtual Machine Management (VMM) is one of the promising tasks aimed at optimal resource utilization within the cloud. The provisioning of hosts capacity to VMs is one of the sub tasks of VM management life cycle. Inefficient provisioning of hosts to VMs may lead to increasing demands for Data Center (DC) resources including power that further enhance cost and SLA violation. This work introduces the problem of VM allocation / provisioning and surveys few VM provisioning approaches present in literature. Also, the performance evaluation of available standard VM Allocation technique is done using famous cloud simulation tool CloudSim-3.0.2. Finally, the future work is given in the field of VM provisioning with an aim to increase Cloud Providers profits while maintaining QoS.

Keywords: Cloud Computing, CloudSim, Quality of Service (QoS), Service Level Agreement (SLA), Virtual Machine Allocation, Virtual Machine Provisioning.

I. INTRODUCTION

Cloud Computing environments is highly dynamic in nature and need to support heterogeneous applications with diverse performance requirements. A large-scale DC may process various kinds of requests with different levels of importance or priority from lots of individual users [1]. The QoS requirements of each user can change dynamically over time. In particular, for a profitable success of this computing paradigm, the Cloud Data Centers (DCs) need to offer a better and strict QoS guarantees mentioned in Service Level Agreements (SLAs).

The cloud DCs that typically hosts thousands of servers utilizes virtualization technology to address these issues of dynamic resource provisioning. Resource provisioning plays a key role in ensuring that the cloud providers adequately accomplish their obligations to customers while maximizing the utilization of underlying infrastructure [2]. So, the dynamic (adaptive) resource allocation schemes are always required to handle dynamic requirements of users in cloud. Hence researchers always aids in implementation of new VM provisioning policies based on optimization goals (user centric, system centric). These schemes must allocate the minimal host resources needed for acceptable fulfillment of SLAs, leaving the remaining resources free to deploy more virtual machines.

The rest of the paper is organized as follows. Section 2 explains the problem of VM Provisioning along with the general architecture of cloud system. Section 3 discusses techniques for VMs provisioning based on various heuristics (Related work). A performance evaluation of available standard VM Allocation technique is done using famous CloudSim tool and meticulously presented in Section 4. Finally Section 5 concludes the paper giving future directions.

II. VIRTUAL MACHINE (VM) ALLOCATION (PROVISIONING)

The clouds are typically large scale DCs having thousands of servers with computing and other resources. These resources are typically virtualized and have several advantages such as on-demand scalability of resources; there are still issues which prevent their widespread adoption in clouds. These issues are potentially addressed by researchers.

To understand the VM provisioning, an Infrastructure-level services (IaaS) cloud model is considered with resources, refer Figure – 1. Data Centers composed a set of managed hosts, which in turn is responsible for managing VMs during their life cycles. VM life cycle includes: provisioning of a host to a VM, VM creation, VM migration and VM destruction. Host is a pre-configured physical processing node in a cloud with one or more CPU cores, capacity in millions of instructions per second – MIPS, memory, storage, and a policy for allocating processing cores to virtual machines. The policy is also known as VM Allocation or Provisioning policies. VM allocation [3] is the process of creating VM instances on hosts that match the critical characteristics (storage, memory), configurations (software environment), and requirements of the SaaS provider.

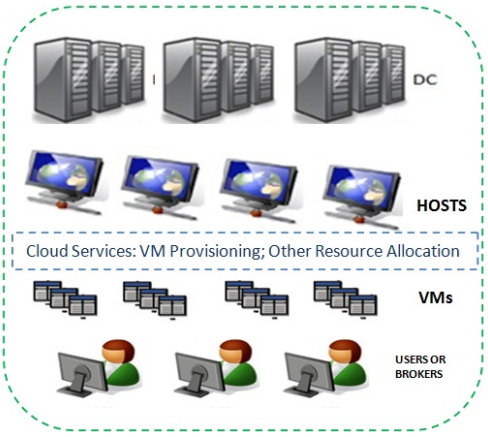

Figure 1: Cloud Computing Architecture

VM provisioning policies are developed based on optimization goals (user centric, system centric). The First-Come-First-Serve (FCFS) is default policy implemented by the VM Provision is a straightforward policy that allocates a VM to the Host. The other system parameters like number of processing cores, memory and storage as requested by the Cloud user form the basis for such mappings. Many complicated policies can also be written by the researchers based on the application needs and Infrastructure. Allocation of application-specific VMs to Hosts in a Cloud-based data center is the responsibility of the VM Provisioning - cloud service.

