ENERGY SAVINGS WHEN MIGRATING WORKLOADS TO THE CLOUD

by

Yan Zheng

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Electrical and Computer Engineering

Spring 2017

ENERGY SAVINGS WHEN MIGRATING WORKLOADS TO THE CLOUD

by

Yan Zheng

Approved: ______________________________

Stephan Bohacek, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: ______________________________

Kenneth E. Barner, Ph.D.
Chair of the Department of Electrical and Computer Engineering

Approved: ______________________________

Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: ______________________________

Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

I wish to express my deepest appreciation from the bottom of my heart to my advisor, Dr. Stephan Bohacek, for his supervision and encouragement. I would also like to thank my friends for their support, in particular Chao Gao, Sue Quick, Tim Slothower, and Kelly Hess. Finally, I must thank my parents, Mrs. Xianping Yu and Mr. Ji Zheng, without whom none of this would ever have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACPI	Advanced Configuration and Power Interface
AWS	Amazon Web Services
CE	Computational Efficiency
CSPs	Cloud Service Providers
DC	Data Center
IaaS	Infrastructure-as-a-Service
IPR	Idle-to-Power Ratio
PaaS	Platform-as-a-Service
PUE	Power Usage Efficiency
SaaS	Software-as-a-Service
SPEC	Standard Performance Evaluation Corporation
TDP	Thermal Design Power
VM	Virtual Machine

ABSTRACT

It is well known that the data centers used by cloud service providers (CSPs) are among the most efficient data centers in terms of energy usage. Consequently, migrating workloads to the cloud can result in a decrease in energy usage. This paper presents results based on a large data set of over 40,000 machines (virtual and physical) spread across over 300 data centers. With this data we quantify the energy savings and the sources of the energy savings. We focus on lift-and-shift migration along with optimal cloud instance size selection, as this type of migration is relatively straightforward. The data indicates that this type of migration should reduce energy usage by an average factor of 4.5 and 7.8. Relatively little of the energy savings is from the efficiencies of CSPs data centers related to efficient cooling and lighting. Instead, most the savings are from using modern CPUs and by correctly sizing the instances so that systems are not underutilized. We also consider potential energy savings from refactoring applications to make use to auto-scaling. While such refactoring has the potential to achieve considerable energy savings, the savings are likely to be less than what is achieved by the initial migration to the cloud. These findings contract the popular belief that one needs to modify applications in order to achieve the benefits of the cloud.

Chapter 1

INTRODUCTION

1.1 Data Center Energy Consumption Overview

Data Center (DC) energy consumption continues to grow in the previous two decades. According to the most recent report [4], US data centers consumed estimated 91 billion kilowatt-hours of electricity, or 1.8% of total U.S. electricity consumption, in 2013 [5]. This is nearly a 1.5 times increase compared to the year 2000 [2]. The estimated annual cost of electricity is likely to be close to $9 billion. On account of the enormous researches on DC efficiency improvements, the annual growth decreased to about 4% since 2010. The report [4] predicts the energy usage will continually increase at this rate until 2020.

However, in addition to improving data center energy efficiency, migrating data to the cloud has great potential to reduce energy consumption on a larger scale, which would result in benefit gains in both environment and economic.

According to a recent report [5], the hyper-scale cloud computer servers use less than 5% of data centers energy to perform a much better efficiency than the other 95% of small, medium, corporate and multi-tenant operations. It is also analyzed that, one of the main challenges of data center energy efficiency is the low server utilization. The average industry server utilization is between 12 to 18 percent, while hyper-scale cloud providers can realize 40 to 70 percent utilization rate [5][6]. In addition, the cloud servers are renewed twice to three times more frequently than the other servers.

1.2 Could Computing Energy Saving Overview

The cloud computing paradigm, specifically, IaaS, offers a wide range of benefits over running workloads in traditional data centers. Benefits include the ease of deployment, highly redundant systems, and access to a globally deployed infrastructure. Todayś cloud service providers (CSPs) have developed and deployed highly efficient computing systems, with many components of the data centers being developed by the CSP [7]. For example, Intel provides custom CPUs that are optimized for the CSP [8]. The result is that the data centers used by CSPs are likely to be among the most efficient data centers in the world, and are far more efficient than data centers used in traditional non-cloud computing. One important implication of the efficiencies of the data centers used by CSPs as compared to traditional data centers is that energy can be saved by moving workloads from traditional data centers to the cloud.

However, cloud migration research is still in early stages of maturity [9]. Our study aims to make contributes for improving the maturity level and consequently trust into cloud migration. Especially, this thesis provides a deep analysis on cost savings of cloud migration.

1.3 Related Work

Cost saving is one of the most important motivations of migrating data from traditional data center to the could. For enterprises to use cloud computing, they normally use case studies to evaluate the benefits, risks and effects of cloud computing on their typical organizations. There are currently many case studies that look into the migration of existing IT systems to the cloud. Armbrust [10] states that elasticity and transference of the risks of over-provisioning and under-provisioning are the important economic benefits of cloud computing. Walker [11] compares the CPU hours of cloud computing and IT infrastructure, and propose an optimal expect CPU utilization of 40 percent. Another case study also shows that the migration of an IT system from an in-house data center to the cloud reduced 37% cost over 5 years[12].

However, these studies are all based on economical or business perspective to help with decision making. There are very few researches investigate into quantifying energy saving of cloud migration.

Lawrence Berkeley National Laboratory provide an open-access energy-efficiency model, Cleer[13], to analyze the estimated energy savings when moving current business software usage to the cloud. The study concludes that the potential primary energy savings from current use to cloud-based use is about 23 billion kWh/yr, which results in an 87% reduction. However, this model only provides the estimation under the three business software use scenarios, including customer relationship management, productivity, and email software. Yet the system specifications are not under consideration. Also, this evaluation only focus on total energy consumption, which is indeed useful, but energy efficiency is also important.

1.4 Objective

We utilize a large data set of several tens of thousands of workloads running in several hundred data centers. With this data, we quantify the energy savings that could be achieved by moving the workloads to the cloud. While the energy savings identified in this thesis are qualitatively known, to the best of our knowledge, this is the first large study to quantify the energy savings. A key conclusion is that energy usage can be reduced by a factor of 4.5 to 7.8 by simply moving workloads to the cloud and selecting optimal instance types, without refactoring the software.

The remainder of the thesis is as follows. Chapter 2 introduces definitions of Cloud Computing, Data Migration and Energy Efficiency.

In Chapter 3, we present the mythology for the evaluation of energy saving. This chapter begins with an overview of the energy reduction of migration to the could, and discusses the data set used in this study. The following sections evaluate the energy saving with different perspective.

In the Section 3.3 , we quantify the energy savings using a lift-and-shift migration, in particular the savings that result from using the CPUs used by CSPs as opposed to CPUs currently used in traditional data centers.

In Section 3.4, we examine CPU utilization and develop a new model for energy usage as a function of CPU utilization.

Section 3.5 quantifies the energy saving that could be achievable by selecting an instance size that meets performance goals, while Section 3.6 studies the case where the instance type is rarely changed, examines the potential energy savings of the computational abilities are frequently adjusted through auto-scaling.

It is well known that data centers used by CSPs are using highly efficient cooling, lightly, and other non-computing systems. Chapter 3.7 provides a summary of findings to help quantify this type of energy savings.

And finally, Chapter 4 provides concluding remarks.

Chapter 2

BACKGROUND AND DEFINITIONS

2.1 Cloud Computing

Cloud is the hardware, software and network that can provide computing resources. Cloud Computing is a method that deliver the resources via the network. The CSPs are the companies that provide the cloud service, which enables the user to use the service as needed, like a utility.

Compared to the traditional DC, cloud computing can save the users from building and maintaining computing infrastructures by themselves. It allows the users to access the computational resources based on demand by offering different levels of services: Infrastructure-as-a-Service(IaaS), Platform-as-a-Service(PaaS), and Software-as-a-Service(SaaS).

2.1.1 SaaS

In the SaaS model, users access the applications with a client interface via Internet. CSPs response for the installation, operation and maintenance of the application. Email is an instance of SaaS.

2.1.2 PaaS

Compare to SaaS, PaaS provide more privilege on a platform, which allows the user to develop new application with the hardware and software that CSPs provide.

2.1.3 IaaS

The CSPs offer more fundamental computing infrastructure, including servers, storage, network and operating system to the users in IaaS. The user has the ability to control these computational resources without the efforts to construct and manage them. Our study focus on the energy savings of IaaS.

2.1.4 Auto-Scaling

Auto-Scaling is a cloud computing strategy that allows users to automatically scale out and in as demand for the services provided by the workload grows and shrinks.

Auto-scaling is implemented that a single workload is spread across multiple virtual machines. Then, when demand for the services provided by the workload is high, many virtual machines are spawned, and as the demand decreased, the number of virtual machines running the workload is reduced. Not only does this approach reduce the cost of running the workload, it also reduces the energy consumed by the workload. We evaluate the energy saving of auto-scaling in Section 3.6. Cloud computing providers, such as Amazon Web Services (AWS) [14], offer this feature.

2.1.5 Data Migration

Data migration is the process of transferring data between computer storage types or file formats. Migrating data from the companies' data centers to the cloud provides significant benefits, including agility, efficiency, performance improvements, and cost savings. There are different data center migration strategies. We introduce lift-and-shift method in this study.

2.1.6 Lift-and-Shift

Lift-and-shift migration refers to simply moving an existing workload software, without significant changes, to cloud hardware.

It is critical to note that lift-and-shift migration is relatively straightforward. The chief requirement is to transfer data, executables, and other files to the cloud-based machines. There are a wide range of tools that automate this type of migration [15][16][17]. We evaluate the energy saving of Lift-and-Shift in Section 3.3.

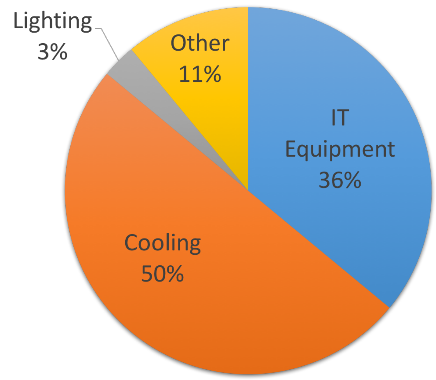

Figure 2.1: A Typical Data Center Energy Breakdown[1]. The IT equipments and cooling infrastructure consume the major energy.

2.2 Power Consumption and Energy Efficiency

2.2.1 Data Center Power Consumption

The energy consumed by a DC can be divided into multiple parts, including server and storage, cooling, power conversion, network and lighting. According to previous researches [18][1], the server and the cooling consume about 80% of the total power of a DC (See Fig 2.1).

2.2.2 Data Center Energy Efficiency

Power Usage Effectiveness (PUE) is defined as the ratio of the total data center energy consumption divided by the energy consumption of the IT equipment, as described in Equation 2.1. PUE accounts for the effectiveness of the energy usage by the IT infrastructure as well as the overhead consumed by the cooling, lighting and the other data center infrastructures.

$$PUE = \frac{Total\ Data\ Center\ Energy\ Consumption}{IT\ Equipment\ Energy\ Consumption} \tag{2.1}$$

A PUE of 1 would mean that all energy is consumed by computing equipment only. However, cooling is a critical component of modern IT, and so PUE is always larger than 1. Section 3.7 compares the PUE of traditional data center and cloud.

