21UCA501 -Cloud Technology Fundaments

UNIT -I Defining Cloud Computing

Cloud computing takes the technology, services, and applications that are similar to those on the Internet and turns them into a self-service utility. The use of the word "cloud" makes reference to the two essential concepts:

Abstraction: Cloud computing abstracts the details of system implementation from users and developers. Applications run on physical systems that aren't specified, data is stored in locations that are unknown, administration of systems is outsourced to others, and access by users is ubiquitous. **Virtualization:** Cloud computing virtualizes systems by pooling and sharing resources. Systems and storage can be provisioned as needed from a centralized infrastructure, costs are assessed on a metered basis, multi-tenancy is enabled, and resources are scalable with agility.

Azure Platform: By contrast, Microsoft is creating the Azure Platform. It enables .NET Framework applications to run over the Internet as an alternate platform for Microsoft developer software running on desktops, which you will learn about in Chapter 10.

Amazon Web Services: One of the most successful cloud-based businesses is Amazon Web Services, which is an Infrastructure as a Service offering that lets you rent virtual computers on Amazon's own infrastructure.

Cloud Types

To discuss cloud computing intelligently, you need to define the lexicon of cloud computing; many acronyms in this area probably won't survive long. Most people separate cloud computing into two distinct sets of models**:**

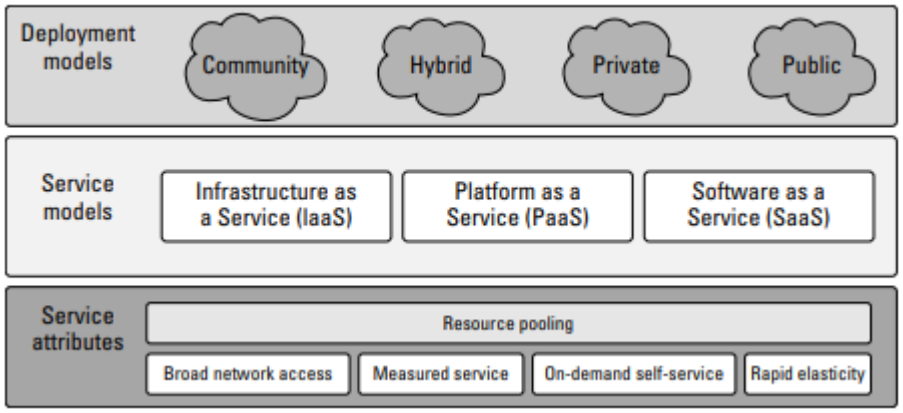
Deployment models: This refers to the location and management of the cloud's infrastructure. l **Service models**: This consists of the particular types of services that you can access on a cloud computing platform.

The NIST model

The United States government is a major consumer of computer services and, therefore, one of the major users of cloud computing networks. The U.S. National Institute of Standards and Technology (NIST) has a set of working definitions (http://csrc.nist.gov/groups/SNS/cloudcomputing/cloud-def-v15.doc) that separate cloud computing into service models and deployment models. Those models and their relationship to essential characteristics of cloud computing are shown in Figure 1.1. The NIST model

originally did not require a cloud to use virtualization to pool resources, nor did it absolutely require that a cloud support multi-tenancy in the earliest definitions of cloud computing. Multi-tenancy is the sharing of resources among two or more clients. The latest version of the NIST definition does require that cloud computing networks use virtualization and support multi-tenancy.

