Energy Efficiency Models and Optimization Algorithm to Enhance On-Demand Resource Delivery in a Cloud Computing Environment

BY

THUSOYAONE JOSEPH MOEMI
(STUDENT NUMBER: 17071100)

DISSERTATION SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE (MSc) IN COMPUTER SCIENCE

DEPARTMENT OF COMPUTER SCIENCE
SCHOOL OF MATHEMATICAL & PHYSICAL SCIENCES
FACULTY OF AGRICULTURE, SCIENCE AND TECHNOLOGY
NORTH WEST UNIVERSITY – MAFIKENG CAMPUS

SUPERVISOR: PROF. O. O. EKABUA

OCTOBER, 2013

DECLARATION

I declare that this Research Dissertation on **Energy Efficiency Models and Optimization Algorithm to Enhance On-Demand Resource Delivery in a Cloud Computing Environment** is my work, and has never been presented for the award of any degree in any university. All the information used has been duly acknowledged both in the text.

Signature _____ Date _____

Thusoyaone Joseph Moemi

Approval

Signature _____ Date _____

Supervisor: **Prof. O. O. Ekabua**
Department of Computer Science
Faculty of Agriculture, Science and Technology
North West University, Mafikeng Campus
South Africa

DEDICATION

This research dissertation is dedicated to Jehovah and my parents:

Monnapule Edwin Moemi.

and

Onalenna Memoria Moemi

ACKNOWLEDGEMENTS

I would like to give my deepest expression of gratitude to the king of eternity, the Almighty God, Jehovah, for allowing me to go through this research process fully resourced and also for the people He placed around me. Jehovah, I thank You for the successful completion of this research work, and for all the wisdom and knowledge you gave to me. Without your help none of this would have been possible.

I am also grateful and very thankful to my supervisor Prof. O.O. Ekabua, for his constructive criticism, helpful support, advice that edifies and all the new things he taught me about research. I am also grateful for the patience he exercised while supervising me and the belief he had in me to finish this degree.

I wish to express my sincere thanks to the lecturers and staff of the Department of Computer Science, North West University, Mafikeng Campus, for their help and support.

Finally, I would also like to express my gratitude to all my family and friends for their encouragent and support – Tumisang Moemi, Onalena Moemi, Edwin Moemi, Maipelo Molemoeng, Tlhalefo Kobue, Michel Mbougni, Francis Lugayizi, Hope Tsholofelo Mogale, Ifeoma Ohaeri, Nosipho Dladlu, Nnenna Eric and Thuto Assegaai. Thank you all.

Abstract

Online hosed services are what is referred to as Cloud Computing. Access to these services is via the internet. It shifts the traditional IT resource ownership model to renting. Thus, high cost of infrastructure cannot limit the less privileged from experiencing the benefits that this new paradigm brings. Therefore, cloud computing provides flexible services to cloud users in the form of software, platform and infrastructure as services. The goal behind cloud computing is to provide computing resources on-demand to cloud users efficiently, through making data centers as friendly to the environment as possible, by reducing data center energy consumption and carbon emissions. With the massive growth of high performance computational services and applications, huge investment is required to build large scale data centers with thousands of centers and computing models. Large scale data centers consume enormous amounts of electrical energy. The computational intensity involved in data centers is likely to dramatically increase the difference between the amount of energy required for peak periods and off-peak periods in a cloud computing data center. In addition to the overwhelming operational cost, the overheating caused by high power consumption will affect the reliability of machines and hence reduce their lifetime. Therefore, in order to make the best use of precious electricity resources, it is important to know how much energy will be required under a certain circumstance in a data center. Consequently, this dissertation addresses the challenge by developing and energy-efficient model and a defragmentation algorithm. We further develop an efficient energy usage metric to calculate the power consumption along with a Load Balancing Virtual Machine Aware Model for improving delivery of no-demand resource in a cloud-computing environment. The load balancing model supports the reduction of energy consumption and helps to improve quality of service. An experimental design was carried out using cloud analyst as a simulation tool. The results obtained show that the LBVMA model and throttled load balancing algorithm consumed less energy. Also, the quality of service in terms of response time is much better for data centers that have more physical machines, but memory configurations at higher frequencies consume more energy. Additionally, while using the LBVMA model in conjunction with the throttled load balancing algorithm, less energy is consumed, meaning less carbon is produced by the data center.

Table Contents

List of Figures

List of Tables

List of Acronyms

AaaS	Applications-as-a-Service
DC	Data Center
DCeP	Data Center Energy Productivity
DCP	Data Center Productivity
IaaS	Infrastructure-as-a-Service
IT	Information Technology
JVM	Java Virtual Machine
PaaS	Platform-as-a-Service
PUE	Power Usage Effectiveness
QoS	Quality of Service
SaaS	Software-as-a-Service
SLA	Service Level Agreement
SPEC	Standard Performance Evaluation Corporation
VM	Virtual Machine

Chapter 1
General Introduction

1.1 Background Information

Cloud computing is a model of computing where massively scalable and elastic IT-related capabilities are provided "as a service" to external customers using the Internet [1]. Cloud computing is also what is being used by organizations that are trying to be more competitive worldwide in recent times, where the organizations are renting for services like hardware, software and data over the Internet; rather than buying the hardware, software or data (intellectual property) for a company and keeping that hardware, software and data inside the boundaries of that specific company asserts [2]. On a broad scale that is what cloud computing is. The new features in this cloud computing paradigm are its acquisition model which is based on purchasing of services; its business model is based on pay for use, its access model is over the Internet to any device and its technical model is scalable, elastic, dynamic, multi-tenant, & sharable [3].

Over the years, there has been a lot of bad service delivery in terms of delivering applications over networks in the Internet and the networks themselves have become more capable and better in delivering applications more efficiently. For example, broadband has become more widespread and some of the popular applications used today by most Internet users in broadband are Facebook, Gmail and Twitter service levels are acceptable in terms of money and time costs deo to the fact that computer facilities are shared by consumers or customers of the services. One of the other advantages of cloud computing is that more than one company can share resources or computer facilities in a data center irrespective of where the data center is located [2]. Some of the services offered in cloud computing environment are Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS) [3]. Figure 1.1 shows the relationship between these services in the cloud environment.

How to pay for services like hardware, software or platform has been a major challenge for companies over the years. Cloud computing has helped organizations that provide services over the Internet to improve their business models, and relationships with their customers such that

companies don't have to buy software, hardware or platform that is going to be out dated in time, but rather rent the services from a cloud provider for some time.

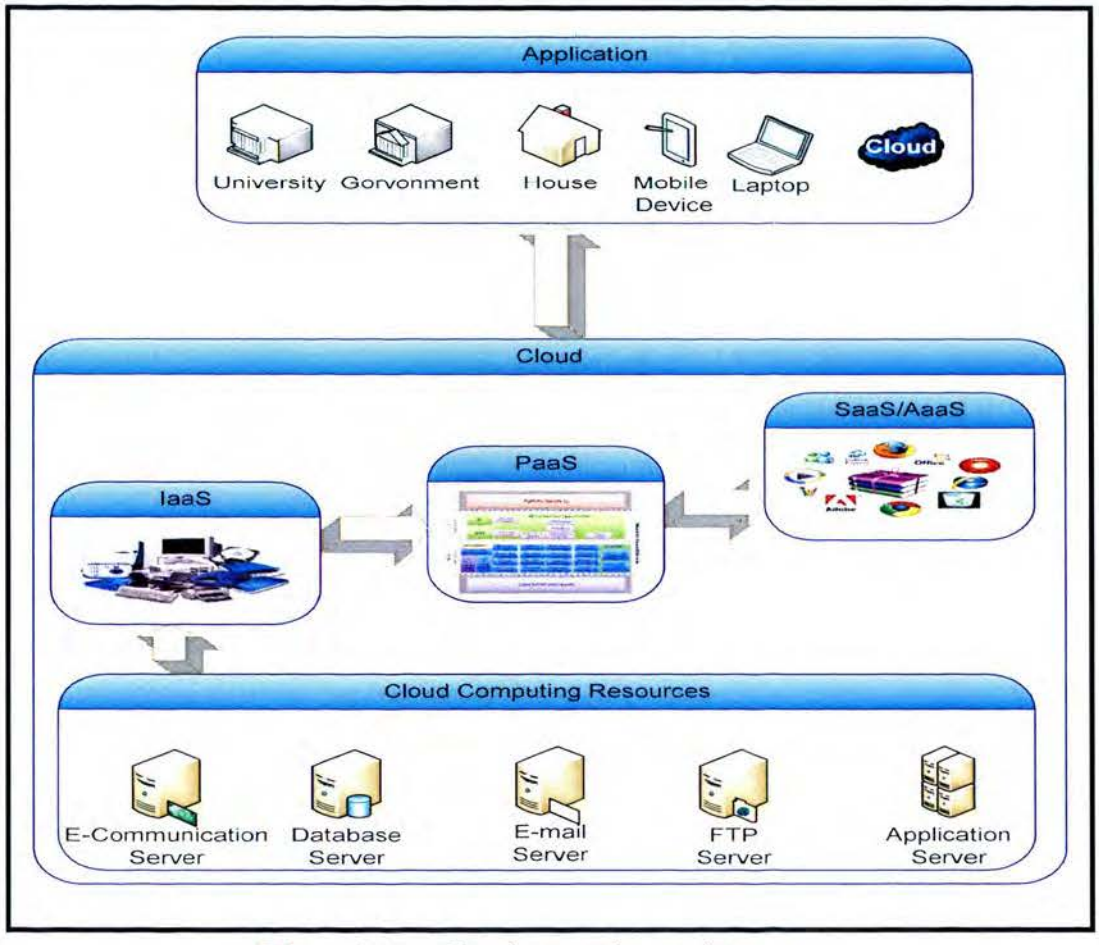

Figure 1.1 Cloud computing environment

Some benefits of cloud computing are; (i) reducing complexity, operation and maintenance of a company or organization by shifting some of the responsibilities outside of the company or organization to a cloud provider that is an expert in that specific field [2]. (ii) Businesses operating in a cloud environment are more agile in the sense that businesses can start new products and services with less risk and less expenditure. (iii) Businesses operating in a cloud environment can be more innovative than businesses operating in the Internet. For example businesses operating in a cloud environment can offer to rent more computing capacity to customers during peak periods.

As good as the idea of operating in a cloud is, the reality is that the services rendered, that is, either software, infrastructure or platform are hosted in buildings across the world, for example in a data center [4]. A data center is a facility used to house computer systems and associated components; physical building that contains multiple servers that store data. A cloud provider is a data center renting services to cloud consumers and consumers mostly according to how much computing power is used. Therefore, to stay competitive, cloud providers have to find more efficient ways of maximizing the use of computing power in their data centers. Hence, many data centers are trying to reduce their energy and power consumption through the implementation of visualization and cloud computing [5]. Hence energy efficiency and the reduction of air pollution are very big reason for providing cloud computing services [1,6]. At the United States of America, 2003, running a single 300-watt server for a year costs about $338, and can emit up to 1,300 kg of carbon dioxide [6].

Some challenges to overcome when implementing cloud computing data centers are power, space, capacity and bandwidth [4,7]. There is also a risk that closed privately owned and controlled cloud computing architectures could suppress innovations [4].

1.2 Problem Statement

Many ways of developing energy efficiency models and optimization algorithms for desktop systems and large scale data centers already exist and have been implemented in cloud computing environments, but there are still issues that slow down their adoption by the rest of the world due to the ever evolving computing industry. For example, the demand for services in real-time by cloud users is increasing and this leads to the need for more power consumption by cloud providers, which increases their carbon emissions [1,6,7]. The more power consumption the more carbon emissions by a data center. Hence, how to provide more energy-efficiency models algorithms for cloud platforms is still a critical problem [7] and there is also a need for Green Cloud computing solutions that can not only save energy for the environment, but also reduce operational costs [1]. The questions that arise from these concerns are:

a) How can we reduce carbon emissions in data centers?

b) How can we reduce energy consumption in data centers?

c) How can we optimize visualization in servers?

1.3 Rationale of the Study

Although there have been improvements in energy efficiency models through the use of virtual machines in cloud computing environments, and many different optimization algorithms have been implemented and better performing algorithms are needed, the problem that still remains is that large scale data centers still produce carbon dioxide as part of their service delivery. There already exists a lot of research on energy efficiency in cloud computing environments but most of the research rarely focuses on the modelling part and therefore does not fully address the reduction of carbon emissions, energy consumption, and load balancing in large-scale data centers.

In all of the research that has been made and the different approaches proposed by the researchers to tackle the problem of carbon emissions in data centers, energy efficiency has played a key role in the provision of different energy efficiency models and optimization algorithms in cloud computing environments. Hence, Justice et al. identified energy efficiency metrics that can be used by IT managers to measure and maintain the implementation of cost savings and green initiatives in data centers [8]. Raj and Shriram investigated the contributing factors to the energy expenditure in a cloud environment [9]. Wang and Wang propose a new energy efficient multi-task scheduling model based on Google's massive data processing framework, etc [10].

When implementing energy efficiency models and optimization algorithms there is a great need for tools to measure the performance of the models and algorithms, and one of the tools to use for performance measurements is performance metrics. These models and algorithms must comply with industry standards such as Leadership in Energy and Environmental Design (LEED) [11].

1.4 Research Questions

This work addressed the following questions:

a) How is energy consumed in data centers?

b) How can we develop efficient models to address energy consumption in data centers?

c) How can we develop efficient algorithm for optimal control of energy consumption in data centers?

1.5 Research Goal

The main goal of this research is to provide an efficient energy model and optimization algorithm to enhance on-demand resource delivery in a cloud computing environment.

1.6 Research Objectives

To achieve the main goal of this research, the following objectives were employed:

a) Developing an energy efficiency model.

b) Developing an optimization algorithm.

c) Developing a metric to measure the performance of the model and algorithm.

d) Implementing the energy efficiency model and optimization algorithm developed.

1.7 Research Contributions

The main contribution of this research is in assisting data centers to reduce energy and power consumptions, and the carbon emissions produced by them. Another contribution is giving data center managers tools to monitor energy consumption in their data center and to choose the most optimal data center configuration. This is achieved by the optimization algorithm, efficient energy

usage metric, and power and load balancing models provided by this research as a validity of the concept and approach.

