To the Cloud

Big Data in a Turbulent World

Vincent Mosco

pp. 15–75

Paradigm Publishers
Boulder • London

Chapter 2
From the Computer Utility to Cloud Computing

> We are on a shift that is as momentous and as fundamental as the shift to the electrical grid. It's happening a lot faster than any of us thought.
>
> —Arthur R. Jassy, head of Amazon Web Services (Hardy 2012a)

Most general accounts of cloud computing attribute the use of the cloud image to its appearance in diagrams that identify key elements in a telecommunications network. The term *cloud computing* emerged in 1996 when technology leaders with Compaq, then a major desktop-computer company, met to discuss the future of computing and especially the Internet. Specifically, they hoped that "cloud computing–enabled applications" would boost sales. Although not entirely clear about this, they concluded that online consumer file storage would likely be among the successful applications. Their prescience was rewarding for the company because it contributed to Compaq's decision to start selling servers to Internet service providers, which became a $2 billion annual business for the company. However beneficial for Compaq, which HP bought in 2002, the server decision was not as successful for one of the meeting's participants, Sean

O'Sullivan, who went on to start a less than successful firm selling file storage and video-on-demand to individual customers. It was just too early for this cloud to rain dollars, even on innovators with foresight. The genuine growth of the cloud awaited the expansion in computer processing power and in telecommunications networks, as well as a general economic recovery following the dot-com collapse of the early 2000s. It was not until 2006 that the term cloud computing came into more general use as companies, led by Google, Dell, and Amazon, started using the term to describe a new system for accessing files, software, and computer power over the Internet instead of from a computer's own hard drive or some other portable storage mechanism (Regalado 2011).

Defining Cloud Computing

There are those who believe that the first use of the term in the twenty-first century was by Eric Schmidt, Google's CEO, when he described the cloud at an August 9, 2006, industry conference: "What's interesting [now] is that there is an emergent new model. I don't think people have really understood how big this opportunity really is. 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.'" The PC maker Dell saw marketing value in the term, and in 2008 the company tried to secure a trademark for "cloud computing." That attempt, which upset many in the industry, ultimately failed. As a result, anyone was free to use the term and many companies decided that the cloud was a great way to capture the next stage in the development of online services (Regalado 2011).

There is no generally accepted definition of cloud computing. Indeed, one overview suggests that twenty-five cloud pundits would likely define it in twenty-five different ways (McFedries 2012). An entrepreneur who teaches programmers how to use the cloud describes it as "a metaphor for the Internet. It's a rebranding of the Internet. That is why there is a raging debate. By virtue of being a metaphor, it's open to different interpretations." But the debate continues because "it's worth money" (Regalado 2011). Most cloud analysts do not equate the Internet with cloud computing. Although cloud systems use the network of networks we know as the

Internet to transmit data and applications, they also make use of private networks that may be linked to the Internet but are separate from it and accessible to only a fraction of users. Moreover, since cloud computing also involves the customized provision of applications and services, it is generally considered to be more than a network of networks. Although the cloud as a defining concept may eventually withdraw into the powerful banality of technologies like electricity, most agree that it has not yet reached the sweet spot of generic universality (Linthicum 2013e).

As of 2013, years after cloud computing began to circulate in public discourse and well after the first mass advertising, including two commercials that aired during the 2011 Super Bowl, Americans remained unclear about what it means. A survey of 1,000 adults carried out in August 2012 suggested that few people had even a rough idea of what cloud computing means. Nevertheless, most indicated that they expect to be working "in the cloud" in the future and, when they had it explained, demonstrated savvy in understanding its potential problems—primarily price, security, and privacy (*Forbes* 2012).

When the U.S. government decided that cloud computing might be a cost-effective way to deliver services, it pushed departments to consider a move to the cloud. However, when department heads expressed little knowledge of cloud computing, the government's chief information officer asked the National Institute of Standards and Technology (NIST) to come up with a definition and description (Regalado 2011). So the closest we have to a generally accepted formal definition is, in the words of a NIST report, "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" (Mell and Grance 2011). To put it in plainer language, cloud computing involves the storage, processing, and distribution of data, applications, and services for individuals and organizations. It is generally viewed as the fastest-growing, or near the fastest-growing, segment of the IT sector, even though in 2012 it represented only 3 percent of all IT spending (Butler 2012b). NIST's definition of cloud computing has been widely accepted throughout the industry as an objective description of the service. But it is important to understand that cloud-computing descriptions, however objective in appearance, are typically conflated with promotion. Whether it is the federal government's chief information officer, NIST, or the National Science Foundation, which in 2012 announced its own commitment to fund cloud-computing research, the goal is to promote the cloud and not just to understand it. So along with the clear definition, NIST proclaims, "The Cloud Computing model offers the promise of massive cost savings combined with increased IT agility. It is considered critical that government and industry begin adoption of this technology in response to difficult economic constraints" (NIST 2013).

The Early Cloud: The Computer Utility and Videotex

To deepen understanding of what cloud computing means, it is useful to consider how it is both an extension of earlier forms of computer communication and, at least in scale, a new development in the use of information technology. In the 1950s, the computer scientist Herb Grosch forecast a world that would share computing resources so that no more than fifteen data centers would be needed to meet the world's information needs. In the 1960s, the concept of the computer utility emerged when Stanford IT expert John McCarthy imagined "computation as a public utility" (C. Ross 2012). This was formalized in 1966 with the publication of Douglas Parkhill's widely read book *The Challenge of the Computer Utility*. Why is it useful to think of cloud computing as a utility? In part it is because some specialists see the cloud as little more than an extension of the computer-utility concept, once referred to as "time-sharing," because usage time on a central computer was shared by multiple users. For example, according to Linthicum, "If you think you've seen this movie before, you are right. Cloud computing is based on the time-sharing model we leveraged years ago before we could afford our own computers. The idea is to share computing power among many companies and people, thereby reducing the cost of that computing power to those who leverage it. The value of time share and the core value of cloud computing are pretty much the same, only the resources these days are much better and more cost effective" (cited in McKendrick 2013a).

Most people are familiar with public utilities for resources like roads, water, and electricity, which provide services to the public over an infrastructure that utilities manage and operate. They can be owned by government or by private enterprise but when it is the latter, utilities are typically subject to some form of local (city, community) or regional (state, county, province) regulation. Without entering the dense thicket of debate over whether they provide a net public benefit over a competitive market arrangement or whether the government-owned or private utility is best, it is sufficient to state that the utility arrangement is typically chosen because it is expensive to build the infrastructure for water and power. When governments conclude that duplicating infrastructure so numerous competitors can enter the market will likely waste resources, they declare a "natural monopoly" and establish a public utility.

As the concepts associated with computer technology, among them cybernetics, information processing, and communication flows, attracted the attention of a wider circle of scholars and policy makers in the 1950s and '60s, some began to think of information as a resource not unlike water and power. The shift from analog to digital methods of processing information provided a tangible or material output that made it easier to think of information in resource terms. The mathematicians Claude Shannon and Warren Weaver (1949) built a widely accepted model of communication flows that emphasized the materiality of communication over the abstract senders and receivers through which communication flowed. They were less concerned with the social forces that made some people senders and some receivers than they were with identifying communication as a tangible flow. When the economists Dallas Smythe and Herbert Schiller began to turn their attention to communication in the 1950s and '60s, they drew connections between their new field of study and the resources, like agriculture and oil, that had occupied economists for many years (Mosco 2009, 82–89). Around this time the computer scientist turned public-policy analyst Anthony Oettinger developed a general resource theory that linked energy and materials to information, and it became the conceptual foundation for the Harvard University Program on Information Resources Policy, which Oettinger chaired for several decades. When the communication scholar Marc Uri Porat (1977) published his influential map of the shift to an economy powered by information workers, it became time to think about an information economy.

These developments gave renewed force to a view that had been debated since the emergence of postal communication and extended to electronic communication technologies, starting with the telegraph and repeated with the telephone, radio, and television. Is it appropriate and useful to employ the concept of a resource to identify the product of these devices and, if so, should this resource be organized in the form of a utility? Over the years, different constellations of political forces produced different policy responses to these questions. But with the foundation of thinking, for example, about the provision of telephone service as a "natural monopoly," experts examining the output of computer technology began to wonder whether the resources propelling the information economy were creating the need for a new utility.

Advancing this discussion of how to organize information resources, Douglas Parkhill wrote about the challenges facing what he foresaw as the coming computer utility. From the start Parkhill recognized that the idea of organizing computer systems as a utility was in the air: "Even now the subject of computer utilities is very much in the public eye, as evidenced by many articles in both the popular and technical press, prognostications by leading industrial and scientific figures and growing signs of interest on the part of governments everywhere" (1966, v). Parkhill took this popular idea and gave it the clear definition and specificity required to move it forward. For him, there were five key components to the computer or information utility:

1. Essentially simultaneous use of the system by many remote users
2. Concurrent running of multiple programs
3. Availability of at least the same range of facilities and capabilities at the remote stations as the user would expect from a private computer
4. A system of pricing based upon a flat service charge and a variable charge based on usage
5. Capacity for indefinite growth, so that as the customer load increases, the system can be expanded without limit by various means

Parkhill envisioned the computer utility to be a public service in the sense that it would make available to anyone, wherever located, a wide range of information resources and services in an online form. With that said, he did not make a commitment to any specific management form, but rather addressed the merits of public, private, and mixed systems because "it is necessary to consider each application of computer utility separately on its merits and balance off in each case the gains and losses resulting from the adoption of the utility concept" (1966, 125). Elements changed as yesterday's computer utility became today's cloud-computing system, but it is worthwhile to reflect on how much of Parkhill's thought is repeated in today's discussions of cloud services. We are now more likely to ask if a system is scalable rather than if it has the "capacity for infinite growth," but new terms should not mask the striking conceptual similarities. Parker would go on to play an important role in implementing his vision of the computer utility through the creation of what bore the discernible yet odd name of *videotex*. This was a computer-based service that delivered information from a central facility to users at terminals in their homes, in public places, and, to a lesser degree, in businesses. Users were able to interact with the service by making specific information requests. Parker helped bring about the most advanced of these systems in a Canadian government-sponsored project named Telidon. Because its use of color images and its processing demands outstripped the capacity of the existing telecommunications network, the system did not advance far out of the starting gate. Nevertheless, simpler systems featuring more manageable services were widely distributed. The best known of these, France's Minitel service, brought terminals to libraries, post offices, and other public places, providing users with basic information like the telephone directory, train schedules, information on government services, stock quotes, and the opportunity to chat with fellow users and have messages delivered to a "mail box." The service provided millions of connections each month and was not retired until 2012 (Sayare 2012). Videotex held great promise as report after report predicted major transformations in every aspect of life, with comparisons made to the automobile and the television (Tydeman et al. 1982).

Videotex was only one of many cloudlike services that emerged in the pre-Internet decades. In fact, what is very interesting to observe, and often lost in the linear histories that see the past as simple precursor to the present, are the vast arrays of different applications that arose under the resource/utility umbrella. Consider the atlas of clouds represented by the Soviet Union's cybernetic systems of the 1960s, Chile's experiment to bring about computerized workplace democracy and economic planning in the 1970s, and the Pentagon's development of a research computer network that helped to create the Internet from the 1970s to the early 1990s.

Cybernetics in the Soviet Union

In spite of World War II's devastating impact, the Soviet Union produced leaders in the burgeoning field of cybernetics, formally the science of communication and control in machines and animals. In the West, the computer scientist Norbert Wiener led the field of luminaries, with a stellar group that in 1953 included John von Neumann, Claude Shannon, William Ross Ashby, Gregory Bateson, and Roman Jakobson, who met regularly under the auspices of the Macy Foundation from 1946 to 1953. Rebelling against established approaches to theory and applied science, they transformed established disciplines and helped to create new ones. Little was left untouched in fields as diverse as biology, communication studies, computer science, linguistics, and psychology. It might only be the gentlest of overstatements to conclude that cybernetics became a Holy Grail of general theory that many believed would revolutionize human thought (Parkman 1972).

These ideas slowly simmered in Soviet science, permitting quiet questioning of rigid theory enshrined in the work of Trofim Lysenko in biology and Ivan Pavlov in psychology while Joseph Stalin retained his iron grip on power. But when Nikita Krushchev consolidated his control as Premier in 1958, change accelerated and the cybernetics that had been officially denounced as "not only an ideological weapon of imperialist reaction but also a tool for accomplishing its aggressive military plans" was by 1961 hailed as the primary technical means to realize the Communist ideal (Gerovitch 2010). In that year the Soviet Academy of Sciences published *Cybernetics in the Service of Communism*, a detailed examination of how cybernetics would transform practically every field of knowledge and application, but especially, to the pleasure of the representatives meeting that year in the Twenty-Second Congress of the Communist Party, the modern Soviet economy.

For its supporters, economic cybernetics would demonstrate the superiority of the Soviet system by applying the new science to the new technology of powerful computers to precisely plan for the production and distribution of goods and services throughout the Soviet Union. In 1962 the chairman of the U.S.S.R.'s Academy Council on Cybernetics made the importance of the marriage between cybernetics and economic planning absolutely clear when he declared that "However unusual this may sound to some conservatives who do not wish to comprehend elementary truths, ***we will be building communism on the basis of the most broad use of electronic machines***, capable of processing enormous amounts of technological, economic, and biological information in the shortest time. These machines, ***aptly called 'cybernetic machines', will solve the problem of continuous optimal planning and control***" (ibid.). In effect, these words announced the birth of the Soviet computer utility. A network of computer centers would be built across the vast expanse of the U.S.S.R., through which a continuous stream of data would flow from shops, factories, and offices. Planners would use the data to assess the success or failure of policies and to plan, in the most minute detail, future economic activity. Regional computer centers would link up in a nationwide network under the auspices of the Central Economic Mathematical Institute, giving the country "a single automated system of control of the national economy" (ibid.; Spufford 2010). This was a plan for state-directed cloud computing in the service of central economic planning, and U.S. intelligence services—already worried about the growth of Soviet military might—feared what might result.

