Carbon Accounting for Sustainable Computing in Cloud Provisioned Data Centers

by

Khaalid McMillan

B.S. Mechanical Engineering, University of Florida (2014)
B.S. Aerospace Engineering, University of Florida (2014)
Minor, Business Administration

Submitted to the System Design and Management Program
in partial fulfillment of the requirements for the degree of

Masters of Science in Engineering and Management

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2023

Authored by: Khaalid McMillan
System Design and Management Program
January 27, 2023

Certified by: Dr. Bruce G. Cameron
Director, System Architecture Group
Senior Lecturer, System Design and Management
Faculty Director, Architecture and Systems Engineering Certificate
Thesis Supervisor

Accepted by: Joan S. Rubin
Executive Director, System Design and Management

Carbon Accounting for Sustainable Computing in Cloud Provisioned Data Centers

by

Khaalid McMillan

Submitted to the System Design and Management Program
on January 27, 2023, in partial fulfillment of the
requirements for the degree of
Masters of Science in Engineering and Management

Abstract

Enterprise digital operations require data-driven and scaleable solutions to monitor and manage the increasing environmental impact caused by manufacturing hardware and operating large data center warehouses. Data centers are complex systems comprising of heating, ventilation, cooling, power distribution, and workspaces to support employees managing the facility's operation. These workloads drive the global financial sectors, critical supply chains, big data analytics supporting consumer buying habits, and managing digital records for healthcare and payrolls that are critical to the function of modern society.

Although these digital operations are hidden within the confines of these large data centers, out of sight and out of mind, their impact on the environment is not negligible. The world's IP traffic is expected to exceed 2.3 zettabytes. Data centers consume significantly more energy per floor space of typical commercial buildings, 1.8% of the total energy used in the United States, and 1% of energy worldwide. This drives the need to measure and understand the impact of these workloads on the environment so that innovation and optimization can be leveraged to allow us to grow these technologies and digital opportunities sustainably.

To sustainably grow our digital presence, we need a method for tracking the environmental impact that the industry can adopt at scale while allowing for continued growth and development of digital technologies in parallel. There is interest from several government organizations to standardize a method to achieve this. This thesis attempts to illustrate a standardized and flexible way to measure the environmental impacts of digital solutions that combine the embodied emissions caused by manufacturing hardware and operational emissions from both the ICT and non-ICT systems in the data center. This solution would provide a means to track the actual ecological influence of digital operations that can be applied to information technology systems at any scale ranging from small technology systems to large cloud-provisioned data center operations. This work will focus on embodied emissions combined with operational emissions, reporting methodologies, and how that information can be used to distribute workloads using a multi-armed bandit algorithm using Thompson Sampling.

This work illustrates that embodied emissions can constitute anywhere from five to thirty percent of a server's total environmental impact. Workloads can be more than 11 times higher between workloads given the same set of hardware in a data center. Choosing the right configuration can reduce emissions by 4 times. Data center location significantly affects operational emissions, simply shifting the region where a workload is executed can reduce emissions by 24% or more. Using a data-driven approach, the spatial distribution of workloads can be implemented to optimize digital operations based on environmental objectives with as low as 30 iterations. The risk from regulation can be reduced, and a competitive advantage can be gained from implementing a sustainability-focused digital architecture.

Thesis Supervisor: Dr. Bruce G. Cameron
Title: Director, System Architecture Group
Senior Lecturer, System Design and Management
Faculty Director, Architecture and Systems Engineering Certificate

Acknowledgments

I want to thank the MIT System Design and Management staff, first and foremost, for creating and supporting a great program. I have learned so much in a short period and wish I could continue this journey, although it must end. I am grateful to Dr. Bruce Cameron for his guidance and patience throughout this process.

I want to thank IBM and my colleagues for supporting my continuing education, and I am grateful for the opportunity to continue my personal growth and professional development. I began with zero data science or machine learning experience and have completed this thesis with over 500 lines of code written in two different programming languages with plans to continue this work. A special thank you to Dan Di Genova, Randy Werner, and Dustin Demetriou for their guidance from start to finish.

I would also like to thank my family and friends who have supported me throughout my time at MIT, the COVID-19 pandemic, and my life to date. I want to thank my mother, Beverly Morgan, who has believed in me from the beginning.

Finally, I would like to thank the love of my life and my rock for sticking with me through it all, my future wife, Roma Rana. I love you more than you can imagine.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the world increasingly dependent on digital technologies, a focused approach to managing environmental impact will be necessary, and innovation is needed [1, 2]. The world's IP traffic is expected to exceed 2.3 zettabytes [3] data centers are consuming significantly more energy per floor space of typical commercial buildings [4, 5] and 1.8% of the total energy used in the United States [6]. The size and quantity of data centers are expected to rise in various forms such as in closets, on-premises, co-located facilities, or hyperscale facilities [7]. This growth has received significant attention from the United States Government starting in the early 2000's [8, 9, 10]. This trend poses a problem of increasing quantities of physical hardware and data center construction is coupled with an increased energy demand [11], manufacturing activities [12], and electronic material waste [13, 14]. The motivation for exploring environmental impacts and manufacturing trends has been further influenced by the Sustainable Development Goals [15]. Some of the goals relevant to emissions reduction are Goal 12: Responsible Consumption and Production, focusing on manufacturing and Goal 13: Climate Action, which focuses on energy consumption related to environmental impact. A goal relevant to establishing partnerships and strategies to identify and tackle these emissions is Goal 17: Partnerships for the Goals, which focuses on encouraging and facilitating implementation through collaboration across the industry.

Gordon Moore's papers in 1965 and 1975 harped on the importance of integrated circuits and the miniaturization of devices. They introduced the idea that trade-offs in integration complexity and manufacturing yield could minimize component cost and drive exponential improvements in cost and complexity aligned with advances in the manufacturing process over time.

Ultimately physical limitations were met, Dennard scaling benefits were no longer realized, and clock frequencies began to plateau in the mid-2000s. Complexity and manufacturing advances are being replaced by constant voltage transistor scaling. Performance gains are anticipated to be achieved through genuinely new devices, integration processes, and overall system architectures leveraging the advances in the industry achieved to date [16].

There is a significant focus on the power consumption of Information Communication Technology (ICT) hardware within the data center [17, 2]. Energy Star certifications are standard among many two-processor servers and other highly utilized hardware in the data center. The characteristics of the hardware this certification focuses on tend to be aligned with the number of processor sockets allocated in the hardware and whether or not the hardware leverages Graphical Processing Units, which tend to utilize more energy than traditional processors [18]. The early models of server power consumption tend to aggregate the power consumed by the drawer as a whole [19, 8]. Eventually, practices grew to measure power usage associated with the machine's state in data center modeling tools such as DCSim and CloudSim to allocate resources to Virtual Machines (VMs) that distribute workloads across servers in the data center[20]. In parallel, various power distribution algorithms have been developed to optimize the allocation of resources based in the cloud [21, 22, 23]. Still, there is no standardized method to monitor and report this information, as noted by work to address energy usage of blockchain technologies [18].

These issues leave the industry with the question, "What carbon accounting model in a data center will promote tangible greenhouse gas emissions reduction driven by digital workloads in the data center." The hypothesis is that a bottoms-up metered approach will be the most influential in driving a standardized solution with the flexibility to scale monitoring and reporting the environmental impact of enterprise workloads in the DC. This research will focus on motivations and trends of environmental impact monitoring in the DC, a review of current accounting methods, gaps in procedures and enabling technologies, and an illustrative case study to understand the value of implementing various strategies and recommendations that could influence the adoption of said methods for sustainability-driven decision making.

1.1 Motivation

Two types of carbon accounting addressed in this work are embodied and operational carbon footprints. Embodied emissions are associated with the energy required to manufacture the hardware itself [12] and can be related to Industry 4.0 initiatives such as green energy usage through electrification of manufacturing, utilization of renewable energies, and the implementation of additive manufacturing for end-use products [24]. These initiatives are already underway in the industry. This paper seeks to explore models to calculate this Life Cycle Analysis (LCA) and have it reported as part of hardware's Vital Product Data (VPD), the self-reported identity of the machine that holds the information such as serial numbers and machine type.

The operational carbon footprint is associated with the energy consumed during the operation of the machine and the energy consumed by the supporting infrastructure in the data center. This thesis will explore various methods to implement power consumption monitoring in the data center guided by best practices established by Lei et al. [18]. The goal is to explore the feasibility of implementation to encourage standardization of newly developed hardware to allow for such monitoring. Based on professional experience, there is the ability to implement active data collection of power Application-Specific Integrated Circuits (ASICs) that could report power consumption to a dashboard in real time and store it on the VPD.

This model would further explore the work done by Siddik et al. by tapping into power management integrated circuits (PMICs) used to regulate various elements within a hardware system, such as Processor, Memory, IO, and Fan Control with time-series data. Furthermore, exploring recommended task scheduling algorithms across computing resources can is explored. Finally, the machine would self-report the carbon consumption associated with the execution of workloads by incorporating the duration required to execute the workload and capture the operational emissions and the embodied emissions associated with the job to be used in a detailed report of the impact of said workload.

New architectures for computing are being explored beyond the traditional sequential model proposed by von Neumann. Previously, hardware and software designers were able to work independently. This came at a cost, computational performance, and energy efficiency, and the utilization of Application Specific Integrated Circuits and other specialized hardware designed for specific workloads showed significant promise. The implementation of "accelerators" is the general term for programming models leveraging a better mapping of the operations of algorithms to a set of physical resources. Examples of this would include the multitude of specialized circuits in smartphones and the broad adoption of GPUs and General Purpose GPUs used for specialized workloads, especially in power-constrained systems.

1.2 General Thesis Objectives

The objective of this thesis is to provide a systemic method to measure the environmental impacts of digital workload that combines the embodied emissions from the manufacturing and transportation of physical equipment as well as the energy associated with running the equipment and use this information to allocate workloads locally in the data center or to various regions based on emission factors of the data center. Once this impact and implementation method is understood, actions can be taken to optimize the distribution and timing of these workloads to achieve sustainability goals established by both the public and private sectors. This lays the foundation for sustainable computing practices to be applied in large computing operations by proposing optimization algorithms that allow for the exploration and exploitation of the ideal mix of technologies to execute digital operations. This will then enable architects to design systems specific to the workloads they are responsible for supporting without risking sustainability commitments.

1.3 Specific Thesis Objectives

This work seeks to test the hypothesis that monitoring power consumption and combining this with embodied emissions reporting can be used to distribute workloads spatially within a data center and among a set of data centers at specific times to meet a business' environmental emissions budget and reduce a static workload's power consumption in a data center with a heterogeneous mix of hardware.

1.4 Thesis Overview

This work will review current methods and feasibility for measuring and recording embodied emissions associated with building the physical hardware supporting digital operations. Review current practices and feasibility for measuring and recording the power consumption of the hardware as aligned to the workloads it supports. The work will conclude with methods to distribute this workload spatially and temporally to meet global enterprise sustainability metrics that align with industry computing trends, influence positive consumer behavior, and meet potential regulations affecting business operations.

Chapter 2

Literature Review

This chapter focuses on current research and trends in the reporting of embodied emissions, the legislation driving detailed reporting of embodied emissions, the energy consumption of data center assets, technologies explored for emissions reporting, how sustainability affects consumer behavior, and gaps in data center emissions reporting.

2.1 The Metric lbCO$_2$e

The main greenhouse gasses in the lower atmosphere that are of concern due to their rising concentrations are carbon dioxide, methane, nitrous oxide, hydrochlorofluorocarbons, hydrofluorocarbons, and ozone, according to the World Meteorological Organization [25]. A metric that describes the global warming effect of these gasses is the term CO$_2$e and has been generally accepted. CO$_2$e is defined by the United States Code of Federal Regulations as "...the number of metric tons of CO$_2$ emissions with the same global warming potential as one metric ton of another greenhouse gas...".

2.2 Embodied Emissions

Embodied emissions define the environmental impacts associated with the production of a physical product and are typically measured in lbCO$_2$e. These include emissions from activities spanning raw material extraction and production, conversion from raw materials into piece parts, assembly, transportation, implementation, service, and end-of-life activities associated with the product [26]. Given the current economic climate, reducing waste and driving cost efficiencies are the number one priority for North American and EMEA manufacturers and third for Asia/Pacific Companies[27]. Operationalizing sustainability requires organizing and disclosing the information needed to make decisions and prioritizing areas where the greatest impact and operations can lead to a competitive advantage. To meet the requirements of regulators and adapt to the pressures from their customers, these manufacturers will need to leverage technology as a critical method to integrate data from a variety of sources to identify gaps and pinpoint opportunities[27]

There are efforts to create a simplified model for a Life Cycle Analysis for Electronic Equipment ranging from consumer electronics to data center Information Communications Technology (ICT) equipment [28]. Larger companies have already started publishing product carbon footprint reports on their websites based on standard configurations and utilization of industry-accepted estimation tools[29, 30, 31].

The electronics industry changes rapidly with continuous innovation and investments that drive growth in data centers [5, 3]. The generation of new technologies generally focuses on performance gains to give players in various spaces a competitive advantage in their specific arenas like drug discovery, reactivity to financial markets, or consumer-facing digital products. This means that little thought has been placed on the environmental impact as sectors react to these opportunities in their self-interest, given the focus on pushing the capabilities of their digital solutions beat their competitors to the market.

This is apparent with the recent move to hyper-scale data centers and the scale-out focus of digital operations to support cloud business models [3]. Examples include shifting from on-premises data centers to co-located or scale-out data centers. This can be seen with large software companies moving to build and manage their own data centers, such as Facebook, Amazon, and Microsoft. Increases in the need for ICT experience have resulted in a consortium called the Open Compute Project, where these software companies are collaborating, standardizing, and documenting best practices to enable data center operators to build a pool of skilled labor focused on supporting ICT equipment deployment, and operations [32]. As these standards mature, there is also a risk that the constraints for skilled labor and costly hardware support actions are reduced, and the data centers will move to the way of the warehouse, where low-skilled staffing is needed in bulk to perform scaled operations. Eventually, this will lead to automation in the industry, enabling significantly larger operations, faster deployment, and consumption of ICT equipment.

Digital workloads will significantly increase the purchase and implementation of ICT hardware in enterprise and edge operations, especially as automation and machine learning enables more use cases and increase the ability to manage large IT infrastructure [33]. As companies begin to focus on their carbon footprint beyond power consumption, they need to track and manage the environmental impact of these operations, as private and public stakeholders focus on the importance of sustainability in day-to-day operations.

Private sector pressures to disclose this information serves as a competitive advantage, and public relations statement affecting purchasing behavior and sustainable marketing initiatives [34]. Regulations may enforce this activity financially through taxes and fees or more aggressively by restricting the import of products into the country that doesn't meet their reporting requirements [35]. The impacts on business operations from the private sector may provide some market benefits depending on consumer motivation for a green and circular economy. There is growing sentiment supporting sustainability by large corporations, but the average consumer may not understand how to evaluate the impact associated with the consumption of digital products. Therefore, the market impact from the private sector consumption may be marginal for the time being. ESG goals have impacted stock market prices, mergers and acquisitions, and financing. Some companies have made capital asset sales and have seen positive stock market behavior [36].

Public sector influences may be seen as more drastic, and a driving factor behind the buying behavior in the stock market and business-to-business interaction [37]. ESG-based regulations derived in the European Union have proven to have an impact on global business operations, termed the "ESG Brussels Effect," where disclosure standards become internationally recognized, similar to how California pioneered stringent car emissions standards that were subsequently adopted by other states in the United States [38]. Furthermore, the cost of implementation can be significant if businesses are forced into a centralized operating and reporting model, which incentivizes companies to establish a reporting system that adheres to open, globalized standards [39].

Companies that own ICT equipment will need to begin tracking the carbon footprint associated with purchasing and implementing physical assets. Because of the complicated mix of materials and manufacturing methods inherent to ICT equipment, the environmental impact analysis is somewhat complex [40]. The manufacturing process comprises traditional manufacturing techniques such as casting, injection molding, machining, and sheet metal forming as well as incorporating advanced production methods such as silicon device production for integrated circuits and Printed Circuit Boards (PCBs) [40]. These advanced manufacturing methods include composites that form dielectric layers intermixed with metal conductive layers and traces. Typically, these advanced manufacturing methods consume rare or hazardous materials that drive significantly more energy, resource consumption, and waste in production.

Emissions are defined by the Green House Gas protocol and are categorized in three scopes [25]. Scope 1 emissions are emissions those that are emitted during the process of converting fuels to energy directly. These would include the gasoline used to power vehicles, or diesel used in power generation. Scope 2 emissions are those that are realized through the purchased and use of energy such as those caused by purchasing electricity from the local electrical grid. Finally, scope 3 emissions are emissions not caused or consumed directly from an actor, but are those that the actor is indirectly responsible for. These emissions are most commonly associated with the products a company use and dispose of from it's suppliers to support it's business operations.