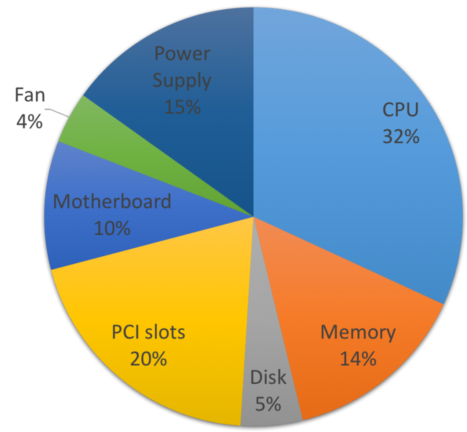

Figure 2.2: A Typical Server Energy Breakdown in the Data Center[2]. CPU accounts for the major share of the energy consumption.

2.2.3 Processor Power Consumption

[2] estimates the power for a typical server, in which the processor is the major power power consume component (See Fig 2.2). Note that the values may different from sever to sever.

2.2.4 Peak Power

Peak power is the maximum power dissipated by the processor under the worst case conditions - at the maximum core voltage, maximum temperature and maximum signal loading conditions.

2.2.5 Thermal Design Power

Thermal Design Power (TDP) is the average maximum power in watts the processor dissipates when operating at *Base Frequency* with all cores active under a manufacturer-defined, high-complexity workload. This parameter is used for designing a components cooling system.

For each CPU, Intel provides the TDP, which estimates the heat that needs to be extracted from the CPU, which is same as the energy consumed by the CPU. Note that TDP is not the peak energy consumed by the CPU, but is the long-term energy usage[19].

Some research[20] states that TDP is usually 20%-30% lower than the CPU maximum power dissipation.

2.2.6 Processor Energy Efficiency

Performance Per Watt measures the rate of computation that can be delivered by a computer for every watt of power consumed. It is used to compare the energy efficiency of a particular computer architecture or computer hardware. In this thesis, we define *Computational Power* (CE) (See section 3.3.1) to evaluate processor's energy efficiency.

A CE of 1 represents the ideal case that all consumed power contributes to the computations. Higher CE indicates higher efficiency and vice versa. As mentioned in Section2.2.5, TDP is not reflecting the actual maximum power of the CPU, which means the measured power consumption could exceed TDP. In other words, CE is possibly larger than 1.

2.3 Benchmark

Measuring performance components is difficult for most users. It requires detailed knowledge of the hardwares implementation, such as simulation, hardware counters, profiling tools, etc. To overcome these difficulties, benchmark enables easy comparison of different system by providing standardized measurements or evaluations.

A computer benchmark is typically a computer program that performs a strictly defined set of operations and returns some form of result describing how the tested computer performed[21].

The Standard Performance Evaluation Corporation (SPEC) benchmarks are widely used to evaluate the performance of computer systems, and published on the SPEC website.

SPEC currently provides more than 20 SPEC benchmarks, in which we use SPEC CPU2006 and SPECpower_ssj2008 in our study.

2.3.1 SPEC CPU2006

SPEC CPU2006[21] focuses on computing intensive performance. This benchmark is widely used to compare processor capabilities.

It includes two benchmark suites: CINT2006, which is used for measuring and comparing compute-intensive integer performance; and CFP2006, used for measuring and comparing compute-intensive floating point performance. Both of them provide the Speed Metrics and throughput metrics (named Rate Metrics). Each metrics has two types of compilation: the base metrics and the peak metrics, in which the base metrics have stricter requirements than the peak metrics.

We use SPECfp2006 Rate base results in our study.

2.3.2 SPECpower_ssj2008

SPECpower_ssj2008[3] evaluates the power and performance characteristics of single server and multi-node servers. It is used to compare power and performance among different servers and serves as a tool set for use in improving server efficiency. The benchmark seeks to measure the relationship between energy usage and CPU utilization. While running the benchmark, the system energy is measured while running a Java-based program that seeks to keep the CPU utilization at a target value. The test is repeated 11 times, where the target CPU utilization varies from 0% (active idle) to 100% (fully utilized).

Chapter 3

METHODOLOGY AND EVALUATION

3.1 Overview

Figure 3.1 provides an overview of the energy savings, the sources of the energy savings, and the tasks required to achieve the energy savings. As shown, we divide the task of migrating workloads to the cloud into three parts, namely, lift-and-shift, optimal instance sizing, and rewriting the application to take advantages of cloud services.

By Lift-and-shift migration, we seek to utilize the exact same computational capabilities in the cloud as the workload is using in the data center. As a result, if the peak CPU utilization in the data center is 50%, then, after lift-and-shift migration, the peak CPU utilization would still be 50%. As a result, we expect that a lift-and-shift migration will achieve energy savings from two sources, namely from the fact that the data centers used by CSPs have less energy waste for cooling, lighting, and other non-computing tasks, and from using more efficient CPUs.

It is well known that low system utilization is an important source of energy waste[5]. Using the CPU utilization measurements in the data set, we find that when combined with using more efficient CPUs, the average energy usage is decreased by around 75%. Again, achieving this reduction in energy usage by optimizing system utilization is relatively straightforward. For example, there exists tools that optimize cloud infrastructure [22].

Lastly, we examine the energy savings that might be achieved when the auto-scaling is utilized. We find that auto-scaling could reduce the energy usage by another factor of 3 or more. We utilize two models to estimate the potential energy reduction form taking advantage of auto-scaling. However, these models only estimate the potential energy reduction. Achieving this potential depends on how suitable the application

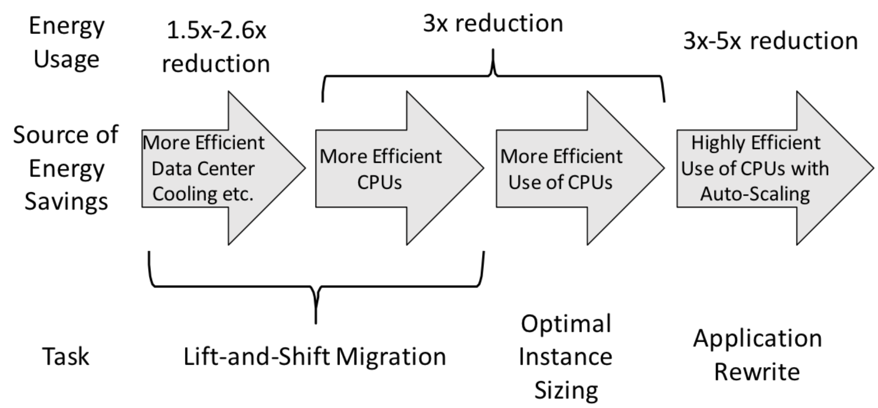

Figure 3.1: Energy Reduction from Migrating to the Cloud and Optimizing Usage of the Cloud Services

is to auto-scaling. Many applications are not suitable for auto-scaling, and hence no energy reduction could be achieved. Moreover, even when an application is suitable for auto-scaling, refactoring an application to support auto-scaling might be a labor intensive task. Therefore auto-scaling, while feasible, is not realistic. On the other hand, we include these values to understand the range of energy reduction.

Specifically, we find significant energy reduction in the straightforward exercise of lift-and-shift migration along with infrastructure optimization. Energy reduction from auto-scaling, while significant, is likely to be less than the initial savings achieved by the migration. This finding differs from the somewhat popular notion that applications need to be re-factored in order to make them cloud suitable and take advantage of the cloud [23].

The key findings of this chapter are the following.

- We quantify the energy savings that results from moving workloads from CPUs currently deployed in data centers to the CPUs used by CSPs
- We quantify the energy savings that results from utilizing optimally sized cloud instance
- We develop a new model for energy usage as a function of CPU utilization that is appropriate for Intel Xeon processors released between 2013 and 2016
- We quantify the potential energy savings from utilizing auto-scaling

3.2 Summary of the Data Set

The data used in this study was collected by Cloudamize Inc.[22] from May 2016 to August 2016. This data set includes measurements from over 40,000 machines (virtual and physical) in approximately 300 data centers. Note that in each data center, not every machine was necessarily monitored. Data was collected for a minimum of 14 days with an average data collection lasting 21 days. For each machine, a wide range of performance metrics where measured, including CPU utilization, which was collected once every 30 seconds or once every 20 seconds, depending on the system. A wide range of information about the underlying hardware was collected, including details about CPU. Finally, network usage information was also collected, including the source and destination IP addresses of packets.

Cloudamize is a business that help companies migrate systems to the cloud and manage systems already on the cloud. The data set used in the study is from businesses that have engaged with Cloudamize or with a partner of Cloudamize. Nearly every business that utilizes traditional data centers are considering migrating to the cloud, because they are actively evaluating migrating to the cloud or beginning to migrate workload. Consequently, the data collected is neither a random sample of data centers nor from the random machines within data centers. Instead, the data is from data centers of companies that are interested in migrating workloads to the cloud.

As a result, the data could be biased toward workloads that are especially suitable or in someway been deemed to require migration to the cloud. While confirmation is difficult, Cloudamize finds that business evaluate moving to the cloud for a large number of reasons, but mostly, the reasons stem from business concerns such as reducing costs, improving application development cycle, improving agility, reducing business focus on maintaining computing infrastructure, and improving focus on core business areas. Therefore, we conclude that the data is useful for drawing conclusions regarding computing in traditional data centers.

Table 3.1: Major Machine Instances in CSPs

CSPs	Machine Types	Code Name	Launch Year
Google	2.6GHz Intel Xeon E5	Sandy Bridge	2012
Google	2.5GHz Intel Xeon E5 v2	Ivy Bridge	2013-2014
Google	2.3GHz Intel Xeon E5 v3	Haswell	2014
Google	2.2GHz Intel Xeon E5 v4	Broadwell	2016
AWS	2.4GHz Intel Xeon E5 v3	Haswell	2015
AWS	2.3GHz Intel Xeon E5 v4	Broadwell	2016
AWS	2.5GHz Intel Xeon E5 v2	Ivy Bridge	2013-2014
AWS	2.9GHz Intel Xeon E5 v3	Haswell	2014
AWS	2.8GHz Intel Xeon E5 v2	Ivy Bridge	2013
AWS	2.6GHz Intel Xeon E5	Sandy Bridge	2012
AWS	2.3GHz Intel Xeon E7 v3	Haswell	2015
Azure	2.6 GHz Intel Xeon E5	Sandy Bridge	2012
Azure	2.3GHz Intel Xeon E5 v4	Broadwell	2016
Azure	2 GHz Intel Xeon E5 v3	Haswell	2015
Azure	3.2GHz Intel Xeon E5 v3	Haswell	2014

3.3 Energy Usage Change When Performing Lift-and-Shift Migration

As mentioned, a lift-and-shift migration attempts to utilize cloud computing resources that have the exact same computational capabilities as the on-premise hardware. Therefore, the change in energy usage from a lift-and-shift migration is a result of utilizing more (or less) computationally efficient hardware in the cloud. In this section, we first quantify and evaluate computational efficiency. Then, we use the computational efficiency to estimate the change in energy usage that results from a lift-and-shift migration.

Over the past several decades, the computational abilities of CPUs have been rapidly increasing, while the power consumed by the CPUs has remained more stable. For example, over the past decade, CPUs include dynamic voltage scaling and ACPI (Advanced Configuration and Power Interface) have allowed energy usage to decrease. Data centers and CSPs periodically refresh their hardware in order to utilize computationally efficient hardware. CSPs frequently offer new families of instance types, which is similar to a hardware refresh. For example, in the past year, Azure has announced four new families, namely the a*v2, d, h, and n families. Table 3.1 shows that the major CPUs used by the 3 popular CSPs are of 2012 to 2016.

The refresh rates for data centers appear to be variable and less frequent. For example, [5] suggests that data center hardware is refreshed every three years, but it is likely that many data centers refresh hardware less frequently. The impact of infrequent hardware refreshes is that data centers utilize less computationally efficient hardware than what is available on the cloud. Consequently, migrating workloads to the cloud can improve computational efficiency and reduce energy usage while achieving the same performance.

