Towards Energy Efficient Cloud Data Centers - A Framework for Evaluation and Analysis of Energy Efficiency

Khalid Ahmed Farah

Thesis submitted for the degree of
Master in Applied Computer and Information Technology - ACIT
(Cloud-based Services and Operations)
30 credits

Department of Computer Science
Faculty of Technology, Art and Design

Oslo Metropolitan University — OsloMet

Spring 2024

Towards Energy Efficient Cloud Data Centers - A Framework for Evaluation and Analysis of Energy Efficiency

Khalid Ahmed Farah

Towards Energy Efficient Cloud Data Centers -
A Framework for Evaluation and Analysis of Energy Efficiency

http://www.oslomet.no/

Printed: Oslo Metropolitan University — OsloMet

Preface

This thesis covers an experiment done with a developed framework to analyze and evaluate cloud data centers, and was developed at Oslo Metropolitan University in the spring of twenty twenty-four as part of the master thesis project.

The supervisor of the thesis is Raju Shrestha, which proposed the interesting thesis topic of energy efficiency in cloud computing. With the project investigating methods of enhancing energy efficiency in cloud data centers through insight and information.

With my experience of the work mostly being analysis, evaluation and management of data extracted from simulated environments in addition to mathematical formulas for estimations.

Abstract

The purpose of the research questions proposed in the thesis were to investigate what methods and combinations could illustrate areas of energy inefficiency at the analyzed and evaluated cloud data centers. However, also investigating how the methods utilized affect other monitored data center devices. With the research questions as a result illustrating of areas of energy inefficiency and what actions a cloud provider could take to reduce energy consumption, environmental impact and operational cost.

The project was scoped to the targeted areas compute, storage, and network. With the compute category targeting the data center physical hosts, while storage was scoped to the data center disks, and for the network scoped to switch devices. With areas such as the data center cooling being left out of the scope of the project due to it's complexity and time requirements on top of compute, storage, and networking.

With the work done in the thesis being exploration of prior work and background information regarding energy efficiency in data centers, with the work further developing an analysis and evaluation framework. Where the resulting framework could analyze a cloud provider's cloud data centers to provide insight into resource utilization and placement of workloads. In addition to further evaluating the cloud provider's cloud data center by estimating the reduction of energy consumption with different methods.

As a result of the work it was discovered that not only did the framework proposed methods which could potentially reduce power consumption at cloud data centers and thus energy consumption. The framework also through the provided insight could inform the user of actions that could improve the energy efficiency through improved compatibility between workloads and devices. Where the placement methods utilized led to greater energy efficiency for the monitored devices with reorganizing placement approach which additionally took into account the over utilization; led in the case of the physical hosts to an increased energy consumption at eight point forty-two percent as supposed to zero point zero two percent reduction in energy consumption. while the affected data center network reduced energy consumption by thirty-three point eighty-five percent and twenty-nine point forty-four percent. With the analysis and evaluation also illustrating a lack of available memory resources as the workloads consumed more RAM than CPU.

The thesis thus concluded by summarizing the work done, but also suggesting improvements to the framework such as the implementing

dynamic power scaling approaches to data center disks and network links. Additionally did the thesis also propose future work such as testing the solution in a real data center environment and implementing analysis and evaluation for the data center cooling.

Acknowledgments

I would like to thank my supervisor Raju Shrestha for giving me the opportunity to work on the proposed thesis topic of energy efficiency in cloud computing and supervising me during the project regarding the thesis work, writing, project structure and planning.

I would also like to thank Hårek Haugerud for providing the thesis latex template utilized when writing the thesis content and findings in Overleaf.

Lastly, I would also like to thank the Institution; Oslo Metropolitan University for providing resources to conduct the experiments done the in the thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Problem Statement

The motivation for the thesis is to improve the energy efficiency of a cloud provider's cloud data centers. However, the motivation stems from the possible confusion of where and how the energy efficiency can be improved at each of the data center's infrastructure.

This is important as the cloud computing paradigm is a growing industry and field of research, which provides virtualized applications and services on demand. Where the virtualized applications and services with their varying resource request require substantial processing capacity leading to an increased energy consumption. (Berl et al., 2010; Hameed et al., 2016; Mastelic & Brandic, 2015; Reddy & Reddy, 2023)

The work will thus attempt to improve the energy efficiency through improved observability and clarity. With the end goal leading a reduced energy consumption to reduce the data center operational cost and the environmental impact from energy consumption.

The problem addressed in the thesis thus is the lack of insight and pathways for energy efficiency improvements through evaluation and analysis of the cloud provider's infrastructure at each data center. More specifically targeting the aspects computing, storage, and networking. (Bruno & Jordan, 2012; Tate et al., 2013)

The reason why this is a problem; is due lack of insight into the cloud provider's current infrastructure. Resulting in a lack observability in possible improvements through optimization of device specification and system function. Another reason why this is a problem is the environmental impact from the energy consumption of data centers and the growing operational cost from the data center energy consumption.

The thesis addresses this as to demonstrates areas of optimization in data centers. With the goal of improving the energy efficiency through dynamic reduction of power consumption and further energy consumption from data centers.

1.2 Objectives

With the problem stated in the aforementioned section, a primary objective of the thesis being to evaluate and analyze each of a cloud provider's cloud data centers with the goal of illustrating areas improvement based on energy efficiency through power and energy consumption of devices by optimizing resource utilization. Another primary objective of the thesis in addition to investigate areas of improvement for energy efficiency; will be to illustrate the results to the user. this is to inform the user of possible actions to reduce their current energy consumption and operational costs.

While a secondary objective of the thesis will be to utilize real workloads, and cloud data centers through real device models. With the goal of simplifying the real adaptation of the thesis work in addition, to the accuracy of the result to physical cloud data centers.

Another secondary objective would be for the solution to produce feedback based on the results. The reason for this would be to get specific results from the evaluation of the cloud provider and their data centers, with short consolidated feedback regarding areas of concern. With the areas of concern being system or devices that are energy inefficient.

1.3 Research Questions

Based on the aforementioned problem statement and objectives of the thesis the following are the defined research question investigated for the thesis:

- What approaches reduce the energy consumption more separately and could approaches be combined to further improve energy efficiency at cloud data centers?
- How does the resulting architecture of different systems affect the energy efficiency of other systems?
- What actions can be taken to improve utilization of resources statistically through detailed cloud data center insight?

1.4 Contribution

Based on the defined objectives; the work contributes to the research of energy efficiency in cloud computing through an analysis and evaluation framework, which utilizes the monitored data from devices. The thesis will thus provide a framework that provides additionally observability in an cloud data center infrastructure of cloud Information Technology equipment such as servers, disks, and network switches. With the framework further contributing by providing information regarding areas of energy inefficiency at each cloud data center and possible improvements to energy efficiency through dynamic energy management approaches.

With the framework mostly contributing by informing cloud providers through more detailed energy reduction for a reduced environmental impact and operational cost.

1.5 Thesis Outline

- The second chapter of the report is the "Background and Literature Review" chapter. The chapter will cover the background information for the thesis and the research within energy efficiency to explain what has been previously done in the research field. Further the chapter will cover the literature review of previous and related work, grouping the works and describing what the groups achieve and contribute. the literature review also describes the limitations of the groups and how the proposed work of the thesis will differ as to justify the work done in the thesis.
- The third chapter of the thesis is the "Methods" chapter. The Methods chapter will cover the research methods utilized to conduct the literature study for the literature in chapter 3. Further the chapter will also cover the tools and techniques utilized to during the project to describe their purpose. Additionally, the will chapter discuss the contingency plan of the thesis and the potential risks.
- The fourth chapter namely "The Proposed Work" chapter will first and foremost cover the idea of the project. The chapter will thereafter note the details of the solution through specifications for the project. Lastly, the chapter will cover formulas and equations used in proposed.
- The fifth chapter is the "Experiment" chapter. The experiment chapter will cover the experiments done throughout the thesis in detail and the subsequent results related to the problem statement of the thesis.
- The sixth chapter is the "Discussion" chapter. In the Discussion chapter the findings and the work presented in chapter five will be discussed. The chapter will then discuss the limitations and issues before discussing future improvements to the proposed work. Lastly, the chapter will answer the problem statement presented and the research questions of the thesis.
- The last chapter is the seventh chapter, namely the "Conclusion" chapter. The conclusion chapter will conclude the thesis by shortly summarizing the work done in the thesis.

Chapter 2

Background and Literature Review

This chapter will first go over background information related to the thesis. The chapter will then cover related work in the literature review where limitations and differences of work are presented to justify the proposed work.

2.1 Background

Cloud computing is a growing industry in IT, where computing resources such as CPU, RAM, Bandwidth and Storage through IT equipment distributed over data centers which are provided on demand to clients by cloud providers with a pay-as-you-go model. Client use these resources to then run as an example a web application, without needing to manage their own servers and IT equipment. Cloud computing thus provides the client to scale their application to the current business demand and providing availability and fault tolerance through replication of data and application in different locations. (Dabbagh et al., 2015)

Alternatives in cloud computing is also the source of computing resources; through the cloud data center models "public", "private", "community", and "hybrid". Where the public cloud data center model regards the use of just publicly available data centers where hardware is shared between unknown clients. While the private cloud data center model entails the use of just privately owned data centers and thus managed and owned data center hardware. The community cloud data center model on the other hand, utilizes a privately owned data center that is shared between known parties. Lastly, the hybrid cloud data center model which combines the public and private models; using the public and private data centers. (Diaby & Rad, 2017; Tate et al., 2013)

Alternatives for how cloud resources are provided are through the models "Infrastructure as a Service", "Platform as a Service", and "Software as a Service" or "IaaS", "PaaS", and "SaaS" for short. Where the Infrastructure as a Service model provides the client of the cloud provider the most responsibility as the resources and services are directly available through

virtual machines managed by the client. Where the virtualized machine is running on the cloud data centers physical hosts with its hardware abstracted and an operating system to run applications and services. While the Platform as a Service model on the other hand lets the cloud provider manage the instances the platform is running on, that a client can customize for their application like Microsoft Azure's SQL database. Lastly, the Software as a Service model the not only the instance but also the platform of the service is being managed by the cloud provider, providing the client an application like Google Docs. (Diaby & Rad, 2017; Tate et al., 2013)

However, with the growing reliance on cloud computing a concern is the growing energy consumption of cloud providers' data centers through equipment like physical hosts, storage servers, switches, cooling and so on. With the data centers reaching a projected twenty percent of global energy consumption and five point five percent of the world's carbon emissions by the year twenty twenty-five. Due to this, research on energy efficiency in cloud computing had begun to reduce energy and power consumption of data center through alteration of systems for managing workloads, storage, network, cooling, and other systems like data center lighting. While attempting to maintain or improve current performance and Quality of Service for the clients of the provider. A reduction of power consumption of devices would thus lead to a reduction in energy consumption of the devices. Further leading to a reduced environmental impact, but also reduced the operational costs of data centers. (Bruno & Jordan, 2012; Buyya et al., 2024; Hameed et al., 2016; Hanafy et al., 2023; Tate et al., 2013)

Regarding the reduction in power and energy consumption; management of power in cloud computing can be split into two main categories for reducing power consumption "Static Energy Management" and "Dynamic Energy Management" or "SEM" and "DEM" for short. Where both methodologies target both the hardware and software level of managed devices. (Karpowicz et al., 2016a)

The static energy management category typically contains methods where energy consumption is optimized statically through the use of low power devices and nano processors. While approaches under the category of dynamic energy management attempt to more dynamically optimize power consumption through mainly power supply modulation and deactivation of idle devices. which are distinguished as "Dynamic Power Scaling" which reduces the power consumption by reducing the performance of the device adapting it to the required load and "Smart Standby" which on the other hand leverages the idle mode capabilities to power down unused components or devices. (Karpowicz et al., 2016a)

Examples of Dynamic Power Scaling are "Adaptive Rate" and "Dynamic Voltage Frequency Scaling" or AR and DVFS for short. Where in the case of AR reduces the energy demand of a network by scaling the processing capabilities of a device, or its transmission or reception speed of its network interface. An example of Smart Standby is "Lower Power Idle" or "LPI", which puts a device or its components into low power mode during short

periods of inactivity. With personal computers implementing "Advanced Configuration and Power Interface" or "ACPI", a combination of both AR and LPI to reduce energy consumption by defining multiple power aware states through the scaling of processor voltage, clock frequency, and idle states when processor is in standby. (Karpowicz et al., 2016a)

2.2 Literature Review

Within the field of energy efficient cloud computing a data center's energy and power consumption can stem from their computing systems, storage systems, networks, cooling, and additional electronics in data centers. With related work exploring approaches to improve energy efficiency by focusing on reducing the energy consumption while maintaining or improving performance by altering the algorithms of the computing systems, storage systems, and networking systems. (Bruno & Jordan, 2012; Buyya et al., 2024)

2.2.1 Computing System

Within cloud computing, according to the authors of "Energy-efficiency and sustainability in new generation cloud computing: a vision and direction for integrated management of data centre resources and workloads" one of the main devices that draw data center power are the data center servers. The authors in this grouping of papers attempt to improve energy efficiency by altering the allocation of resources algorithm by placing client virtual machines in a manner to efficiently utilize a physical host machines. (Buyya et al., 2024)

Within this group an approach to improve energy efficiency is through optimization of resource utilization, the resources mostly targeted within this approach is the CPU, RAM, and bandwidth usage. The goal of this approach is to avoid over utilization of active physical hosts by rebalancing the workload to other hosts and consolidating virtual machines to as few hosts as possible. While over utilization may lead to more energy efficiency researchers avoid it due to the damage it causes to the hardware leading to performance degradation of the host's hardware. (Hameed et al., 2016)

Another goal of the approach was to avoid under utilization of active physical hosts by powering off, putting the device to sleep, or putting host in hibernating mode for under utilized or idle physical hosts to reduce energy consumption. The reason for this by the grouping was because of idle physical hosts consuming fifty percent of the maximum power usage of the host, thus wasting computing power. With other works e.g. "Energy-efficiency and sustainability in new generation cloud computing: a vision and direction for integrated management of data centre resources and workloads" stating that physical hosts when idle consume thirty percent of maximum energy consumption. (Buyya et al., 2024; Hameed et al., 2016; Mastelic & Brandic, 2015)

Another approach to reduce power consumption of computing systems is by altering the hosts' hardware rather than load-balancing the workloads. However, similarly will the approach also put idle hosts in hibernating mode, but for under utilized hosts reduce their power consumption by incrementally powering off redundant hardware parts. Examples of such mechanism are "SpeedStep", "PowerNow", and "Cool'nQuiet". (Berl et al., 2010; Mastelic & Brandic, 2015)

With other similar approaches done by the authors of "Energy-efficiency and sustainability in new generation cloud computing: A vision and directions for integrated management of data centre resources and workloads", which aimed at reducing the energy consumption of computing systems through "Dynamic Voltage and Frequency Scaling". Where the technique was used to manage the power state of CPUs, GPUs, TPUs, and memory to reduce their clock speed thus reducing the power consumed. (Buyya et al., 2024)

The proposed work differentiates from the current literature regarding the computing systems; by targeting the resulting infrastructure arranged by the management system and workload loadbalancer through analysis and evaluation. With the proposed work investigating the physical hosts within each of the cloud provider's cloud data centers, to focusing on the resource utilization from the hosts CPU, RAM, and bandwidth and the servers' power and energy consumption.

With the proposed work more specifically analyzing the cloud provider's servers; targeting the usage of resources and placement of workloads. With the proposed work analyzing the utilization of CPUs, RAM, and bandwidth of the cloud provider's servers. With the workload scheduler being analyzed to investigate the placement of workloads of data center hosts. With the analysis of the work differentiating from related work by illustrating improper placement of workloads and compatibility between workloads and resources; resulting in resource wastage and energy inefficiency.

The proposed work will also differentiate itself from prior work by applying multiple approaches, to investigate the improvements to energy efficiency by evaluating the monitored hosts. With the evaluation further building on the analysis by investigating potential improvements such as the host state and energy consumption. With the evaluation also investigating the resulting energy consumption and resources utilization due to altered placement of allocated workloads through algorithms of consolidation and reorganization of CPU, RAM, and bandwidth usage.

Lastly, the work in regard to the physical host will also differentiate itself from prior work by compare the methods of potential energy reduction in addition to investigating combinations. Which will be done to point out areas of improvement for energy efficiency and present methods with inefficient results.

2.2.2 Storage System

A second group of papers investigated during the literature study of energy efficiency in cloud computing is the storage system of a cloud data centers that is responsible for managing the storage of data for clients. For this system the metrics used as input were the CPU-, RAM-, Disk-, bandwidth-, I/O usage with the number of active and inactive nodes, and additional metrics such as the number of replicas for a file to store. while the metrics evaluated is the power consumption of the work, the data rebuild rate in case of failure, the running time of tasks, number of nodes sleeping, and the number of "SLA" or "Service Level Agreement" violations.

An approach within this grouping is by powering off and redirecting Input and Output load to another storage server as done with the computing system. However, as seen in the paper "eStor: Energy efficient and resilient data center storage" the authors were concern over the structure of storage but also placement of replicated data for resiliency and fault tolerance in the case of disk failure. (Lin et al., 2011)

While the approach done by authors of "Profit-based file replication in data intensive cloud data centers" similarly value the importance of file replicas. the authors instead attempt to use the faster access time of files due to replicas to reduce energy consumption of cloud data storage. (Alghamdi et al., 2017)

On the other hand, authors such the ones from the paper "Energy Efficient Storage Management Cooperated With Large Data Intensive Application" forego the replication of data. Rather attempting to improve energy efficiency through an in-depth storage management system. The proposed system improves the energy efficiency through the placement of data and similarly to the work eStor, the run state of disks during runtime. However, while the work lead to greater energy efficiency than the works it was compared to, the authors note that the solution leads to performance degradation. (Lin et al., 2011; Nishikawa et al., 2012)

Another approach brought up by the authors of "Decreasing power consumption with energy efficient data aware strategies" which focuses on the data placement in regard to the location of the stored data. With the goal of the authors being a reduction in energy consumption by reducing the amount of data transmitted storage required. However, the author did in addition state disks were not power down when idle or underutilized leading of a loss of efficiency. (Vrbsky et al., 2013)

Similar to grouping of computing systems the approach altering the CPU clock rate to reduce power consumption was also utilized in papers such as "A System for energy-efficient data management", which target database management systems to reduce energy consumption. The authors achieved this by altering the query optimizer and utilizing Dynamic Voltage Frequency Scaling. With the limitation of their work being specific to database applications. Additionally, did the authors not power down unused disk but utilize the different modes of their available disks which could lead to energy inefficiency. (Mastelic & Brandic, 2015; Tu et al., n.d.)

Similarly to the papers targeting the physical hosts or data center server, the proposed work of the report will differ from previous work by analyzing and evaluating the monitored infrastructure state. With the proposed work focusing on the capacity utilization, power and energy consumption of persistent storage's disks for each data center.

More specifically, the proposed work would analyze the cloud provider's persistent storage with it targeting the utilization and the scheduling of data placed. With the analysis of the utilization targeting the disk capacity utilized and the power state of the disks. While the analysis of the placement of data will investigate the placement data on the cloud provider's cloud storage per disks. With the proposed work differentiating itself by illustrating areas of improper placement of data leading to energy inefficiency but also areas of improper management of disk power states due to utilization and data placement, illustrating the efficacy of current storage systems and devices.