Two tools to estimate scope 3 emissions introduced and accepted as TCO certified , a predominant sustainability certification for IT products, include the Product Attribute to Impact Algorithm (PAIA) and the iNemi Eco-Impact Estimator tool. These tools leverage industry-accepted databases for carbon footprint estimation based on materials, properties, weights, and geography characteristics for producing parts. Furthermore, there is an effort to drive direct reporting of Scope 1 and 2 emissions from suppliers downstream of the product life cycle, including raw material acquisition, production, and distribution of products. This reporting is currently thought to provide the most direct and reliable estimate of a product's carbon footprint with the least effort.

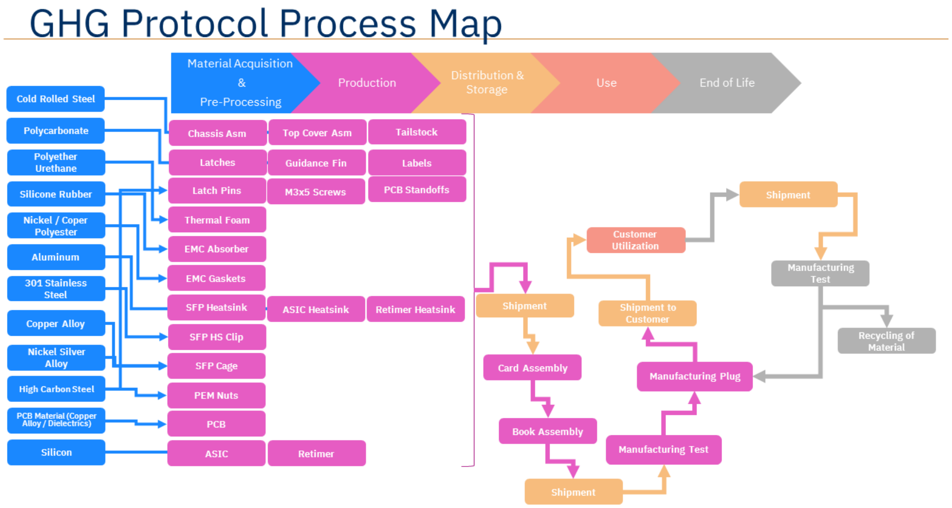

Figure 2-1: GHG Protocol Process Map Example.

2.2.1 Scope 3 Emissions Tools for Manufacturing

Because carbon accounting methods directly associated with manufacturing operations have not been widely adopted, tools have been developed to provide general insight into the environmental impact of producing goods based on raw materials and the type of operations used to produce them. These tools utilize samples of data representing standard production techniques such as injection molding, casting, and machining, to name a few. Widely adopted tools and their assumptions are discussed below.

2.2.2 Generalized Life Cycle Analysis Tools

A tool not specific to the ICT industry is the Simapro tool which leverages the EcoInvent database and provides a user interface and APIs to evaluate the environmental impact of materials and activities associated with sectors that vary from agriculture to metals and polymers. This tool references the ecological impact from similar sectors as the GaBi database, which provides a user interface to interact with the database. Another interface is the OpenLCA tool that references the GaBi database. These databases generally rely on industry surveys for estimated emissions aligned with general practices and then estimate the environmental impact per unit weight for raw material extraction and production for the elements of the product and packaging. For transportation estimates, the databases use factors aligned with varying transportation methods (freight, air, water) based on distance traveled to estimate the emissions aligned with logistics. They then attempt to create an impact per unit of weight and/or distance for materials, products, and transportation to generate a Bill of Activities (BOA) that captures a generalized view of environmental impact. These tools create models for processes like raw material, injection molding, or shipping by a specific method over a set distance.

Many evaluated the viability of these tools, and the models created can vary based on materials, processes, or location. One product's results calculated in one tool may disagree with the results of another [41, 42, 43]. These variations can lie in the weights and calculations used to model the BOAs and how they overcome uncertainties [44].

2.2.3 Eco-Impact Estimator

The iNemi Eco-Impact Estimator tool was proposed in 2012 as a spreadsheet-based method for aggregating the estimated life-cycle impact of ICT hardware using a simplified process, categorizing targeted components, providing reasonable accuracy aligned with the ICT industry needs, and allowing for continuous improvement [40]. This tool is based on a collaboration between industry and academic partners, including Nokia Bell Labs, IBM, TTM, Intel, Purdue University, Logitech, and iNemi. The tool seeks to combine the dominant life cycle impacts from Manufacturing, Transport, Use, and End-of-Life stages to combine Embodied Eco Footprint (Raw Material Extraction, Intermediate Component, and Sub-Assembly Manufacturing), the Operational Eco Footprint (Usage and Servicing), transportation and installation, and recycling and disposition of the ICT product. The tool estimates the environmental impact associated with detailed processes such as printed wiring board (PWB) cutting, lamination, and drilling, as a few examples. The tool intends to offer more availability of information to its members and academic researchers. It provides a method to look at the impact of ICT hardware with more granularity and allows for updates that align with technological advances as the industry evolves. This tool also references the GaBi database. The uncertainty here also include those previously mentioned, like systemic errors, model uncertainty, parameter uncertainty, and uncertainty in the data collected and used [44].

2.2.4 Product Attributes to Impact Algorithm

The PAIA tool was developed by the Massachusetts Institute of Technology Materials Systems Laboratory and Quantis in the Fall of 2016 after being launched in 2009. The PAIA tool has applications specific to servers, displays, notebooks, desktops, thin clients, tablets, storage, and network switches, as well as an All-in-One tool. The tool is a methodology for environmental footprint analysis specific to ICT products. It provides a workflow that reduces the cost and effort associated with calculating the ecological impact of categories of tools in kg CO_2e [45]. There is a consortium surrounding the tool that comprises member companies that allows them to share experiences, stay on top of legislation associated with environmental impact, and shape the direction of the PAIA tool. Current members include IBM, Fujitsu, Dell, Hewlett Packard Enterprise, Cisco, Lenovo, and ViewSonic. This consortium collaborates with organizations like the Environmental Protection Agency and the Semiconductor Industry Association.

The PAIA approach focuses on influential product attributes in data collection and input to understand the environmental impact of ICT products to identify the relative performance of impact reduction strategies. To calculate the carbon footprint of a product, the tool requires the user to provide specific information for parameters that can reduce the estimate's uncertainty and generalizes the rest of the information to alleviate the burden of detailed data mining. For example, the inputs required to calculate the environmental impact of PCB manufacturing include area, additional electronics by weight, manufacturing locations, manufacturing energy, and materials by weight.

The tool is currently utilized to guide and support the development of standards for product carbon footprint analysis in development and procurement initiatives, respond to customer surveys and information requests, help guide material selection, and Green House Gas goal setting while meeting reporting standards such as EPEAT, Grenelle, and CDP. This tool references the GaBi Life Cycle Inventory database that references $lbCO_2e$ data associated with raw material processes, general manufacturing processes, and information specific to electronic component manufacturing. This tool does not consider the effects of meaningful actions a particular actor takes within a product's supply chain. Assuming the database broadly represents the industry activities, it would lump the more sustainable OEM suppliers with the least sustainable ones. These calculations would not measure the real effect of top-down pressure to provide a more sustainable product. A more resource-intensive process would be needed to capture this competitive advantage. Also, the uncertainties above in the data [44] are also present, and the model is obscured from the average user of the web interface.

2.2.5 Gaps in Embodied Emissions Calculations

The tools leveraged currently by the industry are a first step to understanding carbon emissions as a comparison method. Still, they do little to evaluate the actual impact of the specific manufacturing and transportation processes in place today. These tools are more a relative indicator than a real snapshot of the environmental impact of ICT equipment. There is a level of uncertainty built into the analysis done by the PAIA tool, and no uncertainty statement is claimed in the iNemi Eco-Impact Estimator tool. The uncertainty value in the PAIA tool can be quite significant, driving a common declaration in product carbon footprint reports generally stating that "All estimates of carbon footprint are uncertain" [29].

While these tools are great at estimating the product's carbon footprint and identifying parts of the systems where efforts for environmental impact reduction can be taken, there is little this information can do to drive real concerted change in the development of the products and provide companies with a clear competitive advantage in the sustainability arena. Without information specific to the product's full life cycle, it can be unclear where a company should focus its resources, as there will be product-specific measurements to review. Although utilization will always be a large portion of the product's environmental impact, detailed accounting of material composition and supply chain logistics are prime areas of opportunity to tactically reduce the Scope 1 and 2 emissions from partners that contribute to the product's overall environmental impact during production and end-of-life.

2.3 Sustainable Legislation

2.3.1 Digital Product Passports

The European Commission has proposed legislation to establish "a framework for setting eco-design requirements for sustainable products" [46]. This proposal shows the basis for a Digital Product Passport (DPP) that spans a broad base of industries, ICT products being one of them at both the consumer and enterprise level. This plan aims to establish a data-driven circular economy where consumers, regulators, and actors can access information to influence more sustainable decision-making within the product's lifecycle. The included data may include material traceability, carbon footprint information, recycling procedures, and lifecycle guidelines proposed by GS1 in Europe [47]. This proposition establishes the need for industry players to build the required infrastructure to support digital access to product lifecycle information in a standard and practical manner.

Europe has been known to set the standard for stringent environmental reporting, the "ESG Brussels effect," and the proposition of the DPP has begun to set the direction for a digital solution to reporting and evaluating the environmental impact of products over their lifecycle from the cradle to grave. This legislation could prevent exporting products into the EU and even affect internal production. The idea behind the DPP is that the actors in a product's lifecycle would have full access to the information required to create a digital twin of the entire product's lifecycle and subsequently model changes that the actors can take to improve the environmental impact of the product. This can include material and process changes, alterations to the supply chain to enhance the sustainability associated with logistics, the ability to recycle the product, or re-purposing the product to improve its lifespan. But, not all of the effects are considered to be risky. The drive for transparency could allow sustainably minded actors to finally have the infrastructure to display a sustainability-driven competitive advantage and remove the need for timely data gathering and modeling.

To enable this form of digital tracking at the scale anticipated by the regulation, the process would need to integrate with current data generation and storage tools, such as Enterprise Resource Planning (ERP) software that is commonly used to track inventory in industrial manufacturing. This software spans raw material, intermediate, and end product supply chains. With this in place, the required data strings would need to be defined. This allows the information to be integrated into the natural workflow of many businesses using various GHG reporting methods, such as allocation as stated in the Product Life Cycle Accounting and Reporting Standard of the GHG protocol [48].

An example of reporting integrated into the workflow of a product via the ERP system is to utilize a method called Economic Allocation. This method allocates emissions for each unit produced based on its economic value when leaving the multi-output process. This would mean that a company's portion of total revenue leaving a facility for that product would be that portion of that facility's emissions. The assumption is that the cost of resources for the product is built into the price consistent with the company's product portfolio. In this case, this would be data that should already exist somewhere in a company's digital infrastructure since the facility's electricity consumption, the location of the facility, and the composition of the grid should be known [49].

While this is a shortcut to getting to the specific emissions of a product, this would be a very scaleable solution. A genuinely advanced variation of this method would be to directly monitor the power consumption of equipment used to produce this product, the time the product spent in the facility, and downstream emissions could be collected using Internet of Things devices and subsequently reported into the ERP system. This would not be a solution that scales well and could be quite costly to implement, which may deter adoption. Therefore the Economic Allocation method may be the most straightforward method for actors to get as close to the facility data as possible. The data for the two processes mentioned above could be incorporated into GS1 standard barcode information as strings widely utilized today in almost every product and exchange between actors [47].

2.4 Energy Consumption

The digital economy comprises infrastructure in the form of ICT goods and services, e-commerce, priced digital services, and cloud services, according to the Bureau of Economic Analysis [50]. The real gross output of the digital economy grew at an annual rate of 5.2% between 2005 and 2018. Business-to-consumer e-commerce grew at an average rate of 12.7% per year, with cloud services growing at 8.5% and hardware closely behind at 7.9%[50]. This growth is generally driven by the pace at which companies increase their digital footprint to meet the demand for online products and gain an edge over their competitors. These closely follow the advances in computing technology and deployment of Machine Learning, Artificial Intelligence, and Web 3.0 trends to identify novel opportunities to win market share and generate new revenue streams.

There are changes associated with the environmental impact associated with such an expansion [51, 6]. Companies that traditionally had on-premise solutions have transitioned to co-located and cloud architectures, allowing them to physically disassociate themselves from their IT infrastructures. Using this outsourcing method, companies can realize cost savings by virtualizing and placing their workloads in a distributed environment [52, 53] without the need to directly take responsibility for the overhead required to manage the associated ICT equipment.

The distribution of ICT equipment impacts operations where a company's digital presence is sourced to third-party data centers and cloud operators. When a company scales out its operations, it will only incur an expense based upon its Service Level Agreement (SLA) associated with cloud computing resources or buying space in a co-located data center. This means that companies may no longer be able to control the procurement of or monitor the energy consumption associated with the operation of their digital infrastructure. Their workload expansion will rely on third-party organizations to manage the environmental impact related to their digital footprint. Given this lack of control, there would be a disconnect with the reporting standard associated with the Green House Gas Protocol (ISO 14064) [54]. Furthermore, the current reporting structure allows companies to offset their climate impact by paying for energy via Power Purchase Agreements or Time-based Energy Attribute Certificates without considering embodied emissions. These are great methods to spur investments external to the company. Still, this data may not be rich enough to provide insights into internal decision-making processes by obscuring the actual environmental impact associated with their digital footprint.

Enterprise workloads now treat data centers as a pool of computing resources managed by a third-party organization. Ultimately, organizations no longer need to source and manage physical assets in the data center, and deploying digital solutions will bypass hardware considerations. The client will pay for the compute time and type, storage capacity, or execution time on the hardware without considering the power consumed by the hardware and the supporting infrastructure. This means that tracking emissions associated with the physical implementation of hardware will be obscured. Tracking aligned with the greenhouse gas protocol standards to determine their total environmental impact from digital solutions will now be reported to them by the operators of the data centers from which they source computing resources. Evaluating their digital presence by workload requirements will be the most direct means to make informed decisions to improve their sustainable computing strategy. When an organization runs a workload at one data center location, its environmental impact may not be the same as if they were to run it in a different data center.

For example, if a coal-fired power grid predominantly supplies one data center and another is supplied by one primarily hydroelectric, their environmental impact would not be equal. Their claims of a zero-carbon footprint may not be entirely accurate. Understanding the ecological impact of assets, data center location, and time of day could help the company strategically manage its digital architecture optimizing for sustainability goals. Algorithms for machine learning, optimizing business operations, search, and encryption can be evaluated to ensure they run on the ideal hardware resource. This can include identifying when it is most efficient to offload work to a cluster of GPU accelerators, determining when CPU resources may be sufficient to execute decryption algorithms relative to specialized hardware, or determining if there is a benefit to updating large database operations onto new technology as a few use cases.

2.4.1 Operational Emissions

To properly account for the carbon impact of workloads, it is essential to understand the source of the power supplied to the data center. Different types of power grids supply data centers with varying degrees of impact on the emitted greenhouse gasses (GHG) realized per MWh consumed depending on their region [49]. Therefore, the number of carbon emissions a data center owner or workload owner would need to account for would change depending on where the workload is executed.

To determine the mix of energy utilized in those regions, various databases can be referenced to determine the most likely mix of power generation for the locations of the data centers and the associated emissions factors. State data can be used to determine the percent contribution from the various energy supplies as a representation for the regional grid mix, or the data center can identify its region depending on how the grid is structured [55, 56, 57].

Grid compositions can vary regionally within a country and have even more variation from country to country [58, 59]. It is not uncommon for large corporations with global scope to have multiple digital operations spread across varying geographies. One example can be looking at Facebook's and Google's data center operations regionally in the United States.

Some large companies like Facebook and Google post their data center information in their annual sustainability reports. Google and Facebook have various locations spread across the USA, with different energy mixes contributing to their carbon footprint. Data centers in the Mid and Southwest tend to have a more significant energy consumption relative to locations in coastal regions. This could be due to the availability of land for data center construction resulting in more extensive facilities. Data center size was not mentioned in the reports.

Shifting workloads into regions that utilize more carbon-neutral energy sources, such as Oregon, can significantly reduce the carbon impact at the workload level. This is intuitive because certain regions are predominantly supported by fossil fuel-based energy supplies, such as the Fort Worth data center in Texas [60]. Fort Worth's energy mix is dominated by Natural Gas and Coal, with help from some Hydroelectric power and some nuclear power. This contrasts with the Prineville data center in Oregon being powered by a more balanced mix between Biomass, Wind, Solar, Hydroelectric, and some gas.