The NIST cloud computing definitions



The Cloud Cube Model

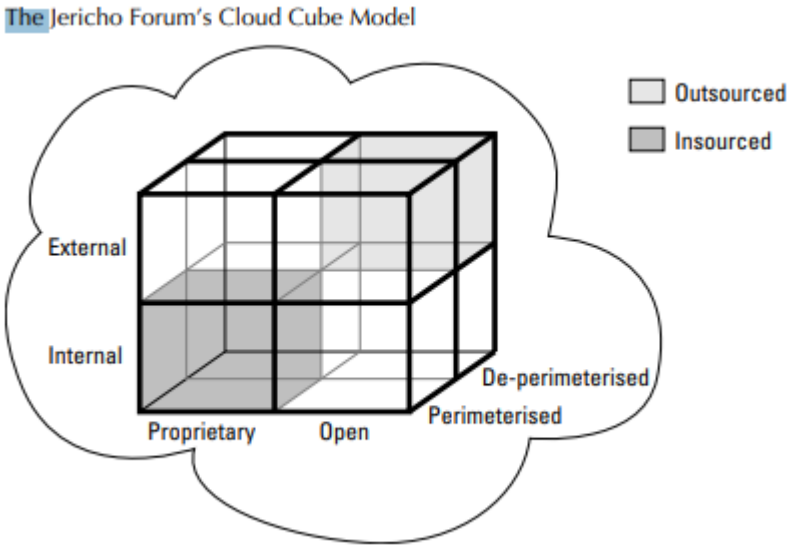
The Open Group maintains an association called the Jericho Forum (https://www.open group.org/jericho/index.htm) whose main focus is how to protect cloud networks. The group has an interesting model that attempts to categorize a cloud network based on four dimensional factors. As described in its paper called "Cloud Cube Model: Selecting Cloud Formations for Secure Collaboration" (http://www.opengroup.org/jericho/cloud_cube_model_ v1.0.pdf), the type of cloud networks you use dramatically changes the notion of where the boundary between the client's network and the cloud begins and ends. The four dimensions of the Cloud Cube Model are shown in Figure 1.2 and listed here: l

Physical location of the data: Internal (I) / External (E) determines your organization's boundaries.

Ownership: Proprietary (P) / Open (O) is a measure of not only the technology ownership, but of interoperability, ease of data transfer, and degree of vendor application lock-in.

Security boundary: Perimeterised (Per) / De-perimiterised (D-p) is a measure of whether the operation is inside or outside the security boundary or network firewall. L

Sourcing: Insourced or Outsourced means whether the service is provided by the customer or the service provider.

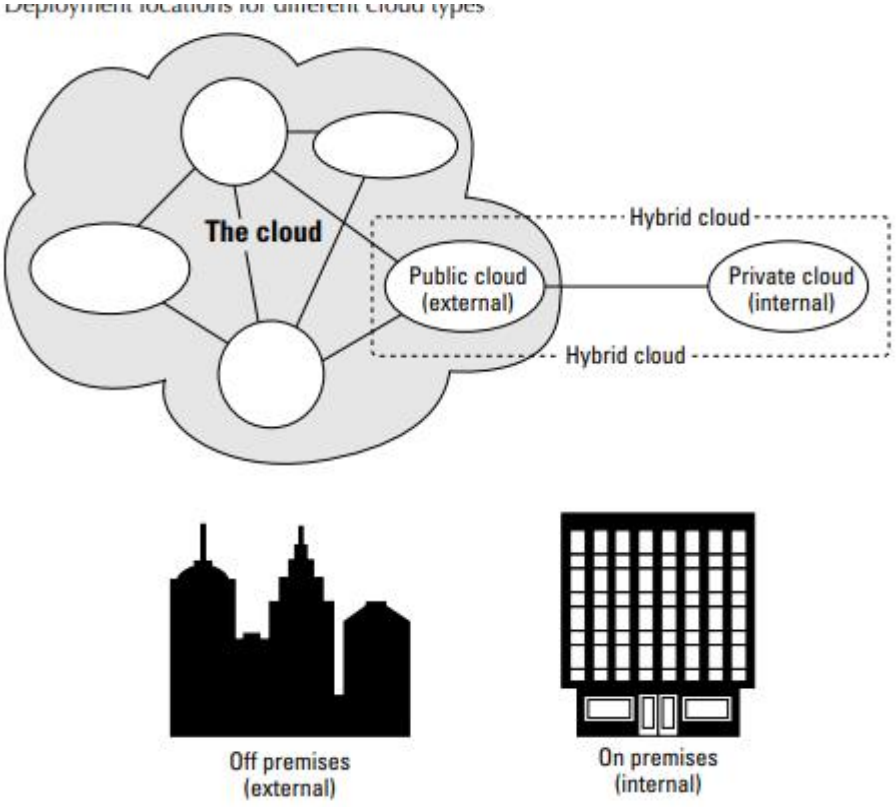


Deployment models A deployment model defines the purpose of the cloud and the nature of how the cloud is located. The NIST definition for the four deployment models is as follows: l

Public cloud: The public cloud infrastructure is available for public use alternatively for a large industry group and is owned by an organization selling cloud services.