In our work, we considered VM provisioning at the host level, as at the host level, we can specify how many processing cores and how much of the overall processing power of each core will be assigned to each VM. For each Host component, the allocation of processing cores to VMs is done based on a host allocation policy [4] and it is possible to assign specific CPU cores to specific VMs. More specifically, if two VMs (VM1 and VM2) each requiring one CPU core can run on host having two CPU cores at the same moment if all other resource requirements are satisfied.

III. RELATED WORK

Saurabh Kumar Garg et al. [2] addressed the resource allocation problem within a datacenter that runs different type of application workloads, particularly transactional and non-interactive applications. They proposed admission control and scheduling mechanism that maximized the resource utilization and profit by ensuring the SLA requirements of users. The proposed resource provisioning mechanism predicts availability of resources in future and schedules the jobs by stealing CPU cycles, which transactional applications may not utilize. The technique can easily integrated with the admission control and handles auto scaling facility.

The authors [5, 6] highlighted and categorized the main challenges, involved to the resource allocation process particular to distributed network clouds. Several internal and external factors that influence the performance of resource allocation models are discussed in these articles. Research Challenges inherent to Resource Allocation are discussed in [6]. The authors [5] stated that resource allocation model has to use computational resources as well as network resources to accurately reflect practical demands.

The authors [7] stated that: it is not possible for a cloud provider to satisfy all the requests due to finite resources at a time. Also it is mentioned that, generally cloud providers use two simple resource allocation policies. One is immediate and other is best effort. Immediate allocation policy allocates the resources if available, otherwise rejects. Best-effort policy allocates the requested resources if available otherwise the request is placed in a FIFO based queue. The work proposed dynamic planning based scheduling algorithm and is implemented in Haizea cloud toolkit.

Rodrigo N. Calheiros et al [8] present a VM provisioning technique that automatically adapts to workload changes related to applications for adaptive management of system while offering end users guaranteed Quality of Services (QoS) clouds. The behavior, performance of applications and Cloud-based IT resources are modeled to adaptively serve end-user requests. An analytical performance (queue based network system model) and workload information is effectively used for efficiency. The results demonstrated that the proposed provisioning technique detects changes in workload patterns (like arrival and resource demands patterns) that occur over time and allocates multiple virtualized IT resources accordingly to attain application QoS targets.

Linlin Wu et al. [9], proposed resource allocation algorithms for SaaS providers who want to minimize infrastructure cost and SLA violations. The parameters like response time (QoS parameter) and service initiation time (infrastructure level parameter) are taken into account. Authors' claims that the proposed algorithms are able to manage the dynamic change of customers, mapping customer requests to infrastructure level parameters and handling heterogeneity of VMs.

Jing Tai Piao and Jun Yan [10] proposed a virtual machine placement and migration approach to minimizing the data transfer time between virtual machines, thus helping optimize the overall application performance.

Ruben Van den Bossche et. al. [11] have discussed about online cost-efficient scheduling. A hybrid cloud scheduling algorithm is used for hybrid clouds for efficient utilization

of resources cost minimization. The hybrid cloud determines what type of workload has to be outsourced and to which cloud service provider. Such decisions must minimize the cost of running the parts of total workload on one or more multiple cloud providers' end. Also, the application's computational requirement and data requirement are taken into account. The computational and data transfer costs as well as network bandwidth constraints are also considered.

Sheng Di et. al. [12] have discussed about error-tolerant allocation and payment minimization. With virtualization technology being increasingly mature, computational resources in cloud systems can be partitioned in fine granularity and allocated on demand. Following are three contributions in this paper: 1) Deadline-driven resource allocation problem is formulated based on the cloud environment aided with VM resource segregation, and also a novel solution with polynomial time is proposed, which could lessen users' payment in terms of their expected deadlines. 2) An error-tolerant method is proposed to guarantee tasks' completion within its deadline is proposed by analyzing the upper bound of task execution length based on the possibly inaccurate workload prediction. 3) Its effectiveness over a real VM-facilitated cluster environment under different levels of competition is validated.

