Energy-Aware Network ResourceAllocation in a Cloud Environment

By

Vahan Artine Yoghourdjian

Thesis submitted to

Notre Dame University

to fulfill the requirements of the degree of

Masters in Computer Science

Examining Committee:

Dr. Rosy Aoun	Advisor
Dr. Hoda Maalouf	Examiner, Chairperson
Dr. Maya Samaha	Examiner

Lebanon, June 2013

This page is intentionally left blank

Energy-aware network resource allocation in a Cloud environment

By

Vahan Yoghourdjian

Approved by:

Rosy Aoun: Assistant Professor of Computer Science
Advisor.

H Maalouf

Hoda Maalouf: Associate Professor of Computer Science
Member of Committee.

Maya Samaha: Assistant Professor of Computer Science
Member of Committee.

Date of Thesis Defense: June 26, 2013

This page is intentionally left blank

I thank my professor and advisor Dr. Rosy Aoun for her support and devotion. I thank Joseph Sahakian for his support and assistance. I thank my brother Hrag Yoghourdjian for his assistance. Most importantly, I thank my parents: SonaTchaparian-Yoghourdjian and Harout Yoghourdjian for their undying love and support.

This page is intentionally left blank

Energy-Aware Network ResourceAllocation in a Cloud Environment

Abstract

In the world of computing, one major statement has always been true: the request for computing power is always greater than the available resources.For this reason, scientists and researchers started considering the concept of sharing resources. In the last decade, the world of computing has seen a big leap forward with the emergence of cloud computing. In this new paradigm, computing resources of data centers spread all over the world can be shared to provide increased power and availability to users. However, with this increase in computing power came along an exponential increase in energy consumption,thus leading to an age where the Information and Communication Technology (ICT) ranks as one of the greatest contributors to Global Warming.

In parallel to making computers powerful, research topics today concentrate on energy efficient computing. The most discussed topic is the energy efficiency of data centers, where huge amounts of energy are wasted and dissipated in the form of heat and cooling overheads. Few articles discuss the energy problem at the level of the cloud network; even though, it is equally power demanding. The huge number of messages, going back and forth through the cloud, needs a huge infrastructure that uses vast amounts of energy. These messages, whether data, requests, or response of services are ever increasing in number and the anticipated increase in the industry's

carbon dioxide footprint is tremendous. The energy efficiency of the ICTplays an important role in sustaining the advancement of Global Warming. The steps towards green computing started with simple guidelines for hardware manufacturers and evolved into complex energy aware management of networks and data centers.

Traditional routing algorithms have mainly focused on maximizing network resources utilization. This approach is not appropriate for a cloud system, due to the huge amount of traffic at hand. For this reason, this thesis proposes an energy-aware routing algorithmbased on an exact MILP formulation. We will discuss the services of the cloud mechanism, their impact on pollution, the power consumption of the components of the cloud, and the attempts undertaken both at the level of data centers and the network to make the cloud "greener".Furthermore, the study conducted herein will show how network resource managementcan significantly enhance energy efficiencyby considering energy metric in the request routing decision making.Our model makes request management decisions based on the overall energy needed by each request and power calculations of the components needed to process a request.

Contents

List of Figures

This page is intentionally left blank

Glossary

AP	Access Point
AMPL	A Mathematical Programming Language
BP	Binary Programming
BIP	Binary Integer Programming
CPU	Central Processing Unit
CSP	Cloud Service Provider
DVFS	Dynamic Voltage and Frequency Scaling
GB	Giga Bytes
GPUE	Green Power Usage Effectiveness
laaS	Infrastructure as a Service
ICT	Information and Communication Technology
ILP	Integer Linear Programming
IP	Internet Protocol
ISO	International Organization for Standardization
kB	Kilo Bytes
MILP	Mixed Integer Linear Programming
MIP	Mixed Integer Programming
MIPS	Million Instructions per Second
MP	Mathematical Programming
Msec	milliseconds
NSFNet	National Science Foundation Network
NSP	Network Service Provider
OAM	Operation and Maintenance
OSI	Open Systems Interconnection

PaaS	Platform as a Service
PUE	Power Utilization Effectiveness
RAM	Random-access Memory
ROM	Read-only Memory
SaaS	Software as a Service
SLA	Service Level Agreement
TB	Terra Bytes
TCP	Transmission Control Protocol
US	United States
QoS	Quality of Service

Chapter 1 – Introduction

Context and Motivation

Cloud computing in its natural form is much greener than traditional corporate and personal computing. The architecture and form of the cloud suggests virtualization and sharing which in itself minimizes the operation and maintenance (OAM) costs and hardware pollution, while maximizing order and organization. A cloud is an expression of distributed computing environment that can be accessed remotely through several heterogeneous media and sends end-users a variety of services based on the requests gathered through a heterogeneous network infrastructure.

The term cloud is deduced from the abstract and remotely accessible nature of the architecture.Cloud computing brought forth tremendous improvements to the computer world; however, as any new innovation several problems followed it. The main problem of cloud computing arose due to the fast growth of cloud computing networks and data centers which demanded further increase in energy requirements; whether inside the data centers or for the transmission and switching of the networks that connect users to the cloud.

Most literature concentrate on improving the power consumption at the data centers, since most of the power loss occurs inside the data centers in the form of heat. Another important part of the power is indirectly lost to manage the excess heat of the data centers which was aresult of initial power loss. The latter is done by powering air conditioning machinery in order to cool down the data centers.

Demand for bandwidth over the internet has been growing exponentially. Growing usage of the internet means advancements in technology and growth for theICT; however, it also means growth in power consumption[5]. Another energy thirsty part of the cloud is the networking and communications; however, this sector has not received as much attention as the data center [6].

At both levels, the data center and the network, enhancements and improvements have been made toboth Hardware and Software. At the data center level, processors have become less power consuming and Universal Power Systems (UPS) have become more efficient in transforming electricity with less waste in the form of heat. Virtualization, parallelism, and other Software solutions have also assisted in making the data centers greener. At the network level, routers and switches have been improved to be more power efficient and use sleep and wake instructions to manage energy. Energy efficient Software has also been deployed to manage these networks and perform routing in a "greener" way[13].

It is worthy to note that a decrease in power consumption does not necessarily mean Green. The method used to produce the energy used also plays a role in the decision. Coal being one of the dirtiest forms of energy production should not be a source for powering the cloud[22].

Many researchers have already studied ways to reduce power consumption at the level of the network. Some propose better routing to ensure the least power consuming path; others suggest adding additional routers and switches in the network to increase the possible path[34]. Many network industries such as CISCO have already deployed solutions that use virtualization and put to sleep devices and components to decrease power consumption. Many Network Service Providers use sophisticated software to monitor their network's energy consumption.This thesis suggests administering power consumption at the early levels of request management. Measuring power consumption before accepting requests would allow preventing waste of energy. The Network layer should impose a protocol that would handshake over a shared metric which takes into consideration the power consumption of the request.

Ourthesis proposes using a Maximum Energy constraint over the network, which could be as a maximum: the energy needed to transfer a message through the longest and most energy consuming path available in the network, and deciding whether to accept or drop a request based on this metric. If the request needs more energy to be delivered than the Maximum Energy of the network, thenthe network will not accept it.Furthermore, other than just deciding upon the acceptance or the rejection of requests; this metric will allow the requests to be routed through the most energy efficient routes possible, thus significantly decreasing the energy consumption of the network, while maintaining the desired acceptance ratio.

Even though many research has been done on energy enhancement at the level of routing and network management; our model is specifically designed to reflect the structure of the cloud and will deal with processing as a service request management based on overall energy utilization and consumption.

Our studies show that a significant amount of energy can be spared by following the proposed formulation and guidelines, without severely affecting the request acceptance ratio and the Service Level Agreement (SLA).This thesis will also show different possibilities to measure and assign a Maximum Energy metric and show its effects and impact on the functioning of the network. The following section of the introduction will discuss in details the different sections of this thesis.

Thesis outline

In the first chapter, we introduced the major topics and the main problems discussed throughout our thesis. We also shared the motive behindthe formulation of this thesis. Along the thesis we will discuss each of the ideas mentioned in the Introduction in further details.

In the second chapter, we provide a general overview of energy consumption of the computer industry. We start by introducingthe cloud approach of computing, which explains how the cloud works and talk in details about the services that cloud computing provides, without going into the energy efficiency or consumption aspect of the cloud. Further into the chapter we briefly talk about energy efficiency in ICT and what green computing means. Then we move along to explain the two main parts of the cloud infrastructure: the Data Center and the Network, and discuss the improvements that could be brought forth to each of these parts to make them more energy efficient.

In Chapter three, we talk about routing problems in general and how traditional request routing problems and algorithms differ from the grid and cloud request routing problems. Wealso explain the significance of the power factor in routing and show the differences in energy efficiency due to the inclusion of this power factor.Then wededuce how the usage of this power factor can lead to a much greener environment. Chapter three also includes descriptions of other works done in this aspect and mention how each differs in style and what benefits they induce.

The first three chapters form a prerequisite for the study of our thesis and the topics discussed within. They further provide necessary information about the methods and models used in our thesis.The fourth and the fifth chapterscan be considered to be the core chapters of our thesis, where we will discuss the work done and present the studies that we conducted.

The fourth chapter has two main sections. In the first section of chapter four, we present the model upon which the study has been conducted and explain the purpose behind using a level of abstraction while designing our model.Further along, the components of the model are given and their properties are explained.In the second section of chapter four,we briefly explain the basic concepts of MILP and the reason behind using MILP to solve ourenergy efficiency problem. We then present our MILP formulas; dissect each formula and explainitseparately. The MILP formulas can be classified into four main groups: The parameters, the variables, the constraints and the objective function. The objective function makes sure that the study returns the results that are favorable and serve the purpose of our thesis.

In chapter five, we present the tests conducted on the model and the outcomes that resulted from the tests. Through the results of the tests that are portrayed and explained throughout chapter five, we prove the effectiveness of the solution that was proposed in chapter four.It is in the fifth chapter where the pieces of the puzzle come together and show how our model increases energy efficiency yet at the same time maintains an acceptable level of SLA.

Finally, we conclude our thesis by providing a summary of the key points addressed in our study. In the conclusion we also provide some perspectives in order to hint some possible paths to continueour work and address possible future improvements.

This page is intentionally left blank

Chapter 2 –Green Information and Communication Technology

a. Cloud Computing

Traditionally computing was achieved through personal computers which consisted of a system unit that had a processor, memory, hard drive(s), and all the necessary cards, which were connected to several peripherals, such as monitors, keyboards, mouse, different types of ROM, Modem, etc.Early improvements affected/targeted the units in the system unit, thus increasing the processing power of the CPU, size of the RAM and the hard drive by either replacing the units with more powerful onesor allowing the addition of more units; i.e. having several RAM cards.

With the enhancement of the internet, the development of heavy applications, and the mandate of decent graphical interfaces, the traditional approaches proved to be insufficient and soon the attention turned to having several system units operating as one and to sharing the resources of more than one system unit to work in parallel on a common process.

