Fortifying the Cloud Computing Using Big Data Security Tools
(Technical Topic)

Investigating the Future of Data Center Construction in Virginia
(STS Topic)

A Thesis Project Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science in Computer Science

Dagim Tekle
Fall, 2023

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

ADVISORS

William Davis, Assistant Professor of STS, Department of Engineering and Society

Elizabeth Oricco, Academic General Faculty, Department of Computer Science

Introduction

Cloud computing is one of the most integral Internet of Things (IoT) infrastructures hosting various services in the categories of entertainment, communication, retail, finance, medical research, business analytics, AI, and more. Unsurprisingly, some of the most successful and innovative companies like Netflix, Fox, Snap, Airbnb, Amazon, CapitalOne, Intuit, AstraZeneca, Philips, and Coca-Cola employ cloud solutions (AWS Solutions Library). The momentous migration to the cloud can be measured in the forecast of increased global annual end-user spending on public cloud services ($490.3 billion in 2022 to $591.8 billion in 2023) and increased global annual spending on construction of new data centers ($32 billion in 2022 to $49 billion by 2023) (DeLis, 2022; Bangalore et. al, 2023).

Despite the economic and technological opportunities presented by the cloud industry, the increased use of data centers poses a host of environmental challenges that need our attention. Data centers are nearly 1% of annual energy-related global greenhouse gas emissions, about 2% of U.S. electricity consumption, and contribute to the annual 60 million metric tons of e-waste (Pratt, 2023). Additionally, the excessive heating from the hardware in data centers requires heavy water consumption: “Google’s data centers in the U.S. alone consumed an estimated 12.7 billion liters of fresh water” (Danleski). Finally, there is an increased protest against the noise from data centers experienced by those in nearby locales (Pratt, 2023). These issues are already alarming in the status quo of maintaining existing data centers. But as the previously established upsurge of expenditure fuels the construction of more data centers, the socio-ecological side-effects – greenhouse emissions, electricity usage, water usage, electronic waste, and noise pollution – will only intensify.

Similarly, PJM in collaboration with Dominion reported that the growing data center load, concentrated in the Northern Virginia region, will result in reliability violations and outage conditions starting in 2024 unless the transmission lines undergo two baseline upgrades and eleven supplemental reinforcements (Abdulsalam, 2022). In other words, the increased presence of data center operations has led to the right conditions that lead to blackouts in the Northern Virginia area, jeopardizing the reliability of electricity for residents. Furthermore, what is the cost of the identified projects? Will residents have increased electricity bills to sustain these substantial upgrades? Or will data center companies bear the cost, dissuading companies from investing in Virginia, and taking economic opportunities away?

For the technical portion of my thesis, I will write a report about my internship experience with a cloud provider and explore some interesting applications of cloud computing in big data computation. For my STS thesis, I hope to explore how different entities affect the construction of data centers in Virginia. To understand the social influence on the construction of data centers, I will utilize a sociotechnical theory called Social Construction of Technology (SCOT) which includes five components: relevant social groups, interpretation, closure, technological frame, and the wider social context (Klien & Kleinman, 2002, pg. 36). Focusing on data center energy usage, I have identified three general categories of relevant groups that would have a vested interest and influence on data center construction: government and legislation, data center owners, and energy companies.

Technical Topic

Cloud computing, according to the National Institute of Standards and Technology, is defined as “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and

services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" (Surbiryala & Rong, 2019, p. 1). Similar to how the widely accessible, minimal set-up and completely self-service electricity lines enable toasters, coffee machines, and smartphone companies to not have to worry about providing electricity for their machines, cloud providers, through internet-based access, enable cloud users to focus on innovation without the hustle of maintaining physical servers and computing resources.

One of the cloud's security best practices is implementing a continuous monitoring and logging system (16 cloud security best practices - crowd strike). The benefits of real-time monitoring are two-fold: 1) having verbose monitoring data can help security investigators properly identify malicious actions when digging through logs for security investigations, and 2) a real-time logging system allows security investigators to respond promptly to the identified malicious activity minimizing the impacts or further damage of a security vulnerability.