2.4.2 Gaps in Energy-Related Emissions Reporting

The major gaps associated with energy-related emissions lie in the degree that power plants report their emissions. Some regions don't have the requirements to, and therefore don't report emissions data making it harder to estimate the environmental impact of the energy sourced from those facilities [49]. The burden for reporting this data may be pushed onto the data center owners to attempt to gather or estimate this data and self-report the emissions associated with the power supplied to their facility.

2.5 Blockchain Based Life Cycle Analysis

Blockchain technology has been leveraged for traceability of a product's full lifecycle from raw material to end-use parts and has been established in the aerospace industry recently [61].

2.5.1 Hyperledger Fabric

Walmart has utilized Blockchain technology through the IBM Food Trust to provide visibility into their supply chain [62]. In this case, the Walmart supply chain wanted to utilize a fully transparent supply chain to establish provenance when food-borne diseases arise; the Hyperledger implementation allowed the company to reduce the time it takes to trace the origin of produce from 7 days to 2.2 seconds. This system utilized GS1 data standards and barcodes across their supply chain to identify the farm that the product originated from in one of their test cases. This system allowed for full traceability and utilization of the data for internal dashboarding efforts to understand the intricacies of their supply chain network. The team found that an open-source vendor-neutral blockchain was necessary as it needed to be used by various stakeholders within the system. The Hyperledger Fabric is a blockchain framework hosted by the Linux Foundation with the intent of a modular architecture and smart contracts as the structure for the logical implementation of the system.

2.6 Sustainability and Consumer Behavior

Recently, sustainability has been seen to affect consumer behavior depending on the product category, demographic, and social factors [63, 64, 65]. 65% of consumers have mentioned wanting purpose-driven brands, but only 26% act on it; this has been coined the "intention-action" gap. Research has shown that public policy is a crucial method in creating habits that form sustainable consumption behavior and that companies can see an increase in consumption if marketing strategies align with the consumer sentiment such as Social influence, Habit formation, Individual self, Feelings and cognition, and Tangibility, coined the SHIFT framework, to strengthen a customer's attachment to the product [64] and subsequently promote consumption [34].

Consumers are willing to purchase based on a corporation's environmental positioning, which has increased as consumers became more aware of the supply chain impact when purchasing locally sourced consumer goods [66]. Deloitte surveyed consumer attitudes and behaviors around sustainability over three years, including years during the pandemic and as pandemic tensions eased. There has been some correlation that the COVID-19 pandemic increased the focus on corporate sustainability and buying habits, but continued consumer spending on sustainable buying is seen in 2022. Consumers embrace circularity within the home, opting to repair rather than replace items. Although only one in ten have begun to purchase carbon offsets and 16% have switched to renewable energy, there seem to be some barriers to adoption perceived by the consumer [63]. Consumers are most likely to leverage sustainable options in categories of products they find essential and utilized more often. This could resonate with this sentiment as digital workloads are leveraged daily. In Deloitte's survey, consumers' top five sustainable topics were sustainable packaging and products, reducing waste in the manufacturing process, commitment to ethical work practices, reducing carbon footprint, and respect for human rights[63]. These topics align with the intent and availability of the proposed digital product passport, which could work to support the SHIFT framework to provide a competitive advantage for companies who do not have a direct high environmental impact and could then drive pressure into the supply chain for more widespread sustainable production[35].

The main barriers to adopting sustainable products include pricing, a lack of interest, and a lack of information. Findings show that showing a price reduction in the utilization of the product and providing more information is key to leveraging sustainability as a competitive advantage rather than considering a regulator burden [63, 35]. Therefore the need to establish accurate reporting and a transparent depiction of the product life cycle could support sustainability actions that promote beneficial consumer behavior and reduce operating costs in the long run, [64, 34] while providing insight into further activities that reinforce these dynamics.

2.7 Enterprise Sustainable Consumption Behavior

Investors, regulators, customers, and employees have driven an evolution in the focus of ESG standards globally throughout organizations. Becoming a mainstream business topic, executives are driving sustainability strategies into their business operations to drive competitive advantage to increase operational and financial performance [2]. In a survey by the IDC, over 70% of organizations with an employee count of over 10,000 have committed to a net-zero target and have begun to focus on verifying their progress toward these goals. This momentum will see a shift in these large organizations' reporting methods and rigor currently in practice, refocusing how they spend and implement their IT structure. Over 51% of respondents already recognize that their sustainability efforts deliver operational and financial benefits through enhanced brand perception, customer loyalty, and the ability to identify inefficiencies beyond general value from an investor-focused ESG sales pitch aligned with reducing regulatory and public-relations risks[2]. These factors lead the IDC to conclude that ESG performance will rank in the top 3 decision factors for IT equipment purchases, with over half of the requests for proposals including metrics targeting carbon emissions, material use, and labor conditions[2]. Furthermore, increasing requirements to provide transparency in data center power consumption, use of renewable energy sources, IT recyclability, and waste reduction is driving trends pushing data center providers to disclose more than ever to their stakeholders. This also drives end-to-end visibility requirements of IT vendors' sustainability process concerning the product that enter these data centers, promoting circularity as a critical component of the product development process.

2.8 Power Aware Workload Placement

Power models have attempted to characterize the potential energy consumption of various workloads based on the type of workload and a model to correlate the consumption of hardware elements such as the CPU, memory, and switching devices [67]. This would allow some ability to predict a somewhat optimal workload placement based on operational emissions, but this has not been widely leveraged or implemented as well.

Recently, research has focused on reducing energy consumption using modern computational methods that mix K-means, ARIMA, threshold methods, and Linear Integer Programs [68, 69]. The goal is to reduce latency driven by the decision-making process for workload placement and computational complexity. The most leverage method found was to predict workloads a period in advance, day-ahead forecasting, in conjunction with compute clusters, power models, and carbon-intensity forecasting, among other constraints to shift workloads temporally. Future work was stated to focus on spatial distribution following a similar framework at Google data centers [70].

2.8.1 Multi-Armed Bandit Algorithms

In this research, it can be seen that a big issue with implementing smart workload placement is the lack of available data to be leveraged for prediction models or forecasting algorithms. The multi-armed bandit algorithms focus on decision-making with uncertainty to attempt to identify the optimal solution. William R Thompson introduced this concept in 1933 and published it in Biometrika. These algorithms have been widely used at scale to explore unknown environments with a large number of possible options like pricing and ad placements in large platforms like Amazon, recommendation engines like those found on Netflix, clinical trial designs, network routing, dynamic pricing, and relevant to this discussion resource allocation [71]. Certain classes of bandit algorithms have been leveraged in the study of reinforcement learning and could be applied to complex learning problems.

These algorithms are a sequential game between a learner and the environment over a horizon of n rounds. The learner chooses options based on historical results, called a policy. The key to these algorithms is that the learner needs to make the best possible decision while the results of the environment class are unknown. In Thompson Sampling, the learner chooses a prior based on their available options. They are practical concerning computational resources and observation. This method focuses on randomization within a given distribution based on previous learning. When the distribution is large, there is a tendency to explore options until a clear winner is established. This leverages a Bayesian or Frequentist approach; Bayesian was chosen to be studied here to focus on establishing priors associated with workloads to "learn" the data center's environment. Poorly chosen priors are left behind, and optimal solutions are generally leveraged while minimizing the theoretical regret, the difference between the average reward from the chosen arm, and the known optimal solution.

2.9 Gaps in Data Center Carbon Accounting

Data centers can access or request information associated with their regional energy mix by contacting their utility provider. This allows them to correlate their aggregate usage to their environmental impact using grid emission factors representing the appropriate portfolio mix associated with the power entering their facilities. Although this is great for the data center owners, this does little to allocate emissions to the customers within the data center to allow monitoring of the contributions to these aggregate emissions. It is not common, although some companies such as Microsoft have started to offer these calculations for their cloud operations [72]. It is unclear whether this includes the emissions from co-located data centers, so it is assumed this is based on Microsoft-owned data centers. The documentation doesn't include the embodied emissions within their data center, contributing to the total environmental impact on top of the operational emissions. This information is essential; the data center facility and manufacturing hardware are only there because of the market's need; therefore, the workloads are the sole reason the resulting emissions exist and need to be accounted for by the end user as part of their downstream scope 3 emissions. This allows a company to understand the full impact of its business on the environment.

The Carbon Trust and the Global e-Sustainability Initiative have developed ICT sector guidance with a sector focused on cloud computing and data center services. This document serves as a guide for industry and academia to perform research. Still, data collection, reliable and consistent sources of secondary data, and actual usage data for ICT hardware are not fully mature. Therefore, these strategies aren't standard in the industry yet [73].

2.9.1 Gaps in Accounting Methods

Given how new emissions reporting is to companies, there will be a delay in adoption, but there is a precedent for shifting to rigorous emissions reporting. Driven by environmental concerns, power plants moved to rigorous reporting methods by the Energy Policy Act of 1992 [59]. This type of reporting is necessary to gain insight into the emissions associated with the production and implementation of ICT equipment due to the scale at which the data centers are operating and decommissioning equipment to keep up with technological advances. Power consumption in one country can be sourced in many ways based on the facility's location and cannot be treated as one portfolio of emissions. Businesses wouldn't know if they are better or worse than the industry standard and wouldn't understand how they can influence their value chain to meet their sustainable goals. In manufacturing, industry-assumed unit process emissions do not include how that facility manages waste or their efforts to introduce efficiency measures such as automation or LED lighting. Streamlined approaches like the PAIA tool tend to obscure the product carbon footprint reporting between products and are generally only able to be leveraged to find how general components in the modeled hardware affect our environment. Product 1 wouldn't be able to differentiate from Product 2 since both could have metals sourced from different parts of the world, have silicon made in a hydroelectric-powered facility or one from a coal-powered facility, or have post-consumer recycled plastics in one versus the other.

Chapter 3

Proposed Method

Organizations interested in reducing their environmental impact to meet their sustainability goals will need to consider more than just power consumption at the processor. They will also need to explore the energy sources supplied to other devices and the bus bars of the data center in which their workloads reside. Where workloads are distributed will affect the level of GHG emissions output to meet that workload. Furthermore, the following study sets a framework for evaluating various IT workloads across many different types of server architectures; the current trend focuses more on balancing loads across a homogeneous server cluster. These architectures can expand beyond focusing on CPU clusters and provide a way to distribute loads in a disaggregated manner to scale the capacity of the data center [74]. This can be further extrapolated to data center locations to target sustainability goals.

GPU utilization is a typical example. Able to process large amounts of data, this technology has gained significant use for networking and data sharing and is now being used to advance the field of machine learning [75]. CPUs are good at facilitating sequential tasks very quickly but struggle with parallel processing across various tasks. GPUs are typically an extensive array of relatively slow processors that work in parallel, enabling them to process large amounts of data.

The initial rise of GPUs used to support the networking performance of video streaming showed that they were still deficient in some applications. CPUs are still utilized in parallel; an example is recognition tasks, where they outperform GPUs.

Similar applications are now used to host machine learning algorithms where tradeoffs between the processing units are leveraged for different aspects of the work to be done. A GPU still needs a CPU to manage operations such as passing data from peripheral cards to the GPU for processing, a sort of command center.

The compute nodes can emulate this in the data center, but across servers rather than singular systems and optimized for sustainability. The military has already driven a need for computing resources to operate reliably and securely, albeit in much harsher climates. Designing the systems for these harsher climates allows data centers to run hotter and more humid. These systems are developed within strict size, weight, and power constraints. All of which can be optimized for sustainability. Reducing the size and weight leads to a reduction in shipping weight and cube size to improve shipping density and power to reduce overall energy consumption.

Google has already made significant efforts to relocate non-mission-critical workloads to locations with larger mixes of renewable energy sources in their portfolio. Data center owners will soon need to deviate from the practice of additionality, where they are only paying an equivalent "tax" or fee to offset their environmental impact on the carbon-intensive energy they are consuming. As technology advances and processors become more power intensive, sourcing renewable energy locally to the data center and encouraging renewable energy production at the regional grid will be crucial in meeting sustainability goals and regulations.

Automating and standardizing reporting will be vital to creating a truly effective system for companies to accurately manage their digital carbon footprint at scale[5, 52, 54].

3.1 Framework

The process of analyzing what machines should be used for workloads and where they should be placed will be explored. Embodied emissions will first be analyzed to understand the general environmental impact of similar servers. Then, a standard workload for power consumption benchmarking will be used to compare the performance per unit of energy for various enterprise servers simulating a containerized platform. Then the total server performance will be examined based on the benchmarked metrics. Finally, the difference in emissions between data center locations will be explored using the emissions factors from their respective electrical grid.

3.2 Calculating Embodied Emissions

The PAIA tool is widely adopted today, with several analyses published by different hardware vendors. Due to the limited availability of detailed hardware information and the idea that this information may only exist with the vendors, the published carbon footprints in lbCO$_2$e values using this tool will be used as a general reference for the environmental impact of the hardware. Table 3.1 shows the 28 servers chosen for analysis and whether or not their carbon footprint reports are published.

3.3 Embodied Emissions

Server Index	Brand Name	Model Name	Product Carbon Footprint (lbCO2e)
18	CISCO	UCSB-B480-M5	Not Available
25	DELL	E41S	Available
32	DELL EMC	E03B	Available
47	DELL EMC	E05B	Not Available
57	CISCO	UCSC-C480-M5	Not Available
61	DELL EMC	E49S	Available
62	DELL EMC	E54S	Available
97	Fujitsu	PRIMERGY RX4770 M5 E-Star Fam4	Not Available
104	Hewlett Packard Enterprise	DL580 Gen10	Available
121	Fujitsu	PRIMERGY RX4770 M6 E-Star Fam4	Not Available
128	Hewlett Packard Enterprise	HPE Superdome Flex 280 Server 4P	Not Available
131	Hewlett Packard Enterprise	HPE Synergy 660 Gen10 Compute Module	Available
149	Hewlett Packard Enterprise	HPE Proliant DL560 Gen 10	Not Available
177	Inspur	NF8260M5	Not Available
178	Inspur	NF8260M6	Not Available
179	Inspur	NF8480M5	Not Available
180	Inspur	NF8480M6	Not Available
191	IBM	Power Systems E950	Not Available
192	IBM	IBM Power E1050	Available
198	HPE Proliant DL560 Gen 10	HPE Proliant DL560 Gen 10	Not Available
201	Lenovo	Lenovo ThinkSystem SR850P Server	Available
204	Lenovo	Lenovo ThinkSystem SR850 V2	Available
213	Lenovo	Lenovo ThinkSystem SR950 Server (Xeon SP Gen2)	Available
214	Lenovo	ThinkSystem SN850	Available
224	Inspur	SA8212H6	Not Available
234	Lenovo	Lenovo ThinkSystem SR860 V2	Available
258	Lenovo	ThinkSystem SR850	Available
261	Lenovo	ThinkSystem SR860	Not Available

Table 3.1: Embodied emissions reported for the servers in the population published by the brands.

3.4 Using the Server Efficiency Rating Tool for Comparing Power Consumption

The Server Efficiency Rating Tool (SERT) suite of workloads [76] serves as a set of standard workloads that can capture power consumption relative to performance for a workload run across almost any type of data center equipment [77, 78]. The Standard Performance Evaluation Corporation, a leading benchmarking organization that is widely recognized throughout the computing industry, owns the SERT suite.

As companies move to cloud-oriented operations, their workloads will sit on top of an architecture and operating system agnostic framework similar to the SERT workload. This level of analysis can become a standard industry practice and drive hardware manufacturers to promote sustainability-oriented development by optimizing hardware for specific tasks such as Tensor Processing Units for tensor multiplication

[79].

The SERT Suite measures the power consumption of the System Under Test for various worklets meant to measure the consumption of CPU, Memory, and Storage operations at different load levels and idle power draw. This information is widely used to characterize the power performances of systems and is publicly reported by the government program that certifies servers and electronic products called Energy Star. This database will be used to understand the power draw of hardware relative to various workloads.

3.4.1 SERT Worklets

The SERT worklets measure performance per watt with metrics specific to their workload, including transactions per second, bandwidth, transactions per second times the square root of the data store size, and IO operations per second. The worklets are meant to represent everyday operations on data center hardware.

All CPU worklet performance metrics are in transactions per second and are described in more detail below. These worklets' performance should increase with more resources or more capable resources. The workloads are Compress, CryptoAES, LU, SHA256, SOR, SORT, and SSJ.

- Compress seeks to understand the power consumption during compression and decompression operations.

- CryptoAES implements a transaction that encrypts and decrypts data using block cipher algorithms, AES or DES.

- LU exercises large mathematical operations and computes the LU factorization of dense matrices to achieve the number of calculations per second.

- SHA256 is an operation focused on security and performs hashing and encrypt/decrypt operations on hardware.

- SOR uses Jacobi Successive Over-relaxation that exercises typical access patterns in finite difference applications; this is another mathematical operation.

- SORT takes a randomized 64-bit integer array and sorts this during each transaction; this is one of the most critical operations in computing.