Similarly will the work also differentiate by applying multiple approaches to investigate improvement to energy efficiency by evaluating the monitored cloud provider's data centers and storage disks. With the evaluation following up on the analysis by investigating potential disks states to be altered with the goal of energy reduction additionally, evaluating the placement of data on the cloud storage disks and how it would disks state and the resulting energy consumption and available resources through algorithms for consolidation and reorganization of disk capacity.

With the proposed work lastly differentiating from previous work by combining and comparing methods. With the goal of determining ways of improving the energy efficiency of cloud data center storage. With the method similarly illustrating energy inefficient methods.

2.2.3 Network System

Another collection of papers regarding energy efficiency in cloud computing is also networking, as networking devices such as routers, switches and their ports and links in data centers consumes up to eight percent of power according to the authors of "Energy-efficiency and sustainability in new generation cloud computing: a vision and direction for integrated management of data centre resources and workloads" and are crucial for Internet access and communication between physical hosts. Where the approaches aim to improve the energy efficiency by targeting the links of networking devices. (Buyya et al., 2024; Yao et al., 2023)

One such group of approaches similarly to the works in computing systems aim to reduce energy consumption by powering off equipment or their hardware components. An example of this is the work done by the authors of "An Energy Saving Routing Algorithm For a Green OSPF Protocol" where the OSPF protocol is reused to get the link states and further used to calculate the shortest path to information communication devices with the unused or unnecessary links being disabled. While the work done by the authors of "An Energy-efficient Networking Approach

in Cloud Services for IIoT Networks" rather make their own algorithms by determining fitness values for each path to determine the most energy efficient paths. (Cianfrani et al., 2010; Jiang et al., 2020)

With the work by "A Dynamic traffic-aware energy-efficient algorithm based on sleep-scheduling for autonomous systems" taking it step further by also reusing the routing protocol "OSPF" to reduce the need for link state monitoring: With the work differentiating itself by also powering down nodes in addition to unused or underutilized links to further reduce energy consumption. However, the author does keep additional nodes available, with the reasoning unclear. (Dabaghi-Zarandi & Movahedi, 2018)

On the other hand researchers have also investigated hardware approaches to reduce energy consumption of networks in general; with one such approach altering the Ethernet adapter to improve its energy efficiency as prior work has focused on improved the input and output of the network adapter. This finally led to the introduction of "Low Power Idle" mode or "LPI" to keep devices synchronized and reduce power consumption when the device is not transmitting data over a link. With the approach rather focusing on reducing the energy consumption without powering off nodes and links. (Zeadally et al., 2012) Another hardware approach is also altering the network adapter but, rather than attempting to reduce the power of a link that is powered on; the approach rather sends a burst of packets to minimize total energy consumption. However, as stated by the authors of "Energy-efficient networking: past, present, and future", this comes at the cost of a higher packet delay as a trade off. (Zeadally et al., 2012)

The works within this grouping typically measures the power consumption of devices through their utilization of resources or more specifically link usage of its bandwidth rather than monitoring the power usage itself. In addition the works focus on the switches rather than devices such as routers. With the work being evaluated in data centers or in simulation with tools such as "NS-3" or "GreenCloud".

The proposed work will also differentiate from related work by analyzing and evaluating the resulting cloud data center network infrastructure state, such as networking devices like switches managed by the data center management system related to the network and their links and ports. With the proposed work focusing on the power and energy consumption, and bandwidth utilization the monitored network devices and their links and ports.

With the proposed work more specifically analyzing the cloud provider's data center networks, targeting the utilization of network resources and the network traffic on the cloud data center network devices. With the analysis of the utilization focusing on the bandwidth utilization and the power state of network device, links, and ports. Where the analysis of the network traffic illustrating the number of data sources transmitting data through a network device. With the proposed work differentiating

from related work by illustrating improper placement of workloads and management of network device power state and their ports and links leading energy inefficiency.

With the evaluation by the proposed work similarly applying multiple approaches with the goal of investigating areas of energy inefficiency by illustrating ways of improving energy efficiency of the cloud provider's monitor data center networks. With the evaluation specifically evaluating the state of network device and the state of their ports. With the evaluation additionally investigating the effect of the scheduling approaches applied on physical hosts and disks on the cloud data center network and the potential network utilization, and energy and power consumption due to the scheduler. While the evaluation utilizes similar approaches, the evaluation differentiates itself from prior work by utilizing the approaches to illustrate areas of energy inefficiency and the potential effect on the utilization of resources and energy consumption if the methods were applied.

With the evaluation further differentiating itself by combining and comparing approaches. Where the comparison done by the proposed work is done to illustrate effective and ineffective approaches. With the combination applied to view the potential reduction in energy consumption.

2.3 Summary

This chapter covered the background information for the thesis covering information such what is cloud computing and the details within it such as the data center models and strategies for providing resources to users, with the section further detailing static and dynamic methods for reducing power consumption and thus energy consumption of devices. Further the chapter covered related work by authors describing their work and it's problems or limitation. With the sections further covering how the proposed work will differentiate from the related work to justify it.

Chapter 3

Methods

This chapter will cover the methodologies used during the project, covering what was used and why. More specifically covering the literature search questions for the literature review, the literature search keywords, the search engines used and the inclusion and exclusion criteria used for including and excluding found papers. Lastly covering the datasets utilized, the simulation tools used, and the programming language and modules utilized

3.1 Tools And Techniques

3.1.1 Literature Search Questions and Keywords

For the specified problem statement of analyzing and evaluating cloud data center infrastructure efficiency and shortcomings, results in the following literature search questions:

- What system in cloud data centers are investigated?
- How is the energy efficiency of data center systems evaluated?
- How is the energy efficiency determined?
- How are solutions evaluated?

As the aforementioned literature search questions regards what systems have been investigated and how they where evaluated and their efficiency determined.

Keywords used to find related work within energy efficiency in cloud computing were "Energy Efficiency", "Cloud Computing", "Computing", "Storage", and "Network". Where the keywords "Energy Efficiency" were used to find papers where the goal were to reduce the energy consumption, improving energy efficiency of the infrastructure used. The keyword "Cloud Computing" on the other hand is used to specify the research field investigated by the researchers, specifying that works should focus on areas such as the cloud environment and data centers. With the keywords "Computing", "Storage", and "Network" further specifying the systems

being investigated in the thesis, with "Computing" as an example leading to papers targeting the energy usage of physical hosts.

3.1.2 Search Engines

A search engine that will be used during the literature study was "Scopus" which is a search engine that specializes in finding peer reviewed papers which are validated by a third party. Another search engine that will be used is "Google Scholar" as it provides more papers than Scopus at the cost of the guarantee of all the papers being peer reviewed.

3.1.3 Literature Search Techniques

A literature search technique that will be used during the literature study is utilization of the literature search questions with the specified search engines to find a list of papers regarding the topic of energy efficiency in cloud computing. Additionally, will also the citations of a paper be investigated to further find prior related work within energy efficiency in cloud computing. While investigating more recent related work by searching for papers that has referenced the paper.

3.1.4 Inclusion and Exclusion Criteria

The first inclusion criteria required to be met for found paper is that the title should be relevant to the topic of the thesis which is energy efficient cloud computing and its systems. This is done by utilizing the specified literature search keywords. If this requirement is not met the paper will not be further investigated.

The second inclusion criteria is that the abstract and conclusion of the paper is investigated to determine its relevance to the topic. If this requirement is met the paper is added to the list of found papers to provide more information regarding a topic or other systems being evaluated and researched for energy efficiency. If the abstract and conclusion is not relevant the paper is excluded from the collected papers.

3.1.5 Datasets

"GWA-T-12 Bitbrains" is a dataset containing performance metrics from one thousand, seven hundred, and fifty virtual machine from a distributed data center run by Bitbrains. The dataset contains the following features a timestamp of the sample's record, number of CPU cores, the amount of CPU usage requested in mega hertz, CPU usage in mega hertz, CPU usage in percent, the amount of memory requested in kilobytes, memory usage in kilobytes, Disk read throughput in kilobytes per second, Disk write throughput in kilobytes per second, network received throughput in kilobytes per second, and network transmitted throughput in kilobytes per second. With a potential use case of dataset being input for analysis and simulated data

center workloads. (of Technology, 2024)

"Financial1" is an online Transaction Process dataset containing traces of read and write to disk made available by Ken Bates, Bruce McNutt, and the Storage Performance Council also known as SPEC. The dataset contains the following features Application Specific Unit, Logical Block Address, amount of bytes being written or read to storage, whether the task is to read from storage or write to storage, timestamp containing the start time of the trace, and lastly an optional field containing additional information regarding the trace. With a use case of the dataset being input of for analysis or simulated data center workloads. (Guo et al., 2023; UMassTraceRepository, 2023)

3.1.6 Simulation Tools

"CloudSim" is an simulation framework that can be used simulate and gather data on cloud computing operations such as the hosts and their running workloads. The framework more specifically is for modeling and simulating cloud computing infrastructure and services, with the goal being a generalized and extensible framework for experimenting and testing emerging cloud infrastructures and services. (Beloglazov & Buyya, 2012; Buyya et al., 2009; Garg & Buyya, 2011; Oracle, 2024)

Additionally, the simulation tool CloudSim is also extendable which has led to numerous frameworks such as "CloudSimDisk" which is a simulation tool to simulate storage systems in cloud environments. The tool thus allows for simulation of storage systems to get information regarding the power and energy usage, the storage usage, the active and inactive storage disks and their performance. (Louis et al., 2015)

A simulation tool which is more targeted towards the networking is "NS-3", The simulation tool NS-3 is a discrete-event open source network simulator developed for research and educational purposes. The simulator can be used to model networking devices and simulate traffic in a defined network infrastructure. (nsnam, 2024)

3.2 Programming Tools

A programming language that will be used is Python, for analysis and evaluation of cloud provider's cloud data centers. Python is a simple, easy to learn and flexible programming language, which provides the option of function based and object oriented programming through classes and methods. (Foundation, 2024)

The Python programming language will also use an assortment of modules for different capabilities, one such module is "Pandas". Pandas is a fast, powerful, and flexible open source module for data analysis and data manipulation, which can be used to managed inputted data from a data source. Another module for analyzing and managing data is "Numpy",

more specifically developed for scientific computing in Python. (Harris et al., 2020; McKinney, 2010; pandas development team, 2024)

On the other hand, the modules "Matplotlib" and "Seaborn" will rather be used for visualization of data. Additionally, will the modules be used for illustration of computed results through visualizations such as plots. (Hunter, 2007; Waskom, 2021)

Lastly with Python "Jupyter Notebook" which is for analysis of data, data science, scientific computing, computational journalism, and machine learning will be used to configure and arrange workflows. Jupyter Notebook provides this through an interface allowing users to split the executable script in parts, present graphs in the editor, and further add markdowns for explanations throughout the code. (Jupyter, 2024)

3.3 Summary

The chapter covered the methodologies used to conduct first and foremost the literature study and the literature review, by detailing the search questions formulated, keywords used, the search engines utilized, and the literature search techniques before stating the inclusion and exclusion requirement for found papers. Further the chapter covered the datasets and the simulation frameworks used during the thesis.

Chapter 4

The Proposed Framework

This the chapter will first and foremost cover the idea of the proposed framework. Further the chapter will cover the specifications of the framework based on the objectives defined in chapter 1.

4.1 The Framework

The framework of the thesis is an analysis and evaluation framework targeting a cloud provider's infrastructure per data center. Where the idea will be to analyze the resource utilization and power consumption of the current infrastructure and evaluate whether the energy efficiency can be improved through different methods. With the goal of informing the user potential improvements and how much of an energy efficient improvement can be done on current infrastructure.

The proposed framework will analyze and evaluate the cloud provider's infrastructure for each data center. More specifically the work will cover physical hosts utilized by workloads, disks used by workloads storing data, and switches transmitting data throughout and out of the data centers. With the proposed framework combining the metrics when discussing the cloud provider.

The proposed framework will obtain the data from the simulations "CloudSim", "CloudSimDisk" and a simple network simulation. Where information and metrics from the simulated devices will be extracted similarly to monitoring where data such as CPU usage is gather from a host and observed as an example. The proposed framework will thus utilize the extracted data from the simulations as a data source to analyze and evaluate.

Lastly, the proposed framework will analyze and evaluate the cloud provider's infrastructure through the programming language Python. With the framework in addition using the data source management module Pandas, the array analysis and management module Numpy, Matplotlib and Searborn for visualization of data, and Jupyter Notebook an interface for analysis and evaluation of data from the data source. (Harris et al., 2020; Hunter, 2007; Jupyter, 2024; McKinney, 2010; pandas development team, 2024; Waskom, 2021)

The proposed framework presented will thus contribute to the research of energy efficiency in cloud computing; by introducing an application providing insight into energy efficiency and resource utilization of a cloud provider's infrastructure at each data center. With the framework further contributing by highlighting areas of energy inefficiency with possible approaches to improve energy efficiency by estimating the potential reduction in energy consumption of monitored devices. With the goal being improved energy efficiency through dynamic energy management approaches in cloud computing from informed decisions.

The proposed framework will thus be structured into two parts, analysis and evaluation of a cloud provider's cloud data center infrastructures. With the structure further dividing into the realms Compute which will focus on physical hosts in data centers, Storage which will target the physical disks from physical storage server in data centers, and Network focus on network devices such as switches and their ports that interconnect physical hosts.

4.2 Specifications

The specification of the proposed framework is the following:

- The analysis and evaluation framework should be able to gather metrics from data center computing, storage, and network servers and devices.
- The analysis and evaluation framework should determine the maximum energy consumption and available resources of devices and the current additionally extracting the current workloads for hosts, data for disks, network traffic for switches.
- The analysis and evaluation framework should be able to evaluate the resource utilization of computing, storage, and network servers and devices to determine if the power consumption can be reduced.
- The analysis and evaluation framework should be able to evaluate the workloads and data placement within cloud data centers to determine if the power consumption can be reduced for physical hosts, disks and switches.

4.3 Formulas

A method that will be used to calculate the power consumption of a physical hosts based on the utilization of the CPU and the physical host's power model. In the case of the host's CPU utilization being a modulus of ten the formula in figure 4.1 will be used, where the "power_model" variable is an array of power values in watts for each tenth utilization percentage. However, in the case of the utilization being between power model values, the framework utilizes the formula 4.2, where the delta

between power models is used to calculate the power consumption of the utilization percentage. (Beloglazov & Buyya, 2012; Buyya et al., 2009; Garg & Buyya, 2011)

$$Utilization = CPU\_instructions\_used / CPU\_instructions\_total$$

$$Utilization\_index = Utilization * 10$$

$$Power = Power\_model[Utilization\_index]$$

Figure 4.1: Equation for getting the power of a physical host with from the utilization of the CPU when the value's remainder is zero when divided by one tenth, and the power model array an array of power in watts per tenth CPU utilization. (Beloglazov & Buyya, 2012; Buyya et al., 2009; Garg & Buyya, 2011)

$$Utilization = CPU\_instructions\_used / CPU\_instructions\_total$$

$$Utilization\_index = Utilization * 10$$

$$Utilization\_ceil\_index = \lceil Utilization * 10 \rceil$$

$$Utilization\_floor\_index = \lfloor Utilization * 10 \rfloor$$

$$Power\_ceil = Power\_model[Utilization\_floor\_index]$$

$$Power\_floor = Power\_model[Utilization\_ceil\_index]$$

$$Delta = (Power\_ceil - Power\_floor)/10$$

$$Power = Power\_floor + Delta * (Utilization - Utilization\_floor/10) * 100$$

Figure 4.2: Equation for getting the power of a physical host with the CPU utilization when not the value's remainder is not zero when divided by one tenth, and the power model array which contains the power consumption per tenth CPU utilization. (Beloglazov & Buyya, 2012; Buyya et al., 2009; Garg & Buyya, 2011)

A third formula used to calculate the power of a physical host linearly rather than with a power model is the formula presented in figure 4.3. Where the maximum power consumption and the static power consumption of a device in addition to it's current CPU utilization is used to calculate the power consumption. Where in the case of the inputted CPU utilization is zero, the returned power consumption is rather zero. (Bel-

oglazov & Buyya, 2012; Buyya et al., 2009; Garg & Buyya, 2011)

$$constant = (max\_power - static\_power)/100$$

$$power = static\_power + constant * utilization * 100$$

Figure 4.3: Equation for getting the power of a physical host with the CPU utilization when calculating the power linearly. Where if the utilization is zero, the power consumption is zero. (Beloglazov & Buyya, 2012; Buyya et al., 2009; Garg & Buyya, 2011)

A Fourth formula used to calculate the power consumption of devices with their processors capacitance, voltage and CPU frequency is the formula presented in 4.4. Where the formula will be used to calculate the power consumption of devices which have varying voltage and CPU frequency for dynamic power scaling. which in the case of the formula targets the method Dynamic Voltage Frequency Scaling. (Ruan et al., 2007)

$$Power = Capacitance * Voltage^2 * CPU\_frequency$$

Figure 4.4: Equation for getting the power of a device when the CPU voltage and frequency varies, where the CPU capacitance is multiplied by the voltage squared and the CPU frequency. (Ruan et al., 2007)

While figure 4.5 on the other hand presents the time required by a disk to write data. Where first and foremost the seek time of the disk is calculated by dividing the file's size by the disk's capacity, before calculating the transfer time which is gotten by multiplying the file's size by the maximum transfer rate of the disks and dividing the value by the disk's capacity. The estimated duration is thus gotten by adding the seek time and the transfer time. (Louis et al., 2015)

$$Seek\_time = File\_size / Capacity$$

$$Transfer\_time = (File\_size * Max\_transfer\_rate) / Capacity$$

$$Transaction\_time = Seek\_time + Transfer\_time$$

Figure 4.5: Equation for time required by a disk to write data to storage. With the formula first getting the seek time and the transfer time before calculating the resulting transaction time. (Louis et al., 2015)

A formula that will be used in the project to calculate the energy consumption in joules from a device's power consumption in watts is the formula presented in figure 4.6. Where the power consumption of the device is multiplied by the time that the device is consuming the power, and as a result getting the energy consumption in joules. With the method utilized to get the energy consumption of physical host with the calculated power consumption and the power modes from devices and components such as disks, switches, and switch ports. (Ruan et al., 2007)

$$Energy = Power(W) / * Time(S)$$

Figure 4.6: Formula for energy consumption, with the device power in watts and time in seconds. (Ruan et al., 2007)

Another formula that will be used in the project is the formula for calculating the "Power Usage Effectiveness" or "PUE" for short as seen in figure 4.7. Where the Power Usage Effectiveness of an area is the resulting value of the facility power consumption divided by the total power consumption of the facility's devices. (Karpowicz et al., 2016b)

$$PUE = Total\ Facility\ Energy / IT\ Equipment\ Energy$$

Figure 4.7: Formula for the Power Usage Effectiveness of a facility where it current energy consumption is divided by its total energy consumption. (Karpowicz et al., 2016b)

4.4 Summary

This chapter covered the idea of the proposed framework of analysis and evaluation for a cloud provider's cloud data center infrastructures.

Further the chapter covered the specifications for the proposed solution for the thesis to specify the capabilities of the framework. Lastly, the chapter covered the formulas that will be used calculate power and energy consumption of devices in addition to the write time for a file, and the Power Usage Effectiveness of a cloud provider.

