A Review on Different VM Scheduling Techniques in Cloud Environment

SachinSoni[1], PraveenYadav[2]
Department of Computer Science, Oriental Institute of Science and Technology, Bhopal, India[1]

Abstract - Cloud computing is one of the well developing fields in Computer Science and Information Technology. The efficient job scheduling increases the client satisfaction and utilizes the system energy in terms of time. Now days cloud computing is one of the fast growing technology in the field of computer science and information technology because of online, cheap and pay as use scheme. The cloud computing mainly a business oriented model to provide on demand computing resource. It is a service oriented design that reduces the cost of access to gather the information of the clients offer greater flexibility and demand based services and so on. The clients need not buy any additional hardware and software cost. The clients are the king of the modern business jargon. Quality is not a single step process. The cloud computing concept is motivated by the idea that information processing can be public utility and can be done more efficiently on large farms of computing resources and storage systems with the availability of all time throughout the world accessible via the Internet.

Keywords: Distributed Computing, On Demand Resources, Cloud Computing, Virtualization.

1. INTRODUCTION

In the cloud computing, the computing resources are provided to the client through virtualization, on the internet. 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]. According to NIST cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources. The computing resources include networks, servers, storage, applications, and services. In cloud computing, the shared pool of computing resources can be rapidly provisioned and released [2]. The management effort or service provider interaction for cloud user is also minimized to increase easiness. This cloud model is basically composed of five essential characteristics, three types of service models, and four deployment models.

As many enterprises, government organization, and other companies begin to start to use cloud computing, security issues came out as a basic problem in computing, as every individual client or user preferred to work on a clear and safe environment where privacy and security of their data is a major concern.

In this paper, we discuss the overview of cloud computing with their components basic model and process scheduling.

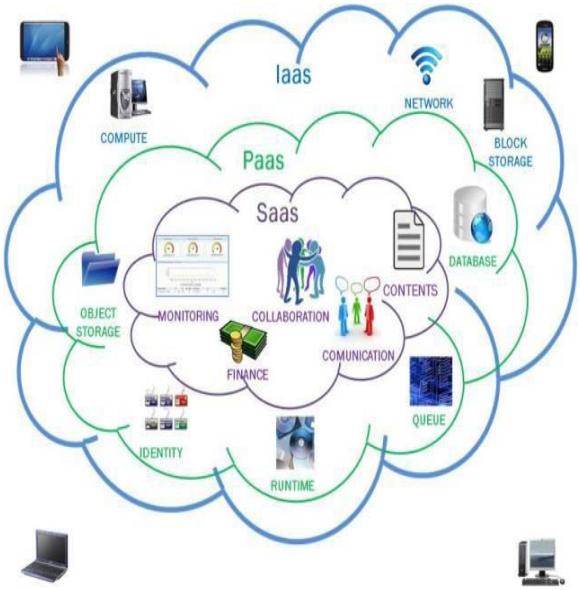

Fig. 1 Cloud Computing

Cloud Components

The cloud computing are mainly five part as shown in figure 1. Service delivery model and service deployment model are main concept of cloud computing. Cloud computing infrastructure, their resources and defining attributes.

1.1 Delivery Model

In the cloud computing there are three types of service delivery model [3] as software as a service (SAAS), platform as a service (PAAS), and infrastructure as a service (also known as hardware as a service).

1.1.1 Software as a service

It provides the client to use the application which is running on the cloud computing service provider infrastructure, client have to purchase only the license of the application required by him and service provider make available it to the client. Client can access those applications through network (browser, PDA). These applications are provided to the client on demand dynamically. High bandwidth is given to client so it starts rapidly and client does not need to worry about maintains and up gradations of its software. Gmail and Facebook is one of the most famous cloud applications.

