SERVICE LEVEL AGREEMENT AWARE ENERGY EFFICIENT APPROACH FOR VIRTUAL MACHINE PLACEMENT IN CLOUD DATA CENTER

Dissertation submitted to the Central University of Punjab

For the award of

Master of Technology

In

Centre for Computer Science and Technology

BY

Nirmaljeet Kaur

Supervisor

Er. Surinder Singh Khurana

Centre for Computer Science and Technology

School of Engineering and Technology

Central University of Punjab, Bathinda

June, 2014

CERTIFICATE

I declare that the dissertation entitled “SERVICE LEVEL AGREEMENT AWARE ENERGY EFFICIENT APPROACH FOR VIRTUAL MACHINE PLACEMENT IN CLOUD DATA CENTER” has been prepared by me under the guidance of Er. Surinder Singh Khurana, Assistant Professor, Centre for Computer Science and Technology, School of Engineering and Technology, Central University of Punjab. No part of this dissertation has formed the basis for the award of any degree or fellowship previously.

Name: Nirmaljeet Kaur
Centre for Computer Science and Technology,
School of Engineering and Technology,
Central University of Punjab,
Bathinda – 151001.

Date:

CERTIFICATE

I certify that NIRMALJEET KAUR has prepared her dissertation entitled “SERVICE LEVEL AGREEMENT AWARE ENERGY EFFICIENT APPROACH FOR VIRTUAL MACHINE PLACEMENT IN CLOUD DATA CENTER”, for the award of M.Tech degree of the Central University of Punjab, under my guidance. She has carried out this work at the Centre for Computer Science and Technology, School of Engineering and Technology, Central University of Punjab.

Er. Surinder Singh Khurana
Assistant Professor
Centre for Computer Science and Technology,
School of Engineering and Technology,
Central University of Punjab,
Bathinda – 151001.

Date:

ABSTRACT

Service Level Agreement Aware Energy Efficient Approach for Virtual Machine Placement in Cloud Data Center

Name of Student:	Nirmaljeet Kaur
Registration Number:	CUPB/M.Tech/SET/CST/2012-13/15
Degree for which submitted:	M.Tech
Name of Supervisor:	Er. Surinder Singh Khurana
Name of centre:	Centre for Computer Science and Technology
Name of School:	School of Engineering and Technology

Key words:	Cloud Computing, Energy Consumption, Data Center, SLA, Virtual Machine Migration, CPU Utilization

Generally, the growth of computing systems has been concentrate on performance improvement or fast processing of applications from the view point of customer, scientific and business perspective. However, this type of computing increases energy consumption. It is continue to consume large amount of energy in the form of electrical power. In cloud computing data center hosting of cloud applications consume huge amount of energy. Optimization of the energy requirements of data center is a very challenging task. This dissertation report describes the energy efficiency of minimization of migration policy in IaaS environment. In this policy VM migration, energy consumption, SLA violations are calculated. VM migration directly influences resource utilization and Quality of Service (QoS) parameters provided by the system. The effect of disturbance cause server overloads, resource shortage, performance degradation etc. So, one of the challenging task is, the placement of VMs, like where, when, which VM could be migrated for saving energy and less SLA violations, with minimum number of migration. In this research work a new approach has been proposed to select VMs to migrate from the over utilized host. Approach migrates those VMs that belongs to the higher category so that that the resources allocated to these VMs have less cost penalties for SLA violations. Proposed approach results are compared to the simulated Minimization of Migration Policy (Beloglazov et al., 2012) and compared its energy requirements, virtual machine migration and SLA violation with Minimum Migration Time and Random Selection policies.

Name and Signature of Student Name and Signature of Supervisor

DEDICATED TO MY BELOVED PARENTS

ACKNOWLEDGEMENTS

“The successful completion of any task would be incomplete without accomplishing the people who made it all possible and whose constant guidance and encouragement secured us the success.”

On the submission of my dissertation report on “Service Level Agreement Aware Energy Efficient Approach for Virtual Machine Placement in Cloud Data center”, I would like to extend my gratitude and sincere thanks to my respected guide Er. Surinder Singh Khurana, Assistant Professor, Centre for Computer Science & Technology, Central University of Punjab, Bathinda, for his constant motivation and support during the course of my work. I truly appreciate and value his esteemed guidance and encouragement from the beginning to the end of this report. I am indebted to him for having helped me shape the problem and providing insights towards the solution.

I would also like to thank Dr. A. K. Jain, Dean, School of engineering & Technology, for their valuable suggestion in the course of my work. I wish to thank all the faculty members of Central University of Punjab, Bathinda for the invaluable knowledge they have imparted me in most exciting and enjoyable way. I am grateful for their constant support and help.

(Nirmaljeet Kaur)

CUPB/M.TECH/SET/CST/2012-13/15

Centre for Computer Science & Technology

School of Engineering and Technology

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Sr. No.	Full Form	Abbreviation
1.	Application Program Interface	API
2.	Basic Input/Output System	BIOS
3.	Central Processing Unit	CPU
4.	Climate Change Servers Computing Initiative	CSCI
5.	Content Provider	CP
6.	Coordinated Voltage Scaling	CVS
7.	Dynamic Voltage and Frequency Scaling	DVFS
8.	Independent Voltage Scaling	IVS
9.	Information Technology	IT
10.	Infrastructure as a Service	IaaS
11.	Internet Service Provider	ISP
12.	Line Replaceable Unit	LRU
13.	Linear Programming	LP
14.	Media Access Control	MAC
15.	Million Instructions Per Second	MIPS
16.	Minimization of Migration	MM
17.	Minimum Migration Time	MMT
18.	Modified Best Fit Decreasing	MBFD
19.	National Institute of Standards and Technology	NIST
20.	Operating System	OS
21.	Pattern Based Allocation	PBA
22.	Personal Digital Assistant	PDA
23.	Physical Machine	PM
24.	Platform as a Service	PaaS
25.	Program Counter Based Access Predictor	PCAP
26.	Quality of Service	QoS
27.	Random Selection	RS
28.	Service Level Agreement	SLA
29.	Software as a Service	SaaS

30.	Structural Constraint Aware Virtual Machine Placement	SCAVP
31.	Total Cost of Ownership	TCO
32.	Uninterruptible Power Supply	UPS
33.	Vary-On Vary –Off	VOVO
34.	Virtual Data Center	VDC
35.	Virtual Machine	VM
36.	Virtual Machine Manager	VMM

CHAPTER 1

INTRODUCTION

Cloud computing, is a type of computing model that came into light around the end of 2007. Basically, it provides a pool of computing resources and services; those are accessed by the users through Internet. The principle of cloud computing is to transfer the computing done from the local computer into the network. This makes the enterprise to use the resources, which includes application, services, network, server, storage and so on, that is required without huge investment on its purchase, implementation, and maintenance and can be uses it for further significant purpose. All the resources are supplied on-demand without any advance reservation and it eliminate over provisioning leads to improve resource utilization (Sadashiv & Kumar, 2011). National Institute of Standards and Technology- define Cloud Computing (Jansen and Grance 2011) as: Cloud computing is a model for enabling ubiquitous, 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 efforts or service provider interaction.

Cloud computing is defined as an advanced computing which computes services over the internet dynamically. It uses virtualized environment for resources allocation to the services in a cloud. Currently, Cloud computing has become a popular technology trend in the IT industry. Many specialists of IT expect that cloud computing will reframe information technology (IT) in the near future (Furht & Escalante, 2010). It has a potent effect on computing because of its On –Demand, pay-as-you-go and utility computing features.

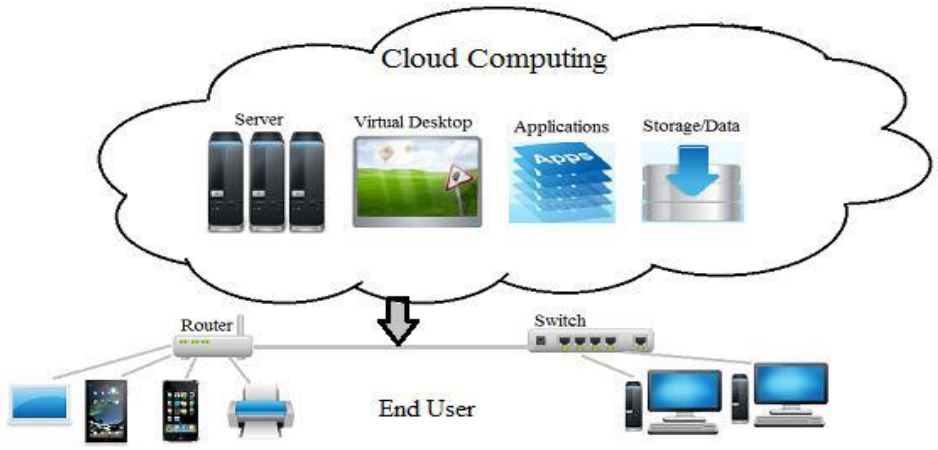

Figure1.1 Cloud Computing

Users can use cloud services with their devices like, laptop, PCs, mobile, PDAs etc. Cloud provider provides services to users according to the requirements of the user or as per requests submitted to the cloud. Users pay to the cloud provider as per they use the services of cloud. Users can access cloud services at anytime from anywhere in the world. Cloud Computing fully depend on internet, also known as internet based technology.

In other context, cloud computing is a model in which data and computation are operated somewhere in the cloud, it is a group of data centers owned and organized by third party. Cloud computing defined as the services delivered over the internet with refer to hardware, software and applications (Antonopoulos & Gillam**,** 2010).

Cloud computing is a computing model, in which a large pool of computing systems are connected with each other in private or public networks, to provide scalable computing infrastructure for services, application, file storage and data. With this technology, the cost of computing, services or application hosting, data storage and delivery is reduced significantly (Harris, 2012).

1.1 Introduction to the Research Area

Cloud computing has been widely adopted by the IT industry, due to its proficiency and exponential growth. This massive adoption of cloud computing has caused the dramatic increase in energy consumption and its impact on the environment in terms of carbon footprints. The high cost of energy and link between energy consumption and carbon emission has raised an issue of energy management. Even the overall energy consumption has been reduced by sharing of resources of cloud by different companies but still a huge amount of electrical energy has been consumed by data centers hosting cloud applications. The increasing cost of energy is a giant difficulty for cloud vendors because it increases total cost of ownership and decreases return on investment. A very small portion of total electrical consumption has been utilized in the cloud data center (International Resources Group, 2009). A lot of work has been done to make the cloud data centre more energy efficient. But still datacenters consume a large amount of energy, there is need to improve the components of cloud computing to make it energy efficient.

1.2 Motivation

Cloud Computing paradigm is most popular because of its capability for provisioning resources quickly. From the perspective of energy efficiency in cloud, there are many techniques and algorithm like server consolidation, virtualization, live migration. Many researchers have been working in developing energy efficient algorithm for cloud computing. The major issue of energy consumption in cloud which includes large size and capacity of datacenters has been getting higher attention in recent years. From the perspective of energy efficiency, a cloud data center can be described as a pool of large computing and communication resources managed in the way to change the received power into computing or data shifting to satisfy user needs. So, to shifting the workload in a minimum set of the computing resources is major task to save energy. In addition, energy consumption in VM migration can be a major percentage of total energy consumption in cloud. In Infrastructure as a Service (IaaS) cloud computing the service and resource requests are given by creating virtual machines of the requested specification on the underlying physical infrastructure.

Many studies conclude that VM migration can cost a large percentage of energy use, as well as cause increase in runtime. Most of the existing algorithms show that it would waste a big percentage of energy (Jansen & Grance, 2011). One method to improve the utilization of resources and reduce energy consumption, which has been shown to be efficient is dynamic consolidation of Virtual Machines (VMs) enabled by the virtualization technology. In VM consolidation typically problem occur to determine what time, which VMs and where should be migrated to minimize the total cost. Most of the work have been done focusing on reduction of allocation cost only. Another method is to optimize the virtual machine migrations. This dissertation report describes the energy efficiency of minimization of migration policy in IaaS environment. In this policy VM migration, energy consumption, SLA violations are calculated. VM migration directly influences resource utilization and Quality of Service (QoS) parameters provided by the system. The effect of disturbance cause server overloads, resource shortage, performance degradation etc. So, one of the challenging task is, the placement of VMs, like where, when, which VM could be migrated for saving energy and less SLA violations, with minimum number of migration.

1.3 Problem Statement

Cloud computing allows hosting of multiple services on a globally shared resource pool where resources are allocated to services on demand. It uses virtualized environment to make services flexible, secure and scalable. But it has some performance degradations of services and also has energy overheads and large amount of power consumption. In past, many researchers worked on making energy efficient algorithm for reducing energy consumption. Many algorithms have been implemented for saving energy of data centers by turning off or by putting idle servers to sleep mode of servers. But these techniques were not so effective because of performance degradations of services and improper resources utilization. (Beloglazov, Abawajy & Buyya, 2012) proposed an idea for making energy efficient algorithm for data centers. They proposed Virtual machine placement algorithm that is minimization of migration (MM), which consider utilization of host CPU and utilization of virtual machines in decreasing order of CPU utilization. The performance of algorithm is better than other placement algorithms but they did not consider SLA parameters while selecting virtual machine for migration, which might be effected by live migration. Most of the violations occur during live migration of virtual machines, migration impact the parameters of SLA. So there is need to develop new approach for SLA aware energy efficient algorithm for resources allocation in data centers.

1.4 Objectives

The objective of this research work is to modify minimization of migration algorithm given in (Beloglazov et al., 2012) so that the no. of SLA violations occurred during execution of user requests will be reduced. To improve the algorithm we will consider SLA parameter (VM category selected during Service Level Agreement) and current load of VMs to select virtual machine for migration.

1.5 Methodology

Step1. Firstly, we simulate the Minimization of Migration approach (Beloglazov et al., 2012) CloudSim.

Step2. We categorize the VMs on basis of VM usage cost and cost of penalties for SLA violations.

Step3. The proposed approach has been simulated in CloudSim.

Step4. By setting up parameters, the results have also been recorded for Minimum Time and Random Selection approaches. These techniques have already been coded in CloudSim.

Step5. Finally the performance of all these approaches including the proposed approach has been compared.

The remainder of the chapters is organized as follows. In the next chapter, cloud computing definitions and terminologies are defined. Chapter 3 discusses the review of literature in the field of energy consumption in computing infrastructure. Chapter 4 presents the proposed work and chapter 5 evaluate the simulation and results. Chapter 6 concludes about the whole research work and its future scope.

CHAPTER 2

CLOUD COMPUTING TERMINOLOGY

This chapter begins with the various terms and technologies used in cloud computing, impact of energy consumption on the computing infrastructure. Cloud computing is one of the most important topics in IT these days, with Microsoft, Google, Amazon, and other big players joining in the fray. However, the technology brings with it new terminology that can be confusing. Here are some common cloud-related terms and their meanings. Cloud computing has its own set of vocabulary.

2.2 Cloud Computing

Cloud computing is continuously change the way of businesses work. It generate a new way to ease collaboration and information access around the great geographic distances along with reducing the overhead and associated maintenance of resources.

According to NIST: “National Institute of Standards and Technology- define Cloud Computing (Jansen and Grance 2011) as: Cloud computing is a model for enabling ubiquitous, 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 efforts or service provider interaction”.