1.8 Research Methodology

The following methodologies were used in this research:

a) Literature Survey

An intensive survey was carried out focusing on existing approaches used by other researchers. The focus areas were energy efficiency models in cloud, grid and distributed computing environment, on demand service delivery, and optimization strategies.

b) Model Formulation

Based on the literature survey an energy efficiency model in a cloud computing environment for on-demand resource delivery was developed in chapter 3.

c) Algorithm Development

Again based on the literature survey and the Load Balancing Virtual Machine Aware (LBVMA) model developed, a defragmentation optimization algorithm in a cloud computing environment for on-demand resource delivery was developed.

d) Metric Development

A performance metric was developed to evaluate the performance of the LBVMA model developed, a defragmentation algorithm.

e) Model Implementation

As a proof of concept, the LBVMA model is implemented and simulated in a cloud environment.

f) Model Evaluation

The performance of the implemented model was evaluated based on energy efficiency of the model on on-demand resource delivery. Some of the parameters used to evaluate the model were response time, throughput, and processed data.

1.9 Included Publications

Part of the research reported in this dissertation has been accepted for publication and another also submitted and is under review by an accredited journal. These papers are:

(i) T. J. Moemi and O. O. Ekabua: Energy efficiency models implemented in a cloud computing environment. *The 4th International Conference on Cloud Computing, GRIDS, and Virtualization (Cloud Computing 2013)*, May 27-June 1, 2013 – Valancia, Spain.

(ii) T. J. Moemi and O. O. Ekabua, O.O. (2013) Energy Efficiency Models for Improving Delivery of On-Demand Resources in a Cloud Computing Environment. *Malaysian Journal of Computer Science.* (A paper submitted and currently under review).

1.10 Chapter Summary

This chapter gave a brief introductory insight of this research work, problem statement, and the aims as well as objectives of the research. It also outlined the research methodology employed in this dissertation and gave a summary of the dissertation in terms of chapters. The next chapter will discuss in detail the components, technologies and concepts of cloud computing and review the literature of other researchers.

Chapter 2
Literature Review

2.1 An Overview of Chapter 2

This chapter presents a review of literature on related works that have been done in cloud computing and energy efficiency of cloud service data centers. It also gives a review of the background information on energy efficiency and optimization in cloud computing environments. The key terminologies used in this research are explained.

2.2 Background Information

Cloud computing is a paradigm that has come to deliver on-demand computing resources as utility to cloud customers. The three main services of cloud computing are Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). These services are provided according to the needs of the cloud consumers by the cloud providers [12]. The services function as follows: Infrastructure-as-a-Service (IaaS) is the most basic service and one of the most important services, because without infrastructure the other services would not exist. This service offers cloud users physical machine and virtual machines. The virtual machines are offered through technologies like hypervisors. Virtualization and the hypervisor technology are the technologies that offer the cloud computing paradigm scalability. Scalability means that cloud providers can scale resources up or down based on the demand of cloud users. The Platform-as-a-Service (PaaS) model provides computing platforms and environments for cloud applications to be hosted. These platforms usually provide computing environments like operating systems and programming languages [2]. Platform-as-a-Service divides the environment it provides into three categories: integrated lifecycle platform, anchored lifecycle platform and enabling technologies as a platform [12]. Software developers don't worry about maintaining the underlying software and hardware layers, they just run their application solutions. The maintenance of the underplaying software and hardware layers are taken care of by the cloud providers. In the Software-as-a-service (SaaS) model, software is installed and managed by the cloud providers [13]. The cloud providers

are the ones that manage the infrastructure and platform on which the software will run. The cloud users have the right of entry to the software through cloud clients.

The consumption rates of electricity have become a subject worth talking about in recent years, as oil prices soar and coal mining becomes more expensive. In time the average data center will consume a lot of electrical energy just to go through a basic day; in fact, in large-scale companies, the very notion of saving electrical energy becomes an important topic altogether. In 2006 IT companies in the United States were said to be consuming about 4.5 billion dollars worth of electrical energy per year [14].

When a lot of electrical energy is used, the transistors and diodes and other parts decrease in their lifespan and thus the equipment becomes less functional with time. This means hardware has shorter life and it has to be replaced a lot, which can be an expensive process. The use of electricity also adds fumes and Carbon emissions into the air, which deplete the Ozone layer and are a threat to the natural environment; so saving energy has more advantages than it appears to have. Using virtualization technology reduces power consumption. Virtualization technology lets one integrate a number of servers to one node and thus the amount of hardware used is reduced by the use virtual machines.

Using the cloud computing paradigm virtualization can help to amplify the potential of getting more work done using fewer resources, which adds to the overall efficiency and provides resources on-demand as utilities over the Internet on a pay-as-you go basis [15]. As a result, maintenance costs of the enterprise computing environment are dropped; enterprises also can outsource computational services to the cloud. As in any other industries, providing trusted Quality of Service is critical for cloud providers. The Service Level Agreements such as time response and throughput with customers are used to define and manage this Quality of Service. Cloud providers ensure resource management that is efficient and provide high resource utilization rates; however, performance-power trade-offs are what cloud providers have to work with, because using too many virtual machines leads to loss of performance [16].

The cloud computing paradigm works more with virtualized resources than with physical resources, this difference grants customers the ability to receive on-demand resources on a pay-as-

you-go basis [13]. Instead of incurring high up front costs in purchasing IT infrastructure and dealing with the maintenance and upgrades of software and hardware, organizations can transpose their computational needs to the cloud. The proliferation of cloud computing has resulted in the establishment of large-scale data centers containing thousands of computing nodes and consuming enormous amounts of electrical energy [17].

The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), projects that by 2014 infrastructure and energy costs would contribute about 75% and IT would contribute 25% to the overall cost of operating a data center [18]. The notion that these resources such as hardware, are consuming all this energy, is not the only factor affecting this great amount of energy they consume. Most of the energy lost here can also be attributed to the poor efficiency with which the resources are used; better management and optimal re-distribution of these resources can alter things altogether [19].

Utilization in normal production servers hardly and rarely reaches100%; data collected from more than 5000 production servers over a six-month period have shown this. It is not that servers are completely idle and not processing anything, but they virtually never maximise their use either nor do they reach their peak operation [20].

Production servers operate at 10-50% of their full capacity for most of their operational time, leading to extra expenses on over-provisioning, and thus extra Total Cost of Acquisition (TCA) [21]. Moreover, managing and maintaining over-provisioned resources results in the increased Total Cost of Ownership (TCO); these implies that the more servers owned will only lead to more money spent in keeping them operational, yet somewhat mostly idle. Another problem is the narrow dynamic power range of servers: even completely idle servers still consume about 70% of their peak power [22]. Therefore, keeping servers underutilized is highly inefficient from the energy consumption perspective; because servers spend and dissipate a lot of electrical energy that never even goes to aiding the actual function of the servers. Gelas et al. [23] have conducted a comprehensive study on monitoring energy consumption by the Grid'5000 infrastructure. They have shown that there exists insignificant opportunities for saving energy via techniques that employ the mechanism of switching servers off or to low power modes so that not much energy is wasted while servers are idle.

In general energy resources are scarce and they ought to be protected for longevity, so there are other problems that come with high power and energy consumption by computing resources and hardware. Power, for example, is still required to feed the cooling system operation. For each watt of power consumed by computing resources, an additional 0.5-1W is required for the cooling system [24]. This means the cooling of processors tags along with itself a notable amount of energy consumption.

In addition, high energy consumption by the infrastructure leads to substantial carbon dioxide (CO_2) emissions; emissions of carbon dioxide contribute to the green house effect [25]. One of the ways to address the energy inefficiency problem is to employ the capabilities of the virtualization and load balancing technologies [26].

Cloud providers can use the technology of virtualization to create multiple Virtual Machine (VMs) instances on a single physical server, and can thus improve the utilization of resources and increase the Return On Investment (ROI) [27]. If energy is saved, then money is saved, this eventually lead to more profit being made, which implies that the invested funds have more return for their worth.

The reduction in energy consumption can be achieved by switching idle nodes to low-power modes (i.e. sleep, hibernation), thus eliminating the idle power consumption. In other words, when systems are not fully operational or not operational at all, it is best they sleep or hibernate, until a request is made for them to perform a task. Moreover, by using live migration [28] the VMs can be dynamically consolidated to the minimal number of physical nodes according to their current resource requirements. However, efficient resource management in clouds is not trivial, as modern service applications often experience highly variable workloads causing dynamic resource usage patterns. When a system is too dynamic, it becomes difficult to predict when it will be idle or not. Therefore, aggressive consolidation of VMs can lead to performance degradation when an application encounters an increasing demand, resulting in an unexpected rise of resource usage. If the resource requirements of an application are not fulfilled, the application can face increased response times, time-outs or failures.

Quality of Service (QoS) is defined via Service Level Agreements (SLAs) established between cloud providers and their customers. It is essential for cloud computing environments provide good QoS; therefore, cloud providers have to deal with the energy-performance trade-off and the minimization of energy consumption, while meeting the SLA demands at the same time. Businesses are established on the basis of customers, so the quality of the services rendered is also an important factor. Saving energy means nothing if services will be poor in quality, so the SLAs help to ensure good quality of services.

2.3 Key Concepts and Terminologies

The basic concepts below are relevant to this research work and they are provided in order to clarify their meaning and to add to the overall clear understanding of this dissertation:

a) Cloud Computing

Cloud computing can be defined as 'a type of parallel and distributed system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned, and presented as one or more united computing resources based on service-level agreements established through negotiation between the service provider and consumers' [13]. The various service models that are commonly used in cloud computing are SaaS, PaaS and IaaS [2, 25].

i) Platform-as-a-Service (PaaS)

PaaS is a service layer in cloud computing that facilitates deployment of applications without the cost and complexity of buying and managing the underlying hardware and software layers; so to speak selling the platform only [14,16].

ii) Infrastructure-as-a-Service (IaaS)

IaaS is a service layer in cloud computing that is responsible of delivering computer Infrastructure-as-a-Service, usually platform virtualization [22].

iii) Applications-as-a-Service (AaaS)/Software-as-a-Service (SaaS)

SaaS or AaaS are used interchangeably in cloud computing and they are service layers in cloud computing that eliminates the need to install and run the application on the customer's own computer [29].

b) Data Center

A data center is a physical building that contains multiple servers. It is also a complex distributed system consisting of a hierarchy of a number of components operating at multiple levels of usage abstractions [13, 28]. A data center encapsulates a set of computer hosts that can either be homogeneous or heterogeneous with respect to their hardware configurations (memory, cores, capacity, and storage). Furthermore, every data center component instantiates a generalized application provisioning component that implements a set of policies for allocating bandwidth, memory, and storage devices to hosts and VMs.

c) Public Cloud

Public cloud [2] services are characterized as being available to clients from a third party service provider via the Internet. The term "public" does not always mean free, even though it can be free or fairly inexpensive to use. A public cloud does not mean that a user's data is publicly visible; public cloud vendors typically provide an access control mechanism for their users. Public clouds provide an elastic, cost effective means to deploy solutions [2].

d) Private Cloud

A private cloud offers many of the benefits of a public cloud computing environment, such as being flexible and service based. The difference between a private cloud and a public cloud is that in a private cloud-based service, data and processes are managed within the organization without the restrictions and requirements that using public cloud services might entail; such as network bandwidth, security exposures and legalities pertaining to protection of the public. In addition, private cloud services offer both the provider and the user wider control of the cloud infrastructure; they improve security and resiliency because user access and the networks used are restricted and designated [30].

e) Community Cloud

A community cloud [13] is managed and used by a group of organizations that have shared interests, such as specific security requirements or a common mission or vision. The members of the community share access to the data and applications within the cloud.

f) Hybrid Cloud

A hybrid cloud [12] is a combination of a public and a private cloud that interoperates into one model; users typically outsource non-business critical information and processing to the public cloud, while keeping business-critical services and data in their control.

g) Network as a Service

The Network-as-a-Service model allows cloud users to interconnect between different clouds [31].

h) GridSim

GridSim is a parallel system simulator, distributed system simulator, grid system simulator and a modeling toolkit [32]. GridSim has many different classes for simulating application resources, resource brokers, users, and resource scheduling.

i) CloudSim

CloudSim [33] is a simulation and modeling toolkit that computes and evaluates resource provisioning algorithms in a cloud computing environment. CloudSim supports modeling of data centers, resource policies, service broker policies, CPU scheduling and virtual machines provisioning.

j) CloudAnalyst

CloudAnalyst [34] is a novel simulation toolkit for cloud computing. Its new approach focuses on simulation of large-scale applications in order to study their behavior in different cloud environment configurations. CloudAnalyst is easy to use because of its user friendly Graphic User Interface (GUI) which is very easy to configure and is flexible; it has a built-in function that saves user configurations, which in turn allows one to run experiments over and over again until a suitable configuration is noted or selected.

k) Load Balancing

Load balancing [17] is a method that tries to achieve maximum throughput, and achieve faster time response while avoiding overhead and overloading. It achieves this by dividing workloads and distributing them across different work stations, different central processing units (CPU), different network links and so on. Load balancing is usually used to supply one Internet service from a server farm. In an Internet service a load balancer is software that is placed at an Internet port where end users have access to the service. There are various methods that can be used to balance loads, most of these involve scheduling algorithms and some of the commonly used ones are round-robin methods, equally spread current execution loads and throttled.

l) Algorithm

An algorithm in its formal sense and meaning is a step-by-step process to solving a problem [16]. Algorithms are used for automated reasoning, which is a sub-field of artificial intelligence. Artificial Intelligence is a study field in computer science that deals with computers that can reason by themselves. Algorithms are also used for processing data and performing hybrid calculations. Mathematically, an algorithm can be thought of as an efficient method of expressing a limited list of instructions that are well defined, and to calculate a function. Algorithms can be expressed using flow charts, pseudo code and natural language statements in point form.

m) Simulation

Simulation is when one imitates or mimics something real or imitates a method or way in which it operates. The action of simulating something usually includes knowing or highlighting key distinctions or behaviors of a chosen material or intangible system. In computer science, simulation has some specialized meanings that have to do with how computers currently run programs [27, 31].

n) Hypervisor

A hypervisor is a term used to describe a virtual machine manager. This virtual machine manager can be a software program or hardware that can create and run virtual machines. the virtual machines created by the hypervisor are known as guest machines and the computer which is running the hypervisor is known as the host machine. Each guest machine created by the

hypervisor can run a different operating system such as Windows, Linux, UNIX, Apple Mac and so on. Hence a hypervisor presents operating systems in the virtual machines and manages their execution; operating systems in the virtual machines share the same hardware recourses of the same physical machine although they are multiple in number in a virtual sense. Hypervisors can be classified in to two types; Native and Hosted [8, 23].

o) Virtualization

Virtualization is a mental image of something [23, 36]. In a computing environment it also works the same way, it is the creation of an intangible version of a tangible entity like hardware platform, network resources and so on. The hypervisor (virtual machine manager) is the firmware or hardware that creates virtual machines [53]; that is the hypervisor virtualizes machines.

p) Cloud Broker

A cloud broker is responsible for mediating negotiations between SaaS and cloud providers; and where those negotiations are governed by the QoS requirements [27, 32]. The broker acts on behalf of SaaS providers. It discovers suitable cloud service providers by querying the CIS and it undertakes online negotiations for the allocation of resources/services that can meet the application's QoS needs. Researchers and system developers must extend this class for evaluating and testing custom brokering policies. The difference between the broker and the cloud coordinator is that the former represents the customer (i.e. decisions of these components are made in order to increase user-related performance metrics), whereas the latter acts on behalf of the data center (i.e. it tries to maximize the overall performance of the data center, without considering the needs of specific customers).

q) Energy Efficiency

Energy efficiency is the goal of minimizing the amount of energy consumed or dissipated when products and services are created or delivered. Most profit based organizations, and industrial organizations, consider energy saving to be part of their mission or vision. For example, automobiles are built to consume less petrol for more distance travelled [9, 37].