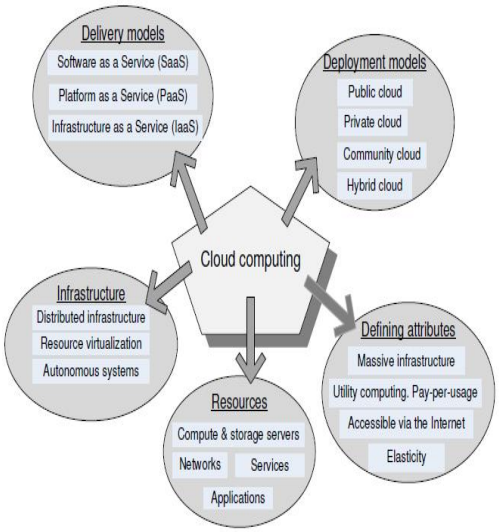

Fig. 2 Cloud Computing

1.1.2 Platform as a service

PaaS layer lies between SaaS and IaaS. It eradicates the costs of buying, configuring, engineering the hardware and software needed for the customer application. PaaS development tools are loaded into the cloud and the customer can get those services from the PaaS vendor by availing services from the PaaS provider. The difference between SaaS and PaaS is that SaaS only hosts completed cloud applications whereas PaaS offers a development platform that hosts both completed and in progress cloud applications. PaaS in addition to supporting application hosting environment, to possess development infrastructure including programming environment, tools, configuration management, and so forth. An example of PaaS is Google AppEngine.

1.1.3 Hardware as a service

Infrastructure as a service delivers basic storage and compute capabilities as standardized services over the network. IaaS is the base layer of cloud stack, it is also sometimes referred to as Hardware as a Service (HaaS) 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.

1.2 Deployment Model

There are four types of cloud deployment model [4] in the cloud computing known as public, private, community and hybrid cloud.

1.2.1 Private cloud

Private cloud computing systems emulate public cloud service offerings within an organization's boundaries to make services accessible for one designated 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.

1.2.2 Public cloud

The deployment of a public cloud computing system is characterized on the one hand by the public availability of the cloud service offering and on the other hand by the public network that is used to communicate with the cloud service. 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.

1.1.1 Community cloud

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.

1.1.2 Hybrid cloud

A hybrid cloud service deployment model implements the required processes by combining the cloud services of different cloud computing systems, e.g. private and public cloud services. It 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.

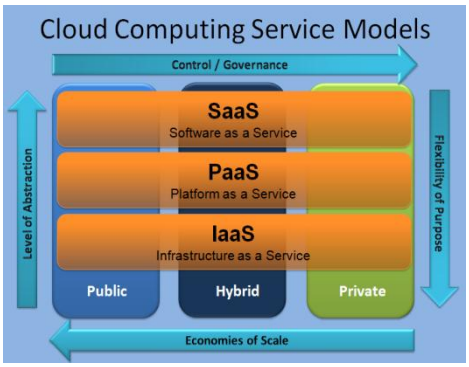

Fig. 3 Cloud Computing Models

1.2 Cloud Infrastructure

The infrastructure of cloud computing is mainly based on distributed server located throughout the world [5]. System resource virtualization plays one of the important roles in the cloud industry. The autonomous system also increases the adaptability of the systems. Virtualization is very useful concept. It allows abstraction and isolation of functionalities of hardware resources. Virtualization also enables portability of computing application and sharing of the physical computing resources. It has been applied to all aspects of computing memory, CPU processing power, storage media, software, networks, as well as services that IT industry offers.

1.3 Cloud Resources

The cloud resources are mainly three types namely storage and computing server, communication network and services applications [5]. In the cloud computing, datacenter is the collection of distributed servers and each server has computing resource to allocate them their users. Datacenter is a large work station in the basement of a building or a house with large number of host computer. Each computer is connected to the other side of the world that is accessible via the Internet to everyone. The cloud users access the resources using virtualization techniques that allow the sharing of computing resources of a host server to many of the users. To implement virtualization there is hypervisor, which allow one machine to be divided into many of the instances generally known as virtual machines. But users don't feel that it is sharing of resources. The distributed servers are in different geographical locations.

1.4 Cloud Attribute

The cloud computing have various types of characteristics and attributes [2]. The first most famous attributes is pay per use concept. The basic characteristics are given below.

On-demand self-service: in the cloud computing there no human interaction. Ever thing is online.

Broad network access: The cloud users have many of the option to access the services. There is many of the service providers who offer the services with effective service cost model.

Resource pooling: The cloud computing resources are pooled to serve large number of cloud user with virtual resources. Resources include storage media, processing capacity, primary memory, and network bandwidth etc are available to their client.

Rapid elasticity: As the cloud is a distributed collection of server and datacenter so capabilities are increasable with minimum effort. When consumer demand increases then cloud provide can increase the resource capacity by adding a new datacenter or server at any time.

Measured service: Cloud systems control and optimize all the installation and configuration related issues with automatically. Resource utilization, process scheduling and other infrastructure management work can be monitored, controlled, and reported only automated system and providing transparency for provider and consumer.