The CIA responded in 1962 by setting up a special unit to study the threat posed by the Soviet cybernetics initiative. One of the most remarkable conclusions drawn from the spy agency's investigation was the expectation, and consequent unease with the idea, that the Soviet plan would actually succeed. According to its task force report, "tremendous increments in economic productivity as the result of cybernetization of production may permit disruption of world markets" (Gerovitch 2010). The CIA concluded that economic success would bring an additional threat: "The creation of a model society and the socio-economic demoralization of the West will be the added ideological weapon" (ibid.). So concerned was the intelligence agency that it continued to discuss the issue with Kennedy administration officials in the period leading up to and throughout the 1962 Cuban Missile Crisis. The president's people were equally worried. In a memo to Attorney General Robert Kennedy, Arthur Schlesinger Jr., historian and special assistant to the president, concluded that the "all-out Soviet commitment to cybernetics" would give the Soviets "a tremendous advantage" and that "by 1970 the USSR may have a radically new production technology, involving total enterprises or complexes of industries, managed by closed-loop, feedback control employing self-teaching computers." Pulling no punches, he concluded that if the United States continued to neglect cybernetics, "we are finished" (ibid.).

Even discounting for the hyperbole that often accompanies the effort to convince those in power to take action, Schlesinger's statement and those of the CIA amount to a declaration that the Soviets' early version of the cloud, with its central planning through cybernetics, would work and might very well defeat the United States. The furor continued as President Kennedy set up a task force to examine the threat of Soviet cybernetics and the CIA continued to sound the alarm. The U.S. military got into the act, too, with the commander of the Air Force Foreign Technology division alarmed that "the system could be imposed upon us from an authoritarian, centralized, cybernated, world-powerful command and control center in Moscow" (ibid.).

As with many U.S. assessments of the Soviet threat, these fears proved exaggerated. Only a small fraction of the Soviet program was implemented because the government diverted available resources to the military, which steadfastly refused to share them with what top commanders believed was the useless project of the economic cyberneticians. This cloud did not vaporize overnight, however. The Soviet Union's cybernetics team was able to patch together a semblance of a computer system for planning and allocating resources, producing less than a robust network, more mist than cloud. Moreover, it took a national network of human "fixers" whose job it was to use whatever means necessary to keep chains of production and distribution working, or, at least, keep them from seizing up entirely, so that the façade of central planning through cybernetics and what Francis Spufford (2010) called the belief in "Red Plenty" could be maintained.

The Soviet Union's dalliance with an early version of cloud computing demonstrated both the potential and the pitfalls of using it for national economic planning. Most analysts have understandably focused on negative lessons, including some combination of the inherent difficulty of developing a cloud model for a massively complex economy, the structural problems built into the Soviet system, and the recognition that computers were not nearly advanced enough to carry the load. Scholars are just beginning to assess the actual potential of the Soviet cybernetics program to meet the government's economic goals (Dyer-Witheford 2013). It would also be interesting to consider the impact of the cybernetics program on the ultimate opening of Soviet life. We know that it permitted scientists and intellectuals to consider alternatives to Stalinist absolutes. Perhaps if more than one generation had continued to work on the program, cybernetic planning might have nudged open more doors in the Soviet Union. We do know that one alternative early computer utility or cloud experiment, Chile's Project Cybersyn, was influenced by the Soviet cybernetics project, but it departed from the Soviet project in significant ways as well.

The Computer Utility Comes to Chile (Almost)

After the people of Chile elected Salvador Allende to the presidency in 1970, he proceeded to carry out social democratic reforms that included increasing the minimum wage and expanding education, public housing, and food programs for the poor. More controversial was the government's decision to nationalize Chile's lucrative copper industry, which had been largely under the control of U.S.-based multinational corporations. In 1973, with the assent and support of the United States, the Chilean military overthrew Allende in a coup resulting in thousands of deaths and imprisonments. The military ruled for the next fifteen years.

During Allende's presidency and with the assistance of an American computer expert Stafford Beer, Chile experimented with computer-assisted economic planning. Arguably the first of the cyberneticians to achieve business success, Beer was dubbed by none other than Norbert Wiener himself as "the father of management cybernetics" (Miller 2002, 3). Soon after Allende's election, Beer accepted the invitation of Fernando Flores, an engineer working in the Chilean State Development Corporation, to establish Project Cybersyn (Proyecto Synco in Spanish), a program to build a computer communications network that would help run the Chilean economy. Like the Soviet system, it would process, organize, and display information on economic activity in real time. But unlike the U.S.S.R.'s system, Cybersyn would use the information to enable workers and local managers to participate by providing information and making decisions. Specifically, the project's developers planned to have workers participate in the development of production models, in the design and implementation of technology, and in economic management at the local and national levels (Medina 2011).

In the 1970s the concept of worker democracy was popular as a means of tapping into the tacit knowledge of skilled workers; as one way to combat what was viewed as pervasive workplace alienation, especially among young workers; and as a means of extending participation from the electoral arena into the modern workplace. Experiments in workplace democracy and worker control were taking place at the time in numerous locations, including prominently in the United States, Israel, and in what was then Yugoslavia (Hunnius, Garson, and Case 1973). With worker democracy in the air, experts in the new technology of computer communication thought about how to apply their technical skills to what was becoming a global movement. As Beer said in 1972, "In Chile, I know that I am making the maximum effort towards the devolution of power. The government made their revolution about it; I find it good cybernetics" (Medina 2011, 3). Allende and his government agreed that cybernetics would enable them to build a computer system that would help "to create a new political and technological reality . . . , one that broke with the strategic ambitions of both the United States and the Soviet Union" (ibid., 3).

Limited computer resources and the short life span of the Allende government did not permit implementation of Project Cybersyn, but it remains important in the history of cloud computing for several reasons. It demonstrated that the history of the cloud contains an important chapter from outside the United States, the Soviet Union, and other centers of world power. Audacious as it was, Project Cybersyn was proposed and designed primarily by engineers and planners in what was then called a third-world country—in the minds of some, a backward nation that should have been concentrating on mining copper for transnational corporations instead of experimenting with computer-assisted planning. Moreover, Cybersyn was consciously designed as an alternative to standard models of economic development on offer from the United States and the Soviet Union. Beer sought a balance between centralized and decentralized control, and between the overall needs of a firm and the autonomy of its component parts. His work tapped into a line of thinking that has found its way into discussions of the cloud. How can we create computer systems that bring about efficiencies through centralization without sacrificing local autonomy? Will big data in the cloud facilitate democracy or overwhelm it? Beer's thinking lined up well with the Popular Unity government's interest in promoting national development without sacrificing civil liberties, a free and open media, and individual autonomy. Finally, the proposal for the Chilean version of a computer utility demonstrates the need to consider the social relations of technology in any discussion of cloud computing. For Chile, the Cybersyn network was important because it would advance national development, but also because it would promote public participation in the political and economic life of the nation. Too valuable to be kept under private control, it would serve society as a whole.

It is easy to question whether Allende's government moved too fast to nationalize resource industries and promote workplace democracy with new information technology. Or perhaps it proceeded too slowly, because the government refused to arm supporters under militant attack from U.S.-backed sectors of the society. It is also easy to brand Beer as an eccentric who got in over his head in a place he did not understand. But before doing so, it is worthwhile to compare Chile's ambitious plans to use a new technology to bring about a thorough democratization of society with two examples from the political uses of today's cloud. The first is generally viewed as an unalloyed success because it is widely seen as a major contributor to returning Barack Obama to the White House. I am referring to his campaign's use of cloud computing and big-data analysis provided by Amazon Web Services (AWS), a division of the online retail giant, to identify potential voters and successfully deliver enough of them to the polls to exceed many pundits' expectations. The campaign built more than 200 apps that ran in AWS, making such heavy use that the company's chief technology officer tweeted his personal congratulations to his counterpart in the Obama campaign once victory was certain. The campaign utilized the Amazon cloud in many ways, but the skilled deployment of databases in modeling, analytics, and integration was key. Specifically, "This array of databases allowed campaign workers to target and segment prospective voters, shift marketing resources based on near real-time feedback on the effectiveness of certain ads, and drive a donation system that collected over one billion dollars (making it the 30th largest ecommerce site in the world)" (Cohen 2012). Another key was a set of tools that helped the campaign determine the most efficient television advertising buys (dubbed the Optimizer) and targeted messages to Twitter and Facebook users (called blasters) (Hoover 2012).

There is nothing especially unusual about these and other strategies in the Obama campaign's partnership with Amazon. It appears that the campaign simply made better use of its data-management resources than did the opposition. What is striking, however, is how little this has to do with practicing democracy, with civic participation, or with activism at any level. In place of democracy, including anything envisioned in the Cybersyn project, we have population management and control.

The second example comes from Great Britain, where Prime Minister David Cameron, a big fan of the iPad and especially the game Fruit Ninja, ordered the creation of an app that would enable him and his inner circle to monitor the British economy. Dubbed No. 10 Dashboard, according to the website of the government's cabinet office, it provides a summary view of national and international information, including housing and employment data and stock prices, as well as data on the performance of government departments. In addition, there is "political context" data drawn from polls, commentary, and a sampling from Twitter. Proud of the app, the prime minister showed it off to newly reelected President Obama at a G8 summit meeting.

It would be easy to draw the conclusion that with Obama's use of the largest cloud-computing company and Cameron's No. 10 Dashboard, we are now light years ahead of Chile's Cybersyn. After all, rooms full of 1970s equipment and software can now fit on a handheld device. But on closer inspection, something substantial has also been lost. The fruits of Cybersyn were to be shared with the entire nation in a transparent process of data production, modeling, display, and distribution. The goal was to advance the Chilean national economy even as it promoted democracy in the workplace and in society. Cameron's app, like Obama's use of AWS, is intended to better manage a population. Neither has much to do with public participation in political decision making. Responding to just this type of criticism, the data director of Obama's campaign felt compelled to declare, "I am not Big Brother." He insisted that "campaigns don't know any more about your online behavior than any retailer, news outlet or savvy blogger" (Roeder 2012). Although it is more than a bit disingenuous to compare a campaign organization that spent over $11 million on technology services with the resources of a savvy blogger, it is accurate to compare what both campaigns knew about online and offline behavior with what Walmart, Target, or any other large, global retailer knows (Gallagher 2012). But what kind of defense is it to maintain that a presidential campaign is no worse than a giant retailer like Walmart when it comes to surveillance? Obama's data director may not be Big Brother, but does this justify the conclusion that "new technologies and an abundance of data may rattle the senses, but they are also bringing a fresh appreciation of the value of the individual to American politics" (Roeder 2012)? What would we think if this came from the data director of Target only with "the American economy" replacing "American politics"? The same holds for No. 10 Dashboard. Indeed, as one commentator noted, Cameron's "app could ... be an apt metaphor for politicians reduced to spectators by the surges and shocks of the globalized world" (Wiles 2012). It does not really empower the inner circle of people for whom it was made. In that respect, it is not dissimilar from a special-purpose iPad app made for the team responsible for restructuring Greece's debt. But this conclusion misses a more important point. Politicians who build apps that take a snapshot of the economy may or may not be powerless to do anything. But there is little, if any, consideration for how such data might empower citizens, nor for how citizens might participate in its creation as workers, voters, or customers. That is why it is important to revisit the precursors of cloud computing, like Project Cybersyn, whatever their outcomes. Moreover, we need to do more than marvel at the advance in technology over the decades because history suggests that technological progress does not necessarily bring about advances in the practice of democracy, and sometimes can result in genuine regression.

The Pentagon and the Internet

Although they left behind important legacies and lessons, videotex, Soviet cybernetics, and Project Cybersyn are no longer around. The work of the Defense Advanced Research Projects Agency (DARPA), on the other hand, is not only important for understanding where cloud computing comes from; it is a significant participant in current military cloud-computing projects. When the Soviet Union successfully placed *Sputnik*, the first operational satellite, into orbit around the earth in 1957, it caught the U.S. government by such surprise that President Eisenhower created an agency within the Pentagon whose job it was to keep these surprises from happening again.

Starting in 1958 the agency, then known as ARPA, was responsible for carrying out research and development on projects at the cutting edge of science and technology. While these typically dealt with national security–related matters, the agency never felt bound by military projects alone. One outcome of this view was significant work on general information technology and computer systems, starting with pioneering research on what was called time-sharing. The first computers worked on a one user–one system principle, but because individuals use computers intermittently, this wasted resources. Research on batch processing helped to make computers more efficient because it permitted jobs to queue up over time and thereby shrunk nonusage time. Time-sharing expanded this by enabling multiple users to work on the same system at the same time. DARPA kick-started time-sharing with a grant to fund an MIT-based project that, under the leadership of J. C. R. Licklider, brought together people from Bell Labs, General Electric, and MIT (Waldrop 2002). With time-sharing was born the principle of one system serving multiple users, one of the foundations of cloud computing. The thirty or so companies that sold access to time-sharing computers, including such big names as IBM and General Electric, thrived in the 1960s and 1970s. The primary operating system for time-sharing was Multics (for Multiplexed Information and Computing Service), which was designed to operate as a computer utility modeled after telephone and electrical utilities. Specifically, hardware and software were organized in modules so that the system could grow by adding more of each required resource, such as core memory and disk storage. This model for what we now call scalability would return in a far more sophisticated form with the birth of the cloud-computing concept in the 1990s, and then with the arrival of cloud systems in the next decade. One of the key similarities, albeit at a more primitive level, between time-sharing systems and cloud computing is that they both offer complete operating environments to users. Time-sharing systems typically included several programming-language processors, software packages, bulk printing, and storage for files on- and offline. Users typically rented terminals and paid fees for connect time, for CPU (central processing unit) time, and for disk storage. The growth of the microprocessor and then the personal computer led to the end of time-sharing as a profitable business because these devices increasingly substituted, far more conveniently, for the work performed by companies that sold access to mainframe computers.