Private cloud: The private cloud infrastructure is operated for the exclusive use of an organization. The cloud may be managed by that organization or a third party. Private clouds may be either on- or off-premises.

Hybrid cloud: A hybrid cloud combines multiple clouds (private, community of public) where those clouds retain their unique identities, but are bound together as a unit. Ahybrid cloud may offer standardized or proprietary access to data and applications, as well as application portability. **Community cloud:** A community cloud is one where the cloud has been organized to serve a common function or purpose. It may be for one organization or for several organizations, but they share common conerns such as their mission, policies, security, regulatory compliance needs, and so on. A community cloud may be managed by the constituent organization(s) or by a third party. Figure 1.3 shows the different locations that clouds can come in. In the sections that follow, these different cloud deployment models are described in more detail.



Service models In the deployment model, different cloud types are an expression of the manner in which infrastructure is deployed. You can think of the cloud as the boundary between where a client's network, management, and responsibilities ends and the cloud service provider's begins. As cloud

Infrastructure as a Service: IaaS provides virtual machines, virtual storage, virtual infrastructure, and other hardware assets as resources that clients can provision. The IaaS service provider manages all the infrastructure, while the client is responsible for all other aspects of the deployment. This can include the operating system, applications, and user interactions with the system.

Platform as a Service: PaaS provides virtual machines, operating systems, applications, services, development frameworks, transactions, and control structures. The client can deploy its applications on the cloud infrastructure or use applications that were programmed using languages and tools that are supported by the PaaS service provider. The service provider manages the cloud infrastructure, the operating systems, and the enabling software. The client is responsible for installing and managing the application that it is deploying.

Software as a Service: SaaS is a complete operating environment with applications, management, and the user interface. In the SaaS model, the application is provided to the

client through a thin client interface (a browser, usually), and the customer's responsibility begins and ends with entering and managing its data and user interaction. Everything from the application down to the infrastructure is the vendor's responsibility.

Benefits of cloud computing

"The NIST Definition of Cloud Computing" by Peter Mell and Tim Grance (version 14, 10/7/2009) described previously in this chapter (refer to Figure 1.1) that classified cloud computing into the three SPI service models (SaaS, IaaS, and PaaS) and four cloud types (public, private, community, and hybrid), also assigns five essential characteristics that cloud computing systems must offer:

On-demand self-service: A client can provision computer resources without the need for interaction with cloud service provider personnel. L

Broad network access: Access to resources in the cloud is available over the network using standard methods in a manner that provides platform-independent access to clients of all types. This includes a mixture of heterogeneous operating systems, and thick and thin platforms such as laptops, mobile phones, and PDA. L

Resource pooling: A cloud service provider creates resources that are pooled together in a system that supports multi-tenant usage. Physical and virtual systems are dynamically allocated or reallocated as needed. Intrinsic in this concept of pooling is the idea of abstraction that hides the location of resources such as virtual machines, processing, memory, storage, and network bandwidth and connectivity.

Rapid elasticity: Resources can be rapidly and elastically provisioned. The system can add resources by either scaling up systems (more powerful computers) or scaling out systems (more computers of the same kind), and scaling may be automatic or manual. From the standpoint of the client, cloud computing resources should look limitless and can be purchased at any time and in any quantity. L

Measured service: The use of cloud system resources is measured, audited, and reported to the customer based on a metered system. A client can be charged based on a known metric such as amount of storage used, number of transactions, network I/O (Input/Output) or bandwidth, amount of processing power used, and so forth. A client is charged based on the level of services provided.

While these five core features of cloud computing are on almost anybody's list, you also should consider these additional advantages:

Lower costs: Because cloud networks operate at higher efficiencies and with greater utilization, significant cost reductions are often encountered.

Ease of utilization: Depending upon the type of service being offered, you may find that you do not require hardware or software licenses to implement your service.

Quality of Service: The Quality of Service (QoS) is something that you can obtain under contract from your vendor. L

Reliability: The scale of cloud computing networks and their ability to provide load balancing and failover makes them highly reliable, often much more reliable than what you can achieve in a single organization. L

Outsourced IT management: A cloud computing deployment lets someone else manage your computing infrastructure while you manage your business. In most instances, you achieve considerable reductions in IT staffing costs.

