The Lack of Green Data Centers

A Research Paper submitted to the department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Samuel McBroom

April 29, 2021

On my honor as a University student, I have neither given nor received aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

signed: ______________________________ date: April 29, 2021
Samuel McBroom

signed: ______________________________ date: ___________
S. Travis Elliot, Department of Engineering and Society

Introduction

The primary goal of this thesis is to contextualize the environmentally harmful effects of massive data center providers and suggest reasons for the absence of greener technologies by focusing on the differences between Lancium's Clean Compute Centers and modern data center designs. Additionally, I will demonstrate that green data center designs will become much more prevalent in the future. To accomplish this, I will use the Social Construction of Technology (SCOT) framework to identify relevant stakeholders, analyze how the current landscape of data center designs serve those groups, and consider how adoption of greener solutions might affect each stakeholder. I will use Lancium, a small, low cost, carbon negative data center provider aimed at the High Throughput Computing market, such as for Machine Learning training or Scientific Modeling (Lancium Compute Documentation, 2021) as a case study from which to draw potential solutions. With its significantly reduced operating costs and positive impact on the environment, Lancium appears to be an ideal data center model for both shareholders, consumers, and the planet. First, I will describe the key points of SCOT. Second, I will present the background of modern data center designs and Lancium's Clean Compute Centers. Third, I will apply SCOT to analyze how data center stakeholders shaped current designs and what the impact of widespread adoption of Lancium's data center model would be on those stakeholders. Finally, I will present some possible closure paths.

Framework Background

SCOT presents the idea that theories do not succeed because they are true, but because they are socially supported. To identify the factors that resulted in the absence of green technologies in modern data center designs, I will employ the core concepts outlined by SCOT. These include the principle of symmetry, interpretive flexibility, and closure (Pinch, T. J., Bijker, W. E., 1984). Pinch and Bijker state that symmetrical comparison of competing technologies, such as traditional and emerging green data center designs, is important because often they have only minor differences in their solution and thus societal factors are the primary determinant for which technology will be successful. Following the principle of symmetry fits well with the current study on the impacts of green data center designs as it might incorrectly be assumed that data centers must necessarily be power intensive because that is how all modern designs function.

SCOT presents the concept of interpretive flexibility, which recognizes that each technology has different interpretations to different social groups. Under the umbrella term of interpretive flexibility are subconcepts to further break down the analysis. Of these, I will primarily focus on the concepts of relevant social groups and design flexibility. Relevant social groups focuses on identifying the different stakeholders most affected by a technology, which provides a logical structure for the comparison of the different data center design approaches. The concept of design flexibility, which considers that multiple designs may produce approximately equivalent solutions, applies well to this study comparing an established, traditional data center design to a relatively novel one. Together, relevant social groups and design flexibility will be used to understand the similarities and drawbacks of traditional data center designs which maximize computer power as opposed to greener solutions that maximize energy efficiency, and their impact on society.

SCOT describes two possible closure paths a technological artifact can face, those being that of rhetorical closure and redefinition of the problem. Rhetorical closure occurs when the best of all possible solutions is reached and recognized by the existing society. Redefinition of the problem appears when the contentious technology solves a problem that overrides other issues it might cause. I will discuss possible developments in the future that could lead to either closure scenario. To accomplish this, it will be necessary to include an additional yet untraditional relevant social group: all living individuals and societies. Today, we know that widely used technological artifacts, such as data center designs, influence Earth's environmental health. This, in turn, impacts individuals and societies across the globe, making them stakeholders in those technological artifacts. SCOT emphasizes that closure is not permanent, and that new social groups and technologies might reopen a previously closed debate. Concluding this study, I will quickly propose a couple future developments that might prompt a reanalysis of data center designs.