Chapter 5

Experiment

This chapter will cover the experiment and its results. Further describing the setup of the simulations utilized and extensions to the simulations to extract results. The chapter also covered the results of the proposed framework where the simulated cloud provider's cloud data centers were analyzed and evaluated.

5.1 Simulation

To get the utilization of computing and storage server and network devices utilization simulations were utilized as to mimic a real cloud provider's cloud data centers. The simulations used are "CloudSim" to simulate physical hosts and running workloads in the form of virtual machines; to get their utilization and power consumption. On the other hand, the "CloudSimDisk" simulation which is an extension of CloudSim was used to more specifically simulate the persistent storage of a data center in the form of hard disk drives to get their utilized capacity and placement of data in the form of files. Lastly, a simplified network simulation tool was used to get the network utilization of devices such as switches to extract their link utilization. As supposed to the network simulation tool "NS-3", due to difficulties with the configuration of devices and setup of cloud data center infrastructure. The simplified network simulation application was thus made to simulate traffic from the physical hosts leaving the data center.

5.1.1 Simulation Workload Input

However to simulate the workloads running on the simulated hosts and the data being stored on the data centers; the datasets "GWA-T-12 BitBrains" and "Financial1" were used. The GWA-T-12 BitBrains will be used to run workloads in the form of virtual machines containing the number of CPUs, MIPS, RAM, Bandwidth and storage used in the CloudSim simulation. While the Financial1 dataset will be used to get files and their sizes to be stored in data centers disks within the CloudSimDisk simulation. With the usage of the datasets being to mimic real data center operations, and

thus device usage. With the CloudSim attribute "MIPS" or "Millions of Instructions Per Second" being equivalent to the CPU frequency of the physical host model, as done in "An efficient energy-aware and service quality improvement strategy applied in cloud computing". (Wang et al., 2023)

5.1.2 Simulation Output

For the CloudSim simulation, the data that will be extracted are the physical hosts' information and their scheduled virtual machines, which is done for each host at each data center in addition to the time out of the twenty-four hour period that the data was gathered from. The host information gathered from the CloudSim simulation is the hosts' identification number and data center it is located at. Further the hosts' total available CPUs, Million of Instructions Per Minute or MIPS, RAM, bandwidth, and storage in addition, to the currently available CPUs, MIPS, RAM bandwidth, and storage. Additionally, the power model of the specific host is also extracted to calculate the host's power consumption at different CPU utilizations, the processors' frequency and voltage pairs, whether the host can utilize DVFS, and lastly the virtual machines running on the host at the monitored time. From the scheduled virtual machines on the host, the extracted data was the virtual machine's allocated CPUs, MIPS, RAM, bandwidth, and storage.

While for the CloudSimDisk simulation the data that will be extracted is information regarding the data center persistent storage in the form of hard disk drives. Other than the data center name and identification number and the disks' identification number, the disks' total capacity and available disks' capacity is was extracted with the disks' average rotation and seek time and their maximum data transfer rate. Additionally, the power consumption of the data center disks' when active and idle were also extracted with their current state. Further the duration of the disks' being active, the duration of the simulation was also extracted with the disks' idle intervals.

From the simplified network simulation aside from the identification number of data centers, switches and their links, the data extracted from the simulation was information regarding the switches such as their roles which could be "Core switch", "Distribution Switch" or "Access switch", and their level in the network infrastructure tree with access switches at level 0. In addition to the the power consumption of the switch ports and the power consumption of switch when active and when idle. With additional information such as simulation duration, whether the port or switch is powered on or off, their utilization of bandwidth, and the history of bandwidth utilization passed through the link and the switch.

5.1.3 Simulation Extensions

An extension done to the CloudSim and CloudSimDisk simulations were implementation of logging the simulation results after completion. In the

case of CloudSim this contained the infrastructure state at different points in time during the cloud provider simulation which lasts twenty-four hours. Additionally, were the entries for the simulated hosts and virtual machines extended to store information regarding their allocated CPU's, RAM, bandwidth, and the scheduled virtual machines at the time. While the CloudSimDisk simulation similarly logged information regarding the disks after the simulation had ended, the extension logged the duration of a disk being active over a time frame with the total duration, the idle intervals, and the duration of the disk' in active mode.

Additional work also went into adding a custom host based on the "Celsius V840" host specifications with an "AMD Opteron 2218" CPU. With the custom host extending the "PowerModelLinear" class which is utilized for linear modeling of power consumption of devices which can describe power consumption where the CPU frequency and voltage of the processors can vary. Within the added class the ability to add pairs of frequencies and voltages were also added for simulation output and further analysis and evaluation by the proposed framework. (Computer, 2007; Tang et al., 2016)

5.1.4 Simulation Infrastructure

For the experimental setup, the cloud provider will contain multiple data centers with different physical hosts, disks, and switches. Additionally will the number of hosts, disks for the storage, and switches for networking also vary as to simulate a realistic cloud provider infrastructure.

for the experiment the cloud provider was configured with four data centers namely "Datacenter_1", "Datacenter_2" "Datacenter_3", and "Datacenter_4". Additionally the physical host models simulated were "HP ProLiant ML 110 G3", "HP ProLiant ML 110 G4", and "HP ProLiant ML 110 G5" from the authors of "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers", with an additional extended physical host "Celsius V840" which uses the "AMD Opteron 2218" CPU. The difference being the Celsius V840 host using a linear power model as supposed to fixed power specifications in watts. As a linear power model also can describe the power of devices with the ability to vary in power due to reduced performance from CPU's with the ability to alter their frequency. (Beloglazov & Buyya, 2012; Tang et al., 2016; Wang et al., 2023) The specification of each host models such as their number of cores, Mega Hertz per core, the host model's available RAM, bandwidth, and available storage space is specified in table 5.1. While power specifications for each host is presented in table 5.2, where for each tenth percentage of CPU utilization the host's power consumption is specified in watts, the table also present the power consumption when the host is unused or idle. Lastly, table 5.3 which illustrates the configuration for the linear power model used for the specified host models such as it using thirty percent of power statically. (Buyya et al., 2024)

Name	CPU cores	MHz	RAM	Bandwidth	Storage
HP ProLiant ML 110 G3	2	3000	4GB	10Gbit	160000
HP ProLiant ML 110 G4	2	1860	4GB	10Gbit	160000
HP ProLiant ML 110 G5	2	2660	4GB	10Gbit	146000
Celsius V840	4	2600	16GB	10Gbit	80000

Table 5.1: Host specification per host model containing the name of the model, number of CPU Cores, the CPU frequency in Mega Hertz, the host's available RAM, bandwidth, and storage space. (Beloglazov & Buyya, 2012; Computer, 2007; Corporation, 2011a, 2011b, 2011c; Tang et al., 2016; Wang et al., 2023)

Name	idle	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
ML 110 G3	105	112	118	125	131	137	147	153	157	164	169
ML 110 G4	86	89.4	92.6	96	99.5	102	106	108	112	114	117
ML 110 G5	93.7	97	101	105	110	116	121	125	129	133	135
Celsius V840	0.0	92.5	110.0	127.5	145.0	162.5	180.0	197.5	215.0	232.5	250.0

Table 5.2: Host power model specification per host model containing the idle power consumption and the power consumption per CPU utilization in watts in addition to the power model per tenth utilization from the linear power model of Celsius V840. (Beloglazov & Buyya, 2012; Computer, 2007; Corporation, 2011a, 2011b, 2011c; Tang et al., 2016; Wang et al., 2023)

Name	Max Power	Static Power Power	Static Power Percent
Celsius V840	250W	75W	30%

Table 5.3: Host power model parameters for linear power models, containing the max power, constant power, and static power percent specified for the host model in the CloudSim simulation. (Beloglazov & Buyya, 2012; Buyya et al., 2024; Computer, 2007)

Name	Frequency and Voltage Pairs						
Celsius V840	Frequency	2.6	2.4	2.2	2.0	1.8	1.0
	Voltage	1.30	1.25	1.20	1.15	1.10	1.05

Table 5.4: Table containing the frequency and voltage pairs for the specified host model; utilized for altering the host's processor frequency and voltage to dynamically alter the CPU's power consumption. (Tang et al., 2016)

In the CloudSimDisk simulation the four data centers rather simulate the data center storage components. In this case the simulated devices are "HGST Ultrastar C10K900", "Seagate Enterprise NAS", and "Toshiba MG04SCA Enterprise"; which are defined hard disk drive drive models from the authors of "CloudSimDisk: Energy-Aware Storage Simulation in CloudSim". The specification for the models including their disk capacity, average seek time, average rotation latency and maximum internal data

transfer is specified for each model in table 5.5. With the disks potential power modes presented in table 5.6 containing the power consumed when the disk is active and idle. (Beeler, 2014; Lab, 2013; Louis et al., 2015; Smith, 2015)

Name	Capacity	Seek Time	Rotation Latency	Transfer Rate
HGST Ultrastar C10K900	900GB	4ms	3ms	198MB/s
Seagate Enterprise NAS	6TB	8.5ms	4.16ms	216MB/s
Toshiba MG04SCA Enterprise	5TB	9ms	4.17ms	215MB/s

Table 5.5: Disk specification per disk model containing the disk model's name, it's disk capacity, the disk's average seek time, the disk's average rotation latency, and the maximum internal data transfer rate. (Beeler, 2014; Lab, 2013; Louis et al., 2015; Smith, 2015)

Name	Idle power	Active power
HGST Ultrastar C10K900	3.0W	5.8W
Seagate Enterprise NAS	6.9W	11.27W
Toshiba MG04SCA Enterprise	6.2W	11.3W

Table 5.6: Disk power model per disk model containing the disk power consumption when idle and when actively reading and or writing data. (Beeler, 2014; Lab, 2013; Louis et al., 2015; Smith, 2015)

For the simulated networking infrastructure on the other hand, the had a layered hierarchical structure where the top layer or level switch in each data center was labeled a "Core switch", which lead traffic out of the data center. Another switch role in each data center was the "Access switch" which are a layer of switches directly connected to physical host or devices potentially requiring internet access. Between the two layers containing core switches and access switches are the "Distribution switches" or "Aggregate switches" which have the responsibility of distributing the traffic from the core switch to the access switches. Further distribution switches may be connected to other distribution switches from the layer above or below in larger networks. Additionally, are devices requiring network connected with to two different access switches, each access switch are also connected to two different distribution switches, and each distribution switch are connected to two different core switches. The goal of the following infrastructure is to create backup links and devices in case of failures. (Bruno & Jordan, 2012; Tate et al., 2013)

For the data center network infrastructure the simulated devices models were "Mellanox SX6056" and "Mellanox SX6036". Where the switch model Mellanox SX6036 is utilized for the role of access switch in the data center networks. While the Mellanox SX6056 model on the other hand, is used as a core switch at data center networks, but also for distribution of the network traffic in the network. The specification for the switch utilized is illustrated in table 5.7 containing device specifications such as it's role, number of available ports, and the available bandwidth per port. While

5.8 specifies the power consumed by the switch model and the power consumption for each of the switch's port. (Technologies, 2013)

Name	Role	Number of ports	Bandwidth per port
Mellanox SX6056	108	Core switch	10Gbit
Mellanox SX6056	108	Distribution switch	10Gbit
Mellanox SX6036	36	Access switch	10Gbit

Table 5.7: Switch specification per switch model and switch role, with the table containing the switch model's name, it's role, it's number of ports, and the available bandwidth per port. (Technologies, 2013)

Name	Device Power	Power per port
Mellanox SX6056	1056W	8W
Mellanox SX6036	132W	8W

Table 5.8: Switch power model per switch model, containing the switches' base power consumption and the power consumption per powered port. (Technologies, 2013)

With the models presented above the number of simulated devices for "Datacenter_1" is illustrated in table 5.9 under the "D1" sub column from the "Amount" column; where the data center contains two hundred HP ProLiant ML 110 G3, eighty HP ProLiant ML 110 G4, eighty HP ProLiant ML 110 G5, and forty Celsius V840 hosts. With the data center containing twenty HGST Ultrastar C10K900 disks. With the network containing four Mellanox SX6056 switch models, two as core switches and two as distribution switches, and lastly twelve Mellanox SX6036 switch models as access switches.

While Datacenter_2 in table 5.9 under the sub column "D2", contains twenty HP ProLiant ML 110 G3, one hundred HP ProLiant ML 110 G4, forty HP ProLiant ML 110 G5 and twenty-five Celsius V840 hosts. With the number of hard disk drives being fifteen Seagate Enterprise NAS disks. With the network similarly including four Mellanox SX6056 switch models with again two as the core switches and two as the distribution switches, and lastly six Mellanox SX6036 models as access switches.

Datacenter_3 on the other hand, as seen in table 5.9 under the sub column "D3"; has forty HP ProLiant ML 110 G4, a hundred and twenty HP ProLiant ML 110 G5, and thirty Celsius V840 hosts. With the number of disks from the table 5.9 illustrating that the data center contains fifteen Toshiba MG04SCA Enterprise disks. With Datacenter_3 also containing two core Mellanox SX6056 switches, two distribution Mellanox SX6056 switches, and six Mellanox SX6036 access switches.

Lastly, Datacenter_4 based on the amount under the sub column "D4" in table 5.9; speciyfing that the data center contains eighty HP ProLiant ML 110 G3, ninety HP ProLiant ML 110 G4, seventy HP ProLiant ML 110 G5, and thirty-five Celsius V840 hosts. While the data center storage

contains seven HGST Ultrastar C10K900, five Seagate Enterprise NAS, and three Toshiba MG04SCA Enterprise disks. With the network infrastructure of Datacenter_4 containing two core Mellanox SX6056 switches, two distribution Mellanox SX6056 switches, and nine Mellanox SX6036 access switches.

Name	Type	Amount			
		D1	D2	D3	D4
HP ProLiant ML 110 G3	Host	200	20	0	80
HP ProLiant ML 110 G4	Host	80	100	40	90
HP ProLiant ML 110 G5	Host	80	40	120	70
Celsius V840	Host	40	25	30	35
HGST Ultrastar C10K900	HDD	20	0	0	7
Seagate Enterprise NAS	HDD	0	15	0	5
Toshiba MG04SCA Enterprise	HDD	0	0	15	3
Mellanox SX6056	Core Switches	2	2	2	2
Mellanox SX6056	Distribution Switches	2	2	2	2
Mellanox SX6036	Access Switches	12	6	6	9

Table 5.9: Table containing the simulated device model Name, Type and Amount at each data center. With the "D1" sub column representing Datacenter_1, "D2" representing Datacenter_2, "D3" representing Datacenter_3, and lastly "D4" representing Datacenter_4.

5.2 Analysis and Evaluation Framework

With the simulation results achieved the simulated cloud environment's infrastructure can now be further analyzed and evaluated. The cloud environment was thus analyzed and evaluated with the extracted data; with this section covering the results from the framework in the areas of compute, storage and network.

5.2.1 Analysis of Compute

An aspect of the cloud provider's infrastructure being targeted for analysis and evaluation was the system managing workloads in the form of virtual machines executing task and the its managed physical hosts. Where within this section the results of the system managing the physical host and their available resources such as the resource utilization, placement of workloads in the form of virtual machines executing tasks are analyzed and evaluated.

Analysis of Computational Resources

From the simulation results out of the available two thousand three hundred and sixty CPUs a thousand six hundred and sixty CPUs was used. This resulted in a utilization of the CPUs seventy point forty-two percent as seen in figure 5.1. With the cloud provider's hosts utilizing 3289525.46

million instructions per second out of the available 5954400, which led to a utilization percentage of fifty-five point twenty-five percent. From figure 5.1 it is also illustrated that the cloud provider has a RAM utilization of eighty point forty-four percent. This was due to the cloud provider's running workloads consuming four thousand six hundred and thirty-three point fifty-nine gigabytes out the available five thousand seven hundred and sixty gigabytes. While the bandwidth and the storage consumed one point one percent and zero point fifty-six percent. Which stemmed from the running workloads consuming a hundred and thirteen point eighty six Gigabits per second out ten thousand and five hundred Gigabits, and eight hundred and sixty-three point four Gigabytes of storage out the available five hundred and fifty-three thousand, two hundred and sixty gigabytes.

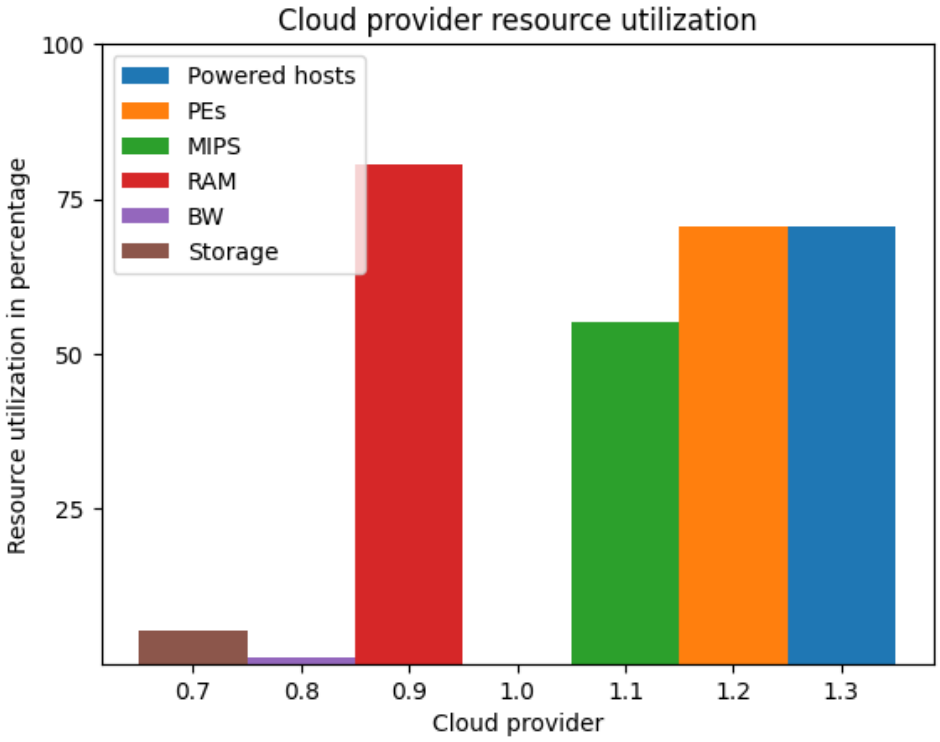

Figure 5.1: Simulated cloud provider resource utilization in percentage.

More specifically for Datacenter_1 which has four hundred of it's physical hosts utilize their six hundred and sixty-seven CPU cores out the available eight hundred and eighty four CPU cores by the simulated virtual machines from the GWA-T-12 Bitbrains dataset of virtual machine performances; resulting in a seventy five point eight percent utilization of CPUs. However, the CPUs had only used fifty-nine point eight percent of their execution capacity as only 1398612.18 Million Instructions Per Second out of the 2339200.0 were carried out. On the other hand, the RAM usage of the data center had instead reached a ninety-six point nineteen percent, as two thousand point one gigabyte of memory was used from the total

of two thousand and eighty gigabytes. While the bandwidth and storage utilization at the data center was at zero point nine percent and zero point sixty-three percent. The resulting utilization was thus derived from the utilized thirty-six point one gigabits out of the total four thousand gigabits, and the storage consuming three hundred and seventy-six point seven gigabytes out of the total fifty-nine thousand and six hundred and eighty gigabytes.

