A Review on Different VM placement Approaches in Cloud Computing Environment

Rohit Chourasia
er.rohcha@gmail.com

Rigvi Roi
r.roi1@hotmail.com

Abstract- **Cloud computing is growing as the next-generation platform for the calculation. Recently, there has been a dramatic increase in the popularity of cloud computing systems because of their attractive features such as rent computing resources on-demand, bill on a pay-as-you-go basis, ubiquitous network access, location independent resource pooling, rapid resource elasticity, transference of risk, etc. Virtualization is the core technology behind the cloud computing, its enable the sharing of physical resources by the multiple client. Hypervisor create the VM for each user and number of VM can be created in a single host. The biggest advantage of the virtualization is to remap the VM in different host to deal with the load balancing. This remapping of the VM is called VM migration. VM migration is a technology that allows the transferring of VM from one host to another host. VM is the fundamental unit in the cloud. Overall performance of the cloud system is depends on the proper placement of the VM, which is handled by the VM scheduler. VM scheduling is a set of rules that determine the host where VM can be placed. It is a key function in any cloud system. Since load on the VM is changed continuously, so VM placement is the challenging task in the cloud. Numbers of VM placement methods have been developed. Main goal of these approaches is to increase the system performance and reduce the energy consumption. In this paper we explain different type of VM placement approaches with their anomalies.**

Keywords: **VM scheduling, VM migration, Cloud, Static VM scheduling, Dynamic VM scheduling, Load balancing.**

I. INTRODUCTION

Cloud computing is emerging as a new technology in the field of computer science [1]. It is become so famous because of their attractive features. According to NIST definition [2] cloud computing is a model for enabling suitable, omnipresent, on-demand network access to a shared pool of configurable computing resources. These resources can be hardware, software, network etc. This pool of shared computing resources can be rapidly provisioned and released. In the cloud, computing resources are provided to the client through virtualization. The large scale computing infrastructure is established by cloud providers to make availability of online computing services in flexible manner so the user find easiness to use the computing services [1]. This cloud model is basically composed of five actors, three types of service models, and four deployment models [3].

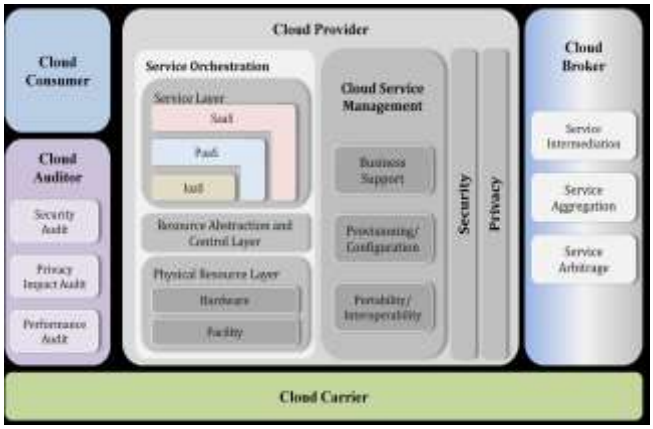

Figure 1: Cloud Computing Models

Cloud Consumer is a person or organization that maintains a business relationship with, and uses service from, Cloud Providers. Cloud Provider a person, organization, or entity responsible for making a service available to interested parties. Cloud Auditor is a party that can conduct independent assessment of cloud services. Cloud Broker is an entity that manages the use, performance and delivery of cloud services, and negotiates relationships between Cloud Providers and Cloud Consumers. Cloud Carrier is an intermediary that provides connectivity and transport of cloud services from Cloud Providers to Cloud Consumers.

Cloud supports three types of services. Software as a service (SaaS) mainly delivered the online software application to the client. These applications are access by the various client using computing devices like a thin or thick clients, interface such as a web browser. The users do not have permission to manage or control the underlying cloud infrastructure including hardware resources or platform infrastructure etc. Gmail and Facebook are one of the most famous cloud applications. Platform as a Service (PaaS) gives the capability to create application services as on their desire. It allows users to develop their software using programming languages and tools supported by the provider.