IV. PERFORMANCE ANALYSIS

4.1 Experimental Setup

The experimental setup involves introduction to Simulation environment, Simulation scenario and Parameters.

A. Simulation Environment

CloudSim [4, 13] is used for simulation of VM Allocation policy. The tool Cloudsim, fronted by Buyya, provides a simple and extensible simulation framework that facilitates seamless simulation, modeling and experimentation of emerging cloud computing services. Its functionalities provides modeling and simulation of large scale data centers, virtualized cloud hosts, profit based and federated clouds, energy-aware computational resources, user-defined policies for allocation of virtual machines and also policies for allocation of host resources to VMs. For our simulations, CloudSim uses Sun's Java version 1.7. Apache Ant is used to compile CloudSim.

B. Simulation Parameters

We attempted to depict the simulation scenario used in our work in Figure – 2. The simulation is performed by characterizing three data centers at different locations owned by an IaaS provider, four hosts, which run varied number of virtual machines. A user will submit and perform 100 tasks (cloudlets) on these VMs. The values of simulation parameters chosen for our simulation are mentioned in Table - 1. The simple VM allocation policy which is simulated for this work chooses the host for a VM with less processing elements (PEs) in use.

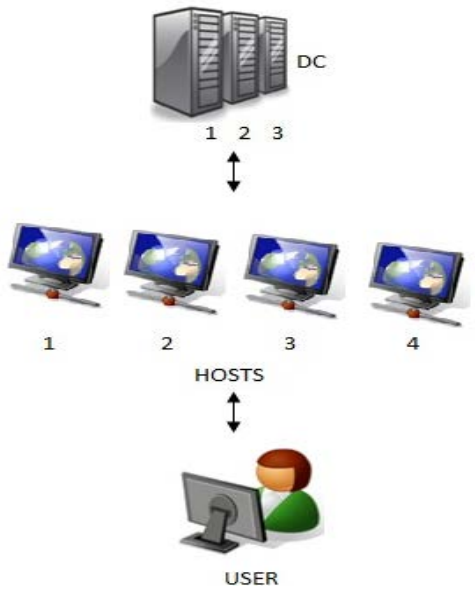

Figure 2: Simulation scenario

Table 1: Cloud simulation parameters

Datacenter	Architecture	x86
	Operating System	Linux
	Virtual Machine Manager	Xen
Host	MIPS	5000
	Processing cores	4 (quad core)
	RAM	2560 MB
	Storage	163840 MB
	Bandwidth	1024 Kbps
	VM Allocation policy	Simple
	VM Scheduler	Time Shared
Virtual Machine	MIPS	1000
	Processing Elements	1
	RAM	512 MB
	Storage	10, 000 MB
	Bandwidth	1000 Mbps
	Cloudlet Scheduler	Space Shared
	Virtual Machine Manager	Xen
Cloudlet / Task	Length	1000
	File size	300MB
	Output size	300 MB
	Processing Elements	1

4.2 Simulation Results and Analysis

Along with the results, we also discuss key performance indicators KPIs [14] (parameters) evaluated for this work.

i. Availability: It is the percentage of time a customer can access the service and it is given by:

$$\frac{(\text{Total servic e time}) - (\text{Total tim e for which servic e was not availabl e})}{\text{Total servic e time}}$$

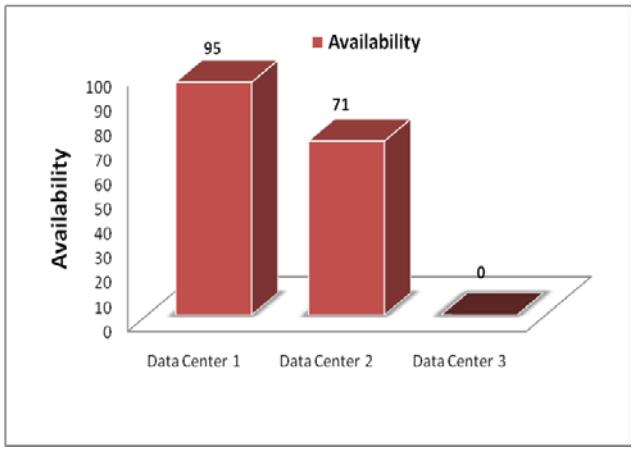

Figure 3: Availability at datacenters

Figure – 3 shows the availability in percentage, of each data center. Here maximum availability achieved is 95%.

ii. Throughput and efficiency: Throughput and efficiency are important measures to evaluate the performance of infrastructure services provided by Clouds. Throughput is defined as the number of tasks completed by the Cloud service per unit of time. Throughput depends on several factors including infrastructure initiation delays and inters task communication delays.

