4-1-2011

Cloud Computing: Architectural and Policy Implications

Christopher S. Yoo
University of Pennsylvania Carey Law School

Cloud Computing: Architectural and Policy Implications

Christopher S. Yoo*

ABSTRACT

Cloud computing has emerged as perhaps the hottest development in information technology. Despite all of the attention that it has garnered, existing analyses focus almost exclusively on the issues that surround data privacy without exploring cloud computing's architectural and policy implications. This article offers an initial exploratory analysis in that direction. It begins by introducing key cloud computing concepts, such as service-oriented architectures, thin clients, and virtualization, and discusses the leading delivery models and deployment strategies that are being pursued by cloud computing providers. It next analyzes the economics of cloud computing in terms of reducing costs, transforming capital expenditures into operating expenditures, aggregating demand, increasing reliability, and reducing latency. It then discusses the architectural implications of cloud computing for access networking (focusing on bandwidth, reliability, quality of service, and ubiquity) and data center interconnectivity (focusing on bandwidth, reliability, security and privacy, control over routing policies, standardization, and metering and payment). It closes by offering a few observations on the impact of cloud computing on the industry structure for data centers, server-related technologies, router-based technologies, and access networks, as well as its implications for regulation.

I. INTRODUCTION

Cloud computing has emerged as perhaps the hottest recent development in information technology. Some observers believe that cloud computing represents a breakthrough development that has the potential fundamentally to transform the nature of computing. Others are more skeptical, arguing that it is nothing more than overhyped repackaging of already extant

* Professor of Law, Communication, and Computer & Information Science and Founding Director of the Center for Technology, Innovation and Competition, University of Pennsylvania. I would like to thank Bill Boebel, Todd Broebsting, Daniel Burton, Robert Krauss, Bill Lehr, Don Norbeck, Jonathan Smith, and particularly Joe Weinman for their help in introducing me to many of the nuances of cloud computing. This work is partially supported by an event on "Antitrust and the Dynamics of High Technology Industries" sponsored by the Technology Policy Institute on October 22, 2010, as well as National Science Foundation Grant CNS-10-40672. It expands on an analysis initially presented in Yoo (2010, pp. 83–87).

technologies.[1] Notwithstanding the divergence of opinions regarding its future prospects, a new cadre of companies has emerged that specialize in cloud computing solutions, including GoGrid, Iland, Rackspace, Saavis, and Sungard. Established computer industry players, such as Amazon, Google, Hewlett Packard, IBM, Microsoft, and Sun, have entered the fray, as have traditional telecommunications companies, such as AT&T, Comcast, NTT, and Verizon. Cloud computing's growing salience is forcing every industry participant and enterprise customer to come to grips with this emerging phenomenon.

Despite all of the attention that has been garnered by cloud computing, many know little about its basic principles or economic foundations. The scant analyses that already exist focus almost exclusively on the issues that surround data privacy. Most importantly, an analysis of the architectural and policy implications has yet to appear in the literature.

This article takes the first exploratory step in that direction. It begins by providing an overview of the key concepts and basic economics underlying cloud computing. It then assesses cloud computing's architectural implications for both access and core networking functions. It closes by offering some observations about cloud computing's impact on market structure and network regulation. As is inevitably the case with any new technology, the novelty and dynamic nature of the subject matter means that any initial take is inevitably incomplete and vulnerable to immediate obsolescence. That said, I hope that my analysis of cloud computing's architectural and policy implications will prove helpful in furthering academic scholarship, business decision making, and policy debates.

[1] Perhaps the most expansive vision of cloud computing is Carr (2008). Leading skeptics include Oracle head Larry Ellison and free software advocate Richard Stallman (Johnson, 2008).

II. KEY CLOUD COMPUTING CONCEPTS

As is the case with many new architectures, a precise definition of cloud computing's key characteristics has proven remarkably elusive.[2] Nevertheless, there is broad agreement that cloud computing is centered on certain core concepts. Some observers have noted that cloud computing has both an outward-looking and an inward-looking face (Birman et al., 2008). From the outward-looking perspective of an end user looking at the cloud, it shifts functions that used to be performed by computers located at the network's edge (such as housing software and data) into data centers residing in the network's core. From the inward-looking perspective of how individual cloud computing elements interact with other cloud computing elements, the focus is on the ability to coordinate and integrate applications and data operating on multiple machines through mechanisms into a seamless whole.

