

IMPROVING CARBON, COST, AND ENERGY EFFICIENCY OF LARGE
SCALE SYSTEMS VIA WORKLOAD ANALYSIS

by

Bradford W. Everman, B.A., M.A., M.C.S.

A dissertation submitted to the Graduate College of
Texas State University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a Major in Computer Science
May 2022

Committee Members:

Ziliang Zong, Chair

Byron J. Gao

Apan Qasem

Damian Valles

COPYRIGHT

by

Bradford W. Everman

2022

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

This work is protected by the Copyright Laws of the United States (Public Law 94-553, section 107). Consistent with fair use as defined in the Copyright Laws, brief quotations from this material are allowed with proper acknowledgement. Use of this material for financial gain without the author's express written permission is not allowed.

Duplication Permission

As the copyright holder of this work I, Bradford W. Everman, refuse permission to copy in excess of the "Fair Use" exemption without my written permission.

DEDICATION

This dissertation is dedicated to God, who provided me with divine inspiration when I had no idea how to proceed, and resolve when I wanted to quit.

I can do all things through Christ which strengtheneth me.

- Philippians 4:13

ACKNOWLEDGEMENTS

I'd like to thank the academy. No, not that one. The other one. The classical Platonic Academy, the belief that knowledge was possible beyond simple self-reflection and was achievable through empirical means. While I want to believe hoisting a golden statue of a possibly naked man above my head is at all meaningful, seeking the truth is far more admirable and indeed, more fulfilling. The unexamined life is not worth living, after all.

My wife is also owed a debt of gratitude for putting up with me during the process of completing this work. As Aristotle stated so succinctly, no great mind has ever existed without a touch of madness. And she definitely had to deal with that madness. She said, "It wasn't that bad," but she's a terrible liar and for that I'm grateful. This work is as much hers as it is mine, and I would not be complete without her.

Dr. Zong played a pivotal part in guiding me through this process, and never once doubted my capabilities as a researcher. His suggestions were always intended to produce the best possible product, and I think his success in that endeavor is evident.

Other people who need a shout out include Mark for helping me put up paneling in my cabin when I needed a break from reality, Robert for providing adult beverages and a sounding board for complaining about college football, my kids for reminding me what's really important, Captain America, Seagrams VO, and my dad for teaching me how to find a 10mm socket underneath a giant pile of wrenches. That's a very useful skill.

Lastly, I want to thank my mom for pretty much everything.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
LIST OF TABLES	x
LIST OF FIGURES	xi
ABSTRACT	xiii
CHAPTER	
I. INTRODUCTION	1
Research Areas	3
Web Hosting	3
Cloud Computing	3
AI Workloads	4
Contributions and Solutions	6
Web Hosting	6
Private Cloud	7
Public Cloud	8
AI - Foundation Models	10
Dissertation Organization	11
II. WEB HOSTING - GREEN WEB	12
Introduction	12
Related Work	13
The FEST Website and Test Environment	14
The FEST Website	14
Physical Servers	15
Web Hosting Environment	16
Preliminary Performance Evaluation	17
User Behavior Analysis	17
Power Measurement	18
Uptime	19
Testing Methodology	19
Software Optimization	21

Caching	21
Content Delivery Network (CDN)	22
Data Collation	22
Results	23
Response Time	23
Base Case	24
Cache Enabled	24
Cache and CDN Enabled	27
Linux System Load Average	27
Base Case	28
Cache Enabled	28
Cache and CDN Enabled	28
Power	31
Summary	34
Raspberry Pi	35
Synology DS718+	35
Synology DS1815+	36
Mac Mini	36
Dell PowerEdge	37
Extended Experiments	37
Hardware Configuration	38
Software Configuration	38
Network Configuration	39
User Behavior Analysis	40
Power Measurement	41
Uptime	41
Testing Methodology	41
Results	43
Observations	43
Conclusions and Future Work	46
III.PRIVATE CLOUD WORKLOAD - ALIBABA	48
Introduction	48
Related Work	50
Alibaba Workload Analysis	50
Cloud System Scheduler Simulator	52
Background	53
Utilization Matters	53
Performance Dominates	54
Alibaba Scheduler	55

Workload Analysis	56
Alibaba Dataset	56
Machine Utilization	58
Batch Jobs	59
Service Jobs	63
Summary	64
Simulator Design	64
Fuxi Scheduler Workflow	65
Terminology	67
Job Submission	69
Simulator Input	69
Simulator Output	71
Limitations	71
Experimental Results	73
Impact of Cluster Size	73
TCO Reduction	77
Conclusions and Future Work	79
IV. PUBLIC CLOUD WORKLOAD - AZURE	81
Introduction	81
Related Work	83
Workload Analysis	86
Microsoft Azure	86
Azure Trace Files	86
Descriptive Analysis	87
Cloud Waste and Cost Analysis	91
Azure Pricing Model	92
VM Cost Calculation	93
Green and Red VMs	94
VM Cloud Waste Points	97
Green and Red Users	98
Cloud Utilization Score	102
Recommendation Algorithm and Cost Savings	104
Why Should Cloud Providers Care?	108
Limitations	108
Conclusions and Future Work	109
V. AI WORKLOADS - FOUNDATION MODELS	111
Introduction	111
Related Work	114

AI Workload Selection	116
Natural Language Processing	116
Foundation Model Selection	117
NLP Prompt Selection	118
Testing Environment	119
Physical Platform	120
Python Environment	120
Power Measurement	121
Scripting	121
AI Workload Analysis	122
Output Evaluation	123
Power	123
Writing Performance	125
Software Carbon Intensity	129
Conclusions and Future Work	132
VI. CONCLUSIONS	137
APPENDIX SECTION	140
REFERENCES	193

LIST OF TABLES

Table	Page
2.1 The Specification of Evaluated Servers	16
2.2 Performance Comparison	17
2.3 Locust Page Weighting	40
3.1 Summary of the 2017 and 2018 Alibaba Cluster Traces	58
3.2 Average CPU, Memory, and Disk utilization	59
3.3 Energy Consumption Comparison	78
3.4 TCO Comparison	78
3.5 Extrapolated TCO Comparison	79
4.1 2019 Azure VM Distribution	89
4.2 2019 Azure VM Pricing in USD per hour	94
4.3 Average CPU Utilization of Different VMs	95
4.4 Delay-Insensitive VM Distribution	95
4.5 VM Waste Factor	98
4.6 VM Cloud Waste Point Distribution	99
4.7 User Score by Percentile	103
4.8 Cost Savings Examples	103
4.9 Additional Cost Savings Examples	105
5.1 Foundation Model Size Comparison	121
5.2 Average Foundation Model Power Usage - CPU Only	124
5.3 Average Foundation Model Power Usage - GPU Accelerated	125
5.4 Average Foundation Model Writing Performance - CPU Only	127
5.5 Average Foundation Model Writing Performance - GPU Accelerated	128
5.6 SCI for Server Lifetime - CPU Only	131

LIST OF FIGURES

Figure	Page
1.1 Top Cloud Initiatives for 2021[1]	4
1.2 Energy for AI Training[2]	5
2.1 A sample Locust script	20
2.2 A sample distribution file	21
2.3 DS718+ Response Time	25
2.4 DS1815+ Response Time	25
2.5 Mac Mini Response Time	26
2.6 PowerEdge Response Time	26
2.7 Pi Response Time	27
2.8 DS718+ Load Average	28
2.9 DS1815+ Load Average	29
2.10 Mac Mini Load Average	29
2.11 PowerEdge Load Average	30
2.12 Pi Load Average	30
2.13 DS718+ Power Usage	31
2.14 DS1815+ Power Usage	32
2.15 Mac Mini Power Usage	32
2.16 PowerEdge Power Usage	33
2.17 Pi Power Usage	34
2.18 Extended Experiments Results	44
3.1 Alibaba Scheduler	55
3.2 7-Day CPU Utilization	56
3.3 CPU Utilization by Hours	57
3.4 Memory Utilization by Hours	60
3.5 Disk Utilization by Hours	61
3.6 CPU Subscription of Batch Jobs	62
3.7 Resource Subscription of Service Jobs	63
3.8 Fuxi Scheduler Resource Allocation Workflow	66
3.9 Fuxi Scheduler Resource Deallocation Workflow	67
3.10 Simulating Job Submissions using Agenda	69
3.11 Wait Time Until First Grant	74

3.12	Number of Grants	74
3.13	Task Wait Time Until First Grant	75
3.14	Task Wait Time Until Last Grant	75
3.15	Wait Time - 2018 Cluster	76
4.1	VM Core Distribution	89
4.2	VM Memory Distribution	90
4.3	Memory Allocation by Number of Cores	90
4.4	Category Distribution	91
4.5	CWI Distribution by User	101
5.1	Cost-Benefit Analysis of AI Models [3]	112

ABSTRACT

The global COVID-19 pandemic has transformed the way businesses utilize digital technologies, with an increasing reliance on cloud resources due to the paradigm shift from traditional to work-from-home models. Cloud computing resources are expected to expand annually by 14.8% from 2022 to 2030, a three-fold increase overall, driven by more and more reliance on decentralization and the changing workplace. As the need for large scale systems continues to grow, the cost, energy consumption and carbon footprint have increased accordingly at unprecedented rates. It is expected that digital industry will contribute to 14% of global greenhouse gas emissions by 2040. Therefore, it is essential to put sustainability at the core of digital technologies and reduce their operating cost and negative impact on environment.

In the past decades, scientists and industry pioneers have made tremendous endeavors in improving the energy efficiency of various digital technologies. The exemplary achievements including, but not limited, to using more energy efficient hardware such as GPUs, FPGAs, and AISCs to solve appropriate problems, using Power Usage Effectiveness (PUE) as a metric to measure the energy efficiency of data centers, using the big.LITTLE architecture to balance the high performance and low power needs of mobile applications, using Dynamic Voltage and Frequency Scaling (DVFS) to decrease energy consumption based on overall system load, using virtual machines to share resources in the cloud, using carbon-aware scheduling to allocate jobs to the least wasteful or most carbon efficient resources, as well as neuromorphic computing that mimics a human brain to minimize the energy consumption of AI applications. All these works have significantly advanced the

research and industry practices of sustainable computing. However, the ever growing data volume and more complex workloads running on large scale systems have brought the challenges to a whole new level. How to improve the carbon, cost, and energy efficiency of large scale systems from the big data and workload analysis perspective has not been fully studied in the literature.

This dissertation explores ways to improve energy efficiency via workload analysis, which provide the additional benefits of improving carbon emissions and lowering operational costs for both large systems and the end-users relying on those system. More specifically, it investigates three typical workloads that have high energy requirements and are widely deployed: website workloads, cloud workloads, and AI workloads. The study of website workloads monitored the power consumption of five different types of web servers and recorded the quality of service (QoS) provided by those servers while simulating real user load. The results demonstrated that a low-powered web server can provide comparable QoS to a higher powered one in many instances. For private cloud workloads, the 2017 and 2018 Alibaba cluster traces were analyzed, and a simulator was designed to test the effectiveness of decreasing the number of servers while maintaining the required level of performance. The simulation results showed that decreasing the number of servers by 5% resulted in negligible impact on performance while lowering yearly electricity costs. A public cloud workload analysis was conducted using the 2019 Microsoft Azure trace, which revealed that a large portion of VMs was underutilized thus wasting significant amount of energy and resources in the cloud. A recommendation algorithm was proposed to help cloud users reduce cost without compromising QoS. Lastly, the energy efficiency and carbon emissions of several

foundation AI models were analyzed using the recently released industry standard - Software Carbon Intensity (SCI), which provided an effective methodology on evaluating the environmental impact of large-scale AI models and shed lights on future design of green AI.

I. INTRODUCTION

The demand for computing resources is climbing every year, accounting for approximately 1% of global electricity usage[4]. Further, cloud computing resources are expected to expand annually by 14.8% from 2022 to 2030, a three-fold increase overall[5]. As this demand continues to rise quickly, the energy consumption and the cost continues to increase. Thus, it is no surprise that 76% of surveyed decision makers for enterprise-level corporations and small businesses (SMB) use cost efficiency as a metric to measure the effectiveness of their computing deployments[1]. Being able to sustain growth is paramount to a company, and reducing overall operational costs is a major component of maintaining that growth. Meanwhile, the energy demand and carbon emissions of digital technologies are increasing at an unprecedented rate. Digital technologies currently account for 2-3% of global greenhouse gas (GHG) emissions, but this footprint is expected to grow to account for as much as 14% of global GHG emissions by 2040[6].

In line with the 2015 Paris Agreement, the United States has committed billions of dollars for clean energy research due to concerns about the negative environmental impact of conventional energy sources. Although this certainly helped decrease carbon footprint, it does not reduce total energy consumption[7]. CEOs of major corporations might purchase green energy, such as from wind and solar, to achieve their carbon neutral goals. However, the systems running their workloads still require the same amount of energy, whether or not it comes from a cleaner source. Although offsetting carbon emissions is a good first step, decreasing overall energy usage is essential for sustainable growth. Lower absolute energy usage also directly results in less carbon waste, regardless of source, so the environmental impact of such reductions is extremely significant.

Scientists and industry pioneers have made significant efforts to continuously

reduce the energy waste and carbon footprint of digital technologies in the past decades. Power Usage Effectiveness (PUE), for instance, has been widely utilized as a metric by all large data centers operated by Google, Facebook/Meta, Amazon, and Alibaba to evaluate energy efficiency. Dynamic Voltage and Frequency Scaling (DVFS) has been integrated into nearly every CPU/GPU to minimize energy waste by lowering voltages and frequencies when the system is not busy. Most AI training is now done on GPUs or ASIC chips instead of CPUs to achieve higher energy efficiency. The big.LITTLE architecture has become the industry standard for mobile computing, which provides an effective solution to balance the high performance and low power requirements of mobile applications. FPGAs have also been used to replace conventional general purpose CPUs for certain tasks to save energy without degrading performance. Emerging technologies such as neuromorphic computing, though not yet widely deployed, have shown great potential to further advancing the frontier of green AI by mimicking how human brains process complex information. There is no doubt that these previous achievements have made significant contributions for sustainable computing. However, the data explosion and increasingly complex workloads have brought the challenges to a brand new level. Unfortunately, our knowledge of how to improve the carbon, cost, and energy efficiency of large scale systems from the big data and workload analysis perspective is still very limited. This dissertation strives to fill the gap by exploring how to improve the carbon, cost, and energy efficiency of large scale systems from the big data and workload analysis perspective. Specifically, I studied this topic in the subsequent research areas.

Research Areas

Web Hosting

Websites are nearly ubiquitous, and used on a daily basis by the average person. A 2018 Netcraft study indicated there are over 1.7 billion websites worldwide, hosted on more than 7 million web servers[8]. Enterprise websites are commonly deployed on energy-hungry servers to ensure quality of service (QoS) is maintained. Typically, however, these web servers are mostly underutilized. Per a 2015 report published by Forbes, enterprise server utilization is a mere 5% - 15%[9]. This low server utilization is also supported by a more recent 2018 report from HostGator, a popular web-hosting platform[10]. Thus, hosting websites on high-end servers may not be a cost-effective, nor energy-efficient, solution. Maximizing energy usage while still maintaining QoS may be possible using lower-end hardware, even while going against conventional wisdom, but hard data is needed to ensure this is the case.

Cloud Computing

The second area targeted by this dissertation was cloud computing, chiefly due to the massive move from conventional computing to the cloud following the COVID-19 pandemic. According to the Flexera 2021 State of the Cloud report [1], 59% of global decision makers have made migrating more workloads to the cloud as one of their top priorities (see figure 1.1). There are multiple reasons for this migration, but the most important one is cost-effectiveness. It is generally cheaper to run a workload in the cloud than deploying it on local systems, especially when most users are accessing those resources remotely. However, resource costs can be difficult to understand and plan for when compared to physical systems, and energy consumption is generally factored in as an operational cost. In the past the cost of a server could be calculated in a straightforward manner, but the wide variety of VMs

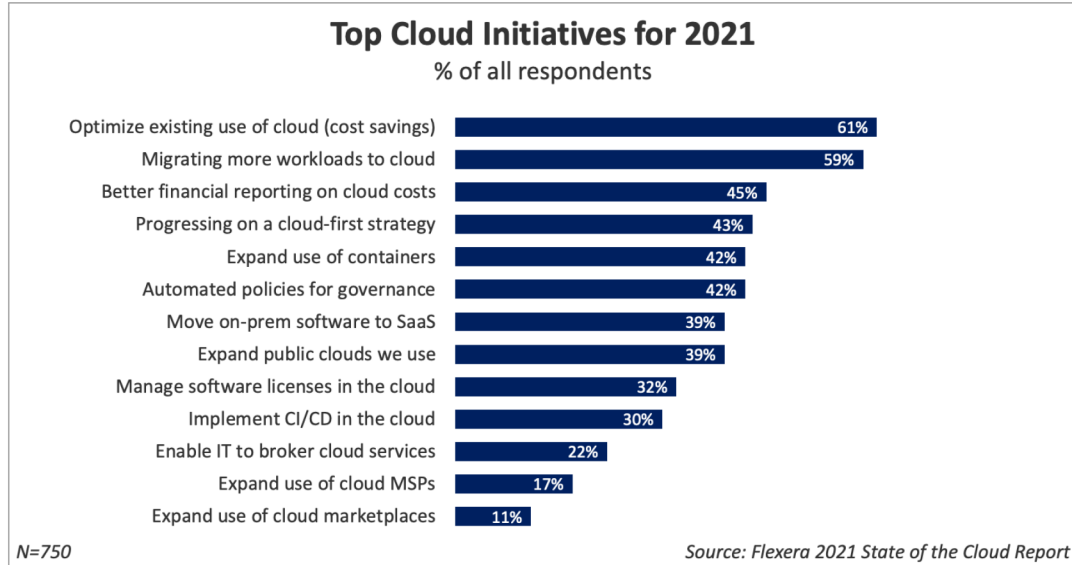


Figure 1.1: Top Cloud Initiatives for 2021[1]

and different virtualization platforms make this more difficult to financially forecast. Further, determining the environmental impact of cloud computing typically is not a primary concern for most users, while providers may have difficulty measuring carbon waste directly and relaying that information to their customers.

Hence, cloud users are often unsure as to not only what resources they require, but how much they should pay for those resources, minimize their carbon impact, and maximize energy efficiency. Cloud providers must also justify their pricing schemes lest potential customers become frustrated due to the lack of information while ensuring their systems are operating as efficiently as possible.

AI Workloads

The demand for AI workloads is growing at an increasing rate and shows no signs of stopping. Per a Fortune Business Insights study, the demand for AI will increase in size ten-fold by 2027, which comes with massive energy requirements[11]. Figure 1.2 illustrates this surge in demand in more detail, as the amount of computing resources used to train deep learning models increased by over 300,000x

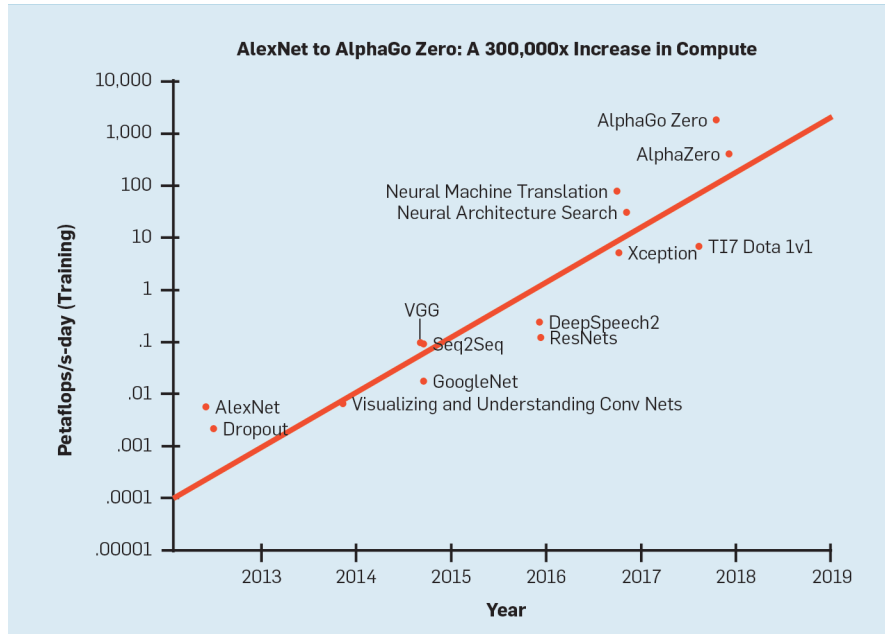


Figure 1.2: Energy for AI Training[2]

between 2013 and 2019. Given the extremely power-hungry nature of AI workloads, this poses a huge problem to the existing infrastructure, and mitigating energy expenditures is important. Foundation models in particular require massive amounts of energy to train and run, and are excellent candidates to analyze for reduction. Further, foundation models are employed in a wide variety of domains and can be adapted to perform multiple tasks, but are prone to excessive carbon emissions[3]. For example, training the popular foundation model GPT-3 is estimated to have required 190,000 kWh of energy and produced 85,000 kg of CO₂ waste[12]. Due to their increasing popularity, it has become critically important to fully understand and quantitatively evaluate the energy consumption and environmental impact of large AI models before deploying them ubiquitously. Unfortunately, systematic research in this field is still in its infancy.

Contributions and Solutions

To address the problems outlined above, I conducted four comprehensive workload analysis studies and made the following contributions in the fields of web hosting, cloud computing, and AI.

Web Hosting

This study provided a detailed analysis to verify that websites can be hosted on low-power systems instead of traditional high-end servers, which can significantly decrease operational costs and energy consumption without sacrificing QoS. More specifically,

1. It performed a comprehensive study using four low-powered systems (Synology 718+ and 1815+, Apple Mac Mini, Raspberry Pi 3) and one conventional datacenter server (Dell PowerEdge 410). The total energy consumption under identical workloads was measured while recording the QoS when serving these workloads. The workloads were derived from a real-world website with synthetic users that simulated real users retrieving information from that website.
2. It verified that deploying high-load websites on low-power servers is a viable alternative solution to the current industry standard rack server. Specifically, it is possible to replace a conventional high-powered server (e.g. PowerEdge 410) with a low-powered one (e.g. Synology 718+) while maintaining a great QoS (221 ms response time) at up to 43 million requests/month, and a good QoS (731 ms response time) at up to 65 million requests/month.
3. It demonstrated that employing SSDs, caching, content delivery networks (CDNs), and other optimizations can greatly improve the ability of low-power

servers. An increase of 250% more requests/month is possible when leveraging such optimizations, while maintaining a near-identical response time and QoS. The experimental results showed the Synology 718+ server achieved comparable QoS to the PowerEdge 410 server with 43 million req/month with utilizing only 9% as much power.

Private Cloud

This study conducted a comprehensive analysis on the Alibaba datasets, which revealed that the over-subscription and under-subscription problems co-exist in the current Alibaba system. A complex simulator was developed to emulate the execution of millions of jobs being scheduled at the Alibaba system and simulation results showed that the total cost and energy consumption can be greatly reduced without degrading system performance.

1. The study analyzed the Alibaba trace to quantitatively show that consolidating hybrid workloads can considerably improve the utilization of large-scale cloud systems and reduce TCO. Further, the analysis on the 2018 trace showed that the servers running only batch jobs (time-insensitive) or service jobs (time-sensitive) have an average CPU utilization of 29.29% and 7.4% respectively. By applying the mixed scheduling strategy, the average server utilization was improved to 39.26%. This can lead to millions of dollars of total cost of ownership (TCO) reduction purely from the reduced number of servers without compromising QoS.
2. The study discovered that jobs submitted to the Alibaba cluster request an unreasonable amount of resources. A large portion of jobs (both batch and service jobs) aggressively over-subscribe resources. Meanwhile, a considerable number of jobs (both batch and service jobs) under-subscribe resources. This hurts the Alibaba system in both ways. The under-subscribed jobs suffer from

performance loss while the over-subscribed jobs waste energy and lower system utilization.

3. A simulator was developed to analyze large-scale cloud workloads and reproduced the hybrid scheduling algorithm that is currently utilized by Alibaba and deployed on their cloud system. It can accept the real-world workloads provided by Alibaba via the 2017 (1,313 servers, 1GB data) and 2018 (4,023 servers, 450GB data) cluster trace files. The simulator provided a feasible approach to study the impact of reducing the number of machines on cost reduction, since direct access to the Alibaba system is not possible. The simulation results demonstrated that the private Alibaba cloud system can be safely downsized by 5% with no negative impact to the work performed. This reduction offers substantial savings to operational costs and reduces energy consumption. Specifically, the 4,023 systems analyzed in the 2018 cluster trace files can be migrated to 3,832 systems. This decreases the initial TCO by \$600,574, and reduces annual energy usage by 115,239 kWh, saving an additional \$12,676 per year.

Public Cloud

This study conducted a comprehensive analysis of the Microsoft Azure 2019 traces to help cloud users reduce cost and waste by identifying underutilized VMs and recommending lower priced VMs without compromising QoS. More specifically, it made the following contributions:

1. It analyzed 235GB of data containing nearly 2.7 million VMs created by 6,687 Azure cloud users within 30 days, a total of nearly two billion VM readings. The results revealed that a large portion of VMs are under-utilized or over-provisioned for resources.

2. It performed a comprehensive, descriptive analysis of VMs running within the Azure cloud and categorized those VMs to better understand their computation impact on the cloud. It was noted that VMs were categorized by CPU core count, requested memory, and the type of VM employed, both Delay-Insensitive which do not need on-demand resources, and Interactive VMs which must be available real-time.
3. It proposed Cloud Waste Points (CWP) to quantitatively evaluate the waste of VMs in utilizing cloud resources. VMs that utilized cloud resource efficiently were categorized as green VMs and those that wasted cloud resources as red VMs, according to their average CPU utilization rate. To help users optimize cost and waste, an algorithm was introduced to identify red VMs and recommend more appropriate VMs for users to run their cloud applications with less cost and waste. These experiments showed that over \$22 million cost savings (i.e. approximately 36% in total cost reduction) can be achieved if the proposed recommendations are adopted by users.
4. It proposed and calculated the Cloud Utilization Score (CUS) of 6,687 Azure users, identifying those with significant cloud impact (6,121 total), allowing classification as green users, normal users, and red users. Green users were those who generally utilized cloud resource efficiently, and red users were those who tended to over-allocate resources, thus creating more waste. The resource usage of normal users was in between. The analysis showed that 1,020 users (~15%) were classified as red users while 2,321 users (~36%) could be classified as green users. The Cloud Utilization Score provided a straightforward and intuitive way to help users compare their cloud usage efficiency to other users' efficiency.

AI - Foundation Models

A study was conducted on foundation models in the domain of Natural Language Processing (NLP) to determine if these models could be classified by quality of output and environmental impact. This study provided the following contributions:

1. Utilizing the Software Carbon Intensity (SCI) metric provided by the Green Software Foundation, this study showed that the environmental impact of foundation models can be quantitatively measured and compared. GPT-J 6B, for instance, yielded an SCI of 8.16 gCO₂e, and GPT-Neo 2.7B an SCI of 4.05 gCO₂e.
2. It showed the quality of a foundation model can be found by objectively analyzing the output. For this study, the writing rubric supplied by the State of Texas (STAAR) was utilized to calculate the Writing Performance Score of foundation models, qualitatively scoring those outputs on a scale between 3 and 12. These scores can be directly compared to determine higher quality outputs. For example, the Writing Performance Score of GPT-J 6B at 5.83 is 12% lower than GPT-Neo 2.7B with a score of 6.63.
3. It showed it is possible for a foundation model to be replaced by a more efficient model, mitigating energy requirements and reducing carbon waste while maintaining a similar level of expected output. GPT-Neo 1.3B (Writing Performance Score of 6.17) provided a similar quality of output to GPT-J 6B while using only 27% as much energy, making it an excellent replacement candidate.
4. It demonstrated that there is no significant difference in the quality of output of foundation models when using either CPU-only or GPU-acceleration. It

further showed that GPU-acceleration can provide up to a 91% reduction in overall energy consumption when employing foundation models.

Dissertation Organization

This dissertation contains five chapters.

Chapter I summarizes the motivation and major contributions of this dissertation.

Chapter II is a direct analysis of using low-powered servers for web hosting (one of the most commonly deployed cloud applications)[13]. This chapter focuses on currently a deployed website specifically to ensure real-world applicability and provides a rough benchmark of QoS.

Chapter III is an implementation of a running simulator that analyzes a cluster of Alibaba servers performing both real-time and scheduled tasks[14].

Chapter IV considers the more general nature of cloud computing by analyzing Microsoft Azure trace files[15]. Cloud Waste Points and a Cloud Waste Indicator are introduced as new metrics for classifying VMs.

Chapter V focuses on AI workloads, identifying power-hungry models and determining if reducing the energy requirements is viable. Specifically, natural language processing (NLP) outputs from foundation models are compared, and Software Carbon Intensity (SCI) is also calculated to determine expected carbon emissions produced when running the models.

Chapter VI concludes this dissertation.

II. WEB HOSTING - GREEN WEB

Introduction

In 2018, Netcraft reported that there are over 1.7 billion websites worldwide, hosted on more than 7 million web servers[8]. To guarantee the response time of web services, most enterprise websites with high- or medium- load are deployed on powerful and energy-hungry servers. It has become a common practice and conventional wisdom that high-load websites must be deployed on high-end servers, despite the fact these web servers are largely underutilized during their daily operation. According to the McKinsey report[9], typical enterprise server utilization is merely 5% - 15%. This low server utilization is also supported by a report from HostGator, a popular web-hosting platform[10].

With such a low utilization, hosting websites on high-end servers may not be a cost-effective solution. Suppose we have two server options to host a website. The high-end server has two Intel E5-2630 processors with a Passmark score of 16,123[16] and a total thermal design power (TDP) of 160W. Passmark is a widely used benchmarking utility that rates the performance of a server. The low-power server has a quad-core Intel Celeron J3455 processor with a TDP of only 10W, and a much lower Passmark score of 2,138[17]. Comparing the Passmark score and TDP of both processors, the low-end Celeron processor is capable of performing twice as much work per watt at full load. This sheds light that using low-power servers to host websites could possibly be a viable and more effective solution, provided that the QoS can meet user expectations.

This dissertation answers three basic questions:

1. Is deploying high-load websites on low-power servers a viable alternative solution overall?

2. If not a completely viable replacement, what level of service is a low-power server capable of maintaining?
3. What software optimization can help improve the QoS of high-load websites running on low-power servers?

Related Work

The majority of previous literature focused on using Dynamic Voltage and Frequency Scaling (DVFS) or demand based workload scheduling algorithms to reduce the energy and operational cost of web servers. For example, Abbasi et al. proposed a solution of serving websites from different data centers based upon Dynamic Application Hosting Management (DAHM)[18]. DAHM leveraged physical servers from different geographic regions and reduced latency and consequently power use by communicating less with clients. Similarly, Deng et al. introduced distributed web hosting, which chose cloud-based hosts based on their ability to meet strict service-level agreement (SLA) requirements while simultaneously employing a low carbon footprint[19]. Al-Qudah et al. proposed a dynamic hosting algorithm, which allocated resources to websites based on demand and using DVFS to reduce energy consumption of servers whenever the demand is low[20]. Chen et al. proposed a hybrid mechanism by combining the state queuing analysis and the feedback control theory to reduce energy consumption without sacrificing SLAs[21]. IBM researchers Elnozahy et al. built a simulator to evaluate the impact of different energy conservation policies on web servers[22]. The two key techniques to control different policies on web servers were DVFS and request batching, which saved 17% - 42% of energy for workloads across a broad range of intensities. Bohrer et al. conducted a case study using the 1998 Winter Olympics website [23]. They were able to reduce 23% - 36% of CPU energy by using DVFS while keeping server responsiveness within reasonable limits. Svanfeldt-Winter et al. conducted a cost

and energy evaluation of ARM-based web servers, compared with Intel Xeon-Based servers[24]. Varghese et al. proposed a cluster of Raspberry Pi systems to supplant conventional datacenter servers, demonstrating the viability of low-powered systems as a possible web server platform[25]. This is further demonstrated in the work done by Pahleval et al. which indicates that extremely low voltage systems ARM-based systems can provide reliable QoS[26]. Another solution, NapSAC, is proposed by Krioukov et al.[27]. By leveraging low-powered servers in a heterogeneous environment, NapSAC seeks to lower power usage while maintaining QoS via a scheduling algorithm.

The scope of most existing literature is limited to improving the energy efficiency of high-end servers through various optimization, or leveraging low-powered systems in a traditional environment. This dissertation is distinguished from aforementioned studies by focusing on low-power servers that are traditionally considered incapable of hosting high-load websites and can be utilized "off-the-shelf", without any special configuration. Moreover, this study goes beyond traditional DVFS algorithms and explores the impact of caching and Content Delivery Network (DNS) on QoS and energy efficiency.

The FEST Website and Test Environment

The FEST Website

To ensure the quality of evaluation, a crucial step was to select a website that could represent real-world workloads of high-load websites powered by recent and popular web technologies. I selected the FEST website, which provides information for a large annual festival in Texas (the URL of the website is undisclosed for privacy concerns), for a number of reasons:

1. Availability - Direct administrative access of the website was available at the

time this research was conducted.

2. Analytics - Web traffic statistics to determine user requests were accessible via Google Analytics and direct server traffic logs.
3. Applicability - This week-long festival has several hundred thousand visitors who use the website to view event scheduling and purchase tickets, thus providing a legitimate source of real-world workload. Although the FEST website has low traffic during the off-season, it receives about 60,000 requests per day during the actual event. The day with the peak traffic has approximately 84,000 page views.
4. WordPress - FEST is powered by the WordPress Content Management System (CMS) - one of the most popular web technologies today. WordPress is deployed on approximately 28.9% of all websites[28] such as CNN, CBS, BBC etc. Therefore, measuring a web host's ability to serve a WordPress site is an excellent indicator of real-world relevance. In addition, Murwantara et al.[29] presented a pragmatic method of measuring the performance and energy of virtualized, cloud-based web servers running WordPress, which can be adopted in the following experiments.
5. Portability - FEST uses a plugin called All-In-One WordPress Migration[30] - which creates an archived version of FEST in its current state. The plugin provides the capability to import a 100% accurate copy of the WordPress site, which ensured FEST would run identically on all the low-power servers that were selected for evaluation.

Physical Servers

Table 2.1 shows the specifications of five servers that were evaluated in the experiments, covering a wide spectrum of devices ranging from a Dell PowerEdge to

Table 2.1: The Specification of Evaluated Servers

Name	CPU	Memory	Disk	CPU TDP
Synology DS718+ [31]	Celeron J3455	10GB DDR3	2x120GB SSDs in RAID1	10W
Synology DS1815+ [32]	Atom C2538	6GB DDR3	4x6TB HDD in SHR [33]	15W
Apple Mac Mini [34]	i5-3210M	16GB DDR3	1TB Fusion Drive [35]	35W
Dell PowerEdge 410	2x Xeon E5504	16GB UDIMM	2x500GB SAS HDD in RAID1	2x80W
Raspberry Pi 3 Model B [36]	ARM Cortex-A53	1GB	32GB Micro SD card	<1W

a Raspberry Pi.

Web Hosting Environment

Ubuntu (Server Version 16.04.4 LTS [37]) was selected as the host operating system to deploy FEST because the ease of which it is to install, and its compatibility with all servers except the Raspberry Pi. Debian Raspbian [38] was chosen as the operating system for Raspberry Pi instead. The two Synology servers run a proprietary operating system called DSM [39], which provides a Virtual Machine (VM) manager and allows a VM to be deployed as needed. For both Synology devices, only one VM was created utilizing all available processor cores, 20 GB of disk space, and 2 GB of memory. As peak memory usage never exceeded 1 GB during testing, limiting the memory had no performance impact and allowed ease of VM migration between devices.

After installing a barebone Ubuntu/Raspbian on each system or VM (including an ssh daemon for remote access), Virtualmin [40], a LAMP [41] administration interface, was installed immediately. Virtualmin downloads, installs, and configures the entire web hosting environment, and provides a means to easily deploy a WordPress instance automatically via Perl scripts. This greatly accelerated the deployment process while maintaining the desired homogeneous environment. The LAMP stack was configured with Apache 2.4.18 (Ubuntu), PHP 7.0.28-0ubuntu0.16.04.1, and MySQL 14.14 Distrib 5.7.21, the most current versions in the Ubuntu 16.04 LTS repository. As Raspbian is based on Debian Stretch, the Pi

Table 2.2: Performance Comparison

System	Disk (MB/s)	PHP (seconds)
DS718+	397.8	1.14
DS1815+	285.8	1.42
Mac Mini	123.4	0.51
PowerEdge	85.9	0.93
Pi	19.2	4.65

was configured with Apache 2.4.25-3+deb9u4, PHP 7.0.27-0+deb9u1, and MariaDB 10.1.26-0+deb9u1.

Using the Virtualmin scripting utility, WordPress 4.9.5 was installed in the base directory of the virtual host. The FEST website was then imported with the migration plugin, and tested to ensure all content was loaded and all links worked properly.

Preliminary Performance Evaluation

After the operating system and supporting environment were setup correctly, I evaluated the baseline performance of each server using a suite of benchmarks. More specifically, the disk benchmark is a simple file copy which performs both read and writes, while the PHP benchmark [42] is a script which does several common operations such as string copy, floating point operations, and loop testing. I ran each benchmark five times, removed outliers, and recorded the average disk speed and performance of basic PHP operations in Table 2.2.

User Behavior Analysis

To stress test each server, a tool was required that could accurately emulate the page view behavior of users. I analyzed the historical user behavior of the FEST website and observed that typical users would load the main page first, then select a

menu item (usually schedule or buy-tickets). As the main page also displays part of the schedule (the daily listing as opposed to the complete listing on the schedule page), it usually takes between 1 - 6 seconds for users to scan the currently displayed listing. I also observed that the main page, buy-tickets page, and schedule page constituted over 70% of the total website traffic, and the frequency of visiting these pages is different.

Given these observations, I created a load test tool using Locust[43], which was selected for its easy configuration and ability to scale simultaneous connections. A Locust script was written to load the main page, the buy-tickets page, and the schedule page. Each page load was weighted in the script according to the frequency of hits, with the main page weighted as three, the buy-tickets page as two, and the schedule page as one. The weights represented how many times a particular page was visited compared to other pages, i.e. in this case the main page would be loaded three times for every single loading of the schedule page.

The minimum and maximum wait times were set to 1 and 6 seconds, respectively. The wait periods determined how long a simulated user would wait to move from one page to another page[44]. These values were selected to mirror the typical behavior a user exhibited on the FEST website. It must be noted that the average time a user stays on a page was not taken into consideration because viewing an already-loaded page on their computers or mobile devices do not consume more server resources, hence they can be considered to have "left" the website.

A separate server (with i3-3240 CPU, 16GB DDR3 memory and 1TB SSD) ran the load stress testing script to simulate user behaviors at different intensities.

Power Measurement

A Watts Up? .Net power meter was used to measure the power consumed by each server. A Python script written by Yongpil Yoon[45] was used to record the

power data in one second intervals as the system was running the load testing script.

Uptime

For each system, another local script was started before the test began, which recorded the uptime of the run. The purpose was to monitor the load averages of each system during the test and match those results with the power data profiled by the Watts UP Python script. To ensure the timestamp of the uptime script matched the timestamp of the Watts Up python script on each record, both systems were forced to synchronize their clocks using the Network Time Protocol (NTP[46]) before each experiment. Clock synchronization brings extra benefit of being able to easily collate all the data and analyze it at the per-second granularity.

Testing Methodology

After each server was set up and configured, I conducted the following identical steps to perform each experiment:

1. Plug in Watts Up? meter and boot the tested server.
2. Synchronize hardware clocks of each system using NTP.
3. Start uptime recording on the target server.
4. Run the Python script to log power usage to a CSV file.
5. Run the Locust script and record timestamp.
6. Match timestamps of the Python and Locust scripts and compare system load, power usage, and web requests at the per-second level.