The High-Performance ComputingthroughCluster computing and the Grid were the results of the necessity to share resources. They are federations of computing resources to compute a common problem. The main difference between the Grid and the Cluster is that the Grid is more heterogeneous and the resources of a Grid are more dispersed geographically than the resources of a cluster.

The Grid technology paved the way to cloud computing. The cloud is an expression of a more reliable Grid that has much fewer Quality of Service (QoS) problems. The cloud relies mainly on abstraction and provides several end-user services, while the Grid provides computational resources.

The cloud mainly comprises of shared servers that work together to form a powerful machine. Virtualization allows the usage of this powerful machine by end-users with no knowledge of the complex structure or details of infrastructure. Users will benefit from services that the cloud provides, not knowing about the underlying hardware and infrastructure complexities.

Users will pay for the services they benefit from on pay-as-you-go terms and do not have to purchase or maintain any specific IT infrastructure. End-users simply need to have a simple computer and an internet access. Users will access the cloud services from wherever they are located in the world through the internet[15]. The location of the user and the location of the computing resources are irrelevant in the cloud environment. Resource Virtualization allows any user with an internet connectivity to access any amount of resources desired and be served without knowledge of any resource's location.

The cloud architecture can be generalized into two main types;public and private. The public cloud as the name suggests is available to the public and offers a variety of services to end-users with different localities. The public cloud is accessible through the internet, while the private cloud is shielded behind a firewall and is accessible by only the members of an organization or an enterprise[6].

The cloud provides three main types of services: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). The cost of these types of services depends on the Quality of Service (QoS) requirements of the user and the Service Level Agreement (SLA) reached that define the Quality of Service Requirements [15], [16].

The benefit of using the cloud is beneficiary for the end-user, since the cloud relieves end-users of the cost and trouble of owning and maintaining systems for several resources of storage and computing, instead the end-users share a centrally managed pool of such resources. Furthermore, the end-users have the ability to access all their assets remotely from anywhere in the world anytime they desire. The cloud also offers scalability; users have the possibility to increase their resources and have access to additional capacity, all equipped with backup facilities and as mentioned earlier, charged per usage[6].

The most popular services of the cloud can be considered to be the Storage as a Service and the Software as a Service.

Storage as a Service providers such as Google with its Google Drive allow end-users to place their data on their servers for a fee. In return users benefit from qualities, such as availability, maintainability, interoperability, manageability, scalability, usability, reliability, etc. while having some drawbacks on security, since the data placed on the cloud is accessible through the internet and the management of the enterprise owning the storage servers, hackers and governments can get access to private and personal data.

Availability allows users to access their data stored on the cloud servers whenever and wherever simply through the internet. Most storage service providers allow their users to share some or all of their data with another user or a group of users, placing interoperability issues at the mercy of a mouse click. Furthermore, storage service providers offer attractive backup packages and safeguard plans to protectthe Cloud data from viruses or other harmful incidents.

Most of the storage service providers allow a certain amount of data to be placed on their servers for free and charge per byte for additional data, thus making their service scalable. Google Drive allows 5GB of data per user to be stored on their servers without any charge and charge fees ranging between 2.49 US Dollars per month for 25GB of storage and 799.99 US Dollars per month for 16TB of storage.

Traditionally Computer Software was sold with a license and had System Requirements such as the specific operating system models and specific framework; furthermore, the license would be functional only on a certain computer and would cease to function if the software is attempted to be executed on another machine. The Software developers would charge users with additional fees for maintenance, upgrade and support. The cloud permits software developers to rent out the latest versions of their software to prospectswhile charging them with a membership fee. Users using Cloud Software as a Service may benefit from the services through any computer connected to the internet; however, the service provider might limit parallel instances of the software running for the same user.

In contrast to the numerous qualities brought by this service, a major drawback of using software as a service over the cloud is availability, while most users today have access to the internet, theymight feel vulnerable by relying on a virtual machine that could disappear and which holds valuable assets. Nonetheless, in the case of common software such as office tools or image editing software this approach would be most preferable. The popularity of the usage of such common software over the cloud, such as Google's Google Docs, proves the latter.

b. Green Computing and Energy Efficiency

There are different types of energy sources and becoming Green does not only mean maintaining low costs of energy or lowering energy usage levels. Becoming Green means being energy efficient while using the cleanest type of energy.

Renewable sources of energy offer the cleanest ways to produce energy and electricity. Hydropower, wind, solar energy, biomass, and geothermal and ocean energy are renewable energy sources. Hydropower plants produce energy using rivers and flowing water. Windmills produce energy using the power of wind that rotates the wind turbines and produces electricity. On the other hand energy produced by coal is the dirtiest and it is estimated that more deaths are associated with air pollution from coal-fired power plants in Poland, Romania, Bulgaria and the Czech Republic than caused by road traffic accidents[27].

Even though, the proposed algorithm of our thesis does not directly deal with the cleanliness of the energy used, it is worth to note that many governments have already initiated processes that enforce their subjects to use renewable and clean energy. One similar process is the tax process on carbon footprint. This tax process has been enforced to encourage the use of renewable energy instead of dirty energy and promote energy efficiency.Carbon taxes are fees levied by governments on the carbon content of fossil fuels such as oil, coal and natural gas [37].

The charge fee is relative to the amount of carbon dioxide each type of fuel emits in order to use its energy. By increasing the fees on the usage of these dirty energy sources, a carbon tax would encourage industries to reduce energy consumption, whilst enhancing energy efficiency, or to consider other natural sources of energy that would cost less in terms of taxes, thus promoting renewable energy. Moreover, this tax policy would allow reductions in the greenhouse effect and global warming [38].Implementation of this strategy has spread throughout the world, including South Africa, China, India, Japan, Australia and many others [39, 40].

Computers in general, as many other industries, have their share in Global Warming and pollution. Computer Hardware contain harmful chemical elements that are non-biodegradable, toxic and consume large amounts of energy in order to be manufactured and to function.

Early computers were not very economical having cathode-ray-tube monitors and huge sizes that wouldn't fit in small rooms. These ancient computers demanded extreme enclosed environments which had to be dust-free and have very low temperatures.

In contrast, computers have reduced pollution in many other domains;documents which in the past were printed on paper and had several hard copies resting on shelves in libraries now are fitted in few kB sized files[22].Meetings of sorts that needed gallons of kerosene to fly partners and associates to meetings are replaced by computers and connections that demand few Watts.Telecommuting is another such example.

Green Computing defines protocols for computing where environmental sustainability is taken into consideration. Different types of measurement metrics are used to evaluate the Greenness of ICT equipment. Such measurements include the magnetic, electrical and hazardous emissions of the devices and their energy consumption while performing on average performance. Further metrics define ergonomics standardsand the impact of the afterlifeof these equipment on the environment.

During the 1990s,the Environmental Protection Agency and the Department of Energy created an international standard trademarked as Energy Star. Energy Star was designed to promote energy efficient products and differentiate them from otherwise non –energy efficient devices. Printers, monitors and other computer devices would be marked by the Energy Star logo if they used 20%-30% less energy than the allowed energy use, which was set by federal standards.

Recent computer manufacturers put tremendous effort in minimizing waste and hazardous materials during the three phases of manufacturing, usage and disposal. The ICT sector and its relevant worlds are putting huge efforts to produce services and platforms that are energy efficient and sustainable; nonetheless, the magnitude of pollution resulting from computer hardware and network infrastructure is grave and the urgency to handle and minimize it is crucial. Statistics conducted in 2008 showed that Computers in general and networking in specific made up 3% of World-wide Energy consumption and 2% of carbon dioxide emission[23].

The following section will deal with energy efficiency issues and possible solutions at the cloud computing level.

c. Energy Efficient Clouds

Figure 1 shows a timeline of the detailed energy usage worldwide of the different sectors of ICT.

The green section of each graph represents the energy consumption of communication networks. The bottom chunk of the green section represents the amount of energy consumed by the Telecom operator networks, the middle chunk of the green section represents the energy consumption of office networks, and the upper most chunk of the green section represents the energy consumption of customer premises equipment.

The electricity consumption of the communication networks has recorded a notable increase between 2007 and 2012 and it is expected to increase even further and at a faster rate due to the advancements and necessity of communications.

The red section of the graphs in Figure 1 represents the amount of energy consumed by personal computers and divides them into four sections: desktops, laptops, Cathode Ray Tubes (CRTs) and Liquid-Crystal Displays (LCDs) respectively from bottom to top.

As it can be seen in the graph, the chunk representing desktops is shrinking, while the chunk representing laptops is growing, due to the fact that desktop users are replacing their huge desktops with easily portable laptops. Similarly, CRTs are being replaced with LCDs and LEDs, thus the shrinking of the electricity consumption of CRTs versus the electricity consumption of LCDs between 2007 and 2012.

The blue section of each graph represents the energy consumption at the data centers. The blue section being the largest section in 2007, helps us understand how energy consuming data centers are. Nonetheless, with the growing equipment and networks of the communication technology, the electricity consumption of communication networks has surpassed the electricity consumption of data centers and will create a grave imbalance during the upcoming years.

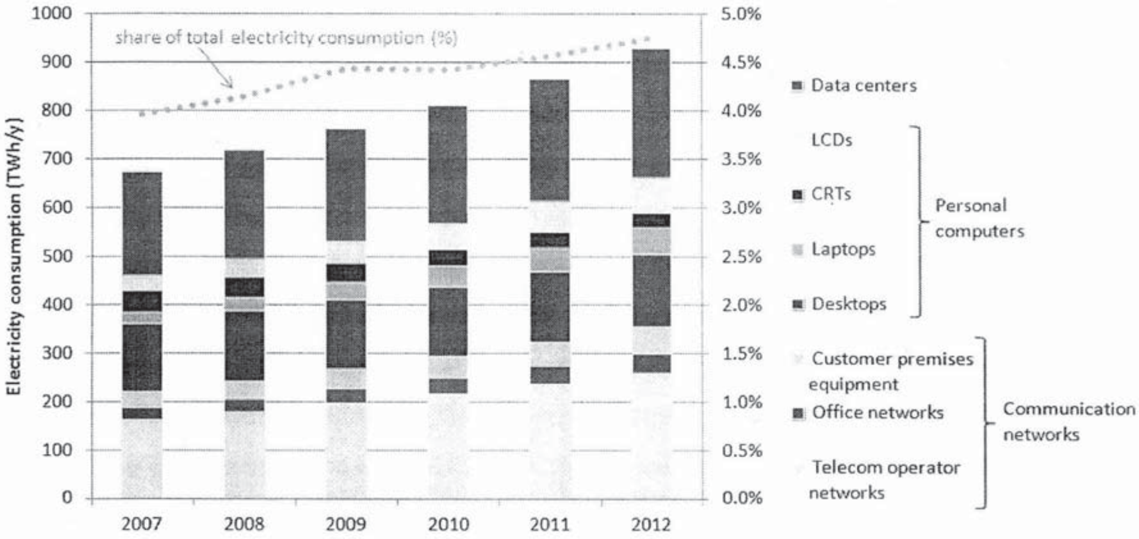

Figure 1 - Worldwide energy use of the ICT [22]

It is evident through Figure 1 that the data centers and the telecom operator networks have recorded a notable increase throughout the years in their electricity consumption. Figure 1 also shows that these two sectors are the most energy consuming sectors and thus need the most attention.

i. Data Center

Data Centers are the brains of the cloud. They are enormous centers filled with air conditioned rooms containing powerful servers racked next to each other. Some servers act as data servers, others have powerful processes; nonetheless, the servers are merged into virtual machines that are designed and managed in a way to conform best to the different types of services provided to the end-users[15].