2.4 Virtual Machines in Cloud Computing

There is a rapid growth in demand for computational power. The growth is driven by modern service applications combined with the shift to the cloud computing model. This growth has led to the establishment of large-scale virtualized data centers. Such data centers consume enormous amounts of electrical energy and this consumption results in high operating costs and carbon dioxide emissions.

The dynamic grouping of virtual machines (VMs) via the employment of live migration and the switching of idle nodes to the sleep mode, allows cloud providers to optimize resource usage and to reduce the amount of energy usually consumed. Providing high quality of service to customers is inevitably linked to the energy-performance trade-off. This is because aggressive consolidation of VMs may eventually lead to slower or poor performance overall. Poor performance would be a violation of the terms pertaining to service provision and, as explained prior, these quality terms are modelled after the SLAs.

Due to the variability of workloads experienced by modern applications, the VM placement should be optimized continuously in an online manner; continuously meaning that it keeps up with the dynamic nature of the system's operations. Virtual machine consolidation is an approach to maximize the utilization of resources while minimizing energy consumption. As shown in Figure 2.1, its basic principle is to maximize the number of inactive physical servers by consolidating the virtual machines on a minimum number of active servers. This means the idle servers are on sleep or hibernation, while the active physical servers carry their processing load. Ideally, due to the static amount of energy consumed by the servers' components (especially the CPU), running servers at the maximum utilization level is more energy efficient. The energy consumed by hardware at 100% peak processing is fairly the same as at lower percentiles, i.e. the energy consumption is static or non-changing with respect to processor usage. This is typically referred to as the lack of energy proportionality in modern server hardware [35]. On the other hand, since the difference in energy consumption between an idle and a suspended server is quite high, suspending an inactive server provides another opportunity for energy saving. A suspended server consumes much less energy than an operational server which is just idle.

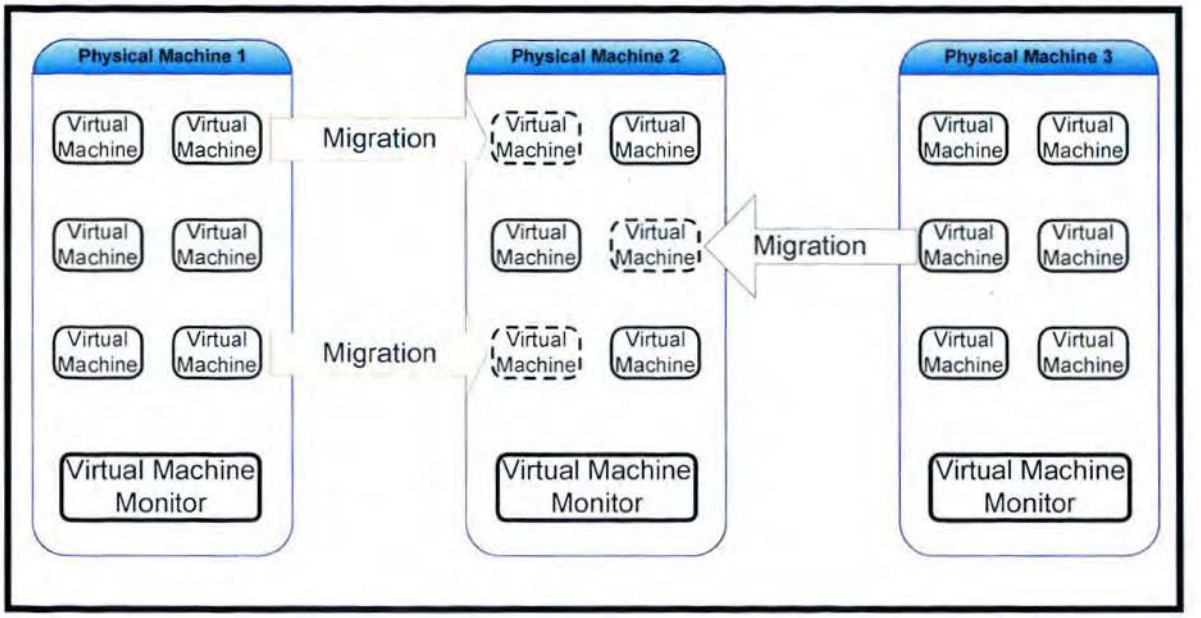

Figure 2.1 Virtual machine consolidation approach [7]

In an IaaS cloud computing environment, virtual machine consolidation can increase the number of suspended servers. Unfortunately, there are several practical problems to solve that may possibly influence the energy consumption of a cloud significantly [7]:

1. Dependencies between energy consumption, resource utilization and performance of consolidated virtual machines: consolidated virtual machines share and compete for resources (e.g., CPU, main memory, I/O) of the physical server they run on. Performance degradations and increased execution times may arise along with energy savings originating from server suspensions [7]. Shared resources imply that these resources may be overloaded if the operation dynamics grow to be unexpectedly demanding. This is not ideal in industries where quality of service affect customers behaviour. Some famous companies have lost money and customers due to overloaded networks or resources [20].

2. Additional overhead for virtual machine consolidation: The key enabling technology for virtual machine consolidation is virtual machine live migration; this essentially means that migration happens 'live' or during operation. However, there is an additional overhead spent for live migration and its preparation and post processing phases, including the migration of a virtual machine's CPU state, main memory, network connections and storage. This means that it takes

memory to process the actual migration and thus it takes time and energy itself to implement, giving an overhead which must be accounted for as it is implemented.

3. Prediction of a server's energy consumption: Typically, a user of an Amazon EC2 instance can control the entire software stack from the kernel upwards and produce arbitrary workloads on physical machines as well as unpredictable user behaviour in terms of starts and terminations of VMs. This complicates virtual machine consolidation considerably, since it is harder to predict a server's energy consumption. On the other hand, energy performance profiling techniques for better predictions can be resource intensive and can produce additional overhead, which is often unacceptable for cloud computing providers [35, 38]. Software that is used to predict more accurately the expenditures is itself going to require the same processing equipment or power to make the more accurate predictions it is installed to make, making it a tricky dynamic to optimize.

These problems arise during operation and they get complex in a practical situation; therefore they should be addressed from a standpoint that is sensitive to the energy efficiency of virtual machine consolidations:

- During normal execution, the performance of a virtual machine should not be affected to avoid increased energy consumption through longer execution times. Faster response and quicker intervals are more important than energy saving, especially for profit organizations.
- During the different phases of virtual machine live migration, the additional overhead should be kept at a minimum. Live migration must not take more memory or processing power than it should, during its implementation.

Other approaches, such as workload aware bin packing algorithms [7], can also benefit from this approach. The technology behind virtual machine consolidation is virtual machine live migration, as supported by all modern Virtual Machine Monitors. For example, the Xen Hypervisor (used by Amazon EC2) can migrate a virtual machine's CPU state, main memory and network connections transparently without a significant downtime. The only functionality not supported is storage synchronization with respect to the virtual machine's root image, swap space and ephemeral local storage.

The Cloud provider is responsible for providing storage synchronization for virtual machine migration. Typically, the virtual machine's root image and its ephemeral local storage are backed on a network file system. There are two negative aspects of this approach. First, the server that the virtual machine runs on needs permanent remote access to the file server via a network connection, leading to network traffic during access to a disk image. Second, random network file system access is less efficient than local disk access. This is due to the higher latency and lower bandwidth of commodity network hardware compared to a local disk.

In contrast, the proposal in this research is based on a storage synchronization approach. It introduces explicit storage synchronization and virtual machine live migration phases instead of permanently synchronizing the disk image via a network.

The proposed storage synchronization approach is based on leveraging the concept of a Distributed Replicated Block Device (DRBD) [36], typically used for high availability data storage in a distributed system. It replicates data storage to different locations in a stable, fault tolerant way. The DRBD module can operate in two modes: stand alone and synchronized. In stand-alone mode, all disk accesses are simply passed on to the underlying disk driver. In synchronized mode, disk writes are both passed on to the underlying disk driver and sent to a backup machine via a TCP connection, while disk reads are served locally.

The DRBD module can be used for live migration in a cloud computing environment [16]: the virtual machine's root image, swap space and ephemeral local storage are used locally during normal execution phases. When a virtual machine is consolidated, i.e., during the storage

synchronization and live migration phases, the DRBD module is set to synchronized mode. After the end of these phases, the DRBD module is switched back to standalone mode.

Furthermore, each virtual machine needs its own working copy of a root file system image due to local modifications, although many virtual machines might share the same basic image. Therefore, the synchronization of gigabyte sized disk images should be reduced to local modifications only. The approach in this research is to use a multi layered root file system (MLRFS) for the virtual machine's root image. The basic image is distributed centrally (e.g., by accessing Amazon S3) and cached on a local disk. The basic image and a separate layer storing local modifications are overlaid transparently and form a single coherent file system using a Copy On Write (COW) mechanism; therefore, only local modifications, i.e., a small, separate layer instead of entire disk images, are transmitted during the disk synchronization phase. This approach has originally been developed by Schmidt et al. [37] to apply security updates to virtual machines. Thus, it is possible to apply modifications to the basic image and to redistribute it in an efficient manner.

2.5 Overview of Load Balancing

Load balancing is a mechanism designed to achieve maximum throughput and faster time response while avoiding overhead and overloading in the process [17]. It mainly does this by dividing workloads and distributing them across different work stations, different central processing units (CPU), different network links and so on.

Load balancing is usually used to supply one Internet service from a server farm [17]. In an Internet service a load balancer is a software that is placed at an Internet port where end users have access to the Internet service [34].

There are various ways through which loads can be balanced; most load balances are achieved via: scheduling algorithms and some of the commonly used ones are round-robin, throttled and equally spread current execution loads [33].

2.5.1 Load Balancing in Cloud Computing

The goal of load balancing is to improve the performance by balancing the load among these various resources (network links, central processing units, disk drives...) to achieve optimal resource utilization, maximum throughput, shorter response time, and to avoid overload. The distribution of load on different systems generally uses traditional algorithms like those used in web servers, but these algorithms do not always give the expected performance with large scale and distinct structure of service-oriented data centers [17]. To overcome the shortcomings of these algorithms, load balancing has been widely studied more carefully by researchers and implemented by computer vendors in distributed systems.

Every data center system has distinct features, which must be carefully studied in order to develop algorithms that are most suited for that particular data center system and its dynamics and nature.

In general, load balancing algorithms follow two major classifications [38]:

- Depending on how the charge is distributed and how processes are allocated to nodes (the system load);
- Depending on the information status of the nodes (System Topology).

In the first case it is designed from a centralized approach, distributed approach or hybrid approach; in the second case from a static approach, dynamic or adaptive approach.

2.5.2 Classification of Load Balancing Technologies

- **Classification According to System Load**

a) **Centralized approach**: In this approach, a single node is responsible for managing the distribution within the whole system [17].

b) **Distributed approach**: In this approach, each node builds its own load vector by itself by collecting the load information of other nodes [17]. Decisions are made locally using local load vectors. This approach is more suitable for widely distributed systems such as cloud computing.

c) **Mixed approach (Hybrid)**: A combination between the two approaches to take advantage of each approach [17].

- **Classification According to System Topology**

a) **Static Approach**: This approach is generally defined according to the design structure or implementation mechanisms of the system [17].

b) **Dynamic Approach**: This approach takes into account the current state of the system during load balancing decisions [17]. This approach is suitable for distributed systems such as cloud computing also.

c) **Adaptive approach**: This approach adapts the load distribution to system status changes, by changing their parameters dynamically and even their algorithms [17]. This approach is designed to offer better performance when the system state changes frequently because it adapts to the changes [7]. It is more suited too for distributed systems such as cloud computing.

2.6 Data Center Productivity

Data Center Productivity (DCP) is the amount of useful work that a data center produces as related or compared to the amount of resources that the center consumes as it performs that work during operational production. This can be mathematically expressed as:

$$DCP = \frac{Useful\ Work\ Produced}{Total\ Quality\ of\ a\ Resource\ Consumed\ Producing\ this\ Work} \quad (1)$$

This implies that productivity is work done per resources consumed.

From this parent equation an entire family of metrics can be derived based on the specific resource that is to be optimized. This equation behaves like a general equation, where specifics settings can be tailored into the computation. For example, one might be interested in the amount of work a data center produces per peak power consumed or per square foot of floor space utilized. The metrics of this family will all be designated by a name of the form data center Productivity (DCxP).

This work focuses on useful work produced relative to the energy consumed producing this work more specifically. This metric is called Data Center energy Productivity (DCeP).

2.6.1 Data Center Energy Productivity Metric

The goal is to define a metric that quantifies the useful work that a data center produces based on the amount of energy it consumes. Mathematically this can be expressed as:

$$DCeP = \frac{Useful\ Work\ Produced}{Total\ Data\ Center\ Energy\ Consumed\ Producing\ this\ Work} \quad (2)$$

Note that since we are considering energy and not power, the period of time over which energy is measured must be specified to make this metric meaningful [22]. This time period shall be called the assessment window. Energy is measured by the integral of instantaneous power over a specific time interval [39].

2.6.1.1 Measuring Energy Consumed

Before examining how to quantify useful work, measuring the quantity that constitutes the denominator in Equation 2 - the energy consumed by a data center during the assessment window will be considered.