Table 1: Static V/S Dynamic approach

Sr. No.	**Type of Approach**	**Based on**	**Addressing Issues**	**Drawbacks**
1.	Static	Fixed values are used, No previous knowledge is required.	Response time Resource utilization Scalability Power consumption and Energy Utilization Make span Throughput/Performance	Not Scalable User can not changed demands at run time
2.	Dynamic	Decisions are made at run time. Run time statics are required to take the decision	Load estimation Minimizing the number of migrations Throughput	Complex Time Consuming

The user does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage. The user has control only over the deployed applications and application hosting environment configurations. Infrastructure as a Service (IaaS) provides the capability to have control over complete cloud infrastructure with CPU processing, storage, networks, and other computing resources. The cloud user is able to deploy and run their software, which can include operating systems and other software applications as website. There are four types of cloud deployment model [4] in the cloud computing known as public, private, community and hybrid cloud.

Private Cloud is a model of cloud computing whose framework is allowed to use with a particular organization. All the resources and services are dedicated to a limited number of peoples. The server and data center is also setup within organization. Sometimes infrastructure is setup by third party but it is in full control of organization. The private clouds are good to privacy and security. Public cloud is model of cloud where all users are allowed to access the services using internet. The user need only internet connection and web browser to access with pay per use scheme. All the services with infrastructure of cloud provider are available on the internet. User need to subscribe the application and make enable to use it. Community cloud includes number of organization to share their services to increase resource utilization of cloud infrastructure. The cloud infrastructure is not limited to only one organization. Hybrid cloud combines both public and private cloud with their advantages. Hybrid cloud offers the benefits of both the public and private cloud. The hybrid cloud is the good solution for purely business oriented concept because many modern businesses have a wide range of concerns to support users demand. User can use these services anywhere at any time using pay-as-you-use model. In cloud environment each data center having a number of host.

Virtualization [5, 6] is the key technology behind the cloud computing. It is a technique which divides the physical resources and allows running multiple OS on a single physical machine at the same time. Virtualization is implemented through the hypervisor also known as virtual machine monitor (VVM), which is a small layer between the physical hardware and operating. VMM is responsible for all the operation related to the VM such as scheduling, VM creation Destruction etc. Each host can create number of virtual machine. And these virtual machines are provided to the client to run their application as pay per use model. When any request comes from any client for cloud service, VMM create new virtual machine and assign the required resources to that virtual machine. Where the VM is created is decided by the VM scheduler, which is defined in VMM. It's defined the set of rules that must be followed by the VM scheduler for creating the VM. VM scheduling is a key function in any cloud system. Overall system performance can be increased by the proper VM scheduling. Since load on the VM can be changed at run time, so VM placement is a very challenging task in the cloud.

II. RELATED WORK

Load on the VM can change dynamically [7], so some resources are left for the future use. To avoid the overloaded and underloaded situation upper and lower threshold are used respectively. Based on the threshold VM scheduling approach can be static or dynamic. In the static approach

fixed lower and upper threshold are used, that can't be changed with time, while in the dynamic approach an dynamic lower and upper threshold are used, that can be changed with time.

B. Sotomayor et al. [8], proposed a Round Robin approach for the VM placement. This approach used the static threshold. In this approach VM are placed on the first come first serve (FCFS). In this approach all process are treated equally and VM scheduler assign the resources to the process that comes first. Most of the cloud provider used this approach.

Y. Fang et al., [9] proposed three layer architecture of cloud computing i.e. application layer, platform layer and infrastructure layer. The application layer is oriented to users, it implements the interaction mechanism between user and service provider with the support of platform layer. Users can submit tasks and receive the results through the application layer in the task scheduling process. The infrastructure layer is a set of virtual hardware resources and related management function. Furthermore, the platform layer is a set of software resources with versatility and reusability, which can provide an environment for cloud application to develop, run, manage and monitor.