Until recently the main purpose of data center enhancement was more power and higher performance and this has been achieved often without paying much attention to energy efficiency. It is believed that a data center on average consumes as much power as 25,000 households would[15]. However, with the performance and increase in demand, energy costs increase as well leading to the attention shift from solely focusing on performance enhancements to making the data centers more powerful yet in parallel decreasing their power consumption and minimizing the amount of wasted energy.

The main trigger for concern for Cloud Service Providers (CSP) was that most of the energy entering the data center was being wasted in the form of heat. Data Center owners started measuring the Power Utilization Effectiveness (PUE) of their facilities. PUE is measured by dividing the total power entering the facility by the power used by the Servers. Green Power Usage Effectiveness (GPUE) portrays in addition to the regular Power Usage Effectiveness (PUE) the greenness of the data centerby measuring the carbon intensity of the energy generating the power[22]. Data Centers are huge carbon producers and consumers of energy. In 2009 data centers worldwide collaborated in emitting more carbon than both Argentina and the Netherlands[15].

In recent years governments have also started pressuring the owners of data centers to reduce their carbon footprint. In Japan, the Japanese Data Center Council has been established to monitor the energy consumption of data centers[15]. Leading computer service providers, such as Google, Amazon and Microsoft have been competing to become more energy efficient. Customers and end-users are directly concerned with climate change and studies show that customers prefer the greener option. Furthermore, such leading companies have joined forces and have formed The Green Grid Consortium to minimize the environmental damage caused by data centers[15].

The two main concentrations of researchers to enhance power consumption of data centers are on the power consumption of the data center cooling system and the power consumed by server processors.

One of the difficulties of managing data centers is the conservation of strict ecosystems for these huge machines. The servers racked in a data center emit immense amounts of heat and need constant cooling and low degree temperature in order to avoid overheating and prevent damaging of the systems. These cooling machineries need high amounts of power to function. Replacing them constantly with more recent and more energy efficient versions would be cost inefficient and rather unmanageable. Instead data center owners should invest in the design of their data centers and plan well where to establish their data centers.

Data Centers should be designed in a way to allow empty space between the racks of servers in order to allow the passage of air. Furthermore, the technique Hot Aisle vs. Cold Aisle should be implemented, which allows the hot air to pass through the hot aisle and be replaced by the cold air which in turn will traverse through the cold aisle.

Moreover, in order to minimize the power wasted on the cooling overhead, data center owners should benefit from natural environmental conditions that are preferred by the servers. Dry and cold regions are the most preferable areas to establish a data center; vents would transfer the cold weather from outside the data center onto the servers and the hot racks. Amazon and Google have already started to shift their attention to such areas and establish most of their data centers in environmentally desirable locations. Natural means could also assist in the cooling of the data center hot spots, many data centers use naturally cold water to pass through pipes inside the data center and act as coolers.

To ensure energy efficiency and hinder the increase of power consumption at the level of the server, CSPs can either enhance the algorithm management and provide the services to the end-users in a way to minimize the processing byreplacing obsolete hardware with more advanced and more energy efficient hardware, such as Dynamic Voltage and Frequency Scaling (DVFS)[13], or they can use Sleep and Wake techniques benefiting from the Virtualization and Dynamic resource Allocation [10]. The Sleep and Wake technique would allow servers that have low traffic to sleep after transferring their processes to another server. The user will not be affected by this shift, since the user only views Virtual Machines.

Most data centers have already started using these techniques to reduce power consumption and new researches are being conducted to find even more methods and alternatives to shift data centers into becoming more energy efficient. After all, money talks and energy efficiency at the level of the data center smells Green in more than one sense.

ii. Network

A Network is an assembly of a variety of equipment that come together to provide a system of transportation for electronic messages. Networks may consist of hubs, edge switches, core switches, bridges, core routers, edge routers, access points (APs), amplifiers, etc. each having different properties that affect their power consumption; furthermore, each type of equipment may have different versions that consume different amounts of power, thus making the challenge of measuring power consumption at the level of theseequipment more difficult [3].

The Internet was initially an academic network; however, with time it became a worldwide communication medium which has become a necessary service for every household, institution and organization. The growth of connection speed and the number of peers connected to the internet has brought forth a massive increase in the number and power of the components forming the internet. Network Service Providers had no choice but to meet this enormous demand which in turn led to a decrease in the cost per byte of traffic and further made the internet desirable by making access more affordable[5].

A network router or switch consists of many different components that affect the overall power consumption of the device. The number of active ports is proportionally related to the power consumption of the device; furthermore, the line speed configured for each port also affects the power consumption[3]. In general there are four main factors that affect the energy efficiency of a network; the traffic rate and congestion, the Quality of Service (QoS) management mechanism, the router configurations, and routing. The total energy consumed by the network worldwide exceeded 250 TWh in 2012, while back in 2007 it was below 200 TWh[22].

To date there are two popular approaches to make networks energy efficient;designing the network in an efficient way and enforcing protocols to ensure energy efficiency. Designing a network is very fundamental since it affects the whole performance and efficiency. Network components aggregate power and the aim of network designers should be to minimize the number of components, while maintaining robustness and maximizing performance[5].To design such a network, network managers and designers should study several possibilities and run tests using several techniques, in order to choose the best possible scenario that fits their model. The efficiency of telecommunication equipment can be measured by a method called Telco Efficiency which measures the total number of bits coming out of each router and network component over an assessment window. Through the results of the Telco Efficiency matrix network designers and managements can pinpoint underutilized components and remove them[22].

Similar to any other electronic equipment, network equipment can consume less power if they are in sleep mode. This can be achieved by energy-aware routing, in order to saturate the routers that are already awake and put them to maximum use, while putting some other routers to sleep, thus saving energy. Researchers have also argued that there should be a maximum threshold that should not be exceeded in order to utilize power most efficiently. This can also be implemented heavily during night hours, week-ends and holidays when many routers can be switched off. According to a CISCO customer care case study, using the CISCO EnergyWise solution at a school in Pennsylvania, USA, to power off network equipment for a duration of twelve hours after school hours has yielded a rough estimate of 10 US Dollars per access point and 5 US Dollars per IP phone[19]. Zuqing Zhu, a member of the IEEE, suggests a network algorithm that would save approximately 18% energy by adjusting Cable Modems\ bonding groups and shutting down poorly utilized Upstream and Downstream (US/DS) ports on the Cloud Modem Termination System (CMTS) causing a maximum of 4.15 msec queuing delay. [21]

As the hardware components of the network affect power consumption and energy efficiency, network software and protocolsaffect energy efficiency equally. In the following chapter we will discuss address the effect of routing algorithms on energy consumption and efficiency.

This page is intentionally left blank

Chapter 3 –Routing Algorithm State of the Art

The network software allows virtualizing the different layers of the network into a hierarchy of protocols, allowing each layer to provide services to the upper layer. Each layer of each device communicates with its specific layer in the other device. The protocols of the layer specify the constraints upon which the communication is established and should be shared by both devices in order for them to understand the messages going through each other. Any message violating the shared protocol will be misunderstood or rejected as a whole [31].

In reality the messages are not directly transferred from a layer of onedevice to the other. They are transferred through the lowest layer which is the Physical layer; however, each layer can interpret the message wrapped by its peer layer. The physical layer is the physical medium and the actual communication occurs through it, whilethe remaining layers do not have physical media connecting to their peers and should deliver their messages through the physical layer and hold a virtual communication with their peers.

Each layer performs a collection of specific functions and it is up to the network designers to specify these layers, their protocols and their functions. There are two main network architectures; the OSI Reference model and the TCP/IP Reference model. The TCP/IP Reference model's protocols are more widely used than the OSI Reference model's protocols; even though, the OSI model's protocols are more valid[31].

The layer that is directly related to routing is the Network Layer.The protocols of the Network layer define how packets are routed through the network. Routing can either be done by using static routing tables programmed into each router, by specifying a route at the start of each conversation, or by being dynamically determined for each packet separately.

a. Traditional Routing Algorithms

In order to perform efficient routing, the network layer should know about the topology of the network and choose appropriate routing paths through the components and nodes of the network. The routing process should also take into consideration congestion issues and maintain a balance in order to avoid the overloading of some nodes while leaving others idle[46]. The optimal path is the path that satisfies the routing metrics specified by the network functionality[33].

Traditionally, Routing Algorithms were required to be correct, simple, efficient, robust, stable, fair and optimal in a sense to minimize the total number of hops. There are two main classes of Routing Algorithms, namely static or non-adaptive and dynamic or adaptive. The non-adaptive algorithms compute the decisions in advance. While the adaptive algorithms compute routing decisions in a way to reflect the changes in topology and traffic of the network [46].

The most common non-adaptive routing algorithms are Flooding, Shortest Path, and Flow-Based. Flooding makes routers forward messages to every other router in the network except to the router from which the message was received; even though, this approach floods the network and has many drawbacks, nonetheless it gets the message through as long as a path exists between the source and the destination. Shortest-Path algorithms measure the shortest path from each node to every other node in the network and when a message needs to be forwarded from one to another, the shortest path is taken [46].

The most common adaptive routing algorithms are Broadcast, Multicast, Hierarchical, Distance-Vector, and Link-State. According to the Distance-Vector algorithm,at regular intervals each node of the network sends its information about the entire network to its neighboring nodes [46]. According to the Link-State algorithm a node sends its information about its neighboring nodes to the entire network only when there is an update in its routing table.Djikstra's algorithm is an example of Link-State algorithms.

b. Cloud Routing Algorithms

Initially routing algorithms' sole target was to deliver a message from a source to a destination. Subsequently networks evolved and more enhanced routing algorithms started taking other characteristics into account. As more and more data was traversing the internet and as networks were increasing in size; routing algorithms had to provide services that did not only ensure delivering the message but also delivering it on time, with the least delay possible.

Further into technology, with the evolution of the cloud, another difference brought to the routing algorithm mechanism was the selection of the destination. The cloud structure provided the users with a level of abstraction. This abstraction hid the location of the servers from the users, thus hiding the destination and creating a challenge for routing and load balancing [47].

