An Energy Efficient Virtual Machine Migration Policy in Cloud Environment

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Software Engineering

Submitted By

Sukhandeep Kaur

(801431030)

Under the supervision of:

Dr. Seema Bawa

Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004, PUNJAB, INDIA

JUNE 2016

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, **"An *Energy Efficient Virtual Machine Migration Policy in Cloud Environment* "**, in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Software Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of ***Dr. Seema Bawa*** and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Sukhandeep Kaur

(Sukhandeep Kaur)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Seema Bawa 23/06/16

(Dr. Seema Bawa)

Professor

CSE Department

Thapar University

Patiala

Countersigned by

(Dr. Deepak Garg)

Head

Computer Science and Engineering Department

Thapar University

Patiala

(Dr. S.S. Bhatia)

Dean (Academic Affairs)

Thapar University

Patiala

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life. It is great privilege to express my gratitude and admiration towards my respected supervisor Dr. Seema Bawa Professor Computer Science & Engineering Department. She has been an esteemed guide and great support behind achieving this task. This work would not have been possible without the encouragement and able guidance of her. I also thank my supervisor for her time, patience, discussion and valuable comments. Her enthusiasm and optimism made this experience both rewarding and enjoyable. I am truly grateful to her for extending her total co- operation and understanding whenever I needed help and guidance from her.

I am also thankful to Dr. Deepak Garg, Associate Professor and Head, Computer Science & Engineering Department and Dr. Rupali Bhardwaj, PG coordinator, for motivation and providing uncanny guidance and support throughout the preparation of the thesis report.

I will be failing in mu duty if I do not express my gratitude to Dr. S. S. Bhatia, Professor and Dean of Academic Affairs, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in he field. I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection, which made my stay at Thapar University memorable.

Last but not the least, I would like to thank my family for their wonderful love and encouragement, without their blessings none of this would have been possible.

Sukhandeep Kaur

Sukhandeep Kaur

(801431030)

Abstract

Due to rapid growth in cloud computing users, managing datacenter resources in an energy efficient way has become a key challenge. To make datacenter energy efficient virtual resources are created on the single physical hardware to fulfill the demands of large number of users simultaneously by using lesser number of active physical hosts. In virtualization, virtual machines are created on single host and these are migrated between different hosts according to current load and resource requirement for VM which results in minimizing the number of active hosts by powering off the idle hosts. Number of virtual machine migration techniques are present such as PSO, MPSO etc, which tries to minimize the energy consumption. Theses approaches minimizes the number of active servers by migration which sometimes results in overloading of remaining active servers and high energy consumption. This thesis work proposes a new virtual machine migration technique which tries to minimize the energy consumption by migrating the load between Virtual machines and running the tasks on the VMs according to VM's load in iterative manner. Cross Pollination Optimization technique is applied to achieve the desired outcome. The proposed algorithm is validated using cloudsim 3.0 simulator with integrated workflowSim. The results from complete validation study proved that Cross Pollination Optimization performs better then Modified Particle Swarm Optimization in terms of energy consumption corresponding to number of virtual machines and CPU utilization.

Table of Content

List of Figures

List of Tables

CHAPTER 1

Introduction

In today's world, internet becomes the daily usage of people for accessing the services. These services are different kind of resources which are provided by the cloud computing. Cloud computing is a platform which guarantees renting of computing and storage infrastructure, remote platform building and customization for business processes and renting of business application as hole. The subscribers made the service level agreement with the platform provider to access the various resources as services and by needs. The large scale data centers are called as clouds. These cloud systems are so complex that they use various kind of huge policies to meet the user requirements.

Example in our routine work we use Gmail accounts which is one of the Google cloud example. In Google cloud all the details related to your Google account, Google docs, Google maps, calendars get stored. To access this data you just need an internet connection by sitting anywhere you can access this data directly from your desktop. As with the invention of Chrome book a new example using cloud you can share the hardware from cloud. Chrome book are the laptops which have only storage to run the chrome OS, which is necessary to run web browser. With this all you can do is online such as apps, media, and storage.

Cloud computing is nothing it just storing and accessing your data from the internet rather than hard drive. It is not storing your data on local drive or on local network to whom you have physical access and it is not about having network attached storage hardware. It is accessing the data on the internet at very high speed and low cost where there is massive data processing taken place on the other side of the internet. So you just need an online connection then cloud computing can be done from anywhere at anytime.

The large scale data centers are called as clouds. These cloud systems are so complex that they uses various kind of huge policies to meet the user requirements. As the popularity

of using cloud services increases these datacenters becomes the backbone of any countries economy, communication, business and online consumer services. From small computers to large mainframe servers these datacenters are used for data processing, data storage devices and networking machines. These datacenters are the main source of electricity consumption.

In 2013 a survey conducted in US identified that the annual energy produced by 34 coal; fired power plants was equal to power consumed by 3 million computer rooms in the New York city for two years. The report from Natural Resources Defense Council finds that 30% of servers are remain idle, which are the main source of data center energy wastage [1]. As most of the servers farms hosting like Amazon, Facebook, Google uses US data centers which results in high increase of economy of country and making the lives of people easier. But if this use of datacenter goes on increasing without controlling the energy wastage then till 2020, the electricity wastage by datacenter annually will be equal to 140 kilowatt per hours which will be equal to the production of 50 large power plants which will results in emitting nearly 150 million metric tons of carbon pollution.

1.1 Evolution of cloud computing

Cloud computing is not so old technology. It evolved through number of phases including grid computing, distributed computing. But the idea of intergalactic Computer Networks called the need of cloud computing which were given in early 60's by J.C.R Licklider. His idea was to interconnect the people on globe to access the programs and data at any site from anywhere. Later John McCarthy feels that computation should be delivered as a public utility. Cloud computing was just the way to extend the services offered by internet [3].

The evolution of cloud computing took place in early 1990, when telecommunication services expands their business from point to point data circuits offered to user to the virtual private network services. From virtual private networks, the resource sharing concept comes to existence which reduces the cost for telecommunication services and increases the speed. Then in late 1990 the cloud computing has expands, its field from

technical to economical, means computing on renting basis where in 1999 salesforce.com provided its first enterprise level application to end users.

In early 2000, Amazon start using cloud computing by providing its web based retail services in 2002. Before it Amazon was utilizing only the 10% of its datacenter capacity at any given time which were serious problem for companies because there can be any time high spikes in capacity needs. To remove this and utilizing the data center capacity, Amazon starts using the new cloud computing infrastructure.

In late 2000s, after Amazon, Google start using new technology by launching Google Docs services for document sharing widely across the globe directly by end users [2].

Table 1.1: Evolution of cloud computing

Year	**Vendors**	**Concept**
2000-2005	Doct-COM Bubble burst	It leads t the introduction of cloud
2006	Amazon	Amazon launched its Amazon web services
2007-2008	Google, Microsoft, Cisco, MySpace etc.	Due to dissatisfaction of people from cloud theses new vendors join the cloud.
2008- 2009	Microsoft , Rackspace	Popularity of private cloud increases due to the security issues of public cloud .
2009- 2010	All who are interested in private cloud	Introduces the concept of EUCALYPTUS which is compatible with Amazon EC2
2009-2012	Netflix	In this period Netflix joined Amazon's AWS which leads

		to generation of Hybrid clouds
2012-2013	Amazon	Increases the demand for IAAS and PAAs

1.2 Cloud Computing

According to the NIST National Institute of standard and Technology definition of cloud is "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. Cloud computing is one of the way to expose both consumer as well as providers business application capabilities. It reduces the cost of owning the infrastructure by giving the services to consumers on pay as u go basis, and for providers it provides benefits of renting the charges from consumers of using their services. There is a agreement between provider and consumer which is called Service level Agreement to make sure to consumer that they will be delivered what they demanded. The major difference between grid computing and cloud is the use of virtual machines which allow the sharing of hardware among different applications running on single machine.

1.2.1 Cloud Computing Characteristics

The NIST has defines the 5 major characteristics of cloud computing which are:

i. On demand self services: In the cloud environment you will be served for what you have paid i.e. in this the consumer can directly increase or decrease his requirements of using services and infrastructure without the human intervention of cloud provider. Because it is working on the pay as u go basis and the SLA is billed between the consumer and provider.

ii. Broad network access: This characteristic defines the network accessibility of cloud. In cloud the consumer can access software, network, storage, services and infrastructure from anywhere at any time just by having an online connection. This is done by the private, public and hybrid clouds. This features attractive for running the business applications because the employee is not with customer all the time, using cloud he can be in touch with his customers during off times also.
iii. Resource pooling: In cloud the consumers have direct access to all the available physical as well as virtual resources of the provider. It can be network, storage and infrastructure. These resources are shared between number of different users using the multi- tenant model independence of the location of resource and consumer.
iv. Rapid elasticity: The cloud is very flexible and scalable. The capabilities can be increased or decreased at any time. You can add or delete the user, resources etc.
v. Measured services: In cloud you will pay only for that what u used. In this there is proper mechanism to track the resource usage by the consumer and they are paid for that only.

1.2.2 Cloud Computing Services

Cloud computing has a potential to change the IT community by transferring local data storage and infrastructure to large scale centralized access. It is the fastest growing option for renting of computers and storage infrastructure services for centralized platform building for customization for business processes for renting of business applications as whole. There is mainly three kinds of services provided by cloud computing are as following:

Software as A Service: In SAAS the consumer can access only the provider's application through a web browser. The consumer does not have any access to the other resources of provider such as network, storage, servers, operating system. The application is installed on the provider's system; the consumer uses it on pay as u go basis.

Platform as A Service: It provides consumer the facility of deploying its application created using provider's libraries, programming languages and tools on the provider's cloud infrastructure. But it has limited access to the cloud infrastructure related to only deploy applications.

Infrastructure as A Service: In this the consumer can access widely, the network, storage, computing and operating system services of cloud to deploy and run software. But it will not manage and control the cloud infrastructure.

1.2.3 Deployment models

Cloud computing is vast technology it uses 4 deployment models. To host cloud into your organization you can choose any deployment model which matches to the requirements of your organization.

Public cloud: It is the kind of cloud in which services are delivered to customers over the internet for public usage. It is the real hosting of cloud, in this there is no distinguish ability and control over the infrastructure for customer. The main technical difference between public and private cloud is the security level provided for services given to public or private subscribers. It is mainly used for business applications providing SAAS. In this the applications are shared by large group of people so it is economical to use. The cost can be shared by users or the provider can provide it in free. Ex is Google.

Private cloud: The private cloud provides some kind of security to the users through firewalls which control the access to data from unauthorized users. It is owned by some by some particular organization and managed and operated by the user of that organization. But in private cloud , there should be some backup plan for natural disaster and internal data theft.

Hybrid cloud: It is the integration of two or more cloud environment. Sometimes it is not feasible to use only private clouds and on the same time due to some security constraints we can't use the public cloud. In such cases we use hybrid cloud which can expand or shrink its services as the workload demands get changed. It is provider's decision that which resources he wants to make private and which are public. Like an e-commerce site,

security and scalability are two major concerns but for broachers site, security is not so critical, so we can place it on the public cloud which is more economical. Hybrid cloud has features like scalability, flexibility and security.

Community cloud: The cloud infrastructure is hosted generally for a particular kind of community such as bank, researchers. It is shared between similar kind of organization because the users understands the security, performance and privacy concerns. It can be hosted internally or externally and can be managed internally or by a third party provider.

1.3 Research Issues in Cloud Computing

Security and privacy: One of the major research issue that cloud computing is facing is maintaining the security and privacy of data stored on cloud. This is the major concern for medical, bank and government institutions which are accessing the cloud facility. During virtualization also security is the major problem faced because there can be the chances of leakage of data when you are migrating one VM from one host to another. Moreover the service level agreement defined between consumer and provider does not define any safety standards. It defines the security issues such as encryption, decryption, confidentiality, integrity, attacks etc. the much research is needed to overcome these issues through efficient mechanism and algorithms.

Virtualization: In virtualization we create number of virtual machine running on single hardware or we can say that many software are running parallel on single operating system. In virtualization the major issues to be resolved are how to allocate the infrastructure to different machines, when to migrate the machine from one host to another, balancing the number of VMs running on the system, allocating the resources to VMs, how to allocate the process to different VMs etc.

Server consolidation: Server consolidation means utilizing to the best of the active server. It is the process to maximizing the resource utilization of each server. The technique used in this is the live virtual machine migration in which VMs are migrated from overloaded hosts to under loaded hosts. It is also used for the energy saving of cloud. To make server consolidation efficient, migrating virtual machines, activating and inactivating the servers are main issues to be solved.

Energy saving: Establishing a tradeoff between energy saving and application performance is also one of the major research issue. As the use of cloud computing getting popular among people the demand for data centers also getting increased. These data centers are the large source of energy consumption, according to survey in US in 2006 the energy consumed by datacenters was 1.5% of the total energy generated and it is likely to be going 18% annually. So the research is going towards green cloud computing. Various approaches are applied to reduce the energy consumption such as switching of the idle servers etc but still this field has some deficiencies which need to be resolved to make the cloud computing fully green cloud computing .

QOS: In cloud the service level agreement is defined between consumer and provider in which the user demands are mentioned and the charges received by provider for those services. If the demands will not be fulfilled then the penalty charges are made by user. So Automated resource provision is one of the key issue to meet the quality of service requirements. In this there is automatically allocation and reallocation of recourses occur according to user current and future demands of application. For this application performance models are constructed for each application used by user. The main techniques used were statistical learning, queuing theory and control theory [7], these techniques can be modified to make application performance model efficient.

1.4 Virtualization

Virtualization is the software through which we run number of different applications using different operating systems on single hardware. It is the fundamental element of the cloud computing. The difference between cloud computing and virtualization is that virtualization is a software which manipulates the hardware and cloud computing is the service which uses the results of this manipulation. It is implemented through live virtual machine migration technique. In simple way we can say that virtualization is the process of creating virtual version of something.

Virtualization hosted architecture has following components:

Host operating system: It is the actual real operating system working on the host. It has the full control of the hardware of the host.

Virtualization layer/ hypervisor: It acts as the intermediate between the host operating system and the guest operating system running on the host. It provides the illusion to each guest operating system that there is no other operating system is running on the same hardware. The hardware is shared between different operating system in transparent manner through this hypervisor.

Guest operating system: These are the guest operating systems which are sharing the hardware of the host. The systems implement the virtualization concept.

