Architecting Computer Vision Workloads on AWS Cloud

(Technical Topic)

The Technopolitics of Content Moderation on Cloud Service Platforms

(STS Topic)

A Thesis 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 for the Degree
Bachelor of Science in Computer Science

By

Arvind Anand

November 1, 2021

Technical Team Member: Dylan Fernandes

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

Sean Ferguson, Department of Engineering and Society

Haiying Shen, Department of Computer Science

Introduction

In the year 2021, few trends in computing come close to the prevalence of cloud computing. Buying, building, and maintaining messy server farms need not apply; deploying code/projects is as easy as signing up for an Amazon Web Services (AWS) account. The obliquity of infrastructure-as-a-service products can allow for the novel application of state-of-the-art technologies in various industries.

For my technical capstone project, my partner and I are developing a scalable web service for Periop Green, a hospital waste reduction tech startup at UVA, that will predict scrub table waste in hospital operating rooms using object detection. The web service will be deployed onto AWS to fulfill the high compute requirements and storage demands of ML workloads. The goal of the end-to-end system is to help perioperative administrators reduce the wastage of unused one-time-use items in the operating room (OR).

Public cloud providers have a role in maintaining the integrity of their customers' data while keeping hate speech and malicious actors off their cloud platform by shutting down malicious applications and harmful data streams. Ultimately, patterns emerge from cloud service providers' systematic shutdown of apps and content; these patterns indirectly give rise to roles and policies for policing the internet. In the STS portion of this paper, I will use the framework of techno-politics to evaluate the emergence of politics in the space of internet-scale applications in the public cloud environment. Specifically, I will examine AWS's role in moderating content on the public internet through its influence and actions to shut down harmful actors.

Technical Topic

Problem Statement

Healthcare in the U.S. is unreasonably expensive for most Americans, even for those who have employer-provided plans. The cost of waste is about $91 billion, or 14% of the U.S.'s total health care spending budget (Lallemand, 2012). A large portion of the high costs in healthcare can be attributed to hospital waste. One source of hospital waste is associated with operating room wastage, accounting for 20 to 30% of all hospital waste (McGain, White, Mossenson, Kayak, & Story, 2012). One type of OR waste is unused one-time-use tools. One-time-use tools range from cheap equipment, like scalpels and scissors, to items that cost several thousands of dollars. Most one-time-use tools go unused in operating rooms and are subsequently disposed of because once a tool is exposed to the operating room, it is deemed unsanitary (Periop Green, 2021). Determining and tracking which tools are necessary for a particular operation is non-trivial for surgeons and perioperative administrators (Periop Green, 2021). Periop Green aims to tackle the problem of tracking and reducing the wastage of one-time-use tools in the OR. Periop Green was founded by Anesthesiologist Matthew Meyer at UVA and has ten primary team members working on business development, product marketing, social media outreach, and engineering. My capstone team member and I are working as cloud engineers on the engineering team.

Technical Implementation

At a high level, Periop Green uses computer vision to reduce waste in the OR by observing the scrub table during a surgical operation, processing the data, and presenting the perioperative administrators with a report containing the frequency of item usage. The

observation stage is performed by a nurse operating an IOT camera that observes the scrub table. The processing and analytics result from processing the observed scrub table video data; we refer to this as the backend system. The focus of my capstone project is implementing the infrastructure, applications, and ML job orchestration in the procession stage (Figure 1); we also refer to this as the Model Engine.

