THE ECONOMICS OF CLOUD COMPUTING

by

Ergin Bayrak, John P. Conley, and Simon Wilkie

Working Paper No. 11-W18

September 2011

DEPARTMENT OF ECONOMICS
VANDERBILT UNIVERSITY
NASHVILLE, TN 37235

www.vanderbilt.edu/econ

The Economics of Cloud Computing[1]

Ergin Bayrak
University of Southern California

John P. Conley
Vanderbilt University

and

Simon Wilkie
University of Southern California

Abstract

Cloud computing brings together existing technologies including service oriented architecture, distributed grid computing, virtualization and broadband networking to provide software, infrastructure, and platforms as services. Under the old IT model, companies built their own server farms designed to meet peak demand using bundled hardware and software solutions. This was time consuming, capital intensive and relatively inflexible. Under the cloud computing model, firms can rent as many virtual machines as they need at any given time, and either design or use off-the-shelf solutions to integrate company-wide data and then easily distribute access to users within and outside of the company firewall. This converts fixed capital costs to variable costs, prevents under or over provisioning, and allows minute by minute flexibly. Consumers are also turning to the cloud for computing service through such applications as Gmail, Pandora, Facebook, and so on. The purpose of this paper is to discuss this new and transformative technology, survey the economic literature, and suggest economic questions that it raises. We find that the literature to date is very thin and fails to address many of the issues raised.

[1]The authors take full responsibility for any errors and may be contacted at ebayrak@usc.edu, j.p.conley@vanderbilt.edu, and swilkie@usc.edu, respectively.
Keywords: Cloud Computing, SaaS, PaaS. IaaS, Economics. Information Technology
JEL Categories: D4, L5, D1, L1,

1. Introduction

Cloud computing is a newly emerging computing paradigm in which computing resources such as software, processing power and data storage are provisioned as on-demand services over broadband networks. Cloud computing enables a shift away from computing as a bundled hardware and software product that is acquired through fixed capital investments, to computing as a location independent and highly scalable service that is acquired on-demand over broadband networks from large-scale computing centers or "clouds" on a pay-per-use basis with little or no fixed capital investment. Moreover, from the supply perspective, economies of scale, distribution of costs among a large pool of users, centralization of infrastructures in areas with lower costs, and improved resource utilization contributes to value creation and efficiency improvements enabled by cloud computing. With these efficiency improvements and large savings in operational costs as well as upfront capital costs of tech-startups, cloud computing carries the characteristics of a disruptive general purpose technology with a potential to greatly impact the economy as a whole.

Although relatively new, cloud computing is already a very significant part of the technology sector. A recent report by IT research and advisory firm Gartner forecasts worldwide cloud services market's revenue to surpass $68.3 billion in 2010 and reach $148.8 billion by 2014[2]. Another 2009 report by IT research and advisory firm IDC predicts worldwide IT spending on cloud services to reach $42 billion by 2012. More broadly, a recent study by Etro (2009) treats cloud computing as a general purpose technology and estimates a conservative[3] 1% to 5% fixed cost reduction across all sectors and estimates the prospective medium term macroeconomic impact of the fast (slow) adoption of cloud computing in 27 European countries to be an incremental GDP growth of 0.3% (0.1%), an increase in

[2] In Gartner's methodology, $32 billion out of the $68.3 billion is accounted under Business Process Services which include online advertising; they are likely to be overestimating the impact of cloud computing. See Gartner, Inc., (2010)

[3] Considering the case study by Khajeh-Hosseini et al. (2010), which found that moving to cloud infrastructure would result in 37% cost saving over 5 years, Etro's estimates of the cost reduction, and the resulting macroeconomic impact remain conservative.

employment on the order of 1.5 (0.5) million workers and incremental business creation on the order of 430000 (83000) small and medium enterprises.

Despite the social and economic significance of cloud computing very little was been written in the economics literature that directly on topic. Most papers seem to appear in various parts of the computer science and to a lesser extent, the formal and informal business literature. Our purpose with this paper is three fold. First is to discuss the basics of what cloud computing is with special attention to its economic implications. Second to survey cloud computing literature as it relates to economic questions. Since very few papers have appeared in economics journals, we extend the survey to include some of the more relevant CS and management literature as well and older economics papers that help us understand cloud computing. Finally, we explore open research questions and how economics might contribute to our understanding of this new and important technology. In section two we review various definitions of cloud computing. In section three we contrast cloud computing with prior general-purpose technologies from an economic perspective and identify characteristics that give cloud computing a distinct economic structure. We also identify policy issues in the cloud computing ecosystem. Section four proposes further research opportunities in the economics of cloud computing. Section five concludes.

2. What is Cloud Computing?

Despite the wide consensus that cloud computing is a new and disruptive general purpose technology, there does not seem to be a comparable consensus in defining what exactly cloud computing is, nor a common understanding of how it affects economic fundamentals. This may be due to the large scope of this new technology as well as its complex, and multi-layered, technical and economic underpinnings. Many definitions and analogies for cloud computing fail to capture this complexity and remain overly simplified such as "moving computer applications and programs to the Internet from the desktops". Other definitions are well received in the computer science literature but remain to be overly

technical and so less useful to economists and other outsides. A collection of twenty two such definitions are summarized in Vaquero et al. (2009)[4] as:

> Clouds are a large pool of easily usable and accessible virtualized resources (such as hardware, development platforms and/or services). These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a pay-per-use model, in which guarantees are offered by the Infrastructure Provider by means of customized Service Level Agreements (SLA).

An evolving definition maintained by United States National Institute of Standards and Technology (NIST) seems to be the most comprehensive and widely accepted definition of cloud computing (Mell and Grance (2009))[5]. The current and 15th revision defines cloud computing as:

> Cloud computing is a model for enabling 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.

There are three widely accepted primary service models where different capabilities enabled by the cloud architecture are utilized to provide different types of services. The term Everything as a Service (XaaS) has been associated with many core services offered by cloud providers, including software (SaaS), development platforms (PaaS), computing infrastructure (IaaS), communication (CaaS), and data storage (DaaS), however, the most widely accepted classifications of service models focus on software, platform and infrastructure layers as the primary service models (Creeger (2009), Durkee (2010), Lin et al. (2009), Mell & Grance (2009), Viega (2009), Vaguero et al. (2009), Weinhardt et al.

[4] Vaquero et al. (2009), in their analysis of 22 expert definitions, also identified ten key characteristics of cloud computing as: user friendliness, virtualization, internet centric, variety of resources, automatic adaptation, scalability, resource optimization, pay per use, service SLAs and infrastructure SLAs. Also, Iyer and Henderson (2010) analyzed over 50 definitions and identified seven key capabilities offered by cloud computing to be: controlled interface, location independence, sourcing independence, ubiquitous access, virtual business environments, addressability and traceability and rapid elasticity.

[5] The complete definition of cloud computing according to NIST can be found in the appendix.