The messages coming into the network had no longer a specific destination; instead the routing algorithm had to direct the messages to the optimum destination through an optimum path.Network Service Providers started dealing with financial properties, such as price and cost. The profit of Network Service Providers became dependent on the Routing Algorithms' efficiency, thus the need to optimize the routing algorithm in order to maximize a profit metric.

c. Energy-Aware Routing Algorithms

Energy efficiency became a metric for cloud networks when the global emission of greenhouse gases due to the cloud network started rising. Researchers started studying the implementation of Routing Algorithms that satisfied the energy efficiency metric along with the other metric.

Routing Algorithms that satisfy the energy efficiency constraint take into consideration both the distance and the power consumption of the destination routers; trying to decide on an optimal destination and maintain an optimal path constrained by both metric.Reducing the number of hops and transmission links would minimize power consumption in all cloud services [6]; however, in the real world the power consumption of the network equipment should also be taken into consideration, since each router or network equipment has different power consumption as discussed in the third section of the second chapter. M. A. Youssef, M. F. Younis and K. A. Arishahave proposed an energy-aware routing algorithm, which uses a constrained minimum number of hops algorithm; using their algorithm they show that for moderate values of the maximum transmission distance the performance is acceptable for all aspects [43].

Z. Guo and B. Malakooti have demonstrated how measuring node energy consumption and using the resulting measurements to conduct fair sharing of energy consumption can lead to alleviation of initial energy differences, thus minimizing overall energy consumption [42].

Similarly, A. Benslimane, R. E. Khoury, R. E. Azouzi and S. Pierre have suggested including energy consideration in routing by using heuristics to compute the multipoint relays and routing table calculation while using min-max energy conversation and other measurements. They have shown how these heuristics help to find an optimal power path, where the maximum energy consumption on that path is the minimum among all the possible paths [44].

Other Energy-Aware routing algorithms deal with protocols that include sleep scheduling strategies, and aim at minimizing energy consumption by putting to sleep underutilized nodes. A. R. Swain, R. C. Hansdah and V. K. Chouhan have presented an energy aware routing protocol that emphasizes on efficiently constructing the broadcast tree with two paths from each node towards the sink, and maintaining a higher energy at each internal node [45].

This page is intentionally left blank

Chapter 4 – Energy-Aware Routing Algorithm

a. Model Schematics

This chapter and the nextwill present the work done and the results obtained. First we present the fictional model that we created based on the NSFNet infrastructure. Our model is a network domain that is represented by a Graph that consists of nodes (N) and links (L) as seen in Figure 2.Our abstract model does not deal with the physical topology of the network and does not include all the details of a real network. Abstractions will be explained in detail with each component of the model. Even in real networks the NSP often hides the underlying details from the CSP.

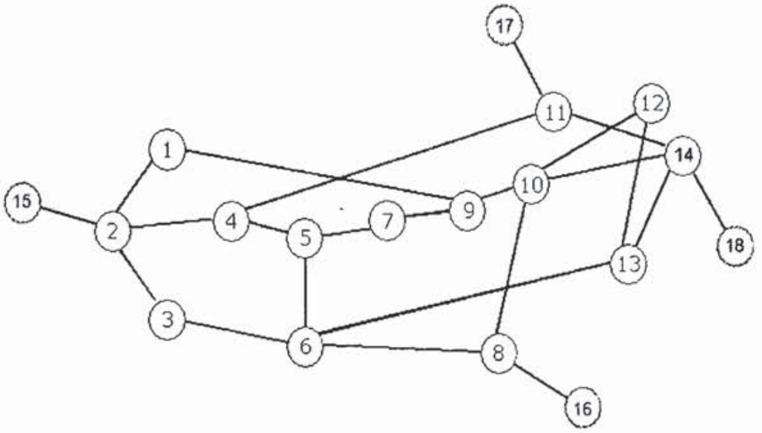

Figure 2 - Model Schematics; the red nodes have processing power while the black nodes do not have any processing power

i. Characteristics of Links

Links $e_{(u,v)}$ connect the nodes of our model to each other. In our model, Linksare characterized by their capacity $B_{(u,v)}$ in bits per second and are of two types. The Links connecting nodes 1 to 14 to each other are core links, and the links connecting node 14 to node 18 ($e_{(14,18)}$), node 11 to node 17 ($e_{(11,17)}$), node 8 to node 16 ($e_{(8,16)}$), and node 2 to node 15 ($e_{(2,15)}$) are edge links and should have higher bandwidths than core links. Edge links areoften assigned higher bandwidths because requests coming from different routes assimilate at the edge routers and traverse through edge links to the processing server nodes.

For the sake of simplicity our model does not include service responses and focuses solely on user requests. In real networks the links connecting the nodes are bidirectional; they transfer the requests to the appropriate destination and in case of a response transfer, the response is made from the server in the data center to the source generating the request.

Due to abstraction our model does not handle or deal with the means by which requests reach the network. In other words our model does not include the access points and the networks through which requests reach our network; whether it is the internet or some private network. In our model requests have already reached the network and reside at one of the routers.

ii. Description of Nodes

Our model includes 18 nodes. These Nodes can be classified into two disjoint subsets: a subset C ∈ N of nodes that have processing power and a subset R ∈ N of nodes that do not have any processing power and merely act as routers. Nodes 1 to 14 do not have processing power and belong to R; nodes 15, 16, 17 and 18 have processing power, provide processing as a service and belong to C.In our model we used IT resource abstraction and did not include data center and cluster details. We specify data centers according to their processing functional capability in MIPS. This has been done for simplicity reasons and in order to have more room for optimization [41]. Each node belonging to R can be a source node for requests; a node belonging to C cannot be a source node for requests, but acts as a destination.

Router Nodesare characterized by the following properties:

- power consumption $\varepsilon_u$in Watts, calculated by the router manufacturer.
- capacity $\varsigma_u$in bits per second
- weight: a constant to replace the several factors that affect the power consumption of network equipment as mentioned in the second section of the first chapter. The greener the router, the less weight it has on the total energy calculation.

In addition to the above properties, server nodes have one more characteristic that specifies how much processing power each server node has. It is defined in Million Instructions per Second (MIPS).

iii. Characteristics of Requests

In our model, requests are scheduled in advance, which means their arrival time and duration are known in advance. They initiate at a source node that acts as the first router which routes the request to a server node, in order to be processed. The request may pass along several router nodes before reaching a server node.

The routing process should satisfy certain constraints and these constraints depend on the properties of the Requests. Requests are characterized by their bandwidth, processing power, start time, finish time and a source which identifies the initial router they reach after entering the cloud.

The bandwidth of a request in bits decides on the path that the request should be sent on in order to reach the server node. The sum of the bandwidths of all the requests traversing through a given link at a given time should not exceed the bandwidth capacity of the link.

The power of a request decides to which servers the request should be sent for processing. Some server nodes will not have enough processing power to process a given request and some others will already be congested with other requests and will not have in time processing power to handle the request. Each request is scheduled to be processed during a time range that extends between the start time and the finish time of the request. Our model is designed to manage the requests in a way to benefit from the maximum possible number of accepted requests.

b. Our Proposed Energy-Aware Routing Algorithm

In this section, we present our mathematical formulation.In this algorithm, we consider computing requests scheduled over multiple data centers. Details of our algorithm will be given along with the equations. We first start by stating the parameters and the variables used in our model. Then we present the constraints that govern our routing algorithm. Finally, we discuss the objective function that reflects the expectations of the NSP.

We have used Mixed Integer Linear Programming (MILP).One of the most powerful modeling techniques used to solve decision making problems is Mathematical Programming (MP)[28].

When dealing with a decision problem, the first step is to state the facts; these givens form the parameters. The second step is to identify the possible variables to be decided upon. The third step is to specify the constraints outside of which a decision cannot be made. These constraints depend on the nature of the decision problem and are written as a set of equations. The final step is to specify an objective function which defines the objective that is needed to be served in the best way possible through the madedecisions. It can either be a maximizing function or a minimizing function.

Linear Programming (LP) is a type of Mathematical Programming (MP) in which the objective functions and constraints are linear[28]. When some of the variables in the model are real-valued and others are integer-valued then the model is mixed, resulting in Mixed Integer Programming (MIP), which in general is equal in meaning to Mixed Integer Linear Programming (MILP) that has a linear objective function and linear constraints.

Mixed Integer nonlinear Programming (MINLP) problems are much harder to solve than Mixed Integer Linear Programming (MILP) problems. Other instances of Linear Programming (LP) also exist, such as Binary Integer Programming, Integer Programming, Binary Programming, etc. [29]

MILP takes as input variables, parameters, constraints and an objective function and returns the optimal solution that would respect all the constraints and follow the objective function of the given problem. Even though, seeking the optimal solution in very large problems would demand an enormous time span; nonetheless, for small problems and designing models it would serve quite smoothly. Even with large problems, heuristics can be used and allow the returning of near optimal solutions in acceptable time.

In our specific model of 18 nodes, it took a fraction of minutes even with thousands of requests; however, with larger network models and having millions of requests this fraction could turn into minutes and even hours. Nonetheless, a NSP does not have to run our proposed algorithm on a daily basis, since our algorithm will help the NSPs in designing their networks and specifying constraints, thus NSPs could run the algorithm when they would like to modify the structure of their networks.

i. Parameters

The following are considered as inputs to the MILP formulations. They represent the characteristics of our graph, including the characteristics of the links and the nodes, the characteristics of the requests, discrete time instants, the matrix between these time instants and the request duration based on the start and the finish times of the requests and the maximum energy allowed for each request.

- A Network Graph $\mathbf{G(N,\ddot{E}), C \in N}$

  C is the set of server nodes that have processing power.
- A Set Γ of computing requests $(\{\boldsymbol{\Gamma} - \boldsymbol{\Gamma_c}\} = \mathbf{0})$

  In our model all requests are computing requests.
- A request tuple $\boldsymbol{\sigma_r}(\boldsymbol{s_r}, \boldsymbol{f_r}, \boldsymbol{b_r}, \boldsymbol{a_r}, \Theta_r)$

  Consisting of start time, finish time, bandwidth, source node and processing power
- A set of discrete time instants

  $$\mathbf{T} = \bigcup_{i=1}^{\Gamma}\{\boldsymbol{s_i}, \boldsymbol{f_i}\} \cup (\boldsymbol{t_0}, \boldsymbol{t_0} + \boldsymbol{\tau}) = \{\boldsymbol{t^m}\}$$

- A $[\mathbf{1} \times \|\boldsymbol{\tau}\|]$ binary vector $< \boldsymbol{t_r^m} >$ proper to σ_r where $\boldsymbol{t_r^m} = \mathbf{1}$ if and only if$\boldsymbol{t^m} \in [\boldsymbol{s_r}, \boldsymbol{f_r}[$
- The Maximum Energy **M** is the amount of energy that a request may use. By default it is the amount of energy that a request may need to traverse through the most energy consuming path of the network.

ii. Variables

The following variables are required by the MILP formulation. They represent the information we need as output. First we need to know if a request has been accepted. If a request has been accepted, we need to know which links did it traverse and at which server node was it processed. Last but not least we need to know how much energy did the request consume while moving from the source node to the sever node.

The first three variables are binary variables. They can have a result of either 0 or 1. The fourth variable is a real variable and can have a real result. The result should be a value represented in Joules.