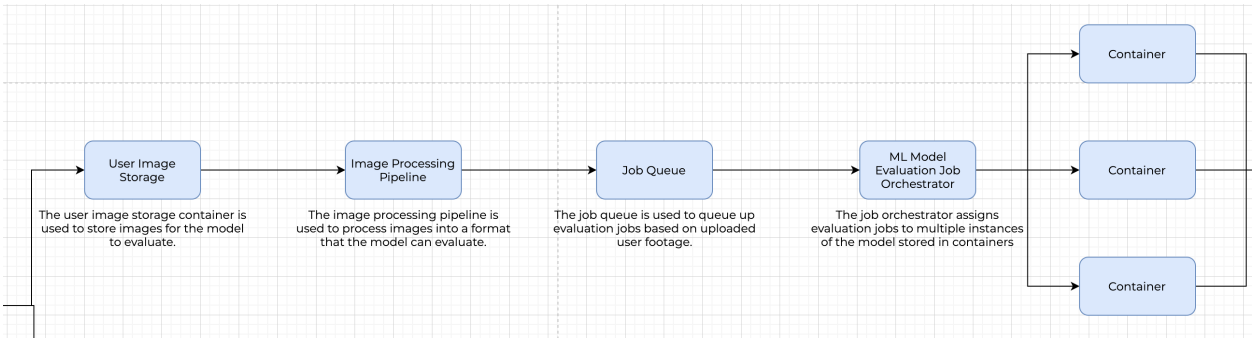

Figure 1: High Level ML Job Prediction and Training Pipeline for Periop Green

The implementation of the Model Engine is broken down into four stages. Stage 0 involves model development, testing in a local environment, and dataset annotation; this stage sets the groundwork for the Periop Green product by showcasing proof of concept for the idea. The computer vision model used is based on the Yolov4 object detection model implemented in the open-source darknet project. Yolo, which stands for "You Only Look Once," is a powerful object detection model that can draw bounding boxes around trained objects in a scene (Joseph Redmon, 2016). Stage 1 involves packaging the model into a portable format using containers, deploying the container onto a virtual machine (VM), and uploading training data using AWS's object storage services. While using a single VM instance is a viable way to train the model, a lot of training and testing would have to be done manually. To make our service more scalable, we need to automate the scheduling of training and testing jobs. Phase 2 utilizes real-time dataset monitoring to schedule training and testing jobs in an automated fashion. Phase 2 leverages

container orchestration to automatically spin up multiple ML model containers concurrently across multiple GPU-enabled VM instances. The automated job scheduling and automated container orchestration will significantly reduce the manual scheduling burden. They will allow for more effective use of computing resources by running many container jobs on a single instance instead of only one. Phases 3 offers refinements to phase 2 along with the development of a frontend interface for monitoring training jobs. Phase 4 mainly focuses on tightening security on our application. Currently, my partner and I are working on phase 2, which we plan to roll out by the end of 2021.

STS Topic

Topic Introduction

In order to explore the political emergences within public cloud services, each player in the public cloud market needs to be classified based on their ability to execute an innovative vision. The largest cloud provider as of 2021 is Amazon Web Services, Microsoft Azure, and Google Cloud (GCP), in descending order of most influential (Raj Bala et al., 2021). In 2019, AWS maintained a healthy lead in the market, which positions it as the most popular, innovative, and influential among all the cloud services providers on the planet (Raj Bala et al., 2021). Amazon now maintains and operates the service for a significant portion of the global internet indirectly by providing infrastructure-as-a-service to hundreds of tech companies around the globe. In terms of influence, AWS outcompetes the rest by virtue of its engineering prowess and innovation (Raj Bala et al., 2021). Observing case studies where AWS exercises its influence to shut down violent content and harmful malware can distill the policy and ethics that indirectly govern internet-scale applications.

Case Study 1: Content and Free Speech

Parler was founded in 2018 as a social media site that is "free speech-driven." (Vengattil, 2021) However, it served as a breeding ground for hate speech and violence, so much so that the activity on the platform directly resulted in the U.S. Capitol riot on January 6, 2021 (Vengattil, 2021). As a result, Apple and Google removed Parler from their respective app stores, and AWS shut down all of the compute resources Parler provisioned for web hosting and data storage (Vengattil, 2021). The removal of Parler from AWS shortly after January 6, 2021, was a display of AWS's influence on the free internet (Dang, 2021). The removal, ultimately, prevented the company of Parler from developing and deploying their applications. Critics of this move say that it showcased how far down the stack big tech companies can restrict free speech; however, others would argue that it is necessary to remove violence from the public internet, wherever it may be (Dang, 2021). AWS's action to remove Parler from using infrastructure is an instance of them exercising their influence to enforce the political idea of restricting free speech in the case of extremism. According to Dang, AWS is expanding its Trust and Safety team to "monitor for future threats (Dang, 2021)." It is not viable for AWS to sift through the massive amounts of content in S3 and applications running on EC2, but they plan to "get ahead of future threats" better to regulate their cloud service (Dang, 2021). Amazon expanding their Trust and Safety team is another example of their influence to enforce their political stance on restricted free speech because it will prevent future extremism from allowing to spread misinformation at the infrastructure level.

While the removal of Parler from AWS did not directly impact the cloud industry, it did force AWS and other players in the tech industry to revisit what is and what is not acceptable on

the public internet within their platforms (Gillespie, 2020). According to Gillespie, many tech platforms "keep an eye on each other" and "act in concert" when it comes to deplatforming individuals and internet-scale applications alike (Gillespie, 2020, p. 6). Gillespie also mentions that Cloudflare was inspired by the actions taken by Twitter and Facebook to no longer host extremist sites like Daily Stormer (Gillespie, 2020). Infrastructure services are also enrolled in the set of players in the tech space; they also respond and influence other players. Cloud companies are a very influential subset of the tech company bubble because of their engineering prowess and reputation: Google and Microsoft's respect in the industry is among the highest. As a result, AWS's move to deplatform Parler will not only influence the policies surrounding content moderation in the tech company bubble but will significantly influence how cloud service providers should moderate content on their platforms. The influence AWS exerts on other cloud providers will result in emerging politics concerning appropriate content for the open internet.