A. Service Oriented Architecture/Thin Clients

Beginning first with the outward-looking view, the dominant vision of computing today involves an application running through software and accessing data that are both stored locally. The data and applications employed by cloud applications, in contrast, reside in a data center rather than in the end user's machine. For example, email has typically been accessed through a software client (such as Microsoft Outlook) that resides on the hard disk of a desktop or laptop computer and stores data on that same hard disk. Someone accessing email through a web-based email service (such as Google's Gmail or Microsoft's Hotmail) does not need to run an email program (known as a client) or store messages locally. Instead, both the application and the underlying data are hosted in Google's data center. A similar distinction can be drawn between

[2] For surveys of proposed definitions of cloud computing, see Geelan (2009); Vaquero et al. (2009); Weinhardt et al. (2009). For other examples, see Buyya et al. (2009); Foster et al. (2009).

an end-user who is running a traditional word processing application (such as Microsoft Word or Word Perfect) and another end–user who is using a cloud based application (such as Google Docs or Microsoft Office Live).

Rather than regarding software and computing power as *products* that consumers purchase in lump sums in advance, cloud computing reconceptualizes software and computing power as *services* that are purchased incrementally on an as-needed basis. Unlike previous iterations of distributed computing (such as grid computing), which required end users to master deployment details and to perform numerous management functions in order to gain access to the shared resources, cloud computing is easily configurable on demand by end users.

Transferring applications and data to data centers drastically simplifies the functions that must be performed by the machines that are owned and operated by end users. Instead, end users only need a *thin client* that is capable of providing nothing more than network connectivity and enough computing power to run a web browser or some other simple interface. The classic example is the netbook, which offers less computing power than the typical laptop or personal computer. The result is an end user device that is much simpler and less expensive.

B. Virtualization

First developed by IBM during the 1960s better to utilize mainframe computing, virtualization allows the computing power of a single machine to be subdivided into a number of smaller virtual machines by permitting a single piece of hardware to run multiple operating systems or multiple sessions of the same operating system. This allows end users to share the same machine while giving the appearance that the end user's application was running on a separate, dedicated machine. The ability to run multiple instances on the same machine permits

finer granularity in provisioning computing services and allows computing resources to be utilized more efficiently.

In addition to subdividing a single machine into multiple smaller virtual machines, modern virtualization techniques also allow cloud computing environments to shift virtual machines from one server to another. Thus, should one cloud computing application need more computing power than initially allocated to it, the software that manages virtualization (called a hypervisor) can seamlessly transfer the task to another server that has more space. The ability to reallocate additional storage and computing power as needed greatly enhances the flexibility and scalability of computing operations. Unlike previous forms of distributed computing, such as grid computing, organizing such virtual resources is done through automatic processes that give the impression that each application is running on a different piece of hardware that is dedicated to a single end user.

C. Delivery Models for Cloud Computing

Cloud computing providers can offer services at different layers of the resource stack, simulating the functions that are performed by applications, operating systems, or physical hardware. Although some commentators categorize cloud computing services somewhat differently, the most common approach segregates services into a three-part taxonomy (see, for example, Foster et al., 2008; Vaquero et al., 2009; Mell & Grance, 2009).

- *Software as a Service* (SaaS) offers finished applications that end users can access through a thin client (typically, but not necessarily, a web browser). Prominent examples of SaaS include Gmail, Google Docs, and Salesforce.com. The end user does not exercise any control over the design of the application (aside from some minor customization and configuration options), servers, networking, and storage infrastructure.

- *Platform as a Service* (PaaS) offers an operating system as well as suites of programming languages and software development tools that customers can use to develop their own applications. Prominent examples include Microsoft Windows Azure and Google App Engine. PaaS gives end users control over application design, but does not give them control over the physical infrastructure.
- *Infrastructure as a Service* (IaaS) offers end users direct access to processing, storage, and other computing resources and allows them to configure those resources and run operating systems and software on them as they see fit. Examples of IaaS include Amazon Elastic Compute Cloud (EC2), Rackspace, and IBM Computing on Demand.

D. Deployment Strategies

The literature distinguishes among at least three different deployment models of cloud computing. In the case of *private clouds*, all of these services are deployed through a privately owned data center that is used exclusively by the organization that builds it. These private clouds may deploy proprietary technologies that are inaccessible to other users of cloud services. In contrast, *public clouds* are provided by third parties that offer their services to a wide range of interested customers. As such, public clouds come the closest to the vision of utility computing that some have advanced since the 1960s. The key difference is that public clouds are more competitive, do not bear a duty to serve, and typically offer a wider range of quality of service and pricing than do traditional public utilities. Rather than commit to one strategy or the other, many enterprise customers employ what have become known as *hybrid clouds*, which focus primarily on proprietary data centers, but rely on public cloud resources to provide the computing and storage that is needed to protect against unexpected or infrequent increases in demand for computing resources.

