Efficiency and Transparency: How a Small Intern Project Can Save Days of Compute Time
(Technical Report)

Analyzing the Effects of COVID-era World Events on the GPU Market
(STS Research Paper)

A Thesis Prospectus
In STS 4500

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 in Computer Science

By
Winston Zhang

October 27, 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

Travis Elliott, Department of Engineering and Society
Rosanne Vrugtman, Department of Computer Science

General Introduction

Computer hardware and software are commonly thought of as fields of science and engineering that have both rapidly advanced in the past century, and facilitated the rapid advance of society, production efficiency, and the human standard of living. With the rise of the Information Era, new developments have caused various social issues to bubble to the surface of social discourse, such as data privacy, censorship, and distribution/denial of service. While some have thought computing was beginning to reach a steady state, recent world events such as supply chain disruptions brought upon by the COVID pandemic, as well as the advent and proliferation of AI models have begun to pave the way for new trends in the computing sphere. One such trend is the cloud computing infrastructure model, which began to gain popularity pre-COVID, and is becoming an increasingly popular model of IT infrastructure and resource allocation. In parallel with the rise of cloud computing in the recent past, the market for computing machinery which the cloud computing industry, as well as the consumer electronics industry, relied on began to experience turmoil due to a tandem shock to the supply of computer chips and surge in demand for electronics. My technical report details an internship project I developed, which highlights the software engineering process in the context of the cloud computing infrastructure model, highlighting its key aspects such as pay-as-you-go pricing, and my STS research paper aims to analyze the factors that led to the sudden onset of market euphoria related to graphics cards, a component found in almost all electronics today. Although these two topics seem disjoint, they have been key developments in the computing industry in the past decade, and are beginning to converge, especially as companies start to shift their services to cloud-based servers, and also employ the use of AI models in a greater capacity, both of which require powerful GPUs to run smoothly.

Technical Report: Efficiency and Transparency: How a Small Intern Project Can Save Days of Compute Time

While many in today's world of e-commerce consider it a prerequisite for companies to have a website to offer a means to order online, many also fail to appreciate the IT infrastructure needs of companies to keep such online services running smoothly, while achieving cost-efficiency. Consider the position of Netflix, which may experience an elevated level of internet traffic in the winter months, relative to the summer, when people stay indoors and stream video for a more significant portion of the day. Under a traditional IT infrastructure model, Netflix would need to buy enough servers to handle the demand, and then let those servers sit unused for the summer, raising their operating costs. Imposing this same model upon thousands of other companies that experience similar seasonal web traffic demands, it becomes clear that the traditional IT infrastructure model results in millions of dollars collectively wasted annually. Cloud computing, an IT solution that has taken the tech sphere by storm in the past decade, alleviates this problem by consolidating resources under a single provider, who then grants access to multiple companies on a pay-as-you-go basis, which can be used in varying quantities (Bigelow, 2022). Returning to the anecdotal example, the computing resources that Netflix doesn't use during the summer could then be elastically reassigned to some other company that experiences elevated traffic in the summer and a lull in the winter, such as a travel booking site. Such a modus operandi provides flexibility and scalability of resource usage while minimizing overhead costs related to providing web-based services.

My technical report documents my project during my second internship at Fannie Mae over the summer of 2023. The software product that my team maintained was a web application intended for use by the company's business division, which provided a graphical interface for users to retrieve information from a large relational database without requiring any knowledge of

SQL queries. I spent the previous year refactoring the software, originally built by my father before my employment, and its supporting assets in assistance to my team as part of a company-wide initiative to migrate software assets to the AWS Cloud to mitigate operating costs. Despite the migration being completed after I left my first internship, cost mitigation goals had not been fully realized, with several features needing to be implemented. My project for my second internship tackled yet another problem relating to unnecessary operating costs. Due to the cloud migration, the software product was now on Amazon's pay-as-you-go billing system for their cloud services. Ordinary use cases of the application would see users attempting to retrieve thousands, if not millions of data points, in one query, which would take the application hours to complete. The website on which the application's user-facing interface offered no way for the user to check on the status of their data retrieval, which would often result in users either enqueueing duplicate jobs that would take hours each, resulting in the application taking several days to process the same job multiple times, or users sending multiple emails to the development team, who would need to stop working to log into the AWS management console to manually check on the status of a job.

To fully address this problem, a full reworking of the software backend was needed to return output to the user more quickly. I did not have enough time in my internship to fully understand how various AWS services were used to make the newly migrated application function as a whole, so I was tasked with implementing a temporary solution: create a new tab on the portal that would show a status board to display the queue and processing status of jobs being processed. My technical report details the technologies I used, the process of development over my ten-week internship, the resulting implementation, and proposed future work to improve my implementation. It has already been completed and is included after this page.

STS Paper: Analyzing the Effects of COVID-era World Events on the GPU Market

The graphics card is a piece of computing machinery that is used to render graphics that are displayed as output to the user, usually via a monitor. The main piece of hardware on a graphics card is the graphics processing unit, or GPU. Despite graphics cards consisting of other hardware components, such as a circuit board and memory, modern vernacular uses the terms graphics card and GPU interchangeably. Unlike the central processing unit, or CPU, the GPU's circuit architecture takes advantage of parallelism, such that its methods of computation accelerate the processing of graphical workloads, which would not be feasible with only a CPU (Intel). First created in the 1970s, graphics cards became more widely used, as demand rose due to the increasing popularity of 3D graphics in arcade and computer gaming (Dally, 2021). Today, GPUs are included in dedicated graphics cards separate from CPUs, are also integrated into processors alongside CPUs, and can be found in almost all personal electronic devices, including PCs, smartphones, and gaming consoles. Other than graphics rendering applications such as video editing and gaming, modern use cases for GPUs also include training machine learning models and mining cryptocurrency (Gillis, 2020).