<u>Binary Variables:</u>

- $\boldsymbol{\beta_r} = \mathbf{1}$if and only if$\boldsymbol{\sigma_r}$ is an accepted request
- $\boldsymbol{H_r^{(u,v)}} = \mathbf{1}$ if and only if$\boldsymbol{\sigma_r}$uses edge$\boldsymbol{e_{(u,v)}}$ for routing
- $\mathbf{P}_{\boldsymbol{r}}^{\boldsymbol{c}} = \mathbf{1}$ if and only if σ_r *uses node* v_c for processing at any time

Real Variables:

- E_r the energy requirement for routing the request from the source node to the server node

The energy $E_r$represents the total energy required to process a request Γ. It is calculated for each request bysumming up the bandwidth of the request by the power of the sending router multiplied by the sending router's weight and divided by its capacity for all the edges traversed by the request.

iii. Constraints

equation 1:

$$\sum_{v_c \in \mathrm{C}} \mu_r^c = \kappa_r \qquad \forall \sigma_r \in \Gamma$$

Equation 1 forces an accepted request to use only one processing node.

equation 2:

$$H_r^{(u,v)} \leq \kappa_r$$

Equation 2 allows any request σ_r to use multiple links for data transfer *equation 3:*

$$\boldsymbol{\alpha_r^n = \begin{cases} \kappa_r & if\ v_n = a_r \\ 0 & otherwise \end{cases}}$$

Equation 3 states that a request can have only one source node

equation 4:

$$\sum_{\sigma_r \in \Gamma} \boldsymbol{\mu_r^c} \cdot \boldsymbol{t_r^m} \leq \text{И}_c \qquad \forall \boldsymbol{v_c} \in \boldsymbol{C}, \qquad \forall \boldsymbol{t^m} \in \boldsymbol{\tau}$$

equation 5:

$$\sum_{\sigma_r \in \Gamma} (\boldsymbol{\mu_r^c} \cdot \boldsymbol{t_r^m} \cdot \theta_r) \leq \phi_c \qquad \forall \boldsymbol{v_c} \in \boldsymbol{C}, \qquad \forall \boldsymbol{t^m} \in \boldsymbol{\tau}$$

Equations 4 and 5 state that a server node can be time-shared by maximum $\text{И}_c$requests at a given time $t^m$and that the power of these requests should not exceed the total processing power of the server node.

equation 6:

$$\sum_{\sigma_r \in \Gamma} \left[\boldsymbol{b_r} \cdot \boldsymbol{H_r^{(u,v)}} \cdot \boldsymbol{t_r^m}\right] \leq \boldsymbol{B_{(u,v)}} \qquad \forall \boldsymbol{e_{(u,v)}} \in \mathbf{N}, \qquad \forall \boldsymbol{t^m} \in \boldsymbol{\tau}$$

Equation 6 states that the bandwidth of a request using $e_{(u,v)}$ should not exceed the capacity of that edge at any time t^m.

equation 7:

$$\sum_{e_{(u,v)} \in \mathrm{N}} H_r^{(u,v)} - H_r^{(v,u)} = \alpha_r^n - \mu_r^c \qquad \forall \sigma_r \in \Gamma, \qquad \forall v_c \in C$$

Equation 7 holds the flow conservation laws, makingsure that router nodes can only send requests to connected nodes, and that the server nodes act as final destinations and cannot act as routers.

equation 8:

$$E_r = \sum_{e_{(u,v)} \in \mathrm{N}} b_r \cdot \frac{\varepsilon_u}{\varsigma_u} \cdot w_u \cdot H_r^{(u,v)} \qquad \forall \sigma_r \in \Gamma$$

Equation 8 measures the energy used to transfer all the requests to a processing server node.

iv. Objective Function

$$\textbf{function } \mathrm{F}_{\alpha}: \qquad \textbf{max} \sum (M \cdot \kappa_r - E_r)$$

Our objective function is intended to maximize the number of accepted requests; however, minimizing the total energy required to transfer the accepted requests to their processing nodes.

M is the highest acceptable energy that a request can spend to be processed at the processing node, positioned at the furthest and most energy requiring path possible throughout the network.

Chapter 5 - Numerical Simulations

In this chapter, we run our MILP model over the NSFNet given in Figure 4.1. We consider several sets of simulations. We have run our simulations over the Neos solver [36], which is a free online accessible server that provides solvers for Mixed Integer Linear Programming supported by A Mathematical Programming Language (AMPL). The results have been plotted using MATLAB and explained in due section.

Each simulation is presented by first presenting the constraints and the setup and then by providing the results through the consecutive figures which are explained thoroughly in due section.

a.Network Performance

We start our simulations by investigating the network performance by checking how the acceptance ratio is affected by increasing/decreasing the available resources: bandwidth and processing power.

In the following two figures we have allowed request acceptance regardless of their energy consumption. The simulation setup will further explain the study.

1. Simulation Setup

We consider several sets of requests ranging from 1 to 100 requests. The requests are generated randomly. Each request is assigned a random duration time, so that our model can schedule different requests at different intervals and assign them to appropriate server nodes based on our constraints. Server nodes can process several requests at each time instant as long as their processing power allows additional processing.

In Figure 3, server nodes are granted large processing power and the maximum energy (M) is assigned a very high value so that no rejection occurs due to energy consumption. The bandwidths of all the links is equal and fixed to a specific value for each curve in Figure 3. The values considered for our curves are 10 and 40 Gbps.

In Figure 4, links' bandwidths and maximum energy (M) are assigned with very large values, so that no rejection occurs due to energy consumption or to bandwidth shortage. Server nodes are granted equal and fixed processing power for each curve. The values considered for our curves are 5, 30 and 400 MIPS.

2. Simulation Results

In Figure 3, we plot the number of accepted requests versus the number of generated requests for two variations of the bandwidth capacity of the links.

The dotted line represents the case where all links have a bandwidth equal to 10Gbps;whereas, the solidline represents the case where all links have a bandwidth equal to 40 Gbps.

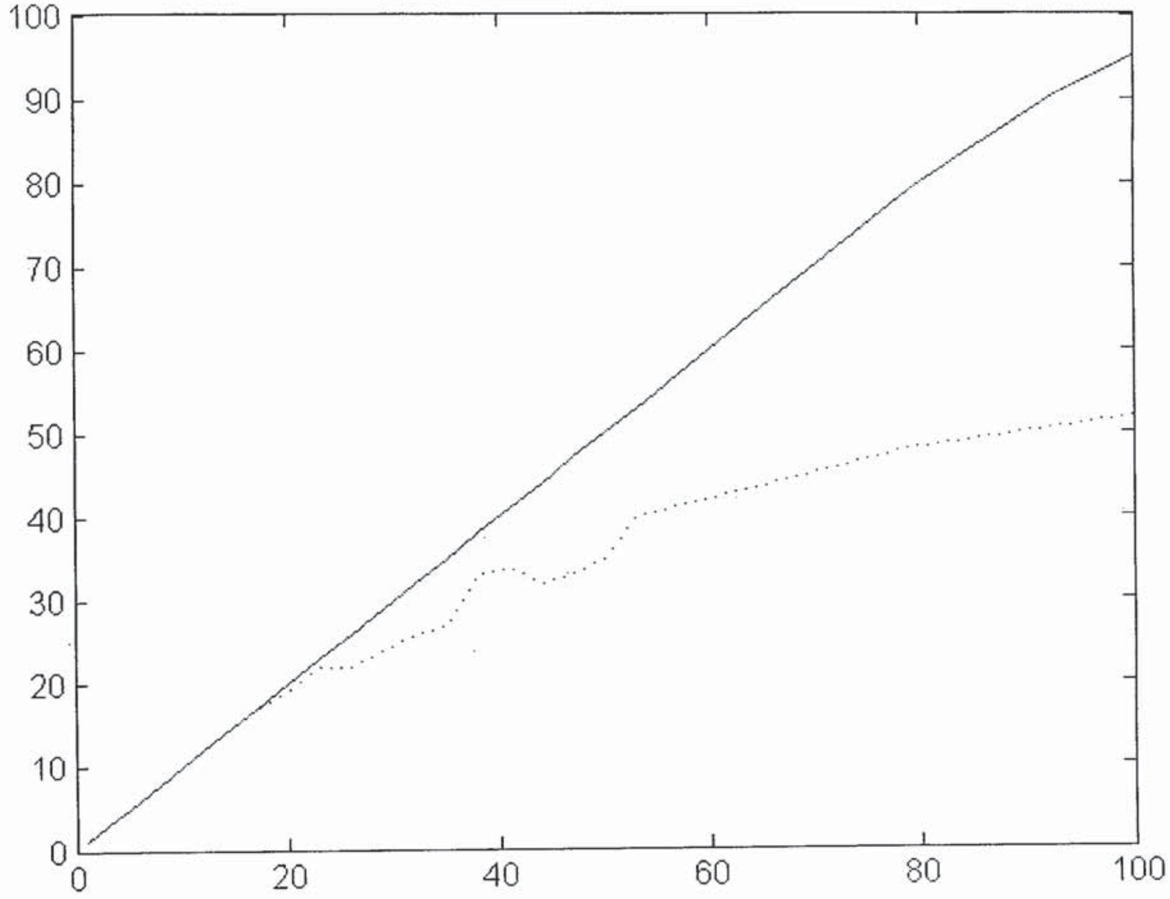

Figure 3 - Number of Accepted Requests vs. Number of Generated Requests for different links bandwidth

We can see that the acceptance ratio of the requests is highly affected by the bandwidth of the links. In the case of high bandwidth availability, the acceptance of the requests is almost linear. However, for low bandwidths, our formulation tries to maximize the sharing of resources among requests. This is why in the case of 50 generated requests the MILP accepts 35 requests; while in the case of 100 generated requests the MILP accepts 50. This is due to having more possible choices for time sharing the resources among the 100 requests; whereas the possibilities are limited in the case of 50 requests.

In Figure 4, we plot the number of accepted requests versus the number of generated requests for three variations of the computing capacity of the server nodes. The dotted line, dashed line, and the solid line represent the case where all data centers have computing capacity of 5 MIPS, 30 MIPS, and 400 MIPS respectively.