In addition, enterprises often stop short of complete hardware virtualization for all applications and instead use cloud computing on a more targeted basis. One classic scenario is *disaster recovery*, in which customers make arrangements to mirror their data in a cloud-based data center and to access the data and computing power to run the applications if the enterprise's internal network or servers should fail. The scalability of cloud computing also makes it well suited to provide overflow capacity to provide insurance against unanticipated spikes in demand (sometimes called *cloud bursting*). Even if such surges are anticipated, cloud computing may nonetheless be a logical strategy if the increase in traffic is sufficiently short lived to render impractical the provisioning of the necessary resources in house, such as occurs during the increasingly popular post-Thanksgiving online shopping event known as Cyber Monday (see Weinman, 2009, for an interesting analysis of the relevant tradeoffs).

III. THE ECONOMICS OF CLOUD COMPUTING

The basic cloud computing concepts that are described above have important implications for the potential economic benefits that are associated with cloud computing.[3] A closer analysis reveals that some of the considerations that are often cited as supporting cloud computing (such as scale economies and converting capital expenditures into operating expenditures) may be less compelling than they initially seem. Instead, the primary advantages are the result of the benefits of aggregating demand.

A. Cost Reductions/Amortization of Fixed Costs

Perhaps the most frequently cited benefit of cloud computing is its ability to reduce cost. End users who shift applications that used to reside on their desktops into the cloud face

[3] My views about the economics of cloud computing are heavily influenced by Weinman (2008), as well as subsequent conversations with him.

considerably lower maintenance costs, as they no longer need to bear the expense of ensuring that the applications run properly. Running software in the cloud also relieves end users of the responsibility for making sure that the software is the most recent version and ensures that all such updates are deployed quickly and easily.[4] The more limited demands placed on the clients that are operating at the edge of the network can also slow down the replacement cycle for end user devices. Cloud computing also obviates the need for enterprises to maintain server and storage capacity.

On the data center side, people often argue that cloud computing facilitates the realization of scale economies by allowing greater amortization of fixed costs. Sharing machines with other users can allow small- and medium-sized firms whose operations are too small by themselves to support a single data center to aggregate their demand with others' demands to achieve the minimum efficient scale that is needed to run data centers in a cost effective manner. Cloud computing also allows the expertise that is needed to run data centers and to update applications to be amortized over a larger number of customers. One must be careful not to overstate the extent to which cloud computing permits firms to realize scale economies. Large enterprises whose computing needs are already large enough to support a single data center are typically able to realize these efficiencies even in the absence of cloud computing.

Those who compare the cost of cloud computing to the cost of creating a proprietary data center that is associated with a private cloud must bear in mind one key difference in their cost structures. The costs that are associated with creating a proprietary data center must be incurred up front and must be sunk regardless of whether the capacity is actually used. When demand is uncertain, a simple comparison of the projected unit cost of a proprietary data center with the incremental cost of cloud computing services can be misleading should the projected levels of

[4] Note that end–users that run PaaS or IaaS instead of SaaS will continue to have to bear these costs.

demand fail to materialize. As a result, any cost comparison must include a risk premium to reflect the possibility that the capacity may not be fully utilized. Indeed, reliance on public clouds becomes more attractive as flows become increasingly unpredictable.

B. The Transformation of Capital Expenditures into Operating Expenditures

Another often touted benefit of cloud computing is its ability to convert capital expenses (CapEx) into operating expenses (OpEx). Although this argument bears considerable intuitive appeal, whether it makes sense in any particular case depends on each firm's individual characteristics (Weinman, 2008). For example, avoiding CapEx improves companies' cash flow positions by reducing their need to make up-front investments in assets that will not generate compensating revenues until later years. Although reductions in the need for cash are generally beneficial, they are less so for companies whose ongoing operations already generate enough free cash flow to cover these capital investments.

Conversely, capitalizing expenses can make an income statement look more attractive by allowing the expenses that are paid during the current year to be transferred into future years (depending on how quickly the applicable depreciation schedule calls for the invested capital to be expensed). Moreover, whether a particular investment will meet a company's internal hurdle for return on invested capital depends in no small part on that company's current cost of capital. The higher is the cost of capital, the more attractive is the transformation of capital expenses into operating expenses.

These considerations can make the benefits from turning CapEx into OpEx less compelling than may appear at first glance. Whether the avoidance of such investments is ultimately beneficial depends not only on the cost savings, but also on any particular company's current financial position.

C. Aggregation of Demand

More compelling than cost reduction or the transformation of capital expenditures into operating expenditures are the potential benefits from aggregating demand. As a general matter, enabling multiple end users to share equipment allows higher utilization of the underlying hardware.