This work assumes that either the electrical power feed to the entire data center is instrumented or that each piece of equipment that makes up the data center including its power conditioning and distribution and cooling infrastructure equipment is separately instrumented, and that it is capable of reporting its current power utilization.

Note that the total data center energy may also be estimated based on a measured value of the total energy consumption of the IT equipment multiplied by the current data center Power Usage Effectiveness (PUE) value given that is if this value is available [7, 15]. Decreasing PUE or equivalently increasing Data Center Infrastructure Efficiency (DCiE) has the effect of improving Data Center Energy Productivity (DCeP).

2.6.1.2 Defining Useful Work

The DCP metric and all its derivative metrics require the quantification of useful work. Useful work may be defined by the equation:

$$\text{Useful Work} = \sum_{i=1}^{M} V_i * U_i(t, T) * T_i \qquad (3)$$

where M is the number of tasks initiated during the assessment window, V_i is a normalization factor that allows the tasks to be summed numerically, $T_i = 1$ if task i completes during the assessment window, and = 0 otherwise. $U_i\ (t,T)$ is a time-based utility function for each task, where the parameter t is elapsed time from initiation to completion of the task, and T is the absolute time of completion of the task [8].

Note that Useful Work is defined to be the sum over i of all tasks 1 through M initiated within the assessment window multiplied by a time-based utility function $U_i\ (t,T)$. The factor V_i assigns a normalized value to each task so that they may be algebraically summed. T_i eliminates all tasks that are either initiated prior to the assessment window or are initiated within the window but do not complete. The following sections will discuss the key terms introduced in Equation 3.

2.6.1.3 Defining a Task

To execute this measurement, all the tasks initiated within the assessment window must be known. Here it is useful to distinguish between the concept of a task type and the concept of a task instance. A task type is descriptive of a specific class of processing that the data center provides and involves the invocation of a specific piece of application software installed in the data center. A task instance is a single invocation of this software with a specific set of input parameters or data and a specific resultant output [10].

While a given data center may in the course of one day carry out a very large (perhaps in the millions) number of task instances, it will normally process a much smaller number of task types.

Task types are defined prior to the assessment of DCeP based on the installed equipment and software within the data center [8]. To simplify the quantification of Useful Work, tasks are aggregated according to task types. Task instances within a given task type will all have the same relative value and must comply with the same service level agreement.

This means that the parameters V_i and U_i (t,T) are determined on a per task type basis instead of per task instance. It further means that certain tasks are the type of tasks that the customer is waiting for, whereas some tasks performed are internal to the organization and unrelated to the customer directly.

The formulation of Useful Work leaves the definition of what is considered a "task" up to the person personalizing the metric for use in a given data center in order to meet the specifics related to the center. This makes the metric applicable to any workload.

However, if a task is defined too broadly, for example, "maintain data base X" it will not be possible to determine if the task completes within the window [8]. This is solved by redefining such a task at a finer level of granularity; increased resolution narrows the computation down. For example, the task "maintain data base X" could be broken down into a number of typical subtasks involving data base X such as "satisfy query against data base X" or "load a new record into data base X," or "run standard report a against data base X."

2.6.1.4 Defining the Assessment Window

To calculate the energy used more sensibly, the time frame for which the energy is consumed must be stipulated; this means a time window must be established or chosen [7]. This time window is called the assessment window.

The length of this window is arbitrary, but to obtain accurate results in executing the measurement, it should be no shorter than about 20 times the mean run time of the any of the tasks initiated in the assessment window. Tasks must be allowed to run long enough to complete, an unduly short time window may lead to results that seem to imply that no task was completed as the time for which they were computed was way too short. An assessment window should be sized in accordance

with the nature of the workload and the purpose of the measurement [40]. For example, a useful assessment window could be as short as a few milliseconds or as long as a month or more.

All these settings are necessary in order to meet the statistical implications of data sampling from the stand-point of Mathematics [21], in what is called a "representative sample" in Statistics that is a data set that suitably encapsulates the spectrum of the workload and task types and instances involved per data center. Note that this methodology as defined, ignores any tasks that may have been initiated prior to the start of the assessment window and those that are initiated within the window but do not complete prior to the end of the window [8]. These tasks will consume energy during the assessment window, but will not contribute to Useful Work. These effects lead to an error in the measurement of Useful Work. This error, however, may be minimized by appropriately sizing the length of the assessment window [8].

2.6.1.5 Assigning a Value to Tasks

It is clear that not every task that the IT equipment in a data center performs has the same value or ranking [40]. Yet, in order to aggregate the useful work that a server or group of servers produces, the tasks must be normalized. This is the purpose of the factor V from Equation 3. The value of V_i must be assigned prior to the assessment of the DCeP metric for each task type so that the value of the task is normalized to some standard task that the data center performs [8]. In this way, more important or valued tasks receive greater weighting in the calculation of Useful Work and the completion of less important tasks receive a lesser weighting.

When appropriate, a straight-forward simplification of this process of determining the V_i weights all of them to the value 1.0. This indicates that all defined task types have approximately the same value to the end user or the owner of the data center. Algebraically, multiplying by the value one (1.0), does not alter a variable's magnitude.

2.6.1.6 Defining a Time-Based Utility Function

Note that the function $U_i(t,T)$ in Equation 3 must be specified for each task type, prior to running an assessment of the metric. This function handles the time dependent nature of the value of each task. The variable "t" is relative run time while T is the absolute time of completion. A given utility function can ignore one or both parameters.

In other words, U_i can be a constant, a function of the relative run time of the task (in which case the function may be denoted by $U_i(t)$), a function of only the absolute completion time ($U_i(T)$) or a function of both the run time and the absolute completion time ($U_i(t,T)$).

If the value of completing a given task is time invariant, U_i for that task should be expressed as a constant. A typical run time based utility function will help in discussing this.

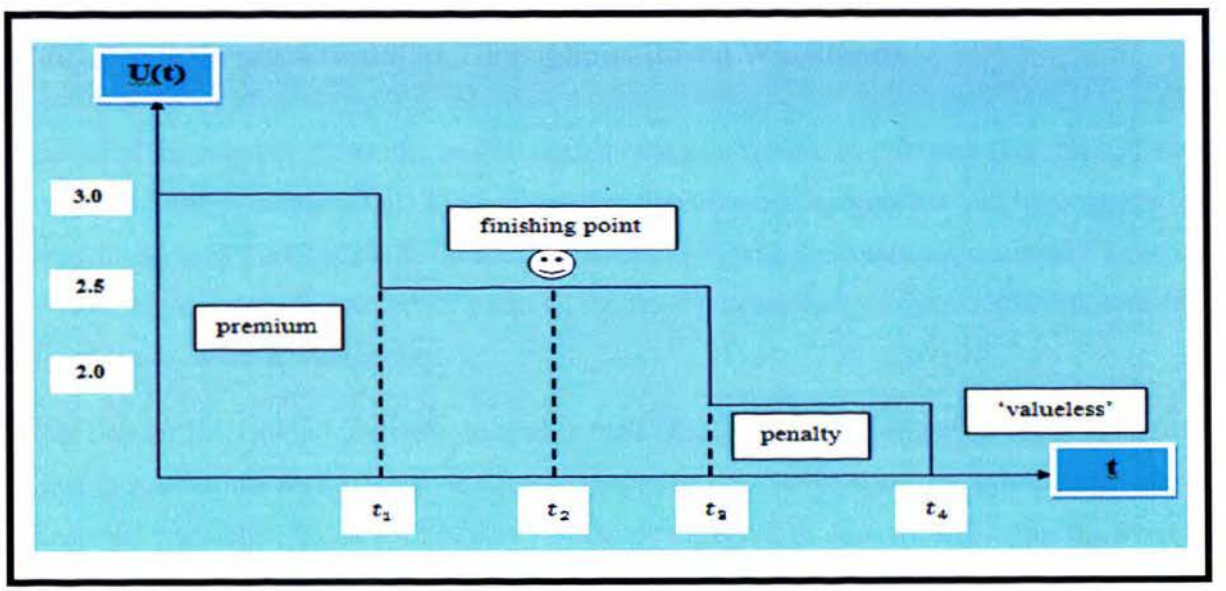

Figure 2.2 Utility function example

The example in Figure 2.2 shows how the value of a task may change as run time increases. Assume that a service level agreement (SLA) applies to this task. In this case, the time based utility function merely represents this SLA in mathematical form. Note that this SLA provides for a premium to be paid to the service provider if the task is completed early (prior to t_1). If it is

completed in the interval t_1 to t_2, the SLA is met. If it does not complete by t_2, the service provider has agreed to a penalty of 50%. At t_3, this penalty goes to 100% [8]. This is, of course, a hypothetical example, but it displays the power of the utility function to capture mathematically the terms of an SLA. If no SLA exists, other means may be used to decide the form of the utility function.

For other tasks, the most important determinant of utility is absolute completion time (e.g., payroll must complete by 6:00 AM Friday morning.). In this case, U_i (T) would be a function of absolute time alone. Note that an easy simplification is to assign a constant value of 1.0 for the utility function; with the aforementioned definition of DCeP we now have a mathematical basis for measuring the total work output of a data center in such a way that it can be related to the energy consumed to produce this work output [8].

2.6.1.7 Transactional or Throughput-Based Workloads

Several of the simplifications that may be made to the equation for useful work (Equation 3) have been discussed in section 2.6.1.2. One specific simplification appropriate for transactional or throughput-based workloads is to set both V_i and U_i to 1.0; with these assumptions made, Equation 3 would represent the simple act of counting the tasks (transactions or operations) that complete during the assessment window [8].

Note that for this case all tasks are deemed to have equal value. This simplification is often used when measuring the performance of a transaction processing application, for example. In typical situations, transaction-based applications will be instrumented to measure and report the average number of transactions processed per second. Other applications may report performance in terms of operations per second, queries per second, web-pages served per second, or jobs completed per second [17]. Using this reported performance value useful work may be estimated by the equation:

$$Useful\ Work = \text{Average Transaction Rate} * \text{Length of Assessment Window} \quad (4)$$

where any rate-based performance metric may be substituted for Average Transaction Rate

Note that the rate of the metric remains the energy consumed during the assessment window. It is also important to note that the length of the assessment window could be any length of time that the data center operations staff deems to be suitable for their business.

2.7 Related Work

Forster et al [12] compared Grid computing and Cloud computing from different angles, in order to find the essential characteristics of both paradigms. The authors started by comparing business models of both paradigms and move on to comparing architectures of both paradigms. They continued to compare programming models, application models and security models of both paradigms. They concluded that Grid and Cloud computing share the same vision, architectures and technologies, but they have different application models, data models, business models, computing models, security models and programming models.

Mach and Schikuta [41] presented an "analytical economic cost model for cloud computing aiming at comprising all kinds of commercial environment." Their model allowed for evolution of business strategies of cloud environment because it includes both fixed and variable costs. Also the model supports decision making processes to be applied with business cases. They concluded that business strategies can be derived for both cloud consumers and cloud providers. More over the authors concluded that, according the economic federations, energy efficiency of cloud systems can be examined and analysed.

Moreno and Xu [42] presented dynamic resource provisioning mechanisms based on customer utilization patterns, to allocate capacity in real-time cloud data centers. They also analyse the impact of their model in fulfilling and energy efficiency. The goal of Moreno's and Xu's model is to "improve energy efficiency by reducing the waste of resource derived from customers' overestimations." They used empirical methods to compare three over allocation approaches over 24 hours. The approaches are over allocation LAF, over allocation FDF and without over allocation. The authors concluded that their model improves data center utilization as compared to simple DRR approaches.

Justice et al [8] identified energy efficiency metrics for reducing costs and implementation of green initiatives in data centers to be used by IT managers for measuring and maintenance purposes. They examined the strengths and weakness of metrics, PUE and DCP, two of the most commonly used metrics data center metrics. From their findings they concluded that there is a need for standards and metrics in the industry and farther recommended that future metrics should be normalized for all data centers across the industry

Xiaoli and Zhanghui [43] proposed a new model which is energy aware that considers the energy efficiency in cloud computing environments. They improved the Bin Packing algorithm. They used simulation with both C++ and Matlab to analyse their results. Based on their results they concluded that their algorithm makes good use of resources and that fragments of the active server can be used well.

Calheiros et al [29] introduced CloudAnalyst, a novel simulation tool for large scale applications in cloud computing environments. They explained in detail the architecture of the CloudAnalyst simulator and the various algorithms and policies contained in it. They further showed how the simulator can be used in different scenarios. They concluded that CloudAnalyst is not a comprehensive solution for all simulation needs and that their approach and tool will evolve over time.

Wang and Wang [10] proposed a new energy efficient multi-task scheduling model based on Google's framework that processes massive data called MapReduce, to improve energy efficiency of servers. For their model, they created a new practical decoding and encoding method for individuals. They used methods from intelligent systems. They also introduced a local search operator in their algorithm to improve its search ability. The authors used simulation methods to validate their model and from their results they concluded that their model is efficient and effective.

Raj and Shriram [9] investigated the impact different parameters have on power consumption by data centers in cloud computing environments. The parameters they manipulated to check the impact are peak power, energy consumed, SLA adherence and average request processing time for

six different configuration. They used simulation to collect their results. Based on their results, Raj and Shriram concluded that their 5th configuration was the best configuration.

Beloglazov and Buyya [16] conducted competitive analysis for both single and dynamic machine migration and consolidation. Based on historical data of resources used by virtual machines, they propose a novel adaptive heuristic for dynamic consolidation of virtual machines. They used real world workloads in their simulation to gather results. Based on their results they found that their proposed algorithm reduces energy consumption significantly while maintaining service level agreements (SLA). Beloglazov and Buyya concluded that it is necessary to develop adaptive or randomized algorithms to improve performance of optimal deterministic algorithms.

Lombardi and Di Pietro [44] proposed a novel security architecture for cloud computing environments aiming to guarantee increase in security and cloud and cloud computing resources, called Advanced Cloud Protection Systems (ACPS). The system is implemented on two open source solutions: Eucalyptus and OpenECP. It was tested for efficiency against attacks and performance evaluated against different workloads. From their results, Lombardi and Di Pietro concluded that their system is good against attacks and that it produces small overheard.

Xing and Zhan [21] discussed and analysed how to improve security in cloud computing environments and also how virtualization is used in cloud computing. They concluded that, because of the many factors that affect cloud computing security, governments, government departments and other industry policy makers should work together to archive one common goal of security.

