CHAPTER ONE

1 Introduction to Cloud Computing

Omkhar Arasaratnam

CLOUD COMPUTING HAS taken the IT world by storm. Often viewed as the utopia of utility computing, cloud computing offers flexibility and financial benefits second to none. It also lowers the entry point to high performance computing, allowing organizations to leverage computing power that they have neither the capital budget nor operational expertise to acquire. This chapter provides background as to where cloud computing came from, what cloud computing is, and discusses some of the advantages and challenges with cloud computing.

HISTORY

Computing has evolved significantly over the last 60 years. In the early days, a large central computer would be used by an entire company. This gradually evolved to departmental computers in the 1970s and later personal computers in the 1980s and 1990s. Although cloud computing is a new term, as a concept it was predicted by computer scientist John McCarthy in the

1960s. McCarthy asserted: ''Computation may someday be organized as a public utility.''

McCarthy had the foresight to predict what we today refer to as cloud computing. In the mid-1960s, Intel co-founder Gordon E. Moore famously predicted that the number of transistors (or computing power) that could be inexpensively placed on an integrated circuit would double every two years. This is commonly known as Moore's law. By the late 1990s, Moore's law had guided computing to heights beyond many organizations' predictions. Much of this demand was fueled by the now popular World Wide Web (WWW), which brought an age of networking and collaboration that had not been seen before.

By the mid-2000s, many companies had discovered that their largest IT purchases were often left idle and only fully utilized during peak demand. These organizations were very large IT or academic organizations. This had researchers wondering how best to leverage the latent processing power. Thus, the initial underpinnings of cloud computing were born.

In 2007, Google, IBM, Carnegie Mellon, MIT, Stanford University, UC Berkeley, the University of Maryland, and the University of Washington collaborated to begin research into cloud computing. Before long, many analyst groups began reporting on the significant market share being established by cloud computing. Many standards organizations and consortiums such as the Open Group, OASIS, and DMTF had also begun working groups to define cloud computing standards.

DEFINING CLOUD COMPUTING

Cloud computing is regarded as an evolutionary rather than a revolutionary step. In other words, cloud computing hasn't drastically altered existing technologies, but rather it has succeeded as a result of the collaboration of several existing technologies.

The actual definition of cloud computing is frequently contested. Most will agree that any computing model that qualifies as cloud computing must at minimum have the following criteria:

Elasticity

Cloud computing is typified by its ability to rapidly scale the capacity of the provided service up or down with little to no interaction from the consumer. This characteristic, known as *elasticity*, is key to cloud computing.

In some delivery models of cloud computing, elasticity is often facilitated through virtualization, although cloud computing does not require virtualization.

Multitenancy

Clouds are inherently multitenanted—even private clouds, which run the workload of a single corporation posses multiple tenants, be they workloads or individual users. This multitenancy and multitenant amortization of the shared compute resource is part of the reason for the economic benefits of cloud computing.

Economics

With cloud computing services, the expectation is that the consumer is charged for the amount of time used on the resource. Cloud computing changes the computing barrier to entry for high performance computing resources, by allowing consumers to use only what they need for the time in which they need it. In turn, this has allowed organizations to effectively respond to peak demand requirements without having excess compute resources sitting idle during dormant periods. Clouds can achieve this by distributing the load across multiple shared resources and relying on economies of scale.

Abstraction

The most significant change with cloud computing is that of abstraction. As we will describe in the following section, most cloud providers provide one or more service layers to their consumers. The operational aspect of the layers supporting the service is insulated from the customer. So, a Software as a Service (SaaS) customer will interact with the application itself, but not with the operating system or hardware of the respective cloud. This key difference allows organizations that do not have the necessary system administration skills or compute facilities to leverage enterprise applications hosted by others.

Many of the technologies that assist in providing these capabilities have been present for many years. Virtualization and autonomic response are areas of computing that have been well understood for decades, as has the Internet. Providers of cloud computing were able to assemble these disparate technologies into the above capabilities, ultimately defining cloud computing.

