Formation of the Cloud: History, Metaphor, and Materiality

Trevor D Croker

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Science and Technology Studies

Janet Abbate, Chair
Daniel Breslau
Saul Halfon
Richard Hirsh

November 14, 2019
Blacksburg, Virginia

Keywords: cloud computing, material culture, infrastructure, internet studies

Formation of the Cloud: History, Metaphor, and Materiality

Trevor D Croker

Abstract

In this dissertation, I look at the history of cloud computing to demonstrate the entanglement of history, metaphor, and materiality. In telling this story, I argue that metaphors play a powerful role in how we imagine, construct, and maintain our technological futures. The cloud, as a metaphor in computing, works to simplify complexities in distributed networking infrastructures. The language and imagery of the cloud has been used as a tool that helps cloud providers shift public focus away from potentially important regulatory, environmental, and social questions while constructing a new computing marketplace. To address these topics, I contextualize the history of the cloud by looking back at the stories of utility computing (1960s-70s) and ubiquitous computing (1980s-1990s). These visions provide an alternative narrative about the design and regulation of new technological systems.

Drawing upon these older metaphors of computing, I describe the early history of the cloud (1990-2008) in order to explore how this new vision of computing was imagined. I suggest that the metaphor of the cloud was not a historical inevitability. Rather, I argue that the social-construction of metaphors in computing can play a significant role in how the public thinks about, develops, and uses new technologies. In this research, I explore how the metaphor of the cloud underplays the impact of emerging large-scale computing infrastructures while at the same time slowly transforming traditional ownership-models in digital communications.

Throughout the dissertation, I focus on the role of materiality in shaping digital technologies. I look at how the development of the cloud is tied to the establishment of cloud data centers and the deployment of global submarine data cables. Furthermore, I look at the materiality of the cloud by examining its impact on a local community (Los Angeles, CA). Throughout this research, I argue that the metaphor of the cloud often hides deeper socio-technical complexities. Both the materials and metaphor of the cloud work to make the system invisible. By looking at the material impact of the cloud, I demonstrate how these larger economic, social, and political realities are entangled in the story and metaphor of the cloud.

Formation of the Cloud: History, Metaphor, and Materiality

Trevor D Croker

General Audience Abstract

This dissertation tells the story of cloud computing by looking at the history of the cloud and then discussing the social and political implications of this history. I start by arguing that the cloud is connected to earlier visions of computing (specifically, utility computing and ubiquitous computing). By referencing these older histories, I argue that much of what we currently understand as cloud computing is actually connected to earlier debates and efforts to shape a computing future. Using the history of computing, I demonstrate the role that metaphor plays in the development of a technology.

Using these earlier histories, I explain how cloud computing was coined in the 1990s and eventually became a dominant vision of computing in the late 2000s. Much of the research addresses how the metaphor of the cloud is used, the initial reaction to the idea of the cloud, and how the creation of the cloud did (or did not) borrow from older visions of computing. This research looks at which people use the cloud, how the cloud is marketed to different groups, and the challenges of conceptualizing this new distributed computing network.

This dissertation gives particular weight to the materiality of the cloud. My research focuses on the cloud's impact on data centers and submarine communication data cables. Additionally, I look at the impact of the cloud on a local community (Los Angeles, CA). Throughout this research, I argue that the metaphor of the cloud often hides deeper complexities. By looking at the material impact of the cloud, I demonstrate how larger economic, social, and political realities are entangled in the story and metaphor of the cloud.

Dedication

For Berenice Escalera. Thank you for walking by my side on this journey.

Acknowledgements

I owe many thanks to all of the people that offered their support, time, and energy as I completed this project. To all of my colleagues at the Department of Science, Technology, and Society at Virginia Tech, I am deeply grateful for the welcoming environment that you fostered and maintained. To my fellow classmates and professors, my education is richer because of your kind spirts.

I wish to send a special thank you to my committee that helped foster my research and sharpen my arguments. In particular, my advisor Janet Abbate was a true supporter of my project and provided invaluable feedback and insight into the development of the dissertation and my own intellectual development. My committee – Daniel Breslau, Saul Halfon, Richard Hirsh – were not only a joy to work with on this project but were advocates throughout my entire graduate career.

Finally, I want to sincerely thank my family for their support. Your love means everything.

Table of Contents

Introduction

The Formation of a Cloud

The Formation of the Cloud is a story about the creation of a new arrangement of computing technologies shaped upon much older visions of computing. This arrangement is called “cloud computing” and is currently the primary growth-area in contemporary networked computing. “The cloud” is a defining feature of the modern web and is the backbone for much of our digital infrastructure. In the majority of public accounts, the story of the cloud is framed as a new type of computing introduced in August 2006. Since 2006, nearly all of the web tech giants have shifted their platforms, users, and business models towards the development and growth of this new vision of computing. The development of the cloud, as presented, reads as the natural evolution of a new technology.

This dissertation attempts to disrupt that deterministic narrative by exploring the history of cloud computing. Contrary to the dominant perception, the idea of the cloud was not invented in 2006. Rather, the notion of the cloud was formed over many decades from multiple visions of computing. These visions of computing are multiple, varied, and often at odds with one another. Two visions of computing in particular, utility computing and ubiquitous computing, have deeply shaped how the cloud is envisioned. Understanding the cloud today requires that we delve into these older histories of computing to see the roots of the cloud’s story.

In addition to the historical unearthing of the cloud, this dissertation also seeks to unpack the cloud by looking at the role of metaphor. The metaphor of the cloud has been, and continues to be, a persistent literary tool in maintaining control over these new technologies. Metaphors are not simply decoration; they are ways of structuring meaning and giving order to systems. Metaphors can hide as much as they reveal. A critical account of the cloud needs to challenge the ease with which we have adopted the metaphor of the cloud. Likewise, the metaphor of the cloud needs to be set against the history of the idea. Often the adoption of a metaphor is uneven; parts of a technology’s history are ignored or overlooked. By reintroducing previous metaphors of computing into our current discussion of the cloud, we can start to reconcile historical imaginings with contemporary frameworks.

The following chapters attempt to pull apart the cloud by looking at the material underpinnings of this new technological system. Underneath the metaphor are material bits that have been spread across the globe. The cloud lives inside data centers, in the cables that deliver information under our feet, and in the devices we carry. These are not immaterial objects; they are often massive infrastructure projects which carry a large amount of technological momentum. Slowly, new networked technologies are becoming enmeshed in, and dependent upon, this new material arrangement. This arrangement raises new social, political, and economic questions as broader society becomes more dependent upon this new era of the cloud. Any attempt to question the values embedded in these systems requires us to look at the actual materials of the cloud.

Each story of the cloud is placed within the broader framework of Science and Technology Studies (STS). STS has a complex history, but one of the prevailing themes concerns the interplay between ideas and materials. The birth of the discipline was rooted in an attempt to place scientific ideology alongside the actual practice of science. As the field expanded to include "technology," in large part due to the influence of historians, the drive to understand the lived experience of socio-technical systems deepened. While STS has a multitude of theoretical frameworks and methodological approaches, there has been a repeated effort to understand ideology in the context of material life. Modern scholarship has eschewed any attempt to create a pure social constructionism, instead embracing the messy and tangible world of actor networks, tacit knowledge, and embodied politics.

This dissertation follows this disciplinary trend, with a particular focus on the role of materiality in digital networks. Research on the internet, in general, has been slow to align digital worlds alongside lived worlds. As the novelty of these digital third-spaces has dwindled with the ubiquity of networked computing, the sociopolitical consequences of networked societies are becoming more obvious. The story of the cloud demonstrates the value of STS research in tackling problems that cannot be fixed to a single research domain. The importance of this approach is that it can follow both the ideology of the cloud and the fiber optic tubes that link it together.

This research could help other related disciplines (such as internet studies, material culture, and the history of computing) see the value of an interdisciplinary approach. For our own field, this dissertation contributes to the tradition of research that unpacks black boxes to reveal the politics of seemingly ordinary scientific and technological objects. Perhaps more importantly, this story of the cloud argues that time, place, and things matter in the construction of an idea. Computing visions do not emerge from the ether. There are people and places that shape these ideas. Metaphors and infrastructures are entwined and are of the same thread. If we ignore this connection, we do both the history of the cloud and the field of STS a disservice.

Contextualizing Cloud Computing

Any interdisciplinary research project needs to set the stage for readers unfamiliar with the subject matter. Before outlining the specific chapters, it is necessary to provide some background information related to computing and the internet. Many of the specific details related to the cloud will be discussed later on, particularly as it relates to data centers and broader internet infrastructures. Additionally, many of the definitions in this introduction will be problematized later when they are presented with more historical context. First, let us start with the birth of computing and the rise of the internet.

Academic historians have documented the history of the computer well.[1,2,3] What started early on as a non-electrical mechanical apparatus[4] (and for a time was biological[5]), morphed into the digital transistor version that we are familiar with today. The digital computer's role has

[1] Campbell-Kelly, Martin, William Aspray, Nathan Ensmenger, and Jeffrey Yost. *Computer: A History of the Information Machine*. Boulder, CO: Westview Press, 2013.
[2] Ceruzzi, Paul E. *A History of Modern Computing*. Cambridge, MA: MIT Press, 2003.
[3] Lee, J.A.N. *Computer Pioneers*. Los Alamitos, CA: IEEE Computer Society Press, 1995.
[4] Standage, Tom. *The Victorian Internet*. London: Phoenix Press, 1999.
[5] Light, Jennifer. "When Computers Were Women." *Technology and Culture* 40, no. 3 (1999).

completely changed from its origins as a tool for specialized calculations in military projectiles or business accounting needs. Today, modern life is dependent upon computers for the maintenance of nearly all complex systems. The advent of mobile computing in the late 1990s and early 2000s has shifted our notion of what a computer is, but the influence of these technologies has only strengthened the role of the computer in everyday life. The past decade has also seen the growth of smart-environments (interconnected appliances and first attempts at digitally connected cities), which expands the reach of computers into our lives.

Historians have also well-documented the rise of computer networks in the 1960s, starting with early packet-switching networks and the creation of the Advanced Research Projects Agency Network (ARPANET).[6] The creation of ARPANET was the first application of TCP/IP (Transmission Control Protocol / Internet Protocol) in a packet switching network. Packet switching provides a means of delivering information over digital networks, and TCP/IP is a protocol that enables a rule-set for the delivery of that information. These two technologies provided the groundwork for the development of early networked computing and are still the bedrock of the modern internet. Many additional protocols and standards have been introduced, eventually culminating in the creation of the internet and the World Wide Web in the late 1980s.

Alongside the development of the computer and networking industries has been the creation of a new global telecommunications infrastructure. Telecommunications companies and investment groups have spent a large amount of capital and time building globally connected networks on top of legacy telegraph and telephone equipment. Computer networks are primarily connected through long-haul fiber-optic cables that are typically buried underground and in submarine cables under bodies of water. Although wireless and satellite connections play an important role in mobile communication, most data still relies upon cables that snake the world and tie into communication hubs. This system is known as the “internet backbone.”

The internet is still a decentralized network in the sense that information can flow without needing to jump through specific routes. However, there are still pieces of critical infrastructure that help give direction or provide information. For instance, Internet Exchange Points (IXPs) are physical locations that direct information between clients before reaching its final destination. Likewise, a small subset of computers acts as “root name servers” which translate written URLs (such as www.google.com) into IP addresses (as part of the Domain Name System).[7] Without these root servers, traffic often would be misdirected.

The history of the cloud is connected to the rise of a new type of decentralization. Broadly speaking, the cloud is an extension of the classic computer-server model. In the past, a user would request information from a single server. In order to log into a university library system, for instance, a single web server would host all of the information, and users would be directed to the same computer. As computer networking became more sophisticated in the late 1990s and early 2000s, administrators of servers were able to allocate resources across the internet more dynamically. Rather than simply hosting a website in a single location, that same resource could be mirrored to different servers in different geographical zones. This helped not only reduce the load on a single server, but also reduced the latency a user might experience by placing servers

[6] Abbate, Janet. *Inventing the Internet*. Cambridge, MA: MIT Press, 1999.

[7] A server is a computer that provides a service to a requesting computer, such as receiving e-mail.

closer to users (network latency is the delay between information being sent from one location to another). It also created redundancy, allowing the same information to be saved across space, an important aspect of backing up information in the case of data loss. As internet speeds increased in the early 2000s and the cost of transferring and storing information online became more economical, new possibilities for the creation of a cloud computing market were opened

These decentralizing networking techniques started to mature in the early 2000s, as the internet became more "re-writable" (generally referred to as Web 2.0). This meant that users were able to more actively create and interact with content (often in the form of commenting or user-directed organization). Rather than simply reading information that was uploaded to the web. Content became more malleable. The idea and location of computing also started to shift during this period.

In the 2000s, the idea of cloud computing fully emerged. This period marked a rebirth of computing as a service that would be provided online. Rather than installing an email client on your computer, web-based email programs started to become the norm. Websites, which used to be primarily managed by individuals, are now increasingly delivered by third-party hosts. Instead of storing files on a local device, services emerged to store information remotely. This shift was in part due to the creation of robust information infrastructure created and operated by newly formed web businesses such as Google and Amazon.

The cloud can be seen as the culmination of many technological and economic developments. These include the maturation of dynamic networking technology, large investments in telecommunications equipment during the 1990s, the wide-scale adoption of broadband, massive growth in e-commerce, and a concentration of diverse information technology (IT) resources into data centers. The creation of the cloud was not a new technology, but a set of technologies and norms that were placed under the umbrella of the cloud. Discussions of the cloud are reflections on the broader shifts that have occurred in the development of modern communication systems.

This is in no way to suggest that the idea of the cloud was a historical inevitability. The metaphor of the cloud was created as a marketing effort and was shaped to reflect certain values. The history of computing and the internet, as I have briefly touched on, is the primary way that the cloud has been framed. This standard view is reflected in the dominant definition of the cloud from the National Institute of Standards and Technology (NIST), which defines the cloud 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. This cloud model is composed of five essential characteristics, three service models, and four deployment models."[8]

This dissertation works to undermine the narrative that the cloud is simply a result of new technologies. Instead, each chapter attempts to pick apart the history of the cloud to understand

[8] National Institute of Standards and Technology. "The NIST Definition of Cloud Computing: Special Publication 800-145." September 2011.

the places and people that produced a preamble for the cloud's eventual rise into the public light. The metaphor of the cloud was produced from early ideas and attempts at building a new computer utility or a ubiquitous computing system. In each of these cases, there were certain values and politics at play. These histories were often drawn upon when justifying the creation of the cloud, but without fully adopting the politics or ideas of the past. Revisiting histories can help the public, researchers, politicians, and the technologists within cloud companies shape the cloud more intentionally.

Looking at the cloud in the context of this history can help the general public grapple with choices regarding which cloud technologies to adopt or reject. Perhaps more importantly, understanding this story can help individuals make technology choices that align with their own values. The easiest choice for most users is to conform to the dominant consumer cloud ecosystem, which typically extracts money or personal data in exchange for access to the cloud. However, with a little bit of effort, users can assert more ownership of their own cloud (or reject the paradigm entirely). These decisions are not without consequences, but many still do not know there are alternative frameworks for understanding the cloud. The history of the cloud may help empower some to take these actions for themselves and potentially push for broader regulatory change.

In addition to this historical approach, I argue that the best way to start opening up the cloud is by looking at its physical impact. This requires looking at the places and organizations that run the cloud. Seeing the footprint of the cloud helps expose what is otherwise invisible. Looking at a data center, for instance, helps bring to mind questions of energy usage, data security, and the integration of these systems into diverse cultures.

Core Terminology

The underpinnings of the cloud are a complicated mixture of computer hardware, software, and human actors. The average user of the internet likely does not know, or have the desire to learn, the specific details of computer networking. Understanding these details, however, is necessary when attempting to dissect the history of the cloud. This dissertation presumes the reader has a basic understanding of what a computer is and a general knowledge of computer networking. There are particular ideas of computer networking which need to be outlined prior to the broader discussion of the cloud.

To start with, there is the issue of attempting to provide the most common definition of the cloud. This definition will be critiqued throughout the chapters, but for now the definition of the cloud will be discussed for the purposes of giving a general idea of what the cloud is. In the most general sense, and for the majority of users, the cloud is simply a way of storing, accessing, and processing information over the internet. The cloud is sometimes flippantly referred to as simply someone else's computer. While this remark is an oversimplification of a complex set of computing infrastructures, the root of the idea rings true. For the majority of users, the cloud is simply a place that exists in the ether of cyberspace.

Most people knowingly access the cloud when they use consumer-facing services that transmit or store personal information. Apple, for instance, offers its iCloud service as a means of remotely storing a user's photos, documents, music, and other files. Similar services, such as Google's Drive or Microsoft's OneDrive, provide similar features and are generally recognized as being a cloud technology. Typically these consumer products are accessed using a software interface, and the user is unaware of where their information is stored (commonly mirrored to multiple data centers) and has no control over the software or hardware that is used.

These consumer-facing cloud products are only one variant of the cloud. Because the cloud encompasses a number of networking technologies, there are many different applications for cloud services. Businesses and other organizations might use the cloud as a supplement or replacement for their own information infrastructure. For instance, the video streaming service Netflix uses servers owned by Amazon Web Services to manage user preferences and host the platform's front-end (the actual videos are delivered at the ISP level). Rather than purchasing and housing the computing hardware themselves, the cloud helps organizations rent the compute power from a cloud provider (the term "compute" is often used in data center discussions regarding the number of computing resources being used across multiple systems). Large cloud providers can provide a wide geographic distribution of cloud server farms, which allows organizations to scale their operations more quickly.

In the previous section, cloud computing was defined by using the NIST definition, which is more specific and has a more technical audience in mind. The definition focuses on a few core concepts, including: on-demand network access, broad network access, resource pooling, rapid measured service, three service models, and five deployment models.[9] This definition will be challenged later, but for the purposes of clarity, each idea will be examined.

To start, the idea of "on-demand network access" refers to the ability to adjust the number of computing resources at will and independent of human intervention. For instance, if a website is receiving a large amount of traffic, a cloud web hosting application could dynamically allocate more compute to serving the content on the website. Likewise, if a user wants to upload a large video file to a cloud storage application, the system will be able to find a storage location that has sufficient space intelligently.

Another important concept is the notion of "broad network access." This refers to a model of access that is independent of the particular device requesting the resources. This means that in most instances, a person is able to access the cloud regardless of the operating system that they are using or type of device that they are accessing the cloud on (desktop, notebook, phone, etc.). This differs from a computing system that is tied into a particular hardware solution or network location.

The cloud also is defined by the notion of "resource pooling" and "rapid elasticity." Resource pooling is analogous to other shared infrastructure projects. Rather than giving each user a separate silo, multiple cloud tenants share computing resources together. If, for instance, you upload a photo to Apple's iCloud, that photo may be saved on the same storage device as another user's photos (although you will not be able to access their photos, let alone know specifically which storage device your photo is stored on). This is a more dynamic form of resource sharing

[9] National Institute of Standards and Technology. "The NIST Definition of Cloud Computing: Special Publication 800-145." September 2011.

that is reliant on a level of software abstraction that separates the base infrastructure (the physical computers and wires in a data center) from the software layer that users access on their own devices (such as an online photo viewer). Likewise, the idea of "rapid elasticity" is the notion that computer resources can be allocated dynamically. Bandwidth, storage, and compute power can be ramped up at a moment's notice.

One of the final essential characteristics of the cloud is that it is a measured service. Like a power system, many cloud systems bill users depending upon the amount of resources they consume, as well as a base fee for access to the service. Typically cloud providers at the infrastructural level bill by the amount of storage, computing power, and bandwidth used during a billing period. For general consumer cloud applications (like web email), usage is either not billed or is billed at different tiers. For instance, Google Drive provides a free tier for online storage but charges for additional space.

These characteristics provide a rough framework of what the cloud is. Typically the cloud is also defined by three different service models and four deployment models. The service models are Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). Each of these models represents a level of abstraction away from the core infrastructure of the cloud. A cloud SaaS product, like a web email client, gives user's access to a software product without allowing them to see the underlying operating system powering the system or the hardware that makes the software function. PaaS products provide the user some access to an operating system, without deep access to the computer's hardware. For instance, an application developer may use a PaaS to host a virtual machine (an emulated computer) to test a program, without having access to the computer that is emulating the machine. Finally, IaaS cloud products provide the deepest access to the bare hardware. An IaaS product allows the user to buy access to computing hardware or bandwidth. In this model, the user has more control over the computing environment they use (such as the type of operating system they use or the speed of the computer they rent access to). Typically, IaaS customers don't have access to the physical building in which their rented resources are housed, but some cloud providers do allow on-site access.

The idea of access ties into the deployment models of the cloud: private, community, public, and hybrid. A public cloud is one that is generally able to be accessed by the broad public. Online storage solutions, like Dropbox or Google Drive, are examples of public clouds because they are open to all, and the computing resources are managed by a single entity. On the opposite end of the spectrum are private clouds. These are computing environments that rely on cloud technologies, but are limited to a particular group of people and not intended for public access. For instance, a business may have an internal cloud for managing in-house documents or communication services.

Between public and private clouds, there are community clouds and hybrid clouds. A community cloud is one that is in some way more limited than a public system. For instance, a university may host a cloud service on a public cloud, but place usage restrictions on who can access the resources (i.e., the university community) and where those resources can be accessed (i.e., only on campus). Finally, a hybrid cloud is one that combines some form of these other deployment models. For instance, a university may have internal resources for the students of a university (a community cloud), but also host resources for the broader public through a general website (a public cloud).

These characteristics and deployment models for the cloud are the dominant way that the cloud is discussed amongst information technologists. Some definitions may emphasize the cloud as a piece of shared infrastructure, whereas other definitions may focus on the role of digital services. However, these definitions rarely look at the role of the cloud as metaphor or the specific material arrangement of the cloud as it relates to the real world. Furthermore, these defining characteristics can be tied to specific computing histories that inform our technological present. The bulk of this dissertation is an attempt to break apart the dominant definition of the cloud by looking at the metaphors, histories, and networks embedded in the story of the cloud.

Chapter Organization

The following chapters are organized to tell the story of the cloud by moving from metaphor to material. The first three chapters speak to the history of the cloud, whereas the final two chapters speak to more contemporary issues using a material culture framework. Each chapter attempts to tell the story of the cloud not as evolutionary, but as a struggle over meaning and control. In doing so, I will unpack the story of the cloud by looking at history, metaphors, and material infrastructures.

The first chapter focuses on the 1960s and 1970s with the creation of utility computing. The vision of utility computing was a primarily academic one. The chapter looks at the concept of computing as a utility by telling the story of two early time-sharing projects (one public and the other private). The rise of time-sharing is discussed, along with the abandonment of utility computing as the Federal Communication Commission started to regulate these new hybrid telecommunication networks. The history of utility computing provides an alternative vision of how we might organize the cloud today.

In the second chapter, visions of ubiquitous computing are discussed. The chapter focuses on a short period of time in the late 1980s to the early 1990s when the concept was coined at Xerox's research firm: PARC. "Ubicomp" is framed against visions of the cloud by looking at how ubiquitous computing was shaped around particular values. In particular, the philosophies of calmness and invisibility prove to be central to the politics of this new vision of technology.

The following chapter looks at the iconology and coinage of the cloud. The chapter starts by discussing the iconography of the cloud and early instances of the cloud as a visual computing symbol. The chapter then moves to the coinage of the term cloud computing and the atrophy of the term in the late 90s. This discussion is followed by the rebirth of the cloud in 2006 and the maturation of the idea in later years. Throughout this chapter, particular attention is paid to the metaphor of the cloud as a marketing tool and the role of symbols in shaping the development of a technology.

In the penultimate chapter, the focus is on the materials of the cloud. Having given the historical context to the cloud's metaphors, this chapter looks to move beyond the language of the cloud to give context to the physical infrastructure that allows the cloud to operate. Particular attention is given to submarine cable networks and the critical role that they play in the expansion of the cloud as a global network. The economic and political implications of the cloud are set against the popular vision of the cloud as a network without a significant material footprint.

The final chapter focuses on cloud computing inside, and nearby, the city of Los Angeles. I visit cable landing locations and look at clusters of data centers to examine the material impact of the cloud. Additionally, I create a mapping of the cloud of Los Angeles and link the cloud to regional histories, economies, and visions of what it means to have a local cloud. The chapter also serves to underscore the material realities of the cloud and to argue that any attempt to regulate the cloud requires that we look at our local environments to understand how metaphor and materials align.

Chapter 1

Utility Computing

Introduction

The cloud should be considered, above all else, an infrastructure system that exists both as a physical and imagined network of computing hardware and software. Like any other infrastructure, the cloud is not simply a collection of parts but emerges as a result of many sociotechnical artifacts interacting with one another. The electrical system is not just the wires, generators, and power stations. It is also the people and the social processes that construct and maintain the network itself. Throughout this chapter and the remainder of the dissertation, I argue that the cloud is inseparable from the earlier computing culture and innovations that helped provide a foundation for the eventual birth of the cloud in the 21st century.

Rather than discuss the history of the cloud in terms of specific technological advancements, such as the creation of a faster computer or a new network protocol, I have opted to consider broader trends in computing that have had a significant influence on the development of the cloud. I borrow from the existing computing history categories to situate the development of the cloud alongside other technological changes.

One of these major shifts was the notion of "utility computing." Utility computing, as a concept, was popularized in 1961 by computer scientist John McCarthy. However, the underlying principles of utility computing stretch back to some of the first mainframe computers. Utility computing, used in the broadest sense, refers to computing that is delivered as an infrastructure resource that can be consumed on-demand. Users accessing the utility system are, typically, charged only for the resources that they consume, rather than needing to purchase and maintain the physical computer itself directly. The metaphor of utility played a significant role in the development of computing in the 1960s. It not only suggested a more efficient means of computing, but it also aligned it with other utilities (such as telephone or electrical networks). The metaphor invited all of the complications that plague any complex sociotechnical system.

Utility computing was inspired by limitations in traditional computing, which limited who could access the computer and for how long. Utility computing emerged first from creative solutions to these limitations. In particular, the creation of "time-sharing" systems most directly inspired the use of the term utility computing. Time-sharing allowed the use of a single computer by multiple users at the same time. The push for, and creation of, utility computing through time-sharing signaled a shift towards broadening the use of computers by multiple publics.

This first chapter looks at the origins, imagining, and implications of utility computing during the 1960s. This period was marked by a transition towards a new relationship with the computer. Mainframe computers started to be used as the backbone for utility computing services. Utility computing provided a vision of computing that encouraged purchasing computer time, rather than ownership of hardware. Public institutions such as the Massachusetts Institute of Technology (MIT) and Dartmouth College led the way in early time-sharing efforts. Private companies followed suit and expanded the reach and impact of utility computing. The contributions of these actors will be discussed throughout this chapter.

Following these actors provides us with a historical backdrop for the creation of cloud computing. The creation of the cloud heavily borrowed upon the framework of utility computing and time-sharing. However, particular ideological and regulatory concerns were present during the 1960s and 1970s that did not emerge during the creation of the cloud. By framing the changes in computing around the notion of "utility," a slew of political issues were raised. Utility computing invites comparisons to public utilities, monopolies, ownership claims, and social commitments. These same issues could be raised for cloud computing, but the metaphor of the cloud strongly influences the type of questions that are asked, deflecting concerns about the public interest. Paying attention to actors' use of metaphors, and therefore subtle ideological commitments, within private and public institutions can reveal the politics behind the creation of early computing network assemblages.

After discussing the transition to utility computing, I will set these advances in computing against the regulatory challenges that disrupted the spread of utility computing. I focus on the FCC's struggle to regulate newly emerging computing services, particularly in the difficulty of regulating hybrid networks.

The bulk of this chapter argues for a reexamination of early utility computing efforts. The history of utility computing is largely a story in which multiple actors envisioned a computer utility, but the visions were never fully put into practice. Early researchers, particularly in academic circles, were largely optimistic about the possibilities of constructing, maintaining, and regulating the computer as a utility resource. Early on, time-sharing seemed to be the technology that would elevate computing to the level of a public utility, similar to electricity or the telephone network. However, as time-sharing became commoditized, there was less desire to frame time-sharing around the idea of a utility. Regulatory pressures, a move towards networked computing, and the rise of the personal computer discouraged encourage further promotion of the computer utility. It was not until the rise of cloud computing that these debates reemerged. Current debates, however, lack the historical context that the history of utility computing can offer.

In the current moment, the idea of a computer utility may seem somewhat antiquated. The spread of decentralized networks makes the comparison to public utilities seem quaint. However, I argue that we should take the notion of utility seriously. The rise of the cloud is the reemergence of a new type of utility and marks a re-centralization of economic power. The history of time-sharing demonstrates the difficulty of taking our relationship to digital infrastructures seriously. That said, if we do pay attention to the imaging and construction of these systems, we can see the semi-opaque process by which political values are interwoven into the code and hardware of these infrastructures.

Infrastructure and Utility

The internet is not a single infrastructure but multiple overlapping infrastructures. This entanglement has made regulating internet access particularly challenging. In the United States, the regulation of the internet is in flux. For much of President Obama's term, the FCC implemented a number of regulations that encouraged net neutrality (non-discriminatory delivery of the internet by internet service providers [ISPs]) and attempted to treat the internet as a utility. [10] [11] President Trump's FCC chairman Ajit Pai has started to challenge and roll-back these previous regulatory goals. Recent mergers have consolidated the power of the ISPs and have left some regions with virtual monopolies. Additionally, ISPs are increasingly owners of media distribution companies, owning not just the pipes but much of the content that flows through them. Despite the move towards less governmental oversight, globally, many countries have adopted stances that frame the internet as a utility. This is also seen in a number of internet governance organizations and the United Nations Human Rights Council that have condemned the restriction of internet access as a violation of human rights law.[12] Amidst this political landscape, there is an indication that digital networks are being increasingly regulated as utilities and infrastructures. Therefore, there is a great deal at stake for how we categorize and define what we mean by utility and infrastructure.

The idea of utility is multidimensional and situated within a broader context. In the most straightforward sense, the term utility refers to usefulness. For instance, hand tools provide a utilitarian function. This is decidedly different than an idea of a larger public/private utility. When we speak about utility in the broad sense, we are invoking issues of regulation, law, and social values. Both the tool and the electrical system provide a utilitarian purpose, but they serve that function at a different scale. In the case of utility computing, the system has been understood in both senses, as simply a tool and also imagined as a larger utility system.

Utilities exist in the context of a broader history over how we ought to manage resources and what role type of ownership model the public should play in managing those assets. Many of the most contested political arenas are debates over how public utilities should be managed. Contemporary regulation theory has been largely informed through debates over electrical, gas, and water ownership.[13] STS researchers and historians of technology, for their part, have used these domains to explore the role of expertise and the development of large technical systems. Importantly, these researchers have looked at how the public reaches a consensus about how the utility ought to be regulated.[14] As these authors have pointed out, in the case of novel utilities (such as the cloud), public acceptance of a new system is often not an issue of functioning hardware but of "non-technical barriers."[15] These barriers are primarily social norms that can be difficult to address from a regulatory standpoint. Generally, this scholarship has focused on the contested process of building and maintaining a utility system.[16]

[10] The White House. "Executive Order -- Improving Critical Infrastructure Cybersecurity." February 12, 2013. https://web.archive.org/web/20130403003414/https://www.whitehouse.gov/the-press-office/2013/02/12/executive-order-improving-critical-infrastructure-cybersecurity (Archived April 3, 2013).

[11] Kang, Cecilia. "Court Backs Rules Treating Internet as Utility, Not Luxury." *New York Times*. June 14, 2016.

[12] Blake, Andrew. "UN Human Rights Council 'Unequivocally Condemns' Internet Shutdowns." *The Washington Times*. July 1, 2016.

[13] Priest, George L. "The Origins of Utility Regulation and the 'Theories of Regulation' Debate." *The Journal of Law & Economics* 36, no. 1 (1993): 295.

[14] Hirsh, Richard. *Power Loss: The Origins of Deregulation and Restructuring in the American Electrical Utility System.* Cambridge, MA: MIT Press, 2001.

[15] Hirsh, Richard. "Historians of Technology in the Real World: Reflections on the Pursuit of Policy-Oriented History." *Technology and Culture* 51, no 1 (2011): 20.

[16] Slayton, Rebecca and Aaron Clark-Ginsberg. "Beyond Regulatory Capture; Coproducing expertise for critical infrastructure protection." *Regulation and Governance* 12, no 1 (2017). 1.

Throughout this dissertation, I will be using the terms infrastructure and utility. There are a number of definitional issues and assumptions that need to be laid out prior to the case study. Generally, utilities can be thought of as a type of infrastructural resource. Public utilities typically include electricity, water, gas, sewage, and communications. Infrastructure overlaps these categories but extends itself to all domains. Even within STS, many scholars have adopted the notion of infrastructure as a means of connecting the broad research domains and topics under a single umbrella.[17] For the purposes of this paper, the term "utility" will be treated as a subcategory of infrastructure. By allowing infrastructure to stand in for utility, there is a greater degree of interpretative flexibility allowed. This is helpful in creating a linkage between utility computing and the cloud.

It is important to recognize that infrastructures are not static. Case studies, such as Paul Edward's research on computer modeling of climate data, demonstrate the complicated linkage between scientific practices, theories, and collection of data.[18] As systems change, humans often attempt to rework or work around the existing system to meet their needs. Edwards and others use the concept of "infrastructure inversion" to refer to the reworking of data in the face of infrastructural change. Attempts to make information homogeneous, such as measuring changes in global weather, often occur in the face of much infrastructure noise and competing values (such as a desire to capture short-term weather changes). The history of utility and cloud computing are subject to change, and, therefore, it is critical that these forms of computing are not read as being static.

My definition of infrastructure is borrowed from law professor Brett M. Frischmann's work on the term. According to Frischmann, infrastructural resources can be defined as meeting the following criteria:

> (1) The resource may be consumed nonrivalrously for some appreciable range of demand.
> (2) Social demand for the resource is driven primarily by downstream productive activities that require the resource as an input
> (3) The resource may be used as an input into a wide range of goods and services, which may include private goods, public goods, and social goods.[19]

Utility and cloud computing fit within this definition of infrastructure. This can be seen in each of the three points.

The first point argues that the resource can be consumed nonrivalrously. A rivalrous resource is one that is a private good that is "finite, nonrenewable, and not sharable," such as an apple.[20] A public good, such as an idea, is infinite, sharable, and non-

[17] 4S Online. "STS Infrastructures: Making and Doing 2015." 4s Online. https://web.archive.org/web/20171104102646/http://www.4sonline.org/md/track/category/sts_infrastructures (Archived November 4, 2017).

[18] Edwards, Paul. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press, 2013.

[19] Frischmann, Brett M. *Infrastructure: The Value of Shared Resources*. New York, NY: Oxford Press, 2012. 61

[20] Frischmann, Brett M. *Infrastructure: The Value of Shared Resources*. New York, NY: Oxford Press, 2012. 62.

congestible. In the middle are "impure public goods." The internet, cloud computing, and utility computing are all impure public goods. They are "impure" because while they are sharable, these resources are finite, congestible, and depreciable. Despite these limitations, these impure public goods are still considered infrastructural resources. The second point speaks to downstream benefits of the resource. The social benefit of the cloud is not the use of the cloud itself, but in the generation of positive externalities. This helps explain the marketing of the cloud as a tool to accomplish end-goals, a resource to build upon. Finally, the third point speaks to the flexibility of infrastructure. This marks infrastructures as general-purpose inputs with a "range of outputs [that] may span private, public, and social goods."[21] Many utility computing systems, as will be described in this chapter, focused on providing general-purpose computing to benefit commercial, public, and social infrastructures.

This definition of infrastructure is important because it makes clear that infrastructure is not simply large public projects. It is a flexible term that encompasses many different domains. This dissertation is examining a particular domain: telecommunications infrastructure. Telecommunications comes with a long history of commons management debates in the form of: "common carrier requirements, antitrust laws, and open-access regulatory regimes."[22] These issues will be hinted at in this chapter and will be discussed in more detail in later chapters. The centrality of infrastructure will be a common theme throughout the dissertation.

Imagining Utility Computing

Time-sharing's history is not just measured by the technological advances made in computing but through the ties to the expanding ideology of the computer as a utility. The idea of utility computing is directly linked to the earliest experiments in time-sharing. Initially, the development of time-sharing and utility computing were tightly linked in a process of co-production.[23] At first, utility computing was discussed only in terms of time-sharing. As time-sharing matured, the theory of utility computing also changed and moved beyond the realm of time-sharing. Utility computing's impact on time-sharing is harder to measure, but can be seen in the way that time-sharing systems were constructed and use of utility as a metaphor.

Metaphors give shape to the invention, construction, and maintenance of new technologies. They are, as Lakoff and Johnsen argued, concepts we live by.[24] The history of technology has long recognized that the ideological framing of a new technology matters. The deployment of specific metaphors or associations can strongly influence the ideology surrounding a specific technology. They also imbue the technology with a political frame. These frames are linked to different ideologies and subject to change. This was certainly the case of time-sharing. The linkage of time-sharing to the notion of "utility" can be credited to the MIT researchers who introduced the notion of time-sharing. It is important, therefore, that we examine how the meaning of utility shifted as time-sharing developed.

[21] Frischmann, Brett M. *Infrastructure: The Value of Shared Resources*. New York, NY: Oxford Press, 2012. 64.
[22] Frischmann, Brett M. *Infrastructure: The Value of Shared Resources*. New York, NY: Oxford Press, 2012. 217.
[23] Jasanoff, Sheila. *States of Knowledge: The Co-Production of Science and Social Order*. London: Routledge, 2010.
[24] Lakoff, George and Mark Johnsen. *Metaphors We Live By*. London: University of Chicago Press, 2003.

Utility can refer to a number of different ideological frameworks and is a type of metaphor itself. There is the most self-evident meaning: to be practical and functional. Certainly, this notion of utility held sway with the early engineers of time-sharing. MIT's Robert Fano and Victor Vyssotsky (two early time-sharing researchers) wrote that the purpose of time-sharing was the management of chaos. Explaining that "instead of chaos…each user enjoys the benefit of efficiency without having to average the demands of his own particular program." [25] Time-sharing would provide a "utility-like view." [26] This use of the utility metaphor is what Information Science researcher Maria Lindh calls the "sense-marking" period in utility computing history. She argues that "in the late 50s and during the 60s, the metaphor's main role was to make sense of the new technology, among IT professionals, economists, and the public, thus revealing its potential."[27] The history of time-sharing reflects this finding, but it rather quickly moved towards what Lindh refers to as a "constitutive" and later "restrictive" phrase.