According to Foster and Garner (Magoulès, Pan & Teng, 2012): Foster: “A large scale distributed computing paradigm that is driven by economics of scale, in which a pool of abstracted platforms, and services are delivered on demand to external customers over the internat.” Gratner: “A style of computing where scalable and elastic IT capabilities are provided as a service to multiple external customers using internet technologies”.

2.2 Energy Efficiency in Computing Systems

It was estimated that energy consumption by large scale industries has been raised from 56% in the year of 2005 to 2010. In 2010, percentage between 1.1% and 1.5 % of the global electricity consumed is calculated.

Energy consumption of a computing infrastructure is not only examines by hardware efficiency, it can be determined by the resources management of a system, which is deployed on the computing infrastructure and the application’s running efficiency on a system.

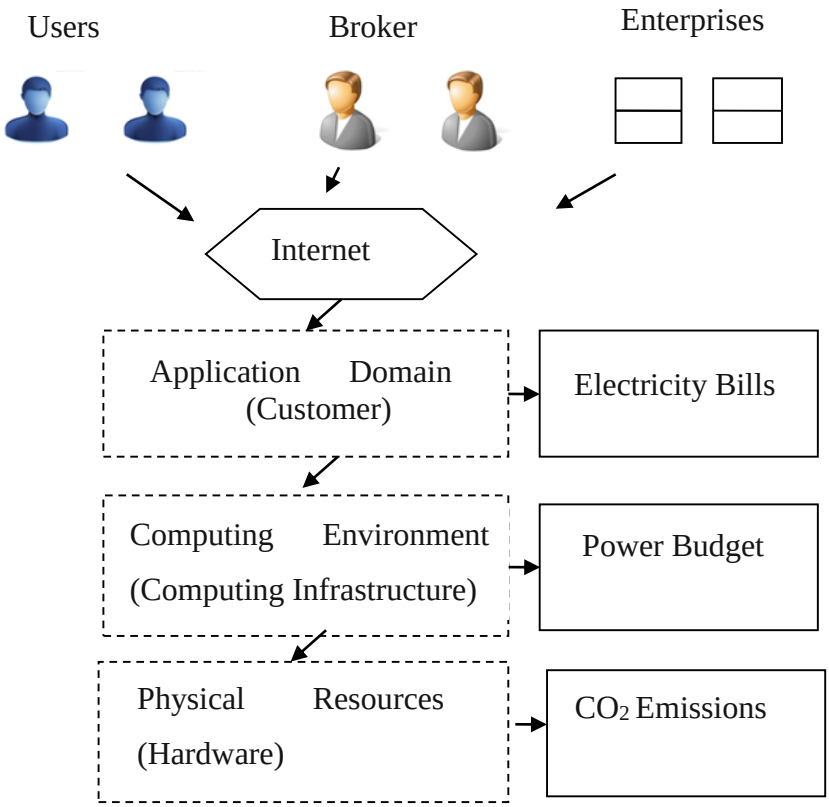

Figure 2.1 Energy Consumption at Different levels of Computing System

Energy Consumption effect customer in a manner of resource usage costs, which are generally calculated by the Total Cost of Ownership (TCO) experienced by the service provider. Increase in power consumption is increase electricity bills and also have some extra operational costs of cooling system and power delivering computing devices such as UPS.

For the reduction of energy consumption in computing infrastructure, many industries try to develop standardized approaches and method. They involved Green Computing, The Green Grid, Climate Savers Computing Initiative (CSCI), Green Electrons Council. With the partnership of large scale industries such as Microsoft, VMWare, Dell, IBM, HP Intel etc.

Energy Efficiency management has been firstly deployed for the mobile devices in the context of battery power. Also these techniques can be deployed on server and data centers, but by applying the specific methods for every computing device (Beloglazov, 2013).

2.3 Energy Efficient Cloud Computing Architecture

In cloud users or customers depend on Cloud providers to requests more of their computing requirements. They have specific Quality of service (QoS) needs, Providers have to provide them according to the requests and sustain their operations. Cloud providers examine and provide different QoS parameters to each individual consumer as agreed in specific Service Level Agreement. To gain this complex system, traditional system-centric resource management architecture are failed to provide incentives for Cloud providers and they no longer continue to deploy it on traditional systems. They need such systems that can share their resources and still maintain all service requests or needs to be of equal importance (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009).

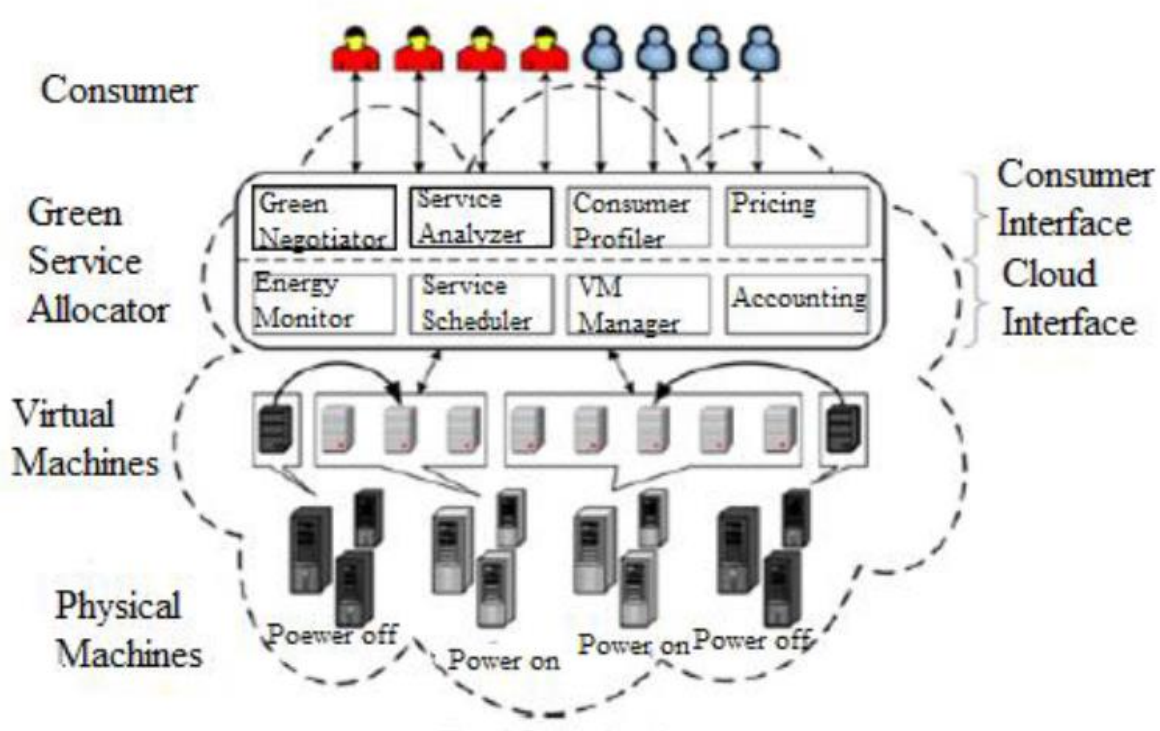

Figure 2.3 The High-Level System Architecture (Sinha, Purohit & Diwanji, 2011)

2.3.1 Consumer

Cloud consumers send service requests to the Cloud from anywhere in the world. It is important to note that consumer can be a Cloud consumers and users/client of deployed services. For a situation, a company can be a consumer which is deploying a web-application, it gives varying workload corresponding to the number of end users accessing it.

2.3.2 Service Allocator

Service allocator is an interface which is lying between the computing infrastructure and consumers. The following components are required to support the energy-efficient resource management:

- **Negotiator:** Negotiator discusses with the consumers/brokers to finalize the SLAs (Service Level Agreement) by specifying prices and penalties of services (for violations of the SLAs) between the Cloud service provider and end user/consumer depending on the consumer's QoS needs.
- **Service Analyzer:** Service analyzer interprets and analyzes the service requirements which may include terms and conditions; of a submitted user request before providing the services. For the acceptance and rejection of submitted user requests, it requires the latest workload and usage of energy data information from Energy Monitor and VM Manager.
- **Consumer Profiler:** This component collects the specific characteristics or needs of consumers/end user so that high priority consumers can be granted with special privileges and prioritized over other consumers.
- **Pricing:** This part of architecture decides charges of provided cloud services. Charges of service requests are applied to handle the supply and demand of specific computing resources that can be facilitate in prioritizing service allocations effectively.
- **Energy Monitor:** Observes energy consumption caused by VMs and physical machines and provides this information to the VM manager to make energy-efficient resource allocation decisions.
- **Service Scheduler:** Assigns requests to VMs and observes resource entitlements for the allocated VMs. If the auto scaling functionality has been requested by a customer.
- **VM Manager:** It keeps track of the availability of virtual machine on hosts and their resource usage. It is in charge of provisioning new VMs as well as reallocating VMs across physical machines to adapt the placement.
- **Accounting:** Monitors the actual usage of resources by VMs and accounts for the resource usage costs. Historical data of the resource usage can be used to improve resource allocation decisions.

2.3.3 VMs (Virtual Machine)

A virtual machine (VM) is a software program implementation of a computing system in which one or more operating system (OS) or programs can be installed and run. Virtual machine usually imitate a physical computing environment, but requests for network, memory, hard disk, CPU and other physical resources are managed or occupied by a virtualization layer.

2.3.4 Physical Machines

The prime physical computing servers provide the hardware infrastructure for creating virtualized resources to meet service demands (Beloglazov et al., 2012).

2.4 Deployment Models

2.4.1 Infrastructure as a Service (Iaas)

This model offers a virtualization environment for computing infrastructure, as-a-service. The client can manage computing resources (data centers, servers, storage and network equipment) as a fully external service instead of purchasing & configuring these resources. The service is usually billed on a utility computing point basis and the amount of resources used (and therefore the cost) will generally reflect the levels of activities. The main purpose of this model is to prevent purchasing, housing, configuring and managing the underlying hardware and software infrastructure, and instead get those resources as virtualized services which are controlled via a service interface (Jansen & Grance 2011). Examples of IaaS are virtual servers provided by Amazon, Rackspace, GoGrid, etc.

It offers web based access to storage and computing power. The consumer does no need to manage or control the basic cloud infrastructure but has control upon the operating systems, storage, and deployed applications (Antonopoulos & Gillam, 2010). Many companies that deploy applications to the cloud will need a specific platform, such as window, .net and microsoft SQL server, or Linux, perl and My Sql. Utilizing a platform as a service solution eliminates the company's need to administer the operating system and supporting software (Jamsa, 2013). Infrastructure as a service provides basic storage and computing capabilities as standardized services over the internet or network. Data center, Servers, storage systems, routers, and switches, other systems are supplied and made available to manage workloads that range from application parts to high performance computing applications utilities (Sun Microsystem, 2009).

2.4.2 Platform as a Service (PaaS)

This model of computing provides a collection of hardware and software services that developers can use to create and deploy those applications within the cloud. By using PaaS, Developers remove the need to buy and manage hardware as well as the requirement to install and handle operating system and database, storage on application demand and the organization can pay for only those services it consumes. It gives tools for developers to build and host web applications, software as a service provider, is built using the force.com platform while the infrastructure is provided by the amazon web service (Antonopoulos & Gillam, 2010). Further, because PaaS removes the developers requirement to worry about servers, deployment of web solutions are going to be more quickly (Jamsa, 2013).The main purpose of PaaS is to decrease the cost of managing, buying, and housing the underlying physical resources and software objects of the platform, including any specific program and database development tools (Jansen and Grance 2011).

There are at least two aspects on PaaS depends upon the views of the provider or consumer of the services according (Sun Microsystem, 2009):

- Someone providing PaaS might supply a platform by integrating an OS, application software, and even a development environment which is then send to a customer as a service.
- Someone operating PaaS would see an enclosed service that is provide to them through an Application Program Interface (API). The customer would interact with the platform environment through the API, platform does work what is necessary to handle and scale it to provide a given level of service.

PaaS services can provide a powerful platform for developers on which to deploy applications, although they may be constrained by the abilities and capabilities that the cloud provider chooses to provide (Sun Microsystem, 2009).

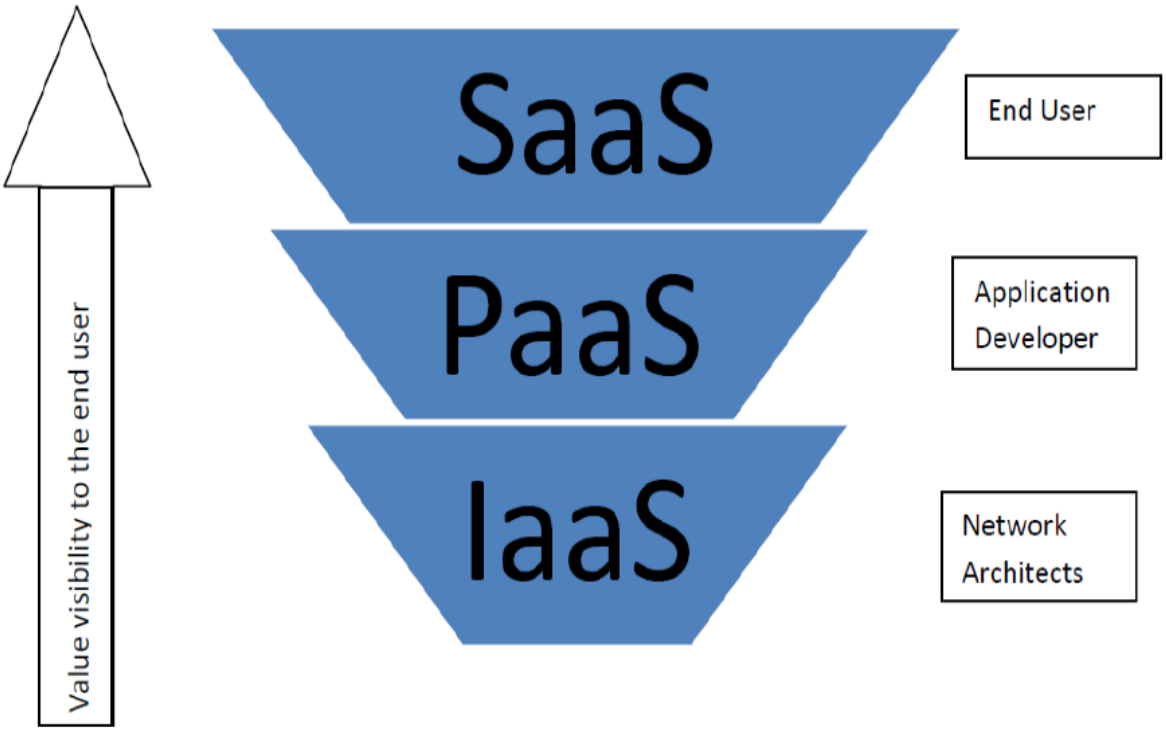

Figure 2.4 Development Models

2.4.3 Software as a Service (SaaS)

Software as a service model provide the complete application as a service on the user demand. A single object of the software resources runs on the cloud and services multiple customers and end users or client organizations (Sun Microsystem, 2009). This model consists applications that are accessible from a number of client devices throughout a thin client (slim clents) interface such as application as thin client, a web browser (Antonopoulos & Gillam, 2010). It is a solution model in which end users use a web browser or network to access software along with the programs implementation and user data and storage in the cloud.