DARPA is even better known for the creation of the Advanced Research Projects Agency Network—ARPANET—the first wide-area network using packet-switching technology. Packet switching breaks up data into blocks or packets, which seek the most efficient network routing. The blocks are reassembled at the end point and, unless there are major network problems, appear to an end user as a unified data, voice, or video transmission. The network was created to link secure military installations and major research facilities and became a direct precursor of today's Internet. In fact, some date the birth of the Internet to January 1, 1983, when for one day ARPANET completely shut off service to the 400 hosts the system served in order to replace the NCP protocol with the TCP/IP network protocol that has defined the Internet ever since (Kerner 2013). The growth of the Internet released the brake on cloud computing that the expansion of the first microcomputers and then personal computers had applied. In addition to requiring significant expansion of distribution capacity in wireline, wireless, and switching capabilities, the Internet's accelerating demand for data storage and processing hastened the arrival of cloud systems.

The precursors of cloud computing demonstrate that what we now call the cloud came from various places that used computing for different goals. Videotex systems aimed to link terminals and television receivers to remote computers that, in practice, provided basic information to people in a handful of nations. The Soviet Union applied its leading role in cybernetics to develop a national system of economic planning. Notwithstanding strong fears in the Kennedy administration, including the CIA, that the program would enable the Soviet economy to overtake its competitors in the West, it was at best a partial success. It fell victim to the limited capacity of computer systems and to the power of the Soviet military, which resisted investing technology resources to build the domestic economy. Chile's Cybersyn sought to bring about a social-democratic version of national development planning by connecting central computers to terminals throughout the country, primarily to establish an interactive system of economic decision making. The short-lived rule of the Popular Unity government of Salvador Allende meant that Cybersyn never made it out of the planning phase. Nevertheless, it demonstrated that cloud computing has historical links to the Global South, where democratic values existed side-by-side with technical visions. Finally, DARPA made use of big military budgets during the Cold War to help bring about time-sharing and the Internet. Perhaps most importantly, unlike the Soviet military, which was hostile to civilian-sector participation, DARPA worked with corporations that developed business applications that eventually led to cloud computing. DARPA continues to be very active in the development of a military cloud.

Anatomy of the Cloud

Today's cloud computing deepens and extends key tendencies established by these and other predecessors. The rise of data centers controlled by a handful of companies continues a process of creating global networks of informational capitalism (Schiller 2014). Companies that once contained an IT department, with its craft tradition, can now move to the cloud, where IT and its labor are centralized and streamlined in an industrial mode of production, processing, distribution, and storage. Furthermore, the cloud takes one more step in a long process of building a global culture of knowing in which information production accelerates through networks that connect data centers, devices, organizations, and individuals. The cloud makes up both a new industrial infrastructure and a culture of knowing, based on digital positivism.

It is easy to lose sight of the significance of cloud computing for informational capitalism and for building a culture of knowing because time and time again in the early years of a technology, there is a tendency to concentrate on those flashy utopian or dystopian visions that make up what has been called the technological sublime (Nye 1994). This is understandable. Just as it was hard to resist the feeling of magic the first time a web page scrolled down a home computer screen, so too was it magical when, for the first time, street lights brightened the night with electricity's illumination and voices emanated from the musical box that came to be called radio. Cloud computing currently resides in this magical sublime phase where transcendent visions of ending space, time, and social divisions tend to dilute our appreciation of the more grounded, long-term, but banal consequences of implementing cloud systems. The experience with electricity is especially relevant because its early days were focused on the capacity to bring light and power, an admittedly significant, if not revolutionary, development. But electricity's sublime allure wore off when people got used to universal lighting, especially when the promised end to crime on the streets did not pan out. The sublime became banal. But the genuine revolutionary power of electricity awaited its withdrawal into the woodwork of banality. It was not until electrical generation was organized into utilities and sent out to power industrial and household applications (yesterday's apps) that one could safely conclude that electrification was a principal participant in an economic and social transformation. From powering automobile assembly lines to turning on vacuum cleaners, electricity's many applications were not terribly sublime, but certainly were transformative (Nye 1990). Indeed, some economists argue that electrification, including centralized power generation and near universal distribution, has been the most significant technological force for economic growth in the modern era (Gordon 2000).

Cloud computing is moving from the sublime stage of infinite promises to what may amount to a similar banality. In this respect, the cloud is a gathering of utilities, certainly not the same as the electrical-power generators that enabled a leap in the industrial revolution, but not so different that it is inappropriate to consider a similar process at work. The sublime cloud is entering a banal phase where there is less focus on it as a discrete entity and more on the transformative applications that it is enabling. As one analyst puts it, "In the mid 19th Century, centralised generation allowed electricity to be provided as a utility, meaning that consumers only had to pay for what they used. Consumption could be scaled up or down to meet demand without the need for capital expenditure. A century and a half on, this is precisely the emancipating effect that cloud computing is now having on the enterprise. Organizations no longer need to build, maintain and renew cumbersome IT infrastructure in order to consume as much, or as little computing resource as they need" (John 2013).

Cloud computing builds on its predecessors, but there are sufficiently significant differences that mark its departure from earlier models. It is useful to consider some of these differences, beginning with the extraordinary growth in the sheer size and scale of cloud facilities. It is no overstatement to argue that cloud centers require a major stretch in our conceptual vision to begin to understand their enormity. Consider the plans for the largest data center (in cost, size, and processing power) now under construction. In September 2012 China's major social-networking firm Baidu, a Chinese version of Google and Facebook combined, announced that it would spend $1.6 billion to build a new cloud center in Yangquan, Shanxi Province, covering 120,000 square meters (about thirty acres), roughly the size of the U.S. Pentagon, one of the largest standalone buildings in the world. The Yangquan facility will contain the capacity to store 4,000 petabytes (PB) of data (1 PB equals 1 million gigabytes; see the following table). When completed in 2016, it will deploy 700,000 central processing units. Drawing comparisons and making estimates is always perilous, but it has been estimated that digitizing the entire collection of print, audio, and video stored in the collection of the Library of Congress would amount to roughly 15 terabytes of data. The storage capacity of the Baidu cloud center would therefore enable it to house the data equivalent of 268,000 Libraries of Congress.

FROM MEGABYTES TO ZETTABYTES

1,000 megabytes = 1 gigabyte (GB)
1,000 GB = 1 Terabyte (TB)
1,000 TB = 1 Petabyte (PB)
1,000 PB = 1 Exabyte (EB)
1,000 EB = 1 Zettabyte

When it opens in 2016, the Baidu center will set a new standard for data facilities, but those operating now are far from small. As of December 2013, the largest existing cloud data center was a 400,000-square-foot structure, part of a 2.2-million-square-foot interconnected collection of data centers operated by the Switch corporation in Las Vegas, where the absence of natural disasters provides a margin of safety. Admittedly, data centers of this size are at the outer edge of typical cloud data centers, but the trend is to build ever-larger ones because size provides efficiencies that are needed as data storage and processing demand continues to grow. In fact, China, in a joint venture with IBM, is in the process of building its own "cloud city" in Langfang, an old industrial district near Beijing, that will cover over 6 million square feet of facilities, including a giant data center and offices to house IT development companies (Zhu 2013).

The corporation Cisco, a major participant in the cloud industry, has put together an index of global data-center traffic. These are estimates, but they provide a general sense of the growth in the sheer amount of data in the cloud, and once again require a stretch of the imagination. Cisco estimates that by the end of 2017, 69 percent of all Internet protocol (IP) traffic will be processed in the cloud as opposed to in facilities operated by a specific organization, like a corporation or government unit, or by individual consumers. Annual global cloud IP traffic is forecast to reach 5.3 zettabytes (ZB) (a single zettabyte is equal to one billion terabytes or, in more concrete terms, 250 billion standard DVDs or 36 million years of HD video) by the end of 2017. Global cloud traffic is expected to grow sixfold by that year (Cisco 2013). This has led some to worry about a cloud "plumbing problem" because the amount of data stored is growing much faster than the bandwidth of network connections needed to process and analyze data (Wegener 2013).

Statistics on the industry are not easily obtained because cloud data centers are either under private control or operated by governments not inclined to share information. Estimates vary, but one census produced a total of 509,000 data centers worldwide at the end of 2011, occupying close to 300 million square feet. Cloud centers are located everywhere in the world but tend to be concentrated in places where land is plentiful but not far from communication and power facilities. This includes what was once agriculture land on the outskirts of population centers, where companies can benefit from low labor costs. These considerations have led Apple to locate its cloud data centers in rural North Carolina and in Oregon. The North Carolina location is especially interesting for both Apple and Google because low labor costs are matched with low energy costs—30 percent lower in North Carolina than the national average. Moreover, North Carolina possesses an increasingly valuable commodity that one would not naturally associate with cloud computing: pig manure, or, as it is referred to more euphemistically, black gold. The state holds 14 percent of the swine population in the United States and pig manure can produce methane gas energy to help meet the massive power-consumption needs of data centers. Apple and Google are not only competing for clicks and customers; they are in a race to determine who can best exploit this unlikely North Carolina resource (Wolonick 2012).

Security is a growing concern, especially as the size and therefore the value of facilities and data have grown. This has led some cloud companies to locate their facilities in mountainous regions that, while quite far from urban areas, offer added protection. Increasingly, the propensity for earthquakes and severe climate events is taken into account in the choice of location. Energy costs for a 24/7 operation are a key consideration and this is leading some cloud companies to explore the novel solution of burying facilities inside mountains close to supplies of cool water to lessen the requirement for air conditioning. For example, Norway's Green Mountain Data Centre is located on the shores of the island of Rennesøy, close to a large fjord. The center itself is contained in concrete buildings within caves built into the mountain. Racks of servers fill halls once used to store ammunition for NATO forces, but what makes the location especially attractive is proximity to a fjord that provides a constant supply of cool water to keep sensitive systems from overheating. Locations like Rennesøy provide both enhanced security and lower energy costs.

It is interesting to observe, and not a little bit ironic, that a technology promising freedom from locational constraints is itself constrained by the need to maximize the ability to house enormous amounts of data and guarantee system reliability. Companies increasingly aim for the sweet spot: cold climate, access to low-cost power, abundant water supply, high-bandwidth Internet connections, political stability, and financial incentives. Several countries meet the requirements, but none more so than Canada, which is increasingly a data-center destination of choice (Perkins 2013). Facilities in Canada take advantage of a technology known as "free cooling" that reduces energy requirements by about half through the use of a cooling circuit that draws on outdoor air to supplement a data center's energy-intensive needs. A specialized heat exchanger uses outdoor air to cool water and glycol that circulate to the server racks, thereby reducing the load on compressors and pumps, which are the big energy hogs in data centers. IBM opened a $90 million data center in a small Ontario community partly because the company can cool the facility for 210 days a year without running energy-consuming chillers. While exotic locations like mountains and fjords attract attention, Canada works for many companies because practically the entire country is in a cold climate, which means there are numerous locations near power and water supplies and close to large cities. According to the head of one IT research company, "The advantage Canada has is it's far cheaper and easier to bring data to power sources, and vice versa. It's much cheaper to stick your data next to a hydro dam" (Stoller 2012). The town of Barrie, Ontario, which houses the aforementioned IBM facility as well as facilities of major banks, has abundant, reliable, and inexpensive supplies of water and power, and benefits from proximity to Toronto, which provides it with excellent Internet connections. Canadian cloud-data-center companies have also pioneered the use of energy-saving systems. OVH.com, a Quebec-based company, uses a unique heat-dissipation and cooling system that has completely eliminated the need for air-conditioning servers in its Canadian locations, and reduced it by 98 percent in its worldwide locations.

Canada, like the Scandinavian nations with which it vies for data-center business, also benefits from political stability and strong data security. Additionally, Canada benefits from proximity to the United States and the additional incentive that data located in Canada is not subject to the USA PATRIOT Act, which permits the U.S. government to intercept and analyze data stored within its borders without a search warrant. In addition to Canadian and Scandinavian locations, Switzerland, with its long-standing political neutrality, is an increasingly favored choice for data centers, but it is expensive. All the discussion of size and proximity to resources makes clear that cloud computing is a very material industry with locational requirements that belie the image of an ephemeral cloud. Cloud-computing data centers are the communication version of those industrial transportation hubs of the past where, for example, the city of Chicago played a large role in America's industrial expansion. It should not be surprising that, until recently, the largest cloud data center in the world was located in that city. Of course, data centers are not rail yards, but just as transportation centers were key nodes in the global industrial grid, cloud data centers are material hubs for global information and communication traffic. Images of invisible data moving through clouds help convey a sense of what the sociologist Zygmunt Bauman (2000) describes as our era's "liquid modernity." Today's iconic products are data, information, and messages, which flow around the world through thin wires or just through the air. But they are rooted in physical structures that make significant material demands on resources and that call to mind the factories of an earlier era. Understanding cloud computing absolutely requires an appreciation of its materiality, of its substantial physicality and its extraordinary demands on the environment.

There are many other ways to describe this dance of petabytes and zettabytes, and we will certainly explore some of these, but suffice it to say that nothing in the history of communication and information processing approximates in scale the levels of storage, processing, and distribution that the cloud makes possible. With that said, it is important to give attention to something missing from cloud computing, but in order to do so we need to address more of its characteristics.

On-demand self-service. Cloud computing allows users to choose their storage requirements and server time automatically without requiring human interaction with the provider of each service.

Broad network access. Users can access the cloud in standardized ways through any platform, such as a tablet, smartphone, or personal computer.

Resource pooling. Resources like storage, processing, memory, bandwidth, network, and virtual machines can be brought together by the provider to serve multiple users with different physical and virtual resources assigned and rapidly redeployed to meet user demand. This enables the provider to engage users without regard to location, unless users demand that the provider specify a location by nation, region, or data center. For example, users in the United States may not want to be served by a data center in China whose "Great Firewall" of censorship limits access to the online world. Or, fearing the application of the PATRIOT Act, users in Europe or Canada may not want to be served from the United States.

