[Unit 1: Introduction]

Cloud Computing (CSC-458)

Jagdish Bhatta

Central Department of Computer Science & Information Technology

Tribhuvan University

Introduction:

What is Cloud?

The term *cloud* has been used historically as a metaphor for the Internet and has become a familiar cliché. This usage was originally derived from its common depiction in network diagrams as an outline of a cloud, used to represent the transport of data across carrier backbones (which owned the cloud) to an endpoint location on the other side of the cloud.

The cloud itself is a set of hardware, networks, storage, services, and interfaces that enable the delivery of computing as a service. Cloud services include the delivery of software, infrastructure, and storage over the internet (either as separate components or a complete platform) based on user demand.

The world of the cloud has lots of participants:

✓ The **end user** doesn't really have to know anything about the underlying technology. In small businesses, for example, the cloud provider becomes the de facto data center. In larger organizations, the IT organization oversees the inner workings of both internal resources and external cloud resources.

✓ **Business management** needs to take responsibility for overall governance of data or services living in a cloud. Cloud service providers must provide a predictable and guaranteed service level and security to all their constituents.

✓ The **cloud service provider** is responsible for IT assets and maintenance.

Overall, the cloud embodies the following four basic characteristics:

- ✓ **Elasticity and the ability to scale up and down**
- ✓ **Self-service provisioning and automatic deprovisioning**
- ✓ **Application programming interfaces (APIs)**
- ✓ **Billing and metering of service usage in a pay-as-you-go model**

Elasticity and scalability:

The service provider can't anticipate how customers will use the service. One customer might use the service three times a year during peak selling seasons, whereas another might use it as a primary development platform for all of its applications. Therefore, the service needs to be available all the time (7 days a week, 24 hours a day) and it has to be designed to scale upward for high periods of demand and downward for lighter ones. ***Scalability*** **also means that an application can scale when additional users are added and** when the application requirements change**. This ability to scale is achieved by providing *elasticity*.** Think about the rubber band and its properties. If you're holding together a dozen pens with a rubber band, you probably have to fold it in half. However, if you're trying to keep 100 pens together, you will have to stretch that rubber band. Why can a single rubber band accomplish both tasks? Simply, it is elastic and so is the cloud. : **Elasticity refers to the ability to flex to meet the needs and preferences of users on a near real-time basis, in response to supply and demand triggers.** In the cloud context, elasticity refers to the ability of a service or an infrastructure to adjust to meet fluctuating service demands by automatically provisioning or de-provisioning resources or by moving the service to be executed on another part of the system.

Self-service provisioning:

Customers can easily get cloud services without going through a lengthy process. The customer simply requests an amount of computing, storage, software, process, or other resources from the service provider. Contrast this on-demand response with the process at a typical data center. When a department is about to implement a new application, it has to submit a request to the data center for additional computing hardware, software, services,

or process resources. The data center gets similar requests from departments across the company and must sort through all requests and evaluate the availability of existing resources versus the need to purchase new hardware. After new hardware is purchased, the data center staff has to configure the data center for the new application. These internal procurement processes can take a long time, depending on company policies. Of course, nothing is as simple as it might appear. **While the on-demand provisioning capabilities of cloud services eliminate many time delays**, an organization still needs to do its homework. These services aren't free; needs and requirements must be determined before capability is automatically provisioned.

Application programming interfaces (APIs):

Cloud services need to have standardized APIs. These interfaces provide the instructions on how two application or data sources can communicate with each other. A standardized interface lets the customer more easily link a cloud service, such as a customer relationship management system with a financial accounts management system, without having to resort to custom programming.

Billing and metering of services:

A cloud environment needs a built-in service that bills customers. And, of course, to calculate that bill, usage has to be *metered* (tracked). Even free cloud services (such as Google's Gmail or Zoho's Internet-based office applications) are metered.

Cloud Computing:

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation.

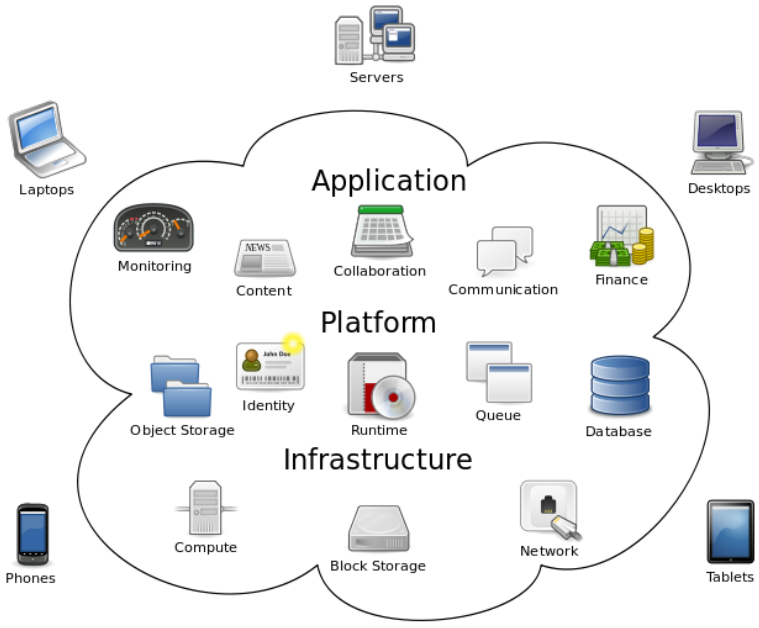
Cloud computing is a general term for anything that involves delivering hosted services over the Internet. These services are broadly divided into three categories: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). The name cloud computing was inspired by the cloud symbol that's often used to represent the Internet in flowcharts and diagrams.

A cloud service has three distinct characteristics that differentiate it from traditional hosting.

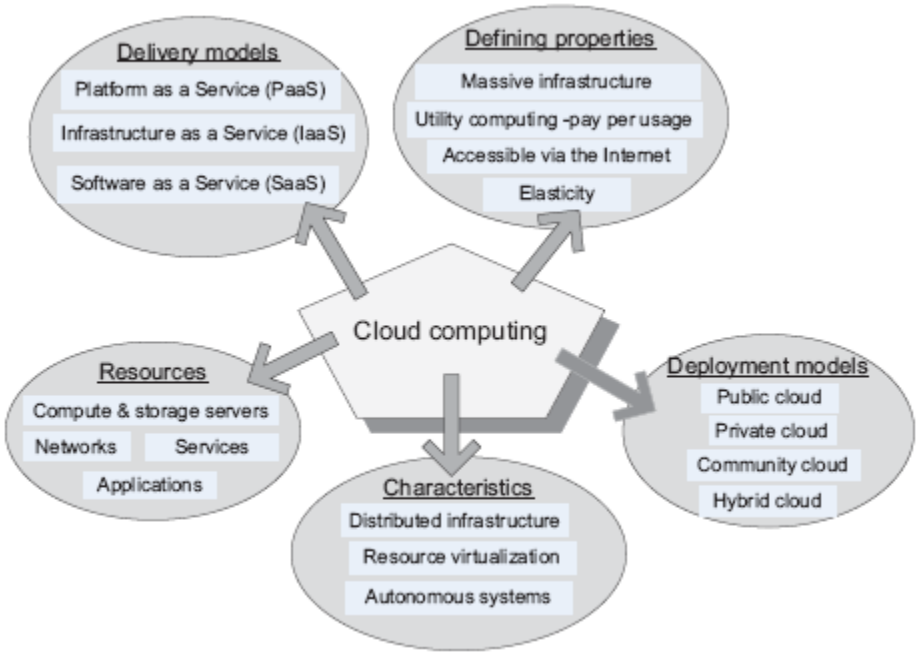
- **It is sold on demand,** typically by the minute or the hour;
- **It is elastic** -- a user can have as much or as little of a service as they want at any given time; and
- **The service is fully managed by the provider** (the consumer needs nothing but a personal computer and Internet access). Significant innovations in virtualization and distributed computing, as well as improved access to high-speed Internet and a weak economy, have accelerated interest in cloud computing.

A more tempered view of cloud computing considers it the delivery of computational resources from a location other than the one from which you are computing.

The service consumer no longer has to be at a PC, use an application from the PC, or purchase a specific version that's configured for smartphones, PDAs, and other devices. The consumer does not own the infrastructure, software, or platform in the cloud. He/She has lower upfront costs, capital expenses, and operating expenses. He/She does not care about how servers and networks are maintained in the cloud. The consumer can access multiple servers anywhere on the globe without knowing which ones and where they are located.



Cloud Computing



Emergence of Cloud Computing:

The origin of the term *cloud computing* is obscure, but it appears to derive from the practice of using drawings of stylized clouds to denote networks in diagrams of computing and communications systems. The word *cloud* is used as a metaphor for the Internet, based on the standardized use of a cloud-like shape to denote a network on telephony schematics and later to depict the Internet in computer network diagrams as an abstraction of the underlying infrastructure it represents. The cloud symbol was used to represent the Internet as early as 1994.

In the 1990s, telecommunications companies, who previously offered primarily dedicated point-to-point data circuits, began offering virtual private network (VPN) services with comparable quality of service but at a much lower cost. By switching traffic to balance utilization as they saw fit, they were able to utilize their overall network bandwidth more effectively. The cloud symbol was used to denote the demarcation point between that which was the responsibility of the provider and that which was the responsibility of the users. Cloud computing extends this boundary to cover servers as well as the network infrastructure.

The underlying concept of cloud computing dates back to the 1950s; when large-scale mainframe became available in academia and corporations, accessible via thin clients /terminal computers. Because it was costly to buy a mainframe, it became important to find ways to get the greatest return on the investment in them, allowing multiple users to share both the physical access to the computer from multiple terminals as well as to share the CPU time, eliminating periods of inactivity, which became known in the industry as time-sharing.

As in the earliest stages, the term "cloud" was used to represent the computing space between the provider and the end user. **In 1997, Professor Ramnath Chellapa of Emory University and the University of South California defined cloud computing as the new "computing paradigm where the boundaries of computing will be determined by**

economic rationale rather than technical limits alone." This has become the basis of what we refer to today when we discuss the concept of cloud computing.