(2009)). The service models resemble a cascading architecture where services on a higher level, as identified by Weinhardt et.al. (2009); "encapsulate functionality from the layers beneath by aggregating and extending service components via composition and mash-up technologies". Similarly, Yousseff and Da Silva (2008) use composability as the criterion to define various layers in the cloud architecture and inter-relations between those layers. They classify a cloud layer to be higher in the architecture (ore at a higher level of abstraction) if its services can be composed from the services of the underlying layer. The three primary cloud computing service models are:

Software as a Service (SaaS) is the service model where the capability provided to the consumer is the ability to use the cloud provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface such as a web browser. This highest layer in the cloud infrastructure is called the Cloud Application Layer in Youseff and Da Silva's (2008) terminology. The applications at this layer are composed from components at the lower layers such as application programming languages, operating systems, network, servers, and storage that the end user does not need to control or manage except limited user-specific application configuration settings. Word processing and email applications such as GoogleDocs and Gmail or Customer Relationship Management (CRM) applications of salesforce.com are examples of this service model as well as backup, recovery and to some extent content delivery and publishing services.

Platform as a Service (PaaS) is the service model where the capability provided to the consumer is a development or runtime environment, programming languages and application programming interfaces (APIs) to design and deploy infrastructure consumer-created applications onto the cloud. This service model resides in the Cloud Software Environment Layer in Youseff and Da Silva's (2008) classification where the consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly

application hosting environment configurations. Examples of these scalable services include Google App Engine, Microsoft Azure or force.com application development platforms.

Infrastructure as a Service (IaaS) is the service model where the capability provided to the consumer is processing, storage, networks, and other fundamental computing resources. This model resides in the Cloud Software Infrastructure Layer in Youseff and Da Silva's (2008) terminology. The consumer is able to bypass the platform layer and the API restrictions therein and is able to run arbitrary software, operating systems and applications. The consumer does not manage or control the underlying cloud hardware but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components. These capabilities are delivered as a single server or as part of a collection of servers integrated into virtual machine environments. In some cases a further disaggregated service provision is possible whereby processing power in the form of virtualized machines is treated as (IaaS), along with communication (CaaS), data storage (DaaS), which can be rented separately (Rimal et.al. (2009), Yousseff and Da Silva (2008)). Amazon's Elastic Computing Cloud (EC2) is an example of IaaS along with the Simple Storage Service (S3) that can be rented separately as DaaS. Microsoft's Connected Service Framework (CSF) on the other hand accompanies its IaaS services as a CaaS component that can be rented separately.

We would like to suggest that while these service models are useful, especially in defining the technological differences between various layers of abstraction, they are not as useful in categorizing their economic impacts. Instead we would like to distinguish between a *retail* and *wholesale* side of cloud computing and argue that these are two qualitatively different uses for the cloud. IaaS ans PaaS consumers are mostly companies using the cloud to outsource their internal IT functions or provide client facing applications including SaaS services. SaaS consumers on the other hand are mostly individuals moving their data (mostly email and networking), simple computation needs (word processing and spreadsheets), and content and entertainment consumption (using Pandora, Hulu, and

World of Warcraft, instead of buying CD's, DVD's or software) to the cloud[6]. Although there are there are certainly significant crossovers where individuals use PaaS or IaaS to run websites on virtualized servers or large companies use SaaS to outsource some functions such as CRM, the qualitative distinction between the retail and wholesale nature of SaaS and PaaS/IaaS remains. We will discuss this in more detail in subsequent sections.

From cloud providers' perspective, these three service models can be set up according to four major deployment strategies. A *private cloud* is the deployment strategy in which the cloud infrastructure is operated solely for a single or organization. A c*ommunity cloud* is similar, but the cloud infrastructure is shared by several organizations and so supports a specific community that has shared interests. These cloud infrastructures may be managed by the organizations themselves or a third party and may exist on premise or off premise. A *public cloud* involves the cloud infrastructure being made available to the general public or a large industry group and is typically deployed by a separate organization selling cloud services. Finally, a *hybrid cloud* is a composition of two or more deployment models that are bound together by a technology that enables data and application portability.

3. A Discussion of the Literature Related to the Economics of Cloud Computing.

Cloud computing started to shape up in the last decade as a result of the convergence of several earlier technologies and IT operating models. From a technical point of view cloud computing is enabled by the culmination of virtualization[7], cluster computing, grid computing, broadband networking and large scale data centers centralized at low cost locations. The development of service-oriented software architectures for creating business processes packaged as services along with service level agreements that contract delivery

[6] Of course, SaaS providers are often consumers of lower level PaaS and IaaS services, if not already vertically integrated to PaaS and IaaS providers.

[7] Virtualization refers to the creation of a virtual machines that are separated from the underlying hardware resources but act like real computers. It enables running multiple operating systems and sets of software on the same hardware.

time or performance of such services further enabled the provision of computing as a service. Recent financial crisis and recessions have also contributed to the accelerating adoption of cloud computing as companies have been forced to find cost-effective IT solutions which is reflected in a IDC (2009) survey of 696 IT executives and CIOs where 51% of respondents report that the key driver behind adopting or considering cloud computing is the need to cut costs, while another 29% report that budget issues force them to find cheaper IT solutions.

Earlier studies in the information technology literature can inform the outlook on cloud computing adoption. For example, a collection of eighteen empirical studies on IT adoption surveyed in Fichman (1992) contribute to the conclusion that IT adoption has different determinants depending on the class of technology in question and locus of adoption. Fichman (1992) uses a class of technology dimension to distinguish type 1 technologies that exhibit a lack of user interdependencies and a lack of a substantial knowledge barriers faced by potential adopters. In contrast, type 2 technologies are characterized by, significant user interdependencies or high knowledge barriers. In this dimension, most cloud offerings such as PaaS and IaaS can be characterized as type 2 technologies that exhibit significant user interdependencies and knowledge barriers for potential adopters, except for simple software as a service application such as word processing, content management and streaming. The second dimension that affects IT adoption according to Fichman (1992) is the locus of adoption, which distinguishes between individual and organizational adoption. Again, with most cloud computing offerings, the locus of adoption is organizational, except for simple SaaS applications that individuals adopt as they move to the cloud. An illustrative classification cloud computing applications according to this framework can be found in appendix 2.

Various empirical studies confirmed that the determinants of classical innovation diffusion dynamics carry over to the personal adoption of type 1 technologies. Firstly, a favorable perception of IT innovation, positively affects the rate and pattern of adoption (Davis et.al.

(1989), Huff and Munro (1989)). Secondly, adopters are differentially influenced both by information channels and sources at various stages of adoption. Furthermore, early adopters tend to be younger, highly educated, involved in mass media and interpersonal communication, and more likely to be opinion leaders (Brancheau and Wetherbe (1990), Leonard-Barton and Deschamps (1988)). Finally, cumulative adoption follows an S-shaped path; starting out slowly among pioneering adopters, reaches "take-off" as a the effects of peer influence set in, and levels-off as the population of potential adopters is exhausted (Brancheau and Wetherbe 1990).

In terms of the organizational adoption of type 1 technologies, in addition to the classical diffusion determinants, Gatignon and Robertson (1990) found that competitive characteristics such as high concentration, high vertical integration of the adopter industry as well as low price intensity, and high incentives in the supplier industry positively correlate with adoption.

According to Leonard-Barton (1987), determinants of individual adoption of type 2 technologies include, besides classical diffusion determinants, adopter's ability in addition to its willingness to adopt. Individuals experienced with IT are found to be more likely to adopt because of a better perception of benefits or their absorptive capacity with respective to innovations Adopter attitudes and preferences, training, accessibility of consulting, and influential peers is also shown to contribute to individual adoption of type 2 technologies.