2. CLOUD SYSTEM

In order to provide services, large-scale data centers are established. These data center contain thousands of running computational nodes providing virtualization by placing many virtual machines (VMs) on each node. There are mainly two types of actors on cloud: end-user and brokers. The end-user requests for the application on cloud and brokers process these request. In the many systems, it is

considered two major roles for brokers: SLA (Service Level Agreement) Negotiation and VM-Monitor. The SLA Manager takes care that no (SLA) is violated and VM-Monitor monitor the current stated of virtual machines periodically at specific amount of time. Figure 4, shows the actual system view cloud computing environment.

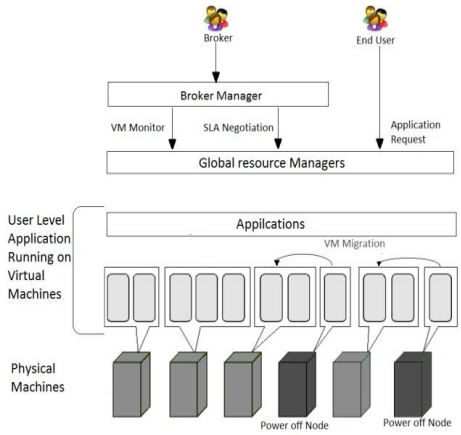

Fig 4: Job Allocation Policy in Cloud Server

All these request are taken by a global resource manager which decides what type of application is been requested and accordingly the VM machine is generated at physical nodes.

3. PROCESS SCHEDULING

The Cloud computing refers an online delivery of computing and storage services to an end-recipients. Cloud system is a collection of distributed data center located throughout the world and these data centers are connected using communication network like as internet. The computing resources are provided to their client using virtualization. Virtualization plays a key role in cloud computing by sharing of single host machine to number of Client's application.

Virtualization technology allows all physical resources to be virtualized and be transparent to client. Client doesn't need to have any information about the hardware type, physical location, level information of computing resources. Client just make a demand for the computing resources and according to the client requirement, the cloud system creates virtual machines inside any one of the host machine of a data center and using that virtual machine clients perform or execute their task or process. These process or task execute and perform their computation on different host machine using separate virtual machine.

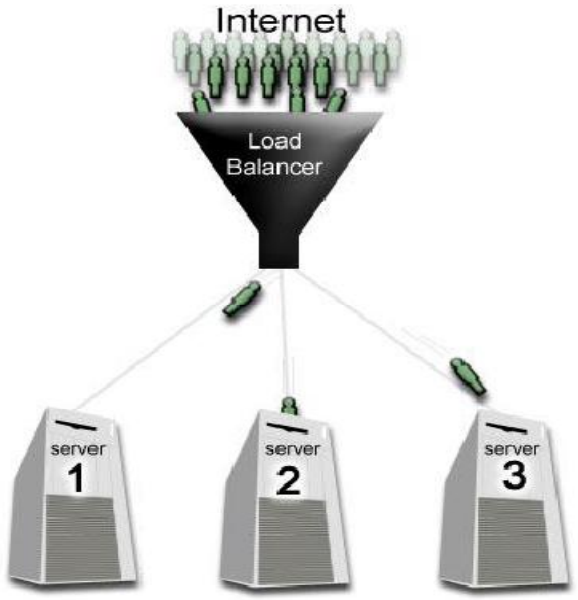

Fig 5: Job Allocation Policy in Cloud Server

But runtime execution and dynamic nature of cloud environment need process scheduling techniques. Process scheduling is one of the important parts of cloud system, in which appropriate resources are assigned to users' task. There are several studies related to the problem of management cloud resources of host machine in various levels like as virtual machine level, host machine level and data center level. Efficiency of process scheduling algorithm directly affects the performance of the whole cloud system performance. As shown in Fig. 5, when server receive the user request for computing resource on demand then it is stored in job pool. In the job pool, many types of the scheduling algorithms may be applied. Using these scheduling techniques, resources of host machine are assigned to user jobs using cloud broker.

4. PROCESS SCHEDULING TECHNIQUIES

It is necessary to do scheduling of cloud tasks in cloud environment to complete the client job. It should be decided which process need to allocated to which host system. If the computing resources are sufficient in current host system to complete the execution of process within deadline then no need to move or share the resources of other host. In this paper we have focused on various task scheduling techniques used in hybrid cloud environment with the aim of cost minimization and optimization for cloud resources. Some of the basic techniques are next part of the paper.

4.1 A Task Scheduling Algorithm Based on Load Balancing in Cloud Computing -

Cloud Computing Architecture includes three layers, 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.