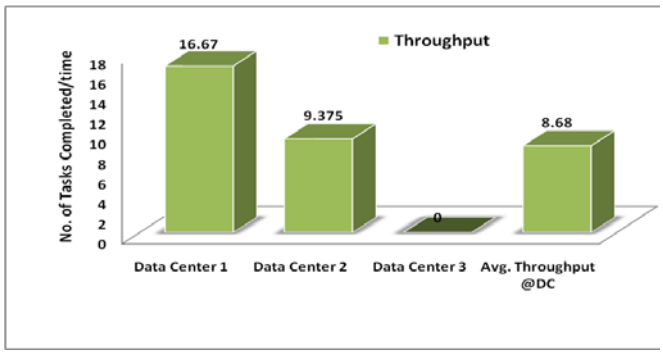

Figure 4: Throughput at datacenters

Let a user application have n tasks and they are submitted to run on m machines from the Cloud provider. Let Te(n,m) be the execution time of n tasks on m machines. Let To be the time overhead due to a variety of aspects such as infrastructure initiation delays and inter task communication delays. Therefore, the total throughput 'α' of a Cloud service is given by:

$$\alpha = \frac{n}{\text{Te}(n, m) + \text{To}}$$

System efficiency is given by:

$$\frac{\text{Te}(n, m)}{\text{Te}(n, m) + \text{To}}$$

The maximum throughput achieved in our simulation is nearly 16 tasks per unit time at DC1 and 9 tasks per unit time at DC2.

iii. Service response time: The efficiency of service availability can be measured in terms of the response time. It means how fast the service can be made available for usage in the case of IaaS. In other words it represents the time taken by the provider to serve this request. Figure – 5 shows four different response times: 30 tasks (cloudlets) are served with no delay, while the delay is 1.2 seconds for 30, 2.2 seconds for 30 and 3.2 seconds for 10 tasks.

One sub factor we measured is Average Response Time and it is given by $\sum T_i / n$ where T_i is time between submission of request for an IaaS service and when it is actually available. The n is the total number of IaaS service requests. In our work average response time is 1.34 seconds.

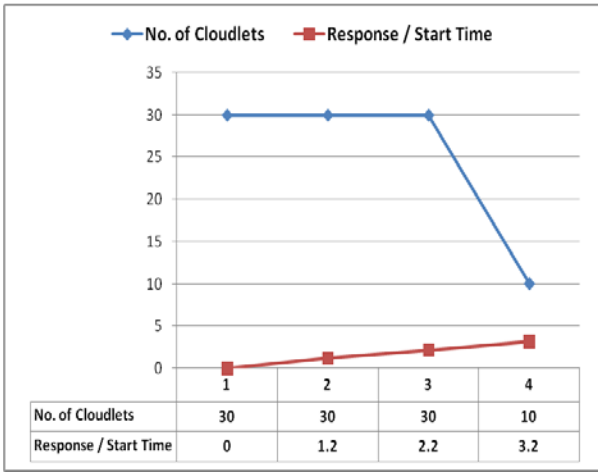

Figure 5: Response Time v/s No. of Cloudlets

5. Conclusion and Future Work

The efficient resource allocation in Cloud Computing Environment is a major challenge. Virtual Machine Management (VMM) is one of the promising tasks aimed at optimal resource utilization within the cloud environment. Inefficient allocation of processing elements of hosts to VMs may lead to increasing demands for data center resources. Hence, this paper mainly focused on putting insight, to importance of VM Provisioning in cloud environment. This work introduces the problem of VM allocation / provisioning and presents a review on VM

provisioning approaches present in literature. Also a performance analysis for simple VM Allocation policy is shown that can help researchers to develop and compare the custom VM provisioning policies based on optimization goals (user centric, system centric). For experimentation purpose, famous CloudSim (version 3.0.2) tool is used on windows platform. The performance evaluation is done on basis of three performance indicators: Availability, Throughput and Service Response Time.