- SSJ exercises the CPU, caches, and memory of the System Under Test to simulate Online Transaction Processing. This worklet identifies the maximum number of transactions that is capable by the system and then runs the workload from a peak of 100% down to system idle.

The Memory worklets measure memory bandwidth across array operations that are important within the system and the impact of scaling across the system's operations. The worklet's performance should be higher for systems with better memory characteristics. These workloads are Flood3, and Capacity3

- Flood3 utilizes four memory bandwidth tests that explore the system's performance across four common and important array operations. The aggregate system bandwidth was averaged across the four tests and multiplied by the amount of physical memory installed. The metric used for this test is bandwidth in gigabytes per second.

- Capacity3 uses input data for a particular transaction instance at random. This information is retrieved pre-computed allowing the workload to focus on memory performance with some CPU performance taken into account. The metric for this workload is transaction per second times the square root of the data store size.

The storageIO workload utilizes four transactions, two random and two serialized, with each pair having a read and write element. This worklet focuses on the data access performance of the system and is measured in IO operations per second. The workload should show higher performance numbers for systems with better storage network characteristics, such as higher bandwidth and lower latency.

3.4.2 Calculating Sever Throughput

$$\text{Server Power Capacity} = \text{Qty Server Power Supplies} \times \text{Power Supply Capacity} \tag{3.1}$$

$$\text{Server Throughput} = EFF_{\text{workload}} \times \text{Server Power Capacity} \tag{3.2}$$

The SERT energy efficiency scores from energy star utilize normalized throughput and power consumption to report a geometric mean of the load level scores referred to as the $\text{EFF}_{\text{workload}}$. The unit of analysis is normalized transactions per Joule and is taken by dividing the worklet's metrics, transactions per second, and multiplying by the Watts to get transactions per Joule as seen in equation 3.1. For simplification, the total power capacity of the system will be calculated to be the power supply capacity in Watts multiplied by the number of power supplies provided by the Energy Star database. This will result in total transactions per second for the entire server represented by equation 3.2.

3.5 Calculating Energy-Related Emissions for each workload

$$Workload_{\text{mWh}} = \frac{\text{Server Throughput} \times \text{Sever Power Capacity}}{3600 \text{ secs} \times 10^6 \text{ Watts}} \tag{3.3}$$

$$CF_{\text{workload}} = DC_{\text{Carbon Intensity}} \times Workload_{\text{mWh}} \tag{3.4}$$

Two data center locations chosen are Dallas and the San Francisco Bay Area. These two locations both have many data centers and some of the largest. These two locations show an interesting difference in their mix of energy sources according to the eGrid database.

The eGrid database reports the lbCO_2e per MWh by state. This factor will determine the workload emissions by the servers for the two data center locations. To get to the MWh from transactions per second, the run time will be calculated using 100,000 workloads as the baseline. This will return a value in seconds. This will be multiplied by the server wattage to calculate a Watt-second value. Then this value will be multiplied by 3,600 to achieve Watt-hours and divided by 100,000 to complete Megawatt Hours. Finally, this will be multiplied by the emissions factor to calculate a value in lbCO_2e.

3.6 Multi-Armed Bandit Implementation

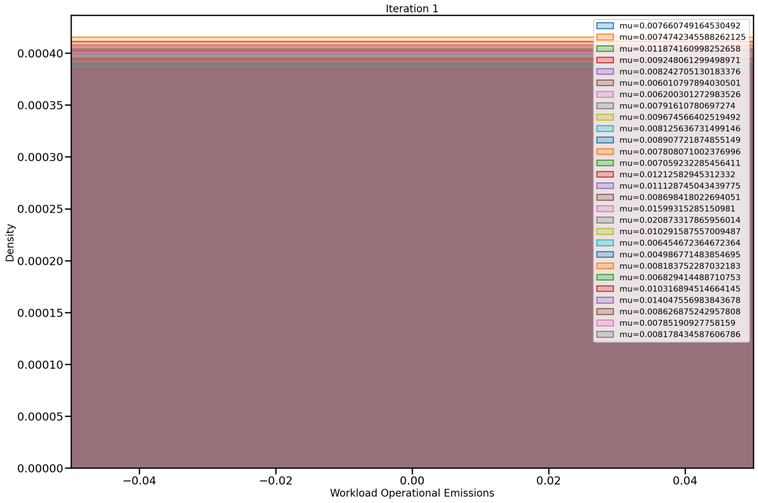

Figure 3-1: Initialization of the bandit algorithm where the distribution is centered around zero and wide, representing an unknown data center environment.

Workloads are typically placed on fixed hardware clusters within the data center. Power consumption data may not exist if the hardware was never run on other systems within the data center, which makes it challenging to identify the optimal ICT asset for that specific workload. Because of the ability to explore an unknown environment and subsequently use that information to find and exploit an optimal solution in real-time, the multi-armed bandit, or k-armed bandit, algorithm was chosen to explore and learn a hypothetical data center portfolio. A series of portfolios were generated to understand the capability of the algorithm to identify the optimal server given three scenarios utilizing Thompson Sampling. The first scenario explores all 28 servers previously mentioned, with their emissions calculated using the California emissions rate. This simulates finding the optimal server within a data center in California. The second scenario randomly places ten servers in a California data center and ten in a Texas data center and then runs the bandit algorithm to identify the best server and data center mix. This scenario aims to simulate a random combination of hardware in two different locations. Finally, the last scenario randomly places ten servers in an Idaho data center, ten in a Washington data center, and eight in a Maine data center. This scenario simulates a random portfolio of servers across the United States.

The emissions were calculated for each scenario's data center portfolio using the emissions rate from the state of the theoretical data center and a set of one million transactions as one batch operation. The emissions numbers were used as the real mean of the server's distribution with a standard deviation of 0.02 lbCO_2e, which provided a somewhat realistic and practically wide distribution for the actual power numbers. The bandit was initialized to have all posterior distributions set to a mean of zero with a standard deviation of 1000 to simulate a wholly unknown and wide distribution. The algorithm's objective was set to minimize the emissions numbers and then run for 1000 iterations. One thousand iterations were chosen arbitrarily; if distributions for power consumption overlap, the bandit algorithm may jump back and forth between multiple sets of resources. This could be caused by fluctuations in operating temperatures of the inlet air into the server, potentially caused by increased workloads on surrounding servers. There are methods such as the Gittins Index or discounting rewards [71] that can be used to determine at what point is the reward for this switching less than the cost of initializing workloads on different pieces of hardware. Because this cost is unknown to the author and the subject of dynamic programming is outside the scope of this work, this was considered but not implemented.

Chapter 4

Results

4.1 Server Population

The Energy Star Enterprise database was used to identify commercially available servers. For this analysis, servers with four processor sockets were chosen to reduce the number of servers in the population for analysis. Table 4.1 shows the server brand and model name and the index used to reference the server in the following charts. A principal component analysis was done to understand the similarity of the servers in the population. A principal component analysis explores the relationships of the features for the server population. It uses statistical measures to identify orthogonal relationships between features for each observation in the population. It then creates a two or three-dimensional representation of each observation that can be used to explore similarities and differences between observations. Most of the server's processor architectures are similar, with the two IBM servers as outliers. This could be because they are the only products not leveraging Intel servers and have significantly higher processor operating frequencies and lower core counts.

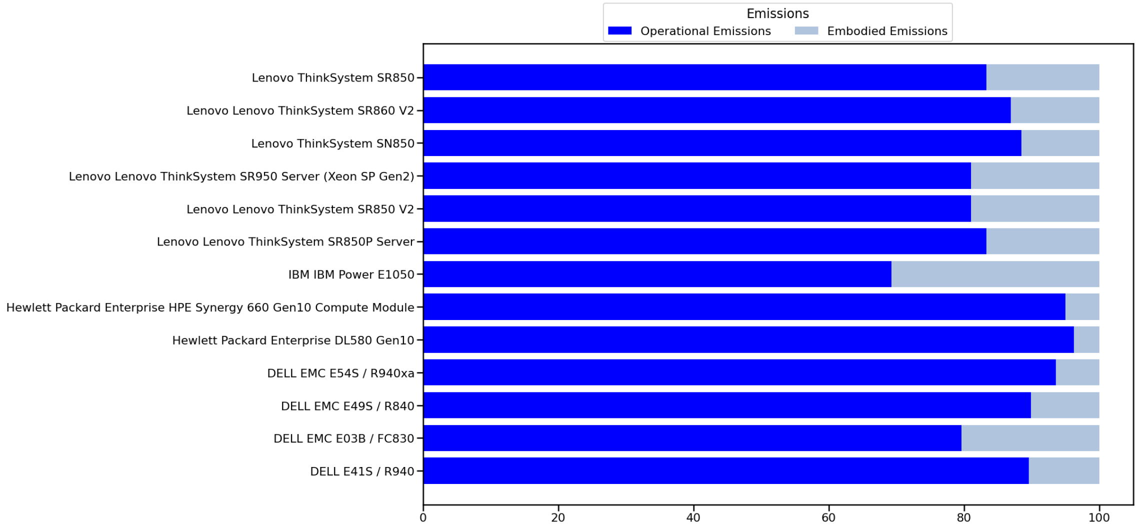

Figure 4-1: Lifetime server operating emissions versus embodied emissions as reported by the manufacturer.IBM, Dell, and HPE model 4-year lifetimes with usage model based on the PAIA EU energy factor. Lenovo models a 5-year lifetime, except for the SR850-v2, all modeled based on the PAIA US energy factor.

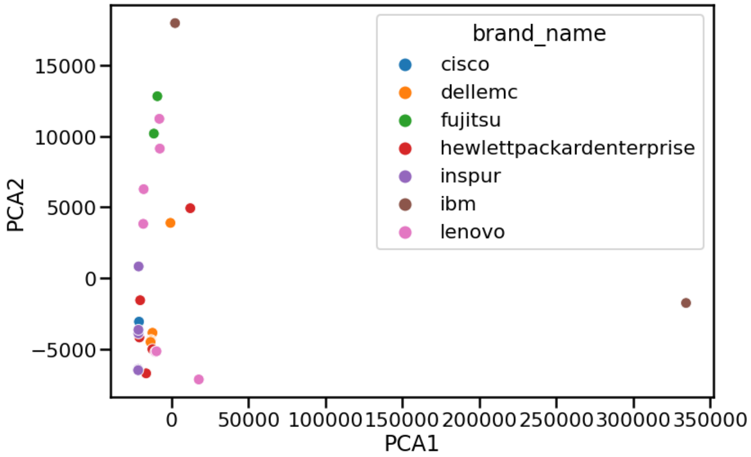

Figure 4-2: Server principal component analysis to understand the likeness of the servers.

After searching for publicly available product carbon footprint reports, it can be seen that over half of the servers in the population did not have a product footprint report published. The PAIA tool was exclusively used for the product carbon footprint of the servers, with a report published. Of the product carbon footprint reports found, the total product carbon footprint from manufacturing, end-of-life, packaging, and transportation was reported as a percentage of the lifetime emissions over the product lifecycle. This represents the product's embodied emissions, but variation in the servers' lifespan affects the volume of operational emissions. The values varied greatly, even for products of a similar size which could allude to varying assumptions, versions of the PAIA tool used, and possible variations in the accuracy of the inputs into the PAIA tool, as mentioned in the literature review.

Due to this variation, the embodied emissions for products are not used in the accompanied analysis since the methods cannot be confirmed to be consistent for producing a carbon footprint number. This would be alleviated given a standard procedure for reporting embodied emissions for ICT assets with consistent assumptions for the analysis. Instead, the percent of embodied emissions concerning operational emissions was explored. These can be somewhat inconsistent since many companies estimate the lifetimes of their systems in different ways, which could lead to variations in operational emissions. Not all companies disclosed the estimated lifetime for their analysis. Embodied emissions make up between 30 and 5 percent of the total emissions for the servers based on the reports used in figure 4-1.

4.2 Server SERT Worklet Capacity

It can be seen that for the suite of SERT worklets, the server emissions vary by worklet greatly. In this analysis, the servers are identified by an index, and the configuration of the top three performing servers is examined in more detail. This analysis utilizes the SERT workload efficiency score to establish transactions per second by multiplying the efficiency score, transactions per Joule, by the system's total power, Watts. The system's total power capacity was calculated by multiplying the number of power supplies by the power supply capacity. The central assumption here is that the system's power architecture allows full utilization of the power available to the system. This could be misleading as some system architectures provision redundant power supplies, and there are no extreme outliers in the performance calculations, so this is assumed to be a reasonable assumption to make.

4.3 CPU Workloads

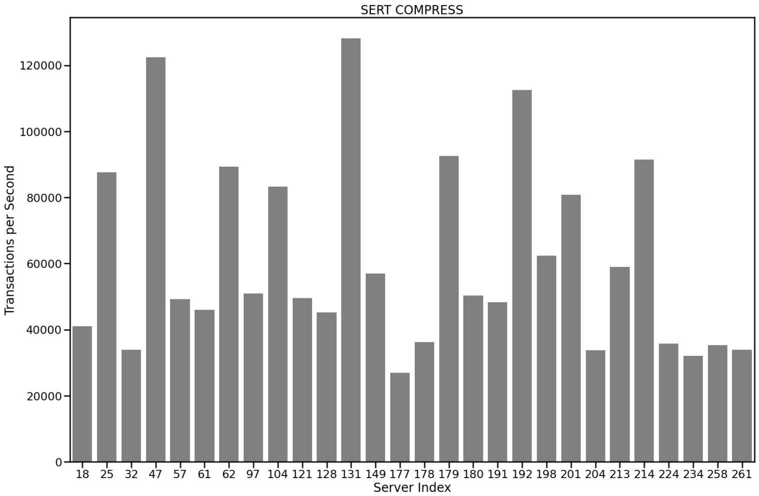

Figure 4-3: Server capability in transactions per second for the SERT Compress workload.

The Compress workload benchmarks the capability of a server during the compression and decompression of data. In this workload, three servers stand out with a higher capacity than the rest. The top-performing server is an HPE blade server running an Intel processor with 20 cores at an operating frequency of 2 GHz. This is followed by a Dell EMC blade server running on an Intel processor with 20 cores at an operating frequency of 2.1 GHz. Finally, this is followed by an IBM Power server running on an IBM 12-core processor operating at 3.35 GHz. This shows a variation in system architecture where the two Intel-based systems are not working at the highest frequency in the available pool of servers. The performance of these servers slightly leads the IBM server performance, which runs on an entirely different architecture, but at a significantly higher operating frequency.

The CryptoAES workload encrypts and decrypts data using a specific block cipher algorithm. In this workload, it is apparent that the magnitude of performance changes between the servers as compared to the Compress workload. Two Dell servers lead in performance, running Intel processors with 20 cores at 2.1 GHz, followed by the HPE server running a 20-core Intel server at 2 GHz. Among the top performers for the Compress workload, the IBM server is one of the worst-performing servers and lands in the bottom three.

The SOR workload exercises typical access patterns in finite difference applications using Jacobi Over-Relaxation to solve Laplace's equation in 2 dimensions. This workload shows variation in performance among the population of servers examined compared to the Compress and CryptoAES servers. Still among the top three servers are the previously mentioned Dell and HPE servers, with another HPE server running a 20-core intel processor at 2 GHz joining the top three in second place. The HPE Synergy 660 (index 131) has consistently been one of the top three performing servers for the workloads discussed but varies between first and third.

The SERT SORT workload uses a sorting algorithm for a randomized 64-bit integer array as a benchmark. Again, there is a variation in performance across the population of servers studied, but the top three performing servers are consistent with the SOR workload. This could mean that the operations tap into similar aspects of the system architecture to realize better performance.

The SHA256 workload performs SHA-256 hashing transformations on a byte array. This workload has the same top three performing servers as the Compress workload, albeit in a different order. The IBM server running on a Power processor leads in performance, followed by the HPE server and Dell servers that consistently perform well in all workloads discussed.

The LU workload computes the LU factorization of a dense matrix that uses partial pivoting focused on benchmarking linear algebra kernels and dense matrix operations. The top-performing workloads are the same three servers as the SOR and SORT workloads, potentially alluding to an ideal architecture to support these mathematical operations.