Datacenter_2 more specifically used three hundred and seventy-eight processors of the total four hundred and twenty available cores; resulting in a ninety percent of CPU cores being used. With utilization of the CPU at seventy-two point eighty-eight percent. With the percent stemming from 703104.60 Million of Instructions Per Second of the available 964800.0. While the data center's workloads consumed ninety-three point forty-two percent; which was calculated from the used data center host memory of nine hundred and seventy-four point fifty-seven gigabytes out of one thousand and forty gigabytes. Additionally, the analyzed cloud data center consumed twenty-seven point forty-six gigabits out of one thousand eight hundred and fifty gigabits; leading to a one point forty eight percent utilization. Lastly the data center storage, which the distributed workloads consumed one hundred and seventy-seven point two gigabytes out of twenty-seven thousand and forty gigabytes, leading to a usage of zero point sixty-six.

Datacenter_3's physical host utilization by it's provisioned workloads resulted in three hundred and seventy-nine of the physical hosts' processors being utilized of the total four hundred and forty available CPU cores. Based on this eighty-six point fourteen percent of Datacenter_3's CPU cores were used. With 751423.44 of the CPUs' Million of Instructions Per Second being used out of 1099200.0. This led to a CPU usage at sixty-eight point thirty-six percent. With the distributed virtual machines at the data center consuming a thousand and sixty-three point twenty-nine gigabytes of memory from the available one thousand one hundred and twenty gigabytes. With the resulting RAM utilization being ninety-four point ninety-four percent. While from the available one thousand and nine hundred gigabits of bandwidth and twenty-six thousand and three hundred and twenty gigabytes of storage, twenty point fifty-three gigabits and one hundred and ninety-nine point nine gigabytes were used. This led to a one point zero and eight percent bandwidth utilization and zero point seventy-six percent utilization of the hosts' storage.

Lastly, Datacenter_4 which had two hundred and thirty-eight out of six hundred and twenty CPU cores at being used; leading to a utilization of thirty-eight point thirty-nine percent. With the workloads consuming 436385.25 of the available 1551200.0 Million of Instructions Per Second from the data center CPUs'. With the percentage of the utilization leading to twenty-eight point thirteen percent. While for the data center RAM the workloads consumed five hundred and ninety-eight point zero-six gigabytes of memory out of one thousand five hundred and twenty gigabytes of RAM. Which resulted in a memory utilization of thirty-nine point thirty-five percent. With the twenty-nine poiny seventy-seven

gigabits out of two thousand seven hundred and fifty gigabits. Leading to a bandwidth utilization of one point zero eight percent. While for the storage utilization; one hundred and nine point sixty-six gigabytes of storage was consumed out of the available forty thousand two hundred and twenty gigabytes. The consumption thus led to a utilization percentage of zero point twenty-seven percent.

The aforementioned results thus are presented in the figure 5.2 which present the resource utilization of the data centers. While figure 5.3 illustrated the resource utilization per host. With table 5.10 containing the specific resource utilization per data center and lastly resource utilization for all data centers in percentage.

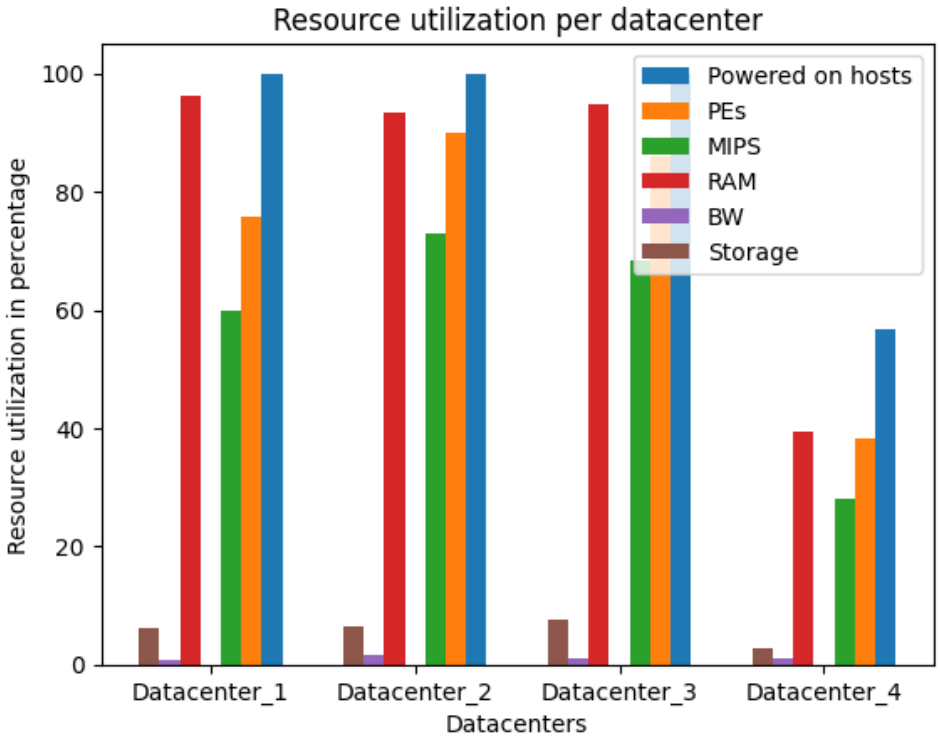

Figure 5.2: Cloud provider's resource utilization in percentage per data center.

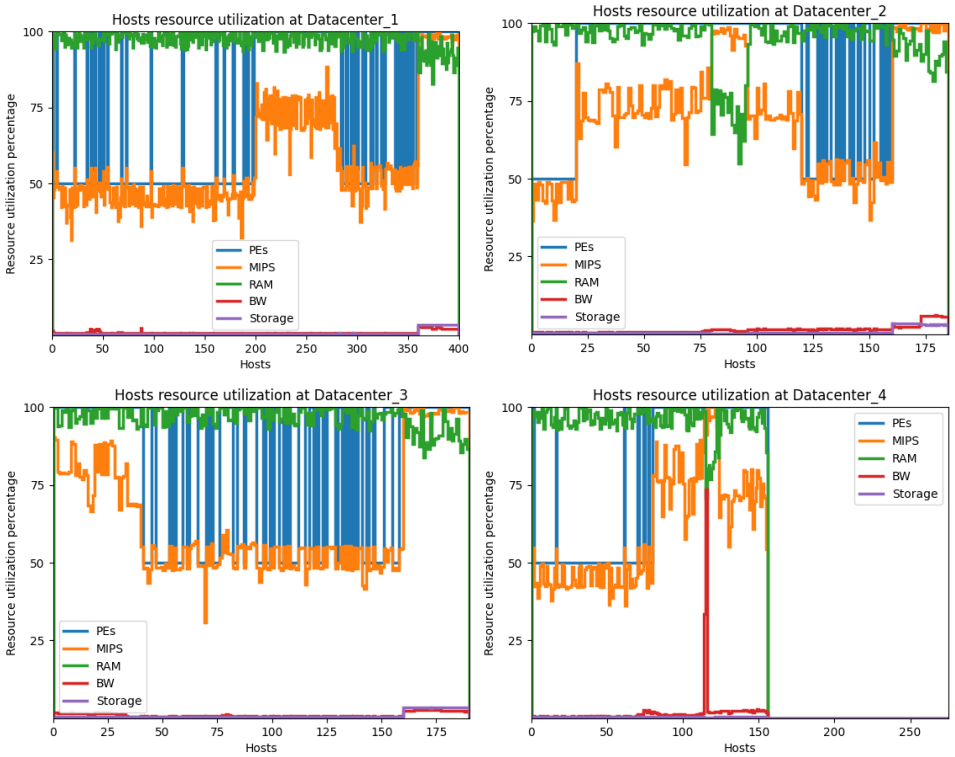

Figure 5.3: Cloud provider's resource utilization in percentage for each hosts with the figure containing four plots, each targeting a data center.

Data center	CPU	MIPS	RAM	bandwidth	storage
Datacenter_1	75.80%	59.79%	96.19%	0.90%	0.63%
Datacenter_2	90.00%	72.88%	93.42%	1.48%	0.66%
Datacenter_3	86.14%	68.36%	94.94%	1.08%	0.76%
Datacenter_4	38.39%	28.13%	39.35%	1.08%	0.27%
TOTAL	70.42%	55.25%	80.44%	1.08%	0.56%

Table 5.10: Table containing the resource utilization per data center and lastly the total resource utilization in percentage.

The Analysis of the framework also presents the power consumption from the extracted simulation results for the cloud provider. Which in this case had one hundred and nineteen of the one thousand and fifty available physical hosts powered down host. This in addition to the utilization of the cloud provider's host led to a power consumption of one hundred and twenty-four thousand four hundred and sixty-eight point eight out of the maximum power consumption of a hundred sixty one thousand three hundred and twenty watts. This has led to the cloud provider saving thirty-six point eighty-five kilo joules per second; saving twenty-two point eighty-four percent of energy from the physical hosts.

Where Datacenter_1 specifically had none of it's hosts powered down, thus the energy and power consumption is stemming from the physical

host utilization. As a result Datacenter_1 consumed fifty-four point ninety-two kilo watts out of the data center's maximum power consumption of sixty-three point ninety-six kilo watts. This has led to the data center reducing the energy consumption by nine point zero four kilo joules per second saving fourteen point fourteen percent of energy.

Datacenter_2 similarly, also had none of it's hosts powered down, thus the resulting power consumption was determined by the hosts' CPU utilization. Which from the analysis of Datacenter_2 showed a power consumption of twenty-four point fifty-five kilo watts out the maximum power consumption of twenty-six point seventy-three kilo watts. Which has led the data center to reduce their energy consumption by two point eighteen kilo joules per second, saving eight point seventeen percent of energy from the physical hosts.

Datacenter_3 also did not have any hosts power down. This led to a power consumption of twenty-five point eighty-five kilo watts from the hosts' CPU utilization out of maximum power consumption of twenty-eight point thirty-eight kilo watts. Leading to a reduced energy consumption of two point fifty-three kilo joules per second, saving eight point ninety-two percent of energy at the data center from it's physical hosts.

Lastly, Datacenter_4 which on the other hand, had one hundred and nineteen of it's physical hosts out of the data center's total two hundred and seventy-five physical hosts powered down. This in addition to the CPU utilization of the remaining one hundred and fifty-six hosts led to a power consumption of nineteen point sixteen kilo watts out of the data center's maximum power consumption of forty-two point twenty-five kilo watts. Resulting in twenty-three point zero nine kilo joules of energy being saved or fifty-four point sixty-six percent.

Analysis of Workload Placement

Besides the resource utilization of data centers, the framework also analyses the placement of workloads. The frameworks thus investigates the placement of virtual machine per hosts, where in the case of figure 5.4 displays the workloads running on all physical hosts. With figure 5.5 illustrating the workload placement per data center.

From figure 5.5 it can be seen that Datacenter_1 has three thousand seven hundred and sixty-seven of the virtual machines of the total of eight thousand six hundred and thirty-four allocated. While Datacenter_2 was provisioned one thousand seven hundred and seventy-two virtual machines. On the other hand Datacenter_3 was distributed one thousand nine hundred and ninety-nine workloads. Lastly, Datacenter_4 which out the total eight thousand six hundred and thirty-four virtual machines running, one thousand and ninety-six virtual machines where allocated to the data center's physical hosts.

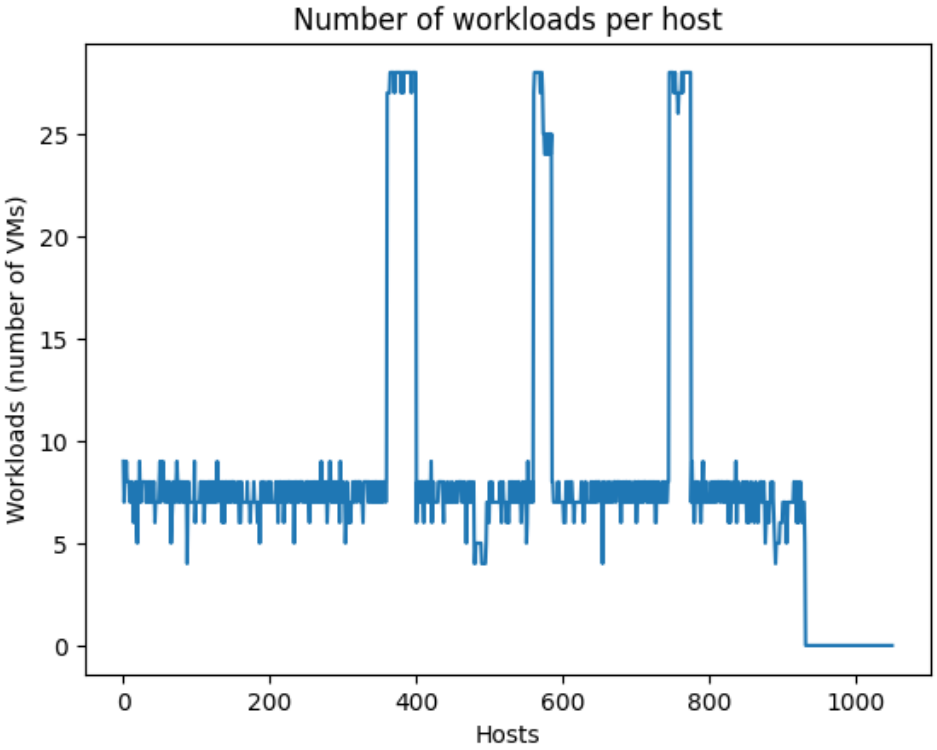

Figure 5.4: Simulated cloud provider workload placement per host.

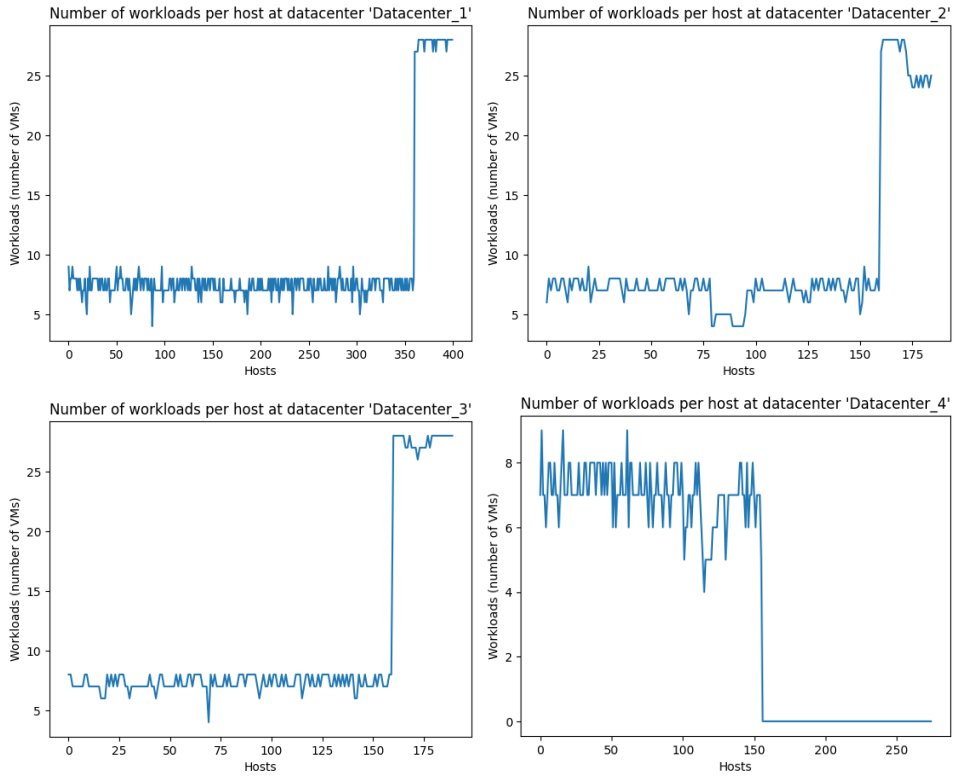

Figure 5.5: Simulated cloud provider workload placement per data center per host.

5.2.2 Evaluation of Compute

Besides analysis of the cloud provider's current infrastructure, the work also evaluates their infrastructure. This is done to review possible methods of improving energy efficiency further to save more energy consumption to reduce environmental impact and operational cost.

Powering Down Physical Hosts

A simple approach taken was to first and foremost investigate whether the data centers had any idle physical hosts. An idle physical host is a host which has no workloads running on it and thus has a zero percent resource utilization. If a data center had an idle physical host the framework would thus analyze the potential reduction of power consumption and further the potential energy being saved.

As a result from the evaluation it was determined that zero additional hosts could be powered down. Thus the utilization of the CPU, MIPS, RAM, bandwidth, and storage remained the same as described in table 5.10.

Dynamic Voltage Frequency Scaling (DVFS)

Another approach used to evaluate the cloud provider's infrastructure over their cloud data centers was through the method of Dynamic Voltage Frequency Scaling or DVFS from the Dynamic Power Scaling group. As the method scales the voltage and frequency of the device to alter power consumption. Which in the case of the cloud provider's infrastructure the host model Celsius V840's voltage and frequency pairs per processor is described in table 5.4. (Buyya et al., 2024; Karpowicz et al., 2016a, 2016b)

As a result of this approach the cloud provider's CPU utilization slightly increasing from sixty-eight point forty-eight to sixty-eight point fifty-eight. With thirty of the three hundred and eighty processors with the ability scale their frequency reducing their utilized frequency and voltage. Where collectively between the thirty CPU cores, thirty-one reductions in available frequencies and voltages were made. Due to this method, the cloud provider's physical hosts consumed a hundred and twenty-three point thirty-six kilo watts. With the cloud provider reducing their energy consumption by thirty-seven point ninety-six kilo joules or twenty-three point fifty-three percent from the cloud provider's maximum energy consumption. As supposed to the current infrastructure which saved twenty-two point eighty-four percent.

With Datacenter_1 specifically altering fourteen of the data center's one hundred and sixty cores that have the ability to dynamically alter the frequency and voltage. Where the frequency and the voltage of the cores in total were reduced fifteen times. Resulting in 1401612.18 Millions of Instruction Per Second being used out the available 2339200.0, leading to a utilization of fifty-nine point ninety-two percent. With the data center consuming fifty-four point thirty-eight kilo watts. Thus saving the data center nine point fifty-eight kilo joules per second or fourteen point ninety-seven percent.

While Datacenter_2 on the other hand only had nine CPU cores that reduced their frequency and voltage, with the frequency and voltage across all the altered cores in total being reduced nine times. This led to a CPU utilization of seventy-three point zero six percent as 704904.6 Millions of Instructions Per Second where used. Resulting in the data center hosts' consuming twenty-four point twenty-two kilo watts. With the data center saving nine point thirty-eight percent of energy consumption or two point fifty-one kilo joules per second.

Datacenter_3, only had seven out of their hundred and twenty CPU processors with DVFS available reduce their available CPU utilization once. As a result the data center's CPU utilization altered to sixty-eight point forty-nine percent, due to 752823.44 Millions of Instructions Per Second out of 1099200.0 were consumed. Because of this the data center's power consumption changed to twenty-five point six kilo watts. Thus the method led to the data center saving two point seventy-eight kilo joules per second or nine point eighty-one percent of the data center's maximum energy consumption.