A more prevalent use of utility in the 1960s was referencing the creation of a "computer utility," analogous to the telephone or electrical system. Evoking the image of a newly forming digital utility system drew upon the imagination of researchers as a new resource for the public(s). This was evident in Computer researcher John McCarthy's first analogy to the telephone system (a claim made in 1961 before time-sharing had truly materialized).[28] Early writings focus very much on the practical nature of infrastructure. For instance, Fano said after developing time-sharing that the goal of MIT's project "was a compute utility" that the user could regard "…as something that was there and reliable."[29] When Project MAC failed one night, user outrage was "the expression of the customer of a public utility."[30] Reliability obviously was a central concern for the builders of early time-sharing systems. A focus on reliability, however, speaks to a deeper issue. One of the central roles of infrastructure is to provide a backbone for the production of other human pursuits. We might substitute the word reliability for invisibility in this instance. Like the move to the cloud, the success of utility was measured (at least by the researchers) by the disappearance of the computer itself.

However, invisibility was not the driving force for the ideological concerns of utility computing. As time-sharing became more of a reality, connections were drawn to the economic, ethical, and policy implications of utility. In a 1966 piece for *Scientific American*, Fano and Corbato wrote:

> Looking into the future, we can foresee that computer utilities are likely to play an increasingly large part in human affairs. Communities will design systems to perform various functions – intellectual, economic, and social – and the systems in turn undoubtedly will have profound effects in shaping the patterns of human

[25] Corbato, Fernando J. and Victor A. Vyssotsky. "Introduction and Overview of the Multics System." *Proceedings – Fall Joint Computer Conference, 1965*: 188.

[26] Corbato, Fernando J. and Victor A. Vyssotsky. "Introduction and Overview of the Multics System." *Proceedings – Fall Joint Computer Conference, 1965*: 188.

[27] Lindh, Maria. "As a Utility – Metaphors of Information Technologies." *Human IT* 13, no. 2 (2016): 78.

[28] Garfinkel, Simson. "The Cloud Imperative." *MIT Technology Review*. October 3, 2011.

[29] Garfinkel, Simson and Harold Abelson. *Architects of the Information Society : 35 Years of the Laboratory for Computer Science at MIT*. Cambridge, MA: MIT Press, 1999. 7.

[30] Garfinkel, Simson and Harold Abelson. *Architects of the Information Society : 35 Years of the Laboratory for Computer Science at MIT*. Cambridge, MA: MIT Press, 1999. 9.

> life. The coupling between such a utility and the community it serves is so strong that the community is actually a part of the system itself. Together the computer systems and the human users will create new services, new institutions, a new environment, and new problems...To what ends will the system be devoted, and what safeguards can be designed to prevent its misuse? It is easy to see that the progress of this new technology will raise many social questions as well as technical ones.[31]

Social, not technical, concerns drove discussions about utility computing. There is a sense in their writings that the value of these systems is found in the community of users. By making the computer a utility, you are opening up the computer, not as a single device, but as an ecosystem. Computing ecosystems are not self-regulating. They need to be managed through rules and norms. Those concerned with utility computing turned to these questions as time-sharing systems were being built.

The deployment of metaphors is rooted in a particular moment. In the case of time-sharing, the reference to utility was quickly linked to the ongoing policy debates over telecommunication regulations. This is a point touched on at the end of the chapter, but it is important to note that without the turbulent telecommunication landscape at the time, it is entirely possible that the metaphor of utility may have been less scrutinized. Without this context, the idea of time-sharing could have been linked to entirely private (non-public) utility systems. Instead, critiques of utility computing were framed against public infrastructures. For instance, public planning frameworks started to be discussed in the context of developing "urban-regional computer [utilities]."[32] Additionally, critical voices were quick to adopt the metaphor of utility as a method of attack. A businessman from a small computer company spoke strongly against the notion that these systems were "natural monopolies," largely due to the fear that the metaphor of utility would be a justification for complete control over a new technological system.[33] At the root of his concern was that the comparison to a traditional utility was ill-suited for an information network. All of these discussions are inexorably linked to the politics of the 1950s and 1960s.

Concerns like these, and the change direction of the computer industry, ultimately dampened the early enthusiasm of the metaphor of utility computing 1970s (and consequently the metaphor of utility dropped out of favor). This chapter walks through the early history of time-sharing to demonstrate how the ideology of utility computing was never rooted in a single understanding of utility. There remain unanswered questions about the type of political commitments that these actors had in mind when envisioning early time-sharing systems. What is clear, however, is that metaphor played a hand in directing the material construction and legal framework for future computer systems.

[31] Fano, R. M. and Fernando J. Corbato. "The Time-Sharing of Computers." *Scientific American* 215, no. 3 (September 1966). 128-140.

[32] Jones, R. D. "The Public Planning Information System and the Computer Utility." *Proceedings of the 1967 22nd National Conference, ACM.*

[33] Denz, Ronald F. "The Case Against the Computer Utility." *Proceedings of the 1967 22nd National Conference, ACM.*

Mainframes Before Utility and Time-Sharing

Before electronic computers were commonplace, computing started as something less mechanical and more biological. The computer, what we typically think as an assemblage of mechanical and digital parts, started as an occupation. People were the first computers. In the 19th and early 20th century, there were a number of tools that calculated or solved problems (from adding machines to punch card looms); however, the person working with these tools was considered the computer. Most of these early computers were women (this was especially true during World War II).[34] Even when the computer moved away from signifying an occupation, the computing industry continued to imbue the computer with differing gender norms.[35]

The identity of a computer started to shift with the coming period of electronic computing. In 1946 the first electronic programmable computer was built. The ENIAC was notable, in part, because it introduced the idea of storing programs in high-speed memory, which has become the hallmark of all future computers.[36] Additionally, the creators of the ENIAC went on to build the UNIVAC, a commercial computer that led the way for non-military uses of computing. After World War II, non-governmental actors started to see the value and practical applications of electronic computing.

In the 1950s, a number of developments spurred the adoption of computers. At this time, companies demanded systems that could store and retrieve large sets of data, with the ability to execute simple mathematic calculations.[37] The United States government needed reliable systems that could be used in the Cold War context (seen in their Whirlwind I and SAGE computer). Academic institutions worked alongside commercial and governmental actors. By the late 1950s, computers started to meet those demands. The move away from vacuum tubes to transistors, and the invention of integrated circuits, greatly increased the reliability of computing systems. Additionally, the adoption of "core memory" (nonvolatile storage of information) increased the speed of systems.

Contemporary scholarship labels the majority of computers during this era as "mainframe computers." [38] However, it is worth noting that this is a small anachronism. The term "mainframe" was not used until much later. Early references to the "main frame" of a computer referred to the physical frame that held the "arithmetic and control elements and the high-speed memory."[39] The term was likely borrowed from the telecommunication industry's use of the term to refer to the frame holding switching equipment.[40] It wasn't until the 1980s, with the

[34] Light, Jennifer S. "When Computers Were Women." *Technology and Culture*. 40, no 3. (1999): 455-483.

[35] Abbate, Janet. *Recoding Gender: Women's Changing Participation in Computing*. Cambridge, MA: MIT Press. 2012.

[36] Ceruzzi, Paul E. *Computing: A Concise History*. Cambridge, MA: MIT Press, 2012. 49.

[37] Ceruzzi, Paul E. *A History of Modern Computing*. Cambridge, MA: MIT Press, 2003. 47-48.

[38] Ceruzzi, Paul E. "The Mainframe Era, 1950-1970." *Computing: A Concise History*. Cambridge, MA: MIT Press. 2012. 54-63.

[39] MIT Computation Center. "Input and Output." In *Coding for the MIT-IBM 704 Computer*. Cambridge, MA: The Technology Press, 1957. XI-1

[40] Computer History Museum. "Mainframe Computers." https://web.archive.org/web/20190105034816/http://www.computerhistory.org/revolution/mainframe-computers/7/166 (Archived January 5, 2019).

advent of the personal computer, that these mainframe computers were clearly marked as distinct. Despite this linguistic absence, it is appropriate to refer to these systems as mainframe computers because they were quite different from the computers of the future.

As computing speed and reliability increased, companies and research institutions started to purchase computers. Mainframes, however, were not without their problems and limitations. One of the main drawbacks of these early computers was that they could only run one program at a time for a single user. The process of feeding punched cards into the system meant that not only would access to the computer be limited, but also the computer's resources would be underutilized.[41] The invention of batch processing allowed multiple jobs to be completed in a row, but a data entry mistake could mean hours or days of lost time. These limitations were felt most strongly in the academic context due to the limited resources available (either due to older computers, lack of dedicated programmers, or limitations in faculty and student time). Consequently, research institutions were the first innovators in general-purpose time-sharing.

Origins of Time-Sharing

Time-sharing was invented to overcome the previously mentioned challenges of limited user access. When time-sharing was in its infancy, it could be interpreted as two things: "One can mean using different parts of the hardware at the same time for different tasks, or one can mean several persons making use of the computer at the same time."[42] The second meaning (several users) eventually became the dominant definition. This chapter continues to use that definition.

In short, the creation of time-sharing transformed the practical uses of a computer. Prior to the invention of time-sharing, computer systems were limited in the number of users that could work on a machine at one time. Because these systems were expensive, it was not feasible to purchase multiple computers. Batch processing worked well for accounting problems (such as billing) because the computing jobs were repetitive and predictable. However, non-accounting jobs posed a challenge for batch processing for a number of reasons. Computer programing, in particular, is a trial and error job that is ill-suited to batch processing. For instance, an error on a punch card would likely ruin a computing job, and that lost time was costly. Programmers complained that computers were poorly designed for these tasks. One argued that "with most computers…the circuits for high-speed multiplication are utilized only a small fraction of the time."[43] It was clear to the programmer that a revolutionary advancement needed to be made to make computing more useful. For many, time-sharing was that revolution.

The first instance of rudimentary time-sharing emerged from the U.S. Navy's missile defense project: SAGE (Semi-Automatic Ground Environment). Built in response to the fragmented radar systems of the time, the Navy commissioned a computer system that could receive real-time information from radar systems via the telephone network. Time-sharing was only one of the many advancements that SAGE made (on-line terminals, computer-driven displays, digital

[41] Arms, William Y. "The Early Years of Academic Computing." Internet First University Press. May 2014.

[42] Corbato, Fernado, Marjorie Merwin-Daggett, and Robert C. Daley. "An Experimental Time-Sharing System." *Proc. Spring Joint Computer Conference* 21 (1962).

[43] Bauer, W. F. "Computer Design from the Programmer's Viewpoint." AIEE-ACM-IRE 1958 Conference. December 3-5, 1958. 48.

signal processing, to name a few).[44] [45] While building SAGE, the Navy worked closely with MIT when developing the system. This connection proved to be crucial to the first general-purpose time-sharing computer.

Most histories about time-sharing start with the work at MIT. As an institution, MIT was no stranger to computing technology. In the 1930s, Vannevar Bush and Harold Locke Hazen had invented an analogue computer (called the differential analyzer). During World War II, MIT started to work with digital electronics. The MIT Radiation Laboratory (Radlab) and the Research Laboratory of Electronics (RLE) worked on radar systems during and after the War, respectively. MIT's role in computing deepened as the Cold War's kindling was lit.

The Soviet Union's first atomic bomb test prompted the Department of Defense and the U.S. Air Force to continue research on radar systems. MIT's Lincoln Laboratory was created in response to the requests of the Air Force.[46] MIT's previous work with Radlab and RLE played a major role in obtaining federal funding. MIT's Servomechanism Laboratory helped build Whirlwind I, the predecessor to SAGE. As mentioned, MIT faculty were also involved with the development of SAGE. These early experiences helped shape the future of computing. The mainframe computers used in these experiments started to acquire elements of later time-sharing, most notability the use of remote computers.

MIT continued to innovate in the 1950s. Federal funding and talented faculty had made MIT a powerhouse in early computing. Roughly 80 percent of MIT's operating budget came from sponsored research.[47] In 1955, the Lincoln Laboratory designed the TX-0, the first large completely transistor-based computer. The TX-0 design was heavily borrowed from SAGE. After a failed successor, the TX-1, the laboratory designed the TX-2 in 1958.

However, by the late 1950s, the limitations of these systems were apparent for certain users. Despite increased speed and reliability, only a single job could operate on the computer at a time. This was not as much of an issue for industrial computing. Companies outside of academia had put a lot of energy into developing effective batch processing. Work on time-sharing was considered a financial drain. As Douglas Ross, a programmer at MIT and collaborator on Whirlwind, put it: "Most of the people…said 'That's great; when you've got all that government money at MIT.'[48]

The idea of time-sharing at MIT likely started with a group of individuals working at the Computation Center. There is debate about who was the first to invent the idea of time-sharing. Outside of MIT, Bob Bemer, a computer scientist who invented ASCII, wrote early on about time-sharing. In a later retrospective, he claimed that time-sharing could operate "the same way

[44] Redmond, Kent C. and Thomas M. Smith. *From Whirlwind to MITRE: The R&D Story of the SAGE Air Defense Computer*. Cambridge, MA: MIT Press, 2000.

[45] Friedewald, Michael. "From Whirlwind to MITRE: The R&D Story of the SAGE Air Defense Computer (Review)." *Technology and Culture*. 43, no. 1 (2002): 203-204.

[46] MIT. *MIT Lincoln Laboratory: Technology in Support of National Security.* Edited by Alan A. Grometstein. Lexignton, MA: Lincoln Laboratory, 2011.

[47] Kaiser, David. *Becoming MIT: Moments of Decision*. Massachusetts, MA: MIT Press, 2012. 105.

[48] O'Neill, Judy. "An Interview with Douglas T. Ross." *Charles Babbage Institute*. November 1, 1989.,13.

that the average household buys power and water from utility companies."[49] In England, programmer Christopher Strachey wrote about multiprogramming, a core component of time-sharing.[50] However, most sources credit John McCarthy with the idea. McCarthy has argued that because most of his work on time-sharing was spoken, Strachey was given undue credit.[51]

Setting aside the question of attribution, and instead looking at implementation, time-sharing was first built at MIT. In 1957 MIT's Computation Center was opened using computers supplied by IBM. A number of researchers demonstrated time-sharing on modified IBM 709 in 1961.[52] Like future time-sharing systems, MIT's computer had to be built around custom hardware and software. It also had to be able to have a "supervisor" computer to manage the jobs on the computer. The result of this work was a demonstration of a computing system that would come to be known as CTSS, or Compatible Time-Sharing System. CTSS proved the viability of time-sharing and provided the groundwork for the development time-sharing at MIT and by other institutions.

Early Time-Sharing

In the early 1960s, time-sharing started to gain traction as a concept. The work at MIT piqued the interest of other universities. Businesses, on the other hand, either did not pay attention to the work or still did not see it as a worthwhile investment. The time-sharing business in 1963 had a collective total of five million dollars in revenue.[53] A decade later, the time-sharing market was approaching one billion dollars.[54] This rapid increase in profitability was almost entirely due to the pioneering work done at universities in the 60s.

Public institutions (primarily research universities) were largely the builders and consumers of time-sharing in the 1960s. As an industry, time-sharing was not commonplace until the 1970s. Very few private companies could afford to spend the money developing time-sharing systems. This left a vacuum for university researchers to embed their own principles and ethics into the code and hardware of these early systems.

After MIT's work, many universities started their own time-sharing efforts that influenced commercial time-sharing. MIT continued its time-sharing efforts with the creation of MULTICS,

[49] Bemer, Bob. "Origins of Timesharing." http://archive.today/2014.10.05-165201/http://www.bobbemer.com/TIMESHAR.HTM (Archived October 5, 2014).

[50] Lee, J.A.N. "Computer Pioneers: Christopher Strachey." IEEE Computer Society, 2015. https://web.archive.org/web/20151117091557/https://history.computer.org/pioneers/strachey.html (Archived November 17, 2015).

[51] McCarthy, John. "Reminiscences on the History of Time Sharing." 1983. https://web.archive.org/web/20150320024550/http://www-formal.stanford.edu/jmc/history/timesharing/timesharing.html (Archived March 20, 2015).

[52] Multicians.org. "History." https://web.archive.org/web/20170312175223/http://www.multicians.org/history.html (Archived March 12, 2017)

[53] Campbell-Kelly, Martin and Daniel D. Garcia-Swartz. "Economic Perspectives on the History of the Computer Time-Sharing Industry, 1965-1985." *IEEE Annals of the History of Computing* (January-March 2008). 20.

[54] Campbell-Kelly, Martin and Daniel D. Garcia-Swartz. "Economic Perspectives on the History of the Computer Time-Sharing Industry, 1965-1985." *IEEE Annals of the History of Computing* (January-March 2008). 20.

which was used by General Electric and later Honeywell.[55] On the West Coast, UC Berkeley's work on "Project Genie" contributed greatly to the commercial success of Scientific Data Systems (SDS) time-sharing systems.[56] In England, Cambridge created "Titan Supervisor," an early time-sharing system that influenced later computing hardware. Additionally, the Michigan Terminal System, a project comprised of multiple universities, was technologically innovative, but licensing arrangements prevented the commercial use of the product.

One of the universities that had the most long-lasting impact on the later development of commercial time-sharing systems was Dartmouth College. Dartmouth presents us with an early example of a time-sharing system that was designed around certain core principles. The story of Dartmouth's work demonstrates most clearly the politics and changing definitions of early time-sharing.

Dartmouth's time-sharing system set itself apart in a few crucial ways. First, the system was designed to be inclusive. The code and hardware were developed around the notion of accessibility. This is reflected in the design of the code that ran the Dartmouth system, as well as the computing infrastructure at Dartmouth. Secondly, the economic structure of the system encouraged broader uses of the system (the resource costs were distributed across the campus and not felt by end-users). Finally, the building of the system was created around the notion of utility computing – which likely contributed to the spread of Dartmouth's later commercial computing success.

The Dartmouth Time-Sharing System

Public universities often act as places of infrastructure building. STS scholars have long recognized the role of intellectual networks in shaping the construction of knowledge and the closing of black boxes.[57] Recently scholars have been interested in unearthing infrastructures in order to make visible what is otherwise a passive and invisible process.[58] Fortunately for researchers, early infrastructure projects often make obvious the design choices of a network. This was certainly the case at Dartmouth, when the notion of utility computing was still in its infancy. The story of DTSS underscores how place (in this case, a public university) influences the design of infrastructure systems.

The idea for a time-sharing system at Dartmouth started with John McCarthy's comment to Thomas Kurtz, a professor in Dartmouth's Math Department. McCarthy's innocuous suggestion,

[55] Multicians.org. "History." https://web.archive.org/web/20170312175223/http://www.multicians.org/history.html (Archived March 12, 2017)

[56] Sprinrad, Paul and Patti Meagher. "Project Genie: Berkeley's Piece of the Computer Revolution." *Berkeley Engineering*. http://web.archive.org/web/20100201212638/http://coe.berkeley.edu/news-center/publications/forefront/archive/forefront-fall-2007/features/berkeley2019s-piece-of-the-computer-revolution (Archived February 1, 2010).

[57] This is a foundational concern of STS, particularly Actor-Network Theory.

[58] Bowker, Geoffrey, Karen Baker, Florence Millerand, and David Ribes. "Towards Information Infrastructure Studies: Ways of Knowing in a Networked Environment." In *International Handbook of Internet* Research, edited by Jeremy Hunsinger. Dordrecht: Springer, 2010. 97-117.

"You guys ought to do time-sharing," set the course for the development of DTSS.[59] Prior to the 1960s, Dartmouth's experience with computing was limited.

Dartmouth's entry into digital computing started in the buildup to World War II. In 1940 George Stibitz demonstrated the idea of remote computing at Dartmouth.[60] Stibitz is known for the use of "digital" in the contemporary sense and the use of relays in computers. In 1937, Stibitz, a research mathematician at Bell Labs', built the first relay binary adder called the "Model K."[61] This adder was a departure from the mechanical calculators of the day. Bell eventually funded Stibitz's work, and the Model K was developed into the "Complex Number Computer."

On September 11, 1940, Stibitz took his Complex Number Computer (then called the Model I) to the *American Mathematical Society* meeting being held at Dartmouth College.[62] At this meeting, a teletype terminal was remotely connected (via telephone line) to Model I at Bell Labs in New York City, the first time a digital computer was used remotely. This demonstration likely inspired many of the computing pioneers in attendance, including John von Neumann.[63] Most importantly, it demonstrated that computing could become detached and abstracted from space.

There was little computing activity at Dartmouth between the years of 1941 to 1955. The impact of the Second World War did not seem to advance computing at Dartmouth. Certainly, this was not the case elsewhere, as WWII was one of the main drivers of computing during the 40s, especially at MIT. However, it was during these years that many of the future faculty at Dartmouth were starting to familiarize themselves with computing. These experiences helped shape Dartmouth's future work. Part of this lull can be explained by the lack of funding from federal sources towards computing at Dartmouth. It is also possible that the faculty at the time were unlikely to have the skills to develop or work on a computer, if they were familiar with the computing advancements being made at all. Furthermore, knowledge about electronics was central to the study of computers between the years of 1940-1955.[64] It wouldn't be until stored program computers became common that Dartmouth's faculty would dive into computing.

In the early 1950s, the faculty and administration at the Dartmouth Mathematics department needed a reworking. There were a large number of retiring professors that needed to be replaced. Additionally, Dartmouth lacked a research-focused math department. In order to address these issues, an outside committee recommended that Dartmouth specialize in the history of mathematics. Although many dismissed the suggestion at first, it was a primary motivator for hiring the two faculty members that would have the most substantial impact on the future of computing at Dartmouth: John George Kemeny and Thomas E. Kurtz.

[59] Daily, Daniel. "Oral History: Thomas E. Kurtz." *Special Collections Dartmouth College*. DOH-44. June 20 – July 2, 2002. 17.

[60] Ceruzzi, Paul E. *A History of Modern Computing: 2nd Edition*. Cambridge, MA: MIT Press, 2003. 70.

[61] Irvine, M. M. "Early Digital Computers at Bell Telephone Laboratories." *IEEE Annals of the History of Computing* 23, no 3. (2001): 22-42.

[62] Hollcroft, T. R. "The Summer Meeting In Hanover." *Bulletin of the American Mathematical Society* 46, no. 11 (1940): 859-868.

[63] Hollcroft, T. R. "The Summer Meeting In Hanover." *Bulletin of the American Mathematical Society* 46, no. 11 (1940): 859-868.

[64] Ceruzzi, Paul. "Coevolution of Electronics and Computer Science." *Annals of the History of Computing* 10, no. 4. 1989.

John Kemeny was hired as a person to "revitalize the math department."[65] Kemeny arrived to Dartmouth in 1953 with a number of significant computing-related experiences. He arrived as a child in the US in 1940 from Budapest at 14 years of age after fleeing from Nazi Germany's imminent invasion.[66] While at Princeton, Kemeny took part in the Los Alamos Project during WWII. At Los Alamos, he was a "computer," using an IBM bookkeeping calculator to obtain solutions for the design of the atomic bomb.[67] He was introduced to a "fully electronic computer based on a binary number system, with internal memory for both data and a stored program," by another Hungarian, John von Neumann.[68] After his experiences at Los Alamos, he finished both his B.A. (1947) and PhD (1949) at Princeton, where he worked as a research assistant to Albert Einstein and continued to meet with von Neumann regarding the electronic computer. He was an assistant professor in the philosophy of science (what he called his hobby) but was never a "pure" mathematician.[69] These early experiences provided the groundwork for Kemeny's later ideas on time-sharing and computer programming.

It was not until two years later that Kemeny met and hired Thomas Eugene Kurtz, another Princeton graduate. Thomas Kurtz received his PhD in 1956 from Princeton in statistics.[70] He discovered computing in 1951 while attending the "Summer Session of the Institute for Numerical Analysis at UCLA." Once he arrived at Dartmouth, he quickly became the liaison for the newly established New England Regional Computer Center at MIT. The Computer Center was a collaboration between MIT, IBM, and many other schools and universities. It allowed campuses like Dartmouth to use IBM computers.

Kurtz's job at the time was to promote the Center and to take punched cards down to MIT to be processed by their computer. Every two weeks, Kurtz would get on the 6:20am train with a box of cards, submit them to MIT around 9:30am and receive the results at the end of the day.[71] This process was laborious and time-intensive. It was clear that this model of computing was not working for the Dartmouth faculty.

Dartmouth received its first computer in 1959 (a Royal McBee LGP-30).[72] Despite having access to a computer, many of the same issues with mainframe computers remained. Punch cards still

[65] Daily, Daniel. "Oral History: Thomas E. Kurtz." *Special Collections Dartmouth College*. DOH-44. June 20 – July 2, 2002. 5.

[66] Ohles, Frederik, Shirley M. Ohles, and John G. Ramsay. *Biographical Dictionary of Modern American Educators*. Westport, CT: Greenwood Press, 1997. 189.

[67] Lee, J.A.N. "Computer Pioneers: John George Kemeny." *IEEE Computer Society*. http://web.archive.org/web/20160322032829/http://history.computer.org/pioneers/kemeny.html (Archived March 22, 2016).

[68] Lee, J.A.N. "Computer Pioneers: John George Kemeny." *IEEE Computer Society*. http://web.archive.org/web/20160322032829/http://history.computer.org/pioneers/kemeny.html (Archived March 22, 2016).

[69] Steen, Lynn Aurthur and Kemeny, John. "John G. Kemeny: Computing Pioneer." *The Two-Year College Mathematics Journal* 14, no.1 (1983): 25.

[70] Wexelblat, Richard L. *History of Programming Languages*. New York, NY: Academic Press (1981). 549.

[71] Daily, Daniel. "Oral History: Thomas E. Kurtz." *Special Collections Dartmouth College*. DOH-44. June 20 – July 2, 2002. 8.

[72] Lee, J.A.N. *International Biographical Dictionary of Computer Pioneers.* London: Fitzroy Dearborn Publications (1996). 411.

had to be fed into the reader, and valuable time was lost when a program was written incorrectly, or there were simply too many users waiting to submit their programs. McCarthy's comment regarding time-sharing sparked the imagination of Kurtz and Kemeny. At the start of the 60s, they set off to develop a time-sharing system modeled around the notion of utility computing.

The LGP-30, like all computers of the time, was not designed for time-sharing, but it provided faculty and students the opportunity to start making steps towards the construction of a time-sharing ecosystem. In order to build a time-sharing system, Dartmouth would need new software and hardware. In 1962, Dartmouth approached the National Science Foundation for a grant to fund the research and purchase a new computer.[73] In the following year, the grant was approved, and an arrangement with General Electric was reached to purchase a GE-225 (later a GE-235), a Datanet-30, and a disk file at a discounted price.[74] With modifications, this hardware provided a slate for the creation of DTSS.

Between the years of 1962-1964, Kemeny, Kurtz, and Dartmouth undergraduate students worked on developing DTSS. Building DTSS required both a novel hardware approach and a new programming language: Beginner's All-purpose Symbolic Instruction Code, BASIC. These two developments occurred simultaneously. As Kurtz put it, "the language of BASIC and the time-sharing system in which it resides were…inextricably meshed."[75] Likewise, in order to understand the success of DTSS, we need to examine the design principles of BASIC.

BASIC was designed around the notion of simplicity. Kemeny remarked that prior to BASIC, programing was designed for "machines and not human beings."[76] Therefore, BASIC was meant to be easily learned by both technical and non-technical users. The syntax of BASIC was written in simple English, and the number of commands was limited.[77] The language was designed for anyone at the college to be able to quickly write computer programs. The original memos dictated that "In all cases where there is a choice between simplicity and efficiency, simplicity is chosen."[78]

Kurtz said, in a retrospective in 1978, that in designing BASIC, they asked: "How can sensible decisions about computing and its use be made by persons essentially ignorant of it?"[79] In designing a language around simplicity and for a broad audience, BASIC was not as flexible as other languages. For instance, BASIC was only meant to be compiled (not modified after the program has been executed).[80] This meant that a certain level of flexibility was lost in favor of speed and ease of use. Audience awareness was key. Kurtz claimed that "we did not design a language and then attempt to mold the user community to its use."[81]

[73] Kemeny, John G. and Thomas E. Kurtz. "Dartmouth Time-Sharing." *Science* 162, no. 3850 (1968): 224.
[74] Kemeny, John G. and Thomas E. Kurtz. "Dartmouth Time-Sharing." *Science* 162, no. 3850 (1968): 224.
[75] Kurtz, Thomas E. "BASIC Session." *History of Programming Languages*, 1981. 521.
[76] Kemeny, John. *Man and the Computer.* New York, NY: Charles Scribner's Sons, 1972. 7.
[77] Slater, Robert. "John Kemeny and Thomas Kurtz – Handling the Computer – BASICally." *Portraits in Silicon*. Cambridge, MA: MIT Press, 1992. 245.
[78] Kurtz, Thomas E. "BASIC Session." *History of Programming Languages*, 1981. 520.
[79] Kurtz, Thomas. "BASIC." *ACM SIGPLAN Notices*. 13, no. 8 (1978): 105.
[80] Kurtz, Thomas. "BASIC." *ACM SIGPLAN Notices*. 13, no. 8 (1978): 108.
[81] Kurtz, Thomas. "BASIC." *ACM SIGPLAN Notices*. 13, no. 8 (1978): 116.

This goal of designing a simple language was probably the most significant design choice and was a key to DTSS's success. Unlike the many other programming languages at the time, BASIC favored ease of use over the flexibility of a more complete programing language. A computer scientist may not have considered BASIC an efficient language because of the limited commands, but it was designed with a broader public in mind. This was a radical departure from the other time-sharing systems at the time. In part, this design choice was a result of Dartmouth's previous attempt at using other available programing languages before inventing their own. The design choices are an early example of computing starting to consider broader publics as potential users.

Quickly the first version of BASIC was developed, along with all that was needed to get the DTSS operating system functioning. The system was first operational in the spring of 1964. In the initial months, DTSS was unstable and regularly crashing. However, by the fall, DTSS had become relatively stable, and the operating system had matured.[82] In the summer, faculty were introduced to the new system, and in the fall many freshmen were introduced to the computer through the introductory mathematics course. In the coming years, nearly all freshmen at Dartmouth – regardless of major – were learning to program using BASIC and DTSS.[83] All university members could access the computer using the teletype machines spread across campus. Additionally, it was accessible by multiple users at a single time (the core principle of time-sharing). A computer interrupt system balanced jobs between different users, allowing for nearly instant communication between the user and the computer.

When compared to other universities that had time-sharing projects, Dartmouth provided generous access to the computer. Even non-academic activities, such as using the computer to simulate the outcome of college football games, were encouraged.[84] The cost of using the system was dispersed (largely through student fees) and not billed to any single individual. This "library model" was threatened by the college administration, who wanted to charge for the service, but it ultimately remained free for students.[85] Furthermore, DTSS reached beyond the university campus. A number of primary schools and smaller colleges gained access to DTSS through the use of remote teletypes. Computer programing was taught remotely across different educational levels. Students and teachers at these schools could access any number of programs that resided at Dartmouth.

DTSS succeeded not simply because of the advancements in computing programing and time-sharing software. Instead, DTSS was a success because it was developed around a few key design choices: simplicity, accessibility, and reliability. Kurtz claims that this led to a symbiosis between the student and the computer "because the design does not discriminate."[86] The meshing of DTSS with BASIC allowed for the development of a shared intellectual infrastructure. These advancements reached beyond computer science; they speak to the "experience [of] what computing can and cannot do" for the individual.[87]

[82] Kurtz, Thomas E. "BASIC Session." *History of Programming Languages*, 1981. 520.
[83] Kemeny, John G. and Thomas E. Kurtz. "Dartmouth Time-Sharing." *Science* 162, no. 3850 (1968): 226.
[84] Kemeny, John G. and Thomas E. Kurtz. "Dartmouth Time-Sharing." *Science* 162, no. 3850 (1968): 228.
[85] Arms, William Y. "The Early Years of Academic Computing." Internet First University Press. May 2014.
[86] Kurtz, Thomas. "The Many Roles of Computing on the Campus." *Spring Joint Computer Conference* (1969). 655.
[87] Kurtz, Thomas. "The Many Roles of Computing on the Campus." *Spring Joint Computer Conference* (1969). 655.

What became clear in the later years of DTSS was that the true value of DTSS was not in selling computing time, but the construction of a new utility.[88] DTSS can be measured as a success by the number of users that it influenced, the contributions to other time-sharing endeavors, and the proof-of-concept for the creation of a computing utility. The work at Dartmouth provides a model for the successful implementation of utility computing at a large-scale. It is unlikely that DTSS would have touched so many communities had it not been treated as an open and shared resource.

Beyond the university community and computing advancements, DTSS quickly contributed to the success of commercial time-sharing companies. First, with GE's adoption of the system, which "became the backbone of the GE service bureau business."[89] Later DTSS and BASIC were used by numerous time-sharing providers. Dartmouth's own DTSS, Inc. provided commercial services for businesses.

Private Time-Sharing

It was not long before commercial companies decided that the time-sharing software had financial potential. Many businesses started to demand access to the computer. For smaller businesses, time-sharing offered access to a computer at a limited cost and access to various software tools. Time-sharing also appealed to large businesses that had, or could afford, a computer, but want to expand their computing division without needing to manage additional in-house technology. The business of time-sharing was predicated on selling access to computing services, without the commitment of a physical computer purchase or lease. The development of the early commercial time-sharing systems brought about numerous technical and regulatory changes. The development of early computer networks was influenced by both public and private time-sharing systems. Public time-sharing companies introduced the notion of utility computing, and commercial providers took the vision and ran with it.

Using Dartmouth's system, GE provided the first commercial time-sharing service.[90] Soon, many other time-sharing providers would start providing services. The development of these time-sharing services occurred under the larger umbrella of the data processing services industry.[91] Companies no longer had to wait to receive their programs back in the mail. Instead, they could remotely access a time-sharing system to execute their jobs quickly. This demand opened the doors for the creation of a new business model – commercial utility computing.

Tymshare

One of the first commercially successful time-sharing companies was Tymshare. Tymshare's business model changed as it grew. As the name implies, Tymshare started as a time-sharing

[88] Kurtz, Thomas. "The Many Roles of Computing on the Campus." *Spring Joint Computer Conference* (1969). 653.
[89] Daily, Daniel. "Oral History: Thomas E. Kurtz." *Special Collections Dartmouth College*. DOH-44. June 20 – July 2, 2002. 42.
[90] Campbell-Kelly, Martin and Daniel D. Garcia-Swartz. "Economic Perspectives on the History of the Computer Time-Sharing Industry, 1965-1985." *IEEE Annals of the History of Computing* (January-March 2008): 20.
[91] Campbell-Kelly, Martin and Daniel D. Garcia-Swartz. "Economic Perspectives on the History of the Computer Time-Sharing Industry, 1965-1985." *IEEE Annals of the History of Computing* (January-March 2008): 18

business. As the business grew, and the time-sharing market matured, the business model shifted towards the management of a packet-switching network. This transformation follows a similar trajectory to Dartmouth's time-sharing efforts. In both cases, as each respective system grew, the importance of maintaining a computing utility became clearer. By examining the development of Tymshare, we can map out different types of values that were embedded in the system.

Tymshare was started by Tom O'Rouke. In the early 1960s, O'Rouke was an electrical engineer and regional manager for General Electric's computing division on the West Coast. At the time, General Electric had not started time-sharing efforts. Despite this, O'Rouke was not a stranger to the idea of time-sharing. As previously mentioned, the Dartmouth time-sharing system initially used GE computers. When O'Rouke was working for GE in Phoenix, Dartmouth approached the company looking for discounted computer systems for DTSS. O'Rouke encouraged GE to give Dartmouth a discount. In turn, O'Rouke became familiar with the idea of time-sharing and was able to use DTSS from San Francisco for his own personal use.[92]

Dartmouth and Berkeley's time-sharing systems were instrumental to the first two years (1965-1966) of the business: Tymshare Associates. Even prior to the creation of the Tymshare software, O'Rouke demonstrated time-sharing to potential clients through DTSS. Furthermore, the development of the software was aided through the use of Berkley and Dartmouth's time-sharing systems. Berkeley gave Tymshare a copy of their time-sharing software, which Tymshare used to make a commercial product. Without these previous efforts, it is unlikely that Tymshare would have developed when it did.

Time-sharing, at the time, faced a number of technical and economic challenges for commercial companies. The infancy of time-sharing software meant that nearly all hardware and software had to be built or modified to work for time-sharing purposes. Tymshare initially purchased a computer from Scientific Data Systems (SDS), which later became Xerox Data Systems. The decision to go with SDS was made after being snubbed by GE, who refused to honor a purchase agreement, citing competition with their new time-sharing business. This system, and later systems, had to be adapted to the specific challenges of long-distance computing.

Prior to the arrival of the SDS 940 in May, O'Rouke hired a number of talented engineers and salespeople to build the system and attract clients. Ann Hardy, previously a programmer at IBM and developer of a time-sharing system at Livermore Labs, developed the monitor (what would now be called the shell or user interface).[93] Verne Van Vlear, from GE, developed the kernel to manage the hardware.[94] Thanks to these efforts, Tymshare had an operational system in July 1966, capable of handling eight simulators users.[95] Many other skilled programmers worked to build a stable time-sharing system, including a collaboration with Com-Share, another time-sharing company from Michigan.

[92] John, Luanne. "Oral History of Tom O'Rourke." *Computer History Museum.* March 13, 2002. 6-7.

[93] Abbate, Janet. "Oral History: Ann Hardy." *IEEE History Center*. July 15, 2002. https://web.archive.org/web/20170103061121/http://ethw.org/Oral-History:Ann_Hardy (Archived January 3, 2017).

[94] Hardy, Ann. "Tymshare Notes: Draft." *Computer History Museum.* November, 17, 2004.

[95] McNown, Rebecca M. "Tymshare, Inc.: 1965-1970. Report for History 10." *Computer History Museum* (August 17, 1970). 9.