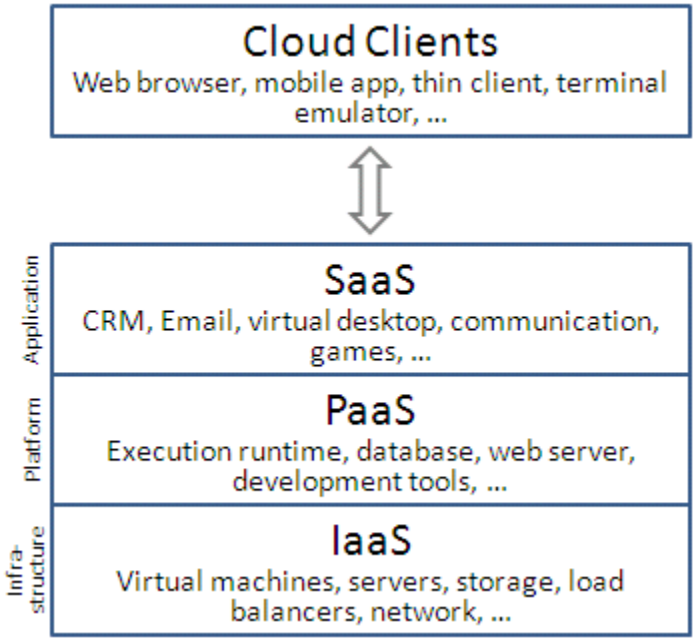
Some people think cloud computing is the next big thing in the world of IT. Others believe it is just another variation of the utility computing model that has been repackaged in this decade as something new and cool.

One of the first milestones for cloud computing was the arrival of Salesforce.com in 1999, which pioneered the concept of delivering enterprise applications via a simple website. The services firm paved the way for both specialist and mainstream software firms to deliver applications over the internet.

The next development was Amazon Web Services in 2002, which provided a suite of cloud-based services including storage, computation and even human intelligence through the Amazon Mechanical Turk. Then in 2006, Amazon launched its Elastic Compute cloud (EC2) as a commercial web service that allows small companies and individuals to rent computers on which to run their own computer applications.

Cloud Based Service Models:

The services given by the service provider to the customers or users through cloud computing technology are said cloud services. Service Provider's server gives both the hardware and software necessary and thus easy in management for both user and the cloud service provider. **Cloud computing providers offer their services according to three fundamental models: Infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) where IaaS is the most basic and each higher model abstracts from the details of the lower models.**



IaaS: In this most basic cloud service model, providers offer computers, as physical or more often as virtual machines, and other resources. The virtual machines are run as guests by a hypervisor, such as Xen or KVM. Management of pools of hypervisors by the cloud operational support system leads to the ability to scale to support a large number of virtual machines. **Other resources in IaaS clouds include images in a virtual machine image library, raw (block) and file-based storage, firewalls, load balancers, IP addresses, virtual local area networks (VLANs), and software bundles.** Amies, Alex; Sluiman, Harm; Tong IaaS cloud providers supply these resources on demand from their large pools installed in data centers. For wide area connectivity, the Internet can be used or—in carrier clouds -- dedicated virtual private networks can be configured

To deploy their applications, cloud users then install operating system images on the machines as well as their application software. In this model, it is the cloud user who is responsible for patching and maintaining the operating systems and application software. **Cloud providers typically bill IaaS services on a utility computing basis, that is, cost will reflect the amount of resources allocated and consumed. STaaS** - STorage As A Service. This service comes under IaaS, which manages all the storage services in cloud

computing. There are many security issues in this service. They are Data Integrity, Confidentiality, Reliability, etc.

IaaS refers not to a machine that does all the work, but simply to a facility given to businesses that offers users the leverage of extra storage space in servers and data centers.

Examples of IaaS include: Amazon CloudFormation (and underlying services such as Amazon EC2), Rackspace Cloud, Terremark, Windows Azure Virtual Machines, Google Compute Engine, and Joyent.

PaaS: In the PaaS model, cloud providers deliver a computing platform typically including operating system, programming language execution environment, database, and web server. Application developers can develop and run their software solutions on a cloud platform without the cost and complexity of buying and managing the underlying hardware and software layers. **With some PaaS offers, the underlying computer and storage resources scale automatically to match application demand such that cloud user does not have to allocate resources manually.**

Examples of PaaS include: Amazon Elastic Beanstalk, Cloud Foundry, Heroku, Force.com, EngineYard, Mendix, Google App Engine, Windows Azure Compute and OrangeScape.

SaaS: In this model, cloud providers install and operate application software in the cloud and cloud users access the software from cloud clients. The cloud users do not manage the cloud infrastructure and platform on which the application is running. This eliminates the need to install and run the application on the cloud user's own computers simplifying maintenance and support. What makes a cloud application different from other applications is its scalability. This can be achieved by cloning tasks onto multiple virtual machines at run-time to meet the changing work demand. Load balancers distribute the work over the set of virtual machines. This process is transparent to the cloud user who sees only a single access point. **To accommodate a large number of**

cloud users, cloud applications can be multitenant, that is, any machine serves more than one cloud user organization. It is common to refer to special types of cloud based application software with a similar naming convention: desktop as a service, business process as a service, test environment as a service, communication as a service.

The pricing model for SaaS applications is typically a monthly or yearly flat fee per user, so price is scalable and adjustable if users are added or removed at any point.

Examples of SaaS include: Google Apps, Microsoft Office 365, and Onlive.

Grid Computing or Cloud Computing?

Grid computing is often confused with cloud computing. Grid computing is a form of distributed computing that implements a *virtual supercomputer* made up of a cluster of networked or Internetworked computers acting in unison to perform very large tasks. Many cloud computing deployments today are powered by grid computing implementations and are billed like utilities, but cloud computing can and should be seen as an evolved next step away from the grid utility model. There is an ever-growing list of providers that have successfully used cloud architectures with little or no centralized infrastructure or billing systems, such as the peer-to-peer network BitTorrent and the volunteer computing initiative SETI@home.

Service commerce platforms are yet another variation of SaaS and MSPs. This type of cloud computing service provides a centralized service hub that users interact with. Currently, the most often used application of this platform is found in financial trading environments or systems that allow users to order things such as travel or personal services from a common platform (e.g., Expedia.com or Hotels.com), which then coordinates pricing and service delivery within the specifications set by the user.

Cloud computing evolves from grid computing and provides on-demand resource provisioning. Grid computing may or may not be in the cloud depending on what type of users are using it. If the users are systems administrators and integrators, they care how

things are maintained in the cloud. They upgrade, install, and virtualize servers and applications. If the users are consumers, they do not care how things are run in the system.

The **difference between grid computing and cloud computing** is hard to grasp because they are not always mutually exclusive. In fact, they are both used to economize computing by maximizing existing resources. However, **the difference between the two lies in the way the tasks are computed in each respective environment**. In a computational grid, one large job is divided into many small portions and executed on multiple machines. This characteristic is fundamental to a grid; not so in a cloud. The computing in cloud is intended to allow the user to avail of various services without investing in the underlying architecture. Cloud services include the delivery of software, infrastructure, and storage over the Internet (either as separate components or a complete platform) based on user demand.

Cloud computing and grid computing are scalable. Scalability is accomplished through load balancing of application instances running separately on a variety of operating systems and connected through Web services. CPU and network bandwidth is allocated and de-allocated on demand. The system's storage capacity goes up and down depending on the number of users, instances, and the amount of data transferred at a given time.

Both computing types involve **multitenancy and multitask**, meaning that many customers can perform different tasks, accessing a single or multiple application instances. Sharing resources among a large pool of users assists in reducing infrastructure costs and peak load capacity. Cloud and grid computing provide service-level agreements (SLAs) for guaranteed uptime availability of, say, 99 percent. If the service slides below the level of the guaranteed uptime service, the consumer will get service credit for receiving data late. The Amazon S3 provides a Web services interface for the storage and retrieval of data in the cloud. Setting a maximum limits the number of objects you can store in S3. You can store an object as small as 1 byte and as large as 5 GB or even several terabytes. S3 uses the concept of buckets as containers for each storage location of your objects. The data is

stored securely using the same data storage infrastructure that Amazon uses for its e-commerce Web sites.

While the storage computing in the grid is well suited for data-intensive storage, it is not economically suited for storing objects as small as 1 byte. In a data grid, the amounts of distributed data must be large for maximum benefit. A computational grid focuses on computationally intensive operations. Amazon Web Services in cloud computing offers two types of instances: standard and high-CPU.

(Refer the research article, provided in the class, entitled "Cloud Computing and Grid Computing 360-Degree Compared", *by the authors; Ian Foster, Yong Zhao, Ioan Raicu, Shiyong Lu, published in 2008).*

Virtualization:

Virtualization is the key to cloud computing, since it is the enabling technology allowing the creation of an intelligent abstraction layer which hides the complexity of underlying hardware or software.

Server virtualization enables different operating systems to share the same hardware and make it easy to move operating systems between different hardware, all while the applications are running.

Storage virtualization does the same thing for data. Storage virtualization creates the abstraction layer between the applications running on the servers, and the storage they use to store the data.

Virtualizing the storage and incorporating the intelligence for provisioning and protection at the virtualization layer enables companies to use any storage they want, and not be locked into any individual vendor. Storage virtualization makes storage a commodity. All this makes for some interesting ways for companies to reduce their costs.

Any discussion of cloud computing typically begins with virtualization. **Virtualization is critical to cloud computing because it simplifies the delivery of services by providing a platform for optimizing complex IT resources in a scalable manner, which is what makes cloud computing so cost effective.**

Virtualization can be applied very broadly to just about everything you can imagine including memory, networks, storage, hardware, operating systems, and applications. Virtualization has three characteristics that make it ideal for cloud computing:

- **Partitioning:** In virtualization, you can use partitioning to support many applications and operating systems in a single physical system.
- **Isolation:** Because each virtual machine is isolated, each machine is protected from crashes and viruses in the other machines. What makes virtualization so important for the cloud is that it decouples the software from the hardware.
- **Encapsulation:** Encapsulation can protect each application so that it doesn't interfere with other applications. Using encapsulation, a virtual machine can be represented (and even stored) as a single file, making it easy to identify and present to other applications.