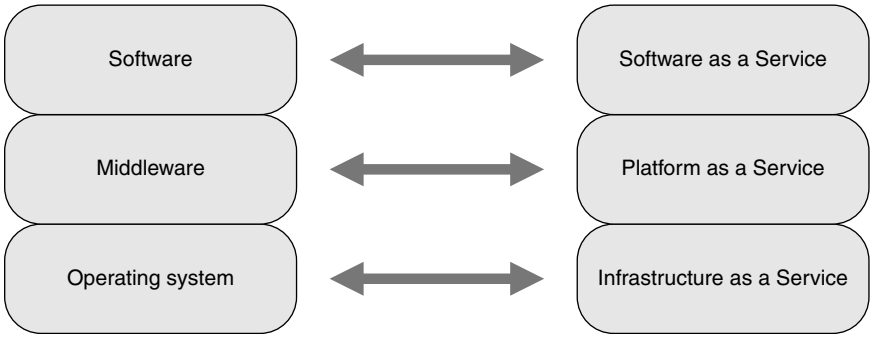

EXHIBIT 1.1 Traditional Model versus Cloud Computing Model

CLOUD COMPUTING SERVICES LAYERS

Cloud computing providers provide different kinds of services to cloud computing consumers. In order to understand the different layers of service, it's important to understand how they would relate in a noncloud computing scenario. See Exhibit 1.1.

The kind of service being provided has many implications on the provider, including how they address concerns such as security, resiliency, compliance, and multitenancy. Cloud computing services fall into one of the following categories, as shown in Exhibit 1.2.

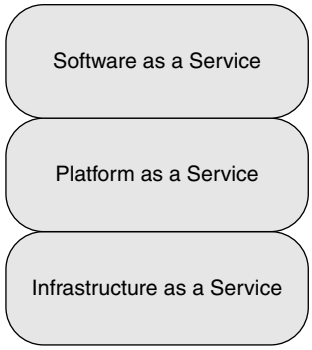

EXHIBIT 1.2 Categories of Cloud Computing Services

Infrastructure as a Service

Infrastructure as a Service (IaaS) providers allow their customers access to different kinds of infrastructure. The provider typically provides this service by dividing a very large physical infrastructure resource into smaller virtual resources for access by the consumer. Sometimes the service provided is a complete virtual machine with an operating system. In other instances the service provided is simply for storage, or perhaps a bare virtual machine with no operating system. In cases where the operating system or other software is included, the cost of the required license is either amalgamated into the cost for the service, or included as an additional surcharge.

IaaS providers are often service providers to other cloud providers (see Integrator). Many current Platform as a Service providers leverage IaaS providers for extra capacity on demand. One of the more popular IaaS providers is Amazon, who provides their EC2 IaaS.

Platform as a Service

Platform as a Service (PaaS) providers extend the software stack provided by IaaS to include *middleware*. Middleware generically refers to software such as a DB2 database, or runtime environments such as a Java Runtime Environment (JRE) or a Websphere application server. This middleware is a prerequisite to running more sophisticated applications, and provides a rich operating environment for the application to exploit. PaaS providers have two methods in which they facilitate the extra capacity needed for a large multitenant system. In some cases, they provide IaaS style virtual machines to the consumer. In other cases they provide an interface through which applications in the case of a runtime environment, or data in the case of a database, can be uploaded. A popular example of a PaaS is Microsoft's Windows Azure platform.

Each method has its advantages and challenges. With an IaaS style approach, the provider typically has more control and stronger separation between tenants. This approach is less efficient, however, as common overhead such as the operating system and the virtual machine itself are duplicated across multiple tenants.

In the second case, the underlying infrastructure is addressed in a much more efficient manner, with a single system image and middleware overhead amortized amongst multiple clients. Conversely, the main challenge with this approach lies in the degree of separation that can be provided between tenants. A runtime environment that is not robust or a misconfigured database can allow one user to adversely affect the quality of service of other users.

Software as a Service

Application as a Service, or Software as a Service (SaaS) providers as they are more commonly known, typically provide a rich web-based interface to their customers. The customer, in most cases, is completely abstracted from the nuances of the application running behind the scenes. Tenant separation is often done at the application layer, leaving a common application, platform, and infrastructure layer underneath. Popular examples of SaaS include Google Apps and Salesforce.com.