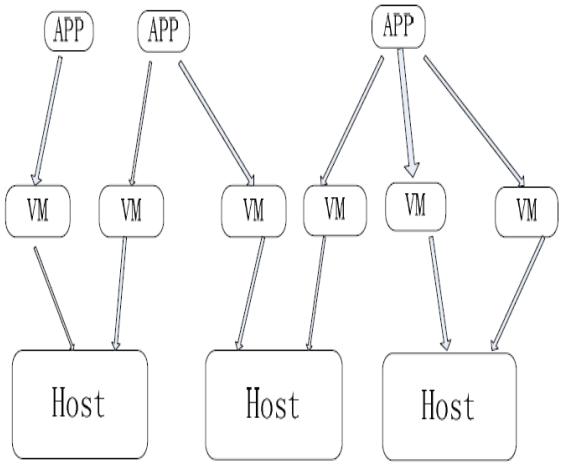

Fig 6: Job Allocation Policy in Cloud Server

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 create 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 Algorithm is divided into two levels scheduling, one is the mapping from task to a virtual machine, another is mapping from the virtual machine to host resources. Only task response time and the demand for resources are considered in this paper. At the same time, because tasks are dynamic, they may arrive randomly. If the tasks arrive at same time, they will be sorted ascending according to the resource applied by users. And if they arrive at different time, they will be sorted according to the time sequence arrived.

In the above algorithm, the virtual machine is scheduled to the host with lightest load each time. The advantage is to avoid overloading for the host hold more resources. Recent studies [2] show that on average an idle server consumes approximately 70% of the power consumed when it is fully utilized. And it only task response time and the demand for resources are considered in this paper.

4.2 A Load Balancing Model Based on Cloud Partitioning for the Public Cloud

There are several cloud computing categories, with this work focused on a public cloud. A large public cloud will include many nodes and the nodes in different geographical locations. Cloud partitioning is used to manage this large cloud. A cloud partition is a subarea of the public cloud with divisions based on the geographic locations. In this paper they 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. The partition load balancer (local resource manager) then decides how to assign the jobs to the nodes.

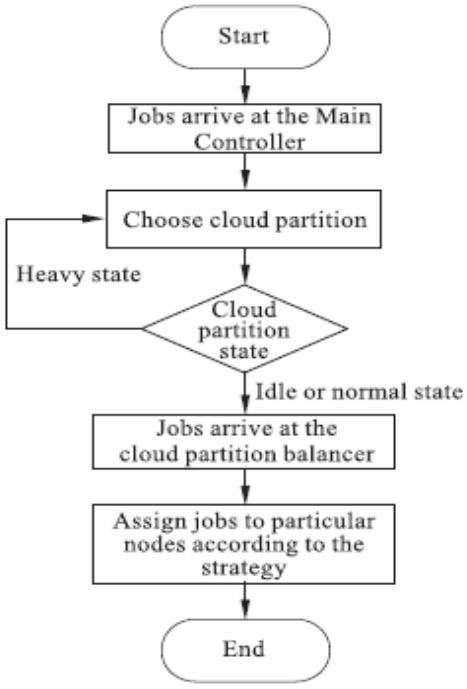

Fig 4: Job assignment strategy

When the load status of a cloud partition is normal, this partitioning can be accomplished locally. If the cloud partition load status is not normal, this job should be transferred to another partition. The load balance solution is done by the main controller and the balancers. The main controller first assigns jobs to the suitable cloud partition and then communicates with the balancers in each partition to refresh this status information.

When a job arrives at the public cloud, the first step is to choose the right partition. The cloud partition status can be divided into three types:

(1) Idle: When the percentage of idle nodes exceeds α change to idle status.

(2) Normal: When the percentage of the normal nodes exceeds β change to normal load status.

(3) Overload: When the percentage of the overloaded nodes exceeds γ, change to overloaded status.

The parameters α, β and γ are set by the cloud partition balancers.

The main controller has to communicate with the balancers frequently to refresh the status information. The main controller then dispatches the jobs using the following strategy: When job i arrives at the system, the main controller queries the cloud partition where job is located. If this location's status is idle or normal, the job is handled locally. If not, another cloud partition is found that is not overloaded. The node load degree is related to various static parameters and dynamic parameters. The static parameters include the number of CPU's, the CPU processing speeds, the memory size, etc. Dynamic parameters are the memory utilization ratio, the CPU utilization ratio, the network bandwidth, etc. In the above algorithm, the virtual machine is scheduled to the cloud partition with lightest load each time. The advantage is to avoid overloading for the host hold more resources. Recent studies show that on average an idle server consumes approximately 70% of the power consumed when it is fully utilized.