To understand how virtualization helps with cloud computing, you must understand its many forms. In essence, in all cases, a resource actually emulates or imitates another resource. Here are some examples:

- **Virtual memory:** Disks have a lot more space than memory. PCs can use virtual memory to borrow extra memory from the hard disk. Although virtual disks are slower than real memory, if managed right, the substitution works surprisingly well.
- **Software:** There is virtualization software available that can emulate an entire computer, which means 1 computer can perform as though it were actually 20 computers. Using this kind of software you might be able to move from a data center with thousands of servers to one that supports as few as a couple of hundred.

To manage the various aspects of virtualization in cloud computing most companies use ***hypervisors, an operating system*** *that act as traffic cop managing the various*

virtualization tasks in the cloud to ensure that they make the things happen in an orderly manner. Because in cloud computing you need to support many different operating environments, the hypervisor becomes an ideal delivery mechanism by allowing you to show the same application on lots of different systems. Because hypervisors can load multiple operating systems, they are a very practical way of getting things virtualized quickly and efficiently.

Cloud Computing Deployment Models (Types):

Public: The infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services. It is also referred to as 'external' Cloud that describes the conventional meaning of Cloud computing: scalable, dynamically provisioned, often virtualized resources available over the internet from an off-site third party provider, which divides up resources and bills its customers on a 'utility' basis.

Public cloud applications, storage, and other resources are made available to the general public by a service provider. These services are free or offered on a pay-per-use model. Generally, public cloud service providers like Amazon AWS, Microsoft and Google own and operate the infrastructure and offer access only via Internet (direct connectivity is not offered).

Private: The infrastructure is operated solely for an organization; it may be managed by the organization or a third party and may exist on or off the premises of the organization. It is also referred to as 'corporate' or 'internal' Cloud, term used to denote a proprietary computing architecture providing hosted services on private networks.

A private cloud could provide the computing resources needed for a large organization, e.g., a research institution, a university, or a corporation. There are some arguments that a private cloud does not support utility computing when the user pays as it consumes resources.

Undertaking a private cloud project requires a significant level and degree of engagement to virtualize the business environment, and it will require the organization to reevaluate decisions about existing resources. When it is done right, it can have a positive impact on a business, but every one of the steps in the project raises security issues that must be addressed in order to avoid serious vulnerabilities

Community: Community cloud shares infrastructure between several organizations from a specific community with common concerns (security, compliance, jurisdiction, etc.), whether managed internally or by a third-party and hosted internally or externally. The costs are spread over fewer users than a public cloud (but more than a private cloud), so only some of the cost savings potential of cloud computing are realized

Hybrid: Here, the infrastructure is a composition of two or more clouds (private, 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). Cloud bursting is an application deployment model in which an application runs in a private cloud or data center and bursts into a public cloud when the demand for computing capacity spikes. The advantage of such a hybrid cloud deployment is that an organization only pays for extra compute resources when they are needed. Experts recommend cloud bursting for high performance, non-critical applications that handle non-sensitive information. An application can be deployed locally and then burst to the cloud to meet peak demands, or the application can be moved to the public cloud to free up local resources for business-critical applications. Cloud bursting works best for applications that don't depend on a complex application delivery infrastructure or integration with other applications, components and systems internal to the data center.

By utilizing "hybrid cloud" architecture, companies and individuals are able to obtain degrees of fault tolerance combined with locally immediate usability without dependency on internet connectivity. Hybrid cloud architecture requires both on-premises resources and off-site (remote) server-based cloud infrastructure.

Hybrid clouds lack the flexibility, security and certainty of in-house applications. Hybrid cloud provides the flexibility of in house applications with the fault tolerance and scalability of cloud based services.

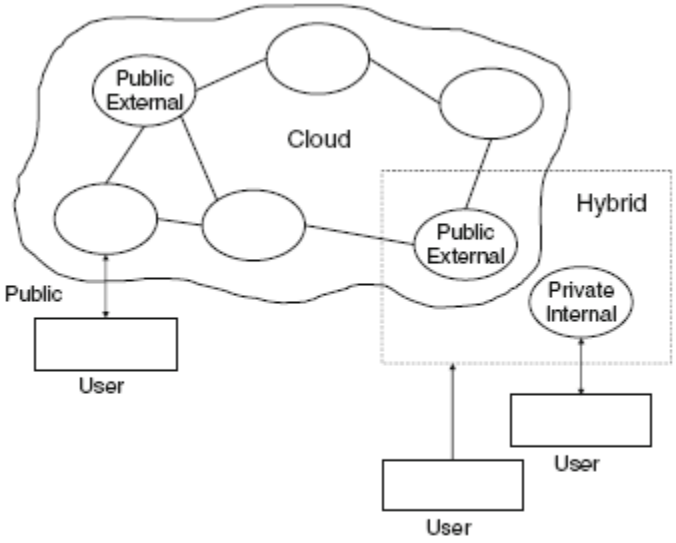


Fig: Types of Cloud Deployment Models

Ethical issues in cloud computing:

Cloud computing is based on a paradigm shift with profound implications on computing ethics. **The main elements of this shift are:**

(i) the control is relinquished to third party services;

(ii) the data is stored on multiple sites administered by several organizations; and

(iii) multiple services interoperate across the network.

Unauthorized access, data corruption, infrastructure failure, or unavailability are some of the risks related to relinquishing the control to third party services; moreover, it is difficult to identify the source of the problem and the entity causing it. Systems can span the boundaries of multiple organizations and cross the security borders, a process called *deperimeterisation*. As a result of de-perimeterisation "not only the border of the organizations IT infrastructure blurs, also the border of the accountability becomes less clear".

The complex structure of cloud services can make it difficult to determine who is responsible in case something undesirable happens. In a complex chain of events or systems, many entities contribute to an action with undesirable consequences, some of them have the opportunity to prevent these consequences, and therefore no one can be held responsible, the so-called "problem of many hands."

Ubiquitous and unlimited data sharing and storage among organizations test the selfdetermination of information, the right or ability of individuals to exercise personal control over the collection, use and disclosure of their personal data by others; this tests the confidence and trust in todays evolving information society. Identity fraud and theft are made possible by the unauthorized access to personal data in circulation and by new forms of dissemination through social networks and they could also pose a danger to cloud computing.

The question of what can be done proactively about ethics of cloud computing does not have easy answers as many undesirable phenomena in cloud computing will only appear in time. But the need for rules and regulations for the governance of cloud computing are obvious. The term *governance* means the manner in which something is governed or regulated, the method of management, the system of regulations. Explicit attention to ethics must be paid by governmental organizations providing research funding; private companies are less constraint by ethics oversight and governance arrangements are more conducive to profit generation.

Accountability is a necessary ingredient of cloud computing; adequate information about how data is handled within the cloud and about allocation of responsibility are key elements to enforcing ethics rules in cloud computing. Recorded evidence allows us to assign responsibility; but there can be tension between privacy and accountability and it is important to establish what is being recorded, and who has access to the records.

Unwanted dependency on a cloud service provider, the so-called *vendor lock-in*, is a serious concern and the current standardization efforts at NIST attempt to address

this problem. Another concern for the users is a future with only a handful of companies which dominate the market and dictate prices and policies

Benefits of using cloud models:

Because customers generally do not own the infrastructure used in cloud computing environments, they can forgo capital expenditure and consume resources as a service by just paying for what they use. Many cloud computing offerings have adopted the utility computing and billing model described above, while others bill on a subscription basis. **By sharing computing power among multiple users, utilization rates are generally greatly improved, because cloud computing servers are not sitting dormant for lack of use.** This factor alone can reduce infrastructure costs significantly and accelerate the speed of applications development.

A beneficial side effect of using this model is that computer capacity increases dramatically, since customers do not have to engineer their applications for peak times, when processing loads are greatest. Adoption of the cloud computing model has also been enabled because of the greater availability of increased high-speed bandwidth. With greater enablement, though, there are other issues one must consider, especially legal ones.

The following are some of the possible benefits for those who offer cloud computing-based services and applications:

- **Cost Savings —** Companies can reduce their capital expenditures and use operational expenditures for increasing their computing capabilities. This is a lower barrier to entry and also requires fewer in-house IT resources to provide system support.

- **Scalability/Flexibility —** Companies can start with a small deployment and grow to a large deployment fairly rapidly, and then scale back if necessary. Also, the flexibility of cloud computing allows companies to use extra resources at peak times, enabling them to satisfy consumer demands.

- **Reliability —** Services using multiple redundant sites can support business continuity and disaster recovery.

- **Maintenance —** Cloud service providers do the system maintenance, and access is through APIs that do not require application installations onto PCs, thus further reducing maintenance requirements.

- **Mobile Accessible —** Mobile workers have increased productivity due to systems accessible in an infrastructure available from anywhere.

Characteristics of Cloud Computing:

Cloud computing has a variety of characteristics, with the main ones being:

- **Shared Infrastructure —** Uses a virtualized software model, enabling the sharing of physical services, storage, and networking capabilities. The cloud infrastructure, regardless of deployment model, seeks to make the most of the available infrastructure across a number of users.

- **Dynamic Provisioning / on demand self service —** Allows for the provision of services based on current demand requirements. This is done automatically using software automation, enabling the expansion and contraction of service capability, as needed. This dynamic scaling needs to be done while maintaining high levels of reliability and security.

- **Broad Network Access —** Needs to be accessed across the internet from a broad range of devices such as PCs, laptops, and mobile devices, using standards-based APIs (for example, ones based on HTTP). Deployments of services in the cloud include everything from using business applications to the latest application on the newest smartphones.

- **Multi-Tenant Capable** - The resources (e.g., network, storage and compute power) can be shared among multiple enterprise clients, thereby lowering overall expense. Resource virtualization is used to enforce isolation and aid in security.

- **Rapid Elasticity** - The consumer should have the ability to rapidly (often automatically) increase or decrease the computing resources needed to carry out their work.

- **Managed Metering / Measured Service** - Uses metering for managing and optimizing the service and to provide reporting and billing information. In this way, consumers are billed for services according to how much they have actually used during the billing period. In short, cloud computing allows for the sharing and scalable deployment of services, as needed, from almost any location, and for which the customer can be billed based on actual usage.

Evolution of cloud computing:

Cloud computing can be seen as an innovation in different ways. From a technological perspective it is an advancement of computing, applying virtualization concepts to utilize hardware more efficiently. Yet a different point of view is to look at cloud computing from an IT deployment perspective. In this sense cloud computing has the potential to revolutionize the way, how computing resources and applications are provided, breaking up traditional value chains and making room for new business models. In the following section we are going to describe the emergence of cloud computing from both perspectives.