4.4 Memory Workloads

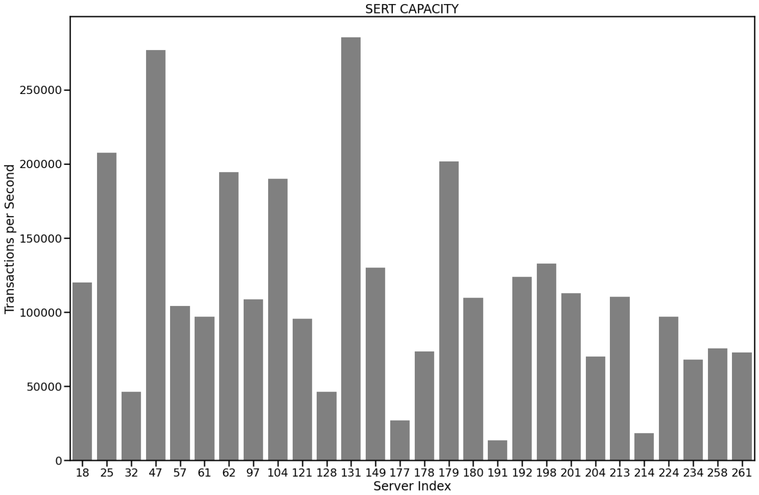

Figure 4-4: Server capability in transactions per second for the SERT Capacity workload.

The Capacity workload performs XML operations on a minimum, and maximum data set and scales with memory capacity leading to higher performance in servers with higher memory capacity. The top three performers in this workload were the HPE and the two Dell servers previously mentioned. These servers did not have the highest memory capacity, ranging from 320 to 380 gigabytes versus the maximum of three terabytes. This could be due to the performance of the system architecture associated with memory access and the interfaces between the processor and memory within the system.

4.5 StorageIO Workloads

The StorageIO workload tests the server's access to internal storage in two methods, read and write to a random file location and then sequential locations. This workload showed the highest degree of variance between the top performers and the general population of servers. The StorageIO sequential workload's top performers were two Inspur rack mount servers. The top Inspur server runs a 20-core Intel processor at 2.4 GHz, followed by a 24-core Intel processor at 2.4 GHz. The third top-performing server for this workload was the IBM server running a Power processor

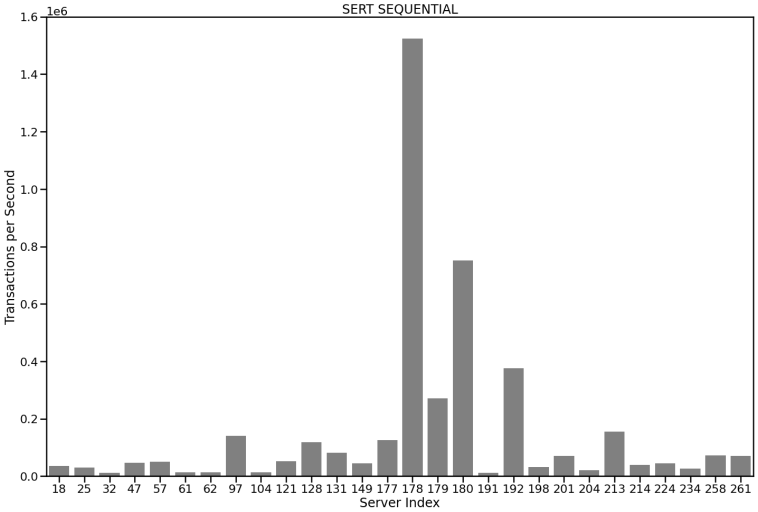

Figure 4-5: Server capability in transactions per second for the SERT Sequential workload.

at 3.35 GHz. The StorageIO random workload's top performers consist of the same two Inspur servers that were the first and second top performers for the sequential workload, but in this case, they were the first and third performers. A Lenovo server running a 28-core Intel processor at 2.3 GHz was the second top performer for this workload. This server was the 6th top performer in the sequential workload. Many top-performing servers were running Intel processors. But, a large portion of the studied population also ran Intel processors. This could allude to the fact that workload capacity could be impacted beyond the processor architecture, which includes core count, operating frequency, and overall system architecture.

4.6 Calculating Emissions Based on Regional Grid

Figure 4-6 displays the range of emissions factors associated with all 50 states in lbCO$_2$e per megawatt hour. It can be seen that states such as West Virginia (WV), Wyoming (WY), and Kentucky (KY) have relatively large emissions factors relative to the population. Vermont (VT), Washington (WA), and Idaho (ID) have the lowest emissions factors relative to the population.

In the analysis, it was seen that the system's architecture showed variation in performance between workloads. Most of the architectures studied were Intel-based,

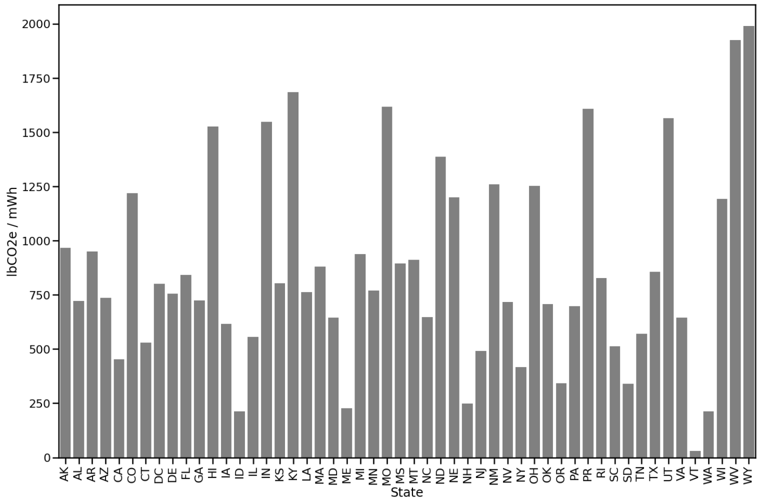

Figure 4-6: Emissions factor in lbCO$_2$e / mWh by US state.

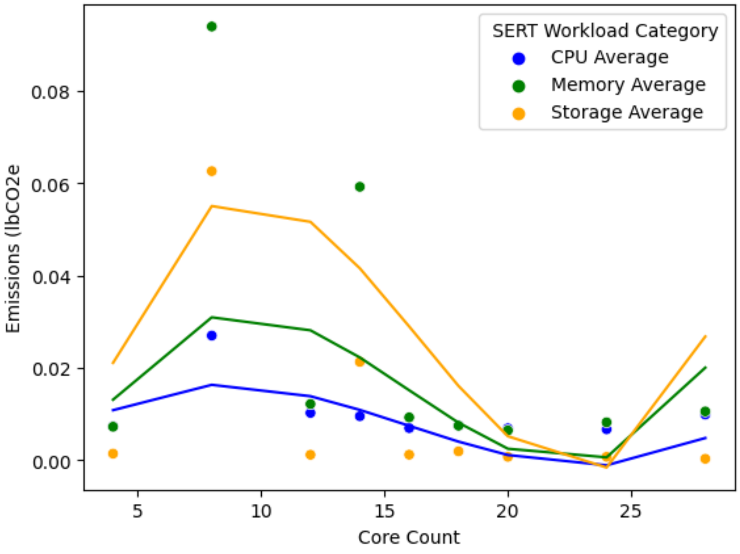

Figure 4-7: Graph of emissions data by core count for each SERT workload type. The highest emitting core configuration for the CPU workloads was four times the lowest emitting configuration. For storage workloads, the highest emitting configuration was 14 times the lowest, and for memory workloads, it was 179 times the lowest emitting configuration.

Workload	Equation	R^2 Value
CPU	$-0.066 + 0.0062x - 0.0005x^2 + 1.047 \times 10^{-5}x^3$	0.48
Storage	$-0.0358 + .017x - .0013 + 2.729 \times 10^{-}5x^3$	0.40
Memory	$-0.0683 + 0.0308x - 0.0023x^2 + 4.719 \times 10^{-5}x^3$	0.46

Table 4.1: Polynomial regression for figure 4-7

and there was still a significant variation in performance, so the core count was explored to understand the difference between the system's processor architectures to see if there were some processor-based architectural considerations to be made. Figure 4-7 shows that 8-core processors could typically execute their workloads with the most operational emissions. Interestingly, increasing the number of cores did not significantly impact the emissions; it could have been mistakenly assumed that reducing the workload run-time would positively correlate with core count since more cores can run the workload in parallel. There is variation between the workloads and the optimal core count for reducing emissions, which alludes to the notion that the server running the architecture impacts the emissions associated with a given workload type. A third-order polynomial regression was calculated. It was seen that the model for the CPU workloads was the best correlated to the data with an R^2 value of 0.48.

eGrid subregions emissions were explored how broader areas in the United States affect the operational emissions and can change regionally, say moving a workload from the East Coast to the Mid-West. It can be seen that there are also more significant variations in regional emissions rates. Shifting a workload from the SERC Mid-West region to the NPCC New England region results in a 24% reduction in emissions. This information can be compared to the energy cost relative to a company's carbon offset prices when placing their workloads and aiming to meet their environmental goals.

Observing emissions from the Compress workload, it is evident that a data center in California will most likely have lower emissions when comparing a server in both locations. But, this assumed that both sites have the same portfolio mix of servers. If the data center's cluster only had access to server type 201 in Texas, 0.009 $lbCO_2e$, and

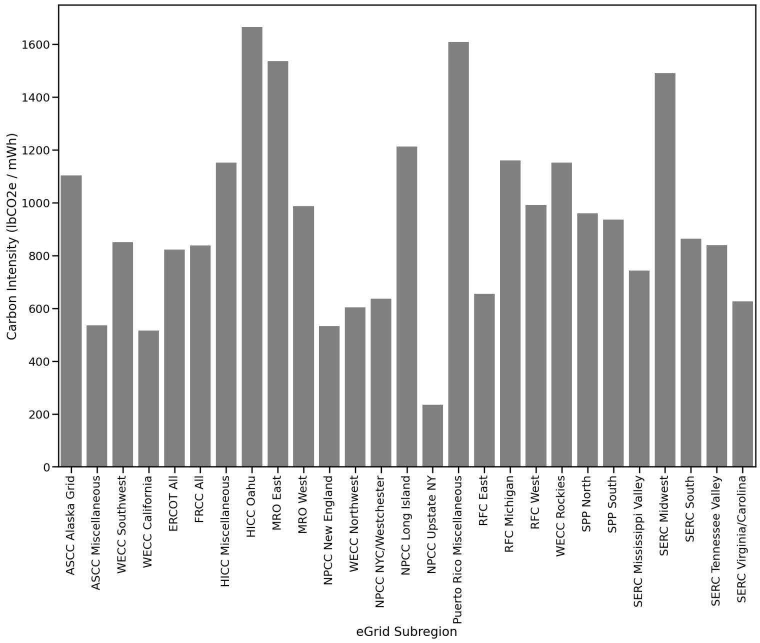

Figure 4-8: Server emissions rates based on regional placement for the SERT Compress workload.

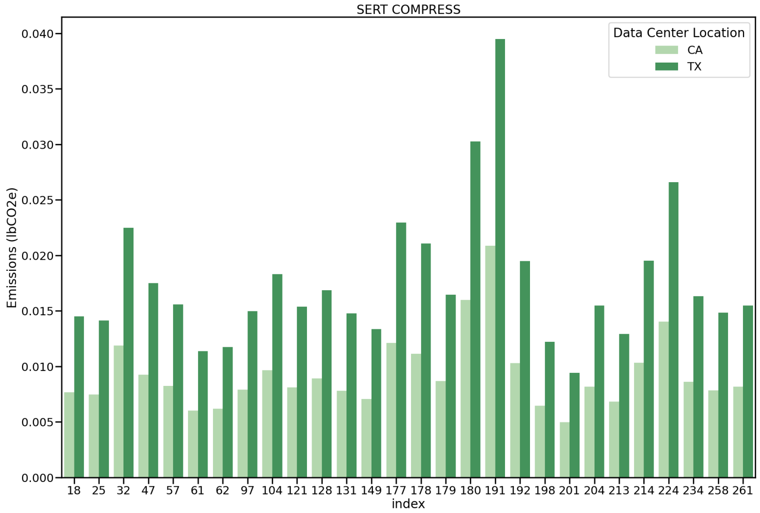

Figure 4-9: Comparing emissions for servers between California and Texas. California's average emissions were 0.009 $lbsCO_2e$ for the SERT workload, while Texas' average emissions were 0.017 $lbsCO_2e$

server type 192 in California, 0.019 lbCO_2e, it would make sense to route the workload to the Texas data center. Therefore, any algorithm that would route workload based on emissions should consider the portfolio mix and the data center location. The price paid for electricity was not considered directly in this analysis; lower energy prices in Texas without considering emissions may be a deciding factor for stakeholders, given their priorities. It can be seen that the Flood and Sequential IO workloads have the highest emissions due to energy consumption. This could be because the flood workload would leverage the system CPU and memory elements. Some of these workloads' energy efficiencies may depend on the periphery devices' efficiency instead of just the CPU.

SERT Workload	Average Emissions in CA (lbsCO_2e/batch process)
Compress	0.0093
Crypto AES	0.0115
LU	0.0074
SOR	0.0107
SORT	0.0102
SHA 256	0.0067
Flood	0.0694
Capacity	0.0089
Sequential	0.0156
Random	0.0062

Table 4.2: Average operational emissions by workload. The emissions from the FLOOD workload are 11x the emissions for the Random IO workload given the same basket of hardware.

Server	lbsCO_2e	Location
Dell EMC r840	0.0003	Vermont
IBM Power e1050	0.0917	Wyoming

Table 4.3: Sensitivity study based on the Compress workload.

A crude sensitivity study was done to explore how the best-performing server in the state with the lowest emissions rate compares to the worst-performing server with the highest emissions rate. The results are shown in table 4.3. There is a significant variation in emissions that are several orders of magnitude. Ideally, these workloads would be able to be placed in the optimal region on the optimal solution, but this may not be feasible.

4.7 Multi-Armed Bandit Workload Distribution

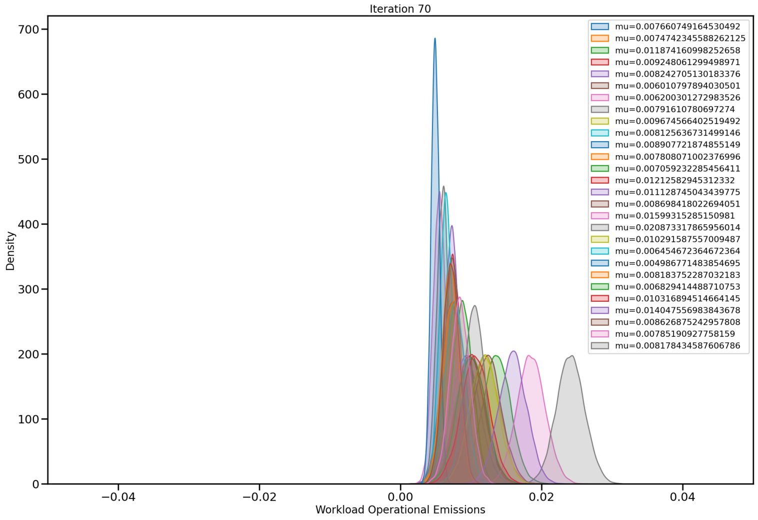

Figure 4-10: The multi-armed bandit algorithm begins to exploit the optimal server in scenario one after only 70 iterations.

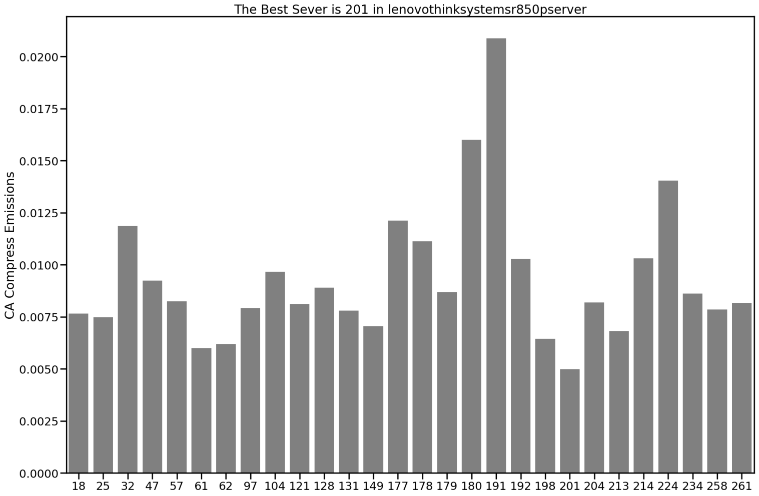

Figure 4-11: Scenario 1 portfolio and algorithm output.

The bandit algorithm was set to minimize the emissions associated with each execution of the batch operation to minimize the total emissions for all runs of the batch operation, 1000 in this case. The results of the bandit algorithm correctly identified the optimal solution for scenario 1 in less than 70 iterations with significant learning across the portfolio. The servers that have not been exploited in this case can be dropped from the portfolio, and the near-top performers can be identified and leveraged if the optimal server is at operating capacity.