4.3 Compare And Balance Algorithms

This algorithm uses the concept of compare and balance approach to reach equilibrium condition and manage resource management. In this algorithm, work load of each host server is calculated firstly and then sum of total load on the all host of the systems is calculated. After that any host with the probability of p[k] (where p[k] is the probability of a host having minimum virtual machine) is selected. If the load of selected host server is less then as compare to current node than the load of current host server is transferred to chosen host. The working approach is presented in algorithm 1. On the basis of probability (number of virtual machine running on the current host and number of virtual machine running in the whole cloud system) each host server randomly selects a host server until it find appropriate host system to transfer their extra load. This process execute continuously so it consume large number of processing power. The above model is only applicable for load balancing of cloud system.

Table 1 Scheduling Algorithms with their Optimization Criteria and Features

ALGORITHMS OR TECHNIQUES	OPTIMIZATION CRITERIA	MULTI CORE AWARE	SUPPORT MULTIPLE WORKFLOWS
Graph based Task Scheduling Algorithm	Minimize Cost	No	No
Cost Effective Provisioning and Scheduling of deadline constrained application	Minimize Cost with Deadline	No	No
Cost Effective Scheduling Heuristics for deadline constrained workload	Minimize Cost with Deadline	No	No
Modified Bees Life Algorithm for Job Scheduling	Minimize Make-span	No	No
Time and Cost Optimization algorithm for multiple workflow	Minimize Cost with Deadline	Yes	Yes

4.4 Adaptive Threshold Based Energy Efficient Consolidation of Virtual Machines in Cloud

The target system, designed by this approach, is an IaaS environment, represented by a large-scale data center consisting of N heterogeneous physical nodes. Each node is characterized by the CPU performance depend in Millions Instructions Per Second (MIPS), amount of RAM, network bandwidth and disk storage. The type of the environment implies no knowledge of application workloads and time for which VMs are provisioned. Users submit requests for provisioning of M heterogeneous VMs characterized by requirements to CPU performance, RAM, network bandwidth and disk storage. This approach focusing on multi-core CPU architectures, as well as consideration of multiple system resources, such as memory and network interface, as these resources also significantly contribute to the overall energy consumption. In order to evaluate the proposed system in a real Cloud infrastructure, besides the reduction in infrastructure and on-going operating costs, this work also has social significance as it decreases carbon dioxide footprints and energy consumption by modern IT infrastructures.

5. CONCLUSION

This paper discussed the basic of cloud computing with their core technology i.e. virtualization. Moreover this paper also discussed some existing VM scheduling algorithm with their anomalies. Cloud computing have large number of resources to distributes their resources on demand. When the user request for the resources, virtual machine manager creates the VM and assign to the user. VM works similar to the PM. Each PM can have number of VM. So the proper placement of these VM is very challenging task due to the unpredictable nature of the VM. Since load on the VM can changed dynamically, so there is a need of an effective dynamic VM scheduling algorithm that place the VM effectively.

REFERENCES

1. Y. sahu et al. “cloud computing overview with load balancing techniques” International Journal of Computer Applications (0975 – 8887) Volume 65– No.24, March 2013, pp: 40-44.

2. Peter Mell,Timothy Grance, "Cloud Computing" by National Institute of Standards and Technology - Computer Security Resource Center-www.csrc.nist.gov.

3. Wenke Ji, Jiangbo Ma “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.

4. Miyuki sato“creating next generation cloud computing based network services and the contribution of social cloud operation support system (OSS) to society” 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises, 2009.

5. Börje Ohlman, Anders Eriksson “What Networking of Information Can Do for Cloud Computing” 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises, 2009.

6. Anas BOUA Y AD, Asmae BLILA T, Nour el houda MEJHED, Mohammed EL GHAZI “Cloud computing: security challenges” IEEE Computer Society, 2012.

7. Wang zong jiang “a new task scheduling algorithm in hybrid cloud”, international conference on cloud computing, 2012

8. Rodrigo N. Calheiros and rajkumar buyya, Cost effective provisioning and scheduling of deadline constrained application on cloud.

9. Ruben ban den bossche kurt jan, “Cost efficient scheduling heuristics for deadline constraints in cloud”, third international conference on cloud computing technology and science, 2011.

10. Modified bees life algorithm for job scheduling in hybrid cloud, international journal of engineering and technology, IJET, ISSN: 2049-3444, volume 2 no 6, june, UK, 2012