Despite the ubiquity of electronic devices, and thus the demand for computing components that have GPUs in them, various social and economic forces in previous years stemming from an unprecedented pandemic in the now globalist economy gave rise to wide-sweeping shortages, scalping, and hysteria that have rarely been seen in the market for technology, or even the greater market for consumer goods. Such social and economic forces induced by the COVID pandemic include disruptions to the fragile supply chain, a sudden rise in demand for electronics devices during lockdown, inflation and increased consumer spending due to drastic stimulus measures taken by the government, and an increase in demand for GPUs in a

bullish market for cryptocurrency. These various forces acting in unison caused what was essentially a complete elimination of supply in the face of rapidly multiplying demand, leading to skyrocketing prices of GPUs and electronic devices that had GPUs in them. The outcome of these market movements shaped the GPU market as it is today, leading to the high prices we see in the consumer GPU market.

The GPU market demonstrates itself to be an endogenous system that responds to the changes in other markets, such as the market for consumer electronics, the market for cryptocurrency, and the market for money and loanable funds. Because there exist so many different contributing factors that led to the current state of the market for GPUs, this sociotechnical system would be best analyzed using Actor-Network Theory (ANT). Actor-network theory is an STS framework that examines systems as networks comprised of participants, human or inanimate, known as actors/actants, who are connected by dynamic relationships that define the network (Muniesa, 2015). ANT is especially useful for the analysis of systems of nested complexity, as networks can themselves be considered actants of a greater network, and conversely, actants can be a smaller network: an automobile can be considered a network whose actants are the engine, the steering wheel, the tires, and even the driver, but each actant can be further broken down as their own network of actants, such as an engine being a network of pistons, valves, and shafts. ANT, in this particular application, considers each separate market force that affected the change in supply or demand in the GPU market to be an actant, which operates within a larger network that affects the GPU market, our main actant of interest. Additionally, since the exorbitant prices and low supply of GPUs in the wake of COVID eventually began to affect the dynamic between producers and consumers, the GPU market would then be further decomposed so that it is examined as its own network, with the individuals

operating within the GPU market as actants. These changes enacted by such forces can be modeled using endogenous supply-and-demand models, a technique commonly employed by macroeconomists to analyze how interrelated markets can affect each other, and reflexively affect themselves.

Conclusion

My technical report details my internship project to provide a window of transparency to a black box software whose costs are directly correlated with its duration of usage in cloud computing resources. It identifies a new kind of problem not commonly encountered in enterprise software before the advent of cloud computing and details the process through which I implemented a software feature to address it.

My STS research paper investigates and analyzes the sociotechnical factors that led to the current market for graphics cards, in the context of Actor-Network Theory. It will use ANT as well as macroeconomic models to qualitatively explain how a concert of various economic forces led to market conditions that have rarely been seen in human history.

Resource efficiency in cloud computing and market conditions of the GPU market, despite seeming disjoint in the computing landscape, are interrelated due to the applications of GPUs. As many companies begin to shift their business models from providing products accessible in perpetuity to live services, they begin also to shift their IT infrastructure to cloud computing to more efficiently scale their operating costs and resource usage with customer demands. Not only are more media companies moving to the cloud, but tech companies are also beginning to increasingly make use of machine learning models after the massive boom in the popularity of AI in 2023 (Mckinsey, 2023). The migration to cloud computing in addition to the increasingly widespread use of AI will mean that the GPU market will see yet another significant

source of demand. My senior thesis portfolio hopes to address the topics of cloud computing and GPU usage in a way that could reveal how changes in computing applications affect the market conditions surrounding the machinery that is used to run said applications.

References

Bigelow, S. (2022, October). *Pay-as-you-go cloud computing (PAYG cloud computing)*. https://www.techtarget.com/searchstorage/definition/pay-as-you-go-cloud-computing-PAYG-cloud-computing

Dally, W., Keckler, W., & Kirk, D. (2021). *Evolution of the Graphics Processing Unit (GPU)*. IEEE Micro, 41 (6), 42-51. https://doi.org/10.1109/MM.2021.3113475

Gillis, A. (2020, December). *Graphics processing unit (GPU)*. TechTarget. https://www.techtarget.com/searchvirtualdesktop/definition/GPU-graphics-processing-unit

Intel. (n.d.). *What is a GPU?*. Retrieved October 3, 2023, from https://www.intel.com/content/www/us/en/products/docs/processors/what-is-a-gpu.html#:~:text=While%20the%20terms%20GPU%20and,board%20that%20incorporates%20the%20GPU.

Mckinsey. (2023, August 1). *The state of AI in 2023: Generative AI's breakout year*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year

Muniesa, F. (2015). *Actor-Network Theory*. In J.D. Wright (Ed.), *International Encyclopedia of the Social and Behavioral Sciences* (2nd ed.). https://www.sciencedirect.com/science/article/pii/B9780080970868850011?via%3Dihub