Traditional Data Center & Lancium Background

Traditionally, massive data center providers such as Google, Amazon, and Microsoft, have built data centers primarily focused on maximizing the amount of computing power available. As echoed in numerous reports, "Data center carbon emissions are a growing global concern" (Bouley, D., 2010). Global temperatures having risen around 1.2ºC since 1850 (EN-ROADS, 2021), with disastrous consequences. Ocean acidity has fallen by nearly 0.1pH (NOAA, 2021), 100 year storms occur more frequently (UCAR Center for Science Education, 2021), sea levels are rising at an increasing rate (NOAA, 2021), and numerous other harmful events have been caused by climate change at present. With temperatures expected to rise nearly 2.0ºC by 2040, unless major energy changes are implemented across the world even more disastrous effects will be seen. The Department of Energy has identified data centers as making up 2 percent of all electricity use in the United States (U.S. Department of Energy, 2021). Google, Amazon, and Microsoft, in an effort to reduced their energy consumption have made various sustainability commitments and long-term goals. Microsoft has committed to going carbon negative by 2030 (Microsoft, 2021). Google has stated that their "ultimate goal is to source carbon-free energy for [their] operations in all places, at all times" (Google Data Centers, 2021). Amazon has pledged to run their business in the most environmentally friendly way possible and achieving 100 percent renewable energy usage (Amazon, 2021), though they have not provided a time frame. While these commitments show the companies are moving in the right direction, their plans are projected to take at least another decade to complete and their data centers still have other harmful effects on the environment.

Lancium, a provider of Clean Compute Centers, demonstrates a number of successful solutions to the rising environmental impact of data centers. Lancium is a carbon-negative data center provider, which it accomplishes by "co-locat[ing] directly at renewable facilities and draw load when power prices are low (or when power would otherwise be wasted) and drop load when power is unavailable or expensive" (About Lancium, 2021). Dropping load by moving, suspending, and resuming compute jobs when needed leads Lancium to be "up to 90% cheaper vs. the leading cloud providers" (About Lancium, 2021). Lancium is also able to avoid the use of air conditioning or refrigerants, which significantly reduces their environmental impact and operating costs. They use older machines with previously expired warranties, and thus extending the longevity of their computers is not a financial concern. Even so, Lancium has found that there are small to negligible negative impacts on their machines while operating them at temperatures above the manufacturer recommendations. Older processors are also less energy dense, and so generate less heat per volume. All of these factors allow Lancium to simply circulate airflow through their data centers using fans to attain an optimal temperature. Purchasing older chips not only has a beneficial impact by cutting refrigerant use and operating costs, but also reduces the amount technological waste sent to landfills. Overall, Lancium is able to reconcile the increasing need for data centers worldwide while also achieving an incredibly small environmental impact at a fraction of the operating costs relative to traditional data center designs.

Analysis

Relevant Social Groups

To properly evaluate and compare two competing technologies, SCOT first introduces the concept of identifying relevant social groups. The most significant social group is the consumers of the data center services. This group can be further divided into two distinct subgroups: general users and developers. General users might include those who do not directly interface with the cloud hosted computers, but depend on the services they provide, such as Google Drive, iCloud, Ebay, and most other websites. Developers, members of the scientific community, and other technically minded individuals, in contrast, work directly with the remote machines. They might be the developers behind user-facing services like Google Drive, individual developers using machines for hobby projects, or scientists who need computing clusters to evaluate mathematical models. The third social group of focus will be the data center providers themselves. This group includes the CEO and other company executives, shareholders, employees, and other individuals directly connected with the company. While social groups, such as governmental bodies, exist, the three identified above are the most relevant when comparing traditional and efficiency-focused data center designs, as they will be the most impacted by changes to the status quo.

Design Flexibility - Computing Power

From the social groups identified above, it is evident why modern data center designs have evolved into the forms they take today. For general users, possibly the most important feature that data centers can provide is fast response times. A study in 2007 found that most users are willing to wait only about two seconds for a webpage to load or information to be retrieved before clicking away or becoming frustrated (Fui-Hoon Nah, Fiona. 2007). 14 years later, users have come to expect even more from technology and might even consider two seconds too long of a load time. Thus, massive data center providers have focused on architectures that provide fast, "always on", machines. Data center providers also need to distinguish themselves from other competing cloud services and meet the increasing demand of developers. To accomplish this, data center providers must continually increase their computing assets. To avoid the high costs of continually building and expanding data centers, many providers purchase newer hardware and upgrade the machinery in existing locations. This results in data center designs built to maximize available computing power. In addition to providing worldwide access to computing services, data center providers are businesses and consequently have the goal of increasing returns for shareholders and salaries for executives or employees. Data center providers purchase new chips from manufacturers to secure warranty protection and avoid potential maintenance costs that might arise if buying older CPUs. The main concern for data center providers over the past decades has been meeting the computing demands of the world, and by analyzing the concerns of relevant social groups it is clear why modern data centers designs have taken the shapes we see today. However, as the principle of design flexibility states: these designs are just one solution to meeting humanity's computing needs, catering to the long-established belief that increasing computing power is the best path for data center advancement.