Rapid elasticity. Cloud resources can be expanded and contracted quickly based on customer needs. Users are not locked into IT investments, but can make use of just what they need. However, it also means that they must rely on a provider that is typically not as familiar as an internal IT department with the history and culture of the organization. Since moving to the cloud increases the likelihood that an organization will shrink its IT department, that leaves the organization with less inside technical expertise or tacit knowledge to help determine its information-technology requirements.

Measured service. Cloud companies can provide and control services efficiently by employing a measurement based on one or more specific services, such as amount of data stored, bandwidth used, or quantity of processing. If the provider is using a metric that reasonably reflects the service provided, then there is transparency for both provider and user.

Types of Cloud Computing

In addition to these characteristics of cloud services, one can identify three different types of cloud service models that focus on infrastructure, platform, and software, with each model providing the customer with different levels of control. These are not inviolate categories but, in spite of large gray areas at the margins, they are nevertheless useful in adding a sense of what different types of cloud models aim to accomplish, from simply providing a storage service to offering additional applications and software for customers to use.

IaaS: Infrastructure as a Service. With this model, the cloud-service provider manages a storage infrastructure for customer data, leaving the customer to deploy its own software, including operating systems and applications. Furthermore, under this model customers can control certain network components, like firewalls. It is ideal for repetitive uses that require elasticity or the capacity to expand or contract quickly depending on use. Examples include online gaming sites, online advertising networks, video-sharing sites, and social-media applications.

PaaS: Platform as a Service. Here, in addition to offering storage facilities, the cloud provider deploys onto the cloud infrastructure applications that the customer has created or acquired using programming languages and tools that the provider supports. Once again, the customer does not manage the infrastructure; all of that is left to the cloud provider. Rather, the customer gets to control the deployed applications. For example, the city of Edmonton, Alberta, contracted with a cloud provider to create its own tool, called Open Data Catalogue, which made information about city services accessible to the public. The U.S. Department of Defense (DOD) used a cloud provider when it needed to emulate battlefield conditions. In the latter case, DOD technical staff developed an application on the Microsoft Azure platform.

SaaS: Software as a Service. Under this model, cloud companies offer their own applications for customers to use on the cloud infrastructure. Customers typically access these applications through what is called a *thin client interface*, such as a web browser that might provide the customer with document processing or web-based email. The customer leaves to the provider control over such infrastructure items as operating systems, networks, server storage, and application capabilities. For example, instead of buying a copy of Microsoft Word, a customer rents the use of the word-processing software for a fixed charge per month or pays a per-use fee. To use the software, the customer logs into the cloud company's system. Similarly, a small business might rent a sophisticated sales database from a cloud company like Salesforce because it would not make economic sense to purchase such a database. Success depends on the quality of the rented software and the reliability of the cloud provider, especially when the software involves multiple tools responsible for running several different units of a business (sales, accounting, administration, and so on). A primary benefit of SaaS is that it minimizes or entirely eliminates the requirement for in-house IT professionals. Companies selling software through the cloud gain from a regular flow of income, especially when they are able to shift popular software from a purchase to a monthly subscription model, as Adobe did with its popular Photoshop (Pogue 2013).

It is also important to distinguish among different models of deploying cloud systems, including private, public, hybrid, and community clouds.

Private cloud. Under this model, the cloud is customized and deployed for a single organization. It may exist on or off the organization's premises, but when it is off premises, the private cloud is protected by the organization's firewall. Private clouds tend to be chosen by organizations, like banks, that have security and regulatory concerns prohibiting them from using cloud services that are widely available to the general public. In essence, the private cloud is a gated community set aside for those willing to pay for an extra degree of security. In this respect, it is a manifestation of a trend, troubling to some, that would dissolve the Internet into a set of privately run networks (Moses 2012). This model is tempting because private clouds can serve as vaults to secure data from snooping eyes, the essence of the business model of companies like Reputation.com (Singer 2012).

Public cloud. This model is typically provided by large cloud-services businesses, such as Amazon Web Services, and offers software, platforms, and infrastructure to the general public or to an industry association. In essence the public cloud is available to anyone who can pay for it and it is expected to grow five times faster than overall IT-industry expansion through 2016 (Lee 2013b). Public cloud SaaS configurations are the most widely known because they include familiar services like Google's Gmail, Apple's iCloud, and the marketing services provided by Salesforce. Organizations that require the greater control but prefer to stay in the public cloud might opt for a PaaS system like Microsoft's Azure or Google's App Engine. Those needing still more control turn to IaaS public-cloud services like those provided by Amazon and Terremark.

Hybrid cloud. When the cloud infrastructure is composed of both public and private clouds that remain unique entities but are linked by technology that allows for data and application portability, we refer to a hybrid cloud service. Many organizations have divided requirements that might lead them to seek out the public cloud for most of their needs and a private cloud configuration to maintain the security of sensitive data. Hybrid-cloud providers who share ownership and management with their customer organizations enable them to enjoy the benefits of both types of deployments. While hybrid clouds appear to be an excellent choice because they can be all things to their customers, they also require careful management to balance the component cloud formations. The company Rackspace has become a leader in the hybrid model.

Community cloud. This model brings together several organizations that have common interests, such as a similar organizational mission, similar set of regulatory requirements, security needs, compliance expectations, or policies. One or more of these organizations might manage the cloud or, what is more frequently the case, they may together hire a third party who runs the cloud in the data center of one of the organizations or houses it off-site. For example, a group of airlines might build a community cloud to house a common reservation system. Community clouds are chosen because they can be customized to meet the specific needs of an organizational group, such as a collection of media firms interested in sharing file-based digital media content. Community clouds are also interesting because they have kept alive the early cloud-computing discussion of building systems that are not primarily under vendor control and operate in a more environmentally sustainable fashion (Briscoe and Marinos 2009).

What's Missing?

Although the words *public* and *community* are used in cloud computing, every cloud model is presumed to be a private service operated by a business with the goal of maximizing profit. Government systems, which often use private provisioning (even the CIA will be using Amazon Web Services for $600 million worth of cloud projects), are primarily employed for management, control, and surveillance (Babcock 2013a). In the context of cloud computing, "public" simply means that vendors will sell to the general public rather than to a single preferred customer, and "community" refers to the common commercial interests shared by users of that cloud model—for instance, they are all airlines. These are very narrow uses, if not outright distortions, of the terms public and community. The public traditionally refers to citizens who participate in the decisions that affect their lives, and a community is a collection of active citizens with common interests. The history of computing has included extensive debates about public and community participation in the construction of networks and in the provision of services. Unless a cloud system is specifically set up to provide information to a public or to a community of citizens, then the vast majority of people do not participate in the cloud as citizens, but rather as consumers who are valued not for their participation in decision making about the cloud, but rather for their propensity to purchase services and to provide information to companies about their consumption patterns.

In addition to being an extraordinary leap in processing and storage power over early cloud-like systems, cloud computing, unlike computer systems that preceded it, is a singularly market-driven project with little consideration of alternatives to the model of management and control that governs it. Where are the debates about using the cloud to expand economic or political democracy? How about worker participation in corporate decision making or greater citizen participation in national or community life? What about public participation in decisions about cloud data centers or cloud systems? Unlike earlier communication systems that, whatever the outcome, sparked vigorous encounters about their potential to expand citizenship and democracy, the cloud is essentially silent on these issues. There appears to be an enormous gap between the prodigious sublimity of the cloud's power to process, store, and distribute information and the banality of its current applications, however practical and profitable.

While almost all cloud systems operate according to a commercial model, there are a few exceptions. For example, grid computing is a means of creating a cloud from below by harnessing the combined power of millions of personal computers to carry out projects. But even these are typically organized by commercial enterprises. Since 2004, IBM has sponsored the World Community Grid, which takes the principle of using the available space on mainly household PCs to address a variety of public-health and environmental research projects. Specifically, it makes this combined computer power available to public and not-for-profit organizations for use in humanitarian research. All results are in the public domain open to the global research community. Research projects cover clean water and energy; the development of drugs to combat malaria and dengue fever; as well as research on muscular dystrophy, cancer in the young, and AIDS. For example, advanced computational methods help to identify candidate drugs that have the right shape and chemical characteristics to block HIV development. Commercial projects are beginning to take advantage of this distributed processing model, including harnessing idle PCs in homes (Novet 2013). These open a door to an alternative form of large-scale computing that does not require a top-down cloud computer model.

Cloud computing therefore distinguishes itself from earlier models in two fundamental ways. First, receiving the most attention is the capacity to store, process, and distribute data beyond anything that preceded it. What were once the exceptional "supercomputers" are now standard in the half million or so data centers worldwide. Second, even as it has exceeded its predecessors, cloud computing operates from a diminished vision that is almost entirely driven by the twin goals of profit and control. There is little interest in using the cloud to bring democracy, citizen-driven design and implementation, worker control, or even worker involvement in decision making. While all of these ideas have been raised in the course of computing's history, they are not part of debates about today's or tomorrow's cloud computing.

Is the Cloud a Utility?

This may change before long because there is a contradiction at the center of cloud computing that will likely heat up the debate. Put simply, some forecasts for computing are coming to fruition and the cloud is taking on more of the characteristics of a genuine utility (Clark 2012a). It is not just the academic and policy communities that are beginning to think of today's IT environment in public-utility language. When asked what his two companies, Twitter and the e-payment firm Square, have in common, Jack Dorsey answers, "They're both utilities." Moreover, Facebook head Mark Zuckerberg has spent years referring to his company not as a social network, but as a social utility. However, when asked if his utility should be regulated, the Facebook founder backed off: "Something that's cool can fade. But something that's useful won't. That's what I meant by *utility*." Of course, most of the cool things we think of that will last are not referred to as utilities. But whatever the definition or the reaction, the concept of the utility is increasingly part of the ongoing debate about the developing structure of the computer universe (Fox 2013). Cloud computing has made it a more frequently used concept.

Taking into consideration the experience of earlier utilities, such as water, gas, and electricity, one energy expert defines the requirements of a utility market as comprising the following:

- A source of energy generation
- A transportation network
- A transmission and distribution capability
- A metering capability
- A pricing mechanism
- A regulator to ensure adherence to rules
- A customer (James Constant, cited in Clark 2012a)

This configuration of characteristics can be debated, but most would agree that they are among the major ones defining a utility. According to Clark, cloud computing meets most of these criteria. It is a source of energy generation in its ability to compute and store data. The Internet and the telecommunications systems connected to it form the transportation network. Data centers handle the transmission and distribution capability because they house the storage and processing capabilities. Cloud services, especially "public" ones, can meter precisely how much storage and processing are being used at any particular time, albeit with different providers applying different pricing methods. Pricing is determined by the cost to receive, process, and respond to a request for storage, processing, and distribution. Although a wide range of factors is involved, cloud providers directly control the costs of hardware and software because they engineer their own systems, and the costs of other factors, such as facilities, staffing, and electricity, depend on the particular market within which they operate. Finally, there is no shortage of customers. Indeed the market most likely will grow to one in which a handful of providers serve practically everyone, just as, for example, water and energy markets do. As Clark concludes, "All that's lacking is a regulator. Whether the cloud computing industry should be regulated is a complex issue that will undoubtedly become a major debate before too long" (Clark 2012a; M. O'Connor 2013).

One can debate whether a government regulator will ever become essential to the cloud computing industry. What cannot be debated is whether cloud computing will be subject to governance; that is, to the need for general management, coordination, and oversight. This can be accomplished by agencies of government, as has been the case historically for most utilities, or it can be accomplished by those with market power. These are both governance structures, notwithstanding the mythology of the market's "invisible hand." The myth tends to focus on the magic of invisible coordination more than on the hand, which, in reality, is quite visible. It is apparent to most observers of cloud computing that, however they might feel about government regulation, there is a growing concentration of power among a handful of cloud providers, most of which are also key players in the production and distribution of software and content. Utility markets often become government regulated because one or a few producers, who use their position to exercise significant power over services and pricing, come to dominate.[1] Historically, this has been the case throughout the history of communication media that was marked in the United States, for example, by Western Union's control over telegraphy, AT&T's over telephony, and the broadcasting networks' domination over radio and then television. In each case regulation was called on to temper the threats of monopoly or oligopoly control. This pattern was followed in other nations, but some of these also turned to public ownership to guarantee widespread, if not universal, access to an essential service. So it is not very surprising that along with the term *utility*, the concept of regulation has entered public debate in today's computer and social-media world (Marshall 2013). The demise of specific cloud services, such as Google Reader, because companies cannot recover fixed costs from their provision, has led some economists to wonder whether government ownership or regulation through public utility status is inevitable for essential but unprofitable services like search (Kaminska 2013). Even as businesses in the less developed world begin to embrace the cloud, they fear that it might do more harm than good without the stability provided by government regulation (Hanna 2013).

One of the key reasons why expert attention is returning to the concept of the utility in cloud computing is that the industry is rapidly becoming dominated by a handful of companies. The power of Amazon, Apple, Google, Facebook, and Microsoft is troubling enough to lead some to doubt that the "invisible hand" will prove adequate to restrain their ability to dominate cloud markets (McKendrick 2013b). Consequently, they maintain, we should begin to think about broader national or even international oversight by elected representatives. As one concerned analyst put it, "The Internet has taken the place of the telephone as the world's basic, general-purpose, two-way communication medium. All Americans need high-speed access, just as they need clean water, clean air, and electricity. But they have allowed a naive belief in the power and beneficence of the free market to cloud their vision. As things stand, the U.S. has the worst of both worlds: no competition and no regulation" (Crawford 2012). According to Crawford, when it comes to the Internet, the United States should follow the historical example of other utilities. When, for example, electricity came under the control of a handful of firms that provided service only to those who paid top dollar, public pressure led to the creation of regulated utilities and public corporations. Opponents of this view argue that the Internet and the cloud are fundamentally different from roads, water, and electricity and that government regulation would stifle the incentive to risk-taking innovation. In 2013, the divide between cloud computing and electrical utilities blurred when research found that a growing number of cloud companies were making significant profits by reselling electricity to customers in addition to providing space to house data. This practice, what has been dubbed the creation of "wildcat electrical utilities," has led to more calls for government regulation of the cloud (Glanz 2013).