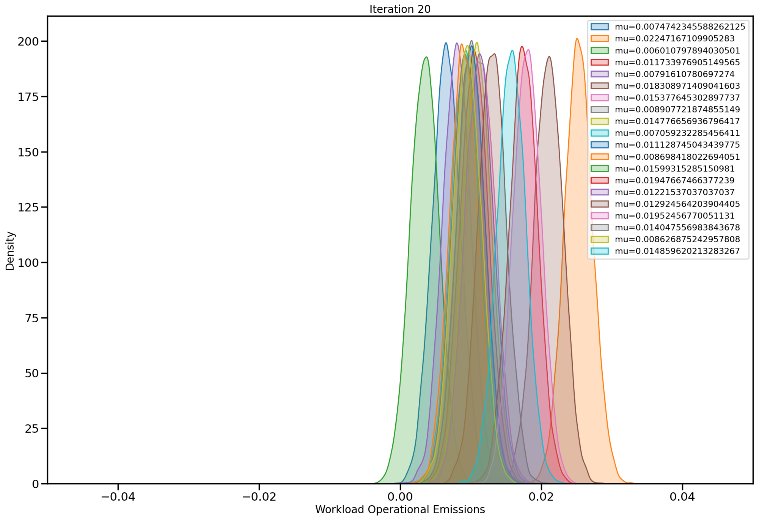

Figure 4-12: An example of the posterior distributions after 20 iterations during scenario2

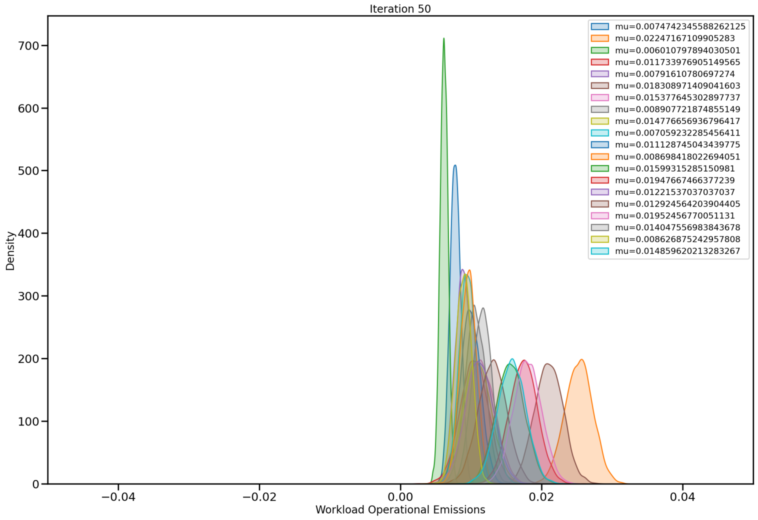

Figure 4-13: The bandit algorithm identifying and exploiting the optimal server after 50 iterations in scenario 2

Figure 4-12 shows the bandit algorithm after 20 iterations as it is learning the optimal server given the set of 20 servers. This shows the algorithm's ability to gather valuable information and exploit the optimal server after only 50 iterations.

The server would have found the optimal server in less than a year if done weekly. The results for scenario two can be found in Appendix B.

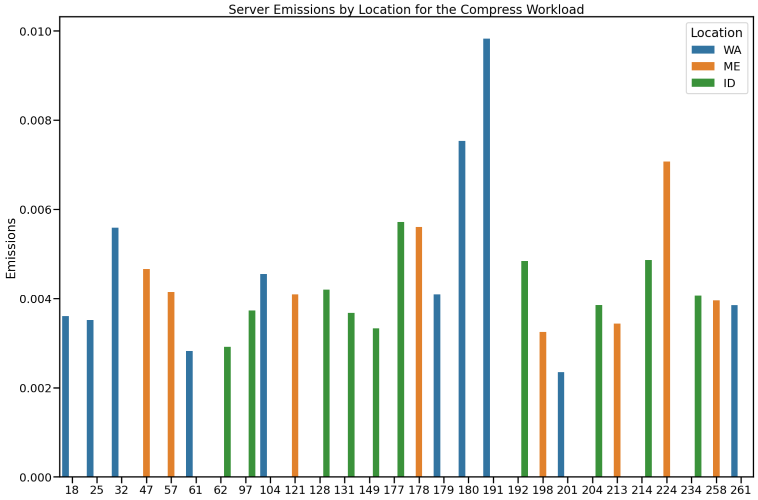

Figure 4-14: Results for scenario 3 showing the algorithm differentiating between server and data center location for the same workload. In this scenario, the three servers with the lowest emissions to complete the same task had emissions within 0.0006 $lbCO_2e$ of each other. Average emissions in Idaho were 0.0041 $lbsCO_2e$. Average emissions in Maine were 0.0045 $lbsCO_2e$. Average emissions in Washington were 0.0047 $lbsCO_2e$.

Finally, figure 4-14 shows the output from the bandit algorithm for scenario 3. These three states had the lowest emissions rates, and the algorithm converged on the lowest emitting servers over 1000 runs, as shown in figure B-9. This drives home the notion that the algorithm can be used to identify a reasonable solution from an uncertain environment and place workloads where emissions are the lowest. The algorithm can be exploited via software and used to orchestrate workloads passively with minimal computational resources, assuming reporting capabilities are present and can operate in real time. In this case, servers with the lowest emissions for this workload can be easily identified by looking at the historical data.

Chapter 5

Discussion

5.1 Potential for Sustainability as a Service

Digital operations as a service have become a popular method to modernize and expand digital operations while lowering the cost and risk associated with managing IT infrastructure. Digitizing processes creates workloads that may not be temporally or spatially constrained by service level agreements, allowing businesses to place workloads strategically to meet specific goals. Workload placement optimized for sustainability is enabled by integrating data that describes the entire scope of an asset's environmental impact. This includes the embodied emissions and operational emissions driven by both idle and utilized states.

By incorporating embodied emissions along with the operational emissions of assets within a data center, an actor can quantify the environmental impact of their digital presence. Following market principles, the assets in the data center are produced and installed to support the demand for the workloads. When a business makes profits and pays for the expansion of digital operations, this is equivalent to a traditional industry where capital equipment is built, installed, and run to support operations in a production or processing facility. Given the scrutiny placed on sustainable business operations, it will take more than estimating and reporting numbers to meet corporate goals transparently and effectively. This data would need to be leveraged and acted upon to reduce energy consumption associated with producing, utilizing, and recycling ICT assets at the pace the technology is advancing.

Enterprises focus more on sustainability goals as they make decisions for technology suppliers, but the scope and timeline are still unclear. This would mean that clear and tangible statements from their supplier base will be needed to realize their environmental goals. In the S&P Global Market Intelligence VotE: Digital Pulse, Environmental Impact 2022 study, 47% of respondents claim that environmental impact is important in purchasing decisions, and 29% claim that it is somewhat important [80]. This alludes to the notion that having sustainability metrics and action items would drive a competitive advantage when companies seek new product offerings and technology solutions.

5.2 ICT Asset Carbon Accounting on System

All systems today have a digital identifier called its vital product data (VPD) that reports the system's model, serial number, configuration, and various telemetry data available for monitoring. Incorporating an integer that corresponds to its embodied emissions is not challenging to do. This bridges the cyber-physical gap by acknowledging the environmental impact of bringing the hardware into existence when it would otherwise be overlooked.

A data center's carbon footprint would include the impact of all elements residing within its walls to support business operations, including hardware and support infrastructure. The data center's carbon footprint would be the downstream emissions for a business's digital presence in the same way that a power plant and fossil fuel utilization would be aligned with the electricity demand for running a facility.

The asset would also be able to report to its VPD the amount of power that piece of hardware has consumed over the lifespan of its operation. This information can be used to assess the emissions associated with the product to prolong its life or determine if it meets the criteria for end-of-life based on metrics to meet its ESG goals.

Most proposed systems have focused on external monitoring or integrating metering into the data center infrastructure. This misses granular data available to the replaceable and upgradeable units within the system by ignoring power management devices internal to the system on devices like drives, fan control units, or networking switches. This leaves room for the standardization of silicon-based technologies reporting their power directly to the VPD and, subsequently, the DCIM Management software with significant granularity. This would allow a considerable data flow to generate a real-time digital twin of the system's operation. Many ASICs allow this today, but the infrastructure to capture and report this data is the responsibility of the system architect, and the ability to aggregate this data has not become a common practice in the industry.

5.3 Monitoring Operational Emissions in the Data Center

Managing power consumption is inherent to most data centers today. The demand for bulk power in the data center is monitored to ensure that all systems within the data center can operate without interruption. During peak demand, there needs to be a large enough power budget so that any spikes in energy usage by ICT equipment or infrastructure do not result in power shortages within the facility. Because of this, most large data center operations have utilized data center infrastructure management (DCIM) software to manage the day-to-day operations of the data center, such as network quality, physical planning, and resource management. Some examples of this software include Nlyte (Nlyte Software), StruxureWare (Schneider Electric), Trellis (Vertiv), and PowerIQ (Sunbird Software), to name a few. DCIM software allows companies to automate, provide visibility, integrate with building management systems, and provide real-time monitoring across the IT infrastructure. DCIM tools are capable of much more, but the features mentioned are of interest to enable sustainability initiatives within the data centers.

Specifically, sustainability within the data center would incorporate space, cooling, power network, storage, and virtualization to optimize workload distribution in the scope of this discussion. To minimize infrastructure costs associated with physical footprints, equipment and building power consumption would be compared to processor utilization. Workload optimization has been a critical focus for many data center operators and has been slowly advancing with the application of Machine Learning to place equipment for thermal, power, communications, and workload placement [81].

Furthermore, with the advent of edge computing and Industry 4.0 initiatives such as the Internet of Things, real-time data center monitoring needs to become more granular as quantities of processing units grow [33]. Data center monitoring and management (DCIMM) aims to use a more data-centric approach to optimize data center operation in a non-invasive way. This includes using equipment with network connections in the data center, like cooling systems, Uninterruptible Power Supplies (UPS), Power Distribution Units (PDUs), and Remote Power Panels (RPPs) in conjunction with Power Management Integrated Circuits (PMICs). PMICs are typically found on elements within the server and manage the power architecture throughout the system, such as voltage regulators, ASICs, and various processors. The PMIC would send power information to the system control structure, which can report the power draw at any time.

Identifying the time series power draw of the elements in the data center associated with a workload would allow operators to determine the hardware within the data center that is being leveraged efficiently. This provides the ability to calculate the environmental impact of the workload in terms of lbCO_2e by aggregating the emissions associated with energy consumption and the manufacturing and procurement of the assets themselves as described by their life cycle analysis.

5.4 Data Center Workload Distribution

After workloads can monitor digital data for asset Carbon Footprint Reporting and power consumption, a system can be implemented for intelligent workload optimization within the DC. Most data centers are assumed to operate with a heterogeneous mix of assets. These assets can include various general compute nodes that operate on different architectures such as Intel, AMD, Power, or Z processor architecture. Various storage devices can range from SSD, Disc, and NvME from other suppliers.

There can also be various types of processors, like Graphical Processing Units, Tensor Processing Units, and Digital Processing Units, that can be leveraged for different kinds of workloads. These elements can be combined and leveraged within a data center to provide services optimized for a client's specific needs.

Currently, workloads are optimized for particular architectures. With the onset of Hyperscale and Cloud/Hybrid Cloud solutions, there is a need to mix various products to provide a holistic solution that meets the client's needs. Examples include the need for Cloud Object Storage, AI Inferencing for Fraud Detection, and secure transaction processing. These solutions may require legacy integration with an Amazon Web Service based on an x86 architecture, an inferencing engine based on a Power Systems architecture, and a transactional engine based on a Z System Architecture. The solution to this lies in integrating various cloud-based services through a platform like Red Hat OpenShift, or other Kubernetes platforms. This is where the underlying systems are foundational and are integrated into the platform architecture. This removes the need for architecture-specific workloads and allows the workload to be run on any system. This freedom supports the rapid deployment of digital operations by enabling the developer to focus on the application but may forgo performance or power consumption considerations. This can allow performance and operational emissions variability based on the hardware utilized.

With a system built on a containerized platform, the workloads can become hardware agnostic, and true "X as a Service" solutions can be optimized to meet almost any business goal. Workloads are distributed, monitored, and subsequently optimized on the fly by mixing and matching various hardware elements within the data center based on objectives and constrained by service level agreements. An example of sustainability would be minimizing the overall workload carbon footprint in CO_2e. Assuming that most large organizations have a consistent set of workloads aligned to batch operations, optimization algorithms such as genetic or multi-armed bandit algorithms can be implemented to identify the optimal set of hardware specific to that workload that minimizes both the energy consumed and the environmental impact of the assets used.

Currently, Google has a product named OVO Vertflow that can be integrated with the workflow management tool called Apache Airflow. This tool allows workflows to be kicked off and run in the greenest Google cloud region available at the time. This means that the Data Center receiving an influx of energy from a solar or wind farm will host the workload that has been kicked off. What is missing is a measure of whether or not the hardware is the most energy-efficient hardware in those data centers.

As the algorithms mature, there will be an ideal path for workloads that enter the data centers until new technology arrives, continuously optimizing the placement of the workload to meet business goals. Through DCIM monitoring, asset utilization optimizes data center operations by identifying the redundant or rarely used hardware to serve the DC's workloads. Alternatively, assets are kept in the mix and explored to determine if they are optimal solutions for new workloads. The goal is to utilize these explore and exploit algorithms to optimize hardware utilization and reduce idle state power draw.

This would allow cloud operators to provide a "Green Premium" for the utilization of the most efficient hardware in the data center. This new product offering lets businesses choose between running their workloads on less efficient systems to save costs or utilizing the most efficient systems to place their workloads. Some companies may choose the lowest cost option and disregard the emissions if there are no negative consequences to this operating model. An incentive for this type of adoption would need to come from external market dynamics. As regulation becomes more aggressive in specific geographies like the United States or the EU, there may be an incentive to adopt a model like this. Furthermore, consumer sentiment regarding sustainable companies would allow some actors to leverage this model for a competitive advantage and drive a market segment in cloud computing aligned with these optimization methods. Because energy costs are baked into the data center operating model, some operators may opt to provide discounts for this operating model or naturally move in this direction to reduce costs. There is an inherent trade-off assuming that the more efficient systems would come at a higher price point. These dynamics are still unclear, but there would need to be creative ways to leverage this opportunity.

5.5 Reporting Energy-Related Emissions

Various databases provide emissions multipliers based on the data center's location, but many power plants in developed nations report emissions by measurement. The data center would be able to inquire as to what the operational emissions associated with their energy consumption are and subsequently report the carbon footprint associated with their operations. These emissions factors are then multiplied by the power consumed, resulting in a greenhouse gas emissions number for the energy used.

5.6 Spatial and Temporal Workload Distribution

Once workloads in the data center are optimized, the distribution of workloads or batches can be placed strategically. Examples of spatial optimization would be relocating workloads that require high energy consumption to a data center where the connected grid utilizes more "green energy" and fewer fossil fuels or to one with more high-efficiency assets that can reduce the physical footprint required to support the execution. Temporal optimization of the workload means that the workload that is not constrained by time or geography can be placed into a buffer queue of a data center that can leverage more Solar or Wind energy at a specific point during the day and then executed when the energy source is significantly more "green."

Various large workloads are periodically run to support enterprise operations [82]. Examples include end-of-day transaction processing, fraud surveillance, data analytics in research applications, training machine learning models, capacity requirements planning, inventory processing, invoicing, and email automation [83]. These workloads generally require significant computing resources leading to a bump in power consumption. The emissions from these operations can be minimized by shifting the execution times or locations to systems and data centers that reduce greenhouse gas emissions during utilization. Some workloads, like data fetching from field operations of large oil rigs, could be done at night on a weekend when there may not be as much demand for commercial computing resources. Another example would be doing your resource capacity planning in the afternoon when there are more contributions to the grid from renewable resources.

Using this information, a data center user can distribute workloads not restricted by regulation or service level agreements to meet local goals. This means workloads will be distributed to regions where the locality, or even the business itself, invested in new technologies to reduce carbon footprints, such as renewable energy or carbon capture. This capability allows companies to directly support sustainable initiatives by investing their digital carbon footprint where it has the most impact. This will create competition focused on sustainable operations between data centers and allow localities to spur investment without the "Not In My Backyard" concerns by driving support for new regional investments.

Pricing for this type of service would need to be explored through market demand. One option for pricing in this space would be tying the fees for this "as a Service" offering to a fraction of the price for a region's carbon credits, power purchase agreement, or energy bill. If the dollars per watt of computing were measured and leveraged in this type of billing scheme, businesses would be incentivized to address the carbon footprint of their digital operations through this method. This would benefit the company by allowing them to meet its sustainability goals transparently while leveraging the competitive advantage gained by marketing these green operations to its customers. This would also spur investment in sustainable energy projects either in the data center or the local region, where the contributions of these fees are used to support sustainable projects directly.