Environmental Concerns & Global Communities

"Climate change is the defining issue of our time" (United Nations Secretary-General, 2018). In recent years, our impact on the climate and the negative ramifications we have seen worldwide has become a primary concern among citizens of all nations. The SCOT framework arose from Pinch and Bijker's paper *The Social Construction of Facts and Artifacts*, published in 1984. At this time, the global impact of technological artifacts was acknowledged but only a select few technologies might ever extend their reach far enough to affect people worldwide. However, we now know that massively scaled businesses and technologies can cause lasting damage to the Earth, and by extension, all who inhabit it. Everyone has a right to the world they want to live in, and are thus stakeholders in any technology which could negatively impact that world. While SCOT's concept of relevant social groups traditionally focuses on the producers and consumers of a technology, I argue we live in a world where all people should be included as stakeholders. New technologies and designs emerging today commonly come with environmental ramifications, and thus either directly or indirectly affect all individuals and communities.

With this newly identified relevant social group, it can be seen that while traditional data center designs solve the demand for globally available computing resources, they make no moves towards solving the emerging global climate change issue. Applying the concept of design flexibility, I will analyze the design of Lancium's Clean Compute Centers and how the design choices made strike a balance between meeting the increasing demand for computational assets and addressing rising climate concerns.

Design Flexibility - Lancium & Green Computing

In his book *Grow a Greener Data Center* (Alger, D., 2009), Douglas Alger defines a green data center as "a computing environment that uses resources in a more efficient manner and has less impact upon people and the environment." To analyze how Lancium operates as a green data center provider, I will focus Lancium's Clean Computer Center design which aims to mitigate three environmentally harmful impacts of traditional data centers: carbon emission cost of power, refrigerant use, and technological waste produced. For each area of impact, the economic effects will be considered alongside the environmental consequences.

The most obvious factor Lancium addresses is lowering the carbon impact of their energy requirements by sourcing electricity from the excess power produced by renewable sources. The benefits of this approach are twofold. First, Lancium operates as a carbon-negative data center because it uses otherwise wasted energy that would go unconsumed. Second, when excess power is produced by renewable sources, energy companies want to get some return for their production investment rather than just letting it dissipate so they price electricity at an extremely cheap rate. In extreme cases, this can even result in negatively priced power where power supplies pay their customers to buy the power. Lancium is able to depend on renewable power because of the market groups they target. Lancium markets themselves as "a cloud grid that is optimized for the execution of High Throughput Computing (HTC)" (Lancium Compute Documentation, 2021). HTC jobs, generally, are loosely coupled, identical programs operating on different sets of input parameters. This class of tasks are optimal for Lancium's power sourcing goals because they can be transitioned to other data center locations if power is temporarily unavailable or expensive at one location. Speed, while a consideration, is not critical and thus machines can be shut down and resumed if power is constrained or if idling. Thus, by identifying their relevant social group as users running HTC jobs instead of generals users or developers expecting always-on services, Lancium is able to provide cheap, green computing resources that meet the demands of their market.

Furthermore, Lancium is able to simultaneously reduce costs and their environmental impact by using airflow to cool server racks rather than air conditioning. Hydrofluorocarbons (HFCs) are the primary molecule used as cooling agents in refrigeration and air conditioning. While they were developed as a greener alternative to chlorofluorocarbons (CFCs) (Benhadid-Dib, S., Benzaoui, A., 2011), "HFCs are potent greenhouse gases with a global warming potential up to 12400 times that of CO2 per mass unit" (Purohit, P., Höglund-Isaksson, L., Dulac, J. *et.al*, 2020). Thus, limiting the use of these modules is a critical factor in combatting climate change. Most data centers are kept at temperatures around 60ºC, in accordance with manufacturer recommendations. This provides some cost protection if the technology fails as manufacturer warranties still apply. However, technology has continued to become more resilient. For example, Google has been experimenting with ambient temperatures in the range of 80ºC to 90ºC. Lancium uses older, cheaper computers allowing them to forgo air conditioning all together and has measured ambient temperatures of up to 108ºC in their data centers since they began operating with little affect on the chips. Choosing a design that prioritizes environmental performance over minor additional maintenance costs and space efficiency has proven to be especially effective for Lancium, as the computers operate consistently at higher temperatures.