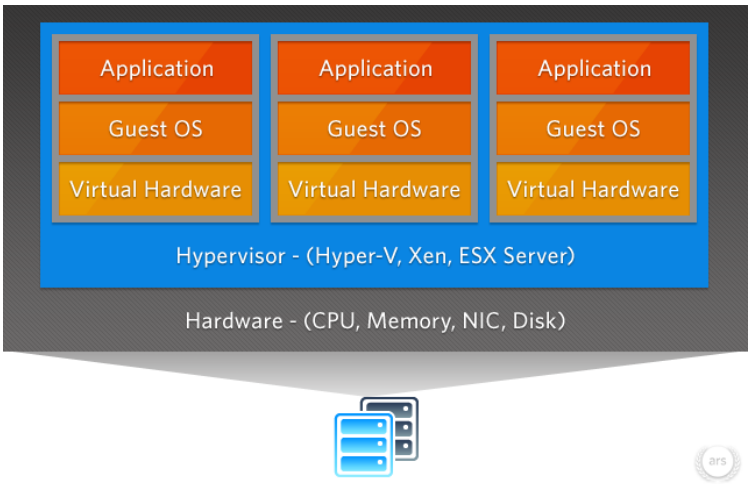


Figure 1.1: Virtualization Architecture [4]

1.4.1 Need of Virtualization

Increasing server utilization: Server consolidation is the process of reducing the active number of servers by live migration of virtual machines. In this we migrate the load from over loaded server to under loaded by server so that server utilization can increase. After migration the ideal servers are switched off.

Saving energy and cooling cost: As the idle servers are increasing the energy consumed by these data center are also increases. The 53% of the total cost of operational expenditure of the datacenter, get wasted on the powering and cooling infrastructure. Virtualization is the the technique which decreases the cooling and powering cost by transferring the physical servers to virtual machines and consolidating them.

Managing datacenter's space and reducing data center footprints: It is related with saving the energy of data center, it also reduces the datacenter footprints using virtualization. Because it provides storage, server and network virtualization to the data center which reduces the data center floor space required, so the cost of owning the data center also get reduced because it require lesser number of servers and storage and network infrastructure through virtualization. Storage virtualization is used to manage the data center storage because the data getting stored and processed on these datacenter servers is big data, the virtualization process is called by these data centers to solve the large storage problem.

Improve disaster recovery: To tackle with any kind of disaster the virtualization has mainly three ways. One is the hardware abstraction, by making running application independence to any particular kind of hardware vendor or server model, during any disaster we can easily use the cheaper hardware placed in DR storage. Second is the server consolidation, by reducing the number of physical servers required we can easily afford replication site if required during any disaster. Third one and most important is the server virtualization which provides software for any disaster recovery which can automate the any disaster recovery process.

Extend lifetime of older applications: Sometime we have such applications which are too old to run on new operating systems and hardware. Moreover the companies or vendor whoever created these are not now ready to update them, then at that time by virtualization and encapsulating the application and its environment we can extend the lifetime of these application.

Help moving things to cloud: The fundamental element of cloud is virtualization, when you started implementing virtualization on servers and underlying hardware, that means you started towards using cloud. Virtualization is the first step towards private clouds. After that when you started using public cloud you will get more comforts by moving your data from outside of your datacenter.

1.4.2 Types of Virtualization

Network virtualization: In network virtualization the single hardware and software network resources are divided in to logical virtual networks by dividing the single bandwidth into number of channel which can transfer data independently for particular server or device.

Storage virtualization: It giver virtualization of storage, in this the data from multiple storage networks gives illusion that it is stored in single storage device which is managed by central server.

Server virtualization: It is the process of hiding the details of server resources from the users running applications on the server. It provides the masking of server resources from user. It can be of operating system masking, hardware masking or Para virtualization.

Operating system virtualization: It gives the illusion to the different users running applications on the guest operating system that there is no other application running the base operating system. It has virtualization layer which isolates the actual operating system from guest operating system. The system get divided into containers where each container maintain the details of process tables, files, storage and network details of each application. The container has a software interface which gives isolation to different applications. but it has restriction that the containers and applications should use the same operating system as base operating system. In this multiple copies of guest operating system get created.

Hardware emulation: It provides the hardware virtualization to the guest operating systems through hypervisor. Hypervisor is the intermediate between the actual operating system and the guest operating system, it intercepts the calls from guest operating systems and allocates the hardware resources accordingly. There is no restriction on container and base operating system, the hardware emulation can run on any operating system. In this virtual machine manager is used which is created by hypervisor. Guest operating system and virtual machines are together stored in this.

Para virtualization: It is the combination of both operating system virtualization and and hardware emulation. It has hypervisor layer but other tools and drivers are also installed. In this the guest operating system can use the hardware directly or it can use the drivers of virtualized environ

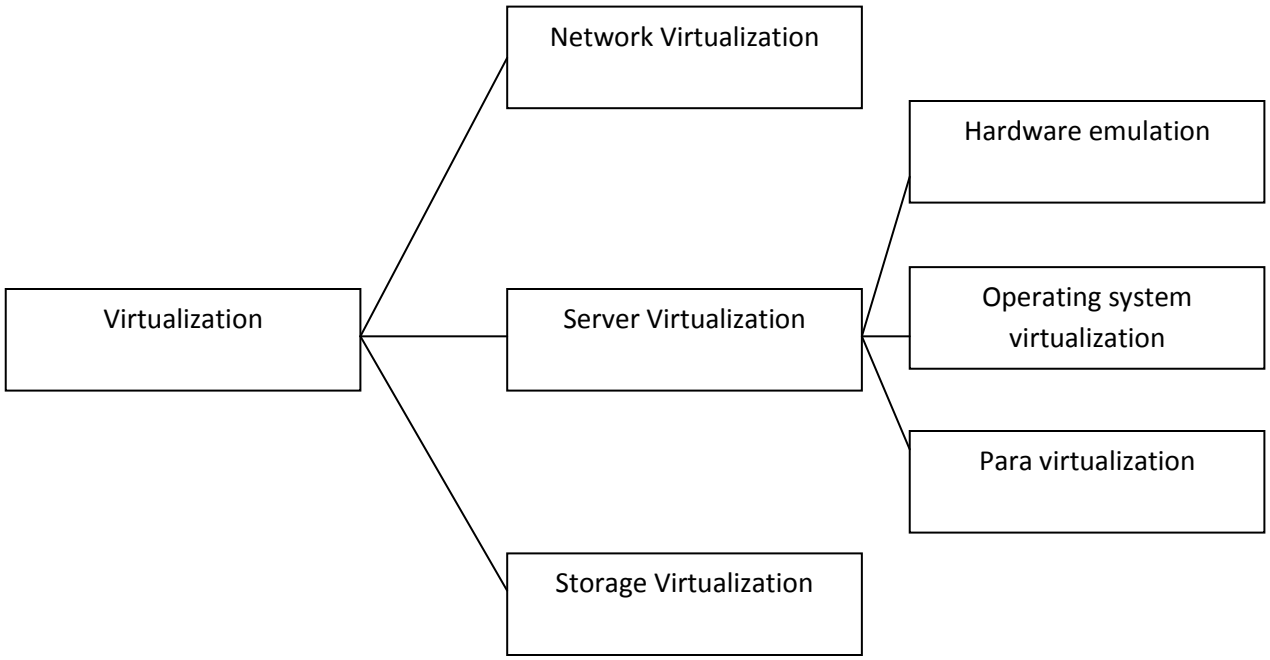


Figure 1.2: virtualization types

1.4.3 Virtual machine migration

Migration is the process of transferring the virtual machine from one host to another. It is responsible for hardware virtualization. The virtual machine manager handles this process of migration. Migrating and copying a VM is different thing, in copying a VM we create the new VM but in migration we move the VM from the old host to new host.

The migration can be off different types as:

Cold migration: Moving a virtual machine from one host to another by powering off it on the source machine is called cold migration. VM can be transferred between different CPU families; the shared storage does not require performing migration. We can migrate VM between different hosts as well as datacenters.

Warm migration: It is also called migration of suspended VM. In this VM is suspended from the host 1 and moved to host 2 by copying all the disk storage and register files across hosts and continuing on destination after some time interval. It can be moved between datacenters but the restriction is that the CPU compatibility requirements should be met n both source and destination side. No shared storage is required to move VM from one host to another.

Live migration: It is also called vMotion migration. The running VM is moved from one host to another by interrupting its availability. In first step copy the storage files such as RAM and disk files from source to destination and then mark dirty RAM pages and recopy these pages, in last make the final move of VM and suspend it for a less then one second to make the final copy. VMs can be moved only within hosts not within datacenters. CPUs on both side should be compatible and shared storage is also required for the migration [5].

Table 1.2: Virtual machine migration types [6]

Migration type	**VM power state**	**Migration between hosts or datacenter**	**Shared storage required**	**CPU compatibility**
Cold	Off	Hosts or datacenter or both	Not	Source and destination can have different CPU families
Suspended	Suspended	Hosts or datacenter or both	Not	CPU on source and destination should be compatible

vMotion	On	Hosts	Yes	CPU on source and destination should be compatible
Storage vMotion	On	Datacenter	Not	N/A
Enhanced vMotion	On	Hosts or datacenter or both	Not	Source and destination should be CPU compatible

1.4.4 Need of Virtual Machine Migration

Load balancing: By migrating VM from one host to another we can balance the load between overloaded and under loaded hosts. In VM migration when VM load goes beyond the VM threshold value that means the host get over loaded, then we move the VM to the host which has VM load value below the threshold value i.e. which is underutilized.

Maintance: Moving VM from one host to another before that host get shutdown helps in maintance and servicing of the source host. There can be many reasons like underutilization or power failure, so in that case rather then keep running the VM on one host we can migrate it from one host to another host.

Power saving: VM migration is one of the techniques which is used in power saving. The hosts which are under loaded , VM migration migrate the VMs from these hosts to the other hosts which can bear the load , so that theses under loaded hosts can be switched off to save the power consumption by idle servers. It keeps minimum number of active servers.

Resource utilization: it improves the resource utilization by efficiently utilizing the server to its maximum. It migrate the load from over utilized to underutilized servers, so that the server can be utilized efficiently and the idle servers can be switched off.

1.5 Energy efficiency

Data centers can host large number of servers so the power consumption by these servers is much greater. A data center spread across 500 square meters can consume power more than the 3500 households [9]. The major energy consumption was due to inefficient server utilization. The main source of data center power consumption is infrastructure load and IT loads such as the power consumed by servers, network management, disk drivers, cooling purposes, lighting etc. out of all theses the major research issue in today is the power consumption by servers. To get the energy consumption by the servers we have to identify first the load on the server and the power consumption of the servers. The servers are classified into various classes shown in table.

Estimating the power uses for servers is very complicated because each server varies in its hardware configuration such as can have different memory installed, different disk drivers, different usage etc. so the power consumption is calculated through the maximum load expected when the server is fully utilized. It is the product of load and the power use per unit. The total energy consumption (E) is calculated as the function of power consumed in terms of CPU utilization u over the time period t (1).

$$E = \int P(u(t)) \qquad (1)$$

The total power consumption by server also includes the power used off the cooling the equipments. Electricity consumption for servers from 2000 to 2005 is get doubled, each year it has growth rate of 14% for US and 16 % for the world. Without involving the electricity used for cooling purpose total electricity consumption in US is 23 billion kWh in 2005 and including for cooling purpose it will go up to 45 billion kWh[10]..

From past decade here is significance improve in the data center energy efficiency as many other server firms operated by companies such as Google, Facebook get involved in these data centers but it contributes only 5% of the total data center efficiency. There is

some small and medium sized firms and multitenant data center are lacking behind from data center energy efficiency benefits.

Table 1.3: Estimated U.S. datacenter electricity consumption by market segment (2011) [1]

Segments	**Number of servers(million)**	**Electrical share**	**Total U.S. data center electricity use (billion kWh/y)**
Small and medium server rooms	4.9	49%	37.5
Enterprise or corporate datacenter	3.7	27%	20.5
Multi- Tenant datacenter	2.7	19%	14.1
Hyper scale cloud computing	0.9	4%	3.3
High performance computing	0.1	1%	1.0
Total(rounded)	12.2	100%	76.4

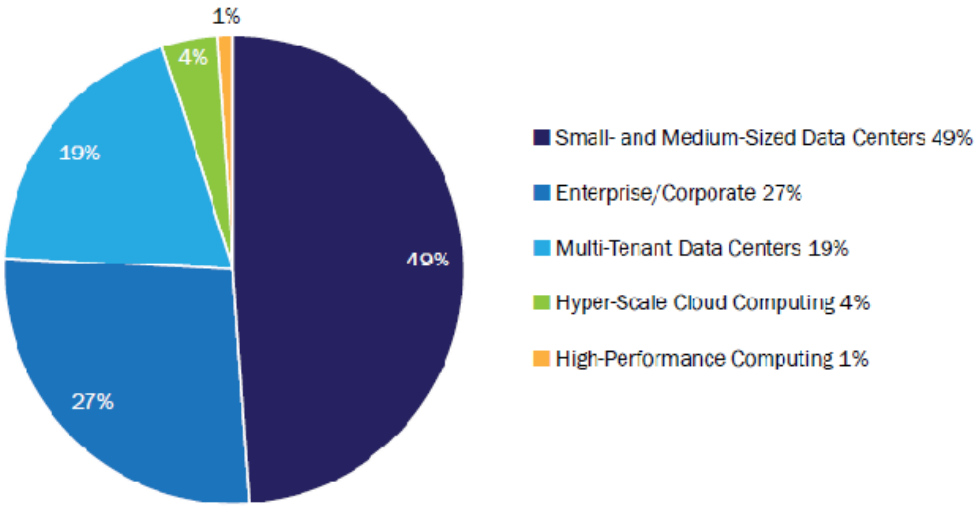


Figure 1.3: Estimated U.S. datacenter electricity consumption by market segment (2011) [1]

1.5.1 Reasons for inefficient use of energy by datacenters

The major reasons for persisting the data center energy efficiency are:

Ideal servers: The 20 to 30% of the servers in data centers are placed unused which are not performing any work but these are consuming the electricity 24/7. The organizations started with the peak provisioning factor that is installed servers according to the peak demand but during the actual processing time most of the servers get wasted due to under loading of workload. These under loaded servers are placed as ideal, the ideal server consumes 60% more power than the server utilizing at its peak. Some data centers does not power down these under loaded server because they have fear that once these get off will take long time to restart.