In addition to these concerns, there is the issue of data preservation. Absent some form of regulation or mutual agreement within the IT industry, and specifically among those who are major cloud-services providers, there is no requirement to preserve the photos, email, videos, postings, data, and files that individuals and organizations believe are securely stored in data centers around the world. As a result, much of the digital evidence from the daily lives of individuals and the decisions and activities of organizations will vaporize, irrespective of how many cloud data centers fill the world. As one concerned tech writer argued, "We're really good at making things faster, smaller, and cheaper. And every step along the way makes for great headlines. But we're not nearly so good at migrating our digital stuff from one generation of tech to the next. And we're horrible at coming up with business models that assure its longevity and continuity" (Udell 2012). Another person who has been active in the online world for years, hosting numerous sites and archives, worried, "Not to be dramatic or anything, but no more than forty days after I die, and probably much sooner, all the content I am hosting will disappear" (Winer, quoted in ibid.). To date, the only reason most of this material has been preserved is due to the heroic efforts of individuals who personally port archives when technology and standards change. Referring to several archives dating from the turn of this century, Udell commented in a *Wired* column, "If I hadn't migrated them, they'd already be gone. Not because somebody died, it's just that businesses turned over or lost interest and the bits fell off the web. Getting published, it turns out, is a lousy way to stay published. With all due respect to wired.com, I'll be amazed if this column survives to 2022 without my intervention" (ibid.). There are some efforts, primarily by governments, to archive and preserve files. The most notable of these may be at the U.S. Library of Congress, which, among other things, is archiving the massive database of Twitter postings. These are all important activities, but they are isolated and much more data disappears than is preserved. Of course, one can argue, there is a great deal of digital content that is not worth paying to preserve. Society has survived in the past without carrying forward from generation to generation the entire weight of the historical record. Nevertheless, since most of that record is now digital, is it not worthwhile to develop strategies to preserve at least some of it in a systematic fashion?

Now it is important to turn to an overview of major participants in the cloud marketplace, starting with the five companies generally considered dominant on the Internet and in the cloud.

Mapping the Cloud Industry: Leaders and Challengers

Arguably the leading force in the U.S. cloud computing industry, and a global giant as well, Amazon began by applying computer power to transform publishing and then the general retail industry. As prodigious as this accomplishment has been, one commentator concluded that these achievements "may be footnotes to the company's larger and more secretive goal: giving anyone on the planet access to an almost unimaginable amount of computing power" (Hardy 2012a). By 2013, according to most accounts, its subdivision, Amazon Web Services, was the leader in U.S. cloud computing. As an analyst for the consulting firm Forrester described it, "Almost every major consultancy supports Amazon; almost every advertising agency runs on Amazon; if I need to hire 10 people tomorrow to help me build my application, it's super easy to find people who have Amazon experience" (Miller and Hardy 2013). While Amazon does not break out revenue for cloud computing, 2012 estimates range from $800 million to as much as $2.4 billion (ibid.; Mims 2013). The company operates its cloud services through the aforementioned AWS, which achieved widespread public attention in 2012 because the Obama presidential campaign used AWS to organize its successful voter analysis and voter-turnout drive. By the middle of 2013, one typically modest industry observer concluded, following the company's thirty-seventh cut of its cloud prices, which sent tremors through the industry, "The proof is in: Amazon fully controls the cloud" (Linthicum 2013c). This conclusion may be premature and a tad overstated, but it does correctly identify Amazon as an increasingly dominant force in the cloud business.

AWS was created in 2004 with about forty employees, and was the first company to rent its data storage and computing power to other companies. Although it is highly secretive about most of its operation, by 2012 Amazon was regularly listing more than 600 job openings on the company's website. It operates several large data centers in the United States, each of which contains multiple buildings with thousands of servers. It also runs data centers outside the United States and has several under construction. AWS is not the largest cloud provider in the United States by quantitative measures such as size of data centers or total number of servers, but it is arguably the most powerful because it is part of the Amazon corporate empire and the relationship marks one of the few times when the often-used buzzword "synergy" is an understatement. AWS benefits from the sheer size of its parent's computing power. For example, while the parent Amazon does not reveal the size of its operations, an executive who knows Amazon well maintains that just one of the company's data facilities in the eastern half of the United States contains more servers dedicated to cloud computing than does the entire operation of one of the major hybrid-cloud companies, Rackspace, which in 2013 served 200,000 clients, mainly business customers, with about 100,000 servers in nine data centers. AWS also benefits from the data that Amazon gathers on its millions of customers whose purchases of books, homeware, clothing, and so on provide information that AWS uses to forecast consumer behavior, a boost for both the parent and the firms that purchase AWS's services. Among its major customers are popular media firms like Netflix, Pinterest, Shazam, and Spotify. Amazon has been so successful in the cloud that company management expects it to become the leading revenue producer for Amazon, topping even its renowned retail division, with sustained growth estimated at 45 percent per year through 2017 (Finkle 2012).

Market power gives Amazon considerable leverage over its competition, large and small. As the head of AWS put it when asked about a stepped-up challenge from Google, "We've always been very good at making everything as low-cost as possible, then we lower it some more" (Miller and Hardy 2013). The company is able to price its services, particularly the storage and data-analysis capacity of its servers, so inexpensively that neither many established nor start-up companies any longer bother investing in their own. Instagram, for example, the highly successful web photo company, which is now a part of Facebook, did not bother investing in its own computers. The start-up Cue, which admits to spending $100,000 a month on AWS services, uses them to scan millions of emails, Facebook postings, and corporate records to provide enhanced data that subscribers can use in all of their online activity. Over 185 federal government agencies also run some part of their services through AWS and Amazon has won a $600 million contract to provide cloud services for the CIA (Babcock 2013a). The company is active internationally; in addition to having data centers located in Asia, Europe, and Latin America, it hosts numerous corporate and government clients outside the United States. For example, a German company used AWS to make digital copies of 20,000 television shows, a job that cost the firm less than it would have spent on the electricity alone if it had done the work in house. AWS servers located in California and Ireland provide people in Africa with the ability to comparison-shop cars using smartphones connected to AWS. There is no gainsaying Amazon's rich database of customer searches and purchases, which adds value to AWS's offerings. As one customer commented, "You can now test a product against millions of users for just a few thousand dollars, or start a company with just one or two people" (Hardy 2012a).

To multiply these success stories, Amazon has to successfully deal with two major challenges: providing continuous reliable service and fending off the competition. AWS has been a generally reliable cloud provider, but a handful of notable outages have damaged the company's reputation. One of the most significant took place over the Christmas holidays in 2012, when Netflix customers lost access for the better part of Christmas Eve and Amazon itself lost service for its own customers on Christmas Day. In 2013 Netflix relied on Amazon for 95 percent of its data-center needs and, in the highly competitive video-streaming marketplace, the company cannot tolerate significant downtime. As one independent analyst concluded, "Netflix and other organizations which rely on AWS will have to reexamine how they configure their services and allocate their service requirements across multiple providers to mitigate over-dependency and risks" (Finkle 2012). Amazon is not alone in experiencing outages. They affect the entire industry, are primarily caused by power problems, and, on average, last for 7.5 hours (Talbot 2013). They also lead to unanticipated consequences and hidden costs (Franck 2013).

Reliability also requires guarantees of security, another problem for public cloud companies, and Amazon is no exception. In 2013, a single security researcher managed to uncover 126 billion files that were left open to the public. From a sample of 40,000 files, he found exposed data belonging to a medium-sized social-media service, the sales records of a car dealership, employee spread sheets, and video game source code from a mobile-games developer. The shockingly exposed files also included unsecured passwords. Amazon took measures to secure the data and warn customers, but this one event left its clients understandably worried that public-cloud data was far more exposed than anyone thought (Brian 2013).

Amazon also needs to overcome competitive pressures, especially from a handful of companies that can also leverage their leadership in new-media hardware, software, and media services. Some of these, like Microsoft, IBM, and Oracle, have more experience than Amazon in the market for large corporate clients. One way for AWS to succeed is by heavily discounting cloud services, then, once the competition is driven out of the market, raising prices once more, a tactic that proved successful in Amazon's retail book-selling operation (Streitfeld 2013). It is not an exaggeration to say that even in this early stage of development, the struggle for competitive dominance in cloud computing, just as across the Internet, is narrowing to a handful of corporations that can marshal a similar degree of leverage (McChesney 2013). These include familiar names: Apple, Google, and Microsoft. Of these leaders, Microsoft is probably the most committed to providing general cloud services, especially to businesses, which have helped the company maintain its elevated position even as the others have successfully challenged its consumer-services market. Businesses and government agencies have long been committed to Microsoft software and the company now aims to move these and new customers from reliance on physical programs to online services for a fee. So far it has been reasonably successful, with over 100,000 businesses using the company's cloud services. It is important to emphasize this point because much of the day-to-day attention in the popular press goes to the others, primarily because Google is the major gateway to search, Apple to music, and Facebook to social media. Even Twitter, a much smaller company, garners more notice than Microsoft. But the company Bill Gates started in 1975 has a very strong foundation in business software and, with software migrating to the cloud, Microsoft has invested heavily in cloud platforms. Over the last few years, the company has quietly built up its Server and Tools division and it now generates $18 billion a year in revenues, with six of its subdivisions topping the $1 billion mark.

Microsoft is counting on the cloud platform offering its Azure service to enable customers to develop applications and otherwise make profitable use of their own information. Azure provides both Platform and

Infrastructure as a Service and once again demonstrates the value of proximity to services and systems within a large company like Microsoft, which developed Azure by using some of the elements of its successful web browser Bing (Wilhelm 2012). In recent years Microsoft has not been as successful in consumer services, but it is also making a big push to take individuals and families, as its advertising slogan repeats, "to the cloud." These include Windows Live, a suite of cloud services that includes file storage, image, video, email, messaging, the Bing search engine (now the second most popular in the United States), and Xbox Live. Finally, Microsoft expects that the cloud version of its very popular suite of word-processing, spreadsheet, and related programs will succeed in the cloud, as what it calls Office 365 begins to deliver them on a subscription basis.

Google's concentration on consumer services pioneered in its search engine has led the company to focus on that side of the cloud market. It has expanded the company's consumer cloud beyond search with document storage (Google Drive), word processing (Google Docs), and entertainment (Google Music) applications. Furthermore, however much it worries tech observers, the company also sells its own devices that are entirely dependent on the cloud for data storage and applications (Gilmoor 2013). These include the familiar Chromebooks as well as Google Glass, which Google hopes to use to sell pay-per-gaze, for which it holds a patent, to advertisers (Bilton and Miller 2013). But with competitive threats from AWS and Microsoft, Google has begun a major push into the business market with Google Compute Engine (GCE), its IaaS unit. Again, as with Amazon, built-in leverage matters a great deal. In this case, Google runs its IaaS on the same technology that powers Google search, which leads the company to claim greater reliability than AWS, especially because of the notable outages the latter experienced in 2012 (Chen 2012). In 2013, Google tied GCE to the Google App Engine and its global network of app developers in the hope of beating the competition by providing customers with a cloud service that includes privileged access to the largest set of apps in cyberspace (Hardy 2013d). This is why Google is not reluctant to boast: "For the most part, GCE is positioned as a way for customers to benefit from years and years of infrastructure investments, which span everything from our datacenter design to our operational practices, our hardware design and software design, [and] includes the software stack"

(Clark 2012b). Reassurances aside, breakdowns lead users to worry that they are not keeping a close enough eye on their own data. Indeed, one of the key challenges for companies like Amazon and Google is to balance the costs of meeting worried companies' demands for geographical proximity to data, even as they make use of a global network of data centers to ensure sufficient network redundancy to support their claims of protection against outages.

Like its rival giants in the industry, Google is comfortable moving into new territory, in this case the business applications market, long dominated by Microsoft. Indeed Google has been so committed to innovative product development that it has been dubbed the General Electric of the twenty-first century (Gapper 2013b). For years, Google Apps was pitched mainly to small firms and start-ups because Microsoft dominated the market for large businesses. But Google has begun to cut into this lucrative segment with major private-sector clients like the pharmaceutical giant Hoffmann-La Roche, where 80,000 employees use the package, and public-sector clients such as the U.S. Department of the Interior, where 90,000 use Google Apps as their staple business-productivity software. Borrowing a page from Amazon's playbook, Google relies on consistent low pricing that Microsoft has difficulty matching (Hardy 2012b). Microsoft fights back, but does not appear to take Google very seriously as a contender in this market. Some might consider this a mistake, but Microsoft is clear that Google is not a threat in the business cloud market because, according to the general manager of Microsoft's business division, Google "has not yet shown they are truly serious" about enterprise applications. "From the outside, they are an advertising company" (Kerr 2012). There is some substance to this view. After all, in 2011 only 4 percent of Google revenue came from its business services, whereas 96 percent came from advertising. Microsoft's cloud-based Office 365 is intended to keep Google's business market share from growing, but Microsoft has yet to demonstrate widespread uptake of the service because businesses, worried about security and outages, still prefer Microsoft's more familiar Office software (ibid.). Early in 2013 Google accelerated a push to challenge Amazon and Microsoft in cloud services. It doubled the size of its office space in the Seattle area, near the headquarters of both rivals, and began large-scale hiring of cloud-computing experts. In addition to opening another in the many revenue streams that Google enjoys, the company expects it will have the multiplier effect of luring app developers and other companies to use Google products and to launch from the Google platform.