Over the last two summers, I interned with a cloud provider in a team dedicated to creating software to enable their security engineers (or security investigators) to perform their investigations wholistically (working with all log data) and efficiently (reduced expense and time to do investigations). Working with all datasets while continuing to maintain or improve the cost and time spent on investigations is especially important when considering the growing and already hefty amount of data the engineers need to comb through and interact with: people create about 328.77 exabytes of data daily over the internet (Duarte, 2023). In other words, people will create 329 million one-terabyte Xboxes worth of data every day. This kind of data is colloquially called "big data" by my co-workers and generally within the computer science world.

I got the privilege to research two big data services for my team. The first service I worked on was big data sanitization. As the name suggests, big data sanitization is a service that

sanitizes (mask, obfuscate, replace, delete columns) big data. Our customers, the security engineers, reported that they would often have to share data with company-external entities as a part of their investigation like a legal service, an external investigator, or the government. Because of the company's internal least privileged access policies and service-level security agreements, however, they cannot share the raw data. To comply with the policies, security engineers would spend much time creating new programs and learning a new language to sanitize datasets. This big data sanitization feature would provide a minimal setup self-service alternative, reducing the overhead for security engineers.

The second project I worked on was investigating a cost-effective scalable big data indexing and querying strategy for security logs. Indexing refers to the process of creating an inverted index, a dictionary used in information retrieval systems to efficiently retrieve files containing the specified search words where searching through all the documents would be prohibitively slow (Inverted index). The inverted index functions like an index at the end of a book listing all the pages that contain a word in question. This is incredibly faster compared to reading through all the pages of a book every time we want to search for a word. On a working cost-effective scalable big data indexing and querying system, security engineers would be able to submit queries and get prompt results (basically a googling system for security logs). Though this strategy is effective, in-house highly scalable indexing and querying systems are costly to store and have some bottleneck issues, which I will explore in-depth in the technical thesis.

STS Topic

What is the cloud anyway? In a 2013 survey asking this question to people in the United States, "29% thought that the cloud in cloud computing was a real cloud … currently sailing by in the sky", "17% pretended to be knowledgeable about the cloud on a first date", and "just 16%

really knew what the cloud was" (So what do you really know about the cloud, anyway?, 2019). Though a decade later, the general public still lacks a clear understanding of what the cloud is and remains unfamiliar with the socio-ecological and economic impacts of data center infrastructure. Thus, the goal of the socio-technical portion of this thesis is fourfold: 1) to provide a simple definition of cloud computing and how it depends on data centers, 2) to investigate the trends and impacts of data center electricity usage in Virginia, 3) to analyze the complex web of negotiations shaping the future of Virginia's data center infrastructure, and 4) to suggest an effective reform for a more sustainable and socially responsible data center industry.

We begin by utilizing SCOT to identify relevant groups by starting with a selection of a few actors, and then observing other relevant actors they identify (Klien & Kleinman, 2002, pg. 32). Starting with government and legislation, first is the Virginia Clean Economy Act (VCEA) which is the local government's attempt to eliminate most CO2 emissions from Virgnina's Electricity sector by 2050 (Ball & Jennings). VCEA is a pivotal legislation that will reshape the electricity market which can thereby affect data center electricity consumption.

Next, the Virginia Economic Development Partnership (VEDP) governmental entity responsible for the state's economic development, reports that Virginia hosts the largest data center market in the world and contains 35% of all known hyperscale data centers worldwide (Virginia Economic Development Partnership). Understanding the governmental forces displays the economic and legal support promoting the data center market. Additionally, the VEDP gives us the perspective of the current state of Virginia's claim to dominance in the data center arena and the global magnitude of this market.