Due to high initial cost of Energy efficient equipments: The small and medium sized firms avoid using the energy efficient equipments such as those certified as Energy Star via Environmental protection Agency due to the high initial cost associated with these. These energy models have significantly improvement on the long term maintance and operation cost and lifetime for these equipments. The multitenant data centers have high

priorities for keeping the cost low and maintain high levels of security, reliability and performance, due to this the energy efficiency is undermines.

Exacerbated split incentives: In multitenant data centers the one who are paying the bills and those who are actually using the IT equipments are working for different companies. Due to lack of contact, it is difficult to know that which equipment is pricing higher. Due to this customer interest get decreases in investing the energy efficient equipments [1].

1.5.2 Green Cloud Architecture

As from the above mentioned study we found that –data centers are not only expensive, these are environment unfriendly also. The cloud providers should provide some mechanism so that the profit they gain from cloud computing should not degraded due to inefficient usage of energy by these data centers. Many governments are also working to reduce the carbon footprints due to data center, to manage the greenery of the environment. As the demand for data centers goes on increasing the number of disk drivers and servers working also goes on high, due to these It is very complicated to control the energy consumption of data centers.

The Green Cloud Computing is the mechanism in which the cloud services are provided in fastest and efficient manner with minimum energy consumption. Existing energy efficient techniques concentrate to minimize the energy consumption, but the green cloud computing consider the dynamic resource requirements of the customer. The target of this is to minimize the total cost of ownership and maximizing the return on investment. The cloud computing does not meet the customer quality requirement with respect to energy savings. The green cloud computing not only allow the services to be allocated to customers on demand bases but also to increases the revenue intake through energy efficient pricing and utility based policies.

Green cloud architecture elements:

To address the problem of energy efficient resource allocation we started moving towards green computing which uses the green cloud architecture. The main elements of the green cloud computing architecture are as:-

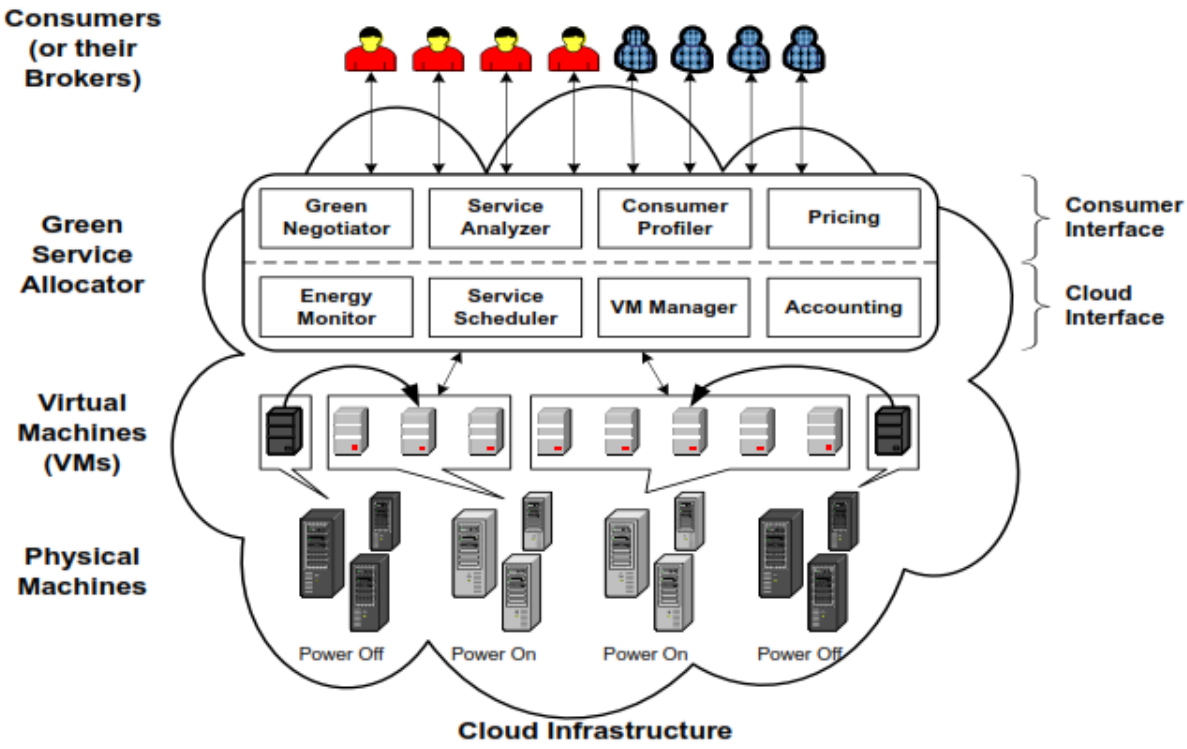


Figure 1.4: Architecture of Green Cloud Computing

Consumer/ broker: Cloud consumer can be the person or the organization which uses the cloud service provider services to interact with cloud computing. Consumer creates the business relationship and makes the agreements with the cloud provider for setting up the service. The consumer and user may not always be the same. The end user is the one who is actually accessing the services which are provided by cloud consumer. For example a company is the cloud consumer who is deploying a service on cloud and the end users are the employees of the company who are accessing that service through cloud.

Green resource allocator: To access the cloud infrastructure the cloud consumer should interact with cloud provider, the green resource allocator acts as intermediate between the cloud provider and the cloud consumer. It has the following six components to perform all the function:

i. Green negotiator: it sets the service level agreement between the cloud provider and cloud consumer to give quality of service. The pricing and penalties are set in case of the violation of SLA agreement. The agreement is based on the

requirements and QOS of the cloud consumer. Such as for web application, for instance the QOS metrics set should be the completion of 95% requests in less then 3 seconds.

ii. Service analyzer: it will analyze all the service requirement details for the request coming and then it will decide whether to approve the request or reject. For this it will consider information regarding VM load , energy consumption by latest etc.

iii. Consumer profile: in this the consumer profile is checked, to identify the different kinds of consumers. According to the details the different consumers are granted with special privileges and priorities.

iv. Pricing: it decides the chargers for different services allocated to consumer. To manage the resources needed for that service and prioritizing these services according to pricing.

v. Energy monitor: it observe the physical machines that which should be power on or off to save the energy consumption by idle servers.

vi. Service scheduler: it schedules the different services to different VMs according to VMs load and decides when to add or delete the VMs. It decides resource requirements for different VMs.

vii. VM manager: it is responsible for the allocation of VMs and analyzing the resource requirements of VMS. It is the main component for migrating the virtual machine from one host to another.

viii. Accounting: it keeps track of resource usage and the cost associated with the resource usage, so that later when the resource request occur the user can be charged according to the previous resource usage. By using the historical information of resource usage, the service allocation policy get improved.

Virtual machines: Numbers of virtual machines are placed on single physical machine. Multiple VMs are running concurrently to serve requests on different operating systems using the single hardware infrastructure. It provides the dynamic resource provisioning by providing the facility of maximum flexibility to configure resources due to portioning of resources on one physical machine. To manage the load on different VMs the migration of VMs take place so that underutilized hosts can be turned off.

Physical machines: these are the actual hardware on which the multiple machines are running. These provide the infrastructure to green cloud and support the virtualization [8]

Chapter 2

Literature Survey

Currently datacenter consume 1.3% of global electricity which is going to be 8% by 2020.in Mckinecy report the cost of electricity bills worldwide due to datacenter was $11.5 billion [11]. The main source of electricity consumption is inefficient server utilization of datacenters because the idle server consumes 60% more power than the server running at peak. The following are the three main energy efficient approaches for cloud computing datacenters:

i. Resource management by switching off idle server and reducing cooling infrastructure for controlling heat dissipation.
ii. Reduce the need of equipment replacement by ensuring permanence of infrastructure
iii. Increasing equipment utilization as computational load is geographically distributed.

The heat dissipation due to these datacenters is the reason of 2% of the global CO2 emission. It increases the operating cost of datacenter as it increases the number of cooling systems required for heat dissipation. In order to reduce the total cost of ownership of these data centers we have to make these data centers energy efficient. The main parameters of interests for cloud service provider are resource utilization details, infrastructure response time, virtualization metrics and transaction metrics [12].

Virtualization is used for energy saving in datacenters through virtual machine migrations. We run number of virtual machines on single physical machine to perform multiple tasks from multiple users which reduces the need of number of physical machines. Virtual machine migration migrate the VM from over utilized server to underutilized server in order to minimize the number of active servers. To perform virtualization the parameters to be considered are number of VMs required by workload,

time taken to initialize new VM, migration time, time taken to allocate additional resources to VM [12].

VM allocation and consolidation

VM allocation is the process of allocating the VM to host initially and consolidation is the prices of optimizing that allocation of VM through VM migration. VM consolidation provides significant benefits to cloud computing by facilitating better use of the available datacenter resources. In static VM allocation, the hypervisor provides resources to VMs based on peak load demand. But this is not the efficient way because the workload requirements keeps on changing. In order to meet the demands of workload, we need to reallocate the VMs dynamically through live virtual machine migration. The main characteristics of VM migration problem are:

i. The workload demands in number of VM are dynamic which results in allocation and removal of VMs.
ii. The resource requirements of particular VM keep on changing.
iii. The cloud provider can use the VMs of its own datacenter or can rent the VMs from external cloud providers.
iv. Migrating VM from one PM to another takes longer time and energy consumption in terms of increasing the load n source side.

Virtual machine migration follows for steps:

i. Initially allocate the host to each VM i.e. virtual machine allocation.
ii. Identify the host which get overloaded or under loaded using some statistical methods and select the VM to be migrated from victim host i.e. virtual machine consolidation.
iii. Perform the migration process by selecting the best suitable migration techniques.
iv. The last but not the least is to select the destination host where the migrated VM has to place. The wrong section of host can lead to excessive power and energy consumption and inefficient resource utilization.

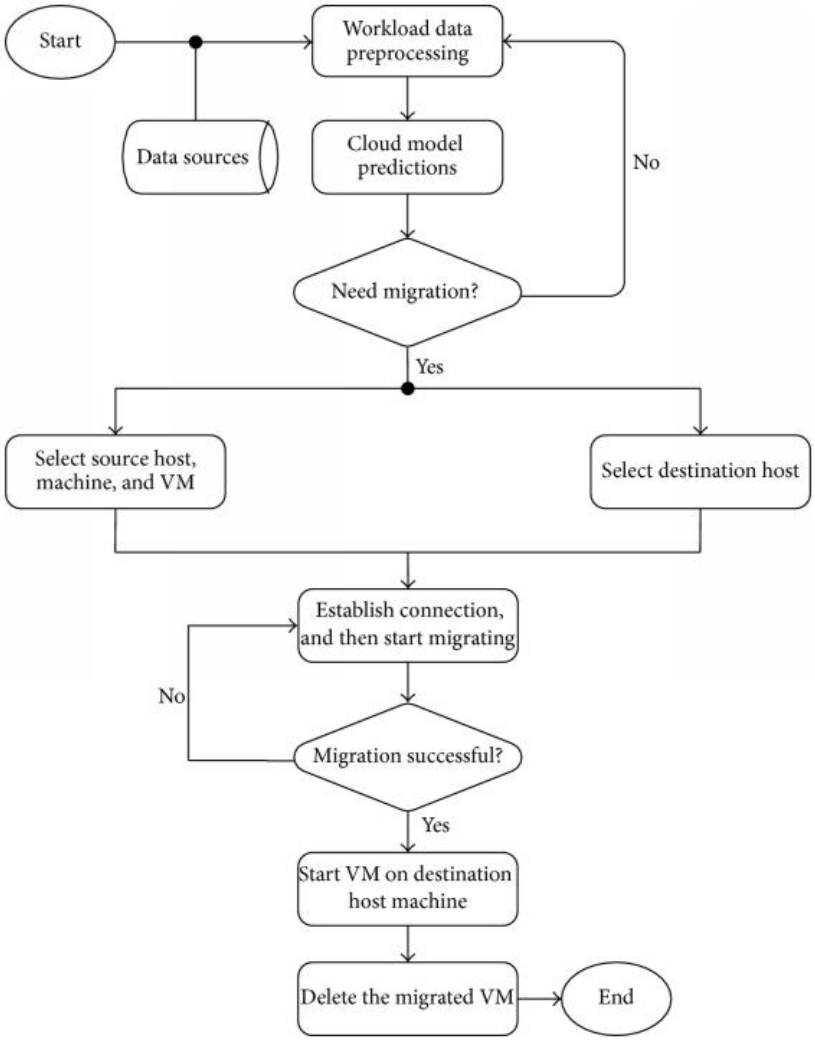


Figure 2.1: Flow chart for Virtual Machine Migration

VM allocation and consolidation problem is an NP hard problem. Initially we place the VM using some bin packing algorithms or VM allocation techniques. After this to utilize the resources efficiently VM consolidation is applied. VM consolidation can have two major goals [13]:

Power saving: in power saving goal VM consolidation target is to minimize the energy consumption and utilizing the resources efficiently. The QOS can be compromised.

QOS: in this the target is to fulfill meet QOS requirements by compromising the energy saving and consumption.

2.1 Virtual Machine Allocation

In virtualization VM allocation is the very first step of virtual machine migration, in this the initial VMs created are allocated to various host. The various techniques are available for the VM placement on hosts such as heuristic based, constraint based, bin packing problem, genetic algorithm, stochastic integer programming etc.

Heuristic based approaches

In heuristic based approaches there are various techniques author gave in [14, 15, 16] such as first fit, volume based fit, single dimensional best fit, etc. these are identified by different kind of resources considered during VM allocation. The first fit approaches are the greedy approach. In this to allocate the VM sequentially all hosts are searched and the one which has sufficient resources will be allocated. The scheduler start searching the PM whenever the request for new arrival of VM occurs, the first PM which has enough resources to run the VM will be allocated. If there is no such PM found which can satisfy the resource requirements of VM, then the new PM is activated and the VM is allocated to this new one. This technique is very simple but the side effect of using this is the unbalancing of load. It does not consider any mechanism to calculate the load on PM before allocating the VM to these hosts.

The next heuristic based approach is the single dimensional best fit. It is one dimensional means only one resource is considered before allocating the VM to host. The PMs are ordered in decreasing capacity of that particular resource which may be CPU, memory and bandwidth. The VM is allocated to the first PM which has maximum capacity for that resource. But due to this again the problem of load unbalancing occurs because the resource considered in one dimensional. In cloud the resources are multidimensional. The load unbalancing occur in such situations when the one host which is good for CPU utilization, but may be other resources such as memory and bandwidth are underutilized.