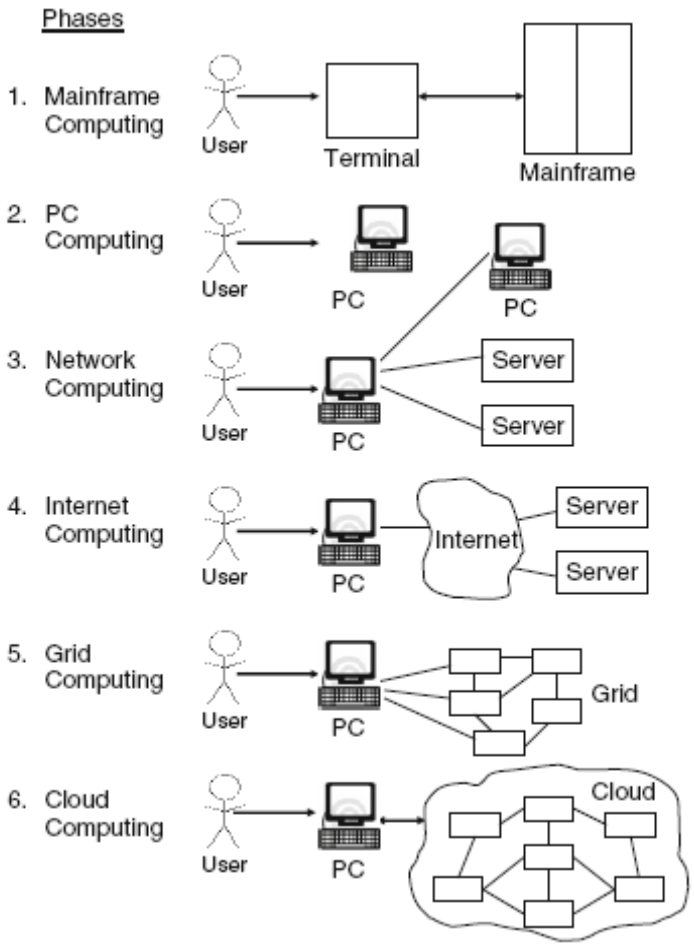


Fig: Evolution of Computing Paradigms from mainframe to cloud computing

Challenges for cloud computing

The development of efficient cloud applications inherits the challenges posed by the natural imbalance between computing, I/O, and communication bandwidths of physical systems; these challenges are greatly amplified due the scale of the system, its distributed nature, and by the fact that virtually all applications are data-intensive. Though cloud computing infrastructures attempt to automatically distribute and balance the load, the application developer is still left with the responsibility to place the data close to the processing site and to identify optimal storage for the data. One of the main advantages of cloud computing, the shared infrastructure, could also have a negative impact as perfect performance isolation is nearly impossible to reach in a real system, especially when the

system is heavily loaded. The performance of virtual machines fluctuates based on the load, the infrastructure services, the environment including the other users. Reliability is also a major concern; node failures are to be expected whenever a large numbers of nodes cooperate for the computations. Choosing an optimal instance, in terms of performance isolation, reliability, and security, from those offered by the cloud infrastructure is another critical factor to be considered. Of course, cost considerations play also a role in the choice of the instance type. Many applications consist of multiple stages; in turn, each stage may involve multiple instances running in parallel on the systems of the cloud and communicating among them. Thus, efficiency, consistency, and communication scalability of communication are major concerns for an application developer. Indeed, due to shared networks and unknown topology, cloud infrastructures exhibit inter-node latency and bandwidth fluctuations which affect the application performance.

The following are some of the notable challenges associated with cloud computing, and although some of these may cause a slowdown when delivering more services in the cloud, most also can provide opportunities, if resolved with due care and attention in the planning stages.

- **Security and Privacy —** Perhaps two of the more “hot button” issues surrounding cloud computing relate to storing and securing data, and monitoring the use of the cloud by the service providers. These issues are generally attributed to slowing the deployment of cloud services. These challenges can be addressed, for example, by storing the information internal to the organization, but allowing it to be used in the cloud. For this to occur, though, the security mechanisms between organization and the cloud need to be robust and a Hybrid cloud could support such a deployment.

- **Lack of Standards —** Clouds have documented interfaces; however, no standards are associated with these, and thus it is unlikely that most clouds will be interoperable. The Open Grid Forum is developing an Open Cloud Computing Interface to resolve this issue and the Open Cloud Consortium is working on cloud computing standards and practices. The findings of these groups will need to

mature, but it is not known whether they will address the needs of the people deploying the services and the specific interfaces these services need. However, keeping up to date on the latest standards as they evolve will allow them to be leveraged, if applicable.

- **Continuously Evolving —** User requirements are continuously evolving, as are the requirements for interfaces, networking, and storage. This means that a "cloud," especially a public one, does not remain static and is also continuously evolving.

- **Compliance Concerns —** The Sarbanes-Oxley Act (SOX) in the US and Data Protection directives in the EU are just two among many compliance issues affecting cloud computing, based on the type of data and application for which the cloud is being used. The EU has a legislative backing for data protection across all member states, but in the US data protection is different and can vary from state to state. As with security and privacy mentioned previously, these typically result in Hybrid cloud deployment with one cloud storing the data internal to the organization.

Distributed Computing in Grid and Cloud :

Distributed computing is a field of computer science that studies distributed systems. **A distributed system consists of multiple autonomous computers that communicate through a computer network. The computers interact with each other in order to achieve a common goal.**

Distributed computing also refers to the use of distributed systems to solve computational problems. In distributed computing, a problem is divided into many tasks, each of which is solved by one or more computers, which communicate with each other by message passing.

The word *distributed* in terms such as "distributed system", "distributed programming", and "distributed algorithm" originally referred to computer networks where individual computers were physically distributed within some geographical area. The terms are nowadays used in a much wider sense, even referring to autonomous processes that run on the same physical computer and interact with each other by message passing. While there is no single definition of a distributed system, the following defining properties are commonly used:

- There are several autonomous computational entities, each of which has its own local memory.
- The entities communicate with each other by message passing

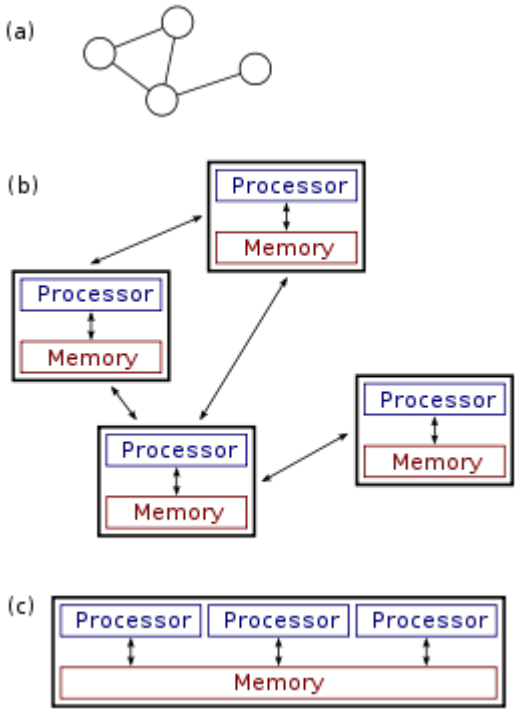
Parallel Vs. Distributed Computing

Distributed systems are groups of networked computers, which have the same goal for their work. The terms "concurrent computing", "parallel computing", and "distributed computing" have a lot of overlap, and no clear distinction exists between them. The same system may be characterized both as "parallel" and "distributed"; the processors in a typical distributed system run concurrently in parallel. Parallel computing may be seen as a particular tightly coupled form of distributed computing, and distributed computing may be seen as a loosely coupled form of parallel computing. Nevertheless, it is possible to roughly classify concurrent systems as "parallel" or "distributed" using the following criteria:

- In parallel computing, all processors may have access to a shared memory to exchange information between processors.
- In distributed computing, each processor has its own private memory (distributed memory). Information is exchanged by passing messages between the processors.

The figure below illustrates the difference between distributed and parallel systems. Figure (a) is a schematic view of a typical distributed system; as usual, the system is represented as a network topology in which each node is a computer and each line connecting the

nodes is a communication link. Figure (b) shows the same distributed system in more detail: each computer has its own local memory, and information can be exchanged only by passing messages from one node to another by using the available communication links. Figure (c) shows a parallel system in which each processor has a direct access to a shared memory.



The last decade, the term 'Grid' has been a key topic in the field of high performance/distributed computing. The Grid has emerged as a new field of distributed computing, focusing on secure sharing of computational and storage resources among dynamic sets of people and organizations who own these resources. This sharing of resources can give people not only computational capabilities and data storage capabilities that cannot be provided by a single supercomputing center, but it also allows them to share data in a transparent way.

Grid Computing can be defined as applying resources from many computers in a network to a single problem, usually one that requires a large number of processing cycles or access to large amounts of data.

At its core, Grid Computing enables devices-regardless of their operating characteristics-to be virtually shared, managed and accessed across an enterprise, industry or workgroup. This virtualization of resources places all of the necessary access, data and processing power at the fingertips of those who need to rapidly solve complex business problems, conduct compute-intensive research and data analysis, and operate in real-time.

Distributed computing was one of the first real instances of cloud computing. Long before Google or Amazon, there was SETI@Home. Proposed in 1995 and launched in 1999, this program uses the spare capacity of internet connected machines to search for extraterrestrial intelligence. This is sort of the cloud in reverse.

A more recent example would be software like Hadoop. Written in Java, Hadoop is a scalable, efficient, distributed software platform designed to process enormous amounts of data. Hadoop can scale to thousands of computers across many clusters.

Distributed computing is nothing more than utilizing many networked computers to partition (split it into many smaller pieces) a question or problem and allow the network to solve the issue piecemeal.

Another instance of distributed computing, for storage instead of processing power, is bittorrent. A torrent is a file that is split into many pieces and stored on many computers around the internet. When a local machine wants to access that file, the small pieces are retrieved and rebuilt.

As the cloud computing buzzword has evolved, distributed computing has fallen out of that particular category of software. Even though distributed computing might take advantage of the internet, it doesn't follow the other tenants of cloud computing, mainly the automatic and instant scalability of resources.