Lastly, Datacenter_4 which did not have any CPU cores with the ability to use DVFS available altering their frequency or voltage. Due to this the data center's the CPU utilization remained at seventy-eight point ninety-six percent alongside the data center power consumption which was at nineteen point sixteen kilo watts.

Resource Aware Workload Placement

An additional evaluation strategy done by framework is a resource aware workload placement strategy. Where the workloads in the case of virtual machine are placed from scratch in a resource aware manner. This is done to compare the efficacy of the management system scheduling workloads to determine if workloads are being placed in resource aware and energy efficient manner.

The resource aware workload placement functions by investigating a potential placement of virtual machines for the data center to reduce power consumption, efficiently use resources and consolidate workloads through scheduling.

As a result of the resource aware workload placement a hundred and nineteen physical hosts were powered down out of the total one thousand and fifty hosts. With the resulting CPU utilization for the new placement utilizing 4077260.31 Millions of Instructions Per Second out of the total 5954400.0 leading to a sixty-eight point forty-seven percent usage. While the resulting RAM utilization was five thousand five hundred and twenty-four point two gigabytes out of five thousand seven hundred and sixty gigabytes; leading to a utilization percentage of ninety-five point nine percent. With the cloud provider's bandwidth utilization at a thousand three hundred and three point seventy-eight gigabits out of the available ten thousand and five hundred gigabits.

More specifically, as a result the placement strategy three thousand seven hundred and sixty-six workloads out the total eight thousand six hundred and thirty four were placed at Datacenter_1, resulting in a CPU utilization of 1398116.72 Millions of Instructions Per Second out the available 2339200.0 Millions of Instructions Per Second from the data center processors, with none of the data center physical hosts being unused and thus not powered down. While the data center's RAM utilization as a result of the new placement of workloads led to a usage at ninety-six point fourteen percent, where one thousand nine hundred and ninety-nine point sixty eight gigabytes were used out of two thousand and eighty gigabytes. Where as the bandwidth utilization of the data center was thirty-six point zero eight gigabits out of the total four thousand gigabits, resulting in a zero point nine percent usage of physical hosts' bandwidth.

Datacenter_2, similarly did not have any of their physical hosts powered, but on the other hand utilized seventy-two point seven percent of the data center's CPUs, by consuming 701403.62 Millions of Instructions Per Second from the total 964800.0, due to this one thousand seven hundred and sixty-eight scheduled workloads. With the data center's workloads consuming nine hundred and sixty point forty-three gigabytes from the

data center's available one thousand and forty gigabytes, with the resulting memory utilization at ninety-three point twelve percent. While one point forty-eight percent of the data center's bandwidth was used by the placed workloads, as twenty-seven point thirty-five gigabits from the available one thousand eight hundred and fifty gigabits.

Datacenter_3, with it's two thousand and two scheduled virtual machines also did not have any hosts powered down. Further the placed workloads consumed 752379.27 Millions of Instructions Per Second out of the available 1099200.0 from the data center CPUs, resulting in CPU utilization of sixty-eight point forty-five percent. While the collective memory utilization of the data center's physical host as a result was ninety-four point eighty-two percent, as one thousand and sixty-one point ninety-six gigabytes of memory is consumed from the total collective memory one thousand one hundred and twenty gigabytes. The workloads on the other hand at Datacenter_4 only consumed twenty point fifty-six gigabits of the available one thousand and nine hundred gigabits.

Lastly, Datacenter_4 which unlike the previous data centers can due to the new placement of workloads power down a hundred and nineteen of the data center's two hundred and seventy-five physical hosts. With the scheduled one thousand and ninety eight workloads utilizing 1225360.72 Millions of Instructions Per Second out the available 1551200.0 Millions of Instructions Per Second from the remaining powered hosts and their CPUs at Datacenter_4. Besides the CPU utilization, ninety-eight point three percent the data center hosts' memory was consumed as one thousand four hundred and ninety-four point twelve gigabytes of RAM was utilized out the the total one thousand five hundred and twenty gigabytes. However, this time the scheduled workloads utilized one thousand two hundred and nineteen point seventy-nine gigabits out the available two thousand seven hundred and fifty gigabits, leading to a forty-four point thirty-six percent data center bandwidth usage.

While the resulting placement of workloads could lead to one hundred and nineteen of the cloud provider's physical hosts could be powered down. The number of powered down host and the CPU utilization from the remaining nine hundred and thirty-one physical hosts led to a power consumption of a hundred and twenty-four point forty-four kilo watts out the maximum one hundred and sixty-one kilo watts. Leading to a reduced energy consumption of thirty-six point eighty-eight kilo joules per second, saving twenty-two point eighty-six percent as supposed to twenty-two point eighty four percent.

Where the resulting power consumption for Datacenter_1 was consuming fifty-four point ninety-one kilo watts out of sixty-three point ninety-six, as the data center has no hosts powered down. Therefore the data center reduced their power by nine point zero five kilo joules per second saving fourteen point fifteen percent of energy consumption. Whereas the workload placement from the simulation saved fourteen point thirteen percent.

While Datacenter_2 also had zero hosts powered down, consuming twenty-four point fifty-two kilo watts out of twenty-six point seventy-three. With the data center reducing their energy by two point twenty-one

kilo joules per second from the maximum data center power consumption. Thus saving eight point twenty-six percent of energy consumption as supposed to the previous where eight point seventeen percent of energy was saved.

Datacenter_3 also had none of the data center's physical hosts powered down. This in addition to the CPU utilization resulted in a power consumption of twenty-five point eighty-five kilo watts out of the maximum twenty-eighty point thirty eight kilo watts. Reducing the data center energy consumption by two point fifty-three kilo joules per second, saving eight point ninety-one percent of energy as supposed to the previous placement which saved eight point ninety-two percent.

Lastly, Datacenter_4 which unlike the previous data centers had one hundred and nineteen of the data center physical hosts powered down. As a result the data center's power consumption was nineteen point fifty-six kilo watts. Thus saving twenty-three point zero nine kilo joules per second or fifty-four point sixty-five percent of energy consumption, unlike the previous placement which consumed twenty-three point zero nine kilo joules per second saving fifty four point sixty six.

Reorganization of Workload Placement

Another approach related to the placement of workloads was reorganization of scheduled workloads, rather than placing the workloads from the GWA-T-12 Bitbrains from scratch. This approach rather targets hosts which are over and under utilized. In the case of the hosts being over utilized, selected workloads are reallocated to under utilized hosts. The approach in addition also attempts to consolidate workloads on under utilized host; as a result powering down unused hosts.

This is done by examining the utilization of the running hosts. When a host with a resource utilization above ninety percent one or more virtual machines allocated to the host are then deallocated from it. Further the approach attempts to allocate the virtual machines to an under utilized or if none possible then powering up a host for allocation. After the virtual machines have been reallocated to prevent over utilization, the framework will then target the under utilized hosts attempting to reallocate virtual machines from one under utilized host to another. If a host as a result has had all of its virtual machines reallocated, the host is powered down.

This is done to prevent under and over utilization. As first and foremost under utilization leads to increased energy usage and thus energy inefficiency. While over utilization on the other hand, leads to performance degradation for a physical host's hardware. (Hameed et al., 2016; Khan & Zakarya, 2021)

As a result of the used strategy seven out of the one thousand and fifty physical hosts were powered down. With the solution leaving zero host over utilized and one thousand and fifty-three hosts under utilized. While the analyzed cloud data centers showed nine hundred and thirty hosts over utilized and one host under utilized. Due to this the cloud provider's power consumption would be one hundred and thirty-four point ninety-

four kilo watts. With the data center saving sixteen point thirty-five percent or twenty-six point thirty-eight kilo joules per second.

With Datacenter_1 containing three thousand two hundred and ninety-two workloads. Additionally is all of the data center hosts would be powered on and under utilized as none of the hosts resource are used optimally as supposed to the previous placement which had all the hosts over utilized. With the allocated workloads reducing the usage of Millions of Instructions Per Second from 1398612.18 to 1222295.35, leading to a fifty-two point twenty-five percent. While the resulting memory utilization became one thousand seven hundred and forty-nine point seventy-nine gigabytes or eighty-four point twelve percent. While the data center's workloads utilized zero point seventy-nine percent as thirty-one point fifty-two gigabits were used. As a result of this the data center consumed fifty-two point eighty-eight kilo watts. With the data center saving seventeen point thirty-two percent or eleven point zero eight kilo joules per second more specifically.

While Datacenter_2 was allocated one thousand five hundred and forty-three workloads on the data center's five hundred and eighty-five physical hosts. Where none of the data center's hosts were powered down additionally, the data center's physical hosts were all under utilized and none over utilized. Due to the new placement of workloads the data center's CPU utilization from it's physical hosts were sixty-three point sixty-six percent with the workloads specifically utilizing 614156.44 Millions of Instructions Per Second out of the total 964800.0. While the utilization of the memory was eighty-one point ninety-five percent or eight hundred and fifty-two gigabytes. With the bandwidth utilization one point three percent or twenty-three point ninety-seven gigabits. As a result the data center consumed twenty-three point fifty-seven kilo watts. Where the data center due to the scheduling algorithm saved three point sixteen kilo joules per second or eleven point eighty-two percent.

Datacenter_3 on the other hand had one thousand seven hundred and fifty-two virtual machines allocated. Because of the allocated the workloads none of the physical hosts were powered down, with all of the hosts being underutilized where they previously were all over utilized. Due to this the data center's physical hosts utilized fifty-nine point ninety-seven percent of their CPUs, eighty three point fourteen percent of their RAM, and zero point ninety-four percent of their bandwidth. With the data center's physical hosts more specifically consuming 659188.97 of their CPUs' Millions of Instructions Per Second, nine hundred and thirty-one point nineteen gigabytes of RAM, and seventeen point ninety-eight gigabits of bandwidth. Due to this the workloads placed on the data center consumed twenty-four point seven kilo watts. With the data center saving twelve point ninety-five percent, or more specifically three point sixty-eight kilo joules per second.

Lastly, Datacenter_4 which had two thousand and forty-seven workloads allocated by the scheduling algorithm. Because of the allocated workloads to the data center's seven of the two hundred and seventy-five phys-

ical hosts were powered down, with the remaining two hundred and sixty-eight physical hosts being under utilized. As the physical hosts CPU, RAM, and bandwidth utilization were fifty point eighty-nine percent, seventy-two point eighty-two percent, and one point forty-one percent. With the data center physical hosts' CPUs, memory, and bandwidth specifically consuming 789368.92 Millions of Instructions Per Second, one thousand one hundred and six point eight gigabytes, and thirty-eight point seventy-two gigabits. This led to the data center consuming thirty-three point seventy-nine kilo watts. With the data center saving eight point forty-six kilo joules per second, or twenty point zero three percent.

5.2.3 Analysis of Persistent Storage

Another system in cloud data centers explored and analyzed was the resource utilization and energy efficiency of the cloud provider's persistent storage at each data center. Where the capacity and placement of data on hard disk drives were investigated.

Analysis of Persistent Storage utilization

An aspect analyzed from the CloudSimDisk simulation was the utilization of the data center hard disk drives. The utilization on the other hand, targeted the storage capacity available and used, but also the duration of the disk being active due to the file transaction time and the idle time of the disks. Where the total storage capacity of the cloud provider was 21740.0 gigabytes, while 81367.55 gigabytes is used. As a result the utilization percentage for the cloud provider's storage capacity is thirty-seven point forty-three percent. With the collective simulated time being 576303.31 seconds; 398871.18 seconds were the collective time of the simulated disks actively reading and writing to storage. While the resulting collective idle or inactive time for the simulated disks were 177432.13 seconds. Where none of the cloud provider's simulated hard disk drives were unused, as none of the disks' had an active duration at zero seconds nor a any of their capacity unused. This utilization of the cloud provider's hard disk drive drives is illustrated in figure 5.6. While figure 5.7 illustrates the utilization in percentage per data center.

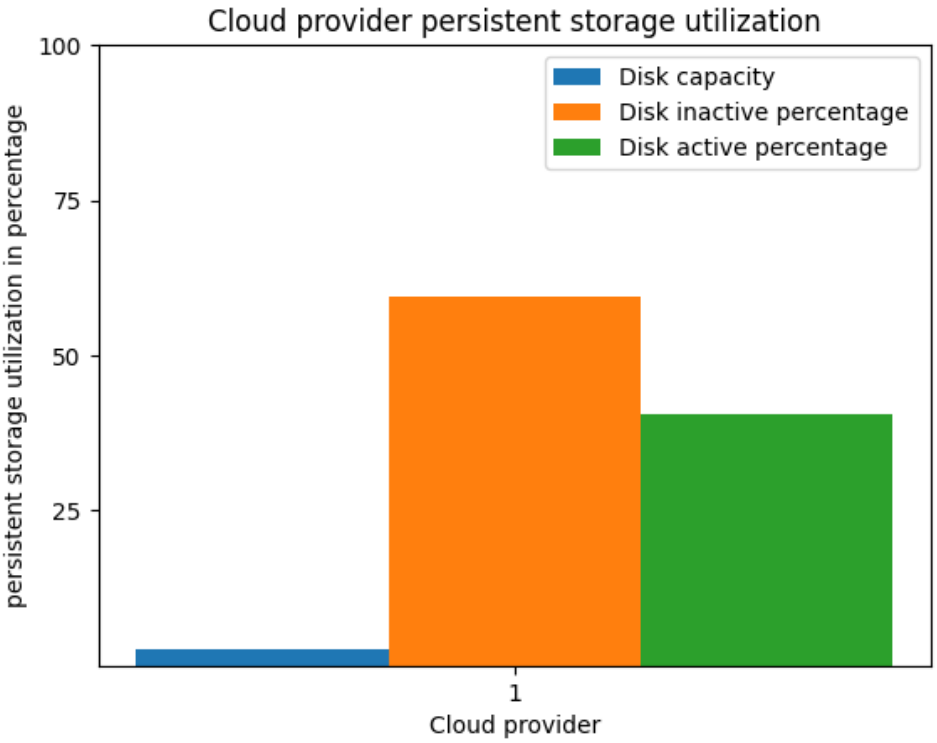

Figure 5.6: Simulated cloud provider storage utilization in percentage.

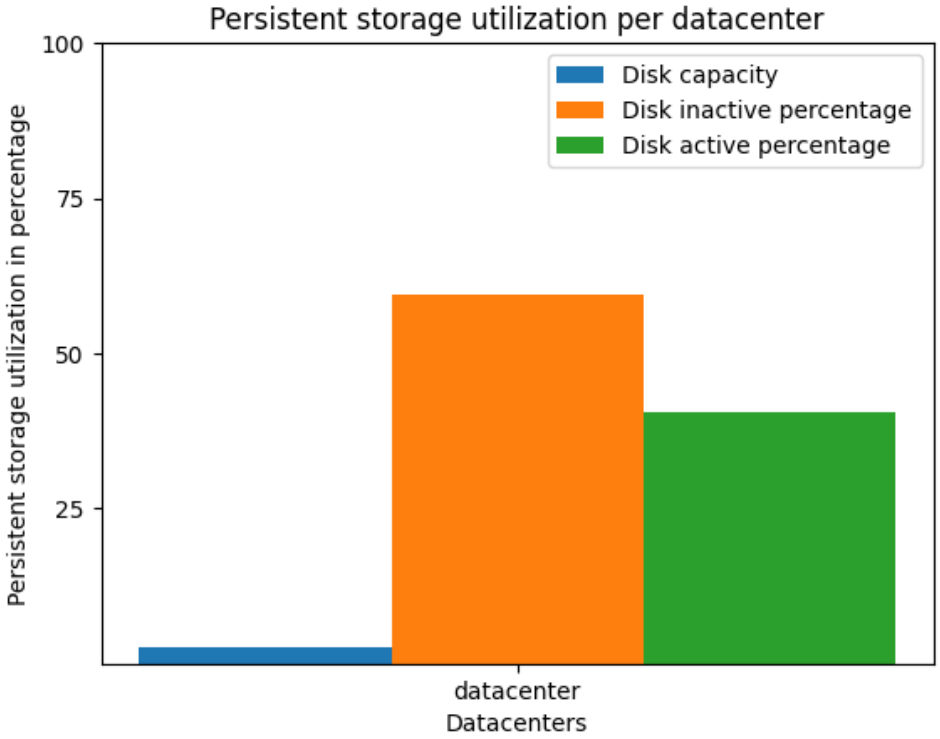

Figure 5.7: Simulated cloud provider storage utilization in percentage per data center.

With Datacenter_1 more specifically consuming 17092.1 of gigabytes out of its available 17100.0 gigabytes. This resulted in a ninety-nine point ninety-five percent utilization of the data center's disk capacity. With the data center's total runtime at 179504.31 seconds from it's nineteen disks, where the data center's hard disk drives were active fifty point zero five percent of the time and inactive forty-nine point ninety-five percent. This was due to the disks being active for 89847.53 seconds and inactive for 89656.79 seconds.

While Datacenter_2 consumed 23350.78 gigabytes out the data center's cloud storage space of 84000.0 gigabytes. Thus from the workloads related to the utilization of storage consuming twenty-seven point eight percent. With the total duration of 132266.33 seconds for the simulated data center's disks. Where the data center's fourteen disks' were active eighty-three point fifty-seven percent and inactive sixteen point forty-three percent of the time. With a collective active time of 110531.32 seconds and a collective inactive time of 21735.01 seconds

Datacenter_3 on the hand consumes thirty-two point sixty-one percent; as 22828.03 gigabytes out of 70000.0 gigabytes of collective disk space was used. With the data center's fourteen disks' collective runtime were 132266.33 seconds, where the collective active duration of disks were 109311.55 seconds and 22954.78 seconds inactive or idle.

Lastly, Datacenter_4 which utilizes thirty-nine point zero eight percent;

as 18096.64 gigabytes out of 46300.00 was used. With the total disk runtime being 132266.33 seconds, where 89180.78 of seconds were the disks actively reading and writing, while the remaining 43085.55 seconds were the disks' being idle.

The presented results regarding the disk utilization for the cloud provider and each of their cloud data center are illustrated in figures 5.8 and 5.9. Additionally, the disk utilization percentage are presented for each data center and the cloud in table 5.11. With the table containing the percentage for capacity, unused disks, active and idle duration.

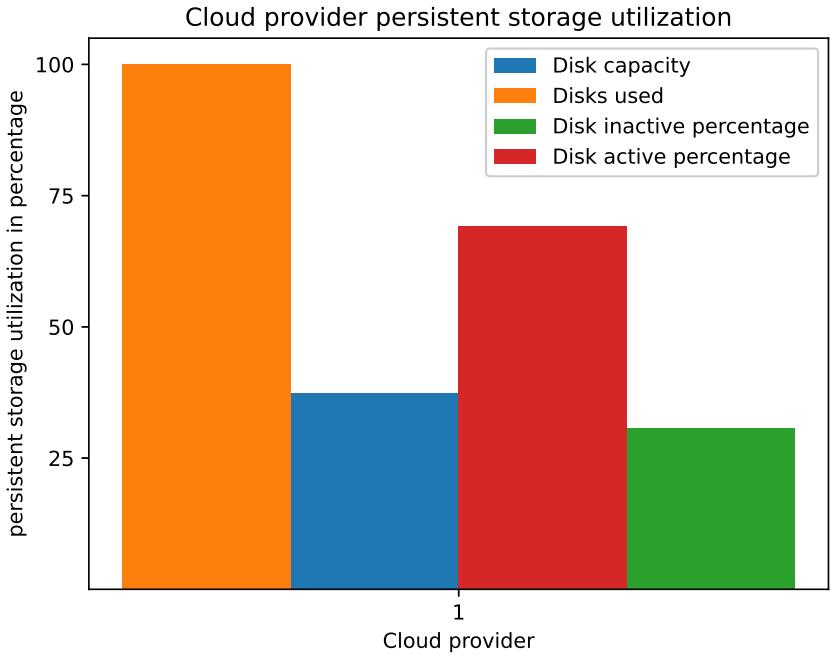

Figure 5.8: Simulated cloud provider storage utilization in percentage per data center.