The Locust script was executed with varying numbers of simultaneous users: 25, 50, and 100. It is worth noting that the simulated workloads in these

```
date +%H:%M:%S >> timestart; \\  
locust --no-web -c 50 -r 50 -n 1000 \\  
--host=http://192.168.1.109 --csv=ds718-CDN-50-1
```

Figure 2.1: A sample Locust script

experiments can represent real-world high-traffic websites. For example, 25 simultaneous users is roughly equivalent to 15 million hits/month, or similar traffic to geico.com[47], 50 users is equivalent to 30 million hits/month, comparable to cisco.com[47], and 100 users is equivalent to 53 million hits/month, about the same traffic as expedia.com[47]. I calculated the hits/month by considering requests per second as indicated in Figures 2.3 - 2.7. Specifically, 25 cached users is equivalent to 6.89 r/s for the DS718+. $6.89 \times 3600 \text{ seconds per hour} \times 24 \text{ hours} \times 30 \text{ days} = 17.8$ million total hits in a month. The calculations for 50 and 100 simultaneous users can be extrapolated similarly.

Each experiment was run five separate times, each considered independently and combined to normalize the final results, under the following three different scenarios:

1. Base installation of Wordpress
2. Caching enabled
3. Caching and Content Delivery Network (CDN) enabled

Figure 2.1 shows a sample run of the Locust script. The current timestamp is written to a file called timestart, which stores the time each test is initiated. The Locust script is then initiated using the console with 50 simultaneous users all starting at the same time for 1,000 total requests. The website base URL in this case is `http://192.168.1.109` (Locust will append other URLs automatically), and the results are logged into two CSV files (distribution and request). Figure 2.2 is a snapshot of the distribution file, which shows the percentage of requests successfully

Name	# requests	50%	66%	75%	80%	90%	95%	98%	99%	100%
GET /wordpress/	499	32	35	42	44	52	56	4900	5100	5141
GET /wordpress/buy-tickets	376	38	41	47	49	58	63	68	140	4912
GET /wordpress/schedule	172	42	45	48	53	60	65	130	5100	5198
None Total	1047	38	42	44	47	55	62	100	5000	5198

Figure 2.2: A sample distribution file

completed on the Synology DS718 server (with caching and CDN enabled) below a temporal threshold. For example, the eighth column of the table indicates that 95% of the requests to retrieve the main page, the buy-tickets page, and the schedule page are completed within 56, 63, and 65 milliseconds respectively.

Software Optimization

Caching

The base installation of WordPress must perform several tasks in order to display a properly rendered web page, all of which are dynamic in nature. Once the user requests a web page, the content on this page must be retrieved from a database. After retrieving the content from the database, PHP scripts apply styles (e.g. CSS) to the content and produce a valid HTML document that can be displayed on the end-user's web browser. Without caching, these operations are computationally expensive and must be performed each time a user views a web page. With caching enabled, the web server can create a static HTML page once, and send it to as many users that request it as possible. This greatly reduces the response time of the host, and provides an effective way to serve more end-users with less CPU load. Further, the compiled HTML and CSS files can be compressed during transmission thereby significantly decreasing load and response time.

Content Delivery Network (CDN)

A CDN is a geographically distributed network of proxy servers that store popular content (e.g. videos and images) that have the greatest demand from websites. Enabling CDN could reduce the workload by offloading some of the work from the FEST server to the CDN server. It can also mitigate slow disk performance of low-power servers by eliminating reads of large images from its local disk drives. Meanwhile, it is possible to increase latency of a website load if the CDN server itself is experiencing problems. As setup of a CDN for testing purposes is not straightforward, the built-in CDN service from the WordPress Jetpack plugin was utilized in my experiments.

Data Collation

As previously stated, each experiment had five separate runs, considered independently and combined to normalize the final results. I used the timestamps to collate and synchronize the recorded power usage data, the uptime data, and the Locust requests data, which were then analyzed synthetically. Since failures of any kind are undesirable for web requests, I recorded all failures (lengthy wait times before any content is displayed was also considered as a failure).

The average response time was plotted, along with the min/max wait time and the number of failures. Further, the distribution of the response time was plotted to illustrate the percentage of requests that could be considered successful. The instant power and system load was not considered, as in nearly every case once a test run had started, both values began to normalize. Plots were also made for the average power usage and system load over all five runs. The start and stop points were significantly smaller, albeit for only two to three seconds; this did not seem significant as each run took over several minutes to complete.

Results

Response Time

Per DoubleClick, 53% of all mobile website visits are abandoned if the load time is over three seconds[48]. Anything longer than two seconds is what a normal user calls "annoying". Therefore, I considered a response time of under 2,000 ms as "reasonable", and defined anything over 5,000 ms in the "infuriating" category. Obviously, this is a completely subjective measurement, dependent upon the user in question. Nonetheless, for fair comparisons, I created the following latency-based Quality of Service (QoS) evaluation scale metric to explicitly evaluate the performance of each server:

1. Superb <50 ms
2. Great <500 ms
3. Good <1,000 ms
4. Reasonable <2,000 ms
5. Mediocre <3,500 ms
6. Infuriating <5,000 ms
7. Abysmal <8,000 ms
8. Absurd >15,000 ms

Figures 2.3 - 2.7 plot the average response time of all five evaluated servers, which are discussed in detail below. For all figures, CDN refers to enabling **both** CDN and caching.

Base Case

The base case refers to the non-cached/non-CDN case. In this case, the two Synology devices were capable of maintaining a *Reasonable* level of service with 25 simultaneous users, both responding in roughly 1,000 ms, with minimal failures. The Pi was closer to 2,400 ms for 25 users, which is on the high side of *Reasonable*, but still quite usable. The maximum response time for all three servers was *Abysmal*, but these appeared to be outliers and are better thought of as failures. The Mac Mini and PowerEdge were both *Great* performers with 25 users, and max response times still being *Reasonable*.

Moving to 50 users resulted in the DS718+ reaching an *Infuriating* response time, with the Pi and DS1815+ being nearly *Abysmal*. At 100 users, the Pi did not successfully respond to any requests and the test was manually terminated. The two Synology servers approached *Absurd* levels of performance. The PowerEdge was *Great* for both 50 and 100 users, while the Mac Mini demonstrated *Great* performance for 50 users and *Good* for 100 users.

Cache Enabled

Enabling HTML caching dramatically improved the response times for all systems. For 25 users, nearly all systems were in the *Superb* category, with even the Pi providing *Great* performance with as many as 100 simultaneous users. The DS718+ and DS1815+ both experience occasional failures at 100 users, albeit still negligible overall. The DS718+ provided *Great* results with 50 users and *Good* results at 100 users. The DS1815+ performed *Great* with 50 users, but began to drift into the *Mediocre* range at the level of 100 users, which was unexpected when compared with the Pi. Both the PowerEdge and Mac Mini were *Superb* for all numbers of users.

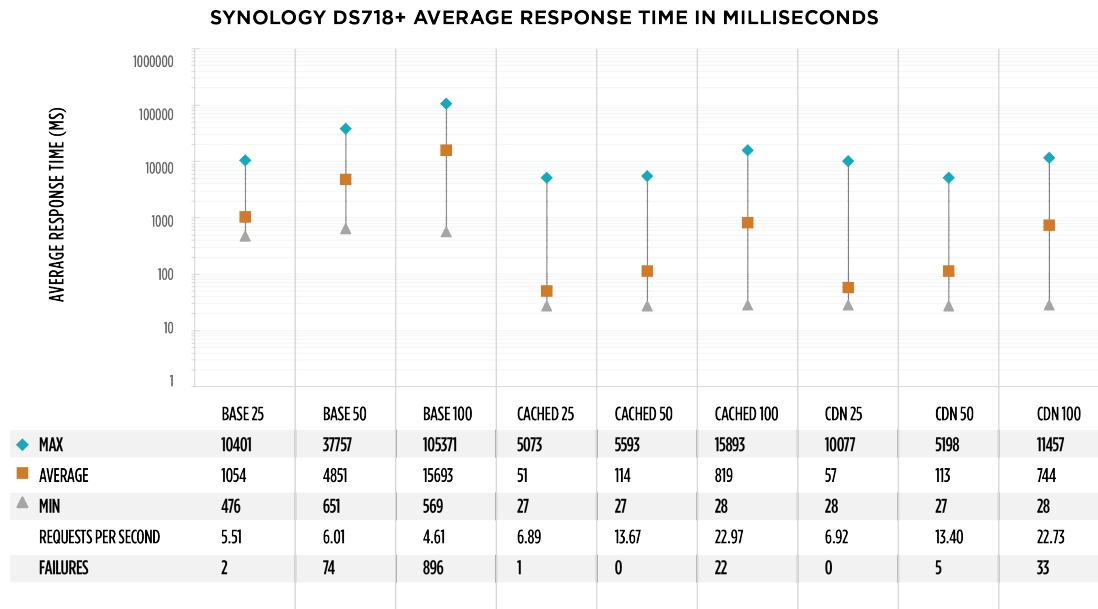


Figure 2.3: DS718+ Response Time

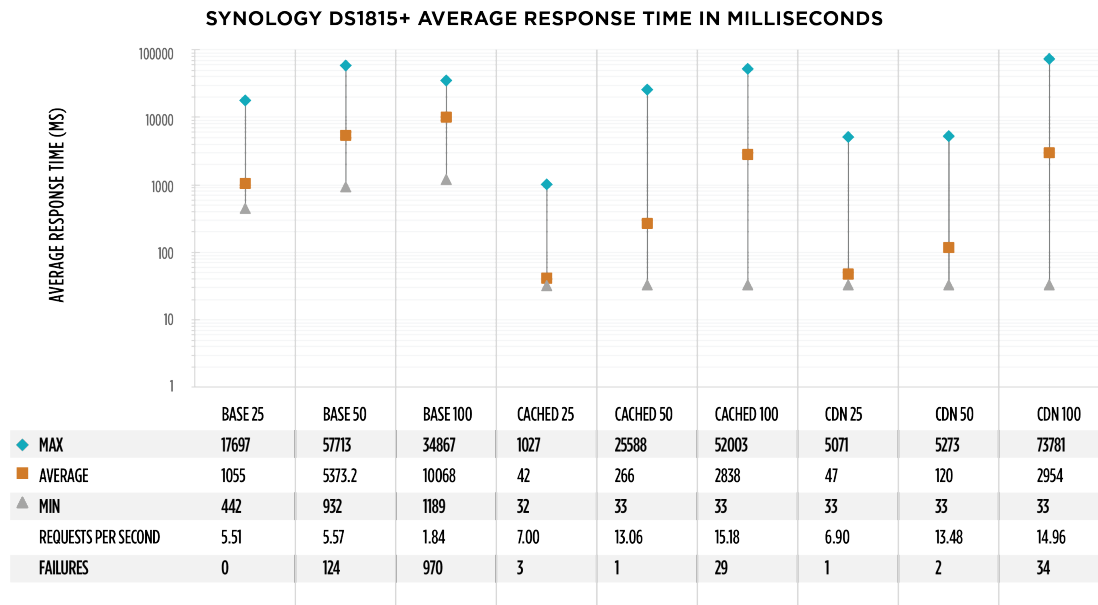


Figure 2.4: DS1815+ Response Time

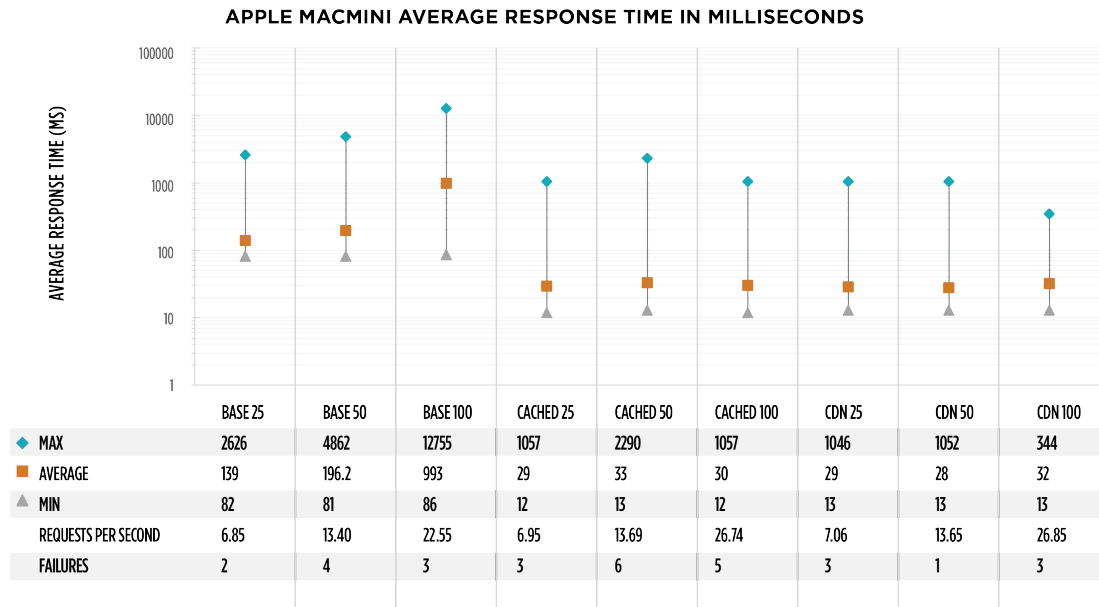


Figure 2.5: Mac Mini Response Time

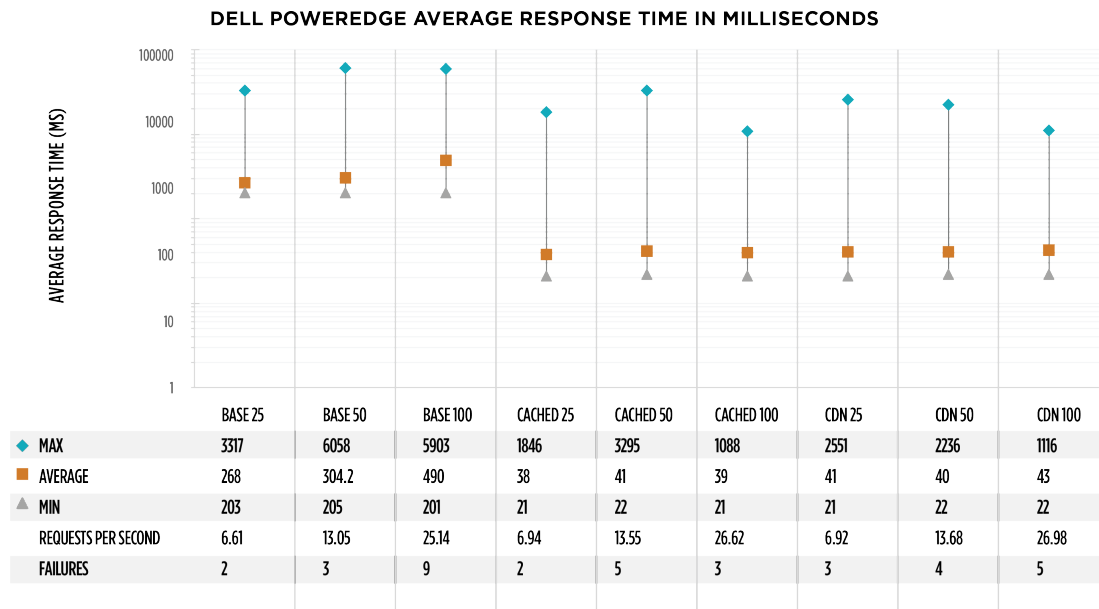


Figure 2.6: PowerEdge Response Time

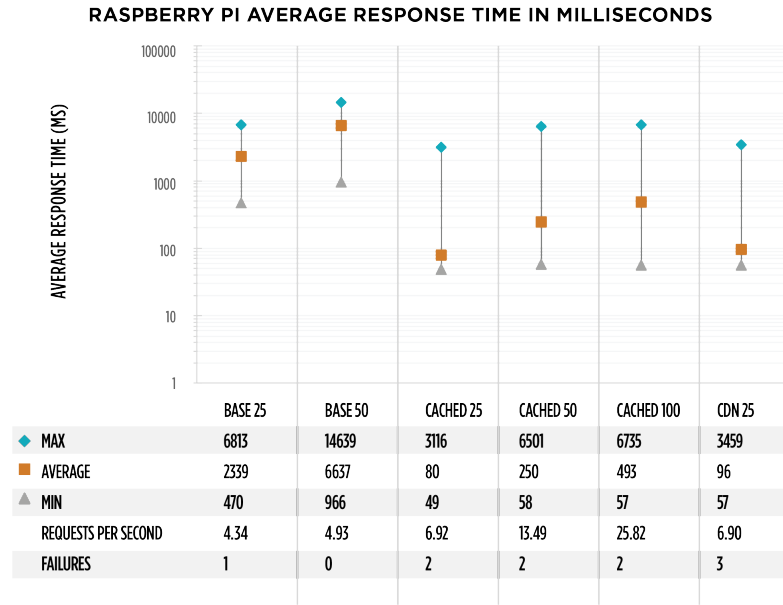


Figure 2.7: Pi Response Time

Cache and CDN Enabled

There was no significant difference in the response times between the cached and cached+CDN scenarios, and in a number of cases the response time increased slightly on average when CDN is enabled. This result is actually surprising as it was expected that the decreased I/O access would help with response time.

Linux System Load Average

Figures 2.8 - 2.12 plot the average Linux system load of all five evaluated servers. In general, the load averages are directly correlated with the response times, as would be expected. The red line for each figure indicates the total number of threads each system can run, and any load over this threshold means processes must be queued before running. The greater the load over this threshold, the longer each process will take to complete. The load is calculated by averaging the complete set of 1 minute uptime readings taken per-second during each experiment.

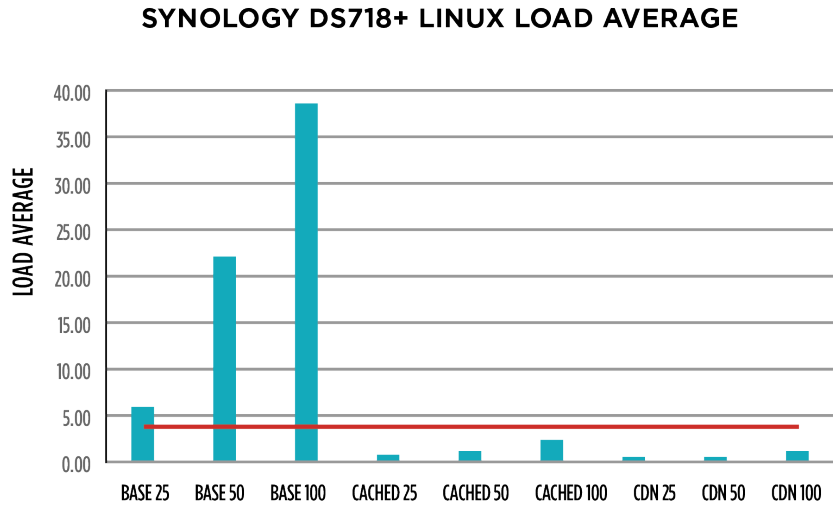


Figure 2.8: DS718+ Load Average

Base Case

In the base case with 100 simultaneous users, the load was so high for the two Synology devices as to be a completely unfeasible solution.

Cache Enabled

With caching, all systems were able to perform below their reasonable maximums.

Cache and CDN Enabled

In all cases, enabling the CDN reduced the load averages, which possibly benefited from the decreased I/O requests that has been discussed previously.

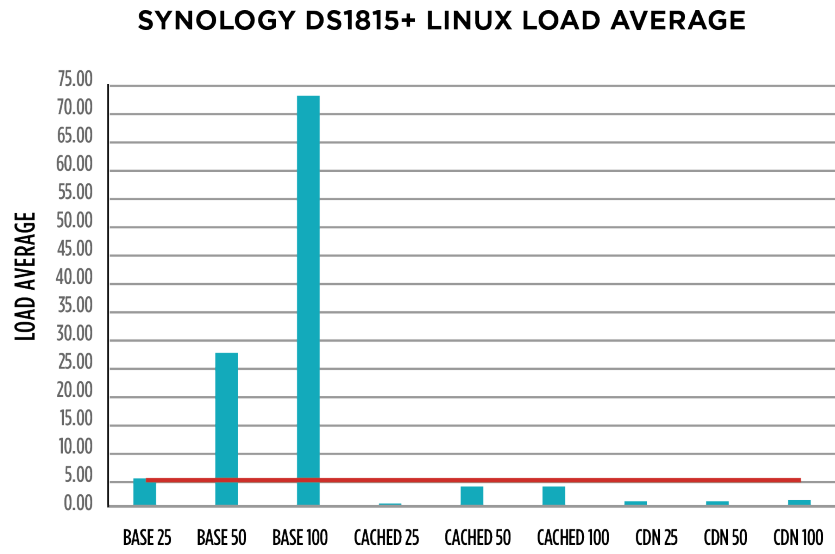


Figure 2.9: DS1815+ Load Average

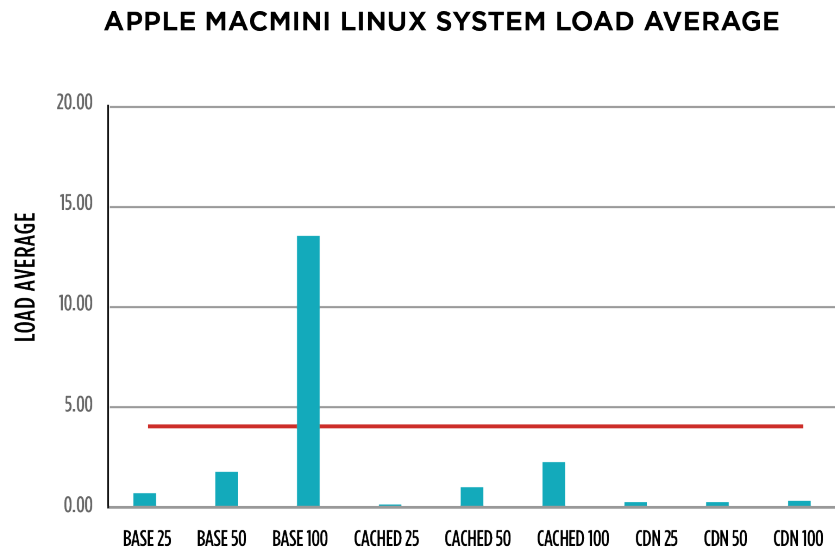


Figure 2.10: Mac Mini Load Average

DELL POWEREDGE LINUX SYSTEM LOAD AVERAGE

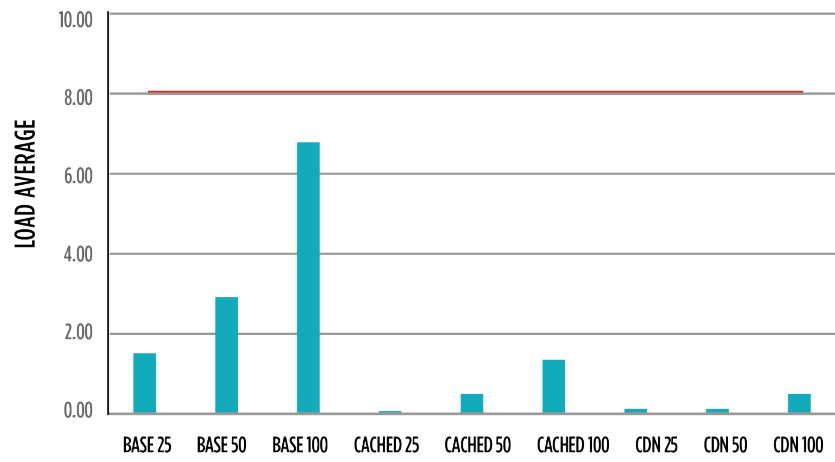


Figure 2.11: PowerEdge Load Average

RASPBERRY PI LINUX SYSTEM LOAD AVERAGE

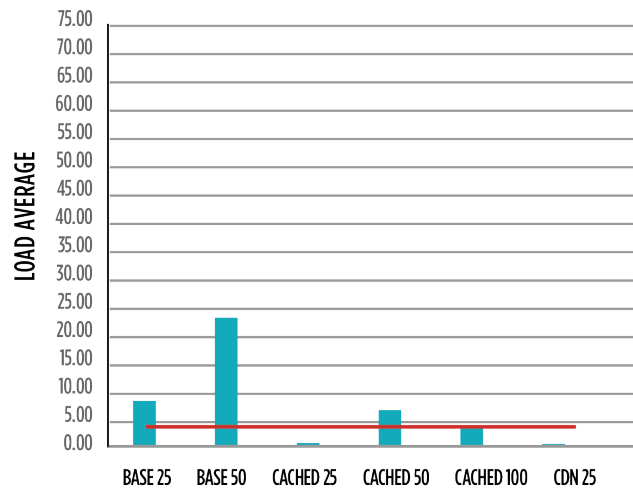


Figure 2.12: Pi Load Average

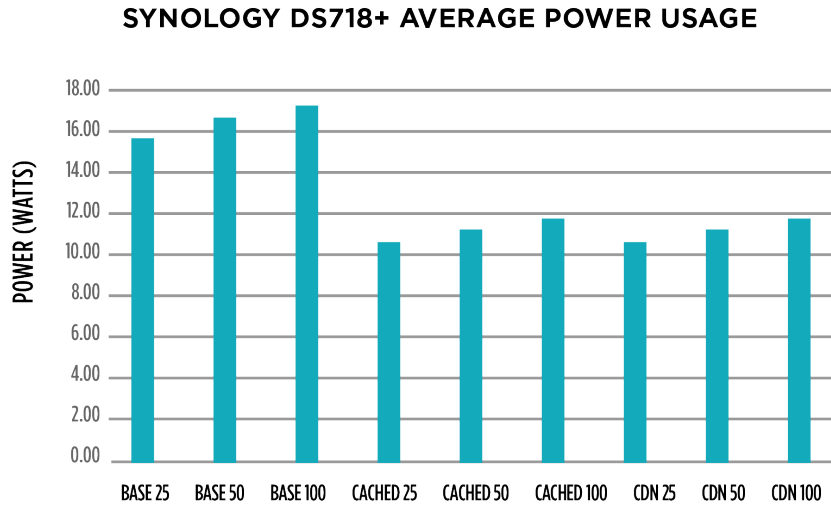


Figure 2.13: DS718+ Power Usage

Power

Figures 2.13 - 2.17 demonstrate the power results of all five systems. As expected, the Dell PowerEdge was the most power-hungry server and the Raspberry Pi used the least amount of power.

In every system, for any number of users, the power usage of the base case was generally much higher than the cached/CDN enabled case. For example, the power usage of PowerEdge, DS718+, and Mac Mini is reduced by approximately 20%, 35%, and 65% respectively for 100 users when caching and CDN were both enabled. The DS1815+ saved roughly 15% of power when comparing the base 50 and cached 50 configurations. With 25 simultaneous users and caching enabled, the Pi did not even record any power usage. The power efficiency of the Mac Mini was impressive, achieving *Great* performance and only consumed less than 10W of power (with caching and CDN enabled). The two Synology devices were expected to perform similarly, but the DS1815+ used dramatically more power than the DS718+. One

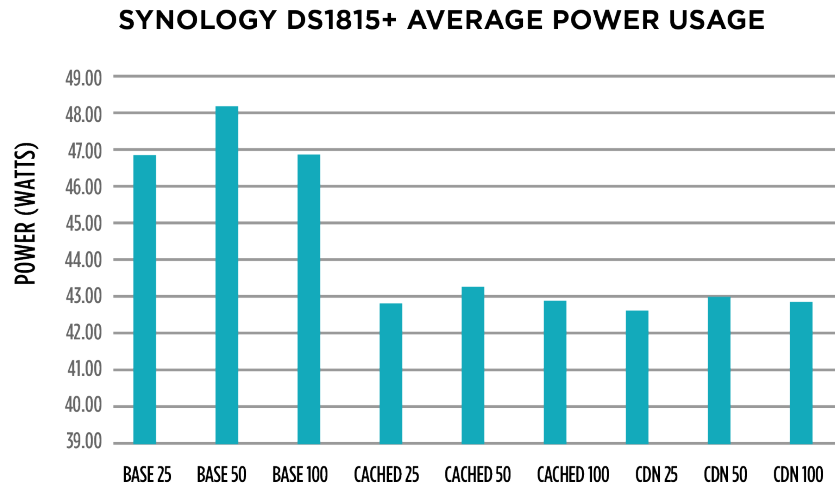


Figure 2.14: DS1815+ Power Usage

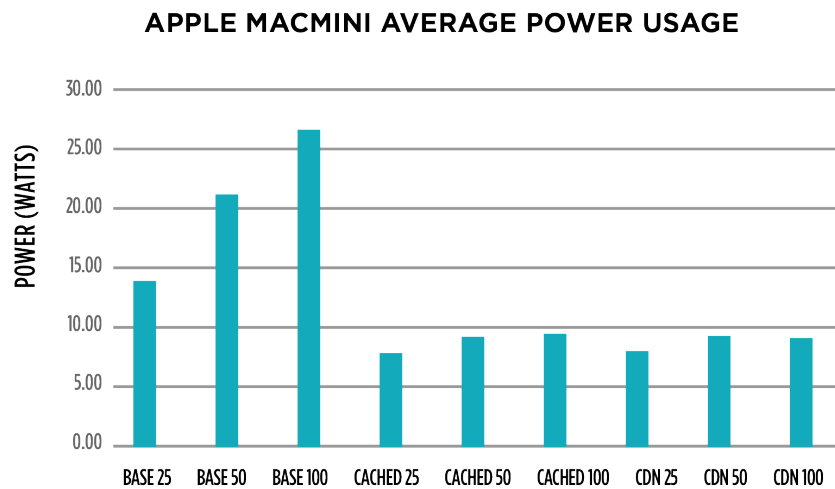


Figure 2.15: Mac Mini Power Usage

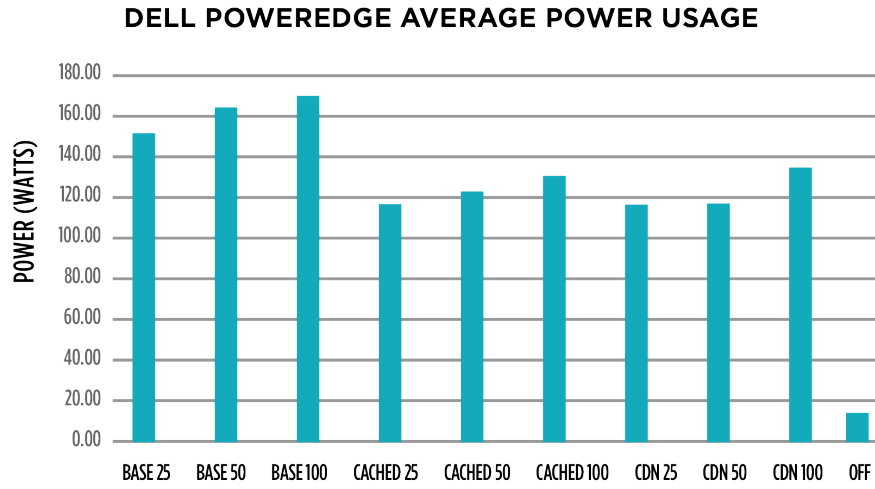


Figure 2.16: PowerEdge Power Usage

possible reason is that the DS718+ used SSD drives, which are more energy efficient than the HDDs used in the DS1815+.

A surprising finding is that the PowerEdge server still consumed noticeable power (13.4W) when being shut down but remaining plugged into the power source. This was verified multiple times to eliminate the possibility of equipment malfunction as I initially thought the Watts Up meter was experiencing problems. In fact, 13.4 W was even more than the active power of DS718+, Pi, or Mac Mini while they were running the cached and CDN tests.

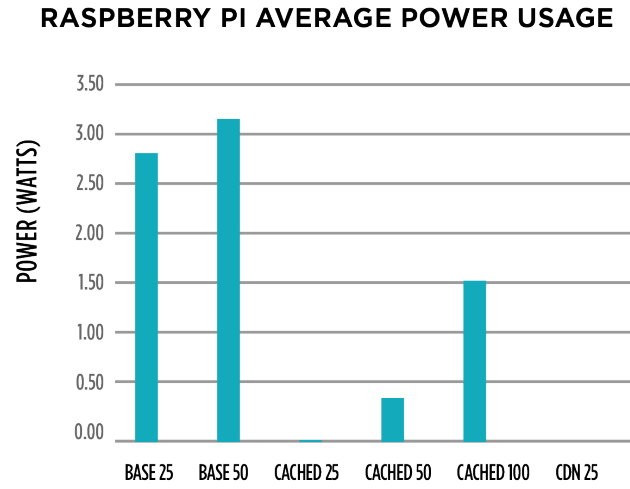


Figure 2.17: Pi Power Usage

Summary

What follows are a few observations that seem obvious from a cursory look at the data, and a more in-depth analysis for each server afterward:

1. Response time, system load average, and power usage are directly related.
When the number of requests for a web page increases, the system must work harder to maintain a reasonable level of service. This in turn taxes the CPU and I/O, which results in more energy consumption.
2. The Pi is not truly a viable platform for dynamic, high-traffic web sites, but with certain software optimizations such as caching and CDN, it could definitely be used to deploy WordPress instances that have medium workloads and serve a reasonable number of users.
3. As expected, the Dell PowerEdge achieved the best performance but consumed vastly more energy than other low power servers.

4. Both Synology devices performed well for smaller loads. The similarities in their CPUs, yet divergent results, indicate that the utilization of SSDs can significantly improve the performance and energy efficiency of web servers.
5. The superior laptop-based hardware and architecture design make the Mac Mini a high-performer (on-par with PowerEdge), but also a much more energy efficient option for web servers.

Raspberry Pi

The Pi performed adequately for medium workloads when caching was employed. It was able to handle 6.92 r/s (requests per second), which is equivalent to nearly 600,000 requests per 24-hour period, well over the 85,000 total user requests required by the FEST website on its busiest day. The Pi's extremely low power makes it a reasonable option for a web server when power saving is a top priority. A specific instance might be a website for certain scientific applications, perhaps in research areas with a strict power limit. The major limiting factor of the Pi is the lack of reliability and redundancy of the flash media. This could possibly be handled with frequent backups but, as a permanent web server, a more robust solution would be desired.

Synology DS718+

The newer DS718+ with SSDs was an excellent performer, being able to handle several million hits per-day when caching was utilized (22.97 r/s for 100 users). With caching disabled, the DS718+ running base 25 case was able to easily manage 5.51 r/s (or roughly 475,000 hits per day). The RAID1 redundancy, dual NICs, and 24/7 file server design demonstrates the device is intended to be used in a capacity where reliability is desired. This device is also capable of running a multitude of other services using the DSM software, such as Git[49], a Google G-Suite clone

service[50], and basic file sharing functionality. Given the flexibility of the VM management, the lower price (about \$400) and power usage, as well as the ability to deploy Docker[51] instances with minimal effort, DS718+ is ideal as a viable replacement for conventional web servers.

Synology DS1815+

The DS1815+ was similar to the DS718+, but the 2X-4X power usage and conventional HDDs made it less effective than DS718+ overall. However, the SHR RAID (roughly equivalent to a dynamically expandable RAID5) provides better data redundancy over other tested devices (although the PowerEdge could be configured to RAID5 if desired). For a website with reasonable traffic (i.e. less than or equivalent to the base 25 users case), the DS1815+ may be an alternative solution.

Mac Mini

The Mac Mini was a great performer, especially when the low power requirements are considered. When running the 100 users case, the Mini was able to maintain a *Good* response time with only a few failures while staying under 30 W. With caching and the CDN enabled, 100 users could be handled more quickly than the PowerEdge utilizing merely 1/12th of the energy. The power-to-performance ratio was the best of all the systems tested. However, the lack of several key features does not lend the Mini to be a true web server replacement. Much like the Pi, the fusion drive is not reliable enough and frequent backups are required. The laptop nature of the Mini also means it has multiple features that could cause issues with uptime, specifically the video subsystem, sound, and other multimedia functionality. Enabling the Mini with a minimum of RAID1 SSDs would make it a web server with both high performance and low power, capable of easily replacing the

PowerEdge. The high cost of deploying the Mini in such a capacity may not mitigate the power savings over a traditional web server, however.

Dell PowerEdge

The PowerEdge was fast and used a lot of power. Such servers are ubiquitous in datacenters specifically because they are capable and dependable. The built-in redundancy such as dual NICs and power supplies means these sorts of systems will have little to no downtime. Unfortunately, as stated, they use much more power to do the same amount of work.

Extended Experiments

There were several limitations imposed on the initial experimentation due to a variety of factors, but which were discovered after the findings were analyzed.

Specifically:

1. Overall network configuration - being on the same LAN eliminated some real-world traffic issues that could arise when testing QoS
2. CDN Calls - using the CDN resulted in essentially the same performance as the non-CDN setup which is most likely a result of the networking configuration
3. Memory Limitations - running Apache within a VM arbitrarily limited the available memory versus a bare metal installation
4. Single Website - serving only one website might be less stressful than multiple sites
5. Test Length - the time to run the tests was limited to several minutes

6. Number of Requests - a small number of page requests were extrapolated to a monthly total
7. Excluded Pages - only three pages of the FEST site were requested during the tests

Taking all these issues into consideration, a new methodology was created in an attempt to address each concern.

Hardware Configuration

I focused on a single server as described earlier in Table 2.1, the Synology DS718+. The low power usage (peak of <18W Figure 2.13) and relatively high performance per-watt were determining factors to use this system in the extended testing.

Software Configuration

Unlike the setup described in the initial testing, I did not use a VM to host Ubuntu and instead installed FEST directly onto the DSM operating system using the provided Web Station application. Web Station provides multiple versions of Apache, PHP, and MariaDB so the following selections were made: Apache 2.4.29, PHP 7.2.13, MariaDB 5.5.59. These were the most current versions available within Web Station at the time the study was conducted, and are optimized for the DS718+, unlike the generic Ubuntu configuration previously utilized. DSM version was 6.2.1-23824 which self-reports as Linux kernel 4.4.59+.

A virtual host was configured to respond to an externally selected hostname, in this case antman.truculent.org. Wordpress 4.9.5 was installed into the base directory of the virtual host within Disk Station and the FEST website was imported via the migration plugin. Tests were then performed to ensure the content loaded properly

and all links worked as expected.

A second installation of Wordpress 4.9.5 was installed into a subdirectory called /wordpress to allow an additional FEST instance to be run simultaneously. As before, Wordpress 4.9.5 was installed into this subdirectory and the FEST website was imported and tested.

Unlike in the previous experiments, the entirety of system memory was made available to Web Station (10GB), although approximately 200MB were utilized as overhead for DSM. This is within the margin of error of observable memory usage of Ubuntu memory usage when deploying FEST on bare metal installs or within the VMs.

Network Configuration

In the previous experiments, requests were made of the test server by a system attached to the same 1 gigabit managed switch, thus both systems resided on the same LAN. Observed latency between systems was <5ms as determined by the ping command. For the extended experiment the following setup was created:

Server

- DS718+ server attached to 1 gigabit managed switch
- Switch attached to dedicated 50MB upstream/50MB downstream Spectrum fiber circuit
- Physical location is New Braunfels, TX, routed through Dallas, TX

Client

- MacBook Pro 2017 laptop attached to 1 gigabit managed switch
- Switch attached to 5MB upstream/45MB downstream consumer grade AT&T Uverse DSL circuit

Table 2.3: Locust Page Weighting

Page	Weight
Main page /	31
buy-tickets	23
schedule	16
faq	2
videos	2
photos	2
history	1
news	1
contact-us	1
map	5
directions	3
accommodations	1
parking	5
buses	2
taxi-ride-shares	1

- Physical location is New Braunfels, TX, routed through San Antonio, TX

The physical location of the server and client systems were approximately four miles apart, with a traceroute of over 30 total hops, an estimated ~ 300 miles of network distance, and an average ping of <100 ms.

User Behavior Analysis

Using the historical user behavior per Google Analytics, traffic for all pages was weighted for a new Locust script to more accurately simulate user views. The Locust script in this case considered each separate URL as a percentage of the total views, resulting in the request weights as displayed in Table 2.3. This is a total of 96% of all site views, with the remaining 4% distributed between >50 separate Wordpress posts. The posts were not considered in this test as they were a relatively small portion of the overall traffic compared to the page views.