It is hard to contend with the view that Apple has succeeded in creating a successful consumer cloud. With iCloud and iTunes Match, Apple has the largest share of the consumer cloud-services market in the United States, substantially ahead of Dropbox, Amazon Cloud Drive, and Google Drive. Moreover, the sheer size of Apple's data centers in the United States (its North Carolina facility alone is one of the largest in the world) and its seemingly constant process of expansion demonstrate the company's continuing popularity. So do the sales of its line of computers, tablets, and smartphones (Fingas 2013). Much of this success can be traced to the vision of its founder Steve Jobs, who recognized the importance of the cloud in 2008 and committed to it in 2011 when, although ill with the cancer that would soon take his life, he announced to a Worldwide Developers Conference the company's "next big insight": "We are going to demote the PC and the Mac to be just a device and we are going to move the digital hub into the cloud" (Isaacson 2011, 533). While Google, Facebook, and Twitter garner attention as media disrupters, Apple has become one of the world's largest media companies by creating cloud versions of traditional media. Apple's iTunes Store and App Store, through which people purchase music, video, and e-publications, earn more money than the combined revenue of the *New York Times*, the Simon & Schuster publishing company (which put out the best-selling biography of Apple's founder), Warner Bros. film studios, and Time Inc. (the largest magazine publisher in the United States). For the fiscal year ending September 2012 Apple's media cloud services earned about $8.5 billion, or $300 million more than the combined revenues of the others (Lee 2012). Because Apple does not clearly break out its pure media sales from those, for example, of its nonmedia apps, not all of its iTunes earnings come solely from media. Furthermore, Apple's content division is still dwarfed by conglomerates like News Corp. and Disney. Nevertheless, Apple's cloud media is increasing at a 35 percent annual rate, making it the fastest-growing commercial media operation in the world.

As successful as it has been in consumer services, Apple has barely made a dent in the business market for cloud services. When it has tried—for example, with iWeb, a website-publishing service—the company has failed to win over customers from its business-services competition. As

Apple backed off from iWeb, its customers needing applications to design websites and a host to serve them were left out of the cloud and in the cold. Unlike that of Amazon, Google, and Microsoft, Apple's business presence is felt only in hardware sales. These are admittedly substantial, but there has been little crossover from hardware into platforms, applications, and services. As one review maintained, "While iCloud, again, is awesome for personal use, businesses will find themselves better served by a terminal server parked in a secure data center, VPN [virtual private network] access to a corporate server, or another cloud-based file sharing solution that ensures only authorized users securely access corporate data" (Eckel 2012). In other words, customers will continue to shop the cloud at AWS, Google, Microsoft, or one of the other cloud business-service companies like Rackspace.

Facebook is also a major player in the cloud computing industry but, like Apple, it uses the cloud to service the gargantuan needs of its own site, which includes about 1.3 billion users. The company learned about cloud computing the hard way when in 2006 its computers came close to literally melting down. At that time Facebook was renting a small space in Santa Clara, California, and filled it with the racks of servers needed to store and process activity on its members' accounts. When electricity powering the growing system overheated critical components, the chief engineer and a few staff headed to a local pharmacy and bought every electric fan in the store. The fans worked, the servers were saved, and the rest, as they say, is history. The company had 10 million subscribers at the time and would not have reached anything close to the billion-plus members who upload 300 million photos a day if it failed to master the cloud (Glanz 2012b). Today, all those photos amount to 7 petabytes of data each month, and a cloud server system that calibrates storage conditions, including temperature, by calculating the likelihood that members will access information and photos. For example, colder storage slows retrieval time, but that works fine for the billion photos a day uploaded around Halloween that members are unlikely to want to retrieve after the costumes are put away for another year. These issues are challenging, but Facebook benefits from keeping all of its data needs in house. As a result, the key pressures facing any cloud provider or user, such as sharing, securing, and syncing, are more easily addressed by Facebook than by companies that are in the business of serving thousands of different businesses.

Amazon, Microsoft, and, to a lesser degree, Google, demonstrate their market dominance to cloud customers through ongoing price cuts that benefit the general user and drive smaller competitors out of business. It is certainly a problem for older companies like Oracle, HP, and IBM, which have significant costs associated with legacy systems that are not as scalable as the latest technology. As a result, these firms are starting to change, either by joining in partnerships with cloud companies or by acquiring promising smaller firms, as all three did in mid-2013 (Hardy 2013b, 2013e, 2013h; Kolakowski 2013). Moreover, IBM, which operates twenty-six data centers around the world, has begun to transform itself into a company that resembles marketing giants like WPP, Omnicom, and Publicis (Waters 2013c). All of this is taking place even as these same advertising firms are transforming themselves into ones driven by the use of big data in the cloud. The merger of Omnicom and Publicis to form the largest advertising business in the world is grounded in the need to take on the new competition from integrated cloud-based information-technology companies (Vega 2013).

Price cuts would appear to be an unqualified benefit to the cloud computing industry and especially to its users, who are increasingly dependent on the service. However, when carried out by industry leaders like Amazon or Google, they are also classic strategies to concentrate power over a market. This has been demonstrated throughout economic history, including in the communication industry from Western Union in telegraphy to AT&T in telephony. For years, AT&T initiated price cuts in telecommunications at the mere whiff of a competitive threat, only to raise them again when the competition was erased. That the company was able to accomplish this even under the regulatory nose of the Federal Communications Commission is evidence of its power and of the continuing failure of government to carry out regulatory responsibilities. It was not until the largest corporate users of telecommunication services organized collectively to fight back that AT&T's grip on the market was broken. Today, analysts wonder if cloud computing will go down that same path. According to one analyst, "There is a race to the bottom when it comes to cloud pricing, as the larger providers try to capture as much share as they can of this exploding market. The downside is that the smaller providers without huge war chests of cash, but with impatient investors, won't be able to make money at the prices that the larger names charge. Many of them will struggle to hang in through the days of low or no cloud computing profits—and many of them will have to toss in the towel or have the towel tossed in for them."[2] The only long-term upside is for the largest providers: "Once the smaller providers are pushed out, you can begin to raise your prices. Hmm, it sounds suspiciously like a page from the big-box stores' playbook—and a warning for cloud adopters not to count on low, low prices as the norm" (Linthicum 2012).

One of the keys to creating and maintaining market control is to exercise power up and down the chain of production. A handful of companies are doing this in one direction through price cuts and in another direction through their relationships with key IT producers—particularly the giant in this market, Intel, the world's largest and most highly valued semiconductor producer. Intel worries that the hardware world it dominated, led by the venerable PC, is in decline. According to one analyst, "Intel still has a lot of dough, but their old world is cracking" (Hardy 2013f). As a result, the company is especially concerned with pleasing what it refers to as the Big Four: Google, Microsoft, Amazon, and Facebook (Apple purchases its chips mainly from Samsung) not only because of their size but also because they lead a critical and growing market. Intel has been losing revenue in the personal-computer market that made it a historic leader. The $25.8 billion it earned from its PC client group in 2012 remains enormous, but that figure represents a decline of 2.25 percent from the first three quarters of the previous year, largely because of the shift to tablets and smartphones from standard personal computers and laptops. On the other hand, the company's revenue shot up by 6.7 percent in its data-center business, where it earned $7.9 billion. That has triggered a serious makeover at the company, which now views itself as more of a cloud-computing company than a client-server business (Hardy 2013f, 2013g).

The head of Intel's data-center group realizes that the company has to change direction, but believes that if it does so successfully, it could boost data-center revenues to $20 billion by 2016. But in order to accomplish this, Intel needs to listen and at times take direction from large, influential companies, something it is not used to doing. As Intel's data-center director described the situation, "The Big Four operate at a very different beat rate, and they are very tech savvy, so they don't want a lot of input. They all get a dedicated salesperson, the same as the others in our Top 40 customers, but there is a lot more direct innovation from them, and a lot of sharing of ideas" (Hardy 2012c). The Big Four are now active in engineering, innovating, and testing new semiconductors, including one Intel installed in September 2011 but did not introduce to the general public until March 2012. Intel admits that its willingness to absorb the potential production problems associated with a new chip that has not yet been released to the general public in order to have the latest semiconductor "was a new thing" for the company (ibid.). Meanwhile Apple, which has been dependent on Samsung for the bulk of its chips, is seriously contemplating manufacturing more of its own, partly because of the Korean manufacturer's announcement in November 2012 that it would boost chip prices sold to Apple by 20 percent. But this is also because Apple simply wants to control more of the production process (Whittaker 2012). Patent battles with Samsung are certainly an issue, but the need for control and the ability to carry it out are even bigger.

Large cloud companies are challenging firms at all points in the chain of production, from small cloud competitors to chip manufacturers. They are also going after companies that manufacture computers. Amazon, Apple, Google, Microsoft, and Facebook all now build their own and challenge companies like Intel and HP to meet or exceed performance specifications. Perhaps the most surprising for its activity in this area is Facebook, because it has not been among those identified with devices. The company has joined with both HP and Intel in the public announcement of a new chip. Google has even developed its own semiconductor but has not patented it because the company is concerned that doing so might reveal too much about its plans (Hardy 2012c). Amazon is building a global computer system including its own customized computers, data storage systems, networking systems, and power stations (Hardy 2013a).

These examples demonstrate some of the ways that large cloud companies are expanding to control the market. They are integrating internally to rationalize production from hardware to software, applications, and pricing. These moves enable companies to extend their control over cloud computing markets and, from there, establish key positions in the development of informational capitalism. One way to look at this process is to see it as a series of steps on the way to the computer utility. That would be accurate but, as was noted earlier, with no regulatory apparatus in place or on the horizon, it is also reasonable to see them as steps on the way to a global cartel, different from but also similar to the oil cartel that

ı engineering,
Intel installed
ic until March
ial production
·eleased to the
s a new thing"
dependent on
nanufacturing
:r's announce-
ld to Apple by
ɔntrol more of
vith Samsung
lity to carry it

ts in the chain
acturers. They
ters. Amazon,
:heir own and
l performance
in this area is
l with devices.
ɔlic announce-
ɛemiconductor
that doing so
ɛon is building
ɔmputers, data
Hardy 2013a).
ge cloud com-
ɛgrating inter-
:, applications,
ir control over
ositions in the
at this process
:r utility. That
tory apparatus
as steps on the
: oil cartel that

influenced global energy-resource markets for many years. Before long, it may be time to think seriously about the implications of a global cartel in information resources. As in oil, such a cartel would provide for the needs of organizations and individuals, using control over various stages in the production and distribution process that powers global capitalism to expand profit and control. Just as in oil or other global commodity markets, there will be small- and medium-sized producers who, from time to time, disrupt the system. Geopolitical upheavals and technological change will also have an impact. In short, cloud computing is rapidly becoming a powerful force in the world because of the quantitative and qualitative leap in information production, processing, storage, and distribution, *and* because of the way the cloud is evolving into a global, private oligopoly, well on the way to becoming a global cartel. It is also interesting to observe the ways that some of the companies making up what might become a cartel are beginning to internalize the appropriate identity for this new role. Consider Google, whose founder, Eric Schmidt, now talks about the need for the company and its competitors to start thinking of themselves as nations, especially when it comes to dispute resolution: "The adult way to run a business is to run it more like a country. They have disputes, yet they've actually been able to have huge trade with each other. They're not sending bombs at each other.... I think both Tim [Cook, Apple's CEO] and Larry [Page, Google's CEO], the sort of successors to Steve [Jobs] and me if you will, have an understanding of this state model" (Lessin 2012).

Schmidt may take this view more seriously than people think. In January 2013 he came under some pretty harsh criticism from the U.S. State Department for traveling to North Korea to meet with its leadership in a round of private diplomacy unsanctioned by the U.S. government. Citing U.S. concerns about a North Korean rocket launch one month earlier, a State Department spokesperson commented, "Frankly, we don't think the timing of this is particularly helpful." Moreover, "They are traveling in an unofficial capacity. They are not going to be accompanied by any U.S. officials. They are not carrying any messages from us. They are private citizens and they are making their own decisions." Coming from the agency responsible for American diplomacy, these are pretty strong words about a prominent U.S. citizen (Gordon 2013; see also Schmidt and Cohen 2013).

Developments like these lead some to wonder whether we are soon to face the problem of monopoly market domination that once led the government to intervene against the power of Standard Oil, IBM, and AT&T. Some have maintained that it was government pressure on IBM, even as it dropped the thirteen-year-old case in 1982, that led the company to unbundle its software from the hardware portion of the business and thereby advance the massive growth of the U.S. information-technology industry. Furthermore, it was likely that the breakup of AT&T around that same time helped make the Internet possible. In addition, the government's 1990s case against Microsoft, which had suffocated innovative companies like Netscape, made it considerably easier for Google and Facebook to appear (Fox 2013).

Not everyone agrees with the view that an oligopoly or a cartel is about to be born. Some maintain that, even with continuous price cuts, Amazon will face stiff competition from within and outside the major cloud providers, including from small innovative companies. There are also concerns about Apple's ability to enjoy elite status in the cloud. Analysts point to the difficulty the company has experienced in making its bedrock iTunes service meet the promise of seamless integration and synchronicity across platforms. Moreover, the company has not expanded its services with offerings that have earned Google and Microsoft the reputation of general cloud-server companies. Also, while everyone agrees that Microsoft has succeeded in building on its success in business services as it has moved to the cloud, doubters wonder whether Windows 8 and SkyDrive will succeed in creating a major cloud-computing presence in the consumer market (*Cloud Tweaks* 2012). Some also insist that many companies, seemingly beaten by the new Big Four (or Five, if you include Apple), have the capacity to fight back and are beginning to do so. These include big broadcasters who have seen their audiences diminish in the expansion of digital social media. According to one analyst, "But as more and more Internet-connected smart televisions find their ways into people's homes, broadcasters see a new opportunity to remain at the center of the global ad industry" (Steel 2012b). They can do so partly because the new wave of Internet-connected televisions permits broadcasters like CBS to sell new forms of advertising to direct marketers who do not typically purchase commercial advertising because they focus on coupons, search ads, and direct marketing. Internet-enabled television receivers permit broadcasters to add web advertisers to the brand advertising that built the industry. Broadcasters now capture only $10 billion of the $60 billion spent annually on direct marketing. But the shift to Internet television has the potential to enable broadcasters to expand that share and enter new markets. So while it is likely that there is some hyperbole in the statement by a CBS researcher that this will usher in "a new golden age of network television," it does indicate that "legacy" companies like NBC, CBS, and ABC will have something to say about the emerging consumer cloud cartel (ibid.).