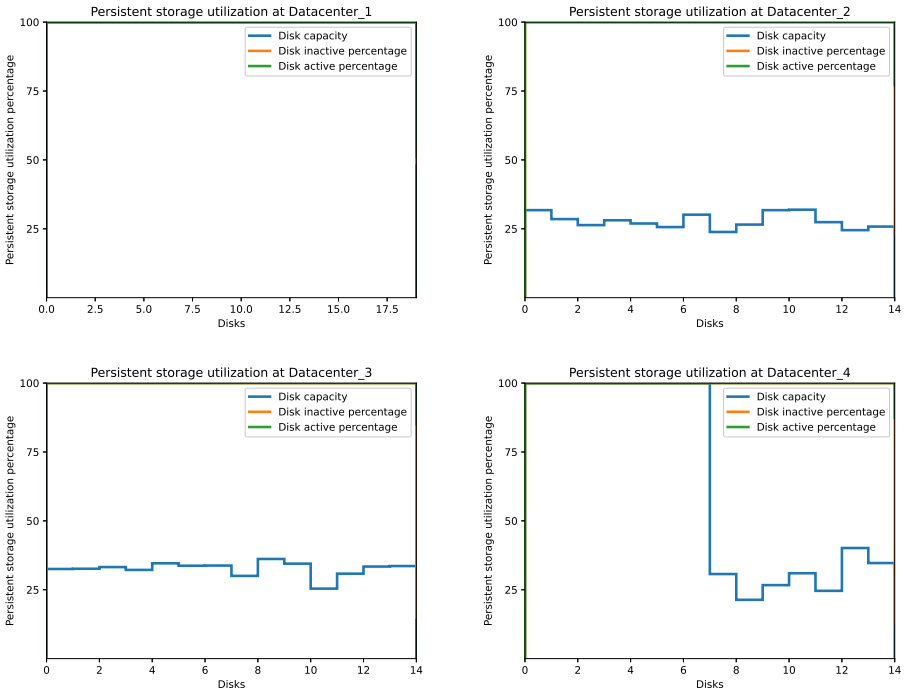

Figure 5.9: Simulated cloud provider storage utilization in percentage for each disk with the figure containing four plots, each targeting a data center. The figure's disk identification numbers are presented in the x-axis of the plots.

Data center	Capacity	Unused disks	Active duration	Idle duration
Datacenter_1	99.95%	0%	50.05%	49.95%
Datacenter_2	27.80%	0%	83.57%	16.43%
Datacenter_3	32.61%	0%	82.65%	17.36%
Datacenter_4	39.09%	0%	67.43%	32.57%
TOTAL	37.43%	0%	69.21%	30.79%

Table 5.11: Table containing the disk utilization per data center and lastly the total disk utilization in percentage.

From the simulation output the disks' power consumption and state was to determined with the disk's active duration. Which in the case of a disk's active duration being above zero, the disk is then considered active, while a disk with an active duration of zero is considered idle. Based on this choice an idle disk is thus unused, with an analysis of the cloud provider's disks illustrating that none of them are unused. With the cloud provider's maximum power consumption being five hundred and forty five watts. However, as analyzed the simulated data center consumed five hundred and forty five watts. Thus the data center based on the current placement of data on persistent storage does not reduce nor save energy consumption.

Analysis of Storage placement

Similarly, to the analysis of workload placement the cloud data center placement of data on disks are also analyzed in the form of storage placement of data on persistent storage. With the framework more specifically, investigating the placement of files from the CloudSimDisk simulation on the data center disks. With the total distribution of the forty thousand eight hundred and fifty-nine files placed by CloudSimDisk in figure 5.10 over all the cloud provider's disks. Whereas figure 5.11 illustrates the placement of files per data center. With data Datacenter_1 containing three thousand five hundred and nineteen of the stored files, Datacenter_2 storing eight thousand two hundred and forty-eight files, Datacenter_3 containing twelve thousand and eight hundred and eight files, and lastly Datacenter_4 which has sixteen thousand two hundred and eighty-four files.

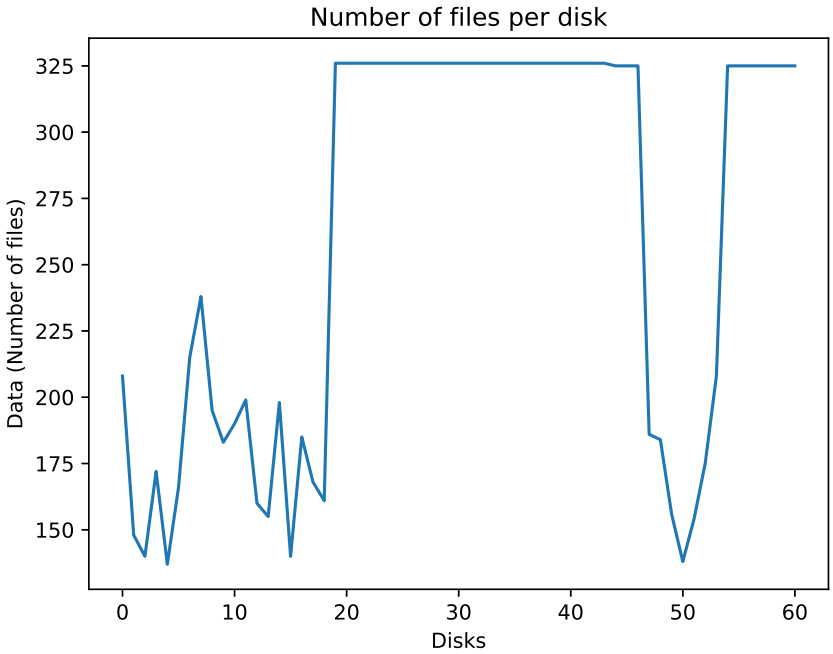

Figure 5.10: Simulated cloud provider storage placement placement per disk.

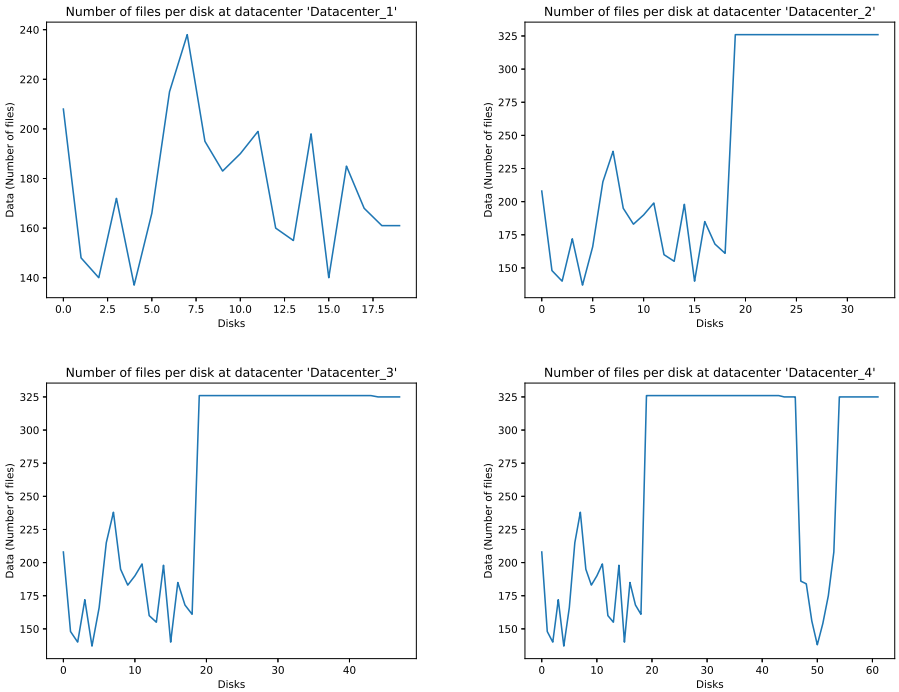

Figure 5.11: Simulated cloud provider storage placement per data center per disk.

5.2.4 Evaluation of Persistent Storage

Along with analysis to gain insight into the resource utilization and data placement at cloud provider's data center, the cloud provider's data centers are in addition also evaluated and reviewed. This is done to determine possible pathways of improvement to the cloud provider's energy efficiency.

Powering Down Disks

An approach used to reduce the power consumption of the persistent storage at each of the cloud provider's data center was to power off unused disk. An unused disk is a disk that has not been active during the duration of the simulation. However, in the case of evaluation of the disk state disregarding the duration, an idle disk will be powered off. the goals of this approach is to power down unnecessary disks to reduce data center energy consumption as idle disks still consume power. As an example the from table 5.5, the hard disk drive "Seagate Enterprise NAS" even when idle consume six point nine watts; powering down such a disk would thus save six point nine watts of power for the cloud data center.

However, from the evaluation of the current placement indicates that the current spread of storage workloads has made use of all the cloud provider's available disks. Therefore the evaluation framework does recommend any number of hard disk drives to power down for the cloud

provider nor at any specific data center. As a result the power consumption for the cloud provider still remains at five hundred and forty five point seventy watts. Additionally, did this not affect the available capacity brought by the cloud storage as none of the disks were unavailable as they were not powered down.

Capacity Aware Data Placement

Similarly to the approach resource aware workload placement which consolidates workloads to be more resource efficient and thus energy efficient, a capacity aware data placement approach is also utilized by the proposed framework. This approach takes data in the form of files placed by the simulation and places them from scratch with the goal of maximizing storage utilization of a disk before placing data on a new disk. This evaluation approach is applied to evaluate the efficacy of the data center management system in regard to the placement of data on the data center disks. Additionally, will the smart standby approach of powering down unused or idle disks also be utilized.

This strategy is applied by firstly getting all the files and disks in addition to their information such as the file size and disk capacity. The evaluation framework will then attempt to place the files on a disk until its capacity or a defined utilization threshold has been met. After the threshold has been met the strategy will then continue to place the remaining data or files on the next available disk.

As a result of the potential placement of data the cloud provider still consumed 81367.55 gigabytes out of total 217400.0 gigabytes as only the positioning of files has changed. However, 209667.55 gigabytes have become unavailable due to the storage consumption and due to disks being powered down. With the difference being apparent in the thirty-one out if the sixty-one disks that can be powered off comparison to the previous placement which didn't allow any of the disks to be powered down. Additionally, the collective run time of powered disks also changed from 576303.31 seconds to 11934.48 seconds. Where the resulting percentage of time the disks were active became fifty-one point fifteen percent as supposed to the prior percentage of sixty-nine point twenty-one percent. With the cloud provider's hard disk drives consuming two hundred and thirty-four point seventeen watts out of the maximum five hundred and forty-five watts. Resulting in the cloud provider saving fifty-seven point one percent of their energy consumption and more specifically zero point thirty-one kilo joules per second compared to the current placement which saved zero joules per second.

While for Datacenter_1 specifically, the data center consumed its maximum power consumption of a hundred and ten point two watts. Similarly, did the previous placement also consume all of the data centers power consumption in addition to utilizing and powering on all of it's available disks. With strategy differentiating itself through the usage of

the disks as the data center's persistent storage resulted in usage of ninety-nine point ninety-five percent, storing three thousand three hundred and fifty-eight files out of fifteen thousand nine hundred and fifty-nine files. Additionally, are the disks collectively active fifty percent of the time as out of the collective duration, the active time is three thousand seven hundred and seventy-nine point twenty-five seconds. With the data center consuming seventeen thousand and ninety-two point one gigabytes out of the total storage of seventeen thousand and one hundred gigabytes. As a result the placement does not save any energy from the previous placement for this particular data center but utilizes the disks more.

Datacenter_2 on the other hand, with the new placements of data resulted in three of its disks being powered down unlike the previous placements which had none of it's disks powered down and twelve thousand and six hundred and one files stored out of fifteen thousand nine hundred and fifty-nine files. With the data center capacity usage from powered on disks being sixty-four thousand two hundred and seventy-five point forty-six gigabytes out of eighty-four thousand gigabytes. Where the data center disks collectively were active two thousand three hundred and twenty-four seconds out of four thousand three hundred and seventy four point ninety-eight seconds to write the delegated twelve thousand and six hundred and one files to the data center cloud storage.

With Datacenter_3 and Datacenter_4 having the most profound changes as no files were stored and none of each of their fourteen disks powered on, making all of the data center's capacity unavailable unless disks are powered on. As a result efficiently saving zero point sixteen and zero point twelve kilo joules per second or one hundred percent for the data centers' storage devices.

Reorganization of data placement

The reorganization of workloads approach was similarly also applied to the cloud provider's cloud persistent storage. With the aim of determining the energy efficiency of data centers' placement of stored data. Where the approach attempts reduce over utilized devices by deallocating data and reallocating to under utilized disks. While for under utilized devices, data is reallocated to other under utilized disks with the goal of consolidating data.

Where the metric used to determine whether a disk's utilization is optimal, over utilized, or under utilized is the disk storage capacity. With selected threshold for over and under utilization being from one hundred percent capacity utilization to ninety-five percent utilization as to maximize the disk capacity being used. With the implemented method thus reallocating data in the form of files from under utilized disks to consolidate and power down unused disks.

As a result of the implemented data placement algorithm, twenty four of the cloud provider's disks could be powered down. Additionally, was zero of the disks over utilized and eleven disks under utilized. With

213367.55 gigabytes of storage capacity consumed by data and unavailable due to disks being powered down, where 81367.55 gigabytes was again consumed by the placed fifteen thousand nine hundred and fifty-nine files. Also, due to the altered placements of data the collective runtime of disk was altered to fifty-eight thousand eight hundred and seventy-six point seventy-seven seconds. With fifty point twenty-nine percent of the duration of the disks being active, as collectively twenty-nine thousand six hundred and eight point forty-eight seconds. where the cloud provider's disks consumed two hundred and seventy-four point eighty-nine watts out of the total five hundred and forty-five point seventy-three watts. Resulting in the cloud provider saving zero point twenty-seven kilo joules per second or forty-nine point sixty-three percent of cloud data centers maximum energy consumption.

Datacenter_1 more specifically, had all of its nineteen disks powered on, with only one disk being under utilized. Where the resulting number of files placed on the data center out of fifteen thousand nine hundred and fifty-nine was three thousand three hundred and fifty-eight files. With seventeen thousand and ninety-two point one gigabytes consumed by the stored files at the data center. Where the data center hard disk drives were active fifty point twenty-nine percent of the time as they were collectively active for seven thousand four hundred and two point twelve seconds. Due to this the data center consumed the maximum power consumption of one hundred and ten point two watts. Therefore not saving any energy at from the data center's persistent storage.

With Datacenter_2 on the other hand powering down nine of it's hard disk drives, with four the five active disks being active were under utilized. While the data center cloud storage had twenty-eight four hundred and ninety-seven point ninety-two gigabytes used by the five thousand five hundred and ninety-six placed files. Where eighty-two thousand four hundred and ninety seven point ninety-two gigabytes were used by the data placed or unavailable due to the disks being powered down. Due to this Datacenter_2's capacity used was thirty-three point ninety-three percent with ninety-eight point twenty-one percent being the capacity unavailable due to the disk usage and power state. With the collective runtime of the data center disks being fourteen thousand seven hundred and nineteen point nineteen seconds, where the data center disk collectively were active for seven thousand four hundred and two point twelve seconds or fifty point twenty-eight percent of the time. With the utilization of the data center disks resulting in a power consumption from the data center persistent storage at fifty-six point thirty-five watts. Due to this the data center saved zero point one kilo joules per second or sixty-four point twenty-nine percent.

While Datacenter_3 had ten of it's fourteen disks powered down, with the remaining four disks being underutilized. However, while the disks were underutilized the data center sixty-eight nine hundred and ninety-nine point three gigabytes were used or unavailable or ninety-eight point fifty-seven percent. Where twenty-seven point fourteen percent or

specifically eighteen thousand nine hundred and ninety-nine point three gigabytes of the storage was utilized by the three thousand seven hundred and sixty-six stored files. As a result of the allocated files on data center disks, the collective total runtime for the disks were fourteen thousand seven hundred and nineteen point nineteen seconds; where the data center disks were collectively for seven thousand four hundred and two point twelve seconds active or active for fifty point twenty-nine percent of the disk runtime. Due to the aforementioned utilization of the data center storage the resulting power consumption regarding the data center disks were forty-five watts out of a hundred and fifty-eight point two watts. With the data center saving seventy-one point forty-three percent of energy consumption or more specifically zero point eleven kilo joules per second.

Lastly Datacenter_4, which had five hard disk drives powered down out the data center's fourteen disks. Where due to the implemented algorithm two of the nine powered on disks were under utilized. With the data center utilizing sixteen thousand seven hundred and seventy-eight point twenty-four gigabytes or thirty-six point twenty-four percent of the cloud storage by the scheduled three thousand two hundred and thirty-nine files. While forty-four thousand seven hundred and seventy-eight point twenty-four gigabytes of the cloud storage was used or unavailable due to the disks being powered down. Due to this the data center's powered disks' collective run time was fourteen thousand seven hundred and nineteen point nineteen seconds; where fifty point twenty-nine percent of time or seven thousand four hundred and two point twelve seconds were the data center disks active. As a result of the data center disk activity, the data center consumed sixty-three point fourteen watts. because of this the data center saved forty-seven point nineteen percent or zero point zero six kilo joules per second.

5.2.5 Analysis of Networks

The last aspect of the simulated cloud data centers analyzed and evaluated was the network infrastructure. With network utilization in the form of bandwidth over the core, distribution, and access switches analyzed and evaluated over four separate data centers over a duration of three point sixty-one seconds, investigating the switch and the switch port states. Additionally, analyzing placement of traffic and evaluating the placement of network traffic as a result of strategies utilized in section 5.2.2.

Analysis of Network Resources

From the simulation results it can be viewed from the analysis of the resulting network of the cloud provider's data centers that none of the switches cloud provider's access, distribution or core switches are powered down. With one thousand seven hundred and forty-three switch ports powered down out of the total two thousand nine hundred and fifteen. However, out of the forty-nine switches zero of them are powered off. With the data center consuming only one point seventeen percent of bandwidth. As only

three hundred and forty one point fifty-nine gigabits out of the available twenty-nine thousand one hundred and sixty gigabits are used. As a result of the cloud provider's powered switches, switch ports, and bandwidth utilization, the resulting power consumption was thirty point ninety-seven kilo watts out the maximum forty-four point fifty-eight kilo watts. With the cloud provider more specifically saving thirteen point sixty-two kilo joules each second or thirty point fifty-four percent of the network's energy consumption due to the powered down switch ports.