To quantify the change in energy usage when migrating to the cloud, we consider the special case of "lift-and-shift," which we define as the exercise of moving a workload to the cloud without any changes in the software or computational capabilities allocated to the workload. Note that while we seek to keep the computational abilities fixed, the computational efficiency is likely to change. For example, if the on-premise hardware utilizes an old CPU, the same computational ability might be achievable with a single core on a new 20 cores CPU. Amortizing the energy used by the hardware across all cores, this single core might use considerably less energy than the old CPU.

3.3.1 CPU Computational Efficiency

We utilize the SPECCPU2006 benchmark[21] to quantify the computational ability of a CPU. As mentioned before, this benchmark is widely used and has been evaluated on nearly every CPU released by Intel. In many cases, a single CPU is evaluated many times on different systems. In such instances, we use the median of the base results.

We define the *computational efficiency* (CE) (Section 2.2.6) of a CPU to be the ratio of the SPECCPU2006 benchmark and the TDP of the CPU. Figure 3.2 shows the CE for Intel Xeon processors included in this data set. The newest CPU observed is Intel Xeon E5-4650 v3, which is released in 2015. The oldest CPU observed is Intel Pentium III Processor(1.00 GHz, 256K Cache, 133 MHz FSB), which is released

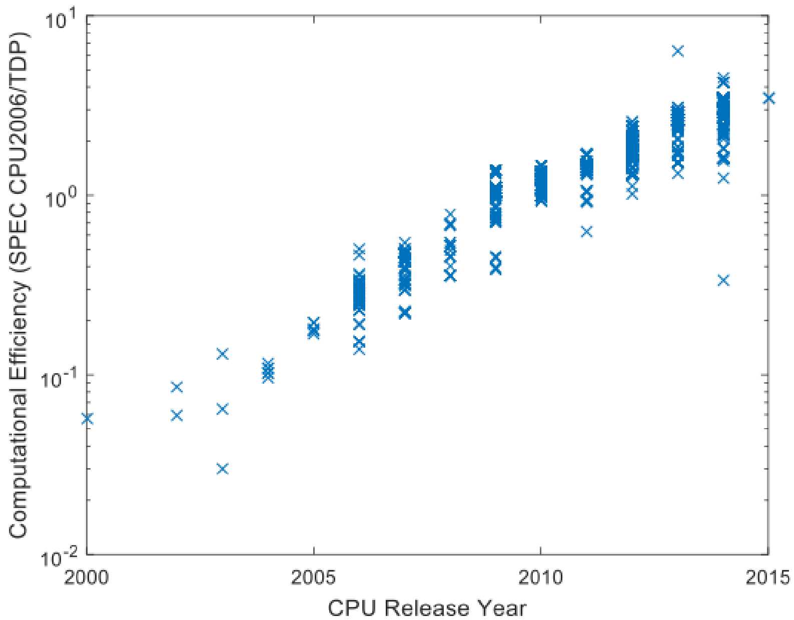

Figure 3.2: Computational Efficiency (CE) for All Processors Measured

in 2000. As expected, CE widely varies over the past 15 years. However, while [5] indicates a hardware refresh occurs every three years, Figure 3.2 shows that CPUs found in data centers can be significantly older, and less efficient. On the other hand, Figure 3.2 also shows that a significant number of CPUs have been recently released and are likely to have state-of-the-art CE.

3.3.2 Computational Efficiency of Cloud Instances

To quantify the change in energy usage when performing a lift-and-shift migration, we need to quantify the CE currently being offered by CSPs. However, CSPs often utilize custom CPUs, for which Intel does not provide official product specification. We use TDP of the CPUs with the most similar specifications compared to the custom CPUs. Table 3.2 shows the CPUs used by AWSs current generation of instances. The CPUs used by AWS are listed in various places in the AWS documentation[24]. Note that older instances families such as M1, M2, and M3 can use older CPUs, and

Table 3.2: Performance and Energy Usage in Processors used by AWS

Custom CPU	Launch Date	Similar to	CFP2006	TDP	CE
Xeon E5-2676 v3	Q2'15	Xeon E5-2680 v3	382.2	120	3.2
Xeon E5-2686 v4	Q2'16	Xeon E5-2697 v4	428.8	135	3.2
Xeon E5-2670 v2	Q4'13	-	301	115	2.6
Xeon E5-2666 v3	Q4'14	Xeon E5-2660 v3	345.9	105	3.3
Xeon E5-2680 v2	Q4'13	-	310	115	2.7
Xeon E7-8880 v3	Q2'15	-	460.7	150	3.1

machines with GPUs currently use older CPUs. However, this study focus on new migrations, which would utilize the latest generation of instances offered by CSPs.

As indicated in Table 3.2, several of the CPUs used by AWS are custom CPUs. For these cases, we search for a similar CPU which the clock rate, number of cores, and cache size is similar to the CPU used by AWS. These similar CPUs are also listed in the table. The average CE over all CPUs is approximately 3. We use this value as the CE for CSPs.

3.3.3 Change in Energy Usage when Performing Lift-and-Shift Migration

Let $CE_{datacenter}$ be the CE for the on-premise machine and CE_{cloud} be the CE for the cloud hardware. Then the change in energy usage for a lift-and-shift migration is estimated as a ratio $CE_{datacenter}/CE_{cloud}$. For example, if $CE_{datacenter} = 1$ and $CE_{cloud} = 2$, then the on-premise machine requires twice the energy to achieve the same computation as the cloud hardware. Therefore, migrating the workload to the cloud will reduce the energy by a factor of 0.5. Hence, the ratio estimates the fraction of energy that will be used after a workload is migrated to the cloud via a lift-and-shift migration, as compared to the energy used when the workload is hosted in the data center.

Figure 3.3 shows 2 curves of the cumulative distribution of this ratio. The upper curve is the distribution of the change in energy usage over all machines in the data set, while the other curve shows the distribution of the average change in energy of

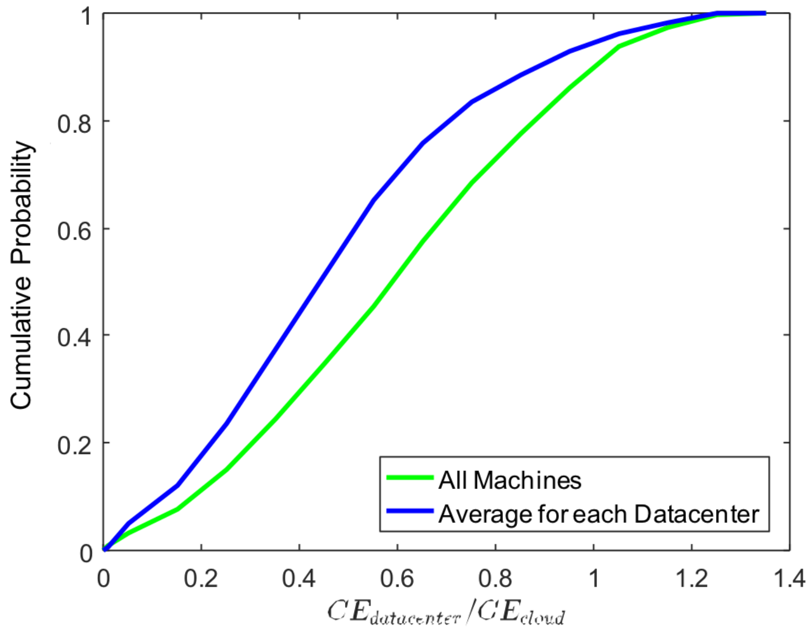

Figure 3.3: Fraction of Energy Used After Moving to the Cloud

each data center. The mean fraction of energy used is 0.64 over all machines and 0.51 for data centers. The figure shows that not all workloads would experience a reduction in energy usage. Around 10% of all machines measured would use more energy after switching to the CPUs used by CSPs, and around 5% of data centers would have an average energy usage increase on these CPUs. The reason is that some CPUs used in data centers are recently purchased and have state-of-the-art CE, which exceeds the CE_{cloud}. However, in Chapter 3.5, we will see that if these machines are sized correctly, the energy usage will decrease for nearly all machines and data centers.

Figure 3.3 also shows that half of the machines in the data set would use around 40% less energy if the workload is simply moved to the cloud. 50% of the data centers monitored would experience around an average 50% reduction in energy usage. One possible explanation of the difference between the average over data centers and the average over individual machines is that the statistic computed over data centers is biased by many small inefficient data centers. For example, if large data centers tend

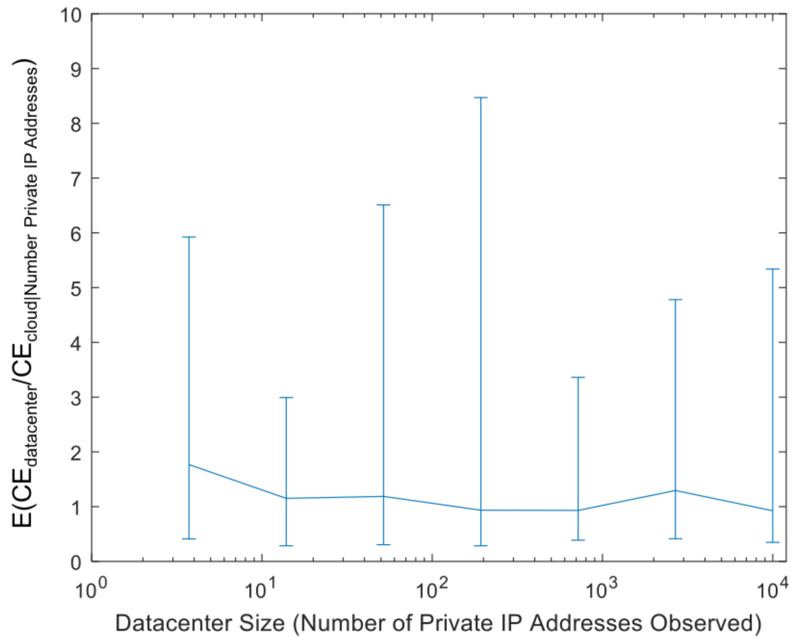

Figure 3.4: Average Fraction of Energy Usage for Each Data Center vs the Size of the Data Center

to be more efficient than smaller data centers, since these large data centers have more machines, per-machine statistics would be biased by the large data centers, while the per-data center statistics would be biased by the small data centers.

Figure 3.4 explores this hypothesis. The data set used in this study does not include the size of the data center, and not all machines in a data center are measured. However, on the machines monitored, network traffic and the source and destination addresses of the packets are collected. From these addresses, the number of unique private IP addresses is computed for each data center, and is used as an indication of the size of the data center.

Figure 3.4 shows the mean value, as well as the maximum and minimum value of $CE_{datacenter}/CE_{cloud}$ as a function of the number of private IP addresses observed. Specifically, the data centers are grouped into seven bins that ranged from under 10 observed IP addresses to several thousands, and to several tens of thousands of observed private IP addresses. For each group, the mean value, the maximum and minimum value of $CE_{datacenter}/CE_{cloud}$ are computed. This figure indicates that there is no significant dependence of $CE_{datacenter}/CE_{cloud}$ on the size of the data center. Rather, regardless of size, data centers can expect similar change in energy usage when migrating workloads to CPUs used by CSPs.

3.4 CPU Utilization and Relationship with Energy Usage

In this section, we examine CPU utilization and develop a new model for energy usage as a function of CPU utilization. The function is used to quantify the savings in the next section.

3.4.1 CPU Utilizations for Workloads in Data Centers

CSPs provide a large number of instance types, each with different computational abilities. For example, Azure currently offers over 70 instance types; Google Cloud Platform also offers a wide range of instance types, as well as custom instance types, where the user can adjust the number of CPU cores and amount of memory. Moreover, a user can easily resize the instance type. For example, most clouds allow the user to resize the machine with a few mouse clicks. This effort is considerably different from the steps required to resize a traditional data center.