Three of the most important challengers to Amazon and other major players in the cloud should be familiar to anyone who has purchased a computer or printer over the last twenty years: IBM, HP, and Dell. These companies hope to profit by building on their established base in data processing and storage to provide services to cloud customers and by serving other cloud-computing companies. It should come as no surprise that IBM is involved in the cloud; the company has had its fingerprints on just about every device associated with the history of computing. In addition to the standard business of hosting providers offering applications over the Internet, IBM is well on its way to, in the words of one analyst, "becoming a sort of arms provider for the cloud, selling customized hardware and software that helps governments, large and mid-sized companies, or Web developers" (Ante 2012). The company is involved in every facet of cloud services, but in 2012 it made a major move to promote its cloud to mid-sized businesses, which meant taking on market leaders AWS and Salesforce. The company was initially successful, posting double-digit gains in its cloud business. However, as with other firms whose history of providing software and other IT services preceded the development of cloud computing, success in the cloud may come at the expense of its core business. This major risk was captured in continuing revenue declines in IBM's Global Services unit and in software sales. The problem for companies like IBM, as well as for HP, Dell, and Microsoft, is that cloud services can cannibalize their own key businesses, including selling software and offering consulting services to help companies run their own IT-linked supply chains. With more and more of IT bumped to the cloud, companies are less likely to require software and services that maintain their own individual IT silos. According to one investment analyst, "We could be seeing the tip of the iceberg on an important deflationary force for traditional packaged applications services" (Ante 2012). That just happens to represent the majority of IBM's global-services business. Compounding the problem is that as long as cloud services live up to their promise of lowered IT costs for companies, and so far they have, cloud revenue for firms with a long history cannot possibly keep up with what they enjoyed in the past when they sold software and services to a host of individual businesses. This is not a problem for companies like Amazon (with its AWS offering), which does not have a legacy business to protect. How IBM, HP, Dell, Microsoft, and now Apple handle this classic case of the "innovator's dilemma" will go a long way to determining whether they have a future of any consequence in or outside the cloud (Bradshaw 2012).[3]

Rackspace represents a set of cloud companies that, unlike IBM, does not have either the advantages or disadvantages of legacy systems to worry about and has moved full bore into providing cloud services. The company, which began in 1998 as a small Internet service provider in founder Richard Yoo's garage, quickly grew to become an established host for customized applications, providing private, public, and hybrid cloud services. Widely recognized as one of the leading cloud companies and with more than 4,000 employees, Rackspace relies on what is called the OpenStack, software that is universally available based on open source principles.[4] In 2012 it approached 200,000 customers using close to 100,000 servers in about 250,000 square feet of data-center space around the world. Demonstrating that it can play with the heavyweights, the company's annual revenues surpassed $1.5 billion. Nevertheless, with long-established firms pouring money into cloud offerings, Rackspace faces an uncertain future. Consider that Dell alone invested $1 billion into its cloud in 2012. How does a firm that takes in not much more in annual revenues keep pace? Additionally, Rackspace benefited from complicated pricing for companies unsure of the technology and the market and unable to gauge pricing well. Now, as the cloud approaches commoditized utility status, with standardized pricing based on hourly use for all customers, Rackspace will have a more difficult time distinguishing itself from large firms like Dell and AWS.

Unlike Rackspace, which has grown to become a leader in general cloud services, companies like Salesforce, which uses the cloud for managing customers, and VMware, which provides cloud services through virtualized

:e 2012). That
:vices business.
live up to their
ıcy have, cloud
ɔ up with what
.ces to a host of
:s like Amazon
ness to protect.
:his classic case
ɔining whether
oud (Bradshaw

:, unlike IBM,
legacy systems
cloud services.
.ervice provider
: an established
ılic, and hybrid
oud companies
n what is called
ɔased on open
ners using close
ıta-center space
: heavyweights,
vertheless, with
Rackspace faces
l billion into its
more in annual
ɔm complicated
ırket and unable
ıoditized utility
r all customers,
itself from large

in general cloud
r managing cus-
ugh virtualized

servers, are leaders among the specialists.[5] The general public became acquainted with Salesforce when it ran two ads costing $3 million during the 2011 Super Bowl. Marc Benioff founded the company in 1999 as one of the first to offer Software as a Service, and the company has since added Platform as a Service to its offerings. Its specialty is customer-relationship management (CRM), a system for managing interactions with clients and prospective clients, primarily to expand sales but also to manage customer service and technical support. CRM has been in use for two decades and is now expanding into the cloud. It operates through software that enables companies to manage their sales and customer-service processes and assess successes and failures. Rather than house CRM internally, companies contract with Salesforce, which provides software and services from its cloud servers. These include storage for all data associated with marketing and sales for a specific company and access to 20 million or so files on business contacts. Companies can also work with Salesforce to develop their own applications and tools in the Salesforce cloud. In April 2012, the company employed close to 8,000 people and generated $2.25 billion in annual revenues. In 2013 it joined a wave of merger and acquisition activity in the industry by spending $2.5 billion on ExactTarget, a company that specializes in managing sales campaigns. As cloud leaders like AWS bulk up with takeover activity, Salesforce felt the need to keep pace. The upside of specialization is that it enables a company to concentrate resources and expertise, but the downside is vulnerability. The company faced this in 2007 when it fell victim to a phishing attack that enabled hackers to lure an employee into revealing credentials that were used to gather customer contact data. The attackers went on to send further attacks to customers through fake Salesforce invoices. Some customers fell for the scam and coughed up more information. For a company specializing in the secure management of customer relations, this was an especially difficult and almost company-destroying failure. Larger firms like Amazon have faced similar challenges, but highly diversified companies like Amazon are better able to weather such storms.

The other challenge for a specialist company is facing genuine competition from one of the giants that can bankroll a major initiative and keep it going in the absence of an immediate boost in profit. Such a challenge came from Microsoft, which moved into CRM after Salesforce but has begun to catch up in customers, markets, and offerings. More importantly, Microsoft Dynamics CRM can draw from users' familiarity with Microsoft products like Office and Outlook to make them feel more secure about taking the leap into cloud-based CRM. Furthermore, because Microsoft has years of experience in servicing on-premises IT departments, it can offer clients a mix of cloud and on-premises data-center services. The key point here is that challenges to leading cloud companies like AWS and Microsoft do come from the diverse set of firms in the cloud marketplace, but big players can also respond powerfully to even substantial inroads from specialist firms (CRM Software Blog Editors 2011). Undaunted, management at Salesforce is rethinking its future by preparing for what it calls Cloud 2, or use of the cloud in social media, especially in mobile communication. In 2012 it took a step in this direction with the $212 million purchase of Heroku, a leading Platform as a Service provider that helps companies develop cloud-based applications.

It is hard to determine whether Salesforce can withstand the competitive push from one of the giants and move into new lines of business. The outcome will also depend on how well Salesforce fends off pressure from other companies making software cloud services a key part of their business. One of the firms to contend with is Oracle, a major business-software provider that until 2012 eschewed the cloud in favor of selling software directly to its business clientele. In fact, its CEO, Larry Ellison, is known to have dismissed cloud computing as a fad. The success of Salesforce and similar companies has changed this view and, after years of foot dragging, the company went on a buying spree that added eleven new companies to the Oracle stable, all but one of which sells software applications through the cloud.[6] In 2013 the company extended its reach into the cloud by launching a set of partnerships, including deals with Microsoft and with Salesforce. These drew a lot of attention, especially among those concerned about growing concentration in the cloud industry (Hardy 2013h). Another challenger to Salesforce, the German software company SAP, has been even more aggressive than Oracle, spending $8 billion on cloud software companies. SAP and Oracle are especially concerned that the cloud will disrupt their traditional model of providing software to business clients (Waters 2013d). All of this amounts to both intense competitive pressure in the growing market to sell software through the cloud and growing consolidation in the cloud software marketplace. Although a number of small firms remain, most are facing amalgamation by choice or necessity. As one industry expert explained, "a wave of deals is likely to leave only a small handful of bigger and more diversified companies standing" (Waters 2012).

Telecommunications Companies Take to the Cloud

For several reasons, telecommunications companies have an enormous stake in cloud computing and they are well positioned to battle the leaders in the industry (Babcock 2013b). It is important to understand that these businesses, especially large companies like AT&T and Verizon, are not just conduits for other firms' data. Through their subsidiaries, they are well integrated into the entire digital economy, including content provision. Consequently, the cloud challenges the entire telecommunications industry because it provides new ways to offer services that have been part of the telecommunications industry for years. The challenge deepens as a handful of integrated conglomerates, the digital giants Google, Apple, Amazon, Facebook, and Microsoft, solidify their hold on cloud services. As these firms build towering silos of their own, once-dominant telecommunications companies are wondering about their place in the cloud economy. Rather than sit back and wait for the industry to settle, firms like AT&T and Verizon have moved quickly to secure a stake. Verizon, in particular, has become a major leader among cloud-telecommunications firms by employing a strategy that has been used over and over again in the industry's history: when the next new thing comes along, buy it. Verizon did so in 2011 by spending $1.4 billion on the major cloud company Terremark, and by acquiring the cloud-application firm CloudSwitch to make the total of the company's cloud investments for the year more than $2 billion. These deals took Verizon to the top of a growing field of telecommunications companies that have moved into the cloud and, in the words of one industry analyst "are prepping Verizon for massive future growth" (Hickey 2012). As important as it was to purchase these assets, Verizon's more important challenge was to integrate them into its other lines of business, especially wireless and FIOS, its bundled Internet access, telephone, and cable service delivered by fiber-optic cable.

For Verizon, the cloud is a key component of a media, telecommunication, and information convergence strategy that will allow the company to control practically all key nodes in the networks that produce, store, process, and distribute services to individual and organizational customers. Moreover, Terremark gives Verizon a significant international presence, something that the company has lacked, particularly in Latin America. It is uncertain whether Verizon can make this strategy work. Many companies, with AOL Time-Warner the most celebrated, have run aground with "can't miss" convergence deals. The outcome will go a long way to determining whether Verizon can join the leaders in the cloud-based communications industry. Complicating matters for Verizon is the expansion of competitive pressures that threaten its comfortable duopoly with AT&T in the United States. The acquisitions of Sprint and of Clearwire have made SoftBank, in the words of one analyst, "a better-funded number three with the spectrum to launch low-priced wireless data products." Moreover, the T-Mobile–Metro PCS merger created a fourth big player in the U.S. market and the ability of the spectrum-rich Dish Network promises to further disrupt the comfortable control of the market that Verizon has enjoyed (*Globe Investor* 2012; Taylor 2013b).

The U.S. Government: Trusting the Cloud and Commercial Providers

Not all cloud computing is controlled by private organizations. But it is interesting to observe the extent to which the U.S. government depends on the private sector for its cloud computing needs, including relationships based on no-bid, sole-supplier contracts with the largest cloud providers. This is significant for several reasons, not the least of which is the amount of money involved. According to one report, the government spends $80 billion annually on information technology and plans to move about 25 percent of its IT budget to the cloud. An example of the movement to cloud services provided on a single-source, no-bid basis is the Naval Supply Systems Command's plan to use Amazon Web Services to store and distribute digital photography and video. The Navy's argument is that AWS offers a single, integrated package that is more reliable and less prone to attack than other cloud services (Foley 2012). Furthermore, NASA, which helped to develop OpenStack, the open source standard that IBM uses for its cloud, also contracted with AWS (Thibodeau 2013). Even the

CIA planned to tender AWS a $600 million contract until IBM blew the whistle, raising questions about how the federal government handles cloud contracts, and a review of the agreement with AWS (Woodall 2013). While waiting to learn whether its bid for the CIA's cloud business would succeed, IBM won the largest government cloud-computing contract, worth $1 billion, from the Interior Department (Miller and Strohm 2013). That helped cushion the blow for IBM when Amazon was officially awarded the CIA contract (Babcock 2013a).

These moves are not very surprising, particularly in light of the history of the U.S. government's relationship to large communication companies (Mazzucato 2013). For years, government agencies, including the Department of Defense, had a very close relationship with IBM for computing and an even closer one with AT&T for telecommunications services. Even as business consumers lined up to support breaking up AT&T and deregulating the telecommunications industry in order to lower prices, the DOD argued that national security required the end-to-end service that AT&T provided. It was not until the Pentagon was assured that security needs would be met that it dropped its opposition to breaking up the telecommunications giant (Schiller 1981). Given this preference for large, stable companies, it is not surprising that the government would turn to AWS to meet some of its cloud-computing needs.

The U.S. government's current move to the cloud is propelled by the belief that cloud computing must become a central means of meeting its information-technology needs. In December 2010 the federal Chief Information Officers Council released a plan to reform government information technology, which included requiring agencies to adopt a "cloud-first" policy for new IT deployments. According to the plan, cloud-first is driven by three interrelated forces. First, large data centers provide economies of scale that are necessary to meet the growing needs of the federal government's "computation infrastructure." For federal IT planners, it is less expensive to centralize data in a few large centers than to retain it in local offices. Second, cloud systems are able to provide almost any type of computation on demand. It is difficult to predict the type and speed of processing and analysis that will be needed and the planners side with those who believe that cloud systems are agile enough to meet their needs, including those they cannot now anticipate. Finally, the cloud unleashes unprecedented analytics capability on large data collections. It is clear from this view that federal IT planners rank big data among the major attractions of cloud computing. Data centers are intended to be not only storage warehouses that agencies can call on when they need data, but also active producers of information that draw on stored data sets (Page 2011). In 2011 NIST released its report defining cloud computing and carefully describing the cloud's specific characteristics to enable managers and staff operating within agencies to have a better idea or, in some cases, their first clear idea of what it was they were being ordered to implement. In 2012 the National Science Foundation (NSF 2012) produced a short report supporting NIST's conclusions and committing the government to fund research into cloud computing. The combination of strong affirmations from the federal government's CIO, from NIST, and from the NSF provided the grounding for strong state support for the cloud.