Beloglazov and Buyya [7] presented a novel technique which guarantees that service level agreements are met at high levels, through dynamically consolidating virtual machines on adaptive utilization thresholds. They used different workload traces for more than a thousand PlanetLab servers to authenticate their technique. The simulation tool they used is CloudSim, and the parameters they used are energy consumed, SLA and virtual machine migration. From their results, they concluded that their technique outperforms other policies in terms of SLA adherence. They also recommend that SLA should be relaxed in order to reduce energy consumption in future.

Wo et al [5] proposed a system for on-line virtual machine cluster provisioning that manages resources efficiently. Their goal for providing such a system was to make available to academic users virtual clusters. They specifically examined intelligent resource mapping and virtual machine image management, and used real world workloads to evaluate their system. The authors examined the overhead, impact from two coefficients and resource utilisation. From their results, Wo et al concluded that their approaches are better than other approaches.

Ahamed and Alexandrov [4] took a closer look and outlined opportunities and problems found in cloud computing, like access management and identity theft of cloud users. They discussed and identified weaknesses in some of the existing models, such as authentication management, access management, monitoring and auditing, trust-relationship, identity service provision, and others in the cloud model. They concluded that for the sake of adequate security safeguard, it is important to align the IAM framework with cloud landscape.

Li et al [45] proposed a novel approach named EnaCloud, an application that allows applications to be placed dynamically live, considering energy efficiency in cloud computing environments. They used the Xen virtual machine monitor to conduct experiments. The parameters they used were time, energy consumed, utilization and number of active notes. From their results they concluded that their approach is feasible.

Beloglazov and Buyya [14] evaluated heuristics to minimize energy consumption using dynamic reallocation of virtual machines. They used CloudSim toolkit to simulate their experiments and get results. From their results, Beloglazov and Buyya concluded that switching idle notes and dynamic reallocation of virtual machine improves QoS and that it is applicable in real world scenarios.

Zhang and Fu [39] presented a framework that uses power profiling and coarse-grain power for energy-efficiency in cloud computing environments called macropower. They implemented a macropower prototype on a test bed to collect their profiled data. They tested different system configurations and the parameters they used were time, power, energy and frequency.

Beloglazov et al [40] presented architectural elements, challenges and vision of environments in cloud computing. They continued to propose a novel software technology for management of energy efficient cloud environments, energy efficient management architectural principles of cloud

computing environments, and energy efficient resource allocation policies scheduling algorithms considering quality-of-service expectations, and device power usage characteristics. The authors evaluated what they proposed using the CloudSim toolkit. Some of the parameters they used were SLA violations, different policies, number of migration, energy consumption and utilization threshold. From the results they concluded that the advances in cloud computing are playing a big role in the reduction of energy consumption of data centers.

Gelenbe et al[46] reviewed the energy efficiency methods and technologies used for operation of network infrastructure and computer hardware. They surveyed some of the policies and energy saving strategies used in cloud computing. They concluded that cloud computing used together with virtualization is one of the key tools to identify trade-offs between energy efficiency, QoS, and performance.

Sato et al [6] designed, implemented and examined a neural network integrated in to the Green Scheduling algorithm in order to optimize energy consumed by servers in cloud computing environments. They performed their experiments through simulation using two load traces, and simulated data centers with 32 single-core, 8 quad-core and 256 dual-core servers. The parameters they used were time and number of requests. They measured power, reduction rate and drop rate. From their results they concluded that 20% additional servers plus their prediction is the best configuration.

Moreno and Xu [27] discuss and debate about how important it is to have energy efficient mechanisms in place inside data centers for the purpose of energy reduction and implementing them in real world cloud environments. Some of the key challenges are energy aware computing, the computing, and so on. From their analysis they concluded that there are still some gaps that need to be filled for more optimal energy savings in cloud computing environment.

Taylor et al [18] assessed the behavioural impact of building occupants from different social domains when provided with personal electricity utilization, using three different groups in their study. They collected their data from electrical meters they installed in a building that has 83 rooms and six flours. From their results they concluded that the group that consumed the least amount of electricity was the one that could view utilization of the network.

Lu and Gu [47] proposed a load-adaptive scheduling model for managing cloud resources based on the ant colony algorithm. They used simulation to test their model. The parameters they used were time, CPU usage, and number of requests. They concluded that their model improves energy efficiency of resources utilized and that it can meet the requirements of self-adaptive resource scheduling in cloud computing environments.

Calheiros et al [29] proposed a simulation tool called CloudAnalyst to fill the gap left by other cloud computing simulators. The gap evaluates geographically located large scale cloud computing applications. The architecture of CloudAnalyst is based on JavaSim and Cloud Sim which is based on GridSim. The authors developed CloudAnalyst using Java Programming Language. From CloudSim they added a graphic user interface using some extensions also from CloudSim, and showed how CloudAnalyst can be used to collect data using different scenarios. The authors concluded that the tool has to grow and evolve and produce better quality results, which in turn will improve performance of up coming cloud computing applications.

Ranjan et al [32] proposed a modelling, simulating and evaluating toolkit for cloud computing environments called CloudSim. They used different use cases to show how CloudSim can be used to model, simulate and analyse different markets. They recommended that cloud providers should use CloudSim to test performance of their resource provisioning and policies of their service delivery and in future they planned on coming up with new pricing and provisioning policies for CloudSim.

Tang and Dai [19] proposed an approach that is consumption related to decrease power consumption of cloud services in cloud data centers. Their approach levels peak and off-peak demands to optimize energy-efficiency. The parameters they used to evaluate their model in their experiments were time, energy consumed and number of users. They concluded that their approach improves generated power utilization and that the gap between peak and off-peak demands is going to be widened in future.

Singh and Hemalatha [31] discussed the value added by simulators to understanding the impact of different scenarios before implementing them in real world. The simulation tools they discussed are GridSim, CloudSim and Cloud Analyst. They concluded that CloudAnalyst is easier to use

than the other two simulators. In future they planned to improve the load balancing algorithms in CloudAnalyst.

Zeng et al [28] proposed a scheduling model for virtual resources and explained it by NSGA11. They developed and implemented three new algorithms in NSGA11 to improve its energy efficiency, and evaluated their model by balancing virtual and physical resource loads. The authors compared their vision of NSGA11 against Random, Static and Rank algorithms. The parameters they used were bandwidth, memory, CPU usage and time. Zeng et al concluded that their vision of NSGA11 is more energy efficient because it processes faster than the other algorithms.

Maciocco et al [35] introduced a low overhead i/o architecture called DirectPath, that optimizes the content within a platform. The goal of DirectPath is to improve throughput performance and energy efficiency. The authors examined DirectPath on a small form-factor SoC based platform and laptops. The parameters they used were CPU utilization, time, and data size to measure throughput and performance percentage. They concluded that DirectPath improves performance by 137% and reduces energy consumption by 50%.

Ghali et at [48] proposed two energy models to minimize energy consumption in providers in data centers, based on a statistical analysis of server operational behaviour and virtual machine migration. One of the models is the server shutdown energy saving model and the other is server sleep energy saving model. The authors used empirical methods to evaluate their models. The parameters they used for evaluation were time, data size and memory. From their experiment they found that the difference between the server models is energy and to put the servers in operational mode from sleep mode to off state mode.

Tan et al [30] proposed a cloud computing management architecture that is energy aware for private cloud data centers, and the authors also presented a power and cloud application cloud model. The goal of their model is to reduce overall energy consumption of data centers. They conducted their experiments using the Xen hypervisor and six DELL PowerEdge C2100 server running different applications. Some of the parameters they used were memory size, VM number, Virtual CPU number. The authors concluded that energy efficiency of cloud applications are efficiently improve their cloud application model.

Kelley et al [49] discussed studies that led to ultra-low power consumption of cloud computing systems, and task allocations in ultra-low power cloud models to optimize cloud computing architectures. They concluded that data centers can use a mix of systems such as xenon nodes and atoms to reduce energy consumption without losing power.

Ayre et al [25] analysed energy consumption in cloud computing environments, covering both public and private clouds. They used the PUE metric to analyse their results. Some of the parameters they used were encodings per week, energy consumed per user, time, energy consumed per process and percentage of total energy consumed. From their analysis they concluded that cloud computing is not the greenest solution and that cloud computing energy consumption needs to be considered as an integrated supply chain logistics problem.

2.7.1 Critical Analysis

It is clear that the need for energy efficient models, energy consumption monitoring tools and power consumption optimizing algorithms captured the attention of a lot of researchers in the area of cloud computing. Most researchers researched the effects of virtual machine migration and consolidation on energy consumption, while others researched the effects that load balancing has on energy consumption. Some researchers used profiling to conduct their experiments and most used simulation. The parameters that are widely used are CPU utilization, time, and number of requests.

2.8 Optimization Algorithms for Resource Sharing in Cloud/Grid

This section discusses different resource sharing/allocation optimization algorithms.

2.8.1 Energy-aware Dynamic Resource Allocation

Recently, virtualization has been greatly developed. This growth and development has led to virtualization being used very prolifically across data centers [14]. In any field, as the efficiency of

a method increases, many organizations begin to show interest. By supporting the movement of VMs between physical nodes, it enables dynamic migration of VMs according to QoS requirements. When VMs do not use all the resources available, they can be logically resized and consolidated on a minimal number of physical nodes, while idle nodes can be switched off or hibernated [7]. Currently, resource allocation in a Cloud data center aims to provide high performance while meeting SLA requirements, without a focus on allocating VMs to minimize energy consumption [50]; that is, to focus more on the satisfaction of the consumers than the technology itself.

In this particular technology, when performance and energy efficiency are to be optimized, three crucial issues must be addressed. First, excessive power cycling of a server could reduce its reliability as it wears out the life of the electronics [22]. Second, turning resources off in a dynamic environment is risky from a QoS prospective [19]. Due to the variability of the workload and aggressive consolidation, some VMs may not obtain required resources under peak load, so failing to meet the desired QoS [7]. Third, ensuring SLA brings challenges to accurate application performance management in virtualized environments [50].

A virtual machine cannot exactly record the timing behavior of a physical machine [26]. This leads to the timekeeping problems resulting in inaccurate time measurements within the virtual machine, which can lead to incorrect enforcement of SLA. All these issues require effective consolidation policies that can minimize energy consumption without compromising the used-specified QoS requirements [7].

2.8.2 QoS-based Resource Selection and Provisioning

Data center resources may deliver different levels of performance to their clients; hence, QoS-aware resource selection plays an important role in cloud computing [28]. This means the quality of service presented is not guaranteed to be standard enough and it therefore needs to be managed so that it may produce optimal results or output.

Additionally, cloud applications can present varying and continually changing workloads during operation times [50]. It is, therefore, essential to carry out a study of cloud services and their workload-dynamics in order to identify common behaviors, patterns, and explore load forecasting approaches that can potentially lead to more efficient resource provisioning and consequent energy efficiency. Although it is complex to carry out such a study, the analysis must be as detailed as possible, in order to achieve the necessary accuracy.

In this context, this research will sample applications and correlations between workloads, and attempt to build energy efficiency models that can help explore the trade-offs between QoS and energy saving. Although energy must be saved, quality must be preserved within the required par.

Further, this research will investigate a new approach to the consolidation strategy of a data center that allows a reduction in the number of active nodes required to process a variable workload without degrading the offered service level. The approach will automatically select a VM configuration while minimizing the number of physical hosts needed to support it.

2.8.3 Optimization of Virtual Network Topologies

In virtualized data centers VMs often communicate between each other, establishing virtual network topologies [20]. However, due to VM migrations or non-optimized allocation, the communicating VMs may end up hosted on logically distant physical nodes providing costly data transfer between each other [14].

If the communicating VMs are allocated to the hosts in different racks or enclosures, the network communication may involve network switches that consume significant amount of power [43]. This data transfer overhead can be eliminated and hence power consumption minimized, by observing the communication between VMs and by placing them on the same or closely located nodes [16].

In the aim to provide energy efficient model and power consumption model of the cloud environment will be developed and the cost of data transfer depending on the traffic volume will be estimated. As migrations consume additional energy and they have a negative impact on the

performance [43], before initiating the migration, the reallocation controller has to ensure that the cost of migration does not exceed the benefit. The topology directly affects virtualization, in that sense, in all attempts to save energy, all the services must remain rendered in their quality.

2.8.4 Autonomic Optimization of Thermal states and Cooling System Operation

A significant part of electrical energy consumed by computing resources is transformed into heat [9]. High temperature leads to a number of problems, such as reduced system reliability and availability, as well as decreased lifetime of devices [22]. In order to keep the system components within their safe operating temperature and prevent failures and crashes, the emitted heat must be dissipated. This is achieved by cooling and ventilation via all the different methods that have been developed over the years [27].

The cooling problem becomes extremely important for modern blade and 1-unit rack servers; the size and structure of these servers leads to the high density of computing resources and complicate heat dissipation [22]. For example, in a 30,000 ft^2 data center with 1000 standard computing racks, each consuming 10kW, the initial cost of purchasing and installing the infrastructure is \$2-\$5 million; whereas the annual costs for cooling is around \$4-\$8 million [27]. Therefore, apart from the hardware improvements, it is essential to optimize the cooling system operation from the software side.

There has been a lot of work done on modeling thermal topologies of data centers; such that it can lead to more efficient workload placement [40]. The new challenges include how and when to reallocate VMs in order to minimize the power drawn by the cooling system while preserving safe temperatures of the resources and minimizing migration overhead and performance degradation [33, 45]. Allocation and re-allocation changes must be directly sensitive to the cooling implications, per data center and specific topology [19].

2.8.5 Efficient Consolidation of VMs for Managing Heterogeneous Workloads

Cloud infrastructure services provide users with the ability to provision virtual machines and allocate any kind of applications on them [15]. This leads to the fact that different types of applications (e.g., enterprise, scientific, and social network applications) can be allocated on one physical computer node. However, it is not obvious how these applications can influence each other, as they can be data, network or computationally intensive thus creating variable or static load on the resources [47]. The problem is to determine what kind of applications can be allocated to a single host that will provide the most efficient overall usage of the resources. There are application types that function more optimally when operated from the same host.

Current approaches to energy efficient consolidation of VMs in data centers do not investigate the problem of combining different types of workload. These approaches usually focus on one particular workload type or do not consider different kinds of applications; it is assumed that although they are different applications, they consume uniform workloads.

A defragmentation algorithm to optimize performance after VM migration with different workload types is proposed in this research. A computationally intensive (scientific) application can be effectively combined with a web-application (file server), as the former mostly relies on CPU performance, whereas the latter utilizes disk storage and network bandwidth. In this sense both applications, although sharing a host, do not get in each other's way, saving more resources.