Finally, in terms of the organizational adoption of type 2 technologies, classical diffusion determinants are again found to be important. The classic S-shaped cumulative adoption pattern was confirmed for this type of adoption by Gurbaxani (1990). Furthermore, Gurbaxani and Mendelson (1990) identified a pattern of cumulative IT adoption at the national level, that starts with and S-shaped path but later followed by an exponential path as a result of price decreases. The BITNET network adoption case they studied is marked by subsidies to early adopters where IBM provided funding for centralized network

management until a critical mass of 200 universities had adopted the network and these subsidies are also found to be a fundamental determinant of adoption.

A recent study by Benlian and Buxmann (2009) on the other hand, considers the adoption of cloud computing by small and medium businesses (SMB), and specifically looks at adoption of software as a service. They empirically study the IT outsourcing and adoption behavior of firms based on a survey of a random selection from 5000 German SMBs, and test 10 hypothesis regarding the drivers of adoption derived from three different theoretical theoretical foundations. Firstly, they turn to transaction cost theory that posit that transactions with high asset specificity are managed less expensively in-house, while the rest should be outsourced for better efficiency. This leads them to their first hypothesis that application specificity is negatively associated with SaaS-adoption, which interestingly is *not* confirmed in the data. Their second transaction cost theory hypothesis is that adoption uncertainty is negatively associated with SaaS-adoption . When business and technology driven uncertainty is high, it is difficult to enforce and monitor performance, inducing the adopter to prefer internal governance for highly uncertain activities. This hypothesis finds support in the data and is confirmed with high significance.

The second set of hypothesis stem from a resource based view framework, which suggests that each firm can be though of as a "bundle of resources," and that these resources are heterogeneously distributed across firms which differentiates them enabling them to successfully compete against others. Two hypotheses that stem from this view are that the application's (i) strategic value and (ii) inimitability are both negatively associated with SaaS-adoption. Despite the correct prediction of the direction of impact in both hypotheses, only the first one finds significant support in the data.

Finally, Benlian and Buxmann (2009) turn to the theory of planned behavior which posits that the cognitive process guiding individual decision making are influenced by the environment as well as decision maker's perception of the environment and ultimately,

individual's intention to perform a certain action is influenced by two main factors: the attitude towards the behavior and the subjective norm. In the context of SaaS adoption, they define attitude toward SaaS-adoption to be the overall evaluative appraisal of an IS executive toward having an IT application provided by a SaaS-provider. Subjective norm or social influence on the other hand denotes the perceived social pressure to perform or not perform the behavior. They derive seven hypotheses from the theory of planned behavior and the two that find support in the data are that, (i) Attitude toward SaaS-adoption, and (ii) Subjective norm (i. e. positive opinion of influential others toward SaaS) are positively associated with SaaS-adoption.

The accelerating transformation of computing to a service oriented model has been envisioned since 1960s and has been likened to public utilities such as the electricity or telephone system[8]. For instance, Martin Greenberger has posited as early as 1964 that; "Barring unforeseen obstacles, an on-line interactive computer service, provided commercially by an information utility, may be as commonplace by 2000 AD as telephone service is today."[9]. Carr (2008) compares the emergence of cloud computing to the rise of electric utilities in the early 20th century and along with his vigorous discussion of this analogy, the term "utility computing" has been resurfaced and popularized by many companies that championed the development of computing as a service[10].

The first similarity between cloud computing and traditional utility models such as the electricity or telephony is that they exhibit characteristics of a disruptive general-purpose

[8] John McCarthy was one of the first computer scientists to refer to utility computing in 1961 by stating that; "If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility… The computer utility could become the basis of a new and important industry." His views were echoed by the chief scientist of ARPANET Leonard Kleinrock in 1969 who stated that; "As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of `computer utilities' which, like present electric and telephone utilities, will service individual homes and offices across the country". (Qian (2009), Buyya et. al. (2008))

[9] Martin Greenberger, "The Computers of Tomorrow" (1964)

[10] Microsoft, Sun, HP and IBM use the term utility computing in reference to the computing services they provide. Nicholas Carr's book "The Big Switch" (2008) also vigorously used the public utility analogy to cloud computing and popularized the term

technology with a potential to unleash a wave of combinatorial innovations. However, compared to traditional utility systems the benefits of which took several decades to be fully internalized, cloud computing is likely to have a relatively faster diffusion of economic impacts as long as the rapid development of cloud computing is matched with an equally fast adoption cycle.

A second characteristic of general-purpose technologies that cloud computing exhibits akin to the utility models is the extensive cost savings that they can attain and pass on to the rest of the economy. Cloud providers' cost savings result mainly from economies of scale through statistical multiplexing, virtualization and clustering that enable higher utilization rates of centralized computing resources. Cloud providers also require large real estate for computing centers as well as uninterruptible power and network bandwidth; therefore locating computing centers in rural areas with proximity to the nexus of power grids and broadband networks is another source of cost savings. The labor cost of maintaining the computing infrastructure is also distributed across a greater number of servers in large computing centers that contributes to the cost savings and economies of scale[11].

However, as identified by Brynjolfsson et.al. (2010), cloud computing exhibits numerous other complexities beyond the traditional utility models that lead to a distinct economic structure. First feature of cloud computing that separates it from traditional utility models is the non-fungibility of the computing resources or software applications. Services offered by cloud computing are neither relatively standardized services like telephony nor highly fungible commodities like electricity. Furthermore, there are no simple mechanisms to support interoperability in the case of cloud computing comparable to the rotary converter technology that enabled fungibility of electricity or interconnection rules that enforced interoperability of telephony services. Although interoperability is likely result in faster

[11] According to Hamilton (2008) One system administrator manages about 1,000 servers in a large data center, whereas a system administrator in a medium-sized data center typically manages 140 servers.

adoption, efficiency and social welfare gains, the possibility of cloud providers losing market power is likely to hinder momentum towards interoperability.

So called "Cloud Neutrality" is an emerging policy issue that mirrors the network neutrality debate in the Internet sphere where questions about whether providers should be allowed to offer services on a prioritized or discriminatory basis are being heavily discussed. Open cloud manifesto[12] is an attempt to establish principles and standards to support an open and interoperable cloud system that is supported by over 400 companies. However the list does not include any of the major cloud providers such as Amazon, Google, Microsoft, or Salesforce.

Scalability, which is one of the main benefits offered by cloud computing is neither unbounded nor error free. According to the "CAP Theorem,"[13] consistency, availability and partition tolerance cannot be achieved simultaneously in a distributed system; therefore as an application increases scale in the cloud system, it loses consistency and creates errors. Because of cloud providers' specialization and expertise system failures and unavailability due to such errors are likely to happen less frequently than in the case of a private computing system. Nevertheless, if such systems failures become frequent and public, this may result is rapid abandonment of the cloud provider and subsequent bankruptcy resembling bank runs. Limits to scalability and errors are most disruptive to high-volume computing applications such as transaction processing or engineering computation that rely on large relational databases and require consistency. Instead of relying on outside cloud providers, companies with sufficiently large computing needs may choose to deploy private cloud computing systems to avoid the risk of cloud provider failure. As noted in Brynjolfsson et.al. (2010), Intel is an example of such a company that chooses to deploy a private cloud to support its mission critical engineering computations by consolidating its

[12] Open cloud manifesto and the list of its supporters can be found at http://www.opencloudmanifesto.org
[13] CAP theorem stems from Eric Brewer's 2000 conjecture that a distributed system can not satisfy consistency, availability and partition tolerance simultaneously, later proved by Lynch and Gilbert (2002)

data centers from more than 100 down to 10 and increasing their utilization resulting in significant cost savings[14] along with increased control.