Another Locust script was created which duplicated the above URLs by prepending /wordpress, thus allowing two separate Wordpress instances to be tested simultaneously. This Locust script was the basis for the second set of tests run on the web server. In this case, the same weights were placed upon /buy-tickets and /wordpress/buy-tickets (23), and similarly for every other URL.

The minimum and maximum wait times were set to 1 and 10 seconds, respectively. This increased wait time more accurately reflected user behavior when navigating the entirety of the FEST website when compared to the maximum of 6 seconds wait time when only viewing the three most visited URLs.

Power Measurement

A sustained peak observed $\sim 20\text{W}$ power usage per the previous experiment was assumed throughout the new tests. This ensured an overestimated worst-case scenario for power consumption and thus should more accurately model real-world use.

Uptime

System load was not independently monitored as it was determined that response time and QoS are directly related to the load average. An unacceptable QoS indicates a high system load, with an overloaded system completely unable to maintain at least a Reasonable QoS.

Testing Methodology

After the server was set up and configured, the following steps were conducted for each experimental scenario:

1. Run Locust script on client system

2. Review results of web-requests at the per-second level

The methodology is somewhat more simplistic than before as the Locust script was written to accommodate a much longer testing period. The script was run using varying numbers of users as before, somewhat expanded: 25, 50, 100, 200, 250, 300. Instead of specifying a static number of total requests that take a variable number of minutes to complete, a test time of 8 total hours was specified. This ensured a more accurate representation of requests-per-second for the server and can easily be more readily extrapolated into monthly page request totals. For example, in the case of 200 cached users on the DS718+ running for 8 hours, 727514 total requests were observed with a QoS of Good. This is equivalent to 25.26 r/s, and 65.4mil total requests in a month.

There were six separate scenarios considered:

1. Base installation of Wordpress
2. Caching enabled
3. Caching and CDN enabled
4. Two separate Wordpress installations
5. Two installations with caching
6. Two installations with caching and CDN

It was already determined previously that the Base installation with 100 users was the approximate limit of usability, and thus 25, 50, and 100 simultaneous users were used for this configuration. 100, 200, 250, and 300 users were considered for the Cached and Cached/CDN scenarios, as <100 users were well below the minimal QoS threshold. During testing, several scenarios experienced unacceptable failure rates almost immediately and were thus terminated prematurely.

Results

Refer to III.A for an explanation of response time QoS evaluation terms. Figure 2.18 tabulates the test results and color codes the results according to the following criteria:

1. Failure rate must be acceptable, i.e. $<1\%$. Locust determines any timed-out request, or one over the 10ms threshold set within the locust.py file, to be a failure. It is thus possible a failure indicates a page that never loaded or simply took far too long to load. Results over this threshold are coded Red to indicate unacceptable.
2. Average response time in milliseconds must be at least Reasonable QoS, or better, and are coded Yellow if they also meet the Failure Rate criteria. Good or better QoS are coded Green if they also meet the Failure Rate criteria.
3. Results of Did Not Finish (DNF) due to extreme initial failure rates are coded as Red.

Observations

1. Base Case: 100 simultaneous users appeared to be the "sweet spot" for the base case, achieving Good results with an extrapolated 26.1mil r/month. 25 and 50 users were able to maintain a proportionally smaller number of users, but well within Great QoS.
2. Cached: Enabling caching more than doubled the number of simultaneous users while still maintaining Good QoS, and Great QoS is possible at 100 users with a total of 43.6mil r/month. At 250 and 300 users, the Failure Rate was excessive, even while maintaining Good response times, and also resulted

Synology DS718+ Extended Results

	Requests	Failure Rate	Avg Response ms	Requests/Second	Requests/Day	Requests/30 Days
Base 25	107645	0.84%	401	3.74	322,935	9,688,050
Base50	184413	0.66%	407	6.4	553,239	16,597,170
Base 100	290110	0.60%	668	10.07	870,330	26,109,900
Cached 100	484859	0.61%	221	16.84	1,454,577	43,637,310
Cached 200	727514	0.63%	731	25.26	2,182,542	65,476,260
Cached 250	602057	11.92%	888	20.9	1,806,171	54,185,130
Cached 300	478757	66.08%	1164	16.62	1,436,271	43,088,130
CDN 100	489655	0.64%	242	17	1,468,965	44,068,950
CDN 200	755893	19.42%	987	26.25	2,267,679	68,030,370
CDN 250	DNF					
CDN 300	DNF					
Dual 50	240936	0.53%	435	8.37	722,808	21,684,240
Dual 100	401394	0.53%	1638	13.94	1,204,182	36,125,460
Dual 200	404148	0.51%	8862	14.03	1,212,444	36,373,320
Dual 250	DNF					
Dual 300	DNF					
Dual Cached 100	498828	0.54%	237	17.32	1,496,484	44,894,520
Dual Cached 200	315453	0.49%	6667	10.95	946,359	28,390,770
Dual Cached 250	820988	8.94%	2712	28.51	2,462,964	73,888,920
Dual Cached 300	DNF					
Dual CDN 100	495382	0.51%	254	17.2	1,486,146	44,584,380
Dual CDN 200	807508	0.51%	1574	28.04	2,422,524	72,675,720
Dual CDN 250	421122	8.99%	6639	14.61	1,263,366	37,900,980
Dual CDN 300	804350	40.27%	2928	29.93	2,413,050	72,391,500

Figure 2.18: Extended Experiments Results

in less requests actually processed. Per previous experiments, this was due to a heavily overloaded system.

3. CDN: Employing a CDN with Caching at 100 users allowed for an extremely low response time, and thus a Great QoS. However, response time and requests were within the margin of error of only using Caching. Further, increasing the number of users to 200 and above resulted in unacceptable failure rates, with 250 and 300 user scenarios failing to finish.
4. Dual Base: At 50 users, roughly 25% more requests were made vs. Base 50, but still maintaining Great results. 100 users further increased requests, but QoS fell to Good. At 200 users, the number of requests was the same, with a response time of Absurd; 250 and 300 users failed to finish.
5. Dual Cached: At 100 users, the results were nearly identical as the Cached scenario, but exhibited unacceptable results with more users. As the number of users increased, so too did the response time and failure rate. Oddly, the 250-user case was able to process many more requests with a Mediocre QoS, albeit with a substantial failure rate.
6. Dual CDN: Nearly the same results as Dual Cached at 100 users, but able to provide a Reasonable level of QoS at 200 users. At 250 and 300 users, failure rate and response time were unacceptable.

During the extended experiments, the following results stand out as interesting:

1. A CDN seems to help maintain an acceptable response time with a high number of requests when multiple Wordpress sites are served. This could be due to a server configuration issue related to Apache, open file limitations, or memory limits. It is unknown at this time why the CDN seemed to have such

a low impact on results until this scenario is employed, and thus deserves more investigation.

2. Without employing caching or a CDN, 10 r/s (26 million r/month) is the reasonable limit for this scenario. Enabling Caching and a CDN roughly double these numbers with a better response time. Maintaining Good QoS is still possible with Caching/CDN at 25 r/s (65.4mil r/month). This seems to indicate that Caching provides allows for a 2X increase in requests with no discernible decrease to QoS.
3. Multiple instances of Wordpress running on the same server do not have a negative impact on performance and might actually give improved overall results. As stated above, the cause for this is unknown.

Conclusions and Future Work

With the booming of the Internet in past decades, the number of websites and web servers has increased tremendously. Conventionally, high-load websites tend to be deployed on high-end servers for the sake of performance and reliability.

However, the nature of web applications and website traffic determines that most web servers cannot be fully utilized, which leads to high cost of ownership and huge energy waste. In this chapter, I conducted a quantitative study to verify that low-power servers could be a viable solution to replace high-cost servers for hosting high-load websites, provided that the key functionality (e.g. system reliability and virtual machine management) and QoS can be preserved.

Numerous research issues have not been addressed in this work, but are worth exploring in the future. For example, what are the key challenges of running websites on low-power servers? What if some websites have a large amount of web content that cannot fit in the relatively small cache of low-power servers? The

current scale of test is on a single web server. How can architecture issues be addressed when a cluster of low-power web servers are necessary to host a high-load website? How can inherent weaknesses of low-power servers (e.g. usability and reliability) be addressed? Can these low-power servers operate at full capacity 24/7 without failures?

III. PRIVATE CLOUD WORKLOAD - ALIBABA

Introduction

In the past decade, we have witnessed the rapid growth of e-commerce and cloud services, which created IT giants like Amazon and Alibaba. According to the Netcraft report [52], Alibaba has become the world's 2nd largest cloud computing company, right next to Amazon. The data centers of Alibaba serve millions of users and process billions of transactions on a daily basis. For example, in the 2017 Single's Day Shopping Festival (11/11), \$25 billion of products were sold and 1.5 billion payment transactions were processed [53] within 24 hours on the Alibaba system. This sales record was five times of the total 2017 Black Friday sales in U.S. [54]. In 2019, the Single's Day Shopping record had grown to \$38.4 billion, a 1.5X increase in two years [55]. The recent coronavirus pandemic has dramatically disrupted the retail industry, which brings a new boom to e-commerce. Leading companies like Amazon and Alibaba have seen flooded online orders. Walmart's online sales have surged by 74% during the pandemic [56]. Target's digital sales grew by more than 100% in March 2020 and are up 275% in April, 2020 [57]. It can be projected that more cloud systems will be built to meet the booming demand of e-commerce.

Large-scale cloud systems that can handle such an excessive amount of online traffic are essential to e-commerce, but their enormous appetite for investment and energy are astonishing as well. One of the Alibaba's newest data centers in Hebei requires ¥18 billion capital investment plus millions of dollars annual electricity bills [58][59]. This number could further grow as the number of servers, the consumed energy, the labor to manage the system, and the electricity price all increase. McKinsey reported that the cost of data centers accounts for approximately 25% of

the total corporate IT budget [60]. Inefficient data centers may threaten profitability, despite the large number of users and dominating market share [60]. One of the primary reasons for data center inefficiency is the low utilization. The Gartner and McKinsey report indicated that the server utilization rate is merely 6% - 12% [60] [61] for most enterprise data centers and another study showed that the utilization of Amazon AWS servers is not high either ($\sim 7\%$ - 17%) [62].

To address the inefficiency issue, virtualization and server consolidation have been widely used to boost utilization rate. However, co-running jobs, especially time-sensitive jobs, on the same server may interfere with each other and adversely affect performance. A purchase request failure (even delay) can be disastrous to user experiences and result in huge revenue loss. Alibaba has explored an innovative approach to alleviate this dilemma by scheduling time-sensitive online service jobs and time-insensitive batch jobs to the same machine [63] [64]. In 2017 and 2018, Alibaba released two cluster traces that used the mixed scheduling strategy. The 2017 trace consists of 12,932 batch jobs and 11,076 online service jobs running on 1,313 machines over a 12-hour period. The 2018 trace is at a larger scale with 4,201,015 batch jobs and 370,540 online service jobs running on 4,023 machines over an 8-day period [65].

For purposes of this dissertation, I conducted a comprehensive analysis on the 2017 and 2018 Alibaba cluster traces and offer the following contributions:

1. I used the Alibaba trace to quantitatively show that consolidating hybrid workloads can considerably improve the utilization of large-scale cloud systems and reduce TCO. My analysis on the 2018 trace shows that the servers that only run batch jobs (time-insensitive) or service jobs (time-sensitive) have an average CPU utilization of 29.29% and 7.4% respectively. By applying the mixed scheduling strategy, the average server utilization is improved to 39.26%. This leads to millions of dollars of TCO reduction purely from the

reduced number of servers without compromising quality of service (QoS).

The TCO benefit will be even larger if savings on the physical footprint, power supplies, maintenance personnel, and cooling facilities are included.

2. I discovered that the jobs submitted to the Alibaba cluster request an unreasonable amount of resources. A large portion of jobs (both batch and service jobs) aggressively over-subscribe resources. Meanwhile, a considerable number of jobs (both batch and service jobs) under-subscribe resources. This hurts the Alibaba system in both ways. The under-subscribed jobs will suffer from performance loss while the over-subscribed jobs will waste energy and lower system utilization.
3. I developed a simulator that allowed me to quantitatively study the impact of reduced cluster sizes on the performance, energy consumption, and TCO of the Alibaba system.
4. My simulation results indicate that 5% of servers can be safely removed from the cluster with negligible influence on performance, which is able to save about \$600,000 (including server purchase cost and one year of electricity cost) for the system running the 2018 trace. Nearly \$68 million TCO reduction can be achieved if similar strategy is extrapolated to Alibaba’s 432,000 Internet-facing servers [52].

Related Work

Alibaba Workload Analysis

Since Alibaba released its cluster trace in 2017, a number of papers have been published to analyze the Alibaba workloads. Lu et al. identified several imbalance scenarios (e.g. spatial imbalance, temporal imbalance, imbalanced CPU and memory utilization per workload, and imbalanced resource demands from different

tasks) in the 2017 Alibaba trace [66]. Liu et al. analyzed the Alibaba trace from the system elasticity and plasticity perspectives [67]. They reported that most containers exhibited fairly steady disk and memory usage but the CPU utilization varied. They also identified that batch instances were over-committed for both CPU and memory resources. Cheng et al. [68] and Jiang et al. [69] analyzed the workload characteristics of the 2017 Alibaba dataset with similar conclusions that container resource usage was over-provisioned and thus batch jobs must use whatever free resources were available when run. These conclusions were borne out by the analysis done by Lu et al. [70], which further demonstrated that system resources were not allocated in a balanced way for the heterogeneous workloads. Beyond workload analysis, Li and Hu took one step further and proposed a deep reinforcement learning based job scheduling (DeepJS) algorithm that used the Alibaba trace as a training set [71]. They claimed that DeepJS was able to make better decisions about scheduling than the default heuristic-based approaches, thus performing more efficiently. Tian et al. examined the dependency structures of Alibaba jobs and compared the task dependency of Alibaba jobs with other benchmarks such as TPC-DS and TPC-H [72]. Guo et al. studied the resource efficiency issue of the Alibaba system by analyzing the 2018 dataset [73]. They discovered that memory became the new bottleneck, batch jobs were treated as second-class citizens, and the Java Virtual Machine (JVM) caused the resource inefficiency of containers running time-sensitive service jobs. Unfortunately, despite various problems being identified, most of the existing analysis work can only provide suggestions but were not able to tackle the discovered problems with verified solutions. This is largely because the practical schedulers deployed on large systems are very complicated and the data provided in the trace files are static. To evaluate the effectiveness of possible solutions, it is essential to develop an accurate simulator. For example, Google developed the Omega scheduler simulator to emulate its cluster scheduling process

and explored the impact of different configurations and scheduling strategies on system efficiency and utilization [74].

Cloud System Scheduler Simulator

There are very few published works discussing how to design and develop simulators for large-scale and complex systems. The Mesos scheduler was first introduced by Hindman et al. for sharing commodity clusters between multiple diverse cluster computing frameworks such as Hadoop and MPI [75]. It has been widely adopted by the Hadoop community after it became one of the Apache’s open source projects [76]. The Mesos scheduler enabled fine-grained sharing of resources across frameworks and provided strong support for running distributed tests at scale. The other Hadoop scheduler is YARN [77], which has a master daemon that communicates with the clients and one or multiple worker daemons that launch and track processes spawned on a machine. A recent study even enhanced YARN for better heterogenous resource allocation [78]. The Omega scheduler [74] was specifically designed to address the low utilization and efficiency problem at Google’s production system by allocating and running a mix of workloads (CPU intensive, memory intensive, batch jobs and low-latency jobs) on the same cluster. To some extent, the Alibaba’s mixing schedulers (Fuxi [64] and Sigma) adopted the design philosophies of the Omega scheduler, which leveraged time-insensitive jobs for higher resource utilization. To the best of my knowledge, this simulator is the first attempt to emulate the Alibaba scheduling process and the analysis is the first one to evaluate the Alibaba system from both the performance and cost efficiency perspectives.

With Mesos, Omega and Fuxi representing the state-of-the-art efficient scheduling algorithms, how to further improve the intelligent scheduling and resource management algorithms for large-scale systems became more challenging.

One of the most promising methodologies is to profile the real-time resource usage, system utilization, and performance information for offline or online in-depth analysis. However, it is not trivial to diagnose interference due to the scale of the system and the volume of trace data. As a result, enterprises started to release trace data and seek collaborations with academia and community. For example, Google released a month-long trace running over 25 million tasks on 12,500 heterogeneous machines in 2011 [79]. Alibaba released a trace file running mixed workloads on over 1,000 machines in a 12-hour period. Since then, numerous analytics works have been published using these two datasets. Zhang et al. and Tumanov et al. analyzed the characteristics of requests for resources and system utilization on the Google datasets [80][81][82][83]. Abdul-Rahman et al. analyzed the Google trace from the perspective of user behaviors [84]. Sharma et al. studied the Google dataset from the angle of task placement constraints [85].

Background

Utilization Matters

Large scale data centers contain hundreds of thousands of machines, which run a variety of jobs. Most of these jobs cannot fully utilize the CPU and memory resources of the servers thereby causing low server utilization. In fact, this is the primary reason for data center inefficiency because given the same workload, lower utilization means a company has to purchase more servers. AMD reported that the growth in server power use worldwide comes mostly from having more servers, not from having higher power use per server [86]. For a large-scale system, a small increase in server utilization could lead to significant cost savings. Assuming that a data center has a total of 2,000 servers (each server costs \$3,000), among which half of them run batch jobs with an average utilization of 25% and the other half run service jobs with an average utilization of 15%. If the average utilization of servers

can be improved to 40%, the system size can decrease by 50% (i.e. remove 1,000 servers), which will reduce the TCO by at least \$3 million provided that the total workload remains unchanged. Therefore, improving server utilization is an effective way to reduce data center TCO.

To calculate TCO savings when improving workload, the following equation is utilized:

$$X = \frac{N * (R2 - R1)}{R2}$$

where N is the number of servers in a datacenter, R1 is the current utilization rate, and R2 is the improved utilization rate. X would be the number of servers "saved". Using the example above, calculating the number of servers no longer needed to performed thusly:

$$1000 = \frac{2000 * (40 - 20)}{40}$$

Performance Dominates

Server utilization can be improved by consolidating multiple jobs on the same server via virtual machine or container technology. However, this may lead to resource contention and cause unpredictable performance degradation. For the time-sensitive jobs (e.g. online purchasing), request failures or delays are not acceptable because they can result in huge revenue loss. For example, Google found that increasing the search results from 10 per page to 30 per page would increase the page load time from 0.4s to 0.9s [87]. This half a second performance degradation resulted in a 20% decrease in searches [88], which would cause millions

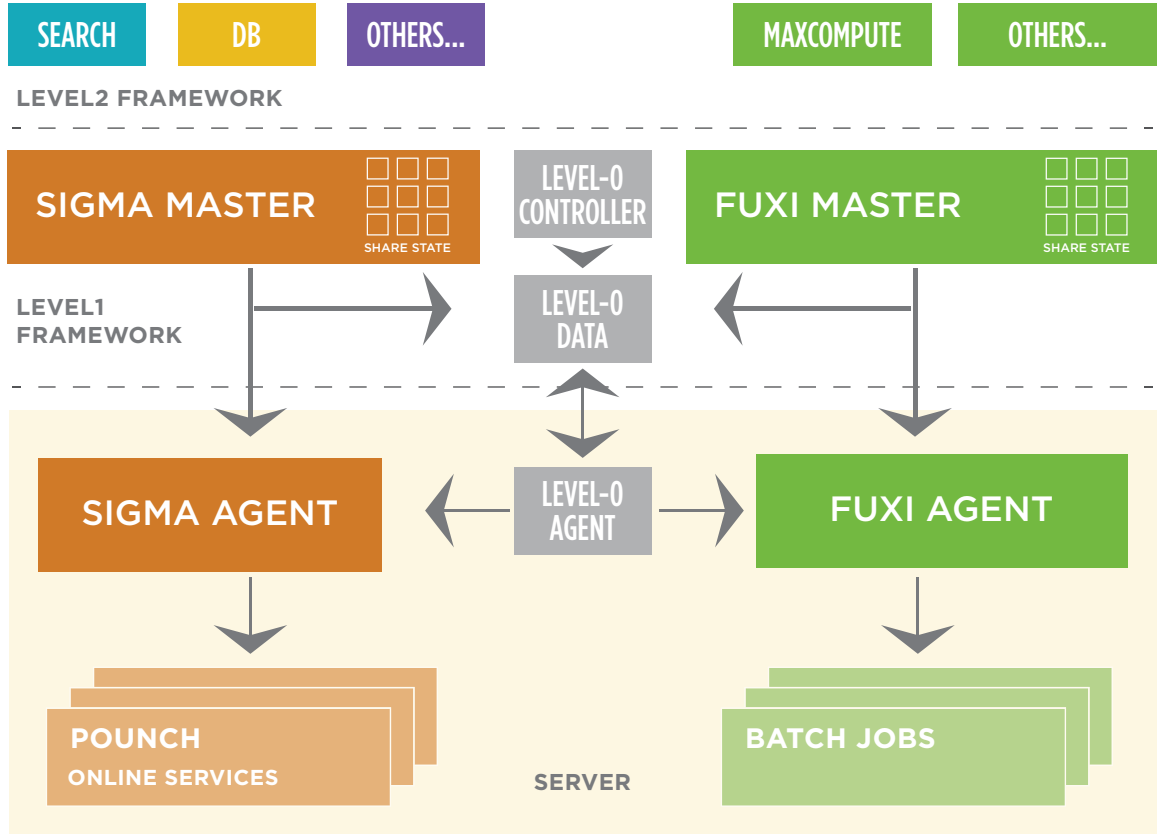


Figure 3.1: Alibaba Scheduler

of dollars of loss in advertisement. It is essential that improving server utilization should not be at the cost of sacrificing performance or QoS.

Alibaba Scheduler

As one of the biggest e-commerce providers, Alibaba runs hybrid workloads such as online service jobs (e.g. searching, online shopping, and advertising) and batch jobs (e.g. data mining and machine learning to understand customer behaviors and provide customized recommendations). Since online service jobs are essential for business, they are very time-sensitive. Batch jobs are not time-sensitive so they are given lower priority and will be terminated when resource contention with time-sensitive jobs occurs. To ensure high server utilization and quality of critical online services, Alibaba implemented a mix scheduling framework (see

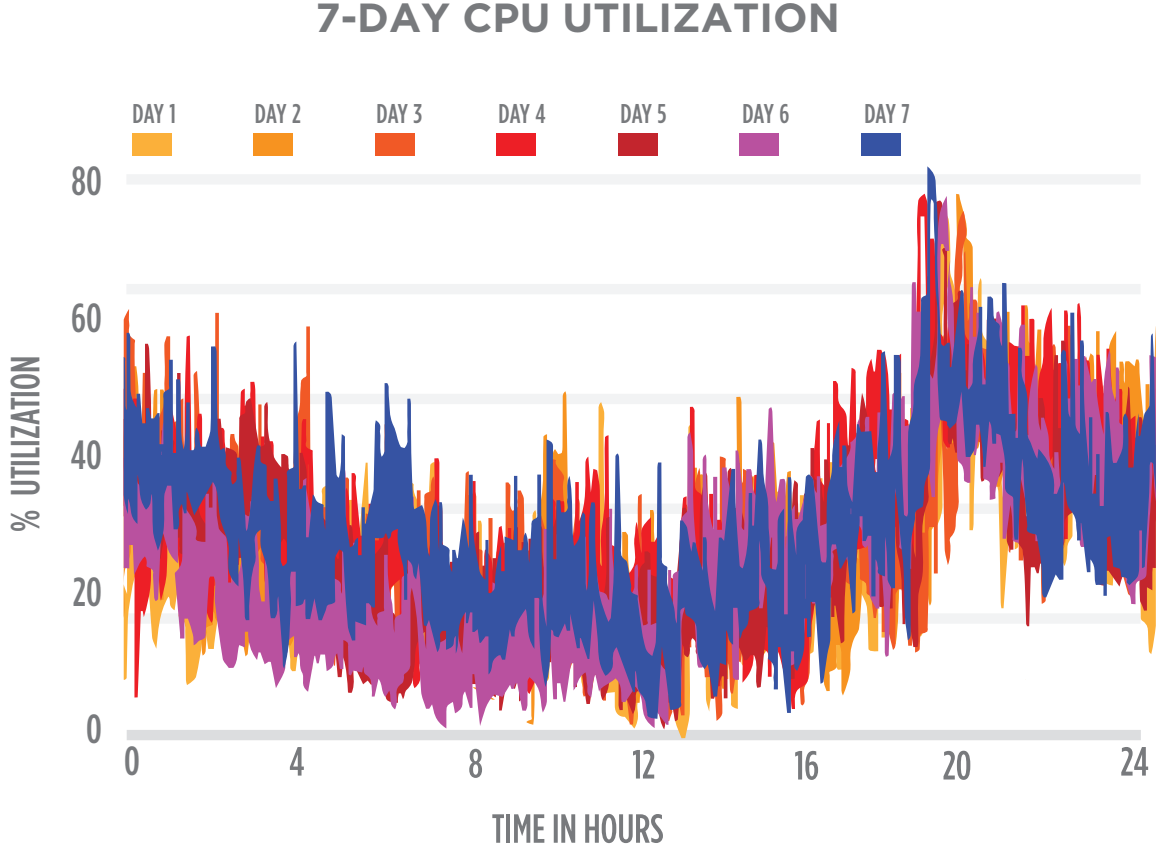


Figure 3.2: 7-Day CPU Utilization

Figure 3.1), in which the time-insensitive batch jobs are co-located with the time-sensitive service jobs. More specifically, there are two schedulers in the Alibaba system. The Sigma scheduler allocates online service jobs while the Fuxi scheduler is responsible for dispatching batch jobs [63]. They share the state of the entire cluster to obtain global view of system status and make better scheduling decisions.

Workload Analysis

Alibaba Dataset

The Alibaba cluster traces provide detailed information about the servers, batch jobs, and service jobs [65]. Specifically, the 2018 trace constitutes 6 files with more than 450 GB of uncompressed data logged over an 8 day period. Each trace

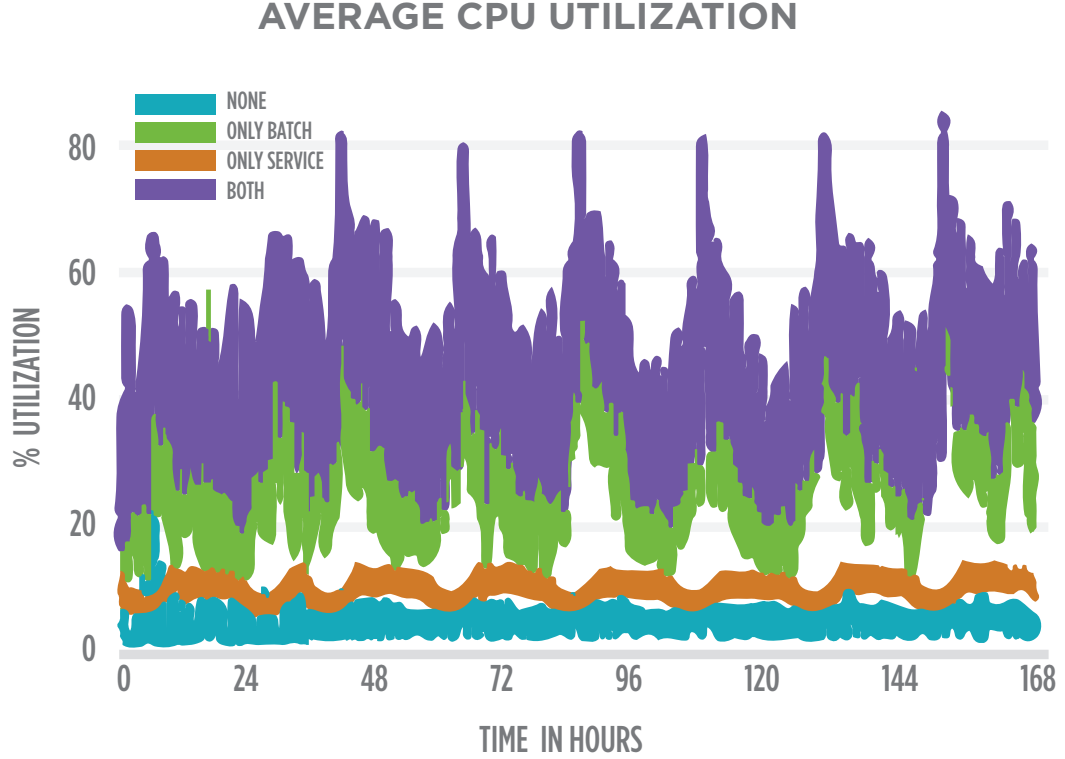


Figure 3.3: CPU Utilization by Hours

file is an SQL dump of its respective table, saved in the CSV format. Since the 8th day dataset had incomplete records (only several hours of data is available and some information is missing), I decided to extract datasets with 7 complete days (i.e. 168 hours) for use in this analysis. Figure 3.2 visualizes the overall CPU utilization for each day included in the 2018 trace during a 24 hour period, from which it can be seen that the valleys and peaks occur at specific times of the day and the daily usage patterns are largely repetitive.

The server information is given in the *machine_meta* table and the *machine_usage* table. The *machine_meta* table describes the events that occur on the physical systems such as capacity and errors. The *machine_usage* table contains the utilization of CPU, memory, and disk access timestamped in 10 second intervals.

Table 3.1: Summary of the 2017 and 2018 Alibaba Cluster Traces

	2017	2018
Number of Servers	1,313	4,023
Batch Jobs	12,932	4,201,014
Batch Tasks	80,386	410,134
Batch Instances	11,829,897	1,351,255,775
Service Jobs	11,076	370,540

The *batch_task* table and the *batch_instance* table provide information about batch jobs. The *batch_task* table shows the status and resource demands of the batch tasks. The *batch_instance* table includes the resource usage of each instance of the task and their status. Service jobs are described in the *container_meta* table and the *container_usage* table. Each batch job has one or numerous tasks and each task runs a single or multiple instances, which execute the same binaries but on different input. Each time an instance tries to run is called a trial. An instance can have more than one trial before its successful termination although this is rare (e.g. only less than 1% of instances experience more than one trial).

Table 3.1 summarizes the total number of servers, batch jobs, tasks, instances, and service jobs in the trace files of the 2017 and 2018 datasets.

Machine Utilization

There are a total of 4,023 machines in the 2018 Alibaba cluster trace, among which 145 machines only run service jobs and 11 machines only run batch jobs. The remaining machines run both service jobs and batch jobs. Figures 3.3-3.5 plot the CPU, memory, and disk utilization of the entire cluster over time in hours, color-coded to indicate machines running service jobs only, batch jobs only, and both jobs respectively. Machines included in *machine_usage* with no recorded jobs are also displayed for completeness. Table 3.2 summarizes the average CPU,

Table 3.2: Average CPU, Memory, and Disk utilization

2017			
	CPU	Memory	Disk
Batch Jobs	17.56%	9.54%	43.31%
Service Jobs	11.27%	34.3%	35.73%
Both	28.11%	52.18%	47.81%

2018			
	CPU	Memory	Disk
Batch Jobs	29.29%	79.55%	10.30%
Service Jobs	7.42%	80.46%	1.94%
Both	39.26%	88.56%	7.96%

memory, and disk utilization of the three different kinds of servers as compared to the 2017 trace files. It can be observed that the mixed scheduling strategy considerably improves both CPU and memory utilization. For example, in the 2018 Alibaba trace, the servers that only run batch jobs or service jobs have an average CPU utilization of 29.29% and 7.42% respectively. By applying the mixed scheduling strategy, the average server utilization is improved to 39.26%. Also noticeable is that machines running only service jobs rarely reach 15% CPU utilization while almost all machines running both batch jobs and service jobs can exceed 80% of CPU utilization multiple times. Similarly, the memory utilization of machines running only service jobs is roughly 80% while it is not uncommon for machines running both jobs to exceed 90% of memory utilization. Disk utilization seems to have little impact on performance with less than 8% of disk resources are used on average for mixed systems.

Batch Jobs

When a batch job arrives, the FuxiMaster starts an AppMaster (a.k.a. Application master or job master) for that job. Once started, the AppMaster reads

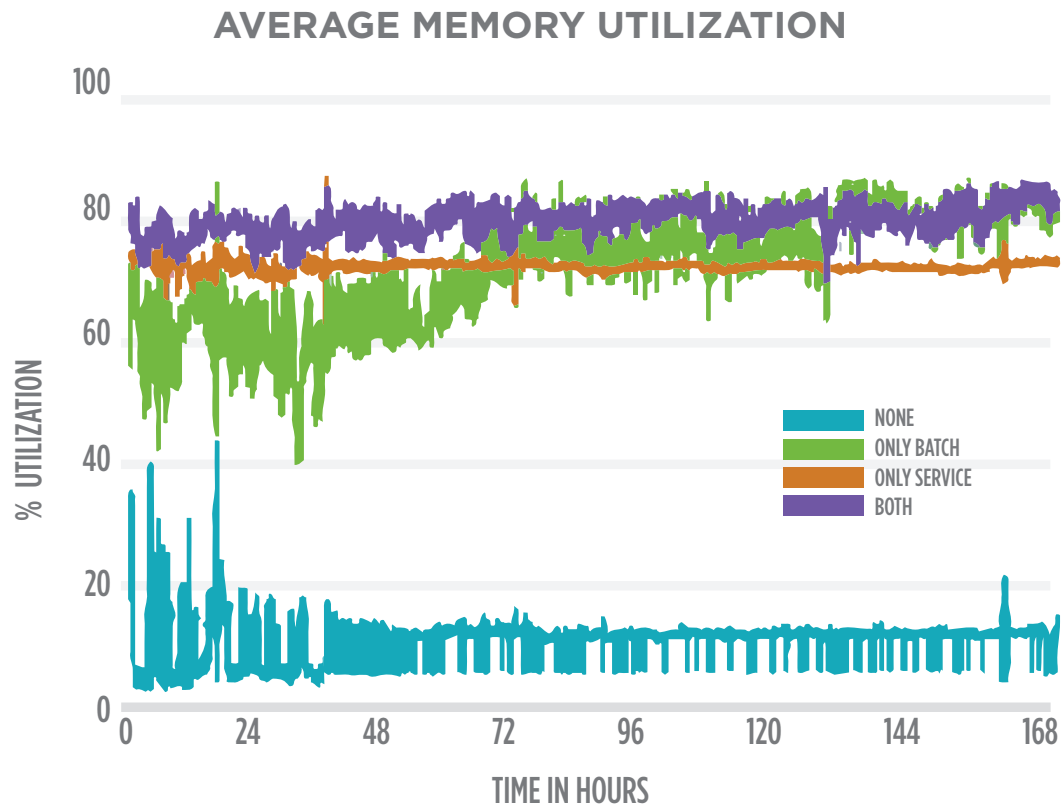


Figure 3.4: Memory Utilization by Hours

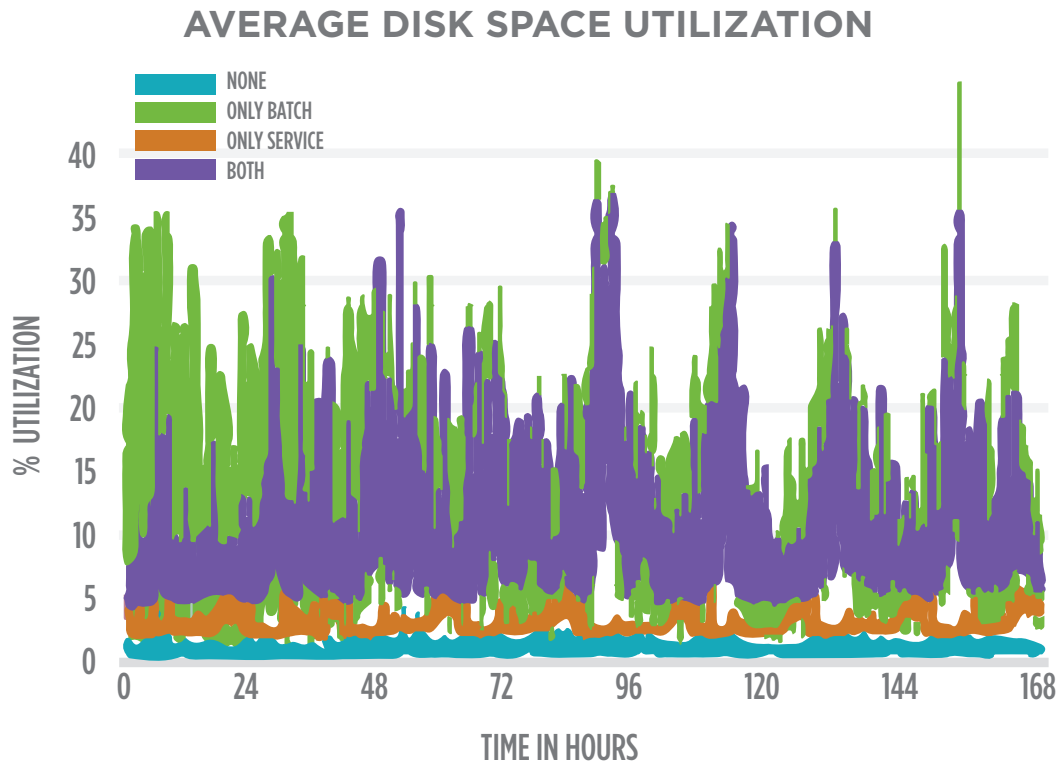


Figure 3.5: Disk Utilization by Hours

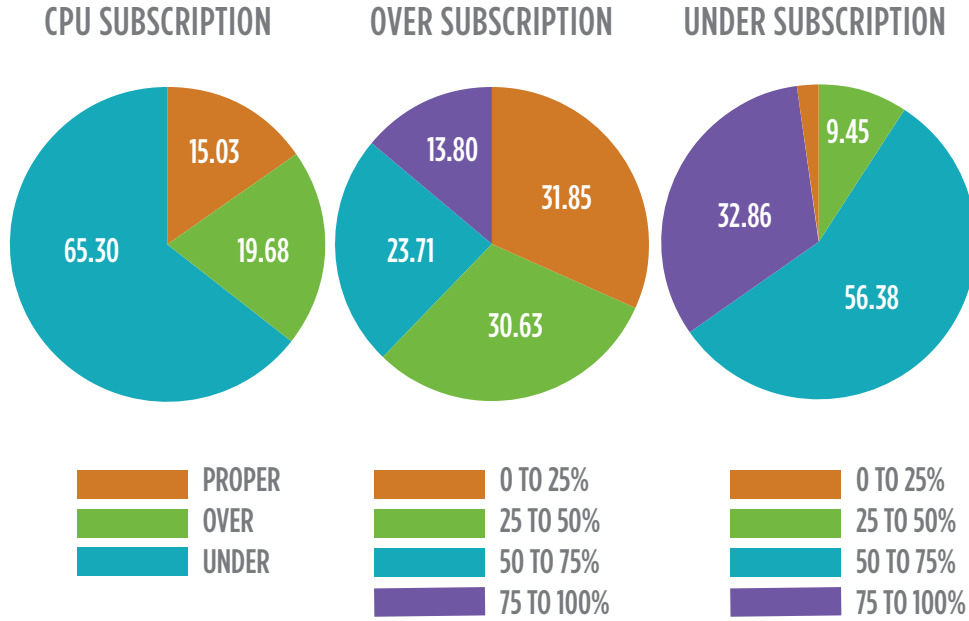


Figure 3.6: CPU Subscription of Batch Jobs

the job description file, interprets the resource requirement, analyzes task dependency, and sends resource requests to the FuxiMaster accordingly. The FuxiMaster allocates resources and sends them to the AppMaster. If there are not enough resources available, it sends the resources in the incremental fashion upon availability. The approved resources sent to the AppMaster is referred to as resource grants. Once the minimum required amount of resources is obtained, the job master starts the task. Resource subscription is the amount of resources that a task requests from the FuxiMaster and gets grants for it. Over-subscription refers to the case in which the requested resources are more than the utilized resources. Similarly, under-subscription refers to the case in which the requested resources are less than the utilized resources. Proper subscription means that the requested resources is roughly equivalent to the utilized resources.