When the computer arrived, access to Tymshare was given to clients for free, on an admittedly unstable version of the software. In November, Tymshare started charging for access to the system. Tymshare offered three services: access to a number of computer languages, industry-specific software applications, and technical support for the clients.[96] Specialized software, particularly for engineering firms, proved to be a valuable asset. However, the initial success of their system was as much about geography as it was services.

The issue of distance and physical space was critical to the long-term development of commercial time-sharing (and later networked computing) and Tymshare's success. The first clients that used Tymshare were individuals in the aerospace industry. Many of these companies were located a short distance from Tymshare's computer in Palo Alto. Consequently, the telephone connection was considered a local call.[97] This meant that Tymshare initially had an advantage between IBM's San Francisco and GE's San Jose locations. However, both GE and IBM were quick to pick up the phone charges for their clients.

What O'Rouke and others realized is that in order to compete, time-sharing had to be geographically accessible, both in terms of access to local calling nodes and distributed computer locations. In August 1966, Tymshare opened an office in Los Angeles and a New Jersey location in the following year. Additionally, Tymshare leased expensive telephone lines to deliver the data over the newly developing network. As the network of sales offices and computers grew, Tymshare looked to expand its reach. The engineering work on multiplexing (allowing multiple connections to share a single telephone line) helped grow Tymshare and made possible the next project: Tymnet.

Tymnet was developed as a network that could link the Tymshare computers and networking nodes. Tymnet developed as a commercial network, drawing inspiration from ARPANET. The network allowed clients of Tymshare to connect remotely to their company's computers, access Tymshare, or any other remote computer that they could dial into. Additionally, it allowed for the creation of online banking and other "transaction applications."[98] The creation of the network allowed for a more centralized management of the network by allowing many IT jobs to be worked on remotely.

While Tymnet remained a subdivision of Tymshare, the growth of Tymnet eventually overtook the value of the original time-sharing system. Regulatory pressures from the FCC meant that time-sharing remained the company's public focus, but throughout the 1970s Tymnet continued to grow. Eventually, Tymnet split from Tymshare, and the network company was sold to the American aerospace company, McDonnell Douglas. Douglas later acquired Tymshare as well.

The history of Tymnet demonstrates some core differences between public time-sharing (in the case of DTSS) and commercial time-sharing. Unlike the work at Dartmouth, the assumed user of Tymshare was a business user who was already familiar with computers. Therefore, the design of

[96] McNown, Rebecca M. "Tymshare, Inc.: 1965-1970. Report for History 10." *Computer History Museum* (August 17, 1970). 2.

[97] John, Luanne. "Oral History of Tom O'Rourke." *Computer History Museum.* March 13, 2002. 9.

[98] Abbate, Janet. "Oral History: Ann Hardy." *IEEE History Center*. July 15, 2002. https://web.archive.org/web/20170103061121/http://ethw.org/Oral-History:Ann_Hardy (Archived January 3, 2017).

Tymshare focused on the economic viability of the system. Accessibility was measured in terms of the difficulty of connecting to the service and the cost associated with the use of the service. Use of the computer was directly measured (how long was the connection, how much data was stored, etc.). This is a different type of accessibility than one that focuses on connecting a broader "public" to computing.

Much of the success of time-sharing at Tymshare can be linked with the development of Tymnet. This speaks to a broader point about time-sharing in general. Time-sharing's value was not simply that it more efficiently used computing resources. Instead, time-sharing demonstrates the economic and cultural value of building computing infrastructures. The rise of computer networking (via the Internet) in the following decades overshadowed contributions that time-sharing made to utility computing.

In the late 1950s and throughout the 1960s, computers seemed to be on track to be the newest utility service. As utility computing started to become a popular concept (backed by the success of public and commercial time-sharing), new regulatory issues emerged.

Regulating the Utility

The idea of the computer as a utility started with computer researchers. Throughout the 60s, the vision of computing was advanced and debated. Time-sharing experiments modeled a type of utility that sparked the academic imagination for larger projects. However, for private data processing companies, this notion of utility computing was worrisome. Many actors, particularly in the regulatory sphere, were suspicious of telecommunication companies owning a new type of digital utility. The later introduction of packet-switching using broader telecommunication infrastructures only invited additional scrutiny from both the owners of the networks, as well as the broader public. In inheriting the metaphor, the data processing companies carried the baggage of utility regulation.

In 1961, at MIT's centennial celebration, John McCarthy said that "computing may someday be organized as a public utility just as the telephone system is a public utility."[99] He went on to say that "the computer utility could become the basis of a new and important industry." Researchers carried on this discussion as the technology for time-sharing was improved. In 1967, Paul Armer of the RAND Corporation wrote an extensive paper on the implications of utility computing on all aspects of sociology. Privacy, in particular, was of concern because the computing industry was growing faster than regulations could adapt. His concern was that "small, widely dispersed puddles" of information about individuals were slowly being organized by this newly forming utility system.[100]

More broadly, the computer utility started to be seen by some as a type "community resource, somewhat analogous to a library."[101] However, the analogy to the electrical or telephone system was the most common comparison for time-sharing systems. The comparison to the telephone system was only strengthened later on as telephone infrastructure was used as the backbone for

[99] Garfinkel, Simson. "The Cloud Imperative." *MIT Technology Review*. October 3, 2011.
[100] Armer, Paul. "Social Implications of the Computer Utility." *Rand Corporation*. 1967.
[101] Fano, R. M. "The Computer Utility and the Community." *IEEE International Convention Record* (1967). 34.

utility computing. Time-sharing, it was argued, is valuable because of the spill-over benefits made possible by the underlying network.[102] As the conversation continued, the discussion of the computer utility moved further away from the feasibility of the idea and towards the possibilities for regulation.

The discussions about utility computing emerged during a turbulent time in American politics, particularly for American telecommunications companies. AT&T's (Ma Bell) monopolistic control over the telecommunications infrastructure was starting to be threatened by small companies that provided added services using AT&T's network (see the use of unauthorized equipment and the FCC's Carterfone ruling).[103] In addition to computer information services, cable TV and microwave-based data transmission threatened the FCC's regulatory strategy "that telephone service be kept structurally separate from other sectors of the telecommunication industry."[104] Despite these regulations, the potential profits of utility computing were too great for many companies to resist.

In 1966, the FCC started looking into the issue of regulation in the first set of computer inquiries: Computer I. There were three separate computer inquires (Computer I, II, and III), each of which attempted to define how these new computing businesses should be regulated. In each new inquiry, the FCC attempted to fit a new type of telecommunication service into an older regulatory box. However, time-sharing and similar services proved to be difficult to regulate because they existed as both a communication service and a data processing medium.

One of the primary issues within the FCC's inquiries was the issue of classification of networks. In Computer I (1966-1970), the FCC attempted to classify networks as either "pure communication" or "pure data processing." Time-sharing services defied this categorization scheme. The FCC labeled these businesses a "hybrid" and decided that each service should be dealt with on a case-by-case basis as to which of the two categories this hybrid network should fall under. [105] This definition proved untenable, and Computer II attempted to clarify this issue with the introduction of a new scheme: "Basic" or "Enhanced" networks. [106] Computer II helped clarify many of the issues by eliminating a hybrid category and labeled nearly all of the services as "enhanced." Previous scholars have highlighted how these regulatory moves treated the computer as a boundary object, and the FCC's role was to create a form of "linguistic engineering" to fit hybridity into an existing regulatory landscape.[107] In 1985 the final inquiry, *Computer III*, took place after the breakup of Bell. In the third inquiry, the FCC focused on removing the "maximum separation requirement" requirement that attempted to keep ownership

[102] Greenberger, Martin. "The Two Sides of Time Sharing." *Working Paper: Alfred P. Sloan School of Management* (1965). 15.

[103] Lesk, Michael. "Son of Carterfone: Network Neutrality or Regulation?" *IEEE Security & Privacy* (May 2010): 77-82.

[104] Zarkin, Michael J. "Telecommunications Policy Learning: The Case of the FCC's Computer Inquiries." *Telecommunications Policy* 27 (2003): 288.

[105] Cannon, Robert. "The Legacy of the Federal Communications Commission's Compute Inquiries." *Federal Communications Law Journal* 55, no. 2 (2003): 174.

[106] Cannon, Robert. "The Legacy of the Federal Communications Commission's Compute Inquiries." *Federal Communications Law Journal* 55, no. 2 (2003): 184.

[107] Lentz, Roberta G. "Dissertation: 'Linguistic Engineering' and the FCC Computer Inquiries, 1966-1989." *University of Tax at Austin*, 2008. 239.

of "basic" and "enhanced" services apart. Yet again, the FCC struggled to deal with the hybridity of these digital networks (for instance, AT&T voice messaging storage[108]). At the end of the inquiries, the separation policy was eliminated, with the stipulation that companies provide competitors with access to their own networks.

However, despite all of the inquiries, the FCC did not get to the root of the issue, whether these new computer businesses were providing a utility service. If they were, should it be considered a public utility or a private one? What are the commitments that the operators of a time-sharing system have to make to the public? As law professor Keven Werbach argued: "The FCC's quarantine approach in *Computer I* allowed it to avoid confronting the hard questions that the computer utility visionaries raised back in the 1960s."[109] *Computer II and III* did not do much to address these questions, except to move towards a private utility system. Instead, these inquiries challenged public trust in this new utility and made the expansion of these systems more difficult (which is a reoccurring theme in the creation of new complex technical systems).[110]

This is not a new issue and is likely to reappear in regard to the regulation of the cloud and the Internet of Things (IoT). Regulatory agencies have skirted around other tricky digital legal issues, from controlling piracy to managing mobile spectrum. This common thread was noted by Sheila Jasanoff, who argued that "in the computer age, it is increasingly difficult to pin down with certainty the places where politically salient events originate, let alone to determine who controls the levers of power."[111] In hindsight, the actors pulling the levers of power in the *Computer Inquiries* may be easier to spot than the current users and developers of the cloud today.

One of the most important lessons of the era of time-sharing is that there were competing ideas of how a public utility should, or could, be built. These ideals are not simply words that computer researchers discussed. Instead, these ideals were built into the software, hardware, and networks of the time. As I have argued, services like DTSS were constructed around a set of design principals that favored accessibility with a more general public in mind. The principles of Tymshare developed around a competing set of ideals. However, for all of their differences, both services raised issues about the role of computer networks in the construction of information utility infrastructures.

Largely, the discussion over-regulating remote computing services has remained dormant since the end of *Computer II*. Discussions about computer utilities have also been minimal. However, looking back at the history of time-sharing in the 1960s, it is clear that there were a number of unresolved visions of computing. These visions are now reemerging with the introduction of cloud computing.

[108] Zarkin, Michael J. "Telecommunications Policy Learning: The Case of the FCC's Computer Inquiries." *Telecommunications Policy* 27 (2003): 295.
[109] Werbach, Kevin. "The Network Utility." *Duke Law Journal* 60 (2011): 1810.
[110] Marchant, Gary, Kenneth Abbott, and Doughlas Sylvester. "What Does the History of Technology Regulation Teach Us About Nano Oversight?" *Jounrnal of Law, Medicine, and Ethics* 37, no 4 (2009): 4-5.
[111] Jasanoff, Sheila. *States of Knowledge: The Co-production of Science and the Social Order*. London: Routledge, 2006. 31.

As seen in the case of utility computing, the metaphor of utility conjures images of control and management, as well as obligations to the public. The creation of cloud computing gives us something more abstract. The story of utility computing is a predecessor to the cloud. Like time-sharing, the creation of the cloud was an attempt to reduce the limitations of computing at the time. Instead of solving the issue of batch processing, the cloud attempted to solve (amongst other things) the inefficiencies of distributed servers. Similarly, cloud computing has allowed users relatively easy access to powerful hardware (while trading off a certain level of control over the hardware itself). However, despite these similarities, the cloud has not directly addressed the politics of regulation in the same way that utility computing did. As this chapter demonstrates, those political issues were never fully answered.

By looking back at the early innovations in utility computing, we might be able to reinterpret the products of the cloud today. The cases I have illustrated demonstrate the impact of designing systems around certain assumptions and audiences. It also shows the importance of thinking about these technological systems as being part of a larger ecosystem. If we think about the cloud as belonging to a larger technological infrastructure or utility, that may give us the tools we need to shape an alternative technological future.

In the next chapter, we will continue to look at the importance of metaphor in computing history by looking at another technological predecessor to the cloud, ubiquitous computing.

Chapter 2

Ubiquitous Computing

Introduction

In the previous chapter, the history of utility computing was discussed in order to reveal one of the defining characteristics of cloud computing. The history of utility computing did not direct the development of the cloud but played an important role in setting the stage and providing the metaphors for a computing model built around the idea of distributed computing. As the cloud becomes further commoditized and embedded into the modern computing landscape, the history of utility computing can continue to inform contemporary debates over the cloud.

Ubiquitous computing features many of the same novel computing advancements, stretching of metaphors, and desires to build new technological infrastructures that the history of utility computing demonstrated. I argue that these two metaphors, utility and ubiquity, encapsulate the core of what the cloud was envisioned as and what it continues to be built towards. Understanding the implications of these metaphors demands that we revisit the original history of these ideas.

This chapter looks at the origin of ubiquitous computing in order to identify the historical and sociological underpinnings of designing computing systems around the notion of ubiquity. By tracing the development of ubicomp (a commonly used abbreviation of the term) at Xerox PARC and later adoption of the idea, the ideology of pervasive computing can be explored. One of the core assumptions made about early ubicomp systems is that they should be designed as "calm technologies." This philosophically inspired idea did not persist in later iterations of ubicomp, but may offer clues about how to design future cloud computing environments. Calm technologies are best understood as technological devices or systems which adapt to the attention of the user by becoming visible when in use and sitting in the periphery when not needed. Calm technologies are attuned to human psychology.

Ubicomp is explored by first examining the origins of "ubiquitous computing" (with particular focus on Xerox PARC). The chapter then looks at the spread of ubiquitous computing and the philosophy behind its earliest form. Finally, the chapter finishes with the shift to the cloud and the tapering off of ubicomp. The chapter offers us an opportunity to read the cloud through the lens of ubicomp. Although the scope of the work at Xerox PARC was small, the culture of product design deserves a closer look as it relates to the development of networked technologies. I argue that as the cloud becomes more ubiquitous, we should attempt to inject the philosophy of calmness that guided early ubicomp technologies.

Ubiquity Beyond Computing

Prior to the coining of the term ubiquitous computing in 1988, computing had already begun to shift toward more pervasive usage. The spread of networked computing helped decentralize the notion of local computing. Early uses of time-sharing in the 1960s let users access remote computing resources from wide geographical areas. After 1988, with the spread of more consumer-friendly internet service options, access to information and other computing resources became more commonplace.

The rise of personal computing, particularly in the 1980s, helped lay the ground for the introduction of ubiquitous computing. This time period saw the price of computing fall and the creation of a personal computing market. Despite the falling prices, computers were still expensive to own for the typical home user. Even with these high costs, many individuals started to imagine the spread of computing beyond institutions and individuals. These imaginings came to be solidified in the coining of the term "ubiquitous computing."

The story of ubiquitous computing continues today with the spread of the cloud and IoT (Internet of Things) devices. In this chapter, we revisit original ubicomp discussions and inventions starting in the late 1980s. The focus here is on the ideological commitments and assumptions made by ubicomp advocates at Xerox PARC and detractors in the technology business more broadly. The history of this metaphor highlights key differences between the types of computing environments envisioned in the early 1990s versus the use of ubicomp today. One of the key differences was an early commitment to designing "calm technologies." This notion of "calm" is explored and offers a potential method of evaluating the current development of cloud technologies. This chapter highlights the importance of distinguishing between calmness from the periphery versus invisibility.

Ubiquitous Computing Today: Core Characteristics

Ubicomp does not have a standard definition that all actors agree upon. This discord between definitions has been true throughout the lifespan of the concept. Today, the field of ubicomp has taken a backseat to the rise of cloud computing and the Internet of Things. Still, ubiquitous computing remains an important idea in academia and as a concept for describing the cloud. After three decades, there is not a single definition of ubicomp. Instead, there is a general agreement of the types of technological characteristics that makes up ubicomp. These characteristics include: pervasiveness, locality, seamlessness, decentralization, calmness, and selective visibility.

The idea of ubiquitous computing has often been used interchangeably with the term "pervasive computing." This linkage was made rather early on. In an early conceptual model, pervasive computing was defined as aspiring "to be ubiquitous…lost-cost, embedded, distributed and non-intrusive."[112] Some authors have argued that ubiquitous computing is a term "used when the emphasis is put on the opportunity of humans to have access to computing and to use multiple computing devices from anywhere," whereas pervasive computing is "used to express that computing is (invisibly) embedded in everything in an all-embracing connectivity."[113] However, both popular[114] and academic articles[115] regularly reverse the definitions of the two terms. What

[112] Ciarletta, Laurent and Alden Dima. "A Conceptual Model for Pervasive Computing." National Institute of Standards and Technology. August 2000.

[113] Horvath, Imre and Regine W. Vroom. "Ubiquitous Computer Aided Design: A Broken Promise or a Sleeping Beauty." *Computer-Adided Design* (2014): 2.

[114] Judge, Jargon. "Ubiquitous? Pervasive? Sorry, They Don't Compute." *ComputerWorld*. March 20, 2000.

[115] Mousa, Assem Abdel Hamed. "Ubiquitous/Pervasive Computing." *International Journal of Innovative Research and Development* 2, no. 10 (2013).

matters, instead, are the core features of each concept and how those concepts intersect with visions of the cloud.

In addition to this idea of ubiquitous computing being pervasive, it is also considered to be a fundamentally local experience. Ubiquitous technologies are generally thought to be technologies that sit in close proximity to us. As I will describe in more detail later, the root of ubicomp emerged from the ability to embed low-cost computers into many small devices. Unlike remote computing, the physical presence of a device is important in designing a unique computing environment.

This notion of computing in the local setting feeds into the final three characteristics: seamlessness, decentralization, and selective visibility. Ubicomp is a vision of computing where devices work seamlessly together through local networking. Through this lens, a computing device on the table can easily interface with the windows on the wall or the phone on the desktop. This ease of operation is imagined through both a notion of seamless networked technologies, as well as a general notion of decentralization. Rather than depending upon a single server, these ubicomp environments have differing degrees of data collection and processing responsibilities. Ubicomp is thought of as somewhat decentralized because it is not built around a hierarchy.

Finally, ubicomp is sometimes (although not as frequently as the other characteristics) associated with a vision of selective visibility. According to this view, technologies should recede into the background when they are not in use and come to the forefront when they are needed. The visibility of a technology, therefore, depends upon the needs of the user. This selective visibility, along with the other dominant characteristics of ubicomp, shares a great deal of overlap with the visions of cloud computing. In the following sections, I describe the emergence of the concept of ubicomp and some of the early visions of the technology. By looking at this early history, we can more easily contrast the dominant characteristics of the cloud against some of the more nuanced differences in ubiquitous computing.

Ubiquitous Computing at Xerox

The concept of ubiquitous computing can be directly linked to the pioneering work at Xerox Corporation’s Palo Alto Research Center, better known as Xerox PARC. The history of Xerox started with the sale of photographic paper and the eventual creation of the “XeroX Copier.” The New York based company worked throughout the 1950s to improve upon the copier design. In the 1960s, Xerox dominated the file copier market. However, many within and outside of the company realized that growth of the business mandated a move towards digital electronics. Both IBM and Kodak challenged Xerox’s dominance in the 1970s through competing products and anti-trust suits.[116]

The building pressure prompted Xerox to purchase Scientific Data Systems (SDS) in 1969. Jack Goldman, a physicist at Xerox, proposed that SDS’s basic research capabilities be bolstered through the creation of a corporate research center. The center, Xerox PARC, was meant to

[116] Elder, Tait. “New Ventures: Lessons from Xerox and IBM.” *Harvard Business Review*. July-August 1989.

challenge IBM's York Town Heights and AT&T's Bell Labs.[117] One of Goldman's motivations for this research arm was to keep up with the emerging computing industry. Eventually, the PARC project was approved, and Goldman was chosen to lead and establish the center in Palo Alto, California.

Xerox PARC's establishment occurred during a unique moment. Michael A. Hiltzik, a journalist who wrote about the development of PARC, argued that this historical moment helped PARC grow in the 70s and 80s. He claimed that Xerox's massive cash flow from the office copier allowed for the hiring of many talented researchers during a period of overall decline in government research budgets (due to the cost of the Vietnam War). Furthermore, the advancements in computing speed allowed for "science's most farsighted visionaries to realize their dreams for the first time."[118] Finally, the management of PARC was largely untethered from the rest of Xerox. Geography also likely had a large impact on the development of PARC. Established in Silicon Valley, the workplace environment at PARC was interconnected with the universities and other IT businesses at the time. Unlike the East Coast offices of Xerox, this workplace environment was casual and informal.

Throughout the 1970s, Xerox PARC worked on developing new computing technologies. Many of these inventions were framed in terms of changing the dynamics between computers and humans. For instance, the development of the Xerox Alto computer introduced a system designed around a graphical user interface environment. Other work, like the development of the programming language Smalltalk, placed emphasis on programming around graphical objects rather than simple text. Additionally, the work at PARC in the 19070s is perhaps most notable for inspiring Steve Job's own efforts at Apple designing a new form of personal computing. Jobs later went on to say that Xerox's failure to make these products commercially successful was due to a lack of "product people" and an overrepresentation of salespeople – calling the management at Xerox's East Coast office "Toner-Heads."[119]

This opinion by Jobs has been echoed by others in the technology industry. Many of these individuals argue that Xerox could have dominated the computing industry. On the surface, this certainly seems plausible. The development of an advanced graphical user interface, prototypes for personal computers, new word processing applications, and Ethernet were significant advances in computing.[120] However, as Hiltzik argued, the ability to turn these ideas into a commercial product was not a given. What PARC excelled at was the development of a vision of computing. One critical account written in 1988, Douglas Smith's *Fumbling the Future*, argued that managerial disagreements (as well as a lack of engineers within management) contributed most strongly to Xerox's failure to dominate the early computer market.[121]

[117] Hiltzik, Michael A. *Dealers of Lightning: Xerox Parc and the Dawn of the Computer Age*. New York, NY: Harper Collins, 1999. 30.

[118] Hiltzik, Michael A. *Dealers of Lightning: Xerox Parc and the Dawn of the Computer Age*. New York, NY: Harper Collins, 1999. xxi.

[119] *Triumph of the Nerds*. Channel 4 / PBS Documentary. 1996.

[120] Hiltzik, Michael A. *Dealers of Lightning: Xerox Parc and the Dawn of the Computer Age*. New York, NY: Harper Collins, 1999. 389.

[121] Smith, Douglas. *Fumbling the Future: How Xerox Invented, Then Ignored, the First Personal Computer*. New York, NY: William Morrow and Company, 1988.

Many of the historical accounts of PARC focus on these early developments and the failure to carry those innovative ideas to market. I have briefly summarized some of these attempts in the previous paragraphs. The rest of this paper is interested in exploring an idea developed later at PARC, that of ubiquitous computing. Unlike the earlier technological advancements, few suggest that PARC failed to commercialize ubiquitous computing. No popular books have been written about PARC "fumbling" the ubiquitous computing market. Perhaps the most obvious reason is that it is still difficult to imagine what ownership of the ubicomp market would look like. The growth in people interested in ubiquitous computing in the 90s, and the establishment of different academic research groups and journals, did not directly translate into a straightforward commercial market. The transition from utility computing to commerical time-sharing was far more straightforward.

I argue that this difference is in large part due to the way that ubiquitous computing was originally framed at PARC. Ubicomp's foundations were infused with a philosophical core that made implementation of the idea difficult. Furthermore, ubicomp was not developed around a single product (like a time-sharing environment or a new technological standard). Instead, all computing devices could be brought under the ubicomp tent. This chapter explores the development of the idea, the source of those commitments, and looks at how those ideas were (or were not) pulled into the cloud.

Weiser and The Idea

Nearly all histories of ubiquitous computing start with Mark Weiser's work at PARC. Weiser's interest in computing started in his junior high school.[122] As an undergraduate student at the New College in Florida, he studied philosophy, a discipline that informed much of his efforts at PARC. After dropping out and working as a programmer in Ann Arbor, he eventually enrolled in the University of Michigan's computer science program earning an M.A. and PhD in Computer and Communication Sciences.[123] After teaching at the University of Maryland (1979-1987), he took a job at PARC in 1987 as a member of the research staff. The following year, 1988, he assumed the role of Principal Scientist and Head of the Computer Science Laboratory at Xerox PARC. Shortly after joining PARC, Weiser talked to a number of PARC researchers, including the anthropologist Lucy Suchman.

Suchman joined PARC in 1979. Her ethnographic work on accounting practices and the use of photocopiers influenced the engineering decisions at Xerox. In describing her role as an anthropologist, she said that "in many ways we think of ourselves more as champions of the mundane. Others dream of far-out widgets. We're saying we really have to give people more useful widgets."[124] This suggestion underscored the need to thoughtfully integrate technologies into everyday life. Weiser picked up on this type of idea while developing his own vision of ubiquitous computing.

During the 1980s, Xerox's financial stability was rocky. The company's stock price fluctuated throughout the decade. Xerox's failure to take advantage of PARC's innovations is frequently

[122] Markoff, John. "Mark Weiser, a Leading Computer Visionary, Dies at 46." *The New York Times*. May 1, 1999.
[123] Weiser, Mark. "Curriculum Vita." February 1996.
[124] Buderi, Robert. "Field Work in the Tribal Office." *MIT Technology Review*. May 1, 1998.

cited as the primary reason for this instability. Furthermore, the departure of many scientists and engineers from PARC was perceived as a "major embarrassment."[125] In addition to the struggles within PARC, broader corporate decisions hurt Xerox stock. For instance, the purchase of an insurance company (Crum & Forster) added to Xerox's decline.[126] Despite these larger issues, the work at PARC continued to develop visions of computing without a direct path to commercialization. However, by the end of the 1980s, Xerox's failings tarnished PARC's reputation. It was under this stormy corporate climate that Mark Weiser introduced his idea of ubiquitous computing.

Weiser first introduced the idea of ubiquitous computing to his colleagues inside the Computer Science Lab at PARC in 1988.[127] Weiser arrived at Xerox in 1987, leaving his teaching job at the University of Michigan. Prior to this career change, Weiser had had an opportunity to work with a few PARC employees on academic projects.[128] In the years leading up to the ubicomp research, he had worked on studying the human dynamic in software design. We can see hints towards ubiquitous computing in his articles regarding the link between humans and programming language. For instance, in 1987 he suggests "people seem to prefer a certain amount of mystery and excitement…dealing with a rich and deceptive world is a basic human skill that should not be denied to humans working as programmers."[129] The notion of mystery ties into Weiser's aspirations of providing an environment that is playful and designed around human psychology. This playfulness extends to the computing code. He suggests that a "ubiquitous source code" might be needed to "discover hidden limitations and hidden strengths" as programing languages continue to develop.[130]

At the start of the 1990s, Weiser started to introduce the idea to the broader public. One of the earliest and most influential papers was the 1991 article in *Scientific American*, "The Computer for the 21st Century."[131] In this article, he articulated his vision of ubiquitous computing and established the design guidelines that many future researchers at PARC adopted. Weiser opened his article with the statement that "the most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it." Throughout the article and in his later work, Weiser underscored the importance of designing technologies with human psychology in mind.

This goal, to design technologies that meld with the human experience, begs the question – to what end? For Weiser, the goal was twofold. First, he seemed genuinely interested in using computers to improve everyday human life. As is often the promise in Silicon Valley, new advancements in (ubiquitous) computing would allow people to work more efficiently and live a more fulfilling life. In the words of Weiser, nothing will be "fundamentally new, but by making

[125] Lueck, Thomas J. "Once a Prodigy, Xerox Faces a Midlife Crisis." *The New York Times*. September, 30, 1984.
[126] Lueck, Thomas J. "Once a Prodigy, Xerox Faces a Midlife Crisis." *The New York Times*. September, 30, 1984.
[127] Weiser, Mark. "Ubi Home." https://web.archive.org/web/20160310012953/http://www.ubiq.com/hypertext/weiser/UbiHome.html (Archived March 10, 2016).
[128] Trigg, Randall H and Mark Weiser. "TEXTNET: A Network-Based Approach to Text Handling." *ACM Transactions on Information Systems (TOIS)* 4, no 1 (1983): 1-23.
[129] Weiser, Mark. "Source Code." *IEEE Computer* 11 (1987): 70.
[130] Weiser, Mark. "Source Code." *IEEE Computer* 11 (1987): 70.
[131] Weiser, Mark. "The Computer for the 21st Century." *Scientific American*. September 1991.

everything faster and easier to do, with less strain and few mental gymnastics, it will transform what is apparently possible."[132]

The second goal, and more pertinent to the development of the cloud, is the development of a technological infrastructure whose strength stems from the ability to oscillate between the visible and invisible. At the center of Weiser's argument is the idea that computers should operate at the periphery until we need them at our attention. In a ubiquitous computing environment, for instance, a user shouldn't have to manually update a computer or retrieve a desired file from another device on the network. Instead, the infrastructure should be smart enough to act without human attention. Conversely, if a user is working on a particular problem, the pertinent information should be brought into the physical world (either through the display or a mechanical apparatus) – what some at PARC called "embodied virtuality."[133]

Early Imagining

Before we take a look at the metaphor and philosophy behind ubiquitous computing, let's first examine early demonstrations and sketchings of how ubicomp was imagined. Throughout PARC's history, the organization has been known for producing ambitious technologies that either failed commercially or were never marketed. The most well-known example is the Xerox Alto, PARC's first attempt at designing a personal computer. In 1973, the Alto was technologically advanced, both in terms of cutting edge hardware and software. However, many later commentators have argued that "Xerox management perceived the Alto venture to be a journey into the unknown, and they failed to seize the opportunity to define and dominate the world of personal computing."[134] This perception of corporate failure was made obvious in the 1980s, in light of Apple's success in the personal computing market.[135] The case of ubicomp follows closely to this previous history.

One of the primary experiments that set the ubicomp projects at PARC into motion was the efforts in 1987 by the Electronics and Imaging Laboratory to fabricate wall-sized flat-panel computer displays.[136] This work was folded into the Computer Science Laboratory in 1988, under the newly formed Ubiquitous Computing program. This large-display project eventually was one of three technologies that formed the foundation of Xerox's efforts. The first, the "LiveBoard" was a large wall-sized display, similar to a television. The other two, the ParcPad and the ParcTab, were smaller computers. The ParcPad was a book-sized computing device, and the ParcTab was a handheld device. These three objects were designed to communicate together to form a computational infrastructure that recognized "the location, situation, usage, connectivity, and ownership of each device."[137] In practice, what PARC was attempting to

[132] Weiser, Mark. "The Computer for the 21st Century." *Scientific American*. September 1991. 104.
[133] Weiser, Mark. "The Computer for the 21st Century." *Scientific American*. September 1991. 98.
[134] Alesso, Peter H. and Craig F. Smith. *Connections: Patterns of Discovery*. Hoboken, NJ: Wiley-Interscience, 2008. 56.
[135] Perry, Tekla S. and Paul Wallich. "Inside the PARC: The 'Information Architects.'" *IEEE Spectrum* (October 1985).
[136] Mark Weiser, Richard Gold, and John Sealy Brown. "The Origins of Ubiquitous Computing Research At PARC in the Late 1980s." *IBM Systems Journal* 38, no. 4 (1999): 693.
[137] Mark Weiser, Richard Gold, and John Sealy Brown. "The Origins of Ubiquitous Computing Research At PARC in the Late 1980s." *IBM Systems Journal* 38, no. 4 (1999): 694.

demonstrate was a model of computing that was detached and abstracted from the traditional computer on the desk.

It would be easy to see PARC's efforts as a precursor to the modern computing environment. There certainly are some similarities. For instance, the rise of "The Internet of Things" (IoT) relies heavily on devices being able to communicate with the local environment. The ParcPad could easily be compared to the iPad. However, this easy connection misses some of the specific design intentions of PARC employees. What set PARC's efforts apart was not the individual components, but how these technologies were building a larger smart environment. Despite this grand vision, PARC did have some limitations. As a company focused on building the office of the future, many of the technologies that PARC demonstrated were based around general office work. The LiveBoard, Pad, and Tab were primitive examples of the types of mobile devices that we have today. The devices were built around productivity tasks (such as sending files to coworkers or increasing other types of office collaboration).

PARC's other projects often strayed into the creative domain. One of PARC's artists in residence, Natalie Jeremijenko, designed the "LiveWire (Dangling String)" in 1995. The artist attached a long string to the ceiling that was attached to a motor. The motor would make the string move when the computer network experienced traffic. A busy computing network would make the string move rapidly. According to the artist, the device was "placed in the spectacularly banal office environment" of Xerox PARC.[138] The project was well-received, in part, because it embodied the spirit of calm technology. Making information physical, while keeping this information on the periphery of human attention, was a clear example of Weiser's aspirations.

While it can be useful to revisit these technologies, I will not do so in depth. Instead, I want to focus on the philosophical underpinnings of ubiquitous computing because they offer the clearest connection to the development of the cloud.

Philosophical Underpinnings

Xerox PARC, from its origins, had a fairly open approach to design. The earliest employee hires at PARC came into the Porter Drive offices (the first PARC location) with a literal blank slate.[139] As the research arm grew, Xerox hired more artists and scholars with backgrounds in the humanities. By the late 1980s, the office culture was well accustomed to incorporating deeper artistic symbolism into their technology. In 1993, this artistic commitment was made formal through the PARC Artist in Residence program (PAIR), which paired artists with scientists to collaborate. In the case of ubiquitous computing, Weiser's idea was inspired by some of the earlier artistic projects at PARC. These artistic projects, along with Weiser's own background in philosophy, led him to inject the idea of selective visibility into ubicomp.

[138] Jeremijenko, Natalie. "Database, Politics, and Social Simulations." *Rhizome Internet.* https://web.archive.org/web/20170602213149/http://tech90s.walkerart.org/nj/transcript/nj_04.html (Archived June 2, 2017).

[139] Hiltzik, Michael A. *Dealers of Lightning: Xerox Parc and the Dawn of the Computer Age*. New York, NY: Harper Collins, 1999. 52-53.

Selective visibility is best defined through the wider lens of visibility. Visibility, as it is understood in social theory, is a way of enabling "the functioning of classificatory infrastructures."[140] Moreover, visibility servers as a means of measuring and exercising power. The pillory that criminals were locked to in 19th century England was visibility enacted as public shaming. Conversely, our walled prisons and solitary confinement use invisibility as a repressive tool, while retaining a clean exterior. Often what we refer to as "clean design" in technology speaks to our ability to engineer aesthetic cloaks around messy technical interiors. Selective visibility, therefore, can be seen as a means of shifting between different states of power, control, and cognitive importance. These elements of power through visibility are at work in the experiments at PARC.

This idea of selective visibility was drawn directly from philosophy. The philosophers that Weiser draws upon share a similar goal of attempting to understand technology's role in unconscious embodied experiences. Previous scholarship has documented how Weiser drew upon these sources. In particular, psychologist Leila Takayama's article[141] on Weiser's philosophical roots looks at Weiser's references in the *Scientific American* publication. In her article, she works on breaking down Weiser's words and describing the philosophical background behind each of Weiser's ideas. After describing these philosophies (which I discuss later), Takayama turns towards the progression of ubiquitous computing after the early experiments at PARC. She suggests that the development of ubicomp after Weiser did not take into account the deeper philosophies and values embedded in Weiser's ideas. By digging into these issues, Takayama suggests ubicomp researchers can refocus on a philosophy that leverages "human experience below the level of focused, conscious attention," while allowing technologies that "simultaneously support and get out of the way of human interpersonal interactions and relationships."[142] This type of analysis can also help us understand ubiquitous computing's relationship to the cloud. Takayama's article offers a solid analysis of Weiser's philosophical influences, to which STS can provide an additional lens layer of analysis.

Ubiquitous computing touches on a number of authors and themes that are commonly found in STS literature. Weiser himself cites a number of authors that could be considered within STS's cannon. Perhaps the most relevant thinker Weiser cites is Michael Polanyi and his book *The Tacit Dimension* or Heidegger's concept of "the horizon" and "ready-to-hand." These philosophies help explain why Weiser wanted to shift the location of computing. Weiser's writings do not go into depth about the specifics of these philosophes and how they apply to ubiquitous computing; however, it is clear that academic theories on the shifting visibility of information were central to the original notion of ubiquitous computing. In particular, the idea of "disappearance."

The notion of disappearance occurs in both contemporary and early STS literature. The history of STS and the core texts of the discipline are largely focused on how scientific knowledge and technological systems are broken down, reconfigured, and packaged once more. An early

[140] Brighenti, Andrea Mubi. *Visibility in Social* Theory. New York, NY: Palgrave Macmillian. 2010. 44

[141] Takayama, Leila. "The Motivations of Ubiquitous Computing: Revisiting the Ideas Behind and Beyond the Prototypes." *Personal and Ubiquitous Computing* 21 (2017): 557-569.

[142] Takayama, Leila. "The Motivations of Ubiquitous Computing: Revisiting the Ideas Behind and Beyond the Prototypes." *Personal and Ubiquitous Computing* 21 (2017): 567.

example of this is Ludwik Fleck's discussions on "thought collectives" as making visible the inherently social aspects of science.[143] After Fleck, the idea of invisibility was probably seen most clearly in the articles and books written on sociotechnical disputes. These case studies are instructive because they expose the rawest moments in the social construction of a technology and scientific theory. Part of the power of these case studies is that they make material what is otherwise ephemeral. These are, at their heart, stories that position the culture of science as disappearing and reappearing.

A focus on materiality in STS occurred perhaps most forcefully in the literature sometimes categorized as "laboratory studies." In Latour and Woolgar's book *Laboratory Life*, the lab is a place where inscription devices transform material into ideas.[144] The specific material arrangements of the laboratory are removed when the experiment is transformed into a research paper. By jumping into the laboratory, STS researchers were able to uncover the everyday performance of science. The later push into Actor Network Theory research expanded these boundaries to materials beyond the laboratory, but they continue to put emphasis on the link between physical environments and ideological construction. This thread runs throughout STS literature. The example of the TEA Laser, for instance, is rooted in the importance of local knowledge and the linkage to the physical laboratory.[145] More contemporary theories, such as Standpoint Theory, place the focus on the embodiment of knowledge.