In volume based best fit the volume of PM is considered to place the VM. To calculate the volume of PM sandpiper [17, 18] automated system is used which considered the

sand volume ratio of PM to detect and mitigated the hot spot. The sand volume formula considers the normalized values for the CPU, memory and bandwidth. The PM are ordered in decreasing order of their sand volume, and the VM is allocated to the first PM. The problem with this approach is that it combines all the resource to calculate the sand volume ratio means the resources are converted from 3D to 1D which may sometimes chooses the wrong PM. Because the PMs which are equal in sand volume ratio, will be equally considered for VM placement which is wrong because the individual resource utilization will not be equal.

Another heuristic based technique is based on dot product. Two vectors are created one is for the resource requirements of the VM and the other one is for the resource utilization or total capacity of the PM for individual dimensions. The dot product is calculated using Vector Dot method [19] for two vectors and then the PMs are arranged according to the dot product . PM which has lowest dot product will be selected for the VM placement. Sometimes this method also chooses the wrong Pm because it does not considered the remaining capacity of the PM. We need such method which can balance the resource utilization such as if one PM has high utilization of CPU and low utilization for PM in that case we will allocate that VM which need high memory requirements and low CPU requirements.

Constraint based approaches

In constraint based approaches there are some specific constraints applied for the VM placement. These are used to solve the combinatorial search problems. The main constraints are capacity constraint, SLA constraint, placement constraint, quality of service constraint. In capacity constraint all the resources are considered to calculate the total capacity of the PM. The VMs are allocated in such way that after placing the VM the resource utilization of VM corresponding to all dimensions should be less then the total capacity of the PM. In SLA constraint the VM will be allocated to those PM which will not violate the SLA requirements. According to placement constraint the VM should be allocated only to the available hosts. In QOS constraint the main target to place the VM is the fulfillment of quality of service attributes such as reliability, availability etc.

the restriction on using the constraint based approaches is predetermination of VM load [20].

Bin packing problem

Placing the VM on PM is NP hard problem which is solved by using the bin packing problem solution. In bin packing problem we have items as well as containers, the problem is to place the maximum number of VM in single container [21]. In virtualization similar concept works jus we have VMs on place of items and PMs on place of container. The difference between both is that in bin packing we can place the VM on one another but in virtualization it is not possible to place the VMs on one another. The various algorithms using this are:

Power aware best fit decreasing: it is kind of first fit heuristic technique, but in this before allocating the VM on host it calculates the increase in power for each PM after allocating the VM and that PM is allocated which for which there is minimum increase in power consumption occur. In global power aware is also similar to this but in this the increase in power is calculated for whole datacenter rather then only for single host.

Stochastic integer programming

This technique is used to solve such problem for which data is unknown but it can be predicted using some distributions. Similar is the case with VM placement where we don’t know the resource demand of VM by using stochastic integer programming we can predict the demand of VM and will be able to find the suitable host which results in less energy consumption [12s].

Static placement algorithm

It is identifies that sometimes the actual VM requirements are less than the half of the maximum value. There are two algorithms correlation based placement and peak clustering based placement. In correlation based placement we will place those VMs on different hosts which have correlation between the loads. It works on the logic that higher the correlation between loads of different VMs then higher the chances of overloading. It is the extension of pMapper algorithm.

In peak clustering based placement we will perform the clustering of VMs. In this Different VMs are clustered according to their peak loads. The similar kind of VM clusters are distributed among PMs [23].

Ajith S. et al[24] gives banker's algorithm for VM allocation in energy efficient way. Sometimes during VM allocation, deadlocks can occur. In cloud computing VM allocation which provides VMs as infrastructure as a service, allocates VMs on demand to users. Due to this sometimes, two or more VM wait for the same resources and deadlock occurs. T avoid this Bankers algorithm is used which detects the safe and unsafe state before actually allocating resources to VMs. It reduces the energy consumption and SLA violation and VM migrations are minimized. It avoids the deadlocks distributed systems. For SLA it gives better results than IQR used for overload detection.

2.2 Virtual machine consolidation

It is the process of migrating the VM from one machine to another in order to consolidate the servers. After initial placement of virtual machines the consolidation is performed. Out of the four steps mention in virtual machine allocation and consolidation above, the three steps will be performed in the virtual machine consolidation i.e. detecting the overloaded or under loaded virtual machine called virtual machine selection, performing the virtual machine migration called VM migration and last is to select the destination host for placing the migrated VM. We will discuss all the three parts as following:

2.2.1 Virtual machine selection

In VM consolidation process the first step is to select or detect the candidate VM to be migrated. To find it we check the host which is over utilized or underutilized then VM from that host is selected. Now to get host is over utilized or underutilized various approaches are used for different resources. A particular pre calculated value is used to detect the hot utilization which is called the threshold value, any host running above this is called over utilized and running below it is called the underutilized. Some systems set two threshold values one is upper threshold and the other is lower threshold. The host whose CPU utilization is above the upper threshold is considered as over utilized and the

one whose CPU utilization is below the lower threshold value is called underutilized which need to be migrated to other host so that the host can be shutdown. The thresholds can be calculated in two ways one is the static threshold and the other is dynamic threshold. In static threshold the single pre calculated value of CPU utilization is set as the threshold value which remains same for all the time. It is not the feasible way because the load on machine gets changes over the time so utilization of CPU also changes. So dynamic threshold is used to calculate the threshold value according to the load for dynamic workload consolidation.

Optimal online deterministic algorithm

Anton B. et al[25] in optimal online deterministic algorithm gave adaptive utilization threshold scheme to set the dynamic threshold for CPU utilization. These techniques use historical data to auto adjust the CPU utilization. First is the median absolute deviation, which is the statistical method it will set the dynamic threshold according to the deviations in CPU utilization. According to this if deviations in CPU utilization are high then chances of overutilization of hosts are also high so it will set the threshold value to some lower value, so that overutilization can be detected. The other statistical method is the inter quartile range which sets the upper threshold based on the statistical dispersion. Some other methods are local regression and robust regression. In local regression we identify the trend line of previous utilizations. So it uses the historical data and by putting some values in the trend line equation it will predict the future utilization to set the dynamic threshold value. The robust regression is the extension to linear regression.

There are various policies to select the VM for migration after finding that the host is overloaded or under loaded such as Anton Beloglazov et al [26] proposed some techniques such as minimization of migration policy, the highest potential growth policy and the random choice policy. In minimization of migration, it selects the minimum number of VMs to be migrated. It also sets the two utilization levels for each host i.e. upper and lower. To select a VM it considers two criteria first is the current utilization of VM should be more then the difference of host utilization and upper threshold value. Second is increase in current utilization of destination host should be minimum.

The highest potential growth policy selects those VM which has the lowest CPU utilization as compared to the actual capacity defined by VM to decrease the overall utilization of host. In random choice policy based on some random number, the subset of VMs to be migrated is chosen randomly. The random number is uniformly distributed variable.

The maximum correlation policy selects those VMs which has higher correlation. It works on the principal that higher the correlation between VM running on host higher will be the overloading of resources. So it select such VMs for migration. The correlation is identified by calculating correlation coefficient between VMs.

Zhibo C. et al[27] in dynamic VM consolidation proposes the improved MC policy for VM selection. It provides the range of correlation coefficient between -1 and 1. The positive correlation leads to the overloading of hosts because in this both parameters increases and decreases simultaneously. In positive correlation there is positive relation between variable and in negative correralation coefficient there is negative relation between variable,

2.2.2 Virtual machine migration

After selecting the VM to be migrated from over utilized host next step is performing the VM migration. In VM migration the running VM is migrated from one host to another in transparent manner such that there should be invisible shutdown of VM occurs. The virtual machine monitor performs this task. During VM migration there is different kind of state need to be transferred such as virtual devices states i.e. CPU, motherboard disk etc, network connections for VM to be migrated with devices and physical memory of VM.VM migration process consist of mainly theses steps [28]:

i. Identifying the source and destination for VM migration.
ii. Initiate the migration by pre copying the VM memory state from source to destination.
iii. Transfer the control of CPU state from source to destination by terminating at source and resuming on destination. The termination time should be less.

iv. Identifying the changed content on destination from source and transferring these again to completely shut down the VM on source.

Pure stop and copy

The virtual machine migration should have two goals. First is to minimize the downtime during VM migration for which the services are not available second is the total migration time which includes synchronizing the source an destination hosts[29]. Christopher C et al [29] studied the two approaches pure stop and copy and pure demand paging. In pure stop and copy approach the VM on source is stopped and all the memory pages are transferred to destination. In the end the VM on destination is started. It is the simple approach but it has limitation of large downtime.

Pure demand paging

In pure demand paging the VM on destination is started and the pages are transferred from source to destination as the demand occurs. In starting only OS data is sent using stop and copy approach to start the VM on destination. Due to this downtime get decreased but the performance get degraded and the total migration time also gets increases.

Writable working set based live migration

To overcome these problems author suggested new technique which combines these above techniques. It uses the Writable Working set which consist of the set of pages which are mostly get dirtied. There are some pages which often get modifies so these can be transferred using the pre copy approach. In the end to send the writable working set it uses stop and copy approach for short duration. To handle it in more efficient way we can add dirty bit in page table for those pages which get updated mostly. Those pages which has dirty bit set will be send in the stop and copy phase and are the candidate for writable working set.

Improved pre copy based using LRU technique migration

Ei P. Z. et al[30] gives improved live migration of VM as the total migration time get increases in pre copy approach. To reduce the total migration time it introduces a one more phase i.e. the pre processing phase during live migration of VM. In this pre processing phase those pages are identified first which are updated frequently to reduce pre copy approach overheads. It uses LRU least recently used and splay tree algorithm to identify the working set. In this improved technique those pages which get used recently will be the candidates for working set and will be transferred in the last. To apply LRU it constructs the stack which consist the list of recently used pages on the top of the stack. Splay tree algorithm is used to group the memory pages based on process id which results in fast prediction of working set.

Author [31] gives the optimized live migration for memory intensive applications to improve the performance degradation due to live migration. Optimized approach the migration time is increases but the downtime is decreases which ultimately improves the performance. In the first face all the pages are marked as dirty which get modified. In second face those dirty pages are transferred again to destination and the dirty bit is reset, it can be set again if any modification takes place. The second phase estimates the migration time, it is the longest pre copy phase. The page copying process continues till the transfer rate to destination host does not exceed the pre defined value. In end the pages which are modified but not copied are transferred by shutting down the VM on source machine. This time of migration is called the downtime which is inversely proportional to the migration time.

The improved pre copy

Author in [32] gives the improved pre copy based live migration of VM. In improved technique we use page bitmap in which frequently updated pages are stored and will be transmitted only in the last round. The page bitmap sets the bitmap bit if it found any page which gets modified in previous iteration and rather than sending it again, it will send it in the last round which results in less migration time. Because it assures that the

modified pages should be sent only once in the process, due to this the transmission data and total transmission time get reduced.

Compression based VM migration

As in previous migration techniques we identified that the data to be migrated is large as number of iterations are high which results in network overheads. Author in [33] suggested the memory compression based live migration of virtual machine. In which to reduce the amount of data transfer memory compression is used. The pages to be transferred are firstly compressed using adaptive zero aware characteristic based algorithm for compression and the on the destination side these pages are decompressed using same decompression technique. Due to small data transfer the network traffic get reduced but compression time introduces some overheads. To remove such overheads we use word similarity technique for compressing similar kind's words in a page. This algorithm has only benefit is the compression time is less than the data transfer time. To decrease the compression time it uses some regularity between pages like pages with high intensity of zero bytes. Such pages are stored in LRU cache memory for future references and decompression.

Post copy VM migration

Another technique for live migration of VMs is introduced by author [34] is post copy VM migration. In this the VMs CPU states are transferred first then the memory states are transferred. It eliminates the limitation of pre copy approach which increases the number of duplicate pages on the destination host and increases the transfer overhead of data. In post copy it is assured that each page should be transferred only once. So it is based on purely demand paging, in which the CPU states are transferred on destination host, each page is transferred as the demand for that page occur on destination host. It increases the initial number of page faults which can be overcome by pre paging, pre paging is the technique which predicts the future page demand and transfer it directly on destination before its access. It also introduces dynamic self ballooning technique which avoids the transfer of free memory pages which occur in pre copy approach. This technique of post copy is useful for write intensive applications.

2.2.3 Virtual machine placement on destination

Makhloug Madji[35] proposed online algorithm for server consolidation in cloud data centers which is based on b matching theory. The main objective is to eliminate the SLA violations and to minimize the cost by providing the optimal amount of resources for particular VM. It provides the best tradeoff between convergence time which is 5 seconds for 3000 VMs and 9000 PMs. To reduce SLA violations it increases the number of servers required, so energy management is not efficient.

Ant colony based energy aware placement

Louis, et al[36] give the ant colony based energy aware workload placement approach which were based on Ant colony Based approach. Till now mostly the workload consolidation approaches use single dimension approaches to place the VM after consolidation. In this author give the multi dimensional bin packing problem to place the VM. It was first kind of work in which the ant colony is applied on MBPP due to this it places VMs dynamically according to current resource requirements. The results were evaluated with the traditional greedy approaches which show that ACB gains high power saving by server consolidation.

Distributed parallel ant colony optimization algorithm

GaochaoXu et al[37] provides VM placement approach based on distributed parallel ant colony optimization algorithm. In Ant Colony optimization algorithm runs in iterative manner parallel in the distributed environment on number of hosts, to select the best host. In first iteration it will run on several hosts to give the solution for second stage. In second stage the ACO run on first stage solution to give the optimal solution from first stage it has one problem that it may get stopped before giving the optimal result sometimes when number of VMs and PMs are large. To solve this problem we can increase the number of iteration for large scale VMs. The advantage of running the ACO in parallel manner is the easy detection of any failure of physical machine and adjusting it

by giving the solution simultaneously, which results in low failure rate. But to run this algorithm all the physical hosts should be in the same and fast LAN environment.

KNN based regression algorithm

G. Motta et al[38] gave KNN based regression algorithm for energy aware consolidation of cloud data centers it is based on the statistical method of calculating the load on the k nearest neighbors. This approach has significant impact on saving the energy and reducing the number of active servers. In this to perform dynamic consolidation, the load on k nearest hosts is calculated using the k-nearest neighbor regression algorithm. It results in the list of hosts which get over utilized and underutilized. Thus theses hosts can be considered in performing the VM migration of VM consolidation. This algorithm runs in four phases. First phase detect the host which get over utilized i.e. when the predicted load get increased from the capacity of host. In second phase the VM is selected to be migrated from the host. In third phase the host is selected for placing the VM. The last is the prediction of underutilized host over which the VM has to be allocated. The migrated VM is placed on destination host using power aware best fit decreasing. It minimizes the SLA violations.