The problem of latency adds a layer of complexity to cloud systems. Relatively lower cost of moving photons over fiber compared to the cost of moving electricity over the grid favors locating computing centers away from the clients and closer to power sources, however, the savings in power cost can be outweighed by the cost of information latency due to distance and network congestion. For instance, the need for instantaneous computing in the case of financial services and trading may dictate that computing be local.

Data security is another concern in cloud computing systems. On the one hand, consumers of cloud services may be apprehensive about trusting outside companies with their private, mission-critical data since much of the measures taken to ensure data security on the part of the cloud provider are obscure to the consumer. Numerous examples of security and privacy breaches most famously by the Facebook platform or the Gmail system create grounds for consumer concern. On the other hand, cloud providers value successful security implementation as one of the most important assets for gaining positive reputation, hence have a strong incentive to maintain successful security practices. Furthermore, security can also benefit from economies of scale; and is likely to be handled more effectively by large cloud providers that have a vested interest in doing so.

Another related problem is the design and implementation of platforms that enable users to combine, share and trade computing resources. Some progress on this front has been made in the Grid Computing sphere where market mechanisms have been proposed and implemented to trade computing resources in the form of virtualized server instances. Altmann et.al. (2010) for example proposes such a market mechanism called GridEcon and contrast it with other market mechanism such as Grid Architecture for Computational

[14] Brynjolfsson et.al. (2010) note that Intel saved about $95million by consolidating its data centers from more than 100 down to 75, with an eventual target of 10. According to Intel's co-CIO Diane Bryant, 85% of Intel's servers support engineering computation, and those servers run at 90% utilization.

Economy (GRACE), Tycoon, Popcorn Project[15]. GridEcon, the details of which are explained in a series of articled collected in Altmann and Klingert (2010), is another mechanism for allocating commoditized computing resources, i.e. standardized virtual machines. It employs a classic double auction design with extensions that allow extra attributes to be offered such as duration (i.e. a time interval), price, urgency (i.e. the point in time when the computing resource is needed), and maximum validity time of the offer. It then matches buy and sell offers with a many-to-many approach with consideration of issues like scalability and not favoring large providers of computation.

The relative fungibility of resources in the grid computing case allow some progress in the design of market mechanisms, and can be useful to inform the design of lower level IaaS that share similarities with grid architecture. However, for the wider cloud computing environment including higher level PaaS and SaaS, nonfungibility of service components is likely to hinder progress in the design of market mechanism.

The literature directly on topic written by economists is extremely small. In addition of Etro (2009), we find Ghodsi, et al. (2010), Friedman, et al. (2011) and Langford et.al. (2009) and we conclude this section by discussing each in turn. These last three papers are applications of cooperative bargaining theory and mechanism design to particular aspects of cloud computing. Surprisingly, there does not seem to be any other empirical literature nor applications of theoretical industrial organization or other methods written by economists on this topic.

[15] GRACE is a generic framework/infrastructure for grid computational economy, details of which can be found at http://www.buyya.com/ecogrid/

Tycoon is a market-based system for managing compute resources in distributed clusters developed by HP. More info can be found at http://tycoon.hpl.hp.com/

The Popcorn project provides an infrastructure for globally distributed computation over the whole internet and a market-based mechanism of trade in CPU time. More info at http://www.cs.huji.ac.il/~popcorn/

Cloud environments are composed of multiple resources, CPU, memory, network bandwidth, etc. The allocation of these limited resources among competing users is a question that has started to receive increased attention in the literature. In cloud systems, many different tasks compete for different resources. Some tasks are CPU-intensive; while others are memory or bandwidth intensive, and most tasks, regardless of their dominant resource requirement, need to use some amount of other resources too. This leads to a situation where cloud users *prioritize* resources differently. An important question is what constitutes a efficient and fair allocation of resources in such situations, and can we come up with an algorithm or decision rule that implements an efficient and fair allocation rule. Ghodsi et.al. (2010) propose dominant resource fairness (DRF) as a criterion to ensure fairness in resource allocation problems and provide an efficient algorithm that implements it. They define dominant resource of a task to be the resource that it percentage-wise demands most of relative to the total resource available, and provide an algorithm that aims at equalizing each user's fractional share of its dominant resource. The also prove that DRF has various desirable properties such as single resource fairness, bottleneck fairness, share guarantee, population monotonicity and envy-freeness. They also show that Kalai-Smorodinsky (KS) solution with celebrated desirable properties corresponds to DRF, and despite the possible existence of envy at KS solutions depending on utility profiles, DRF always results in an envy-free allocation.

Friedman et al. (2011), on the other hand proves that DRF and KS are equivalent and the percentage-wise demands on the part of users produce a lexicographic preference profile that results is envy-free allocations. Furthermore, Friedman (2011) studies allocation problems in cases where users demand a minimum number of virtualized servers above which they are indifferent to the allocation. He shows that this demand structure results in a resource allocation problem over a polymatroid bargaining set. He then shows that the Nash bargaining solution can be achieved with an additional continuity axiom and making use of homotopies of the polymatroid bargaining set.

Bandwidth is another essential cloud resource, which makes traffic management another important problem in the cloud architecture that has received some recent attention. For example, Langford et.al. (2009) develop an admission control (AC) mechanism in an Intranet traffic management context, where urgent and delayable traffic coexist and occupy network capacity. They design an admission control mechanism that collects feedback on utilization of communication links and allocates bandwidth to competing (but delayable) network traffic, whereas urgent traffic automatically bypasses the AC mechanism. In their model, sources request AC for permission to send bits and the AC mechanism grants a nonexclusive lease of bandwidth with a rate limit and a time limit equal to or less than the time interval with which utilization feedback is collected. By injecting delayable traffic into the network with an objective to maintain a uniform flow and maximize network utilization, they show that their admission control mechanism achieves a reduction in capacity that is close to first best. They also run simulations with real world network traffic data and confirm the capacity reducing impact of AC.

4. An Economic Perspective

As we mention above, we see retail/wholesale as a better way to organize an economic analysis of cloud computing than the XaaS structure prevalent in the IT and CS literatures. In this section we outline some of the economic questions that cloud computing raises and suggest how economic analysis might contribute to our understanding this new technology.

On the retail side of cloud computing, we can begin by looking the problem from the perspective of consumer theory. The development of an effective technology to make micropayments for content and services has the potential to revolutionize the web. In practice, however, consumers already do make micropayments indirectly (and in a second best way) by giving up personal information to providers of cloud services in exchange for access. For example, Facebook is an SaaS that provides storage and networking services to its users. Users have a choice of how much information to give to the site, such as one's

email address, physical address, religion, age, relationship status, pictures of you and your friends, links to content you like, content you have created, etc.. The further a user reveals information, the more sharing he can do with his friends. Similarly, when we use Google, our searches are tracked and indexed, Hulu and Netflix create records of our viewing habits, Visa and MasterCard know how much liquor we buy, Ebay, Expedia, and especially Amazon know a great deal about our consumption habits, and many other sites deposit tracking cookies. Content and service providers monetize this data directly by serving ads to users based on this data or by selling or renting the accumulated information on a user to third parties.