The rationale underlying this insight is straightforward: A basic principle of statistics is that the sum of the variance of two numbers is always greater than or equal to the variance of the sum of those numbers. Stated slightly more formally: $(\sigma_{1+2})^2 = (\sigma_1)^2 + (\sigma_2)^2 + 2\rho\sigma_1\sigma_2$, where $(\sigma_{1+2})^2$ represents the variance of a bundle of goods 1 and 2, and $(\sigma_1)^2$ and $(\sigma_2)^2$ represent the variance of each component. Since $(\sigma_1 + \sigma_2)^2 = (\sigma_1)^2 + (\sigma_2)^2 + 2\sigma_1\sigma_2$, this implies that $\sigma_{1+2} \leq \sigma_1 + \sigma_2$ for all $\rho \leq 1$. So long as the demands for the components are not perfectly correlated, the standard deviation of the bundle will be less than the sum of the standard deviations of the components (Schmalensee, 1984). The reduction in variability that is associated with the aggregation of demand does not directly lead to cost reductions, since operating costs are often negligible. Instead, the primary cost derives from the fixed cost of providing sufficient capacity to handle peak demand. Nonetheless, unless all of the peaks are perfectly correlated, the aggregation of demand should reduce peak variability as well as overall variability (Weinman, 2011a).

The reduction in the variability that results from aggregating demand helps cloud computing firms achieve higher utilization rates than can individual companies achieve on their own. The fact that firms in different industries systematically see their traffic peak at different times makes it possible for combinations of firms to increase the efficiency of hardware utilization, as compared with what would be possible on their own. Moreover, the scalability of

cloud computing allows end users to guard against the possibilities that demand might either be unexpectedly high or unexpectedly low.

Although the benefits from aggregating demand are theoretically inexhaustible, they are subject to the principle of diminishing marginal returns. Given the risk premium that cloud computing firms must charge, it is quite possible that a large firm may find its own traffic large enough by itself to achieve the necessary reduction in peak variability. Nevertheless, the fact that the flows that are generated by a single firm is likely to display a degree of correlation limits the speed with which individual firms realize the benefits that are associated with aggregating demand. In any event, firms that fall below the size that is necessary to reach the relevant point of diminishing marginal returns will find that cloud computing firms are in a better position to bear the risk that is inherent in the uncertainty of traffic flows than are the smaller firms. This is particularly important because of the indivisibilities and long lead times that are associated with expansions of data center capacity.

D. Increased Reliability

The manner in which virtualization permits the same data to be hosted at multiple data centers greatly enhances the reliability of public cloud solutions. A single data center with reliability r faces a failure rate of $(1 - r)$. So long as the likelihood of failure rates is independent across data centers, simultaneously hosting data across n data centers causes the failure rate to drop to $(1 - r)^n$. To use a specific example, if individual data centers have reliability of two nines (99%), storing those same data across two data centers increases reliability to four nines (99.99%). Mirroring the data in a third data center increases reliability to six nines (99.9999%) (Weinman, 2008). Even if some correlation exists, operating multiple data centers should cause some net increase in reliability unless the failures are perfectly correlated.

E. Reduced Latency

Aggregating traffic with other firms also provides firms with a cost-effective way to reduce latency. Consider the simplest case, in which latency is entirely a function of the proximity of a data center. The agility of modern cloud computing can reduce latency by continuously migrating data to the data center that is located closest to where it is currently being demanded.

Sharing data center capacity with other end users also makes it easier for cloud computing customers to reduce latency by increasing the number of data centers. In a one dimensional space, one needs only double the number of data centers in order to halve the distance between them. In a two-dimensional space, however, significantly larger numbers of additional data centers are required to reduce distances by one half. If the data centers are currently laid out in a square grid, halving the distances would require doubling the number of data centers along both axes. In effect, cutting the latency in half would require quadrupling the number of data centers. Indeed, this analysis can be generalized to conclude that a k dimensional space would require 2^k more data centers in order to halve the distance between them. There are other factors expressed in other terms of the occasion, and the precise numbers change somewhat depending on the precise configuration, but the basic intuitions remain (Weinman, 2011b).

The nonlinear relationship between distance and the number of locations represents another source of economies of scale that further increase the benefits of cloud based solutions. This conclusion is, of course, subject to a number of caveats. In many cases, network latency is more the result of queues in the routers than the distances between them. Nevertheless, in an increasingly global world in which traffic often travels long distances, industry participants have

found that locating data centers closer to end users plays a key role in reducing latencies and in improving the economics of negotiating low-cost interconnection agreements (Yoo 2010).