Simplified maintenance and upgrade: Because the system is centralized, you can easily apply patches and upgrades. This means your users always have access to the latest software versions.

Low Barrier to Entry: In particular, upfront capital expenditures are dramatically reduced. In cloud computing, anyone can be a giant at any time. This very long list of benefits should make it obvious why so many people are excited about the idea of cloud computing. Cloud computing is not a panacea, however. In many instances, cloud computing doesn't work well for particular applications.

Measuring the Cloud's Value:

Cloud computing presents new opportunities to users and developers because it is based on the paradigm of a shared multitenant utility. The ability to access pooled resources on a pay-as-you-go basis provides a number of system characteristics that completely alter the economics of information technology infrastructures and allows new types of access and business models for user applications. Any application or process that benefits from economies of scale, commoditization of assets, and conformance to programming standards benefits from the application of cloud computing. Any application or process that requires a completely customized solution, imposes a high degree of specialization, and requires access to proprietary technology is going to expose the limits of cloud computing rather quickly. Applications that work with cloud computing are ones that I refer to as "low touch" applications; they tend to be applications that have low margins and usually low risk. The "high touch" applications that come with high margins require committed resources and pose more of a risk; those applications are best done on-premises. A cloud is defined as the combination of the infrastructure of a datacentre with the ability to provision hardware and software.

A service that concentrates on hardware follows the Infrastructure as a Service (IaaS) model, which is a good description for the Amazon Web Service described in Chapter 9. When you add a software stack, such as an operating system and applications to the service, the model shifts to the Software as a Service (SaaS) model. Microsoft's Windows Azure Platform, discussed in Chapter 10, is best described as currently using SaaS model. When the service requires the client to use a complete hardware/software/application stack, it is using the most refined and restrictive service model, called the Platform as a Service (PaaS) model. The best example of a PaaS offering is probably SalesForce.com. The Google App Engine discussed in Chapter 11 is another PaaS. As the Windows Azure Platform matures adding more access to Microsoft servers, it is developing into a PaaS model rather quickly.

Scalability: You have access to unlimited computer resources as needed. This feature obviates the need for planning and provisioning. It also enables batch processing, which greatly speeds up high-processing applications.

Elasticity: You have the ability to right-size resources as required. This feature allows you to optimize your system and capture all possible transactions.

Low barrier to entry: You can gain access to systems for a small investment. This feature offers access to global resources to small ventures and provides the ability to experiment with little risk.

Utility: A pay-as-you-go model matches resources to need on an ongoing basis. This eliminates waste and has the added benefit of shifting risk from the client.

The laws of cloudonomics Joe Wien man of AT&T Global Services has concisely stated the advantages that cloud computing offers over a private or captured system. His article appeared on Gigaom.com at: http:// gigaom.com/2008/09/07/the-10-laws-of-cloudonomics/. A summary of Wien man's "10 Laws of Cloudonomics" follows and his interpretation: 1. Utility services cost less even though they cost more. Utilities charge a premium for their services, but customers save money by not paying for services that they aren't using. 2. On-demand trumps forecasting. The ability to provision and tear down resources (de-provision) captures revenue and lowers costs.

The laws of cloudonomics

Joe Wienman of AT&T Global Services has concisely stated the advantages that cloud computing offers over a private or captured system. His article appeared on

Gigaom.com at: http:// gigaom.com/2008/09/07/the-10-laws-of-cloudonomics/. A summary of Wienman's "10 Laws of Cloudonomics" follows and his interpretation:

1. **Utility services cost less even though they cost more.**

   Utilities charge a premium for their services, but customers save money by not paying for services that they aren't using.

2.On-demand trumps forecasting.

The ability to provision and tear down resources (de-provision) captures revenue and lowers costs.

3. The peak of the sum is never greater than the sum of the peaks.

A cloud can deploy less capacity because the peaks of individual tenants in a shared system are averaged over time by the group of tenants.

4. Aggregate demand is smoother than individual.

Multi-tenancy also tends to average the variability intrinsic in individual demand because the "coefficient of random variables" is always less than or equal to that of any of the individual variables. With a more predictable demand and less variation, clouds can run at higher utilization rates than captive systems. This allows cloud systems to operate at higher efficiencies and lower costs.

5. Average unit costs are reduced by distributing fixed costs over more units of output.

Cloud vendors have a size that allows them to purchase resources at significantly reduced prices. (This feature was described in the previous section.)