Even though virtualization in traditional data centers allows resizing the machines, it does not solve the basic problem that computing capacity must be purchased and installed before usage. As a result, when sizing a data center, one typically purchases excess capacity. As computing requirements grow, VMs are resized and moved within the data center to utilize the excess capacity. Once the excess capacity is fully utilizes, a data center resizing effort is initiated. Data center resizing is often a complicated process that might require approval from several levels of management along with considerable planning to purchase and deploy the capacity, and perhaps with the requirements on expanding other capacities such as power and cooling.

Since the deployment of new capacity in a data center is a complicated process, sufficient excess capacity is purchased so that data center resizing is infrequent. More

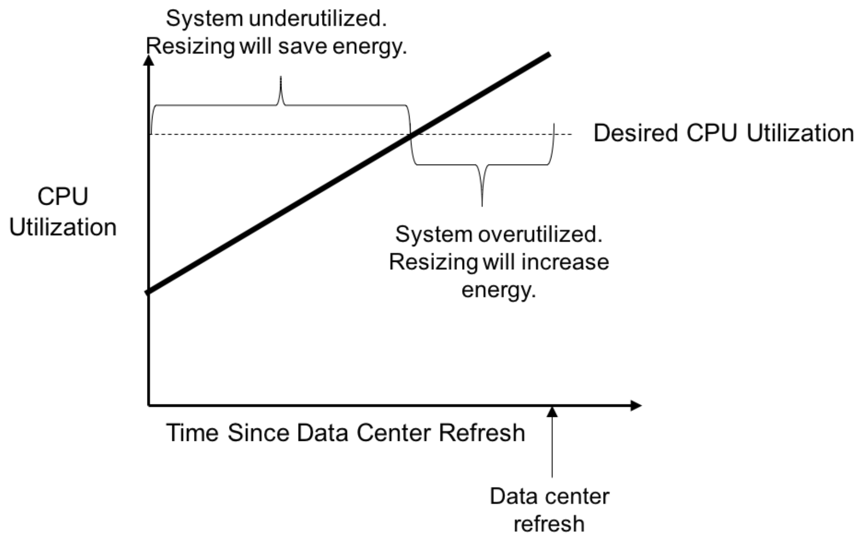

Figure 3.5: Periodic Data Center Refresh and Its Potential Impact on Energy Savings when Migrating to the Cloud.
Initially, just after a data center refresh, the systems are underutilized. Over time, as usage increases and applications grow more complicated, the systems become more heavily utilized. Eventually, the systems become over utilized shortly before a data center refresh.

specifically, when sizing a data center, system architects might design to achieve a low CPU utilization for workloads in a data center. The target CPU utilization is not based on performance objectives, but based on performance goals, anticipated growth, and the desired time between data center resizing episodes.

The result is that systems in traditional data centers can experience cyclic CPU utilization. Initially, after the data center is sized, the CPU utilization is low and the hardware is underutilized. Over time, the hardware becomes more utilized, eventually, reaches a level of high utilization before another data center resizing exercise is performed and the cycle begins again. Figure 3.5 illustrates how system utilization might vary between data center refreshes. Of course, this is only one possibility. Often there are a wide range of factors that impact data center utilization and performance over time. Unfortunately, performance of the applications is only one of many such factors. In any case, we can expect that system utilization can be both too low as well as too

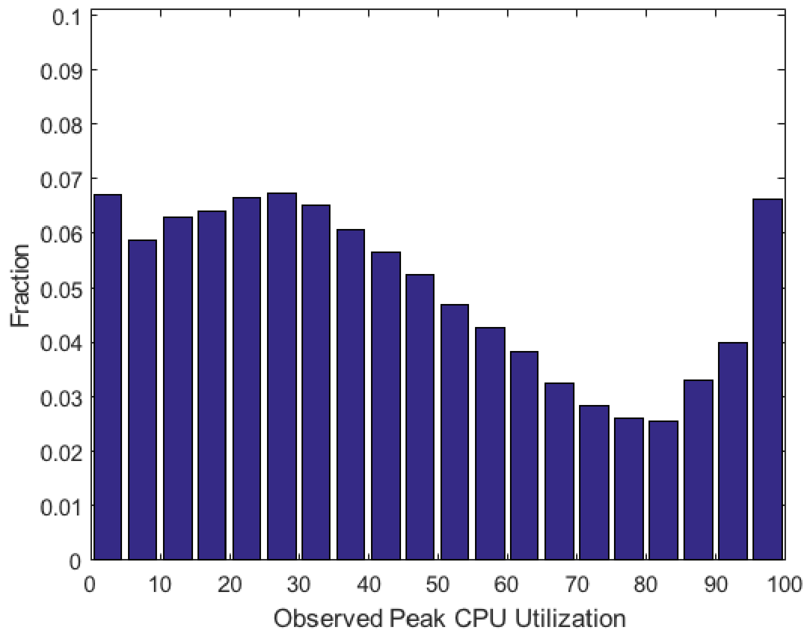

Figure 3.6: Distribution of Observed Peak CPU Utilization

high.

Figure 3.6 and 3.7 show the *peak CPU utilizations* observed. Specifically, the CPU utilization is measured every 20 or 30 seconds (depending on the system). These high-frequency measurements are smoothed with a 5-minute smoothing window. The maximum value of this utilization is collected for each day. The *peak CPU utilization* is the 95th percentile for these daily maximum values. Note that this method is selected so that single busy periods have little impact on the estimate of whether the system is over or underutilized. Instead, we seek to provision based on a typical peak.

Figures 3.6 and 3.7 show that the *peak CPU utilization* is frequently and fairly low. For example, Figure 3.7 shows that the peak CPU utilization is below 50% for around 60% of the machines measured. On the other hand, a non-negligible fraction of machines have high peak CPU utilization. To explore this further, consider Figure 3.8, which shows the mean CPU utilization along with the 10th, 25th, 75th, and 90th percentiles of the CPU utilization as a function of the release date of the CPU. This

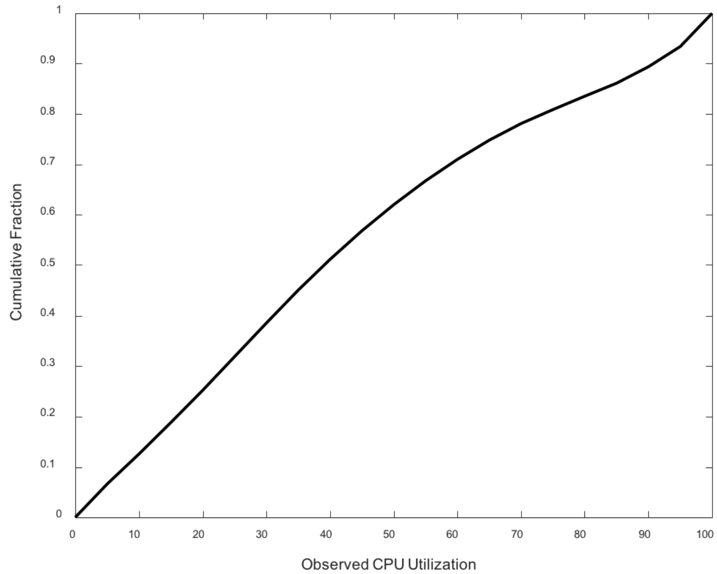

Figure 3.7: Cumulative Probability Distribution of Observed CPU Utilization

figure indicates that the *peak CPU utilization* of CPUs released since 2013 is typically lower than the peak utilization of CPUs released before 2013. This observation agrees with the data center refresh model described above, where machines in data centers with newer hardwares are expected to have lower utilization than machines with older hardwares.

From Figures 3.6-3.8, we can expect that many machines are over-provisioned, so that energy can be saved by moving the workload to the machines with less computational capabilities. On the other hand, a reasonably large fraction of machines have high CPU utilization. During migration, these machines would be moved to the machines with more computational abilities, which would increase the energy usage as compared to a lift-and-shift migration that seeks to keep CPU utilization unchanged. In order to quantify the change in energy usage from resizing the machine, we need to

1) Select a target peak CPU utilization, and

2) Understand how energy usage changes with CPU utilization.

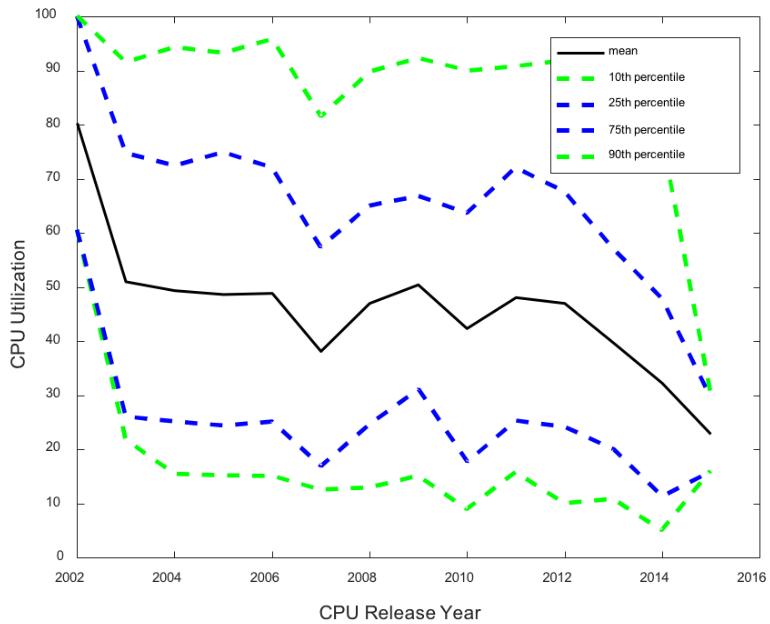

Figure 3.8: CPU Utilization versus the CPU Release Date

Determining the optimal CPU utilization is a complicated issue that is outside the scope of this thesis. Moreover, system designers select the target peak CPU utilization based on various "rules-of-thumb." Therefore, in the following analysis, we consider several target CPU utilizations, ranging from 50% to 90%. The second issue, determining how energy usage varies with CPU utilization, is addressed in Section 3.4.2 - Section 3.4.4.