Companies or organizations that are using SaaS solutions removes the need for in house applications, maintenance and administrative support for the applications, and cloud data storage. Because SaaS solutions located within the cloud, the solutions can easily scale up to meet customer demands. Furthermore organizations can pay only for the SaaS services on demand means that the organizations pay only for the resources they use, generally on a per-user basis. SaaS services exist to a large range of applications and provide customers and end users with a cost effective way to access services with better started and an affordable long –term solution (Jamsa, 2013).

Software as a Service refers to software delivered over a browser. This model eliminates the requirement to install, configure and run applications on the user's own computers/servers. It simplifies maintenance, configuration, upgrades and support. Main purpose of this model is to decrease the total cost of physical resources and software development, installation, maintenance, configuration and operations. The cloud user does not manage or control the prime cloud infrastructure or individual applications, except for choice selections and limited administrative application settings (Jansen & Grance, 2011).

2.5 Types of Cloud

2.5.1 Public Cloud

In general definition when a cloud is made available in a pay-as-you-go manner to the general public is known as public cloud (Antonopoulos & Gillam, 2010). The general public is defined in this scenario as either individual consumers or corporations. This cloud infrastructure is owned by a cloud vendor organisation (Krutz & Vines, 2010). Public clouds are organized and run by third parties, and services and applications from different customers (such as different companies), are mixed together on the public cloud's servers, networks and storage systems. Public clouds are mostly hosted away from user premises, and they produce a way to reduce user risk and cost by supplying a flexible, level of temporary appendix to enterprise infrastructure. If Implementation of a public cloud has done by considering some issues such as security, performance, and data locality as major aspects, then the existence of other applications, services running in the cloud should be visible to cloud architects, developer and end users. Indeed, public clouds can be much bigger than a company's private cloud and also might be, offering the abilities and capabilities to scale up and down on needs, and transfer infrastructure risks from the company, enterprise to the cloud provider.

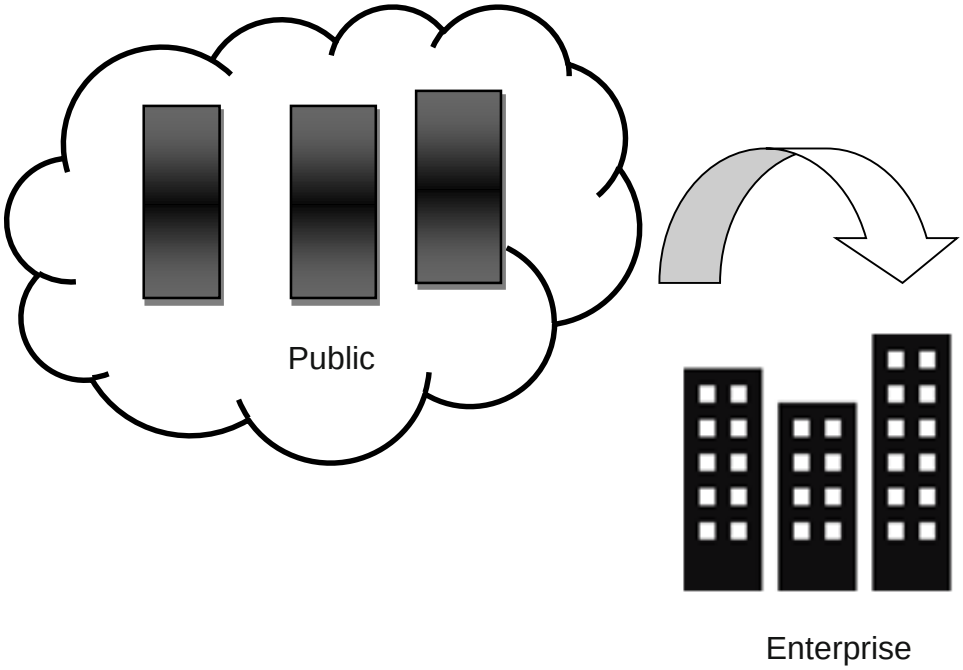

Figure 2.5 Public Cloud

Portions of a public cloud can be consumes for the total use of a single client on a priority basis, by creating a (VPD) Virtual Private Data center. A virtual private data center provides end users greater transparency into its infrastructure instead being restricted to deploying VM images in a public cloud. Now a day, customers can utilize not just virtual machine images; also they can manipulate servers, network devices, storage systems, and network topology. Creation of a virtual private data center with all elements located in the same service helps to learn the issue of data locality, because of abundant bandwidth and generally free when connecting these resources within the same service (Sun Microsystem, 2009).

2.5.2 Private Cloud

The term private cloud is used for the computing infrastructure which is operated specially for a business or an organization (Antonopoulos & Gillam, 2010). This type of clouds is creating, exclusive use of one client, quality of service and supplying the utmost control over data, and security. The company keeps the infrastructure and has control over the cloud that how applications, services are deployed on it. Private clouds might be deployed in a company data center, as well as may be deployed at a co-location facility for providing better performance of application and services.

This model of clouds can be built and handled by a company's, IT organization or by a particular cloud provider. It is also called as hosted private model, for example a company such as Sun can install, run, and configure the infrastructure to support a private cloud along with company's enterprise data center. This model provides a high level of control to companies over the consumption of cloud resources while conducting the expertise needs or demands to establish and operate the environment (Sun Microsystem, 2009).

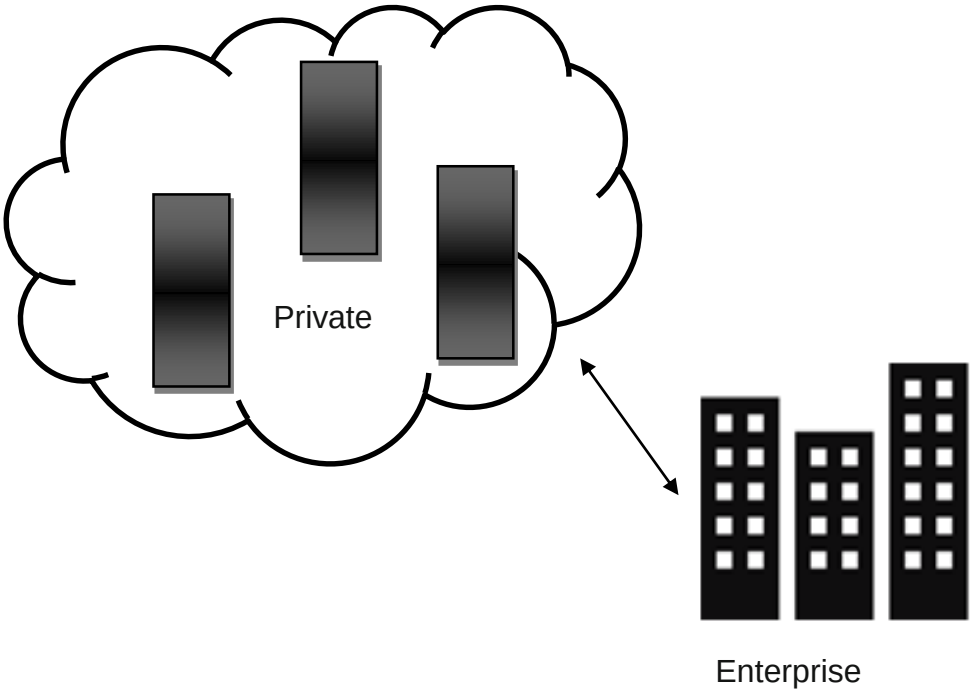

Figure 2.6 Private Cloud

NIST describes "a private cloud as a cloud infrastructure operated solely for an organization, organized by the organization or a third party, and existing either on premise or off-premise".

It is different from public cloud with two ways. Firstly, a private cloud is basically dedicated to an individual enterprise. Secondly, security is the major part to be considered tighter in private cloud deployment than it is in public cloud (Krutz & Vines, 2010).

2.4.3 Hybrid Cloud

A composition of two types of cloud is called a hybrid cloud, where a private cloud is able to maintain high service availability by scaling up their system with externally provisioned resources from a public cloud when there are rapid workload fluctuations or hardware failures (Antonopoulos & Gillam, 2010). Hybrid cloud is a

combination of both public and private cloud designs and models. Same functionality is given by this model as given by public and private cloud, it can help to provide resources on-demand and externally provisioned scale. The ability to enlarge a private cloud by using the resources of a public cloud can help to maintain quality of service levels in the face of rapidly workload fluctuations. A hybrid cloud also can be easily handling planned workload spikes. Sometimes called as “surge computing,” in this a public cloud can be monitor and perform periodic tasks that can be created and deployed easily on a public cloud (Sun Microsystem, 2009).

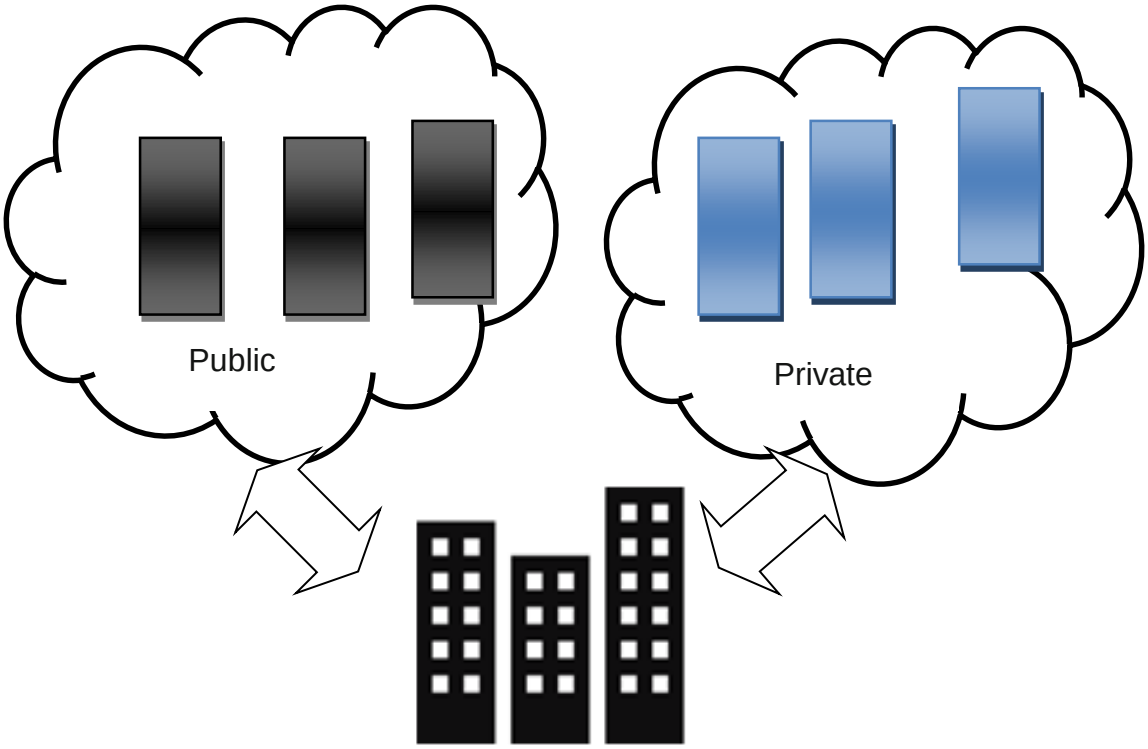

Figure 2.7 Hybrid Cloud

NIST defined it as “a composition of two or more clouds that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability”. For example, hybrid cloud may consist of an organization which is deploying software applications in the public cloud, at the same time keeping sensitive applications in a private cloud. (Krutz & Vines, 2010). Hybrid clouds generate the complexity of observing how to distribute applications, services, security etc. for both public and private cloud. Among all the issues that need to be examined is the relationship between the data and processing resources.

2.4.4 Community Cloud

In community cloud infrastructure is shared by two or more organizations, companies and supports a particular community that has shared resources (e.g., mission, policy, security requirements and compliance considerations). It may be managed by the organizations or a third party and may exist on premise (YS, 2013). According to NIST, "The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns. It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises" (Mell & Grance, 2009). Services of Community Cloud are multi-tenant shared same computing resources, giving managed infrastructure and platform services with the virtual machine automated provisioning abilities.

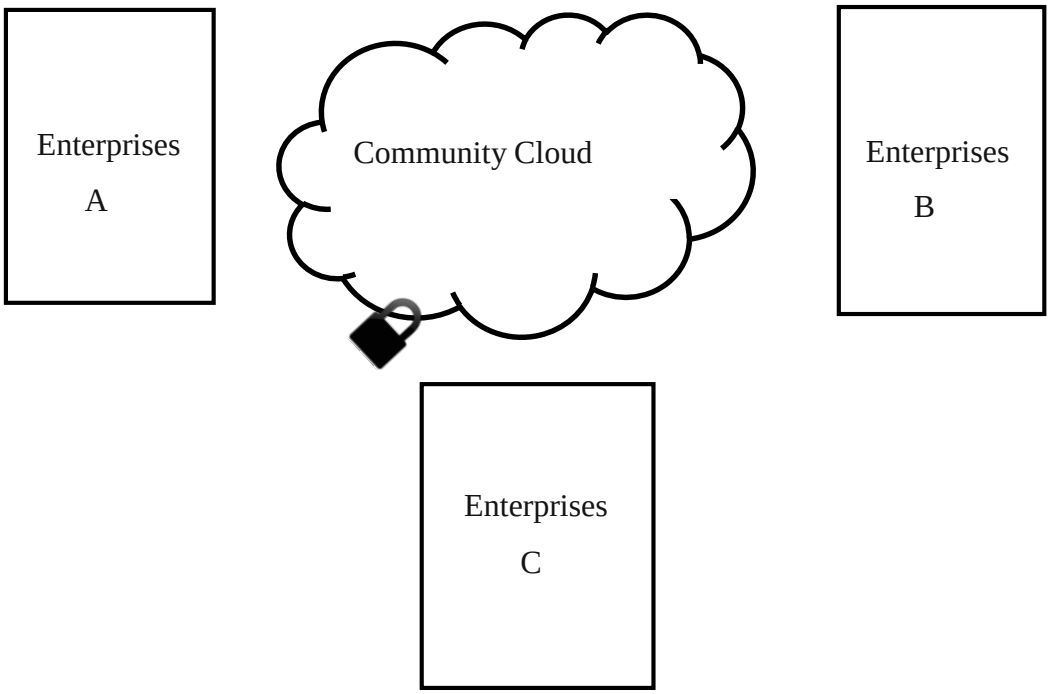

Figure 2.8 Community Cloud

This model gives privilege to customers for managing the size of major investments and generate bursting ability to meet peak demand level at a lower cost than a private cloud model.

The Community Cloud aspires to combine distributed resource provision from Grid Computing, distributed control from Digital Ecosystems and sustainability of Green Computing, with the utilities of Cloud Computing, while creating greater use of self-management advances from Autonomic Computing. Replacing vendor Clouds by establishing the underutilised resources of customer machines to form a Community Cloud, with nodes potentially fulfilling all roles, consumer, producer, and most importantly coordinator (Marinos & Briscoe, 2009).