Another design decision made by Lancium was to source older machines for their data centers, which reduces technological waste sent to landfills. According to one study, "Yearly e-waste generation relies on the average lifespan, no. of total units in check, and mass of the electronic device. Three years is the common lifespan of computers" (Arkam, R., Natasha, Fahad, S. *et al*, 2019). This design choice might seem to put Lancium at a computational disadvantage, as Moore's Law states "that the speed of computers; as measured by the number of transistors that can be placed on a single chip, will double every year or two" (Mollick, E. 2006). However, in recent years there has been a decline in the performance improvement of integrated circuits. An article titled *The End of Moore's Law* (Theis, T. N., Wong, H.-S. Philip, 2016) "shows a very important exponential trend in information technology that already shows a sharp slowing of progress." The trend being discussed is the switching energy, which is a form of quantitatively measuring the power of different processors. Thus, chips produced today are, on average, only about 25% to 30% more efficient than chips produced three or four years ago. Lancium purchases the older chips, which reduces both e-waste and operating costs because the older computers are priced significantly cheaper than newer machines. To achieve throughput values comparable to a data center with cutting edge chips, Lancium must purchase and store around 30% more chips, but the savings on older machinery still outweigh the higher volume costs. Thus, from a modern perspective it can be understood why Lancium made the design decision to use outdated equipment, as e-waste is kept out of landfills, costs are lowered, and with the stagnation of chip performance there is little impact to the available computational resources.

Lancium Massively Scaled

As the number and processor density of data centers grow, the impact of their carbon, refrigerant, and technological waste will continue to adversely affect the environment. However, many studies show that this outcome is not inevitable, and Lancium has demonstrated a number of successful solutions to the rising environmental impact of data centers. Why then, are larger data center providers not also adopting these greener measures? There are several plausible factors including limited access to clean energy sources, costs of transitioning existing data centers, marketing, and demand for always-on power.

Lancium is a comparatively small provider and markets to a much narrower customer base relative to giants like Amazon and Google. As such, providing always-on power is not a main priority. To run entirely on renewable sources, massive providers would need to reserve the intermittent power using “energy storage systems such as pump hydro, compressed air, molten salt, flywheel, etc.” (Kao, W., 2015). However, these options would create an added cost and might offset the energy savings of using renewables, resulting in lower returns for the providers and higher costs for customers. In today’s world, these relevant social groups have much more direct control over a company's operations than those affected by its environmental impacts, which might explain why data centers continue to rely primarily on fossil fuels. In addition, massive data centers may simply out-demand the supply of available clean energy, resulting in the need for more renewable generators.

Green design choices, such using basic airflow cooling systems and adopting older chips are tightly coupled. Designs consisting of tightly packed, cutting edge machines must be water-cooled to prevent overheating. Traditional data centers likely use cutting edge processors for three primary reasons. First, to meet the computing demands of customers, data centers would either need to expand in size to house increasing numbers of older computers, which is significantly more costly than utilizing already available space. Second, with higher numbers of less robust chips, and at the scale massive providers operate, added maintenance could incur significant costs to both the providers and customers. Marketing might also play a role as developers are more likely to find data centers using cutting edge technologies as more appealing than seeming outdated providers. As stated above, these groups have direct influence over company decisions and operations, unlike individuals around the globe struggling to deal with Earth's changing climate. To overtake traditional data center designs, either Lancium needs to meet current data center performance benchmarks or the concerns of the three social groups need to be more equally addressed.

Closure Paths

The first closure path described by SCOT is rhetorical closure. Under this path, the relevant social groups must agree on an optimal data center design. With the groups identified above (data center providers, general consumers, developers, and societies worldwide), an optimal solution would result in high returns for shareholders, low maintenance, low costs for consumers, always-on availability, and negligible impact on the environment. It appears that Lancium's model is a significant step towards achieving this over traditional data center designs. However, to be viable at a worldwide scale, improved battery technologies would need to be designed and modern chips need to be more energy efficient and heat robust.

The second closure path that could be followed is a redefinition of the problem. This could take two forms. Either, concern for the environment dominates the demand for increasing computational resources or new technologies are developed to moderate our impact on the planet and traditional data center designs can continue to operate will little concern to greenhouse gas emissions and waste.

Finally, while SCOT states that interpretive and design flexibility collapse over time, closure is not permanent. Possible situations may arise that redefine the optimal data center design, which might include controlling our impact on the environment in the future, discovering new, clean sources of energy, moving data centers from Earth to orbiting satellites in space, and others.