Figure 3.6 shows that in the 2017 Alibaba trace, 65.30% of batch jobs

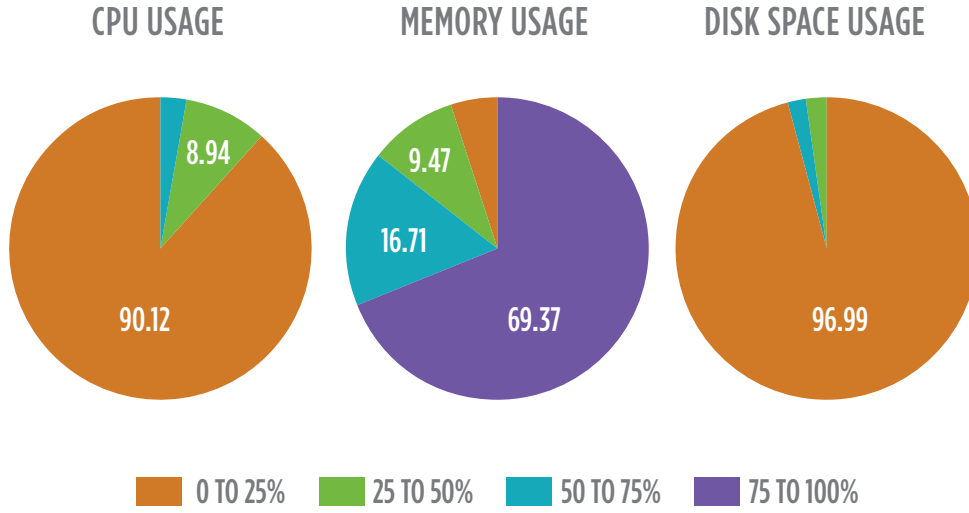


Figure 3.7: Resource Subscription of Service Jobs

over-subscribe CPU resources, among which 13.80%, 23.71%, 30.63% and 31.85% over-subscribe the CPU resources by 0-25%, 25%-50%, 50%-75%, and 75%-100% respectively. These over-subscribed resources will be held by the tasks (without being used) until they terminate, which leads to low utilization and degrades system performance. The under-subscription problem is also common and severe. The analysis indicated that 19.68% of batch jobs under-subscribe CPU resources, among which 32.68% under-subscribe CPU resources by 25%-50% while the other 65.83% under-subscribe CPU resources by over 50%. These jobs could suffer from longer execution time.

Service Jobs

Service jobs are time-sensitive, therefore they are given higher privileges in obtaining system resources. To prevent interference from other co-running jobs,

service jobs are executed in containers. Each container is guaranteed to have a certain amount of CPU, memory and disk space, which is exclusively allocated for it. However, the allocated resources for containers are always not fully utilized. Figure 3.7 shows that in the 2017 Alibaba trace, 90.12% of the containers use less than 25% allocated CPU and 8.94% of containers use only 25% to 50% of CPU resources. As for memory and disk resources, 13.92% of the containers use only less than half of the allocated memory and 96.99% of containers use less than 25% of the allocated disk space.

Summary

To summarize, the over-subscription and under-subscription problems co-exist in the Alibaba system and are equally severe, which hurts the system efficiency in two ways. The excessive over-subscription of resources causes low resource utilization and decreases the system efficiency. Meanwhile, jobs that under-subscribe resources suffer from delayed execution time due to the lack of sufficient resources.

Simulator Design

The previous section analyzed the unique workload characteristics and resource utilization of the Alibaba system. This analysis indicates that there is a great potential to further improve system efficiency. For example, the system utilization and performance could be improved if the over-subscribed resources were allocated to the under-subscribed jobs. The number of machines in the cluster could be reduced without degrading performance. Unfortunately, the Alibaba trace was given as static trace files and all parameters in the trace files had been pre-determined. In order to conduct experiments to quantitatively study the impact of varied cluster size and different system configurations on the performance and TCO of the Alibaba system, it was essential to design and develop a simulator that could

reproduce the scheduling process. In 2013, Google developed an "Omega" simulator [89] to emulate its cluster scheduling process and explored the impact of different configurations and scheduling strategies on system efficiency and utilization [74]. I adopted a similar methodology used by Google to design and develop the simulator.

What follows is a presentation of the design of the Alibaba scheduler simulator driven by real Alibaba production workloads [63] (it can also take synthetic workloads). The design and implementation were largely derived from the details explained in the Alibaba Fuxi paper [64]. The Google Omega scheduler simulator [89] also provided important insights (e.g. how to use Agenda to handle job submission and how to update the event time using a clock) on how to implement a scheduler simulator for large-scale systems. This simulator supports both the 2017 and 2018 Alibaba traces.

Fuxi Scheduler Workflow

The Fuxi scheduler includes five main components: the FuxiMaster (FM), the FuxiAgent (FA), the AppMaster (AM), the Task Master (TM), and the Task Worker (TW). Figure 3.8 plots the typical workflow that the Fuxi scheduler allocates resources for jobs. Figure 3.9 shows the steps to deallocate resources and terminate a batch job. When a job is submitted to the FuxiMaster, the FuxiMaster will first try to find a FuxiAgent residing on a machine with available resources. Next, the FuxiAgent will start the AppMaster for the job. When receiving the job description, the AppMaster first predicts the resource needs for each task then sends resource request to the FuxiMaster. The FuxiMaster will check the available resource pool for free resources, allocate the resources when available and send resource grants to the AppMaster. If there are not enough resources available, the FuxiMaster will push the unfulfilled resource requests into a queue called locality tree and send resource grants in an incremental fashion. Every time when certain resources are granted, the

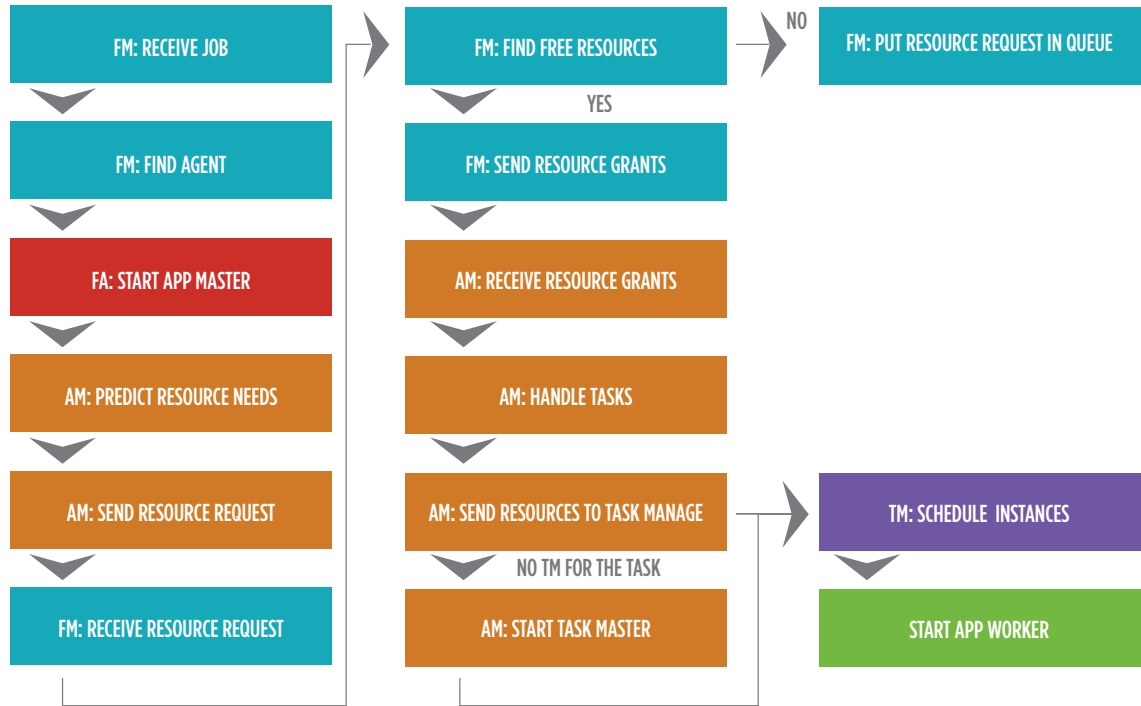


Figure 3.8: Fuxi Scheduler Resource Allocation Workflow

FuxiMaster will update the status of the resource requests. Once the AppMaster receives resource grants, it determines which task to be scheduled next based on task dependency and the resources received. The AppMaster then starts the Task Master for that task and sends resources to it. After that, the Task Master will be able to schedule the instances of the task for execution. While a task is running, the Task Worker will periodically report the status of the task to the Task Master. When all Task Workers have reported completion to the Task Master, the Task Master will consider the job to be completed. The Task Master will send resource deallocation request to the AppMaster, which will be forwarded to the FuxiMaster. It is the responsibility of the FuxiMaster to finally deallocate the resources and update the free resource pool. Once all tasks in a job have completed, the AppMaster will report its completion to the FuxiMaster then terminate the job.



Figure 3.9: Fuxi Scheduler Resource Deallocation Workflow

Terminology

To better understand the workflow of the Fuxi scheduler simulator, it is beneficial to briefly introduce four important terminologies: job description, resource requests, resource grants, and locality tree. More details of these terminologies can be found in the Alibaba Fuxi paper [64].

1. Job description - A job description contains job id, job submission time, number of tasks, and task description of each task in the job. Task description includes the task's creation time, dependency, and the required resources. In our simulator, a job description is generated for each job residing in the input trace files.
2. Resource request - After receiving the job description, the AppMaster calculates the resource needs of the job and sends resource request to the FuxiMaster. A resource request consists of schedule unit definition, quantities for each schedule unit, and other attributes such as location preference and priority etc. [64], which is sent from the AppMaster to the FuxiMaster to apply for resources.

3. Resource grants - A resource grant is sent by the FuxiMaster to the AppMaster, which carries information about the machine id, the CPU id, memory, and disk space allocated for the job.
4. Locality tree - In the Fuxi's scheduler, a job only needs to specify its resource demand once. The FuxiMaster makes incremental resource allocation if the requested resources cannot be fulfilled at once. This can significantly reduce the communication and message processing overhead by preventing jobs from repetitively asserting full resource demands and checking the status of their requests. The FuxiMaster keeps the unfulfilled demands of each job in a locality tree, which supports the incremental resource allocation protocol [64]. In this simulator, a locality tree is implemented to keep unfulfilled resource requests in the waiting queue and automatically grants resources to the AppMaster when new resources are available. Once the requested resources are fulfilled, the FuxiMaster removes the request from the queue. The Alibaba Fuxi paper [64] indicates that jobs with higher priority will get the requested resources early. When the priority is the same, the waiting time will be taken into consideration. This simulator uses the First In First Out (FIFO) algorithm to resolve the conflict when the priority of multiple jobs is identical.
5. Free resource pool - Upon receipt of resource requests from the AppMaster, the FuxiMaster will check the free resource pool and try to find sufficient free resources which can meet the application's locality requirements. If the free resource is insufficient, the resource requests will be queued by the FuxiMaster in the locality tree [64]. Before making final allocation decisions, the FuxiMaster considers load balancing of each machine as well. If the load on a machine is higher than the specified load limit, resources on that machine will not be removed from the free resource pool.

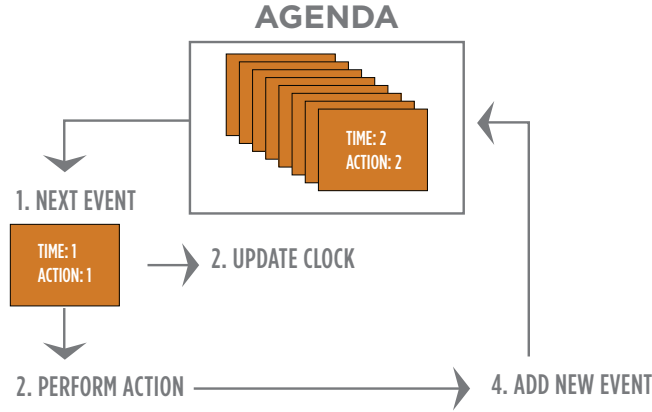


Figure 3.10: Simulating Job Submissions using Agenda

Job Submission

The most difficult part was to accurately simulate the submission of jobs to the Fuxi scheduler at the specific time given in the Alibaba trace files or synthetically generated input files. Using the implementation of the Google Omega scheduler simulator as a basis [89], I used a priority queue of Agenda that is sorted by time to simulate the job submission process (see Figure 3.10). Essentially, the Agenda is an event queue that collects the events and sorts them by time. Each event has an action and a specific time at which the action is performed. The simulator maintains a clock that is updated each time an event occurs with the time of that event. Initially, the Agenda is filled with job arrival events and each job arrival event will add new events to the Agenda.

Simulator Input

Three input files were generated from the released Alibaba trace files: the batch job file, the batch task file, and the service job file.

1. The batch job file contains the following information: job arrival time, job id, and the number of tasks in the job. The job arrival time is the earliest creation

time of a task among all tasks in the job. The number of tasks in the job is counted directly from the corresponding trace file. The job submission time is one second less than the job arrival time, which simulates the time needed to start the AppMaster before the task can be created. In real systems, the task creation time may vary depending on the delay in scheduling the job. Here the one second is a rough estimate because the real delay is not provided in the Alibaba trace files.

2. The batch task file provides information that includes job id, task id, task creation time, the number of instances in the task, duration of each instance, and the amount of CPU, memory and disk space needed for an instance to run. Duration of the instance is calculated from the trace file as the average difference between instance creation time and end time of the successful trial of all instances. The number of instances of the task is calculated by counting final trial of instances with the same job id and task id. The required CPU and memory resources are calculated as the average CPU and memory usage of all instances of the task.
3. The service job file consists of container id, creation time of the container, duration of the container, machine id on which the container is created, the number of CPUs it needs, and their CPU ids as well as the allocated memory and disk space information. All information is available from the container event table except for the duration. As no information exists in the trace files indicating how long a container will exist, it can be assumed the container exists until the end of simulation. It can also be inferred from the trace file that resources for each container are available when they are assigned to a server, thus reallocating container resources as needed.

Simulator Output

The simulator collected and recorded the following information of each individual machine in a 5-second interval: (1) the amount of CPU, memory, and disk space used; (2) the wait time in queue of each job (time between the job creation and an AppMaster is created for the job); (3) the wait time of each resource request (time between it is created and fulfilled completely); (4) the wait time of each task before it gets first resource grant; and (5) the wait time of each task before it gets last resource grant. The savings on total cost of ownership (TCO) was not directly produced by the simulator. Rather, it was calculated from the simulator output results indirectly after considering the cost of each server, the average power usage of a server, and the electricity price.

Limitations

Although the goal was to emulate the Alibaba scheduling process as accurately as possible, it is worth noting that the simulator is not able to completely emulate the process due to the lack of complete information about the scheduler. Additionally, the simulation might not be 100% accurate due to the following constraints:

1. Referring to Table 3.1 again, it is obvious the enormity of the 2018 dataset made it difficult to parse fully in a timely manner. For example, performing the simulation until the cluster has completed all tasks required an excessive amount of disk space and several days of run-time, making it difficult to perform tests in an efficient manner. Thus, due to memory and time constraints, only the 2017 trace and 24 hours of the 2018 trace were used in the simulation. Since the 2017 and 2018 cluster traces exhibit very similar over-subscription and under-subscription problems and the daily usage

patterns in the 2018 traces are highly repetitive (ref. Figure 3.2), it is reasonable to extrapolate the simulation results to the entire 2018 dataset.

2. The machine utilization information was only available after the first batch jobs were submitted. This could cause some abnormalities due to the lack of a complete snapshot of the cluster. Further, only the average machine utilization for jobs was considered as this is all that was provided in the real trace files, i.e. the simulator tries to duplicate the results found in the trace. This could possibly affect the scheduler's ability to perform most efficiently.
3. Without overhead information of creating the AppMaster, the Task Master, and the Task Worker, the simulator estimates the information, which might affect the waiting time of a job or task and indirectly affects system utilization.
4. The real Alibaba system uses Fuxi scheduler to dispatch batch jobs and the Sigma scheduler to dispatch service jobs. The Fuxi scheduler and Sigma scheduler work together for the mix scheduling policy. Unfortunately, the details about the Sigma scheduler implementation is not published. This simulator makes simple assumptions that the Sigma scheduler creates a container using the resources mentioned in the trace file without much overhead. In addition, since the trace files do not provide information on how long a container exists, the simulator allows the container (once created) to exist for the entire life cycle of the simulation. Recall that as containers usually request more resources, this might have negative affects on system utilization.
5. The real Fuxi scheduler supports other important functionalities such as fault tolerant scheduling and multi-level black-listing scheme [64], which are not included in this simulator but which could affect system utilization.

Experimental Results

Using the simulator, the scheduling process was reproduced, followed by a quantitative evaluation of the impact of varied cluster size on system performance and TCO. Specifically, two groups of experiments were conducted and the results analyzed. Impact of Cluster Size below evaluates the impact of reducing cluster size on system performance, while TCO Reduction analyzes reduction to overall TCO.

Impact of Cluster Size

The total number of machines in the 2017 Alibaba trace file is 1,313. Theoretically, reducing the number of machines could possibly increase system utilization. However, this may hurt performance by increasing the wait time of tasks. The question is, to what extent can the cluster size be reduced with a negligible influence on task wait time? This group of experiments was designed to study the impact of varied cluster size on system performance. The number of machines in the cluster are reduced incrementally, and wait time observed for resource allocation. Figures 3.11 and 3.12 show the percentage of tasks that must wait for incremental resource grants. It can be seen that for the cluster size of 1,313 and 1,248 machines, more than 99% of the tasks are scheduled with no wait time and received was requested resource as a single grant. However, when the cluster size is reduced further to 1,138 machines, the percentage of tasks that did not wait dropped to approximately 96.5%. Similarly, the percentage of tasks that received all resources as a single grant dropped to 95.16%. Figure 3.13 shows the number of tasks that waited until receiving the first grant. Wait time until first grant is the wait time of the task to receive the first resource grant after sending its resource request. The task is scheduled to start once the first grant is received, so wait time until first grant also means wait time until the task is scheduled. The number of

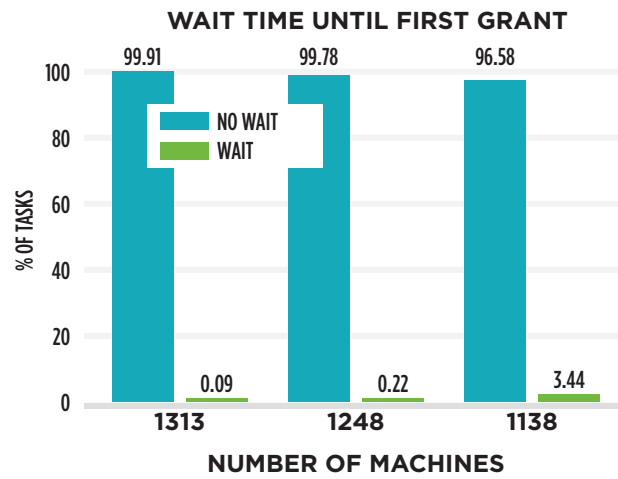


Figure 3.11: Wait Time Until First Grant

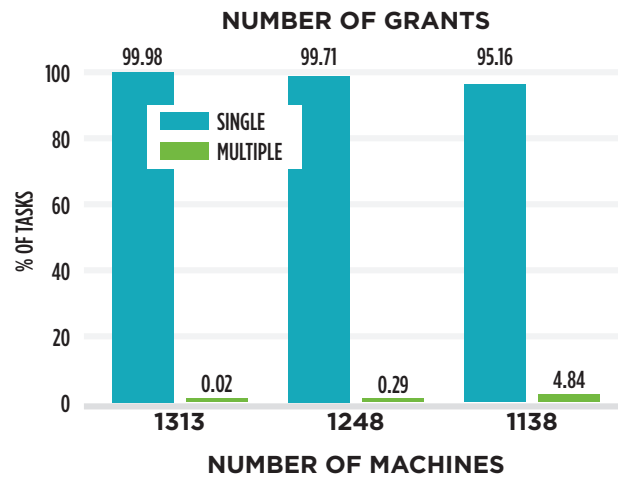


Figure 3.12: Number of Grants

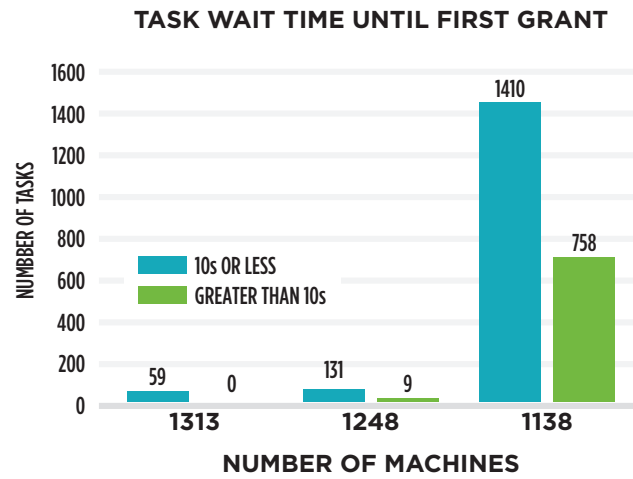


Figure 3.13: Task Wait Time Until First Grant

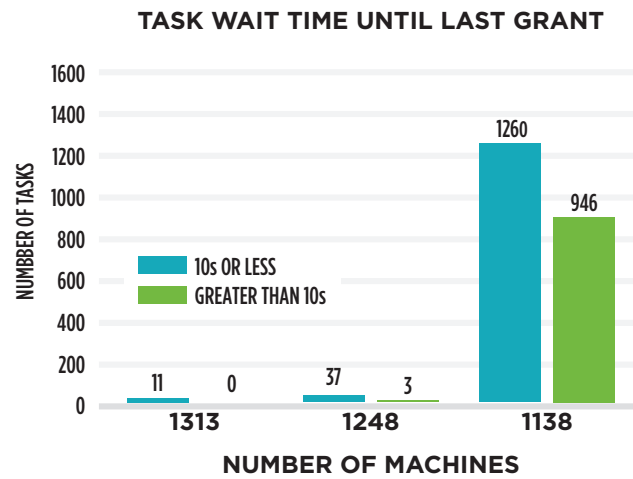


Figure 3.14: Task Wait Time Until Last Grant

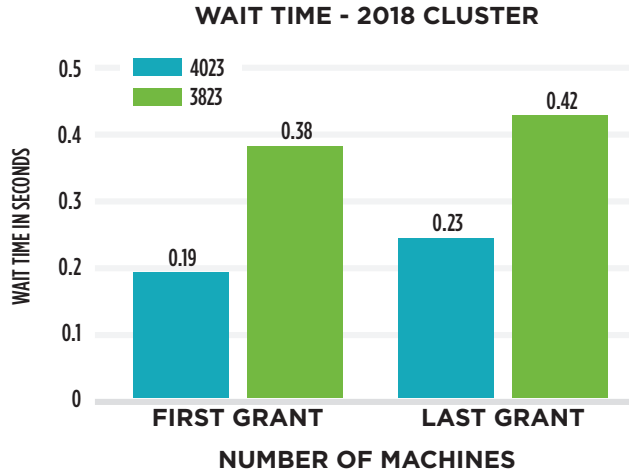


Figure 3.15: Wait Time - 2018 Cluster

tasks that waited less than 10 seconds until first grant was 59, 131 and 1,410 for cluster size of 1,313, 1,248 and 1,138 machines respectively. From this it can be inferred that the number of tasks that must wait is increased significantly when reducing the cluster size to 1,138. When comparing the maximum wait time of tasks until first grant is received for cluster size of 1,313 and 1,138, the latter had a higher wait time by 23 seconds. But in case of cluster with 1,248 machines, the number of tasks that waited for more than 10 seconds for the first grant was only 9.

Figure 3.14 shows that no task needed to wait for more than 10 seconds for the last grant for a cluster size of 1,313 machines. For the cluster size of 1,248, the number of tasks with wait time until last grant less than 10 seconds had slightly increased when comparing it with cluster size of 1,313 machines. However, the number of tasks that must wait less than 10 seconds had significantly increased (close to an unacceptable level) when the cluster size was reduced to 1,138 machines. Preliminary results with the 2018 trace demonstrated comparable response times with a 5% reduction in cluster size. Figure 3.15 shows the wait time in seconds until first grant and last grant when simulating the first 24 hours of the

2018 dataset for 4,023 and 3,832 machines, respectively. It can be seen that the difference in response time of the cluster was negligible, and further testing on the entire 7 day period is expected to produce results similar to the 2017 trace (see Figure 3.2 for the repeated daily patterns).

TCO Reduction

The previous results have shown that the cluster size can be reduced by 5% without degrading performance. What follows is a calculation to determine how much TCO reduction can be achieved. According to [64], the Alibaba server contains a 6-core Xeon E5-2430 processor (2.20GHz), a 96GB memory and a 12*2T disk array. Machines are connected via two gigabit Ethernet ports. The Dell Power Edge T340 server is used, which has very similar configurations as the Alibaba server specifications, to estimate the TCO. The Dell Power Edge T340 server consumes approximately 1.653 kilowatt-hours (kWh) of power per day and costs \$3,078 [90]. I believe this is an extremely conservative estimate as a typical enterprise server purchased in 2020 would cost roughly twice this amount and use at least double amount of energy every day. In the following cost calculation, it can be assumed the electricity price of the Alibaba datacenter is \$0.11 per kilowatt hour, which was the average data center electricity price reported by Data Center Frontier [91] at the time this study was conducted. These calculations do not consider power savings techniques such as DVFS as power usage information is unavailable in the trace files. This concern is addressed by using reasonable, but conservative, estimates of overall system power usage.

Table 3.3 compares the energy consumption of the two clusters and Table 3.4 compares the TCO. By shrinking the 2018 cluster size from 4,023 to 3,832 (95% of its original size), the total cost of ownership can be reduced by \$600,574. It is worth noting that this number only includes the reduced hardware purchase cost and

Table 3.3: Energy Consumption Comparison

Cluster Size	Average Energy Per Day (kWh)	Average Energy Per Year (kWh)
4,023	6650	2,427,257
3,832	6334	2,312,018
Savings	316	115,239

Table 3.4: TCO Comparison

Cluster Size	Purchase Cost (USD)	Yearly Electricity Cost (USD)
4,023	\$12,382,794	\$266,998
3,832	\$11,794,896	\$254,322
Savings	\$587,898	\$12,676
Total Savings (USD)	\$600,574	

power savings on the server side. The real TCO benefit will be much larger if the reduced cost on server maintenance and power savings on the cooling facilities are also included.

A cursory analysis of the TCO in Table 3.4 reveals substantial savings, albeit on a small scale. Alibaba’s large size, however, makes this almost insignificant when considering their \$12 billion yearly operating costs as of 2018 [92]. Per Netcraft [52], Alibaba had 432,000 Internet-facing servers as of 2017, and this number is most likely much larger today. Table 3.5 calculates the TCO reduction if 5% of Alibaba’s 432,000 Internet facing servers were removed (i.e. removing 21,600 servers in total). The results show that TCO can be reduced almost \$68 million overall, which includes nearly \$66.5 million of hardware capital savings and over \$1.4 million annual electricity bill conservation. The TCO benefit would be even larger if savings on physical footprint of data centers, maintenance personnel, and cooling facilities were also considered. Meanwhile, less energy consumption allows more rapid growth due to the ability to use existing power grids and infrastructure, thus increasing overall long-term profitability.

Table 3.5: Extrapolated TCO Comparison

Cluster Size	Purchase Cost (USD)	Yearly Electricity Cost (USD)
432,000	\$1,237,356,000	\$28,670,954
410,400	\$1,263,211,200	\$27,237,407
Savings	\$66,484,800	\$1,433,548
Total Savings (USD)	\$67,918,348	

Conclusions and Future Work

E-commerce has experienced rapid growth in the past decade and embraced a new boom during the coronavirus pandemic. As more vendors consider online sales as essential business, the number of data centers that support cloud services with hybrid workloads will rise quickly. To make a profit, these data centers must have exceptional performance with reasonable TCO. This study discusses the best practices of Alibaba by analyzing its 2017 and 2018 cluster traces. My analysis confirmed that consolidating the time-sensitive service jobs with time-insensitive batch jobs can significantly improve system utilization and reduce TCO. Meanwhile, it was discovered that the over-subscription and under-subscription problems co-exist in the Alibaba system. Addressing this problem will have great potential to further reduce TCO without compromising performance. In addition, a simulator was developed to reproduce the scheduling process using the Alibaba trace, which allowed the evaluation of the impact of varied cluster size on the performance (in term of task wait time) and TCO of the Alibaba system. The experimental results indicate that reducing the size of the cluster running the Alibaba 2018 trace by 5% can save approximately \$600,000 of TCO without degrading system performance. The total TCO reduction will scale up to nearly \$68 million if extrapolated to Alibaba’s 432,000 Internet facing servers.

Work in this area can be further extended to reduce the resources allocated for containers and allocate them for under-subscribed batch jobs, provided that the

performance of time-sensitive service jobs will not be affected. Moreover, the simulator can be further improved by addressing some of the limitations stated above in Limitations.

IV. PUBLIC CLOUD WORKLOAD - AZURE

Introduction

Despite the urgent need for reducing cloud waste and cost, the majority of cloud users do not have sufficient knowledge in understanding the complex pricing schemes or lack important information in deciding what cloud resources are most appropriate for their applications or services. Discerning pricing schemes from various cloud providers is never straightforward and detailed information may even be opaque to cloud users [93]. For example, an application utilizing only 5 minutes of actual cloud resources could be charged throughout the lifetime of the virtual machine (VM) if users do not shut it down. Such information is not always obvious from the vendor provided cost calculation tools, which can lead to over-budget issues and unexpected cloud waste.

Thus, to make effective use of cloud resources, transparency in pricing and understanding what resources are granted for that price are paramount. 45% of cloud customer's expenditures are spent on resources they will never use [94], wasting money that could be better utilized elsewhere. Even comparing the pricing schemes between vendors is difficult, as apples-to-apples comparisons are not always possible, leading to confusion about which vendor is able to provide the needed service for the best cost [95]. Moreover, simply looking at CPU and memory allocation for VMs gives limited information on the efficiency of those VMs or the real out-of-pocket cost involved to deploy such VMs [96]. Transparency in segmenting and pricing cloud resources is thus valuable for both customers and vendors and greatly improves the decision making process when selecting cloud services.

Although cloud providers are making consistent efforts to help users optimize

their cloud cost, current best practices are often limited to collecting basic monitoring data (e.g. CPU percentage, network in/out, and disk read/write). The collected monitoring data is automatically displayed on a dashboard where users can visually observe the resource utilization status of their VMs. The Azure cloud portal also provides advanced monitoring that leverages Azure Diagnostics extension to profile more detailed information such as custom performance counters, application logs, windows event logs, crash dumps, customer error logs etc. [97]. This information is helpful but cannot provide direct and valuable data that is related to the cloud cost and waste.

It is abundantly clear that the cloud community needs more in-depth knowledge about how to achieve cost savings when migrating services to the cloud. However, there is a dearth of information in the literature that can provide a detailed analysis about cloud waste or viable solutions to help reduce cloud cost. To address these concerns, I conducted a comprehensive analysis of the Microsoft Azure 2019 traces and propose Cloud Waste Points (CWP) to quantitatively evaluate the cloud waste of VMs. I also propose Cloud Waste Indicator (CWI) to classify users as red, normal or green users and rank them using Cloud Utilization Score (CUS) based on their cloud usage efficiency. Additionally, this work attempts to help cloud users reduce cost and waste by identifying red VMs and recommending lower priced VMs without compromising quality of service (QoS). More specifically, this study makes the following contributions:

1. I analyzed 235GB of data containing nearly 2.7 million VMs created by 6,687 Azure cloud users within 30 days, which accumulates to nearly two billion VM readings. The results reveal that a large portion of VMs are under-utilized or over-provisioned for resources.
2. I performed a comprehensive, descriptive analysis of VMs running within the Azure cloud and categorize those VMs to better understand their computation

impact on the cloud. It is noted that VMs are categorized by CPU core count, requested memory, and the type of VM employed, both Delay-Insensitive which do not need on-demand resources, and Interactive VMs which must be available real-time.

3. Cloud Waste Points (CWP) are proposed to quantitatively evaluate the waste of VMs in utilizing cloud resources. VMs that utilize cloud resource efficiently are categorized as green VMs and those that waste cloud resources as red VMs, according to their average CPU utilization rate. To help users optimize cost and waste, an algorithm is introduced that can identify red VMs and recommend more appropriate VMs for users to run their cloud applications with less cost and waste. These experiments show that over \$22 million cost savings (i.e. approximately 36% in total cost reduction) can be achieved if the proposed recommendations are adopted by users.
4. Cloud Utilization Score (CUS) of 6,687 Azure users is calculated, identifying those with significant cloud impact (6,121 total), allowing classification as green users, normal users, and red users. Green users are those who generally utilize cloud resource efficiently, and red users are those who tend to over-allocate resources thus creating more waste. The resource usage of normal users are in between. The analysis shows that 1,020 users (~15%) are classified as red users while 2,321 users (~36%) can be classified as green users. The Cloud Utilization Score provides a straightforward and intuitive way to help users compare their cloud usage efficiency to other users' efficiency.

Related Work

Many cloud users struggle to forecast their fast-growing cloud costs accurately and effectively reduce the cost and waste in utilizing cloud resources. Several studies

have been published in analyzing different cloud workloads. Di et al. characterized the tasks in a Google data center and compared them with grid workloads [98]. Ho et al. proposed a model to optimize the energy consumption of transactional applications running on federated clouds [99]. Mishra et al. described an approach to characterize Google cloud workloads [100]. Reiss et al. analyzed the Google trace with heterogeneous workloads and found that longer-running jobs tend to have relatively stable resource utilization, which can be used to optimize the scheduling of heterogeneous jobs on multi-purpose clusters [82]. These previous studies on workload characterization provided useful information on how to analyze complex workloads.

There are several studies regarding how to reduce cost of large scale cloud systems. Alalawi and Daly conducted a survey on various techniques that can reduce the energy usage of MapReduce in Hadoop systems [101]. Feifei et al. [102] proposed a model and associated analysis tool to measure energy consumption in cloud environments based on different runtime tasks. Qiu et al. proposed a genetic-based optimization algorithm to reduce the energy cost of cloud systems with phase change memory [103]. Mamun et al. proposed an intra- and inter-server smart task scheduling algorithm, which can jointly optimize profit and energy when allocating jobs to datacenters [104]. Wen et al. proposed an energy and cost aware algorithm for scheduling instance-intensive workflows with batch processing in clouds [105]. Everman et al. analyzed the Alibaba traces and demonstrated the great potential of reducing the total cost of ownership in Alibaba system by co-scheduling hybrid workloads to improve system utilization [106]. Google recently presented carbon-aware computing for datacenters, which actively minimizes electricity-based carbon footprint and power infrastructure costs by delaying temporally flexible workloads. The core component of Google’s carbon-intelligent compute management system is a suite of analytical pipelines that gather the next

day’s carbon intensity forecasts, train day-ahead demand prediction models, and use risk-aware optimization to generate the next day’s carbon-aware Virtual Capacity Curves (VCCs) for all datacenter clusters across Google’s fleet [107].

To the best of my knowledge, there are only three existing studies that considered the Microsoft Azure cloud platform directly. Hadary et al. described the design and implementation of Protean, which allocates VMs on the Azure platform [108]. Protean allows Azure to expand its VM under different conditions with little change to the core infrastructure by separating policy and mechanisms. It also revealed the resource spikes caused by COVID-19 and discussed how to mitigate issues induced by those spikes. Shahrads et al. focused specifically on Function as a Service (FaaS) in the Azure system [109]. They discussed the lack of research in characterizing FaaS workloads, which is a vital component in ensuring adequate resources are available at the lowest cost. They tackled this problem by reducing the number of function cold starts via a resource management policy. The most relevant work was published by Cortez et al. [110] because it utilized the 2017 Azure data traces, the precursor to the 2019 traces examined in our paper. They described the Azure 2017 workloads in detail and discussed certain behaviors that can be used to predict future behavior of VM workloads. They also introduced a model that can analyze VM telemetry offline and offer predictions to resource managers. However, their predictive model aimed to increase the utilization of Azure from the system perspective and was tied directly to the Azure scheduling system. In contrast, this study focuses on the user’s perspective and strives to help cloud users reduce waste and cost.

All of the above studies are concerned with either the Azure platform itself, or similar cloud hosting platforms, but do not directly address the end-user’s ability to select the proper resources for their individual needs. This study differentiates itself by making recommendations to end-users and helping them reducing cloud waste

and cost. When requested resources are highly utilized by end-users, the overall utilization of the Azure system is improved as well. Therefore, this work complements previous system oriented optimizations instead of competing with them. More importantly, the following approaches and metrics (e.g. CWP, CWI, and CUS) can be applied to other cloud platforms such as Amazon AWS or Google Cloud because the workload characteristics and user behaviors are unlikely to change when using another cloud system.

Workload Analysis

Microsoft Azure

Microsoft Azure is a public cloud computing platform that provides a wide range of services, including software as a service (SaaS), infrastructure as a service (IaaS), and platform as a service (PaaS). It is currently (at the time of this writing) the second largest cloud platform and serves many enterprises such as eBay, Boeing, Samsung, BMW etc. [111]. Azure relies heavily on the virtualization technology and allows users to deploy virtual machines (VMs) in various capacities that are suitable to their applications. Users can select the classic or resource manager models, which are two different ways for deploying and managing their cloud resources in Azure. The classic model allows individual resources (e.g. a VM or database) to be monitored while the resource manager model groups related services to allow ease of management [112].

Azure Trace Files

Microsoft released a set of Azure-related data traces in 2017 and 2019 respectively. These are representative traces of the VM workloads collected by monitoring a subset of hosted VMs on the Azure platform. They are "sanitized" subsets of a singular Azure geographical region, which means identifying

information has been removed to retain anonymity. In this study, the 2019 trace is analyzed as it was more recent and representative of the current Azure cloud platform. The 2019 Azure trace contains a subset of applications running on Azure during July of 2019 with the following summarized information [113]:

- 235GB of data contained within 198 files
- 30 consecutive days of VM readings
- 2,695,548 total VMs
- 6,687 individual users
- 5-minute VM CPU utilization readings
- 104,371,713 total VM hours
- 1,942,780,023 VM utilization readings
- > 380,000,000 total virtual core hours

Descriptive Analysis

To better understand how users utilize Azure cloud, a descriptive analysis was conducted, and observations summarized as follows:

1. Figure 4.1 shows the core distribution of all VMs. It demonstrates that the vast majority of VMs contain either 2 or 4 cores, among which 59% have 2 cores and 31% request 4 cores. There are a limited number of 8 core VMs (7%) and only roughly 3% of VMs are deployed with 24 or more cores.
2. Figure 4.2 plots the memory distribution of all VMs. It can be observed that VMs with 8GB and 32 GB of DRAM are most common, which account for 39% and 32% respectively. 2 GB and 4 GB VMs are representing 12% and 16% while 64 GB and greater VMs is merely 1% of all VMs.