That's not to say that a distributed system couldn't be built to be a cloud environment. Bittorrent, or any P2P system, comes very close to a cloud storage. It would require some additional protections like file ownership and privacy across all nodes but it could probably be done. Privacy like that is not quite what P2P is all about though.

The Cloud Computing paradigm originates mainly from research on distributed computing and virtualization, as it is based on principles, techniques and technologies developed in these areas.

[Unit 2: Cloud Service Models]

Cloud Computing (CSC-458)

Jagdish Bhatta

Central Department of Computer Science & Information Technology

Tribhuvan University

Cloud Service Models:

Cloud Communication:

Cloud communications are Internet-based voice and data communications where telecommunications applications, switching and storage are hosted by a third-party outside of the organization using them, and they are accessed over the public Internet. Cloud services are a broad term, referring primarily to data-center-hosted services that are run and accessed over an Internet infrastructure. Until recently, these services have been data-centric, but with the evolution of VoIP (voice over Internet protocol), voice has become part of the cloud phenomenon.

Cloud communications providers deliver voice & data communications applications and services, hosting them on servers that the providers own and maintain, giving their customers access to the "cloud." Because they only pay for services or applications they use, customers have a more cost-effective, reliable and secure communications environment, without the headaches associated with more conventional PBX system deployment.

Companies can cut costs with cloud communications services without sacrificing features. The success of Google and others as cloud-based providers has demonstrated that a cloud-based platform can be just as effective as a software-based platform, but at a much lower cost. Voice services delivered from the cloud increases the value of hosted telephony, as users can equally well turn to a cloud-based offering instead of relying on a facilities-based service provider for hosted VoIP. This expands their options beyond local or regional carriers.

In the past, businesses have been able to do this for IT services, but not telecom. Cloud communications is attractive because the cloud can now become a platform for voice, data and video. Most hosted services have been built around voice, and are usually referred to as hosted VoIP. The cloud communications environment serves as a platform upon which all these modes can seamlessly work as well as integrate.

There are three trends in enterprise communications pushing users to access the cloud and allowing them to do it from any device they choose, a development traditional IT communications infrastructure was not designed to handle. The first trend is increasingly distributed company operations in branches and home offices, making WANs cumbersome, inefficient and costly. Second, more communications devices need access to enterprise networks – iPhones, printers and VoIP handsets, for example. Third, data centers housing enterprise IT assets and applications are consolidating and are often being located and managed remotely.

Communication-as-a-Service (CaaS):

Communications as a Service (CaaS) is an outsourced enterprise communications solution that can be leased from a single vendor. Such communications can include voice over IP (VoIP or Internet telephony), instant messaging (IM), collaboration and videoconference applications using fixed and mobile devices.

The CaaS vendor is responsible for all hardware and software management and offers guaranteed Quality of Service (QoS). CaaS allows businesses to selectively deploy communications devices and modes on a pay-as-you-go, as-needed basis. This approach eliminates the large capital investment and ongoing overhead for a system whose capacity may often exceed or fall short of current demand.

CaaS service offerings are often bundled and may include integrated access to traditional voice (or VoIP) and data, advanced unified communications functionality such as video calling, web collaboration, chat, realtime presence and unified messaging, a handset, local and long-distance voice services, voice mail, advanced calling features (such as caller ID, threeway and conference calling, etc.) and advanced PBX functionality. A CaaS solution includes redundant switching, network, POP and circuit diversity, customer premises equipment redundancy, and WAN fail-over that specifically addresses the needs of their customers. All VoIP transport components are located in geographically diverse, secure data centers for high availability and survivability.

CaaS offers flexibility and expandability that small and medium-sized business might not otherwise afford, allowing for the addition of devices, modes or coverage on demand. The network capacity and feature set can be changed from day to day if necessary so that functionality keeps pace with demand and resources are not wasted. There is no risk of the system becoming obsolete and requiring periodic major upgrades or replacement.

CaaS requires little to no management oversight from customers. It eliminates the business customer's need for any capital investment in infrastructure, and it eliminates expense for ongoing maintenance and operations overhead for infrastructure. With a CaaS solution, customers are able to leverage enterprise-class communication services without having to build a premises-based solution of their own. This allows those customers to reallocate budget and personnel resources to where their business can best use them.

Advantages of CaaS

From the handset found on each employee's desk to the PC-based software client on employee laptops, to the VoIP private backbone, and all modes in between, every component in a CaaS solution is managed 24/7 by the CaaS vendor. Let's look as some of the advantages of a hosted approach for CaaS;

Hosted and Managed Solutions: Remote management of infrastructure services provided by third parties once seemed an unacceptable situation to most companies. However, over the past decade, with enhanced technology, networking, and software, the attitude has changed. This is, in part, due to cost savings achieved in using those services. However, unlike the "one-off " services offered by specialist providers, CaaS delivers a complete communications solution that is entirely managed by a single vendor. Along with features such as VoIP and unified communications, the integration of core PBX features with advanced functionality is managed by one vendor, who is responsible for all of the integration and delivery of services to users.

Fully Integrated, Enterprise-Class Unified Communications: With CaaS, the vendor provides voice and data access and manages LAN/ WAN, security, routers, email, voice mail, and data storage. By managing the LAN/WAN, the vendor can guarantee consistent quality of service from a user's desktop across the network and back. Advanced unified communications features that are most often a part of a standard CaaS deployment include:

- Chat
- Multimedia conferencing
- Microsoft Outlook integration
- Real-time presence
- "Soft" phones (software-based telephones)
- Video calling
- Unified messaging and mobility

Providers are constantly offering new enhancements (in both performance and features) to their CaaS services. The development process and subsequent introduction of new features in applications is much faster, easier, and more economical than ever before. This is, in large part, because the service provider is doing work that benefits many end users across the provider's scalable platform infrastructure. Because many end users of the provider's service ultimately share this cost (which, from their perspective, is miniscule compared to shouldering the burden alone), services can be offered to individual customers at a cost that is attractive to them.

No Capital Expenses Needed: When business outsource their unified communications needs to a CaaS service provider, the provider supplies a complete solution that fits the company's exact needs. Customers pay a fee (usually billed monthly) for what they use. **Customers are not required to purchase equipment, so there is no capital outlay.** Bundled in these types of services are ongoing maintenance and upgrade costs, which are incurred by the service provider. The use of CaaS services allows companies the ability to collaborate across any workspace. CaaS can also accelerate decision making within an organization. Innovative unified communications capabilities (such as presence, instant messaging, and rich media services) help ensure that information quickly reaches whoever needs it.

Flexible Capacity and Feature Set: When customers outsource communications services to a CaaS provider, they pay for the features they need when they need them. **The service provider can distribute the cost services and delivery across a large customer base. This makes the use of shared feature functionality more economical for customers to implement. Economies of scale allow service providers enough flexibility that they are not tied to a single vendor investment.** They are able to leverage best-of-breed providers such as Avaya, Cisco, Juniper, Microsoft, Nortel and ShoreTel more economically than any independent enterprise

No Risk of Obsolescence: Rapid technology advances, predicted long ago and known as Moore's law, have brought about product obsolescence in increasingly shorter periods of time. Moore's law describes a trend he recognized that has held true since the beginning of the use of integrated circuits (ICs) in computing hardware. Since the invention of the integrated circuit in 1958, the number of transistors that can be placed inexpensively on an integrated circuit has increased exponentially, doubling approximately every two years. Unlike IC components, the average life cycles for PBXs and key communications equipment and systems range anywhere from five to 10 years. With the constant introduction of newer models for all sorts of technology (PCs, cell phones, video software and hardware, etc.), these types of products now face much shorter life cycles, sometimes as short as a single year. CaaS vendors must absorb this burden for the user by continuously upgrading the equipment in their offerings to meet changing demands in the marketplace.

No Facilities and Engineering Costs Incurred: CaaS providers host all of the equipment needed to provide their services to their customers, virtually eliminating the need for customers to maintain data center space and facilities. There is no extra expense for the constant power consumption that such a facility would demand. Customers receive the benefit of multiple carrier-grade data centers with full redundancy—and it's all included in the monthly payment.

Guaranteed Business Continuity: If a catastrophic event occurred at your business's physical location, would your company disaster recovery plan allow your business to

continue operating without a break? If your business experienced a serious or extended communications outage, how long could your company survive? For most businesses, the answer is "not long." **Distributing risk by using geographically dispersed data centers has become the norm today. It mitigates risk and allows companies in a location hit by a catastrophic event to recover as soon as possible. This process is implemented by CaaS providers because most companies don't even contemplate voice continuity if catastrophe strikes.** Unlike data continuity, eliminating single points of failure for a voice network is usually cost-prohibitive because of the large scale and management complexity of the project. With a CaaS solution, multiple levels of redundancy are built into the system, with no single point of failure.

Infrastructure-as-a-Service (IaaS):

Infrastructure-as-a-Service (IaaS) is the delivery of computer infrastructure (typically a platform virtualization environment) as a service. IaaS leverages significant technology, services, and data center investments to deliver IT as a service to customers. Unlike traditional outsourcing, which requires extensive due diligence, negotiations ad infinitum, and complex, lengthy contract vehicles, IaaS is centered around a model of service delivery that provisions a predefined, standardized infrastructure specifically optimized for the customer's applications.

Infrastructure as a Service (IaaS) **includes the capability to provision processing, storage, networks, and other fundamental computing resources; 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). Services offered by this paradigm include: server hosting, web servers, storage, computing hardware, operating systems, virtual instances, load balancing, Internet access, and bandwidth provisioning. **IaaS clouds often offer additional resources such as; images in a virtual machine image library, raw (block) and file-based storage, firewalls, IP addresses, virtual local area networks (VLANs) and software bundles.** Examples of

IaaS providers include **Amazon CloudFormation, Amazon EC2, Windows Azure Virtual Machines, DynDNS, Google Compute Engine, HP Cloud, Rackspace Cloud, ReadySpace Cloud Services, and Terremark.**

The *IaaS* cloud computing paradigm has a number of characteristics such as: **the resources are distributed and support dynamic scaling**, **it is based on a utility pricing model and variable cost**, and **the hardware is shared among multiple users**. This cloud computing model is particularly useful when the demand is volatile and a new business needs computing resources and it does not want to invest in a computing infrastructure or when an organization is expanding rapidly.