Exact allocation and migration algorithm

F Ghirbi et al[39] proposed the exact allocation and migration algorithm, which combines the allocation of VMs with migration algorithm. In first step VMs are allocated to PMs using bin packing algorithm, then the VMs are migrated using the consolidation algorithm using integer linear programming algorithm. It reduces the number of active servers using bin packing algorithm and energy saving through consolidation.

Power aware resource provisioning technique

V. K. Mohan et al[40] give power aware resource provisioning technique in which the workload dispatcher controls all the requests to server by predicting each server cost for the request and SLA constraints. It uses the LP formulation to identify the optimal server for VM request. It uses some assumptions such as idle server consumes 70% more power then active machines, power consumption is directly proportional to number of VMs

instantiated and intensity of workload directed at VMs does not effect the power consumption and CPU utilization of host. The parameters considered are peak power consumption, SLA violation, average request processing time etc. but the implementation was performed on homogeneous workload environments.

MCC and fuzzy AHP based placement

N. Kord et al[41] combines the PABFD with minimum correlation coefficient MCC concept and fuzzy AHP method to reduce SLA violations. In order to select the VM for migration it uses the local regression method to detect source host and minimum migration time to select the VM. It considers two criteria: increase in power consumption after the VM allocation to host and correlation coefficient between VM which is going to be migrated and the VMs running on target host. Because higher the correlation coefficient between resource usage higher will be the SLA violation. In the end to get the optimal solution it uses Fuzzy Analytical approach which selects the best host. In future this techniques can be applied on different workload models and reducing number of VM migration.

Enhanced weighted round robin algorithm

Alnowiser et al[42] give the enhanced weighted round robin algorithm, it keeps on checking the running VMs states to identify the over utilized processing elements and if any found then hibernate that element by sending some signal such as Wake-On-LAN. It uses dynamic frequency voltage frequency signal to set the minimum VM frequencies for each task. To meet the task deadlines the DVFS selects the best CPU frequency out of available frequencies. The energy consumption is minimized using VM reuse. it results in 49% saving in energy then greedy approach and 31% than of round robin.

ENACLOUD

J. Sekhar et al[43] give an energy saving approach called ENACLOUD. In this workloads are tightly aggregated to minimize the number of open boxes. As the workload demands keeps on changing, ENACLOUD provides workload remapping through consolidation. It has two goals one is to decreasing the number of open boxes and other is

decreasing the migration time. When the resources and sequences of workload are provided then there are 3 events that triggered: workload arrival event, departure event, resizing event.

Adaptive Live Migration

P. Lu. et al[44] proposed Adaptive Live Migration technique to improve load balancing. In this strategy firstly it collects the load values on each host to determine whether to apply the migration or not. To check the load at each host it uses Linux commands. It will perform the migration when the mean value of the sum of the maximum and minimum distances from the average utilization is greater than the threshold. Now schedule the live migration by checking the load balancing history records. It uses the previous history to select the source and destination for VM migration, if the similar kind of record exist then we use the same source and destination in VM migration as used in previous one. If several records exist them select the latest one. If no such record exists then choose the source whose load value has maximum distance from the threshold value and choose the destination whose load value has minimum distance from threshold value. It uses dirty bit map to perform migration, compression technique for generic applications and check pointing for memory intensive applications.

Measure-Forecast-Remap

`Z. A. Mann in his survey [45] founds the energy efficient solution which uses cycle of Measure-Forecast-Remap. In this the measure part determined the current workload on the host and perform the past analysis of workload. According to this it will predict the future demand of machine by using some statistical methods this is called the forecasting phase. In last from the solution of measure and forecast passes, it remaps the VMs to hosts. This was the technique identified by IBM for the workload consolidation. The limitation is that it considers only single dimension resource and work for single data center. There is time intervals set after which the load on VMs is measured.

pMapper

Another technique is pMapper[46] which were used for VM consolidation gives better energy efficient results. It also minimizes the number of migrations but again the problem is single dimensional resource it uses i.e. CPU capacity. It has three forms i.e. min Power Parity, min Power Placement with history and pMap. In min Power Parity there is increase in number of migrations. Because all the VMs are order in decreasing order of CPU utilization. It places the VM on that PM which has least increase in power consumption. To check this, it will iteratively place the VMs on PM which ultimately increases the number of migrations. To overcome this problem of number of migrations in min Power Parity, a new version of this was found which is called min Power Placement with history. In this the firstly VMs are allocated before performing the consolidation. The initial placement is based on some criteria which reduces the number of migrations during VM consolidation. The further improved version to MPPH is pMAP, in this also initial plane of VM placement is generated and for VM migration also generated but it will perform only those migrations which results in high energy efficiency.

Balanced minimum K-cut approach

Balanced minimum K-cut approach [47] is also used for VM consolidation which considers the network communication cost to consolidate VMs. Each PM has slots which are reserved for each VM running on the PM. The consolidation is done by mapping the VMs on to slots provided by PM. In this the clusters are formed of VMs according to communication intensity and slots are formed based on communication cost. Then the mapping is performed from clusters to slots and this process goes on for each new generation of clusters.

Maximum correlation coefficient and migration control based approach

Mohammad A. et al[48] gives energy aware consolidation using combination of heuristic and migration control. The heuristics technique that is used in this is maximum correlation coefficient. Maximum correlation technique works that higher the correlation between VMs running on the host, higher the possibility of overloading the host. At the

same the migration control method says that don't select the steady and resource consuming VMs as source for VM migration. The reason for not choosing such VMs is that the one which is consuming high resources now will definitely consume high resources on destination host which results in overloading of destination host. So by combining both the methods we conclude that migrate those VMs which has high correlation with other VMs on that host. It results in less energy consumption and improves the number of VM migrations. It also reduces the network traffic and number of host shutdowns. In future there is need to work on the VM placement, handling the under loaded and overloaded hosts based on real workload traces.

Performance to power ratio aware VM allocation

Xiaojun R. et al[49] gives performance to power ratio aware VM allocation . it is first kind of VM allocation which gives performance to power ratios for various hosts i.e. it maintains the balance between host utilization and energy consumption. It is called PPRGear in which we calculate mainly 11 levels of different utilization for each host which are called gears. The best gear will be the one have highest PPR value. The top n gears with the highest PPRs are chosen as preferred gears. Our target is to run the host at best gear, any host running at lower then preferred gear is called over utilized and running at lower then preferred gear is called underutilized. It has same SLA violation rate as with DVFS. The problem with this approach is selecting the number of preferred gears. If we consider the large number of gears then there we will not be able to detect the over utilized host and VM migration will not occur. Same is the case in underutilization; if gears will be large then we can't find the underutilized host due to large range of preferred gears. It was compared with static threshold based, medium absolute deviation and found that it reduces the 69.31 % energy consumption by reducing the migration and shutdown times.

CPU overhead aware VM placement

Pushtikant M. et al[50] defined median absolute deviation as the median of absolute deviation from the data's median for a univariate data set in their energy efficient VM allocation policy. It uses the last outcome and load for estimating the workload and VM

allocation unit for a given load. It suggested a new VM allocation policy by calculating variation and gives a new formula for calculating the median absolute deviation.

Multiobjective VM placement algorithm

Ali P. et al[51] gives two phase multiobjective VM placement algorithm along with dynamic migration technique for geo distributed data centers. It considers data and CPU correlation. Data correlation refers to data dependency between two VMs and CPU load correlation depicts the CPU utilization coinciding during a certain time interval. So due to this data center management is difficult. The VMS with high correlation are clustered together, while high data correlation VMs should be placed apart. This research gives energy performance tradeoff by combining CPU load and data correlation. It gives two phase controller along with migration technique which divides VM placement problem into clustering and allocation phase. In global phase the VMs are clustered according to data and CPU load means highly data correlation are clustered together and high CPU correlation are placed apart. In local phase VMs of each cluster are allocated to server of their corresponding data center and optimal frequency for each server is computed. The green allocator after placing VMs on servers manages the energy source at each datacenter.

VM placement and migration scheme

Thuan D. et al[52] gives Joint VM placement and migration scheme which does not only reduces the energy consumption but can also the cross network traffic. It balances the load among hosts by minimizing the number of active servers. Minimizes the communication among hosts in order to save energy consumption by network devices and links. In JPM algorithm to minimize the energy consumption of data center, we keep record of VMs currently being served and VMs waiting to be served. The optimization process is iterative in nature, at every optimization iteration we keep the VM placement or migration decisions that have been made in previous iterations. To optimizing the communication cost Graph Portioning Based Rank Minimization is used because the communication loads and attraction among VMs is shown by graph. An optimal solution based on inter server communication can place the VMs on different hosts. If

optimization scheme places all the VMs on single host, in that case inter server communication cost and unsatisfied attractive force among VMs get reduced. It is multiobjective as it considers both cross traffic and power consumption and jointly optimizes these. But there is still not any particular solution to the problem of load balancing among different zones.

Network and power aware VM placement

Kwonyong L. et al[49] found in their research that most of the research lack to identify the CPU overhead required for the communication between VMs. Sometimes the network bandwidth allocated to VM can't be fully utilized due to lack of CPU resources of PM required to meet CPU overhead for communication. This work will identify the exact PM which has sufficient CPU resources to meet CPU overhead. A linear regression based algorithm is used to identify the CPU overhead required to allocate necessary network bandwidth. To calculate CPU overhead it uses split drive model given by Xen. To calculate the bandwidth requested for each VM it uses ipref tool. This algorithm calculates the network bandwidth capacity for each PM and for virtual cluster. The candidate key will be the one which ha remaining capacity for bandwidth after allocating the currently running VMs more than the network bandwidth requested. This approach is reactive and static in nature.

Simulated annealing based VM consolidation

As virtualization become popular in cloud computing, its main target is to reduce the consumption and making the cloud infrastructure cost effective. Antonio M. et al[50] in his research introduces a new technique which is based on heuristic technique i.e. on simulated annealing for the consolidation of VMs. The proposed model is based on mixed integer linear programming and simulated annealing. In this the initial parameters to be considered is current allocation of VMs to host, resource requirements of each VM, number of active hosts. In consolidation it normalizes the total initial power of servers to the linear combination of power consumed by active servers and normalizing number of migrations to number of VMs. It gives cost effective model which calculates VMs

migration based on resource utilization and energy efficiency of server. It can also be extended to calculate the network cost and network power consumption.

Gamal E. et al[51] gives efficient resource utilization technique for VM consolidation. In resource utilization technique the consolidation is based on CPU utilization variance. During consolidation of VM, it will calculate the host utilization variance for each migarable VM on the allocated host and other hosts. This step is repeated for N times for each migrable VM to allocate on N hosts and check the CPU variance. Now check the minimum utilization variance of each VM on corresponding hosts allocated to these VMs. It calculated the live migration cost, SLA and energy consumption.

Network power aware approach

Till now most energy efficient techniques are based on server utilization through CPU. But there are some network issues occur during VM consolidation which needs to be resolved. Suheib A. et al[53] suggested a technique which considers network overhead during VM migration and reduces the network traffic. Network power aware approach uses two models during VM consolidation i.e. power model and network model. The power model is used to detect the over utilized and underutilized host based on some threshold values. The network model added to this selects VM based on transmission throughput. The transmission throughput is number of bits transferred in particular time between two VMs. It will trigger VM migration when this throughput is less then some predefined value. The destination host will be the one which has least transmission time to transfer one bit from one VM to another VM. If it selects the source machine on power model based then the destination host will be the one which has enough resources to run the VM. It reduces the total cost and increases realibity and utilization of cloud.

Chapter 3

Problem Formulation

To make the datacenters energy efficient we use the concept of virtualization in cloud computing so that maximum number of users can access the services at the same time using minimum actual cloud infrastructure. Sometimes the requirements of one user process cant be satisfied by the current resources allocated to that virtual machine, in such cases rather then waiting for the long time for resources the VMs are shifted to some other host which has enough resources to run that VM efficiently. So this whole process is done by Virtual Machine migration technique. In virtual machine migration. Various virtual machine migration techniques are studied in the literature survey, which includes the virtual machine allocation policies and virtual machine consolidation policies.

3.1 Research Gap in virtual machine Migration

The major issues to be resolved in virtual machine migration are:

i. Energy efficient migration: in virtualization migration is performed to make the data centers energy efficient as it balances the over utilized and underutilized servers through live migration of VMs. But migration itself is a overhead because large number of migration results in large data and CPU state transfer which results in network as well as power consumption overheads. Due to large migrations sometimes the performance of application also gets degraded as the downtime increases during migration. So to make the data center energy efficient we need some tradeoff between the number of migrations and the number of active servers.
ii. Number of active servers: virtual machine consolidation is used to reduce the number of active servers to make the data center energy efficient. As in literature survey we studied that the idle server consumes 60% more energy then the server which if fully utilized, we need some techniques which can shutdown these

servers when they are underutilized. VM migration should perform migration from over utilized server to underutilized server. Sometimes VM migration results in minimum number of active servers which ultimately increases the load on available servers which get over utilized easily. So there should be some tradeoff between load and the number of active servers

iii. Total migration time: as the virtual machine migration balances the load on server by compromising the performance time of application. During virtual machine migration the migration time is the time during which the VM memory states are copied from source to destination and the downtime is the time when the machine on source get stopped and the CPU states are transferred from source to destination. Numbers of migration techniques are listed in literature survey which are used for migration. These techniques have different migration time and downtime. The pre copy approaches has migration time less but downtime is high which degrades the performance of application. The post copy migration technique has small downtime because the CPU states are transferred before memory states. But it creates the overhead of page faults. So to minimize the overall migration time of virtual machine migration we need some appropriate technique according to application requirements.

iv. Selecting source and destination: To migrate the virtual machine there should be need of migration which can be identified by checking the servers utilization. If server is over utilized means some VMs should be migrated to the server which can run these VMs. To shutdown any server, check its utilization if it is underutilized then running VMs on this should be transferred to other server and it should be completely shut down. To identify the over utilized and underutilized server we use some pre calculated load values which can be handled by server efficiently. These load values are called threshold values. Static and dynamic threshold values are used according to workload demands. Selecting the appropriate technique for calculating threshold value is one of the research issue in virtual machine migration.