This already active exchange provides a basis for developing a theory of the "economics of privacy". As a currency that can be exchanged for services, privacy has some interesting features. First, the property rights are not clearly defined. What can be done with information we post or give to a site depends both on regulations and the terms of service. Since many providers of cloud services, especially social networking sites, have significant network externalities and lock-in effects, consumers are not in a good position to the resist changes in the terms of service which effectively raise the privacy cost of using the service. Thus, there is a case to use regulation to prevent such abusive monopoly practices. Note, however, that even though Facebook can sell a user's information, that user can give the same information to another provider. Thus, Facebook does not "own" our information in a conventional sense. Privacy is like a ten dollar bill that we can use again and again, economists would say that private information is *excludable* - I can refuse to give it to site, but non-rivalrous - I can give it again and again to many sites. However a user may think that once he has spent "ten dollars" worth of privacy at one site, he might as well spend the same ten dollars with anyone who offers anything useful to us. However, if the privacy price is twenty dollars (meaning we have to give up additional information) he would have to think again. Once the information is released, it can never be called back. The cloud never forgets.

A theory of privacy might include the following factors. Not all privacy is equally valuable to service providers. Advertisers would prefer to market to richer people all else equal. This means that the rich are being underpaid and the poor, overpaid for their private information. Providers may try to enrich the quality of the user-base by offering services that selectively appeal to valuable demographics, but we would expect to see the rich under participating in the cloud. If it were possible to certify that certain users were from a valuable demographic, providers could offer them "gold memberships" with enhanced services in exchange for their information. For now, this is a market failure. One might also think about time inconsistent preferences. For example, young people value social networking and other cloud services more than older people, and also value their privacy less. Who cares if an 18 year old posts something stupid or a 22 year old shares an embarrassing picture? No harm done. Obviously, as one gets into one's thirties, these things have more serious consequences, but old postings cannot be called back. Thus, one can think of Facebook as buying low and selling high. Young people who revealed a lot of data, later on are older and join a more valuable demographic from whom is harder to get private information.

Cloud technologies also open many questions in labor economics. Mainly these stem from large scale data integration within firms that cloud infrastructure make possible, and the consequent ability to interact with this data through clients ranging from desktop computers to mobile devices. Many workers can now work at home or on the road. This reduces the capital needs of companies to provide office space. It is even possible for companies to further reduce capital needs by leveraging the use of the employees' computers, cell-phones, cars, and other equipment. Many jobs, customer support, for example, don't need to be done in continuous blocks of time, but workers are needed or can be used at all times of day. This capacity creates flexibility that makes it more feasible for the physically disabled and those with family responsibilities to participate in the labor force. At least for some types of work, the allocation of tasks and monitoring of productivity can be largely

automated by integrated business systems. There is little need to have an employee supervised at a central location by a physical person. Thus, one might empirically study whether we see greater work force participation, especially by women and the disabled in certain industries as cloud technologies get more ubiquitous. From a theorical standpoint, one could explore what kinds of jobs lend themselves to this sort of decentralized system of labor, what kind of incentive and monitoring schemes will be efficient, and what types of workers and businesses will be attracted to these new models. These issues are driven by individual interaction with SaaS for the most part and thus are part of the retail effects of cloud computing.

A phenomenon that SaaS in the cloud greatly facilitates is the collection of free labor that many people provide to all sorts of activities through "crowd sourcing". People write reviews of products at Amazon, Chowhound, and TripAdvisior, post blogs and comments on blogs, volunteer to be moderators at sites they like, add bookmarks to Reddit and Stumbleupon, add content to Wikipedia, contribute code to Linux and other open source software projects. These contributions are valuable and in many cases, the cloud service providers monetize them in various ways[16]. Why and how people make these contributions, whether due to altruism, a desire for reputation, pleasure or other reward, is not well explored in the context cloud computing. See Conley and Kung (2010) and references therein for a theoretical discussion of these motivations for FOSS projects like Linux, Archak and Sundararajan (2009) and Leimeister, Huber, and Bretschneider (2009) for a discussion of efficient crowd sourcing mechanisms, and Kittur, Chi and Suh (2008) for a example of research on Amazon's Mechanical Turk micro-task markets

Turning to the wholesale side of cloud computing, consider that before cloud services were available, businesses had to build private server capacity to be able to meet peak user demand. One of the major reasons cloud service provider can offer a cheaper alternative is

[16]Think of the recent sale of the Huffington post as an example of how this free labor from bloggers is monetized.

that the peak demands of different companies are not perfectly correlated. Thus, private servers typically run at an average of 20% capacity and only rarely hit their peak. Large cloud providers, on the other hand, can provision servers to run at 80% capacity and still provide the needed peak load capacity to the fraction of their customers experiencing demand spikes. To an economist, this is just a version of portfolio optimization. Instead of the classic finance problem of choosing a bundle of stocks based on the price covariance matrix to maximize expected return for any given risk level, one wants to choose customers based on the capacity demand covariance in order to maximize revenue for any given probability of not having sufficient capacity. Thought of in this way, customers with correlated demand spikes (such as retailers who expect web traffic to peak in December) are less desirable unless they can be offset with customers whose peaks are negatively correlated. Customers with high variance are no problem provided the portfolio of such customers is large and so aggregate variance is small. Knowing how to solve this problem would tell cloud companies which customers are cheap and which are expensive (in terms of the percentage of excess capacity needed to serve them at the level specified in any given Service Level Agreement). This might lead to cloud companies attempting to price discriminate in favor of, directly advertising to , or offering value added services of interest only to customer types who are cheap to serve.

Cloud computing resources also make feasible large-scale data integration projects. This involves putting together the data and software systems for a company's payroll, sales, purchasing, shipping, customer support, and other functions in one unified package. The goal is to create one interoperable system to better manage and control inventory, work flow and quality, human resources management and so on. Unfortunately implementing such Master Data Management (MDM) and Enterprise Systems Management (ESM) solutions is quite difficult. Most go over budget, take longer than anticipated and do not meet expectations. In fact estimates based on surveys of enterprises suggest that 30-70% of

information technology projects are judged internally to be failures.[17] These failures extend from a variety of sources, but lack of executive leadership, failure to get buy-in from stakeholders, lack of technological expertise, underfunding, and the existence of too many legacy systems are often highlighted.

The rewards to successfully implementing ESM projects in the cloud are significant, though some are hard to document. Lower operating costs due to better management of inventory and human resources, better sales through use of Customer Relations Management (CRM) software, and the uses of data mining and business intelligence systems, and especially the flexibility and scalability that such cloud based solutions provide are most often cited.

Clearly undertaking an ESM project is risky but also has significant rewards. The degree of risk relates to quality of management, the company culture, the level of technical expertise, and the state of existing data systems. Thus, one can think of this as a "signaling game" in which a company undertakes a project to show that it has agile management, a high level of technical expertise, or new product lines or marketing ideas that will require rapid expansion. Benefits of such signaling might include; causing less well-placed competitors to exit or sell-out at reduced prices, inducing venture capitalists to invest, or raise the company's value at an IPO. Since ESM is scalable, it makes rapid company expansion much easier and cheaper. Thus, it is very similar to installing excess but unused capacity in an oligopoly game. Since legacy systems and employees used to the current way of doing things are an impediment to successful deployment of ESM, new companies who can build both systems and company culture from the ground up have an inherent advantage over incumbents. This suggests that there might be more churning of companies as the cycle of technological advance speeds up, or that incumbent companies would do well to continuously spin off new product and business lines to independent subsidiaries.