F. Other Benefits

Cloud computing provides numerous other benefits that bear mentioning: Enabling individual end users to access data from any Internet-enabled location facilitates remote and mobile access and makes it easier for end users to remain productive. It also facilitates collaboration between multiple end-users. Storing data in locations other than laptops and desktop PCs protects against certain types of security and privacy threats (although, as discussed below, it raises other types of concerns at the same time).

Relying on cloud services may yield management benefits as well. Aggregating demand across multiple companies may allow data centers to realize the benefits of learning by doing more quickly. Seeing data from multiple sources may allow data centers to recognize patterns in data in ways that will allow them to identify threats and to improve the performance of certain applications. It also provides greater leverage over the overhead costs of keeping abreast of the latest technological developments. Lastly, by reducing the need for corporate management to devote time and energy to overseeing the company's IT resources, cloud computing can conserve on one of the dearest resources in any organization: management focus.

IV. Architectural Implications

For all of cloud computing's potential advantages, it also possesses several potential disadvantages. Whether the advantages outweigh the disadvantages depend largely on the configuration of the access and core networking functions on which cloud computing depends. The underlying architecture can be disaggregated into two separate components: The first is

with respect to the access network through which end users connect to data centers. The second is with respect to the links that connect data centers to one another.

A. Access Networking

As an initial matter, cloud computing will place considerably more onerous demands on the access networks through which end users will gain access to the cloud. The access networks' ability to meet these demands will go a long way in determining cloud computing's attractiveness as an option.

1. Bandwidth

Cloud computing is likely to increase the demands that are placed on the local access network. As an initial matter, new cloud computing customers must have some means for uploading their data to the data centers when setting up new applications. At this point, however, the access network does not have sufficient bandwidth to support this level of utilization. Because datasets in the terabyte range would take weeks to upload, cloud computing providers currently recommend that customers download their data onto a physical storage medium and send it via an overnight mail service (Brodkin, 2010). Eventually, the hope is that network capacity will increase to the point where large datasets can be provisioned through the network itself. Even after data has been provisioned to a new cloud computing facility, the fact that processing that used to occur locally is now being performed in the data center typically means that a greater volume of traffic must pass to and from the client that is being operated by the end user.

2. Reliability

A related concern is access network reliability. The availability of an access network connection is meaningless if it is not functioning properly. Even when the application and the data reside on the end user's hard disk, the absence of a network connection can severely limit the end user's ability to perform productive work. Network failure becomes an even more insuperable obstacle when these elements are hosted in the cloud. Indeed, Gmail, Salesforce.com, and Amazon's S3 (Simple Storage Service) and EC2 (Elastic Compute Cloud) have suffered from well-publicized service outages that had severely adverse effects on their customers. The higher stakes mean that some customers are likely to demand that access networks offer higher levels of guaranteed uptime.

3. Quality of Service/Network Management

End users' willingness to offload services that used to be provided locally into the cloud depends in no small part on how quickly the cloud is able to perform those functions. As a result, cloud computing customers are likely to insist on service level agreements that guarantee them certain minimum levels of quality of service. These demands will likely vary from company to company. For example, financial services companies typically require perfect transactions with latency guarantees measured in microseconds. In addition, these companies will require the cloud provider to verify the delivery time of every transaction after the fact.

One way that cloud computing systems can improve the quality of service of network services is by taking advantage of the presence of multiple connections between two points. The Internet currently relies on protocols such as the Border Gateway Protocol (BGP) to determine the route that any particular stream of packets may take between domains. BGP is limited in its

ability to manage multiple paths, routing all traffic along a single route instead of balancing traffic across both paths. BGP, moreover, is controlled by the core routers rather than by end-users. A new architecture for cloud computing could improve network performance by providing greater ability to allocate traffic across multiple paths and to allow faster recovery from congestion and network failure. It could also increase functionality by giving end-users control over the particular routes taken by their traffic. Cloud computing is also likely to require sophisticated network management techniques to provide minimum levels of quality of service.

4. Ubiquity

For mobile users, cloud computing requires a higher degree of ubiquity than traditional computing solutions. When the software and the data needed to run a particular application reside on the end user's hard disk, the unavailability of a network connection may inconvenience the end user and reduce the application's functionality, but it does not necessarily stop that end user from being productive. When the software and data reside in the cloud, however, the absence of a network connection has more serious consequences, effectively preventing end users from running the application at all. As a result, cloud computing customers regard ubiquitous access to network connections as critical.

5. Privacy and Security

Cloud computing necessarily requires large amounts of data that previously did not leave a company's internal network to be transported via the access network. The fact that this data must pass outside the company's firewall and through the access network renders it vulnerable to attack vectors that are different from those that plague corporate campuses. Moreover, the law holds all institutions that maintain health and educational records responsible for maintaining

their privacy. The fact that such records are now housed in the cloud does not obviate those responsibilities.