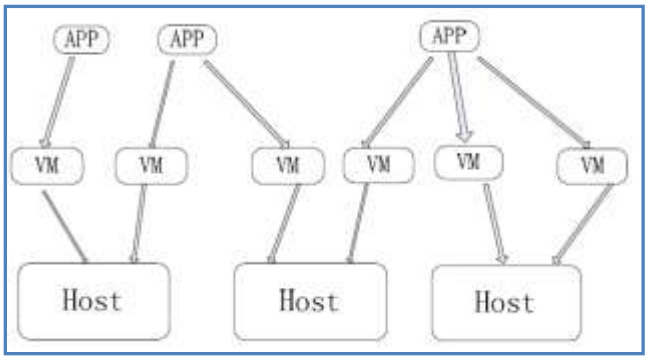

Figure 2: Job Allocation Policy in Cloud Server [9]

So according to the above architecture, they are using two levels scheduling model. The first level scheduling is from the users' application to the virtual machine, and the second is from the virtual machine to host resources. In this two levels scheduling model, the first scheduler creates the task description of a virtual machine, including the task of computing resources, network resources, storage resources, and other configuration information, according to resource demand of tasks. Then the second scheduler find appropriate resources for the virtual machine in the host resources under certain rules, based on each task description of virtual machine. Main problem with this approach are, virtual machine is scheduled to the host with lightest load each time to avoid overloading. So it will increase the number of active server. Recent studies [10] show that on average an idle server consumes approximately 70% of the power consumed when it is fully utilized. They also only considered task response time and the resource demand by the task.

S.Selvarani et al. [11], proposed cost based scheduling algorithm. The scheduler accepts number of tasks, average MI of tasks, deviation percentage of MI granularity size and processing overhead of all the tasks. Resources are selected.Tasks are sorted according to their priority, and they are placed in three different lists based on three levels of priority namely high priority, medium priority and low priority. Now job grouping algorithm is applied to the above lists in order to allocate the task-groups to different available resources. After assigning the task to the list, they are allocating the process according to the priority that means high priority task are process first. This method is good, but problem with this approach is that it's not deal with the dynamic nature of the process.

G. Xu et al. [12]., proposed a scheduling algorithm which are using global and local resource manager. Global resource manager (Main controller) installed in the main server and local resource manager (Partition controller) installed into the local host. When a job arrives at the system, Global resource manager decide which cloud partition should receive the job. Then local resource manager decides how to assign the jobs to the nodes.

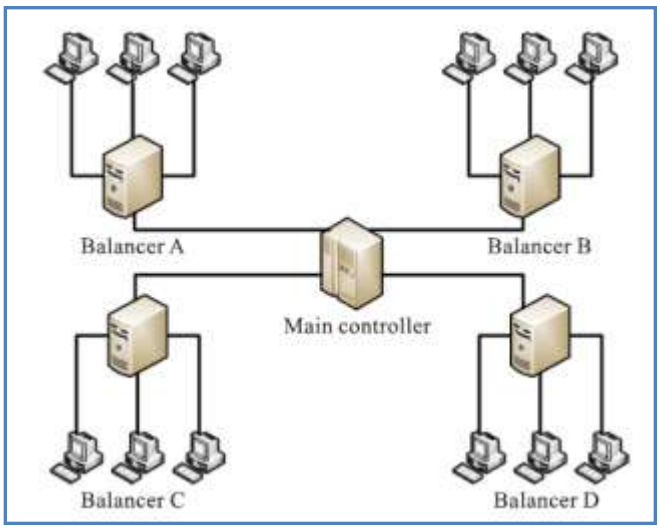

Figure 2: Relationships between the main controllers, the balancers, and the nodes

Mayank Mishra et al. [13], Proposed a method for placing the VM which is based on Vector theory. They are using resource vectors TCV, RUV, RCV and RRV in the 3-D

space which we will use for making different VM placement decisions. One of our prime goals while placing VM is to make the resource utilization of PMs as balanced (along each resource dimension) as possible, i.e., the RUV of a PM should be as closely aligned to the TCV as possible. This would require that we have a way of finding complementary VM for a PM. They proper balance the resources but not focus to the server consolidation. So energy consume by the data center is high.

Y. Song et al. [14], proposed a RAINBOW model for the resource allocation. Resources are allocated on the base of priority. VM whish having the higher priority firstly assign to the host compare to the VM which have a lower priority. This approach seems good, but there may be starvation if VM with higher priority comes regularly. This approach also not support to the VM migration.