[17]This data comes from a variety of sources most of which seem to be reports from consulting companies. See Krigsman (2010) or Golorath (2007) for more discussion.

On the other hand, we can imagine that companies might follow the strategy of secretly undertaking ESM projects. Since these can take one to two years to complete, a company can steal the march on its competitors. If the project is successful, a company has as many as two years to enjoy production with lower operating costs and cheap and easy expandability. An aggressive company might be able to achieve a scale that would make it hard for competitors to survive especially in sectors that enjoy network externalities. Thus, we might see innovation races where the first across the post gets the network externality and thus monopoly. One can also imagine companies heading for bankruptcy going all in on ESM projects. If they fail, they are out of business anyway, if they succeed, the company might be saved. This implies we might see more ESM projects in recessions than in boom years, for this reason, and also because the opportunity cost of disruption to business and of employee time are lower, ceteris paribus, and because the advantages of being able to cheaply scale up as the business cycle turns positive again are greater.

One big problem with ESM is that one size does not fit all. Not only do different industries have different needs, but each company has its own management, culture, legacy systems, solvency, and so on. Especially for companies with older cultures and less technological savvy, deciding how to move forward is difficult. A common strategy is to look for examples of successful implementations and copy these solutions. This tends to create a second-mover advantage. The first innovator takes significant risks, but if he is successful, his competitors can simply copy him. Thus, the advantages to taking the risk are temporary. This implies that innovation might be slower in sectors where there are no network externalities or other factors that would allow a low cost firm to rapidly expand its market share. We might also expect that the most successful incumbent in a sector would have little incentive to rock the boat and risk teaching his competitors how to innovate.

Companies starting big ESM projects can proceed in a variety of ways. The cheapest and fastest way is to put together off-the-shelf solutions and do as little customization as possible. This might mean using SaaS such as Salesforce.com for CRM and Workforce.com for HR

functions and Oracle Web Services for data integration, for example. Putting these together with legacy systems may require building applications using PaaS so as to focus on the direct software needs rather than worrying about the details of the infrastructure. The alternative is to build a customized system from scratch using IaaS, which is more expensive, time consuming and requires greater technical knowhow.

Aside from the obvious advantage that custom systems can be tailored to the exact needs of the company there are several reasons for companies to choose this path. Perhaps the most important is that "lock-in" is a big concern when you use SaaS and PaaS. Software service providers each store data in their own proprietary way.[18] Extracting such data and building a new software system around it is an expensive and difficult task. In addition, employees get used to the workflow and interfaces of these proprietary systems. This also makes it costly to switch systems. In a similar way, the greater the degree of abstraction in the PasS platform – for example special API's for interactions between components like databases and email, proprietary libraries of code that the users can build applications with, the more difficult it is to move to a new provider. Thus, building ESM on more abstracted layers of XasS makes users of cloud services more vulnerable to price increases. In addition, users are not in a good position to enforce high service quality. More specifically, as technology and markets change, cloud providers may choose not to continue to support functions or features of their services that are highly valued by a subset of customers. Updates to cloud systems may affect the way that they interact with the rest of a company's ESM solution and so crashes may result that require time to fix.

Of course, the possibilities of lock-in makes such cobbled together systems less valuable to users and so less profitable to providers. Thus, we might consider a game between cloud providers in which they choose how easy to make it for customers to cleanly move to another provider. While this would decrease their market power, it would increase their

[18] Lock-in is a has been extensively studied in economics. See Farrell and Klemperer (2007) for a recent survey

value and thus the price they could charge for services. Especially if service providers do not plan to take advantage of lock-in by suddenly switching from a low price/customer-base building phase to a high price/rent extraction phase, it would seem to be a dominant strategy to make it easy for customers to leave. These same considerations give an advantage to rapidly growing companies and very large companies like Google, Oracle and Amazon. Such companies are less likely to switch to a rent extraction phase since this would deter future customers. Growing companies get more benefit from maintaining a good customer reputation, while large companies suffer more damage if they try to extract rent from a subset of customers.

The information technology revolution of the 80's and 90's in part replaced human workers with computers and automated systems. As such, it tended to convert variable costs into fixed costs. This made entry by poorly capitalized challengers more difficult. The cloud computing revolution has the opposite effect. Companies no longer need to build server farms and computing services become pay-as-you-go. This is especially significant in sectors where most costs are computing services such as content delivery, web based gaming and entertainment, social networking and so on. Since companies are not saddled with legacy systems, they can build integrated cloud applications from the ground up. These factors make it easier for entrants to offer entertainment and other services that compete with incumbents like big networks, Hollywood studios and music companies. Since scale increases both revenue and costs at the same time, profitability can be maintained at variety of scales. This in turn allows new companies to offer services of particular interest to relatively small groups and also to scale up rapidly to take advantage of network externalities if they exist. Thus, the value of the intellectual property of large incumbents is reduced. The democratizing effect of the cloud on entrepreneurship means that more options available that compete for consumer interest with existing content. The claims that piracy facilitated by P2P networks built on cloud infrastructure is the primary cause of the decline of these incumbents must therefore be taken with a grain of salt.

Industry lobbying to protect their position has led to the passage of such measures as the Sonny Bono Copyright Term Extension Act and the Digital Millennium Copyright Act (DMCA) which extend copyright terms and increase penalties for piracy. The justification is that artists will not produce art if it is not protected. It would be interesting to see reliable estimates of how much IP would be lost if copyright protection were reduced. Musicians make most of their money from performances and touring, plenty of people write blogs and post their fiction and poetry, YouTube posts millions of both user and company created video, even textbooks and lectures are posted for free.

One may argue that a large part of the reason for this is that many artists and other content creators are not primarily motivated to produce by prospect of monetary reward. At the same time, content creators are generally not in a position to spend much to distribute their works and thus access the non-monetary benefits of reputation, praise, fame, etc. In the old days, the only way to distribute works was to publish books, records, tapes, movies, and so on. These were expensive both to produce and to distribute, and as such, content creators had no alternative but to go through company gateways to get their products to the public. With new technologies, many cloud based, it is cheap both to produce and distribute content. Thus, monetizing content might have been necessary to give incentives to publishers to distribute new works, but not get artists to create them. Now that artists can distribute their own works, the policy justification for extensive copyright protection may be much weaker. Of course, one size does not fit all. Big expensive productions like major Hollywood movies would not be made without copyright protection. This suggests that the solution might be to make copyright costly (perhaps $1000). This would mean that for works that are expensive to produce and require payback, copyright protection would be available and would be purchased. On the other hand, for works of small commercial value where money is not the primary motivating factor, creators would choose not to copyright and the creative commons would be enriched as a result.