The existing system is based upon the PSO [54] based VM migration. The existing model utilizes the particle swarm optimization method for the process migration decisions

between the virtual machine resources in the cloud. The existing system has been deployed as the swarm intelligent solution for the process migration and consolidation solution. The inter-VM process migration method using the PSO considers the reduced number of active physical nodes. The reduces active physical node based VM migration system is intended to increase the load over the selected VM resources which may cause the unexpected delay and overloaded VMs. Also the particle swarm optimization is the swarm intelligence algorithm, which severely suffers from the particle optimism problem, which forces it to rely upon the favorite particle irrespective of its velocity and direction. Also the degree of the load balance is decreased due to the limited or reduced number of active physical nodes in the active cloud environment.

3.2 Objectives

The objectives of this thesis work is :

i. To study and analyze existing VM migration and consolidation techniques and models.
ii. To propose an energy efficient model for VM migration and consolidation.
iii. To implement and validate the proposed model for energy efficiency.

3.3 Methodology

The methodology used to achieve these objectives is:

i. The existing methods, techniques and models for VM migration and consolidation are studied and analyzed.
ii. Energy efficient model for VM consolidation :
    a. Uses the concept of cross pollination theory for virtual machine migration.
    b. The cross pollination theory with match based weight age can solve the problem of optimism.
    c. It can improve the load balance degree to achieve higher performance than MPSO.
    d. Concept of static threshold is used.

e. Compare energy consumption with threshold values and number of virtual machines.

iii. To implement and validate the proposed model for energy efficiency virtual machine migration simulator workflowSim is used and to get simulation results netbeans is used. The parameters considered are load on physical host, CPU utilization, overall response time of workflow execution .

Chapter 4
Proposed technique CPO

In order to improve the MPSO based VM migration model, we should reduce the load over the individual resource by using the resource load migration rather than giving the overload to the lesser number of the resource nodes as per suggested in the existing model. The PSO mechanism suffers from the problem of particle optimism, which causes the less exact at the regulation of its speed and direction. The particle optimism carries the drawback in the case of scattering and optimization mostly and rules out the major problem solving because its drawback of creating the non-coordination systems such as solution to cloud energy efficiency by VM migration. The cross-pollination theory with match-based weightage can solve the problem of optimism problem. The weighted cross-pollination can overcome the problems created by the use of PSO. Also the load balance degree can be improved by using the cross-pollination theory to achieve the higher performance than the MPSO solution proposed in the exiting model.

Our technique cross pollination is based on the natural phenomenon of flowser pollination. Flower pollination is the process of making new seeds by transferring pollen grains from one flower to another of same species. The pollen grains are transferred by the use of pollinators which are animals. Pollinators plays an important role in fertilization process. It can be of two types i.e. self pollination and cross pollination. In self pollination, the plant does not need pollinator for the fertilization process, it can grow by yourself. In cross pollination, pollination process requires external pollinators to occur suc as water, wind etc. in this the pollination process occur between flowers of different plants. The cross pollination process is used to solve many computing problem such as scheduling, load balancing etc. in cloud computing. The cross pollination technique used in our thesis solves the problem of virtual machine allocation by managing the resources energy efficiently..

Cross Pollination Algorithm

Step 1) all the jobs coming for processing are stored in job queue. Calculate the length of the queue.

Step2) apply the random function on queue length and generate the initial bee.

Step3) for each initial bee, extracts job from job queue and assign to the virtual machine .

Chapter 5

Experimental design

The proposed model is aimed at solving the problem of the energy consumption in the cloud platforms using the CloudSim simulator. The VM (virtual machine) migration is to migration the task of a VM to another VM. Task of placing the runtime tasks in the memory in the perfect placement or sequence in order to reduce the total tasking and communication overhead in terms of load. The VM migrations the major part of the cloud architectures to improve the performance of the cloud platform. The VM management faces the major challenges from the bias-free dynamic resource allocation while keeping the cloud performance on the maximum in terms of execution time and computational overhead. The proposed model is balanced vm migration scheduling algorithm with the intelligent technique of cross-pollination for the high-performance to reduce the energy consumption among the cloud platform.

5.1 Simulation Environment

Cloud computing provides services such as infrastructure as a service and software as a service etc. In these services virtualization is used in which number of VMs run on the host the provider can host application services on the cloud which can be utilized by the users. But evaluating the performance of theses services before actually deploying is one of the most difficult and complex task. Cloudsim provides a simulation tool kit for for the modeling and simulation of application services. It can model the datacenters, Vms and resource allocation policies. It provides to major benefits time effectiveness and flexibility during application modeling. By evaluating the performance of application on cloudsim the time get reduced because it takes less time then evaluating the performance in original cloud environment. It also provides the flexibility because the changes can be easily made in the application service if it does not satisfy the required functionality.

The main features supported by cloudsim are:

i. Modeling and simulation of large scale cloud datacenters on single physical host.
ii. It provides platform for the simulation of allocation policies and service brokers.
iii. Automatically simulate the network environment with the simulation system elements.
iv. Simulate the private and public cloud element into the federated cloud environment.
v. Provides virtualization engine to create and manage the virtualization services.
vi. Provides simulation for energy aware computational resources.

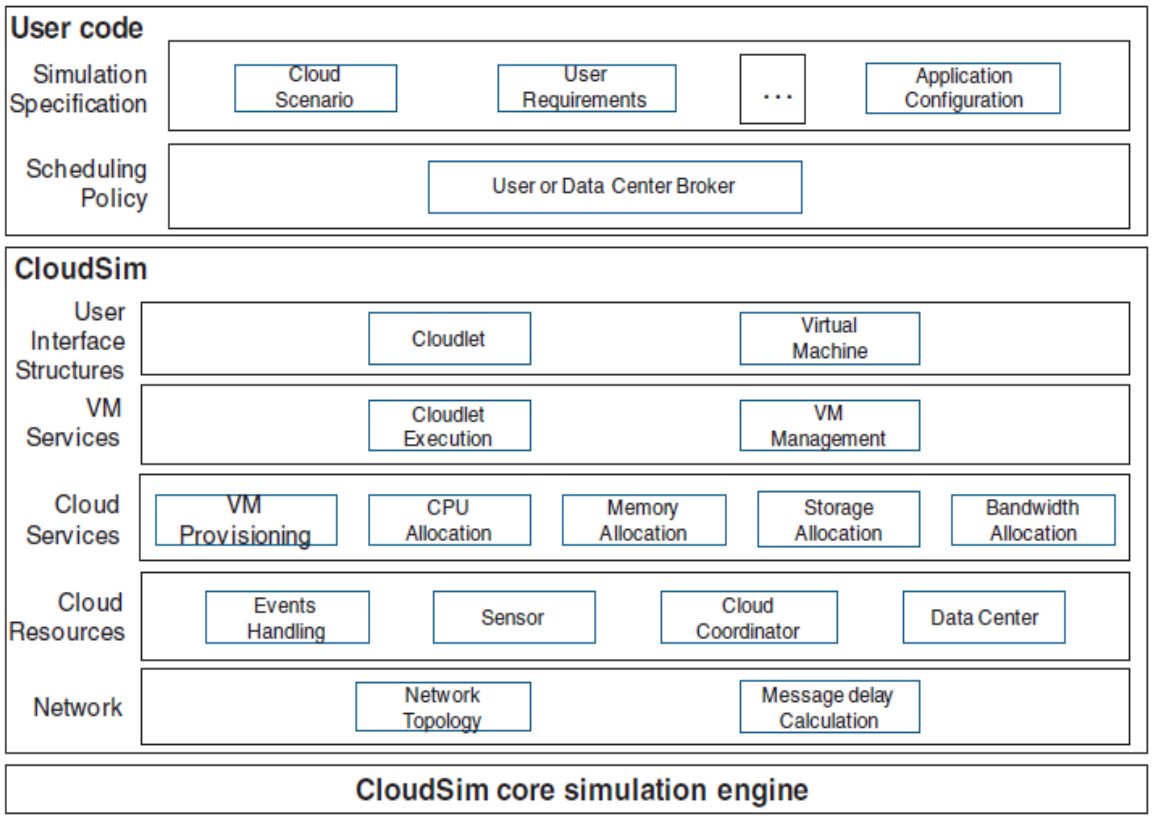


Figure 5.1: CloudSim Components[54]

Cloudsim simulator layer is responsible for the modeling and simulation of virtualized environment including VM provisioning, network bandwidth etc. to implement your own extended functionalities the cloud provider will extend programmatically to core VM

funcationality at this layer. The top most layer is the user code which specifies the basic functionalities for hosts, VM, applications etc

5.2 CPO BASED OPTIMAL VM MIGRATION ALGORITHM

In this paper, the optimal load sharing approach based on the CPO (cross-pollination optimization) has been proposed for the load balancing approach over the cloud environment in the case of VM migration. The path A defines the resource 1 and path B defines the second resource. The other resources can be assigned with the further alphabets with the assumption that all of the resources are logically capable of executing the given tasks in the cloud computing environment. The resource selection must be done on the basis of availability CPU processing powers, which must make the whole process efficient in terms of response time. The traditional methods are known to allot the random resources for the given task, which dent the performance of the cloud scheduling model and hence slower down the query processing procedure, while results in the higher response time. The cross-pollination is the probability-based procedure to choose the appropriate resource in the available list of VMs. The proposed model is aimed at lower the task response time in order to maximize the number of tasks processing in the span of one milisecond. The proposed has been made capable of subdividing the task, which facilitates the quicker process and processes the smaller tasks faster than the hefty ones to reduce the overall load and to increase the number of successful request processing every second. The subdivision of the tasks is based on the length of the task. A task is usually divided in the t slots, where t is smallest time unit available for the task length calculation in our proposed model. A task smaller than or equal to t will be processed in one round, where the tasks larger than t can be scheduled in queue or on different VMs according to the load and time calculation for the faster processing. The random proportional rule is used to recognize the ratio of task in processing on the given VM has been presented in the following equations:

$$P_1 = \frac{(R_1+k)^h}{(R_1+k)^h+(R_2+k)^h} \qquad (2)$$

$$A_1 = P_1 * T_{Ri}$$

Where A_1 is the number of tasks assigned on the resource A, P_1 is the probability of the resource, R_1 is the pheromone value based on the available ratio of RAM and CPU on VM under consideration, Tr_i depicts the resource availability required to process task i. The k and h are the coefficients used for the choice of probability among the available resources for the scheduling of the tasks among the available resources. The value of k and h is calculated on the basis of the VM load and resource availability on all of the available VMs. The variation in the values of k and h will define the variability on the basis of the current processing load on the different VMs, which inspires the task assignment decision of the CPO algorithm. The used rule for the probability calculation has been given in the following equation:

$$P_j = \frac{(R_i + k)^h}{\sum_{i=1}^{n}(R_i + k)^h}$$

In this thesis, the Meta tasks are used for the testing of the proposed model. The Meta tasks does not carry any dependency on the other tasks in the processing queue, which means the response time will be calculated for the each individual task by calculating the difference between the finish time and start time. The waiting time is also considered as the response time delay, which is caused due to the waiting period spent in the queue. The proposed model has been described in the detail in the following steps:

ALGORITHM 1: CPO ALGORITHM

1. The input parameters for the CPO (cross pollinated optimization) method are initialized.
2. The available VM list is loaded into the runtime memory with CPU details.
3. The available resource list is updated with the VM list prepared on the step 2

$$VM_l = \{V_1, V_2, V_3, V_4 \dots \dots \dots V_n\}$$

4. The resource capacity is calculated for each VM and has been given as the following list:

5. Loop is started for every resource

    a. Read and load the available resources on each VM and update the resource list with current values.

$$VM_R = \int_{i=1}^{N} VM_i$$

$$VM_i = \frac{VM_{CPU_u}}{VM_{CPU_t}}$$

$$VM_R = \{VM_1, VM_2, VM_3, VM_4 \dots \dots \dots VM_n\}$$

    b. The seeding pheromone value is initialized for the particular VM.

6. End the loop on step 5

7. Load the task list into the runtime memory.

$$T = \{t_1, t_2, t_3, t_4 \dots \dots \dots t_n\}$$

8. Calculate the length of the tasks in terms of CPU burst time.

9. Subdivide the each task j and create the sub-task list where each sub-task is assigned with i.

10. Scheduling iterations are initialized for the number of tasks or sub-tasks with i.

    a. Calculate the probability value Pi for each VM.

    b. Calculate the VM Load in the form of resource usage percentage using the following formula, where Aj depicts the availability of the VMs

$$A_{ij} = P_{i1j} * T_{Ri}$$

    c. Compare the task length with probability value and resource usage on all available VMs.

    d. Shortlist the list of available VMs which can process the given task i.

e. Assign the task i to the VM with minmum response time and highest probability.

$$\text{If} \quad T_c(i) < Vc_j$$

$$VM_L(MAX\ P_i) \leftarrow T_c(\text{i})$$

f. Update the resource list with the current values of VM load

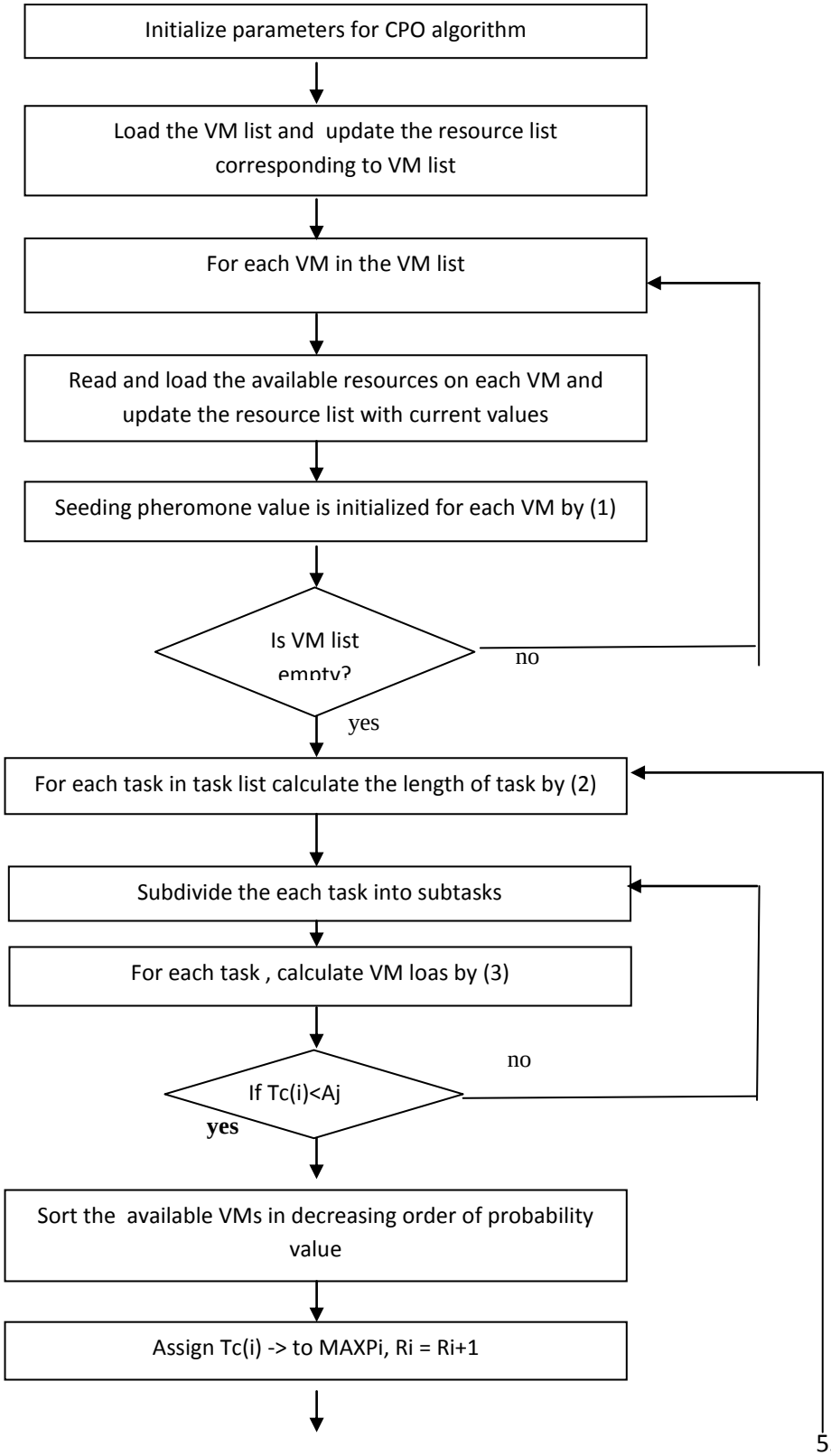
g. Update the pheromone value for each available resource.

$$R_i = R_i + 1$$

h. Go the step 8 if not end of task list.