The attention to the materiality of science and technology is relevant to the story of Xerox. This can be seen in the vision of ubicomp. The ubiquitous computing environment was imagined as more than simply a new type of technology. Instead, the creators of this vision focused on how to build environments that could modify and uplift the human experience. The demonstrations at Xerox PARC focused on the typical office workspace, but the ideology of ubicomp attempted to push beyond the conference room. These technologies were meant to be devices that worked alongside the user, smartly moving into a user's focus when needed. An office with enough technological integration could, it was hoped, increase user productivity and enjoyment. For example, wireless technologies have long been the focus of ubiquitous computing projects because they allow offices to be organized differently by untethering the computer from the desk and allowing more creative organization of labor within the workspace.

By reading the early ubicomp texts, it becomes clear that there was an imagination of a broader technological future. For instance, in a joint article by Weiser and John Seely Brown (then a Chief Scientist at PARC), they discussed the importance of "the periphery" in technologies. In the article, they use the example of a car's engine noise. Typically, when the engine is functioning properly, the noise sits in the periphery. However, an irregular sound quickly brings our focus to the mechanical issue. Weiser and Brown pushed for technologies that could "move easily from the periphery of our attention, to the center, and back."[146] The ability to keep things in the periphery allows humans to do multiple things at once, without the mental exhaustion of

[143] Fleck, Ludwik. *The Genesis and Development of a Scientific Fact*. Chicago, IL: University of Chicago, (1936) 2008.

[144] Latour, Bruno and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton, NJ: Princeton University Press, 1986.

[145] Collins, Harry M and Robert G Harrison. "Building a TEA Laser: The Caprices of Communication." *Social Studies of Science* 5 (1975): 441-450.

[146] Weiser, Mark and John Seely Brown. "The Coming Age of Calm Technology." Xerox PARC. October 5, 1996.

trying to maintain focus on each item in the periphery. The desire for innovations around the periphery was the main motivation for the label "calm technologies."

How successful were they at their goal to design calm technologies? Like previous experiments at Xerox, that depends on how you measure success. In the 1990s, Weiser and his colleagues wrote a number of papers and invented a few demos of ubicomp devices. PARC was the center of innovation of ubiquitous computing and dominated the conversation about the future of ubicomp. This is reflected in a 1994 *Wired* article, where the author proclaimed, "PARC is Back!"[147] The *Wired* journalist was impressed by the fresh, philosophically inspired ideas at PARC. However, the remaining question at the end of the article is whether Xerox would "fumble the future" again. In the article, John Seely Brown suggested that through a process of "co-evolution" PARC could work collaboratively with Xerox's corporate strategy to mesh the research and commercial aims of each side.

As it turned out, the intellectual scope of PARC was deep and spread quickly into academic circles, but the actual material impact on Xerox's business or any other technology firm was smaller than expected. No commercial products were made as a result of the research projects at PARC. Slowly, over the years, the interest in ubicomp within Xerox faded. In 2001, after years of declining profits, PARC was spun-off into an independent company. They were charged with focusing on building profits and a move from "open innovation" to "collaborative innovation."[148] This change in company culture has placed a focus on research ideas that must account for commercial viability.

This is not to say that the history of PARC prior to 2001 was without commercial motives. The history of ubicomp should be viewed with a critical eye. For instance, anyone familiar with Xerox's history could understand that the appeal of this new computing environment was not entirely altruistic. From a business standpoint, being able to sell new technologically-enhanced office machines outside of the computer could be a new untapped market. This is not to mention the potential for an omnipresent panoptic tool in the office. An "always-on" environment could just as easily double as a new tool for squeezing productivity out of workers.

Despite this potential critique, I believe that as far as Weiser and his team were concerned, their efforts were truly benevolent. Weiser's writings clearly show a desire to make human life more enjoyable. Sometimes these experiments manifested themselves in small ways, for instance, a system that could monitor the coffee pot to notify users over the local network when a fresh pot of coffee was made in the office.[149] In other more substantial visions, Weiser dreamed of a playful vision of computing. In a concise letter writing on the importance of invisibility, Weiser wrote: "I propose childhood: playful, a building of foundations, constant learning, a bit mysterious and quickly forgotten by adults. Our computers should be like our childhood: an invisible foundation that is quickly forgotten but always with us, and effortlessly used throughout our lives."[150] This vision of childhood and the disappearance of computing points to Weiser's

[147] Rheingold, Howard. "PARC is Back!" *Wired*. February 1, 1994.
[148] Talbot, David. "The Comeback of Xerox PARC. *MIT Technology Review*. December 21, 2011.
[149] Schilit, Bill N., Norman Adams, and Roy Want. "Context-Aware Computing Applications." *IEEE Workshop on Mobile Computing Systems and Applications*, December 8-9, 1994. 4-5.
[150] Weiser, Mark. "The World is Not a Desktop." Perspectives Article for ACM Interactions. November 7, 1993.

desire to reduce the complexity of computing (particularly interfaces that distract our attention), while still trying to build systems that engage the human spirit. The focus to bring humans into harmony with computers through selective visibility can be seen throughout his writings.

It is important to note, however, that for as much as the name ubiquitous computing suggests a world of computing everywhere, Weiser did not express much desire to inject smart technologies into all landscapes. The ubiquity implies an intentional approach to technological environments over a deluge of smart devices that add little value to the everyday experience of life. Looking at the references to Polanyi and Heidegger, we can clearly see the importance of the tacit side of ubicomp. The focus on the materials of everyday life is a more grounded approach to a rather lofty technological goal.

What is perhaps the most instructive lesson from the work at PARC is the intentionality of the ubicomp projects. It seems to me that Weiser is acutely aware of the power of black-boxing a technology. Ubiquitous computing, as Weiser imagined it, was trying to create a technological environment that draws upon the power of invisibility. Weiser asked designers of these technologies to reimagine the relationship between the user and the computer. In doing so, he also wanted these designers to weave his own philosophy into these new technologies. In Weiser's mind, it was not simply enough to build a "smart" coffee maker. Instead, to be truly ubiquitous computing, that new invention needed to mesh into the broader network of technological devices. Additionally, that coffee maker would integrate itself into the human environment, disappearing when coffee is not needed and appearing when a caffeine jolt is necessary. There is a touch of paternalism in this design logic. The employees of PARC seemed to assume that their technology could be shaped for individuals rather than individuals being active in shaping their own intention relationship with devices.

Even on a very simple level, Weiser was concerned primarily with reducing the amount of noise in the human environment. In "The Computer for the 21st Century," he states that "most important, ubiquitous computers will help overcome the problem of information overload. There is more information available at our fingertips during a walk in the woods than in any computer system, yet people find a walk among trees relaxing and computers frustrating. Machines that fit the human environment, instead of forcing humans to enter theirs, will make using a computer as refreshing as taking a walk in the woods."[151] The comparison to a walk in the woods is an attempt to align ubicomp with a more natural and intuitive way of computing. The metaphor of walking in the woods is perhaps a bit awkward, as it presumes a leisurely pace and safety without needing to pay attention to potential dangers around you. However, the larger point that Weiser is attempting to make is that ubiquitous computing should be designed to give users information when they request it, but not overwhelm them. Weiser's push for a calmer computing future, however, was not fully adopted in the years following the introduction of the term.

Ubicomp Beyond Xerox

Xerox PARC's experiments with ubicomp never died, but the energy and enthusiasm for the projects diminished over time. Weiser continued to push for the idea until his untimely death

[151] Weiser, Mark. "The Computer for the 21st Century." *Scientific American*. September 1991.

from cancer in 1999.[152] Interestingly, Weiser, for as much as he has been honored as the "father" of ubiquitous computing, started to be more protective of the phrase towards the end of his life as he witnessed the way the term could be misused. In a letter to a professor in the MIT Media Lab working on physical computing, he said, "my request is that you help me stop the spread of misunderstanding of ubiquitous computing based simply on its name. Ubicomp was never just about making 'computers' ubiquitous. It was always, like your work, about awakening computation mediation into the environment."[153] The revolutionary idea, as Weiser described it, was found in the structural changes to the infrastructure that enhanced everyday life.

Academics and mobile technology companies were some of the first to pick up the idea of ubicomp. In 1997, the journal *Personal & Ubiquitous Computing* was founded and took up the idea of detaching computing from the desk. Initially, there was a real focus on everyday technologies. As the journal expanded, the focus turned to broader issues, but Weiser's influence is still felt in the journal. In 1999, one of the first ubicomp conferences was held (then called Handheld and Ubiquitous Computing). At the first symposium, the majority of articles discussed ubiquitous computing in terms of building connected environments. A theme that ran throughout the first conference was a question of how to build a connected environment. Many of the speakers addressed the potential for adding sensors to new environments and measuring life outside of a traditional computing office space, such as the other spaces in a home or measuring the flow of information in a city.[154] Weiser is mentioned in many of the articles, but his broader vision was largely ignored.

Outside of academia, the vision of ubicomp that PARC proposed largely was ignored. Instead, the focus of commercial business was, and continues to be, increasing the connectedness of different technologies. Mobile devices, such as Personal Digital Assistants, became the new site for spreading computing everywhere. For instance, in 1999, Nokia dominated the mobile device market and sponsored various ubicomp research (including the previously mentioned conference). In an interview with *Wired*, the CEO of Nokia framed pervasive computing in terms of shifting computing from the desktop to the mobile device. In the same article, a researcher from Nokia sees these new mobile technologies as connecting us to our "herd," suggesting that "pervasive wireless communication will bring us back to the behavior patterns that were natural to us and destroy behavior patterns that were brought on by the limitations of technology."[155] This vision of computing sits uncomfortably alongside Weiser's vision of a similar "natural" relationship to technology.

As mobile devices continued to grow, the focus on pervasive or ubicomp faded from both the academic and commercial arenas. Journals that focused on ubicomp increasingly grew specialized and less connected with the philosophy that inspired PARC's idea. Mobile devices continued to develop along with the maturation of Web 1.0 service. The dawn of Web 2.0 (a term used to describe an interactive and re-writable internet) pushed many companies to focus on

[152] Want, Roy. "Remembering Mark Weiser: Chief Technologist, Xerox PARC." *IEEE Personal Communications*. February 2000.

[153] Ishii, Hiroshi. "Bottles: A Transparent Interface as a Tribute to Mark Weiser." *IEICE Trans. Inf. & Syst.* 87, no. 6 (2004): 1310.

[154] Gellersen, Hans-W. *Handheld and Ubiquitous Computing: First International Symposium, HUC'99*. Lecture Notes in Computer Science. Germany: Springer-Verlag Berlin Heidelberg, 1999.

[155] Silberman, Steve. "Just Say Nokia." *Wired* September 1, 1999.

building new digital environments. In the early 2000s, it seemed that ubicomp had fallen by the wayside.

From Ubicomp to the Cloud

As I have alluded to, the importance of the history of ubiquitous computing reemerged with the creation of the cloud. The cloud had repeatedly been framed as a new type of pervasive computing. Increasingly, the cloud is becoming married to the idea of the Internet of Things. In this union, we see Xerox's initial vision starting to spread but with a modified narrative about the harmony between humans and technology. The cloud is seen to be the new computing environment that negotiates the conversations between all of our computing devices. The integration of the cloud with our local materiality has greatly increased in recent years. Smart home devices, like the Google Home or the Amazon Alexa, are the new physical embodiments of the cloud that act as a local bridge between the cloud and our local environments.

I argue that the cloud has developed with little focus on the actual history of ubiquitous computing. I have attempted to capture some of that history in this chapter. However, as much as I think the specific history is worth retreading, perhaps more important is looking at the fundamental ideas and commitments that the researchers at PARC introduced. By looking back at the work at PARC, we can highlight three interconnected concepts that tie into the history of ubicomp and the cloud.

One of the first concepts and design narratives is that these technologies should pay attention to the disappearance of technologies. It is perhaps more accurate to describe these technologies in terms of "selective visibility." PARC researchers were focused on designing technologies around the idea of the periphery. Many products that are connected to the cloud are marketed in terms of their ability to be invisible from the user. However, invisibility is not the same as technologies in our periphery. There is a value in being able to selectively view infrastructure, to be able to move back and forth between our direct attention and our background. Weiser suggested that a technology should come into focus when it is useful for a user. However, many cloud services are designed to obscure the messiness that might expose some inconvenient truth. There are, of course, benefits of being able to hide the messy aspects of a technology from a user. Clean designs can be more usable and approachable; however, for many consumers of the cloud, this type of selective visibility is not an ability that is offered.

Related to this notion of invisibility is PARC's focus on the materiality of ubicomp. The importance of a physical connection to these smart technologies was critical to the researchers at PARC. Many of the experiments at PARC were focused on the connection between the user and the environment that he or she existed in. The value of the technology was not found in an external dataset but in a desire to create a harmonious local network of devices. Much of the cloud today is still not focused on the actual materiality of computing environments. Instead, off-site data centers have shifted the space of computing to a more centralized location (an ironic twist from the original narrative about the cloud's ability to decentralize computing). When the cloud does focus on the materiality of local spaces (such as smart sensors), many of these devices are unable to communicate with one another because of siloed ecosystems.

Finally, the history of ubicomp is often concerned with the concept of harmony in the design of "calm technologies." A commitment to designing calm technologies is rarely present in modern technologies, including the cloud. However, perhaps this idea from Weiser needs to be reintroduced into conversations regarding the cloud. The cloud is marketed as a smart technology that enables greater efficiency. Perhaps it is worth considering if the cloud could not be updated to include a calming component. A calm cloud could be defined as a distributed computing system that adapts to the expectations and desires of a user while remaining flexible to expansion or shrinkage. A calm cloud would need to focus on the interoperability of competing cloud systems so that a user could easily migrate his or her information between clouds. This calmness would be reinforced through standards and intelligent user experience design choices. This would require cloud businesses to rethink how information is captured and presented to the users. In turn, it would also require a reworking of the business model of cloud computing, as the current design of the cloud may interfere with the creation of calmness.

That said, I believe this notion of calmness actually meshes easily with the push for smart devices that are meant to use the cloud to facilitate everyday life. However, any move to build calm clouds should also take into account PARC's design philosophy that these technologies ought to give users control over their own computing environment. The solution to calmer technologies may not be in adding more sensors to build a smarter cloud. Instead, the focus could be on building more intentional sensors that allow for a calm cloud. This would mean perhaps being clearer about the type of data gathered and limiting where and how that information is collected. Furthermore, it may slow down the flow of data, but it would also be a system that respects user preferences and larger geographic laws and customs.

Any effort to build a better cloud by revisiting the history of Xerox PARC should keep these ideas of disappearance, materiality, and calmness in mind. PARC's vision of ubicomp offers insights into the way that the cloud can be modified to create a calmer, more intentional, technology. For as much as the cloud is described to be a realization of ubiquitous computing, the adoption of the philosophy behind the concept is lacking. The ubiquity of the cloud as it sits today is not the type of ubiquity that Weiser or many researchers at PARC imagined. It is important to highlight this distinction as the cloud continues to mature.

The following chapters take this history of ubiquitous computing and utility computing to address the sociological and political ramifications of the development of the contemporary cloud.

Chapter 3

Iconography, Coinage, and Formation of the Cloud

In the previous chapters, I highlighted two forms of computing that set the stage for the emergence of the cloud. In the first instance, I looked at the notion of utility computing. Utility computing offered a metaphor of computing beyond the local workstation. The metaphor of the computer utility presented many with a vision of computing that mirrored other large infrastructural systems. However, the idea lay dormant after the 1970s. In the second instance, I examined the way that ubiquitous computing developed in the late 1980s and early 1990s. Xerox PARC's work on a new form of computing focused on developing intelligent computing environments, but their efforts were largely were limited to small research prototypes. Despite this, the metaphor of ubiquity had a lasting effect on the development of the cloud.

In this chapter, I turn towards the birth and early development of the cloud. I start by discussing the iconography of the cloud and early instances of the cloud as a visual computing symbol. I then move to the coinage of the term cloud computing and the atrophy of the term in the late 90s. This discussion is followed by the rebirth of the cloud in 2006 and the maturation of the idea in later years. Throughout this chapter, I pay particular attention to the metaphor of the cloud and the ideology behind the early cloud.

There is particular focus in this chapter on the transition from the visual cloud, to "cloud computing," and finally, "the cloud." This is an important shift from visual, to adjective, and noun. The creation of "the cloud" is both a result of the culmination of networking technologies and a marketing campaign to flatten complexities and package a new computing marketplace. The positioning of the cloud, as a noun, gives this new computing infrastructure a sense of agency and importance. Curiously, even as cloud computing technologies become more clearly defined and delineated, the idea of "the cloud" remains mysterious. I often use these terms interchangeably because their meanings are linked together. However, the reader should keep in mind the distinction between "the cloud" as symbol, descriptor, and noun.

The primary purpose of this chapter is to point to the ambiguities found in the early representation of the cloud (in symbol form) as a means of dismantling the notion that the cloud's meaning is fixed and stationary. Rather, the visual language of the cloud was in flux before the term was coined. Additionally, even after the cloud was introduced as a concept, there were numerous social actors who attempted to come to terms with what this new technology category ought to represent. Through these dialogue and early cloud experiments, a new vision of computing was melded together. In the process, the ideas of utility and ubiquitous computing were deployed, but ultimately the nuance of these early histories was not included in our modern notion of the cloud.

Cloud computing is an umbrella term that has attempted (and continues to attempt) to bring together multiple technologies, metaphors, and ideologies. This ability to expand and encompass a multiplicity of realities is one of the reasons that the idea of the cloud has persisted even as the underlying technologies have changed. By looking at the first examples of cloud computing, it is evident that the cloud was never created to be a static object. Instead, like an atmospheric cloud, it can expand, shrink, and move with relative ease. Much of this flexibility seems to stem from the loose adoption and implementation of the cloud.

The original iconography of the cloud often stood to mean anything outside the control of a local computing network. As this chapter attempts to argue, much of the contemporary rhetoric about the cloud mirrors this original use of the symbol. The cloud is rarely a technology that individuals interact with. Instead, the connection to the cloud is a mediated experience that can intentionally or unintentionally obscure the connection between the user and the physical cloud. The language of the cloud, as well as some of the early cloud symbols, borrowed heavily from a marketing approach. I follow the evolution of the symbol and the term, to the implementation of the idea to demonstrate how many of the promises of the cloud are premised on certain assumptions about how users ought to use remote computing services. These assumptions are rooted in the notion that users should consume the cloud through a new type of computing relationship. In this agreement, the user is considered a temporary consumer of computing resources with the implicit bargain that a certain level of control will be released in exchange for an easily managed product. The iconology of the cloud feeds into this narrative by providing a visual symbol of computing outside of a user's control.

Ultimately, the historical period captured here demonstrates the fading of visible geography and the establishment of a new cloud identity. The invisibility of cloud computing emerges from the creation of "the cloud" as a marketing tool. Even as new data centers were established to fortify the infrastructure's backbone, the materiality of the cloud was not reflected in the public view. Rather, the materials of the cloud only helped cloak the growth of the term. Whereas the older symbols of the cloud presented us a vagueness without direct intention, the new symbol is intentional and full of action.

I argue that early cloud computing was conceived as a technology that meshed many technological visions while retaining a large amount of interpretative flexibility. It is important to pay attention to this moment in computing history because it helped frame the limits of what the cloud is and to what extent the cloud exercises control over the users. The early cloud was largely structured as a technology designed for individuals, but controlled by large technology companies. In part, this corporate control over the cloud was a byproduct of the emerging technological infrastructure that companies like Google and Amazon were building in the early 2000s. How the concept was introduced also demonstrates that the rush to the cloud was also motivated by a desire to control an emerging market in the networked computing space. This conversation is framed by a particular focus on the role of ideology in symbols, maps, and metaphors.

Iconography

In order to trace the origins of the cloud, it is important to look at both terminology and visual symbols. For cloud computing, the origins of the idea can be found in previous computing technologies. However, the visual representation of the cloud happened in a much more subtle fashion. The symbol of the cloud can be traced back to early maps of computer networks. These early network maps were largely a collection of lines, circles, and boxes.

Maps and symbols are, of course, political, and different disciplines have looked at the meaning behind them. Critical theorists have long taken issue with cartography as a means of naturalizing the cartographer's perspective as universal.[156] Reacting against modernism's objectivism, many scholars have turned towards a study of "postmodern geographies," which reject the universality of maps.[157] Critical cartography tries to read the deeper meanings behind the symbol to deconstruct the dominant discourse.[158] These discussions of cartography feed into a larger discussion of the role of symbols in constructing identity. For instance, Benedict Anderson's discussion of national identity as an "imagined community" speaks to this point. This can be seen in Anderson's discussions of maps as a form of print capitalism that binds a diverse community into an imagined whole.[159] Theories similar to Anderson can be found throughout the social sciences.

Broadly speaking, the social sciences have looked at the importance of how we visualize metaphors. Metaphors and symbols are ways of grouping incoherent systems into an understandable whole. In George Lakoff and Mark Johnson's *Metaphors We Live By*, they make the argument that "spatialization" is often a key feature of metaphors. [160] Abstract art, for instance, is regarded as being "high art," whereas crude humor is found at the "bottom of the barrel." This is important to keep in mind as we look at the history of the cloud. The iconography of the cloud is tied closely to the metaphor of the cloud. The visual language often places the cloud on a different spatial level. The written metaphor of the cloud only helps to solidify this distinction. Today, the visual and written languages of the cloud continually influence each other in a process of co-production. Early icons of the cloud, however, existed as purely visual culture in the form of maps and diagrams.

Within STS, the role of symbols and knowledge maps are central to core theories. A classic example of STS dealing with these topics is found in the construction of a scientific fact or theory. Theories map the work of numerous actors into a solid whole. In understanding how an idea is turned into a scientific fact, STS researchers unpack the social dynamics that link different actors together. Kuhn's theory of "scientific revolutions" is one example of how different paradigms drive towards ideological consensus. [161] Other scholarship has looked more closely at the politics of scientific citations (another form of a knowledge map).[162] More broadly, STS has continually turned towards networks as a way of understanding how sociotechnical symbols can be deconstructed. This effort to understand networks is found throughout STS literature, most obviously in works utilizing Actor Network Theory. ANT, along with other STS work, has made the claim that symbols are political and bound to a historical moment unremarkable.

[156] Wood, Denis. "The Map as a Kind of Talk." *Visual Communication* 1, no. 2 (2002): 139-161.

[157] Soja, Edward W. *Postmodern Geographies: The Reassertion of Space in Critical Social Theory*. New York: Verso, 1989.

[158] Harley, John Brian. "Deconstructing the Map." *Cartographica* 26, no. 2 (1989): 1-20.

[159] Anderson, Benedict. *Imagined Communities: Reflections on the Origin and Spread of Nationalism.* London: Verso, 1983.

[160] Lakoff, George and Mark Johnson. *Metaphors We Live By*. University of Chicago Press: Chicago, 1980. 17.

[161] Kuhn, Thomas S. *The Structure of Scientific Revolutions. Chicago:* University of Chicago Press, 1962.

[162] Small, Henry G. "Cited Documents as Concept Symbols. *Social Studies of Science* 8, no. 3 (1978): 327-340.

Looking at the meaning of a symbol is important for how we link metaphor to imagery. A symbol, in the most simple definition, is a marking or object that holds meaning. The power of a symbol is found in the layers of meaning that sit behind the symbol itself. The defining characteristic of a symbol is that the meaning is relational. Whereas a metaphor is a comparison, a symbol is a representation. Symbols have no meaning outside a social context and often differ depending upon the audience. The study of symbols has been central to many academic fields, most obviously in semiotics. For the purposes of this chapter, ANT provides the primary theoretical framework for understanding symbols. ANT, sometimes called a material-semiotic method, does not make a clear distinction between language and real "things." Rather a hybrid approach is favored, which extends "semiotics to things instead of limiting it to meanings."[163] A focus on networks is particularly useful when looking at the meaning of the cloud as a symbol as part of a map. Maps are ways of stitching symbols together to generate meaning. The lines on a map create meaning through symbols. These symbols on maps suggest who is inside of a network, point to relationships between actors, and exclude others. ANT's focus on enrollment of actants and a commitment to symmetry are additional reasons why it is a useful tool for unpacking the history of the cloud.

In the field of Internet Studies, these theories are starting to be put into action. Maps of digital networks are starting to be critiqued. [164] How computer networks represent the (in)visibility of data is of particular interest in the case of early cloud mapping. Because the rhetoric of cloud computing draws upon a selective visibility, it is important to treat the cloud as a symbol that has a real material impact on geography. Additionally, this idea of a shifting visibility is important because it speaks to the themes of knowledge and control. The early symbols of the cloud presented one version of control for a specific set of actors (primarily those involved in early computer networking). As the symbol of the cloud is transferred over to the modern cloud, the iconology of the cloud is placed in a new context with a different set of actors and economic dynamics. This transposition of the cloud allows for a new type of invisibility for cloud marketing and adoption.

In order to discuss the politics of internet maps, it is useful to start with the first packet-switched wide-area computer network, ARPANET. The first maps of ARPANET were simple drawings that resembled a wiring diagram. Black lines connected each node to the larger network. In a map from 1973, these lines connect boxes and circles.[165] Zig-zag lines, similar to a lightning bolt, represented a satellite link between two universities. Other ARPANET maps placed these lines on top of a map of the internet. Each university or network hub was accurately placed, but the lines connecting each point were only an approximation of the actual wires that connected the computers together. These maps served a few important functions. Most importantly, they allowed the viewer to see the actors on the network clearly. If you wanted to understand how UCLA and MIT were linked, the map offered a clear path. However, these maps also served an

[163] Latour, Bruno. "On Actor-Network Theory. A Few Clarifications Plus More Than a Few Complications." *Philosophia* 25, no 3 (1990). 11.

[164] Dodge, Martin. "Understanding Cyberspace Cartographies: A Critical Analysis of Internet Infrastructure Mapping." Thesis UCL. April 2008.

[165] Cheng, Selina. "What the Entire Internet Looked Like in September 1973." *Quartz*. December 12, 2016. https://web.archive.org/web/20170630052044/https://qz.com/860873/a-1973-map-of-the-internet-charted-by-darpa/ (Archived June 30, 2017).

alternative purpose. They give a clear structure to a physical network and a sense that the network is tied into itself.

It is also important to remember that the network maps represent a particular viewpoint of the map maker. In the case of ARPANET, many of the maps were made by network engineers from a defense contractor. The contractor, BBN, helped develop some of the core technologies of the network and used the data from their Network Control Center to produce the maps. BBN developed logical maps (lines and boxes that show the connections between the computers) and topological maps (geographic approximations of where in the United States the network stretched). These maps hide as much as they show. Internet historians have critiqued these maps for not showing a sense of flow, gateways, or hierarchy in the network.[166] In the early ARPANET maps, there is little indication of the icon of the cloud. The lack of detail of a local computing environment (instead focusing on the broader network) points to a location where a cloud symbol could start to be used. Other technologies and networks were the first to employ the rounded edges of a cloud.

In the early history of networked computing, the majority of network maps held onto the rigid lines and rectangles. Still, there were instances that demonstrated the potential for artistic flexibility. We see this with the introduction of curved lines that eventually would transform into the cloud symbol. These clouds were first represented as fluid lines. One of the first instances of fluid lines was used in AT&T's video conferencing service called "Picturephone" (1964).[167] The map, in this instance, was visually less like a cloud and more like a wavy circle.[168] The curves in the map represented a short-hand for describing all of the networks and systems that were outside the control of the network operator. The cloud started to take more shape in the decades following the Picturephone's release. The significance of the Picturephone's map, and later iterations of curved lines, was the way that it starts to disrupt the dominant visual map of cables linking different locations together.

In the 1970s, researchers in the United Kingdom started to develop the Joint Academic Network (JANET). The goal of the researchers at the time was to link the UK's computing centers into a single network.[169] In 1984 a network topology map showed the connection between JANET and the UK's Rutherford Appleton Laboratory (RAL).[170] In this map from the Science and Technology Facilities Council, there is a cloud-like symbol inside the RAL Local Area Network. The network map shows a line connecting JANET to RAL. In this instance, the cloud is the way that the computers on the RAL network connect to one another. However, there are no cloud symbols in the early JANET maps (they do appear later). This suggests that the cloud symbol is a form of abstraction. Computer that exist outside of the cloud are specific and identifiable. Conversely, the details of the equipment inside the cloud shape are considered unimportant.

[166] Fidler, Bradley and Morgan Currie. "Infrastructure, Representation, and Historiography in BBN's Arpanet Maps." *IEEE Annals of the History of Computing*. July-September 2016.

[167] Hu, Tung-Hui. *Prehistory of the Cloud* Cambridge, MA: MIT Press. X.

[168] Dorros, Irwin. "The Picturephone System: The Network." *Bell System Tehcnical Journal* 50, no 2 (February 1971). 232.

[169] Wells, Mike. "JANET – the United Kingdom Join Academic Network." *Serials* 1, no. 3 (November 1988).

[170] JANET. "Network Maps." *http://jam.ja.net/marketing/janet30years/network-maps/*

The symbol of the cloud was not used exclusively in network maps; it was also used in more general computer literature. Tim Berners-Lee, the inventor of the World Wide Web, used the symbol of a cloud in his original proposal distributed to his colleague at CERN.[171] In his proposal, he uses arrows, nodes, and clouds. Berners does not explicitly mention the use of the cloud, but his diagram uses the symbol to represent a general idea or topic. Non-cloud circles represent something specific, such as a location or specific technology. The cloud, in this instance, does not have any inherent meaning. Instead, it is a category that points to something other than itself.

The cloud icon did not see real popularity until the 1990s. A clear example of a cloud icon can be seen in the National Science Foundation Network (NSFNET) map from 1991. NSFNET was originally developed as a way to link the NSF's supercomputer centers together in the 1980s.[172] In 1991 the NSFNET produced a map of the United States with a cloud symbol linking the locations together.[173] The cloud is labeled "Backbone Infrastructure" to indicate the role that the network had in connecting these research locations together. The cloud stood in as the heart of the network. The symbol spoke to a broader change occurring in the management of the internet. The successful development of NSFNET marked the end of ARPANET (ARPA helped the transition to NSFNET) and a transition towards a commercial internet.[174] In a map made four years later (1995), after the transition towards commercial ownership of the internet backbone, the NSFNET was still represented as a cloud that hovered over the nation.[175] However, the symbol of the cloud stood in to be a Network Service Provider that provided the connections to smaller regional providers. The cloud, both as a national symbol and a means of grouping regional providers, suggests a symbolic language of power and an attempt to control the boundaries of computing. This power is enacted through the placement of the cloud above the nation and as omnipresent force that links individual computing locations together. Through these maps, the symbol of the cloud establishes computing boundaries by defining all other forms of networked computing in relationship to the cloud.

Usage of the cloud symbol increased on the verge of the term being coined. One example can be found in the National Press Building Network Topology map. Lines and symbols linking computers show the connection between the internal office technologies.[176] Four clouds can be found over networks outside of the building. In this instance, the cloud represented a connection to an external network. Another example can be seen in the Energy Sciences Network high-speed computer map in 1994, which introduced a cloud around the letters "ATM." ATM stands for asynchronous transfer mode, which is a network transferring approach for linking the

[171] Berners-Lee, Tim. "Information Management: A Proposal." *CERN*. March 1989.

[172] Rogers, Juan D. "Internetworking and the Politics of Science: NSFNET in Internet History.´ *Information Society* 14 (1998): 213-228.

[173] Dodge, Martin. "An Atlas of Cyberspaces: Historical Maps of Computer Networks." https://web.archive.org/web/20171025085724/http://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/atlas/historical.html (Archived October 25, 2017).

[174] Abbate, Janet. *Inventing the Internet*. Cambridge, MA: MIT Press. 1999. 194-195.

[175] Goldstein, Steven N. "Future Prospects of NSF's International Connections Program Activities." ISOC.org May 3, 1995.

[176] Archive.org. "National Press Building Network Topology." 1995. https://archive.org/details/RTFM-D-19950115-itr

computers on ESNET.[177] Here the cloud represents a technology, rather than a specific place (in later maps, it was specified that the cloud represented the Internet Service Provider QWEST). Despite these early instances, the symbol of the cloud was not commonly used. In a 1994 report that focused on methods of mapping networks, the symbol isn't found.[178] The vast majority of computer networking maps continued to use non-cloud symbols. Still, these early examples point towards the meaning that sits behind the symbol of the cloud. They also provide an interesting contrast to recent NSF maps of the cloud, which lack detail or location specificity.[179] This lack of detail suggests cloud's spread across the entire network.

The representation of the cloud in computing has shifted as different actors have embraced the iconography of the cloud. In the case of these early symbols, the cloud was used to highlight a number of different meanings. In certain instances, the cloud symbol is used to indicate a system that is beyond the control of the viewer. In other instance, the symbol is a tool of abstraction that removes the specificity of technologies within a network. As I will discuss later, the use of the cloud is less diverse today. Large cloud arrays are privately held and are not built to link smaller clouds together. Rather, the plurality of clouds seems to be shrinking as large cloud providers consolidate the majority of the cloud market. Maps of cloud computing centers today are visual representations of infrastructural tools that can be tapped into (rather than traversed upon). These early visual symbols of the cloud are important to pay attention to because it demonstrates the malleability of visual symbols and the impact those symbols can have on our understanding of technological politics. Furthermore, these early visual images demonstrate that the issue of the cloud as a place of control and power is not a new occurrence.

The metaphor of the cloud, as I have discussed in the previous chapters, is a way of giving flexibility to a complex technological system. Clouds are seen as both material and ephemeral, depending on who is drawing the picture. Symbols shift in meaning as different actors imprint their own interpretation on the system. In the same way that utility or ubiquitous computing has taken on different meanings from their historical origins, looking at the changing visual nature of clouds can potentially help us analyze, contextualize, and critique the continued development of the cloud today.

Coinage

Despite these early visual symbols of the cloud, the term "cloud computing" was not coined until 1996.[180] Two men, George Favaloro and Sean O'Sullivan, are generally credited with the term cloud computing. Favaloro worked for the computer hardware company Compaq. Favaloro, a marketing executive, was looking for a way to sell Compaq computers to Internet providers. O'Sullivan, the founder of a startup called NetCentric, was in discussions with Favaloro about a partnership between the two companies. In their negotiations about a five million dollar

[177] ESNET. "Historical Network Maps: ESnet Backbone Topology Map Winter 1994." https://www.es.net/engineering-services/the-network/network-maps/historical-network-maps/

[178] Suhadi. "Thesis: Design of a Computer Network to Improve Information Quality for the Indonesian Army." Naval Postgraduate School. September 1994.

[179] National Science Foundation. "Learning Cloud Providers Join with NSF to Support Data Science Frontiers." *National Science Foundation*. February 7, 2018. https://www.nsf.gov/news/news_images.jsp?cntn_id=244450&org=NSF

[180] Regalado, Antonio. "Who Coined 'Cloud Computing'?" *MIT Technology Review*. October 31, 2011.

investment, they imagined a new way of selling computing services over the Internet. NetCentric's "Point-of-Presence software" was meant to introduce a new way of running "software inside the internet."[181] In describing this new approach to remote software, they coined the term cloud computing.

The cloud was initially framed as a broad marketing term. In a draft of a press release, Compaq stated that the partnership with NetCentric would allow "ISPs to better address business communications needs and dramatically increase revenues by offering metered pay-as-you-go services, such as Internet-based fax, voice, video conferencing, and file management services."[182] The press release also underscored the value of cloud computing as being a "universal platform" that could run on any system. This universality seems primarily focused on providing ISPs a choice in how to deploy cloud computing services, not on a notion of egalitarian access. ISPs, which did not sell computing hardware to users, saw potential in being able to indirectly sell access to computing power remotely on top of their traditional internet utility service. The primary user of the cloud was meant, at least initially, to be a business user. The notion that the cloud was initially a public resource is decidedly untrue.

Much of this focus on targeting business users can be attributed to Compaq's own business goals. In a 1996 internal document, Compaq's Internet Solutions Division sought to target Network Service Providers as a way to "transition from providing basic services today (access only) to providing value-added services."[183] Of course, it is important to note the context in which the idea of the cloud was being proposed. These discussions occurred just prior to the dot-com bubble. Compaq was looking to invest in new compelling technology companies. In general, there was a new acceptance of these startups as a new growth mechanism. Personal computer ownership had risen in the early 90s, which helped fuel these investments in novel ideas. In 1997, 36.6% of American households owned a computer, and 18% of those had access to the internet.[184] These numbers, while small today, were a dramatic increase from the start of the 90s.

Compaq wanted a way of managing chaos during this turbulent time. In a later interview, Favaloro said that "computing was bedrock for Compaq, but now this messy cloud was happening and we needed a handle to bring those things together."[185] The cloud, from its inception, was a term that was meant to pacify and control an unwieldy technological environment. The term was, and is, a way of linking together many different services under a broad umbrella. Many of the services proposed in 1996 were aspirational and not truly possible at the time. Nearly all consumer internet users were using slow modems before the spread of broadband access. The information infrastructure to support these demanding applications simply would not be robust enough until a few years later. Favaloro and O'Sullivan's idea of the cloud, however, would prove to be the antecedent to the public emergence of the cloud in the middle 2000s.

[181] Smith, Gordon. "O'Sullivan Credited with Coining 'Cloud Computing." *Irish Times*. November 4, 2011.
[182] Compaq. "Draft 3: Compaq and NetCentric Partner to Deliver Universal Internet Platform." January 27, 1997.
[183] Compaq. Internet Solutions Division Strategy for Cloud Computing. CST Presentation. November 14, 1996.
[184] U.S. Census Bureau. "Home Computers and Internet Use in the United States: August 2000." U.S. Department of Commerce Economics and Statistics Administration. September 2001.
[185] Regalado, Antonio. "Who Coined 'Cloud Computing'?" *MIT Technology Review*. October 31, 2011.

Silent 00s.

From 1997 to 2006, there are very few references to cloud computing. During this time period, however, a number of large technological changes were occurring that helped set the strange for the emergence of the cloud. Some of the most important changes during this period involved shifting network technologies. Two of these technologies that I will highlight are the spread of high-speed internet access and virtualization.