SaaS providers typically increase the capacity of their systems through scale up or scale out methods—depending on the characteristics of the application. SaaS applications that scale up are usually moved to larger platforms as their capacity requirements grow. SaaS applications that scale out are typically run on large clusters of servers. As additional capacity is required, the provider adds additional machines to the cluster.

As there is a significant amount of shared resources used between tenants in an SaaS environment, the ability of one tenant to affect the quality of service of other tenants is always a concern. The ability for an SaaS provider to adequately *fence* or insulate one tenant from another is key to maintaining quality of service.

ROLES IN CLOUD COMPUTING

The cloud-computing paradigm defines three key roles. These roles each have different responsibilities and expectations relative to one another. Any party might have multiple roles depending on the context. See Exhibit 1.3.

Consumer

Simply defined, a consumer consumes any service that is provided. In Exhibit 1.3, the SaaS provider exposes an SaaS to the SaaS consumer. The consumer is permitted access to this service for a fee of some sort, though in many instances this fee is augmented or replaced through advertising revenue. The consumer has no responsibility, nor access beyond the SaaS provided to them.

Provider

The providers in this case are both the PaaS provider and the SaaS provider. The PaaS provider provides a PaaS to the SaaS provider. The SaaS provider in

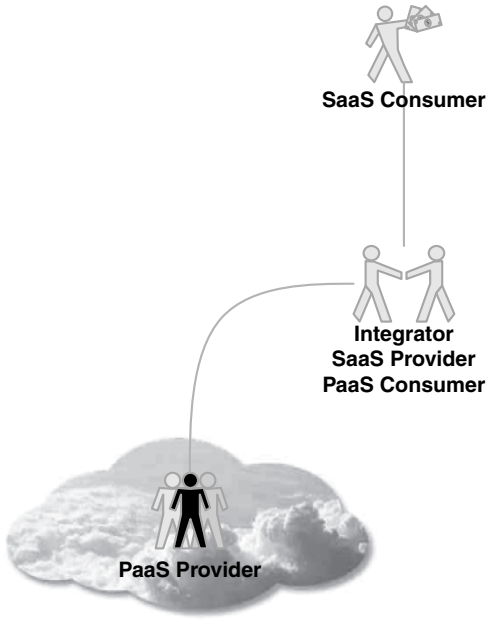

EXHIBIT 1.3 Three Key Roles

turn provides an SaaS to the consumer. Ultimately, the provider is anyone who provides a service to one or more consumers.

Integrator

The integrator role is sometimes referred to as a broker. The integrator essentially assembles the services of many providers under a new service. In some cases, this might involve integrating multiple providers of the same service—for example, the integration of multiple IaaS providers to provide a more resilient or fully-featured IaaS service. In other cases, per our diagram, the integrator might consume another provider's service (in this case PaaS) in order to run a service of their own. The integrator's service is ultimately exposed as an SaaS to the SaaS consumer.

Depending on the perspective, we will see that each party can have multiple roles. The SaaS provider is ultimately a consumer of the PaaS provider, and an integrator of the PaaS service with its SaaS.

CLOUD COMPUTING DEPLOYMENT MODELS

Cloud computing has a number of different deployment models. A deployment model is a particular method of delivering a service. In the case of cloud computing, these are unique methods of deploying a cloud computing service. Deployment models often have particular characteristics that suit them to appropriate workloads. The most commonly used deployment models are as follows.

Private

In a private cloud, the cloud computing services are provided by an internal organization for use by other internal organizations. The provider in this case is most often the internal Information Technology (IT) or Information Systems (IS) department.

Consumers vary, but typically these consumers are consumers of other IT services. As both the consumer and provider are internal organizations, private clouds allow the consumer greater control over quality of service provided by the cloud. For instance, an internal customer can more easily assert the relative priority of a particular workload. In another instance, the internal consumer might assert specific characteristics regarding how critical the workload is, and therefore, the availability requirements of the cloud.