In this research the particular kind of applications that can be effectively combined and what parameters influence the efficiency will be investigated, and an optimization algorithm for managing them per heterogeneous system created will be developed. This knowledge can be applied to energy efficient resource management strategies in data centers in order to achieve more optimal allocation of resources and, therefore, improve the utilization of resources and reduce energy consumption.

For the resource providers, optimal allocation of VMs will result in higher utilization of resources and, therefore, reduced operational costs. Cloud users will benefit from decreased prices for the resource usage as operational costs are reduced.

2.9 Chapter 2 Summary

This chapter is a review of existing and related literature in energy efficiency and optimization in cloud computing environments. It also explains the key concepts and terminologies as used in this project. In addition, virtual machines which form a fundamental technology for the rising demand in computational power are also explained. Data center productivity with regard to energy consumption metric and an overview of energy of related work have also been explained, to buttress the direction of this research.

Chapter 3
Model, Algorithm and Metric Development

3.1 Chapter Overview

In this chapter the developed model, algorithm and metric are presented, giving in each case a full explanation of its components and variables. In addition the chapter will give details of the experimental set up and also demonstrate how the model, algorithm and metric work together to archive the main goal of this research.

3.2 Energy Efficiency (Power Model)

One of the objectives to achieve the main goal for this research, was to develop an energy efficiency model. A linear relationship between CPU utilization and electrical power is assumed for this model. For example, say for a given job j1, information of the processing time and the processor utilization is enough to calculate its power consumption. We define the consumption of a resource r_i at any given time as:

$$C_{\mathrm{i}} = \sum_{j=1}^{\mathrm{n}} c_{\mathrm{i,j}} \qquad (4)$$

Where n = number of task running at that time and $c_{i,j}$ is the resource usage of job j_j.

We also define energy usage, P_i, of a resource at any time as:

$$P_{\mathrm{i}} = (P_{\mathrm{max}} - P_{\mathrm{min}}) * C_{\mathrm{i}} + P_{\mathrm{min}} \qquad (5)$$

Where P_{max} refers to the peak load consumed energy and P_{min} refers to active mode minimum energy consumption usually as low as one (1%) percent.

3.3 Optimization Algorithm

Another objective for this research is to develop energy optimization algorithm. The development of this algorithm requires the development of two different algorithms to effectively reduce the rate of energy consumption in the data centers. These two algorithms are load balancing virtual machine aware algorithm and defragmentation algorithm.

3.3.1 Load Balancing Virtual Machine Aware Algorithm

Load balancing is the process of taking complex or large work load that needs a lot of processing power to be processed and dividing it into modules then distributing it to different machines or nodes for processing. In so doing, the processing time and processing power are reduced. For example, taking a large mathematical equation and using a distributed system to compute it just like in Grid computing. In cloud computing, the process is the same, but the only difference is that the process is done on a virtual plain, which is at virtual machine management or hypervisor level.

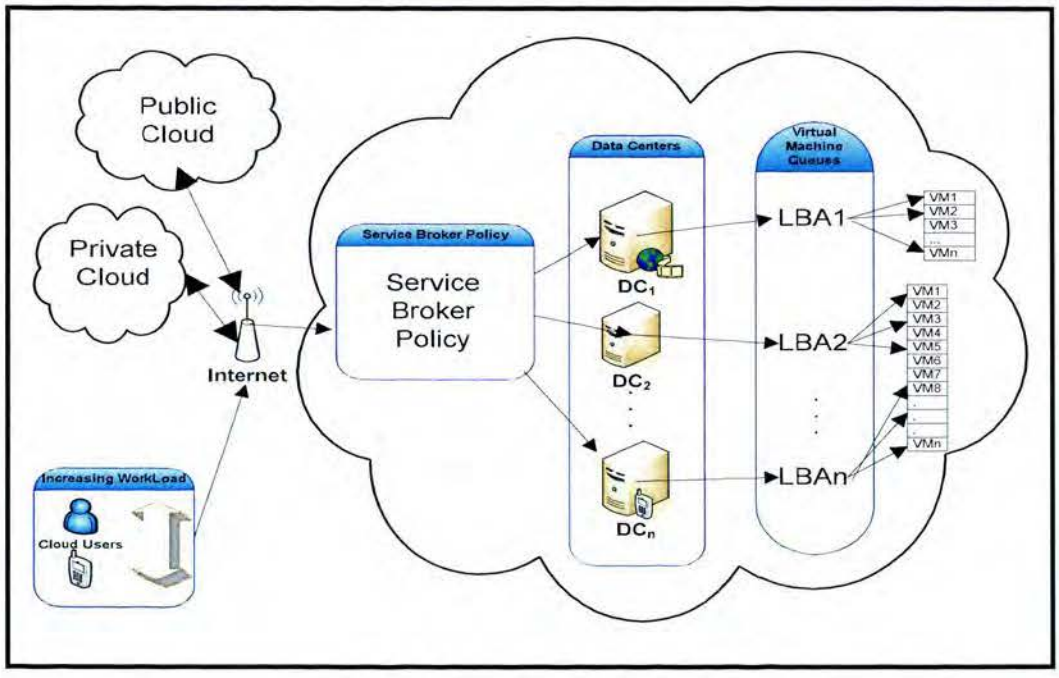

Figure 3.1 LBVMA model

In this work the load balancing algorithm that is more energy efficient in a virtual machine management level is determined. Virtual machine management level is referred to as virtual machine queues in the designed model as can be seen in Figure 3.1 (Load Balancing Virtual Machine Aware, LBVMA Model).

There are different types of clouds in the Internet, as cloud computing is using computing resources over a network as a service. Hence, the LBVMA model shows how different clouds and cloud users interact through the Internet. More importantly, the proposed cloud configuration includes a virtual machine queue which manages the distribution of loads to different virtual machines.

Increasing workload in the model is the cloud customers requesting for services in the cloud through the use of the underlying network, the Internet. Public and private clouds behave the same way as they do as discussed in section 2.3, they interact with other clouds and cloud customers through the Internet as well. In cloud configuration in this dissertation there is a Service Broker Policy which ensures that the cloud customers and cloud providers understand SLAs among themselves before they proceed with the transaction of money and services. All the services, whether they be SaaS, PaaS or IaaS, are hosted in data centers, which is what follows the Service Broker Policy. Data centers have servers that host the services and most importantly the virtual machines. So, just before distributing the cloud customer requests to the virtual machines there are virtual machine queues that are managed by load balancers. The load balancers use different load balancing algorithms to distribute work loads to the virtual machines.

3.3.2 Defragmentation Algorithm

A server is a computer that manages centrally stored data or network communication resources. A server also provides and organizes access to these resources for other computers linked to it. Storage devices in most servers are a collection of hard drives or hard discs, and as people or nodes connected to the server in the network insert and delete into the server, the hard drives get fragmented. This fragmentation causes processing speed to be slow because the Central Processing Unit (CPU) collects and stores its data and information in the hard drive or Random Access Memory (RAM). The reason for this is that the hard drive, after fragmentation, has variable

distance between data or information stored in it in terms of space in bytes. Defragmentation is a process of removing this distance and bringing information in the hard drive closer together, and in so doing making processing time faster because searching time is reduced. Virtual machine migration is when virtual machines are moved from one physical machine to another. This migration also causes fragmentation. So to solve this problem a defragmentation algorithm is proposed as in Figure 3.2. The algorithm should be active after virtual machine migration, just before the server turns off, to optimize processing time and energy consumption.

```
1.      Quicksort(HD,p,t) {
2.       if (p < t) {
3.        q <- Partition(HD,p,t)
4.        Quicksort(HD,p,q)
5.        Quicksort(HD,q+1,t)
   }
}


6.      Partition(HD,p,t)
7.       x <- HD [p]
8.       i <- p-1
9.       j <- r+1
10.      while (True) {
11.      repeat
12.           j <- j-1
13.       until (HD [j] <= x)
14.       repeat
15.          i <- i+1
16.       until (HD [i] >= x)
17.      if (i<-""> A[j]
18.       else
19.          return(j)
  }}
```

Figure 3.2 Defragmentation algorithm

The algorithm is based on the assumption that the hard drives found in servers of the data centers are designed like an array. So, basically the Defragmentation Algorithm is a quicksort algorithm. The reason for choosing the quicksort algorithm is that it has a worst case performance of $O(n^2)$ and a best case performance of $O(n \log n)$ in terms of time complexity. From line one to line five

(in Figure 3.2) is the main function of the quicksort algorithm, which contains a pivoting element and three functions, one to sort the left hand side of the pivot and the other to sort the right hand side of the pivot. The last function is the function that will reposition the pivot during the sorting process. Lines 6 to 14 show how and in what conditions the function that will reposition the pivoting element will do so.

Input: HD fragmented array HD, a pivot element p and a traversing element t.

Output: HD defragmented array HD.

3.4 Efficient Energy Usage (EEU)

The total energy consumed in cloud data center together with other necessary resources needed for service in a cloud data center are what we referred to as EEU.

$$EEU = \frac{\sum_{n=1}^{t} \text{Yearly Consumption}}{\text{Yearly Consumption}} \quad (6)$$

$$1 <= \text{EEU} <= \infty \quad (7)$$

EEU can be varied from 1 to ∞ as seen in (7). This means that every data center with an EEU value of 2 or more shows that for every kWh consumed by the server, one kWh is consumed by the supplementary services like the cooling system of the data center, the lighting system and other supporting systems.

3.5 Experimental Setup

For the purpose of the experiments in this research, CloudAnalyst was used to simulate the data. Cloud Analyst is a simulation tool designed to simulate real cloud environments and scenarios. It is built on CloudSim designed on a java programming language and iText 2.1.5. On the other hand, cloud analyst has all the capabilities of CloudSim with a user friendly Graphic User Interface (GUI) [28, 54]. The experiment was run on a machine with a core *i*5 intel processor and 4Gig RAM. The experimental design map is as shown in Figure 3.3.

(in Figure 3.2) is the main function of the quicksort algorithm, which contains a pivoting element and three functions, one to sort the left hand side of the pivot and the other to sort the right hand side of the pivot. The last function is the function that will reposition the pivot during the sorting process. Lines 6 to 14 show how and in what conditions the function that will reposition the pivoting element will do so.

Input: HD fragmented array HD, a pivot element p and a traversing element t.

Output: HD defragmented array HD.

3.4 Efficient Energy Usage (EEU)

The total energy consumed in cloud data center together with other necessary resources needed for service in a cloud data center are what we referred to as EEU.

$$EEU = \frac{\sum_{n=1}^{i} \text{Yearly Consumption}}{\text{Yearly Consumption}} \quad (6)$$

$$1 <= \text{EEU} <= \infty \quad (7)$$

EEU can be varied from 1 to ∞ as seen in (7). This means that every data center with an EEU value of 2 or more shows that for every kWh consumed by the server, one kWh is consumed by the supplementary services like the cooling system of the data center, the lighting system and other supporting systems.

3.5 Experimental Setup

For the purpose of the experiments in this research, CloudAnalyst was used to simulate the data. Cloud Analyst is a simulation tool designed to simulate real cloud environments and scenarios. It is built on CloudSim designed on a java programming language and iText 2.1.5. On the other hand, cloud analyst has all the capabilities of CloudSim with a user friendly Graphic User Interface (GUI) [28, 54]. The experiment was run on a machine with a core *i*5 intel processor and 4Gig RAM. The experimental design map is as shown in Figure 3.3.

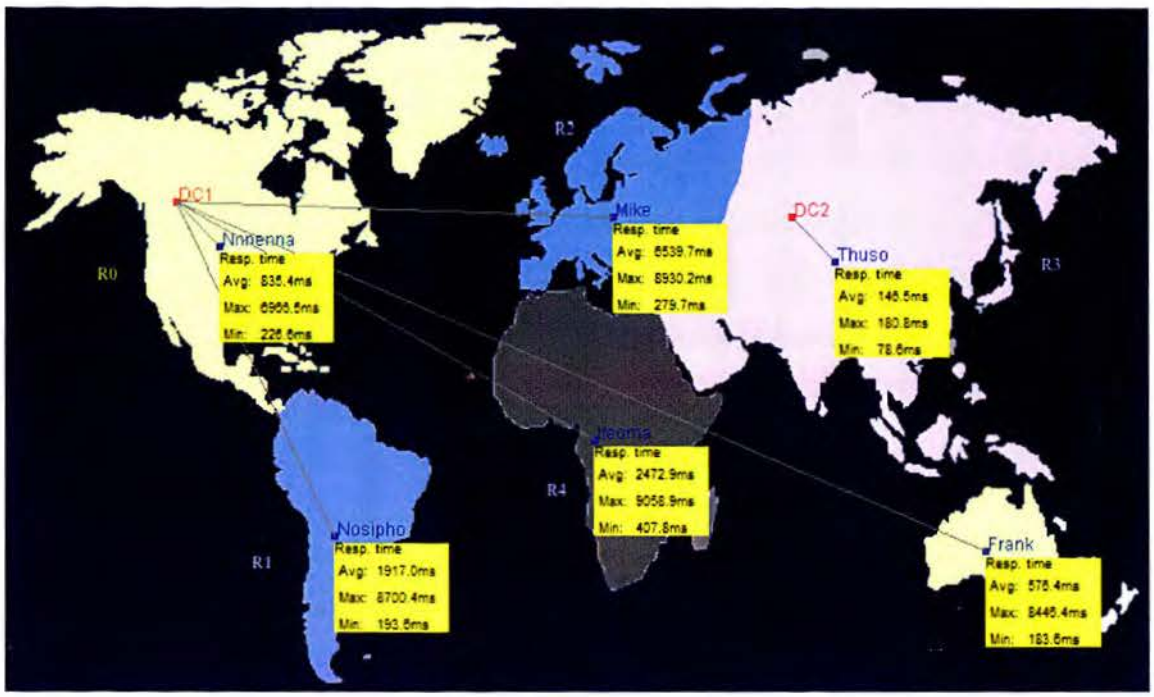

Figure 3.3 Map showing region

Six geographically located user bases were created and two data centers as shown in Tables 3.1 and 3.2, respectfully.