In the previous chapters, I have argued that utility computing and ubiquitous computing were two major predecessors to the cloud. In the case of utility computing, one of the primary limitations was the lack of a strong information infrastructure. The success of many early time-sharing companies, like Tymshare, depended on access to the networks of telecommunication companies. Even if a timesharing company leased telephone lines, access to the network could be slow and expensive. Furthermore, these new information networks could not be accessed by the majority of people or small businesses. Those with the financial means to access the service were still limited by the amount of bandwidth that could be carried over the telephone lines. It wouldn't be until the start of the 21st century that broadband service would become commonplace in the United States, and the network would have sufficient bandwidth to handle data intensive tasks.

Similarly, in the case of ubiquitous computing in the early 1990s, many of the technological standards and network infrastructure were not yet robust enough to support common cloud computing tasks. Xerox's work on ubicomp played an important role in envisioning a densely interconnected local computing environment. What this vision didn't fully articulate was the wide-scale networks that would make these connections both possible and valuable. For instance, when Weiser imagined ubicomp, Wi-Fi computing standards had not been put into place, and the majority of desktop or laptop devices were hardwired. The expense of producing small computers also limited the commercial possibility of ubicomp devices. Perhaps more important, most computing tasks were localized. The software that Xerox imaged was local and not built to exist "in the cloud."

Shortly before the term cloud computing was reintroduced to the public, a number of key cloud technologies were invented or refined. The refinement of "virtual machines" (VM) in particular helped spur the development of early cloud applications. Unlike a traditional computer, a virtual machine can exist entirely as a separate container within a computer. One powerful computer can run multiple virtual machines that can be used by multiple users (an end-user may not even be aware they are sharing the same computing hardware). Virtualization was a core component of early timesharing and proved to be an efficient means of sharing computer processing and storage capabilities. In the early 2000s, virtualization expanded beyond the rudimentary uses that early time-sharing systems employed. One key improvement was the ability to place a gap between the physical hardware of the computer and the user, which allowed a number of important applications. One of these applications was the creation of robust VMs that could emulate the experience of using a computer locally. Multiple instances of a computer's operating systems could be run on the same computer.

VMware, which was later acquired by Dell, released their influential VM software in 1998.[186] By 2001, their *VMware Server* software allowed a single processor to run up to 20 virtual machines at once.[187] Developments in virtualization helped push the industry towards a consolidation of computing resources. Without virtualization, the cloud could not exist in its current form. A core feature of the cloud is that cloud services are dynamic. What virtualization allows is a separation of hardware from software and for those resources to be allocated dynamically. On a technical level, this is significant because it allows for flexibility in allocating computing resources. The more significant political aspect of this separation speaks to the untangling of the user from control over remote resources (typically servers). This move away from local computer management has had a large impact on the IT workforce that historically has worked with computers locally and is consequently changing the location and labor pool of these professions. Ordinary users have also seen large shifts away from local control of programs and computing resources, even as computing infrastructures attempt to erase the boundaries between the local and remote.

The development of virtualization occurred alongside major changes in the public internet. Changing network architectures made remote connections more commonplace, even amongst non-business users. In the 1970s, telecommunication companies used a network protocol called X.25 to carry voice other telephone lines.[188] This allowed for more users to share infrastructure to make "virtual calls," which could be billed depending upon the usage of the network.[189] The X.25 protocol suite was eventually supplemented by the frame relay standard. These standards gave way to the use of Virtual Private Networks that could transmit data securely while using the public internet infrastructure. This meant that the emerging web could serve as a secure means for remote computing. As more users took to the internet, web-based businesses started to disrupt many industries, both for technology firms and non-digital businesses.

Part of this change was a new wave of "Web 2.0" companies that could challenge the dominance of older software and hardware companies such as Microsoft. Google and Amazon, two titans of the cloud today, were developed with the help of these more dynamic technologies created during the Web 2.0 boom. The economics of the information technology industry changed. Companies that relied upon computer hardware and software sales were challenged by these new internet companies. In turn, search and e-commerce businesses demanded a more flexible computing infrastructure in order to continue growing. These businesses started to work on internal systems that resemble the cloud infrastructure that we are familiar with today. That said, neither Amazon nor Google made an effort to sell their internal services to outside organizations. Instead, small companies started selling cloud computing services without the cloud terminology.

[186] Yoffie, David B. "VWware, Inc., 2008." *Harvard Business School Case* 709-435. 2008.

[187] VMware. "Tool for Server Consolidation." October 2001. https://web.archive.org/web/20170830032954/https://www.vmware.com/pdf/vmware-dhbrown.pdf (Archived August 30, 2017).

[188] Mitchell, Bradley. "A Guide to X.25 in Computer Networking." May 28, 2018.

[189] International Telecommunication Union. "D.11: General Tariff Principles." May 31, 1991. https://web.archive.org/web/20181130124635/http://www.itu.int/rec/T-REC-D.11-199103-I/en/ (Archived November 30, 2018).

In 1999 the startup Salesforce offered a modern preview of the cloud. Salesforce was founded with the stated goal of "delivering essential business applications and services via the Internet."[190] Today Salesforce is a cloud computing company. The primary aim of Salesforce was to develop a cloud business around the idea of "software as a service" (SaaS). Businesses could tap into Salesforce's servers running software remotely. An early slogan that the company promoted was "no software." The slogan implied that software would move away from local desktops and towards a continually updating software product in the cloud.

Salesforce marketed their product as a way of empowering and building a continuing relationship with consumers. In 2004 the Chief Customer Officer at the time, Jim Steele, pitched the benefits of Salesforce. In an interview, he said, "when people buy software, they have no choices. By renting software as a service as we do, they always have a choice."[191] Steele's comments came in reaction to the traditional desktop software business model where the software is sold once, and there is less incentive to support the software on a long-term basis. Salesforce's actions were an early signal of the move away from ownership of software. This was one of the company's primary goals, to create a business around perpetual rental of software over ownership.[192] This particular notion of "choice" is tied to the idea that consumer freedom comes from the ability to switch between software, while always being under the umbrella of a large virtualized computing environment. This rental economy has since become one of the cornerstones of the cloud.

Alongside the growth of Web 2.0 companies, and the push for SaaS at Salesforce, the physical infrastructure of the internet was being improved. Major investments from the dotcom boom increased investment in fiber optic cables, both on land and at sea. Many of these investments were financial risks, but they have enabled many companies to build their cloud on top of a data-rich resource. These previous projects often are overlooked in the history of the cloud but are a critical part of the story. Writing about submarine cable investment in 1996, writer Neal Stephenson argued that "once a cable is in place, it tends to be treated not as a technological artifact but almost as if it were some naturally occurring mineral formation that might be exploited in any number of different ways."[193] In the case of the cloud, these resources were drawn heavily upon.

These improvements to the internet's infrastructure and the changing economic possibilities for online businesses helped create an environment where remote computing could thrive. This period also marks the start of a more widely commercialized internet and a new pool of users. At the start of the 21st century, many jobs that could previously only be practically done locally (such as storing or transferring large files) could now be executed remotely. The web was rapidly changing, and the second emergence of the cloud loomed just around the corner.

[190] Salesforce.com. "Company Profile." March 2, 2000. https://web.archive.org/web/20000302221044/http://www.salesforce.com:80/info/company.html (Archived March 2, 2000).

[191] CRMBuyer. "Salesforce.com President Jim Steele: Why 'No Software' Is Good Business." February 16, 2004. https://web.archive.org/web/20170426091452/http://www.crmbuyer.com/story/32866.html (Archived April 26, 2017).

[192] Kirby, Carrie. "Marc Benioff / Swimming against the dot-com tide." *SFGate*. September 3, 2002.

[193] Stephenson, Neal. "Mother Earth Mother Board." *Wired*. January 1, 1996.

(Re)Introducing the Cloud

The first real public introduction of the cloud occurred at the Search Engine Strategies Conference in August 2006. Eric Schmidt, then the Chairman and CEO of Google, discussed a new model of computing. He argued that most IT companies "were build up around a model where you had a PC client and then a set of services, Unix, OS2, Windows, etc., and a lot of proprietary protocols between those."[194] This type of computing is a more direct sales relationship between the vendor and the consumer. Schmidt then reintroduced the idea:

> It starts with the premise that the data services and architecture should be on servers. We call it cloud computing – they should be in a "cloud" somewhere. And that if you have the right kind of browser or the right kind of access, it doesn't matter whether you have a PC or a Mac or a mobile phone or a BlackBerry or what have you – or new devices still to be developed – you can get access to the cloud. There are a number of companies that have benefited from that. Obviously, Google, Yahoo!, eBay, Amazon come to mind. The computation and the data and so forth are in the servers.[195]

In 2006 this proposal symbolized a turn away from the dominance of older IT business models and towards a new type of network architecture. It was a shift that many, including Schmidt, predicted. In 1993, Schmidt had said that "when the network becomes as fast as the processor, the computer hollows out and spread across the network." He added that the true profits will not go to "the companies making the fastest processors…but to the companies with the best networks and the best search and sort algorithms."[196]

Google was not alone in seeing the benefits of shifts. Amazon, for instance, started working on building their internal network as an eventual public infrastructure resource. From 2003 to 2004, a small team at Amazon started to work on a plan to standardize Amazon's retail infrastructure and potentially sell virtual servers to the public.[197] Two years later, Amazon introduced its "S3" service, which is a cloud storage solution (although the language of the cloud is absent in the initial press release).[198]

[194] Schmidt, Eric. "Conversation with Eric Schmidt hosted by Danny Sullivan." *Search Engine Strategies Conference.* August 9, 2006. https://web.archive.org/web/20180430231055/https://www.google.com/press/podium/ses2006.html (Archived April 30, 2018).

[195] Schmidt, Eric. "Conversation with Eric Schmidt hosted by Danny Sullivan." *Search Engine Strategies Conference.* August 9, 2006. https://web.archive.org/web/20180430231055/https://www.google.com/press/podium/ses2006.html (Archived April 30, 2018).

[196] Wired. "The Information Factories." October 1, 2006. https://web.archive.org/web/20180504162934/https://www.wired.com/2006/10/cloudware/ (Archived May 4, 2018).

[197] Black, Benjamin. "EC2 Origins." January 25, 2009. https://web.archive.org/web/20170801095615/http://blog.b3k.us/2009/01/25/ec2-origins.html (Archived August 1, 2017).

[198] Amazon. "Amazon Web Services Launches." March 14, 2006. https://web.archive.org/web/20181208153555/https://press.aboutamazon.com/news-releases/news-release-details/amazon-web-services-launches-amazon-s3-simple-storage-service/ (Archived December 8, 2018)

In the months and years following Schmidt's interview, a number of companies adopted the language of the cloud. In August of 2006, Amazon announced its "Elastic Compute Cloud" service that has become the bedrock for many cloud computing environments today.[199] Google released its App Engine in 2008, Google Storage in 2010, and eventually its Google Cloud Platform in 2011. Nearly all major tech companies announced some type of cloud product during this time frame (such as Microsoft's Azure, Apple's iCloud, and IBM Cloud). The language of the cloud gave technologists and the public a term that encapsulated this wide-scale transformation. This language was not formed by accident. The cloud, from the first coining of the term, was a marketing decision. The language of the cloud makes the business model of the cloud more abstract and detached from the actual services being rendered and the infrastructure that is being drawn upon.

During the early 2000s, the implementation of the cloud was primarily focused on taking advantage of existing corporate computing infrastructures. This was especially true for "Web 2.0" companies with large amounts of underutilized compute. As the concept started to take hold in the late 2000s, the idea of the cloud started to spread beyond the idea of simply rentable virtual computers. In 2008, the vice president of Hewlett-Packard was one of many who saw the cloud as a new computing landscape where "the search for information will be done for you, not by you."[200] In this viewpoint, the transformation is about making "everything a service."

Much of this change in rhetoric came from pursuing new markets. While the market for cloud computing services continued to grow during this time period, the overall cloud market started to open up to a non-technical audience. Statements, like the one made by HP's vice president, were early attempts to frame the cloud as beneficial for the general public because they offer a more managed and intelligent delivery of computing resources. Of course, this comes at the cost of turning more control over to the operator of the cloud.

For the most part, the cloud was largely unproblematized during this period. General consumer knowledge and adoption of cloud services were comparatively small compared to today. Still, there were some who warned of the changing computing landscape. Richard Stallman, the creator of the operating system GNU and the Free Software Foundation, spoke harshly in 2008 of the cloud. He warned that the use of the cloud leads to a loss of individual control and is driven by marketing logic.[201] Likewise, Larry Ellison, founder of Oracle, bashed the idea of the cloud as a "fashion-driven" fad which amounts to "complete gibberish."[202] Ironically, Oracle's main business model is in the cloud, and Ellison has since changed his opinion.

A Changing Linguistic and Visual Landscape

[199] Amazon. "Announcing Amazon Elastic Compute Cloud (Amazon EC2) – beta." August 24, 2006. https://web.archive.org/web/20141027130331/http://aws.amazon.com/about-aws/whats-new/2006/08/24/announcing-amazon-elastic-compute-cloud-amazon-ec2---beta/ (Archived October 27, 2014)

[200] Robison, Shane. "The Next Wave: Everything as a Service." *HP*. February 2008. https://web.archive.org/web/20171214015942/http://www.hp.com/hpinfo/execteam/articles/robison/08eaas.html (Archived December 14, 2017).

[201] Johnson, Bobbie. "Cloud Computing is a Trap, Warns GNU Founder Richard Stallman." *The Guardian*. September 29, 2008.

[202] Reuters. "Larry Ellison Sounds Off on Cloud Computing." *CIO Insight*. September 26, 2008.

By adopting the language of the cloud, organizations also tie themselves to the history and visual culture of the cloud. Turning towards the metaphor of the cloud was not a neat, simple, or organized process. Instead, the move to the cloud was (and continues to be) a messy and complicated process. The initial adoption of the cloud was focused on developers and businesses that wanted access to a robust computing infrastructure. Later, as the cloud developed, it was marketed to general consumers as a product to be lived in or consumed (rather than buying access to). Through the early years of the cloud, the meaning of the cloud has been contested and tweaked. Despite this complexity, there has been a move towards accepting the symbol of the cloud as part of the internet's makeup.

When organizations and companies first introduced the cloud to their potential audience, they framed the cloud in different ways. Much of the initial marketing of the cloud was aimed at developers. One of the possible reasons that developers were the initial audience was the perception that they would adopt new technologies more readily. Perhaps, more importantly, is that many of the initial cloud services were products that appealed to developers' needs. Developers and other IT professionals often use the cloud differently than a typical user because their work draws more upon virtualization and other large-scale applications. This is reflected in how the cloud was marketed and draws upon the history of utility computing. For instance, early into the rebirth of the cloud, the CEO of Amazon Jeff Bezos posed the question: "You don't generate your own electricity. Why generate your own computing?"[203] Comparisons to the electrical grid are ways of treating the cloud as an infrastructural resource and is also a way of speaking to the previous histories of utility computing.

By taking a look at the first cloud products, it is clear that the cloud was marketed differently to technologically savvy developers over a general consumer. The developer's cloud is transparent and specific, whereas the general user is vague and abstract. This was certainly the case for Amazon's EC2 service. Interestingly, the symbol of the cloud was not particularly prominent in early marketing materials. In 2008 Amazon's webpage describing the service spent little time describing what the cloud is and was more focused on the benefits of their service for developers.[204] Unlike contemporary cloud services, the description of what the service is and the technological underpinning of the service are straightforward. The primary marketing message is focused on reliability and cost-effectiveness. The cloud "instances" are described in terms of the hardware specifications of the computer and where geographically the cloud service is located. This is important for a number of reasons, primarily because it reduces latency to the end-user. The location of the server is also important when building platforms to conform to different legal systems (either on a state or national level).

This approach to marketing the cloud resembles the original imagination of a computer utility service. Computer service is billed by usage, and the user has more freedom to do what they would like on the system. In some ways, this vision of utility computing is more expansive because users are not limited to a single computer but can easily scale-up. Additionally, like the relatively open time-sharing platforms from the 1960s, Amazon offered many open-source

[203] McFedries, Paul. "The Cloud is the Computer." *IEEE Spectrum* (August 2008).
[204] Amazon. "Amazon EC2." December 8, 2008.
https://web.archive.org/web/20081208084613/https://aws.amazon.com/ec2/ (Archived December 8, 2008).

software choices for developers on EC2. By partnering with Sun Microsystems and Red Hat Linux, Amazon was able to offer an open platform for users and not force users into a closed ecosystem.[205] These features suggest that perhaps certain values of a utility computer vision were built into these early cloud computing products.

Where the example of Amazon's EC2 system doesn't match the story of utility computing is the political issues that plagued early utility systems and the open dialogue about how the system ought to be regulated. The academic researchers in the 1960s imagined a computing infrastructure service that looked quite different than the private service that Amazon offers. Questions of ownership, regulation, and access are absent from the contemporary conversation. Instead of universities operating and directing the fate of these networks, these new infrastructures are wholly owned by private capital. Due to changing regulatory frameworks, these private networks were able to grow quickly during the early 2000s. One of the primary stumbling blocks of the utility computing conversations in the 1970s occurred as the FCC moved to regulate the telecommunication services during that time. The cloud of today seems to escape these types of regulatory investigations because it doesn't appear on the outside to be a natural monopoly, largely because it is geographically dispersed (not a product of a single country) and not seen as a critical backbone of the modern web. Ultimately the question of regulation was put aside as the internet turned towards commercialization and towards a new type of utility computing.

In the late 2000s, the cloud started to mature, and the audience for the cloud expanded to general consumers. During this period, there was a noticeable shift in how the metaphor of the cloud was deployed. This shift can be seen in the new consumer-facing products that were marketed at the time. Products like Google Apps, Dropbox, or Apple's iCloud were all released around the same time. These services fall into the Software-as-a-Service (SaaS) category. Unlike IaaS or Platform-as-a-Service (PaaS), the user of SaaS applications is completely separated from the hardware and underlying applications of the cloud. Instead, the user is totally enclosed in a virtualized space. While SaaS solutions were not first invented during this period, the significance and ubiquity of SaaS solutions grew as large technology companies build out their cloud infrastructures. At the same time, users were more likely to own multiple devices and wanted a way of syncing their files across computers and accessing their data on-demand.

In 2011, Apple announced its iCloud service to replace its similar, but less functional, MobileMe product . During iCloud's unveiling Steve Jobs, CEO of Apple, painted a picture of the cloud as a means of creating harmony between various devices. In describing the service, he said that "we are going to demote the PC and the Mac to be a device…and we are going to move the digital hub, the center of your digital life, into the cloud."[206] Arguing against the idea that the "the cloud is just a hard disk in the sky," iCloud is the product that will weave all of your products together so that "it all just works." [207] This vision of computing speaks to the dreams of some ubiquitous computing advocates who wanted a seamless connection between multiple devices.

[205] Stanford University School of Engineering. "Amazon Enters the Cloud Computing Business." May 20, 2008. 11.
[206] Apple. "WWDC 2011 – iCloud Introduction." 2011.
[207] Apple. "WWDC 2011 – iCloud Introduction." 2011.

The visual symbol of the cloud played a prominent role in Apple's announcements. In the presentation, the icon of the cloud hovers above all of the computers and mobile devices of the potential consumer. Instead of drawing permanent lines between the devices and the cloud, small lines beam up and down information from the cloud and the hardware. The cloud ties the ecosystem together, but the visual language is an effortless, invisible connection. The cloud symbol can also be found in individual apps. The "Documents in the Cloud" app has a silver cloud inside a document. In other apps, like Photo Stream, clouds can be seen floating in the background. The entire presentation gives the feeling that the specter of the cloud floats with unrelenting persistence in the Apple ecosystem.

Not mentioned in Apple's presentation are the underlying systems of control that this vision of the cloud offers. The cloud not only allows for the easy syncing of files and preferences, but it also hooks Apple's technological infrastructure into your personal devices (assuming they are devices made by Apple). The convenience of the iCloud comes at a cost, both a yearly fee and a broader social cost. By uploading your digital life to the cloud, there is a large incentive to remain tied to the cloud. As the iCloud system has matured, it has become necessary to attach oneself to the cloud to have access to the latest features. The allure of the cloud is not exclusive to Apple, and the story of iCloud can be seen to a lesser degree in Google and Microsoft's own cloud offerings.

In these SaaS examples, we get the strongest indication that the symbol of the cloud serves as a shorthand for describing technology that users interact with but is largely outside their control. The cloud industry is full of the cloud symbol floating over devices and locations. For instance, in an advertisement for Amazon's "cloud player," a smiling cloud sits between the consumer's workplace and home. The cloud player is touted as "your own online, secure, personal music space."[208] In another example, Google Drive does not use a cartoon cloud but represents the Google SaaS applications as floating in an abstracted grey background.[209] The narrator tells the listener that "now all your stuff is in one place, easy to find, and easy to share." It is not mentioned where this place is. In other cloud advertisements, the Google Play store (where Google sells movies, music, television shows, books, and applications) is shown linked to the Google cloud drive without reference to the actual materiality of the Internet and the servers that allow the distribution of this content.[210]

Changing Weather

Looking back at the development of the cloud from its early beginning as a visual symbol, we can start to understand the political implications behind the strategic deployment of the cloud. It is important to note that throughout this chapter there has never been a monolithic symbol. There were, and continues to be, contradictions about where the cloud sits visually and how the metaphor should be interpreted. I have attempted to give a sketch of the broad changes in the visual and rhetorical culture of the cloud. Rather than weakening the impact of the symbol, I

[208] Amazon. "Introducing Amazon Cloud Player." November 1 2012
http://www.youtube.com/watch?v=77ZU92tltOY&feature=plcp
[209] Google. "Go Google: Google Drive." Apr 24, 2012.
http://www.youtube.com/watch?v=wKJ9KzGQq0w.
[210] "Google Play Test #0923." Jul 13, 2012. http://www.youtube.com/watch?v=gqzoX280-yQ.

think that this ambiguity is what gives the cloud its power and flexibility. That said, I believe there has been a noticeable shift in the dominant ways that the cloud symbol is deployed.

In the earliest examples, the cloud did not float over the entirety of devices and locations. Instead, the symbol often stood in for something that the network map designer did not have control over. The wavy lines in AT&T's picturephone gave way to the geographic placement of the cloud over the early computer networks. In the UK, JANET connected to a remote cloud at another research site. Here, the symbol of the cloud indicated either a connection to another network or a network that was controlled by a different network operator. Early network maps, before the cloud was coined, did not use the symbol in the same way, but no instance that I have come across uses the cloud as an omnipresent symbol as it is commonly used today.

After the coinage of the cloud, the metaphor and the symbol took a turn towards a more universal approach. A cloud started to erase geography. The data centers and computers that the cloud sits above are hidden behind the shine of new consumer products. In the instances where the materials of the cloud are not hidden, those products (primarily IaaS and PaaS services) target technologically savvy users that want access to the power technological infrastructures that were developed in the early 2000s. This difference in the way that these services are marketed speaks to the politics of the cloud. Those that pay for access to computing power are more closely tied to the geography of the cloud. In this instance, the symbol of the cloud plays a diminished role. In those SaaS programs, users have less control over the underlying infrastructure and little idea of where the infrastructure is located.

Understanding the history of how the symbol and the language of the cloud have changed reveals the power of metaphors in directing technological possibilities. The use of the cloud was not a historic inevitability. The history demonstrates that the cloud, as an idea, grew from a marketing term. The marriage of the symbol and language after 1996, along with changes in technological standards and internet infrastructure, helped make the symbol of the cloud a powerful metaphorical tool. The ambiguity of the cloud hides the complex histories of the computer legacies that it is built upon (primarily utility and ubiquitous computing). By recognizing the social construction of this symbol, we can start to unpack the quickly closing cloud-shaped box.

Turning towards the contemporary moment, the metaphor and symbol of the cloud have only become more dominant as more companies have made aggressive turns towards the cloud. Nearly all major tech companies have recognized the economic necessity of controlling the information infrastructure that connects businesses to their customers. In the following chapter, I will discuss the significance of this shift by examining the material infrastructure that is hidden behind the rhetoric and visual symbols of the cloud.

Chapter 4

Materials of the Cloud

The politics of the cloud are found in the materials of the system, as well as the representations of that materiality. The cloud, both as an idea and a set of technologies, could not exist without some type of underlying physical infrastructure. The idea of the cloud does not exist in a vacuum. The visions of computing that the previous chapters have addressed have largely been discussed in terms of ideas, but it is important to point out that these visions of computing could not have been realized without the invention and deployment of computing hardware. The history of the cloud follows this same story. Without the production and installation of hardware, the cloud would not exist in its current form today.

Understanding the history of the cloud requires looking beyond specific hardware inventions. Many histories have been written about specific computers or companies that had a hand in promoting the idea of the cloud. These histories are valuable but are somewhat ineffective at studying large-scale changes. In order to address the development of the cloud, we need to look at the broad changes in information infrastructure that allowed the different actors to model what the cloud is or what it ought to be. This involves an understanding of the networks that the idea of the cloud attaches itself to. These points of attachment provide us with a framework of the cloud, a way of understanding the actors in the network, and the politics hidden under the promise of an abstracted form of detached computing.

Highlighting this attachment is important because it makes the ephemeral solid. To understand the material history of the cloud is to see the cloud as something inextricably linked to the internet's infrastructural history. The cloud is spread across geographies but exists in real spaces. The data centers that hold the main processing and storage power of the cloud are the many hearts of the cloud's anthropomorphized body. To carry the metaphor further, the arteries of the cloud are seen in the thousands of miles of fiber optic cables that tie different digital actors together. Throughout this technological organism, there are also organic actors that work to maintain the health and well-being of the network.

It is impossible to draw the boundaries of the cloud accurately. This chapter makes no attempt to do so. Instead, this section seeks to understand the relationship between the ideology and the materials of the cloud. Previous chapters have attempted to dismantle the natural order of the cloud by highlighting the historical incongruences of the cloud's prehistory. This contribution supplements those discussions by looking at the material impact of the cloud. Pointing out the physical nature of the cloud doesn't discount the strong impact that visions of computing can have. Instead, the focus on the material only highlights further the way that the metaphorical cloud and the cloud infrastructure are continually co-producing each other. As technologies change, the vision and boundaries shift as well. Likewise, changing technological visions impact the real material of the cloud.

This chapter starts by first addressing how scholars in material culture, STS, and internet studies have understood the relationship between materiality and ideology. After addressing those theoretical roots, I turn towards an overview of cloud infrastructure and prior work that has been done to address the materiality of the cloud. Finally, to illustrate how the material infrastructure of the cloud is embedded in the history of politics, I focus on recent investments in underwater submarine communication cables by technology companies. Throughout this chapter, I underscore that grappling with the politics of the cloud requires diving into the broader network that links multiple clouds together.

Previous chapters have argued that cloud is a blend of computing ideologies that have been meshed into a contemporary whole. The way that the cloud has been framed, either in the case of ubiquitous computing or utility computing, has often overlooked the material reality of these computing technologies. The metaphor of the cloud continues to suggest a form of computing detached from space. In this chapter, the physical aspects of the cloud are discussed; I argue that the material infrastructure of the cloud is central to the current and future possibilities that the cloud allows. If we continue to separate the cloud from the networks that maintain its function, we will have an incomplete understanding of the potential possibilities, issues, and challenges. The politics of the cloud resides not only in what the cloud represents. Instead, the politics live in the network, in the cables that carry the signals and the human labor that maintains those infrastructures.

Public Clouds

Before delving into the specifics of the material cloud, it is worth looking at the dominant cloud discourse. There have been different attempts to define what cloud computing is. Perhaps one of the most pervasive definitions comes from the National Institute of Standards and Technology (NIST), which identifies five key features: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.[211] The NIST definition further clarifies what the cloud is by looking at different service models of cloud computing (Software/Platform/Infrastructure as a Service [SaaS/PaaS/IaaS]) and the deployment model (Private, Community, Public, and Hybrid). This rather wide definition of the cloud has left quite a bit of flexibility for different actors to model their own idea of what the cloud is.

The NIST's definition is a product of many technical researchers and information technology professionals working on the boundaries of the term. The stated purpose of the definition is to "serve as a means for broad comparisons of cloud services and deployment strategies, and to provide a baseline for discussion from what is cloud computing to how to best use cloud computing."[212] The definition is primarily focused on the structure of services (such as how usage is billed or what IT product is being sold). What this approach does not capture are the larger infrastructural characteristics that make up the cloud. For instance, there is not a clear distinction between "the cloud" (as a global network) and "clouds" (multiple competing networks). This distinction is important because it brings up issues of how the cloud is actually deployed and to what extent interoperability should be expected between providers. Rather, the definition focused primarily on the way that cloud computing can function on many layers.

[211] Mell, Peter and Tim Grance. "SP 800-145." September 2011. https://csrc.nist.gov/publications/detail/sp/800-145/final

[212] Mell, Peter and Tim Grance. "SP 800-145." September 2011. 1. https://csrc.nist.gov/publications/detail/sp/800-145/final.

IT professionals have long used the language of “layers” as a metaphor for understanding different aspects of computer networking.[213] This community has also been responsible for much of the clarification regarding SaaS, PaaS, and IaaS. In both the NIST’s definition, and definitions found more broadly, there is a fuzzy relationship between the software and the hardware of the cloud. Cloud infrastructure, as the NIST sees it, is both “a physical layer and an abstraction layer” where the “abstraction layer sits above the physical layer.”[214] This is similar to the Open Systems Interconnection model that segments computing and telecommunication systems into different conceptual layers (typically seven) from the physical to the application.[215] Likewise, the NIST sees the cloud through this lens but the layers are not as neatly segmented. The physical layer, according to the NIST, includes any hardware resources to support the cloud services. These physical resources, however, continually bleed into the definitions of the cloud’s software layers (especially in the case of IaaS).

When we start looking at the language of the cloud outside of a technical framing, the definitions are more general. The definition of the cloud changes largely depending upon the audience. For instance, when talking to users of AWS (a developer-focused product), Amazon defines cloud computing as “the on-demand delivery of compute power, database storage, applications, and other IT resources through a cloud services platform via the internet with pay-as-you-go pricing.”[216] This differs from more consumer-focused definitions which often touch upon some of these criteria, but are less nuanced. In the most causal of definitions, the cloud is a place that exists beyond your local computer.

One of the common trends that can be noticed throughout the technical and more casual definitions of the cloud is an underappreciation or recognition of the materials of the cloud. The discussions of the cloud as physical often take place amongst information technologists that are either deploying their own private cloud or have the need to visit their local cloud data center physically. For certain jobs, such as working physically inside an Internet Exchange Point or managing the hardware of a data center, the cloud becomes immediately physical.

The following section of the chapter looks at the materials of the cloud through academic theories from STS and infrastructure studies. The purpose of this section is to build a picture of how the cloud exists in the real world alongside the common definitions of the cloud. What I attempt to show is that an understanding of how the cloud sits upon the natural landscape can have an impact on the ideological debates about the future of computing. Without at least a somewhat clear picture of the infrastructure of the cloud, the conversations about the cloud will ultimately miss the material impact of this new technological arrangement.

[213] Arms, William. “The Early Years of Academic Computing.” *The Internet-First University Press.* 2017.

[214] Mell, Peter and Tim Grance. “SP 800-145.” September 2011. https://csrc.nist.gov/publications/detail/sp/800-145/final

[215] Zimmermann, Hurbert. “OS1 Reference Model-The IS0 Model of Architecture for Open Systems Interconnection.” *IEEE Transactions on Communications* 28, no 4 (1980).

[216] Amazon. “AWS: What is Cloud Computing.” https://web.archive.org/web/20180717200739/https://aws.amazon.com/what-is-cloud-computing/ (Archived July 17, 2018).

Theorizing A (Physical) Landscape

It is helpful to look at some of the literature in STS and infrastructure studies as a means of giving context to the creation of the cloud. This section starts by looking at the materiality of infrastructure and moves towards a more general overview of materials in STS. A common theme that flows throughout these different academic disciplines is the idea that to understand how politics are embedded in objects and systems, we need to look at the relationship between networks. Recognition that "artifacts have politics" is nothing new in STS, but the recognition of the idea is still powerful when coupled with the understudied side of technological infrastructures.[217]

Within STS, the question over materiality has been one of the central themes and points of fracture within the discipline. In the 1960s and 70s, the sociology of scientific knowledge (SSK) emerged as a discipline. Following the influential works of Merton and Kuhn, sociologists of science began to question the extent to which social judgments constructed sociotechnical systems. The "strong programme," primarily lead by sociologists at the Edinburgh School, looked at how scientific theories developed from social positions (not universal truths).[218] Critiques of the strong programme argued the core tenants of the theory (causal, impartiality to truth and falsity, symmetry, reflexivity) leave too much room for absolute relativity.[219] Perhaps most saliently, some argued that the strong programme ignored too much of the materials of science.[220] Early theories on the social construction of science have been criticized for failing to account for science in practice.[221] Moves towards the study of the lab and other ethnographic work were, in part, responses to this critique.[222]

The co-discipline of STS, the History of Technology, has historically tried to inject discussions of the physical into theory. In *Technology and Culture's* introductory article, Melvin Kranzberg, lamented the humanities focus on ideas over things.[223] Instead, Kranzberg and his contemporaries shifted towards the material aspects of things. This focus on looking at the materials of the past helped inform the dominant view of their field. This viewpoint, to put it simply, states that technological histories have a weight and resonance that needs to be grappled with. This can be seen especially in the literature on failure. It is not so simple to write the "failure" of a technology, whether that is a wooden airplane or video conferencing system from the 1960s.[224] [225] Instead, you need to look long-term at these inventions to measure the ripples these technologies produce. Historians of technology have long attempted to show how culture and technology co-produce each other. Furthermore, understanding the process of co-production

[217] Winner, Langdon. "Do Artifacts Have Politics?" *Daedalus* 109, no. 1 (1980).

[218] Bloor, David. *Knowledge and Social Imagery*. Chicago: University of Chicago Press, 1976.

[219] Serin, Funda Neslioglu. "The Strong Programme and the Rationality Debate." *Kilikya Felsefe Dergisi* 2 (2017): 41-50.

[220] Bloor, Daivd. "Anti-Latour." *Studies in History and Philosophy of Science* 30, no. 1 (1999): 81-112.

[221] Latour, Bruno. *Science in Action*. Cambridge: Harvard University Press, 1987.

[222] Hess, David. "Ethnography and the Development of Science and Technology Studies." In *Handbook of Ethnography* (ed. Atkinson, Coffey, Delamont, Lofland, and Lofland) Los Angeles, CA: SAGE, 2011. 6.

[223] Kranzaberg, Melvin. "At the Start." *Technology and Culture* 1, no. 1 (1959): 4.

[224] Schatzberg, Eric. *Wings of Wood, Wings of Metal: Culture and Technical Choice in American Airplane Materials, 1914-1945*. Princeton, NJ: Princeton University Press, 1999.

[225] Lipartito, Kenneth. "Picturephone and the Information Age: The Social Meaning of Failure." *Technology and Culture* 44, no. 1 (2003): 50-81.

isn't simply a matter of analyzing the "black box" and seeing what is inside. The general consensus has been that "to explain history with technology in it, you had to explain the technology…" and "even if the technology was socially constructed, one had to know how it was constructed, what it did, and why it operated that way. In short, one had to unpack the black box. That was what historians of technology do."[226]

Unpacking the black box isn't simply a metaphor for taking apart the ideas of a technology. In certain cases, it requires literal unpacking of dusty boxes to understand how social meaning was infused into the materials of a technology. More contemporary scholarship in STS and the history of technology have carried their work forward by continuing to demonstrate how science and technology are socially constructed, using various techniques, methods, and ideological lenses. For instance, Actor Network Theory has done a great deal to demonstrate how politics are embedded into networks and widen the area of study for many. ANT can be thought of as a "material-semiotic tool," where the "materials" are the people, technologies, and other non-human actors.[227] Likewise, feminist theories such as standpoint theory[228] or Haraway's conception of the cyborg[229] have repeatedly emphasized the importance of the lived experience as a means of understanding broader politics in and outside of science. Other movements within STS have placed the spotlight on the nature of things, places, and the significance of those physical realities on sociotechnical ideologies.

Within this literature, there are a few key ideas that are of particular use when looking at the materiality of the cloud. Actor Network Theory, as a whole, is a useful tool for understanding the development of the cloud as a new global digital network. In particular, the concept of "translation" can help set the stage for the spread of the cloud. Translation, as Callon originally framed the concept, involves "…creating convergences and homologies by relating things that were previously different."[230] The process of translation is largely rooted in a geographic approach because the analysis is not fixed to a single place.[231] Instead, the concept speaks to the never-ending process of seeing actors and networks in a dance of co-construction. This can be seen in the creation of the cloud, which was not a natural creation. Instead, the mapping of the cloud was a process by which different actors had to translate different networking standards, computing equipment, and software into an amalgamated whole. In this chapter, the idea of translation can be seen in all of the attempts to marry the cloud to a particular region, while still attempting to sell the vision of the cloud as a unified whole.

Finally, it is important to consider how ideology manifests itself physically. Ideology refers generally refers to a system of normative ideals. The role of ideology in shaping our world can be traced back to early arguments from Karl Marx that ideology works to uphold social orders by

[226] Roland, Alex. "What Hath Kranzberg Wrought? Or, Does the History of Technology Matter?" *Technology and Culture* 38, no. 3 (1997): 700.
[227] Law, John. "Actor Network Theory and Material Semiotics." In *The New Blackwell Companion to Social Theory* edited by Bryan S. Turner. Oxford: Wiley-Blackwell, 2008.
[228] Hekman, Susan. "Truth and Method: Feminist Standpoint Theory Revisited." *Signs* 22, no 2 (1997): 341-365.
[229] Haraway, Donna. *Simians, Cyborgs, and Women: The Reinvention of Nature*. New York: Routledge, 2015.
[230] Callon, Michel. "Struggles and Negotiations to Define What is Problematic and What is Not." In *The Social Process of Scientific* Investigation, ed. Knorr, Krohn, and Whitley. London: Dordrecht, 1980. 211
[231] Barry, Andrew. "The Translation Zone: Between Actor-Network Theory and International Relations." *Millennium* 41, no 3. (2013): 414-415.

making the ruling class's norms dominant.[232] The materialism of ideology is underscored in later critical theory research, which demonstrated the role of particular ideologies in constructing new technologies.[233] STS has followed in this tradition by attempting to compare the ideology to the actual practice of science and technology. For the purposes of this chapter, it is important to highlight the role of ideology in reinforcing material realities. The ideology of the cloud can be manifested physically through the expansion of the network. This process is cyclical, where ideology builds itself into the network (for example, through particular hardware or software arrangements), and the network itself supports the ideology of the cloud as ubiquitous.