5.6.1 Implementing Sustainability Goals Using Today's Tools

Utilizing software to control workload distribution requires a central computing center. The sensors feed data into this compute node, which needs to be standardized, stored, and processed. This involves database management, an interface to implement optimization, and a connection to the network that manages the flow into and out of the data center through hypervisors and into virtual machines. The software-defined distribution for hypervisor clusters is currently in practice with many commercial offerings and includes redundancy across servers for reliability. It periodically checks the state of the machine to ensure up-time via heartbeat messages. It would use this to establish the embodied and operational emissions associated with the devices required to support the architecture. This would then need to consider the other servers in the cluster to determine the embodied emissions associated with that virtual machine and the workloads it supports. This method can then leverage the optimization algorithms for consistent workload distribution across the cluster and enable virtual machines on different servers throughout the cluster to support the operations within it.

These tools are commonplace in current cloud operations but have not been tailored to optimize for sustainability goals. Most of these tools generally focus on balancing processor utilization and memory capacity to meet the needs of the business. Modifying the objective of these tools would allow for the ability to shift workloads to idle or more efficient machines and move the workloads between data centers to meet their needs.

5.6.2 Hardware Defined Sustainable Workload Distribution

The concept of software-defined distribution above can be done by implementing various logic gates on the physical hardware, where a cluster spanning several physical servers is connected via a network switch. This network switch utilizes pseudo-boolean logic and the reporting provided by the VPD of the systems to ping, read and write memory to a cache and distribute workloads to specific machines based on a given goal. This then is placed into a more extensive network, providing a hierarchical ability to distribute the workload across clusters and manage the data center. This method has not been implemented or experimented with and would offer a new generation of sustainable smart switches that would reduce latency, complexity, and power draw required to distribute workloads over a large set of hardware.

5.6.3 Hyperledger Reporting

Both methods of workload distribution would allow the hardware to report serial number, equipment type, embodied emissions, category of embodied emissions (Chips, PCB, Metal, Plastic), power consumption to date, years in service, data center, client category, SLA category, Location, and the DPP Unique product identifier. This information can subsequently be reported to a hyperledger blockchain supported by idle hardware in the facility and log transactions in the form of workloads. This allows a generally large and immutable database to run efficiently and provide access to performance, workload, consumption, and emissions data. The hyperledger fabric offers access control to elements on the blockchain, allowing regulators, data center owners, and workload owners access to the information pertinent to them. After a period of normal operations, data centers can generate a "standard" operating state for the hardware that can be used by machine learning to track the environmental impact of the server and either reroute workloads or promote decommissioning of the hardware based on its green payback period or limit.

5.6.4 "Green" Hardware

The notion of Green Hardware is to be defined by the operator. Still, it would include leveraging the information stored and transmitted in a manner considered to have justified its existence. One example is to charge a small fee for sustainable workload distribution, then have that information stored on the VPD as a subset of the total emissions. This assumes the infrastructure is in place that minimizes the carbon footprint of the client's system. The hardware then has simple logic implemented to ensure that the subset is a fixed percentage of the total emissions of the hardware until the value of the accumulated fees equals the total monetary value needed to pay a carbon offset tax or power purchase agreement.

This would then influence consumer behavior throughout the value chain of the environmental impact of a business's digital presence. Green hardware would drive a competitive advantage for companies that wish to maintain an ecological marketing presence, allow enterprises to leverage sustainable certifications, and meet regulations that can open access to countries focused on environmental management. Furthermore, beyond the operations of the data center, this would drive the adoption of sustainable technologies at the local level to support investment and job creation for new and current facilities.

5.6.5 Diurnal Task Shifting

Shifting workload execution to align with the local power grid's carbon intensity can reduce greenhouse gas emissions for the given workloads. Some work done in this space is to increase or decrease CPU utilization inversely with the carbon intensity of the grid [69]. There have been efforts to characterize workload efficiency in cloud data centers by introducing a new metric focused on a set of services, called payloads, that encompasses the IT and non-IT power consumption within the data center. This is called the Cloud Power Usage Efficiency, which aims to establish the ability to characterize the carbon intensity of workloads and provides the logic to distribute the workloads based on the carbon intensity factor.

This workload placement method could reduce the carbon footprint of workloads that are well known and would be a method to alleviate some of the constraints of specific workloads that are restricted to certain geographies or facilities [70]. Several of the work being done has addressed the notion that the carbon intensity varies throughout the day, and the biggest hurdles to implementation are predicting which workloads can be shifted, indicating where to place workloads at certain times, prioritizing workloads, and modeling the power consumption for all of the elements utilized for a given workload.

5.6.6 Multi-Armed Bandit Implementation

The multi-armed bandit algorithm using Thompson Sampling worked well for the theoretical scenario, typically finding the optimal solution in less than one hundred iterations. If batch operations can be characterized or previous-day predictions can be used, like in Google's Carbon Intelligent Computing System, this would be an ideal method to distribute workloads across clusters of like hardware to identify spatially optimized workload distribution within the data center regionally between data centers. The algorithm does not compute intensively and can create interpretable learning data that computers and human operators can leverage.

Chapter 6

Conclusion

Although the ability to distribute workloads spatially and temporally is not a readily available solution, the required elements for such a system are generally available and can integrate with the vital product data already reported by the ICT hardware. They are necessary as the growth of cloud computing and hyperscale data centers increases the ecological impact of manufacturing and operations. A method is needed that provides for a standardized and comparable calculation of embodied emissions, the ability to monitor and report power consumption by workload, the ability to report operational emissions dynamically, and the ability to distribute workloads to minimize combined embodied and operational emissions constrained by service level agreements and regulation.

The ICT sector needs to collaborate to identify a harmonized method to calculate embodied emissions for data center assets. There have been some proposals and guidelines, but a standard body has rigidly formalized nothing. Although attempts have been made with tools such as the PAIA tool, iNemi Eco-Impact Estimator, and general lifecycle analysis tools, there will not likely be mass adoption of a rigorous method. Efforts are required to combine elements of the PAIA tool, such as uncertainty calculations, with the granularity of the iNemi tool to allow for comparative and realistic measurement of embodied servers that are unique to the hardware. This is necessary for a generation-to-generation evolution of hardware for internal metrics. It also provides a competitive incentive for hardware producers to reduce their environmental impact and for customers to make educated decisions. Furthermore, data centers should begin to focus on their data center's entire "carbon load" to account for the volume of embodied emissions associated with its operations. Scaling out operations should be reflected in carbon footprint reporting beyond just the energy consumed during operation.

ICT suppliers can implement intelligent monitoring at little to no system power or performance cost. This would lead to intelligent operations in the data center, allow their customers to identify cost reduction opportunities during operation, and provide their clients the ability to meet their environmental goals through their products. This would further simplify the infrastructure needed to monitor and report power consumption and offer granular data to drive digital twin modeling and operational optimization. This can be done at the device level to drive insight into the operation of the data center infrastructure in real time with system-level operations inside the ICT assets of the data center.

The analysis has shown that operational emissions can be optimized with the above data, given a set of hardware and data center locations. With the data, large data center operators can report their emissions in a manner that supports progress toward environmental goals, cost reduction, and risk mitigation, possibly opening the door for new and novel business opportunities that the world is looking for. The ability to gather and act on data in real time allows for temporal and workload distribution in many ways, such as through K-means, previous-day predictions, linear integer programs, or bandit algorithms. These methods have shown the potential to reduce the environmental impacts associated with the growth of the digital age. They could potentially be applied beyond the data center as manufacturing, mining, and other industries become more digitized with the advent of edge computing.

As all of these digital operations shift away from on-premises and self-managed solutions, businesses must be aware of their digital carbon footprint that scales with their digital solutions. The market, regulators, and other stakeholders are becoming aware of digital operations' increasing hardware and energy consumption. Businesses must prepare to deal with the risks to their operations and reputations. This research has shown a crude way of addressing these issues. Future work will continue to try to implement the methods discussed in a scaleable manner that supports dynamic and ever-changing operations. Hopefully, it will influence how ICT hardware is designed and deployed to benefit the environment and innovation.

Appendix A

Tables

Appendix B

Figures

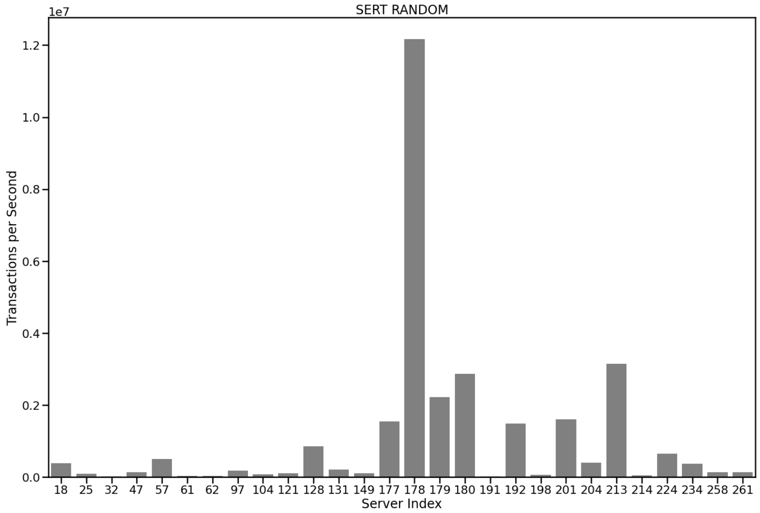

Figure B-1: Server capability in transactions per second for the SERT Random workload.

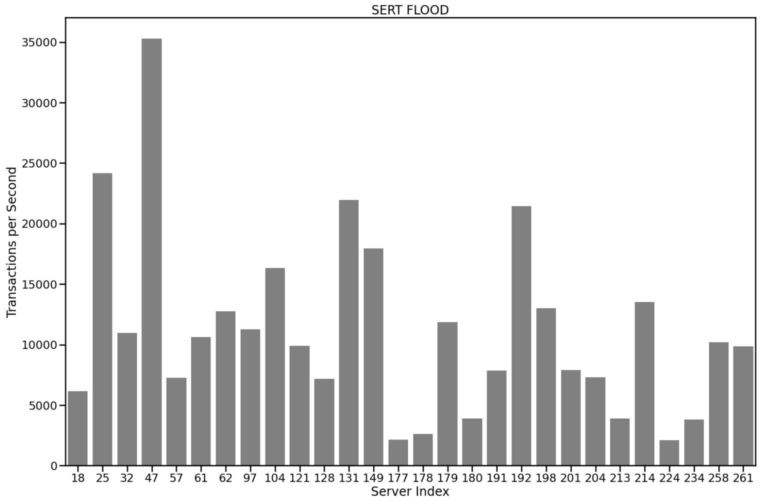

Figure B-2: Server capability in transactions per second for the SERT Flood workload.

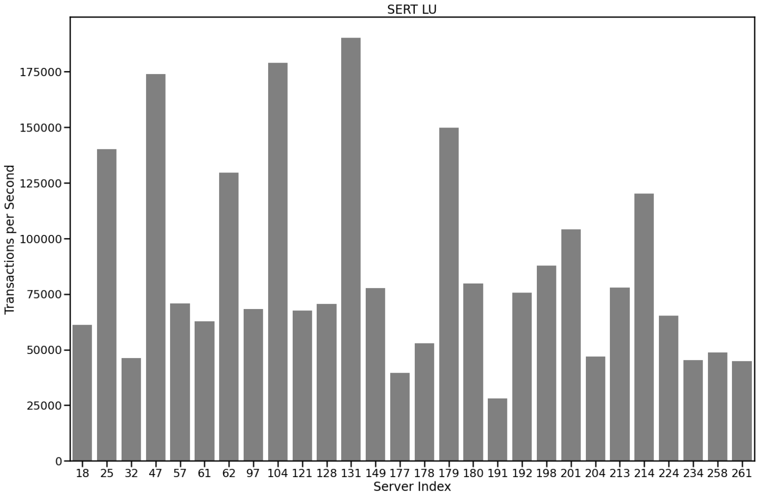

Figure B-3: Server capability in transactions per second for the SERT LU workload.

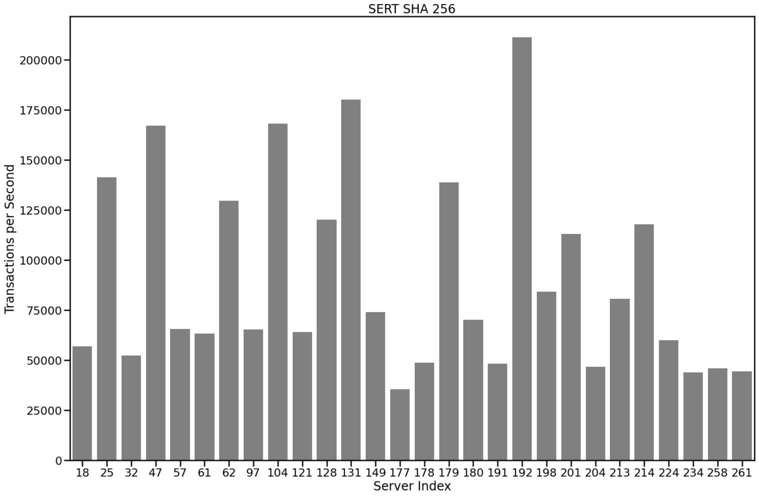

Figure B-4: Server capability in transactions per second for the SERT SHA256 workload.

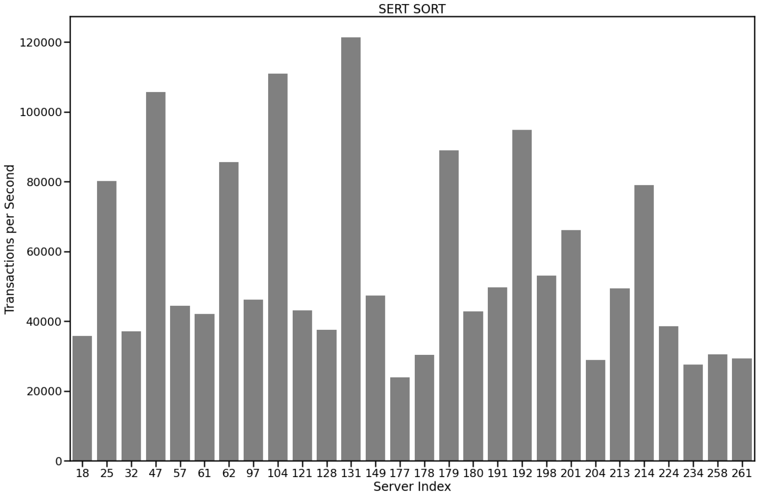

Figure B-5: Server capability in transactions per second for the SERT SORT workload.

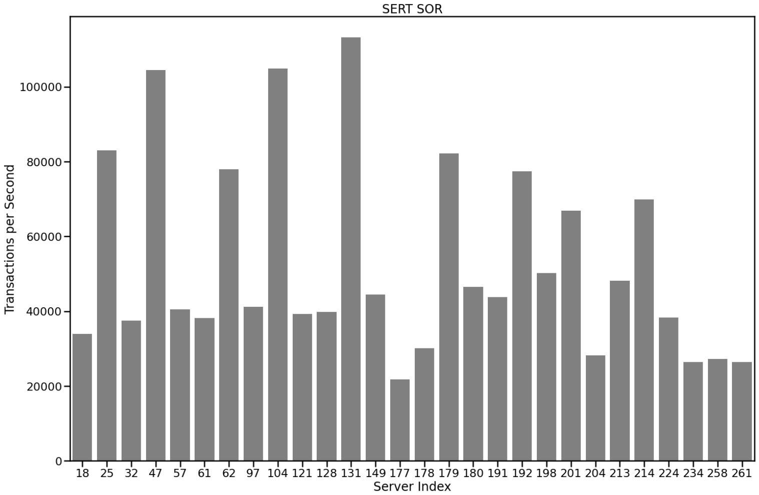

Figure B-6: Server capability in transactions per second for the SERT SOR workload.

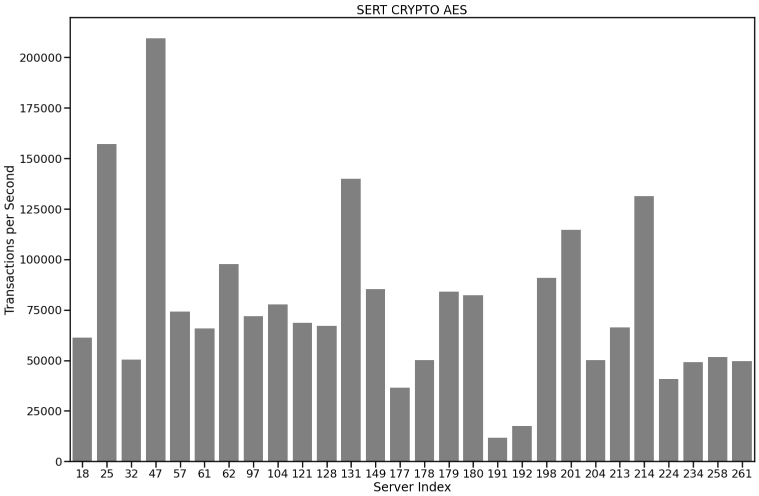

Figure B-7: Server capability in transactions per second for the SERT CryptoAES workload.