3.4.2 Energy Efficiency Metrics

To understand the relationship between energy usage and CPU utilization, we utilize SPECpower_ssj2008 benchmark[3]. Currently, 512 benchmark results have been uploaded to the spec.org data set. These submissions include Intel Xeon CPUs (which are the focus in this study) along with CPUs from Intel Core and AMD Opteron families. Currently, 488 samples of Intel Xeon have been submitted, covering 86 different

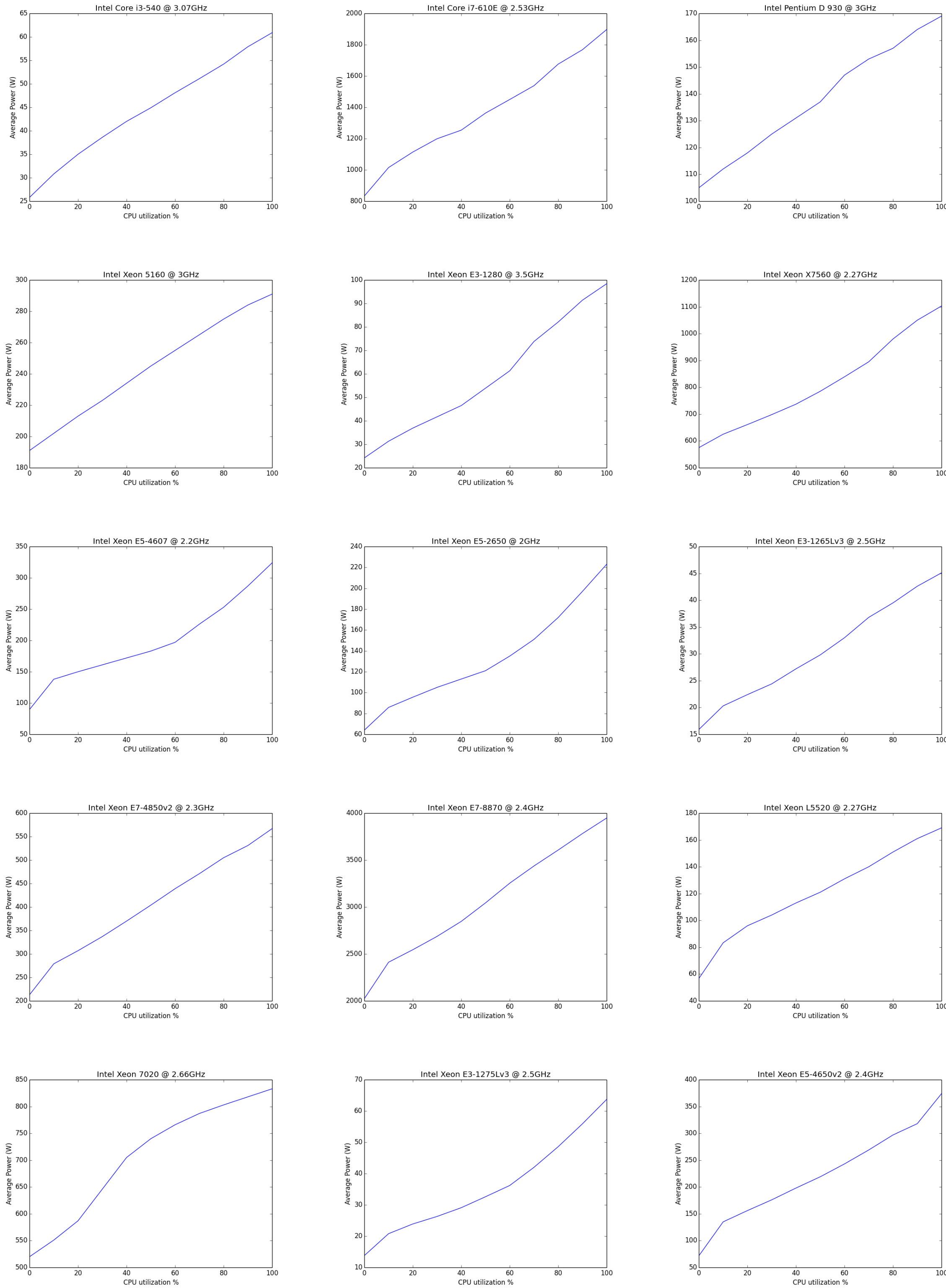

Figure 3.9: Power Consumption to CPU Utilization Ratio, using data from SPECpower_ssj2008[3]

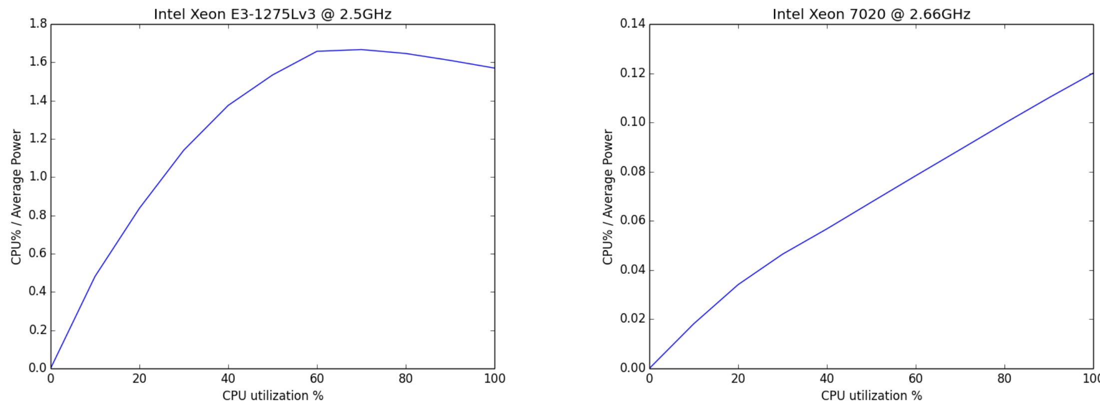

Figure 3.10: Compare Computational Efficiency of Intel Xeon E3-1275L v3 and Intel Xeon 7020, using data from SPECpower_ssj2008

Intel Xeon CPUs. The included CPUs are released from 2005 to 2016. The data reveals that the contemporary processors are not energy proportional. Especially, their idle power consumption is surprisingly high compare to the full load power consumption. Section 3.4.3 shows detailed analysis on idle-active energy consumption. To avoid energy waste caused by the low CPU utilization, finding the most "power efficiency point" becomes critical.

For the power efficiency aspect for different CPU utilization, we propose the *Average Power-to-CPU Utilization ratio (CPU%/Average Power)*, which is defined as the ratio of the CPU utilization over the average power at each CPU utilization level. The examples in Figure 3.9 shows that the power consumption is not linear with CPU utilization, and the idle power is not zero.

The observation in Fig 3.10 shows that the peak power efficiency is different for different CPUs. For example, the most efficient utilization for Intel Xeon E3-1275L v3 is at 60%, while Intel Xeon 7020 is most efficient when fully utilized. Note that the peak power efficiency are all happened above 50% utilization. The result from all available data from SPECpower_ssj2008 is in Fig 3.11, which shows that all CPUs achieve the peak performance when running over 50%, and most CPUs can achieve

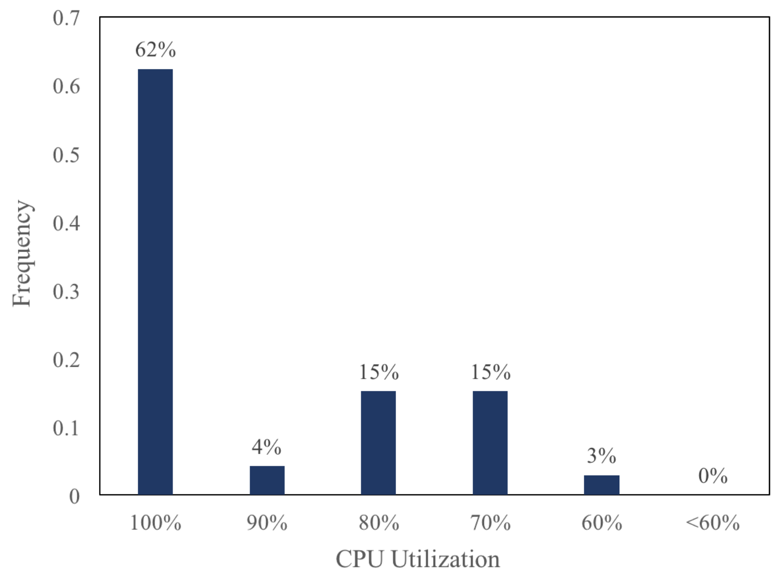

Figure 3.11: Distribution of Observed Peak Power Efficiency

peak performance when fully utilized. We can conclude that the CPU is more energy efficient with higher CPU utilization (above 50%).

3.4.3 Idle-Active Energy Consumption

Varsamopoulos [25] proposes *idle-to-peak power ratio* (IPR) as the ratio of P_{idle} and P_{max}, and states that the ratio reduced for the systems from 2007 to 2012. The research also predicts the system will become more energy-proportional in the future. Except for the energy proportionality, the ratio can also denote the energy efficiency. Our study confirms this prediction as shown in Figure 3.12. Compared to Figure 3.2, we can see that the ratio has the reversed trend as CE, that is, the IPR decreases as the computational efficiency increases. Figure 3.13 demonstrates this property. We also find out that the decrease of the idle power is the main factor to get the ratio decrease, this assumption is in accordance to the comparison between Figure 3.14 and Figure 3.15. The figures show a clear trend of P_{idle}/P_{max} decrease as P_{idle} decreases, but shows no similar trend with P_{max}.

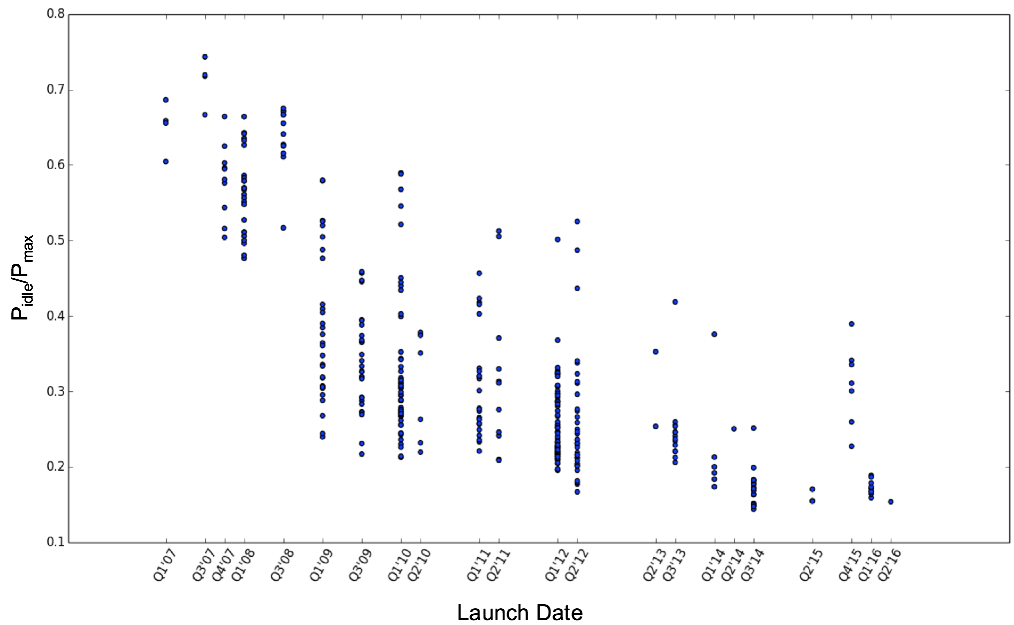

Figure 3.12: P_{idle}/P_{max} for all Measured CPUs, using data from SPECpower_ssj2008

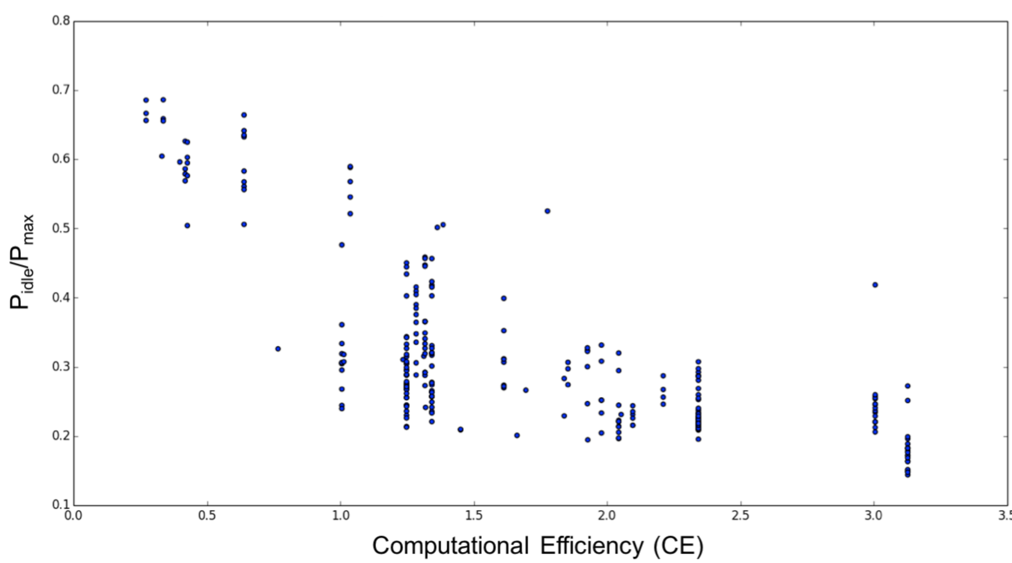

Figure 3.13: P_{idle}/P_{max} vs Computational Efficiency (CE) for all Measured CPUs, using data from SPECpower_ssj2008