11. End the iteration and process the tasks.

FLOW CHART OF CROSS POLLINATION OPTIMIZATION ALGORITHM



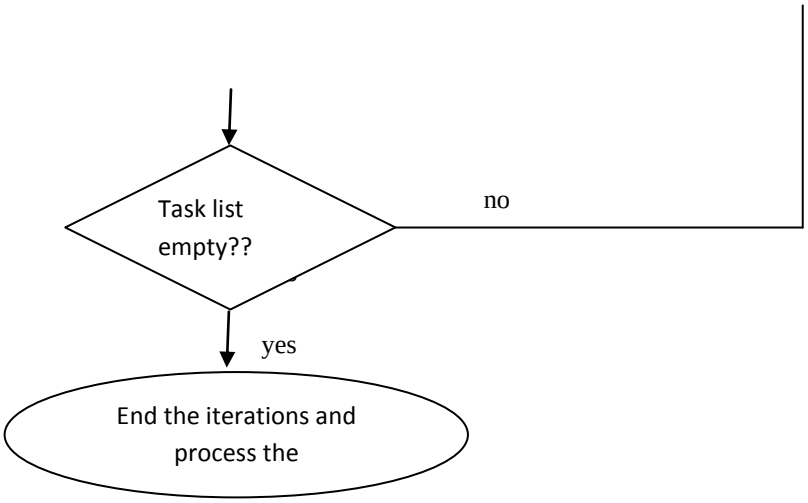


figure 5.2: flow chart of cross pollination optimization algorithm

Chapter 6

Testing and results

6.1 Test Plan

As the cross pollination algorithm is designed to reduce the energy consumption during VM migration. The proposed algorithm is tested by comparing it with some of the existing energy efficient algorithms. Such as Modified particle swarm optimization: this algorithm defines the fitness function to make the approach energy efficient as the total Euclidean distance between resource utilization and energy consumption. It removes the problem of falling into local optima of particle swarm optimization.

The test plan designed to confirm the cross pollination algorithm considers the similar input data for both the algorithms. The two parameters considered are energy consumption in terms of number of VMs and threshold values. If the proposed algorithm performs better then the modified particle swarm optimization algorithm for these parameters with same input data, then we will consider our algorithm energy efficient. To validate the results input values are changed such as the energy consumption at different threshold values is calculated. The overall response time for all the process during each iteration is also calculated. The parameters to be compared are number of VMs, threshold values and energy consumption, overall response time.

6.2 Test results

To test the algorithm we run cross pollination on cloudsim which consist of 500 virtual machines. The threshold value also varies for the data input. The total number of process are 10 which run in iterations. Both the algorithms run for same data input and the results are calculated for energy consumption in terms of number of virtual machines and CPU utilization.

The first comparison is done between number of virtual machines and the energy consumption for each. The energy consumption increases with increasing number of virtual machine on the system. As the number of datacenter goes on increasing, the energy overhead also goes high. To make the data center more energy efficient our algorithm CPO provides better results then the MPSO [55] algorithm. We started from 50 VMs, as the process starts execution the VMs number get increases which goes till 410. For all the different number of VMs at different instances the energy consumption is calculated for both MPSO and CPO. From results CPO gives better results then MPSO. Table 6.1gives the data of energy consumption with respect to number of virtual machines corresponding to each algorithm.

Table 6.1: Testing result for energy consumption for different VMs

No of Virtual Machines	Existing	Proposed
50	50	46
100	75	70
150	125	122
200	160	155
250	200	193
300	225	210
350	260	254
400	310	304
450	350	341
500	410	406

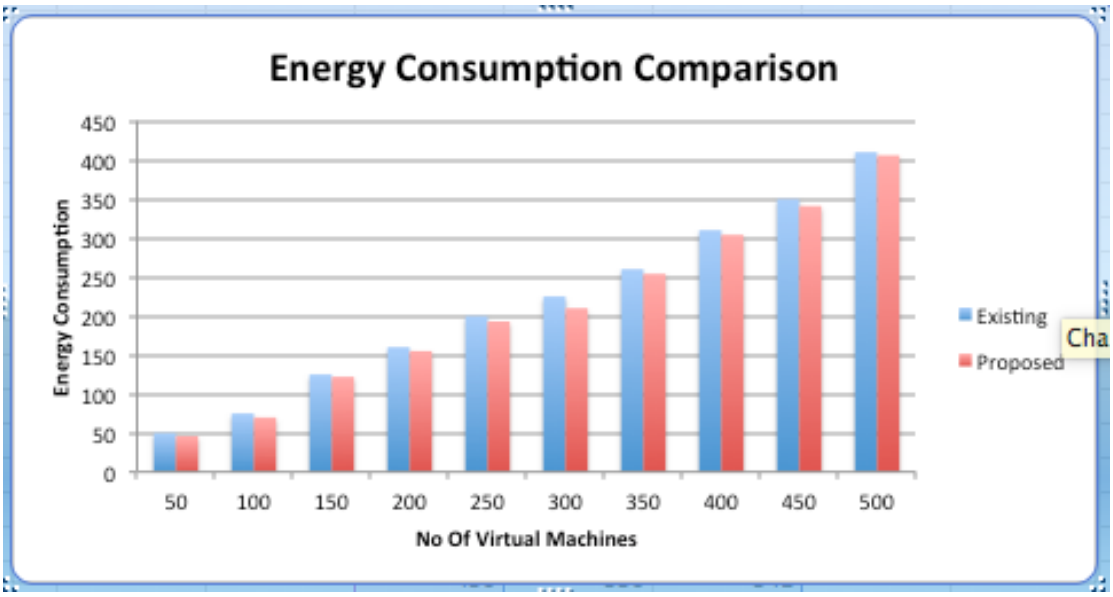


Figure6.1: energy consumption for different number of virtual machines

Energy consumption get varied as the CPU utilization varies. At different CPU utilization system has different energy consumption. By calculating the energy consumption at different CPU utilizations we can choose the best CPU utilization for the virtual machine migration process. In our thesis we choose the threshold value from 0 to 0.9 and calculated the energy consumption for each threshold value corresponding to CPO and MPSO. From the results we conclude that the CPO has lesser energy consumption then MPSO.

Table 6.2: Testing results for energy consumption at different threshold values

Threshold	Existing	Proposed
0	41	20
0.1	41	72
0.2	41	64
0.3	40	36

0.4	42	30
0.5	45	30
0.6	61	52
0.7	90	84
0.8	113	74
0.9	120	80

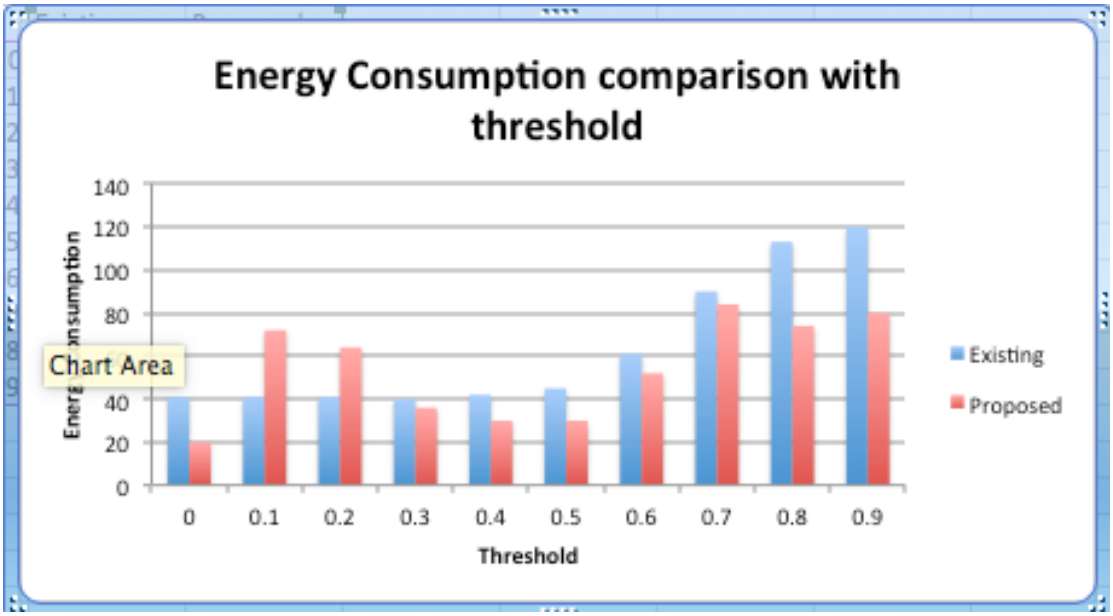


Figure 6.2: Energy consumption at different threshold values

The overall delay for each process in total iterations is calucluated

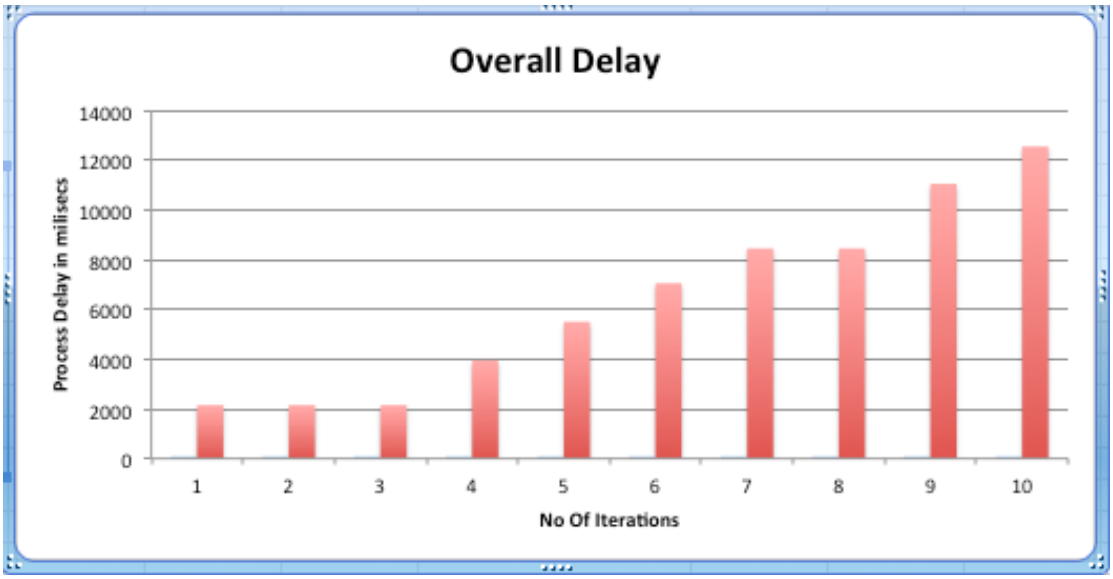


Figure 6.3: overall delay for each process

Chapter 7

Conclusion and future work

7.1 Conclusion:

In this thesis work our target was to minimize the energy consumed by datacenters running worldwide. From survey conducted it is concluded that datacenters consumes 1.3% of the global electricity. The reasons for datacenter's such huge energy consumption are idle server, high cost of energy efficient equipments and execrative split incentives etc. The main cause which we resolved in our thesis is idle servers. The idle servers consumes 60% more power than the one which is running at its full potential.

The virtual machine migration is applied to reduce the number of idle hosts by balancing the load between underutilized and over utilized servers through migration. In literature survey we mentioned some of the VM migration techniques. The VM migration is performed in two steps such as VM allocation and VM consolidation which includes, choosing the source for migration, selecting the migration technique for performing migration of data and states, last but not the least is deciding the destination host for placing the migrated virtual machine. The various VM allocation techniques studied and analyzed are heuristic based approaches, constraint programming, stochastic integer programming and static placement etc. The overloading of host is decided by threshold values of resource utilization. The selection of virtual machine form source host is done by random choice policy, highest potential growth and maximum correlation coefficient etc. The migration techniques studied to transfer data and states are post copy, pre copy, improved pre copy and pre copy based on LRU etc. To select the destination host for placing the migrated virtual machine major VM placement techniques studied are KNN based regression, MCC and fuzzy Analytical Hierarchy Process placement and Enhanced weighted round robin algorithm etc.

The major issues found in virtual machine migration process are energy efficient migration, number of active servers, total migration time and selecting source and destination host etc. We introduce an energy efficient virtual machine migration technique. The existing techniques uses the modified particle swarm optimization algorithm, we tried to minimize the energy consumption by virtual machine through cross pollination optimization.