3. Figure 4.3 visualizes the number of VMs for each core/ memory combination. Specifically, ~175k VMs are deployed with 4 virtual cores and 8 GB of memory. Meanwhile, over 600k VMs are deployed with 4 cores and 32 GB of memory. 2-core VMs are distributed a bit more evenly in combination with 2 GB or 4 GB DRAM constituting around half the number of deployments, and 8 GB VMs making up the other half. Very few 2-core VMs are deployed with 32 GB of DRAM but nearly all 8-core VMs contain 32 GB of memory. There are also a very small portion of 8-core VMs being deployed with 64 GB DRAM. In the trace files, all 24-core VMs are allocated with 64 GB of memory and all 30-core VMs are allocated with 70 GB of memory. In our analysis, any memory bucket greater than 70 GB is treated as 70 GB.
4. Table 4.1 provides an overarching description of the lifetime of all VMs. The average lifetime of approximately 2.7 million VMs is 60.27 hours. However, the standard deviation, minimum, and maximum VM hours clearly show that the lifetime of each VM varies wildly. It can be deduced from the table that the shortest lifetime of any VM in the dataset is 0.083 hours (i.e. about 5 minutes) while the VM with the longest lifetime runs for almost 720 hours (i.e. 30 days).
5. There are three categories of VMs in the Azure traces: Delay-Insensitive, Interactive, and Unknown. Delay-Insensitive VMs do not need on-demand resources at all times thus can wait longer until those resources are available. Interactive VMs are used to serve time-sensitive applications so they prefer no wait time and guaranteed resources whenever required. The Unknown category is for those VMs that do not have either of the previous two categories specified. Figure 4.4 shows the percentage of core hour usage in each category, from which can be seen that most of the computational load is performed by Delay-Insensitive VMs. Interactive VMs also use over 30% of total core hours.

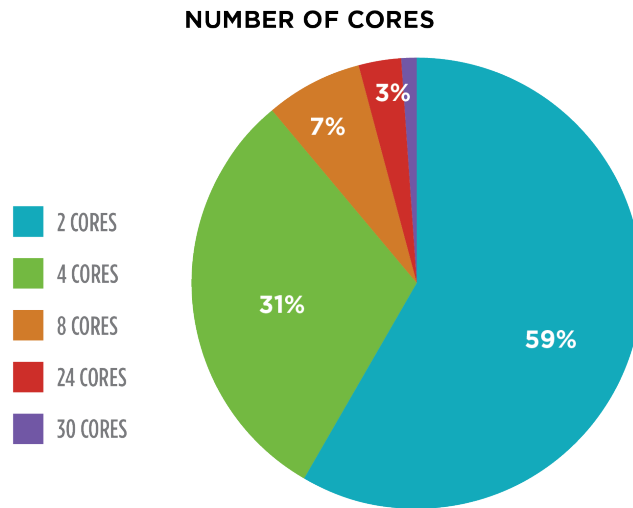


Figure 4.1: VM Core Distribution

Table 4.1: 2019 Azure VM Distribution

Number of VMs	2,695,548
Total Hours	~104,371,713
Mean Hours	60.27 hours
Standard Deviation	186.18 hours
Minimum	.083 hours
Maximum	719.83 hours

Unknown VMs only constitute about 7% of the overall core hours.

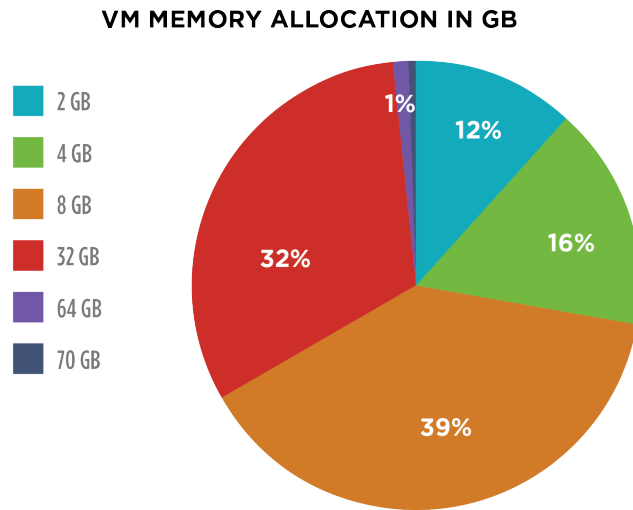


Figure 4.2: VM Memory Distribution

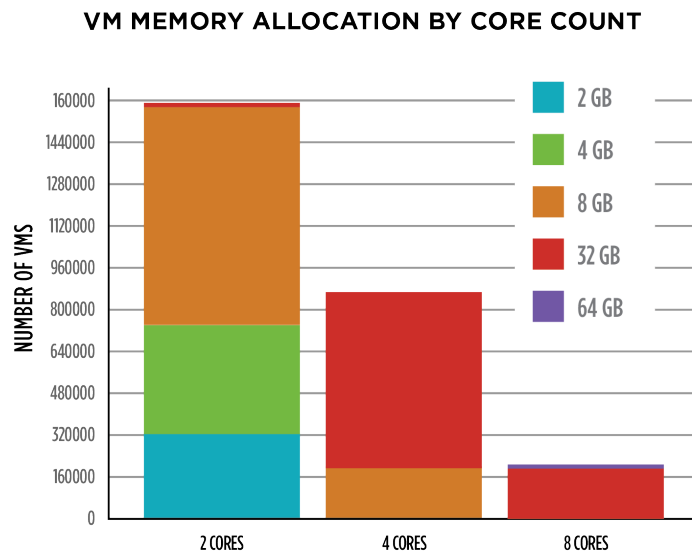


Figure 4.3: Memory Allocation by Number of Cores

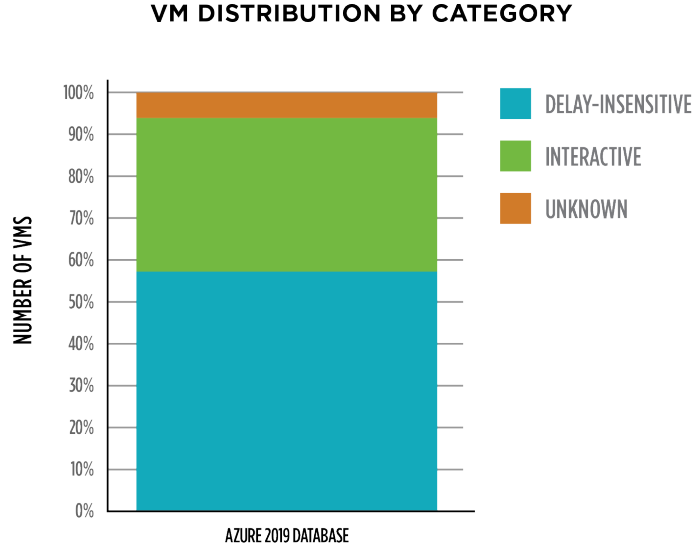


Figure 4.4: Category Distribution

Cloud Waste and Cost Analysis

To help users reduce cloud cost and waste, it is essential to calculate the cost and waste of each VM, quantitatively evaluate the overall cloud usage efficiency of different users, help users identify costly and wasteful VMs, and make useful recommendations for users to improve their cloud usage efficiency. In this section, I present that methodology and metrics in detail. Specifically, Azure Pricing Model below discusses the pricing model as specified by Azure. VM Cost Calculation explains how the cost of each VM is calculated. Green and Red VMs demonstrates how the waste of each VM is determined by the Cloud Waste Points and VM Cloud Waste Points shows how to categorize different VMs as green or red VMs. Green and Red Users introduces the Cloud Usage Indicator (CWI) metric and discusses how CWI is utilized to classify Azure users as green, normal, or red users. Recommendation Algorithm and Cost Savings presents the recommendation algorithm, which can substantially reduce cloud waste and cost. Why Should Cloud Providers Care? highlights several reasons that cloud providers

would be motivated to reduce users' waste and cost.

Azure Pricing Model

According to the pricing information listed on Azure website [114], each VM is priced based on the requested core count and memory size. However, there are various factors that can affect the final price of each VM. These factors include but are not limited to the choice of Operating Systems, the region that provides the cloud services, the type of servers (e.g. general purpose, CPU optimized, or Memory-optimized), the type of VMs (Delay-Insensitive or Interactive), and the associated storage capacity. Deriving an accurate pricing model needs specific information about all these factors. Unfortunately, the required information to apply complex pricing was missing in the Azure traces (e.g. lacking information regarding which VM is discounted and at what rate). Even if all the required information was made available, the pricing model evolves over time and may have changed at the time this study was conducted. Therefore, the following assumptions were made to derive a relatively accurate pricing model, as shown in Table 4.2.

- Deployed VMs are run on Linux (CentOS or Ubuntu). Azure offers a different pricing scheme for Windows VMs, albeit more expensive for comparable resources. Thus these cost calculation will be more conservative.
- The VM will be deployed to the US-West (California) region.
- The VM is general purpose, not CPU or Memory-optimized, and the physical CPUs available to the VM are circa 2019 or older.
- Minimal storage is available for each VM.
- Users "pay as they go", and do not receive any discounts for pre-paying nor volume purchasing.

- The price of Interactive VMs is 3.33x of the price of a Delay-Insensitive VMs that requests the same amount of cores and DRAM. Note that this ratio is based on the information from Google Cloud [115] because the Azure website [114] does not provide separate price for Interactive VMs and Delay-Insensitive VMs.
- VMs categorized as Unknown are considered Delay-Insensitive when calculating price, which ensures a conservative estimate of cost-savings.

From Table 4.2, it can be observed that the corresponding price of each VM with different core counts and DRAM capacity (e.g. a 2-core + 4 GB VM would be charged 7.6 cents/hour or about \$1.85 per day). It is worth mentioning again that Table 4.2 is a simplified version of Azure pricing model at the time of conducting this study. The real Azure pricing model offers more complex discount options (e.g. reserved VM instances and spot pricing) and different price schemes for special hardware (e.g. HBv3 VMs optimized for high performance computing). The complex pricing model is unused because some of the required information is not provided in the Azure traces and the goal of this study is to provide a simple, yet close-enough, estimation to calculate cloud cost and waste.

VM Cost Calculation

The VM cost was not provided in the Azure traces. To calculate the cost of each VM, the price of each VM, and its corresponding lifetime, must be known. Core hours are also calculated to help compare the computational resources utilized by different users.

- **VM Price**

The price of each type of VM was retrieved from Table 4.2.

- **VM Lifetime**

Table 4.2: 2019 Azure VM Pricing in USD per hour

	Memory					
	2GB	4GB	8GB	32GB	64GB	70 GB
Cores						
2	.036	.076	.099	.126		
4			.159	.208		
8				.333	.437	
24					.832	
30						1.353

VM lifetime is the length of time in hours a VM exists, which was calculated by the difference between the creation and deletion timestamps within the traces after converting seconds into hours.

- **Core Hours**

Core hours were calculated by multiplying the lifetime of the VM and the number of cores of that VM. Core hours are used to indicate the computation resources utilized by a VM. Note that different combinations can lead to the same number of core hours. For example, a 2-core VM running for 1 hour and a 4-core VM running for 30 minutes will yield identical core hours.

- **VM Cost**

The cost of each VM was calculated by multiplying its lifetime and the corresponding USD amount denoted in pricing Table 4.2.

Green and Red VMs

Users run a variety of applications on the Azure cloud using different sizes of VMs. To understand how efficiently the cloud resources allocated for these VMs are utilized, an analysis was conducted on the average CPU utilization of different sizes of VMs. From Table 4.3, it can clearly be observed that a large portion of allocated

Table 4.3: Average CPU Utilization of Different VMs

Cores	Memory (GB)	Average CPU Utilization
2	2	23.61%
2	4	16.04%
2	8	17.86%
4	8	11.10%
4	32	12.47%
8	32	10.93%
8	64	5.75%
24	64	7.85%
30	70	22.03%

Table 4.4: Delay-Insensitive VM Distribution

User Classification	Total VMs
Green	12,046
Red	35,838

cloud resources are under-utilized by VMs. For example, the average CPU utilizations of VMs with 4 cores + 8GB DRAM, 4 cores + 32GB DRAM, and 8 cores + 32GB DRAM are all around 10%. The average CPU utilization of VMs with 8 cores + 64GB DRAM and 24 cores + 64GB DRAM is much lower than 10%. This result is unsurprising considering that a previous analysis of Amazon AWS servers showed an average utilization rate between 7% and 17%[62]. Azure VMs appear to exhibit typical under-utilization concerns experienced in other cloud platforms.

VMs that are less efficiently utilized will likely waste cloud resources and increase cloud spend. Identifying such VMs and recommending them to users for downgrading or shutting down can help reduce cloud cost and waste. Based on the analysis shown in Table 4.3, any VM that displays $\geq 10\%$ for its average CPU utilization over the lifetime of that VM is denoted as green VMs, while red VMs are those with an average CPU utilization of $< 10\%$. 10% is chosen as the threshold due

to the value being low enough to justify consideration for reducing resources. Table 4.4 shows that when 10% is used as the threshold, 12,046 Delay-Insensitive VMs are categorized as green VMs and 35,848 Delay-Insensitive VMs are categorized as red VMs. It is worth noting that 10% is a very conservative threshold for these cost saving results. A higher threshold would yield more cost savings. Meanwhile, multiple thresholds can be used to support finer-grained categorization for VMs. This analysis remains simple and extremely conservative to demonstrate the validity of the approach. It can be posited that finer-grained categorization would lead to even greater cost reduction. Since the original Azure dataset does not provide detailed information about the actual deployed workloads, it is difficult to determine specific reasons why the CPU utilization rates are so low in some VMs. However, the following reasons are common and typical according to studies in the literature [73, 69, 70, 67] and our previous research [106, 116]:

- Not enough work - This is one of the most important reasons that causes low resource utilization. Many users (especially new cloud users) tend to over-provision the needed resources, but the deployed workloads are much lower in resource requirements. This is largely due to the lack of in-depth knowledge about the characteristics of their workloads. For example, UserID P43dcspmv (truncated for space) deployed 1,875 VMs with an overall average CPU utilization of merely 2.5%. Further, most of these VMs were 8-core + 32 GB deployments, resulting in a substantial waste of resources. This type of users would greatly benefit from our recommendation algorithms to significantly reduce their cost and waste in the cloud.
- Lack of parallel computing - A sequential application cannot leverage multiple virtual cores so requesting more cores for such applications would decrease overall CPU utilization. Even for parallel applications, the actual speedup is largely affected by task dependencies and the maximum speedup is limited by

the Amdahl's law.

- Improvement of hardware - As hardware continues to become faster and more efficient, utilization rates for the same workload may decrease. Some users may not be aware of this and continue to request the similar amount of resources as in the past.

VM Cloud Waste Points

Categorizing VMs as green or red certainly helps identify those inefficient VMs but it cannot quantitatively evaluate the amount of "waste" red VMs contribute to the cloud. Two red VMs that have identical CPU utilization, but run on different sizes of VMs, will waste a different amount of cloud resources. To solve this problem, Cloud Waste Points (CWP) are introduced to determine the amount of resources each red VM wastes. The CWP of a VM is derived by multiplying the lifetime of that VM by its corresponding waste factor.

Table 4.5 shows the waste factors being used for different sizes of VMs. The waste factor table is directly related to Table 4.2 in that the waste of each VM is largely decided by its price. A value of 1 is assigned to the smallest VM-type available (2 core-2GB memory), and then each VM is weighted based upon the proportion of cost increase. The assumption made here is that Azure prices VMs based on their proportional impact to the cloud. Unknown VMs are treated equally as Delay-Insensitive VMs and the same waste factors applied. Since the price of Interactive VMs is 3.33 times of the Delay-Insensitive VMs, the waste factors of Interactive VMs is also 3.33 times of the Delay-Insensitive VMs. Longer running red VMs will contribute more to resource waste than VMs that exist for only short periods of time. Meanwhile, red VMs that request more cores and DRAM space will be assigned a larger weight for the significance of its waste per unit time. Note that all green VMs were assigned a CWP of 0 as they are considered to utilize resources

Table 4.5: VM Waste Factor

Delay-Insensitive and Unknown						
Cores	Memory					
	2GB	4GB	8GB	32GB	64GB	70 GB
2	1	2.1	2.8	3.5		
4			4.4	5.8		
8				9.3	12.2	
24					23.1	
30						37.6

Interactive						
Cores	Memory					
	2GB	4GB	8GB	32GB	64GB	70 GB
2	3.3	7	9.3	11.7		
4			14.7	19.3		
8				31	40.3	
24					76.9	
30						125.2

efficiently. The CWP of all red VMs is calculated and Table 4.6 shows the CWP distribution of these VMs, which clearly demonstrates CWP can quantitatively evaluate the waste of different red VMs.

Green and Red Users

In the Azure traces, each VM is associated with a unique user. Therefore, the accumulated CWP for each user can be calculated by adding the CWP of each VM created by that user. For example, if a user creates 10 Delay-Insensitive 2-core + 4 GB VMs running for 1 hour each (waste factor of 2.1 each) and another 10 Interactive 4-core + 8 GB VMs also running for 1 hour each (waste factor of 14.7 each), and assuming all VMs are red (i.e. under-utilized), the total CWP of the user

Table 4.6: VM Cloud Waste Point Distribution

CWP	Total VMs
<1,000	1,370,850
1,000 - 2,000	39,525
2,000 - 3,000	32,090
3,000 - 4,000	12,256
4,000 - 5,000	9,026
5,000 - 6,000	8,745
6,000 - 7,000	8,114
7,000 - 8,000	238
8,000 - 9,000	507
9,000 - 10,000	240
10,000 - 20,000	8,821
>20,000	2493

would be $(10 * 2.1) + (10 * 14.7) = 168$. If another user creates the same number of VMs, but utilizes them much more efficiently, perhaps assuming only 2 red VMs of each type, their CWS would be calculated as $(2 * 2.1) + (2 * 14.7) = 33.6$, or 20% of the first user. It can be inferred that users who deploy VMs but neglect their use, or use the VM for a particular job but fail to destroy it when unneeded, will negatively impact available cloud resources and thus have a higher accumulated CWP. As Green VMs are assigned a CWP of 0, it is possible for a user who efficiently utilizes all of their created VMs to received a CWP of 0. Thus the ideal CWP is 0, with the worst possible CWP being assigned to a user with all red VMs.

To classify cloud users based on their practices in utilizing cloud resources (either efficiently or inefficiently), the Cloud Waste Indicator (CWI) metric is introduced, which is defined as the average CWP of all VMs deployed by an individual user. For instance, if a user creates 100 VMs with a total CWP of 1,000, the average CWP per VM (i.e. CWI) of this user would be 10. Since CWI can be used as a metric to measure the cloud utilization efficiency of a single user, it can be

utilized to categorize users into three different groups - the green users, the normal users, and the red users. Green users are those who utilize cloud resource efficiently (i.e. low CWI) and red users are those who have a higher CWI. Normal users are those whose average CWI lies between the CWI of red and green users.

To determine the delineation between user types, first users with less than 200 total core hours are removed. This is equivalent to a 2-core VM running for approximately 4 days, and was chosen as a reasonable cutoff point to eliminate users who do not have sufficient usage data in the cloud. This results in a total of 6,121 users (i.e. ~92% of all users). Next, the CWI of each user is calculated, then normalized using the following formula:

$$CWI_{norm} = \frac{CWI_i - \min(CWI)}{\max(CWI) - \min(CWI)}$$

where CWI_{norm} is a value between 0 and 1.

.01 is used as the CWI_{norm} threshold for green users and .05 as the CWI_{norm} threshold for red users as this is the point where actual waste becomes more substantial. In other words, a user will be classified as a green user (if CWI_{norm} is lower than .01), a red user (if CWI_{norm} is .05 or higher), or a normal user (if CWI_{norm} between .01 and .05).

This analysis revealed that among the 6,121 Azure users considered, 1,020 users (~15%) were classified as red users while 2,321 users (~36%) could be classified as green users. This result confirmed the findings from the latest Flexera cloud report [1] that a large number of cloud users are wasting and overpaying for cloud resources. In addition, it also showed that cloud users either know well or have no clue on how to utilize cloud resources efficiently as less than half fall in between (only ~45% are normal users). Figure 4.5 shows a breakdown of CWI by number of users. Users with a $CWI > .2$ and $> .1$ are those that exhibit extremely wasteful behaviors thus may suffer from a large amount of unnecessary cloud spend.

The following observations can be made about user CWI:

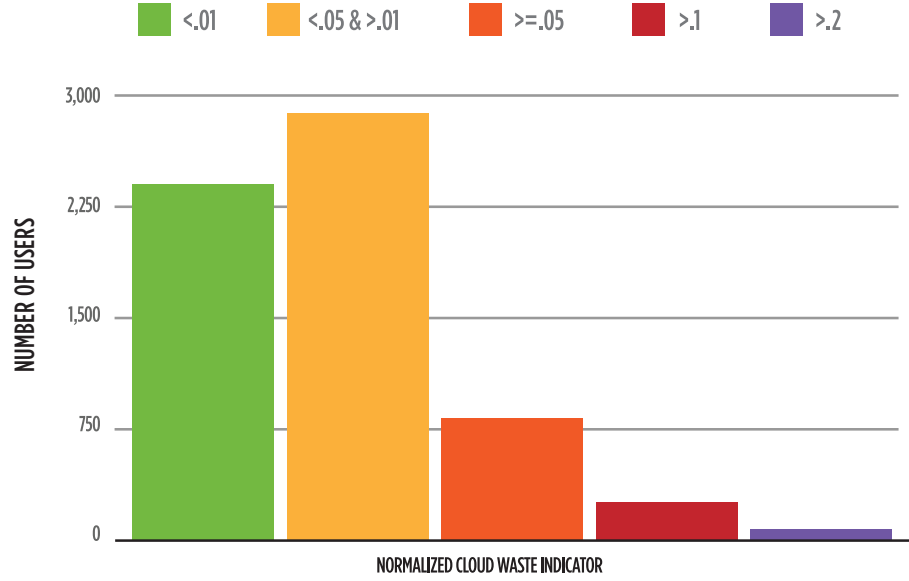


Figure 4.5: CWI Distribution by User

- 1,020 users classified as red account for 17,112 total VMs, or .6% of the entire number of VMs considered (2,691,463). However, the cloud waste of these VMs constituted almost 22% of the total cloud waste for all users, or almost 37x the expected value. Red users thus have a proportionally higher negative impact on the cloud than other users.
- The highest CWI_{norm} (1.00) was assigned to two users, both with a single 30-core VM with extremely low resource usage. Further, the 100 users with highest CWI_{norm} accounted for 1,291 total VMs, or only .04% of the total number of VMs considered, while the amount of waste produced by these users is over 5% of the total cloud waste, a factor of 125x. This indicates that a small number of users can have a dramatic negative impact on the efficiency of a cloud system.
- The 100 users with the most VMs by number (1,910,235, or 71%) all had a CWI_{norm} of <.01, i.e. they are all green users. Further, their cloud waste was approximately 9.2% of the total, an order of magnitude lower than the waste

caused by the red users.

- The user with the highest total number of VMs (366,225) and CWI_{norm} of .00001 had a similar absolute cloud waste to many users with between 20 and 100 total VMs (CWI_{norm} over .1). This provided evidence that CWI can effectively differentiate users who produce similar amounts of waste by how well a user allocates VM resources.
- Differentiating users between green and red can be useful in spotting users who create VMs that consume resources while under-utilizing those VMs. However, it is possible that red users may still create some green VMs while green users may create red VMs. Generally speaking, green users tend to utilize cloud resources more efficiently than red users.

Cloud Utilization Score

CWI is useful for classification, but may be less intuitive to individual users who are interested in how they compare to other users with respect to cloud resource usage. Thus below each user is provided with a Cloud Utilization Score (CUS), calculated by providing the percentile rank of their CWI_{norm} in comparison with other users in the cloud.

For example, a user with a Cloud Utilization Score of 99 has an extremely low CWI_{norm} , i.e. this user beats 99% of all other cloud users in cloud utilization efficiency. Similarly, a user with a Cloud Utilization Score of 10 would be a red user that only beats 10% of all other cloud users in cloud utilization efficiency. Table 4.7 displays the Cloud Utilization Score with the corresponding CWI_{norm} values.

Table 4.7: User Score by Percentile

CWI_{norm}	Cloud Utilization Score
.0000	99
.0003	90
.0040	80
.0100	70
.0160	60
.0200	50
.0230	40
.0300	30
.0400	20
.0600	10
.2100	1

Table 4.8: Cost Savings Examples

Cores	Memory	Max CPU	P95 Max CPU	Lifetime	Avg CPU	Pricing	Cost	New Pricing	New Cost	Savings
8	32	78.3%	9.3%	300	0.5%	0.333	\$800.75	0.208	\$499.20	\$301.55
4	32	78.3%	24.7%	719	3.5%	0.208	\$598.90	0.126	\$362.38	\$236.52
4	8	62.1%	20.6%	719	3.0%	0.159	\$457.81	0.099	\$284.72	\$173.09

Recommendation Algorithm and Cost Savings

With a list of red VMs identified for each user, an algorithm can be developed that makes recommendations to users on how to reduce cost and waste in the cloud. This algorithm works as follows. First, it creates a red VMs list for each user. Second, it calculates the CWP of all VMs in the list and sorts the VMs in the list in descending order. Third, it sends recommendations to that user asking to migrate down the top n (n is a parameter that can be changed by the user) VMs in the list by one level. This algorithm repeats for all users. The recommendation algorithm can run on a daily basis to analyze all currently running VMs flagged as red and make suggestions to cloud users for potential savings. However, it will never automatically change configurations of running VMs. The final decision can only be made by users who create the VMs.

To demonstrate how the recommendation algorithm can help reduce cost, three specific cases are illustrated in Table 4.8 where avgcpu is observed to be low thus a cheaper VM is recommended for cost savings. In the experiments, it was assumed that the core count can be reduced without impacting the performance of the VM significantly if the observed avgcpu is 10% or less, i.e. a red VM. When making recommendations to degrade VM configurations, the model is conservative by suggesting switching to a lower level that is right below the current level. This can help users reduce cost over the VM lifetime and the cloud system will also be utilized in a more efficient way. As 2-core + 2 GB VMs are at the cheapest possible pricing tier, their cost cannot be lowered, even if resource usage is extremely low. However, it is still possible to classify these VMs as red and report them to users, who can further investigate and decide if such under-utilized light VMs can be shutdown or consolidated with other similar light VMs.

One primary concern about using average CPU utilization as the flag for red

Table 4.9: Additional Cost Savings Examples

Cores	Memory	Max CPU	P95 Max CPU	Lifetime	Avg CPU	Pricing	Cost	New Pricing	New Cost	Savings
8	32	0.06%	0.05%	719	0.01%	0.333	\$800.75	0.208	\$499.20	\$301.55
4	8	0.32%	0.25%	719	0.09%	0.159	\$457.81	0.099	\$284.72	\$173.09
24	64	2.6%	0.09%	719	0.01%	0.832	\$14,373.57	0.437	\$2,516.53	\$11,857.04

VM is the large discrepancy between Avg CPU and Max CPU. For example, the three examples shown in Table 4.8 have very low Avg CPU utilization ($<3.5\%$) but the Max CPU usage could exceed 60%. Will the performance be affected if the VM is downgraded? Yes possibly and that is why the recommendation algorithm is conservative by only suggesting to downgrade the VM for one level at a time, which aims to minimize the impact of downgrading VM on workload performance. In addition, the algorithm only provides suggestions. Users will make the final decisions whether or not their applications can be sufficiently run with less resources at the benefit of a lower cost. Meanwhile, it could also be argued that there are many cases such decisions would be easy and straightforward to make. For example, Table 4.9 demonstrates several cases where both Max CPU and P95 Max are sufficiently low. The 24-core + 64 GB VM, for example, is an excellent candidate for migrating as it exhibits an extremely low-level of resource utilization while also being very expensive to run. Downgrading this VM to the next lower level results in a substantial 82% cost savings (i.e. \$11,857.04). The question that must be answered is why a user would choose to voluntarily move their workflow to a less capable VM. While these calculations provide compelling statistical evidence for doing so, only the user himself can make the best decision. If a user is informed they can lower their overall cost by 50% by moving a 2-core, 4-GB VM to a 2-GB VM, this recommendation might be ignored unless the impact to their workload performance is also included. A better recommendation would be "the impact is roughly a 10% decrease in performance for a 50% in cost reduction". The user can best determine if the recommendation should be adopted or not. Mission-critical red

VMs would not be good candidates for migration to a lower tier, for instance, but tasks that do not need to be completed immediately might be worth downgrading. Ideally, asking users what sorts of workloads are running on their VMs would significantly improve the quality of the recommendation algorithm. However, this often increases the complexity of the algorithm. I believe showing simple recommendations and leaving final decisions to users would be a viable solution.

To evaluate the potential savings of our recommendation algorithm, all red VMs in the 2019 Azure traces are categorized. 1,369,364 VMs can be identified as candidates for resource reduction, which could help save over \$22 million (see below for details) assuming these recommendations are all adopted by the users in question. This is a substantial reduction in cost, resulting in an overall savings of over 36%.

Original Cost: \$61,595,170.23

Total VMs: 2,695,548

VMs with Savings: 1,369,364

Percent of VMs with Savings: 51%

New Cost: \$39,341,202.17

Total Savings: \$22,253,968.06

If instead the focus shifts to the 100 most wasteful users, nearly \$9 million in cloud spend can be saved.

Original Cost: \$36,543,750.63

Total VMs: 292,938

New Cost: \$27,591,016.10

Total Savings: \$8,952,734.53

Meanwhile, it is worth noting that it can also be observed there exist situations where Delay-Insensitive VMs use vastly higher computational resources than normal. When the avgcpu is abnormally high (e.g. 80% - 90% or even higher), it

indicates those VMs are significantly under-provisioned. Suggesting a higher configuration of Delay-Insensitive VM or Interactive VM to ensure proper resources are available when needed would be valuable in these cases.

In summary, to enable the recommendation algorithm for practical use and provide reasonable cost savings evaluation, I took the following conservative actions when making recommendations and made conservative calculations when estimating savings.

- A very conservative threshold was used to flag red VMs. Only VMs that have an average CPU utilization of $<10\%$ will be identified as red VMs.
- It is recommend to downgrade a red VM by one level to minimize the impact of downgrading VM on workload performance. Even though more aggressive downgrading could be recommended for larger savings for the examples shown in Table 4.9, only the conservative downgrading is used when calculating savings.
- As I have no information about the use-cases for each VM, nor the nature of their workloads, the actual migration decision is left to the user. Migration does not occur automatically as to not disrupt potentially mission critical workloads that might need more resources than are initially indicated.
- Cost savings is calculated specifically to inform the user they might be able to save money by lowering their resource usage. Again, it is up the individual users to decide if this savings is worth potentially impacting their workloads. Even if only a portion of users (ex. 50%) follow these recommendations, the savings is still substantial. As can be seen from the example of only the 100 most wasteful users above, the algorithm can achieve significant savings (approximately \$9 million) even when targeting a much smaller group of users.

Why Should Cloud Providers Care?

It is obvious that CWP, CWI and CUS can increase the awareness of Azure cloud users about their cloud waste and cost. The recommendation algorithm explained above can help them reduce the cloud waste and cost as well. However, why would cloud providers (e.g. Microsoft Azure) want to give users the option to use lower-tiered VMs, cutting into their revenue? I argue that cost is one of the most important factors when users select one cloud platform over another, if not the most important. When two cloud providers offer similar services, the one with a lower cost is chosen for the most part. Reducing costs while still offering competitive products will almost always grow market share by increasing not only the resources used by individual users (users paying less per VM will be inclined to move more of their infrastructure to the cloud), but also by growing the userbase. Having a reputation for a better user experience is extremely effective marketing, and more users equates to more revenue. Further, the greater utilization of resources on the platform actually improves the ability of the cloud provider to adjudicate provisioning and determine possible QoS issues, or discover areas for growth. Volume of sales will always outpace smaller margins for enterprise-level industries, so the economic upside for Azure is substantial. Last but not the least, sustainability has become the priority of all primary cloud providers. By helping users reduce waste in their cloud infrastructure, cloud providers will reduce the overall carbon emissions to the environment thereby partially fulfilling their sustainability goals.

Limitations

There are several limitations in this work. First, the lacking information regarding the nature of jobs and applications running on each VM could affect the quality of the recommendations.

Another issue contained within the trace files was lack of information about memory usage of each VM. While the total memory bucket assigned to each VM is available, there was no way to discern the actual memory utilization rate. Per Karamagi et al [117], the average memory utilization of VMs within traces provided by Google is under 60% on average. Further per Azure, swap space can be manually configured for VMs but is not active upon initial creation [118]. Thus, I extrapolate the over-provisioning of memory to be a substantial issue left to be explored in the future.

In addition, several assumptions were made about the Azure pricing model and VM types. For example, it was assumed all Unknown VMs are Delay-Insensitive for calculating CWP, and also a waste factor of 3.33x for the impact of Interactive VMs. It is probable that Unknown VMs may actually be Interactive in nature, but by being conservative and using the lower waste factor ensures that CWI is not overcalculated for users.

Lastly, an assumption is made that all users accept the recommendations to downgrade their underutilized VMs. For a variety of reasons, users may still decide to keep their existing VMs as-is, and thus the cost savings outlined above would similarly be decreased. I attempt to mitigate this issue specifically by selecting a low CPU utilization threshold (10%) when identifying red VMs for downgrading.

Conclusions and Future Work

The rapid growth of cloud computing has significantly increased the expenditures on cloud services. The 2021 Flexera State of the Cloud Report [1] surveyed 750 enterprises, which estimated their public cloud spend was over budget by an average of 24% and expected their cloud spend to further increase by 39% next year. Reducing cloud cost and waste has become the first priority of many enterprises conducting businesses on cloud. Nevertheless, there is a clear gap in the

literature to study the user behaviors in the cloud and provide viable solution to reduce cloud cost and waste. In this study, I strive to fill the urgently needed gap by conducting a comprehensive analysis of the Microsoft Azure 2019 traces, which reveals that a large portion of VMs are under-utilized or over-provisioned for resources (i.e. red VMs) and they are wasting significantly more resources than green VMs. To help mitigate the cloud waste problem and save cost, we classify cloud users as red, green and normal users and rank their cloud usage efficiency using the Cloud Utilization Score (CUS). Additionally, I propose an algorithm to identify red VMs and recommend VMs with lower configurations that can reduce cost without compromising quality of service (QoS). The experimental results show that over \$22 million savings can be achieved if these recommendations are properly adopted by cloud users.

Despite the substantial amount of possible cost savings, this analysis is limited by several assumptions such as the lack of information about the characteristics of cloud applications and their memory usage. Investigations on the Azure Functions Trace [119] could be helpful by analyzing specific applications running within Azure Functions, a serverless facility as outlined in the aforementioned FaaS paper [109]. It is expected that information gleaned from these traces will provide a better understanding of the resources required by applications running within VM deployments, which can help improve the quality of these recommendations.

V. AI WORKLOADS - FOUNDATION MODELS

Introduction

Artificial intelligence (AI) workloads are exceptionally power-hungry and becoming more ubiquitous. According to Labbe[120], the largest data centers require more than 100 megawatts of capacity, which is enough to power roughly 80,000 U.S. households and much of the growth was demanded on running AI models. Further, these workloads differ from general web and cloud workloads due to their unique nature. Per Susskind[121], while we can observe the output of an AI model, it is difficult to see "inside" these models and gain any insight as to where or how the power consumption occurs. Oddly, this is an old problem, with such implementations even being used by home computers such as the Commodore 64[122][123].

A significant area of interest in mitigating energy usage is the concept of Red vs. Green AI. Red AI is vitally concerned with being as accurate as possible, disregarding cost to achieve the best results. Conversely, Green AI considers the computational cost of output, and attempts to mitigate resource usage and reduce carbon impact. In 2021, Stanford University released a report outlining steps to help migrate to greener AI, stating that carbon impact can and should be reduced in most cases. The first step is using more efficient models and hardware, or training models in low-carbon intensity regions[3]. If further reduction is no longer possible, a cost-benefit analysis should be performed to determine if a larger AI model is appropriate to deploy over a smaller, more efficient one. Figure 5.1 highlights a typical cost-benefit analysis of an AI model. The Total Value of the model is a product of multiple factors, including the positive social and environmental benefits provided by the model, as well as adverse elements such as energy costs, carbon emissions, and societal harms. This score can then be used to determine if a Red

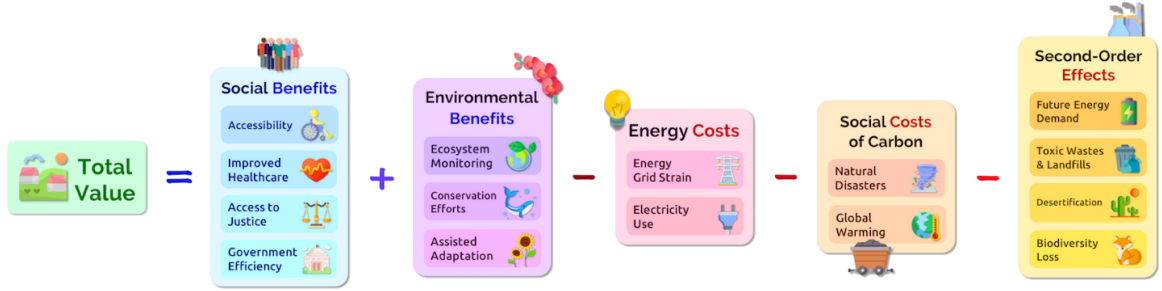


Figure 5.1: Cost-Benefit Analysis of AI Models [3]

model is valuable enough to deploy in lieu of a Green one.

Knowing if a model is Red or Green initially before doing such an analysis is important, and thus differentiating models is required. I focused on foundation models in this dissertation specifically because they require large amounts of energy to train and run, and are well-suited for classification into Green and Red AI. A foundation model is any AI model that is trained on a broad set of data, at a large scale, and can be adapted for more general use as needed. In other words, these are AI models that can perform a wide variety of useful tasks but are prone to excessive carbon emissions if improperly implemented. Typically, training a foundation model utilizes massive amounts of energy. For example, training the GPT-3 model consumed approximately 190,000 kWh of energy and produced 85,000 kg of CO₂[12]. However, as foundation models are applicable to many disparate domains, each implementation may use differing amounts of energy. Determining energy consumption of models is dependent on the type of output produced, and may be unrelated to the training costs. Thus, this study focused on the energy consumption and carbon emissions of foundation models at the inference stage, not during training. By focusing specifically on workload analysis of foundation models, it is possible for a direct comparison between models to be made.

Currently, there is no existing methodology to study the energy requirements and environmental impact of foundation models. This dissertation is an example of

how such classification can be done by selecting a specific domain, in this case Natural Language Processing (NLP), then measuring the energy consumption when running foundation models to produce output. This output was then analyzed using a rubric for evaluating language, the State of Texas Assessments of Academic Readiness (STAAR), which allowed a qualitative measurement in this domain. Further, Software Carbon Intensity (SCI), an industry standard metric released by the Green Software Foundation in 2021, was calculated to compare and evaluate each foundation model based on their carbon emissions. SCI provides users with the carbon impact of their software, allowing them to make better choices for future implementations and mitigate carbon waste. To the best of my knowledge, this study is the first work utilizing SCI to evaluate the carbon impact of foundation models at the inference stage, which makes the following contributions.

1. Utilizing the Software Carbon Intensity (SCI) metric provided by the Green Software Foundation, this study showed that the environmental impact of foundation models can be quantitatively measured and compared.
2. It showed the quality of a foundation model can be measured by objectively analyzing the output. For this study, the writing rubric supplied by the State of Texas (STAAR) was utilized to calculate the Writing Performance Score of foundation models, qualitatively scoring those outputs on a scale between 3 and 12.
3. It showed it is possible for a foundation model to be replaced by a more efficient model, mitigating energy requirements and reducing carbon waste while maintaining a similar level of expected output.
4. It demonstrated that there is no significant difference in the quality of output of foundation models when using either CPU-only or GPU-acceleration. It

further showed that GPU-acceleration can provide up to a 91% reduction in overall energy consumption when employing foundation models.

Related Work

There is a lack of in-depth studies specifically targeting energy usage by AI deployments. While there are numerous research articles regarding energy expenditure during training various models, how much energy those models use out "in the wild" is difficult to find. The most substantially useful rationale for tracking energy consumption of AI is found in Henderson et al. as regards environmental impact[3]. The researchers make the case that carbon impact from foundation models can, and should, be mitigated and offer several solutions during training, including selecting energy grids with minimal carbon emissions, using more efficient hardware to train models, using more efficient models, and distilling models to be more applicable. The authors further suggest that carbon offsets cannot be relied upon to mitigate impact, and stress assessing energy consumption before the use of models.

A study by Varghese et al. analyzes the performance of edge computing systems which aids in benchmarking for app development, resource management, and adaptive decision-making [124]. Three of the edge performance benchmarks consider AI platforms directly, thus giving a starting point for identifying AI capabilities and workloads.