6. Superiority in numbers is the most important factor in the result of a combat

(Clausewitz). Weinman argues that a large cloud's size has the ability to repel botnets and DDoS attacks better than smaller systems do.

7. Space-time is a continuum (Einstein/Minkowski).

The ability of a task to be accomplished in the cloud using parallel processing allows real time business to respond quicker to business conditions and accelerates decision making providing a measurable advantage.

8. Dispersion is the inverse square of latency.

Latency, or the delay in getting a response to a request, requires both large-scale and multi-site deployments that are a characteristic of cloud providers. Cutting latency in half requires four times the number of nodes in a system.

9. Don't put all your eggs in one basket.

The reliability of a system with *n* redundant components and a reliability of r is 1-(1-*r*)*n*. Therefore, when a datacenter achieves a reliability of 99 percent, two redundant datacenters have a reliability of 99.99 percent (four nines) and three redundant datacenters can achieve a reliability of 99.9999 percent (six nines). Large cloud providers with geographically dispersed sites worldwide therefore achieve reliability rates that are hard for private systems to achieve.

10. An object at rest tends to stay at rest (Newton).

Private datacenters tend to be located in places where the company or unit was founded or acquired. Cloud providers can site their datacenters in what are called "greenfield sites." A *greenfield site* is one that is environmentally friendly: locations that are on a network backbone, have cheap access to power and cooling, where land is inexpensive, and the environmental impact is low. A network backbone is a very high-capacity network connection. On the Internet, an Internet backbone consists of the high-capacity routes and routers that are typically operated by an individual service provider such as a government or commercial entity. You can access a jump page of Internet backbone maps at: http://www. nthelp.com/maps.htm.

Cloud computing obstacles

Cloud computing isn't a panacea; nor is it either practical or economically sensible for many computer applications that you encounter. In practice, cloud computing can deviate from the ideal described in the previous list in many significant ways. The illusion of scalability is bounded by the limitations cloud providers place on their clients. Resource limits are exposed at peak conditions of the utility itself. As we all know, power utilities suffer brownouts and outages when the temperature soars, and cloud computing providers are no different. You see these outages on peak computing days such as Black Monday, which is the Monday after Thanksgiving in the United States when Internet Christmas sales traditionally start. The illusion of low barrier to entry may be pierced by an inconsistent pricing scheme that makes scaling more expensive than it should be. You can see this limit in the nonlinearity of pricing associated with "extra large" machine instances versus their "standard" size counterparts. Additionally, the low barrier to entry also can be accompanied by a low barrier to provisioning. If you make a provisioning error, it can lead to vast costs. Cloud computing vendors run very reliable networks. Often, cloud data is load-balanced between virtual systems and replicated between sites. However, even cloud providers experience outages. In the cloud, it is common to have various resources, such as machine instances, fail. Except for tightly managed PaaS cloud providers, the burden of resource management is still in the hands of the user, but the user is often provided with limited or immature management tools to address these issues.

Measuring cloud computing costs

As you see, cloud computing has many advantages and disadvantages, and you can't always measure. You can measure costs though, and that's a valuable exercise. Usually a commodity is cheaper than a specialized item, but not always. Depending upon your situation, you can pay more for public cloud computing than you would for owning and managing your private cloud, or for owning and using software as well. That's why it's important to analyze the costs and benefits of your own cloud computing scenario carefully and quantitatively. You will want to compare the costs of cloud computing to private systems. The cost of a cloud computing deployment is roughly estimated to be

CostCLOUD = Σ(UnitCostCLOUD x (Revenue – CostCLOUD))

where the unit cost is usually defined as the cost of a machine instance per hour or another resource. Depending upon the deployment type, other resources add additional unit costs: storage quantity consumed, number of transactions, incoming or outgoing amounts of data, and so forth. Different cloud providers charge different amounts for these resources, some resources are free for one provider and charged for another, and there are almost always variable charges based on resource sizing.