2.5 Need of Cloud Computing

In traditional computing, a number of servers are used to store data and information. It includes high cost to maintain servers in order to work properly. Many type of costs are present in traditional computing, such as, cost on cooling systems, heat production in servers, power failure, overload, hard to get backup of servers, less portability of servers from one server to another server while transferring application. But cloud computing is a fully virtualized technology. This model uses virtual servers, data centers, storage etc. for performing computing. It can easily reframe, reuse and re-design the applications and services as per the requirements.

The aim of this computing model is to use distributed resources in better way, and to put them together for achieving high throughput, and making it capable in managing large-scale computation (Antonopoulos & Gillam, 2010).Cloud computing is viewed as a web based computers, resources and services which help the developers to make complex web based systems. Cloud oriented resources are virtual which can give more resources to the systems if need arise suddenly. All resources are in the form of virtual structure, e.g. virtual server, disk space, hosts, application etc., which can simply be added on demand and generally to the application that uses them. Due to its virtual nature, it makes easier to resize the cloud based solutions. Companies that rely on cloud only pay for resources consumed by them and thus effectively reduce the cost generated by data centers and processing resources (Jamsa, 2013).Prime attribute of cloud computing is to remove and disguise complexities for enabling agility, scalability, flexibility and on demand service delivery (Schulz, 2012).

Cloud Computing is an eminent approach to supply computing resources. Six paradigm of computing is given in (Furht & Escalante, 2010). In first phase mainframe computing lies, in this user used mainframe and terminal computing in a

resource sharable manner. In PC computing only PCs are used by the user as per their requirements.

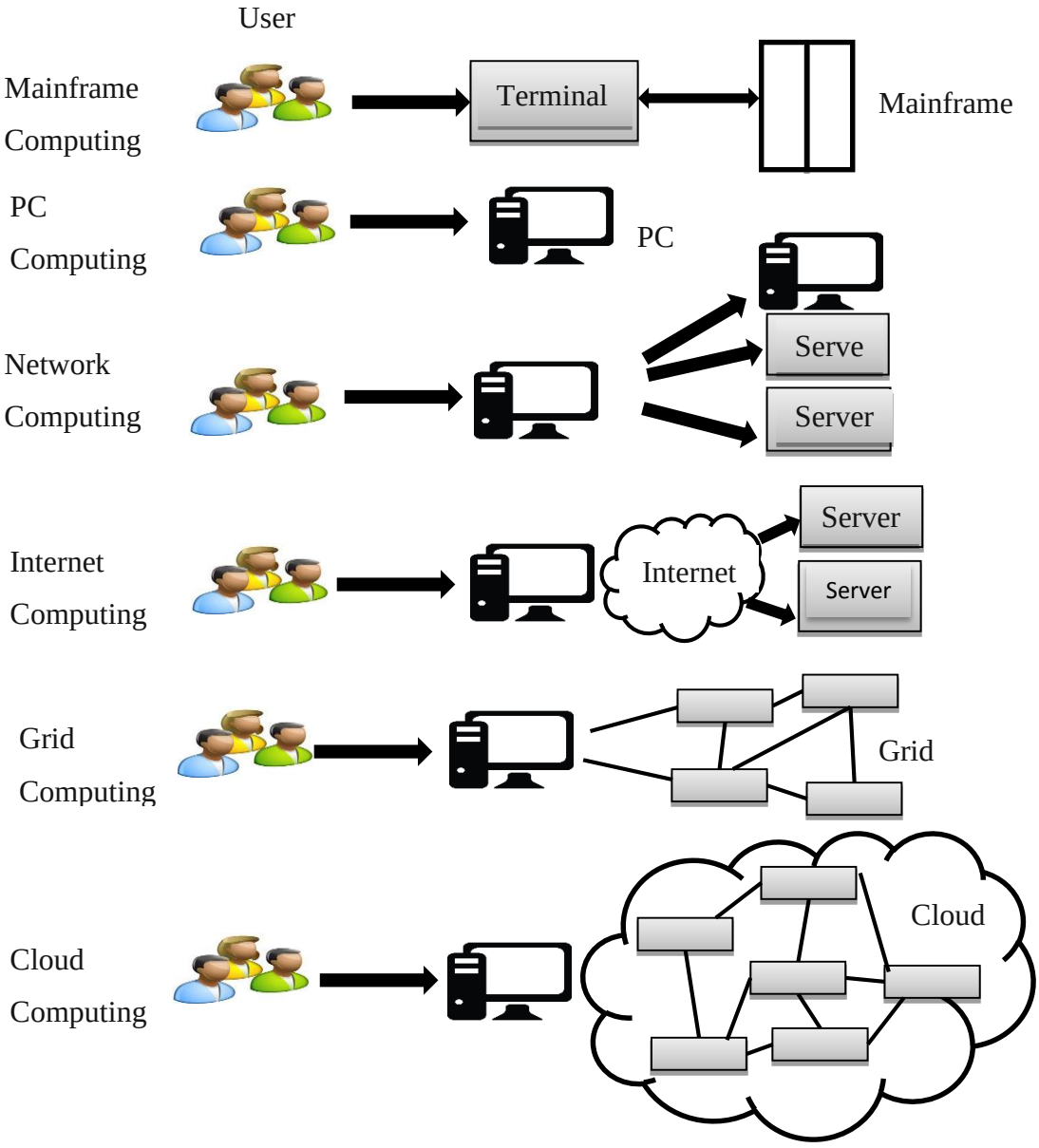

Figure 2.9 Six Computing Paradigms (Voas and Zhang (2009))

Network computing is involved computing devices such that PCs, servers etc. In next phase internet computing is scales up the performance of computing. It is consists internet connections and servers for completing the needs of users. The growth of internet gives origin to grid computing, along with the availability of high speed networking in the computing world.

Computational power of resources, data and services readily available to the users with different priority levels of diverse area. Services can interact with users' minimal interference of humans (Parashar, & Lee, 2005). In the next generation cloud computing is comes, it is developed from many computing technologies and business approaches that employed from number of year. It gives ability to use computing infrastructure and storage resources on a pay per you use basis and reduce the cost of infrastructure in industry sector.

The Major benefit of cloud computing is that all the users can access an eminent variety of resources without any implementation or configuration of the whole infrastructure. It is a disruptive technology which is getting momentum (Antonopoulos & Gillam, 2010).

2.7 Characteristics of Cloud Computing

2.7.1 On-Demand Self-Service

Cloud Computing enables users to use resources on-demand from anywhere in the world, without the direct interferences of human and cloud provider. In order to provide highly effective and sustainable resources to users, the interface should be user-friendly and flexible to maintain the service requests and responds (Krutz & Vines, 2010). Users can access cloud services through online control panel and they can scale up their infrastructure up to an acceptable stage without disrupting the host operations and configurations. These services are provided to the user by cloud provider on pay as you use basis, users have to bill the amount for resources used in a particular time or for particular amount resources.

2.7.2 Availability and Mobility

Users from all over the globe can access cloud services at anytime from anywhere. They have the capability to request resources through a standard internet connection. The high availability of cloud services makes enterprises to use more amounts of resources for their complex infrastructure. It provides mobility by giving access of data and applications to the user from around the globe.

2.7.3 Scalability

In cloud computing users have freedom to access a large amount of services that scale based on their demand. Cloud provider provides services to the user in the peak times and must also fill the demands of multiple applications during busy situations of host.

2.7.4 Elasticity

Elasticity means a cloud have the ability to expand and reduce services as need arrives. Resources efficiently have changed to meet the needs of users at any time. This dynamic allocation of resources might be done automatically and it seems to the user as a large scale of resources that paid as you use when needed. The service provider is transparently manages users resource need based on dynamic provisioning of resource allocation.

2.7.5 Cost Effective

Cloud Computing is a pay per usage model which allows customers to pay for the resources they required with generally no investment in the hardware infrastructure present in cloud. There is no need of hardware maintenance, configuration and updating at the customer side. The cloud is available at much cheaper rates and consequently, can significantly lower the organization's IT expenses. In addition, there are many one-time-payments, pay-as-you-go and other scalable, reliable options available, which make it very reasonable for an organization.

2.7.6 Location Independent Resource Pooling

The cloud has a large and acceptable resource pool to provide the customer's requirements and provide economic scale, meet the acceptable Service Level Requirements of large group of customers. All the resources are efficiently allocated to wide range of applications for optimum performance execution. These resources (server, data center, storage etc.) are placed at any geographic location, but can be accessed virtually by the users from anywhere around the globe. According to NIST "There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (eg. Country, state, or data center)" (Krutz & Vines, 2010).

2.7.7 Collaboration

Cloud gives the opportunity to user to communicate and share easily outside of the traditional methods. There is no issue of different locations, user can access data and application through cloud and if a company is applied some policies or other changes in their own field than outsider (vendors, contractors etc.) can easily access their confidential files or terms and conditions by getting privilege from the owner from different locations. Cloud makes easy to share records, information with

third party or advisors. It is basically provide common data and information to work simultaneously for users.

Software companies are more adaptable, flexible and cost-effective, they allow consumers to choose only what they use or need rather than off the shelf becomes as a large initial investment. In current computing environment, despite how many companies and other organizations regularly use cloud-based services, number of terminology associated with cloud computing.

2.8 Component of Cloud Computing

2.8.1 Data Center

A data center (sometimes called a server farm) is a centralized depository for the storage, management, and distribution of data and information. Typically, a data center is a possibility used to hold computer systems and associated parts, such as telecommunications and storage systems. Mostly, redundant, unneeded or backup power supplies, redundant data communications connections and security devices are installed in data centers (Stryer, 2010). A data center takes much longer to get started and can cost businesses $10 million to $25 million per year to run (Angeles, 2013). Data centers can absorb up to 100 times more energy than a standard office building. Data center energy consumption doubled from 2000 to 2006, extending more than 60 billion kilowatt hours per year. This number doubled again in 2011 (EERE, 2013). Less than half the power used by a typical data centre is actually required to operate its IT equipment and the other half goes to support infrastructure including cooling system, UPS inefficiencies, power distribution losses and lighting (International Resources Group, 2009).

2.8.2 Server

Cloud server provides services to customers on demand via the Internet. Rather an individual server, group of connecting servers provide cloud services (Srikantaiah, Kansal& Zhao, 2008). The proper utilization of these servers need to be considered to reduce energy requirements. As large-scale cloud providers tend to operate their infrastructure at upper and more stable utilization levels than corresponding on-assumption operations, the same tasks can be performed with far fewer servers (WSP, 2010). On the consolidation side, the operating system and applications of multiple underutilized physical servers are shown being consolidated

onto a single or, for redundany, many servers in a virtual environment along with a separate virtual machine accmodating a physical machine (Schulz, 2012).

2.8.3 Virtualization

Virtualization is the technique of partitioning the resources of a server into multiple VMs. Virtualization can offer dramatic benefits for a computing, including improved utilization, energy saving, fast deployment, enhanced maintenance capability, segregation, and encapsulation. Virtualization permits applications to move from one server to other server while applications are still running, without downtime, providing flexible workload management, and high availability during planned maintenance or unplanned events (Antonopoulos & Gillam, 2010). It provides a level of autonomy of choice for the client, and effectiveness and cost savings for providers. Moreover, in many cases migration to virtualized clouds will be primarily through hybrid clouds as according to customers test and estimate the cloud standard in the perspective of their own corporate requirements (Krutz & Vines, 2010). For applications, services and data that do not add themselves to consolidation, there is a different form of virtualization technology is to allow transparency of physical resources to keep inter-operability, coexistence, locality between new and existing one software tools, storage, servers and networking technologies, for example, creating new more energy efficient storage or servers with improved performance of resources to coexist with already existing resources and applications (Schulz, 2012).

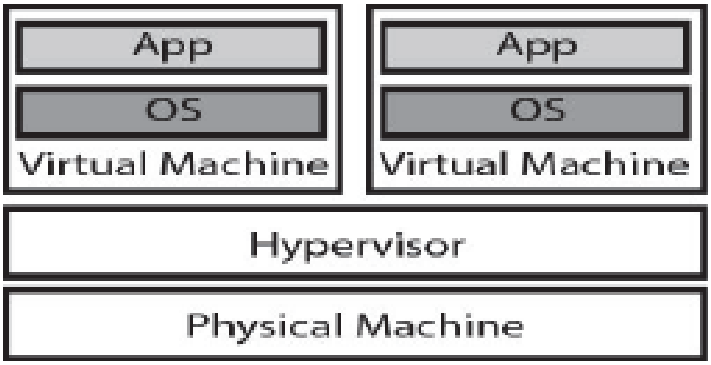

Figure 2.10 Virtualization (Strunk & Dargie, 2013)

In a virtualization environment a virtual machine act as a physical machine, this can run in parallel and in isolation manner, for sharing the same physical resources.

A middleware called a hypervisor abstract these virtual machines from the hardware and observes the current use of resources by VMs (Marinos & Briscoe, 2009).

2.8.3.1 Server Virtualization

Server virtualization can be implemented in hardware or assisted by hardware, implemented as a standalone software running bare metal with no underlying software system required, as a component of an operating system, or an application on a running existing operating system (Schulz, 2012).Mostly server virtualization is done by the usage of a hypervisor to reasonably assign and dispersed physical resources. Basically, the hypervisor permits a guest operating system, which is running on the virtual machine, to function it as solely in the control of the hardware, guest operating system is unaware that many other guests are sharing it. Every guest operating system is saved from the others and so it is unaffected by any variability or configuration matters of the others (Antonopoulos & Gillam, 2010). The motivation is to reduce the complexity of managing the so called server sprawl (Hill, Hirsch, Lake & Moshiri, 2013).

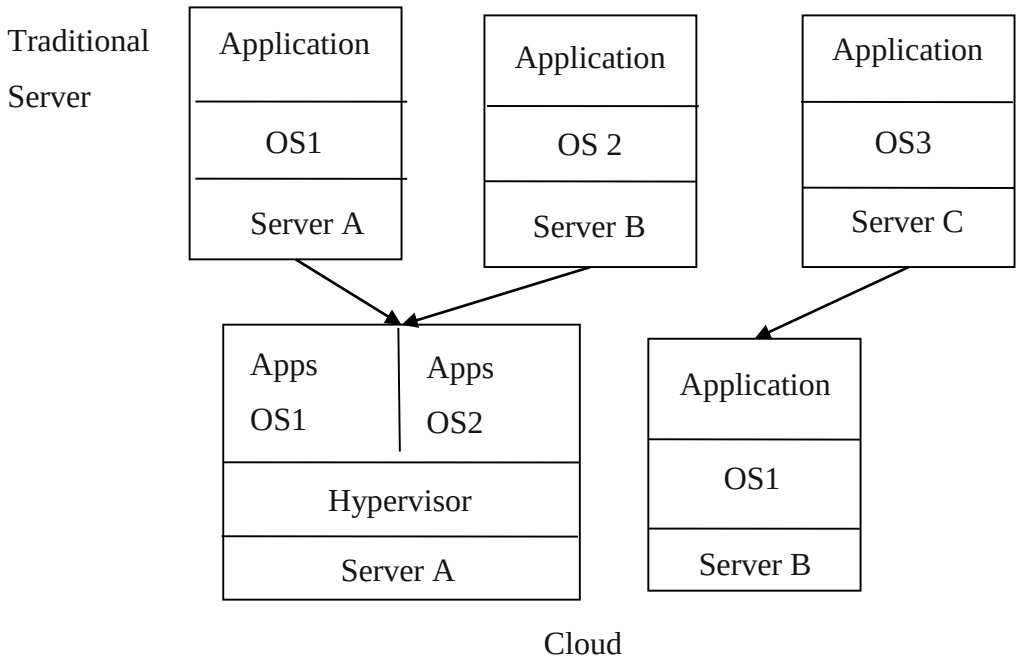

Figure 2.11 Server Virtualization

An example of virtualization: in non-cloud computing there is a need for three servers; in the cloud computing, two servers are used.