With Datacenter_1 more specifically utilizing four hundred and fifty-six switch ports out of eight hundred and sixty four, leaving the remaining four hundred and twenty-two powered off. Where twenty of the powered down ports are from the Datacenter_1's access switches, a hundred and ninety are from their distribution switches, and two hundred and twelve are from their core switches. While the data center's bandwidth consumption is a hundred and eight point twenty-eight gigabits out of the data centers maximum bandwidth of eight thousand six hundred and forty gigabits. With the data center's power consumption as a result of their active network devices and active ports is nine point forty-six kilo watts out of the maximum of twelve point seventy-two. With four point ninety-eight kilo watt power consumption stemming from the data center's access switches, while two point thirty-four kilo watts and two point fourteen kilo watts of power consumption are from the data center's distribution and core switches. As a result the data center network is saving three point twenty-six kilo joules each second or twenty-five point sixty-six percent of data center energy consumption.

While Datacenter_2 has four hundred and thirty-nine of its switch ports powered down, with twenty-five of the shutdown ports stemming from the data center's access switches, two hundred and two from the distribution switches, and two hundred and twelve from the core switches. Additionally, is the data center utilizing one point twenty seven percent of the data center bandwidth. With the data center consuming eighty-two point thirty-nine gigabits out of six thousand four hundred and eighty. Due to this, the data center network is in total consuming six point sixty-seventy-five kilo watts, with two point thirty-seven kilo watts stemming from the access switch, while two point twenty-four kilo watts are consumed by the distribution switches, and lastly two point fourteen kilo watts from the core switches. With the data center as a result saving three point forty-five kilo joules per second or thirty-three point eight percent.

Datacenter_3 on the other hand, utilizes sixty-one point six gigabits out of the available data center bandwidth six thousand four hundred and eighty. With the data center out of its total of six hundred and forty-eight ports two hundred and twenty-two ports are powered on, while the remaining four hundred and thirty-four are shutdown. Where twenty of the shutdown ports are from the data center's access switches, two hundred and two from the data center's distribution switches, and two hundred and twelve from the core switches. The data center network

utilization thus led to a power consumption of six point seventy-nine kilo watts out of ten point two kilo watts. With two point forty-one kilo watts of the data center network consumption coming from the access switches at the data center, while two point twenty-four and two point fourteen kilo watts of the data center's power consumption from the distribution and core network switches. As a result, the data center is saving three point forty-one kilo joules per second or thirty-three point forty-one percent of energy consumption.

Lastly, Datacenter_4 which is utilizing one point eighteen percent of the data center's available bandwidth from it's simulated switches, as the workloads from the GWA-T-12 Bitbrains dataset of virtual machine are consuming eighty-nine point thirty-two gigabits by sending traffic out of the data center through the access, distribution and core switches. In addition, are four hundred and forty-eight switch ports powered down from the total seven hundred and fifty-six, where forty of the switch ports are attributed to the data center's access switches, a hundred and ninety-six to the distribution switches, and two hundred and twelve to the core switches. which in comparison to the prior data centers, Datacenter_4 consumed seven point ninety-six kilo watts out the data center's maximum power consumption of eleven point forty-six kilo watts. Where the distribution switches two point twenty-eight kilo watts and the core switches two point fourteen kilo watts, unlike the access switches at the data center which consumed three point fifty-three kilo watts. As a result saving the data center three point five kilo joules a second or thirty point fifty-one of the data center's operational energy consumption.

Analysis of Network Traffic

Besides analysis of the data center network utilization, the framework also analyzed the network usages on the switch links from the connected devices. Which in the case of the included network infrastructure are the previously analyzer workloads and physical hosts. With the total amount of sources transmitting over the data center network being three thousand one hundred and fifty. With Datacenter_1 having one thousand and two hundred sources transmitting data, Datacenter_2 containing five hundred and fifty-five sources, Datacenter_3 which has five hundred and seventy sources utilizing bandwidth, and lastly Datacenter_4 which has eight hundred and twenty five devices utilizing the data center's network.

5.2.6 Evaluation of Network

Besides analysis of the network, the framework also evaluates the cloud provider's network for their specific data centers. With the section reviewing and evaluating the current network infrastructure, but also the potential power consumption saved by the new workload placement and reorganization algorithms presented in section 5.2.2.

Powering Down Switches and Switch Ports

An approach utilized to dynamically reduce power consumption of network devices was powering down unused ports and switches. As a result of the approach the framework detected twelve switches which could be powered down by, while one thousand seven hundred and two switch ports could be powered off. Due to this the cloud provider's network infrastructure of switches consumed thirty point ninety-six kilo watts. Where the cloud provider saved thirteen point sixty-two kilo joules per second or thirty point fifty-four percent of energy consumption.

Where Datacenter_1 specifically could keep fourteen of its sixteen switches powered on. With four hundred and eight of the data center's eight hundred and sixty-four switch ports could be powered down. As a result of number of powered down switches and ports the data center consumed seventy-three point forty-four kilo watts. As such the data center is saving five point thirty-eight kilo joules per second or forty-two point twenty-six percent.

With Datacenter_2 similarly powering down two of the ten switches and four hundred and thirty-one switch ports at the data center. Due to this the data center is consuming four point sixty-four kilo watts. Where the data center saves five point fifty-six kilo joules per second from the total ten point two kilo joules per second reducing used energy by fifty-four point fifty-one percent.

Datacenter_3, also have two out of the data center's ten switches powered down as they were unused. Additionally, could also four hundred and twenty-six of the switch ports powered down out the total six hundred and forty-eight. As a result of the powered down ports and switches, the cloud data center's power consumption would be four point sixty-eight kilo watts. Where the data center saves five point fifty-two kilo joules per second or fifty-four point twelve percent of data center network's energy consumption.

Lastly, Datacenter_4 which on the other hand can have six of the data center's switches powered off as they are not used. Further the data center could also more specifically power down four hundred and thirty-seven of it's seven hundred and fifty six switch ports. Because of this the cloud data center's power consumption would be five point thirty-two kilo watts. With the data center's network saving six point fourteen kilo joules per second or fifty-three point fifty-four percent.

Analysis of effect from resource aware workload placement

In addition to evaluating the current data center network state, the framework also evaluates the effect of the evaluation method used when placing workloads on the physical hosts from scratch on the cloud data center network. The framework thus investigates the relational effect through the effected network from the resource aware workload placement at each of the data centers. With the method in addition applying a

dynamic energy management approach by powering down unused links and switches.

This is done by first by implementing the strategy described in section 5.2.2, under the subsection "Resource Aware Workload Placement". From there the framework then determines the resulting utilization of links between physical host and access switches, where unused links are powered down. If all links are unused and thus powered down the framework determines the switch unused and powers the switch down. Similarly is the process repeated between access switches and distribution switches, and between distribution switches and core switches.

As a result of the resource aware workload placement the cloud provider's cloud data center networks could have twelve of their forty-nine switches powered down. In addition, to one thousand eight hundred and ninety of the two thousand nine hundred and sixteen switch ports powered down. As a result of this, the cloud data center network's power consumption was twenty point forty-eight kilo watts out of forty-four point fifty-eight kilo watts. With the cloud provider in total saving fifty-four point zero five percent or twenty-four point one kilo joules per second more specifically.

Datacenter_1, which more specifically powered down two of the data center's sixteen switches. With the four hundred and twenty-six ports out of eight hundred and sixty-four powered down. Resulting in a power consumption of seven point two kilo watts. With the data center saving forty-three point four percent or five point fifty-two kilo joules per second out of the total twelve point seventy-two kilo joules per second.

While Datacenter_2 on the other hand, still had eight of the data center switches ten switches powered on. With four hundred and forty-three ports shutdown out the total six hundred and forty-eight. Due to this the data center consumes four point fifty-four kilo watts out of the maximum power consumption of ten point two kilo watts. With the strategy saving the data center five point sixty-six kilo joules per second or fifty-five point forty-five percent.

While Datacenter_3 similarly powered down two of their ten switches, where four hundred and thirty-eight out of six hundred and forty-eight ports could be powered down. This led to a power consumption of four point fifty-eight kilo watts out of ten point two kilo watts. Saving five point sixty-two kilo joules per second or fifty-five point zero six percent of energy consumption.

With Datacenter_4 powering down six of the cloud data center's thirteen switches. This led to five hundred and eighty-three out if seven hundred and fifty six ports able to be powered down due to the new placements. Resulting in a power consumption of four point sixteen kilo watts out of eleven point forty-six. Therefore, saving seven point three kilo joules per second or sixty-three point seventy-three percent.

Analysis of effect from resource aware workload reorganization

In addition to the effect of the resource aware workload placement from scratch on the data center network, the framework also evaluates the effect of the resource aware reorganized workload placement. With the goal of investigating the energy efficiency on the data center network as a result of the reorganized placement placement of workloads on the physical hosts. Additionally applying a dynamic energy management approach on the network by powering down unused links and switches.

How this is done, is first by implementing the reorganized placement algorithm presented in section 5.2.2. Further the resulting network is structured, with the strategy investigating what ports are necessary and powering down unused ports. Similarly, if all ports are unused the whole switch is thus powered down.

As a result of the scheduling algorithm reorganizing host workloads, the cloud provider's data centers had nine of its forty-nine switches powered down. With one thousand seven hundred and sixty-nine switch ports being powered down out the total two thousand nine hundred and sixteen. With the cloud provider consuming twenty-one point eighty-five kilo watts out of forty-four point fifty-eight kilo watts. Resulting in the cloud provider saving twenty-two point seventy-three kilo joules per second or fifty point ninety-nine percent.

With Datacenter_1 specifically powering down two of the data centers switches. Where four hundred and twenty-six ports were powered down out the eight hundred and sixty four switch ports. As a result of the data center network utilization, the cloud data center consumed four point fifty-four kilo watts out of twelve point seventy-two kilo watts. Where the data center saves fifty-five point forty-five percent of data center energy consumption or five point sixty-six kilo joules per second more specifically.

With the second data center or Datacenter_2 also powering off two of the it's switches. Where four hundred and forty-three ports shutdown out the total six hundred and forty-eight ports with the remaining two hundred and five ports powered on. Due to this the reorganized workload placement at the cloud data center consumed four point fifty-four kilo watts of power from the maximum ten point two kilo watts. Because of this, the data center saved five point sixty-six kilo joules per second or fifty-five point forty-five percent.

Similar to Datacenter_1 and Datacenter_2, Datacenter_3 also powers down two of it's ten switches. With four hundred and thirty-eight of the switches' ports being powered down out of the total six hundred and forty-eight. Due to this the data center's network out of its maximum power consumption of ten point two kilo watts; consumed four point fifty-eight kilo watts. With the data center saving five point sixty-two kilo joules per second or fifty-five point zero six percent.

Lastly, Datacenter_4 which on the other hand had three of it's thirteen switches powered down. With four hundred and sixty-two of the switch

ports at the data center being unused and thus powered down out the total seven hundred and fifty-six ports. As a result of the reorganized placement, Datacenter_4 consumed four point sixteen kilo watts out the data center network's maximum power consumption of eleven point forty-six kilo watts. With the data center saving sixty-three point seventy-three percent or seven point three kilo joules per second of total data center network's energy consumption.

5.3 Summary

This chapter first and foremost described the configuration of the simulations that were run during the project, with the goals of describing their setup, configuration, and devices simulated per simulation. Further the chapter went into the what was extracted from the simulation for the cloud data center evaluation and analysis framework. With the remainder of the chapter covering the results from the strategies used by the framework developed for the realms Compute, Storage, and Networking.

Chapter 6

Discussion

This chapter will discuss the results presented in Chapter 5. The chapter first cover observations made during the project regarding the solution in addition to the adaptability of the solution for real cloud data centers. The chapter will then compare the methods for reducing energy consumption used and the combinations of them to further improve energy efficiency for the cloud provider. Further will the chapter discuss the assumptions and limitation of the proposed framework, afterwards discussing future work to improve the proposed framework. Lastly, concluding the chapter by addressing the thesis problem statement, research questions, objectives, and the contribution of the thesis and the framework.

6.1 Observations

An observation during the analysis of the cloud data center's physical host from the insight provided by the framework is the available bandwidth and storage. As seen in figure 5.3, it can be observed that for data center four however, the physical host number a hundred and fifteen utilizes over a quarter of it's available bandwidth leading to a spike in usage at twenty-five percent.

Another observation from the resource aware re-organization of workload placement targeting the servers for computation was the lack of available RAM. As results illustrated a trend of workloads from the dataset GWA-T-12 Bitbrains utilizing more RAM than the host CPU, bandwidth and storage.

Similarly is an observation from the extracted resulting from the proposed framework regarding the placement of workloads. Which based on figure 5.3 illustrates for Datacenter_1, Datacenter_2, and Datacenter_3 a large number of placement of workloads on the Celsius V840 simulated hosts. This could be due to the host model's sixteen gigabytes of memory as supposed to four gigabytes for HP ProLiant ML 110 G3, HP ProLiant ML 110 G4, and HP ProLiant ML 110 G5 specified in table 5.1. As prior observations observe a high RAM demand by the allocated workloads.

A potential solution in this case would be to further extend the number of Celsius V840 hosts at each data center to improve operations. Similarly

could a potential solution be through more heterogeneous physical hosts with differing amount of the CPU, RAM, bandwidth and storage resources at the physical hosts. Thus diversifying the physical host models to optimize resource utilization for varying workloads, or specify workloads for a data center.

Another observation from the results of the framework is from the reorganized placement method. Which data illustrates gaps of utilized disks, which could effect a data center's by leaving switches powered on causing an energy inefficiency. Unlike the data placement strategy which placed data from scratch, as it did not result in gaps of utilized disks but rather consolidated them. This could similarly could also happen to the data center's network connected to physical hosts, in the case where there were more available physical data center hosts.

Another observation made was the evaluation framework powering down unused backup links, and network devices. A potential remedy to this would be to either keep the link or network device active or to instead power it on when the primary link or network device is powered down. With second approach only utilizing the secondary link or network device when required to further reduce cloud data center energy consumption while maintaining a backup solution for fault tolerance.

6.2 Real Adaption of Solution

Regarding the adaption of the solution to real environments, the solution was first and foremost tested using simulation however, despite being a simulation, it utilized real device models based on specifications and virtual machine performances and storage traces. Additionally, several data center's and several device models of different at each data center were simulated. This was done to make the simulation more realistic taking into account larger data center networks and heterogeneous physical hosts, disks, and network switches with real workloads.

Additionally were obtainable data such the resource usage of hosts but also the resource consumption of running workloads, disks and the size of data contained, and bandwidth utilization from switches and their ports used with static data such as the power specifications of a device. This was done to not use information not obtainable from cloud data center environments such as what tasks the workload is executing.

However, improvement to also take into account the network infrastructure regarding the storage servers containing the data center disks and the resulting switches connected, should also be implemented. This would thus make the solution align more with realistic data center infrastructures.

An adaptation required for the solution is for the framework to update in real time and using updated logs rather than illustrating the result at a particular time in addition to over a duration. This would give insight whether a data center is energy efficient at particular times or over a

specified time frame.

Additionally, would an adaptation be use of alternative data sources, with the goal of utilizing pre-existing software to reduce power consumption from additional data collector. This was similarly done from in the paper "An Energy Saving Routing Algorithm For a Green OSPF Protocol" by re-using gathered and used routes determined by the Dijkstra's Algorithm to power down unused links. (Cianfrani et al., 2010; Jiang et al., 2020)

With the application requiring static metrics regarding the monitored devices such as the device's specifications. In addition to the dynamic metrics such as the CPU usage for physical hosts, the disk capacity for disks and the bandwidth utilization for switches as part of the provided data source.

6.3 Comparing and Combining Approaches

6.3.1 Compute

Regarding the cloud data centers' physical hosts, the framework utilized smart standby, dynamic power scaling, and workload scheduling algorithms and methods to evaluate. With the framework powering down unused hosts rather than having them be idle, scaling down available Millions of Instructions Per Second per core, rescheduling workloads from scratch in a resource aware manner, and lastly reorganizing workloads in a resource aware manner taking into account over and under utilization.

With the approach which just powers down unused hosts saving zero percent more energy consumption at the data center. While the approach which just reduces the available Millions of Instructions Per Second reduces the cloud provider's energy consumption by zero point eighty-nine percent. With a combination of the two methods similarly reducing the current infrastructure energy consumption by zero point eighty-nine percent.

With the placement strategy that places workloads from scratch powering down hosts and utilizing DVFS saving zero point zero two percent more energy consumption. While the placement strategy which reorganizes allocated workloads but also attempt to keep the utilization withing a range to avoid over and under utilization resulted in an increased energy consumption of eight point forty-two percent. Which could be due to the method reallocating workloads to avoid over utilization however, not able to optimally place the virtual machines on the hosts. This could potentially be due to the compatibility between the physical hosts' available resources and the scheduled workloads, as prior observation showed that the workloads required more RAM than CPU utilization.

6.3.2 Storage

Regarding the cloud data center persistent storage, the framework made use of smart standby techniques and data placement algorithms. With the framework powering down unused disks rather than keeping them idle,

placing data from scratch to more efficiently utilize disk capacity, and data re-organization to mostly consolidate data placed on disks.

Where based on the results in chapter 5, the powering down disk approach saved zero percent energy for the cloud provider's persistent storage. This was due to the simulated data centers' not having any disks which were unused, as each disk had actively read and written data. On the other hand, approaches like capacity aware data placement and capacity aware reorganizing placement rather than affecting the disks themselves directly scheduled the placement of files on disk at each cloud data center storage. With the simple capacity aware placement of files in addition to powering down unused disk saved fifty-five point zero three percent more energy consumption than the previous placement on the persistent storage. While the reorganized approach in addition to powering down unused disks saved forty-nine point sixty-three percent more energy consumption than the prior placement of data at the cloud provider's data centers.

However, while the capacity aware approach which places data files from scratch reduces more energy consumption, the reorganized approach reduces the number of underutilized disks from thirty-one to eleven.

6.3.3 Network

The network devices and their ports unlike the data center hard disk drives used did not have an idle mode specified which would reduce power consumption from the active operating mode when unused or idle. However, the framework did use the smart standby approach of powering down the unused ports, links and further unused switches, with the goal of reducing the cloud provider's energy consumption regardless of having an idle mode. Which was further used when analyzing and evaluating the effect on the data center network by the resource aware workload placements strategies.

Where first and foremost powering down switch ports from the results in chapter 5 resulted in a zero percent reduction in energy consumption from the current data center. While when just powering down switches resulted in a reduction in energy consumption by twenty-eight point ninety-nine percent. With the strategies combined resulted in a twenty-eight point ninety-nine percent energy reduction.

Further the effect on the network from the resource aware workload placement strategy which placed workloads on physical hosts from scratch; led to the cloud provider's data centers reducing energy consumption by thirty-three point eighty-five percent. While the resource aware workload reorganization strategy led to the cloud provider's data center networks' reducing energy consumption by twenty-nine point forty-four percent.

While in the case of the reorganization approach led to the cloud provider's physical hosts to consume more energy; the approach reduced energy consumption for the cloud provider's network switches. However, similar to the comparison in section 6.3.1, the resource aware workload placement approach implemented without concern of over utilization reduced more energy consumption.