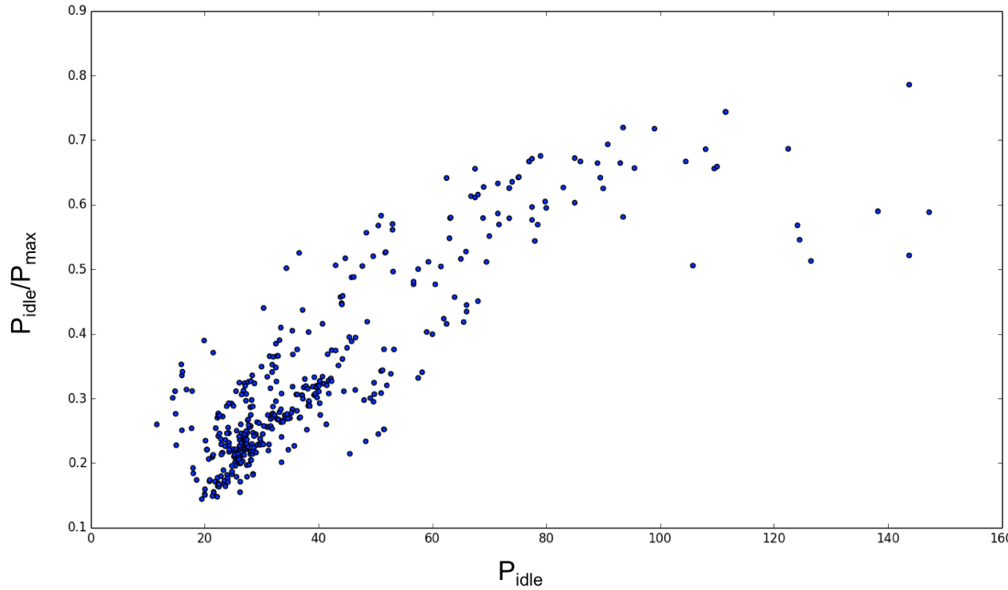

Figure 3.14: P_{idle}/P_{max} vs P_{idle} for all Measured CPUs, using data from SPECpower_ssj2008

There are massive researches [26][27][28][29] on energy efficiency effectively contributes to the reduction of the idle power waste during the past decades. The idle power for the newest CPUs is as low as about 15% of the max power consumption, which is reduced from 60% [30] [31] ten years ago. The implemented technologies have been used by many hardware manufacturers, for example, Intel SpeedStep [32], AMD Cool'n'Quiet and AMD PowerNow! [33].

3.4.4 Energy Usage as a Function of CPU Utilization

During the past decades, since CPUs have undergone significant changes that impact the relationship between energy usage and CPU utilization. For example, [34] and Intel Turbo Boost have been implemented and refined. However, more recently, the changes appear to have been less significant. Specifically, Figure 3.16 shows the relationship between CPU utilization and normalized energy usage, where the energy usage is normalized by the maximum energy usage. As expected [19] and the analysis in Section 3.4.2, energy usage is not a linear function of CPU utilization. First, there

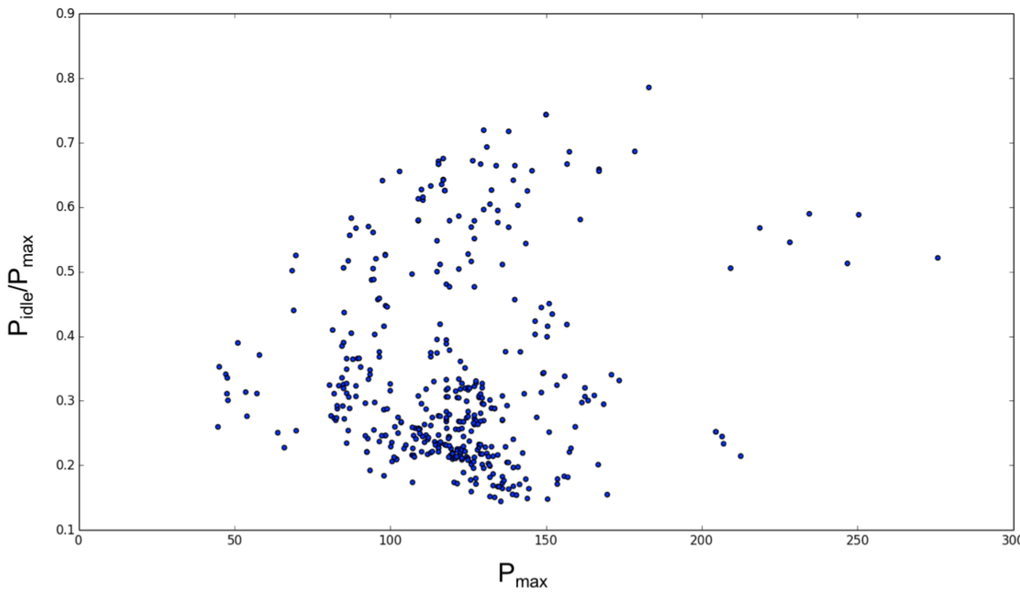

Figure 3.15: P_{idle}/P_{max} vs P_{max} for all Measured CPUs, using data from SPECpower_ssj2008

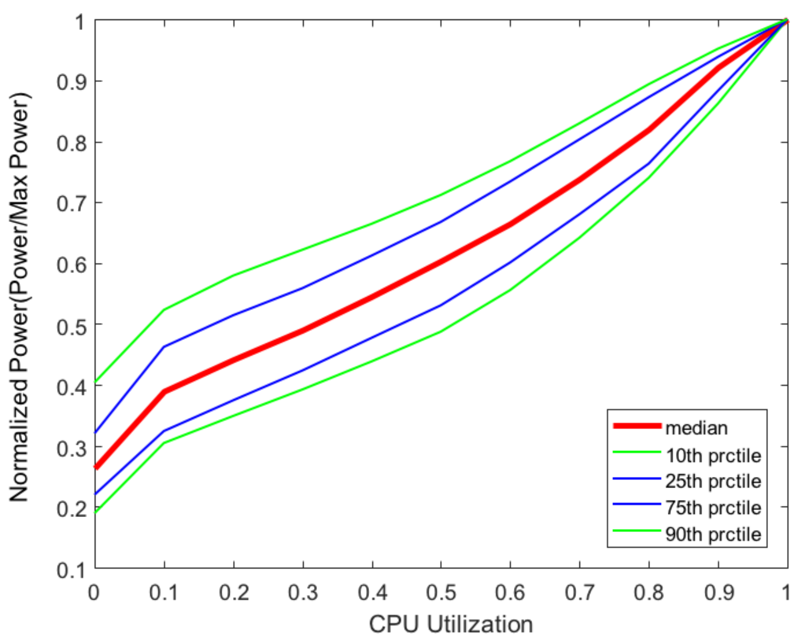

Figure 3.16: System Energy Usage vs CPU Utilization from the SPECpower_ssj2008

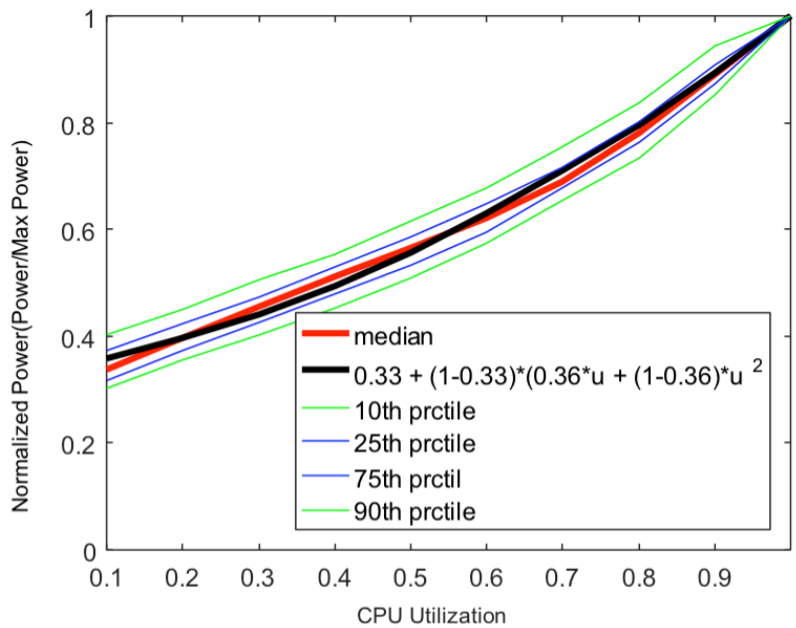

Figure 3.17: System Energy Usage vs CPU Utilization from the SPECpower_ssj2008 for CPUs Released after 2012

is a non-zero energy usage when the utilization is zero (Section 3.4.3). Second, the energy usage is not well approximated by a linear or even an affine function. Moreover, there is a significant spread in the energy usage for a given CPU utilization.

Although, notice that in Figure 3.16 there is a deviation at 10% utilization, before which the slope is larger than the after. It is because the advanced energy saving strategies for idle servers mentioned in Section 3.4.3 that cause relatively low idle power consumption. And this phenomenon is more obvious in the modern CPUs. Consider our data base, we only fit the power model with CPU utilization greater than 10%. Figure 3.17 shows the relationship between energy usage and CPU utilization for CPUs released after 2012. This figure also includes the graph of the function

$$E\left(u\right) = 0.33 + \left(1 - 0.33\right)\left(0.36u + \left(1 - 0.36\right)u^2\right). \tag{3.1}$$

For the CPUs released after 2012, this model gives a good fit of the normalized energy usage as a function of CPU utilization. Note that this model is different from the model developed in [31]. One possible source of the differences is that our model focuses only on recently released CPUs, while the CPUs used in [31] were older.

Note that the model (3.1) is only applicable to recently released CPUs. However, the CPUs used by current CSPs such as Azure, GCP, and AWS are all recently released, and therefore this model can be used to estimate the relative energy usage as a function of CPU utilization for systems that have been migrated to the cloud.

3.5 Energy Usage With Optimal Sized Instances

Several studies have indicated that CPU over-provisioning is an important component of energy waste in data centers [5][10][35]. For example, [5] states that CPU over-provisioning results in 50% of energy being wasted. However, the conclusion is not based on direct measurements of CPU utilization. The cloud offers several methods that can greatly reduce energy waste (and cost) due to over-provisioning. This section studies the case where the instance type is rarely changed, and examines the potential energy savings by choosing the optimal sized instances.

With the model of energy usage as a function of CPU utilization given by (3.1), we can estimate the change in energy usage that results from resizing the machine, and focus on statically sized machines. By this we mean that we select a single machine size, and examine the energy usage where the computational resources allocated to a workload are allowed to varying dynamically. Using a single machine size is considerably less administrative effort than dynamically sizing machines. However, statically sizing machines does not mean that the machine size is never changed, but only that the size is rarely changed. In this study, we assume that the machine is not resized to accommodate the workload in the collected data. That is, we will select a single instance size for each workload, where the instance size must be suitable for the entire data set for that workload. This is a reasonable assumption because the typical data set is only for a few weeks and does not include any data sets that span longer than 2 months.