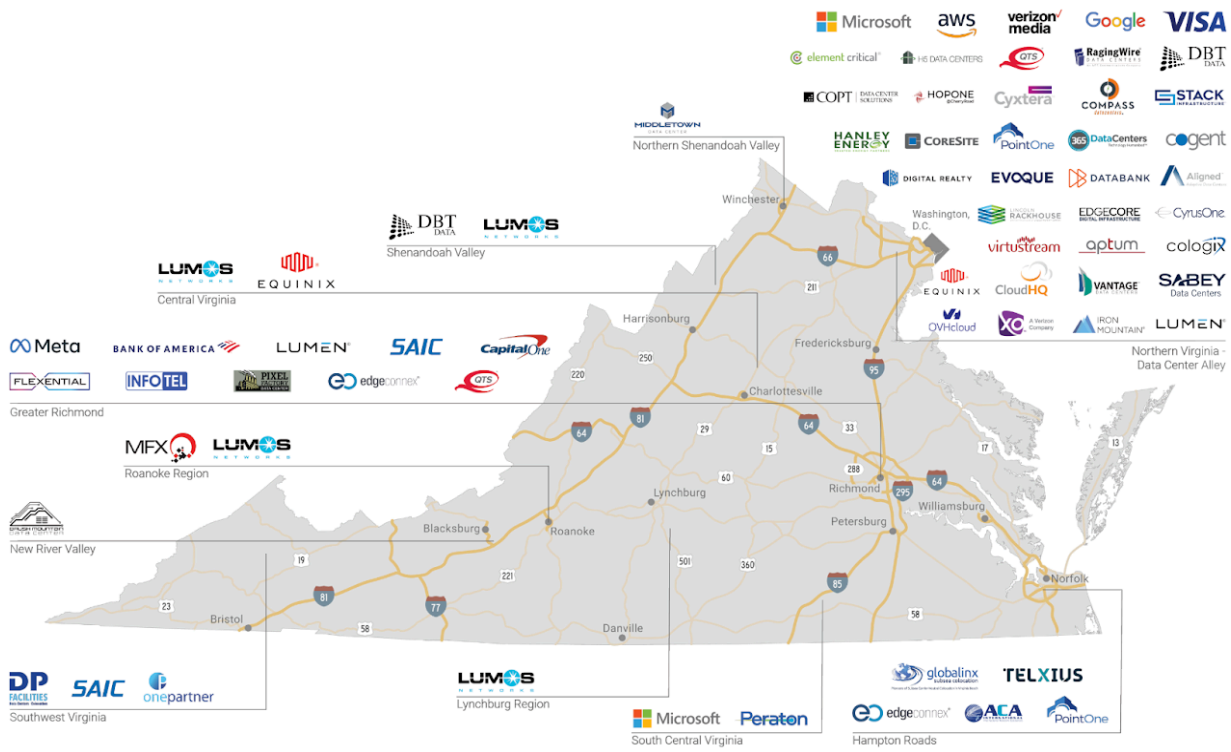

Figure 1: Data Centers in Virginia (Virginia economic development partnership)

The VEDP also identifies the next set of relevant groups: companies that own and operate data centers in Virginia, and Amazon Web Services (AWS) is one of these groups (Figure 1). AWS has been in Virginia since 2006 and owns 72 of the 179 data center sites in Virginia (AWS in your community, 2023; Virginia Data Centers; AWS US East). In addition to operating 40% of Virginia's data centers concentrated in Northern Virginia, AWS has a new headquarters in Arlington. Within the realm of companies owning and operating data centers in Virginia, AWS takes center stage as a representative force.

Lastly, the electricity market introduces two key players, PJM as an electricity transportation company and Dominion Energy as Virginia's primary energy generating company (Abdulsalam, Dominion Energy). Their roles in providing electricity to the data center market

and their obligations to send usage reports under legislation like VECA highlight the electricity market's intricate interplay between the government and the data center owners.

Though this is not an exhaustive list, it is a representation of the main relevant groups involved in data center infrastructure—namely, the government, the data center market, and the electricity market. By tracing the interactions among these entities listed above, I intend to apply the SCOT framework to gain insights into how the relevant groups 1) interpret data centers, and 2) negotiate with one another to reach a consensus. By gaining this deeper understanding of how these groups are shaping the landscape of data center construction in Virginia, we can examine socio-ecologically consequential changes in the form of new policies and/or improved sustainability practices.

Overall Conclusion

The rapid expansion of cloud computing and data center construction presents a dual narrative of technological advancement and environmental challenges. The staggering growth in global annual spending on public cloud services and the construction of new data centers underscores the economic and technological opportunities in the cloud industry. However, the social, environmental, and economic repercussions centered around energy usage, demand urgent attention. Addressing these challenges requires collaboration among industry leaders, policymakers, and the public to ensure a sustainable and socially responsible future for the evolving data center landscape.

The socio-technical aspect of the thesis unveils a complex web of negotiations involving various stakeholders like the government, the data center market, and the electricity market. The application of the SCOT framework offers a deeper understanding of the background and trajectory of the data center landscape in Virginia. The insights from this framework will enable

us to investigate the types of reforms that will effectively address the socio-ecological disparities from data center infrastructure, allowing citizens to collaborate effectively towards a socially responsible and sustainable future.