IaaS provides the hardware and the software for servers, storage, networks, including operating systems and storage management software. The Infrastructure as a Service poses the most challenges.

IaaS providers manage the transition and hosting of selected applications on their infrastructure. Provider-owned implementations typically include the following layered components:

- Computer hardware (typically set up as a grid for massive horizontal scalability)
- Computer network (including routers, firewalls, load balancing, etc.)
- Internet connectivity (often on OC 192 backbones)
- Platform virtualization environment for running client-specified virtual machines
- Service-level agreements
- Utility computing billing

Rather than purchasing data center space, servers, software, network equipment, etc., IaaS customers essentially rent those resources as a fully outsourced service. Usually, the service is billed on a monthly basis, just like a utility company bills customers. The customer is charged only for resources consumed. The chief benefits of using this type of outsourced service include:

- Ready access to a preconfigured environment that is generally ITIL-based (The Information Technology Infrastructure Library [ITIL] is a customized framework of best practices designed to promote quality computing services in the IT sector.)
- Use of the latest technology for infrastructure equipment
- Secured, "sand-boxed" (protected and insulated) computing platforms that are usually security monitored for breaches
- Reduced risk by having off-site resources maintained by third parties
- Ability to manage service-demand peaks and valleys
- Lower costs that allow expensing service costs instead of making capital investments
- Reduced time, cost, and complexity in adding new features or capabilities

Some of the most popular IaaS solutions are discussed as below;

Compute as a service: One of the most ubiquitous IaaS offerings today, compute as a service provides compute capacity that includes servers, operating system access, firewalls, routers and load balancing on demand. These systems have management interfaces, and their capacity can be either shared or private. Depending on the provider and the options an enterprise chooses, compute as a service also can include automated patch management, management of infrastructure software, storage management, security management, dedicated customer support and customized SLAs.

Web hosting: Many organizations rely on their websites for marketing and revenue, and any glitch in operations can mean a loss of business. **Moving a website to an IaaS based model ensures that the website won't get bogged down during peak traffic times** — and that organizations won't have to overpay for capacity to manage those traffic spikes. What's more, loads will always be balanced, and uptime is guaranteed, thanks to SLAs. Other perks include offsite backup and fast connections for eliminating slow page and content downloads, no matter how much rich media a site includes.

Storage as Service: Storage as a Service is a business model in which a large company rents space in their storage infrastructure to a smaller company or individual. In the

enterprise, SaaS vendors are targeting secondary storage applications by promoting SaaS as a convenient way to manage backups. The key advantage to SaaS in the enterprise is in cost savings -- in personnel, in hardware and in physical storage space. For instance, instead of maintaining a large tape library and arranging to vault (store) tapes offsite, a network administrator that used SaaS for backups could specify what data on the network should be backed up and how often it should be backed up. His company would sign a service level agreement (SLA) whereby the SaaS provider agreed to rent storage space on a cost-per-gigabyte-stored and cost-per-data-transfer basis and the company's data would be automatically transferred at the specified time over the storage provider's proprietary wide area network (WAN) or the Internet. If the company's data ever became corrupt or got lost, the network administrator could contact the SaaS provider and request a copy of the data. Storage as a Service is generally seen as a good alternative for a small or mid-sized business that lacks the capital budget and/or technical personnel to implement and maintain their own storage infrastructure. SaaS is also being promoted as a way for all businesses to mitigate risks in disaster recovery, provide long-term retention for records and enhance both business continuity and availability.

Disaster recovery and backup as a service: The idea behind moving disaster recovery to the cloud is to ensure that organizations have uninterrupted access to data and applications, regardless of emergencies, such as power outages, natural disasters or system failures. These solutions always include redundancy and automatic failover to ensure ongoing access, reducing downtime to nearly zero. Many solutions also employ continuous data protection (CDP), which allows for multiple versions of all data sets to be recovered. This gives users the ability to restore data to any point in time. Data and applications are stored in secure offsite facilities. **There are two basic options when it comes to disaster recovery as a service: backup and restore from the cloud and backup and restore to the cloud**. With the first option, organizations retain applications and data on their own premise, but back up data to the cloud and restore it to hardware on their own premise when a disaster occurs. With the second option, data is restored to virtual machines in the cloud. For mission-critical applications and resources that must be

recovered quickly and completely, the best Schoice is often to replicate data to virtual machines.

Desktops as a service: DaaS is, in essence, an IaaS cloud created solely for hosting and serving virtual desktops. Essentially, it's pay-as-you-go computing that allows enterprises to quickly provision, access, run and deactivate virtual desktop machines as needed. Organizations can choose to connect through a private network service instead of the public Internet. In most cases, the service provider offers storage for the virtual computers, ensures security and data protection, and controls the network bandwidth to ensure uptime. Most solutions come with a self-service portal for provisioning and multitenant monitoring, reporting and billing. **DaaS is a way to make sure that there are always enough desktop environments available to new workers, with enough storage and all the right applications. And because the desktops can be accessed via the Internet, users can log in and access their familiar workspaces from any location.**

Servers as a service: **Accessing servers in the cloud means that no matter what the project, or even if it's the busy season, there will always be enough compute power to go around**. It's useful for one-time projects that require additional capacity, or for handling spikes in transactions. And because it's a service, enterprises can rest assured that they'll never be paying for more server capacity than they need. Accessing servers as a service also means organizations can cut their IT administrative, maintenance and service workloads. That's particularly important with servers, which can require complex and expensive system administration. **The servers are restricted to secure, private areas dedicated to the organization's use, so security is ironclad.**

Networking as a service: This is the newest entrant in the IaaS category. **The idea is to offer networking resources on demand in order to support virtual networks — resources such as firewalls, load balancing and WAN acceleration services.** Simply put, NaaS provides unified connectivity across storage, networking and servers that changes to meet the demands of virtualized infrastructures. In some cases, a networking service can support quality of service (QoS) and other network-based auditing and

monitoring services. As with other IaaS services, NaaS involves no upfront costs and supports full scalability, flexibility and security.

Modern On-Demand Computing:

On-demand computing is an increasingly popular enterprise model in which computing resources are made available to the user as needed. Computing resources that are maintained on a user's site are becoming fewer and fewer, while those made available by a service provider are on the **rise. The on-demand model evolved to overcome the challenge of being able to meet fluctuating resource demands efficiently**. Because demand for computing resources can vary drastically from one time to another, maintaining sufficient resources to meet peak requirements can be costly. Overengineering a solution can be just as adverse as a situation where the enterprise cuts costs by maintaining only minimal computing resources, resulting in insufficient resources to meet peak load requirements. Concepts such as clustered computing, grid computing, utility computing, etc., may all seem very similar to the concept of on-demand computing, but they can be better understood if one thinks of them as building blocks that evolved over time and with techno-evolution to achieve the modern cloud computing model we think of and use today. One example that we can examine is Amazon's Elastic Compute Cloud (Amazon EC2). This is a web service that provides resizable computing capacity in the cloud.

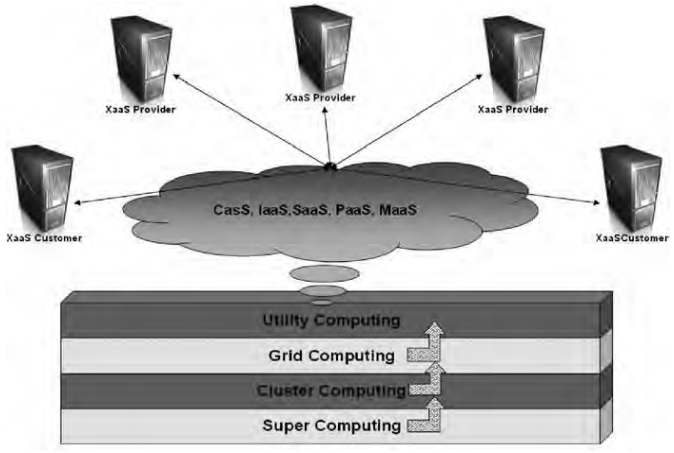


Fig: Building Blocks to the cloud

Amazon Web Services:

Amazon was the first providers of cloud computing (*http://aws.amazon.com*); it announced a limited public beta release of its Elastic Computing platform called *EC2* in August 2006.

AMAZON EC2 SERVICE:

Elastic Compute Cloud (EC2) **is a Web service with a simple interface for launching instances of an application under several operating systems, such as several Linux distributions, Microsoft Windows Server 2003 and 2008, OpenSolaris, FreeBSD, and NetBSD.**

EC2 **allows a user to load instances of an application with a custom application environment, manage networks access permissions, and run the images using as many or as few systems as desired.** *EC2* instances boot from an AMI (Amazon Machine Image) digitally signed and stored in *S3*; one could use the few images provided by Amazon or customize an image and store it in *S3*.

A user can,

i. lunch an instance from an existing AMI and terminate an instance;
ii. start and stop an instance;
iii. create a new image;
iv. add tags to identify an image; and
v. reboot an instance.

EC2 is based on the Xen virtualization strategy. **In *EC2* each virtual machine functions as a virtual private server and is called an *instance;* an instance specifies the maximum amount of resources available to an application, the interface for that instance, as well as, the cost per hour. This is a web service that provides resizable computing capacity in the cloud.** It is designed to make web-scale computing easier for developers and offers many advantages to customers:

- It is a web service interface that allows customers to obtain and configure capacity with minimal effort.
- It provides users with complete control of their (leased) computing resources and lets them run on a proven computing environment.
- It reduces the time required to obtain and boot new server instances to minutes, allowing customers to quickly scale capacity as their computing demands dictate.
- It changes the economics of computing by allowing clients to pay only for capacity they actually use.
- It provides developers the tools needed to build failure-resilient applications and isolate themselves from common failure scenarios.

Amazon EC2 presents a true virtual computing environment, allowing clients to use a web-based interface to obtain and manage services needed to launch one or more instances of a variety of operating systems (OSs). Clients can load the OS environments with their customized applications. They can manage their network's access permissions and run as many or as few systems as needed. **In order to use Amazon EC2, clients first need to create an Amazon Machine Image (AMI). This image contains the applications, libraries, data, and associated configuration settings used in the virtual computing environment. Amazon EC2 offers the use of preconfigured images built with templates to get up and running immediately. Once users have defined and configured their AMI, they use the Amazon EC2 tools provided for storing the AMI by uploading the AMI into Amazon S3.** Amazon S3 is a repository that provides safe, reliable, and fast access to a client AMI. Before clients can use the AMI, they must use the Amazon EC2 web service to configure security and network access.