Case Study 2: Malware Cloud Hosting

Other bad actors on the internet exist only to exploit others, like spyware, ransomware, and generic network-based malware. In July 2021, Pegasus was used by the Israeli government to spy on high-profile figures and journalists by exploiting a bug on iOS. Furthermore, they used several cloud service providers to store data: AWS, Linode, Digital Ocean, and OVHcloud (Johansson, 2021). According to Johansson, these cloud providers cannot tell if a malicious actor is using their cloud service because it is impossible to distinguish one running service from another. However, AWS can associate cloud resources with account names, the only deterministic way to shut down a non-compliant account. AWS came to know of NSO's Pegasus via a report in the media, which prompted them to shut down accounts and resources allocated to

NSO (Johansson, 2021). While AWS took this approach, the other cloud providers "did not commit" to shutting down Pegasus servers on their platform (Johansson, 2021). This shows that while AWS might have the most influence among the cloud service providers, their actions do not necessarily translate into unanimous industry-wide adoption. However, AWS's action to take down NSO's AWS account did spark debate on preventing malicious actors from using the public cloud. Though AWS did not influence other service providers to follow suit, it forced them to rethink and strengthen their policies. Gillespie's ideas about tech company influence still hold here, with Digital Ocean and OVHcloud revisiting their terms of use policies.

Concluding Thoughts

The political emergence in the public cloud space is not only a free-speech debate; it is a constant cycle of iteration with the ultimate aim to influence the open internet. White nationalism, hate speech, and malware will live on so long as the internet remains open. However, the sheer domination of cloud entities like AWS keeps a large subset of the internet free of the aforementioned bad actors. The Parler case study shows that AWS uses its influence to shut down violent platforms. The Pegasus case study shows that AWS also uses its influence to clean the network of directly harmful bad actors. Ultimately, examining these case studies paints a broad picture of the emerging internet politics concerning large-scale cloud providers: there is constant iteration within the network of tech platforms, and the emergent politics influence other companies and could influence legislation.

Next Steps

In the current state of the techno-politics involving large public cloud service providers, few instances of take-down actions have occurred, but the few that have been indicative of how ethical actions cloud service providers play a role in regulating the open internet. As a follow-up,

I would evaluate how actions taken by AWS Trust and Safety have influenced international legislation. This would be an analysis of how various actors influence the legislation used to regulate cloud service providers. Furthermore, I would explore current regulations on tech companies and how actions taken by other actors have influenced said regulations to ultimately determine the trajectory of public cloud services related to the regulation of the open internet.

As for my technical capstone project, the future road map involves implementing phases 2-4, as mentioned earlier. I also want to have a high-level document detailing the technical vision for the company. This document will contain the tech stack, a top-down list of tools and technologies, and the fundamental design philosophy for implementing future products and phases. Phases 3 and 4 will happen next year, with phase 3 finishing by early Q2.

Bibliography

Dang, S. (2021, September 5). *Amazon considers more proactive approach to determining what belongs on its cloud service*. Retrieved from Reuters: https://www.reuters.com/technology/exclusive-amazon-proactively-remove-more-content-that-violates-rules-cloud-2021-09-02/

Flexera. (2019). *STATE OF THE CLOUDREPORT.* Retrieved from Flexera: https://resources.flexera.com/web/media/documents/rightscale-2019-state-of-the-cloud-report-from-flexera.pdf

Gillespie, T. A.-F. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, 29-41.

Johansson, E. (2021, July 20). *Amazon actually does something against Pegasus spyware.* Retrieved from Verdict: https://www.verdict.co.uk/amazon-pegasus-nso/#:~:text=Amazon%20has%20shut%20down%20cloud,journalists%20and%20human%20rights%20activists.&text=AWS%20unplugged%20its%20services%20to,wires%20on%20Monday%2C%20Vice%20reports.

Joseph Redmon, S. D. (2016). You Only Look Once: Unified, Real-Time Object Detection . *arXiv*.

Lallemand, N. C. (2012). Reducing Waste in Health Care. *Health Affairs Health Policy Brief*.

McGain, F. M., White, S. F., Mossenson, S. M., Kayak, E. B., & Story, D. M. (2012). A Survey of Anesthesiologists' Views of Operating Room Recycling. *Anesthesia & Analgesia*, 1049-1054.

Periop Green. (2021). *Periop Green Mission Page*. Retrieved from PeriopGreen.com: https://www.periopgreen.com/about/mission

Peter G. Peterson Foundation. (2020). *WHY ARE AMERICANS PAYING MORE FOR HEALTHCARE?* . Peter G. Peterson Foundation.

Petre, M. B. (2019). A national survey on attitudes and barriers on recycling and environmental sustainability efforts among Canadian anesthesiologists: an opportunity for knowledge translation. *Canadian Journal of Anesthesia*, 272-286.

Raj Bala, B. G. (2021, July 27). *Magic Quadrant for Cloud Infrastructure and Platform Services.* Retrieved from Gartner: https://www.gartner.com/doc/reprints?id=1-271OE4VR&ct=210802&st=sb

Vengattil, E. C. (2021, January 21). *Reuters*. Retrieved from reuters.com: https://www.reuters.com/business/media-telecom/what-is-parler-why-has-it-been-pulled-offline-2021-01-12/