Private clouds generally provide the most control, as both provider and consumer are part of the same organization. This control comes at a price, as the organization ultimately bears the full cost of the cloud infrastructure.

Community

Community clouds have membership in one or more organizations. Community clouds are often groups of individuals and organizations collaborating for the purpose of a particular mission or concern. This might be an industry consortium, an awareness group, or another group altogether. In some instances the community cloud is a shared responsibility, either financially or from a compute resource perspective. In other instances, one member of the community provides all the funding and resources with other members contributing as appropriate. The defining factor for the community cloud is that the different constituents are all assembled for a common cause. As community clouds are provided and supported by consortia with a common cause, there is some influence that members have over the quality of service. The ultimate decision usually applies to the majority and/or veto members.

Public

Public clouds are clouds where the provider delivers a cloud service to any customer who wishes to access it. Unlike a private or community cloud, there are no stipulations regarding the consumer's ownership or cause in using the cloud; it is simply provided for any customer who wishes to support the payment model. One challenge of public clouds is the assurances regarding quality of service. In many instances, today's cloud providers offer little in compensation for missed service level agreements (SLAs). Most often, the only compensation offered is the reimbursement of fees paid by the consumer. For many organizations, the business impact of application downtime can be magnitudes above the cost of the cloud service. Consider the extreme example of a stock exchange where downtime is measured in millions of dollars per hour versus a cloud service measured in cents per hour. This has caused many organizations to avoid using public clouds for their mission-critical workloads, and instead to relegate their usage to noncritical assets.

Further concerns regarding security issues, such as data residency requirements, have prevented the adoption of public cloud computing across some countries. When leveraging public clouds, consumers should exercise due diligence to ensure that their use of the cloud does not violate legislative, regulatory, or industry requirements.

Hybrid

Hybrid clouds are not a separate deployment model, but rather an intersection of two or more—for example, a private SaaS application that is based on a public IaaS.

Hybrid clouds are usually focused on leveraging the economies of scale present in public cloud offerings, but also on driving workloads that have more stringent quality of service requirements. For this reason, in many cases hybrid clouds are internal clouds, which turn to the capacity of public clouds for peak demand.

CHALLENGES

Although there are many benefits to cloud computing, it is not without its own challenges. Traditional computing models have permitted a high degree of control over compute resources. Cloud computing, by virtue of abstraction, prevents the consumer from having the same level of influence over the computing resource.

Of great concern is the ability of consumers to assert quality of service. Quality of service refers to aspects of a service that are not functional (press the red button to submit your form) but are important considerations—for example, how quickly the form submission takes to process. This leads to some of the following challenges with public cloud computing.

Availability

One of the cited benefits to cloud computing is the lowered barrier to entry. This is facilitated financially by sharing a significant amount of infrastructure between consumers. This allows for economies of scale from both a service management and computing resource perspective. Unfortunately, this also causes challenges regarding system availability.

A nefarious user within the cloud can adversely affect the performance or availability of other consumers by attempting to over-consume resources. A similar effect can be attained by using the significant elasticity of the cloud to launch a denial of service attack against other consumers or organizations.

Although most cloud providers have methods of defending against these attacks launched by criminals, law enforcement can be just as disruptive. Consider a recent case where law enforcement officials confiscated racks of servers out of a data center as a result of alleged illegal activity from one of the consumers. Unfortunately, due to the use of virtualization many tenants' information was confiscated at the same time.

Equally significant, in late 2007, Rackspace suffered a 36-hour outage due to a damaged transformer outside of their data centers. This was a significant issue that affected all of its customers.

Cloud consumers should continue to consider availability requirements as well as business continuity plans when using cloud computing. Cloud computing does not alleviate these requirements.

Private cloud users are equally susceptible to this challenge. In scenarios where the directive is to migrate all workloads to a single data center, or to leverage virtual machines over separate physical machines, IT departments may inadvertently increase their single points of failure. This occurs as critical infrastructures are moved to a common area, or previously redundant systems are shifted onto single pieces of hardware.