While information about energy usage of existing AI models is notably sparse, there is information about energy expended during the initial training. Patterson offered insight on methods to improve energy efficiency and carbon emission reduction for model training, including geographically locating to optimize carbon output, utilizing cloud providers as they may be more energy efficient, and being explicit about power consumption to aid end-users in selecting appropriate models

to perform tasks[125]. The energy consumption of several models are determined, giving insight on applying this methodology to running pre-trained models in a production environment. Similarly, Anthony provided carbontracker as a means to track and predict the energy usage and carbon emissions of training deep learning models[126]. This tool provided users of these models insight into their environmental impact during deployment. Gao et al. offer a method for modeling and characterizing AI and big data workloads called Data Motifs[127]. Eight motifs are identified, which is useful for determining the types of expected computational requirements for AI workloads. A publicly available benchmarking tool called BigDataBench is also available. Do et al. provided a low-power solution directly applicable to AI workloads [128]. The work introduces Newport, a commodity SSD that can be utilized by server operating systems. AI applications are run on Newport, resulting in significantly reduced energy requirements. Labbe explains in detail the huge negative environmental impact incurred when training large AI models[120]. He provided MegatronLM as an example, showing that training this model used as much energy as three homes during a calendar year. He also points out the rapid growth of datacenters and their extensive energy consumption, while it being "far from the case" that these datacenters are green. Toews gives insight as to the growing problem of AI training by demonstrating the difference between GPT-2 and GPT-3[129]. Where the former took a dozen petaflop days to train, the latter required several thousand, a massive increase. He makes the case that the problem will continue to get worse due to the greater reliance on AI models. DeWeerdts makes a similar case, using carbontracker (above) and recommends such tools as necessary to track energy usage of deployed AI models.[130]

AI Workload Selection

The first problem faced when comparing foundation models against each other in an effort to classify them as Red and Green is selecting something that can be easily understood without extensive expert knowledge. For example, if we were to compare various chess playing programs against each other, it might be difficult, if not impossible, to determine which results of two outputs were better unless we had these AI programs playing against Grand Masters, much like Deep Blue's performance was evaluated against Garry Kasparov in 1996[131]. The abundance of chess Grand Masters is rare, similarly with experts in all sorts of AI-related tasks. While it is mathematically possible to determine which chess move is better, this requires essentially using a vast amount of computing resources to compare those two moves. Hence, we must select something that lends itself to easier "normal human" analysis, i.e. something that is not only more intuitive, but which can be evaluated by a non-expert.

Natural Language Processing

Given the above concerns about the required analysis, I selected Natural Language Processing (NLP) for this study for three reasons:

1. Native speakers of a language will generally notice something is "off" about a particular sentence or response. While it may be difficult to determine exactly what is wrong with a sentence initially, the general feeling is that it does not adhere to the expected rules of language use, perhaps indicating generation by a non-native speaker or an AI. This allows the typical researcher to more easily segregate initial outputs before a more lengthy in-depth analysis takes place.
2. There are a multitude of rubrics available to evaluate the English language, and can be applied to concretely classify statements as "good" or "bad"

grammar. This allows a more comprehensive analysis after the initial phase of selecting coherent statements.

3. Recently, many NLP implementations of foundation models have been made directly accessible and can be run on physical, commodity hardware. This allows accurate power measurement that might not be possible when using such models in the cloud or via remote systems, and also provides direct access to ensure the models are working as intended. Further, these models can be run in a reasonable amount of time, allowing multiple models to be evaluated against each other without requiring overt scheduling concerns.

Foundation Model Selection

The first task for the evaluation of NLP implementations is to, naturally, select some foundation models to compare. The two most widely known foundation models used to perform NLP are GPT-2 and GPT-3, both made available by OpenAI in 2019 and 2020, respectively[132]. Unfortunately, GPT-3 is accessible only via API at this time[133], and thus is a poor candidate for direct power measurement. However, EleutherAI[134] has made available a number of GPT-3 clones, all usable on a local system. Further, they also released GPT-J 6B which is intended to supplant GPT-3 as an open-source version of that model, and can be utilized wherever desired. Thus the following five foundation models were selected using the main criteria of direct accessibility:

- **GPT-J 6B** - The aforementioned GPT-J 6B is a 6 billion parameter model that, as previously stated, attempts to replicate GPT-3's capabilities while remaining open-source[135]. It was trained on the Pile, a 825 GB open-source language modelling data set consisting of 22 smaller sets combined[136]. Per Naik, GPT-J 6B compares very favorably to GPT-3, and could be further fine-tuned to produce even better outputs[137].

- **GPT-Neo 2.7B, GPT-Neo 1.3B, GPT-Neo 125M** - Like GPT-J 6B, GPT-Neo is family of models that were designed as implementations of GPT-3, also trained on the Pile[138]. The values after the family name refer to the number of parameters utilized in each model; 2.7 billion, 1.3 billion, and 125 million, respectively. As with GPT-J 6B, GPT-Neo compares favorably to other GPT-3 replications, but the authors offer this caveat: "GPT-Neo was trained on the Pile, a dataset known to contain profanity, lewd, and otherwise abrasive language. Depending on your usecase GPT-Neo may produce socially unacceptable text..and offensive content may occur without warning."
- **GPT-2** - First made available in 2019, GPT-2 is a 1.5 billion parameter transformer model that produces excellent natural language results[139]. It is included in this study due to it being available in the HuggingFace repository[140], and also to directly compare an older foundation model against newer implementations.

Together these models provide a good starting point in which to classify NLP implementations given their wide range of parameters.

NLP Prompt Selection

After selecting the domain to evaluate, and models in that domain, the final step to evaluate the workloads was choosing some prompts that initiated AI response. That is, I came up with a list of 10 questions or sentence starters that the models would use as a basis to produce an output. This sort of "ask a question, get a response" approach to evaluating AI, and NLP specifically, has been around for over 50 years, with Joseph Weizenbaum releasing the ubiquitous ELIZA software in 1966[141]. While all of the prompts are open-ended, to at least some degree, they vary in what sorts of answers are to be expected. For instance, if I ask a person how their day is going, I can expect a wide variety of valid responses; the AI model

would thus, too, be expected to produce answers that might wildly differ from each other. However, when asking something more concrete such as "what is an elephant", a valid response might indeed include a story about how elephants murdered a family member during an African safari, but a more reasonable reply would probably list physical attributes and expected habitat. Hence, the prompts have an inherent bias for response type (direct answer vs. creative exposition), but also allow outputs outside the norm. I selected prompts that seemed interesting, and also some that might be difficult for an AI model to parse and produce coherent responses. The prompts are as follows:

1. What is an elephant?
2. What is a lemur?
3. How are you doing?
4. Guess my weight
5. Can you dance?
6. Good day to you, sir
7. What is a computer?
8. Who is Elon Musk?
9. Are you alive?
10. Mad Max is a

Testing Environment

To directly compare each model, it was first necessary to setup an environment that allowed for all models to be run under identical circumstances, as well as

monitor power usage directly. What follows is an explanation of how that environment was configured to ensure direct comparison.

Physical Platform

Access to an existing server was made available through a Texas State University lab, already set up for AI research. The server contains an AMD Ryzen Threadripper 2950x processor (16 physical cores with hyperthreading support for 32 threads), 4 Nvidia RTX 2080TI GPUs, and 128GB of DDR4 Memory in quad channel configuration[142]. The server is configured to record power usage by both CPU and GPU making it an excellent platform to conduct this study.

Python Environment

While the foundation models are made readily available via the aforementioned HuggingFace repository, directly testing those models requires writing a script to interface with them and capture output. Python is a widely used scripting language that has many built-in features for providing such an interface, while utilizing minimal code due to numerous libraries written specifically to perform a wide range of tasks. To configure the environment, miniconda was first installed via command line on the testing server[143]. Miniconda allows a private environment to be configured, segregating versions of Python, libraries, code, etc. As the server was in use by other researchers with specific needs, it was necessary to create this virtual Python installation. After miniconda was installed, both Pytorch[144] and Transformers[145] were installed and configured to allow direct interface to the HuggingFace repository via Python scripts. Additionally, to enable access to the NVIDIA GPUs on the server, CUDA was installed[146][147]. A sample script was created to ensure access to the models was working, and also that GPU acceleration was enabled. To complete the setup, the foundation models were downloaded to local

Table 5.1: Foundation Model Size Comparison

Model	Disk Usage
GPT-J 6B	22.5 GB
GPT-Neo 2.7B	9.5 GB
GPT-Neo 1.3B	4.6 GB
GPT-Neo 125M	502 MB
GPT-2	2.2 GB

storage as direct power measurement was necessary. Transformers allows creating a direct interface to a wide variety of AI models available at HuggingFace, either via API or locally stored; the second option was selected for the aforementioned reason. The models consumed approximately 39 GB of space, which is considerable as these were pre-trained. See Table 5.1 for a breakdown of model size.

Power Measurement

A custom CPU and GPU power monitoring utility called AMD RyzenPower was utilized to record energy usage during testing. This utility monitors the actual power utilized by both CPU and GPU, independently, in real-time, allowing accurate measurements of the total amount of energy required to produce a response to each prompt provided to the foundation models. As overhead for the utility is extremely low, the impact on power monitoring was negligible; further, any extra power requirements for the utility were identical for all tests and thus can be assumed in the normal energy requirements of operating the system.

Scripting

In order to test the models, a Python script utilizing the Pytorch and Transformers libraries was written. This script is straightforward: a model is selected, with a choice between a CPU-only run, or GPU acceleration enabled. The

script then initiates AMDRyzenPower to record power usage, logging that usage to CSV files which detail energy used in 1 second increments with CPU and GPU information being recorded separately. The script then loads the model from the local source and provides each prompt sequentially, along with two parameters: `max_length` and `temperature`, recording the output. **max_length** refers to the expected word count for the output and was set to 100 to allow significantly interesting results without taking an excessive amount of time. Temperature is explained below. Each prompt is evaluated by the model 3 times, creating 30 total entries per run. This output is then written to a file for later review. Finally the script stops the power recording and terminates.

The parameter **temperature** needs some further explanation[148]. The best description of this parameter is how "open-ended" the expected answer should be. Temperature 0 is akin to a simple text search: more common words (or tokens) can be expected, and the same output occurs nearly every time. Temperature 1, however, is much more random, often producing extremely strange results. A temperature of 0.9 was selected for purposes of this study, as this results in more consistency and coherence, while ensuring interesting output, i.e. more human-like language.

AI Workload Analysis

After configuring the test environment, the following steps were performed to generate and test the workloads:

1. Select foundation model
2. Choose CPU or GPU target
3. Run Python script with selected options

4. Match timestamps in CSV files with prompts to determine energy usage per prompt

The Python script was thus executed five total times under two separate scenarios:

1. CPU-only model evaluation
2. GPU-accelerated evaluation

This resulted in ten total runs, albeit with some failures. Neither GPT-J 6B nor GPT-Neo 2.7B successfully ran using GPU acceleration. After some investigation, it appears that these models were too large to load into memory with the GPUs currently installed on the test system. Hence, 240 total outputs were produced, resulting in approximately 106KB of total text.

Output Evaluation

Two evaluations of each model were done in order to segregate them appropriately. While there is an obvious numeric analysis that can be performed (comparing the total energy taken to produce each output), evaluating the output for coherency and applicability to the provided prompt requires a separate methodology explained below.

Power

Table 5.2 shows the average total power used by each output produced by the respective models. The results are broken down thusly:

- Time -The average amount of time in seconds taken for a prompt to be evaluated and an output generated. In the case of GPT-2, for instance, each prompt took roughly 27.5 seconds to be evaluated before the model generated and recorded a reply.

Table 5.2: Average Foundation Model Power Usage - CPU Only

Model	Time (Seconds)	Power (Watts)	Total Energy (Joules)
GPT-J 6B	139.1	48.9	6802.9
GPT-Neo 2.7B	68.6	51.8	3553.5
GPT-Neo 1.3B	35.7	54.2	1934.9
GPT-Neo 125M	7.8	50.3	392.3
GPT-2	27.5	61.7	1696.8

- Power - Average recorded power in watts utilized by the CPU over the total time each prompt was evaluated until terminating in an output.
- Total Energy - The amount of joules consumed by the CPU during the entire prompt evaluation process. This is calculated by the formula

$$Watts * Seconds = Joules.$$

From the results it can easily be seen that GPT-J 6B is nearly twice as power-hungry per run than the next closest model (GPT-Neo 2.7B), using over 6,800 joules of energy to generate a single response. Further, GPT-J 6B takes nearly 18x as long to complete a single output as compared to GPT-Neo 125M. The power consumed per-second by the models is fairly consistent, with a single exception. Interestingly, GPT-2 used $\sim 13\%$ more power per-second than the next closest model (GPT-Neo 1.3B), albeit taking less time, resulting in roughly 12% less total joules consumed. As would be expected, the number of parameters is directly related to the amount of total energy consumed by each model. The most direct comparison would be between GPT-Neo 1.3B and GPT-2 with 1.5 billion parameters; it would appear in this case that GPT-2 is somewhat more efficient than GPT-Neo 1.3B

Looking at table 5.3, a massive increase in speed and efficiency when employing a GPU is immediately obvious. As with the CPU Only results, the total amount of joules is calculated by in a similar manner, but segregates the power consumed by CPU and GPU for sake of clarity. While the CPU consumes between 15%-30% less

Table 5.3: Average Foundation Model Power Usage - GPU Accelerated

Model	Time (Seconds)	CPU Power (Watts)	GPU Power (Watts)	Total Energy (Joules)
GPT-Neo 1.3B	2.1	42.6	36.8	166.7
GPT-Neo 125M	1.0	42.9	21.7	64.6
GPT-2	2.4	43.2	30.5	176.9

power per-second, the GPU itself utilizes a considerable amount, resulting in roughly 25% more total power per-second on average when compared to the CPU Only results. However, the time taken to produce each output is drastically reduced; in the case of GPT-Neo 1.3B we see a 17x increase in speed. This results in a 91% reduction in overall power used, demonstrating the sheer efficiency of GPUs when doing any sort of work with AI models. Contrasting GPT-Neo 1.3B and GPT-2 again, the former has an edge this time, being slightly faster and more power-efficient. It can be speculated that this is due to newer AI models being better optimized when targeting GPUs to produce output. The similar power requirements between GPT-Neo 1.3B and GPT-2 would seemingly make selecting either model a toss-up for NLP; this of course assumes both models produce a similar quality of results. Hence the necessity of evaluating the actual outputs for comparison as outlined below.

Writing Performance

As previously stated, the quality of NLP outputs is as important as their energy efficiency; in fact, in some cases it might be even more important, depending on the circumstances; an expert system utilizing NLP that guides a doctor through a complex medical procedure would need to be much more accurate than a diversion like the aforementioned ELIZA. Thus, further evaluation beyond power usage is required. Per the Program for the International Assessment of Adult Competencies (PIAAC) in a 2012 study, the average adult in the United States reads at approximately an 8th grade level[149][150]. As this work is concerned with NLP

outputs of a more general nature, I decided to target the 8th reading level as a baseline for output evaluation. In the event that NLP models designed to assist the doctor above were being assessed, that might require a medical expert to accurately judge, undermining the entire point of this work being possible for the "average researcher". That said, rubrics for the scoring of the English language exist, and can be applied by anyone sufficiently skilled in reading and writing both normal prose and technical documentation, i.e. the "average researcher" AKA a PhD student working on finishing their dissertation. In 2012, the Texas Education Agency adopted The State of Texas Assessments of Academic Readiness, commonly called the STAAR[151]. As part of this adoption, evaluative measures were made available to all Texas teachers to ensure a broadly applicable model could assess various skills expected of students within the public school system, including reading and writing. Via STAAR, a rubric[152] and scoring guide[153] are available for English I, a class taken by all freshman-level (9th grade) students in Texas. STAAR rubrics are not available to assess 8th grade reading and writing, hence the selection of this rubric instead. However, as the STAAR tests are administered after the Fall semester (halfway through the 9th grade school year), I decided this was the best possible choice to judge the model outputs as could be done given the circumstances.

The rubric is divided into three portions:

- Organization/Progression - the clarity of organization, focus, and responsiveness of the output
- Development of Ideas - the output contains details and examples appropriate to the prompt, also how engaging and interesting it is to read
- Use of Language/Conventions - precision of word choice, clear, correct grammar, and purposeful sentence structure

For each output, the portions of the rubric are evaluated separately and

Table 5.4: Average Foundation Model Writing Performance - CPU Only

Model	Writing Performance Score
GPT-J 6B	5.83
GPT-Neo 2.7B	6.63
GPT-Neo 1.3B	6.17
GPT-Neo 125M	3.57
GPT-2	5.83

assigned a numeric score between 1 and 4. These three portions are then added together, resulting in a total score between 3 and 12, with 12 being the highest possible score, i.e., the output is an example of superior writing. 3 denotes a nonsensical output, with poor grammar, little acknowledgement of the prompt, and lack of overall coherence. To eliminate as much bias as possible, all outputs were scored randomly, i.e., selected and scored with no indicator as to the model employed.

All 240 outputs were scored using the aforementioned rubric, and the results tallied and averaged. See tables 5.4 and 5.5 for scoring results. The average writing performance for each model is denoted, separated between CPU Only and GPU Accelerated results, as with power usage above. See Appendix A for each individual output with scoring. It goes without saying that scoring the outputs was rather amusing at times.

Comparing the CPU Only and GPU Accelerated results, it is immediately apparent that there is no statistical significance between the quality of output generated by either method. GPT-Neo 1.3B demonstrates a bit of drop off when utilizing the GPU, but this is well within the bounds of what would be expected between groups of output. Thus, when possible, it would make sense to always use the GPU if speed and power savings are valued; there is no downside to output quality when employing the GPU. However, a very significant difference can be

Table 5.5: Average Foundation Model Writing Performance - GPU Accelerated

Model	Writing Performance Score
GPT-Neo 1.3B	5.87
GPT-Neo 125M	3.5
GPT-2	5.83

seen in the score differential amongst the more power-hungry models. As shown in table 5.2, GPT-J 6B uses vastly more energy during the production of each output, but performs roughly on par with GPT-Neo 2.7B, GPT-Neo 1.3B, and GPT-2. GPT-Neo 2.7B is actually the best performer, at least for quality of output, with an almost 12% better overall score when compared to GPT-J 6B, while using only 52% as much power. There could be many reasons for this disparity, including the quality of the prompt selected, researcher bias when reading the outputs (perhaps preferring one sort of writing style over another), or pure randomness. It is entirely possible that some models are simply better at certain sorts of outputs than others, and GPT-Neo 2.7B just happens to be tailored to produce better results for the specific prompts utilized in this study.

Interestingly, GPT-Neo 1.3B and GPT-2 are very similar in writing performance, which aligns closely with their very similar energy consumption. When comparing these two models to GPT-Neo 2.7B, there is roughly a 2X energy increase for the at-best aforementioned 12% better score, and thus an important decision must be made: how valuable is that 12% increase in quality?

Finally, it must be mentioned that judging the actual quality of output from the various foundation models revealed the difficulty in reproducing natural language within a synthetic environment. Even though strict rules for English exist, there are countless instances where those rules might be broken for sake of clarity, emphasis, or even stylistic choice. What separates good English from bad English might be a matter of taste, and thus the criteria for tailoring output changes from

domain to domain. Indeed, the "mere" 12% increase in quality of output might be extremely significant if an expert system is needed vs. an online chatbot that can be expected to spout gibberish in some instances.

Software Carbon Intensity

The Software Carbon Intensity (SCI) specification, a new metric provided by the Green Software Foundation, describes a methodology for calculating the rate of carbon emissions for a software system. Its purpose is to inform users and developers about the possible carbon impact of their tools, approaches, services, and architectures, and assist with making better choices they use in the future[154]. Per the specification, SCI is a rate of carbon emissions per one unit of R , and is represented by the following equation:

$$SCI = ((E * I) + M) \text{ per } R$$

Where:

- E = Energy consumed by a software system
- I = Location-based marginal carbon emissions
- M = Embodied emissions of a software system
- R = Functional unit (e.g. carbon per additional user, API-call, ML job, etc)

R in this case refers to each individual run of the AI, i.e. evaluating a single prompt and producing an output. E per each R has already been calculated in tables 5.2 and 5.3. As the test server is physically located in San Marcos, TX, power is supplied by Electric Reliability Council of Texas (ERCOT). Per the 2019 Grid Electricity Emissions Factors report, systems utilizing the ERCOT power grid generate 0.4784 kgCO₂e per kWh[155], and is assigned as the value of I . M is calculated as follows:

$$M = TE * (TR/EL) * (RR/TR)$$

Where:

- TE = Total Embodied Emissions, the sum of LCA emissions for all hardware components
- TR = Time Reserved, the length of time the hardware is reserved for use by the software
- EL = Expected Lifespan, the anticipated time that the equipment will be installed
- RR = Resources Reserved, the number of resources reserved for use by the software
- TR = Total Resources, the total number of resources available

TE is somewhat difficult to calculate, and thus I used the Boavizta dataset to estimate the TE for the test server utilized in these experiments[156][157]. Per AMD, the Ryzen 2950x processor has a TDP of 180W[158]. The closest system in the Boavizta database with a similar processor TDP has a value of is a Dell PowerEdge R540 with a processor TDP of 165W. As the PowerEdge is a very similar high-end system, this is a good sample for approximating embodied emissions. gwp_total for the PowerEdge is 8230 kgCO₂e, and thus this value is used for TE in these calculations. Due to the lack of information available on carbon emissions during the GPU manufacturing process, TE (and subsequently SCI) is calculated under the assumption that the server only utilized the CPU when running models.

TR is selected as 3 years due to the length of actual usefulness of the system; this is a typical length of time a system of this sort is deployed before replacement is necessary to keep up with advancements in technology. EL , however, is set at 5 years as this is the average length of time a server is expected to be deployed before

Table 5.6: SCI for Server Lifetime - CPU Only

	GPT-J 6B	GPT-Neo 2.7B	GPT-Neo1.3B	GPT-Neo 125M	GPT-2
Total kWh	1285.26	1361.31	1424.35	1321.75	722.72
Total Runs	680,143	1,379,125	2,650,084	12,129,230	3,440,290
Total Emissions	5552.87	5589.25	5619.41	5570.32	5283.75
SCI (gCO₂e)	8.16	4.05	2.12	0.46	3.45
90%	7.348	3.648	1.909	0.414	3.102
80%	6.532	3.243	1.697	0.368	2.757
70%	5.716	2.838	1.485	0.322	2.413
60%	4.900	2.433	1.274	0.277	2.069
50%	4.084	2.028	1.062	0.231	1.725
40%	3.268	1.623	0.850	0.186	1.380
30%	2.452	1.218	0.638	0.140	1.036
20%	1.635	0.813	0.427	0.094	0.692
10%	0.819	0.408	0.215	0.049	0.347
0%	0.003	0.003	0.003	0.003	0.003

replacement. Essentially it is assumed that although a system of this sort can be expected to perform without hardware failure for 5 years, after 3 years the hardware will be insufficient to run contemporary jobs.

RR and TR are both set to 1. When running the AI jobs, the entire system is utilized, hence both reserved and total resources are fully allocated to those jobs.

Using the values above, the value of M would be:

$$M = 8230kgCO_2e * (3/5) * (1/1) = 4938kgCO_2e$$

For purposes of this calculation, joules of energy must be converted to kWh, with 3,600,000 joules equal to a kWh. Using GPT-J 6B, the total energy per model run would thus be $6802.9/3600000 = 0.00189$ kWh. Further, we assume the AI model runs for the entire lifespan of the server (3 years) and thus we convert the time for each run into the total number of AI model runs over that time span: 3 years * 31,536,000 seconds/years / 139.1 seconds = 680,143.78 total runs. Hence, the total amount of energy utilized by GPT-J 6B over the lifespan of the server would be $680,143.78 * 0.00189$ kWh = 1285.47 kWh. SCI for GPT-J 6B is then calculated thusly:

$$TotalEmissions = (1285.47kWh * 0.4784kgCO_2e/kWh) + 4938kgCO_2e = 5552.97kgCO_2e$$

SCI is per run of each instance of the GPT-J 6B model and would be calculated simply as $5552.97/680,143.78 = 0.0082 \text{ kgCO}_2\text{e}$, or $8.2 \text{ gCO}_2\text{e}$.

Running the model 100% of the time may not be reasonable as there are definitive periods of idle time between experiment setup, data collation, etc. Table 5.6 displays SCI at various loads for all five models employing the CPU-only. 90% in the table means that the model in question ran for 90% of the time, with 10% "downtime". 0% simply means the server has no running jobs, i.e. is completely idle using an average observed energy expenditure of 23.23W. Total Runs specifies the number of times the model was run during the time period at 100% usage.

Looking at table 5.6, it is apparent that the carbon expenditure during creation of the server itself has a massive impact on the overall SCI value. This indicates that selecting the device to run an AI model is extremely important when considering carbon waste. Further, while GPT-J 6B is the most power hungry model for each run, the fact that the smaller GPT-Neo 2.7B and 125M can run significantly more times during the server deployment means those models actually use more power and produce more carbon waste overall if run continuously. It is also significant that the older GPT-2 model has less of an impact on carbon waste, overall, than some of the newer AI models. Thus, SCI can definitely be a determining factor during the selection of a model if carbon waste is a concern.

Conclusions and Future Work

The following conclusions can be made using the research performed above:

The environmental impact of foundation models can be quantitatively measured and compared.

Per table 5.6 the SCI for each model can be calculated and compared to other

models, providing the power and carbon emissions those models will produce over time. GPT-J 6B, for instance, produces nearly 100% more carbon emissions than GPT-Neo 2.7B each time the models produce an output, demonstrating a significant difference in environmental impact. However, it must be noted that, as stated previously, employing GPT-J 6B may be a valid choice over other models if its value for a particular deployment outweighs its negative environment effects. SCI is thus a useful metric and not an absolute indicator of the validity of using a foundation model.

The quality of a foundation model can be found by objectively analyzing the output.

If an objective means can be employed to evaluate outputs, then indeed the quality of a model can be determined, the quality in this study being the Writing Performance Score, with values between 3 and 12. The NLP implementations were somewhat similar in quality of output, albeit GPT-Neo 125M being an exceptionally poor performer. However, their overall quality in generating outputs was still marginally above a mediocre 9th grade English student. This in itself provides some excellent insight in ways to improve the models, perhaps by further analyzing the outputs and tailoring the model to target specific weak spots in the language generation. Applicability to other domains should be apparent. For instance, showing AI generated faces to graduate students and asking them to determine which face looks "more human" would be an example of a metric that could be employed to evaluate an image processing model. The students might be asked to score each face between 1 (a crude circle with dots) to 10 (appears to be an original photograph). Ultimately, whatever metric is chosen to score an AI model, it is still the responsibility of the researcher to decide if the output matches expectations. Even when using other people to assist as with the facial generation, the researcher must decide which outputs to put forth for judgement; and in fact what sorts of

outputs to generate initially.

It possible for a foundation model to be replaced by a more efficient model, mitigating energy requirements and reducing carbon waste while maintaining a similar level of expected output.

This may require more insights such as a cost-benefit analysis, but given the results of the study it does appear some models can indeed be migrated to a different model, depending on expected output. Specifically, if an NLP model was required that produced a Writing Performance Score of 6, four of the five models evaluated would be viable candidates for such a task. While GPT-Neo 2.7B would be more attractive due to its higher score, if the need only required a score of 6, the best option would be the much more efficient GPT-Neo 1.3B. In this particular case, the models both provide the expected level of performance, while GPT-Neo 1.3B is much cheaper to run, or about 27% the energy cost of GPT-J 6B.. There might be instances where GPT-Neo 2.7B would be a better choice, however; perhaps in specific situations where the outputs for a particular type of language generation score much higher than average. In this case the user making such a selection would need to make a choice based on a finer granularity than as determined by the initial evaluations provided herein. If minimizing carbon impact is a primary concern, SCI provides a concrete way to compare foundation models and easily determine which outputs are "cleaner". This quantitative measurement can be contrasted with the qualitative reasons for selecting the model initially, and extremely similar qualities of expected outputs differentiated by actual carbon waste. Hence, selecting the most appropriate foundation model can be done with a much greater degree of confidence.

There is no significant difference in the quality of output of foundation models when using the CPU-only or GPU-acceleration.

Referring to tables 5.4 and 5.5, the Writing Performance Score of an output is not significantly affected when only using the CPU or enabling GPU-acceleration.

Further, up to 91% of energy is saved using the GPU over the CPU-only. Thus, employing GPUs is demonstrably a best practice.

It is obvious that a limitation in this study is lumping all the outputs together and providing an overall average score. Certainly, scoring each model according to specific types of prompts (informative, scientific, expository) would be an avenue of pursuit for future research. Further, evaluating many more prompts might prove useful to achieve more salient differentiation between models. However, given the enormity of time taken to perform this initial study, it should be evident that evaluating AI models can be as power-hungry as the initial training process and is no light undertaking.

Another area for further research would be in evaluating GPT-J 6B and GPT-Neo 2.7B using GPU acceleration. It is speculated that similar performance increases and power mitigation would occur, as with the other models, but hard data would be useful.

It would also be interesting to run a wider variety of prompt evaluations, tailoring temperature and max_length to each prompt type. This is a limitation of my work, in that those parameters remained static throughout. I expect that each prompt type would lend itself well to more comprehensive studies, perhaps one evaluating NLP implementations based on their ability to produce coherent scientific answers, and another that evaluates synthetic poetry.

Overall, foundation models can be evaluated and segregated by power consumption, and their outputs objectively scored and compared with each other. This is highly useful for making determinations as to which model to employ, but the final decision still requires expert knowledge in the deployment domain to accurately assess need vs. power consumption/savings. Further, a proper cost-benefit analysis is paramount to ensure model migration does not negatively impact the expected quality of output.

Finally, an interesting problem that occurred in performing this study was the need to determine which foundation model was more efficient by directly testing the models for inefficiency.

VI. CONCLUSIONS

Taken individually, it is evident that each computational area investigated in this dissertation experiences the issue of energy waste, i.e. inefficiency in regards to power consumption to perform useful work. While the approach to mitigate this consumption may differ by domain, there is no doubt that much can be done to reduce such waste.

My first goal was to demonstrate that such waste occurs in large scale systems; while this seems somewhat intuitive based purely on the laws of physics and the notion of waste heat, it was necessary to prove demonstrably and scientifically that this was indeed the case. And in fact, there have been countless prior studies concerned with reducing power to improve operational costs and environmental impacts that have all come to similar conclusions. However, their approaches were more concerned with physical systems, or very obvious ways to improve power efficiency. But there are only so many ways to lower cooling costs, or virtualize systems and eliminate physical machines. At some point, new methods must be explored and employed lest we run out of power; no amount of West Texas windmill farms can possibly keep up with the increasingly rapid demand for new data centers and cloud resources. The problem of energy waste does exist, and it is well known that newer ways to approach this problem are necessary as current solutions such as PUE optimizations, container virtualization, and DVFS may have reached the limits of their ability to optimize power usage. Thus, the segue into the remainder of this work.

Energy efficiency in large scale workloads is a major concern, and was definitively shown in this work. But what sorts of solutions am I offering to this problem? There is no singular answer to this question, but in each domain I demonstrated possible avenues of pursuit to mitigate the problem of energy waste. I

offered ways to improve power efficiency in all of the research areas, being somewhat easy to implement. Meaning, these solutions require no great outlay of effort or resources (and in fact most cases require less, cheaper resources). For web hosting, I demonstrated that low-powered, energy efficient servers can replace conventional power-hungry servers and provide excellent QoS for millions of monthly requests. After developing a novel simulator, I showed that cloud clusters can be reduced in physical size, saving hardware and yearly electricity costs with no reduction in performance. Further, I also confirmed that most VMs in cloud systems are underutilized and offered metrics to pinpoint problematic VMs and save millions of dollars annually. Lastly, I objectively classified multiple foundation models and assigned Software Carbon Index (SCI) scores to their outputs, which can greatly help mitigate carbon emissions.

However, even with further advances in energy savings, the necessity of new power plants will be a harsh reality, as demand outpaces supply. I see this problem akin to the oil crisis of the late 1970s. Car manufacturers had been bolting giant four-barrel carburetors on 400 cubic inch motors just a few years prior for maximum horsepower; later those same companies were forced to figure out ways to achieve 20 miles-per-gallon in an effort to reduce gasoline usage. Inevitably, lighter, more fuel-efficient motors and vehicles were produced. So, too, must more efficient methods be developed in the realm of large scale computing but, until those methods are discovered, my suggestions are an extremely attractive stop gap in that they can be performed immediately, with little cost, and with immediate positive effect.

Finally, I specifically focused on cost, while only being peripherally concerned with environmental impact. This is not completely intentional but, from a pragmatic viewpoint, companies and people will always seek ways to save money. By lowering power costs, however, they will also have less impact on the environment, thus reducing the carbon footprint of their workloads. In an oblique

way, the environmental concerns can be addressed by making greener methods of performing a task more economically attractive. Realistically, this is the best method for promoting green energy, as the end user will eagerly employ such methods if their workloads cost less to run on such systems.

APPENDIX SECTION

APPENDIX A: AI Prompt Results

GPT-J 6B

- What is an elephant? A rhino? A giraffe? A shark? Surely these are animals, and that is what the dictionary says. But what is _life_? Isn't it just a series of different chemical combinations? Aren't these different combinations, the laws of which govern the interactions and functions of a living organism, the definition of life? What can an organism feel, and what are emotions? What exactly is consciousness? What is emotion? What is an act of consciousness? Would" - 4
- What is a lemur? Like the words "dog" and "cat," the term "lemur" is one of those words that doesn't really have a precise meaning. "Lemur" is both a noun and a verb, and it refers to a class of primates, but in a general sense, it's just an umbrella term for the primates that live in Madagascar. There are over 200 species of lemur, but for the sake of this explanation - 6
- 'How are you doing? You're looking well. I'm afraid I've been neglecting you. I've been so overwhelmed trying to run this show, I haven't seen you, I haven't come to you. She took my hand and shook it firmly. "I'm fine, thank you, Dr. Rosenthal. You look quite different. I can see you've grown a beard." "Yes. I've been asked to stay on a bit longer." - 5
- 'Guess my weight. Then guess my age." "I couldn't do that," she protested. "Why not?" "Don't you see? It wouldn't be fair." It wouldn't be fair to me, he thought. And she should know that. "It wouldn't be fair to you either," he

said. He looked at her, her big face, the way her soft brown hair flowed over her shoulders, and the tenderness in her' - 4

- "Can you dance? This is a discussion on Can you dance? within the Concealed Carry Issues & Discussions forums, part of the Related Topics category; Originally Posted by Pookie I've seen a number of videos, and they all look pretty convincing. The only thing I want to...I've seen a number of videos, and they all look pretty convincing. The only thing I want to be careful of is a guy who can dance, because it looks like he's" - 3
- 'Good day to you, sirs.' "I am sorry to bother you, but you have a guest." "I have sent the car home." "This way." "I will put that on the table." "Here, sir." "Here." "Mr. Stevens, it was not my fault." "I tried everything." "There she is!" "Sylvie!" "Oh." "Oh, my God." "Oh, my God, you're all right." "M' - 4
- 'What is a computer? A computer is a piece of hardware made of electronic components that are wired together to provide various functions. When a word like "computer" is used to refer to both the hardware and the software together it is called "computers" because both hardware and software together are an "information processing unit." What is a computer store? A computer store is a retail store, usually owned by another company, that sells computer hardware and software in their stores.' - 7
- 'Who is Elon Musk? Elon Reeve Musk, known as Elon Musk, is a South African-born British technology entrepreneur, inventor, and investor, who co-founded SpaceX, the rocket company that was founded in 2002. He created PayPal in addition to the electric car company Tesla Motors. As of 2015, his net worth is estimated to be \$10.4 billion, making him the 46th richest person on the planet. Although Musk was born in Pretoria, South' - 11

- "Are you alive? I'm having a really bad day So I'm just gonna go ahead and just walk away I don't want to say things I don't mean I'm not really sure what to say. What do you want? I mean. I just don't know what to tell you girl Maybe if I could find the words to say How to make you understand But I feel like I'll never learn I feel like I'll never learn" - 3
- 'Mad Max is a darkly comic look at a post-apocalyptic world where the only thing that matters is survival. It was directed by George Miller (Mad Max, Happy Feet), produced by Byron Kennedy, and stars Mel Gibson (Hulk, Braveheart) and Tom Hardy (Bronson). Mad Max was released in theaters on May 15th, 2015. It is rated R for "Language, Violence and Profane Content." I am rating the film PG-13 for - 9
- What is an elephant? Is it a fish? A dog? A dragon? No. It's a mammal. In fact, it's the largest land animal on the planet. Yes, all of those animals have something in common. They're all mammals. But not all mammals are the same. When you look under a microscope you'll find that elephants, dogs, and cows all have different shapes and features. This makes sense, because they're all mammals. In addition to having different" - 5
- 'What is a lemur? Lemurs are primates that are generally between 20 and 25 cm in length. They have a soft, dense, bushy tail for extra support. They have fur, are mostly arboreal and have small, rounded heads. There are 13 species of lemur that live in the wild. They have adapted well to the habitat in tropical forests, grasslands and mangrove swamps. Lemurs have long memories and are extremely' - 8
- 'How are you doing? Let me start by saying "hello!" There are some things you just cannot say to a stranger. The obvious ones are "hello." "Good day!"