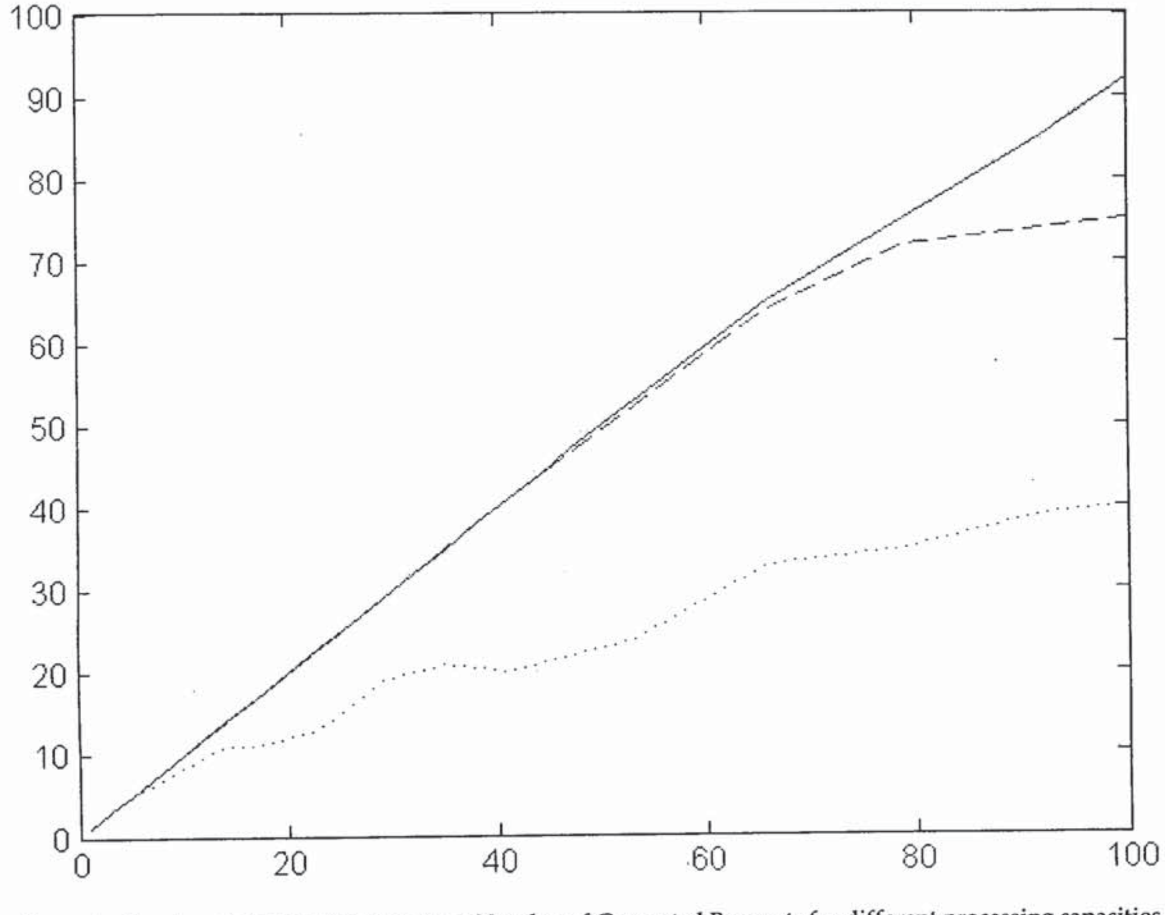

Figure 4 - Number of Accepted Requests vs. Number of Generated Requests for different processing capacities

Similar to the previous simulation, we can see that the acceptance ratio is directly affected by the capacity of the data centers. We can see that for the case of 400 MIPS (solid line) the acceptance ratio is almost linear. This is due to the fact that the bandwidth is sufficient enough so that no rejection occurs based on bandwidth limitations. On the other hand, this case provides processing power more than the power needed by the requests, which renders it equivalent to having unlimited processing power.

The 30 MIPS case (dashed line) provides similar results as the former as long as the number of requests does not exceed 70. Above this number of generated requests, the available power becomes insufficient and leads to rejection of some requests. As the number of generated requests increases above 75 (from 75-100), we can see that the acceptance of requests becomes almost constant. This is due to the fact that the data centers become saturated.

In the last case (dotted line) the data centers are under provisioned and have been granted a capacity of 5 MIPS. This lack of capacity leads to a high number of rejected requests and allows a maximum of 40 accepted requests in the case of 100 generated requests. One might wonder why 40 generated requests would yield only 20 accepted requests and not 40 accepted requests, since we saw that in the case of 100 generated requests the model yielded 40 accepted requests! The reason behind this result lies in the fact that for 100 generated requests our model has more possibilities to choose 40 requests that can time-share the available processing power.

b. Energy-Aware Simulation

As a second studywe investigate the behavior of our model due to energy restrictions. In this section we study how the acceptance ratio is affected at different values of maximum allowed energy consumption. The simulation setup will further explain the study.

1. Simulation Setup

We consider two sets of simulations. In the first simulation, represented in Figure 5, we randomly generate 100 requests and we study the network's behavior over different values of the maximum energy (M) ranging from 2 till 70 Joules.

In the second set, represented in Figure 6, we fix the maximum energy (M) at four different values: 5, 10, 20 and 50 Joules, and vary the number of generated requests from 1 till 100.

2. Simulation Results

In Figure 5, we plot the number of accepted requests versus the maximum energy (M) allowed.

We can see that the acceptance ratio of the requests is highly affected by the maximum energy (M) allowed. After 40 Joules we can see that the number of accepted requests stops increasing. This is due to bandwidth limitations of this scenario. The network becomes congested with 85 requests that have been accepted. For this reason a CSP can still accept the maximum number of requests allowed by its network, while not having to allow very high energy consumption.

This scenario shows us that if the maximum energy was limited to 40 or allowed to 70 Joules this will not affect the number of accepted requests.

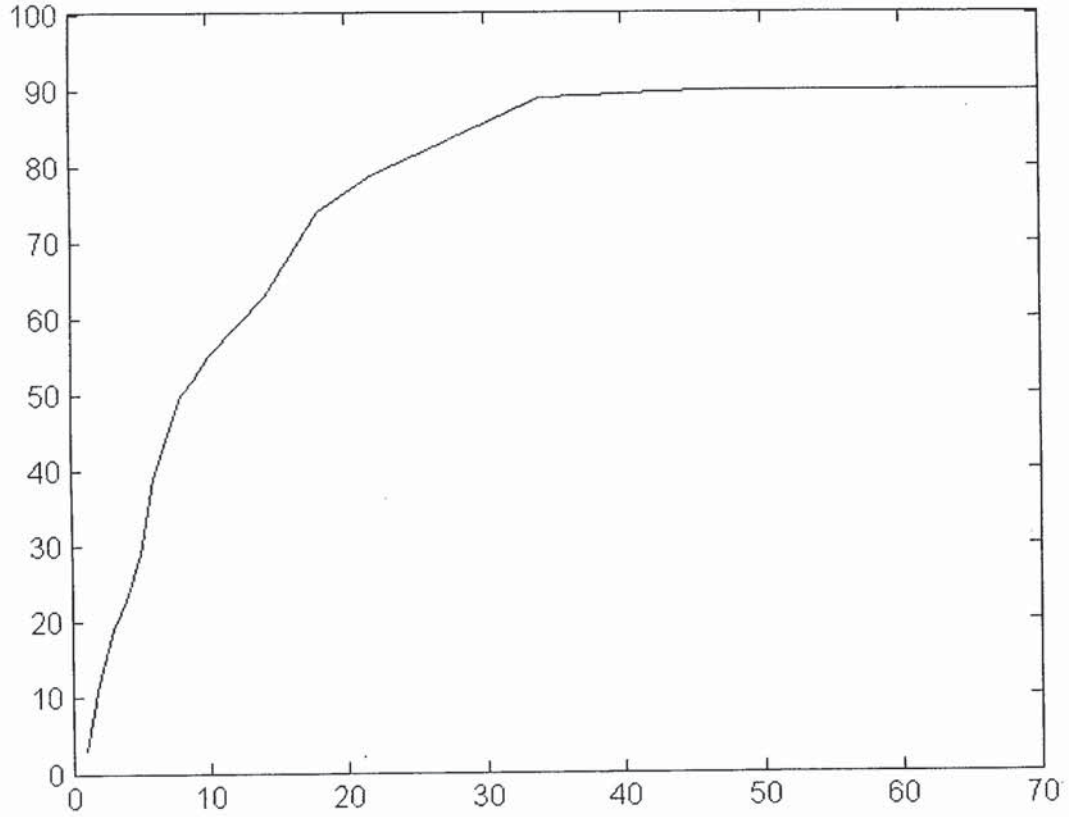

Figure 5 - Number of Accepted Requests vs. the Maximum Energy Allowed

In Figure 6, we plot the number of accepted requests versus the number of generated requests for four variations of the maximum energy (M) allowed. The solid line, dotted line, dashed line and thedash-dot line represent the cases where the maximum energy has values of 50 Joules, 20 Joules, 10 Joules and 5 Joules respectively.

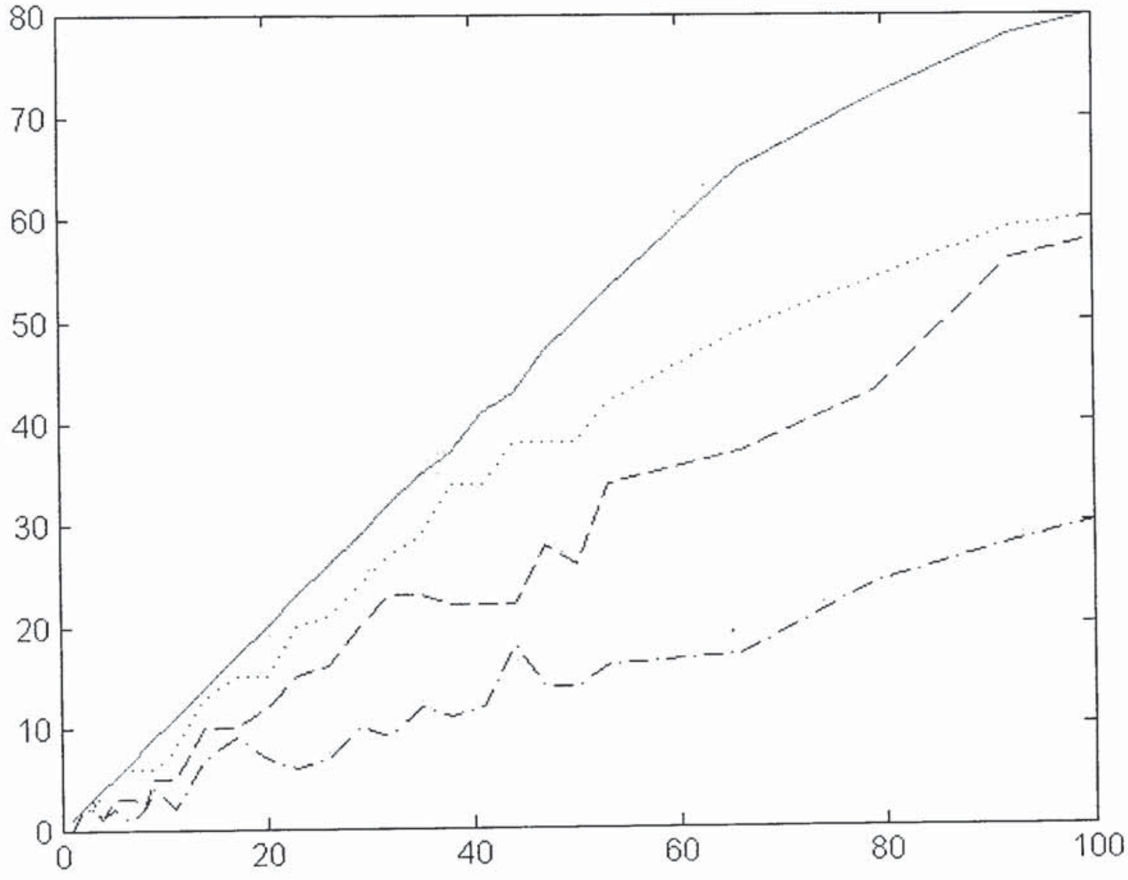

Figure 6 - Number of Accepted Requests vs. Number of Generated Requests for several Maximum Energy Values

At the start, where the number of generated requests is less than 20 the dash-dot line yields the least number of accepted requests due to the limitation on the maximum allowed energy consumption of 5 Joules. The dashed line (10 Joules)has somewhat higher yield; nonetheless it remains to be around 50% of an acceptance ratio. The dotted line (20 Joules) yields a value approximately close to that of the solid line (50 Joules), due to the fact that the number of generated requests is still small and 20 Joules is quite enough to accept most requests.