2.8.3.2 Desktop Virtualization

In desktop virtualization, desktop runs a hypervisor that also downloads the virtual machine image from the server and launches it on the end point client. In this manner the processing power of the end point is fully exploited, a VM image can be cached for efficiency and only incrementally updated when needed, and finally user data, which can be large, need not be centrally maintained but mounted from the local disk as soon as the virtual machine boots (Hill et al., 2013). With desktop virtualization, each user gets a virtual machine that contains a separate instance of the desktop operating system and whatever applications have been installed.

Desktop virtualization however, breaks the direct connection between physical hardware, operating system, application and display (Gupta, Nayak & Pattnaik, 2012). A single desktop computer can run simultaneously multiple operating systems. It consumes less power and reduced the need for duplicate hardware (Krutz & Vines, 2010).

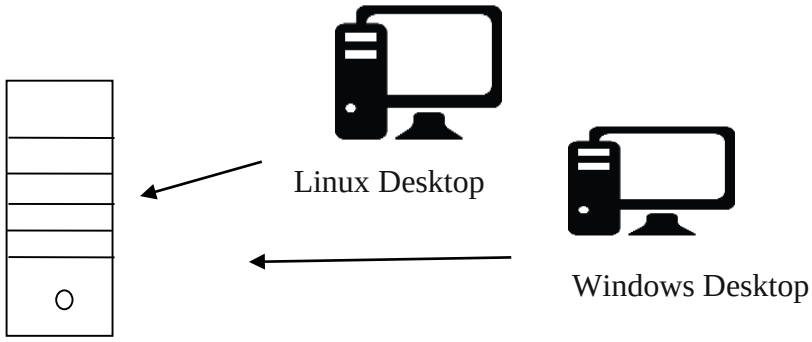

Figure 2.12 Desktop Virtualization

There are many other type of virtualization such as network virtualization, application virtualization, presentation visualization and storage virtualization.

Need of Virtualization (Jamsa,2013), (Krutz, & Vines, 2010):

- Increased device utilization
- Decreased device footprint
- Decreased power consumption
- Simplified operating system and application administration

- Ease of software provisioning
- Device and storage stability
- Increased user access to key resources
- Increased flexibility in supporting multiple operating system environment
- Improved use and management of software licences
- Improved disaster recovery and business continuity
- Rapid deployment of additional servers
- Promotion of economic scale
- Fault tolerance
- Usage based on service level agreement

2.8.4 Virtual Machine

A virtual machine is a software program implementation of a virtualized computing environment which includes an operating system or other related program for installing and run. The virtual machine usually imitate a physical computing environment, but requests for network, memory, hard disk, CPU and other physical resources are handled by a virtualization layer, it translates these queries and requests to the underlying physical hardware resources. Many VMs can be dynamically start and stop on a single physical machine according to incoming requests, hence providing the flexibility of configuring several partitions of resources on the similar physical machine to different requirements of service requests. Multiple VMs can simultaneously run applications based on many operating system environments on a physical machine. With dynamically migrating VMs across the physical machines, loads can be build up and unused resources can be switched to a low-power mode, turned off to operate at low-performance levels to save energy (Beloglazov et al., 2012).

Each virtual machine can have a different operating system and a virtual machine monitor is used to manage the VMs on a physical node. A Virtual Machine Migration is often referred as a hypervisor (Antonopoulos & Gillam, 2010). It is the virtual machine monitor that enables a physical machine to be virtualized into different VMs. Virtual Machine can have different instructions sets from the physical hardware if needed. Even if the instructions sets are the same, the size and number

of the physical resources seen by each virtual machine and in fact will usually be different (Hill et al., 2013).

2.8.5 Virtual Machine Migration

VM migration is one of the important capabilities of system virtualization, which allows applications to be migrated along with their execution environments across physical machines. Live migration further allows the VM to be migrated almost without any interrupt to its application's execution. VM migration is an important means for managing applications and resources in a large virtualized system. As the scale of virtualized systems such as clouds continue to grow, the use of VM migration, particularly live VM migration, becomes increasingly more important and frequent for optimizing application executions and resource usages in the systems. VM migration can be resource intensive on its own and specifically it can consume substantial CPU cycles and network bandwidth. Hence, one VM's migration would compete for resources with other VMs' application executions as well as their migrations. Meanwhile, the resources available to perform a VM migration would also affect the performance of the migration and consequently the performance of the VM's application (Wu & Zhao, 2011).

Virtual machine migration has been investigated in the systems research community also in related areas such as clustering, grid computing. Migration of virtual machine involves keeping and copying the whole state of the machine at a snapshot in time, involving processor and memory state along with all virtual hardware resources such as devices or network MAC addresses, BIOS. In basic principle, this also involved the entire disk space, including system and customer directories, files along with swap space used for virtual memory operating systems scheduling (Hill et al., 2013).

2.8.6 Service Level Agreement

Service Level Agreement addresses the legal binding contract, which states that Quality of Service guarantees that the provide promises to provide the execution environment to its hosted application on the client side (Gupta, Nayak & Pattnaik, 2012). Providers use them to describe their supply of resources. Customers use them to describe their demand for resources. SLA is a component of a service contract where the levels of services are properly defined. In reality, the term SLA is sometimes used to mention the contracted delivery time or performance. The SLA is report a common understanding about services, priorities, penalties,

responsibilities, guarantees and warranties. Each area of service scope should have the "level of service" defined. The SLA may define the levels of availability, serviceability, reliability, performance, operation, or other attributes of the service (Karadsheh, 2012). Allocation of Cloud computing resources is based not only on functional requirements but also on different non-functional requirements, e.g., application execution time, reliability, and availability also termed as quality of service (QoS) requirements and are expressed by means of service level agreements (SLAs). Despite the existence of SLAs, providers and customers of computing resources face the problem of varying definitions of computing resources in Cloud computing markets. Computing resources are described through different non-standardized attributes, e.g., CPU cores, execution time, response time, inbound bandwidth, outbound bandwidth, and processor type (Maurer, Emeakaroha, Brandic & Altmann, 2012).

The SLA (Schulz, 2010) specifies the quality of service (QoS) between a service provider and service consumer, and usually includes the service price and the level of QoS adjusted by the price of the service. For instance cloud provider can charge a higher price to a consumer who requires a high level of QoS (quality of service). Flexible and reliable management of SLA agreements is of vital importance for both Cloud providers and consumers. On the one hand, prevention of SLA violations avoids penalties that providers have to pay and on the other hand, based on scalable, flexible and timely reactions to possible SLA violations, enables Cloud computing to take roots as a flexible and reliable form of on-demand computing (Emeakaroha, Netto, Calheiros, Brandic, Buyya, & De Rose, 2012). As the utilities are delivered over the cloud to the end user it frames the complete system into a pure business model and the cost trade-off is also to be a primary factor whose transparency is to be mentioned o the customers (Gupta et al., 2012).

CHAPTER 3

REVIEW OF LITERATURE

The previous chapters have given brief introduction of cloud computing technologies and terminologies. This chapter reviews the work done in the area of power management and energy efficiency. This chapter continue with explaining the earlier power management approaches in clusters, server, data center until cloud computing has come into light. This discussion surveyed energy/power aware approaches with hardware, software, servers, data centers, virtualization and operating systems.

Cloud Computing has praised for its service deliveries and for its cost effective features. But growing cost of energy is a major threat to cloud computing. As data and computing are increasing quickly, there is a need of large amount of data centers and servers to process them. It leads to increase the power and energy usage of computing. Most of the researchers focused on this emerged problem that how to reduce the energy consumption of data centers. In the area of power management the first work is done by (Pinheiro, Bianchini, Carrera & Heath, 2001), they came into light the idea to save power and energy for clusters of workstations. They have developed such type of systems that can dynamically change the state of cluster nodes from on to off or off to on. In ON state is efficiently manage the load imposed on the system and OFF state is to save power under low load. This technique is followed by algorithm which made decisions for cluster configuration and load distribution. Decisions are totally depend on total load imposed on the cluster and the power and performance of different cluster configurations. Scheme is considered throughput and execution time of applications by considering the load of resources like CPU, disk storage and network interface for two types of servers, named as power aware cluster based network server and power aware operating system server. The algorithm can adds or removes one node at a time of instant, this can leads to performance degradation of applications in a large scale computing environment. This work was helpful in saving energy of data centers but have performance reduction Service Level Agreement (SLA) and operations are time consuming.

Bohrer, Elnozahy, Keller, Kistler, Lefurgy, McDowell& Rajamony, (2002) have considered power management in server systems. Web server systems have major

use of data center server. Dynamic Voltage and Frequency Scaling techniques is used to measure the energy of web servers, usually this technique is used for portable devices. For getting energy efficient web server, there are many techniques, which include condition of CPU in halted and runnable state in accordance with CPU can go to inactive or active state for saving power in terms of energy. CPU has high as well as low utilization of processing. At low utilization CPU component can be idle, that leads to spun down them and rest of the components can work in less power. More power consuming components are CPU, disk or memory. Paper examines the power consumed by server with the measurements of three web sites and results shows CPU consume the large fraction energy. Energy consumption is reduced up to 36% by using Dynamic Voltage and Frequency Scaling in CPU. So, there is a higher possibility to save energy with CPU power management.

Elnozahy, Kistler & Rajamony, (2003) investigates the power management in servers by employing five different policies and calculates energy usage and response time. Dynamic Voltage and Frequency Scaling and Vary - on/ Vary - off techniques are deployed in policies. Energy usage and response time of web servers are measured by considering these policies; Independent Voltage Scaling (IVS), Coordinated Voltage Scaling (CVS), Vary-On Vary-off (VOVO), Combined Policy (VOVO-IVS), Coordinated Policy (VOVO-CVS). Paper has used Olymoics98 like in (Bohrer et al., 2002) for obtaining workload from internet server. Paper investigated that voltage scaling is a better mechanism than vary-on/vary-off, because it decreases cluster capacity and consumes less amount of power. Voltage scaling has no effect on hard drives during frequency power cycle and it can be used openly in many energy saving techniques. Paper considered only response time and less amount of requests, in other words system have less amount of workload, but this capacity may vary from low to high, cause performance degradation in the form of service level agreement. Its importance is less significant with heterogeneous workloads.

Till now it is clear that CPU consumes the maximum power in computing. The next paper discusses the energy saving of I/O devices.

Gniady, Hu & Lu, (2004) studied the problem of energy saving of I/O devices by proposing Program Counter based Access Predicator (PCAP). It predicts the idle periods dynamically when I/O devices get shut down to save energy. PCAP apply

path-based correlation to examine a specific sequence of program counter for each idle period and information is accurate. It can continuously improve the several invocations of the same program and also can predict the future events of an idle period. Paper uses correlation between the application, I/O operations and users. Paper shows that 76% energy saving is done by PCAP along with the low mispredictions.

Femal & Freeh, (2005) have done work on the problem of power saving in circuits of components by over provisioning. Power management is a primary issue in circuit capacity, not to increment it from its original structure, due to the present needs. Paper proposed host-based and network centric models for the distribution of power and to achieve maximum throughput (considered only throughput in service level agreement).

Poellabauer, Singleton & Schwan, (2005) has been worked. Paper focused on the problem of energy saving in real time applications while considering only memory bound applications. Proposed work introduced the future utilizations of the system with new feedback based DVFS approach by monitoring memory access of real time applications. DVFS calculates CPU utilization, frequency and voltage combinations. Feedback approach followed run time of tasks and cache miss rates. Here, cache miss rates are used as indicators for memory access rates which computes future task run time efficiently. This techniques is applied on six applications for those cases it saves 27% energy, But also have some cons. Inaccurate predictions can cause exceeded utilization and this leads to performance degradation of system and re-computations of frequency and voltage. This approach does not work on I/O bound application and communication-intensive applications.

Kotla, Ghiasi, Keller & Rawson, (2005) have worked processor voltage and frequency. Paper focused on the frequency and voltage of processor that are going to assign different type of workload. It dynamically assign the processor power to various workloads. It make easy to migrate work from one processor to other and also prevent overhead. For saving energy in servers (Chen, Das, Qin, Sivasubramaniam, Wang & Gautam, 2005) presented a problem of saving energy in servers with multiple application running on it, in accordance with the service level agreement. Like (Kotla et al., 2005), this approach apply server provisioning to different applications running on severs. It employ DVS technique for managing

power in servers. Paper introduced proactive, reactive techniques for predicting future workloads and managing time for different workloads. Due to the overwork load techniques can give performance degradation and may increase the operational costs due to the on and of servers.

Heath, Diniz, Carrera, Meira & Bianchini, (2005) discusses the problem of energy conservation in servers, they suggests that the need of computing is to generate self-configure power for heterogeneous servers, which can distribute power; throughput and can optimize configuration modelling. But this type of structure has some limitations like operational costs, cooling costs, performance degradation etc.

Isci, Contreras & Martonosi, (2006) have formulated a framework which includes runtime phase predictor for observing dynamic power management in the real time systems by using DVFS algorithm. But have no significant effect on power management and also in terms of service level agreement (Considered only execution of application.

Nathuji& Schwan, (2007) have investigated that how virtual technologies are used to decrease power consumption data center. They proposed virtual power technique for online power management that supports isolation and independence in operation of guest virtual machines and it controls the effect of power management of virtual machines in virtualized environment. They introduced a new power management technique that is soft resource scaling, which contain hard and soft provisioning of power management. Hardware scaling is generated by the virtual machine monitor by providing less resource time to virtual machine. Due to the restricted states of hardware scaling, soft and hard technique can give high amount of power saving. Author suggests that system designers should make power efficient systems because not every part of a computing system consumes peak amount of energy, it varies according to the workloads. Some components have high amount of energy consumption and some have very low. Paper gives a new idea for resource management, at local level and at global level, but it have limitation that at the guest operating system might be uniform power management or have non uniform power management.

As earlier discuses that each component of system consumes varying amount of energy. Memory devices have different states of consuming energy, each state consume different amount of energy.

Diniz, Guedes, Meira & Bianchini, (2007) formulated techniques for restricting the power consumption of main memory devices. They proposed four techniques; Knapsack, LRU-Greedy, LRU Smooth, LRU ordered, these techniques are modified to some extent, so that if device crosses the limited budget of power than new state of devices is originated. Expect Knapsack, rest of the three techniques uses time thresholds for changing the state of a device. Enough space consumption, overheads and complexity are the drawbacks of these approaches. They also accept that work is not suitable for multiprogramming events and limiting the power budget for different states of device is a major issue problem.