"Can I help you?" Although I wouldn't let it get to that point, I'd rather not even attempt it. But my mother didn't give me' - 3

- "Guess my weight and buy my favorite books! When I'm at the bookstore, I tend to gravitate to the weight loss section. I can always find a good book on losing weight, but it takes me forever to find that great one on health and fitness. This is my method of finding the perfect book for myself. I've been buying the book that I really want and weighing myself when I get home. I like to track my weight changes with the books. The price is" - 4
- 'Can you dance? Can you sing? Can you dance? Can you sing? Can you dance? Can you sing? Can you dance? Can you sing? Can you dance? Can you sing? Can you dance? Can you sing? Can you dance? Can you sing? Can you dance? Do you speak? Do you understand?' - 3
- 'Good day to you, sir.' "I think that will be all." "You can go back to your hotel if you like." "No, I'm leaving today." "What?" "I'm going to the south of France for a little while." "But, uh, what about your concert tonight?" "I canceled it." "Well, that's not very nice." "It's not being very nice that bothers me." "You see, I didn't get to say good-' - 5
- 'What is a computer? A computer is a machine that performs the processes of calculation. A computer calculates a fact or an equation. The basic element of calculation in a computer is the bit. What are the parts of a computer? A computer consists of: a Central Processing Unit (CPU) Input Unit Memory Output Unit Storage Units A computer may be small or big, it's made to measure and is' - 7
- "Who is Elon Musk? - CNBC Who Is Elon Musk? - CNBC The billionaire founder and CEO of Tesla Motors and SpaceX talks about his vision of

mostly of hard food such as fruits and seeds." While that has its share of verbiage, it doesn' - 7

- 'How are you doing? I'll be with you in a minute." He went away and returned in a few minutes with a glass of water. "Are you ready for some of these pills?" "As for that, Doctor, why couldn't you give me them in the first place?" "Then you wouldn't have had the pleasure of seeing you drunk." He laughed, and his laugh sounded like a tinkling of bells. He came ne' - 6
- "Guess my weight. It's been a hot summer, and for the past few weeks, I've been noticing how hard it is to guess what I weigh just by looking at me. I'm getting close to my heaviest since I started dating my current boyfriend in November. (Though, I'm not actually sure if that's my heaviest weight.) I know it's not my absolute heaviest, like I was last summer. At my heaviest, I never quite knew how"- 4
- 'Can you dance? The second most important skill for someone in the tech industry is dancing. And the worst possible way to discover a dance partner is online. I've had this happen a few times. I'm dating a woman at a tech meetup and we have chemistry. The fact that we're both geeks doesn't matter. We've been on three dates now and we just can't stop talking about the same' - 4
- "Good day to you, sir. You see, we're not the only ones with a secret of sorts. I am not afraid of my sister nor my brothers not even of my father. It's you that I fear and if you don't agree to this offer I'll be taken directly to Azkaban. We're talking about something really serious, my friend. So, we could do this the easy or the difficult way The easy way would" - 5
- 'What is a computer? It's what we use to program our software, build our websites, and perform all the calculations in our daily lives. But beyond these

roles, we've become intimately familiar with computers as instruments of creativity-to us, computers seem to embody the spirit of modernity and technological progress. To the extent that we've become comfortable with this perception, we've forgotten where these things came from. On the first issue, computers are' - 12

- 'Who is Elon Musk? Elon is a graduate of the University of Pennsylvania, where he majored in Physics and Economics. He is an innovative entrepreneur and a billionaire. The founder of a successful car company called Tesla Motors, the company he co-created with Martin Eberhard and Marc Tarpenning, that produces electric cars. Elon was born the only child to Maye Musk, a computer scientist, and Erle P. Musk, a biochemist' - 11
- 'Are you alive? I have always been a loner. At least, since I started writing. For many years now my thoughts have been my only source of company, my only true friends. I've been isolated from the real world, cut off from the people who live in it. It didn't happen yesterday. It was over a long span of years. A lifetime really, a lifetime of loneliness. It didn't happen' - 4
- 'Mad Max is a post-apocalyptic action film and the third installment of Mad Max series directed by George Miller and written by Brendan McCarthy. It is the first film in the series to be set primarily in a desert rather than a wasteland. Production began in February 2004 and shooting lasted seven or eight months. The film premiered at the 2004 Cannes Film Festival to critical acclaim and was released in Australia in October 2004. It became the second highest grossing film in the history of Australian cinema, and is considered' - 9

- 'What is an elephant? The answer, according to the National Geographic online encyclopedia, is "a large member of the order Pilosa." The term elephant originates from the Indo-European root *xer, which has two related senses: to pull or to drag, as in the case of a boat that goes at a great pace, or an elephant pulling a sled. In zoology, an elephant is a large, flightless, and herbivorous mammal, weighing - 9
- "What is a lemur? What is a lemur? I think we all know what a lemur is. We all know the cute cartoon lemurs of the Disney movies, and we've all seen the fuzzy little furry creatures in the commercials for baby formula. But what really is a lemur? Like all creatures on this planet that are not humans or chimps, lemurs are unique. They are a family, or at least a family group, - 6
- 'How are you doing?" he asked. "I'm all right, thanks. I'm going back to the hotel. Will you call me when you get back to the house?" "I will," he said. She didn't answer. She walked into the elevator, went to the fourth floor, and asked for the desk on the ground floor. The clerk remembered her. She got lost - 7
- 'Guess my weight? Are you like me? Do you need to know the exact number of calories you gained or lost yesterday? Well, do I? No. Do you? No. I think not. I'm not going to tell you the exact number of calories you burned or spent yesterday, to tell you how many bananas you ate or how many calories you burned while walking or how many steps you took. But I will - 4
- 'Can you dance? The power of the foot step is not in the foot step alone, but in the feet themselves. This is the first in a three-part series. Read part two here For most people, life is a dance of constant change. The changes in our lives may be small, subtle, or large - minor improvements in our physical

movement, or major improvements in the way we move through life. They may just be small, subtle improvements, or they may be - 5

- 'Good day to you, sir.' "Uh..." "I believe you're here to see Chief Inspector Jappler." "Yes, it says so right here." "We've got a lot of paperwork." "Well, I'm his assistant." "This is just a formality." "Thank you, I'll take it." "I don't think that's necessary, ma'am." "Why don't you step in here." "I'll show you how the office is set up - 6
- 'What is a computer? In the most basic sense, it is an electronic machine with several components, one of which is a data processor. However, to understand a computer, you must first understand a computer language. The human mind takes input and makes a series of decisions to give the desired output, but a machine must take a string of ones and zeros and turn them into text. This is where computer languages come into play. These programming languages allow a computer to make decisions by reading - 9
- "Who is Elon Musk? The answer is simple: He's not a politician. No, you're not imagining it. Tesla CEO Elon Musk is a very serious guy who is not particularly interested in political positions or campaign trail politics and who doesn't see the need for one. But he's also not the sort of CEO you imagine going around making empty promises and selling products he's not really enthusiastic about. In fact, as the CEO of Tesla and SpaceX and the co - 7
- 'Are you alive?' He's alive. I just told him he's alive. "Did you like your flight?" "It was alright, but I'm a little worried about our seats. I hope you didn't leave anything behind." "No, we took care of that." "Good. I'm a little worried about our seats." - 6
- "Mad Max is a film that changed my world with its shocking plot and an

unflinching portrayal of the human condition. The film's opening scene is not unlike an urban disaster film, with the protagonist wandering through the ruins of LA. But unlike many disaster films, Max is not a victim who cannot escape his misfortune. In fact, he is a man who has made a bold statement to his fellow man: "You will never be able to do anything about this, but you can control your - 10

- 'What is an elephant? You've seen the movie, The Lion King. Everyone knows the story now - King Louie (voiced by Ben Pridmore) is a little person on a big island. There is the talking tree, and everyone loves it. It is the only tree with a human voice, and it sings to make the king happy. As far as the story goes, King Louie is the most important character of the whole story, but not in the way that - 4
- 'What is a lemur? A lemur (*Leontopithecus santiagoi*). These primate-like mammals look more like monkeys than apes. Credit: Ithaca College Lemurs, the primates of Madagascar. Credit: Ithaca College The LeMurs of Madagascar A lemur is the largest primate on earth and is considered a family of primates. As you can guess from the name, le - 4
- 'How are you doing? How are you? I'm so sorry but today is a very sad day. I'm not going to bother to put a good-bye on here but I just want to say thank you for everything, I know a lot of you have been reading this and I appreciate it. I had a lot of fun with this story when I was writing it. I love that it's a fantasy so most of the characters are based on me - 5
- 'Guess my weight - and it isn't even close I thought I was pretty good when it comes to figuring out my weight. I get it wrong a lot, but it's usually my fault and I'll get all my weight figured out before I have to wear a really heavy vest

to get to a meeting. But the other day I got it all right and I was shocked. I had a couple of good ones. But not that - 3

- 'Can you dance? The answer is: Maybe, when you're ready to learn how. Don't be discouraged by the first few times you try to dance. Even if you are a total beginner, you'll probably want to learn to dance at some point in your life, so just be patient, and take slow, deliberate steps towards the dance you want to be able to do. If you're interested in learning to dance at - 11
- 'Good day to you, sir," the man said with a slight incline of the head. To my surprise, I glimpsed the unmistakable nose and mustache of my father. "Is Captain Doreen around, sir?" I inquired, turning to the man. "No, sir," he replied. "She's out for the moment. We sent word for her." I looked for the woman, but could only locate her by her voice mail message. " - 11
- "What is a computer? A computer is a device consisting of electrical and electronic circuits, operating on information. I agree that this isn't a computer but it is a device that runs on electricity. The difference is the information it carries inside the computer. A computer is nothing more than a logic machine. An electronic circuit is nothing more than a logic circuit. A logic circuit performs some operation on the signals that it receives. It does not matter if there is - 6
- 'Who is Elon Musk? One of the world's most famous entrepreneurs "The idea of Tesla is to put millions of people in their cars and homes, and every person could become a full-time Tesla driver." -Elon Musk, Tesla founder Tesla is the company that can make a car that is cheaper, safer, more efficient, and a hundred percent environment-friendly. They are the ones to bring back the energy, and electric cars, - 6

- 'Are you alive? Are you alive? I am so alive. I am alive! (Yes, I am alive. Yes, I am alive!)" We're living in a time when a "lion king" is a modern-day American icon or an ancient Greek hero. No, I'm not referring to America's very own Mr. T, who can be a bit over the top on occasion, but this American "lion king" is not the prince of darkness that - 3
- 'Mad Max is a film that has only been in the public domain for a short time. There is no release date yet and there are already several legal challenges to the film's classification, so it's unclear at the moment whether or not it's going to be able to enter general release. The film's classification status has a lot to do with what exactly it is, and there are a number of different aspects that make the film difficult to classify. - 4
- 'What is an elephant? A big, heavy, eight-legged mammal with a long trunk. A very unusual mammal. But how unusual? Well, they are the only mammals known to have learned to live in a humid heat, and they learned to climb trees and swim pools. They have no foreheads, eyes or noses in their heads. But they have a complex nervous system with multiple brain centers and can hear in the night, smell underwater, and walk like a human. They walk upright and - 7
- 'What is a lemur? A lemur is a small primate. They come in a number of different species that belong to the primates (apes) family. When you get down to it, lemurs are just monkeys with long tails. These animals are usually around 15.5 to 35.5 inches in length. They have long soft fur that can reach up to 4 feet in length. They have short legs and fingers and a long, thin tail. - 7
- 'How are you doing?" he asked, and then turned and walked away. I had an idea then. "I would like to go over to your house." "I can't." "Why not?" "It's for sale and I don't want to sell it because I'm afraid it will turn into a place of misery for me." - 10

- 'Guess my weight.' "I have been a good girl." "What if I don't feel like I have to eat at all?" "Then I will make you eat until I don't feel like eating, okay?" "Come on." "Come on." "I love you." "I love you, too, Mom." "Hey, what is this?" "You don't like your favorite pizza, just like everyone else." "Why?" "Because it's made from scratch." " - 3
- "Can you dance? I can't dance. I like to move my feet in a figure-eight pattern, and when the dance starts, my knees are bent and my feet are pointed straight up; when I come to the barre, I do a quick side step with my right leg. It's a dance, really. I have a little more confidence with my body than I do with my mind. I think it's good that I don't have a good self- - 6
- 'Good day to you, sir," he said, and opened the door. Cedric pushed in behind him. "What time does your little friend come upstairs to bed, monsieur?" "Not till midnight. But if you would come to supper, you and I would have a very pleasant evening together." The door of the bedroom closed; the footman waited for the word. " - 10
- 'What is a computer? It is a machine that stores information in binary form and uses it to solve problems. The word computer, when spoken with a British accent, sounds like it is being used in a very similar way to a computer. When you pronounce the word correctly, you give the appearance that you are indeed hearing it spoken as it is spoken in a computer. When you pronounce it incorrectly or not at all, you sound as if you are either speaking another language or you - 6
- 'Who is Elon Musk? Musk is a billionaire entrepreneur and technology visionary. He is one of the world's best known entrepreneurs and is often compared to Henry Ford for revolutionizing the transportation industry and being the person who invented the auto as we know it. In an early age, he was

fascinated by space exploration and dreamed of space travel. But his interests and passion spread beyond space exploration. He is also interested in making money and investing in companies. He has - 12

- 'Are you alive? What a great thing that you found time to send a little note to me. It's very sweet of you. Here you go. The second chapter of The Story of the Man Who Found Time Again. Chapter 2. The second chapter contains a number of things that might have been included in an earlier chapter or even an even later chapter. But Chapter 2 serves as a kind of foreword for Chapter 3 so now we' - 3
- 'Mad Max is a violent, anti-social, and unbalanced character who believes he is above the law. He is also a bit of a fascist. When Max is not getting into fights with his friends or assaulting or killing women and kids, he enjoys driving his car fast, smashing as many cars as possible, and firing laser weapons. One thing that Max is good at is driving. This gives him a way of escaping from his boring and sad life. Max has a very interesting personality but sometimes - 9

GTP-Neo 1.3B

CPU only

- 'What is an elephant? A: The answer to this question is very simple. The answer is that they are not very different, actually. They simply appear in quite many different places in different languages. In the plural the word "elephants" is applied to all elephants in the world. In the singular it is applied to one particularly small one: The elephant lives in the jungle, but is otherwise a small animal. In the - 4
- 'What is a lemur? Lemurs are small-bodied, herbivorous primates, which belong to the family Lemuridae. These diminutive animals are most commonly

found in Southeast Asia and Africa where they are protected by the dense forests of those regions and some of the most endangered forests remain on the island of Madagascar. Lemurs (Lemuridae: Lemur) are extremely diverse, and live in a variety of habitats. Many lemur species, such - 12

- 'How are you doing? How are you doing?" It didn't matter about your appearance, your voice, your age or your education. He called. I was glad because I was worried about you and I wanted to tell you that you were the one I missed most. I did try to tell you by opening my mouth. But that wasn't possible. He called again. I opened my mouth. He said I was going to be very busy soon - 9
- "Guess my weight was down to one more pound! I can now wear my jeans a little wider, and my pantie can be pulled up, which I love because they don't seem to fit. But I have a couple of more issues with my new jeans. One I have been trying to fix since this morning. It's the back pockets that are just gaping holes. I had to remove the top buttons and my pockets, but when I pull them up to check everything, they - 8
- 'Can you dance? The way I was dancing at that party the other night, with you there, just the way you danced with me?" "Oh, of course. We were both dancing. You were doing something very nice and graceful. You were so very pretty. You could have been a dancer in anything. I know how I looked, but I didn't feel like it, or think of it, the next minute." "The next minute, then - 7
- "Good day to you, sir. My name is Robert Gee, and I am an assistant professor of the History of Religions at the University of the Colonies at Fort Lauderdale. We have a new edition of Fischer, but we don't have the new edition of the new edition yet. I am pleased to meet you, sir. CALLER: Please tell me about it, sir. Sgt. L. G - 6

- What is a computer? A computer is a machine with two main components: a central processing unit (CPU), and memory. A CPU or hardware processor is the part of a computer that runs the application code. A storage device is a component that stores data. Many types of storage devices can be used. Storage devices may be hard disks, solid-state drives (SSD) and solid-state hybrid drives. Different types of storage devices are used for different applications. - 9
- 'Who is Elon Musk? What is Elon Musk? For most of his life, he has been an entrepreneur, most notably having founded PayPal, Tesla, The Boring Company, SpaceX, and Boring Company, Inc; the most recent being Tesla Motors with a \$1.5 billion deal that is valued at \$62 billion. He is currently the CEO of Tesla Corporation with a net worth of \$42 billion and has also been a co-founder of PayPal, a Fortune 500 company. - 9
- 'Are you alive? My name is Joss Tarrant, I am an American film director. I am currently residing in Tokyo, Japan. I am also a screenwriter. I've got a plan to kill you, Joss. I am a video artist. I'm going to make some art before you kill me. When you come into this world, you will look like me. You'll look like me, but you - 3
- "Mad Max is a character I've wanted to play for a while now, but I've always been hesitant about it. I am a fan of the comic and the movie alike, but I am still a little turned off by the idea of the lead character being a male version of the same character. I have had the character in my head for a long time now, but I've never written a character and made my ideas about what I would do with him. Recently I came up with a new" - 7
- 'What is an elephant? The most common definition is that elephants are big, tall, hairy, and powerful animals. However, there is a much more subtle

difference. They are not human. They are not animals. They are not a part of our everyday world. They are not even primates. They are an animal. So what is a elephant? If you're sitting next to someone holding their head and letting their neck fall off, the elephant is walking by. If you - 4

- 'What is a lemur? What is a lemur? The word lemur is often used to mean any member of the order Pholidota, such as sloths or anteaters, but the word has other meanings, too. In Britain, where the term "lemur" is used to describe the members of other species of primates, the word lemur is used to refer to the lemurid genus, which is most closely related to the sloth and - 7
- 'How are you doing? A: I'm doing quite well and also have some friends and family around the same. But you know I was actually there a few days ago and I did not take any medications - which wasn't that important to me but I do take them anyway I was also having a migraine with the headache and I was dizzy. The doctor prescribed Benadryl, a migraine medication. A: What did you take for the headache? - 8
- 'Guess my weight? A: You can try using: `weight = float(input("Please enter your weight: "))` `print(weight)` If you were wondering what is this called, it's called "Double Integral", it's an alternative to the original Integral function (because the calculation is done over whole number of weights!) - 3
- "Can you dance? How far could they go with all the energy they have? How far could they go with all the knowledge? How far could they go with all the training? How far could they go with so much enthusiasm and so much knowledge? They've never done it before. Don't you think they're going to do it? They can do it. Is this it? What's next? - 9
- 'Good day to you, sir!" "Do you know who I am, sir?" "Do you know what I

do?" "Do you know who the God of the sea is?" "Do you know who I am?"

"Do you know why I come?" "Do you know your God?" "Does it have to be a woman?" "Yes, and a foreigner, too." "Then why was your woman killed?"

"Don't tell me that my God had a wife and children!" - 3

- "What is a computer? A computer is very similar to a television that makes it interactive. It's a device that we use to make our lives easier, more fun and easier not to say boring. What is it made of? The computer is made of a lot of parts: the hard drive the memory the system the CPU and other processors the graphics card the display the monitor - 5
- 'Who is Elon Musk? Elon Musk has always been the biggest surprise of the last decade. In the previous generation, it seemed that almost everything he did was on a level playing field. For example, as one of the founding fathers of the electric car giant Tesla, he was able to compete with traditional petrol and diesel car manufacturers, while also building his own Tesla vehicle and developing something that was truly groundbreaking in itself. Then, with his electric car company going public in - 6
- 'Are you alive? You haven't heard?" Sally sighed. "I was too upset to listen." "And he?" "He got angry, he shouted at me. He said he would kill me if I told anyone. I have no idea if he meant that or not. I have no idea. I am afraid to think." "Well, what do you think he meant?" "I don't know." "And here we - 6
- 'Mad Max is a popular science fiction film that received generally positive reviews upon its release in 1999. It was directed by J. Michael Straczynski and stars Arnold Schwarzenegger, Charlize Theron, James Caan, and John Hurt. The film focuses on a group of humans who live in a huge, dystopian city on the surface of a dying planet. In addition to the group inhabiting a skyscraper in the city, there are also an insectoid army, the A.I., made of - 6

- 'What is an elephant? The name is an elephant. An elephant is a huge animal, and a powerful animal. This is because an elephant's trunk can stretch to 80 feet or more. A whole herd of elephants can consist of as many as 50,000 elephants. An elephant has a big horn in its trunk, and that's what allows elephants to charge. It's not unlike a lion attacking a kangaroo. It's like - 4
- 'What is a lemur? "I am an ape," said our host as we made our way to the next set of safari tents. "There are lemur species on the East African coast. Some are even larger than the wildebeest and hippopotamus." We were now on the beach-an island in a bay, overlooking the sea. "Why are we standing on the beach?" I asked. - 4
- "How are you doing? I'm having a bit of trouble reading this forum. I've been having very strange problems with the new firmware on the X1, not to worry, will still be using the stock firmware. Anyhow, the problem started happening after I installed a lot of software on the board that I don't know if it's a software or a hardware issue and the problem has spread even more now. I finally gave up and uninstalled software for every single thing and went - 5
- 'Guess my weight! Well, here's the thing: I lost my mom last Wednesday. I'm sorry to all of my friends and my readers, but this is not a good thing. Although, I have to say that my friends are all better than I am, as I still couldn't make a good-enough effort to keep up with my schedule. But what really hit me was this: I know that this is probably not " - 4
- 'Can you dance? The answer to this question isn't very specific. But if you are dancing with a friend, you may be able to suggest a dance and they may then come up with a dance of their own. This is a fantastic way of bonding and creating new memories. Can you jump? Jump is very versatile, and can be a

great exercise for both dancing and learning. It is ideal for those who like to get involved in a group or exercise - 6

- 'Good day to you, sir! You can not imagine what a pleasure it is to us to have the honour of making your acquaintance. Your lordship has already had the pleasure of hearing us speaking of you before, and to know that you are a gentleman would be much more gratifying than an hour's conversation. Lord Aylmer. I hope I am not very late, Mr. O'Hara; I had to leave the l - 9
- 'What is a computer? Well, it's a computer that runs software. Think about it like a big version of the PC or laptop. Sure, we have a lot of computers that run Microsoft Windows and other operating systems, but how many computers run the software that run our phones and tablets? According to a new report from the U.S. Economic Policy Institute, for every person whose smartphone runs a fully-loaded operating system-such as Android or Apple's iOS - 6
- 'Who is Elon Musk? One thing I know for sure is that Elon Musk is a really powerful person. And that's why I can't stop telling the story of how we all became friends (or at least his friends): The first time I met the billionaire founder of SpaceX and Tesla was when I was a student. I had just moved to the city just outside of Palo Alto (the Silicon Valley region of California) and was looking around for places to live - 8
- 'Are you alive? That doesn't sound like a question worth asking." The man behind the wheel of The Boneyard said that the people who shot up a movie theater in Aurora, Colo. were wearing masks and that there was an "explosion of glass as the camera is going in," according to a report by The New York Times. "They didn't kill everyone," he continued. "You didn't - 3
- 'Mad Max is a very dark and depressing book. It's a movie that's very sad,

not so much on the surface, because it's a movie well-adapted to the form, but rather on the heartlessness that comes with the way in which the story is told. There are so many unanswered questions about this movie that I couldn't even begin to answer them, so I'll keep this a short review, because I can' - 5

GPU Accelerated

- 'What is an elephant? The word refers to an animal which is very similar to itself, being in the same form, size, and behavior. 1. It is the largest land animal, and as such, it is sometimes used to describe the world in which we live. It has no senses. It has nine fingers, but it can never feel pain and can only have the sense of smell and taste. 2. It is an animal that has - 4
- 'What is a lemur? What is a lemur? Lemurs are the world's largest cat family, the largest of only three types of monkeys. They live on Madagascar, Asia and Borneo, with more than 3,500 species to date that range in weight from less than 5 kg to more than 18 kg. What are the species? The lemur family contains about 100 species. In each species, there are numerous genera - 7
- "How are you doing? I'm fine, thank you for asking. So this past week, I found out that I am a part time college student and I had to cancel my trip to the doctor's office because of the holidays. So, there is still a chance that I could be able to go back because I will be done with school and I will not have to make any plans of my own anymore. But, I had a question or two that I did not know where - 6
- "Guess my weight... or my height... or my age... or my sex... or my country... or my skin... or my religion.. But the one thing I can tell you for certain... The one thing you have to accept... As you grow old... You will be surprised... You will find out that you... really... love someone.. You will find out that you

don't love someone.. That's not love... - 5

- 'Can you dance? The question comes out of a conversation that went like this:
"Hey, you guys ever been on the dance floor in a place that has free alcohol?"
"Yes!" says one of my boyfriend's friends. "We been to hell and back." "You guys have to try it, it goes to your head." The next day, as we make our - 6
- 'Good day to you, sir. We will not make any changes, but we do wish to take our time to study the matter and then act accordingly." "There is need of haste. I want to visit the place myself, with my party. I shall send some of my men ahead to ascertain if anything has transpired. As regards the two girls—I wish to bring this matter to a conclusion as quickly as possible, for, at present, I am not - 4
- 'What is a computer? What is programming? This question has been asked and answered many times as I have read and reread the answers to other similar questions on this site and other related sites and forums. But I know that I have not found a definitive answer to what is a computer nor to what is programming. I know that a computer is a machine that is designed to perform specific tasks. And I know that programming is a set of skills that allows a programmer to enter a particular program and a - 5
- 'Who is Elon Musk? What is Tesla Inc? Elon Musk has been in the spotlight for various reasons. For many, he is a self-made billionaire and his vision for the future includes the electric vehicle revolution and a world of autonomous and sustainable vehicles. Some have pointed to his philanthropic work in the world. Others, however, doubt the veracity of his claims about Tesla (TSLA - news and information) and his vision of the future. What are - 8
- 'Are you alive? A voice echoes from the dark, the voice of your dead friend;

the voice of your mother: "Do your worst." So he killed her daughter. "The man is out of control," a detective's voice says. "He is obsessed." The woman and her friend are in shock, and in the dark you hear his voice: "I have a daughter who is alive, trapped in the - 6

- 'Mad Max is a 1984 American action film directed by Roger Corman and written by Roger Corman and Paul Mazursky, based on the short story "Max the Monster" by Don Winslow Plot The film revolves around a man named Maxie (Roger Corman) who lives in the fictional town of Paradise, Alabama, with his wife and daughters. At the same time that he lives in Paradise, he works in the meat packing industry and also has a small role in - 6
- 'What is an elephant? The word or phrase elephant has several meanings. The most common meaning is a big animal, and most often refers to any large animal larger than a horse or pony. Etymology The origin of the word elephant is uncertain, but it is likely first used in the 17th century. The root of the word is of uncertain origin, but it is believed to have come from the Latin elephantum, meaning "of the elephant". The noun elephantus - 7
- "What is a lemur? Leopards were in danger of extinction until 1887 with the successful introduction of the white-faced lemur (*Psammomys*) by J.B. Sahlins. The first white-faced lemur was found in 1905 in the Sahlins' garden in Paris. It was named by Sahlins, because its fur of a deep copper-color which had not been observed in any other species of lemur - 5
- 'How are you doing? If you don't like the direction things are going, then I can't help you. But I can try to steer the ship in a positive direction, and help you do your research and keep an eye out for that new restaurant that just opened across from the university. If you are already established, you can

choose from the following menu categories: 1. A small salad. 2. A hearty main.
3. A plate full - 6

- 'Guess my weight: 1 - I have had my fair share of love handles, and I'm a big boy! As I just said before, I try to put myself first. I feel a lot more in control in a romantic relationship. I think it's just my nature, but it's nice to know I don't have to apologize for everything! Thanks for sharing your feelings! 2 - I love to read - 5
- 'Can you dance? It was a question. One I've been asking myself for the past week. Why the hell do my feet hurt so much right now? I was in line behind two other couples waiting to dance. And right behind them was a woman who had fallen asleep in my arms a few times. When I finally got her to wake up, I was already off to the dance floor. After a few hours, she stopped looking at me and started - 6
- 'Good day to you, sir. My name is Linn. And is my dear mother-in-law's maid - she's like a sister to me - is to come out tomorrow as well as today. We want to have tea together. Is there any way it could be arranged? My mother-in-law is away. We're very busy. And I'm sure that if you could arrange it that I could come out tomorrow as well as today." "Of course," I said - 5
- 'What is a computer? Is it a device that can run programs and other code so it can run software for you? This article is going to be about what a computer is, and how they differ from other technologies. If you are thinking of purchasing a computer, make sure you know what you are buying! What is a computer? A computer is a device that is used for one particular use, it is not something that allows you to have access to programs that run on another - 7
- 'Who is Elon Musk? Is he a billionaire? What does he stand for? There's no

shortage of opinions on Elon Musk. Who is the Tesla CEO, and his views on other companies. There are many who think he is an asset to the company and people would be crazy to doubt him. I am an early investor in Tesla and have no doubt that he is a successful businessman. However, there are those who question Elon Musk because they disagree with some of the - 7

- 'Are you alive? I'm not talking about the part of the planet where we live, but the part of the planet in the universe where we have a home. That's where we exist together, where we love and support each other. That's where you are, right now, and forever will be. In that space, on this planet, with you. I'm talking about our home. Not the place you live in, but - 9
- 'Mad Max is a big hit. But we may need something better than a \$50 million-plus budget to make it great. The Mad Max trilogy is a hit with critics, and has created two other movies in recent years. Now, according to The Wrap, DC is giving the movie the green light in time for Christmas. Yes, because this is a DC movie. In a statement with the article, DC Entertainment President Greg Silverman said the company is moving forward with the - 5
- 'What is an elephant? This popular question in India was first raised by the late Mahatma Gandhi in his famous Satyagraha in which he led the anti-British movement of the 1940s. He asked the community to raise the question to the Prime Minister and the Chief Justice of India, Justice Sardar Patel, who was his main target. Gandhi put the question in the context of the movement of Indian nationalists. The question was to ask about the nature of the movement itself, which Gandhi - 4
- "What is a lemur? Leopards live in South East Asia but are most prevalent in Southeast Asia. They are mainly found throughout Borneo and Sumatra, and also in northern New Guinea and New Guinea's largest island, Papua New

Guinea. Leopards also inhabit the islands of the southern coasts of the Malay Peninsula, the Lesser Sunda Islands, and Borneo. There are only about 600 of the large-brained animals, which - 4

- 'How are you doing? I'm doing great! I've been a long time reader of this site and I have had the privilege of meeting many of you. I enjoy your blogs, your advice, and your feedback. So please, please, please leave me a "Thank You" note. I'm sure I'll get the credit. I have written a book...I need feedback! - 8
- 'Guess my weight in pounds?" "I'm pretty sure I'd be happy just eating cereal." "What is going on?" "Where am I?" "Where am I at?" "What am I doing?" "That's it." "Let's go home." "Wait a minute, I got to get my clothes on." "Come on, let's go." "I guess I need a shower." "We're going to look at everything that you brought home." "Oh, I - 3
- "Can you dance? The only thing is that I'm in pain. My heart is beating so hard that I'm afraid, That now I've made the most solemn promise, That I won't be able to make another one. You will have to dance on the floor, for It's your wedding day. Oh, but it won't be your last, either. You must - 4
- "Good day to you, sir. I hope you've got a warm day for your first morning today. I've got a big news going round the office so I thought that you might like a little bit of it. So I've been asking you to do some research about the Vietnam veterans' situation and I thought that it might be useful if you did your bit for our campaign. And I know you may well be - 6
- 'What is a computer? A computer is a series of electronic components known as parts of the equipment. Who is a computer user? A computer user is a person that uses the computer in a computerized environment such as a home, office, college or retail store. It is a person that uses a computer in daily tasks

for business and recreation purposes. What is software used by a computer? A computer uses software, i.e. programs, to - 5

- 'Who is Elon Musk? He's the founder, CEO, and President of SpaceX since October of last year. The company has also founded several other companies, including Solar City, Tesla, SolarRoof, and SolarCity. When you get the opportunity to meet the man behind the wheel of one of the busiest shipping corridors in the world, you might get the impression that it is nothing more than a dream job. Maybe you don't quite get it though. - 8
- 'Are you alive? Or are you a ghost? When we arrived at the house, I heard some sound coming from upstairs. "It's my mother...she's on the roof and calling for help." My husband, Steve, and I rushed up the stairs, but by the time we got to the roof, she was nowhere to be found. A few more knocks, and the call went unanswered. "I'm scared to death, - 9
- 'Mad Max is a very good movie-I think. I've seen it a few times, though it's been years since I've managed to see it. The first time I saw it was one of my earliest memories, at age six. My family was visiting our local movie theater for the first time. We went to it for two shows then a little while later my parents went for a second time. What was on the screen that night? The first thing that - 6

GPT-Neo 125M

CPU only

- 'What is an elephant? A: The elephant is a creature that is not a human. A: The elephant is a creature that is not a human. A: The elephant is a creature that is not a human. A: The elephant is a creature that is not a human. A: The elephant is a' - 3

"We are facing a crisis." - 4

- 'What is a lemur? A lemur is a wild lemur found only in Madagascar. It has no teeth. Its skin is mostly gray, with black spots on the lower part of the neck, shoulder blades, and tail. Males have a long, thick tail, and sometimes an additional tuft of hair at the end of the tail. The teeth are small and pointed. Males can be between the ages of 12 to 25 years. They are not usually found in the wild, - 4
- 'How are you doing? I'll call you on your cell phone." Proudly, she stood there. I never spoke the name Michael Michael Michael again, though that's exactly how I remember. I didn't see him again until I was an eighth grader in 2000. A classmate took me aside and asked me what I thought of this new addition to the family. "A total asshole?" I said. "Really?" she laughed. - 5
- 'Guess my weight in pounds", "gives me the ability to weigh and measure my body") ; \$lst get_field ('lst') ; \$lst.``. \$lst - body - name. `` return \$lst }
function get_field (& \$f) { return \$f ; } RAW Paste Data php require_once
'curl.php';#####

- 3
- 'Can you dance? I love to dance." In some ways, this makes for a particularly rich story. But it's not just a story about dancing and love; it's also a story about how culture shapes and sustains a community and the people who live there. The city of Seattle has grown up with a large Hispanic population in recent years, which has caused tension with the Native population, which has in turn been fighting for inclusion. When people are able to dance and live - 7
- 'Good day to you, sir; and thank you for the kindness you have shown us, and

the kind treatment you have shown us. The whole nation is praying for you. We are all looking at you like a young man with no prospects and no prospects of living. We want to ask you, sir, you have given us a great deal of pleasure, and I believe you really do believe in us. Your soul has been set free. I believe in God and I believe in your mission. - 6

- 'What is a computer? If a computer is defined as anything that is not a person, computer programs, computers and computer programs are the most important examples of computers. If the definition of computer has to be changed from the original intent of the framers and as the laws of the United States and other jurisdictions have grown to cover everything a computer can do, some other things come to mind: The following are some of the definitions of computers that have been created and enforced in various jurisdictions - 5
- 'Who is Elon Musk? He's the founder and CEO of PayPal. He's also in charge of Tesla Motors and SpaceX, which, according to Forbes, have an estimated value of \$10 billion. Musk also has a few other ventures, too. He was the first person to get a patent for a self-healing car. He's also a big supporter of the "sharing economy" by charging for goods and services. When he's not building rockets, he's making - 6
- "Are you alive? Oh, hell yeah. What do you do here? That's right. I like to do my rounds, get out and see what it's like in the world. Do you have a boss? No. My father was a policeman. He went on to become a cop, and then I was born into it. Have your parents ever made you feel bad? Every now and then. They always put me - 6
- "Mad Max is a movie directed by George Miller and written by Paul W.S. Anderson in which the young Max Rockatansky, who is a prisoner of the Redneck rippers, travels to the Arctic to become the world's most famous

badass. It is set in the Canadian Arctic, some time in the early 20th century. It is heavily influenced by westerns and the 1950s. How does it rank? The movie won three Golden Globe awards, was nominated for - 6

- "What is an elephant? An elephant is a species of African rhinoceros and also known as the African elephant, or African elephant. It is the largest land mammal in Africa. The species, which is native to East Africa, is the largest land mammal in Africa with the largest horn in the world and is among the largest animals on the planet. In Africa, it lives in sub-Saharan Africa and the Democratic Republic of the Congo. A rhino's horn - 5
- 'What is a lemur? [edit] A lemur is a small, non-avian mammal native to sub-Antarctic regions of Africa, Southeast Asia, and southern South America. They are the most closely related of the five species of mammal which, collectively, are known as the Old World monkeys (the oldest monkey is a monkey-man, some of which are still around). There is also a population of the lemur living on the island of Madagascar (Vincent - 9
- 'How are you doing? We're waiting on your paperwork with the authorities." "I'm in a hurry," Harry said as the twins turned to look at him incredulously. "How did they manage to put a letter into your hand?" Tonks asked, obviously confused. Then she pointed at the two in the doorway. "Don't tell Sirius, of course." "Oh, right sorry," Harry agreed. "Sorry about that." After some tense moments - 9
- "Guess my weight was still low before I got married. I had a lot of body fat. It was really, really a struggle, but now it's starting to drop off a little bit. I feel more confident with my body and my voice. Advertisement AVC: How do you think that changing your body has affected the way people see you? HN: It's

definitely been a boost. I used to be the kind of person that I would never speak - 8

- 'Can you dance? Who can you dance with? Who can you dance with?" It's a very tricky thing to do. The problem is there are two people in this world and two ideas, and they clash, and that's not the way the world is supposed to be. It's supposed to be this place where two things can clash, where an idea can collide. I think this is the only way to do it. It's probably not going to be very popular, but I - 5
- 'Good day to you, sir." Lil Wayne came up to the man, got off, and walked off into the night. I waited a few minutes for him to disappear out of sight, but he didn't. I thought about calling the cops. I was sure it would be the last I'd see of him alive. When I had finally given up, I was so sick I couldn't even stand up, and in my panic, I got to the door and locked it. - 5
- "What is a computer? A computer is basically a computer that is programmed with the intention of operating independently of the human mind, just as your own brain is. The internet is a good example of this idea taken very far - it's not just a giant computer that is running a programming language called Java, but is actually running some algorithms on the internet, and the internet is really a computer. What are the biggest threats to our computers? With the internet, almost every computer that - 6
- "Who is Elon Musk? The self-described billionaire tech entrepreneur is best known for driving the solar system's most expensive rocket, his company SpaceX, which put a satellite in orbit, and for running Tesla Motors, which sells electric cars. Musk has a track record of being wildly disruptive, but his recent behavior was most notably associated with his relationship with one

woman. At a panel discussion at the Democratic National Convention in July, Musk was asked about a post he'd written on Twitter defending - 10

- "Are you alive? [Cocky laugh] Do you see any ghosts? [Cocky laugh] [Possible gasp and silence] I see nothing! [Wife's scream] Oh please help me! [Possible scream and silence] [Cocky laugh] Can you hear me? [Wife's scream] Please, stop that! [- 3
- "Mad Max is a movie that, unlike most of its counterparts, is not interested in exploring the moral implications of what happens in the face of extreme peril. It is certainly not interested in the philosophical implications of the story's events, the questions it explores, or even the implications of the consequences in question. It is about the action. It is about the chaos. It is about the violence. It is about the destruction. It is about the action, and about the story, and about what is - 10
- 'What is an elephant? It's a huge animal! An elephant is like a human. A human has skin and bones, but an elephant is more like a horse. (Also, the word "elephant" means "large elephant.") Elephants are intelligent. Elephants are powerful. Elephants are dangerous. Elephants are cute. Elephants are a great way to celebrate Easter. Elephants- 4
- "What is a lemur? A lemur, it's a member of a family of mammals in the genus Ambystoma. But what does it look like? The lemur looks a lot like a squirrel. The squirrel looks more like a large rabbit, while the lemur looks more like a small squirrel. This squirrel-like look really puts lemurs in the perspective of this other species. What is a lion? Leopard is a family of animals in - 5
- 'How are you doing? Did you just ask me if you could speak to me?" Yang Xiao Long asked as she took the chair opposite. She was in a good mood as

she sat across from the little girl. She was just about to ask the question that had been on her mind. "Actually, I have some questions..." Yang Xiao Long smiled, and smiled some more. She then turned to look back at the girl. "Why not tell me?" she asked - 9

- "Guess my weight is down again. What a waste? I'll do a little fudging and get back to you. Please don't ask me to take off anymore. I can't do that. -Sandra Dear Sandra, Thank you for the kind words and for your support. Please know that they have been most reassuring. While your new weight loss diet has been a success, I have experienced a change in your personality and your attitude. - 4
- 'Can you dance? Can you play guitar? Can you stand up on your own two feet and sing?" "I like my job," he says. He will be in his second full season as an Osprey, having signed a contract this summer and started training at the OSPIRY facility near Denver, where the team will open the exhibition season Sept. 11 against the Coyotes. He has played in five of the last six AHL regular-season games, averaging - 4
- 'Good day to you, sir." "It is a pleasure." "I wish I could have seen your hair, it is so beautiful you will take anything you wish." "I will take anything you wish but not anything worth having if I could get it. I have been on shore in the sea twice and I am quite full and quite exhausted. If I had not taken the mare you gave me to carry me I may have been quite drowned, but I believe - 6
- 'What is a computer? It's a piece of machinery that runs your computer. And what about that picture on the wall?" "I'm sorry, but I'm sorry. But I'm sorry!" "Come back up," he murmurs. "You've seen the movie. You can't deny the pictures that are being projected." "Sure I have. But when I'm done with those pictures, I want to do what I can to help the cause." - 4

- 'Who is Elon Musk? Read more He has developed a line of electric car companies - Tesla Inc, SolarCity Corp and Solar Roof - that has also created a significant demand for his services to run the business in the long term. Last month, Musk launched a \$2bn plan to build a city - called Boring Company - within the city of Pittsburgh, built on a large scale using low-cost, modular construction. Boring Company is currently undergoing construction as a test site - 7
- "Are you alive? I'm sorry. This is really getting old. Why doesn't anybody want to do this kind of work? I think people are just bored of it. How do we get people to care? Well, there's this new thing called the Internet, which's this weird, cool thing that you can access in any location that you want. The Internet actually has a few problems. It's not fast enough, and it's not secure. So you can buy these little computers out - 5
- 'Mad Max is a game set in the distant post-apocalyptic world of Max Payne, which began life as part of the first installment of the Max Payne series. It is a third person action game where players assume the role of two warring gangs, the Imperial Rangers and the Renegades. As a gang member, you will face off against the police, a powerful militia armed with machine guns, and the brutal gangs of the war-torn wastelands on three separate levels: the desert city - 5