The proposed technique is based on the biological phenomenon of flower pollination i.e. cross pollination for virtual machine migration. In this technique the whole VM allocation process is carried for 10 processes on 500 virtual machines. The threshold values are changed during migration process and the energy consumption is calculated. The results have proved that Cross pollination Optimization algorithm performs better then modified particle swam optimization algorithm in terms of energy consumption corresponds to number of virtual machines and CPU utilization threshold values. The algorithms correctness is proved through implementing it on CloudSim and workflowSim for energy efficiency and the results are compared with the existing Modified Particle Swarm Optimization.

7.2 Future work:

In this algorithm the motive was to reduce the energy consumption during virtual machine migration. The energy also gets consumed during migrating virtual machines. Sometimes the wrong selection of destination host can increases the number of migrations further, so the number of virtual machines migrated can be decreased in future by choosing more appropriate technique. Cross pollination uses static threshold values for CPU utilization. The CPU utilization gets changed as the workload changes. So to handle the dynamic workload demands we need some logic to set the CPU threshold value dynamically according to workload.

The SLA violation and the cost due to energy consumption can be the main challenge for future to be considered in Cross pollination algorithm.

Energy consumption for different VMs can be further reduced. The work is performed only for 10 processes in future we can extend it to large number of processes. The processes considered are meta processes in future the dependent tasks can be considered.

References

1. NRDC article, "America's Data Centers Are Wasting Huge Amounts of Energy", Available at, http://www.nrdc.org/energy/data-center-efficiency-assessment.asp.
2. http://www.eci.com/cloudforum/cloud-computing-history.html
3. http://www.computerweekly.com/feature/A-history-of-cloud-computing.
4. Virtualization Architecture [http://arstechnica.com/business/2011/02/virtualization-in-the-trenches-with-vmware-part-1-basics-and-benefits/]
5. https://pubs.vmware.com/vsphere50/index.jsp?topic=%2Fcom.vmware.wssdk.pg.doc_50%2FPG_Ch11_VM_Manage.13.2.html
6. **http://geek-university.com/vmware-esxi/vm-migration-types/ ,)**
7. Y. Ghanam, J. Ferreira, F. Maurer**,** "Emerging Issues & Challenges in Cloud Computing— A Hybrid Approach," in *Journal of Software Engineering and Applications*, 2012, 5, 923-937 http://dx.doi.org/10.4236/jsea.2012.531107 Published Online November 2012 (http://www.SciRP.org/journal/jsea).
8. R. Buyya, A. Beloglazov, and J. Abawajy, "Energy-Efficient Management of Data Center Resources for Cloud Computing: published in https://www.researchgate.net/publication/45921163..
9. A Vision, Architectural Elements, and Open Challenges T. Masteli´c, and I. Brandi´c, "Recent Trends in Energy Efficient Cloud Computing," JOURNAL OF LATEX CLASS FILES, VOL. 11, NO. 4, DECEMBER 2012.
10. Jonathan G. Koomey, "ESTIMATING TOTAL POWER CONSUMPTION BY SERVERS IN THE U.S. AND THE WORLD," …..
11. A. Hameed,A. Khoshkbarforoushha, R. Ranjan, P. P. Jayaraman, J. Kolodziej,P Balaji, S. Zeadally, Q.M. Malluni, N. Tziritas, A Vishnu, S.U. Khan, A. Zomaya, " A Survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems," Springer- verlag wien 2014.

12. D. H. Khan, Prof. D. Kapgate, Prof. P.S. Prasad, “A Review On Virtual Machine Management Technique and Scheduling in Cloud Computing”, International Journel of Advanced Research in Computer Science and Software Engneering 3(12), December-2013, PP. 838-845.

13. R. K. Gupta, R. K. Pateriya, “Survey on Virtual Machine Placement Technique in Cloud Computing Enviornment”, International Journel on Cloud Computing: Services and Architecture, vol. 4, No. 4, August 2014.
14. A. Beloglazov et al., “Energy efficient allocation of virtual machines in cloud data centers”,proceeding in 10th IEEE/ACM Intl. Symp. on Cluster, Cloud and Grid Computing, pp. 577-578, may2010.

15. Lei Xu,Wenzhi et al., ”Smart-DRS:A Strategy of Dynamic Resource Scheduling in Cloud Data Center”, IEEE International conference on Cluster Computing Workshops, pp. 120-127, Sept..2012
16. N. Bobro et al., “Dynamic placement of virtual machines for managing SLA violations”, proc. in 10th IEEE international symposium on integrated network management pp. 119-128, May 2007.
17. C. Clark et al., “Live migration of virtual machine”, proceeding of the 2nd conference on symposium on network system design and implementation, vol. 2, pp. 273-286, 2007.
18. T. Wood et al., “Black-Box and Gray-Box strategies for virtual machine migration”, NSDI'07Proceedings of the 4th USENIX conference on Networked systems design & implementation, pp. 7-17, 2007.
19. Mayank 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.
20. Constraint programming. http://en.wikipedia.org/wiki/Constraint_programming.
21. http://en.wikipedia.org/wiki/Bin_packing_problem.

22. Z. A. Mann, “Allocaton of Virtual Machines in Cloud Data Centers- A Survey of Problem Models and Optimization Algorithms”, Published in ACM Computing Surveys, Vol. 48, issue 1, 2015.
23. A. Singh. N, M. Hemalatha, “Energy Efficient Virtual Machine PlacementTechnique using Banker Algorithm in Cloud DataCentre,” published in 2013 International Conference on Advanced Computing and Communication Systems (ICACCS -2013), Dec. 19 – 21, 2013, Coimbatore, INDIA.
24. A. Beloglazov_ and R. Buyya, “Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation ofVirtual Machines in Cloud Data Centers,” CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE, Wiley Press, New York, USA,pp. 1397-1420, Sep. 2012.
25. A. Beloglazov , J. Abawajyb, R. Buyya, “Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing,” Future Generation Computer Systems, pp. 755-768, May. 2012.
26. Z. Cao and S. Dong, “Dynamic VM consolidation for energy-aware and SLA violation reduction in cloud computing,” 13th International Conference on Parallel and Distributed Computing, Applications and Technologies, pp. 363-369, IEEE, 2012.
27. M. Nelson, B. Lim, and G. Hutchins, “Fast Transparent Migration for Virtual Machines,” in Proceedings of USENIX Annual Technical Conference(USENIX’05), Marriot Anaheim, Anaheim, CA, USA, PP. 391-394, April 10-15, 2005.
28. C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, A. Warfield, “Live Migration of Virtual Machines,”in proceedings of 2nd Symposium on Networked Systems Design & Implementation, pp. 273-286, 2005.
29. E. P. Zaw and N. L. Thein, “Improved Live VM Migration using LRU and Splay Tree Algorithm,” International Journal of Computer Science and Telecommunications , Volume 3, Issue 3, pp. 1-7, , March 2012.
30. K. Z. Ibrahim, S. Hofmeyr, C. Iancu, E. Roman, “Optimized Pre-Copy Live Migration for Memory

Intensive Applications,"in international Conference for High Performance Computing, Networking, Storage and Analysis(SC), pp. 11-20, September 2007.

31. F. Ma, F. Liu, Z. Liu, "Live Virtual Machine Migration based on Improved Pre-copy Approach," in Proceedings of the 2009 IEEE International Conference on Cluster Computing(Cluster 2009), pp. 1-10, 2009.
32. H. Jin, L. Deng, S. Wu, X. Shi, X. Pan, "Live Virtual Machine Migration with Adaptive Memory Compression," published in IEEE international Conference on Software Engneering and Service Sciences ICSESS, pp. 230-233, 2010.
33. M. R. Hines, U. Deshpande, and K. Gopalan, "Post-Copy based Live virtual machine Migration using adaptive pre-paging and dynamic self-balloning," in Proceedings of the ACM/Usenix International Conference on Virtual Execution Enviornment, pp. 51-60, ACM, 2009.
34. H. I. G, R. C, "A Survey on VM Consolidation For Energy Efficient Green Cloud Computing", International Journal of Emerging Technology in Computer Science and Electronics ISSN: 0976-1353, Vol. 19.
35. M. Hadji, P. Labroage, " Online Algorithm for Servers Consolidation in Cloud Data centers," Technology Research Institute- IRT SystemX8, Avenue de la vauve, 91120 Palaiseau, France.
36. L. Rolling, C. Morin, " Energy Aware Ant Colony Based Workload Placement in Cloud",- 2011 GECCO 14, July 12-16, 2014, ACM 978-1-4503-2662-9/14/07.
37. G. Xu, Y. Dong, X. Fu, "VM Placement Strategy Based On Distributed Parallel Ant Colony optimization Algorithm," in An International Journel of Applied Mathematics and Information Sciences 2015.
38. G. Motta, N.S fondrini, and D.Sacco, "Cloud computing: An architectural and technological overview," in Proc. Int. Joint Conf.Serv. Sci., 2012, pp. 23–27.
39. V.K. Mohan Raj, R. Shriram, " Power Aware Provisioning in Cloud Computing Enviornment", International Conference on Computer Communication and Electrical Technology 2011.
40. N. Kord, H. Haghighi, "An Energy Efficient Approach for Virtual Machine Placement in Cloud Based Data Centers", 2013 5th Conference on Information and Knowledge Tyechnology.

41. A. Alnowiser, E. Aldhahri and A. Alahmadi, " Enhanced Weighted Round Robin Scheduling with DVFS technology in Cloud", 2014 International Conference on Computational Science and Computational Intelligence
42. J. Sekhar, G. Jeba, S. Durga, "A Survey on Energy Efficient Server Consolidation Through VM live Migration", International Journal of Advances in Engneering and Technology, Nov. 2012, Vol.5, Issue 1, PP. 515-525.
43. P. Lu, A. Barbalace, R. Palmieri and B. Ravindran, "Adaptive Live Migration to Improve Load Balancing in Virtual Machine Enviornment".
44. Z. A. Mann, "Allocaton of Virtual Machines in Cloud Data Centers- A Survey of Problem Models and Optimization Algorithms", Published in ACM Computing Surveys, Vol. 48, issue 1, 2015.
45. M. A. H. Monil, R. Qasim, R. M. Rahman," Energy-aware VM Consolidation Approach Using Combination of Heuristics and Migration Control,"in IEEE 2014.
46. X. Ruan and H. Chen, "Performance-to-Power Ratio Aware Virtual Machine (VM) Allocation in Energy-Efficient Clouds", in International Conference on Cluster Computing,pp. 264-273,2015 IEEE.
47. P. Malviya, S. Agrawal and S. Singh, ""**"**An Effective Approach for Allocating VMs to Reduce the Power Consumption of Virtualized Cloud Environment**",in** Fourth International Conference on Communication Systems and Network Technologies, pp. 573-577, 2014 IEEE.
48. A. P. Pablo, G. d. Valle, D. Atienza, "Exploiting CPU-Load and Data Correlations in Multi-Objectiv VM Placement for Geo-Distributed Data Centers," in 2016 EDAA.
49. T. Duong-Ba, T. Nguyen, B. Bose, and T. Tran, "Joint Virtual Machine Placement and Migration Scheme for Datacenters," published in, IEEE conference on Globecom 2014 - "Symposium on Selected Areas in Communications: GC14 SAC Cloud Networks", in 2014 .
50. Kwonyong Lee and Sungyong Park, "A CPU Overhead-aware VM Placement Algorithmfor Network Bandwidth Guarantee in Virtualized Data Centers," 2015 International Conference on Cloud and Autonomic Computing

51. Antonio Marotta and Stefano Avallone, “A Simulated Annealing based Approach for Power Efficient Virtual Machines Consolidation”, 2015 IEEE 8th International Conference on Cloud Computing
52. *Gamal Eldin I. Selim1, Mohamed A. El-Rashidy2, Nawal A. El-Fishawy*3, “An Efficient Resource Utilization Technique for Consolidation of Virtual Machines in Cloud Computing Environments,” in proceedings of 33rd NATIONAL RADIO SCIENCE CONFERENCE, pp. 316-324, Feb 22- 25, 2016,
53. Suheib Alhiyari and Ali El-Mousa, “ A Network and Power Aware Framework for Data Centers Using Virtual Machines Re-Allocation,” 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT).
54. R. N. Calheiros1, R. Ranjan, A. Beloglazov, C. A. F. D. Rose and R. Buya, “CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms”, published in Wiley Online Library. Pp 23-50, 24 August 2010.
55. A. Xiong and C. Xu, “ Energy Efficient Multiresource Allocation of Virtual Machine Based on PSO in Cloud Data Center,” published by Hindawi in , Mathematical Problems in Engneering Volume 214, available at http://dx.doi.org/10.1155/2014/816518.

Abbreviations

VM	Virtual Machine
CPO	Cross Pollination Optimization
PM	Physical Machine
CPU	Central Processing Unit
QOS	Quality of Service
SAAS	Software as a Service
SLA	Service Level Agreement
IQR	Inter Quartile Range
LRU	Least Recently Used
MC	Maximum Correlation
OS	Operating System
ACO	Ant Colony Optimization
LAN	Local Area Network
KNN	K- nearest Neighbor
MCC	Minimum Correlation Coefficient
AHP	Analytical Hierarchy Process
DVFS	Dynamic Voltage Frequency Scaling
PABFD	Power Aware Best Fit Decreasing

PPR	Performance to Power Ratio
JPM	Joint Placement and Migration
PSO	Particle Swarm Optimization
MPSO	Modified Best Fit Decreasing

List of Publications

1. S. Kaur and S. Bawa, “A Review on Energy Aware VM Placement and Consolidation Techniques,” has been accepted for IEEE Sponsored International Conference on Inventive Computation Technologies (ICICT 2016).
2. S. Kaur and S. Bawa, “ Energy Efficient Virtual Machine Migration Policy in Cloud Enviornment”, communicated in IEEE Sponsored International Conference on Inventive Computation Technologies (ICICT 2016).

YOUTUBE VIDEO LINK

shttps://youtu.be/8FtHlJ4uQ10

Plagiarism Report

ORIGINALITY REPORT

9%	5%	7%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.nrdc.org Internet Source	1%
2	Duong-Ba, Thuan, Thinh Nguyen, Bella Bose, and Tuan Tran. "Joint virtual machine placement and migration scheme for datacenters", 2014 IEEE Global Communications Conference, 2014. Publication	<1%
3	Sinha, Indrajit, and Milind Kumar Sharma. "Cloud computing in small and medium sized enterprises: an architectural model", International Journal of Enterprise Network Management, 2015. Publication	<1%