As the number of generated requests increases, the gap between each line increases. This is due to the incapability of the models with lower allowed maximum energies to handle large number of generated requests.

For the same available resources, it is evident, through Figure 6, that the maximum energy (M) allowed affects the acceptance ratio drastically. For instance, if we take the case of 80 generated requests, we can see that between M = 5 (dash-dot line) and M = 50 (solid line) the gain in terms of number of accepted requests is 60%. Similar to previous studies, the reason behind the increase in all the lines, even in the case of the dash-dot line which has a very low energy metric fixed at 5 Joules, is that for an increased number of generated requests our model has more possibilities to choose among diverse requests that can time-share the available resources of our model.

c. The Influence of the Maximum Energy Allowed on the Total Energy Consumption

1. Simulation Setup

The following set of simulations show the effect of the maximum energy allowed (M) on the Total Energy Consumption of all the requests accepted. All the resources of our model have been reset to high values, in order not to affect the request acceptance ratio in any way. We have run 2000 requests at 9 different times, each time modifying the value of the maximum energy allowed.

During the first round, we set the value of the maximum energy allowed to the most energy consuming path of the model, thus not limiting our model by any metric. During the following runs we set the value of the maximum energy allowed to values of 24, 34, 39, 44, 49, 61, 65, and 75 Joules respectively.

2. Simulation Results

In Figure 7, we plot the total energy consumed versusthe maximum energy (M) allowed. We can see that the total energy consumed is highly affected by the maximum energy (M) allowed.

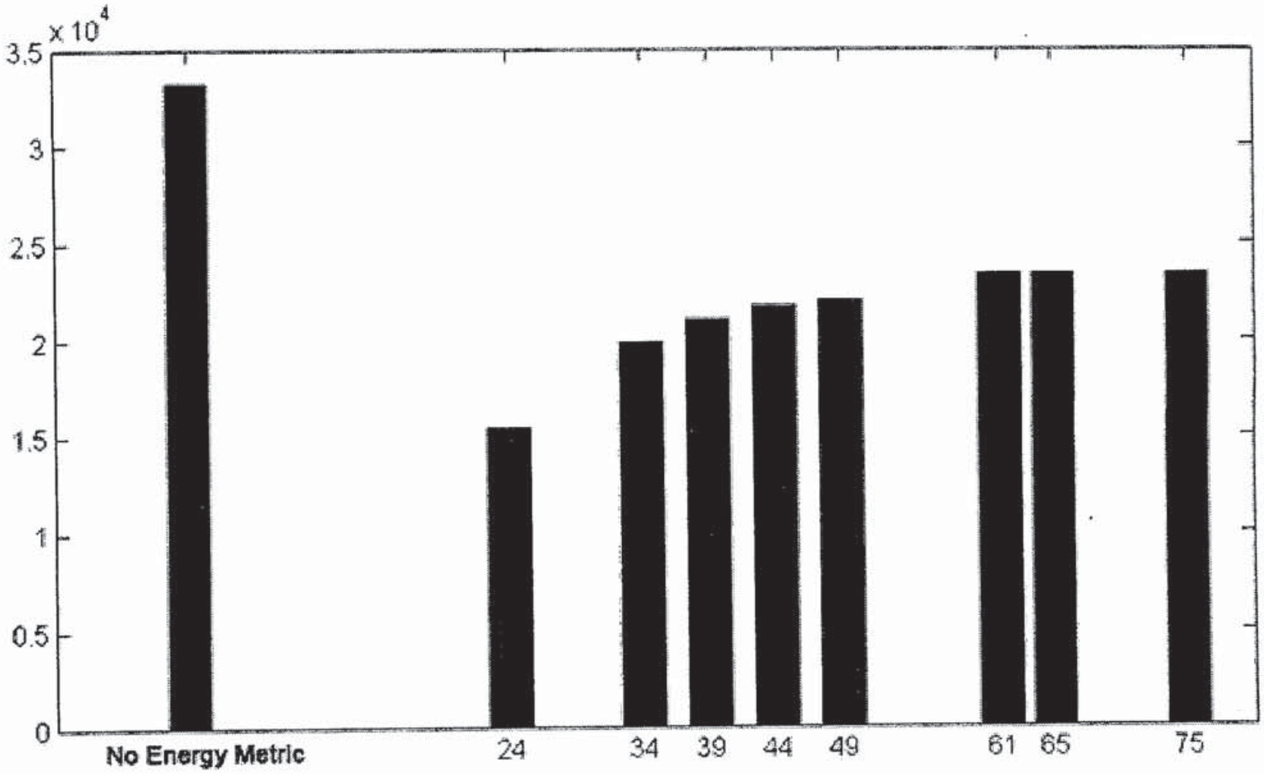

Figure 7-Total energy consumed vs. Maximum Energy Allowed

In the first case we start by setting a benchmark, where we do not limit the model with any energy metric. This shows that without the energy metric, the total energy consumed by the network has reached a high value of 35,000 Joules in order to handle 2000 generated requests. This benchmark is represented in Figure 7 by the first bar, labeled by "No Energy Metric" on the x-axis. The rest of the bars have a maximum energy constraint of 24, 34, 39, 44, 49, 61, 65 and 75 respectively.

The bar labeled by 24 results in 15,000 Joules of energy consumption, which is the lowest value plotted, as seen in Figure 7. The next four bars labeled 34, 39, 44, 49 and 61 yield total energy consumptions of around 20,000 Joules being the lowest, respectively increasing until reaching a value of 23,406 Joules at a maximum energy allowed of 61 Joules.

As seen in Figure 7, the total energy consumed by all the requests does not change above maximum allowed energy of 61 Joules. That is because 61 Joules is the bestmaximum energy allowed which yields a 100% acceptance ratio (as seen in Figure 8), while minimizing the energy consumption. Increasing the allowed energy M above 61 Joules provides no additional improvement.

To conclude we should stress on the fact thatwhile not having an energy metric,the total energy consumption almost reached a value of 35,000 Joules; whereas, having an energy metric fixed at a maximum energy allowed value of 61 Joules reached a maximum of 23,406 Joules.

There exists a difference of 12% between 23,406 and 35,000, hence we can conclude that utilizing a maximum energy constraint is 12% more energy efficient than not using our proposed energy metric.

d. The Influence of the Maximum Energy Allowed on theRequest Acceptance Ratio

1. Simulation Setup

The following set of simulations show the effect of the maximum energy allowed (M) on the number of accepted requests. Similar to the previous set of simulations, all the resources of our model have been reset to high values, in order not to affect the request acceptance ratio in any way. We have run 2000 requests at 9 different times, each time modifying the value of the maximum energy allowed.

During the first run we set the value of the maximum energy allowed to the most energy consuming path of the model, thus not limiting our model by any metric. During the following runs we set the value of the maximum energy allowed to values of 24, 34, 39, 44, 49, 61, 65, and 75 Joules respectively.

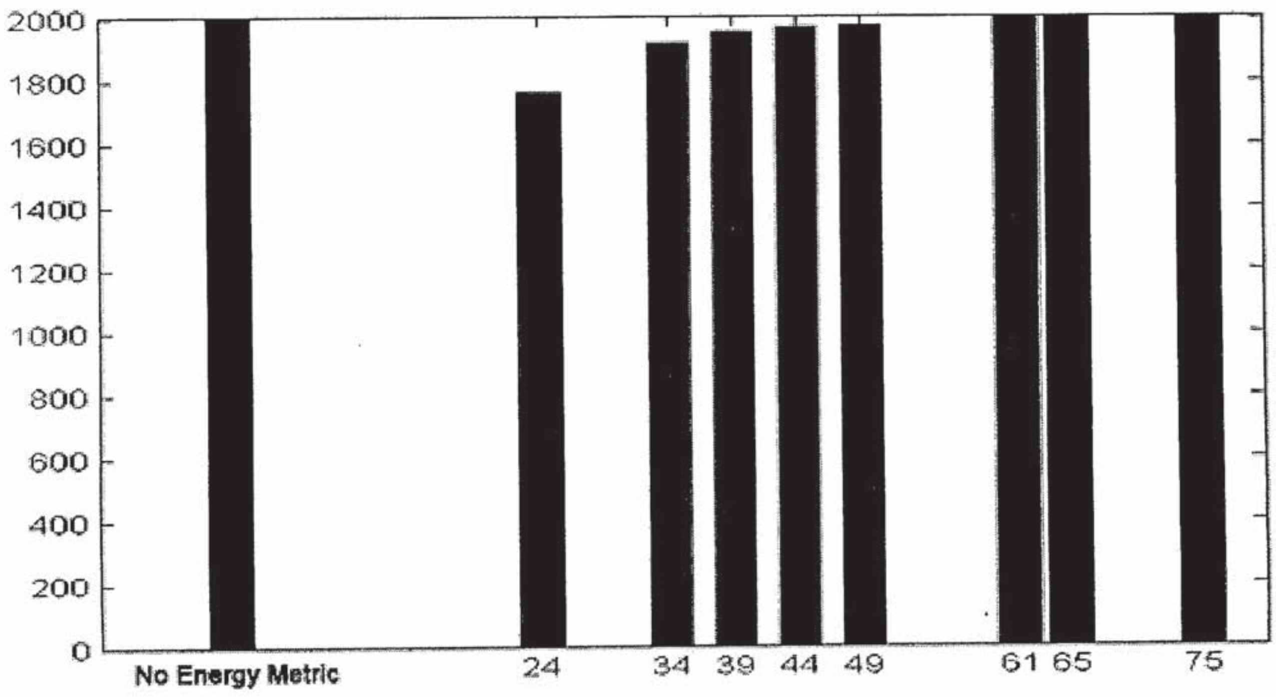

Figure 8 - Number of Accepted Requests vs. Maximum Energy Allowed

2. Simulation Results

As seen through the previous figure, constraining the routing algorithm with a specific maximum energy (M) (in our case 61 Joules) minimizes the total energy consumption. In Figure 8, we can see that at a maximum energy of 61 Joules,all of the requests have been accepted, thus deducing that the Maximum Energy metric does not hinder the acceptance ratio of the network.