A user can interact with *EC2* using a set of SOAP messages, (*The Simple Object Access Protocol (SOAP) is an application protocol developed in 1998 for Web applications. Its message format is based on the Extensible Markup Language. SOAP uses TCP and more recently UDP transport protocols; it can also be stacked above other application layer protocols such as HTTP, SMTP. The processing model of SOAP is based on a network consisting of senders, receivers, intermediaries, message originators, ultimate receivers,*

and message paths. SOAP is an underlaying layer of Web Services), and can list available AMI images, boot an instance from an image, terminate an image, display the running instances of a user, display console output, and so on. The user has root access to each instance in the elastic and secure computing environment of *EC2*. The instances can be placed in multiple locations in different Regions and Availability Zones. *EC2* allows the import of virtual machine images from the user environment to an instance through a facility called *VM import.* It also distributes automatically the incoming application traffic among multiple instances using the *elastic load balancing* facility. *EC2* associates an *elastic IP address* with an account; this mechanism allows a user to mask the failure of an instance and re-map a public IP address to any instance of the account, without the need to interact with the software support team.

To be able to connect to a virtual machine in a cloud, a client must know its IP address. For security reasons public IP addresses are mapped internally to private IP addresses. For example, a virtual machine running under Amazon's *EC2* has several IP addresses:

---

1. *EC2 Private IP Address:* The internal address of an instance; it is only used for routing within the *EC2*Cloud
2. *EC2 Public IP Address:* Network traffic originating outside the *EC2*network must use either the public IP address or the elastic IP address of the instance. The public IP address is translated using the Network Address Translation (NAT) to the private IP address when an instance is launched and it is valid until the instance is terminated. Traffic to the public address is forwarded to the private IP address of the instance.
3. *EC2 Elastic IP Address:* The IP address allocated to an AWS*EC2* account and used by traffic originated outside the *EC2*cloud. NAT is used to map an elastic IP address to the private IP address. Elastic IP addresses allow the cloud user to mask instance or availability zone failures by programmatically re-mapping a public IP addresses to any instance associated with the user's account. This allows fast recovery after a system failure; for example, rather than waiting for a cloud maintenance team to reconfigure or replace the failing host, or waiting for DNS to

propagate the new public IP to all of the customers of a Web service hosted by EC2, the Web service provider can re-map the elastic IP address to a replacement instance.

Amazon EC2 Service Characteristics:

- **Dynamic Scalability:** Amazon EC2 enables users to increase or decrease capacity in a few minutes. Users can invoke a single instance, hundreds of instances, or even thousands of instances simultaneously. Of course, because this is all controlled with web service APIs, an application can automatically scale itself up or down depending on its needs. This type of dynamic scalability is very attractive to enterprise customers because it allows them to meet their customers' demands without having to overbuild their infrastructure.

- **Full Control of Instances:** Users have complete control of their instances. They have root access to each instance and can interact with them as one would with any machine. Instances can be rebooted remotely using web service APIs. Users also have access to console output of their instances. Once users have set up their account and uploaded their AMI to the Amazon S3 service, they just need to boot that instance. It is possible to start an AMI on any number of instances (or any type) by calling the *RunInstances* API that is provided by Amazon.

- **Configuration Flexibility:** Configuration settings can vary widely among users. They have the choice of multiple instance types, operating systems, and software packages. Amazon EC2 allows them to select a configuration of memory, CPU, and instance storage that is optimal for their choice of operating system and application. For example, a user's choice of operating systems may also include numerous Linux distributions, Microsoft Windows Server, and even an OpenSolaris environment, all running on virtual servers.

- **Integration with Other Amazon Web Services:** Amazon EC2 works in conjunction with a variety of other Amazon web services. For example, Amazon Simple Storage Service (Amazon S3), Amazon SimpleDB, Amazon Simple Queue

Service (Amazon SQS), and Amazon CloudFront are all integrated to provide a complete solution for computing, query processing, and storage across a wide range of applications. **Amazon S3** provides a web services interface that allows users to store and retrieve any amount of data from the Internet at any time, anywhere. It gives developers direct access to the same highly scalable, reliable, fast, inexpensive data storage infrastructure Amazon uses to run its own global network of web sites. The S3 service aims to maximize benefits of scale and to pass those benefits on to developers. **Amazon SimpleDB** is another web-based service, designed for running queries on structured data stored with the Amazon Simple Storage Service (Amazon S3) in real time. This service works in conjunction with the Amazon Elastic Compute Cloud (Amazon EC2) to provide users the capability to store, process, and query data sets within the cloud environment. These services are designed to make web-scale computing easier and more cost effective for developers. Traditionally, this type of functionality was provided using a clustered relational database that requires a sizable investment. Implementations of this nature brought on more complexity and often required the services of a database administer to maintain it. **Amazon Simple Queue Service (Amazon SQS)** is a reliable, scalable, hosted queue for storing messages as they pass between computers. Using Amazon SQS, developers can move data between distributed components of applications that perform different tasks without losing messages or requiring 100% availability for each component. Amazon SQS works by exposing Amazon's web-scale messaging infrastructure as a service. Any computer connected to the Internet can add or read messages without the need for having any installed software or special firewall configurations. Components of applications using Amazon SQS can run independently and do not need to be on the same network, developed with the same technologies, or running at the same time. **Amazon CloudFront** is a web service for content delivery. It integrates with other Amazon web services to distribute content to end users with low latency and high data transfer speeds. Amazon CloudFront delivers content using a global network of edge locations. Requests for objects are automatically routed to the nearest edge server, so content is delivered with the best possible performance. An edge server receives a

request from the user's computer and makes a connection to another computer called the origin server, where the application resides. When the origin server fulfills the request, it sends the application's data back to the edge server, which, in turn, forwards the data to the client computer that made the request.

- **Reliable and Resilient Performance Amazon Elastic Block Store (EBS):** is yet another Amazon EC2 feature that provides users powerful features to build failure-resilient applications. Amazon EBS offers persistent storage for Amazon EC2 instances. Amazon EBS volumes provide "off-instance" storage that persists independently from the life of any instance. Amazon EBS volumes are highly available, highly reliable data shares that can be attached to a running Amazon EC2instance and are exposed to the instance as standard block devices. Amazon EBS volumes are automatically replicated on the back end. The service provides users with the ability to create point-in-time snapshots of their data volumes, which are stored using the Amazon S3 service. These snapshots can be used as a starting point for new Amazon EBS volumes and can protect data indefinitely.

- **Support for Use in Geographically Disparate Locations:** Amazon EC2 provides users with the ability to place one or more instances in multiple locations. Amazon EC2 locations are composed of Regions (such as North America and Europe) and Availability Zones. Regions consist of one or more Availability Zones, are geographically dispersed, and are in separate geographic areas or countries. Availability Zones are distinct locations that are engineered to be insulated from failures in other Availability Zones and provide inexpensive, low-latency network connectivity toother Availability Zones in the same Region. By launching instances in any one or more of the separate Availability Zones, one can insulate their applications from a single point of failure. Amazon EC2 is currently available in two regions, the United States and Europe.

Elastic IP Addressing:

Elastic IP (EIP) addresses are static IP addresses designed for dynamic cloud computing. An Elastic IP address is associated with your account and not with a particular instance,

and you control that address until you choose explicitly to release it. Unlike traditional static IP addresses, however, EIP addresses allow you to mask instance or Availability Zone failures by programmatically remapping your public IP addresses to any instance in your account. Rather than waiting on a technician to reconfigure or replace your host, or waiting for DNS to propagate to all of your customers, Amazon EC2 enables you to work around problems that occur with client instances or client software by quickly remapping their EIP address to another running instance. A significant feature of Elastic IP addressing is that each IP address can be reassigned to a different instance when needed. Now, let's review how the Elastic IPs work with Amazon EC2 services. First of all, Amazon allows users to allocate up to five Elastic IP addresses per account (which is the default). Each EIP can be assigned to a single instance. When this reassignment occurs, it replaces the normal dynamic IP address used by that instance. By default, each instance starts with a dynamic IP address that is allocated upon startup. Since each instance can have only one external IP address, the instance starts out using
the default dynamic IP address. If the EIP in use is assigned to a different instance, a new dynamic IP address is allocated to the vacated address of that instance. Assigning or reassigning an IP to an instance requires only a few minutes. The limitation of designating a single IP at a time is due to the way Network Address Translation (NAT) works. Each instance is mapped to an internal IP address and is also assigned an external (public) address. The public address is mapped to the internal address using Network Address Translation tables (hence, NAT). If two external IP addresses happen to be translated to the same internal IP address, all inbound traffic (in the form of data packets) would arrive without any issues. However, assigning outgoing packets to an external IP address would be very difficult because a determination of which external IP address to use could not be made. This is why implementors have built in the limitation of having only a single external IP address per instance at any one time.

Monitoring-as-a-Service (MaaS):

Monitoring-as-a-Service (MaaS) is the outsourced provisioning of security, primarily on business platforms that leverage the Internet to conduct business. MaaS has become increasingly popular over the last decade. Since the advent of cloud computing, its

popularity has, grown even more. Security monitoring involves protecting an enterprise or government client from cyber threats. A security team plays a crucial role in securing and maintaining the confidentiality, integrity, and availability of IT assets. However, time and resource constraints limit security operations and their effectiveness for most companies. This requires constant vigilance over the security infrastructure and critical information assets.

Monitoring as a Service (MaaS) is at present still an emerging piece of the Cloud jigsaw but an integral one for the future. In the same way that businesses realised that their infrastructure and key applications required monitoring tools that would ensure the proactive elimination of any downtime risks, Monitoring as a Service provides the option to offload a large majority of those costs by having it run as a service as opposed to a fully invested in house tool. So for example by logging onto a thin client or central web based dashboard which is hosted by the service provider, the consumer can monitor the status of their key applications regardless of location. Add the advantages of an easy set up and purchasing process and MaaS could be a key pay as you use model for the de-risking of applications that are initially being migrated to the Cloud.