Raghavendra, Ranganathan, Talwar, Wang& Zhu, (2008) have investigated data center power consumption by applying coordination of five diverse power management approaches. They use nested structure of multiple feedback controls at different levels. Paper described control theory for applying feedback control loop to coordinate the controller's events. Work performs mathematical analysis and provides an opportunity for dynamic changes. Author suggests that virtual machine technology can be a more efficient approach for decreasing power consumption in terms of saving energy. This work is not suitable for strict SLAs or for varying SLAs to different types of application. So, cannot applicable in cloud enterprises because of the performance degradation of services.

Chen, He, Liu, Nath, Rigas, Xiao & Zhao, (2008) have worked on energy aware server provisioning in internet hosting services. They have used real data from windows live messenger for measuring performance and power models. Approach followed server transient behaviour and favoured different load dispatching algorithms. Work saves a significant amount of energy excluding the user experiences.

Kusic, Kephart, Hanson, Kandasamy & Jiang, (2009) implemented a dynamic resource provisioning framework in heterogeneous virtualized environment in a way of sequential optimization and employed it using limited lookahead control. It calculated the costs of virtual machine provisioning and encodes optimization problem. The purpose of this works is to increase the profit of providers with reduced power consumption and less SLA violations. A kalman filter is used to predicts the future arrivals, due to the time varying nature of workloads, it's impossible to determine priori distribution resources. This work needs simulation based learning for implementing it on specific applications. Due to the long execution time (30 min

for 15 nodes) of optimization controller, it is impossible to deploy it on large scale systems.

Mahadevan, Sharma, Banerjee & Ranganathan, (2009) have applied energy saving techniques on networks. Networks include data center and IT enterprises networks. This scheme has saved closed to 75% energy, also with better performance and reliability. it has SLA violations.

Srikantaiah et al. (2008) have discussed the problem of scheduling requests for web applications in virtualized heterogeneous environment. Author wants to reduce energy consumption with considering service level agreement actively. Virtual machine migration may generate additional costs for initiating the required configurations on systems. At high utilization of various resources performance degradation incurred during workload consolidation. Experimental work shows that energy consumption per transaction gives a U shape curve which leads to conclude that an optimal point of utilization can be formed. They introduced a heuristic for the multidimensional bin packing (NP problem) problem to manage the optimization of multiple resources for the workload consolidation. This approach is considered workloads and application dependent. Also dynamic resource allocation leads to on and off servers during consolidation, sleep and wake up modes may incur extra costs.

Cardosa, Korupolu & Singh, (2009) have proposed a technique for virtual machine placement and power consolidation of VMs in data centers. They studied the problem of power efficient allocation of virtual machines in heterogeneous virtualized computing environment. Min, Max and share techniques are used in virtualization technologies. Resource allocation of VMs is done on the basis of available resources, power costs and application utility. They have implemented the min, max and share parameters of VMM. Approach is represents minimum, maximum and proportion of the CPU, which is further allocated to VMs for sharing the same resources. But this approach has some limitations in the sight of SLAs and VMs placement. It does not follow the strict SLA and have some limitations to prioritise the sharing parameters of applications. During virtual machine allocation, VMs are not placed at run time and only CPU is considered in this approach, none of other resources are included in this process.

Chiaraviglio, & Matta, (2010) have introduced an approach for power management between Content Provider and Internet Service Provider to decrease

the total power consumption. Internet Service provider is the owner of computing infrastructure, it handles physical topology which consists a set nodes and links. On the other hand, a content Provider consists a number of servers which is further connected to Internet Service Provider. Nodes are assigned to the server on the basis of preferential degree placement. ISP nodes are grouped according to the decreasing links with cities at different locations. CP servers are provided to only those nodes which high connection group. They have achieved he gaol of minimizing the energy consumption in network during allocation of resources, but considered simple traffic, an additional traffic complexities can produce obstacle in the allocation of resources and energy saving. Also they haven't paid attention to quality of server parameters, which have an important part of Service. Level Agreement.

Gyarmati & Trinh, (2010) have investigated the power saving architecture of data center. They suggest that data center architecture should consists additional switches, eg., number of ports. Although, optimization of network architecture can be deployed only at the data center design time and cannot be implemented dynamically.

Guo, Lu, Wang, Yang, Kong, Sun & Zhang, (2010) implemented virtual data center (VDC) and SecondNet architecture is used to support its design, implementation and evaluation. The presented architecture is assuming that a data center network is owned by a single entity. A virtual cluster management system allocates the resources for satisfying bandwidth guarantees. The allocation resources is observed by a heuristic which minimized the total bandwidth utilization. Virtual Machine allocation is done when some VMs are de-allocated. Also allocation is not dynamic and it depends on present state of network load. However, this approach has achieved high network utilization but does not minimize the energy consumption by the network.

Rodero, Vaquero, Gil, Galán, Fontán, Montero & Llorente, (2010) have formulated a new layer closer to the lifecycle of the service in the cloud that allows automatically deployment of service along with multi virtual machine configurations. This layer is stands at the top of the cloud and manages the resources execution of the different services. Proposed approach specified the major rules for scaling the VM configuration in and out. It does not consider the network communication between virtual machines.

Ye, Huang, Jiang, Chen & Wu, (2010) work represents a virtual machine based architecture for energy-efficient data center for cloud computing. Efficient provisioning, consolidation and migration policies can improve the energy efficiency. Then they studied the potential performance overheads occurred due to server consolidation and migration of virtual machine policy. Their experiments results conclude that both the two approaches can effectively employ energy-saving goals. But live migration of virtual machine incurs performance overheads.

Chetan, Geevarghese & Karthik, (2010) presents a placement algorithm which allocates various resources such as memory, bandwidth, MIPS, processing power, etc. from a physical machine (PM) to VMs such that the number of PMs (Physical Machine) used is minimized. They proposed two algorithms, Reduction to the Set Cover problem, LP formulation. But the problem of placing virtual machine is still there, that is when, where and how to migrate virtual machine for energy efficiency of cloud.

For virtualized cloud data an energy efficient resource management system has been proposed (Beloglazov & Buyya, 2010) proposed. Approach reduces operational costs and provides Quality of Service (QoS). By consolidation of VMs helps to savings energy according to current utilization of resources, on the other hand virtual network topologies created between VMs and thermal state of computing links. They present results of simulation-driven analysis of heuristics for dynamic reallocation of virtual nodes using live migration according to current needs for CPU performance. The heuristics have been evaluated by simulation using the extended CloudSim toolkit. One of the method leads to significant reduction of the energy consumption by a Cloud data center by 83% in comparison to a non-power aware system and by 66% in comparison to a system that applies only DVFS technique but does not adapt allocation of VMs in run-time. Moreover, MM policy enables flexible adjustment of SLA by setting appropriate values of the utilization thresholds: SLA can be relaxed leading to further improvement of energy consumption.

Jayasinghe, Pu, Eilam, Steinder, Whally & Snible, (2011) focused, on how to improve performance and availability of services hosted on IaaS clouds. They arrive with structural constraint aware virtual machine placement (SCAVP), supports three types of constraints: demand, communication and availability. But there is a problem

of Virtual Machine placement, they formalized that Virtual Machine placement is a NP hard problem.

Man& Kayashima, (2011) have given the idea of finding optimal placement of virtual machines. Virtual Machine placement on number of physical servers is one kind of bin packing problem, as discussed by (Ye et al., 2010), which is known as an NP-hard. Author proposed the use of a heuristic algorithm to solve this problem where the virtual machines play the roles of personal desktops. Proposed algorithm that is PBA (pattern-based allocation) can be used to reduce the number of physical servers required for hosting a certain number of virtual desktops. With this algorithm, the number of servers required for placing virtual desktops is reduced. But, research work is mainly concentrate on desktop infrastructure, not on the overall cloud environment for implementing virtual machine placements.

Esfandiarpoor, Pahlavan, & Goudarzi, (2013) considers four energy-aware resource management algorithms for virtualized data centers so that, total energy consumption of data center is minimized. Author proposed a new algorithm (OBFD) that sorts list of VMs in decreasing order of their required MIPS instead of current CPU utilizations. After ranking VMs, it tries to find the best server for each VM that leads to the minimum increasing of power consumption of data center. They give placement algorithm by considering rack utilization, Place virtual machines rack by rack. But there is a drawback of these algorithms is that, algorithm saves data center power consumption but not overall network power consumption and it considers only racks.

Lin, Liu & Wu, (2011) proposed a method, that is Dynamic Round-Robin method, to deploy the virtual machines for power saving purpose. Dynamic Round-Robin is an extension from original Round-Robin, in this First-Fit is used as baseline algorithm. In this method VMs are assigned to servers uniformly. Thereby, it avoids servers overloading, and network traffic is balanced but the disadvantage is that no server or network switch is left idle so none of them can be turned off to save power.

Wei, Lin, Chuang, Kong Xiangzhen, (2011) proposed an approach to optimizing energy consumption by determining the best candidate of migrating virtual machine (VM) and the best candidate of destination physical machine (PM) for new increased workload and migrating VM in an energy cost efficient way, two models are established to settle these two problems: energy aware migration model and load dispatching model. Paper presented a comprehensive method combining the

energy aware migration model and the load dispatching model. Based on the analysis of pre-copy process during migration and the knowledge that energy is linearly proportional consumed to data volume copied, They established the energy aware migration model with required performance constrains which determines the best migrating VM candidate. The load dispatching model indicates that each PM has an optimal point of resource utilization enables this PM operating in a most energy efficient way, the workload should be dispatched across PMs to obtain the overall optimal energy cost, and the load dispatching model is in charge of making decision of the best destination PM for increased workload as well as migrating VMs, and it also suggests the optimum time to migrate which is the time when resource demand of total VMs in a PM exceeds the optimal point of this PM.

But this paper did not consider the performance metrics, which might be affected by live migration and also different types of workload have different performance degradation during migration as discussed in (Ye et al., 2010).

Sinha, Purohit & Diwanji, (2011) have proposed a dynamic threshold based policy for CPU utilization for host at data center in cloud. This consolidation policy will work on dynamic and unpredicted workload avoiding useless power consumption. This paper gave the idea of energy efficiency requirement and would also ensure quality of service to the end user by reducing the Service Level Agreement violation. Paper considered different metrics for performance measurements like Energy cost, power cost, operational cost, virtual machine to be migrated and also SLA violations, which occurs during migrations. Best Fit Decreasing Algorithm is used for placements for virtual machines. But there are some limitations lies in Best Fit Decreasing algorithm, that is, it consider only utilization of VM, and list of VM is formalized according to the decreasing order of VM utilization, it doesn't consider others metrics like CPU, memory, bandwidth and SLA parameters etc.

So, there is a need to make such an energy efficient algorithm which can consider all those performance metrics which may or may not be affected by migrations.

Goudarzi, Ghasemazar & Pedram, (2012) represents a resource allocation problem in which the aim is to reduce the total energy cost of cloud computing system along with meeting the specified client-level SLAs in a sense of probability. In cloud computing environment, cloud provider pays penalty for the percentage of a client's requests or needs that do not meet a decided upper bound on their service time period. They generate an algorithm imply on convex optimization and dynamic

programming is generated to solve the aforesaid virtual resource allocation problem. It have seen that considering the SLA with efficient VM placement can help to reduce the operational cost in the cloud computing system.

Borgetto, Maurer, Da-Costa, Pierson, & Brandic, (2012) present an integrated approach for VM migration and reconfiguration, and PM power management. They incorporate an autonomic management loop, where proactive actions are advised for all three areas in a hierarchically structured way. The efficacy approach is evaluated by considering classical algorithms like Vector Packing, First Fit and Monte Carlo adapted for energy-efficient reallocation. The algorithm results shows energy savings up to 61.6% while having some SLA violation rates.

Beloglazov et al., (2012) define an architecture and major principles for energy-efficient Cloud computing. Author presents, open research challenges, allocation algorithms and resource provisioning for energy-efficient resource management of Cloud computing environments. This proposed work on energy-aware allocation heuristics based provisioning data center resources, improves energy efficiency of the data center. This paper conducts a survey on energy efficiency in computing and proposes; a architecture for energy-efficient management of Clouds, energy-efficient resource allocation strategies and scheduling algorithms by considering QoS expectations and power consumption characteristics of the computing devices and open research challenges, which can bring significant benefits to both the resource providers and consumers. Author proposed algorithm for Energy-aware allocation of data center resources:

Firstly for virtual machine placement that is MBFD (Modified best fit decreasing) algorithm, sort all VMs in decreasing arrangement of their current CPU utilizations, then it allocate each VM to host that delivers the least growth of power consumption owed to this allocation. Secondly, for virtual machine selection, to determine when and which VMs should be migrated, introduce three double-threshold VM selection policies. i) The minimization of migrations policy ii) The highest potential growth policy iii) The random choice policy. From the simulator results, provided by author, this is concluded that MM (minimization of Migration) is best from other policies, because it is least violate SLA(Service Level Agreement).In this paper SLA is calculated, by considering MIPS (million instructions per second) of services or applications or tasks. Now this paper considers parameters which are necessary for

better performance of virtual machine placements. But it did not include SLA parameters in placement algorithm.

Strunk & Dargie, (2013) prove that migration of virtual machine have consumed a significant amount of energy. In this work, they consider energy cost of VMs when love migration occurred. They investigate the energy cost live migration and impact of VMs size and network bandwidth on it. Power consumption of source and destination is increased when there is an increment in utilized bandwidth. The time of VMs migration is varied according to the VMs size and network bandwidth.

Wang, Huang, Charles, Wen & Li-Chun Wang, (2014) proposed a mechanism for virtual machine placement by considering energy efficiency and Quality of Service for reducing the problem of traffic imbalance in on and off state of virtual machines for saving energy. They Consider three techniques; hop reduction, energy saving and load balancing. Hop reduction technique group the VMs to reduce the traffic load; by partitioning the virtual machine in groups in such a manner that each group of VMs is balanced and have minimum cost between the groups. Load balancing mechanism updates the information of virtual machine periodically. Still these mechanisms are not energy efficient from the perspective of virtual machine placement, because random placements can lead to lack of resources for some servers. Due to this SLA violations can occur.

Shuja, Bilal, Madani, Othman, Ranjan, Balaji & Khan, (2014) discusses the issues of quality of services for the services given by the cloud provider at same time decreasing the energy consumption of datacenters. They surveyed that some energy efficient computing device like ARM and PCM processors and drives respectively have no comparable performance with the underlined hardware. Paper suggests that there is a need of energy optimization technique which can consider workload profile of datacenter and end user SLAs for balancing the performance requirement and computing energy efficiency.

Knowledge Gap

Cloud computing is facing a major problem of energy consumption by datacenters. In earlier year, authors basically concentrated on the hardware devices and DVFS technique. DVFS basically used for minimizing the power consumption of processors. (Bohrer et al, 2002) used DFVS for measuring the energy of web servers, similarly (Elnozahy et al, 2003) used it for measuring energy usage and response time. (Poellabauer, 2005), (Isci, 2006) also use DVFS for implementing

new energy efficient techniques. There is a major point which have to be consider is service level agreement, which is considered by (Beloglazov, 2010). They consider CPU utilization, virual machine migration, and have less SLA violations. But still there is a need to consider end user SLAs for better service delivery.

Above discussion shows different energy efficient techniques in cloud computing. Further he work will extended to develop a new approach for minimizing the virtual machine migration and saving energy in data center under the IaaS environment.