As future work, we will deal with the challenging issue of optimal resource allocation (including VM), considering various parameters like performance, availability, profit maximization, reliability, response time, energy efficiency etc. depending on the application needs. In future, an adaptive VM Allocation policy for dynamic cloud environment has to be proposed in order to satisfy cloud providers' and users' needs without violating SLAs and maximizing profit.

REFERENCES

[1] Min Chen, Hai Jin, Yonggang Wen, Victor C. M. Leung, "Enabling Technologies for Future Data Center Networking: A Primer", IEEE Network, July/August 2013, IEEE.

[2] Saurabh Kumar Garg, Srinivasa K. Gopalaiyengar, and Rajkumar Buyya, "SLA-Based Resource Provisioning for Heterogeneous Workloads in a Virtualized Cloud Datacenter", Springer-Verlag Berlin Heidelberg 2011, ICA3PP 2011, Part I, LNCS 7016, pp. 371–384, 2011.

[3] Quiroz A, Kim H, Parashar M, Gnanasambandam N, Sharma N. Towards autonomic workload provisioning for enterprise grids and clouds. Proceedings of the 10th IEEE/ACM International Conference on Grid Computing (Grid 2009), Banf, AB, Canada, 13–15 October 2009. IEEE Computer Society: Silver Spring, MD, 2009; 50–57.

[4] Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov, C´esar A. F. De Rose, and Rajkumar Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms", 2010 John Wiley & Sons, Ltd., SOFTWARE – PRACTICE AND EXPERIENCE, 2011; 41:23–50.

[5] Mohamed Abu Sharkh, Manar Jammal, Abdallah Shami, and Abdelkader Ouda, "Resource Allocation in a Network-Based Cloud Computing Environment: Design Challenges", CLOUD NETWORKING AND COMMUNICATIONS, IEEE Communications Magazine, November 2013, 0163-6804/13, 2013 IEEE.

[6] Patricia Takako Endo, André Vitor de Almeida Palhares, Nadilma Nunes Pereira, Glauco Estácio Gonçalves, Djamel Sadok, Judith Kelner, Bob Melander and Jan-Erik Mångs, "Resource Allocation for Distributed Cloud: Concepts and Research Challenges", IEEE Network July/August 2011, 0890-8044/11, 2011 IEEE.

[7] Amit Nathani, Sanjay Chaudharya, Gaurav Somani, "Policy based resource allocation in IaaS cloud", Future Generation Computer Systems 28 (2012) 94–103, 2011 Elsevier B.V.

[8] Rodrigo N. Calheiros, Rajiv Ranjany, and Rajkumar Buyya, "Virtual Machine Provisioning Based on Analytical Performance and QoS in Cloud Computing Environments", 2011 International Conference on Parallel Processing, IEEE Computer Society, 0190-3918/11, 2011 IEEE.

[9] Linlin Wu, Saurabh Kumar Garg and Rajkumar Buyya, "SLA-based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments", 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, IEEE Computer Society, 978-0-7695-4395-6/11, 2011 IEEE.

[10] Jing Tai Piao, Jun Yan, "A Network-aware Virtual Machine Placement and Migration Approach in Cloud Computing", 2010 Ninth International Conference on Grid and Cloud Computing, IEEE Computer Society, 978-0-7695-4313-0/10, 2010 IEEE.

[11] Ruben Van den Bossche, Kurt Vanmechelen, Jan Broeckhove, "Online cost-efficient scheduling of deadline-constrained workloads on hybrid clouds", Elsevier, Future Generation Computer Systems, Vol 29, Issue 4, June 2013, pp 973-985.

[12] Sheng Di, Cho-Li Wang, "Error-Tolerant Resource Allocation and Payment Minimization for Cloud System", IEEE Transactions On Parallel And Distributed Systems, Vol. 24, No. 6, June 2013, pp 1097-1106.

[13] Rajkumar Buyya, Rajiv Ranjan, and Rodrigo N. Calheiros, "Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities", 978-1-4244-4907-1/09, 2009 IEEE.

[14] Stefan Frey, Claudia L¨uthje, Christoph Reich, "Key Performance Indicators for Cloud Computing SLAs", EMERGING 2013 : The Fifth International Conference on Emerging Network Intelligence, ISBN: 978-1-61208-292-9, IARIA, 2013.