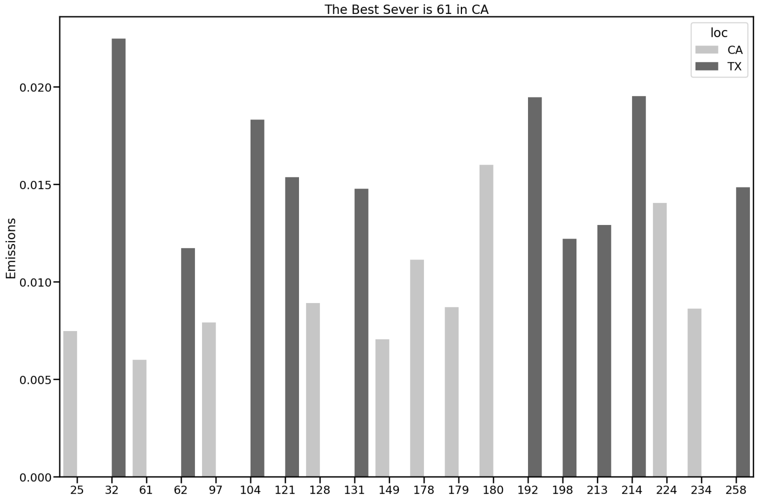

Figure B-8: Results from the bandit algorithm for scenario 2.

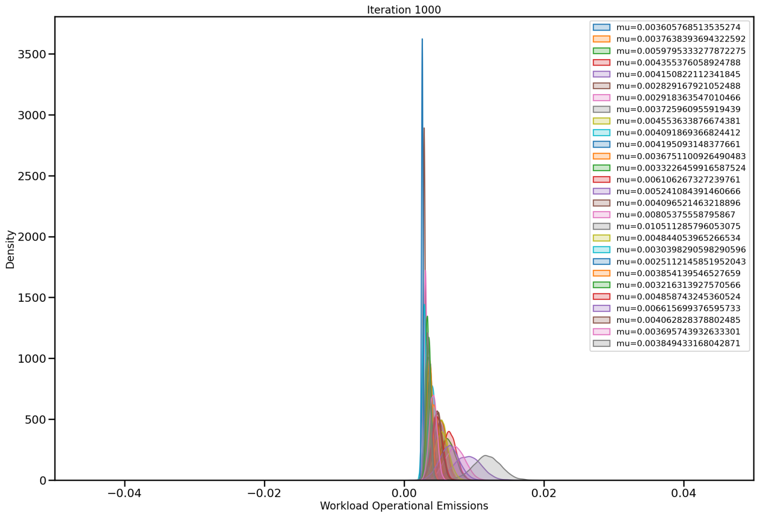

Figure B-9: Results from the bandit algorithm for scenario 2.

Bibliography

[1] K. Pucker, "Overselling Sustainability Reporting," 2021.

[2] B. Stengel, D. Anesini, R. Brothers, A. Cravens, Z. Chertok, S. Krishnan, V. Kroa, W. Lee, C. L. Marshall, S. G. Middleton, R. Paquin, C. Price, P. Reymann, W. Schuster, K. Shikita, G. Trinidad, J. K. Speer, D. Versace, and J. Westcott, "IDC FutureScape: Worldwide Sustainability/ESG 2023Predictions," tech. rep., 2022.

[3] CVN Index, "The Zettabyte Era: Trends and Analysis," 2016.

[4] "Data Centers and Servers | Department of Energy."

[5] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, "Comparing datasets of volume servers to illuminate theirenergy use in data centers," 2020.

[6] M. A. B. Siddik, A. Shehabi, and L. Marston, "The environmental footprint of data centers in the United States," *Environmental Research Letters*, vol. 16, 6 2021.

[7] CISCO, "Cisco Global Cloud Index: Forecast and Methodology, 2016–2021," tech. rep., 2016.

[8] I. Incorporated, E. Incorporated, R. E. Brown, R. Brown, E. Masanet, B. Nordman, B. Tschudi, A. Shehabi, J. Stanley, J. Koomey, D. Sartor, P. Chan, J. Loper, S. Capana, B. Hedman, R. Duff, E. Haines, D. Sass, and A. Fanara, "Report to Congress on Server and Data Center Energy Efficiency: Public Law 109-431," 8 2007.

[9] A. Shehabi, S. J. Smith, E. Masanet, and J. Koomey, "Data center growth in the United States: decoupling the demand for services from electricity use," *Environ. Res. Lett*, vol. 13, p. 124030, 2018.

[10] M. Cosar, "CARBON FOOTPRINT IN DATA CENTER: A CASE STUDY," *Fresenius Environmental Bulletin*, vol. 28, pp. 600–607, 2019.

[11] A. Micarelli and S. Conti, "Reducing the Carbon Footprint
}of Computing," 2020.

[12] X. Chen, M. Despeisse, and B. Johansson, "Environmental Sustainability of Digitalization in Manufacturing: A Review," *Sustainability 2020, Vol. 12, Page 10298*, vol. 12, p. 10298, 12 2020.

[13] T. R. Miller, J. Gregory, H. Duan, R. Kirchain, and J. Linnell, "Characterizing Transboundary Flows of Used Electronics: Summary Report," tech. rep., 2011.

[14] S. M. Abdelbasir, C. T. El-Sheltawy, and D. M. Abdo, "Green Processes for Electronic Waste Recycling: A Review," *Journal of Sustainable Metallurgy 2018 4:2*, vol. 4, pp. 295–311, 4 2018.

[15] United Nations, "TRANSFORMING OUR WORLD: THE 2030 AGENDA FOR SUSTAINABLE DEVELOPMENT UNITED NATIONS UNITED NATIONS TRANSFORMING OUR WORLD: THE 2030 AGENDA FOR SUSTAINABLE DEVELOPMENT," tech. rep., 2015.

[16] T. N. Theis and H. S. Philip Wong, "The End of Moore's Law: A New Beginning for Information Technology," *Computing in Science and Engineering*, vol. 19, pp. 41–50, 3 2017.

[17] N. Lei, "A robust modeling framework for energy analysis of data centers,"

[18] N. Lei, E. Masanet, and J. Koomey, "Best practices for analyzing the direct energy use of blockchain technology systems: Review and policy recommendations," *Energy Policy*, vol. 156, p. 112422, 9 2021.

[19] J. G. Koomey, "ESTIMATING TOTAL POWER CONSUMPTION BY SERVERS IN THE U.S. AND THE WORLD," 2007.

[20] P. Bhandia, R. S. Anupindi, P. Yekbote, N. Singh, H. Phalachandra, and D. Sitaram, "DCSim: Cooling Energy Aware VM Allocation Framework,"

[21] R. Ghafari, F. H. Kabutarkhani, and N. Mansouri, "Task scheduling algorithms for energy optimization in cloud environment: a comprehensive review," *Cluster Computing 2022*, pp. 1–59, 1 2022.

[22] M. Inam and M. Z. Nayyer, "Energy-Aware Load Balancing in a Cloudlet Federation," *Engineering Proceedings 2021, Vol. 12, Page 27*, vol. 12, p. 27, 12 2021.

[23] J. L. Berral, Goiri, R. Nou, F. Julià, J. Guitart, R. Gavaldà, and J. Torres, "Towards energy-aware scheduling in data centers using machine learning," *Proceedings of the e-Energy 2010 - 1st Int'l Conf. on Energy-Efficient Computing and Networking*, pp. 215–224, 2010.

[24] M. D. Monzón, Z. Ortega, A. Martínez, and F. Ortega, "Standardization in additive manufacturing: activities carried out by international organizations and projects,"

[25] "Greenhouse gases | World Meteorological Organization."

[26] "1 - Embodied Carbon 101 - Carbon Leadership Forum," 2020.

[27] G. Bassi, "How Are Manufacturers Approaching Supply Chain Sustainability?," 2022.

[28] X. Lin and A. Dissertation, "MODERNIZING LIFE CYCLE ASSESSMENT VIA INFORMATIC TECHNIQUES," tech. rep., 2021.

[29] "Regulatory Compliance | ECO Declarations | Lenovo US."

[30] "HPE Product Carbon Footprint frequently asked questions."

[31] "Product Carbon Footprints | Dell USA."

[32] "About » Open Compute Project."

[33] M. Levy and J. O. Hallstrom, "A new approach to data center infrastructure monitoring and management (DCIMM)," in *2017 IEEE 7th Annual Computing and Communication Workshop and Conference, CCWC 2017*, Institute of Electrical and Electronics Engineers Inc., 3 2017.

[34] H. Han, "Consumer behavior and environmental sustainability in tourism and hospitality: a review of theories, concepts, and latest research," *Journal of Sustainable Tourism*, vol. 29, no. 7, pp. 1021–1042, 2021.

[35] S. B. Banerjee, E. S. Iyer, and R. K. Kashyap, "Corporate environmentalism: Antecedents and influence of industry type," *Journal of Marketing*, vol. 67, pp. 106–122, 4 2003.

[36] Wachtell, R. . Lipton, and Katz, "The Coming Impact of ESG on M&A,"

[37] T. H. L. S. F. o. C. Governance, "ESG and M&A in 2022: From Risk Mitigation to Value Creation," *https://corpgov.law.harvard.edu/*, 1 2022.

[38] S. Boubaker, I. Derouiche, H. Farag, M. Elnahass, R. Redondo Alamillos, and F. de Mariz, "How Can European Regulation on ESG Impact Business Globally?," *Journal of Risk and Financial Management 2022, Vol. 15, Page 291*, vol. 15, p. 291, 6 2022.

[39] J. Patorska, A. Laszek, J. Leoniewska-Gogola, D. Maciborski, and A. Fusiara, "Impact of international, open standards on circularity in Europe," tech. rep., Deloitte, 2022.

[40] T. Okrasinski, F. Zhao, L. Dender, E. Helminen, D. Kline, X. Lin, P. Murphy, A. Peterson, and M. Schaffer, "Modernizing a Life Cycle Eco-Impact Estimator for ICT Products," tech. rep., 2020.

[41] E. Pauer, B. Wohner, and M. Tacker, "The Influence of Database Selection on Environmental Impact Results. Life Cycle Assessment of Packaging Using GaBi, Ecoinvent 3.6, and the Environmental Footprint Database," 2020.

[42] I. T. Herrmann and A. Moltesen, “Does it matter which Life Cycle Assessment (LCA) tool you choose? – a comparative assessment of SimaPro and GaBi,” *Journal of Cleaner Production*, vol. 86, pp. 163–169, 1 2015.

[43] J. J. Conway and M. Herson, *Evaluation of Environmental Foot Printing Techniques*. PhD thesis, Massachusetts Institute of Technology, 2012.

[44] S. Patanavanich, “Exploring the Viability of Probabilistic Underspecification as a Viable Streamlining Method for LCA,” tech. rep., 2011.

[45] E. Olivetti and R. Kirchain, “A Product Attribute to Impact Algorithm to Streamline IT Carbon Footprinting,” in *Design for Innovative Value Towards a Sustainable Society*, pp. 747–749, Springer Netherlands, 2012.

[46] D. Donnellan, A. Lawrence, and J. Dietrich, “Critical regulation: the EU Energy Efficiency Directive recast,” tech. rep., 2022.

[47] “Proposed Architecture and Principles for Digital Product Passports,” tech. rep., 2022.

[48] W. Callahan, S. A. James Fava, S. Wickwire, J. Sottong, J. Stanway, and M. Ballentine, “Product Life Cycle Accounting and Reporting Standard GHG Protocol Team,” tech. rep.

[49] K. Ummel, “CARMA Revisited: An Updated Database of Carbon Dioxide Emissions from Power Plants Worldwide Kevin Ummel CARMA Revisited: An Updated Database of Carbon Dioxide Emissions from Power Plants Worldwide,” tech. rep., 2012.

[50] J. R. Nicholson, “New Digital Economy Estimates,” tech. rep., Bureau of Economic Analysis, 3 2022.

[51] F. Moore, “Data Center Energy Consumption Enormous Data Centers Creating a Hyperscale Heat Wave,” tech. rep., 2020.

[52] E. Masanet, A. Shehabi, J. Liang, L. Ramakrishnan, X. Ma, V. Hendrix, B. Walker, and P. Mantha, “The Energy Efficiency Potential of Cloud--Based Software: A U.S. Case Study,” tech. rep., 2013.

[53] Microsoft, “The carbon benefits of cloud computing,” tech. rep., 2017.

[54] D. Mytton, “Assessing the suitability of the Greenhouse Gas Protocol for calculation of emissions from public cloud computing workloads,” *Journal of Cloud Computing*, vol. 9, 12 2020.

[55] “24/7 Clean Energy – Data Centers – Google.”

[56] “Data Centers - Meta Sustainability.”

[57] D. Wheeler and K. Ummel, "Calculating Carma: Global Estimation of Co2 Emissions from the Power Sector," *SSRN Electronic Journal*, 5 2008.

[58] "Where does our energy come from?."

[59] R. P. Borthwick, J. Whetstone, J. C. Yang, and A. Possolo, "Examination of United States Carbon Dioxide Emission Databases," tech. rep., NIST, 2011.

[60] "THE EMISSIONS & GENERATION RESOURCE INTEGRATED DATABASE eGRID Technical Guide with Year 2020 Data Clean Air Markets Division,"

[61] M. Rolinck, S. Gellrich, C. Bode, M. Mennenga, F. Cerdas, J. Friedrichs, and C. Herrmann, "A Concept for Blockchain-Based LCA and its Application in the Context of Aircraft MRO," *Procedia CIRP*, vol. 98, pp. 394–399, 1 2021.

[62] "How Walmart brought unprecedented transparency to the food supply chain with Hyperledger Fabric," 2019.

[63] "Shifting sands: The changing consumer landscape | Deloitte UK."

[64] K. White, R. Habib, and D. J. Hardisty, "How to SHIFT consumer behaviors to be more sustainable: A literature review and guiding framework," *Journal of Marketing*, vol. 83, pp. 22–49, 5 2019.

[65] "The Elusive Green Consumer."

[66] "Sustainability & Consumer Behaviour 2022 | Deloitte UK."

[67] R. Mehrotra, A. Dubey, S. Abdelwahed, and A. N. Tantawi, "A Power-Aware Modeling and Autonomic Management Frame-work for Distributed Computing Systems," in *Handbook of energy-aware and green computing*, vol. 2, pp. 621–648, 2012.

[68] W. Lin, Y. Zhang, W. Wu, S. Fong, L. He, and J. Chang, "An adaptive workload-aware power consumption measuring method for servers in cloud data centers," *Computing*, pp. 1–24, 5 2020.

[69] T. Bahreini, A. Tantawi, and A. Youssef, "An Approximation Algorithm for Minimizing the Cloud Carbon Footprint through Workload Scheduling," *IEEE International Conference on Cloud Computing, CLOUD*, vol. 2022-July, pp. 522–531, 2022.

[70] 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," *IEEE Transactions on Power Systems*, 6 2021.

[71] T. Lattimore and C. Szepesvári, "Bandit Algorithms," tech. rep.

[72] "Calculating My Carbon Footprint - Microsoft Sustainability."

[73] "ICT Sector Guidance built on the GHG Protocol Product Life Cycle Accounting and Reporting Standard Chapter 1: Introduction and General Principles," tech. rep., 2017.

[74] K. Bergman, J. Shalf, G. Michelogiannakis, S. Rumley, L. Dennison, and M. Ghobadi, "PINE: An Energy Efficient Flexibly Interconnected Photonic Data Center Architecture for Extreme Scalability," 2018.

[75] "CPU vs. GPU: The Paradigm Shift - Systel," 2021.

[76] "The SERT ® 2 Metric and the Impact of Server Configuration The SERT 2 Metric and the Impact of Server Configuration," tech. rep., 2021.

[77] "Standard Performance Evaluation Corporation (SPEC) Power and Performance Benchmark Methodology V2.2," tech. rep., 2014.

[78] "Standard Performance Evaluation Corporation (SPEC®) SERT® Suite User Guide 2.0.6," tech. rep., 2022.

[79] "Cloud TPU | Google Cloud."

[80] L. Eagle, "Enterprises Increasingly Select Technology Suppliers Based on Sustainability Needs Analysts-Liam Eagle Enterprises Increasingly Select Technology Suppliers Based on Sustainability Needs," tech. rep., S&P Global Market Intelligence, 2022.

[81] "DCIM-For-Dummies_3rd-Edition,"

[82] "What is Workload? - Definition of Workload in Cloud Computing."

[83] "What is Batch Processing? - Enterprise Cloud Computing Beginner's Guide - AWS."