CHAPTER 4

PROPOSED WORK

In this work we proposed a new approach to select VMs to migrate from overloaded host. Our approach will work on the basis of VM categories that are negotiated in service level agreement. In this chapter we firstly VM categories have been discussed and after that we discuss about the proposed scheme.

4.1 Categorization of Virtual Machines

As the cloud computing is utility based computing, we proposed to categorize VM based on the VM usage cost. We divide the VMs into five categories: category 1 to 5 with following assumptions:

- The cost of category-1 VMs is lowest while the cost of category-5 VMs is highest i.e. higher the category, higher the cost.
- In case of SLA violation, cloud service provider have to higher penalty amount for the category 5 VMs as compared to other Categories i.e. higher the category, higher the penalty.

Cloud users will select the VM category they want to use and mention in the Service Level Agreement.

4.1 Proposed Approach: Sla Aware Energy Efficient Approach for Virtual Machine Placement in Cloud Data Center

Virtual machine manager gives response to all requests and allocate VMs according to requests. During allocation of VMs, a number of migrations occur to achieve maximum response time and to avoid service level agreement. It also consumes lots of energy in the form of power consumption. Due to the dynamic nature of cloud computing, sometimes servers will overloaded and that may be cause performance degradation due to the lack of resources at the guest operating systems. To handle the situation few of the VMs need to shifted from such overloaded hosts. In this section we purpose a new approach that selects the VM based upon the CPU utilization and category of VM (selected by cloud user during service level agreement). The proposed approach prefers to migrate the VMs that belongs to the higher category so that that the resources may be allocated to these VMs to avoid costly penalties for SLA violations. Our approach optimizes the energy consumption with lesser amount of penalties paid for SLA violations.

Proposed Algorithm

The algorithm for the proposed approach is given below:

1. *Prepare the list OH of Hosts having CPU utilization > Upper-Threshold*
2. *For each Host H in list OH do*
   2.1 Until CPU utilization of H > Upper-Threshold
      2.1.1. Find the list UCL of VMs that belong to the uppermost category among VMs running on H.
      2.1.2. Find the VM V from list UCL such that Absolute ((CPU utilization of H – Upper-Threshold) – CPU utilization by V) is minimum.
      2.1.3. Add V to list of machines to be migrated.
      2.1.4 Set CPU-Utilization of H = CPU-Utilization of H - CPU utilization by V
   [End of step 2.1].

[End of step 2].

3. *Migrate the VMs selected in step 2.*

[End of the Algorithm]

Firstly, algorithm finds out the overloaded host. Then for each overloaded host, the algorithm select VMs to migrate by considering the category of VMs. The algorithm stops when new utilization of host is lower than the upper utilization threshold. Flow chart given in fig-13 depicts the working of proposed algorithm.

.

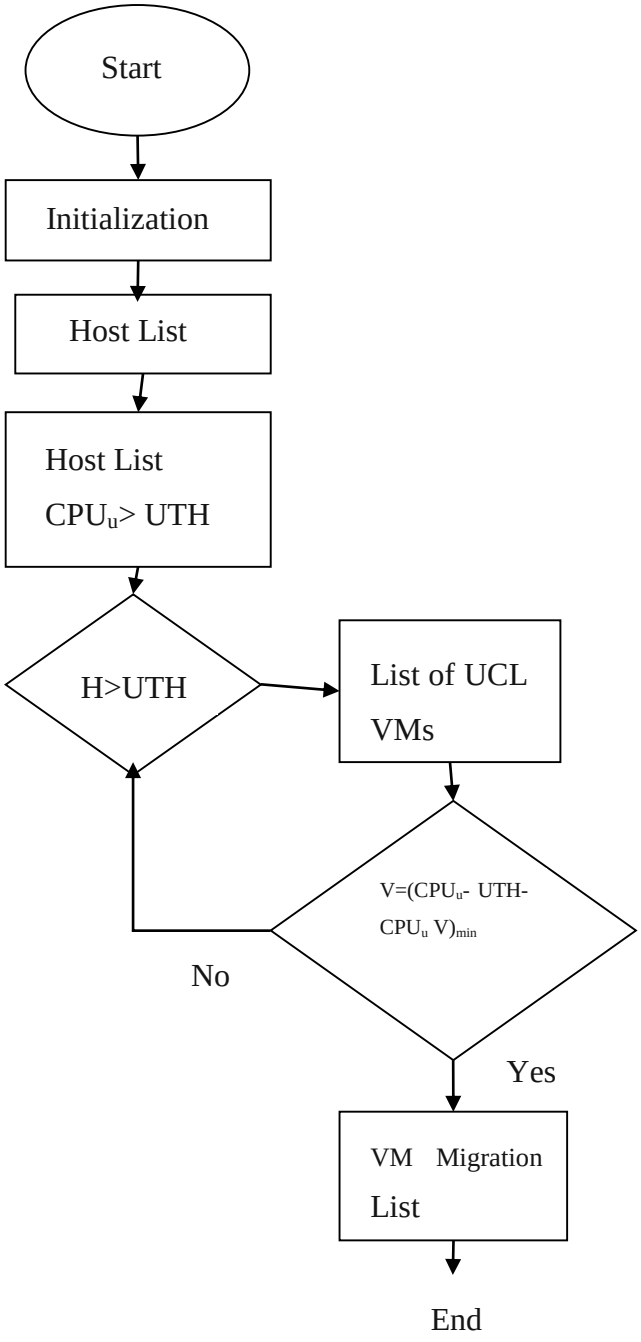

Figure 4.1 Flow chart of proposed work

CHAPTER 5

SIMULATION AND RESULTS

The performance of the proposed approach has been evaluated by simulating it in CloudSim--3.0.3 (Calheiros, Ranjan, Beloglazov, De Rose & Buyya, 2011). We have also simulated other scheme (Random Selection Policy, Minimum Migration Time Policy and Minimization of Migration Policy) used for the same purpose and compare the performance with the proposed approach. This chapter presents the simulation parameters. Further, the results have been also discussed in this chapter.

5.1 Simulation Parameters

The proposed approach and existing approaches have been simulated in CloudSim with simulation parameters as mentioned in Table-5.1.

Table 5.1 Simulation Parameters

Parameters	Value/s
Physical Node	200
Host Type	2
MIPS	1860, 2660
Host Storage	1GB
Host RAM	4 GB
Host CPU Core	1
Number of VMs	300
VM Type(based on configuration)	4
MIPS(VM)	2500, 2000, 1000, 500
VM Size	2.5GB
RAM(VM)	870, 1740, 613MB
CPU Core (VM)	1
Simulation Length	10*60*60 seconds (10 hours)

5.2 Power Model

As in our simulation setup there are two types of host machines, to measure the energy consumption two models have been used. The details of these models have been given in table-5.2.

Table 5.2: Power Models used to measure energy consumption

Host Type	Power Model
MIPS- 1860 RAM- 4GB	PowerModelSpecPowerHpProLiantMl110G4Xeon3040
MIPS- 2660 RAM- 4GB	PowerModelSpecPowerHpProLiantMl110G5Xeon3075

5.3 Results

In order to evaluate the performance of Minimization of Migration Policy, Minimum Migration Time, Random Selection Policy and proposed work, several performance metrics have been observed. First metric is the total energy consumed by the physical servers of the data center. Total energy consumed by data center while operated with various approaches has been mentioned in table-5.3

Table 5.3: Energy Consumption by MM, MMT, RS and Proposed Approach

CPU Utilization	MM	RS	MMT	Proposed Approach
1	40.847453	184.66	174.29	40.847453
0.9	77.05	215.48	212.83	77.906899
0.8	83.552243	222.43	221.29	87.109620
0.7	91.928023	239.82	230.63	93.237517
0.6	97.031112	271.01	261.92	100.666787
0.5	111.724611	302.37	293.29	114.919009
0.4	123.346810	330.82	325.63	126.394323

As represented by graph given in figure-5.1, the proposed scheme consumes least energy than the other approaches for all values of upper utilization except for 100% utilization. The graph also depicts that by using 100% upper utilization threshold for various hosts the energy requirements are minimum for all approaches.

The upper utilization threshold affects the other parameters, SLA violations is the one of such parameters. To find the optimal approach we also observe number of

SLA violation occurred and the result for the same have been mentioned in table 5.4.

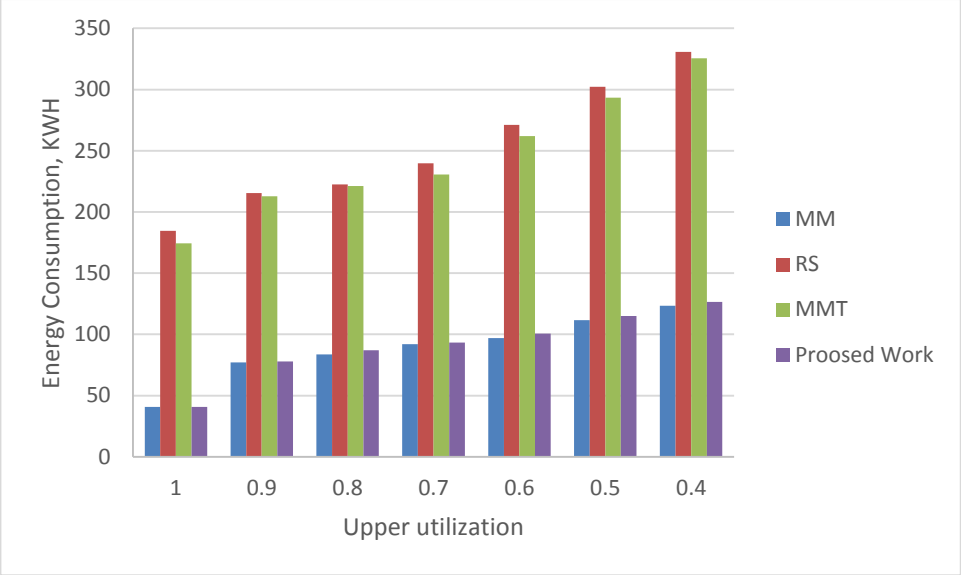

Figure 5.1: Energy Consumption by MM, MMT, RS and Proposed Approach

Table 5.4: SLA Violations occurred due to MM, MMT, RS and Proposed Approach

CPU Utilization	MM	RS	MMT	Proposed Approach
1	32.67	18.90	19.16	32.67
0.9	15.56	15.02	15.27	15.45
0.8	12.76	13.63	13.63	11.49
0.7	9.54	12.36	12.63	10.34
0.6	11.94	11.44	11.64	9.69
0.5	11.60	10.65	10.49	12.41
0.4	12.37	10.58	10.35	11.89

As shown in graph (figure-5.2), in most of the cases, number of SLA violations occurred with Minimization of Migration (MM) technique is minimum. The number SLA violations occurred with the proposed approach are marginally more than MM technique but less than other approaches.

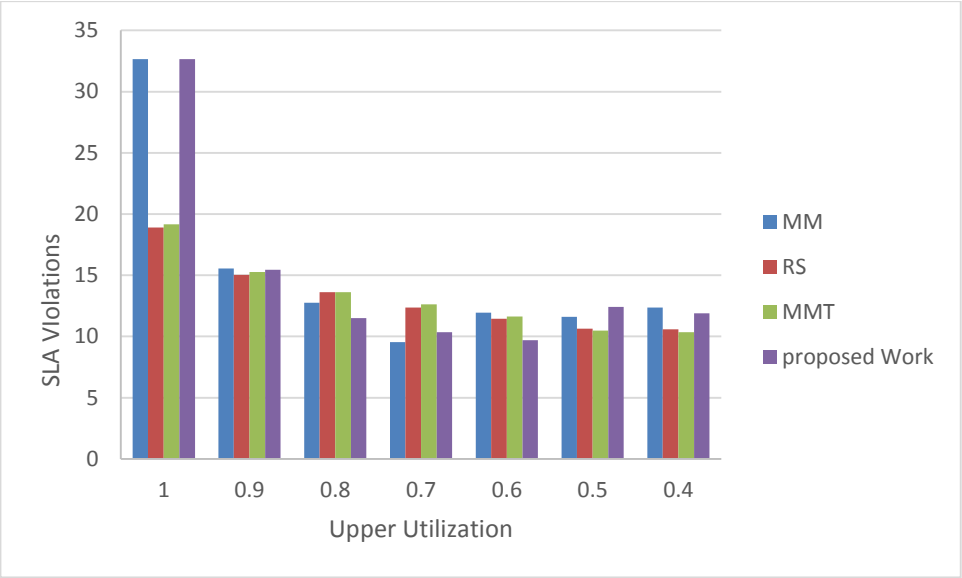

Figure 5.2: SLA Violations occurred due to MM, MMT, RS and Proposed Approach

To reduce the SLA violations these techniques migrate VMs from over-utilized hosts. The number of VM migrations with various schemes have been mentioned in table-5.5.

Table 5.5: No. of VM Migrations in case of MM, MMT, RS and Proposed Approach

CPU Utilization	VM Migrations(RS)	VM Migrations (MMT)	VM Migrations (MM)	Proposed Work
1	35544	28869	458	458
0.9	25794	23181	6874	7071
0.8	31228	27260	7506	8406
0.7	35181	30337	8604	9849
0.6	35143	31834	9382	10806
0.5	31890	25940	8616	9693
0.4	27659	23455	6765	7563

The graph in figure-5.3 represents that the number of migrations occurred are minimum in case of Minimization of Migration approach. However, the proposed approach migrate far less number of VMs as compared to rest of the approaches.

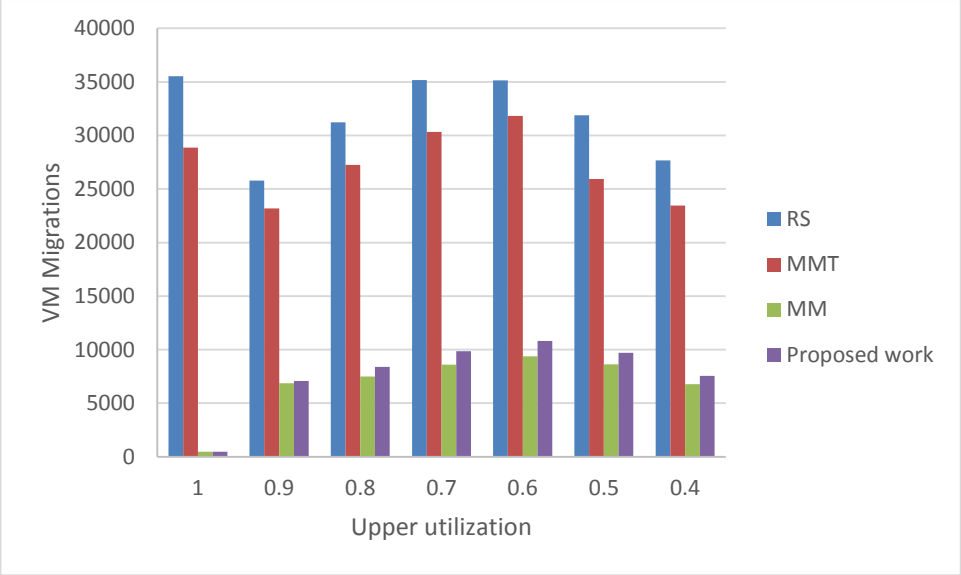

Figure 5.3: VM Migrations in case of MM, MMT, RS and Proposed Approach

The proposed approach demands slightly more energy as well as lead to slightly more SLA violation than Minimization of Migration approach. However, the cost paid in lieu of SLA violations can be reduced by using the proposed approach as it reduces the SLA violations belonging to VM categories for which penalties are higher. The category wise SLA violations occurred with our approach have been mentioned in table-5.6.