III. CHALLENGING ISSUES IN VM SCHEDULING

Virtual machine scheduling can be applied in various scenarios, depending on the user requirement. Numbers of challenges are faced during the development of VM scheduling methods, some of them are:

a. When the problem size is large it is very difficult to find the suitable parameter values for the proposed model. For example if there are n physical machine and each physical machine having m VM. Then there are n*m possibility for placing each VM. Therefore a mechanisms are required that automatically capture those values.
b. There are no standard model exists, which represent various scenarios of VM Scheduling. Applications have various QoS constraint such as response time, latency, consistency, throughput, and transaction rate. These requirements are vary according to the type of application and how they used. Modeling and quantifying these requirements are very challenging task.
c. User requirement in the cloud are change dynamically. So it is very difficult to calculate the accurate VM utilization. Therefore there should be methods that automatically resize the VM according to the user requirement.
d. The initial placement of VM is a NP hard problem. Proper placement of VM is the complicated task in large cloud center. Amazon EC2, 40,000 servers and schedules 80,000 VMs every day [15]. These numbers may increase as the cloud popularity increased. Finding suitable solutions for a large-sized data centers in an acceptably short time, with high scalability, is known to be difficult.

IV. CONCLUSION

VM scheduling is one of the critical task in the cloud environment, due to the dynamic nature of the VM load. It is a NP- hard problem, so appropriate placement is a very challenging task. An efficient VM placement can increase the overall system performance. Recent studies [] shows that a wrong VM placement can increased the number of migration as well as energy consumption. This paper explained the different type of exiting VM scheduling approach with their anomalies. VM scheduling can be static or dynamic. Because of the dynamic nature static approach are not suitable for the cloud.

REFERENCES

[1.] R. Buyya et al., "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility", Future Generation Computer Systems, vol. 25, no. 6,june 2011, pp 599-616.
[2.] Mell, P. et al., "The NIST Definition of Cloud Computing". NIST Special Publication, 2011
[3.] P. Mell et al "Cloud Computing" by National Institute of Standards and Technology - Computer Security Resource Center- www.csrc.nist.gov.
[4.] Wenke Ji et al., "A Reference Model of Cloud Operating and Open Source Software Implementation Mapping" in 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises, 2009
[5.] Borja Sotomayor et al., "Enabling cost-effective resource leases with virtual machines," Research Gate article may 2014..
[6.] L. Cherkasova et al. "When virtual is harder than real: Resource allocation challenges in virtual machine based it environments" in proc. 10th conference on hot topic in operating system, Vol. 10, pp.20-20, 2005.
[7.] M. Katyal et al., "A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment", International Journal of Distributed and Cloud Computing, vol. 1, pp. 1-14, 2013.
[8.] Sotomayor, B., Montero, R. S., Llorente, I. M. & Foster, I. (2009). Virtual infrastructure management in private and hybrid clouds. IEEE Internet Computing, 13(5), 14-22.
[9.] Yiqiu Fang et al., "A Task Scheduling Algorithm Based on Load Balancing in Cloud Computing" Springer, pp: 71-77, 2010.
[10.] R. K. Gupta et al., "A Complete Theoretical Review on Virtual Machine Migration in Cloud Environment", IJ- Closer, vol. 3, pp. 172-178, 2014.
[11.] S. Selvarani et al., "IMPROVED COST-BASED ALGORITHM FOR TASK SCHEDULING IN CLOUD COMPUTING", in the proceeding of IEEE int. conf. on Computational Intelligence and Computing Research (ICCIC), pp. 1-5, 2010
[12.] G. Xu et al. "A Load Balancing Model Based on Cloud Partitioning for the Public Cloud" in the proceeding of IEEE conference on TSINGHUA SCIENCE AND TECHNOLOGY, pp. 34-39, 2013..
[13.] M. Mishra et al., "On Theory of VM Placement: Anomalies in Existing Methodologies and Their Mitigation Using a Novel Vector

Based Approach", IEEE/ACM 4th international conference on cloud computing, pp 275-282, July 2011.

[14.] Y. Song, "Multi-Tiered On-Demand resource scheduling for VM-Based data center" In Proc. of the 2009 9th IEEE/ACM Intl. Symp. on Cluster Computing,155, 2009.

[15.] Hu. inhua et al., "A Scheduling Strategy on Load Balancing of Virtual Machine Resources in Cloud Computing Environment," Third International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), pp.89-96, Dec. 2010.