Cloud resource pricing doesn't always scale linearly based on performance. To compare your cost benefit with a private cloud, you will want to compare the value you determine

in the equation above with the same calculation:

CostDATACENTER = Σ(UnitCostDATACENTER x (Revenue – (CostDATACENTER/Utilization))

Notice the additional term for Utilization added as a divisor to the term for CostDATACENTER. This term appears because it is assumed that a private cloud has capacity that can't be captured, and it is further assumed that a private cloud doesn't employ the same level of virtualization or pooling of resources that a cloud computing provider can achieve. Indeed, no system can work at 100 percent utilization because queuing theory states that as the system approaches 100 percent, the latency and response times go to infinity. Typical efficiencies in datacenters are between 60 and 85 percent. It is also further assumed that the datacenter is operating under averaged loads (not at peak capacity) and that the capacity of the datacenter is fixed by the assets it has. The costs associated with the cloud model are calculated rather differently. Each resource has its own specific cost and many resources can be provisioned independently of one another. In theory, therefore, the CostCLOUD is better represented by the equation:

CostCLOUD = 1

(UnitCostCLOUD x (Revenue – CostCLOUD))INSTANCE*n* +

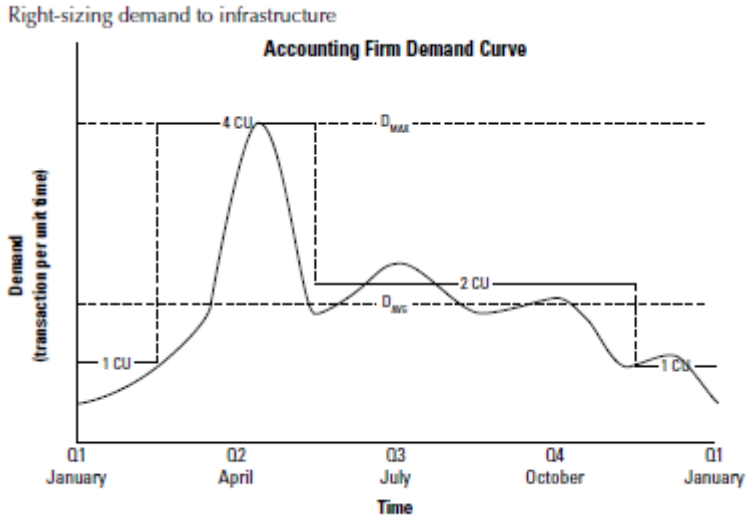
(UnitCostCLOUD x (Revenue – CostCLOUD))STORAGE_UNIT*n* +.

(UnitCostCLOUD x (Revenue – CostCLOUD))NETWORK_UNIT*n* + …

Right-sizing

Consider an accounting firm with a variable demand load, as shown in Figure 2.2. For each of the four quarters of the tax year, clients file their quarterly taxes on the service's Web site. Demand for three of those quarters rises broadly as the quarterly filing deadline arrives. The fourth quarter that represents the year-end tax filing on April 15 shows a much larger and more pronounced spike for the two weeks approaching and just following that quarter's end. Clearly, this accounting business can't ignore the demand spike for its year-end accounting, because this is the single most important

portion of the firm's business, but it needs to match demand to resources to maximize its profits.



Buying and leasing infrastructure to accommodate the peak demand (or alternatively load) shown in the figure as DMAX means that nearly half of that infrastructure remains idle for most of the time. Fitting the infrastructure to meet the average demand, DAVG, means that half of the transactions in the Q2 spike are not captured, which is the mission critical portion of this enterprise. More accurately using DAVG means that during maximum demand the service is slowed to a crawl and the system may not be responsive enough to satisfy any of the users. These limits can be a serious constraint on profit and revenue. Outsourcing the demand may provide a solution to the problem. But outsourcing essentially shifts the burden

of capital expenditures onto the service provider. A service contract that doesn't match infrastructure to demand suffers from the same inefficiencies that captive infrastructure does.

The cloud computing model addresses this problem by allowing you to right-size your infrastructure. In Figure 2.2, the demand is satisfied by an infrastructure that is labeled in terms of a CU or "Compute Unit." The rule for this particular cloud provider is that infrastructure may be modified at the beginning of any month. For the low-demand Q1/Q4 time period, a 1 CU infrastructure is applied. On February 1st, the size is changed to a 4 CU infrastructure, which captures the entire spike of Q2 demand. Finally, on June 1st, a 2 CU size is applied to accommodate the typical demand DAVG that is experienced in the last half of Q2 through the middle of Q4. This curve fitting exercise captures the demand nearly all the with little idle capacity left unused.