Let $u(t)$ be the CPU utilization observed at time t. Then, the normalized energy usage is $\int E(u(t))\,dt$; this is the normalized energy usage if the system is migrated (i.e., a lift-and-shift migration) to an instance that has the exactly same computational capabilities as the hardware used in the on-premise data center.

We assume a simple scaling model, where the computation capabilities varies linearly with the amount of computation resources applied. For example, if a system is allocated 2 cores, then allocating one core will exactly double the CPU utilization (up to a maximum of 100%). This is a reasonable model for modern multi-threaded applications of which the requests arrive at random. Moreover, we assume that fractional resources can be allocated, e.g., where 2.4 cores can be allocated. Further, we assume that for a fixed CPU utilization, the energy usage scales linearly with the number of resources allocated. This means, for example, if a system is 50% utilized and if we double the computational resources allocated to the system while the load is doubled (so that the CPU utilization is the same on the larger system), then the energy usage also doubled.

Combining these assumptions, we have the following model of energy usage when the computational capabilities are increased by a factor of c

$$E(u(t)/c) \times c,$$

that is, the CPU utilization decreases by a factor of c, which leads to a change in energy usage by the machine. However, the number of machines has also increased by a factor of c, which increases the energy usage by a factor of c. Note that if the energy usage is linear in CPU utilization, then $E(u(t)/c) \times c = E(u(t))$, and correctly sizing cloud instances would not impact energy usage. However, despite efforts, energy usage is not a linear function of CPU utilization[19].

The *peak CPU utilization* is defined in Section 3.4. We denote u_p as the observed *peak CPU utilization* and u_T as the target *peak CPU utilization*. Since we assume a simple scaling model, we can achieve u_T by adjusting the computational resources

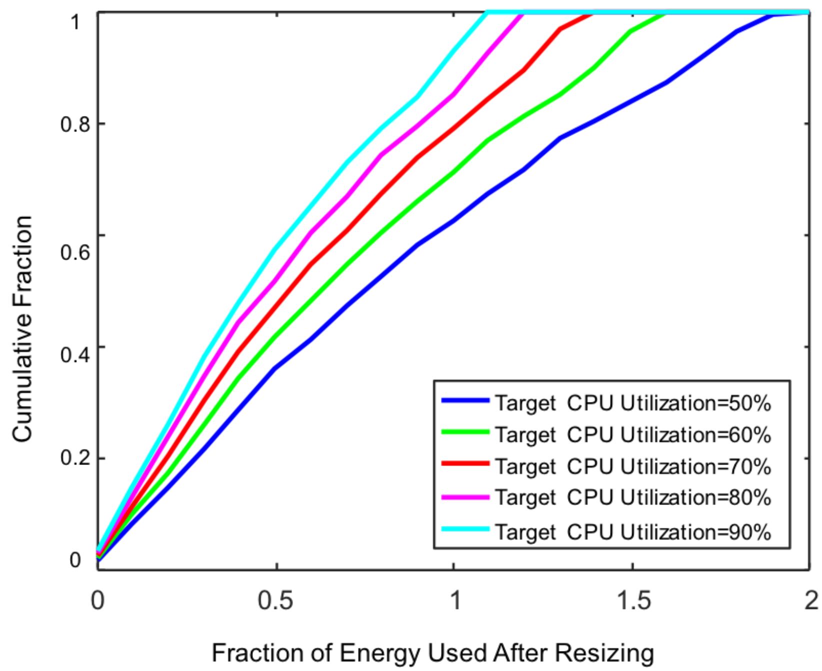

Figure 3.18: The fraction of energy usage after resizing to an optimally size instance, where the size of the optimal instance achieves the target peak CPU utilization

allocated by a factor of $c = u_p/u_T$. In this case, the fraction of energy usage after optimally resizing the cloud instance is

$$\frac{\int E\left(u\left(t\right) / \left(u_p/u_T\right)\right) \times \left(u_p/u_T\right) dt}{\int E\left(u\left(t\right)\right) dt}.$$

Figure 3.18 shows the cumulative distribution of the change in energy usage over all machines measured for a target *peak CPU utilization* of 50% to 90%. Note that resizing the instance type does not reduce the energy in all cases. This is expected since the figures in Section 3.4 showed that energy usage is quite large on some machines. For example, the *peak CPU utilization* exceeds 50% in around 40% of the measured machines. As a result, if the target CPU utilization is 50%, then, as expected, Figure 3.18 shows that the energy usage increases in 40% of the machines. For larger target CPU utilization, a larger fraction of system experience a reduction in energy usage as a result of resizing.

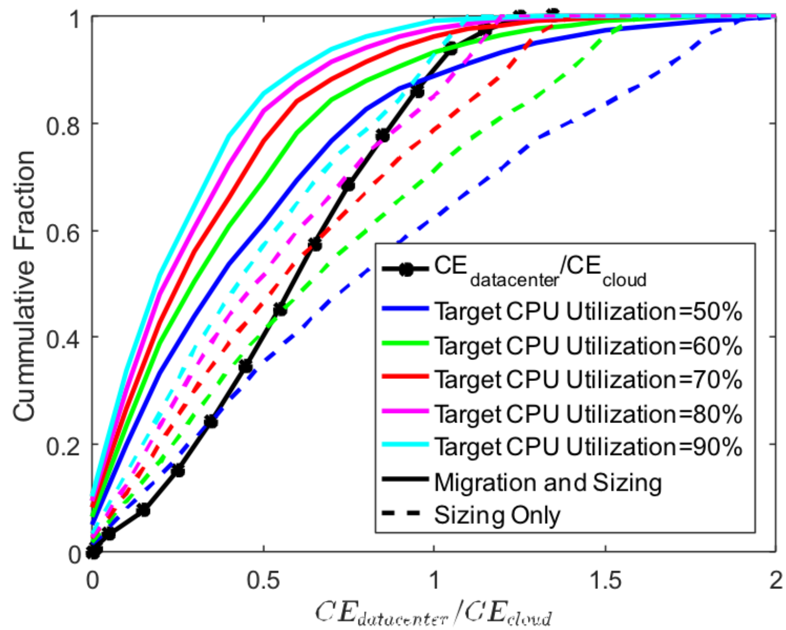

Figure 3.19: Fraction of energy used after migrating the workload to the cloud via a lift-and-shift migration and then resizing the an optimally sized instance

It is important to note that Figure 3.18 shows the ratio of the energy usage after resizing and the energy usage when on the cloud, but before resizing. Figure 3.19 shows the fraction of energy used after first performing a lift-and-shift migration, and then resizing. For reference, this plot also includes the fraction of energy used after resizing alone and the ratio $CE_{datacenter}/CE_{cloud}$, which the fraction of energy used from performing a lift-and-shift alone. Figure 3.19 shows that a combined lift-and-shift and instance resizing results in significant energy savings, and greatly exceeds the energy saving of either of the steps, namely lift-and-shift migration and resizing. Table 3.3 shows the mean fraction of energy used after this type of migration.

Recall in Section 3.3, we see that when utilizing a lift-and-shift migration, the energy usage does not decrease for all workloads. Specifically, newer CPUs are already energy efficient, perhaps more efficient than those used by CSPs. However, in Section 3.4, we saw that the CPU utilization of newer CPUs is typically low, that is, newer machines tend to be underutilized. Therefore, for the workloads on these newer

Table 3.3: Mean Fraction of Energy Used after Migration and Resizing

Small Target Peak CPU Utilization	Small Mean Fraction of Energy Used
50%	0.51
60%	0.43
70%	0.37
80%	0.33
90%	0.30

machines, even through there is no energy saving under a lift-and-shift, we see that the energy usage is reduced when resizing. On the other hand, older CPUs tend to have higher utilization, and resizing to meet performance goals will increase energy usage. However, the CPUs used by CSPs are far more efficient than these older CPUs. Therefore, in these cases we also see energy reduction when combining lift-and-shift migration and optimal resizing.

The energy saving identified is significant. For reasonable values of target peak CPU utilization of 70%-80% (the other values are included for references), the lift-and-shift with an optimal sized instance reduces energy usage by nearly a factor of 3. Moreover, there exists a wide range of tools that nearly automate this type of migration [15] [16] [17] [22].

3.6 Energy Usage with Idealized Auto-Scaling

The IaaS paradigm has many useful characteristics. One useful characteristic is that users are charged based on the duration that machines are running. As a result, there are significant financial advantages to utilizing auto-scaling. This section examines the potential energy savings of the computational abilities are frequently adjusted through auto-scaling.

Here is a simple idealized model for energy usage when auto-scaling is employed. Let $c(t)$ be the resources allocated to the workload at time t. Then, let $u(t)$ be the observed CPU utilization of the workload in the data center. The CPU utilization in the cloud at time t would be $u(t)/c(t)$, while the energy usage would be $E(u(t)/c(t)) \times$

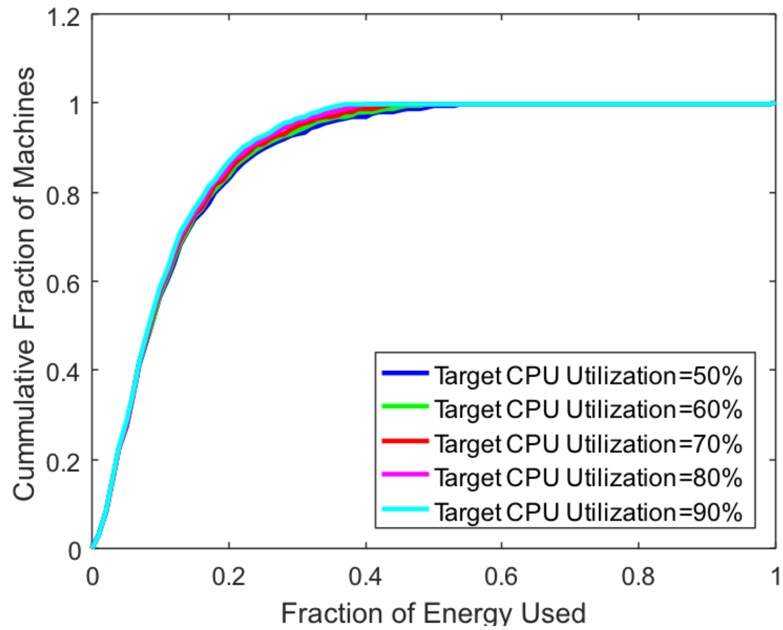

Figure 3.20: Fraction of Energy Used After Employing Idealized Auto-Scaling

$c(t)$. Let u_T be the target CPU utilization. Then, we set c so that $u_T = u(t)/c(t)$. Therefore, the energy used is $E(u_T) \times u(t)/u_T$. The fraction of energy used by auto-scaling as compared to the energy usage after a lift-and-shift migration is

$$\frac{\int E(u_T) \times u(t)/u_T dt}{\int E(u(t))\, dt}.$$

Figure 3.20 shows the cumulative distribution of the fraction of energy used after auto-scaling is implemented, as compared to the energy used after migrating to the cloud and optimally sizing the instance type. Clearly, significant energy savings are possible. However, there are significant drawbacks of this model. First, not all workloads are suitable for auto-scaling. Even workloads are suitable might require a significant rewrite. Second, the model that the allocated computational resources, $c(t) = u(t)/u_T$, is highly idealized. Specifically, it assumes that the computational resources are changed instantaneously and continuously. While this type of energy

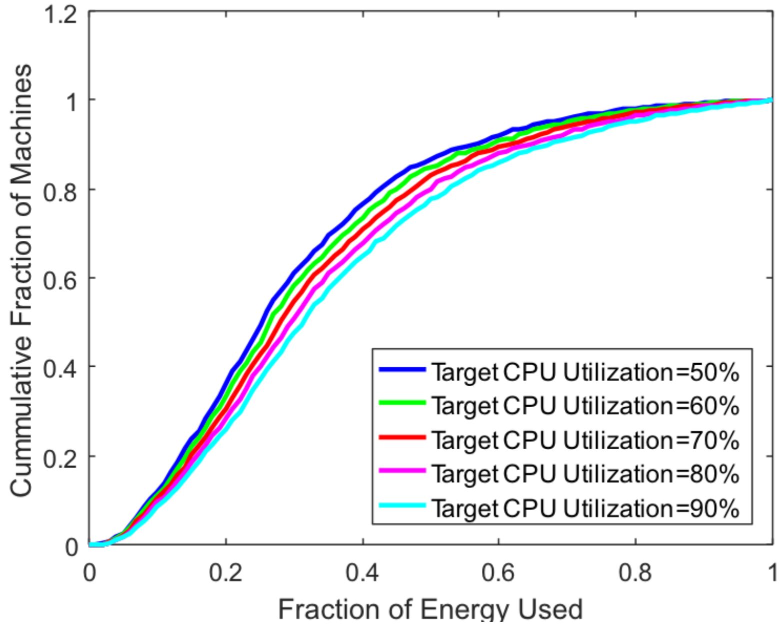

Figure 3.21: Fraction of Energy Used After Employing Idealized Auto-Scaling where the Instance is Resized Once Each Hour

usage might be reasonable for a PaaS system such as AWSs Lambda service. It perhaps too idealized for most workloads and IaaS.