Table 3.1 User bases

Name	Region	Requests per User per Hr	Data Size per Request (bytes)	Peak Hours Start (GMT)	Peak Hours End (GMT)	Avg Peak Users	Avg Off-Peak Users
Ifeoma	4	20	150	15	20	300000	80000
Mike	2	20	150	12	24	450000	300000
Nnnenna	0	20	150	14	18	250000	20000
Nosipho	1	20	150	16	20	500000	60000
Thuso	3	20	150	13	22	350000	5000

Shown in Table 3.1 is a collection of input parameters that form parts of managing user bases. These input parameters are manipulated to produce the outcome. The output results obtained from these input parameters are reported and discussed in Table 4.1 of chapter 4.

Table 3.2 Data centers

Data Center	# VMs	Image Size	Memory	BW
DC1	5	10000	512	1000
DC2	5	10000	512	1000

Shown in Table 3.2 is a collection of input parameters that form parts of managing user bases and these input parameters are manipulated to produce the outcome. The output results obtained from these input parameters are reported and discussed in Table 4.2 of chapter 4.

As important information, for each data center, the physical machine uses x86 architecture while running in a Linux operating system and Xen virtual machine manager. Each physical machine has four processors and their speed is 10 000Hz. Tables 3.3 and 3.4 show the characteristics of the delay and bandwidth matrix.

Table 3.3 Delay matrix

Region\Region	0	1	2	3	4	5
0	25	100	150	250	250	100
1	100	25	250	500	350	200
2	150	250	25	150	150	200
3	250	500	150	25	500	500
4	250	350	150	500	25	500
5	100	200	200	500	500	25

Table 3.3 shows the delay matrix between the different geographically located regions as shown in Figure 3.1. At different regions, the inputs of delay data to be transmitted were varied. The varied inputs are as shown in Table 3.3.

Table 3.4 Bandwidth matrix

Region\Region	0	1	2	3	4	5
0	2,000	1,000	1,000	1,000	1,000	1,000
1	1,000	800	1,000	1,000	1,000	1,000
2	1,000	1,000	2,500	1,000	1,000	1,000
3	1,000	1,000	1,000	1,500	1,000	1,000
4	1,000	1,000	1,000	1,000	500	1,000
5	1,000	1,000	1,000	1,000	1,000	2,000

Table 3.4 shows the inputs bandwidth matrix between the different geographically located regions as Figure 3.1. At different regions, the input of bandwidth data to be transmitted was varied. The varied inputs are as shown in Table 3.4.

3.6 Chapter Summary

This chapter discussed the developed algorithm, power model, proposed efficient energy usage metric, load balancing model and also explained in detail the experimental set up. The next chapter examines the experimental design presented in this chapter, and discusses the detailed results obtained from the experiments conducted.

Chapter 4
Results and discussions

4.1 Introduction

This chapter presents a detailed discussions and interpretations of the results obtained from the experimental set-up reported in chapter 3 of the research work carried out.

4.2 Experimental Results

Using the develop energy efficiency model and the various inputs from the user bases, the delay and bandwidth matrix, the simulator generates the user base response time from the two data centers. The generated result is as shown in Table 4.1. The output contains the minimum, maximum and average user base time response in terms of mille seconds.

Table 4.1 User base response time

User Bases	Min (ms)	Max (ms)	Avg (ms)
Frank	172.83	1718.89	461.95
Ifeoma	401.4	7688.57	1526.57
Mike	254.16	7304.98	2464.76
Nnnenna	49.23	736.05	292.74
Nosipho	160.75	2453.54	848.1
Thuso	50.43	5461.83	370.14

The output produced after running the configuration in Tables 3.1, 3.2, 3.3 and 3.4 in terms of response time is shown in Table 4.2.

Table 4.2 Data center response time

Data Centers	Min (ms)	Max (ms)	Avg (ms)
DC1	0.11	7076.86	1831.7
DC2	8.71	3295.29	1106.04

4.3 Discussion

When the amount of data processed at user base level (as shown in Table 3.1) is compared to the amount of user base response time shown in Table 4.1 and Figure 4.1, it is evident that quality of service in terms of response time is much better for data centers that have more physical machines.

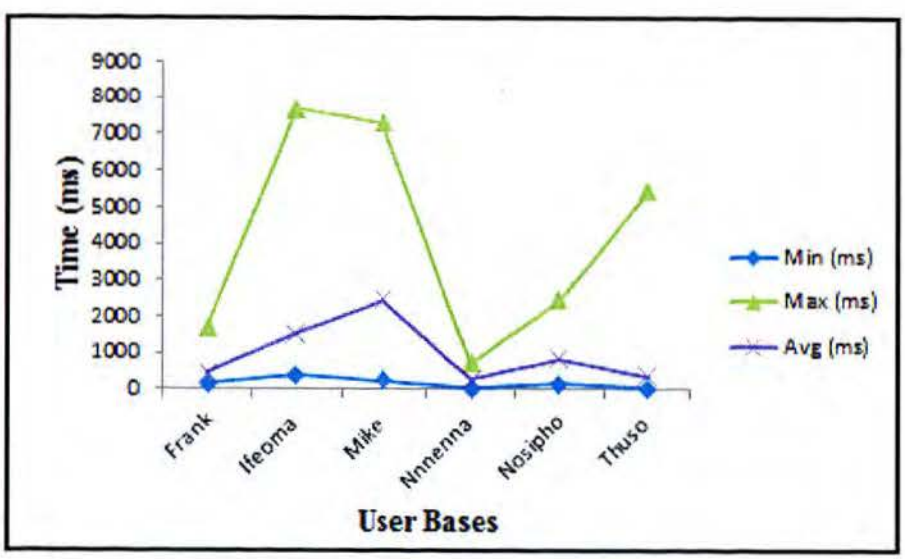

Figure 4.1 User base response time

Figure 4.1 shows a correlation of user bases and their corresponding response time, which is used to determine the minimum, maximum and average response time for the two (DC1, DC2) data

centers reported in Table 4.2. With the results obtained, it was possible to graphically show the minimum, maximum and average response time for the data centers as shown in Figure 4.2.

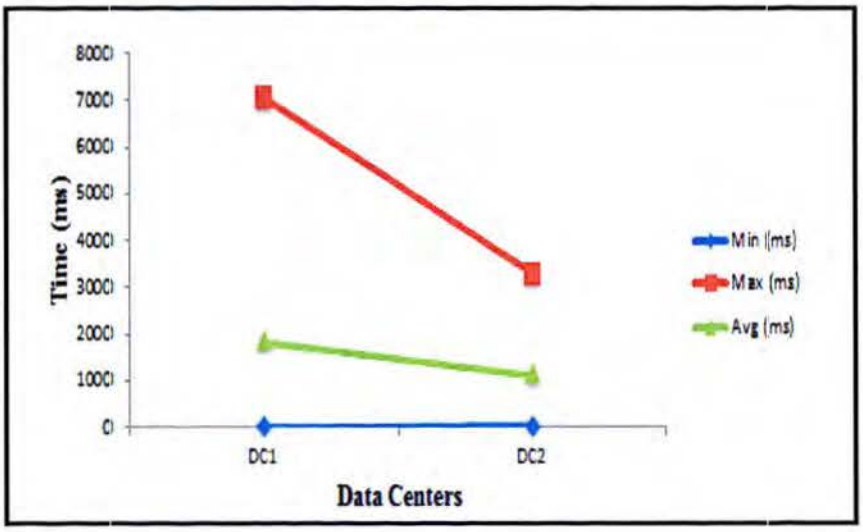

Figure 4.2 Data center response times

Using the same configuration in Tables 4.1 and 4.3, Figure 4.3 shows a comparison of energy consumed by a data center facility using the LBVMA Model, where the load balancing policy is throttled and service broker policy is set to optimal time response. Therefore, Figure 4.3 shows a correlation between execution time and the energy consumed. From the LBVMA model and the chosen parameters, it can be seen that energy consumption is less.

Figure 4.4 shows power consumption due to different memory configurations. The memory configurations are 1000 MHz, 1200 MHz and 1500 MHz respectfully. The X-axis shows time in milliseconds and the Y-axis shows the power in kilo Watts. From Figure 4.4 it is clear that memory configuration at higher frequencies consumes more power.

In section 3.3 it was mentioned that the load balancing algorithm that is more energy efficient in virtual machine management level would be determined in this research. This next experiment shows these results. The same configuration in Tables 4.1 and 4.3 were used for all algorithms.

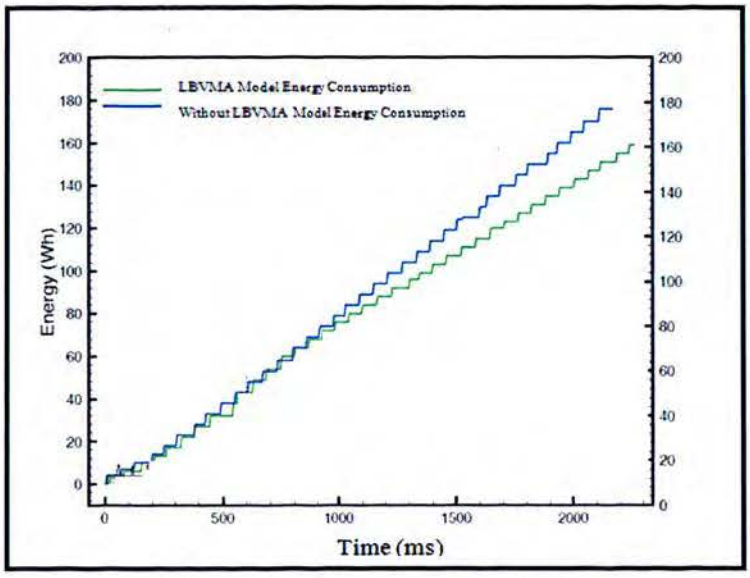

Figure 4.3 Consumed energy for power model

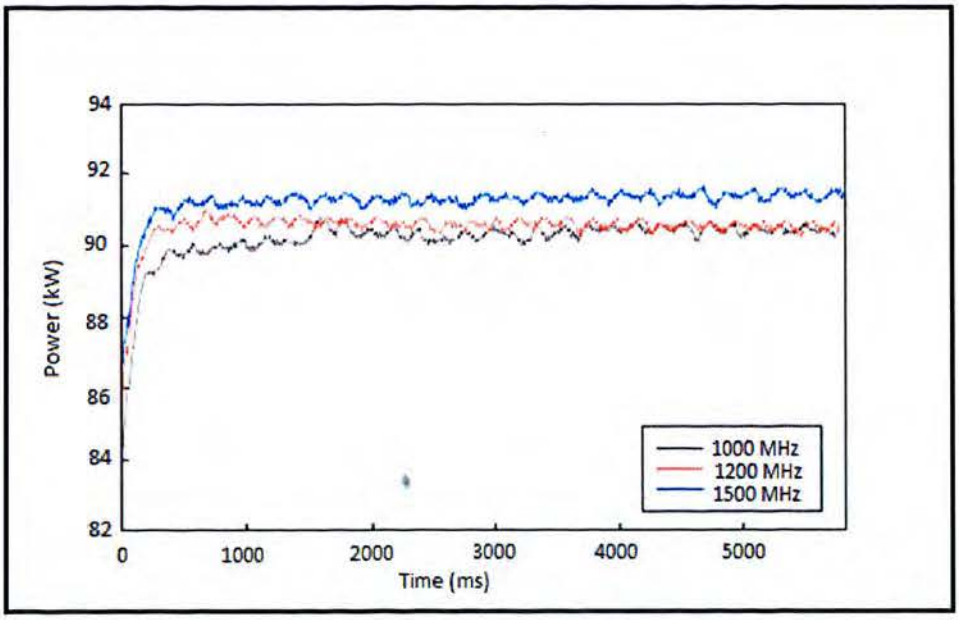

Figure 4.4 Power consumed by memory

Table 4.3 Overall time response

	Round Robin	Equally Spread Current Execution Load	Throttled
Avg (ms)	3739.52	3613.4	1996.66
Min (ms)	75.28	72.77	49.23
Max (ms)	7673.36	7721.21	7688.57

Table 4.3 shows overall response time of requests processed from user bases to data centers and vice versa. Figure 4.5 shows a graphical representation of the data in Table 4.3 and presents a correlation of the load balancing algorithm with respect to response time represented in milliseconds. From Figure 4.5 it is clear that the throttled load balancing algorithm performs better in terms of response time.

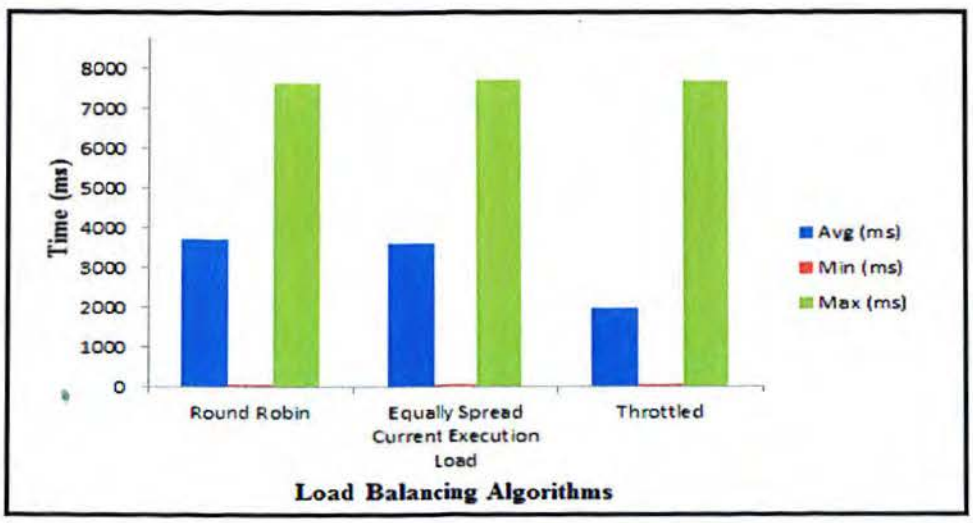

Figure 4.5 Overall time response

4.4 Chapter Summary

This chapter presents a discussion and interpretation of the experimental results obtained in Chapter 3. The results show how the models, metric and algorithm work together to enhance an on-demand resource delivery in a cloud computing environment.

Chapter 5
Summary, Conclusions and Future Work

5.1 Summary

The service model of cloud computing involving the provision of large pools of high performance computing resources and high capacity storage devices by a service provider that are shared among end users as required has gained considerable attention. Although there exist many cloud service models, end users subscribing to the service have their data hosted by the service, and have computing resources allocated on demand from the pool. The service provider's offering may also extend to the software applications required by the end user. To be successful, the cloud service model also requires a high-speed network to provide connection between the end user and the service provider's infrastructure.