A big question is how to price and allocate resources over customers. At any given time, a cloud provider has fixed physical resources. The installed servers provide a certain amount of storage, network connectivity, and CPU cycles. Thus, there is a significant fixed cost of providing capacity that must somehow be shared over users. Users, on the other hand have variable demands. Some of this is beyond the user's control, such as page accesses by customers, but others, such as database updates, payroll processing are more flexible. Users also have differing costs of latency. Delaying the completion of a task by several seconds may cost a provider of real time customer oriented applications lost sales, while it might not make much difference to a firm that creates architectural renderings or animations.

Cooperative game theory can provide some guidance here. For example, one could use the Shapley value to allocate the fixed costs. The basic idea is to choose a customer and then consider every possible subset of the remaining customers. For each of these subsets, find the additional fixed cost expenditure required to meet the chosen customer's Service level agreement (SLA). Take the average of these and this is in a sense the average fixed cost expenditure required to serve a given customer if we are agnostic about the order of customer arrival. This way of allocating costs satisfies appealing sets of axioms. See Chun (1998) and references therein. SLA's can be viewed as a "claim" on CPU, bandwidth and storage that cannot be met at every instant by the cloud provider. Thus, we might allocate the shortage using a bargaining with claims approach. See Aumann and Maschler (1985) and Chun and Thomson (1992), for example. This would require using either a loss function to aggregate the failures to meet claims in the three dimensions of CPU, bandwidth and storage into one dimension, or more interestingly, to expand the bargaining with claims approach to three dimensions. Similar extensions of other axiomatic bargaining solutions would further inform this problem. As we mention above, Ghodsi et.al. (2010), and Friedman (2011) are the only papers of which we are aware that use such techniques for cloud computing applications.

SLAs, however, are very crude instruments and don't give users an incentive to internalize the costs that their momentary demands impose on other users. Having users and providers bid for and offer services should make the allocation more efficient if correctly implemented. This suggests that Auction theory may also provide an approach to allocate and price cloud services. One of the key problems is getting agents to reveal how valuable services are and how costly latency is. Of course, there is a large auction literature in economics (see Klemperer 1999 for a survey) but it does not seem to have been directly applied to cloud computing. The computer science literature is much more active in this area. See Buyya, Yeo and Venugopal (2008) and references therein. A significant real world challenge is to come up with real time automatic bidding programs to participate in these markets. Langford, Li, McAfee and Papineni (2011) is the only paper written by economists along these lines of which we are aware. In that paper they set up the problem of finding the optimal traffic shaping mechanism, which requires that we can first identify the different classes of traffic and then induce time shifting through discriminatory prices of traffic throttling. They solve this optimization problem and show that the solution is relatively simple- pricing "through rate limits." They then simulate their results using data obtained from Yahoo!.

A final economic issue of concern is that bankruptcy, reorganizations or even changes in business focus on the part of cloud service provides pose significant risks for users. In the worst case, users may lose data stored in the cloud. At a less catastrophic level, features and functionality may be dropped or no longer supported, service levels of companies distracted by internal problems may decline. Seeing this, users who can may withdraw their data and find new providers of services. This further weakens the company, and induces more customers to leave. The dynamics are very much like a bank-run. There may be a role for regulations and standards to give users some assurance of stability. This also may provide a strategic advantage for established firms with a long run reputation fro reliability.

5. Conclusion

Cloud computing is the next step in the on-going evolution of Information Technology. From a technical standpoint, very little that currently is done on cloud platforms could not have been done with previously available technology. However the cost-reductions, rapid scalability and flexibility of cloud solutions give them a revolutionary potential in many economic sectors. These factors also open many significant economic questions in industrial organization, labor economics, and other areas on both the theoretical and empirical side. Despite this, the the economics literature is exceeding thin. Most of work is confined to few papers in cooperative and non-cooperative game theory. We survey this literature as well as the more relevant parts of the much larger computer science and business literature on this topic. We argue that the technological categorizations used tin these fields (Software, Platform and Infrastructure as a Service) do not correlate well to the economic impacts cloud technologies. Instead, we propose that it is more useful for economists to think about "wholesale" and "retail" cloud applications. Wholesale cloud applications are primarily aimed at businesses. They facilitate large-scale data integration projects, rapid low-fixed cost entry and expansion of new start-ups, and the outsourcing of many non-core aspects of a given firm's activities. Retail cloud applications, on the other hand, are primarily aimed at consumers. They move applications and content off personal computers to various types of cloud platforms and make both consumer produced and purchased content available in increasingly device and location agnostic ways. In addition, retail applications transform the way the consumers socialize, communicate, and consume entertainment. We conclude by suggesting several directions for new research.

Appendix

A1. The NIST Definition of Cloud Computing

National Institute of Standards and Technology, Information Technology Laboratory Peter Mell and Tim Grance Version 15, 10-7-09

Note 1: Cloud computing is still an evolving paradigm. Its definitions, use cases, underlying technologies, issues, risks, and benefits will be refined in a spirited debate by the public and private sectors. These definitions, attributes, and characteristics will evolve and change over time.

Note 2: The cloud computing industry represents a large ecosystem of many models, vendors, and market niches. This definition attempts to encompass all of the various cloud approaches.

Definition of Cloud Computing:

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models.

Essential Characteristics:

On-demand self-service. A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.

Broad network access. Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).

Resource pooling. The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.

Rapid elasticity. Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.

Measured Service. Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported providing transparency for both the provider and consumer of the utilized service.

Service Models:

Cloud Software as a Service (SaaS). The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface such as a web browser (e.g., web-based email). The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

Cloud Platform as a Service (PaaS). The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations.

Cloud Infrastructure as a Service (IaaS). The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls).

Deployment Models:

Private cloud. The cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on premise or off premise.

Community cloud. The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission,

security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on premise or off premise.

Public cloud. The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.

Hybrid cloud. The cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load-balancing between clouds).

Note: Cloud software takes full advantage of the cloud paradigm by being service oriented with a focus on statelessness, low coupling, modularity, and semantic interoperability.

A.2.1. Classification of Cloud Computing Applications via Fichman's Framework

		Locus of Adoption	
		Individual	Organizational
Class of Technology	Type 1 (low user interdependencies and knowledge barriers)	Personal adoption of simple SaaS applications such as email, word processing, data management	Organizational adoption of SaaS applications such as CRM or enterprise email.
	Type 2 (high user interdependencies and knowledge barriers)	Personal adoption of PaaS for web development and IaaS for hosting	Organizational adoption of PaaS for application development, and IaaS for high volume computing

A.2.2.Determinants of Adoption via Fichman's Framework

		Locus of Adoption	
		Individual	Organizational
Class of Technology	Type 1 (low user interdependencies and knowledge barriers)	Classical diffusion variables: Perceived Innovation Characteristics Adopter Characteristics Information Sources and Communication Channels Change Agents and Opinion Leaders	Classical diffusion variables Organizational characteristics Organizational decision processes Stage of implementation Competitive effects (adopter industry) Supply side factors Economic factors (price) IT group characteristics
	Type 2 (high user interdependencies and knowledge barriers)	Classical diffusion variables Managerial influences Critical mass Absorptive capacity Implementation characteristics Institutions lowering knowledge barriers	 Combination of variables 

References

Aumann, R., Maschler, M., (1985) "Game theoretic analysis of a bankruptcy problem from the Talmud", *Journal of Economic Theory*, Volume 36, Issue 2, Pages 195-213.