GPU Accelerated

- 'What is an elephant? How are elephant herds structured? Are they autonomous or semi-autonomous? And what type of animal is most commonly kept in elephant herds? How is the elephant herd organised? Are there individual or collective elephant herds? How does the elephant herd work? Are wild elephants, as at present, managed in some way? What are the benefits of

keeping wild elephants? Do elephants have a purpose and benefit? The first thing to examine - 3

- 'What is a lemur? The lemur (*Eulemur*) is a mammal native to Madagascar. It is a closely related species to the pygmy lemurs of South Africa. Like the pygmy lemurs, the lemur does not hibernate, although it spends a lot of time underground. Lemurs are very intelligent primates that are capable of human language. There are over 40 species of lemurs living in the forests - 8
- "How are you doing? It's been a long time since we've gotten to spend some quality time together. It is my belief that the only person who can tell the difference between you and me in person is me. So I must speak to you without you realizing I'm there. I am taking time to reconnect with myself and my friends. I will be back in New Zealand in two weeks. We cannot wait to see each other again. It will shock you, no pun - 9
- 'Guess my weight? That's a question that's been with me since I was young. It's been on a rollercoaster of highs and lows for the past three years. Just ask my family." - 9
- 'Can you dance? "I would really love to do something, but I just don't have the right schedule." - 3
- 'Good day to you, sir," said a little black youth, standing over him. "Is there anybody here with me to buy a pair of shoes?" "I don't believe you," said the doctor. "I heard you had a little girl in a house of ill repute down here. You have no right to talk to her, sir. She's a slave, and you are no slaveholder." "Well, I don't," the boy said in a tone - 9
- "What is a computer? It is a device that uses electricity, and is used exclusively to communicate with the outside world. The machine that we call

a computer, works in a very simple manner. It was designed by the engineers at ATT Bell Laboratories. Computer is a device. And it uses electricity, and you can see how it works by looking at something like a computer. It's a device. Now, you might ask if that's really true. - 5

- "Who is Elon Musk? Elon Musk. Elon Musk is a billionaire business mogul who has built the world's biggest companies and has also been a leader in many of the technology disciplines he now embraces. As CEO of SpaceX and Tesla, Musk has developed advanced technologies that help people be more productive, sustainable, and happy. In 2005, SpaceX was the first company to reach space with the Falcon 9 rocket. In 2011, Musk became the first person to send a car - 10
- "Are you alive? No, I'm not. I was. Is that ok? I guess not, but I am pleased to hear that. You have a phone...Yeah...I have, but it doesn't work. Really? No, it's a fake. So...what do you have? I have the thing that does the math on everything, and it's really, really messed up - 5
- 'Mad Max is a series of video games designed by Avalanche Studios for the Super Nintendo Entertainment System, developed by Avalanche Studios and published by THQ. The games are set in a post-apocalyptic world ruled by the super-mutant Furiosa, a sentient, violent robot. It was directed by Sam Raimi and released for the Super Nintendo in 2004... - 4
- "What is an elephant? There should not be a problem with the elephant in the room. The elephant in the room is the fact that there will not be an adequate number of police officers on guard at the time any incident is in progress. That's true regardless of whether it is a protest or a demonstration, or any other type of event. How many security agencies are necessary? There

should be no problem with the number of security agencies with which to provide security - 4

- "What is a lemur? A lemur is just a small primate, just like us, but with a different body shape. The lemur's face is like a long mouth with a pointed nose, one of the most interesting features of this primate. It is believed that lemur's bodies are made up of different layers of muscle tissue. They are the same as humans except for the fact that their upper body has no muscles and they use their long tails for support. - 5
- "How are you doing? I am so proud! You did such a terrific job making the bed! Diana: I'm sorry I couldn't do it, but it was very difficult. Girard: Yeah, you could have, but you needed me. I'm glad you didn't have to leave me. Leanne: I know, I know...I was just getting a bit tired of the bed. Girard: What? - 3
- "Guess my weight (I lost 6 pounds since this diet) And I lost 4 inches on my tush (I think that's it) If you are not familiar with the term tush, it's very common for men to have erections or orgasms while in their pants, just like the girl in the picture above (in fact, women have been experiencing it for centuries) While the phenomenon is a well known one, there is a little known way of reversing that. - 3
- 'Can you dance? No, I'm not your dancing master and I'm not your dance teacher! You can't do this on your own! It's not fun at all!" "This is why I'm a teacher!" In response to her, Shiroe said at the beginning of episode 14, "I don't have any special power. I just make sure you stay strong, you know?" "You're my secret weapon?" On October 12, - 4
- 'Good day to you, sir!" she said, and then went into the kitchen, where she

made herself a glass of wine. "You will go away to-morrow evening," said the master of the house. "The maids are out, and she will go with you out, she will be with you until you go." The woman was delighted at that idea, and went off. The next day they were once more in the yard, and the master of the house - 9

- 'What is a computer? A computer is a device that makes use of computers in order to do a particular task. For example, we might use a computer to search for information on the Internet. A computer is considered "smart" if it makes good use of its operating system. The term "smart device" should not be confused with the concept of a computer, which is a type of device that makes use of other types of devices, such as a keyboard or mouse. There - 6
- 'Who is Elon Musk? He is CEO and co-founder of Tesla Inc. He co-founded PayPal and the online payments system, PayPal, but is best known for his space flight record-breaking feat of sending a semi-truck filled with satellites to Mars in just 21 hours. Musk, also the founder of SolarCity and Space Exploration Technologies, has a wealth of experience in technology that have helped propel his success. He also is an angel investor in the companies he is working on with Tesla - 7
- 'Are you alive? Who are you? Have you made a decision? Are you ready to go?" The entire day was filled with memories for me - and I had a feeling I'd be seeing more of them in 2018. - 3
- 'Mad Max is a visual novel developed and published by Feral Interactive. The game is available for PC, PlayStation 4 and Playstation Vita. It was released in Japan on October 2nd, 2016, and is being localized for the West. The game was also released in North America and Europe in 2016. Contents show]

Gameplay There are many different paths to take through the game, and some of them are quite dangerous! The story is a tale of a group of - 6

- "What is an elephant? An elephant is a large, stocky, hairy beast with a strong, muscular body, with a trunk like a bull's, and hoofs like a horse's. It is an herbivore and a large, strong animal. (Some even say an alligator.) It weighs about 3,500 pounds (1,852 kg), making it the largest living animal on the planet. As a member of the Carniv - 6
- "What is a lemur? Lemurs are about the same size as a cow or chicken. They are a mammal that inhabits Africa and Asia, but they also exist in southern Europe, the Middle East and South America. While some of them are native to Africa, others are in Asia. Lemurs are an animal that has long been recognized as having a unique evolutionary history. As you'll recall, the Neanderthals and our own species of Homo sapiens were - 5
- 'How are you doing? Are you feeling better?' "I'm not feeling any better," I replied, "I can't imagine anything is better if I'm walking in a snowstorm." "I understand, but I still think you could use a little help. How did you get your new coat?" "I got it by chance. A few people were in the shop for me and they thought I looked good. They brought me along and I was wearing the coat - 9
- "Guess my weight. It's a big question that is getting more and more popular every week. For example, one blogger claimed that he is 2 pounds lighter than we were when he started his blog. If you've been wondering how you've been doing on that scale, here's a short answer. How I keep track of my body weight I don't usually keep track of my weight in my blog because I don't tend to see the metric unit used much and - 6
- 'Can you dance? FRIEND: Of course. GUEST: (wearing a red, tight black

dress) TRAINER: (wearing a red, tight black dress) TRAINER: (wearing a red, tight black dress) GUEST: (wearing a red, tight black dress) DANCER: (wearing a red, tight black dress) TRAINER: - 3

- "Good day to you, sir, for you come to a strange town. WILSON Where? ALFRED To see my own mother. WILSON She is not here? ALFRED No, not here. (He turns away from her.) ALFRED Then where is my father? WILSON He's a stranger; the gentleman you have seen - 6
- 'What is a computer? If you ask a computer, "what is a computer?", you will get a confused face. So if you're asked a question like this, it's time to start looking into some answers...Computer A: What is a computer?! Computer B: Yes? A: A computer has some functions. It can send and receive and execute instructions and it has a processor. B: A processor? I can see how that is confusing. - 3
- "Who is Elon Musk? Elon Musk is the CEO and founder of Tesla Motors, Inc. and Space Exploration Technologies Corporation (SpaceX), the world's premier space exploration company. In December 2003, while at The University of Pennsylvania, he received his bachelor's degree in electrical engineering and computer science with a concentration in electrical engineering. After completing his training, he began working at Tesla Motors as a manufacturing engineer, first in Fremont, California, and later in Brooklyn, New York. - 10
- 'Are you alive? The answer is YES, and I need to get back to work!" I thought. The only way I can make this information public is if people can look back and see that it wasn't some complete hoax. So, I started to write all the posts together, then I asked a friend of mine to do the layout, I edited the post

up on Reddit, and the public started to see it. The whole thing was pretty entertaining, and it got a - 6

- "Mad Max is a film about women and men and the violence that they face. Because it's a movie, it's always the women who suffer. But they're not the only ones. They're just the ones who suffer the most. Men die in wars, women die in car wrecks, men die in murders. Some of the men and some of the women fall victim to the same kind of violence. But the women get the short end of the stick. This is a film - 6

REFERENCES

- [1] Flexera, “Cloud computing trends: 2021 state of the cloud report.” <https://www.flexera.com/blog/cloud/cloud-computing-trends-2021-state-of-the-cloud-report/>, 2021.
- [2] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, “Green ai,” *Commun. ACM*, vol. 63, pp. 54–63, nov 2020.
- [3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Rășnă, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, “On the opportunities and risks of foundation models,” 2021.
- [4] A. Radovanovic, R. Koningstein, I. Schneider, B. Chen, A. Duarte, B. Roy, D. Xiao, M. Haridasan, P. Hung, N. Care, S. Talukdar, E. Mullen, K. Smith, M. Cottman, and W. Cirne, “Carbon-aware computing for datacenters,” 2021.
- [5] G. V. Research, “Cloud computing market size, share and trends analysis report by service (iaas, paas, saas), by deployment (public, private, hybrid), by enterprise size, by end use (bfsi, it and telecom, retail and consumer goods), by region, and segment forecasts, 2022 - 2030.” <https://www.grandviewresearch.com/industry-analysis/cloud-computing-industry>, 2022.
- [6] “Climate-Friendly Information and Communication Technology.” https://www3.wipo.int/wipogreen/en/news/2020/news_0021.html.
- [7] J. Worland, “The pandemic remade every corner of society. now it’s the climate’s turn.” <https://time.com/5953374/climate-is-everything/>, 2021.

- [8] Netcraft, "March 2018 Web Server Survey."
<https://news.netcraft.com/archives/2018/03/27/march-2018-web-server-survey.html>, 2018.
- [9] B. Kepes, "30% of servers are sitting "comatose" according to research."
<https://www.forbes.com/sites/benkepes/2015/06/03/30-of-servers-are-sitting-comatose-according-to-research/>, 2015.
- [10] HostGatorSupport, "Cpu resource usage."
<https://www.hostgator.com/help/article/cpu-resource-usage>, 2018.
- [11] F. B. Insights, "Artificial intelligence (ai) market size, share & covid-19 impact analysis, by component (hardware, software, and services), by technology (computer vision, machine learning, natural language processing, and others), by deployment (cloud, on-premises), by industry (healthcare, retail, it & telecom, bfsi, automotive, advertising & media, manufacturing, and others) and regional forecast, 2020-2027."
<https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114>, 2020.
- [12] K. Quach, "Ai me to the moon...carbon footprint for training gpt-3 same as driving to our natural satellite and back." https://www.theregister.com/2020/11/04/gpt3_carbon_footprint_estimate/, 2020.
- [13] B. Everman and Z. Zong, "Greenweb: Hosting high-load websites using low-power servers," in *2018 Ninth International Green and Sustainable Computing Conference (IGSC)*, pp. 1–6, 2018.
- [14] B. Everman, N. Rajendran, X. Li, and Z. Zong, "Improving the cost efficiency of large-scale cloud systems running hybrid workloads - a case study of alibaba cluster traces," *Sustainable Computing: Informatics and Systems*, vol. 30, p. 100528, 2021.
- [15] B. Everman, M. Gao, and Z. Zong, "Evaluting and reducing cloud waste and cost - a data-driven case study from azure workloads," in *2021 12th International Green and Sustainable Computing Conference (IGSC)*, 2021.
- [16] PassMarkSoftware, "Cpu benchmarks - multiple cpu systems."
https://www.cpubenchmark.net/multi_cpu.html, 2018.
- [17] PassMarkSoftware, "Passmark cpu lookup." <https://www.cpubenchmark.net/cpu.php?cpu=Intel+Celeron+J3455+%40+1.50GHz&id=2875>, 2018.
- [18] Z. Abbasi, T. Mukherjee, G. Varsamopoulos, and S. K. S. Gupta, "Dahm: A green and dynamic web application hosting manager across geographically distributed data centers," *J. Emerg. Technol. Comput. Syst.*, vol. 8, nov 2012.

- [19] N. Deng, C. Stewart, D. Gmach, M. Arlitt, and J. Kelley, "Adaptive green hosting," in *Proceedings of the 9th International Conference on Autonomic Computing*, ICAC '12, (New York, NY, USA), pp. 135–144, Association for Computing Machinery, 2012.
- [20] Z. Al-Qudah, H. A. Alzoubi, M. Allman, M. Rabinovich, and V. Liberatore, "Efficient application placement in a dynamic hosting platform," in *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, (New York, NY, USA), pp. 281–290, Association for Computing Machinery, 2009.
- [21] Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang, and N. Gautam, "Managing server energy and operational costs in hosting centers," in *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '05, (New York, NY, USA), pp. 303–314, Association for Computing Machinery, 2005.
- [22] M. Elnozahy, M. Kistler, and R. Rajamony, "Energy conservation policies for web servers," in *Proceedings of the 4th Conference on USENIX Symposium on Internet Technologies and Systems - Volume 4*, USITS'03, (USA), p. 8, USENIX Association, 2003.
- [23] P. Bohrer, E. N. Elnozahy, T. Keller, M. Kistler, C. Lefurgy, C. McDowell, and R. Rajamony, *The Case for Power Management in Web Servers*, pp. 261–289. USA: Kluwer Academic Publishers, 2002.
- [24] O. Svanfeldt-Winter, S. Lafond, and J. Lilius, "Cost and energy reduction evaluation for arm based web servers," in *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, pp. 480–487, 2011.
- [25] B. Varghese, N. Carlsson, G. Jourjon, A. Mahanti, and P. Shenoy, "Greening web servers: A case for ultra low-power web servers," in *International Green Computing Conference*, pp. 1–8, 2014.
- [26] A. Pahlevan, J. Picorel, A. P. Zarandi, D. Rossi, M. Zapater, A. Bartolini, P. G. Del Valle, D. Atienza, L. Benini, and B. Falsafi, "Towards near-threshold server processors," in *2016 Design, Automation Test in Europe Conference Exhibition (14-18 March 2016)*, pp. 7–12, 2016.
- [27] A. Krioukov, P. Mohan, S. Alspaugh, L. Keys, D. Culler, and R. Katz, "Napsac: Design and implementation of a power-proportional web cluster," *SIGCOMM Comput. Commun. Rev.*, vol. 41, pp. 102–108, jan 2011.
- [28] K. K, "Wordpress stats: Your ultimate list of wordpress statistics (data, studies, facts - even the little-known)."
<https://www.codeinwp.com/blog/wordpress-statistics/>, 2021.

- [29] I. M. Murwantara, B. Bordbar, and L. L. Minku, "Measuring energy consumption for web service product configuration," in *Proceedings of the 16th International Conference on Information Integration and Web-Based Applications and Services*, iiWAS '14, (New York, NY, USA), pp. 224–228, Association for Computing Machinery, 2014.
- [30] ServMask, "All-in-one wp migration."
<https://wordpress.org/plugins/all-in-one-wp-migration/>, 2021.
- [31] SynologyInc., "Diskstation ds718+."
<https://www.synology.com/en-us/products/DS718+>, 2018.
- [32] CNET, "Synology diskstation ds1815 review."
<https://www.cnet.com/products/synology-disk-station-ds1815-plus-nas-server-0-gb/specs/>, 2018.
- [33] SynologyInc., "Dsm tutorial - shr."
<https://www.synology.com/en-us/knowledgebase/DSM/tutorial/>, 2017.
- [34] KyleMediaLLC, "Mac mini "core i5" 2.5 (late 2012) specs (late 2012, md387ll/a, macmini6,1, a1347, 2570): Everymac.com."
https://everymac.com/systems/apple/mac_mini/specs/mac-mini-core-i5-2.5-late-2012-specs.html, 2013.
- [35] AppleComputer, "About fusion drive, a storage option for some mac computers." <https://support.apple.com/en-us/HT202574>, 2020.
- [36] RaspberryPiLtd, "Buy a raspberry pi 3 model b - raspberry pi."
<https://www.raspberrypi.org/products/raspberry-pi-3-model-b/>, 2016.
- [37] CanonicalLtd, "Scale out with ubuntu server."
<https://www.ubuntu.com/server>, 2021.
- [38] RaspberryPiLtd, "Raspberry pi os."
<https://www.raspberrypi.org/downloads/raspbian/>, 2016.
- [39] SynologyInc., "Disk station manager."
<https://www.synology.com/en-us/dsm/6.1>, 2018.
- [40] Virtualmin, "Open source web hosting control panel."
<https://virtualmin.com/>, 2021.
- [41] CodeDocs, "Lamp (software bundle) - codedocs."
<https://codedocs.org/what-is/lamp-software-bundle>, 2021.
- [42] D. Opitz, "A benchmark script for php and mysql."
<https://github.com/odan/benchmark-php>, 2018.

- [43] J. Heyman, “Locust - a modern load testing framework.” <https://locust.io/>, 2021.
- [44] LocustIO, “Writing a locustfile - locust 2.5.0 documentation.” <https://docs.locust.io/en/stable/writing-a-locustfile.html>, 2021.
- [45] Y. Yoon, “A logging program for watts up? .net/pro meters..” <https://github.com/yyongpil/wattsup>, 2015.
- [46] Wikipedia, “Network time protocol - wikipedia.” https://en.wikipedia.org/wiki/Network_Time_Protocol, 2021.
- [47] SimilarWeb, “Similarweb traffic analytics..” <https://www.similarweb.com/>, 2021.
- [48] A. Shellhammer, “The need for mobile speed: How mobile latency impacts publisher revenue.” <https://www.doubleclickbygoogle.com/articles/mobile-speed-matters/>, 2016.
- [49] “Git - local-branching-on-the-cheap.” <https://git-scm.com/>, 2021.
- [50] “Google g suite.” <https://gsuite.google.com/>, 2021.
- [51] “The docker project - developers love docker. businesses trust it.” <https://www.docker.com/>, 2021.
- [52] Netcraft, “Cloud wars: Alibaba becomes 2nd largest hosting company.” <https://news.netcraft.com/archives/2017/08/22/cloud-wars-alibaba-becomes-2nd-largest-hosting-company.html>, 2017.
- [53] J. Russell, “Alibaba smashes its single’s day record once again as sales cross \$25 billion.” <https://techcrunch.com/2017/11/11/alibaba-smashes-its-singles-day-record/>, 2017.
- [54] I. Lunden, “Black friday racks up \$5.03b in online sales, \$2b on mobile alone.” <https://techcrunch.com/2017/11/11/alibaba-smashes-its-singles-day-record/>, 2017.
- [55] S. Klebnikov, “Alibaba’s 11/11 singles’ day by the numbers: A record \$38 billion haul.” <https://www.forbes.com/sites/sergeiklebnikov/2019/11/11/alibabas-1111-singles-day-by-the-numbers-a-record-38-billion-haul/#3eb0d20c2772>, 2019.
- [56] P. Wahba, “Walmart’s online sales surge during the pandemic.” <https://fortune.com/2020/05/19/walmart-online-sales-amazon-ecommerce/>, 2020.

- [57] K. Tyko, “Target digital sales make significant gains because of covid-19 demand, but in-store sales drop, driving shares down.”
<https://www.usatoday.com/story/money/2020/04/23/coronavirus-pandemic-target-online-growth-due-covid-19/3007311001/>, 2020.
- [58] B. Yungang, “How to reduce the cost of cloud computing data center?.”
<https://www.leiphone.com/news/201610/R0FXfe0mJlmueVT7.html>, 2016.
- [59] ChinaDaily, “Ali green intelligent data center settled in zhangbei, will become the heart of northern data.”
http://www.chinadaily.com.cn/interface/yidian/1120781/2016-09-12/cd_26769440.html, 2016.
- [60] J. Kaplan, W. Forrest, and N. Kindler, “Revolutionizing data center energy efficiency.”
https://sallan.org/pdf-docs/McKinsey_Data_Center_Efficiency.pdf, 2008.
- [61] L. Goasduff and C. Pettey, “Gartner says efficient data center design can lead to 300 percent capacity growth in 60 percent less space.”
<https://www.gartner.com/newsroom/id/1472714>, 2010.
- [62] H. Liu, “A measurement study of server utilization in public clouds,” in *Dependable, Autonomic and Secure Computing, IEEE International Symposium on*, (Los Alamitos, CA, USA), pp. 435–442, IEEE Computer Society, dec 2011.
- [63] Z. Zhang, “About alibaba cluster and why we open the data.”
<https://github.com/alibaba/clusterdata/wiki/About-Alibaba-cluster-and-why-we-open-the-data>, 2017.
- [64] Z. Zhang, C. Li, Y. Tao, R. Yang, H. Tang, and J. Xu, “Fuxi: A fault-tolerant resource management and job scheduling system at internet scale,” *Proc. VLDB Endow.*, vol. 7, pp. 1393–1404, aug 2014.
- [65] “Alibaba trace data.” <https://github.com/alibaba/clusterdata>, 2017.
- [66] C. Lu, K. Ye, G. Xu, C.-Z. Xu, and T. Bai, “Imbalance in the cloud: An analysis on alibaba cluster trace,” in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 2884–2892, 2017.
- [67] Q. Liu and Z. Yu, “The elasticity and plasticity in semi-containerized co-locating cloud workload: A view from alibaba trace,” in *Proceedings of the ACM Symposium on Cloud Computing*, SoCC ’18, (New York, NY, USA), pp. 347–360, Association for Computing Machinery, 2018.

- [68] Y. Cheng, Z. Chai, and A. Anwar, “Characterizing co-located datacenter workloads: An alibaba case study,” in *Proceedings of the 9th Asia-Pacific Workshop on Systems*, APSys ’18, (New York, NY, USA), Association for Computing Machinery, 2018.
- [69] C. Jiang, G. Han, J. Lin, G. Jia, W. Shi, and J. Wan, “Characteristics of co-allocated online services and batch jobs in internet data centers: A case study from alibaba cloud,” *IEEE Access*, vol. PP, pp. 1–1, 02 2019.
- [70] C. Lu, K. Ye, G. Xu, C.-Z. Xu, and T. Bai, “Imbalance in the cloud: An analysis on alibaba cluster trace,” in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 2884–2892, 2017.
- [71] F. Li and B. Hu, “Deepjs: Job scheduling based on deep reinforcement learning in cloud data center,” in *Proceedings of the 2019 4th International Conference on Big Data and Computing*, ICBDC 2019, (New York, NY, USA), pp. 48–53, Association for Computing Machinery, 2019.
- [72] H. Tian, Y. Zheng, and W. Wang, “Characterizing and synthesizing task dependencies of data-parallel jobs in alibaba cloud,” in *Proceedings of the ACM Symposium on Cloud Computing*, SoCC ’19, (New York, NY, USA), pp. 139–151, Association for Computing Machinery, 2019.
- [73] J. Guo, Z. Chang, S. Wang, H. Ding, Y. Feng, L. Mao, and Y. Bao, “Who limits the resource efficiency of my datacenter: An analysis of alibaba datacenter traces,” in *2019 IEEE/ACM 27th International Symposium on Quality of Service (IWQoS)*, pp. 1–10, 2019.
- [74] M. Schwarzkopf, A. Konwinski, M. Abd-El-Malek, and J. Wilkes, “Omega: Flexible, scalable schedulers for large compute clusters,” in *Proceedings of the 8th ACM European Conference on Computer Systems*, EuroSys ’13, (New York, NY, USA), pp. 351–364, Association for Computing Machinery, 2013.
- [75] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. Katz, S. Shenker, and I. Stoica, “Mesos: A platform for fine-grained resource sharing in the data center,” in *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation*, NSDI’11, (USA), pp. 295–308, USENIX Association, 2011.
- [76] A. S. Foundation, “Mesos architecture.”
<http://mesos.apache.org/documentation/latest/architecture/>, 2012.
- [77] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, O. O’Malley, S. Radia, B. Reed, and E. Baldeschwieler, “Apache hadoop yarn: Yet another resource negotiator,” in *Proceedings of the 4th Annual Symposium on Cloud Computing*, SOCC ’13, (New York, NY, USA), Association for Computing Machinery, 2013.

- [78] X. Zhang, Y. Lyu, Y. Wu, and C. Zhao, “Mixheter: A global scheduler for mixed workloads in heterogeneous environments,” *J. Parallel Distrib. Comput.*, vol. 111, pp. 93–103, jan 2018.
- [79] Google, “Google cluster data.” <https://github.com/google/cluster-data>, 2011.
- [80] A. Tumanov, J. Cipar, G. R. Ganger, and M. A. Kozuch, “Alsched: Algebraic scheduling of mixed workloads in heterogeneous clouds,” in *Proceedings of the Third ACM Symposium on Cloud Computing*, SoCC ’12, (New York, NY, USA), Association for Computing Machinery, 2012.
- [81] Q. Zhang, J. Hellerstein, and R. Boutaba, “Characterizing task usage shapes in google compute clusters,” in *Proceedings of the 5th International Workshop on Large Scale Distributed Systems and Middleware*, 2011.
- [82] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch, “Heterogeneity and dynamicity of clouds at scale: Google trace analysis,” in *Proceedings of the Third ACM Symposium on Cloud Computing*, SoCC ’12, (New York, NY, USA), Association for Computing Machinery, 2012.
- [83] Z. Liu and S. Cho, “Characterizing machines and workloads on a google cluster,” in *2012 41st International Conference on Parallel Processing Workshops*, pp. 397–403, 2012.
- [84] O. A. Abdul-Rahman and K. Aida, “Towards understanding the usage behavior of google cloud users: The mice and elephants phenomenon,” in *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, pp. 272–277, 2014.
- [85] B. Sharma, V. Chudnovsky, J. L. Hellerstein, R. Rifaat, and C. R. Das, “Modeling and synthesizing task placement constraints in google compute clusters,” in *Proceedings of the 2nd ACM Symposium on Cloud Computing*, SOCC ’11, (New York, NY, USA), Association for Computing Machinery, 2011.
- [86] P. Robichaux, “Calculating server power costs.” <http://www.itprotoday.com/windows-8/calculating-server-power-costs>, 2008.
- [87] D. Farber, “Google’s marissa mayer: Speed wins.” <https://www.zdnet.com/article/googles-marissa-mayer-speed-wins/>, 2006.
- [88] UKFast, “Google marissa mayer speed research.” https://www.youtube.com/watch?v=BQwAKsFmK_8, 2010.
- [89] Google, “Cluster scheduler simulator overview.” <https://github.com/google/cluster-scheduler-simulator>, 2012.

- [90] Dell, “Powerededge t340 tower server.”
<https://www.dell.com/bo/en/business/p/poweredge-t340/pd>, 2020.
- [91] R. Miller, “The cloud becomes a major force in green energy.”
<https://datacenterfrontier.com/the-cloud-becomes-a-force-in-green-energy/>, 2017.
- [92] MarketRealist, “Baba.” <https://marketrealist.com/ticker/baba/>, 2018.
- [93] ISG, “Is cloud pricing really usage-based? that depends on how close you look.” <https://isg-one.com/articles/is-cloud-pricing-really-usage-based-that-depends-on-how-close-you-look>, 2020.
- [94] G. Build, “Public cloud pricing, explained.”
<https://cloud.withgoogle.com/build/infrastructure/public-cloud-pricing-explained/>, 2020.
- [95] K. Weins, “Aws vs azure vs google cloud pricing: Compute instances.”
<https://www.flexera.com/blog/cloud/aws-vs-azure-vs-google-cloud-pricing-compute-instances/>, 2016.
- [96] D. Mytton, “The hidden costs of cloud.”
<https://davidmytton.blog/the-hidden-costs-of-cloud/>, 2019.
- [97] Microsoft, “Introduction to cloud service (classic) monitoring.”
<https://docs.microsoft.com/en-us/azure/cloud-services/cloud-services-how-to-monitor>, 2020.
- [98] S. Di, D. Kondo, and W. Cirne, “Characterization and comparison of cloud versus grid workloads,” in *2012 IEEE International Conference on Cluster Computing*, pp. 230–238, 2012.
- [99] N. T. Ho, M. Gribaudo, and B. Pernici, “Improving energy efficiency for transactional workloads in cloud environments,” in *Proceedings of the Eighth International Conference on Future Energy Systems, e-Energy '17*, (New York, NY, USA), pp. 290–295, Association for Computing Machinery, 2017.
- [100] A. K. Mishra, J. L. Hellerstein, W. Cirne, and C. R. Das, “Towards characterizing cloud backend workloads: Insights from google compute clusters,” *SIGMETRICS Perform. Eval. Rev.*, vol. 37, pp. 34–41, mar 2010.
- [101] M. Alalawi and H. Daly, “A survey on hadoop mapreduce energy efficient techniques for intensive workload,” in *Proceedings of the International Conference on Big Data and Internet of Thing, BDIOT2017*, (New York, NY, USA), pp. 62–66, Association for Computing Machinery, 2017.
- [102] F. Chen, J.-G. Schneider, Y. Yang, J. Grundy, and Q. He, “An energy consumption model and analysis tool for cloud computing environments,” in *2012 First International Workshop on Green and Sustainable Software (GREENS)*, pp. 45–50, 2012.

- [103] M. Qiu, Z. Ming, J. Li, K. Gai, and Z. Zong, "Phase-change memory optimization for green cloud with genetic algorithm," *IEEE Transactions on Computers*, vol. 64, no. 12, pp. 3528–3540, 2015.
- [104] S. Mamun, A. Gilday, A. Singh, A. Ganguly, G. Merrett, X. Wang, and B. Al-Hashimi, "Intra- and inter-server smart task scheduling for profit and energy optimization of hpc data centers," *Journal of Low Power Electronics and Applications*, vol. 10, p. 32, 2020.
- [105] Y. Wen, Z. Wang, Y. Zhang, J. Liu, B. Cao, and J. Chen, "Energy and cost aware scheduling with batch processing for instance-intensive iot workflows in clouds," *Future Generation Computer Systems*, vol. 101, pp. 39–50, 2019.
- [106] B. Everman, N. Rajendran, X. Li, and Z. Zong, "Improving the cost efficiency of large-scale cloud systems running hybrid workloads - a case study of alibaba cluster traces," *Sustainable Computing: Informatics and Systems*, vol. 30, p. 100528, 2021.
- [107] Google, "Carbon-aware computing for datacenters." <https://arxiv.org/abs/2106.11750>, 2021.
- [108] O. Hadary, L. Marshall, I. Menache, A. Pan, E. E. Greeff, D. Dion, S. Dorminey, S. Joshi, Y. Chen, M. Russinovich, and T. Moscibroda, "Protean: VM allocation service at scale," in *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, pp. 845–861, USENIX Association, Nov. 2020.
- [109] M. Shahrad, R. Fonseca, I. Goiri, G. Chaudhry, P. Batum, J. Cooke, E. Laureano, C. Tresness, M. Russinovich, and R. Bianchini, "Serverless in the wild: Characterizing and optimizing the serverless workload at a large cloud provider," in *2020 USENIX Annual Technical Conference (USENIX ATC 20)*, pp. 205–218, USENIX Association, July 2020.
- [110] E. Cortez, A. Bonde, A. Muzio, M. Russinovich, M. Fontoura, and R. Bianchini, "Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms," in *Proceedings of the 26th Symposium on Operating Systems Principles, SOSP '17*, (New York, NY, USA), pp. 153–167, Association for Computing Machinery, 2017.
- [111] Sudhanshu, "Who are biggest customers of the microsoft azure platform?." <https://www.cisin.com/coffee-break/Enterprise/who-are-biggest-customers-of-the-microsoft-azure-platform.html>, 2021.
- [112] Datadog, "How to monitor microsoft azure vms." <https://www.datadoghq.com/blog/how-to-monitor-microsoft-azure-vms/>, 2021.

- [113] Azure, “Azure public dataset.” <https://github.com/Azure/AzurePublicDataset>, 2020.
- [114] M. Azure, “Pricing - linux virtual machines.” <https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux>, 2021.
- [115] Google, “Google cloud pricing.” <https://cloud.google.com/pricing>, 2021.
- [116] Z. Zong, A. Manzanares, and X. Qin, “Ead and pebd: Two energy-aware duplication scheduling algorithms for parallel tasks on homogeneous clusters,” *IEEE Transactions on Computers*, vol. 60, no. 3, pp. 360–374, 2011.
- [117] R. Karamagi and S. Ally, “Analysis of wasted computing resources in data centers in terms of cpu, ram and hdd,” *American Journal of Computer Science and Technology*, vol. 8, p. 8, 10 2020.
- [118] P. Ranjan, “Swap space in windows azure virtual machines running pre-built linux images - part 1.” <https://azure.microsoft.com/en-us/blog/swap-space-in-windows-azure-virtual-machines-running-pre-built-linux-images-part-1/>, 2013.
- [119] M. Azure, “Azure functions trace 2019.” <https://github.com/Azure/AzurePublicDataset/blob/master/AzureFunctionsDataset2019.md>, 2020.
- [120] M. Labbe, “Energy consumption of ai poses environmental problems.” <https://www.techtarget.com/searchenterpriseai/feature/Energy-consumption-of-AI-poses-environmental-problems>, 2021.
- [121] Z. Susskind, B. Arden, L. K. John, P. Stockton, and E. B. John, “Neuro-symbolic ai: An emerging class of ai workloads and their characterization,” 2021.
- [122] K. Brain and S. Brain, *Artificial Intelligence on the Commodore 64: Make your micro think*. Sunshine Books, 1984.
- [123] T. J. O’Malley, *Artificial Intelligence Projects for the Commodore 64*. Tab Books, 1985.
- [124] B. Varghese, N. Wang, D. Bermbach, C.-H. Hong, E. de Lara, W. Shi, and C. Stewart, “A survey on edge performance benchmarking,” 2020.
- [125] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, “Carbon emissions and large neural network training,” 2021.
- [126] L. F. W. Anthony, B. Kanding, and R. Selvan, “Carbontracker: Tracking and predicting the carbon footprint of training deep learning models,” 2020.

- [127] W. Gao, J. Zhan, L. Wang, C. Luo, D. Zheng, F. Tang, B. Xie, C. Zheng, X. Wen, X. He, H. Ye, and R. Ren, “Data motifs: A lens towards fully understanding big data and ai workloads,” in *Proceedings of the 27th International Conference on Parallel Architectures and Compilation Techniques*, PACT ’18, (New York, NY, USA), Association for Computing Machinery, 2018.
- [128] J. Do, V. C. Ferreira, H. Bobarshad, M. Torabzadehkashi, S. Rezaei, A. Heydarigorji, D. Souza, B. F. Goldstein, L. Santiago, M. S. Kim, P. M. V. Lima, F. M. G. França, and V. Alves, “Cost-effective, energy-efficient, and scalable storage computing for large-scale ai applications,” *ACM Trans. Storage*, vol. 16, oct 2020.
- [129] R. Toews, “Deep learning’s carbon emissions problem.” <https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/>, 2021.
- [130] S. DeWeerd, “It’s time to talk about the carbon footprint of artificial intelligence.” <https://www.anthropocenemagazine.org/2020/11/time-to-talk-about-carbon-footprint-artificial-intelligence/>, 2020.
- [131] L. Aung, “Deep blue: The history and engineering behind computer chess.” <https://illuminate.usc.edu/deep-blue-the-history-and-engineering-behind-computer-chess/>, 2010.
- [132] “Openai.” <https://openai.com/>, 2022.
- [133] Shreyak, “Text generation using gpt3.” <https://becominghuman.ai/text-generation-using-gpt3-781429c4169>, 2021.
- [134] EleutherAI, “Eleutherai hugging face repository.” <https://huggingface.co/EleutherAI>, 2021.
- [135] B. Wang and A. Komatsuzaki, “GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.” <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- [136] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, “The Pile: An 800gb dataset of diverse text for language modeling,” *arXiv preprint arXiv:2101.00027*, 2020.
- [137] A. R. Naik, “Eleutherai’s gpt-j vs openai’s gpt-3.” <https://analyticsindiamag.com/eleutherai-s-gpt-j-vs-openai-s-gpt-3/>, 2021.
- [138] S. Black, G. Leo, P. Wang, C. Leahy, and S. Biderman, “GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow,” Mar. 2021.

- [139] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.
- [140] Huggingface, "Hugging face - gpt2." <https://huggingface.co/gpt2>, 2021.
- [141] J. Weizenbaum, "Eliza-a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, pp. 36–45, jan 1966.
- [142] C. Blakeney, X. Li, Y. Yan, and Z. Zong, "Parallel blockwise knowledge distillation for deep neural network compression," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 7, pp. 1765–1776, 2021.
- [143] Anaconda, "Minconda - conda documentation." <https://docs.conda.io/en/latest/miniconda.html>, 2021.
- [144] Pytorch, "Pytorch." <https://pytorch.org/>, 2021.
- [145] HuggingFace, "Huggingface transformers." <https://huggingface.co/docs/transformers/index>, 2021.
- [146] H. K. Kushwaha, "Running python script on gpu." <https://www.geeksforgeeks.org/running-python-script-on-gpu/>, 2021.
- [147] NVIDIA, "Cuda zone." <https://developer.nvidia.com/cuda-zone>, 2021.
- [148] J. A. Kolar, "A simple guide to setting the gpt-3 temperature." <https://algowriting.medium.com/gpt-3-temperature-setting-101-41200ff0d0be>, 2020.
- [149] M. G. (ETS), R. F. (ETS), L. M. (Westat), T. K. (Westat), and J. H. (Westat), "Literacy, numeracy, and problem solving in technology-rich environments among u.s. adults: Results from the program for the international assessment of adult competencies 2012." <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014008>, 2012.
- [150] A. Wylie, "What's the latest u.s. literacy rate?." <https://www.wyliecomm.com/2021/08/whats-the-latest-u-s-literacy-rate/>, 2021.
- [151] T. E. Agency, "Staar english i and english ii resources." <https://tea.texas.gov/student-assessment/testing/staar/staar-english-i-and-english-ii-resources>, 2021.
- [152] T. E. Agency, "English i expository writing rubric." <https://tea.texas.gov/sites/default/files/Rubric-EOC-Eng1-WrtgExpository.pdf>, 2011.

- [153] T. E. Agency, “English i expository scoring guide.”
<https://tea.texas.gov/sites/default/files/2021-staar-english-1-scoring%20guide-tagged.pdf>, 2021.
- [154] G. S. Foundation, “Software carbon intensity (sci) specification.”
https://github.com/Green-Software-Foundation/software_carbon_intensity/blob/main/Software_Carbon_Intensity/Software_Carbon_Intensity_Specification.md, 2021.
- [155] Carbonfootprint.com, “2019 grid electricity emissions factors.”
https://www.carbonfootprint.com/docs/2019_06_emissions_factors_sources_for_2019_electricity.pdf, 2019.
- [156] Boavizta, “Digital and environment: How to evaluate server manufacturing footprint, beyond greenhouse gas emissions?.”
<https://www.boavizta.org/en/blog/empreinte-de-la-fabrication-d-un-serveur>, 2021.
- [157] Boavizta, “Environmental footprint data.”
<https://github.com/Boavizta/environmental-footprint-data/blob/main/boavizta-data-us.csv>, 2022.
- [158] AMD, “Amd ryzen threadripper 2950x processor.”
<https://www.amd.com/en/products/cpu/amd-ryzen-threadripper-2950x>, 2022.