As a result, cloud-based solutions must be able to assure these institutions that their data are being handled in a way that preserves confidentiality by giving end users greater ability to control the manner in which their traffic passes through access networks. In addition, cloud computing may require an architecture that permits the exact routes that particular traffic takes to be auditable and verifiable after the fact.

B. Data Center Interconnectivity

The advent of cloud computing will change the nature of demand that is placed on data centers. In addition to heightening data centers' seemingly unquenchable thirst for electric power, cloud computing is placing ever increasing demands on the ways that data centers are configured and on the links that connect data centers to one another.

1. High Bandwidth Networking

The need to augment computing resources on demand requires the ability to move large amounts of data between data centers very quickly. These demands are heightened still further by the needs of virtualization, which depends on the ability to knit together instances operating on several different machines into a coherent whole. As a result, cloud computing will require that all data centers be linked by high capacity connections. In addition, these connections are likely to employ technologies that are able to guarantee a higher level of quality of service than the level that is enabled by the Internet's current best efforts architecture.

2. Reliability

Just as cloud computing requires greater reliability from the access network, hosting the software and the data needed to run application in the cloud also requires greater reliability in the data centers. As a result, data center operations will need a high level of redundancy, both in terms of computing power and in terms of the interconnections between servers. In addition, because cloud computing often requires that a particular application be run in parallel processes that operate on multiple servers, the system must be "self healing" in that it must be able to tolerate and recover from the failure of one of those servers.

The difficulty of the management problem is heightened by the fact that tier 1 Internet service providers (ISPs) can now support data forwarding rates that exceed the processing speed of a single CPU. This means that the functions that the network regards as being performed by a single router will actually be performed by multiple chassis, each with multiple line cards, forwarding processors, and control processors. As a result, what appears to the network as a single router is actually a large distributed system. This federated approach to routing is driven not only by the realities of network engineering, but also by the product of the security and reliability demands of cloud computing, which require both scalability and redundancy. The complex systems must also be able to process upgrades and configuration changes seamlessly.

3. Security and Privacy

As noted earlier, cloud computing necessarily requires large amounts of data that used to reside on an end-user's hard disk or behind a corporate customer's firewall to reside instead on a server in a data center. Moreover, virtualization envisions that these data will often be located on the same servers as other companies' data, including those of channel partners and competitors.

As a result, the hardware located in these data centers and the networks interconnecting them require guarantees that other companies will not gain access to their data.

In addition, customers are likely to require assurance that failure of one virtual machine operating on a server will not compromise other processes operating on the same server. Industry participants are also often very protective of information about the volume and pattern of their transactions. They are thus likely to impose stringent requirements on what data can be collected about their operations and how those data are used.

4. Control over Routing Policies

Cloud computing is also placing new demands on the network's approach to routing. As noted earlier, the BGP-based system responsible for routing traffic on the current Internet employs an algorithm that by default sends traffic along the path that transverses the fewest autonomous systems. The Internet's protocols do not provide any basis for verifying a packet's source or the particular path that it traversed.

Most cloud computing providers need greater control over the paths that are taken by traffic that passes between data centers. As a result, many rely on MultiProtocol Label Switching (MPLS) or some other protocol to exercise control over the precise paths that are taken by particular traffic. Such control mechanisms are essential to ensuring that flows between data centers maintain the required levels of quality of service, protect network security, and maintain the privacy of end users' data.

The fact that data may be shifted from one data center to another also potentially makes those data subject to another jurisdiction's privacy laws. Because customers are ultimately responsible for any such violations, they are likely to insist on a significant degree of control over where their data reside at any particular moment.

5. Standardization

The approach discussed so far implicitly presumes that all of the virtualization of particular cloud computing application will be performed by a single provider. There is no reason that this has to be the case as a theoretical matter. It is quite possible that multiple cloud computing companies could work together to provide a single cloud computing application. For providers to be able to interoperate with one another, the industry would have to develop standards under which different cloud computing providers can exchange traffic and jointly interact with data as well as protocols for joint coordination and control.

In addition to allowing multiple providers to provision a single application, the emergence of standards could also permit the integration different applications provided by different providers. Instead of relying entirely on a single application, cloud computing could integrate multiple applications provided by multiple sources and integrate them on a dynamic basis. Such cross-application coordination would depend on the availability of standards to govern the interactions among these applications.