There are also major implications in a number of government demonstration projects in education and research. One of the most significant is a program operated out of the National Endowment for the Humanities Office of Digital Humanities. It demonstrates how government's use of the cloud and big data is contributing to the restructuring of education, and not just in the areas where we would expect change, such as computer science and the disciplines associated with the sciences. It is also reaching into the social sciences and even the humanities. One can learn a lot about the direction of change from the size of a force creating it, but one can also learn a great deal from its reach, as when government projects extend to fields traditionally kept outside the scope of computerization. Chapter 5 examines the digital humanities in the context of assessing big data in the cloud. Suffice it to say here that the digital humanities project represents an important initiative that is often lost in the understandable focus on larger military and civilian projects. Its significance for the future of education and research far outweighs the size of its budget (Gold 2012).

In spite of the enthusiasm for the cloud in government, there remain several issues that have the military and intelligence sectors especially concerned about moving data to corporate-owned cloud systems. Arguably the most important is security. At the very least, there is concern about moving classified data and computer power essential for combat missions to off-site locations. Formal concerns have already been raised with respect to the security of data in NASA's cloud systems (Kerr 2013). Furthermore, the size and complexity of government and especially military computer systems make the prospect of moving to the cloud very expensive. It would not be a matter of simply relying solely on available technologies because many government departments, and especially the military and intelligence sectors, require customized systems that are integrated within and across units. Finally, government, and especially defense, requires a very high level of support and, while some of the major providers have developed excellent backup for their customers, it is uncertain whether the necessary support is available in the current cloud industry (Gangireddy 2012).

Even in the face of these worries, the government is showing a level of faith in private cloud companies that has surprised some experts. This extends to using private cloud firms to provide security for the government's systems. For example, the Naval War College awarded a single-source contract to the SaaS vendor CloudLock to safeguard the implementation of online tools like Google Docs and Google Drive. Given the concern with security, one analyst responds to this use of the cloud to protect the cloud with the conclusion that "it's remarkable that agencies are defying conventional wisdom in this way" (Foley 2012). In a more significant step, intelligence agencies are beginning to make use of commercial cloud computing, including the public cloud, which serves all customers. Furthermore, according to one IT leader in the intelligence community, agencies now have enough confidence in the public cloud "to bring some commercial cloud capabilities inside our fence lines" (ibid.).

The alternative to this use of commercial cloud services is to retain IT activity on-site or to develop a government, military, or intelligence-agency cloud capability. This is certainly taking place too. In 2011, Los Alamos National Lab began providing IaaS services from its own data center and has joined with the National Nuclear Security Administration to develop a community cloud that extends to the entire Department of Energy (ibid.). Of greater strategic significance is the Department of Defense decision to create a military cloud as a means to fend off cyber-attacks that have been proliferating in recent years. These include the April 2010 attack emanating from China that redirected 15 percent of Internet traffic through China's networks for eighteen minutes and the 2011 virus attack on U.S. drone weapons. The latter used malware to record keystrokes and required continuous deletion and rebuilding of hard drives. To avoid these attacks, DARPA set up Cloud to the Edge (COE) in 2011, which began by opening a set of hotspots for secure communication. According to one analyst, COE looks a lot like Google's suite of online services, minus the search engine (Tanaka 2012). It is hosted on a secure system of servers by the Defense Information Systems Agency, which has itself given out a $45 million sole-source contract to the Alliance Technology Group for a data-storage facility to provide four exabytes of storage capacity (Hoover 2013). To back up its cloud initiative, the Department of Defense committed another $5 million to advance its cyber-battleground project, with the auspicious title of Plan X, that would allow the agency "to rehearse and manage what officials call 'cyberwarfare in real-time, large-scale, and dynamic network environments'" (Nextgov 2013).[7] To implement its plan the Pentagon will hire and deploy 4,000 military and civilian technology specialists to the U.S. Cyber Command, but that is not likely to be enough (Brannen 2013). This prompts some to anticipate a near-term shortfall in cloud experts (Weisinger 2013).

It is not just security that prompts interest in the cloud. The DOD also wants to better manage its IT budget and hopes the turn to cloud computing will go a long way to saving 30 percent by 2016. Already engaged in the consolidation and modernization of data centers, the DOD has eliminated many and cut the number of technical support desks in half. Overall, it would like to reduce the number of networks, data centers, and help desks by 80 percent (Tanaka 2012). Storing everything from unclassified to top-secret information, the military cloud began with a test case led by the National Security Agency, which gathers, stores, processes, and analyzes huge amounts of data. Typically sheltered from the public attention that is more typically directed at the CIA and the FBI, the NSA, which is three times the size of the CIA and has one-third of total U.S. intelligence spending, burst onto the front pages of newspapers worldwide in the spring of 2013. A series of leaks and newspaper accounts revealed that, contrary to previous claims, the agency worked closely with U.S. telecommunications providers and the largest Internet companies to gather data on Americans and foreigners by scooping up and analyzing telephone conversations, emails, social-media postings, and other electronic communication. With the $20 million Prism program that included major Silicon Valley and telecommunications companies that shared information on users with the NSA, the spy agency hoped to better target threats to the United States by analyzing metadata—that is, who was contacting whom, as well as content whose keywords big-data analysts could use to root out suspected terrorists (Luckerson 2013). Nevertheless, many critics took issue with what appeared to be an unprecedented and, until the leaks, secretive attack on the privacy of users (Wilson and Wilson 2013). Controversies aside, government policy makers hope that cloud computing will enable the NSA to meet its goals with greater security and at lower cost, thereby demonstrating the value of moving other government agencies to the cloud. Nevertheless, experts worry that concentrating military information in one large cloud system, however well secured, provides an inviting target for cyber-attackers around the world. One expert worries that the move to the cloud is the equivalent of "painting a cyber bullseye" on the NSA and the military: "Cloud computing, in military terms, fosters a target-rich environment because the very things that make the cloud appealing also make it a tempting mark. Because of this and the high probability that a vast amount of data will be stored on a cloud, attackers only need to be lucky once as compared to having to be lucky multiple times when attacking a legacy system. With this in mind, a more appropriate question for the NSA would be 'what kind of information would your organization refuse to place on a Cloud?'" (Tanaka 2012)

It is not as if military planners are unaware of the security problems of cloud computing. According to DARPA, "Cloud computing infrastructures, in particular, tightly integrate large numbers of hosts using high speed interconnection fabrics that can serve to propagate attacks even more rapidly than conventional networked systems. Today's hosts, of course, are highly vulnerable, but even if the hosts within a cloud are reasonably secure, any residual vulnerability in the hosts will be amplified dramatically" (ibid.). Nevertheless, like many other agencies, it is convinced that, with appropriate security measures, the military benefits of cloud systems outweigh the risks because "clouds and distributed computing environments can: provide redundant hosts, correlate attack information from across the ensemble, and provide for diversity across the network" (ibid.). What matters for the military is whether it can develop what it calls "mission-oriented resilient clouds" that can be deployed effectively in combat.

Clouds over China

Cloud computing systems have a firm foothold in the United States, where about 40 percent of the world's data centers are located, but they are also spreading internationally. Outside the United States, Scandinavia has become a major data-center venue and the cloud is no stranger to the Middle East, but China has made the most significant progress in the overall development of cloud computing (Horn 2011; Glover 2013). By the end of 2012 China represented about 3 percent of the global cloud marketplace, but it is expected to grow at a 40 percent annual rate, reaching $18.6 billion in annual revenue by the end of 2013. Led by China, the Asia region is expected to lead the world in cloud traffic and workloads by 2016 (Ong 2012). China's burgeoning cloud industry benefits from minimal competition with the major U.S. providers. Amazon is not there and Microsoft has just begun to introduce its Azure cloud service in China. This has left lots of room for the development of indigenous cloud services, including the Alibaba Group, which provides both cloud infrastructure and services to a variety of national and multinational clients over its Aliyun network. In addition, Baidu, known in the West as the "Google of China" for its prowess in search services, has invested heavily in cloud storage and processing, evidenced in a 2012 investment of $1.6 billion in a new data center and a deal to offer free personal cloud storage on Android phones. Baidu's major competitive challenge comes from Tencent, an instant-messaging and online-gaming company with 400 million users, making it one of the largest consumer-application cloud companies in the world, with a valuation in 2012 of $60 billion. In 2013 Tencent took a major leap in the cloud marketplace when it announced that it would be the first to build a center in the western China city of Chongqing, where planners expect significant new growth in the cloud (*People's Daily Online* 2013). In 2012, the world's leader in telecommunications equipment production, Huawei, also moved into cloud computing and storage, a decision that led to a significant growth in company profit (Reuters 2013a). China's cloud development is helped by the presence of Asian firms like Pacnet that benefit from having developed network and data-center services in the Asian region, including Hong Kong, Singapore, and Australia (Powell 2013).

In 2013, Baidu demonstrated that it does far more than provide service to China when it signed a deal with France Telecom to offer its mobile browser throughout Africa and the Middle East on the French company's smartphones (Thomas 2013). In addition to these network-driven cloud providers, companies have emerged that provide storage services. A leader in this area is MeePo, a storage service similar to Dropbox. The company has experienced remarkable growth, with capacity in 2012 reliably estimated at 50 terabytes (Chou 2012).

One of the most ambitious cloud projects in the world is China's commitment to build cloud cities. The goal is to construct giant data centers connected to firms that provide value-added services, as well as research and development for domestic and international markets. Some of these involve working with major international partners who provide capital and expertise, even as local companies control the project. For example, China-based Range Technology is teaming up with IBM to construct a 6.6-million-square-foot cloud-computing center in Langfang, near Beijing. It will provide cloud services to government and private-sector organizations, as well as host cloud systems and mobile devices (Bundy and Haley 2012). In addition to linking up computer-service providers like Baidu and computer companies like Lenovo, cloud centers also welcome the involvement of China's large telecommunications companies. For example, in 2011 China Telecom formed a partnership with the global cloud-services company SAP to offer cloud services to small and medium-size businesses in China. In 2012 the country's three giant telecommunications firms, China Telecom, China Mobile, and China Unicom, agreed to invest $47 billion to develop data centers, including one of the world's largest, to help create an economic hub in Chengdu, a city in China's southwestern province of Sichuan. Chengdu already builds one-fifth of the world's computers and the plan is to expand the Tianfu Software Park around the cloud data centers. In this way Chengdu will move up the value ladder from computer manufacturing to data storage, processing, and transmission, on the way to becoming a center for research and development (Evans-Pritchard 2012). With fifty-one universities graduating 200,000 scientists and engineers each year, Chengdu has the foundation to take these steps to higher-value production.

China certainly appears to be poised to become a world leader in cloud computing. It is building enormous cloud data centers, including some of the world's largest, at a feverish pace. Not satisfied with the construction of cloud facilities, it is creating entire cloud cities. Of equal significance, China is carrying out a detailed cloud-computing strategy that is most significant for integrating all the major participants, including hardware manufacturers who are becoming leaders in server production for the global marketplace, software designers, application developers, business-service providers, and telecommunications companies. But there is another side to this success story. China faces technical challenges, including connectivity problems and the absence of certification programs for cloud companies and their staff, something that has been institutionalized among leading companies like Amazon. Moreover, as Chapter 4 describes, cloud computing faces numerous environmental, social, and labor challenges. These are all greatly heightened by the size and speed of cloud development in China, as well as by the unsettled nature of its political and legal infrastructure (Qian 2013).

Cloud computing creates significant environmental problems associated with its massive energy requirements and, secondarily, with construction and disposal of materials and equipment. These are all exacerbated in China because the country is already plagued by widespread air pollution as energy needs have spiked across the country, and reliance on coal-fired power plants deepens the problem. Building the world's largest cloud facilities, including entire cloud cities, will only add to an already critical problem. The same holds for security, surveillance, and privacy issues. These pose challenges everywhere, but nowhere more prominently than in China, where there is no guarantee that if they build it, the world will come. China has long been mired in controversies about the security of personal and organizational data. Will Western companies and governments that have complained about the theft of data store their information in China's data centers? A society that practices massive surveillance of its own citizens and routinely censors information can hardly be surprised to find very low trust in the security of its cloud systems. It is not only foreign businesses that worry about surveillance issues. A 2013 Forrester Research report documented concerns among Chinese entrepreneurs who are reluctant to take to the cloud. Some of this results from the lack of experience with outsourcing or externally managed services. With little to prepare them for the cloud, companies are understandably cautious. But security worries loom large and this accounts for a distinct preference for private-cloud services as the less risky cloud option (Qing 2013). Finally, China's hyper-accelerated industrialization has created massive labor problems that include but extend well beyond the notorious practices of the electronics manufacturer Foxconn. Annually producing 200,000 scientists and engineers in one city alone is an outstanding achievement, but managing them and the millions of new workers who constitute China's army of knowledge workers is an entirely different challenge. From construction to operation, from maintenance to support, cloud computing makes enormous demands on labor markets and workplace practices. To add these demands to a society already in the throes of labor upheavals across the country will certainly tax China's leadership for years to come.

This overview of cloud computing has covered key features of its genealogy, defining elements, key characteristics, and major exemplars. The next chapter builds on this foundation by examining how cloud computing is promoted in marketing and myth, and describes why it is important for supporters to fashion this complex, but nonetheless banal, technology into the technological sublime.