Figure 3.21 shows the energy savings in a slightly more realistic scenario. Here we assume that

$$c(t) = \frac{1}{u_T} \max_{s \in hour(t)} u(s),$$

where $hour(t)$ is the hour of time t, i.e., $hour(8:15) = \{t : 8:00 \leq t < 9:00\}$. That is, this model assumes that the computational resources allocated to the workload are adjusted only once an hour, and they are adjusted based on the CPU utilization over that hour. Clearly, this model suffers from the drawback where it assumes the CPU utilization can be accurately predicted. Methods to predict CPU utilization and the impact of inaccuracy are out of scope of this study. However, Figure 3.21 shows that there are likely to be significant energy savings by refactoring applications to take advantage of auto-scaling. For example, the mean fraction of the energy used is around 0.3, implying a reduction of energy usage by a factor of 3.

3.7 CLOUD DATA CENTER EFFICIENCY

It is well known that the data centers used by CSPs are highly optimized and are far more efficient than what can be accomplished by all but the largest companies [7]. A key characteristic of this optimization is a significant reduction in the energy used by systems not directly related to computing, such as cooling and lighting. As mentioned in Section 2.2.2, PUE is a widely used metric that evaluates this type of efficiency.

Unfortunately, computing PUE is difficult and is not always computed by companies. For example, only 27% of companies reported that they compute PUE [36]. As a result, it is difficult to estimate the value of PUE in traditional data centers. While there have been surveys on PUE, they tend to be contradictory. In particular, the Uptime Institute publishes results of an industry survey each year. This survey includes PUE until 2014, but not includes PUE in 2015 and 2016. The surveys from prior to 2015 report rather low PUE as compared to the surveys performed by Digital Realty Trust. In particular, the Uptime Institute survey of 2014 reports an average PUE of 1.7 [37] while Digital Realty Trust reports an average PUE of 2.9 [36]. Moreover, Digital Realty Trust reports that only one in five companies have a PUE below 2, indicating a significant discrepancy between these two studies. One possible source of the difference between these two results is that the Uptime Institute survey includes data from CSPs as well as very large and highly efficient companies with hyper-scale computing facilities such as Facebook ,which reports a PUE of 1.09 [38]. In fact, Uptime Institute reports that half of those surveyed are from CSPs or similar companies. Another indication is that the Uptime Institute survey samples CSPs and companies with hyper-scale computing facilities and reports only 7% of the respondents are using the cloud. This value is far smaller than other studies such as [39], which reports that 20% of workloads are on the cloud.

While estimating current PUE for traditional data centers is difficult, PUE of CSPs is slightly easier. Specifically, Google reports a PUE of 1.12 [40] and Microsoft reports 1.125 [41]. AWS does not report their PUE. Nonetheless, we assume that the

PUE for workloads in the cloud are 1.123.

We conclude that by migrating to the cloud, the PUE experienced by the workload decreased from a value of between 1.7 and 2.9 to a value around 1.123. That is, the PUE is effectively reduced by a factor of 1.5 to 2.6. Notably, these values imply that while PUE is a critical component of energy efficiency, the other components, using a more efficient CPU and correctly sizing the instance, play a bigger role in energy reduction.

Chapter 4

CONCLUSION

This paper presented results regarding energy savings that will be achieved by migrating workloads to the cloud. The results are based on a large set of data collected from over 40,000 machines spread across over 300 data centers. We focus on lift-and-shift migration along with optimal instance sizing, as this type of migration is relatively straightforward to perform. The data indicates that such migrations will reduce energy usage by a factor of between 4.5 and 7.8. Much of the savings are from two sources, namely moving workloads off of old and inefficient CPUs to newer CPUs used by CSPs, and by correctly provisioning the cloud infrastructure. These sources of energy savings are considerably larger than the energy savings that results from moving from less efficient data centers to more efficient data centers owned by CSPs. It is critical to note that lift-and-shift and optimal instance sizing can be preformed nearly automatically [15][16][17][22].

This study focused on energy usage. The sources of the energy are also important to consider. For example, CSPs often utilize solar and wind farms to offset their energy usage. As a result, migrating workloads to the cloud could reduce carbon emissions beyond what would be achieved by only reducing energy usage. However, further study is required to clarify and quantify the impact that cloud migration has on carbon emission.

REFERENCES

[1] Dayarathna M, Wen Y G, and Fan R. Data center energy consumption modeling: A survey. *IEEE Communications Surveys & Tutorials*, 18(1):732–794, 2016.

[2] Brown R and et al. Report to congress on server and data center energy efficiency: Public law 109-431. *Lawrence Berkeley National Laboratory*, 2008.

[3] SPEC. SPECpower_ssj2008. https://www.spec.org/power_ssj2008/.

[4] Shehabi A, Smith S J, Horner N, and et al. United states data center energy usage report. *Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775 Page*, 4, 2016.

[5] Whitney J and Delforge P. Data center efficiency assessment. *Issue Paper, Aug*, 2014.

[6] Benik A and Ventures B. Gigaom — the sorry state of server utilization and the impending post-hypervisor era. https://gigaom.com/, 2013.

[7] Amazon. AWS innovation at scale. https://www.youtube.com/watch?v=uIjh79NeSVQ, 2015.

[8] Sverdlik Y. Intel designs custom chips for aws' new c4 instances. http://www.datacenterknowledge.com/archives/2014/11/13/intel-designs-custom-aws-cpu-for-fastest-ec2-instances-ever/, 2014.

[9] Jamshidi P, Ahmad A, and Pahl C. Cloud migration research: a systematic review. *IEEE Transactions on Cloud Computing*, 1(2):142–157, 2013.

[10] Armbrust M, Fox A, Griffith R, and et al. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, 2010.

[11] Walker E. The real cost of a cpu hour. *Computer*, 42(4), 2009.

[12] Khajeh-Hosseini A, Greenwood D, and Sommerville I. Cloud migration: A case study of migrating an enterprise it system to iaas. In *2010 IEEE 3rd International Conference on cloud computing*, pages 450–457. IEEE, 2010.

[13] Masanet E, Shehabi A, Ramakrishnan L, and et al. The energy efficiency potential of cloud-based software: A us case study. *The Energy Efficiency Potential of Cloud–Based Software: A US Case Study*, 2013.

[14] AWS. https://aws.amazon.com/.

[15] CloudEndure. https://www.cloudendure.com/.

[16] RiverMeadow. http://www.rivermeadow.com/.

[17] Microsoft. Microsoft Azure. https://docs.microsoft.com/en-us/azure/.

[18] Pelley S, Meisner D, Wenisch T F, and et al. Understanding and abstracting total data center power. In *Workshop on Energy-Efficient Design*, 2009.

[19] Ganapathy D and Warner E J. Defining thermal design power based on real-world usage models. In *Thermal and Thermomechanical Phenomena in Electronic Systems, 2008. ITHERM 2008. 11th Intersociety Conference on*, pages 1242–1246. IEEE, 2008.

[20] Goyal H and Agrawal V D. Characterizing processors for energy and performance management. In *Microprocessor and SOC Test and Verification (MTV), 2015 16th International Workshop on*, pages 67–72. IEEE, 2015.

[21] SPEC. SPEC CPU2006. https://www.spec.org/cpu2006/.

[22] CloudaMize. http://www.cloudamize.com/.

[23] Linthicum D. Beware shortcuts in your journey to the cloud. http://www.infoworld.com/article/3136124/cloud-computing/beware-shortcuts-in-your-journey-to-the-cloud.html, 10 2016.

[24] Amazon. AWS Documentation. https://aws.amazon.com/documentation/.

[25] Varsamopoulos G and Gupta S KS. Energy proportionality and the future: Metrics and directions. In *Parallel Processing Workshops (ICPPW), 2010 39th International Conference on*, pages 461–467. IEEE, 2010.

[26] Hua S X, Qu G, and Bhattacharyya S S. Energy reduction techniques for multimedia applications with tolerance to deadline misses. In *Proceedings of the 40th annual Design Automation Conference*, pages 131–136. ACM, 2003.

[27] Choi K, Soma R, and Pedram M. Fine-grained dynamic voltage and frequency scaling for precise energy and performance tradeoff based on the ratio of off-chip access to on-chip computation times. *IEEE transactions on computer-aided design of integrated circuits and systems*, 24(1):18–28, 2005.

[28] Xian C J, Lu Y H, and Li Z Y. Energy-aware scheduling for real-time multiprocessor systems with uncertain task execution time. In *Proceedings of the 44th annual Design Automation Conference*, pages 664–669. ACM, 2007.

[29] Meisner D, Gold B T, and Wenisch T F. Powernap: eliminating server idle power. In *ACM Sigplan Notices*, volume 44, pages 205–216. ACM, 2009.

[30] Barroso L A and Hölzle U. The case for energy-proportional computing. *Computer*, 40(12), 2007.

[31] Fan X B, Weber W D, and Barroso L A. Power provisioning for a warehouse-sized computer. In *ACM SIGARCH Computer Architecture News*, volume 35, pages 13–23. ACM, 2007.

[32] Intel product specifications. http://ark.intel.com/.

[33] Products we design — graphics, processors and apus — amd. http://www.amd.com/en-us/products.

[34] Mittal S. A survey of techniques for improving energy efficiency in embedded computing systems. *International Journal of Computer Aided Engineering and Technology*, 6(4):440–459, 2014.

[35] Judge J, Pouchet J, Ekbote A, and et al. Reducing data center energy consumption. *ASHRAE Journal*, 50(11):14, 2008.

[36] DigitalRealty. North america campos survey results. https://www.digitalrealty.com/resource-center/, 2014.

[37] UptimeInstitute. 2014 data center industry survey. https://journal.uptimeinstitute.com/2014-data-center-industry-survey/.

[38] Facebook. Facebook data center PUE/WUE. https://www.facebook.com/PrinevilleDataCenter/app/399244020173259/.

[39] Elumalai A, Starikova I, and Tandon S. IT as a service: From build to consume. http://www.mckinsey.com/industries/high-tech/our-insights/it-as-a-service-from-build-to-consume, 2016.

[40] Google. Efficiency: How we do it. https://www.google.com/about/datacenters/efficiency/internal/.

[41] Microsoft. Microsoft's cloud infrastructure: Datacenters and network fact sheet. https://azure.microsoft.com/en-us/overview/datacenters/, 2015.