As shown in this research, cloud computing potentially offers an overall financial benefit where end users share a large, centrally managed pool of storage and computing resources, rather than owning and managing their own systems. Often using existing data centers as a basis, cloud service providers invest in the necessary infrastructure and management systems, and in return receive a time-based or usage-based fee from end users. The end user in turn sees convenience benefits from having data and services available from any location, from having data backups centrally managed, from the availability of increased capacity when needed, and from usage-based charging. The last point is important for many users in that it averts the need for a large one-off investment in hardware, sized to suit maximum demand, and requiring upgrading every few years.

In summary, while noting the financial benefits, the shift in energy usage in a cloud computing model has received little attention. Through the use of large shared servers and storage units, cloud computing can offer energy savings in the provision of computing and storage services, particularly if the end user migrates toward the use of a computer or a terminal of lower capability and lower energy consumption. At the same time, cloud computing leads to increases in network traffic and the associated network energy consumption. In this research dissertation, the balance

between server energy consumption, network energy consumption, and end-user energy consumption was explored, to present a fuller assessment of the benefits of cloud computing.

The issue of energy consumption in information technology equipment has been receiving increasing attention in recent years and there is growing recognition of the need to manage energy consumption across the entire information and communications technology (ICT) sector. The management of power consumption in data centers has led to a number of substantial improvements in energy efficiency. Cloud computing infrastructure is housed in data centers and has benefited significantly from these advances. Techniques such as sleep scheduling and virtualization of computing resources in cloud computing data centers, for example, improve the energy efficiency of cloud computing.

From the outcome of the research reported in this dissertation, one can see that the level of utilization achieved by a cloud service is a function of the type of services it provides, the number of users it serves, and the usage patterns of those users. Large-scale public clouds that serve a very large number of users are expected to be able to fully benefit from achieving high levels of utilization and high levels of virtualization, leading to low per-user energy consumption. Private clouds that serve a relatively small number of users may not have sufficient scale to fully benefit from the same energy-saving techniques. This analysis is based on the view that cloud computing fully utilizes servers and storage for both public and private clouds. However, it is not clear whether in general the energy consumption saving during data migration with a cloud offsets the higher energy consumption due to lower utilization of servers and storage. But what is very clear as achieved from this research is the fact that the logical unification of several geographically diverse data centers assists cloud computing to scale during periods of high demand. However, energy-efficient data migration between these data centers is necessary to ensure that cloud computing is energy efficient. In this analysis, public clouds consumed more energy than private clouds because users are connected to the public cloud through the public Internet.

5.2 Conclusion

In conclusion, this research dissertation reports is the development of a power or energy efficiency model, a load balancing virtual machine aware (LBVMA) model and an efficient energy usage (EEU) metric to enhance the calculation of power consumption in data centers and to determine whether or not the energy used in a data center is used efficiently. Additionally, the development and implementation of a defragmentation algorithm as an optimization algorithm to optimize processing time in cloud data centers after virtual machine migration is reported.

Results obtained during the experimental set-up to implement the models and algorithms developed indicate that, the quality of service in terms of response time is much better for data centers that have more physical machines than for those with less machines, but there was an observable higher energy consumption for memory configuration with higher frequencies.

Consequently, it is clear that the level of energy utilization achieved by a cloud service is a function of the type of services it provides, the number of users served, and the usage patterns of the users. What is not significantly clear is whether the energy consumption saving during virtual machine migration with a private cloud offsets the higher energy consumption due to lower utilization of servers and storage.

5.3 Future Work

While it is important to understand how to minimize energy consumption in data centers that host cloud computing services, it is also important to consider the energy required to migrate data to and from the end user and the energy consumed by the end-user interface. The research studies of energy consumption in cloud computing have focused only on the energy consumed in the data center. However, to obtain a clear picture of the total energy consumption of a cloud computing service, and understand the potential role of cloud computing to provide energy savings, a more comprehensive analysis is required. To minimize the energy consumption in transport, cloud computing data centers should be connected through dedicated point-to-point links incorporating optical bypass where possible. Indeed, reducing the number of routings hops and transmission

links would yield benefits to all services. This would serve as another angle of consideration by related research as future work.

In addition, developing a tool to monitor activities at virtual machine level that would enhance energy optimization processes is also another planned future activity.

References

[1] J. H. Sienkiewicz. (2009, Apr.) Cloud Computing: A Perspective. Workshop lecture.

[2] J. Hermans and M. Chung. (2012, Mar.) KPMG's 2010 Cloud Computing Survey. PDF.

[3] J. Lin. (2008, Sep.) What is Cloud Computing? Class lecture Presentation.

[4] K.E.U. Ahamed and V. Alexandrov, "Identity and Access Management in Cloud Computing," *Cloud Computing for Enterprise Architectures*, pp. 115-133, 2011.

[5] T. Wo and J.Li Y. Chen, "An Efficient Resource Management System for On-Line Virtual Cluster Provision," in *IEEE International Conference* , 2009, pp. 72-79.

[6] Y. Sato and Y. Inoguchi T. V. T. Duy, "Performance evaluation of a green scheduling algorithm for energy savings in cloud computing," in *Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium*, 2010, pp. 1-8.

[7] A. Beloglazov and R. Buyya, "Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Vertual Machines in Cloud Data Centers," in *8th International Workshop on Middleware for Grids, Cloud and e-Science*, 2010, p. 4.

[8] J. Justice, M. Krampits, M. Letts, G. Subramanian and M. Thirumalai T. Daim, "Data Center Metrics: An Energy Efficiency Model for Information technology managers," *Management of Environmental Quality: An International Journal*, vol. 20, pp. 712-731, 2009.

[9] V. Mohan Raj and R. Shriram, "Power aware provisioning in cloud computing environment," in *ICCCET*, 2011, pp. 6-11.

[10] X. Wang and Y. Wang, "Energy-efficient multi-task scheduling based on MapReduce for cloud computing," in *Seventh International Conference on Computational Intelligence and Security (CIS)*, 2011, pp. 57-62.

[11] J. Cidell, "A Political Ecology of the Build Environment: LEED Certification for Green Buildings," *Local Environ.*, vol. 14, pp. 612-633, 2009.

[12] Yong Zhao, Ioan Raicu, Shiyong Lu Ian Foster, "Cloud Computing and Grid Computing 360-Degree Compared," , Chicago, IL, USA, 2008, pp. 1-10.

[13] Y. Zhao, I. Raicu and S. Lu I. Foster, "Cloud computing and grid computing 360-degree compared," in *Grid Computing Environments Workshop*, 2008, pp. 1-10.

[14] R. Buyya, Y. C. Lee and A. Zomaya A. Beloglazov, "A taxonomy and survey of energy-efficient data centers and cloud computing systems," in *Advances in Computers*, 2011, pp. 47-111.

[15] A. Beloglazov and R. Buyya, "Energy efficient allocation of virtual machines in cloud data centers," in *10th IEEE/ACM International Conference*, 2010, pp. 577-578.

[16] C. S. Yeo, S. Venugopal, J. Broberg and I. Brandic R. Buyya, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Comput. Syst*, vol. 25, pp. 599-616, 2009.

[17] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," in *Concurrency and Computation: Practice and Experience*, 2011.

[18] M. Zbakh, H. El Bakkali and D. El Kettani A. Khiyaita, "Load balancing cloud computing: State of art," in *Network Security and Systems (JNS2)*, 2012, pp. 106-109.

[19] J. E. Taylor and J. A. Siegel G. Peschiera, "Response–relapse patterns of building occupant electricity consumption following exposure to personal, contextualized and occupant peer network utilization data," *Energy Build*, vol. 42, pp. 1329-1336, 2010.

[20] C. J. Tang and M. R. Dai, "Dynamic computing resource adjustment for enhancing energy efficiency of cloud service data centers," in *System Integration (SII*, 2011, pp. 1159-1164.

[21] R. Nathuji and K. Schwan, "VirtualPower: coordinated power management in virtualized enterprise systems," vol. 41, pp. 265-278, 2007.

[22] Y. Xing and Y. Zhan, "Virtualization and Cloud Computing," in *Future Wireless Networks and Information Systems*, 2012, pp. 305-312.

[23] Weber W.D., and Barroso L.A. Fan X., "Power provisioning for a warehouse-sized computer," in *34th Annual International Symposium on Computer Architecture (ISCA 2007)*, New York, 2007, pp. 13-23.

[24] Gelas JP, Lefevre L, and Orgerie A de Assunao MD, "The Green Grid'5000: Instrumenting and using a Grid with energy sensors," in *5th International Workshop on Distributed Cooperative Laboratories: Instrumenting the Grid (INGRID 2010)*, Poznan, 2010.

[25] Leech P., Irwin D., and Chase J. Ranganathan P., "Ensemble-level power management for dense blade servers," in *33rd International Symposium on Computer Architecture (ISCA 2006)*, Boston, 2006, pp. 66-77.

[26] R. W. A. Ayre, K. Hinton and R. S. Tucker , "Green cloud computing: Balancing energy in processing, storage, and transport," in *Proc IEEE*, 2011, pp. 149-167.

[27] B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt and A. Warfield P. Barham, "Xen and the art of virtualization," *The case for energy-proportional computing*, vol. 37, pp. 164-177, 2003.

[28] I. S. Moreno and J. Xu, "Energy-efficiency in cloud computing environments: towards energy savings without Performance degradation," *International Journal of Cloud Applications and Computing (IJCAC)*, vol. 1, pp. 17-33, 2011.

[29] W. Zeng, M. Liu and G. Li J. Zhao, "Multi-objective optimization model of virtual resources scheduling under cloud computing and it's solution," in *Cloud and Service Computing (CSC)*, 2011, pp. 185-190.

[30] R. N. Calheiros and R. Buyya B. Wickremasinghe, "Cloud Analyst: A CloudSim-Based Visual Modeller For Analysing Cloud Computing Environments and Applications," in *dvanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference*, 2010, pp. 446-452.

[31] G. Tan, X. Zhang and J. Zhou L. Xu, "Energy aware cloud application management in private cloud data center," in *International Conference on Cloud and Service Computing (CSC)*, 2011, pp. 274-279.

[32] N. A. Singh and M. Hemalatha, "High performance computing network for cloud environment using simulators," in *ArXiv Preprint arXiv:1203.1728*, 2012.

[33] R. Buyya and M. Murshed, "Gridsim: A toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing," *Concurrency and Computation: Practice and Experience*, vol. 14, pp. 1175-1220, 2003.

[34] R. Ranjan, A. Beloglazov, C. A. F. De Rose and R. Buyya R. N. Calheiros, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and Experience*, vol. 41, pp. 23-50, 2011.

[35] B. Wickremasinghe, "CloudAnalyst: A CloudSim-based Tool for Modelling and Analysis of Large Scale Cloud Computing Environments," MEDC Project Report, 2009.

[36] C. Maciocco, T. Y. C. Tai, R. Yavatkar, L. K. Lu and A. W. Min R. Wang, "DirectPath: High performance and energy efficient platform I/O architecture for content intensive usages," in *Proceedings of the 3rd International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet*, 2012, p. 3.

[37] DRBD Software Development for High Availability Clusters. (2012, Nov.) DRBD. [Online]. http://www.drbd.org/

[38] Matthias Schmidt, Christian Strack, Simon Martin and Bernd Freisleben Roland Schwarzkopf, "Increasing virtual machine security in cloud environments," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 12, no. 1, pp. 1-12, 2012.

[39] S. S. Waraich, "Classification of dynamic load balancing strategies in a network of workstations," in *IEEE Fifth International Conference on Information Technology: New Generations*, 2008, pp. 1263-1265.

[40] Z. Zhang and S. Fu, "Macropower: A coarse-grain power profiling framework for energy-efficient cloud computing," *30th IPCCC, IEEE*, pp. 1-8, 2011.

[41] A. Beloglazov and J. Abawajy R. Buyya, "Energy-Efficient Management of Data Center Resources for Cloud Computing: Avision, Architectural Elements and Open Challenges,"

Arxiv Preprint arXiv: 1006.0308, 2010.

[42] W. Mach and E. Schikuta, "A consumer-provider cloud cost model considering variable cost," in *EEE Ninth International Conference*, 2011, pp. 628-635.

[43] I. S. Moreno and J. Xu, "Customer-aware resource overallocation to improve energy efficiency in realtime Cloud Computing data centers," 2011.

[44] W. Xiaoli and L. Zhanghui, "An energy-aware VMs placement algorithm in cloud computing environment," *ISDEA*, pp. 627-630, 2012.

[45] F. Lombardi and R. Di Pietro, "Secure virtualization for cloud computing," *Journal of Network and Computer Applications*, vol. 34, pp. 1113-1122, 2011.

[46] J. Li, J. Huai, T. Wo, Q. Li and L. Zhong B. Li, "Enacloud: An energy-saving application live placement approach for cloud computing environments," in *CLOUD'09. IEEE International Conferenc*, 2009, pp. 17-24.

[47] E. Gelenbe, M. Di Girolamo, G. Giuliani, H. De Meer, M. Q. Dang and K. Pentikousis A. Berl, "Energy-efficient cloud computing," *The Computer Journal*, vol. 53, pp. 1045-1051, 2010.

[48] X. Lu and Z. Gu, "A load-adapative cloud resource scheduling model based on ant colony algorithm," in *Cloud Computing and Intelligence Systems (CCIS)*, 2011, pp. 296-300.

[49] C. Ghali, A. Chehab and A. Kayssi I. Sarji, "CloudESE: Energy efficiency model for cloud computing environments," in *International Conference on Energy Aware Computing (ICEAC)*, 2011, pp. 1-6.

[50] B. Kelley, J. Prevost and M. Jamshidi K. M. Nagothu, "Ultra low energy cloud computing using adaptive load prediction," in *World Automation Congress (WAC)*, 2011, pp. 1-7.

[51] S. K. Garg and R. Buyya L. Wu, "SLA-based resource allocation for software as a service provider (SaaS) in cloud computing environments," in *Cluster, Cloud and Grid Computing (CCGrid)*, 2011, pp. 195-204.

[52] University of Melbourne. (2012, May) Melbourne Clouds Lab. [Online]. http://www.cloudbus.org/cloudsim/

[53] B. Pfaff, J. Chow, M. Rosenblum and D. Boneh T. Garfinkel, "Terra: A virtual machine-based platform for trusted computing," in *ACM SIGOPS Operating Systems Review*, 2003, pp. 193-206.

[54] A. Kansal and F. Zhao, "Fine-grained energy profiling for power-aware application design," vol. 36, pp. 26-31, 2008.