Infrastructure

In the previous chapter on utility computing I used a definition of infrastructure that underscored the broad framing of infrastructure.[234] This definition spoke to the social benefits of infrastructure and the political implications of adopting one type of public system over a private one. According to this view, infrastructure is a resource that can be consumed nonrivalrously for a range of demands. The demand for this resource is driven by downstream activities. It can also be used as an input for other goods and services. Utility computing fits under this definition of infrastructure, and I argue that the cloud, as an extension of utility computing, also falls under this definition.

There are other definitions of infrastructure that help speak to the lived experience of large socio-technical systems. For instance, Susan Leigh Star and Geoffrey Bowker's work on the ethnography of infrastructure has looked at technical assemblages as a living system, set within a particular context. They describe information systems as "linking experience gained in one time and place with that gained in another, via representations of some sort."[235] Star offers nine properties of infrastructure: embeddedness, transparency, learned as part of membership, linked with convention of practice, embodiment of standards, built on an installed base, becomes visible upon breakdown, and is fixed in modular increments.[236] These properties of infrastructure attempt to break down the master narrative of infrastructure that is impersonal and moves from system to user.

This is a theme that has been explored in systems perspective theory. Thomas Hughes and others in the history of technology have continually discussed large technical systems as being dynamic. The hybridity of a digital network that depends on the expansion of a global network invites the language of technology momentum.[237] Digital networks seem to suggest a perpetual energy stemming from changes in software and the rerouting of data across networks. These networks also seem to be slowing in momentum as the cloud's infrastructure matures and

[232] Leopold, David. "Marxism and Ideology: From Marx to Althusser." In *The Oxford Handbook of Policical Ideologies*, ed. Michael Freeden and Marc Stears. New York: Oxford University Press. 2013.

[233] Feenberg, Andrew. "Critical Theory of Technology and STS." *Thesis Eleven* 138, no 1. (2017): 3-12.

[234] Frischmann, Brett M. *Infrastructure: The Value of Shared Resources*. New York, NY: Oxford Press, 2012. 61

[235] Star, Susan Leigh and Geoffrey Bowker. "Sorting Things Out: Classification and Its Consequences." Cambridge: MIT Press, 1999. 290.

[236] Star, Susan Leigh. "The Ethnography of Infrastructure." *American Behavioral Scientists* 43, no. 3 (1999): 1999. 381-382.

[237] Hughes, Thomas. "Technological Momentum." In *Technology and Society: Building Our Sociotechnical Future*. Edited by Debora Johnson and James Wetmore. Cambridge, MA: MIT Press (2009): 141-150.

becomes more standardized. The internet seemingly gives us a loosely coupled network, even as the interlocking of the cloud becomes tighter.

Star and other authors have acknowledged that studying infrastructure is not a simple task because the subject can scale from the micro to the macro. This is even more difficult in terms of information systems. Much of the initial discussion on the topic stems from scholarship in the 1990s that started to address the broader impacts that the internet was having on global economies. One prominent example is German sociologist Ulrich Beck's *Risk Society* that offered a picture of a world in which risks are increasingly difficult to calculate because of the complexity of modernity. An important aspect of his argument is that many of these new risks are invisible, cannot be seen, and require the "sensory organs" of science to understand.[238] Like the risk of nuclear power, the risks of information systems are difficult to see in everyday life. This type of concern was reflected later in Manuel Castell's discussion of the network society and the idea of "space of flow" against the "space of places."[239] Instead of focusing on any one detail, Castell asks scholars to look at the spaces between infrastructures to understand the meaning that ties the things together. This is reminiscent of many classic STS works that were rooted in uncovering the social dynamics in a lab or a factory. Castell's scholarship isn't that different from any of the others that I have mentioned. The common theme in all of these works is that the focus on networks, relationships, and flows help provide us with the context we need to study these systems.

This brings us to the specific books and articles that have been written on the construction of information infrastructures and the cloud. The cloud, and the larger internet, are primarily made up of people, hardware, and code. The environment plays an important role, which I will touch on later, but the core of the cloud is information (minds and code) and materials (bodies and hardware). This dissertation, like other academic work before it, is interested in how these components interact to create an organic whole. The author Lawrence Lessig helped set the stage in 1999 by claiming that "code is law." In his book, he argues that computer code is a regulatory tool. [240] How a cloud computing platform is programmed will have an impact on the limits of what is possible. This is reflected in the literature on DRM and other copyright control schemes.[241][242] Lessig's important contribution made clear that "changes in the code [are] (unlike the laws of nature) crafted to reflect choices and values of the coders."[243]

Other research has looked at what politics are at stake in the management of contemporary information networks. Much of the early books looked at the struggles of users and the possibility for a new type of citizenship online (netizens).[244] Later on, scholars interested in Internet governance questioned the ability of internet users to resist control and challenged the

[238] Beck, Ulrich. *Risk Society: Towards a New Modernity*. London: Sage, 1992. 27.

[239] Castells, Manuel. *The Rise of the Network Society*. Cambridge: Blackwell, 1996.

[240] Lessig, Lawrence. *Code and Other Laws of Cyberspace*. New York: Basic Books, 1999.

[241] Jenkins, Henry. *Convergence Culture: Where Old and New Media Collide.* New York, NY: University Press, 2006.

[242] Lessig, Lawrence. *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*. New York, NY: Penguin Press, 2004

[243] Lessig, Lawrence. *Code: Version 2.0*. New York: Perseus Books, 2006. 110

[244] Hauben, Michael and Ronda Huaben. *Netizens: On the History and Impact Of Usenet and the Internet*. Los Alamitos: IEEE Computer Society Press, 1997.

utopic views of a completely open and uncensored web.[245] Corporate and governmental control over new information infrastructures has been a popular topic. Scholars have been interested in how new infrastructures can control dissent[246], limit a range of viewpoints[247], and generally become closed ecosystems.[248] These conversations bleed into a large literature on the telecommunication industry, which has often had a contentious relationship between users and regulators.[249]

Of this group of scholars, some have also studied the physical construction and maintenance of the internet. For instance, Paul Ceruzzi's *Internet Alley* is an extensive case study of the buildings, companies, and environmental histories of Tysons Corner, Virginia. By looking at the history of the region, Ceruzzi offers the reader a story about how the past history of the region contributed to it becoming a regional technology powerhouse. The military and telecommunication history he describes gives the context to the emergence of a newly formed internet backbone. Critically, the story of Tysons Corner is about the intersection of materiality and geography. The buildings and people stand in relation to a political center (Washington D.C.). Likewise, the places of the cloud have their own materiality (in the buildings, technologies, and people that make up these centers) that stand in relation to larger geopolitical places.

One of the main themes that is found throughout this literature is the notion that technological infrastructures are charged with ideology, history, and politics. The turn towards the cloud and the deployment of submarine cables is a story that echoes these previous findings.

When the Cloud Becomes Physical

One of the challenges of describing the cloud is that it exists both as an idea and a physical network. The idea of the cloud is founded on the visions of computing that I have described previously. These visions are often alluded to in order to make the cloud seem ephemeral. A casual and uncritical interpretation of ubiquitous technologies does not focus on the materiality of a technology. A more nuanced reading, as I attempted to provide through the history of Xerox PARC, shows that the lived experience of a technology matters. In this chapter, I will continue focusing on the lived experience of the cloud.

For the purpose of clarity, I want to focus on the aspects of the cloud that are primarily physical. This section looks at how the cloud is built and what it means for the cloud to be considered a knitting of physical objects. This section looks specifically at the cables, servers, buildings, people, and environments where the cloud resides. This turn to the physical doesn't discount the important ideological role that the cloud plays. Instead, by turning our attention to the roots of the cloud, we can better understand how the physical deployment of the cloud helps contextualize the images of the cloud that we carry in our heads.

[245] Goldsmith, Jack and Tim Wu. *Who Controls the Internet?: Illusions of a Borderless World*. New York, Oxford University Press, 2006.

[246] Morozov, Evgeny. *The Net Delusion: The Dark Side Of Internet Freedom*. New York: Public Affairs, 2011.

[247] Pariser, Eli. The Filter Bubble: What The Internet is Hiding From You. New York: Penguin Press, 2011.

[248] Wu, Tim. *The Master Switch: The rise and Fall of Information Empires*. New York: Alfred A. Knopf, 2010.

[249] Crawford, Susan. *Captive Audience: The Telecom Industry and Monopoly Power in the New Gilded Age.* New Haven: Yale Press, 2013.

In the first two chapters, ubiquitous and utility computing were discussed. In both of these cases, how the backbone of the network was set up mattered. In the example of Dartmouth's time-sharing network, the computer running the calculations and storing data was easily identifiable and locatable. As the notion of utility computing expanded, the sites of computing became more numerous (as seen in Tymshare's expanded network). Even as the network expanded, where the computers were located played a significant role in the adoption and continued use of these services. In the case of commercial timesharing, companies didn't want to pay long-distance charges to connect remotely to the network. Furthermore, accessing remote networks introduced more latency to the network and slowed down computing tasks. Likewise, the history of ubiquitous computing demonstrated the limitations of building networks beyond the local environment. PARC's work on ubiquitous computing focused on increasing the productivity of the workspace by taking away distractions. Both of these visions of computing have been attributed as a source of inspiration for the cloud, but the physical nature of these visions is often unacknowledged or underplayed.

I argue that the same case can be made for the cloud. The materials of the cloud matter just as much, if not more, than how the ideology of the cloud is deployed. The cloud, unlike other computing technologies, is dependent upon a global network. The cloud is primarily a distributed computer network, which is primarily focused on dynamically providing computational power, data storage, and information delivery. Producing and delivering these services requires both a robust technological infrastructure, as well as established service models, which outline how the resources are monitored and used within an organization or sold to an end-user. What distinguishes the cloud from other models of computing is the scale and long-term ramifications of these systems.

There are many different ways that the cloud can be studied as a physical object. A popular method has been to look at the construction and operation of data centers. As mentioned, data centers are the heart of the cloud and contain the bulk of computing power. Data centers have been criticized and supported by academics on a number of different fronts: from data privacy to environmental impact.

Some of the most salient articles written have been centered on the ecological impact of data centers. Data centers consume a great deal of natural resources, both in the form of electricity and in the building materials needed to construct and build these spaces of information. One author made the claim that writing about data centers is a process of "unpacking the green box" and that despite the significant energy costs, data center companies are invested in offsetting environmental damage by investing in renewable energy sources.[250] Some have viewed critically those efforts to "green-wash" data centers by highlighting the large carbon footprint that an always-on data center requires.[251] Determining the actual impact of data centers has been a

[250] Ipsen, Heather. "Catching the Cloud and Pinning It Down: The Social and Environmental Impacts of Data Centers." Thesis Syracuse University, May 2018.
[251] Pearce, Fred. "Energy Hogs: Can World's Huge Data Centers Be Made More Efficient?" *Yale Environment 360*. April 3, 2018.

challenge methodologically.[252] Despite many academic and popular articles on the topic, more micro-level research is needed to measure the actual environmental impacts of data centers.[253] The history of computing, in particular, is just beginning to start addressing the environmental impacts of computing as a whole.[254]

The physical impact of data centers is not limited to environmental issues. Another area of scholarship has looked at the construction of data centers as new places of corporate ownership. One author compared cloud infrastructure to data centers which "…are rooted in excess, redundancy, and contingency, governed by the looming specter of worst-case scenarios." [255] International legal questions are continually raised regarding the storage of data in foreign countries and the ability of the nation-state to control the data of its citizens. For instance, the European Union's General Data Protection Regulations (GDPR) has brought into question how to ensure where data is stored and what rights EU citizens have in controlling how that data is deleted. Outside of the European Union, countries are attempting to regulate the cloud within the context of national borders. Both nations and corporations are dealing with the challenge of integrating the cloud into the global marketplace. For instance, China's heavy regulatory environment has forced cloud companies, such as Apple, to build data centers in close partnership with Chinese companies.[256]

There are many other aspects of data centers that could be addressed in regards to privacy, economic impacts, and human capital. However, instead of looking directly at data centers, this chapter studies the cables that link cloud data centers together. In specific, this section unpacks the construction and deployment of submarine cables. I believe that a focus on oceanic cables provides a unique method of understanding the cloud's connection to the environment and society. Far too often, depictions of cloud computing simply draw a direct link between a data center and an end-user. What is left out of this portrait are miles of cables linking networks together and the people working to construct these networks. In this shadow of the cloud are larger questions about the economic and political incentives for expanding the cloud globally.

Submarine's Beginnings

Nearly all of the internet's backbone operates using fiber optic cables. These cables transmit data at high speeds using light inside a glass fiber. Most internet traffic is still carried via fiber optic cables, despite the increase in wireless use. Cables are, and will likely continue to be, the most efficient method of delivering information reliably. Deployment of cables is a capital intensive project and is a long term investment by the cable owner(s). Most fiber optic cables are buried underground, in a similar fashion to other utility services. The majority of cloud data centers

[252] Lykou, Georgia, Despina Mentzelioti, and Dimitris Gritzalis. "A New Methodology Toward Effectively Assessing Data Center Sustainability." *Computers and Security* 76 (2018): 327-340.

[253] Carruth, Allison. "Ecological Media Studies and the Matter of Digital Technologies. *PMLA* 131, no. 2 (2016). 369.

[254] Ensmenger, Nathan. "Computation, Materiality, and the Global Environment." *IEEE Annual of the History of Computing* 35, no. 3 (2013).

[255] Holt, Jennifer. "Where the Internet Lives: Data Centers as Cloud Infrastructure." In *Signal Traffic: Critical Studies of Media Infrastructures* (edited by Lisa Parks and Nicole Starosielski). Urbana: University of Illinois Press, 2015. 83.

[256] Reuters. "Apple Sets Up China Data Center to Meet New Cyber-Security Rules." *Reuters*. July 11, 2017.

depend upon the construction of reliable and redundant fiber backbones. Most often, cloud data centers are located in locations that are close to these cables, within a reasonable distance of an internet exchange point, and near a pool of information technology professionals.

Major cloud providers are not simply selling access to a single data center. Instead, they are selling consumers access to a global network with a great deal of flexibility in how resources are distributed globally. As the cloud market has grown, it has been increasingly important for providers to build out their network to ensure consumers have access to a regional cloud in whichever country the user or their respective nation's law demands. Some nations, as mentioned before, are demanding that these computing services are located within a nation's borders. Even in countries that do not demand regional control, an organization may prefer (for any number of reasons) to have the source of cloud computing closer to its users. Consequently, cloud providers are starting to make investments in ensuring that their globally distributed data centers can quickly communicate between each node of the network.

Many of the major cloud companies (including Microsoft, Amazon, and Google) have made significant investments in submarine communications cables. These cables carry the majority of the internet's traffic, including traffic between data centers.[257] The drive to own, or at least have a stake in the cable ownership, is part of the larger narrative the cloud. The history of the cloud is not simply a story about the development of a new form of computing; there is also a battle over the infrastructure of the larger internet and an open question about whether open platforms can coexist alongside the modern cloud. Submarine cables are just one site in which these questions are being debated, but may offer insight into how the web might develop in the future.

Before addressing submarine cables built for the cloud directly, let's start by looking at the history of submarine cables. Much of the history of cables demonstrates that ownership of a cable is a political tool. The first submarine cables were constructed and deployed in the nineteenth century. The initial use of these cables was transmitting short telegraph messages. The first cable, laid in 1850, was a simple construction of copper wiring covered by the gutta-percha tree leaf.[258] Slowly, the telegraph cables grew in size, adding additional layers of protection. Materials like hemp dominated early cables, but were eventually replaced by more durable shields, like iron and steel.[259]

For the bulk of the late nineteenth century, the British government controlled and managed much of the global submarine cable system. These submarine cables allowed the British Empire to govern distant territories. The cables were considered so fundamental to British security that an "All-Red Line" (secure lines for communication across the British Empire) was planned and constructed.[260] World War I and II highlighted the need for cable security. While the design and construction of the cables did not change drastically, with the exception of the addition of polyethylene in the 1930s, there was a clear recognition that control and management of the

[257] Alazri, Aisha Suliaman. "Telecommunication Traffic Through Submarine Cables: Security and Vulnerabilities." *IEEE* 2016 11th ICITST. December 5-7, 2016.

[258] John Brett, "Letter from John W. Brett." Library of Congress (Ref: rbpe 23302900). London, 1863.

[259] Edward J. Malecki and Hu Wei. "A Wired World: The Evolving Geography of Submarine Cables and the Shift to Asia" *Annals of the Association of American Geographers* 99, no. 2 (2009).

[260] George Johnson. *The All Red Line; The Annals and Aims of the Pacific Cable Project*. Ottawa: James Home & Sons, 1903. 1.

cables was a major security concern. In WWI and WWII, both sides of the conflict actively sought to tamper with and destroy the enemy's cables.[261]

It was not until the late 1980s that fiber optic cables were deployed for use in submarine cables. Up until this point, the primary material used was coaxial cable. In 1956, TAT-1, or Transatlantic Number 1, was the first transatlantic telephone submarine cable deployed. At the time, TAT-1 supported thirty-six voice channels.[262] Despite this accomplishment, the coaxial cable was not a very efficient carrier of data across long distances. TAT-1 cable and other early cables were built primarily because of various economic and political pressures between the American Telephone & Telegraph Company (AT&T), Canada's Overseas Telephone Corporation, the International Telephone and Telegraph, and the British Post Office.[263] More importantly, TAT-1 signaled a turn towards more institutionalized domestic state monopolies. It also was a part of a series of moves to privatize the majority of the international telecommunications network.

As more coaxial submarine cables were built, the telephone system was able to support more bandwidth for telephone calls. However, with the growth of the Internet, telecommunication companies started to invest in fiber optic cable routes. Fiber optic cables have numerous economic and computational benefits over coaxial cable. Fiber optic cables provide higher throughput of data, resistance to signal interference, and are lightweight. For these reasons, the past twenty-five years have seen a complete shift to fiber optic cables. The "dot com" boom in the 1990s saw an explosion in the amount of fiber optic cables laid down.[264] The eventual market crash resulted in an excessive amount of bandwidth that was underutilized.

The excesses of the early 2000s provided ample headroom for the growth of the international data markets. As the decade progressed, the amount of available "dark fiber" (fiber optic cables which are non-operational and waiting to be made operational) decreased. Slowly, starting around 2008, it became clear that increasing traffic meant that additional cables would need to be deployed. In particular, the Pacific link between North America and Asia had very few cables linking the continents, which prompted heavy investments to meet growing bandwidth demands. [265]

A Plurality of Data Centers

When looking at the history of submarine cables today, we need to keep in mind the changing geographic requirements of data centers. The primary job of most traditional data centers is to

[261] Alfred Price. *The History of U.S. Electronic Warfare Vol 1: The Years of Innovation – Beginnings to 1946*. Arlington: Association of Old Crows, 1984.
[262] IEEE Global History Network. "Milestones: The First Submarine Transatlantic Telephone Cable System (TAT-1), 1956." https://web.archive.org/web/20160304033834/https://ethw.org/Milestones:The_First_Submarine_Transatlantic_Telephone_Cable_System_(TAT-1),_1956 (Archived March 4, 2016).
[263] Jill Hills. "Regulation, Innovation and Market Structure in International Telecommunications: The Case of the 1956 TAT1 Submarine Cable," *Business History* 49, no. 6 (2007): 868-885.
[264] Barney Warf. "Engineering Time and Space with the Global Fiber Optics Industry" in *Engineering Earth*, ed. Stanley D. Brunn. New York City: Springer, 2011. 115-129.
[265] Gardiner, Bryan. "Google Hunts for Submarine Bandwidth as Traffic Surges." *Wired*. September 25, 2007.

process information and store information and, when requested, to deliver that information. In a typical data center model, a remote server exists in a single location, and that information is sent to a user wherever he or she is. For example, if a user lives in Texas and a server is located in Seattle, the user will connect through the internet, often bypassing submarine cables. If that user is located in South Africa, that connection will use at least one underwater cable. The physical distance between the server and the user can have a negative impact on performance.

The development of the cloud has altered this traditional model. For large technology companies, it is more common to have multiple data centers located across physical space. Instead of a centralized location, users connect to the closest regional server. This reduces latency and reduces the amount of traffic that needs to be requested from a distant location. On the face of it, it would seem that submarine cables would become less important for cloud applications. This shift in computing has actually had the opposite effect because of the enormous amount of data mirroring, traffic-balancing, and syncing needed between cloud servers.

This significant shift can be seen in the spike in cable usage by cloud computing companies in the past decade. In 2010, general internet usage accounted for 80% of the total share of submarine cable utilization.[266] By 2016, that figure had declined to 54%.[267] This change in usage can be explained by a drastic increase in non-public traffic between data centers. Shrinking availability of bandwidth has prompted cloud companies to rapidly increase investments in submarine cables. By relying on existing cables, the speed, reliability, and network competitiveness of these corporate clouds could start to be jeopardized. Whereas in the past telecommunication cables were infrastructural tools that made money through leasing of bandwidth, today they are critical infrastructural components of larger business systems. Provider and route diversity, on top of economic potentials, are now important factors in determining where to lay new cable.[268]

While cloud computing continues to rely upon the open internet for connecting users to the cloud, connections between data centers are becoming increasingly closed. The closing of the cloud is largely related to the creation of global cloud networks. This new landscape has given rise to what some researchers call "cloud paths."[269] These cloud paths are any connection made between cloud data centers that are used almost-exclusively for inter-data center connectivity. The stated logic behind direct connections between data centers is rather clear. Direct connections will reduce the latency between servers, in part by reducing the number of hops that the information will need to travel. Another aspect is security. Although the majority of information in the cloud is encrypted, there are still concerns with data security when passing information over multiple networks. Cloud-paths reduce the possibility of multiple attack vectors by keeping the data in-house.

[266] Mauldin, Alan. "Shaping the Global Wholesale Bandwidth Market." *Telegeography*. July 28, 2017.

[267] Mauldin, Alan. "Shaping the Global Wholesale Bandwidth Market." *Telegeography*. July 28, 2017.

[268] Mauldin, Alan. "Three Things Investors Should Know About the Submarine Cable Market." *Telegeography*. June 12, 2018.

[269] Haq, Osama, Mamoon Raja, and Fahad R. Dogar. "Measuring and Improving the Reliability of Wide-Area Cloud Paths." *International World Wide Conference 2017.* April 3-7, 2017.

One aspect of these emerging cloud paths that is not often discussed is the broader impact that these networks have on communication networks as a whole. The rush to invest in submarine cables is not entirely driven by performance and security. Cable investments are also about exercising control over the broader infrastructure of modern communication networks. As more websites and information services rely upon the cloud in order to function, the backbone of the cloud starts to be seen as synonymous with the internet in general. Who control the cables, and how those cables are regulated, can potentially have a massive impact on the future of the web.

Alphabet's Cables

The story of large, so-called "hyper cloud," technology companies investing in submarine cables has not been central to the narrative of the cloud. Instead, most observers have looked towards the construction of data centers, developments in cloud applications, and user adoption of cloud services. While these are important sites for research, more weight should be given to the oceanic projects that link the cloud data centers together.

As previously mentioned, large technology companies have shifted towards the cloud as the central aspect of their business. Microsoft, Google (via parent company Alphabet), Amazon, and Facebook have all publically indicated that the cloud is part of their long term business model and, consequently, each company has made investments in submarine cables. These types of investments are large capital projects that can take years to be completed. Even with the interest in these cables, most technology companies are still partnering with other telecommunication investors to help fund the construction of cable and provide expertise in deploying and working within regulatory frameworks.

There are many notable investments in submarine cables that pertain to the development of the cloud. At the start of 2018, Facebook invested in a new cable linking Hong Kong and the United States (joined by Asian and Australian investors).[270] Just a few years prior, Facebook jointly worked with Microsoft to back the transatlantic "Marea" cable, which added additional bandwidth to an already competitive data route.[271] In other efforts, Facebook partnered separately with Amazon and Google to connect Pacific routes. Investments in submarine cables by technology companies are a relatively recent development but are starting to have a significant impact on the development of future systems.

This impact can be seen in the type of investments that cloud companies are making and the type of rhetoric that is used to justify these investments. Looking at Microsoft and Facebook's Marea cable, the companies place a large focus on control. Microsoft's "guiding principles" for their cloud networks, including Microsoft's primary cloud product Azure, are geographic closeness,

[270] Sverdlik, Yevgeniy. "Facebook Invests in US-Hong Kong Submarine Cable." *Data Center Knowledge*. January 22, 2018. https://web.archive.org/web/20181113002220/https://www.datacenterknowledge.com/facebook/facebook-invests-us-hong-kong-submarine-cable (Archived November 13, 2018)

[271] Bach, Deborah. "Microsoft, Facebook and Telxius Complete the Highest-Capacity Subsea Cable to Cross the Atlantic." *Microsoft*. September 21, 2017. https://web.archive.org/web/20181003141648/https://news.microsoft.com/features/microsoft-facebook-telxius-complete-highest-capacity-subsea-cable-cross-atlantic/ (Archived October 3, 2018).

maintaining control over capacity, and a proactive network management strategy. [272] When discussing these principles, Microsoft advertises that "Azure traffic between our datacenters stays on our network and does not flow over the Internet."[273] Microsoft's corporate strategy clearly focuses on building out a robust information infrastructure as a means of ensuring control over their cloud.

Like Microsoft, Amazon has made moves to ensure this control over the network. In 2016 Amazon invested in the Hawaiki Cable (as a capacity purchaser) linking the United States, Australia, and New Zealand.[274] The cable lands in the town of Hillsboro, Oregon, close to where Amazon AWS data centers are located. Additionally, Amazon has invested (as a part-owner) in two additional Pacific cables (Jupiter and BtoBE) along with Facebook and other telecom companies.[275] Publically Amazon has not said much in terms of cable ownership and has placed more of a focus on building out the edges of their cloud in terms of data center connections.

Google is the only company of the major cloud providers to be the sole owner of submarine cables, owning a total of 14 (3 of which are completely owned by Alphabet).[276] Google has also been investing in these cables for longer than most companies. In 2008 Google, now the subsidiary of Alphabet, first invested in the $300 million dollar "Unity" cable that linked Japan to California. At the time, Google's initial move was viewed with suspicion. Analysts did not predict other non-telecom companies to start investing in submarine cables, citing the instability of the market and falling data costs.[277] However, Google continued to partner with telecommunication investors and became a bellwether for other cloud platforms.

Even in cases in which Google is not the sole funder of the cable, there are still shifts in the development of these new networks. For instance, Google partnered with South American telcos to connect Brazil to Florida. The Monet cable (made operational in 2017) is unique in that the termination point is directly in a data center.[278] Almost all submarine cables first land at a location called a cable landing station along the coast. The decision to end the cable directly inside a colocation point helps reduce the latency between the landing station and the data center. It also is a signal that the cloud is at the heart of these investment decisions.

[272] Khalidi, Yousef. "How Microsoft Builds Its Fast and Reliable Global Network." March 15, 2018. https://web.archive.org/web/20181230185126/https://azure.microsoft.com/en-us/blog/how-microsoft-builds-its-fast-and-reliable-global-network/ (Archived December 30, 2018).

[273] Khalidi, Yousef. "How Microsoft Builds Its Fast and Reliable Global Network." March 15, 2018. https://web.archive.org/web/20181230185126/https://azure.microsoft.com/en-us/blog/how-microsoft-builds-its-fast-and-reliable-global-network/ (Archived December 30, 2018).

[274] Sverdlik, Yevgeniy. "Amazon's Cloud Arm Makes Its First Big Submarine Cable Investment. May 13, 2016. https://web.archive.org/web/20181112204733/https://www.datacenterknowledge.com/archives/2016/05/13/amazons-cloud-arm-makes-first-big-submarine-cable-investment (Archive November 12, 2018).

[275] Telegeography. "A Complete List of Content Providers' Submarine Cable Holdings. November 9, 2017. https://web.archive.org/web/20180211142535/https://blog.telegeography.com/telegeographys-content-providers-submarine-cable-holdings-list (Archive February 11, 2018).

[276] Telegeography. "A Complete List of Content Providers' Submarine Cable Holdings. November 9, 2017. https://web.archive.org/web/20180211142535/https://blog.telegeography.com/telegeographys-content-providers-submarine-cable-holdings-list (Archive February 11, 2018).

[277] Gardiner, Bryan. "Google's Submarine Cable Plans Get Official." *Wired*. Feburary 25, 2008.

[278] Moss, Sebastian. "Equinix to Land US-Brazil Monet Fiber Submarine Cable." *Data Center Dynamics*. September 12, 2016.

Most recently, Google is working on deploying three new routes linking Asia, Europe, and South America to the United States.[279] The Curie cable will be solely owned by Google and will connect Valparaíso, Chile to Los Angeles. Google is providing all of the funding for the cable, but will work with the submarine deployment company SubCom to lay the cable.[280] This 10,000 km cable will have a possibility of connecting to Panama via a branding unit, but for now is focused on building Google's cloud connections between North and South America.

The decision to land in Los Angeles and Chile was a strategic decision in both the location of the cables and the proximity to Google's existing and future data centers. Google's only data center in South America is located near Chile. The Quilicura location (near Santiago) was made operational in January 2015 and is touted by Google as "one of the most efficient and environmentally friendly data centers in Latin America" and exists alongside Chile's "ideal combination of reliable infrastructure, skilled workforce, and a commitment to transparent and business-friendly regulations."[281] The Santiago region is currently connected to two submarine cables, the South American Crossing (owned primarily by Level (3)) and South America-1 (owned by Telxius), both of which circle the Western and Eastern sides of South America.[282] Google Curie cables add a third link and the first direct connection to the Western United States.

The actual data center site sits on the exterior of Santiago and roughly 60 miles from the submarine landing station in Valparaíso. Across the street from Google's data center is a location for Latin America data center company Sonda. Four miles down the road is another large data center owned by Level(3), one of the world's largest fiber-optic carriers. Although agreements over internet backbone use are private (sometimes called "peering agreements"), it is not difficult to imagine that Level(3)'s ownership of one of the two submarine cables connecting Chile to the world may have encouraged Google's decision to build the Curie cable.

Google's decision to deploy Curie also coincides with a move by the technology investment industry into Latin America. Venture capitalist money into the region doubled to 1.1 billion in 2017 and is estimated to reach 2.5 billion in 2018.[283] Also in 2018, Google has followed this trend by investing 140 million additional dollars into its Quilicura location. At the announcement of additional investment, Chile's president Sebastián Piñera signaled the move as participation in "the fourth industrial revolution."[284]

[279] Sverdlik, Yevgeniy. "Three New Submarine Cables to Link Google Cloud Data Centers." *Data Center Knowledge*. January 17, 2018. https://web.archive.org/web/20190410225748/https://www.datacenterknowledge.com/google-alphabet/three-new-submarine-cables-link-google-cloud-data-centers (Archived July 8, 2018).

[280] SubSea World News. "TE SubCom to Build Curie Subsea Cable for Google." January 17, 2018. https://web.archive.org/web/20180117161856/https://subseaworldnews.com/2018/01/17/te-subcom-to-build-curie-subsea-cable-for-google/ (Archived January 17, 2018).

[281] Google Data Centers. "Quilicura." https://web.archive.org/web/20181026034845/https://www.google.com/about/datacenters/inside/locations/quilicura/ (Archived October 26, 2018).

[282] Telegeography. "Submarine Cable Map." 2018. https://web.archive.org/web/20180519021150/https://www.submarinecablemap.com (Archived May 19, 2018)

[283] Costa, Gonzalo. "The Tech Investment Wave Has Reached Latin America." *Tech Crunch*. July 23, 2018 https://techcrunch.com/2018/07/23/the-tech-investment-wave-has-reached-latin-america/

[284] Laing, Aislinn. "Google To Invest $140 Million To Expand Data Center in Chile." *Reuters*. September 12, 2018.

There are, of course, deeper economic issues at play. A few months prior to Google's announcement, President Piñera was under pressure from Chile's lead export, copper, falling in value.[285] Investments from technology companies are attractive from a policy perspective, and these companies also see Chile as a location for cheaper labor and untapped customers. This can be seen in Amazon's pitch to the Chilean government to provide the cloud computing resources to mine the country's telescope data.[286] Amazon also signed an agreement with Chile to modernize governmental services using AWS and possibly build a data center within the country.[287] Pair this move by Amazon with a new submarine cable from Chinese company Huawei to connect the Patagonian region to mainland China, and it is clear that the expansion of the cloud is part of larger geopolitical shifts.[288] The cloud is not immune from the politics of uneven geographical development.[289] This is made true as the cloud expands its "nodes" to the digital periphery. Cables, when investigated, start to make material the networks that surround the cloud.

Moving to the Northern Hemisphere, on the other side of the proposed cable sits Los Angeles. The story of the submarine cable in Chile was framed as a move to the future, a diversification of economic opportunities, and an opening up of global markets. The story in Los Angeles was framed differently, both by popular tech publications and Google itself. Los Angeles's cloud narrative speaks to a city attempting to retain its dominance as a creative capital for entertainment. This story also attempts to draw upon the notion of California as a leader in technological development. These types of stories are powerful because they link the metaphor of the cloud to the cultural climate of a region. As Bowker and Star suggest, infrastructure is a lived experience. In the case of the cloud, the cloud exists as the same technology across the globe but is felt as a localized emotion. This is also an instance of translation in action, where the specificities of a local network become meshed into the large whole.

Google's decision to deploy the Curie cable came six months prior to its announcement of a Google Cloud Platform data center in the same region.[290] Los Angeles has a number of submarine cables that link the city to the rest of the globe, so the decision to build the data center in Los Angles was most likely not heavily motivated by the new Curie cable. That said, the additional bandwidth that the cable could provide would only be an incentive for Google's new data center. This data center, referred to as "us-west2," was touted by local politicians as a

[285] Mander, Benedict. "Chile Is Canary In Copper Mine as Price of Metal Falls." *Financial Times.* July 30, 2018.
[286] Garrison, Cassandra. "Amazon Eyes Chilean Skies As It Seeks to Datamine the Stars." *Reuters*. September 4, 2018.
[287] Moss, Sebastian. "AWS Pitches Chilean Data Center For Virtual Observatory In The Cloud." Data Center Dynamics. September 5, 2018.
[288] The Santiago Times. "Huawei to Build Chile's 2,800KM Subsea Cable Project." March 27, 2018. https://web.archive.org/web/20180624232215/http://santiagotimes.cl/2018/03/27/huawei-to-build-chiles-2800km-subsea-cable-project/ (Archived June 24, 2018).
[289] Harvey, David. *Space of Global Capitalism*. Verso: London, 2005.
[290] Tsidulko, Joseph. "Google Goes Hollywood: Tech Giant Launches 17th Cloud Data Center Region In Los Angeles." *CRN*. June 26, 2018. https://web.archive.org/web/20181004130147/https://www.crn.com/news/cloud/300105701/google-goes-hollywood-tech-giant-launches-17th-cloud-data-center-region-in-los-angeles.htm (Archived October 4, 2018).

positive step forward to ensure Los Angeles as a "center of invention and creativity."[291] This is the fifth cloud data center for Google in the United States and is a move that looks to capture a valuable marketplace.

When announcing the creation of a Los Angeles cloud, regional politics and geography were part of the sales approach to the region. During the public unveiling event, Google employees pitched the cloud to the creative industry in particular.[292] The event featured creative professionals, as well as Google staff. Paul-Henri, the President of Global Customer Operations, framed the cloud as a site for creativity and as a "cloud of the future." He also underscored that Google carried 40% of global internet traffic and that a Google cloud is 80% faster than other clouds. Other employees from Google played to the typical imagining of Southern California as a place of creativity. In the presentation, there was also a broader narrative at play of California as the endless frontier and the cloud as a tool for continuing the endless American frontier. The announcement hit upon connections to movie studios and other creative projects. Google's new product "Cloud Firestore" was marketed to production houses as a way of locally backing up video projects to the cloud by physically mailing a data device to a Los Angeles location.

Los Angeles is also a region that Amazon does not currently have any cloud data centers in. Northern California, the heart of Silicon Valley, has a number of cloud data centers, but this is the first play by a major cloud company to expand into Southern California. Google's office location in Venice, California and its new location in Playa Vista signal the importance of the region. In particular, products like YouTube are increasingly being meshed with traditional media production. Establishing a corporate location, while also deploying additional bandwidth to the region via the Curie cable, starts to paint a picture of how clouds are constructed around identity, people, and materials.

One of the most interesting missing pieces of information from the announcement was the actual location of the data center(s). Google does mention three different cloud zones within Los Angeles, so it is likely that these data centers are spread out over the city. This lack of physicality is an interesting absence in a presentation that spoke clearly to the geography of Los Angeles (from their proximity to studios and the mention of new Google campus locations). This lack of precision in terms of location is indicative of the larger narrative regarding cloud data centers and submarine cables. The location of the cloud is both an asset and a liability. The cloud is ideally located close to where you are (to reduce latency), but at the same time, the location should not matter. The submarine cable connecting Los Angeles to Chile is part of a larger plan to build a global network but the details of where the cables land and where the data centers are located don't typically factor into a public discussion of the cloud – until they align with the marketing goals of the cloud owner. There is a selective visibility about what infrastructure edges matter and which are insignificant.

If you trace the physical infrastructure of the cloud, as I attempt to do in the following chapter, you will see that these details can matter in significant ways. There are environmental, economic,

[291] Tropin, Kirill. "Our Los Angeles Cloud Region is Open For Business." *Google Cloud*. July 16, 2018. https://cloud.google.com/blog/topics/infrastructure/our-los-angeles-cloud-region-open-for-business

[292] Google Cloud. "Google Cloud Los Angeles Region Celebration." August 3, 2018. https://www.youtube.com/watch?v=hq1Vioh5nuY

and political questions at each junction of the cloud. The ways in which the Cloud is presented to the public, even to those with a strong technical background, often downplay the locality of the cloud unless it is a marketing advantage. The connection of submarine cables seems like a rather boring topic until you frame the cable investment as a larger project to control the backbone of the internet and maintain the dominance of one cloud over another. As the cloud becomes more central to the internet, who controls these cables and which markets they serve will have a large impact on the politics of the modern communication systems.