Table 5.6: Category-wise Average SLA violation occurred with the proposed approach.

CPU Utilization	Category 0	Category 1	Category 2	Category 3	Category 4
1	32.27	30.06	29.77	25.31	23.93
0.9	20.04	19.27	18.34	13.88	11.98
0.8	16.68	15.79	16.55	9.85	7.44
0.7	13.01	13.23	11.39	8.65	7.30
0.6	12.76	12.54	12.86	6.64	6.50
0.5	18.93	18.26	15.17	10.24	8.98
0.4	17.64	19.78	16.93	5.93	7.03

5.4 Analysis of Results

During the simulation it has been observed that Minimization of Migration technique (Beloglazov et al., 2012) consumes minimum power and also lead to minimum no. of SLA violations for most of the upper utilization values. However, the proposed approach demand slightly more power (as 8.3% additional power in case of 70% upper utilization) and lead to slightly more SLA violations than this technique. Although performance of proposed approach is comparatively degraded than Minimization of Migration approach, yet the cost paid in lieu of SLA violations can be reduced. The proposed approach reduces such cost by reducing the SLA violations belonging to VM categories for which SLA violation penalties are higher.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

The research work is developed to support the energy-efficient management and allocation of Cloud data center resources. We have categorized the VMs based upon their usage cost and cost paid in lieu SLA violations. In this work, a new approach has been proposed to select VMs to migrate from the over utilized host. The proposed approach prefers to migrate the VMs that belongs to the higher category so that that the resources may be allocated to these VMs to avoid costly penalties for SLA violations. Our approach set a trade-off between the cost paid for consumed energy and the cost paid in lieu of SLA violations.

The proposed approach has been simulated in CloudSim. During the simulation, energy consumption, SLA violations and no. of VM migrations have been observed and recorded. For comparative evaluation, we have also simulated Minimization of Migration approach. The results by simulating MMT and RS approaches have also been recorded.

During the simulation it has been observed that Minimization of Migration technique consumes minimum power and also lead to minimum no. of SLA violations for most of the upper utilization values. However, the proposed approach demand slightly more power (as 8.3% additional power in case of 70% upper utilization) and lead to slightly more SLA violations than this technique. Although performance of proposed approach is comparatively degraded than Minimization of Migration approach, yet the cost paid in lieu of SLA violations can be reduced. The proposed approach reduces such cost by reducing the SLA violations belonging to VM categories for which SLA violation penalties are higher.

In this work we have only considered SLA violations in terms of Million Instruction Per Second (MIPS), in future new approaches can also be developed by considering SLA violations on other parameters such as RAM and UP-Time etc.

REFERENCES

Antonopoulos, N., & Gillam, L. (2010). Cloud Computing Principles,Systems and Applications. Springer, US.

Beloglazov, A. (2013). *Energy-Efficient Management of Virtual Machines in Data Centers for Cloud Computing*. PhD Thesis,The University of Melbourne, Australia.

Beloglazov, A., & Buyya, R. (2010, May). Energy efficient resource management in virtualized cloud data centers. Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing. Pp. 826-831.

Beloglazov, A., Abawajy, J., & Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Generation Computer Systems, 28(5), 755-768.

Bohrer, P., Elnozahy, E. N., Keller, T., Kistler, M., Lefurgy, C., McDowell, C., & Rajamony, R. (2002). The case for power management in web servers. In Power aware computing (pp. 261-289). Springer US.

Borgetto, D., Maurer, M., Da-Costa, G., Pierson, J. M., & Brandic, I. (2012, May). Energy-efficient and sla-aware management of iaas clouds. Proceedings of the 3rd International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet. Pp. 25.

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems*, *25*(6), 599-616.

Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A., & Buyya, R. (2011). CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and Experience*, *41*(1), 23-50.

Cardosa, M., Korupolu, M. R., & Singh, A. (2009, June). Shares and utilities based power consolidation in virtualized server environments. IFIP/IEEE International Symposium on Integrated Network Management, pp. 327-334.

Chen, G., He, W., Liu, J., Nath, S., Rigas, L., Xiao, L., & Zhao, F. (2008, April). Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services. In *NSDI* (Vol. 8, pp. 337-350).

Chen, Y., Das, A., Qin, W., Sivasubramaniam, A., Wang, Q., & Gautam, N. (2005, June). Managing server energy and operational costs in hosting centers. In *ACM SIGMETRICS Performance Evaluation Review*,Vol. 33, No. 1, pp. 303-314.

Chetan, S., Geevarghese, J. J., & Karthik, R. (2010). Placement Algorithm for Virtual Machines.

Chiaraviglio, L., & Matta, I. (2010, April). Greencoop: cooperative green routing with energy-efficient servers. Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking, pp. 191-194.

Diniz, B., Guedes, D., Meira Jr, W., & Bianchini, R. (2007). Limiting the power consumption of main memory. ACM SIGARCH Computer Architecture News, 35(2), pp. 290-301.

Elnozahy, E. M., Kistler, M., & Rajamony, R. (2003). Energy-efficient server clusters. In Power-Aware Computer Systems, pp. 179-197. Springer Berlin Heidelberg.

Emeakaroha, V. C., Netto, M. A., Calheiros, R. N., Brandic, I., Buyya, R., & De Rose, C. A. (2012). Towards autonomic detection of sla violations in cloud infrastructures. Future Generation Computer Systems, *28*(7), pp. 1017-1029.

Esfandiarpoor, S., Pahlavan, A., & Goudarzi, M. (2013). Virtual Machine Consolidation for Datacenter Energy Improvement. Green Computing and Communications (GreenCom), 2010 IEEE/ACM International Conference on & International Conference on Cyber, Physical and Social Computing (CPSCom). Pp. 171 – 178. Hangzhou.

Federal Energy Management Progarm, (2013). [EERE]. Homepage,< http://www1.eere.energy.gov/femp/program/dc_energy_consumption .html> Accessed 2014, March 15.

Femal, M. & Freeh, V. (2005). Safe over-provisioning: Using power limits to increase aggregate throughput. In Power-Aware Computing Systems (PACS), pp 150-164. Springer Berlin Heidelberg December.

Femal, M. E., & Freeh, V. W. (2005, June). Boosting data center performance through non-uniform power allocation. Proceedings, Second International Conference on Autonomic Computing. Pp. 250-261, ICAC 2005.

Furht, B., & Escalante, A. (2010). Handbook of Cloud Computing, pp.656. Springer, Boca Ratan.

Gniady, C., Hu, Y. C., & Lu, Y. H. (2004, February). Program counter based techniques for dynamic power management. IEEE Proceedings on Transactions on Computers. Volume. 55, Pp. 24-35.

Goudarzi, H., Ghasemazar, M., & Pedram, M. (2012, May). Sla-based optimization of power and migration cost in cloud computing. 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp. 172-179.

Guo, C., Lu, G., Wang, H. J., Yang, S., Kong, C., Sun, P., & Zhang, Y. (2010, November). Secondnet: a data center network virtualization architecture with bandwidth guarantees. Proceedings of the 6th International Conference. Pp. 15.

Gupta, P. K. D., Nayak, M., & Pattnaik, S. (2012). Cloud Computing: Based Projects Using Distributed Architecture, pp.340.PHI Learning Pvt. Ltd, Delhi.

Gyarmati, L., & Trinh, T. A. (2010, April). How can architecture help to reduce energy consumption in data center networking? Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking, pp. 183-186.

Harris, Torry. (2012). Cloud Computing-an Overview. *SDA*.

Heath, T., Diniz, B., Carrera, E. V., Meira Jr, W., & Bianchini, R. (2005, June). Energy conservation in heterogeneous server clusters. Proceedings of the tenth ACM SIGPLAN symposium on Principles and practice of parallel programming, pp. 186-195.

Hill, R., Hirsch, L., Lake, P., & Moshiri, S. (2013). Enterprise cloud computing. In *Guide to Cloud Computing*, pp. 209-222. Springer, London.

International Resources Group. (2009). Energy Efficiency in Building. Lowerence Berkeley National Laboratory, U.S.

Isci, C., Contreras, G., & Martonosi, M. (2006, December). Live, runtime phase monitoring and prediction on real systems with application to dynamic power management. Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture, pp. 359-370.

Jamsa, K. (2013). Cloud Computing SaaS,PaaS,IaaS,Virtualization,Business Models,Mobile,Security and More. Jones & Barlett Leraning, Burlington.

Jansen, W., & Grance, T. (2011). Guidelines on security and privacy in public cloud computing. NIST special publication, 800, 144.

Jayasinghe, D., Pu, C., Eilam, T., Steinder, M., Whally, I., & Snible, E. (2011, July). Improving performance and availability of services hosted on iaas clouds with structural constraint-aware virtual machine placement. IEEE International Conference on Services Computing (SCC), pp. 72-79.

Karadsheh, L. (2012). Applying security policies and service level agreement to IaaS service model to enhance security and transition. Computers & Security, 31(3), 315-326.

Kotla, R., Ghiasi, S., Keller, T., & Rawson, F. (2005, April). Scheduling processor voltage and frequency in server and cluster systems. Proceedings 19th IEEE International on Parallel and Distributed Processing Symposium, pp. 8-pp.

Krutz, R. L., & Vines, R. D. (2010). *Cloud security: A comprehensive guide to secure cloud computing*. John Wiley & Sons, Canada.

Kusic Parashar, M., & Lee, C. A. (2005). Grid Computing: introduction and overview. Proceedings of the IEEE, special issue on grid computing, 93(3), pp.479-484.

Kusic, D., Kephart, J. O., Hanson, J. E., Kandasamy, N., & Jiang, G. (2009). Power and performance management of virtualized computing environments via lookahead control. *Cluster computing*, *12*(1), pp.1-15.

Lin, C. C., Liu, P., & Wu, J. J. (2011, July). Energy-aware virtual machine dynamic provision and scheduling for cloud computing. IEEE International Conference on Cloud Computing (CLOUD), pp. 736-737.

Magoulès, F., Pan, J., & Teng, F. (2012). *Cloud computing: Data-intensive computing and scheduling,* pp.231. CRC Press.

Mahadevan, P., Sharma, P., Banerjee, S., & Ranganathan, P. (2009, April). Energy aware network operations. INFOCOM Workshops 2009, IEEE, pp. 1-6.

Man, C. L. T., & Kayashima, M. (2011, September). Virtual machine placement algorithm for virtualized desktop infrastructure. IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), pp. 333-337.

Marinos, A., & Briscoe, G. (2009). Community cloud computing. In *Cloud Computing* (pp. 472-484). Springer Berlin Heidelberg.

Maurer, M., Emeakaroha, V. C., Brandic, I., & Altmann, J. (2012). Cost–benefit analysis of an SLA mapping approach for defining standardized Cloud computing goods. Future Generation Computer Systems, 28(1), pp. 39-47.

Mell, P., & Grance, T. (2009). The NIST definition of cloud computing. National Institute of Standards and Technology, 53(6), 50.

Pinheiro, E., Bianchini, R., Carrera, E. V., & Heath, T. (2001, September). Load balancing and unbalancing for power and performance in cluster-based systems. Workshop on compilers and operating systems for low power,Vol. 180, pp. 182-195.

Poellabauer, C., Singleton, L., & Schwan, K. (2005, March). Feedback-based dynamic voltage and frequency scaling for memory-bound real-time applications. Real Time and Embedded Technology and Applications Symposium, RTAS 2005, 11th IEEE, pp. 234-243.

Raghavendra, R., Ranganathan, P., Talwar, V., Wang, Z., & Zhu, X. (2008, March). No power struggles: Coordinated multi-level power management for the data center. In ACM SIGARCH Computer Architecture News, Vol. 36, No. 1, pp. 48-59.

Rodero-Merino, L., Vaquero, L. M., Gil, V., Galán, F., Fontán, J., Montero, R. S., & Llorente, I. M. (2010). From infrastructure delivery to service management in clouds. Future Generation Computer Systems, 26(8), pp. 1226-1240.

Sadashiv, Naidila, & Kumar, SM Dilip. (2011). *Cluster, grid and cloud computing: A detailed comparison.* 2011 6th International Conference on Computer Science & Education (ICCSE).

Sara Angeles (2013), Business News Daily on Cloud vs. Data Center: What's the difference? Homepage, <http://www.businessnewsdaily.com/4982-cloud-vs-data-center.html> Accessed 2014 April 26.

Schulz, F. (2010, June). Towards measuring the degree of fulfilment of service level agreements. 2010 Third International Conference on Information and Computing (ICIC), Vol. 3, pp. 273-276.

Schulz, G. (2012). Cloud and Virtual Data Storage Networking. CRC Press, Boca Raton.

Shuja, J., Bilal, K., Madani, S. A., Othman, M., Ranjan, R., Balaji, P., & Khan, S. U. Survey of Techniques and Architectures for Designing Energy-Efficient Data Centers. IEEE Systems journal.

Sinha, R., Purohit, N., & Diwanji, H. (2011). Power Aware Live Migration for Data Centers in Cloud using Dynamic Threshold. International Journal of Computer Technology and Applications, 2(6).

Srikantaiah, S., Kansal, A., & Zhao, F. (2008, December). Energy aware consolidation for cloud computing. Proceedings of the 2008 conference on Power aware computing and systems, Vol. 10.

Strunk, A., & Dargie, W. (2013, March). Does live migration of virtual machines cost energy? 2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA), pp. 514-521.

Stryer, P. (2010). Understanding Data Centers and Cloud Computing. Global Knowledge Training, CCSI, CCNA.

Wang, S. H., Huang, P. P. W., Wen, C. H. P., & Wang, L. C. (2014, February). EQVMP: Energy-efficient and QoS-aware virtual machine placement for software defined datacenter networks. In *Information Networking (ICOIN), 2014 International Conference on* (pp. 220-225).

Wei, B., Lin, C., & Kong, X. (2011, December). Energy optimized modeling for live migration in virtual data center. 2011 International Conference on Computer Science and Network Technology (ICCSNT), Vol. 4, pp. 2311-2315.

WSP. (2010). Cloud Computing and Sustainability: The Environmental Benefits of Moving to the Cloud. Accenture.

Wu, Y., & Zhao, M. (2011, July). Performance modeling of virtual machine live migration. 2011 IEEE International Conference on Cloud Computing (CLOUD), pp. 492-499.

Ye, K., Huang, D., Jiang, X., Chen, H., & Wu, S. (2010, December). Virtual machine based energy-efficient data center architecture for cloud computing: a performance perspective. Proceedings of the 2010 IEEE/ACM International Conference on Green Computing and Communications & International Conference on Cyber, Physical and Social Computing, pp. 171-178.

YS Voas, J., & Zhang, J. (2009). Cloud computing: New wine or just a new bottle? *IT professional*, *11*(2), pp.15-17.

YS, R. (2013). A Novel way of Improving CPU Utilization. In Cloud M.Tech Thesis, National Institute of Technology Rourkela, Odisha.