Even though, without using the maximum energy constraint, the request acceptance ratio is also at 100%, but as seen in Figure 7 the difference in total energy consumption is grave. The total energy consumption for the case with no specified energy constraint is 35,000 Joules, while the total energy consumption for the case with a 61 Joules constraint is 23,406 Joules.

Furthermore, at a maximum energy of 34 Joules, our algorithm yields 1920 accepted requests over 2000 generated requests; i.e. 95% acceptance ratio. Yet it decreases the total energy consumed from 23,406 Joules to 19,854.6 Joules; i.e. roughly a 16% decline in total energy consumption. Thus, concluding that utilizing the energy metric will significantly decrease the total amount of energy consumed; nonetheless, maintaining a decent acceptance ratio.

It is up to the service providers and network administrators to decide upon the value of the energy metric, bearing in mind the performance of their network and its desired characteristics.

This page is intentionally left blank

Chapter 6 - Conclusion

In this thesis, we have presented ICT as one of the main stakeholders of Global Warming and have introduced some of the approaches that have been used in order to minimize the footprints of ICT on environmental degradation and pollution. In the beginning of chapter two we have introduced some of the main services that cloud computing provides and discussed the energy consumption of the cloud and its different sectors.

We have shown that the cloud network is one of the main contributors to ICT's carbon footprint, hence weintroduced research germane to this topic anddiscussed about solutions that have been applied in order to minimize energy consumption at the network and telecommunications level.

Chapters one to four form an introduction and serve to provide a background to better understand our study. They also reveal the motivation and context upon which the studies were thought of and conducted. There on, Chapter four and five present our model and our studies and prove the efficiency and purpose of our proposal.

In this thesis, we haveproposed a way to minimize power consumption by including power consumption metric in network resource allocation algorithms and have proposed an exact approach based on MILP formulation, where we evaluate the energy consumption per request and compare it to a maximum allowed value (M).

In chapter four we have shown that our model relies on the fact that requests have different arrival times, so that they get scheduled with the minimum energy consumption possible.

We have simulated our model over a modified NSFNet 18-nodes model with 4 data centers, which is described in details in chapter four. Our results, in chapter five, have shown that the value of M highly affects the total energy consumption of the network, while slightly affecting the acceptance ratio of requests.

This allows NSPsto study the performance of their networks and adjust the value of M so that a maximum number of requests are satisfied with minimum energy consumption.

Some NSPs might maintain a 100% acceptance ratio; however, using our metric to route the requests efficiently; others, might want to decrease the acceptance ratio further, thus achieving lower energy consumption levels.

As future work, the model presented in this model could be enhanced to include the power consumption of the processing server that would be residing in the respective data center.

Furthermore, the requests could be detailed further into being either storage requests or processing request. The links in the model could be enhanced to deliver both request and response messages; thus include the response's power consumption in the energy calculation.

Green Computing will be one of the key factors in the fight against Global Warming since the Information and the Communication Technology is advancing rapidly and the sector is exhibiting tremendous growth which will record even more success along the years.

This success, as stressed in this thesis,will sadly be accompanied by a great demand of energy and any carelessness could lead to a dark legacy imprinted on the planet.

It is our responsibility as computer scientists, and technicians and engineers in the related fields to make sure that our legacy has the constructive color of green and not the destructive color of grey.

With technology in general and communication networks in specific being the decisive factor of our present era, such energy saving and efficient proposals and methods should be employed on international levels and as protocol in order to achieve a common international goal and avoid any environmental crisis. Instead of being left to the business oriented decisions of service providers.

Reduce is as essential as the other two R-s in the Reduce, Recycle, Reuse trinity.

This page is intentionally left blank

List of References

[1] A. Beloglazov, R. Buyya, Y. Choon Lee, A. Zomaya, "A Taxonomy and Survey of Energy-Efficient Data Centers and Cloud Computing Systems", Proceedings of the Advances in Computers, Vol. 82, 2011.

[2] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared", Proceedings of the IEEE GCE Workshop, 2008.

[3] M. Priya, Sh. Puneet, B. Sujata, R. Parthasarathy, "A Power Benchmarking Framework for Network Devices", Proceedings of HP Labs, 2010.

[4] A. Adelin, P. Owezrski, Th. Gayraud, "On the Impact of Monitoring Router Energy Consumption for Greening the Internet", Proceedings of the 11th IEEE/ACM International Conferece on Grid Computing, 2010.

[5] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, S. Wright, "Power Awareness in Network Design and Routing", Proceedings of Cisco Systems.

[6] J. Baliga, R. W. A. Ayre, K. Hinton, R. S. Tucker, "Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport", Proceedings of the IEEE, Vol. 99, No. 1, 2011.

[7] S. K. Garg, Ch. Sh. Yeo, R. Buyya, "Green Cloud Framework for Improving Carbon Efficiency of Cloud", Proceedings of the Euro-Par, 2011.

[8] M. Scott, R. Watson, "The Value of Green IT: a Theoretical Framework and Exploratory Assessment of Cloud Computing", Proceedings of the 25th Bled eConference, 2012.

[9] L. Xu, Z. Zeng, X. Ye, "Multi-objective Optimization Based Virtual Resource Allocation Strategy for Cloud Computing", Proceedings of the 11th International Conference on Computer and Information Science, 2012.

[10] Ch. Yang, K. Wang, H. Cheg, Ch. Kuo, W. Ch. C. Chu, "Green Power Management with Dynamic Resource Allocation for Cloud Virtual machines", Proceedings of the IEEE International Conference on High Performance Computing and Communications, 2011.

[11] A. Beloglazov, R. Buyya, "Adaptive Threshold-Based Approach for Energy – Efficient Consolidation of Virtual Machines in Cloud Data Centers", Proceedings of the ACM, 2010.

[12] A. Beloglazov, R. Buyya, "Energy Efficient Allocation of Virtual Machines in Cloud Data Centers", Proceedings of the IEEE, 2010.

[13] A. Beloglazov, R. Buyya, "Energy Efficient Resource Management in Virtualized Cloud Data Centers", Proceedings of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010.

[14] A. Beloglazov, J. Abawajy, R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing", Proceedings of Elsevier B.V., 2011.

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

[16] A. Beloglazov, R. Buyya, Y. Ch. Lee, A. Zomaya, "A Taxonomy and Survey of Energy-Efficient Data Centers and Cloud Computing Systems", Proceedings of the Advances in Computers, Vol. 82, 2011.

[17] A. Beloglazov, R. Buyya, "Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers", Proceedings of Concurrency and Computation: Practice and Experience, 2011.

[18] M. A. Salehi, P. R. Krishna, K. S. Deepak, R. Buyya, "Preemption-aware Energy Management in Virtualized DataCenters".

[19] "Council Rock Schools in Pennsylvania Save $8.8M on Energy", CISCO Customer Care Study, 2012.

[20] "Mobile's Green Manifesto 2012", GSMA Association, 2012.

[21] Z. Zhu, "A Novel Energy-Aware Design to Build Green Broadband Cable Access Networks", Proceedings of the IEEE Communications Letters, Vol. 15, No. 8, 2011.

[22] B. Lannoo, "D8.1. Overview of ICT energy consumption", Proceedings of the EINS Consortium, 2013.

[23] Parliamantary Office of Science and Technology, "ICT AND CO_2 EMISSIONS", proceedings of Postnote, No. 319, 2008.

[24] Greenpeace International, "COOL IT LEADERBOARD VERSION 6: APRIL 2013", JN 445, 2013.

[25] Greenpeace International, "How Clean is Your Cloud?", JN 417, 2012.

[26] Greenpeace International, "A Clean Energy Road Map for Apple. How AppleCan Meet its Coal-free Goal", JN 417 Update,2012.

[27] Greenpeace International, "Silent Killers. Why Europe must replace coal power with green energy", JN 449, 2013.

[28] E. Castillo, A. J. Conejo, P. Pedregal, R. Garcia, N. Alguiacil, "Building and Solving Mathematical Programming Models in Engineering and Science", Pure and Applied Mathematics Series, Wiley, New York, 2002, pp. 1-42.

[29] J. W. Chinneck, "Practical Optimization: a Gentle Introduction", 2004, Chapter 18.

[30] Th. S. Ferguson, "Linear Programming. A Concise Introduction".

[31] A. S. Tanenbaum, "Computer Networks, Fourth Edition", Prentice Hall, 2003.

[32] D. Mehdi, K. Ramasamy, "Network Routing: Algorithms, Protocols, and Architectures", Morgan Kaufmann Publishers, 2007.

[33] R. Baumann, S. Heimlicher, M. Strasser, A. Weibel, "A Survey on Routing Metrics", TIK Report 262, Proceedings of the Computer Engineering and Networks Laboratory, 2007.

[34] I. Stojmenovic, X. Lin, "Power-aware localized routing in wireless networks", Proceeding of SITE University, 2000.

[35] L. M. Feeney, "Mobile Ad Hoc Networking", IEEE Press, 2004, pp. 301–327.

[36] http://www.neos-server.org

[37] P. Hoeller, M. Wallin, "OECD Economic Studies No. 17, Autumn 1991. Energy Prices, Taxes and Carbon Dioxide Emissions", OECD website, p. 92, 1991.

[38] M. J. Poterba, "Tax Policy to Combat Global Warming: On Designing a Carbon Tax", Cambridge, MA:MIT Press, 1991.

[39] "South Africa Gears Up for Carbon Tax", CPC News, June 2010.

[40] "China Ministries Propose Carbon Tax", Alibaba News, May July 2011.

[41] C. E. Abosi, R. Nejabati, D. Simeonidou, "A Novel Service Composition Mechanism for the Future Optical Internet", IEEE/OSA Journal on Optical Communications and Networking, Vol. 1, No. 2, July 2009.

[42] Z. Guo, B. Malakooti, "Energy Aware Proactive MANET Routing with Prediction on Energy Consumption", IEEE, 2007.

[43] M. A. Youssef, M. F. Younis, K. A. Arisha, "A Constrained Shortest-Path Energy-Aware Routing Algorithm for Wireless Sensor Networks", IEEE, 2002.

[44] A. Benslimane, R. E. Khoury, R. E. Azouzi, S. Pierre, "Energy Power-Aware Routing in OLSR Protocol", IEEE.

[45] A. R. Swain, R. C. Hansdah, V. K. Chouhan, "An Energy Aware Routing Protocol with Sleep Scheduling for Wireless Sensor Networks", 24th IEEE International Conference on Advanced Information Networking and Applications, 2010.

[46] D. Mehdi, K. Ramasamy, "Network Routing Algorithms, Protocols, and Architectures", 2007.

[47] A. K. Sidhu, S. Kinger, "Analysis of Load Balancing Techniques in Cloud Computing", International Journal of Computers & Technology, Volume 4 No. 2, March-April, 2013, ISSN 2277-3061