Closing The Submarine Links

Submarine cables are not the heart of the cloud, but they are the invisible arteries that allow the data to be pumped across the globe. They are forms of invisible infrastructure that are critical to the sustained growth of the cloud and will help shape the future of modern communications. Despite their impact, their significance is underappreciated. Furthermore, nearly all definitions of the cloud downplay any mention of the physicality of the cloud. When the materials are injected into the metaphor of the cloud, it is a representation of a data center on a very abstract level. The details, such as where the hard drives are located or how servers are cooled, are not part of the discussion. This lack of attention to the materiality of the cloud has been even truer in regards to the fiber optic cables that are necessary for large-scale cloud growth. Despite this, the deployment of submarine cables is having lasting impacts on the shape and purpose of the cloud.

The story of the cloud's encroachment upon the submarine cable market ties into a larger body of scholarship that is concerned with technological imaginaries, invisible infrastructures, and the politics of technological progress. In STS, researchers have long recognized the importance of opening black boxes. The data center is one such box that researchers are starting to open up. Submarine cables, I argue, are the pieces of twine holding these boxes together. As hyper-scale clouds grow and submarine cables continue to be deployed, the relationship between the data center and the submarine cable will only become more tangled. This chapter starts to undo the knot by suggesting the investments in submarine cables by cloud companies are part of a larger project to control the entire cloud ecosystem. Looking at where these cables are deployed, and the politics of geography, we can start to frame these projects as something with values attached. These projects are not simply infrastructural investments but are the physical embodiment of a form of translation and an attempt to build a new type of computing network from disparate places and technological building blocks.

Unpacking the cloud's black box is critical in understanding how and why the cloud is developing in the manner that it is. The story of the Curie cable is a snapshot of a larger project to bind multiple clouds together into a seamless one. In each instance, however, we can see how regional politics and identity are framed around the idea of the cloud. For Chile, the cloud represents a connection to a globalized technology industry, advancement in scientific research, and a modernized government. For Los Angeles, the cloud speaks to a creative identity and the economic pressures of Hollywood. This duality speaks to the previously mentioned research on infrastructure and identify. The identity of an infrastructure isn't found only in the materials, but in the "space of flows" that outlines the possibility of what the system can provide.

One method of addressing this opaque metaphor is to start digging into the infrastructure itself. Just as the original telegraph cables were full of politics, both on and off the lines, today's cables are a continuation of that story. Scholars should start thinking about submarine cables as the political tools that they are. As countries, businesses, and individuals continue to put their digital lives on the cloud, we need to think about what implications building these global networks will have on the possibilities for our digital futures. Submarine cable projects have cost investment groups billions of dollars and cannot be easily reversed. For the most part, cable operators have been good stewards of the lines and have not been discriminatory in what data flows in the open oceans. However, the shift towards cable ownership by cloud companies should warrant a second look by the public. In the long term, what will be the implications of these cables on the internet in 20 or 50 years? Of course, the answer to that question is unclear. Hopefully this chapter has started that conversation.

Chapter 5

Clouds in Los Angeles

The cloud is a lived experience. What that experience is, and how that experience will be negotiated, will depend upon the individual user. Anyone who participates in modern digital communication actively participates with cloud computing. In certain cases, this participation is obvious. Uploading files to Apple's iCloud is an active engagement in a branded cloud service. More frequently, the cloud acts as the underlying invisible infrastructure of digital life. The majority of video delivery services, like Netflix or YouTube, are using the cloud to move large amounts of traffic around the world. Cloud servers also power the majority of high traffic websites. As more systems are moved to the cloud, our experience of what the web is will also be framed by the outlines of the cloud.

At times the cloud can feel like a monolithic experience. The consumer-facing solutions present a vision of uniformity. Cloud applications look and feel essentially the same regardless of where you are in the world. In part, this is by design; the cloud is meant as an infrastructural resource to be built upon. However, this uniformity can make the local seem unimportant and belie the complicated realities of the cloud. Our collective vision of the cloud as "computing everywhere" underscores the feeling of aimlessness as we attempt to marry the idea of the cloud with its actual materials.

Previous chapters have attempted to demonstrate how the idea of the cloud was built upon older metaphors of computing. I argued that these metaphors were unevenly translated into our current understanding of what we ought to expect from the cloud. Today's cloud is not a natural evolution of or simple addition to previous metaphors. Instead, the idea has been shaped around these older ideas in order to model a new form of computing with its own set of ideological commitments. The ideologies of the cloud are not abstract, but they are diffused. My discussion of investment in submarine cables demonstrates that the ideology of the cloud is deeply intertwined with shifting ownership models and planned corporate futures. The submarine cable example started to open up the cloud as a place that could be touched and seen.

This chapter extends that conversation to look at what it means to consider the cloud as a network of places and lived experiences. The materiality of the cloud is composed primarily of the networks and computers that are stitched together. In between these technological systems are human bodies and ideas that go largely unrecognized. Seeing the humanity in these systems requires an unearthing of the networks and an examination of the maintenance of the cloud. The deeper point behind the move to "uncover" the cloud is to identify the points of ideological construction that have helped foster an ambiguous metaphor.

The developments in cloud computing over the past decade have introduced a type of computing visibility alongside this new metaphor. The cloud has become simultaneously all-encompassing and also mundane. Visions of flashy cloud data centers, powered by green energy and filled with a million blinking lights, are set in stark contrast to the non-descript grey exteriors of data centers nested in office parks or hidden within office buildings. The everyday markings of the cloud, from fiber-optic conduit buried under the street to the small human communities managing and upgrading servers, go largely unnoticed. It is difficult to merge this dualism, between the omnipotence of a digital cloud and the tangible tangling of cables. However, this task is necessary and worth the effort for all those interested in how to manage our relationships to a new form of computing.

The issue of the visibility of materials raises questions of locality. What would it mean to consider the cloud as a local object? In other words, can we think of the cloud as a neighboring actor? The question seems to defy the logic of the cloud, which purports itself to be infinitely expandable and separate from geographic limitations. This is a question that I think researchers should take seriously because it offers an avenue into the lived experience of the cloud. By looking at the cloud as a local phenomenon, it frames it as something that can be accessed and worked with.

In an attempt to start opening up this question, this chapter starts by looking at the connection between the local and remote clouds. The focus of this chapter is looking at the process of translation and uneven geography between the heart of the cloud and the end-user. In the following sections, I look at the concept of local knowledge in STS and the history of computing. Then I apply these themes to the clouds of Los Angeles. I map out two major cloud hubs in Los Angeles and their connections to the area's history and linkages to other cloud hubs. By blending the local, the regional, and the global together, I attempt to demonstrate the complicated mixture of physical materials and metaphors that have been unevenly woven into our story of the cloud. I conclude by arguing that these lived experiences are important sites for understanding the politics and ideological commitments we make to the cloud.

Local Knowledge

Before addressing the issue of the cloud's materiality, it is helpful to look at the notion of local knowledge in the context of STS and related disciplines. Many, if not most, of STS's core theories and insights have some type of connection to how knowledge is produced, maintained, and modified. The foundational texts of our discipline discuss scientific knowledge as a fundamentally social process. The injection of social information into the conversation will always entail some type of lived experience. Looking broadly, it is difficult to look at the canon of STS and not find a link to local knowledge.

In the case of the history of science and STS, researchers were able to contextualize experiments, theories, and scientific laws through the lens of a particular worldview. For example, to understand the "air-pump" experiment, historians need to consider how local experiences influence what a "correct" way of doing science is.[293] Likewise, STS has explored how scientific objectivity is always linked with the cultural context of the scientific process.[294] STS theory has continually shifted towards more local knowledge. The move towards laboratory studies was a sharp theoretical turn to address gaps in knowledge by looking at "science in action."[295] The

[293] Schaffer, Simon and Steven Shapin. *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Oxford: Princeton University Press, 2017.

[294] Daston, Lorraine and Peter Galison. *Objectivity*. New York: Zone Books, 2007.

[295] Latour, Bruno. *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge, MA: Harvard University Press, 1988.

ethnography reply has been useful in pulling apart the different knowledge groups that participate in the production of sociotechnical ideas. For instance, research on the context of science might look at a community of energy physicists to see how specific environmental factors help shape what type of knowledge is legitimate.[296]

Lay knowledge has been of particular interest to STS researchers. In part, the focus on "everyday" experience is a methodological tool of setting the scientific method against other ways of producing knowledge. More importantly, perhaps, is the way in which non-scientific communities challenge deterministic models of science. Lay expertise often stems from lived experience of a community and can help challenge or influence scientific bodies of knowledge.[297] Researchers well versed in anthropological theory have been particularly skillful in connecting the concept of indigenous knowledges to technical and scientific issues. [298] Likewise, feminist theory has long spoken to the importance of looking at embodied knowledge to counterbalance hegemonic systems of "legitimate" knowledge. Works on feminist standpoint theory speak most clearly to the tacit knowledge discussion.

STS has also faced the issue of becoming tangled in, whether intentionally or not, the 1990's Science Wars. The backlash against postmodern critiques of science has made many in the STS community reframe their critique. In regards to local knowledge, this has taken form by more carefully considering what the role of embodied knowledge should be in the public domain. This can be seen in works by Harry Collins and Robert Evans, both of whom have written extensively regarding lay knowledge. Their argument for a third wave of science studies – "Studies of Expertise and Experience (SEE)" is one response to the current theoretical climate.[299] According to Collins and Evans, SEE attempts to address the gap between how scientific consensus is formed and how political policy is constructed. SEE, and related scholarship,[300] argue that policymaking today often relies upon the public to weigh the validity of competing expert testimony. How lived experience sits alongside scientific knowledge production and the policy arena is still an issue being worked out in STS.

To fully describe how STS is linked to theories of local knowledge would be an exhaustive exercise that would ultimately include the majority of STS work. However, one of the main themes that can be pulled from this body of knowledge is that understanding the material context of a person or institution is central when trying to see how social values are embedded in ideas. Commonly, STS texts point to dominant scientific culture undervaluing of layperson experience, whereas the material lives of technical professionals are understudied. This dualism is of particular interest for this chapter because it maps closely to the lived experience of the cloud and computing more broadly.

[296] Trakweek, Sharon. *Beamtimes and Lifetimes: The World of High Energy Physicists*. Cambridge: Harvard Press, 1988.

[297] Epstein, Steven. "The Construction of Lay Expertise: AIDS Activism and the Forging of Credibility in the Reform of Clinical Trails." *Science, Technology, & Human Values* 20, no. 4 (1995): 408-437.

[298] Watson-Verran, Helen and David Turnbull. "Chapter 6: Science and Other Indigenous Knowledge Systems." In *Handbook of Science and Technology Studies*. Thousand Oaks, CA: Sage, 1995.

[299] Collins, Harry and Robert Evans. "The Third Wave of Science Studies: Studies of Expertise and Experience." *Social Studies of Science* 32, no 2. (2002): 235-296.

[300] Jassanoff, Sheila. *The Fifth Branch: Science Advisers as Policymakers*. Cambridge, MA: Harvard University Press, 1990.

Local Computing Knowledge

Computing has typically been considered a local experience. Most historical accounts of the computing experience are generally looking at the connection that a human has to a singular computing device. As the cloud and the internet of things grow, our notion of computing is shifting. Much of this shift has to do with the expansion of computing beyond the desktop. Mobile phones, tablets, and smart devices are challenging our default notion of what it means to use a computer. In developing nations, mobile technologies are often the most accessible forms of computing.[301] Even in this example, where the hardware has shifted, the dominant paradigm is still thought of as a local experience.

Historically, computing has not always been framed as the connection between a human and a machine. This can be seen even by looking at the first instances of computing. Computers in the early 20th century were human (most of whom were women).[302] These human computers were needed to calculate complex mathematical problems prior to the invention of more advanced digital computers. Even in the early example of the programmers of the ENIAC, computing can be read as a lived experience. It is only once computing lost its biological connection that people started to think of computing as an experience between users and machines. This type of relationship has been documented heavily in different fields, perhaps most notably in the entire canon of Human-Computer Interaction.

The most common framing of computing has been between a human and a local device. In the age of desktop computing (prior to mobile computing), the conceptual model between user and machine was rather simple. The experience of computing was primarily a local experience. Mobile and integrated smart devices (IoT) are starting to shift the conventional understanding of the location of computing. That said, even early networked computing experiments in the 1960s demonstrated how space impacts both the materials of computing as well as the user's own experience. For instance, in the chapter on utility computing, I discussed the impact that timesharing had on Dartmouth's campus. In the Dartmouth example, the idea of "the computer" started to shift as computer terminals produced a gap between the core hardware of a computer and the terminal. The user of the computer, in this case, may view computing as an experience that moves beyond the desktop. In the 1960s, some scholars started to imagine the computer as an infrastructural resource, taking this notion of remote computing to its logical extreme. Likewise, my discussion of ubiquitous computing at Xerox underscores the role that physical presence has had in the experience of computing. The design philosophy of ubiquitous computing at Xerox recognized the importance of local context for implementing smart devices into everyday life.

The changing theories of locality can be seen in thematic shifts in academic work. The early cultural histories of the internet often upheld the binary between the virtual world and the physical.[303] The phrases "cybersphere" or "netizen" seem overly simplistic in today's technological landscape, but they speak to the delineations that were drawn in the 1990s between

[301] Pew Research Center. "Internet Seen as Positive Influence on Education but Negative on Morality in Emerging and Developing Nations." March 19, 2015.

[302] Light, Jennifer. "When Computers Were Women." *Technology and Culture* 40, no. 3 (1999).

[303] Woolgar, Steve. *Virtual Society? Technology, Cyberbole, Reality*. Oxford: Oxford University Press, 2002.

a virtual culture and material world. The focus on online communities as new cybercultures was emblematic of this viewpoint.[304] As networked computing has bled into all aspects of contemporary life, this distinction lost some of its theoretical power. Still, the core theme in this set of literature still resonates today. These authors were discussing the rise of digital culture and communities, which have increasingly been enmeshed with everyday life. More contemporary works in the sociology of computing have recognized this change and have not embraced this binary view of computing, instead choosing to focus on more specific components of computing (e.g., the impact of social media[305], political influence[306], or media consumption[307]).

More recent literature has looked at the rise of mobile technologies and the impact of computing outside of the desktop. In particular, in countries without sufficient ground-level infrastructure, mobile computing has raised interesting new questions for researchers. Likewise, the development of "smart-environments" harkens back to the issues previously raised at Xerox PARC and is currently challenging our established computing norms on all fronts: from the privacy of devices in the confines of the home to the Orwellian concerns of a digitally-empowered totalitarian state.

Turning our attention to the cloud's place in this shift in computing, the cloud is the backbone of much of this shift. It is easy to identify the ways in which the cloud is working alongside traditional political borders. More recently, companies have built clouds around national norms and laws. These localized clouds adapt to the place in which they are built. This can be seen in the creation of cloud infrastructures that conform to specific governmental demands (for instance, the United States works with Amazon's "GovCloud").[308] Complying with local, national, and international laws has also pushed cloud providers to provide more individualized solutions, which has splintered the cloud's claim of a globalized marketplace.[309] In regions with a culture of higher consumer protections, such as the European Union, many governmental figures have been slow to adopt a cloud primarily controlled by companies in the United States.[310] Likewise, in countries with more permissive privacy laws, cloud providers have been careful to not move into regions where the integrity of their consumers' data could be compromised.[311]

[304] Turner, Fred. *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism.* Chicago: University of Chicago Press, 2006.

[305] Greenhow, Christine, Julia Sonnevend, and Colin Agur. *Education and Social Media: Toward a Digital Future*. Cambridge, MA: MIT Press, 2016

[306] Morozov, Evgeny. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York, NY: Public Affairs, 2013.

[307] Jenkins, Henry, Sam Ford, and Joshua Green. *Spreadable Media: Creating Value and Meaning in a Networked Culture*. New York, NY: New York University Press, 2013.

[308] Amazon. "AWS GovCloud (US)." https://web.archive.org/web/20181231135523/https://aws.amazon.com/govcloud-us/ (Archived December 31, 2018).

[309] Hoover, J. Nicholas. "Compliance in the Ether: Cloud Computing, Data Security and Business Regulation. *Journal of Business & Technology Law* 8, no 1 (2013).

[310] Billings, John T. "European Protectionism in Cloud Computing: Addressing Concerns Over the Patriot Act." *Commlaw Conspectus* 21 (May 2013).

[311] Menon, Gowri. "Regulatory Issues in Cloud Computing – An Indian Perspective." *Journal of Engineering, Computers, & Applied Sciences* 2, no 7 (2013).

In the past decade, the place of computing has become both distant and local. For most users, data stored in the cloud is not guaranteed to be in a particular location. The information can flow and be replicated across the globe. Smartphones are constantly pinging different servers and lean on these remote resources to make the smartphone experience seamless and integrated. The operating systems that power these devices are coded with the cloud as a critical design component, one which is difficult to untangle. In this sense, the metaphor of the cloud projects itself as a universal force.

The place of computing has also become increasingly local. When service to a local cloud is disrupted, whether it be an AWS server or a content delivery network for Netflix, massive issues can arise. As city planners attempt to integrate smart sensors into cities, the public will rely more heavily on the smooth operation of a nearby cloud server. This shift in computing has led to a new type of distributed centralization. A handful of major cloud providers control the majority of information processing, storage, and distribution capabilities. These providers are building out their networks to more cities and are attempting to shape these clouds around local needs and requirements.

These issues with geography, coupled with the changing relationship between users and computing devices, have created a new type of computing landscape. This new landscape is one of selective distance, where clouds are anywhere and invisible. Where computing "happens" has dramatically shifted as the ubiquity of devices has exploded, and data is being stored and processed remotely. In order to come to grips with this change, it is worth exploring the impact that these shifts have had on both the discourse of computing and the material world around us. In particular, that means looking at the manifestations of the cloud in real space. In the following section, I take that challenge to heart by exploring my own technological backyard.

The Clouds of Los Angeles

Cloud computing is rarely thought of as belonging to the domain of material culture, let alone a local experience. This disconnect is largely aided by the metaphor of the cloud that suggests a lack of a physical connection. The metaphor asks users to gaze rather than hold. In order to inject more materialism into the discussion of the cloud, I decided to turn towards my local environment to see how the cloud has a physical impact on the ecology of the city. In this section, I study the materials of the cloud by looking at the impact of the cloud on the city of Los Angeles. By looking at a cloud that is local to me, I attempt to tie the notion of local knowledge to an otherwise monolithic cloud environment.

This investigation is not the first look at the material impact of computing. The history of computing, in general, regularly entangles itself with material culture. Early on, historians of technology argued that in order to understand computing, the researcher needs to understand the social context outside of the hardware.[312] These historians have also argued that despite the initial cheers from Internet optimists, geography has always been a factor in directing the type of

[312] Mahoney, Michael. "The History of Computing in the History of Technology." *IEEE Annals of the History of Computing* 10 (1988). 113-125.

computing that occurs on a local level.[313] One example of this scholarship in practice is Paul Ceruzzi's book on the history of computing in Virginia's Tysons Corner.[314] More specifically, in regards to data centers and the cloud, there have been a few books that have addressed some of the material culture of the cloud. Andrew Blum's book, *Tubes*, demonstrated for non-academic audiences the people and buildings that help the internet function.[315] A more academic approach can be found in Tung-Hui Hu's book, *The Prehistory of the Cloud*, which touches on some of the material aspects of the cloud's history.[316] Hu's argument spends much of the time discussing the cloud as a cultural phenomenon, but his work also looks at the cloud's relationship to physical infrastructure such as railroad tracks and data centers. All of these works approach the issue of material culture in a slightly different way depending upon the methodological approach and the type of computing culture they are studying. In order to study the cloud, it is useful to look at the assemblages of the cloud by looking at not only the data centers but also the larger physical infrastructure that links these remote data centers together.

The Cloud Freeway

In Los Angeles, the major freeways act as arteries for the sprawling city and lines of traffic. Alongside these concrete highways is another network. Under the surface of these streets are networks of fiber optic cables that route themselves from the beach, towards downtown, and to the rest of the US. These cable routes, like freeways, are placed on public land for the benefit of private infrastructure. However, the invisible nature of these cables makes this public debt less obvious. Traffic, both above ground and in these data pipes, keeps growing each year. The spread of the cloud has had a measurable impact on the Southern Californian economy and the physical makeup of the environment. This impact, however, is largely unrecognized and underappreciated. This section looks to open up this discussion by studying how the cloud's present resonates with the area's past.

Not all clouds are built alike because they exist in the terroir of a place. While the underlying computing hardware may be more or less universal, the deployments of these resources are tied to the physical geography and the ideology of a region. The introduction of a cloud to a particular region attempts to measure the needs and desires of a location. For instance, when Google established a regional cloud in Sydney, Australia, the company placed focus on the importance of local access by contrasting their cloud to data centers located in South East Asia (focusing primarily on a reduction in latency). Cloud computing has raised a number of regulatory and technical challenges in the face of national interests, which have attempted to control the shape of new regional clouds.[317] These legal questions often intersect with discussions of the materials of a cloud (such as the location of the cloud, energy use, and how humans are permitted to move within a data center).

[313] Goldsmith, Jack and Tim Wu. *Who Controls the Internet? Illusions of a Borderless World*. New York: Oxford University Press, 2006. 46.

[314] Ceruzzi, Paul. *Internet Alley: High Technology in Tysons Corner, 1945-2005*. Cambridge, MA: MIT Press, 2008.

[315] Blum, Andrew. *Tubes: A Journey to the Center of the Internet*. New York: Ecco, 2013.

[316] Hu, Tung-Hui. *A Prehistory of the Cloud*. Cambridge, MA: MIT Press, 2015.

[317] Singh, Jatinder, Jean Bacon, Jon Crowcroft, Anil Madhavapeddy, Thomas Pasquier, W. Kuan Hon, and Christopher Millard. "Regional Clouds Technical Considerations." *University of Cambridge Computer Laboratory: Technical Report #863*.

Measuring the material impact of a data center is traditionally discussed in terms of energy usage, both to operate and to cool the computers. This environmental concern can be seen in cloud data center design, which often takes into account the natural world. For instance, data centers located near the Arctic Circle utilize both colder ambient temperatures and more readily available hydroelectric power to increase their power efficiency.[318] More experimental data centers by Microsoft have attempted to place data centers underwater as an alternative cooling method.[319]

In light of these green visions, data centers in Los Angeles have not been presented or designed according to more environmentally progressive standards. For the most part, data centers in the region are standard designs that are cooled using typical air-cooled designs drawing upon the region's power grid. In press releases and promotional materials for local data centers, discussions of energy consumption are minimal. In terms of environmental certifications, the response is mixed. Some data centers in the region have obtained lower-tier LEED certifications (LEED is one of the most common certification models for environmentally responsible building design). Looking at other data center regions, Los Angeles appears to lag behind in these green certifications.[320]

The region's power grid still relies primarily upon natural gas for the majority of electrical generation.[321] Energy usage seems, as a whole, to be an afterthought for the area's data centers. This lukewarm approach to data center design sits alongside the city's infamous legacy as one of the most air-polluted places in the United States.[322] Efforts to reduce this pollution have mostly occurred in larger political arenas. The mayor of Los Angeles and the previous governor of California both implemented ambitious initiatives to curb California's use of energy and reduce air-pollution (Los Angeles's "Green New Deal" and California's SB 100, respectively).[323] Neither of these legislative moves directly address the energy usage of data centers.

This focus on energy usage often overshadows other material impacts that the construction of a cloud data center has on a region. These impacts include everything from regional economic changes to shifts in human labor. While much of a cloud data center's operations can be handled remotely, there is still a need for in-person technical and administrative support. On-site labor is needed for the smooth operation of a cloud data center. If there is an issue, manual swapping of

[318] Sverdik, Yevgeniy. "Facebook Data center in Sweden Has New Hydro-Powered Neighbor." *Data Center Knowledge*. October 22, 2015. https://web.archive.org/web/20160320084719/http://www.datacenterknowledge.com/archives/2015/10/22/facebook-data-center-sweden-new-hydro-powered-neighbor/ (Archived March 20, 2016).

[319] Kehe, Jason. "Undersea Servers Stay Cool While Processing Oceans of Data." *Wired* December 17, 2018.

[320] Equinix. "Green Certifications." https://web.archive.org/web/20170609062317/http://www.equinix.com/company/green/green-certifications/ (Archived June 9, 2017).

[321] California Energy Commission. "Total System Electic Generation 2017." https://web.archive.org/web/20190518095425/https://www.energy.ca.gov/almanac/electricity_data/total_system_power.html (Archived May 18, 2019).

[322] Amercian Lung Association. "State of the Air 2019." https://web.archive.org/web/20190421231001/https://www.lung.org/our-initiatives/healthy-air/sota/city-rankings/most-polluted-cities.html (Archived April 21, 2019).

[323] Roth, Sammy. "L.A. Mayor Garcetti's 'Green New Deal' Would Phase Out Gas-Fueled Cars." *Los Angeles Times*. April 29, 2019.

computer components may be needed. Furthermore, on-site assistance is often needed for third parties who lease computing space within the building's infrastructure. Typically, these operations need to be staffed (at least in terms of security guards) on a continual basis.

As a whole, data centers support a small number of jobs. According to federal estimates, most large data centers (roughly 20,000-75,000 square feet of compute space) in the United States will support 157 local jobs after the building is completed.[324] Relative to other projects of a similar scale, data centers produce fewer jobs for the amount of land used and financial investments made. The jobs that are produced are largely high-paid technical professionals. Depending upon the region, data centers can either bolster existing labor markets or create new ones. In the case of Los Angeles, a large labor pool (along with a rising tech-startup culture) likely did not have to draw upon external workers. Other cases, particularly in other states which offer tax incentives for the construction of a data center, may create new jobs but not for the residents that currently live there.[325]

Despite the small increase in jobs, changes in labor markets can have an impact on the material culture of a region. Data centers may encourage the construction of nearby industries that draw upon the data center's resources. Conversely, data centers that are constructed near existing industrial hubs (as is the case in this chapter) may help strengthen existing businesses and jobs. At the same time, the introduction of a data center may also challenge smaller data centers or technical-support companies (due to the fact that the businesses may be incentivized to shift their computing workload to the cloud and therefore need reduced technical support). These shifts in human labor are difficult to measure but are important to pay attention to when considering the role of the cloud in labor markets.

There are two major hubs of cloud computing in the Los Angeles area. The first is located in downtown Los Angeles. Centered in, and adjacent to, the One Wilshire Building is a collection of cloud data centers. The second hub sits roughly ten miles away in the coastal city of El Segundo. Both of these locations are part of a larger fiber-optic loop that cuts under the city streets. While there are other data centers in the region, these two locations are the densest and most important locations of cloud computing in Southern California. Before looking at the data centers of Southern California, I need to briefly touch on how the history of the region is tied to the development of the cloud.

The clouds of Los Angeles have their own regional flavor. When looking at the culture of a region, it is useful to start by looking at regional mythologies. Southern California, in specific, belongs to the mythology of American expansion. In this tradition, the West is thought of as a place of growth and potential. The ideology of manifest destiny towards the coast had a lasting impact on the psychology of the Pacific states. The history of computing in California has been tied to this mythology of improvement, and more recently, a critique of California utopianism. The dominant expressions of this improvement ideology have been mostly seen in Northern California, arising from the birth of the computer industry in Silicon Valley during the

[324] U.S. Chamber of Commerce. "Data Centers: Jobs and Opportunities in Communities Nationwide." *U.S. Chamber of Commerce Technology Engagement Center*, 2017.

[325] Tarczynska, Kasia. "Money Lost To the Cloud: How Data Centers Benefit from State and Local Government Subsidies." *Good Jobs First*, 2016.

counterculture movements. Despite living under the shadow of Northern California, Southern California has its own distinct history of computing. Government spending in the late 1940s and 50s helped fuel many of the early computing projects in the military and aviation business. In many coastal cities surrounding Los Angeles, early forms of digital computing (especially as part of the aerospace industry) were critical to the economic development of the region.[326]

Southern California's role in networked computing emerged in full with the early ARPANET experiments. UCLA featured prominently as the location where the first message was sent over the network.[327] Despite some important contributions, Southern California has often lacked a place in the history of computing. Related industries, such as aerospace and computer graphics within Hollywood, have had a large influence on the development of computing as a whole. Aerospace organizations have steadily contributed to advancements in computing. Early supercomputers and development in computer-aided design were deeply influenced by the Southern California aerospace industry.[328] Likewise, Hollywood has been a driving force behind a number of advanced computing graphic improvements.[329]

The previous chapter started to open up the question of how a new regional cloud enters a culture with a distinct history. The rest of this chapter fleshes out that question by looking at the cities that house the two clusters of cloud data centers. Before looking at these areas, I had to meet the cables at the beach.

Ocean

It was a cold and wet morning when I arrived at Dockweiler State Beach in Playa del Rey. At the time of my visit, the Pacific Ocean was choppy and the normally crowded beach was sparsely populated. The beach was once a popular surfing destination, but various coastal management policies to prevent erosion calmed the waters. Looking northward from the shore, you can see the white buildings of Santa Monica in the distance. From this vantage point, you can see the playfully named "Silicon Beach," which refers to the numerous tech companies that have set up corporate campuses in the affluent beach town. Many of the major cloud providers call this region home, from YouTube's campus to Salesforce. The companies were drawn to the region largely because of the climate, a new labor pool, and the proximity to the entertainment industry. Part of the rise of the Los Angeles cloud can be tied to the rise of this new technology center. For the moment, my concern was with what was below my feet.

Dockweiler Beach sits directly adjacent to the final sections of runway for the Los Angeles International Airport. Consequently, the area is continually filled with the sounds and sights of jets taking off to their destinations, flying just a few hundred feet above visitors' heads. Although

[326] Westwick, Peter J. *Blue Sky Metropolis: The Aerospace Century in Southern California*. Berkeley, CA: University of California Press, 2012.

[327] Kromhout, Wileen Wong. "UCLA, Birthplace of the Internet, Celebrates 40th Anniversary of Network's Creation. UCLA Newsroom. October 15, 2009.

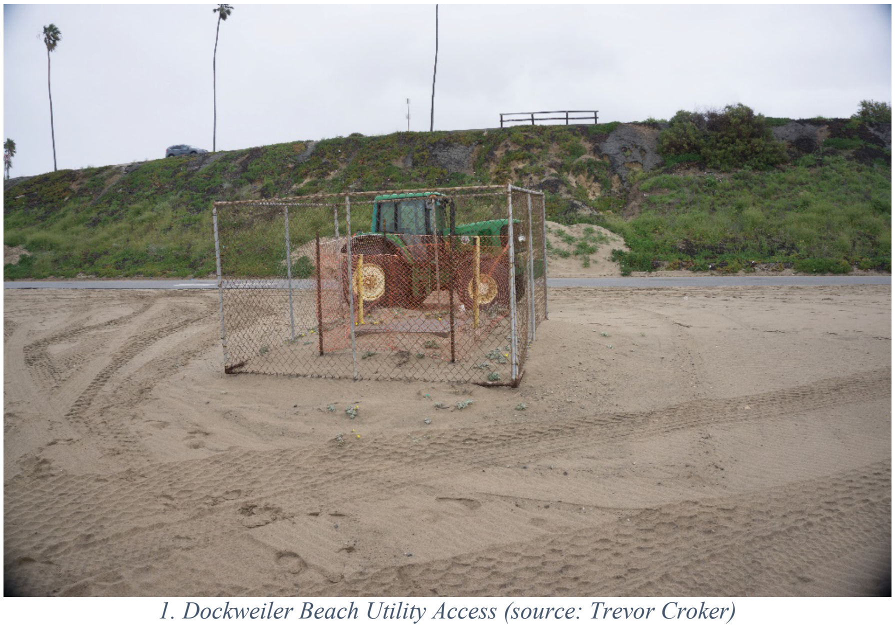
[328] Jameson, Antony. "Computers and Aviation." In *From Physics to Daily Life* (edited by Beatrice Bressan). Weinheim, Germany: Wiley Blackwell, 2014.

[329] Ryu, Jae Hyung. "Reality and Effect: A Cultural History of Visual Effects." Dissertation, Georgia State University. 2007.

you wouldn't know by looking, the amount of human traffic flying above is matched only by the amount of information traffic traveling below.

I made a trip to the beach to examine one of the newest cable landing spots in the region. On this section of coastline, multiple submarine cables have been brought ashore by specialized cable-laying ships and buried underground. One of the more significant cable projects at Dockweiler Beach is the Trans-Pacific Telecommunications Cable Hub. This hub is currently serving as the conduit for the Pacific Light Cable Network (PLCN), which links Hong Kong to the United States. This is significant to the cloud as both Google and Facebook have partial ownership in the PLCN. [330] The hub can support three additional fiber optic cables through the steel bore pipes that guide the cables to the beach.

When I first arrived at the location, a light rain started to come down from the clouds above. The irony, of course, was not lost upon the researcher. Despite the obvious clouds above, there were almost no visual clues of the important infrastructure buried under the beach. Between the RV parking lot and the Los Angeles maintenance facilities sit children's playground equipment and various machines for raking and moving sand. The only clue of any cabling is a utility access hole. Two fences, one metal and the other plastic, surround access to the hole (figure 1). The environmental reports on the project mention these manholes as the locations where the ocean route reaches the terrestrial system. [331]



1. Dockweiler Beach Utility Access (source: Trevor Croker)

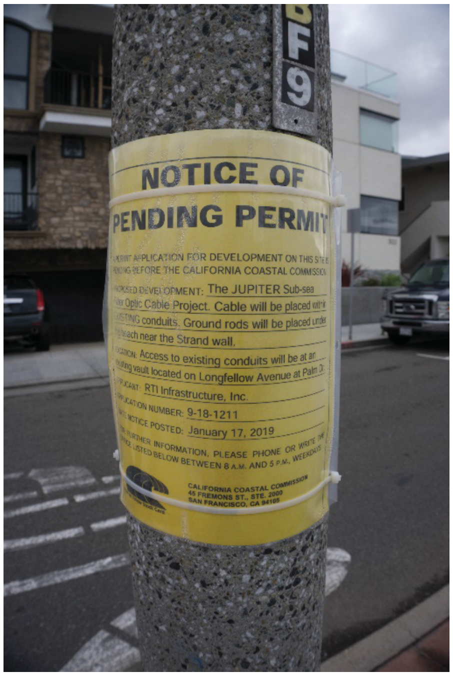
My experience at Dockweiler Beach was similar to my visit to another popular cable landing spot. Five miles south, in the city of Hermosa Beach, multiple trans-pacific cables are buried under the sand. Hermosa Beach is a popular landing site for submarine cables, in part, because the city (rather than the county) owns the beach. This helps clear up some of the regulatory

[330] TE Connectivity. "TE News: Facebook, Google, PLDC and TE SubCom to Co-Build the PLCN Submarine Cable Network." *TE Connectivity*. October 12, 2016.

[331] City of Los Angeles Department of Public Works – Draft EIR Public Meeting. *Coalition Court Reporters*. June 6, 2017.

issues that other locations might face because the city has a history of dealing with these projects. In exchange, the city receives an initial payment by the cable owners and annual usage payments. Over five years, this ultimately nets the city millions in revenues.[332] The city snakes the cables through the streets to inland power feeding equipment to "regenerate" the signal.[333] At this point, the cables are routed towards downtown LA to an unnamed data center.

When I visited the Hermosa Beach location, the only sign of the cloud was another pair of manholes. As I was starting to leave the site, I spotted a yellow paper for a submarine cable permit (figure 2). The notice from the California Coastal Commission announced the pending permit for the JUPITER cable system using existing cabling conduit. This 8,700-mile cable will connect Los Angeles to Japan and the Philippines. The JUPITER cable is just one of many other cables that are buried along the city's sand. Like the cables at Dockweiler, JUPITER is touched by the cloud in that it is owned by a consortium of tech giants: Amazon, Facebook, and SoftBank.[334]



NOTICE OF PENDING PERMIT

PERMIT APPLICATION FOR DEVELOPMENT ON THIS SITE IS PENDING BEFORE THE CALIFORNIA COASTAL COMMISSION

PROPOSED DEVELOPMENT: The JUPITER Sub-sea Fiber Optic Cable Project. Cable will be placed within EXISTING conduits. Ground rods will be placed under the beach near the Strand wall.

LOCATION: Access to existing conduits will be at an existing vault located on Longfellow Avenue at Palm Dr.

APPLICANT: RTI Infrastructure, Inc.

APPLICATION NUMBER: 9-18-1211

DATE NOTICE POSTED: January 17, 2019

FOR FURTHER INFORMATION, PLEASE PHONE OR WRITE THE OFFICE LISTED BELOW BETWEEN 8 A.M. AND 5 P.M., WEEKDAYS

CALIFORNIA COASTAL COMMISSION
45 FREMONS ST., STE. 2000
SAN FRANCISCO, CA 94105

2. Notice of JUPITER Cable (source: Trevor Croker)

My trip to both of these cable landing sites was unremarkable in visual appeal, but revealing in what it says about the life of the cloud. The "boring" locations, if I may evoke the pun, are monotone by design. The everyday nature of these cabling systems sits in stark contrast to the flashy promise of the cloud. The installation of these cable systems is uncontroversial. Looking at the public feedback for both the Dockweiler location and the Hermosa Beach location, the

[332] McDonald, Ryan. "Undersea Cable Installation Continues in Hermosa Beach." *Easy Reader News*. January 14, 2017.

[333] Hermosa Beach. "Draft EIR: Power Feed Equipment Facilities." December 2015.

[334] Buckley, Sean. "Amazon, Facebook, and SoftBank to Build New Submarine Cable System." *FierceTelecom.* October 30, 2017. https://web.archive.org/web/20180731024142/https://www.fiercetelecom.com/telecom/amazon-facebook-and-softbank-to-build-new-14k-submarine-cable-system (Archived July 31, 2018).