6. Metering and Payment

One of the primary advantages of cloud computing is that it permits the provisioning of computing resources on demand on a pay-as-you-go basis. As a result, cloud computing requires some means for metering resource usage. In so doing, cloud computing inherently supports commercial deployment to an extent that the project-oriented approach that is associated with grid computing was never able to achieve. Moreover, the fact that different customers require different levels of quality of service means that such prices are likely to vary widely. Any cloud

computing system must provide the basis for setting prices, monitoring and accounting for activity levels, and obtaining the appropriate amount of monetary compensation.

V. POLICY IMPLICATIONS

The future success of cloud computing thus depends on the development of new architectural features. These architectural features also have policy implications, both in terms of the likely industry structure as well as their interaction with regulation.

A. Industry Structure

Perhaps the most important structural impact of cloud computing is to change the roles of key industry participants. Cloud computing's distinctive architecture will give new prominence to certain technologies, while simultaneously downplaying the importance of others.

1. Data Centers

In general, the market for data centers is likely to remain quite competitive. As noted earlier, the minimum efficient scale for data centers is sufficiently small that many enterprises generate sufficient volume to deploy their own data centers cost effectively. The ready availability of a bypass option through private clouds effectively limits the market power of cloud computing providers.

At the same time, factors exist that may cut in the opposite direction. As noted earlier, data center operations may require certain levels of scale if they are to bear risk efficiently, aggregate the information necessary to optimize applications, and support the optimal number of data centers to support reliability and to minimize latency. In addition, data centers may incorporate proprietary technologies that may further concentrate the market.

2. Server-Related Technologies

As an initial matter, the emergence of cloud computing shifts the key role of organizing computing resources away from the operating system on the end user's device to the hypervisor that is responsible for setting up and reallocating virtual machines. The industry structure of the hypervisor market is quite different from that of the operating system market, in which Microsoft has long been dominant. VMWare launched its product in 2001 and used its head start to establish a dominant position. Later entrants, such as Citrix's XenServer (launched in 2007) and Microsoft's Hyper-V (launched in 2008), have made substantial gains in small and medium sized enterprises that are only now in the process of virtualizing, but have struggled to dislodge VMWare's position with large companies. In March 2010, Citrix and Microsoft entered into a partnership that was designed to help them compete more effectively with VMWare. However, VMWare's market-leading position has proven remarkably hard to dislodge.

The shift to cloud computing also means that applications providers will place less emphasis on the operating system that runs on individual desktops and place greater focus on the operating system running on the relevant servers. The market for server operating systems is less concentrated, with Linux constituting a more effective alternative to Microsoft in the server market than in the operating system market for desktop computers. At the same time, the shift toward thinner clients reduces the importance of the integration between end user operating systems and devices.

3. Router-Based Technologies

The greater dependence on network functionality that is associated with cloud computing places greater emphasis on the capacity and reliability of routers. The growing complexity of

routers may cause the market for routers, which already encompasses only a small number of participants -- such as Cisco, Alcatel-Lucent, Juniper, Huawei, Ericsson, and Tellabs -- to become even more concentrated.

4. Access Networks

Cloud computing's implications for the market structure of access networks is less clear. By its nature, the access networks that are associated with cloud computing employ dedicated high-speed lines to connect corporate campuses with key network interconnection points. Regulators regard this portion of the network as already being the most open to competition. The increase in traffic that is associated with cloud computing should only make it more so. At the same time, structural factors exist that may provide incumbent telecommunications providers some competitive advantages. As an initial matter, the extent to which ubiquity is needed for reliability or reduced latency clearly favors more established providers.

B. Regulation

The development of the cloud computing industry is also hamstrung by the fact that cloud computing providers face considerable regulatory uncertainty. Cloud computing presumes that providers will provide guaranteed levels of quality of service and offer different levels of reliability and that people will pay for these services. In order to deliver these services, network providers will have to establish redundant connections and use sophisticated network management techniques to implement routing policies that guarantee that the network can satisfy the demands of cloud computing users. To the extent that customers' needs vary, one would expect some customers to pay a premium for higher levels of reliability and network service.

Such pricing differentials for prioritized service implicate the debate over network neutrality that has dominated telecommunications policy for the past five years. Particularly troublesome is the proposal floated by the Federal Communications Commission during the summer of 2010 to reclassify the transport component of broadband access services as telecommunications services even when they are not offered on a standalone basis. Indeed, some cloud computing providers report that concerns about being subject to telephone-style regulation has led them to forgo providing their own data center interconnections and instead to outsource those services.

On a broader note, the analogy between cloud computing and public utilities should not be taken too far. History has shown that public utility regulation is best suited when the product being regulated is relatively uniform, the primary network elements have already been deployed and thus do not require additional investment, and technologies and market shares are stable. As such, it is poorly suited to environments like cloud computing, in which the product varies widely, investment incentives play a critical role, and in which the underlying technology and market positions are in a state of dynamic flux.