6.4 Limitations and Assumptions

A limitation of the solution in regard to the analysis and evaluation of the storage system, is that the solution does not take into file replicas. Similarly, does the solution not take into account duplicate virtual machines for availability and fault tolerance regarding placement of workloads on physical hosts. As such the solution when analyzing and evaluating the placement of files and virtual machines on the Storage Area Network's hard disk drives and physical hosts could place duplicates on the same device hurting the availability and fault tolerance originally provided.

Another limitation of the solution is the possible inaccuracy of the power consumption based on the equations 4.1, 4.2, and 4.3, which depended on the specified power model or constants from e.g. "Standard Performance Evaluation Corporation" in the case of the physical host and "Storage Review" in the case of the hard disk drives presented in the tables 5.2, 5.6. Where the power consumption is estimated based on CPU utilization and in the case of hard disk drives whether it is actively reading and or writing or in low power mode.

An assumption regarding the inaccuracy of the power consumption, an assumption is that the utilization of memory and bandwidth does not affect the power consumption of the physical hosts. Additionally, it is assumed that the provider does not utilize GPUs and TPUs for their physical hosts; to affect the power consumption of the physical hosts.

6.5 Addressing the Problem Statement

The thesis addressed the problem of a lack of observability into resource utilization of cloud provider's cloud data centers through analysis of consumed or unavailable resource out of the total available. With the thesis further addressing the problem of areas of energy inefficiency by utilizing approaches to evaluate current cloud data center infrastructure; providing graphs and data to inform the cloud provider methods to optimize energy efficiency in areas of compute, storage, and network.

6.6 Addressing the Research Questions

Research question presented in section 1.3 was "What approaches reduce the energy consumption separately and could approaches be combined to to further improve energy efficiency at cloud data center"; where the employed methods reduced energy consumption for the cloud provider's but not necessarily at each cloud data center in the case of the placement strategies. By combining the methods as illustrated in section 6.3 methods combined or separately led to reduction from one or more methods in energy consumption as powering down switches and ports with the placement strategies led to greater energy efficiency.

While for the research question "How does the resulting architecture of different systems affect the energy efficiency of other systems?" was through evaluation of the network. This was done by investigating how the alteration of workload placement of virtual machines on physical hosts would effect the cloud data center networks. With unused switches and switch ports as a result being powered down to reduce power.

The third research question presented in Chapter 1 was "What actions can be taken to improve utilization of resources statistically through detailed cloud data center insight?". Where the actions were presented through the results shown in chapter 5 and in section 6.3 were the possible reduction of energy consumption is presented. With additional actions based on the insight described in section 6.1 to reduce energy consumption with physical hosts that have more tailored resources like RAM to the running workloads.

6.7 Addressing the Objectives

The primary objectives stated in section 1.2, was to evaluate and analyze the each of a cloud provider's cloud data center and to present the results for users. Which based on the results presented in chapter 5; was achieved through the analysis and evaluation framework. As the solution not only analyses the current infrastructure, but also evaluates potential methods to improve energy efficiency. With the results illustrated through graphs and data.

The secondary objectives on the other hand were regarding the cloud environment being as realistic as possible, in addition to providing short feedback based on the framework's findings. With the thesis attempting to achieve the objectives by utilizing the GWA-T-12 Bitbrains and Financial1 datasets of virtual machine performance and disk read and write traces, simulation tools, and simulated devices based on real device model specifications. However, the proposed framework does not achieve the secondary objective of providing short and consolidated feedback from results rather providing more and detailed information.

6.8 Thesis Contribution

The thesis contributed to the research field of energy efficiency in cloud computing through an analysis and evaluation framework. With the proposed framework providing insight into device utilization at cloud data centers, and further proposing areas of improvement through Smart Standby and Dynamic Power Scaling approaches. With the observations from the section 6.1 in the thesis, illustrating potential static energy management techniques to employ better resource usage based on results and insight from the framework. As a result the framework and the thesis presents areas of improvement to reduce cloud data center energy efficiency to reduce environmental impact and operational costs.

6.9 Summary

The chapter covered a discussion of the resulting findings. With the chapter first discussing the observations from the results, before discussing the adaptability of the solution to real world environments. Further the chapter evaluated the simulations utilized to get the input data for the analysis and evaluation framework. Afterwards, the chapter compared and discussed the combination of the methods and algorithms utilized to reduce energy consumption. The chapter then discussed the limitation of the solution and assumptions made throughout the project, before covering future work and improvement to the proposed solution. Lastly the chapter addressed the problem statement, research questions, and objective; before concluding with the resulting contributions of the thesis.

Chapter 7

Conclusion

The thesis covered the topic of energy efficiency in cloud computing by first and foremost introducing it in the introductory chapter of the thesis, defining with it a problem statement, research questions, objectives, and the contribution. With the thesis afterwards detailing the prior work with background information and a literature review. Leading to the thesis covering the methodologies used during the project such as the used literature study tools and techniques. With the thesis finally proposing the solution in chapter 4, detailing the idea and it's specifications and formulas utilized through out the project. Where chapter 5 detailed the simulated infrastructure and the results from the analysis and evaluation at the different areas compute, storage, and network for each cloud data center. With the thesis lastly discussing the results and findings in the discussion chapter of the thesis; addressing the problem statement, research questions, objectives, and what the thesis as a result contributed.

Where as discussed the proposed framework met the placed specifications and primary objectives but, sadly failed to provide concise feedback of areas regarding energy inefficiency, which was one of the secondary objective. However, through it's detailed insight provided information such as the compatibility between workloads and their allocated devices. With the thesis also addressing the research questions by comparing and combining the methods used to determine improvements to energy efficiency but also investigating how the workloads on physical hosts can affect the energy efficiency of the data center networks.

7.1 Future Work

From section 6.2 or "Real Adaption of Solution", future work should first and foremost entail improvements like taking into account the data center network for cloud data center storage. With the framework in addition connecting the physical hosts, physical storage servers, and the internet. This improvement would then extend the number of network devices monitored, but also would investigate the effect the different data placement strategies on the added switches and their subsequent ports.

With additional work based on the points mentioned in section 6.2 being implementation of real time monitoring of energy efficiency, monitoring of energy efficiency over a time frame, and utilization of alternative data sources. The goal of this work would be to view the energy efficiency and resource utilization at different times as the consumption of resources and energy may vary over time. with improvement regarding alternative data sources utilize pre-existing metric collection software like the re-use of routing paths in the paper "An Energy Saving Routing Algorithm For a Green OSPF Protocol"; to reduce workloads consuming additional power. (Cianfrani et al., 2010; Jiang et al., 2020)

While future improvements based on the limitations and assumptions stated in section 6.4; would entail awareness regarding the replicas of files and data in cloud data storage or duplicate virtual machines. As to not compromise the availability and fault tolerance provided by the services of the cloud provider. With the work goal of the future work being avoiding the placement replica files and virtual machines on the same disks and physical hosts; to maintain fault tolerance and availability for clients of the cloud provider.

Additionally, will future work based on the limitations and assumptions stated target the estimated power consumption. Where the goal of the work would be researching and implementing more detailed power consumption from additional components e.g. GPU and TPU or the power consumption from link utilization of monitored switches. Further the work would also compare the monitored power with the estimated power consumption to illustrate the degree of inaccuracy of the calculated power and thus the energy consumption.

Besides the future work based on the real adaption of the solution it's limitations and assumptions, future work could also investigate the performance and "Service Level Agreement" violations of the current infrastructure. With the work additionally investing the potential performance impact and Service Level Agreement violations from the proposed improvements. This would illustrate the trade-off between energy efficiency and performance; to further inform the cloud provider regarding the potential consequences of implementing the methods.

However, future work could also entail digging deeper into energy efficiency by implementing Dynamic Power Scaling approaches like Dynamic Voltage Frequency Scaling for data center persistent storage and network devices. With the goal being throughput of the disks and switch ports being dynamically scaled to dynamically adjust the power usage, as similarly done for the processors of the evaluated physical hosts.

In addition to larger works such as an extension of the analysis and evaluation framework targeting the cooling of the data center. The framework would then analyze the effect of the workloads running and utilization of data center devices and the resulting energy consumption of machines such as a "Computer Room Air Cooling" unit from the cooling

system; to ensure temperature is within an optimal temperature to avoid device failure. With the evaluation aspect of the framework investigating the effect of the workloads on the energy consumption of the cooling. Further the framework could also experiment with estimating task that will create more heat and group them and similarly for the network with workloads requiring more bandwidth being group together; to utilize fewer switches and Computer Room Air Cooling units.

Lastly, future work would entail testing the developed analysis and evaluation framework in a real cloud data center environment. With the goal of testing its efficacy in real world environment in regard to reducing the energy consumption of the data centers of a cloud provider. With the framework being tested through it's estimation and illustration and proposed strategies and the implications of it's selected strategies.

Appendix A

Appendix

A.1 The Evaluation Framework Code

The source code for the analysis and evaluation framework is available on the GitHub repository "https://github.com/KhalidAFarah/cloud-energy-efficiency/tree/main". Where the Python scripts for the analysis of the simulation results is available under the "analysis" folder. While the evaluation scripts for the simulations is available under the "evaluation" folder. With the simulation data extracted stored under the "data" folder.

A.2 Simulation Source Code

In addition to the framework source code, is the simulation source codes also available. However the projects are rather available through the uploaded zip folder which contains the CloudSim, CloudSimDisk and the simple network simulation projects. With the file path to the utilized java for the CloudSim simulation being "cloudsim-6.0-pre/modules/cloudsim-examples/src/main/java/org/cloudbus/cloudsim/examples/ CloudSimImportedDataset.java". While the file path for the utilized java for the CloudSimDisk simulation being "CloudSimDisk/sources/org/cloudbus/cloudsimdisk/examples/Test.java". Lastly, the simple network simulation which is named "app.py" under the "simple-network-simulation" folder.

Bibliography

Alghamdi, M., Tang, B., & Chen, Y. (2017). Profit-based file replication in data intensive cloud data centers. *2017 IEEE International Conference on Communications (ICC)*, 1–7. https://doi.org/10.1109/ICC.2017.7996728

Beeler, B. (2014). Seagate enterprise nas hdd review. https://www.storagereview.com/review/seagate-enterprise-nas-hdd-review

Beloglazov, A., & Buyya, R. (2012). Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency and Computation: Practice and Experience*, *24*, 1397–1420. https://doi.org/10.1002/cpe.1867

Berl, A., Gelenbe, E., Girolamo, M. D., Giuliani, G., Meer, H. D., Dang, M. Q., & Pentikousis, K. (2010). Energy-efficient cloud computing. *Computer Journal*, *53*, 1045–1051. https://doi.org/10.1093/comjnl/bxp080

Bruno, A., & Jordan, S. (2012). *Ccda 640-860 official cert guide*. Cisco Press.

Buyya, R., Ilager, S., & Arroba, P. (2024). Energy-efficiency and sustainability in new generation cloud computing: A vision and directions for integrated management of data centre resources and workloads. *Software - Practice and Experience*, *54*, 24–38. https://doi.org/10.1002/spe.3248

Buyya, R., Ranjan, R., & Calheiros, R. N. (2009). Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: Challenges and opportunities. *2009 International Conference on High Performance Computing & Simulation*, 1–11. https://doi.org/10.1109/HPCSIM.2009.5192685

Cianfrani, A., Eramo, V., Listanti, M., Marazza, M., & Vittorini, E. (2010). An energy saving routing algorithm for a green ospf protocol. *2010 INFOCOM IEEE Conference on Computer Communications Workshops*, 1–5. https://doi.org/10.1109/INFCOMW.2010.5466646

Computer, F. S. (2007, July). Celsius v840 top-end workstation power. https://www.nts.nl/files/product/3198_%5Ben%5D_Celsius%20V840_Sales.pdf

Corporation, S. P. E. (2011a). Hewlett-packard company proliant ml110 g3 (historical). https://www.spec.org/power_ssj2008/results/res2011q1/power_ssj2008-20110127-00342.html

Corporation, S. P. E. (2011b). Hewlett-packard company proliant ml110 g4. https://www.spec.org/power_ssj2008/results/res2011q1/power_ssj2008-20110124-00338.html

Corporation, S. P. E. (2011c). Hewlett-packard company proliant ml110 g5.

Dabaghi-Zarandi, F., & Movahedi, Z. (2018). A dynamic traffic-aware energy-efficient algorithm based on sleep-scheduling for autonomous systems. *Computing*, *100*, 645–665. https://doi.org/10.1007/s00607-018-0589-6

Dabbagh, M., Hamdaoui, B., Guizani, M., & Rayes, A. (2015). Toward energy-efficient cloud computing: Prediction, consolidation, and overcommitment. *IEEE Network*, *29*, 56–61. https://doi.org/10.1109/MNET.2015.7064904

Diaby, T., & Rad, B. B. (2017). Cloud computing: A review of the concepts and deployment models. *International Journal of Information Technology and Computer Science*, *9*, 50–58. https://doi.org/10.5815/ijitcs.2017.06.07

Foundation, P. S. (2024). Python. https://www.python.org/

Garg, S. K., & Buyya, R. (2011). Networkcloudsim: Modelling parallel applications in cloud simulations. *2011 Fourth IEEE International Conference on Utility and Cloud Computing*, 105–113. https://doi.org/10.1109/UCC.2011.24

Guo, B., Yu, J., Yang, D., Leng, H., & Liao, B. (2023). Energy-efficient database systems: A systematic survey. *ACM Computing Surveys*, *55*, 1–53. https://doi.org/10.1145/3538225

Hameed, A., Khoshkbarforoushha, A., Ranjan, R., Jayaraman, P. P., Kolodziej, J., Balaji, P., Zeadally, S., Malluhi, Q. M., Tziritas, N., Vishnu, A., Khan, S. U., & Zomaya, A. (2016). A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems. *Computing*, *98*, 751–774. https://doi.org/10.1007/s00607-014-0407-8

Hanafy, W. A., Liang, Q., Bashir, N., Irwin, D., & Shenoy, P. (2023). Carbonscaler: Leveraging cloud workload elasticity for optimizing carbon-efficiency. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, *7*. https://doi.org/10.1145/3626788

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jiang, D., Wang, Y., Lv, Z., Wang, W., & Wang, H. (2020). An energy-efficient networking approach in cloud services for iiot networks. *IEEE Journal on Selected Areas in Communications*, *38*, 928–941. https://doi.org/10.1109/JSAC.2020.2980919

Jupyter. (2024). Jupyter. https://jupyter.org/

Karpowicz, M., Niewiadomska-Szynkiewicz, E., Arabas, P., & Sikora, A. (2016a). Energy and power efficiency in cloud. https://doi.org/10.1007/978-3-319-44881-7_6

Karpowicz, M., Niewiadomska-Szynkiewicz, E., Arabas, P., & Sikora, A. (2016b). Energy and power efficiency in cloud. https://doi.org/10.1007/978-3-319-44881-7_6

Khan, A. A., & Zakarya, M. (2021). Energy, performance and cost efficient cloud datacentres: A survey. *Computer Science Review*, *40*, 100390. https://doi.org/10.1016/j.cosrev.2021.100390

Lab, S. E. (2013). Hgst ultrastar c10k900 review. https://www.storagereview.com/review/hgst-ultrastar-c10k900-review

Lin, B., Li, S., Liao, X., Wu, Q., & Yang, S. (2011). Estor: Energy efficient and resilient data center storage. *2011 International Conference on Cloud and Service Computing*, 366–371. https://doi.org/10.1109/CSC.2011.6138549

Louis, B., Mitra, K., Saguna, S., & Ahlund, C. (2015). Cloudsimdisk: Energy-aware storage simulation in cloudsim. *2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC)*, 11–15. https://doi.org/10.1109/UCC.2015.15

Mastelic, T., & Brandic, I. (2015). Recent trends in energy-efficient cloud computing. *IEEE Cloud Computing*, *2*, 40–47. https://doi.org/10.1109/MCC.2015.15

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Nishikawa, N., Nakano, M., & Kitsuregawa, M. (2012). Energy efficient storage management cooperated with large data intensive applications. *Proceedings - International Conference on Data Engineering*, 126–137. https://doi.org/10.1109/ICDE.2012.47

nsnam. (2024). Ns-3. https://www.nsnam.org/

of Technology, D. U. (2024). Gwa-t-12 bitbrains. http://gwa.ewi.tudelft.nl/datasets/gwa-t-12-bitbrains

Oracle. (2024). Java. https://www.java.com/en/

pandas development team, T. (2024, April). *Pandas-dev/pandas: Pandas* (Version v2.2.2). Zenodo. https://doi.org/10.5281/zenodo.10957263

Reddy, P. V., & Reddy, K. G. (2023). An energy efficient rl based workflow scheduling in cloud computing. *Expert Systems with Applications*, *234*. https://doi.org/10.1016/j.eswa.2023.121038

Ruan, X., Qin, X., Zong, Z., Bellam, K., & Nijim, M. (2007). An energy-efficient scheduling algorithm using dynamic voltage scaling for parallel applications on clusters. *2007 16th International Conference on Computer Communications and Networks*, 735–740. https://doi.org/10.1109/ICCCN.2007.4317905

Smith, L. (2015). Toshiba mg04sca enterprise hdd review. https://www.storagereview.com/review/toshiba-mg04sca-enterprise-hdd-review

Tang, Z., Qi, L., Cheng, Z., Li, K., Khan, S. U., & Li, K. (2016). An energy-efficient task scheduling algorithm in dvfs-enabled cloud environment. *Journal of Grid Computing*, *14*, 55–74. https://doi.org/10.1007/s10723-015-9334-y

Tate, J., Beck, P., Clemens, P., Freitas, S., Gatz, J., Girola, M., Gmitter, J., Mueller, H., O'Hanlon, R., Para, V., et al. (2013). *Ibm and cisco: Together for a world class data center*. IBM Redbooks. https://books.google.no/books?id=DHjJAgAAQBAJ

Technologies, M. (2013). Power saving features in mellanox products. https://network.nvidia.com/pdf/whitepapers/WP_ECONET.pdf

Tu, Y.-C., Wang, X., Zeng, B., & Xu, Z. (n.d.). A system for energy-efficient data management.

UMassTraceRepository. (2023, January). U mass trace repository. https://traces.cs.umass.edu/index.php/Storage/Storage

Vrbsky, S. V., Galloway, M., Carr, R., Nori, R., & Grubic, D. (2013). Decreasing power consumption with energy efficient data aware strategies. *Future Generation Computer Systems*, *29*, 1152–1163. https://doi.org/10.1016/j.future.2012.12.016

Wang, J., Yu, J., Song, Y., He, X., & Song, Y. (2023). An efficient energy-aware and service quality improvement strategy applied in cloud computing. *Cluster Computing*, *26*, 4031–4049. https://doi.org/10.1007/s10586-022-03795-w

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Yao, W., Wang, Z., Hou, Y., Zhu, X., Li, X., & Xia, Y. (2023). An energy-efficient load balance strategy based on virtual machine consolidation in cloud environment. *Future Generation Computer Systems*, *146*, 222–233. https://doi.org/10.1016/j.future.2023.04.014

Zeadally, S., Khan, S. U., & Chilamkurti, N. (2012). Energy-efficient networking: Past, present, and future. *Journal of Supercomputing*, *62*, 1093–1118. https://doi.org/10.1007/s11227-011-0632-2