Archak, N and Sundararajan, A, "Optimal Design of Crowdsourcing Contests" (2009). *ICIS 2009 Proceedings*. Paper 200. http://aisel.aisnet.org/icis2009/200

Benlian, A., Hess T. and Buxmann, P.,(2009) "Drivers of SaaS-Adoption- An Empirical Study of Different Application Types", *Business & Information Systems Engineering*, 5: pp. 357-368.

Brancheau, J. C.; and Wetherbe, J. C. (1990) "The Adoption of Spreadsheet Software: Testing Innovation Diffusion Theory in the Context of End-User Computing." *Information Systems Research*, Volume 1, 1990, pp. 115-143

Buyya, R., Yeo, C. S., 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*, Vol. 25, No. 6, pp. 599—616.

Chun, Y., (1998) "A new axiomatization of the shapley value, Games and Economic Behavior, Volume 1, Issue 2, pages 119-130,

Chun, Y., and Thomson, W., (1992) "Bargaining problems with claims", *Mathematical Social Sciences*, vol. 24(1), pages 19-33., August.

Conley, J., and Kung, F.-C., (2010), "Private Benefits, Warm Glow, and Reputation in the Free and Open Source Software Production Model", *Journal of Public Economic Theory*, vol. 12 pages 665–689.

Creeger, M. (2009) "CTO Roundtable: Cloud Computing" *Communications of the ACM*, Vol. 52, No. 8, pp. 50—56.

Davis, F. Bagozzi, R. and Warshaw, R. (1989) "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models." *Management Science*, Volume 35, 1989, pp. 982-1003

Durkee, D. (2010) "Why Cloud Computing Will Never Be Free". *ACM Queue*, Vol. 8, No. 4, pp. 1-10.

Etro, F., (2009) "The Economic Impact of Cloud Computing on Business Creation, Employment and Output in Europe", *Review of Business and Economics*, Vol. 54, 2, pp. 179-218

Farrell J. and Klemperer P. (2007) "Coordination and Lock-In: Competition with Switching Costs and Network Effects" in *The Handbook of Industrial Organization, Volume 3*, Chapter 31 pp. 1967-2072, Elsevier.

Fichman, R. G., (1992) "Information Technology Diffusion: A Review of Empirical Research," *Proceedings of the Thirteenth International Conference on Information Systems*, Dallas, , pages195-206

Friedman E., Ghodsi, A., Shenker, S. and Stoica, I. (2011) "Bargaining Theory in the Cloud", unpublished working paper.

Gartner, Inc., (2010) Press release: Gartner Says Worldwide Cloud Services Market to Surpass $68 Billion in 2010, http://www.gartner.com/it/page.jsp?id=1389313.

Gatignon, H. and Robertson, T. S. (1989) "Technology Diffusion: An Emperical Test of Competitive Effects." *Journal of Marketing*, Volume 53, 1989, pp. 35-49.

Ghodsi, A., Zaharia, M., Hindman, B., Konwinski, A., Shenker S. Stoica I. (2010) "Dominant Resource Fairness: Fair Allocation of Heterogeneous Resources in Datacenters" Technical Report No. UCB/EECS-2010-55.
http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-55.pdf

Golorath, D, (2007) "Software Project Failure Costs Billions.. Better Estimation & Planning Can Help" http://www.galorath.com/wp/software-project-failure-costs-billions-better-estimation-planning-can-help.php.

Gurbaxani, V. (1990) "Diffusion in Computing Networks: The Case of BITNET." *Communications of the ACM*, Volume 33, 1990, pp. 65-75.

Gurbaxani, V. and Mendelson, H. (1990) "An Integrative Model of Information Systems Spending Growth." *Information Systems Research*, Volume 1, 1990, pp. 23-47.

Hamilton, J. (2008) "Internet-scale service efficiency" *Proceedings of the Large-Scale Distributed Systems and Middleware (LADIS) Workshop.*

Huff, S. L. and Munro, M. C. (1989) "Managing Micro Proliferation." *Journal of Information Systems Management*, 1989, pp. 72-75.

Iyer, B. and Henderson, J. C. (2010) "Preparing for the Future: Understanding the Seven Capabilities of Cloud Computing", *MIS Quarterly Executive*, Vol. 9, No. 2, pp. 117-131.

Khajeh-Hosseini, A., Greenwood, D., and Sommerville, I. "Cloud Migration: A Case Study of Migrating an Enterprise IT System to IaaS," *Cloud Computing, IEEE 3rd International Conference on Cloud Computing*, pp. 450-457,

Kittur, A. , Chi, E. H., Suh, B. (2008) "Crowdsourcing user studies with Mechanical Turk" in *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pp 453-456

Klemperer, P. (1999), "Auction Theory: A Guide to the Literature". Journal of Economic Surveys, 13 pp. 227–286.

Krigsman, M, (2010) "Enterprise 2.0 IT Project Failure", in *The Next Wave of Technologies: Opportunies from Chasos*, Chapter 18, Paul Simon, ed., John Wiley and Sons, Inc.

Langford, J. Li, L., Mcafee, P., and Papenini, K.,(2009) "Cloud Control: Voluntary Admission Control for Intranet Traffic Management", *Information Systems and e-Business Management Special Issue*: *2009 Workshop on E-Business*

Leimeister, J., Huber, M., Bretschneider, U., and Krcmar, H. (2009) "Leveraging Crowdsourcing: Activation-Supporting Components for IT-Based Ideas Competition" *Journal of Management Information Systems*, Vol. 26 No. 1 , pp. 197 - 224

Leonard-Barton, D. (1987) "The Case for Integrative Innovation: An Expert System at Digital." *Sloan Management Review*, Volume 29, pages. 7-19.

Leonard-Barton, D. and Deschamps, I. (1988) "Managerial influence in the implementation of new technology." *Management Science*, Volume 34, 1988, pp. 1252-1265.

Lin, G, Fu, D., Zhu, J. and Dasmalchi, G. (2009). "Cloud Computing: IT as a Service" *IT Professional*, Vol. 11, No. 2, pp. 10-13.

Lynch, N., and Gilbert, S., (2002) “Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services”, *ACM SIGACT News*, Volume 33 Issue 2 , pp. 51-59.

Qian, L., Luo, Z., Du, Y. & Guo, L. (2009) “Cloud Computing: An Overview” *Proceedings of 1st International Conference on Cloud Computing Conference*, pp. 626-631.

Rimal, B.P., Eunmi Choi, Ian Lumb, (2009) ”A Taxonomy and Survey of Cloud Computing Systems,” Networked Computing and Advanced Information Management, Fifth International Joint Conference on INC, IMS and IDC , pp. 44-51,

Vaquero, L. M., Rodero-Merino, L., Caceres, J. & Lindner, M. (2009) “A Break in the Clouds” *Towards a Cloud Definition. Computer Communications Review*, Vol. 39, No. 1, pp. 50-55.

Viega, J. (2009) “Cloud Computing and the Common Man” *Computer*, Vol. 42, No. 8, pp. 106-108

Weinhardt, C., Anandasivam, A.. Blau, B.. and Stößer, J. (2009) “Business Models in the Service World” *IEEE IT Professional*, Vol. 11, No. 2, pp. 28-33.

Youseff, L., Butrico, M. and Da Silva, D. (2008) “Toward a Unified Ontology of Cloud Computing” *Proceedings of 2008 IEEE Grid Computing Environments Workshop*, pp. 1-10.