As of now, the fact that the cloud computing application programming interfaces (APIs) are largely proprietary limits customers' ability to switch to a different provider should they become dissatisfied with their current provider or if their current provider become insolvent. Standardizing APIs would facilitate data portability and would allow greater reliability by allowing the same functions to be performed by multiple cloud computing providers. It remains to be seen if the problems with the lack of standardization are any worse in the context of cloud computing than in other contexts and if they can be solved without governmental intervention.

Perhaps the greatest challenge facing cloud computing providers is with respect to security and privacy. In addition to the business concerns that are raised by these issues, privacy- and security-related mandates vary widely across jurisdictions. The variability of these requirements subjects end users to different types of potential liability depending on where their data are hosted. As a result, end users are likely to insist on being able to verify where their data have been hosted after the fact and may well insist on a degree of ex ante control over where their data are hosted.

VI. CONCLUSION

Cloud computing holds considerable promise as a transformative technology that can change the very nature of computing. Assessing its relative merits requires a clear understanding of its key concepts and its underlying economics. What is perhaps most interesting is that the primary economic benefits of cloud computing do not come from the scale economies from amortizing fixed costs over large volumes. Instead, the primary benefits result from the reduction in variability resulting from the aggregation of demand.

Cloud computing also requires networks, data centers, and routers that are more ubiquitous, reliable, efficient, and secure. It also demands operational support for controlling and verifying the routes that particular traffic takes as well as for setting and charging prices. Satisfying these new demands will require major changes to the architecture, both in terms of access networking and in terms of data center connectivity. It also implicates important questions of public policy.

Only time will tell if cloud computing turns out to be the transformative technology that many predict. In the meantime, both the engineering and the policymaking community should

take steps to ensure that industry participants have the tools and the latitude to explore cloud computing's full potential.

REFERENCES

Birman, K., Chockler, G., & van Renesse, R. (2008). Towards a cloud computing research agenda, http://www.cs.cornell.edu/projects/quicksilver/public_pdfs/SIGACT2.pdf.

Brodkin, J. (2010, June 10). Amazon cloud uses FedEx instead of the Internet to ship data, Network World, http://www.networkworld.com/news/2010/061010-amazon-cloud-fedex.html.

Buyya, R., Yeo, C., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems, 25, 599–616.

Carr, N. (2008). The big switch. (New York: W.W. Norton).

Foster, I., Zhao, Y., Raicu, I., & Lu, S. (2008). Cloud computing and grid computing 360-degree compared. (In Proceedings grid computing environments workshop: GCE 2008 (pp. 1–10)). DOI 10.1109/GCE.2008.4738445.

Geelan, J. (2009, January 24). Twenty one experts define cloud computing. Cloud Computing Journal, http://cloudcomputing.sys-con.com/node/612375.

Johnson, B. (2008, September 29). Cloud computing is a trap, warns GNU founder Richard Stallman. guardian.co.uk, http://www.guardian.co.uk/technology/2008/sep/29/cloud.computing.richard.stallman.

Mell, P. & Grance, T. (2009, October 7). The NIST definition of cloud computing (version 15), http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc.

Schmalensee, R. (1984). Gaussian demand and commodity pricing. Journal of Business, 57, S211-S230.

Vaquero, L., Rodero-Merino, L. Caceres, J, Lindner, M. (2009). A break in the clouds: Toward a cloud definition. ACM SIGCOMM Computer Communication Review, 39, 50–55.

Weinhardt, C., Anandasivam, A., Blau, & Stösser, J. (2009, March/April). Business models in the service world. IT Pro, 28–33.

Weinman, J. (2008, September 7). The 10 Laws of Cloudonomics. GigaOm, http://gigaom.com/2008/09/07/the-10-laws-of-cloudonomics/.

Weinman, J. (2009, November 30). Mathematical proof of the inevitability of cloud computing, http://cloudonomics.wordpress.com/2009/11/30/mathematical-proof-of-the-inevitability-of-cloud-computing/.

Weinman, J. (2011a, February 27). Smooth operator: The value of demand aggregation, http://www.joeweinman.com/Resources/Joe_Weinman_Smooth_Operator_Demand_Aggregation.pdf.

Weinman, J. (2011b, April 12). As time goes by: The law of cloud response time, http://www.joeweinman.com/Resources/Joe_Weinman_As_Time_Goes_By.pdf.

Yoo, C. (2010). Innovations in the Internet's architecture that challenge the status quo. Journal on Telecommunications and High Technology Law, 8, 79–99.

Yoo, C. (2010). The changing patterns of Internet usage. Federal Communications Law Journal, 63, 67–89.