Conclusion

Lancium's data center design, while offering many financial and environmental benefits over traditional designs, is a socially constructed technology, and traditional designs dominate the market because they are socially supported. However, recognizing that data center designs have a non-trivial impact on the environment, and accounting for needs and impact on the social group consisting of all individuals and societies, reveals that green data center designs will play a larger role in the future. The negatives of Lancium's greener designs may be accepted in favor of protecting the Earth and hopefully, together, we can apply our engineering ingenuity to solving the remaining shortcomings and leave society and our blue dot stronger and healthier than we found them.

References

Pinch, T. J., Bijker, W. E. (1984). The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. Retrieved March 9, 2021, from https://www.jstor.org/stable/285355

Bouley, D. (2010). Estimating a Data Center's Electrical Carbon Footprint. Retrieved March 9, 2021, from https://sg.insight.com/content/dam/insight/en_US/pdfs/apc/apc-estimating-data-centers-carbon-footprint.pdf

EN-ROADS Climate Interactive. (n.d.). Retrieved March 9, 2021, from https://en-roads.climateinteractive.org/scenario.html?v=2.7.37

Ocean Acidification. (n.d.). Retrieved March 9, 2021, from https://www.noaa.gov/education/resource-collections/ocean-coasts/ocean-acidification

Hurricanes and Climate Change. (n.d.). Retrieved March 9, 2021, from https://scied.ucar.edu/learning-zone/climate-change-impacts/hurricanes-and-climate-change

Data Centers and Servers. (n.d.). Retrieved March 9, 2021, from https://www.energy.gov/eere/buildings/data-centers-and-servers

Microsoft Corporate Social Responsibility. (n.d.). Retrieved March 9, 2021, from https://www.microsoft.com/en-us/corporate-responsibility/sustainability?activetab=pivot_1%3aprimaryr3

Google Data Centers. (n.d.). Retrieved March 9, 2021, from https://www.google.com/about/datacenters/renewable/

Sustainability in the Cloud. (n.d.). Retrieved March 9, 2021, from https://sustainability.aboutamazon.com/environment/the-cloud?energyType=true

About Lancium. (n.d.). Retrieved March 9, 2021, from https://portal.lancium.com/about/company

Fui-Hoon Nah, Fiona. (2007, February 3). A study on tolerable waiting time: how long are Web users willing to wait?. Retrieved March 9, 2021, from https://doi.org/10.1080/01449290410001669914

United Nations Secretary-General. (2018). Secretary-General’s remarks on Climate Change [as delivered]. Retrieved March 9, 2021, from https://www.un.org/sg/en/content/sg/statement/2018-09-10/secretary-generals-remarks-climate-change-delivered

Alger, D. (2009, August). Grow a Greener Data Center. Retrieved March 9, 2021, from https://learning.oreilly.com/library/view/grow-a-greener/9781587059919/

Benhadid-Dib, S., Benzaoui, A. (2011, April 14). Refrigerants and their impact in the environment. Use of the solar energy as the source of energy. Retrieved March 9, 2021, from https://doi.org/10.1016/j.egypro.2011.05.040

Purohit, P., Höglund-Isaksson, L., Dulac, J., Shah, N., Wei, M., Rafaj, P., and Schöpp, W. (2020, October 5). Electricity savings and greenhouse gas emission reductions from global phase-down of hydrofluorocarbons. Retrieved March 9, 2021, from https://doi.org/10.5194/acp-20-11305-2020

Akram, R., Natasha, Fahad, S., Hashmi, M. Z., Wahid, A., Adnan, M., Mubeen, M., Khan, N., Asif Rehmani, M. I., Awais, M., Abbas, M., Shahzad, K., Ahmad, S., Hammad, H. M., Nasim, W. (2019, April 25). Trends of electronic waste pollution and its impact on the global environment and ecosystem. Retrieved March 9, 2021, from https://doi-org.proxy01.its.virginia.edu/10.1007/s11356-019-04998-2

Mollick, E. (2006, August 21). Establishing Moore's Law. Retrieved March 9, 2021, from https://doi.org/10.1109/MAHC.2006.45

Theis, T. N., Wong, H.-S. Philip. (2016, May/June). The End of Moore's Law. Retrieved March 9, 2021, from https://doi.org/10.1109/MCSE.2017.29

Kao, W. (2014, October 24). Renewable and Clean Energy for Data Centers. Retrieved March 9, 2021, from https://doi.org/10.1002/9781118937563.ch30