Data Residency

Different countries and regions have different requirements regarding how its citizens' information should be handled. In some areas, such as the European

Union (EU), there are specific requirements regarding protection of personally identifiable information (PII) of EU residents. In other areas, such as the United States (US), there are directives regarding protected health information (PHI), such as the Health Insurance Portability and Accountability Act (HIPAA).

The challenge is that without knowing where all the cloud provider's assets reside, it is difficult to know with which legislation the consumer needs to comply. Furthermore, if the cloud provider has multiple centers worldwide, in many instances it is impossible to tell where in the world a particular consumer data set might be at any one point in time.

Cloud consumers should consult with their providers regarding the countries in which they operate, and if possible restrict them to a subset that is congruent with their security and compliance requirements.

Multitenancy

Although multitenancy affords cloud consumers unprecedentedly low prices, security and compliance considerations need to be taken into account. Depending on the layer of cloud service being provided, the appropriate security controls should be employed. This can include host intrusion detection systems (HIDS), hypervisor based security agents, host firewalls, and many others.

Some cloud providers will enforce particular controls on their consumers to ensure a minimal level of security. Others will allow their clients full flexibility.

Cloud consumers who have concerns regarding the security of their workload in a multitenant environment should consult their provider regarding their security standards. If the provider's standards are deemed insufficient the consumer should augment this with compensating controls, or defer to an alternate provider.

Depending on the layer of cloud service provided, multitenancy can sometimes cause challenges with obtaining audit logs. Audit logs in a multitenant environment might contain other tenant information, which may or may not be cleansed properly prior to release. If specific audit logging requirements need to be met, the consumer should validate what the provider is capable of providing.

Performance

Many cloud providers assert a particular level of performance based on the service purchased. The main challenge regarding performance is the customer's recourse for degraded performance.

The penalty in an SLA provided by most cloud providers, more often than not, entails a refund of fees for services rendered. For many IaaS services, this is measured in cents per hour. There is no consideration given to the business impact of the service degradation, which may be many orders of magnitude higher. Consider a cloud consumer who might, in turn, provide a stock trading service to their customers with SLA penalties measured in thousands of dollars per second; in such a case, the penalty offered by the cloud provider is wholly inadequate.

Data Evacuation

In cloud environments, data evacuation can be a significant concern. Data evacuation focuses on how sensitive information is cleared from physical storage due to the suspension or deletion of a consumer's resources. This might be a table in a database, a virtual disk on a Storage Area Network (SAN), or virtual memory on a suspended disk. In the highly elastic world of cloud computing, memory is usually de-allocated, but not cleared. So a virtual machine containing sensitive information wouldn't be *zeroed out* prior to deletion, but rather its disk space would be released back to the SAN as is.

Depending on the consumer's requirement regarding data security, this might be a point of concern. Consumers should discuss this requirement if applicable with their provider. In instances where providers do not have advanced capabilities, the consumer should employ their own controls, such as encryption, when possible.

Supervisory Access

One of the main challenges with public cloud computing is that the highest level of access to the system, supervisory access, is maintained by the cloud provider. A cloud provider can inspect the activities of every virtual machine in an IaaS cloud. An SaaS provider has access to the information of all tenants in their cloud.

This is an important consideration for cloud consumers, especially as it pertains to the storage of highly regulated, confidential, proprietary, or other information deemed appropriate for a limited audience.

In such scenarios, it is recommended that the consumer seek contractual obligations and potentially define specific controls for the provider to prevent undue exposure as well as an agreed upon response in the event of an exposure.

IN SUMMARY

Cloud computing is the latest evolution in the delivery of computing power. It lowers the point of entry, permitting access to computing power previously only available to the largest organizations. It also permits smaller organizations to leverage fully-managed computing infrastructures, reducing the requirements for highly skilled IT staff.

Cloud computing is not without its challenges. Sensitive workloads that have strict quality of service requirements may not be appropriate for deployment to the cloud. Of particular concern are workloads with strict security and compliance requirements.

As cloud providers evolve their service offerings to the subsequent generations of the cloud, more sensitive workloads may eventually be permitted to leverage the advantages of cloud computing.