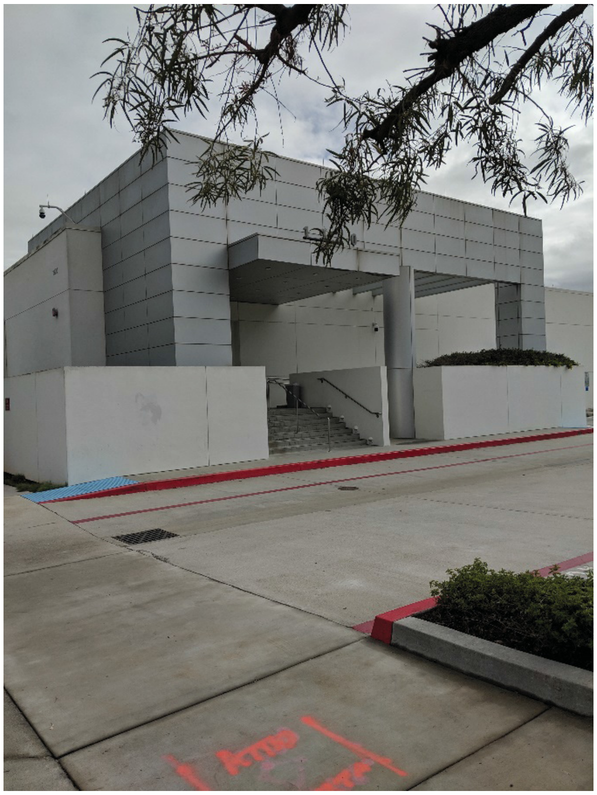
feedback was mild.[335] Most were concerned with the impact that the projects would have on traffic and noise levels during construction. There was also a general confusion about the purpose of the cables. Ultimately the projects were approved by both cities, with minimal changes to the initial planning documents. For my next step, I followed these cables to their final destination.

In my quest to discover what it means to think of the cloud as a material technology, I attempted to follow the submarine cables into the data centers. This journey took me to two different regions: the city of El Segundo and downtown Los Angeles. For as much as the cloud is thought of as dispersed, the concentration of these data centers helps demonstrate that clouds are linked to broader assemblages of technology and networks of people.

El Segundo

My first visit took me from the sand at Dockweiler to the streets of El Segundo. Following the cable route, as seen in planning documents, I moved down the main road and through a quiet residential neighborhood. The trail eventually ended at a cloud data center operated by the American company Equinix. Equinix operates two data centers in El Segundo. Their "LA3" data center is the point at which submarine cables emerge from the ground and feed into Equinix's equipment (figure 3). As I approached the building, I was struck by the lack of color or signage that indicated what the building housed. The grey concrete exterior had no business name, with only the address numbers "1920" giving an indication of the location. Numerous security cameras lined the exterior of the building, spaced roughly ten feet apart. To any passerby, the building is a blank box.

[335] City of Los Angeles Department of Public Works. "Public Review of the Draft EIR 1-2." *Final Environmental Impact Report Los Angeles Trans-Pacific Telecommunications Cable Hub*. August 2017.



3. Equinix LA3 (source: Trevor Croker)

Upon closer inspection, the building had touches of the cloud on the exterior. Looking at the sidewalk outside the data center, the paint markings sprayed onto the street from utility work give a clue about the purpose of the building. The orange paint sprayed on the ground is a marker of communication cables and conduits buried underneath (figure 4).[336] Next to these sidewalk markings are little orange flags from AT&T that warn of buried fiber optic cables. On the backside of the facility are cooling ducts that wrap around the exterior walls. Outside of these details, the building blends into a bland cityscape.

[336] Stamp, Jimmy. "Decoding the City: The Road Graffiti Placed by Utility Workers." April 26, 2013. *Smithsonian.com*. https://web.archive.org/web/20170207110741/http://www.smithsonianmag.com/arts-culture/decoding-the-city-the-road-graffiti-placed-by-utility-workers-42822014/ (Archived February 7, 2017).

4. Cable Sidewalk (source: Trevor Croker)

As much as I would like to describe the interior of the data center, I was unable to obtain permission to tour any of the facilities I mention. After explaining the purpose of my research, none of the data center operators were willing to let me take a peek behind the walls. This unwillingness to open their doors was not unexpected, as the security of these buildings is taken rather seriously. Fortunately, other journalists and academics have documented the layout and logic behind data center design, and I have used this research to supplement my own work. In much of this literature, the argument is that the interior of data centers is reflective of the outside environment.[337] The interiors of data centers are often designed around the needs of the clients and how the building is cooled. Furthermore, this research has argued that data centers tend to form as geographic clusters due to competition from other locations and proximity to nearby infrastructure (like submarine cables or internet exchange points).[338] In lieu of personally being able to step inside the Equinix data center, I pull upon this literature to describe the purpose of the building.

While I may have been unlucky or not persistent enough to be granted access inside the data center, this secrecy could be a more recent development. In the past five years, technology companies have been reluctant as a whole to be public about the expansion of their data centers across the United States. Recent reporting has revealed the use of nondisclosure agreements and shell companies to sidestep public objection to data center projects.[339] Local critique often focuses on the lack of jobs these projects provide, the concentrated use of natural resources, and favorable tax breaks the company may get. There are clear financial reasons why the details of these projects are kept in the dark.

[337] Alger, Douglas. *The Art of the Datacenter*. Upper Saddle River, NJ: Prentice Hall Press, 2012.

[338] Blum, Andrew. *Tubes: A Journey to the Center of the Internet*. New York: Ecco, 2013. 232-233.

[339] Dwoskin, Elizabeth. "Google Reaped Millions In Tax Breaks As It Secretly Expanded Its Real Estate Footprint Across The U.S." *The Washington Post*. February 15, 2019.

Equinix's buildings serve a special function because they are classified as a "colocation" point. Colocation is an important concept in modern networking. Rather than spreading out many different servers across space, colocation points allow multiple servers to be housed in the same space. Many cloud services are hosted in colocation facilities because they give easy access to other networking paths. These specialized locations are generally "carrier-neutral," which refers to a computing space that is open to any number of technology tenants. Colocation data centers allow companies to place their computing equipment in close proximity to one another. This permits quick and reliable sharing of traffic in a local environment, without incurring additional fees. It is also an example of how the success of the cloud as a marketplace is dependent upon local material connections in real-space. In this colocation model, the owner of the building acts as the landlord and manages the facilities, where the clients are responsible for the maintenance of their own servers. Access to the first stop for submarine cable traffic makes Equinix's LA3 an obvious choice for major clouds from the likes of Microsoft, Google, and Amazon. It is also a reminder of how clouds are linked to specific physical networks.

Colocation centers are central to the function of the cloud. While large cloud providers do have dedicated data centers that are exclusively for that company's usage, most clouds are hosted from shared environments. For instance, when Google marketed its cloud coming to Los Angeles, it only mentioned Equinix in passing. In Google's whitepaper on infrastructure security, they mention, "Google additionally hosts some servers in third-party data centers, where we ensure that there are Google-controlled physical security measures on top of the security layers provided by the data center operator."[340] In presenting a cloud to the public, there is rarely a mention of where the data center will be physically located, but the reality of many clouds is that they are part of a shared pool of infrastructural investments.

Turning back to the specific geography of the city of El Segundo, it is clear that the location of the Equinix data center was not accidental. The data center location is tied to a longer history of corporate development. El Segundo, as a city, was formed as a quickly rising industrial town. In 1911, Standard Oil built a refinery on the undeveloped land that was close to nearby oil fields.[341] The oil facilities led to the creation of "black gold suburbs," and an oil pipeline to link the refinery to the ocean.[342] The refinery still is operational today and sits only blocks away from the current data centers. As the city developed, Standard Oil's influence (aided by the development of the nearby airport) shaped the development of the city as an industrial town. In the middle of the century, a number of aerospace companies set up shop. Today, Raytheon, Aerospace Corporation, and Northrop Grumman are the largest employers in the region.[343]

Driving around the city, a number of highly secret facilities sit alongside these secured data centers. Equinix's other location, a $95 million dollar location opened in 2009, sits across the

[340] Google Cloud. "Google Infrastructure Security Design Overview." January 2017. 3. https://web.archive.org/web/20190509131651/https://cloud.google.com/security/infrastructure/design/ (Archived May 9, 2019).

[341] Davidson, Ronald A. "Before 'Surfurbia': The Development of the South Bay Beach Cities through the 1930s. *Yearbook of the Association of Pacific Coast Geographers*. Volume 66 (2004). 80.

[342] Davidson, Ronald A. "Before 'Surfurbia': The Development of the South Bay Beach Cities through the 1930s. *Yearbook of the Association of Pacific Coast Geographers*. Volume 66 (2004). 87.

[343] City of El Seguno. "Comprehensive Annual Financial Report: Fiscal year 2016-2017." El Segundo Finance Department.

street from Northrop Grumman on the East. Across the street to the west is another data center operated by T5. T5's marketing underscores the building's superior seismic engineering to counter the area's frequent earthquakes.[344] A few blocks away sits the office building of Level 3 Telecomm and another data center from Digital Realty (figure 5). The entire commercial zone is tightly linked. Walkways connect the United State's Air Force facilities with Aerospace's offices. Software development companies sit next to Boeing's Satellite Systems campus.

5. Digital Realty Data Center (source: Trevor Croker)

In visiting the data centers in the city, none of them stood out visually. The buildings have no large signs and are difficult to recognize as data centers without prior knowledge. Driving along the backsides of the buildings reveals the large machines that regulate the temperature of the computers inside. My experience mirrored that of other researchers that have looked at computer facilities in industrial parks. El Segundo's feeling mirrors closely the historian Paul Ceruzzi's book on the technology firms in Tysons Corner in Northern Virginia.[345] Years later, journalist Ingrid Burrington, inspired by Ceruzzi's research, visited the nearby city of Ashburn, Virginia to visit new cloud data centers. She concluded her piece by stating that "the incoherent banality of northern Virginia also felt like a fitting aesthetic conclusion to this journey to see the cloud."[346]

My experience in El Segundo reflects this banality. This invisibility of infrastructure seems by design. The image of the data center as a colorful computing space is not universal. Virtual tours of data centers are almost always bright, green, and playful. This belies the reality of most data centers being boring, functional spaces that do not seek attention. There are obvious functional reasons why data centers sit at the end of a submarine cable's route and nearby important governmental and commercial centers. For the clouds of this region, El Segundo's history as an industrial town, access to a large power plant, many corporate neighbors, and access to multiple

[344] T5. "New State-of-the-Art T5@LA Facility in El Segundo Is Now 'Server-Ready'." *T5datacenters.com.* November 28, 2012.
[345] Ceruzzi, Paul. *Internet Alley: High Technology in Tysons Corner, 1945-2005*. Cambridge, MA: MIT Press, 2008.
[346] Burrington, Ingrid. "Why Amazon's Data Centers Are Hidden in Spy Country." *The Atlantic*. January 8, 2016.

submarine cables likely influenced the development of multiple clouds within this small city. These historical and economic realities are not part of the narrative of the cloud but are central to the cloud's story.

Downtown Los Angeles

The second hub of cloud computing is located in downtown Los Angeles. Like El Segundo, submarine cables flow from the ocean into the city center (this time from Hermosa Beach) and extend out to the rest of the United States. Unlike El Segundo's industrial roots, however, the clouds of downtown Los Angeles have been mostly built in response to the city's telecommunication history. In order to understand these clouds' materials, I drove towards the skyscrapers.

Downtown Los Angeles is a space of dualities. Disparities of wealth are evident when looking at the proximity of space. Postmodern theory has regularly picked up on this theme. Political geographers Mike Davis and Edward Soja have previously offered rather bleak interpretations of the city.[347] [348] In all of these accounts, the primary argument is that the use of space matters in the politics of city life. The architecture of a city is, in part, a statement of its values. This is why postmodern geographers, such as Fredric Jameson, have been so interested in downtown Los Angeles buildings such as The Bonaventure Hotel. In his analysis, and my own experience, the hotel is a "bewildering immersion" for the senses. [349] The site of my research, and the home of the clouds of Los Angeles, sits only blocks from this location and offers its own form of immersion and invisibility.

My first time passing by 601 South Grand Avenue in downtown Los Angeles California was an uneventful experience. Looking across the street, you can see an Irish pub, a shoe repair store, and a copy shop. The tall buildings on both sides create a closed-in feeling, which is amplified by the sound of cars and buses slowly creeping along the one-way street. The pavement looks in rough shape, appearing to be in need of a complete resurfacing. The office buildings lining the block did not capture my attention; instead, I hurried along to another destination.

Years later, I revisited the same street to understand what I overlooked the first time I passed by. Suddenly, the nondescript building and the road in disrepair made sense. The building that I passed at 601 South Grand was The One Wilshire Building (figure 6). The building sits in the heart of downtown LA, alongside other office complexes and local landmarks. Strangely, the One Wilshire Building doesn't actually sit on Wilshire Boulevard, but like so many things in this city, projections of beauty may be more important than the reality.

Before touching on the significance of the building to the cloud, let's look at the building's history prior to the internet boom. Built in 1964, the original location of One Wilshire was an empty parking lot that was turned into an office building. At the time, the modern thirty-story

[347] David, Mike. *City of Quartz.* London: Verso. 1990.

[348] Soja, Edward. *Thirdspace: Journeys to Los Angeles and Other Real-and-Imagined Places*. Oxford: Basil Blackwell, 1996.

[349] Jameson, Fredric. "Postmodernism: Or, The Cultural Logic of Late Capitalism." Duke University Press: Durham, NC, 1991.

building was the tallest in that section of downtown, a feature that would have a substantial impact on the building's future.[350] Shortly after construction, the floors were quickly filled by law firms. [351] The office building remained rather unremarkable as far as computing is concerned until the telecommunication deregulations in the 1980s came into effect (as alluded to in the chapter on utility computing).

In order to understand the relevance of One Wilshire, both as it relates to the cloud and the broader internet, we need to look at the breakup of AT&T. In 1984 AT&T was broken up into seven smaller regional telecommunication companies. One of these new companies was Pacific Bell (one of the "Baby Bells" and a holding of Pacific Telsis), which served the Southern California region. As a result of this breakup of AT&T in the 1980s, Pacific Bell enjoyed newly granted market protection as a public utility. Feeling emboldened by this new regulatory environment, the newly formed Pacific Bell took steps to ensure a new regional monopoly. One of these moves was to ban competitors from using equipment on the rooftop of the Pacific Bell regional office on the 400 block of South Grand. [352] At the time. MCI Communications, a rival telco, needed access to rooftop space for microwave communications. Without access to Pacific Bell's rooftop, they made an agreement with One Wilshire to install their equipment on the tallest building in downtown Los Angeles.[353]

[350] LA Conservancy. "One Wilshire."https://www.laconservancy.org/locations/one-wilshire (Accessed December 2018)
[351] One-Wilshire. "History." 2016. https://web.archive.org/web/20180217112035/http://www.one-wilshire.com/explore-one-wilshire/history/ (Archived February 17, 2018).
[352] Hartz, Peter. "L.A.'s Telecom Hotel." *LA Weekly.* September 8, 1999.
[353] Varnelis, Kazys. "Centripetal City." *Cabinet Magazine* 17 (Spring 2005).

1. One Wilshire (source: Trevor Croker)

The decision to install equipment on One Wilshire dramatically changed the future of the building. Multiple telecommunication tenants moved into the building following MCI. Close proximity to Pacific Bell's switching station, along with rooftop access, made the building an attractive location. Floor by floor, networking equipment started to replace the law firms and other businesses. By the 1990s, the building had developed into a major telecommunications hub. Within the walls of the building are miles of interconnected cabling, servers, and other networking gear that link Los Angeles to the rest of the globe.

One Wilshire is now known as a major telecom hotel (or carrier hotel). A telecom hotel is a computer networking location in which multiple telecommunication carriers (typically internet service providers or backbone providers) install their equipment in a shared space. Telecom hotels are similar to "collocation sites," but typically, telecom hotels are purpose-built to house multiple telecommunication service providers.[354] Simply put, the telecom hotel is a neutral location where information-related companies can share infrastructural resources and talk to one another.

[354] NSTAC. "Vulnerabilities Task Force Report Concentration of Assets: Telecom Hotels." *National Security Telecommunications Advisory Committee.* February 12, 2003.

One of the building's distinguishing features is a space called a "meet-me room." This room, along with the rest of the building's infrastructure, is run by the company CoreSite. CoreSite's meet-me room is a physical space inside the telecom hotel that lets different customers connect to one another via a direct data connection. In the meet-me room, literal cables are strung from one company's computers to another company's computers. For instance, if internet service providers want to share information with each other, the room allows for physical cables to be strung between companies. The close proximity not only reduces the amount of latency between networks but it also avoids the cost of sending information across networks (so-called "local loop fees"). The meet-me room at One Wilshire is a useful infrastructure tool for companies to be able to have fast and secure interconnections between multiple providers.

The inside of One Wilshire, as seen in photos and videos of the interior, lives up to this image of Los Angeles's postmodern confusion.[355] The inside of the building is a testimonial to controlled chaos. Networking gear and cabling litter each floor, tying the building together. Miles of multicolored cabling spills out of their cable trays on each floor.[356] Those who have visited the interior of the building point to the density of the equipment and the general mess that this form of computing creates. Interestingly, this vision of a messy computing environment is not part of CoreSite's promotional materials for the building. Instead, the space is presented through images of color-specific routed cabling, clean walls of server racks, and open office spaces.[357] Yet again, this is a vision of the cloud as a controlled and managed space of computing that does not often match the reality of computing where clouds interconnect and legacy equipment sits next to new gear.

Moving from the inside of the building to the outside, the common markings of the cloud can be seen. The street is covered in more utility markings showing the routes of cables into and out of the building (figure 7). Looking closely at the building's edifice, the glass windows give no clue of the massive amount of computing hardware inside. Some of the electrical equipment can be seen on the southern side of the building. The eleven diesel generators are buried inside and on top of the building.[358]

[355] The Center for Land Use Interpretation. "One Wilshire: Telco Hotel Central." 2002. https://web.archive.org/web/20190503142326/http://clui.org/section/one-wilshire-telco-hotel-central (Archived May 3, 2019).

[356] Bullock, Dave. "A Lesson In Internet Anatomy: The World's Densest Meet-Me Room." *Wired*. March 3, 2008.

[357] CoreSite. "One Wilshire Data Center Gallery." https://web.archive.org/web/20180917094848/https://www.coresite.com/data-centers/locations/los-angeles/one-wilshire (Archived September 17, 2018).

[358] One Wilshire. "Explore One Wilshire." https://web.archive.org/web/20180226174406/http://www.one-wilshire.com/explore-one-wilshire/infrastructure-specifications/ (Archived February 26, 2018).

7. Outside One Wilshire (source: Trevor Croker)

The building is critical to the regular maintenance of the cloud and much sought-after space amongst technology companies. The value of One Wilshire has not been lost on the owners of the building. In 2013, the building sold for $437.5 million. At $660 per square foot, it is the most expensive piece of downtown LA real estate.[359] Nearly all of the major telecommunication carries in the United States have their equipment in the building. The success of One Wilshire spurred the development of nearby telecommunication hotels and cloud computing locations.

One Wilshire is the largest data center in downtown, but many other data centers are located only blocks away. This proximity to One Wilshire is by design. Marketing materials from these other data centers mention how close they are to the building and how they are tied into that building's infrastructure. One of the closest data center is Equinix's LA1 downtown location, which has a direct fiber-optic connection between the two buildings (figure 8). AWS, Google Cloud Platform, and Microsoft Azure are housed in this building as part of Equinix's "Cloud Exchange" (mirroring One Wilshire's "Meet-Me Room"). Equinix occupies the sixth and seventh floors but their presence is invisible on the street-level. Digital Realty also operates its equipment out of this building. Further down the same street is Equinix's LA-2 location, Navisite (a cloud data center run by the ISP Spectrum), along with a number of other smaller collocation companies that tie into One Wilshire.

[359] LA Times. "One Wilshire Sells for Record $437.5 Million." *LA Times*. July 18, 2013.

8. Home of Equinix's Cloud Exchange (source: Trevor Croker)

Despite this deep interconnection, the physical environment is almost completely devoid of markings of the cloud. It is difficult to find an entrance to these buildings, let alone signage to direct you. This absence stands in stark contrast to the legacy telecommunication offices and equipment. For instance, AT&T Switching sits a few blocks east. The building is most notable for a massive, now antiquated, microwave tower (figure 9). This tower was used for long-distance calls up until the 1990s.[360] The AT&T building, putting aside the stylized tower, reflects an architectural brutalism in its lack of windows and concrete exterior. The corporate logo is clearly displayed and is reflective of an older era of telecommunications history.

[360] LA Conservancy. "SBC Madison Complex." https://web.archive.org/web/20170419090211/https://www.laconservancy.org/locations/sbc-madison-complex (Archived April 19, 2017).

92. AT&T Switching Center (source: Trevor Croker)

Downtown Los Angeles, as a whole, is the most important location for the cloud in all of Southern California, but there are few physical markings of the cloud. For most internet users in the area, part of their internet traffic will flow through one of the buildings that I have mentioned. Millions of photos, documents, website requests, and emails are stored inside these blank buildings. One Wilshire, in particular, is one of the most important locations for the internet's backbone and the smooth operation of the cloud. In this sense, the cloud is absolutely material and connected to the culture of the region. At the same time, the larger narrative of the cloud attempts to negate this material logic.

Conclusion

My research on the materiality of the cloud points to the challenges of analyzing, engaging with, and potentially making changes to this emerging computing arrangement. I started this chapter by asking the question, what does it mean to consider the cloud a local object? In short, I believe starting to answer to this question requires that we, as a broader public, first acknowledge that the cloud is material and that materiality has consequences. Introducing the idea that the cloud is physical is an important move because it allows us to start to ask questions of the cloud's arrangement and helps break down deterministic ideologies of computing advancement. My trip to see the cloud in person is one attempt to start breaking down the notion of the cloud as immaterial.

A large part of this chapter has been focused on the infrastructure of the cloud. The visibility of infrastructure is often the catalyst for public debate. This is a theme that runs throughout various academic literature, especially STS. The literature on power infrastructure has argued that more visible forms of energy (such as wind farms) meet steeper resistance because they do not blend into the landscape.[361] The cloud, by and large, is hiding in plain sight. When clouds are deliberately made visible, they are often the most beautiful examples. Looking at plain and boring buildings is a step towards breaking down the sanitized vision of the cloud.

The issue of the cloud will only grow more pressing as these networks become more deeply integrated into our lives. One of the most recent trends is the increasing number of smart devices that are being placed in homes, businesses, and cities. These devices are creating a burden upon the cloud. Furthermore, devices like smart traffic lights are more impacted by latency.[362] From these concerns, an extension of the cloud is being formed. Within IT groups, a new concept has been proposed as a layer between the devices and the cloud. This conceptual metaphor is called "fog computing."[363] Like the cloud, the metaphor of "fog" is loaded with its own meanings and interpretations. The idea of fog computing is relatively new and outside of the scope of this dissertation. However, fog computing offers one of the many signals that point to the importance of physical geography in the operations of the cloud. Before tackling the fog, we first need to truly see the cloud.

Unpacking the cloud's black box requires digging through an uneven history and looking carefully at its trace. The cloud is a utility, but it is unlike the visions of utility computing that were imagined in the 60s. The cloud is ubiquitous, but it is not formed around the human-centered design of the 90s. The cloud is shaped by an ever-changing metaphor that adapts itself to the audience that it envelops. It is dependent upon a network of fiber optic cables, but it appears to float above the demands of physics. It lives among us, even as it belongs nowhere. These are the enigmas that need breaking down if we are to seriously grapple with the consequences of this new networked society.

[361] Hirsh, Richard F. and Benjamin K. Sovacool. "Wind Turbines and Invisible Technology: Unarticulated Reasons for Local Opposition to Wind Energy. *Technology and Culture* 54, no. 4 (2013): 705-734.

[362] Bonomi, Flavio, Rodolfo Milito, Jiang Zhu, Sateesh Addepalli. "Fog Computing and Its Role in the Internet of Things." *Mobile Cloud Computing Workshop*. August 17, 2012. 130.

[363] Bonomi, Flavio, Rodolfo Milito, Jiang Zhu, Sateesh Addepalli. "Fog Computing and Its Role in the Internet of Things." *Mobile Cloud Computing Workshop*. August 17, 2012. 130.

Conclusion

Developing Clouds

Since the mid-2000s, cloud computing has been of interest to the technology industry, but it did not become the bedrock of many of the largest content and information platforms until recently. Both financially and culturally, the cloud has meshed itself into the fabric of our digital lives. Cloud services, as a whole, have seen massive growth in revenue, reaching 175.8 billion dollars in 2018 and estimated to hit 278.3 billion dollars in 2021.[364] Today, devices are being designed around the logic of the cloud. This can be seen in the creation of new operating systems that tie user interface to an external cloud. Similarly, mobile computing has been designed with the assumption that much of a user's computational and storage needs will be shifted to off-device servers. Many generally regard the Internet of Things (IoT) and the improvement of artificial intelligence as the future of computing. Both of these domains will be dependent, long-term, upon large, powerful cloud arrays. The story of the cloud is integral to the future of computing, and how we interpret its history will influence how we see our current actions.

The tech giants have bet a large portion of their financial futures on the ownership of this new arrangement of computing. Microsoft, once fueled nearly entirely by software sales, now is fully committed to selling access to its products and services on any platform or hardware device. Facebook and Google, both of which rely heavily upon advertising revenue, continue to build out their own clouds as a way of diversifying their revenue sources and maintaining their ad-platform dominance. Apple, which has often stumbled in providing services, has made a renewed push to sell access to cloud-powered products in light of their waning hardware sales. Amazon, the last of the big five, has continued to deeply invest in its cloud offerings, using that infrastructure as a means to drive its other business goals.

In present-day industry conversations, artificial intelligence (AI) and "edge computing" are seen as the next growth areas. In both of these examples, the cloud is a core component of these ventures. Many AI projects, particularly those under the arm of "deep learning," have to process massive amounts of information in order to train the system. It is often impractical to use a single computer to process this information, and therefore AI systems are generally trained using cloud server farms. Smart assistants, like Apple's Siri, use the cloud to continually train and improve their recognition of human commands and to execute even more complicated requests. Human requests are continually fed into the cloud to refine the system.

Likewise, edge computing is closely linked to the cloud. The "edge" in edge computing refers to the location of servers that can process data. One of the issues that I have continually raised in the previous chapters is the issue of geography. Edge computing is an attempt to move the cloud closer to the users. As more IoT devices are communicating with remote servers, there has been a rise in bandwidth usage as local devices look for answers from far away data centers. Edge computing is an attempt to bring the cloud closer to users by shifting some of that computational

[364] Gartner. "Gartner Forecasts Worldwide Public Cloud Revenue to Grow 17.3 Percent in 2019." September 12, 2018. https://web.archive.org/web/20190808092059/https://www.gartner.com/en/newsroom/press-releases/2018-09-12-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2019 (Archived August 8, 2019).

power to "fog nodes" (small-scale services located in closer physical proximity to the devices at the edge).

These imagined technological futures of smart cities, intelligent assistants, and advanced AI systems are driving current conversations about the future of computing. Importantly, cloud computing is at the root of all of these growth areas. The expansion of AI or IoT necessitates a robust network of high-performance cloud data centers, run by a limited number of private companies.[365] Likewise, the metaphors of edges and fog nodes only reinforce the linguistic power of the cloud as metaphor. As technology innovators attempt to bend our technological imagination towards these new ideas, it is critical that we do not lose grip on the history of the cloud.

Summary

In the previous chapters I told the story of the cloud by framing its creation as a confluence of actions that span multiple decades. The cloud was never born "naturally," it was a creation that was pieced together from fragments of computing history and spun into a whole. The metaphor of the cloud is the glue that binds the disparate parts together. Underneath the metaphor of the cloud are histories situated in specific spaces, made by people with their own idea of how technology should be designed, and existing within real geographies.

In the first two chapters, I provided a historical look back at the formation of the cloud. These chapters served as a means of understanding how the contemporary cloud came to be. The first of these stories looked at the early developments surrounding timesharing and utility computing. Here we see another metaphor (utility) being deployed. Computer scientists at Dartmouth saw the metaphor of utility as a way of imagining computing in the same terms as previous utility systems. In their work on timesharing and building a computer system at Dartmouth, the scientists were successful in spreading their network to other educational systems and some business markets. This same story demonstrates that metaphors are not static. As timesharing became more widespread and commercial enterprises like Tymshare expanded, the metaphor of utility started to be taken seriously by regulators at the FCC who saw the rise of a new information utility that could be placed into a new regulatory framework. Soon after the FCC started looking into regulating these new timesharing businesses, the conversations about building a national computer utility died down. The principles of utility computing, shared and measurable computing resources that can be purchased and consumed at will, live on in the cloud, but the issue of regulation has not followed.

The chapter on ubiquitous computing offers a similar example of metaphor at work. By looking at the creation of ubiquitous computing at Xerox PARC, we can see how metaphors are connected to specific places and ideas. In the case of Mark Weiser's use of "ubiquity," he uses the term in a broader philosophical context. Weiser's desire to create "calm technologies" and technology at the periphery were part of a larger psychologically-driven design philosophy that put human desire at the center of product development. The metaphor of ubiquitous computing,

[365] Lohr, Steve. "At Tech's Leading Edge, Worry About a Concentration of Power." *New York Times*. September 26, 2019. https://web.archive.org/web/20190928165613/https://www.nytimes.com/2019/09/26/technology/ai-computer-expense.html (Archived September 28, 2019).

in that context, played an important role in building experimental computing devices for the work environments. As the idea of ubiquitous computing left PARC and moved into the broader world, the network of concepts broke down, and the meaning of the metaphor shifted. Ubiquity has come to mean devices everywhere, rather than a network of smart devices attuned to the demands and limitations of humans. The cloud, which grew in part from the idea of ubiquitous computing, is growing divorced from this broader context. Bringing this history back to focus may help us develop calmer clouds that are centered on human needs.

The remaining chapters bridge the gap between the cloud's past and the present. The cloud first appeared as a visual tool used in early computer networking maps. In these maps, we see the image of the cloud being deployed as shorthand for a computer system outside our control. Alternatively, we also see the symbol of the cloud representing all computers within a network. There is a duality at play in this visual imagery that is later seen in the coinage of the cloud as a term.

The cloud as both an image and a concept has demonstrated flexibility in application and scale. The technologies of the cloud have been shrouded in a metaphorical framework that obscures itself from easy critique. There is a vagueness to the language of the cloud that helps conceal specific technologies at work, economic motivations, bodies in action, and the larger shift in how computing is structured. In the early 2000s, we see technology companies building large distributed computing platforms for their own purposes. Excessive amounts of idle computational power, alongside the growth of broadband and new virtualization software, created the foundation for a new hybrid of utility and ubiquitous computing (without adopting the messy history from computing's past).

These chapters work against the narrative that the cloud is singular and without a place. Rather, I argue that the language of the cloud only works to obfuscate the larger infrastructural shift that is happening. The expansion of submarine cables demonstrates the intense material realities of constructing a new global digital infrastructure and the intensely local nature of the physical cloud. Cables and data centers require enormous amounts of capital and are long-term investments that impact different communities. Cloud companies must consider a number of variables when deciding where to build out their infrastructure. Issues of environmental impact, labor demands, and local politics are always at stake but rarely are at the forefront of our conversation about the cloud. The materials of the cloud work to make the current system invisible. Making the cloud visible requires looking at the fiber optic cables buried under the ocean and the data centers hidden behind a generic edifice.

Arguments

Before addressing practical steps to build a more equitable cloud, I want to reiterate the primary arguments running through each chapter. This starts by looking at the relationship between metaphor and materiality. It is far too easy to think about the cloud in simple monolithic terms. Even the phrase, "the cloud," implies a singular system rooted in an unknown place. In reality, there are competing computer networks with differing technical standards, hardware, and software. There are multiple clouds, each with different layers and services provided to different actors. The "growth of the cloud" is not the spread of a singular cloud but is actually the result of many companies attempting to gain infrastructure supremacy by establishing more interconnected, powerful, and geographically diverse remote computing networks. These material struggles rarely enter into our conversation about the cloud, even as they are the driving force behind its expansion. Instead, most public understanding of the cloud comes from metaphor.

The metaphor of the cloud has always been, and continues to be, a marketing tool for cloud operators. The use of the metaphor was not created for the benefit of the end-user. The cloud metaphor works primarily in the service of marketing departments as shorthand to disguise the politics and values embedded inside networks. The use of metaphor often removes important details about how information is stored, shared, and moved across networks. Putting data "in the cloud" lacks a specificity that makes it harder to ask questions about control, regulation, and possibilities for alternative cloud arrangements. I argue that the history of the cloud has demonstrated that major cloud providers have purposely embraced the metaphor of the cloud over other metaphorical frameworks in order to discourage critique.

One of the promises of the cloud is that it frees us from having to think about where computing resources should be located. Information that was once stored locally can now be pushed to a remote server that can be accessed on-demand. Ironically, this push to liberate us from the need to think about local storage has also caused the infrastructure of the cloud to become more material. In order to store and transmit all of our information remotely, cloud companies are struggling to build new data centers and lay new fiber-optic cabling to manage this new resource. Only through a massive expansion of the physical infrastructure of the cloud are we able to maintain the illusion that our information is weightless and disconnected from a place.

The global expansion of the cloud presents us with another opportunity to consider the cloud as a utility system and the possibilities for regulation. The cloud has developed largely outside regulatory pressures and has consequently faced little oversight. The debates over utility computing in the 1970s have not reemerged even as the cloud has become more central to computing. This is likely to change in the coming decade as the largest cloud providers will face increased scrutiny as the cloud entrenches itself into the fabric of the web. The metaphor of the cloud has helped dampen some of this pressure, but regulators and the public are increasingly worried about issues of privacy and data control in ways not seen before. If these debates do reemerge, they will be in the context of a more fully developed cloud computing network. Arguments over how the cloud "ought" to operate will need to be more forceful in the face of a system with a high degree of technological momentum. I argue that looking at the material infrastructure of the cloud in the context of the cloud's history will give all parties the best chance of interrogating the politics of the cloud.

The arguments that I have made have all been developed alongside the challenge of grappling with both the power of metaphor and materiality. To say that the cloud exists only in language is to overlook the massive physical changes that have happened to make such a network possible. Conversely, it does not make sense to discuss the infrastructure of the cloud without considering the role of metaphor in influencing the development of the network. We need to think about how these two domains (language and materials) work simultaneously. Ultimately, this starts by looking at the values imbedded in ideas and objects.

Praxis

The focus on metaphors and materials helps provide a general framework for understanding the cloud today. Currently, the metaphor of the cloud remains a powerful tool in being able to sell computing services without delving into the messy interior of the cloud. As the term starts to stabilize, we risk making the black box more difficult to open. I suggest that we work on building a new mental framework for the cloud by reinjecting material into metaphor.

Practically, this means explicitly discussing the actual technology, places, and people that develop and maintain the cloud. This task is easier than it may first appear. If we look at the way that large businesses adopt the cloud, we can see one possible model for grappling with the cloud. As I have already argued, businesses often demand more information about the materials of the cloud and have greater bargaining power over what type of cloud they purchase. For instance, a business will be able to negotiate what level of control they want to manage and what they would rather leave in the hands of the cloud provider. If a company wants to have limited access to their servers, that is something that can be built into a cloud contract. Conversely, a business can choose a less-privacy or security-focused solution if they do not deem it important. In short, enterprise users are given granular control over how the cloud enters their domain and can set limits on the expansion of the cloud.

Policymakers and general users ought to use the enterprise cloud as a model if they are interested in making changes to the current cloud arrangement. The computing industry could build these changes into the development of the cloud itself. This could take many forms, but the most straightforward approach would be to build in the same level of user control into all cloud products, regardless of who the audience is. For instance, a cloud-powered email service could give users a choice about where their data is located (geographically), levels of data security (mirroring to multiple locations), and levels of data access (what level of encryption and control is allowed).

As a general model, the cloud should be viewed as a newly developing technological infrastructure. We should discuss issues of public/private control, energy usage, distribution systems, legal frameworks, and what types of commitments we make by maintaining this system. This requires thinking long-term about the type of digital networks we are building. We need to consider the cloud as one type of modality, a way of living digitally. If we value privacy, for example, we can decide to build walled gardens that model a precautionary approach to the expansion of our digital futures. The cloud can be rejected, modified, or embraced. We start having these conversations by recognizing that the construction of the cloud was an intentional decision but the public has had little say in the design philosophy.

For the general public, it is important that the consumer is aware of his or her options when deciding to use the cloud. Firstly, consumers should make themselves active participants in managing their own data. Free online cloud services are, of course, not truly free. Users pay for these services by giving their data to a company that can monetize that information. The consumer rarely, if ever, is given information about where their data is stored and how it is safeguarded. Avoiding the cloud is becoming more difficult as devices and services are built with the assumption of an always-on connection to a remote server. Consumers can decide to store their information locally, or host their own email server, but these options have their own costs. Trying to remove oneself from participating in the cloud may be infeasible for most, but it remains an option for those willing to put in additional resources.

What is Left

Although I have attempted to tell the story of the cloud in the most complete sense, there are still many avenues for future research and expansion. Legal scholars, or STS researchers interested in sociotechnical regulation, could delve more deeply into the implications of managing the cloud at the level of public policy. This research could look at the role of the nation-state in the growth of the cloud and what complexities the cloud injects into international legal systems. Similarly, STS scholars could consider the story of the cloud as a jumping-off point for new research into cloud-powered devices (many of which would provide ample material for case-studies and network analysis).

Critical geographers, material studies researchers, and those in infrastructure studies could expand upon this research by looking into other examples of digital technologies intersecting with the physical geography. How digital infrastructure compares to other traditional infrastructures needs much more refinement and time to develop fully. It would be fruitful to look at how other aspects of the cloud's infrastructure are built and maintained (for instance, the supply of electricity to data centers or how humans work to maintain the cloud on a micro-scale). There are numerous intersection issues at play when considering the place of the cloud in different communities.

There is also a need for a more diverse and global view of the cloud. My research only touched briefly on the development of the cloud in urban and industrialized countries. Access from rural locations, or countries without established wired internet infrastructure, would provide an important perspective into the relative importance of the cloud in other areas of the globe. How cloud providers attempt to modify systems around these communities, and in turn how those people negotiate their relationship to the cloud, would be a highly valuable contribution to the conversation.

As a whole, more work needs to be done on the role of metaphor in driving forward our technological futures. Metaphors are powerful tools that allow us to dream futures. As the story of the cloud also demonstrates, metaphors can simplify and remove complexities. When metaphors are applied to large sociotechnical systems, this combination of imagination and simplification can have serious consequences. We can discard, retool, or invent new metaphors as we see fit. First, however, we need to look through the clouds to see materials at work.