The technical aspect of the thesis delves into the critical role of security measures in handling the colossal volume of data. The internship experience with a cloud provider sheds light on innovative tools designed to enhance security engineers' efficiency in navigating security logs. These tools, such as big data sanitization and scalable indexing, contribute to the ongoing efforts to fortify the cloud and mitigate potential threats.

References

16 cloud security best practices - crowdstrike. crowdstrike.com. (2023, August 21). https://www.crowdstrike.com/cybersecurity-101/cloud-security/cloud-security-best-practices/

Abdulsalam, S. (2022, July 12). Dominion Northern Virginia Area Immediate Need. PJM. https://www.pjm.com/-/media/committees-groups/committees/teac/2022/20220712/item-08---dominion-northern-virginia---immediate-need.ashx

AWS in your community: Here's what's happening in Northern Virginia. US About Amazon. (2023, October 12). https://www.aboutamazon.com/news/aws/aws-data-centers-virginia

AWS Solutions Library. Amazon. https://aws.amazon.com/solutions/

AWS US East (N. virginia) Data Centers & Colocation - Baxtel. (n.d.). https://baxtel.com/data-center/aws-us-east-n-virginia

Ball, B., & Jennings, A. (2021, December 14). Modeling Decarbonization: Report Summary and Policy Brief for Virginia Governor's Office Administration and Policymakers (Chapter 1194, 2020). REPORTS TO THE GENERAL ASSEMBLY. https://rga.lis.virginia.gov/Published/2021/SD17

Bangalore, S., Bhan, A., Miglio, A. D., Sachdeva, P., Sarma, V., Sharma, R., & Srivathsan, B. (2023, January 17). Investing in the rising data center economy. McKinsey & Company. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/investing-in-the-rising-data-center-economy

Danleski, D. (2023, April 28). AI programs consume large volumes of scarce water. News. https://news.ucr.edu/articles/2023/04/28/ai-programs-consume-large-volumes-scarce-water

DeLis, M. R. (2022, October 21). Gartner Forecasts Worldwide Public Cloud End-user spending to reach nearly $600 billion in 2023. Gartner.

Dominion Energy. (2023, February 8). Q4 2022 earnings call. https://s2.q4cdn.com/510812146/files/doc_financials/2022/q4/2023-02-08-DE-IR-4Q-2022-earnings-call-slides-vTC.pdf

https://www.gartner.com/en/newsroom/press-releases/2022-10-31-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-reach-nearly-600-billion-in-2023

Duarte, F. (2023, April 3). Amount of data created daily (2023). *Exploding Topics*. https://explodingtopics.com/blog/data-generated-per-day#how-much

Inverted index. GeeksforGeeks. (2023, July 5). https://www.geeksforgeeks.org/inverted-index/

J. Surbiryala and C. Rong, "Cloud Computing: History and Overview," 2019 IEEE Cloud Summit, Washington, DC, USA, 2019, pp. 1-7, doi: 10.1109/CloudSummit47114.2019.00007.

Klein, H. K., & Kleinman, D. L. (2002). The social construction of technology: Structural considerations. *Science, Technology, & Human Values*, *27*(1), 28–52. http://www.jstor.org/stable/690274

Monserrate, S. G. (2022, January 27). The cloud is material: On the environmental impacts of computation and data storage. *MIT Case Studies in Social and Ethical Responsibilities of Computing*. https://mit-serc.pubpub.org/pub/the-cloud-is-material/release/1

Pratt, M. K. (2023, June 7). Cloud computing's real-world environmental impact: TechTarget. *Sustainability and ESG*. https://www.techtarget.com/sustainability/feature/Cloud-computings-real-world-environmental-impact

So what do you really know about the cloud, anyway?. CloudWedge. (2019, February 11). https://cloudwedge.com/resources/so-what-do-you-really-know-about-the-cloud-anyway/

The Piedmont Environmental Council. (n.d.). https://www.pecva.org/

Virginia Data Centers - Colocation and Cloud. Data Center Map. (n.d.). https://www.datacentermap.com/usa/virginia/

Virginia economic development partnership. VEDP. (n.d.). https://www.vedp.org/