Many industry regulations require organizations to monitor their security environment, server logs, and other information assets to ensure the integrity of these systems. However, conducting effective security monitoring can be a daunting task because it requires advanced technology, skilled security experts, and scalable processes—none of which come cheap. MaaS security monitoring services offer real-time, 24/7 monitoring and nearly immediate incident response across a security infrastructure—they help to protect critical information assets of their customers. Prior to the advent of electronic security systems, security monitoring and response were heavily dependent on human resources and human capabilities, which also limited the accuracy and effectiveness of monitoring efforts. Over the past two decades, the adoption of information technology into facility security systems, and their ability to be connected to security operations centers (SOCs) via corporate networks, has significantly changed that picture. This means two important things: (1) The total cost of ownership (TCO) for traditional SOCs is much higher than for a modern-technology SOC; and (2) achieving lower security operations costs and higher

security effectiveness means that modern SOC architecture must use security and IT technology to address security risks. **Typical services provided by many MaaS vendors are described below;**

Early Detection: An early detection service detects and reports new security vulnerabilities shortly after they appear. Generally, the threats are correlated with thirdparty sources, and an alert or report is issued to customers. This report is usually sent by email to the person designated by the company. Security vulnerability reports, aside from containing a detailed description of the vulnerability and the platforms affected, also include information on the impact the exploitation of this vulnerability would have on the systems or applications previously selected by the company receiving the report. Most often, the report also indicates specific actions to be taken to minimize the effect of the vulnerability, if that is known.

Platform, Control, and Services Monitoring: Platform, control, and services monitoring is often implemented as a dashboard interface and makes it possible to know the operational status of the platform being monitored at any time. It is accessible from a web interface, making remote access possible. Each operational element that is monitored usually provides an operational status indicator, always taking into account the critical impact of each element. This service aids in determining which elements may be operating at or near capacity or beyond the limits of established parameters. By detecting and identifying such problems, preventive measures can be taken to prevent loss of service.

Intelligent Log Centralization and Analysis: Intelligent log centralization and analysis is a monitoring solution based mainly on the correlation and matching of log entries. Such analysis helps to establish a baseline of operational performance and provides an index of security threat. Alarms can be raised in the event an incident moves the established baseline parameters beyond a stipulated threshold. These types of sophisticated tools are used by a team of security experts who are responsible for incident response once such a threshold has been crossed and the threat has generated an alarm or warning picked up by security analysts monitoring the systems.

Vulnerabilities Detection and Management: Vulnerabilities detection and management enables automated verification and management of the security level of information systems. The service periodically performs a series of automated tests for the purpose of identifying system weaknesses that may be exposed over the Internet, including the possibility of unauthorized access to administrative services, the existence of services that have not been updated, the detection of vulnerabilities such as phishing, etc. The service performs periodic follow-up of tasks performed by security professionals managing information systems security and provides reports that can be used to implement a plan for continuous improvement of the system's security level.

Continuous System Patching/Upgrade and Fortification: Security posture is enhanced with continuous system patching and upgrading of systems and application software. New patches, updates, and service packs for the equipment's operating system are necessary to maintain adequate security levels and support new versions of installed products. Keeping abreast of all the changes to all the software and hardware requires a committed effort to stay informed and to communicate gaps in security that can appear in installed systems and applications.

Intervention, Forensics, and Help Desk Services: Quick intervention when a threat is detected is crucial to mitigating the effects of a threat. This requires security engineers with ample knowledge in the various technologies and with the ability to support applications as well as infrastructures on a 24/7 basis. MaaS platforms routinely provide this service to their customers. When a detected threat is analyzed, it often requires forensic analysis to determine what it is, how much effort it will take to fix the problem, and what effects are likely to be seen. When problems are encountered, the first thing customers tend to do is pick up the phone. Help desk services provide assistance on questions or issues about the operation of running systems. This service includes assistance in writing failure reports, managing operating problems, etc.

Note: *For detail Refer the article entitled* ***"Enhanced Monitoring-as-a-Service for Effective Cloud Management"*** *by the authors Shicong Meng, and Ling Liu, IEEE, 2012 (provided in the class)*

The Traditional On-Premises Model:

The traditional approach of building and running on-premises applications has always been complex, expensive, and risky. Building your own solution has never offered any guarantee of success. **Each application was designed to meet specific business requirements. Each solution required a specific set of hardware, an operating system, a database, often a middleware package, email and web servers, etc.** Once the hardware and software environment was created, a team of developers had to navigate complex programming development platforms to build their applications. Additionally, a team of network, database, and system management experts was needed to keep everything up and running. Inevitably, a business requirement would force the developers to make a change to the application. The changed application then required new test cycles before being distributed. Large companies often needed specialized facilities to house their data centers. Enormous amounts of electricity also were needed to power the servers as well as to keep the systems cool. Finally, all of this required use of fail-over sites to mirror the data center so that information could be replicated in case of a disaster. Old days, old ways—now, let's fly into the silver lining of todays cloud.

The New Cloud Model : PaaS

PaaS offers a faster, more cost-effective model for application development and delivery. PaaS provides all the infrastructure needed to run applications over the Internet. Such is the case with companies such as Amazon.com, eBay, Google, iTunes, and YouTube. The new cloud model has made it possible to deliver such new capabilities to new markets via the web browsers. PaaS is based on a metering or subscription model, so users pay only for what they use. **PaaS offerings include workflow facilities for application design, application development, testing, deployment, and hosting, as well as application services such as virtual offices, team collaboration, database integration, security, scalability, storage, persistence, state management, dashboard instrumentation, etc**.

Platform-as-a-Service (PaaS):

Cloud computing has evolved to include platforms for building and running custom web-based applications, a concept known as Platform-as-a- Service. PaaS is an outgrowth of the SaaS application delivery model. **The PaaS model makes all of the facilities required to support the complete life cycle of building and delivering web applications and services entirely available from the Internet, all with no software downloads or installation for developers, IT managers, or end users.** Unlike the IaaS model, where developers may create a specific operating system instance with homegrown applications running, **PaaS developers are concerned only with web based development and generally do not care what operating system is used.** PaaS services allow users to focus on innovation rather than complex infrastructure. Organizations can redirect a significant portion of their budgets to creating applications that provide real business value instead of worrying about all the infrastructure issues in a roll-your-own delivery model. The PaaS model is thus driving a new era of mass innovation. Now, developers around the world can access unlimited computing power. Anyone with an Internet connection can build powerful applications and easily deploy them to users globally.

PaaS provides the capability for consumers to have applications deployed without the burden and cost of buying and managing the hardware and software. In other words these are either consumer created or acquired web applications or services that are entirely accessible from the Internet. Usually created with programming languages and tools supported by the service provider these web applications enable the consumer to have control over the deployed applications and in some circumstances the application-hosting environment but without the complexity of the infrastructure i.e. the servers, operating systems or storage. Offering a quick time to market and services that can be provisioned as an integrated solution over the web, **PaaS facilitates immediate business requirements such as application design, development and testing at a fraction of the normal cost.**

Key Characteristics of PaaS:

Chief characteristics of PaaS include services to develop, test, deploy, host, and manage applications to support the application development life cycle. Web-based user

interface creation tools typically provide some level of support to simplify the creation of user interfaces, based either on common standards such as HTML and JavaScript or on other, proprietary technologies. Supporting a **multitenant architecture** helps to remove developer concerns regarding the use of the application by many concurrent users. **PaaS providers often include services for concurrency management, scalability, fail-over and security**. Another characteristic is the **integration with web services and databases**. Support for Simple Object Access Protocol (SOAP) and other interfaces allows PaaS offerings to create combinations of web services (called mashups) as well as having the **ability to access databases and reuse services maintained inside private networks**. The **ability to form and share code with ad-hoc, predefined, or distributed teams greatly enhance the productivity of PaaS offerings**. Integrated PaaS offerings **provide an opportunity for developers to have much greater insight into the inner workings of their applications and the behavior of their users by implementing dashboard- like tools to view the inner workings based on measurements such as performance, number of concurrent accesses**, etc. Some PaaS offerings leverage this instrumentation to enable pay-per-use billing models

Software-as-a-Service:

The traditional model of software distribution, in which software is purchased for and installed on personal computers, is sometimes referred to as Software-as-a-Product. Software-as-a-Service is a software distribution model in which applications are hosted by a vendor or service provider and made available to customers over a network, typically the Internet. SaaS is becoming an increasingly prevalent delivery model as underlying technologies that support web services and service-oriented architecture (SOA) mature and new developmental approaches become popular. **SaaS is also often associated with a pay-as-you-go subscription licensing model.** Meanwhile, broadband service has become increasingly available to support user access from more areas around the world.

International Data Corportation identifies two slightly different delivery models for SaaS. **The hosted application management model is similar to an Application Service Provider (ASP) model. Here, an ASP hosts commercially available software for**

customers and delivers it over the Internet. The other model is a software on demand model where the provider gives customers network-based access to a single copy of an application created specifically for SaaS distribution.

SaaS is most often implemented to provide business software functionality to enterprise customers at a low cost while allowing those customers to obtain the same benefits of commercially licensed, internally operated software without the associated complexity of installation, management, support, licensing, and high initial cost. Most customers have little interest in the how or why of software implementation, deployment, etc., but all have a need to use software in their work. Many types of software are well suited to the SaaS model **(e.g., accounting, customer relationship management, email software, human resources, IT security, IT service management, video conferencing, web analytics, web content management).** The distinction between SaaS and earlier applications delivered over the Internet is that SaaS solutions were developed specifically to work within a web browser. **The architecture of SaaS-based applications is specifically designed to support many concurrent users (multitenancy) at once.** This is a big difference from the traditional client/server or application service provider (ASP)-based solutions that cater to a contained audience. SaaS providers, on the other hand, leverage enormous economies of scale in the deployment, management, support, and maintenance of their offerings.

SaaS Implementation Issues:

Many types of software components and applications frameworks may be employed in the development of SaaS applications. Using new technology found in these modern components and application frameworks can drastically reduce the time to market and cost of converting a traditional on-premises product into a SaaS solution. According to Microsoft, SaaS architectures can be classified into one of four maturity levels whose key attributes are ease of configuration, multitenant efficiency, and scalability. Each level is distinguished from the previous one by the addition of one of these three attributes. The levels described by Microsoft are as follows;.

SaaS Architectural Maturity Level 1—Ad-Hoc/Custom. The first level of maturity is actually no maturity at all. Each customer has a unique, customized version of the hosted application. The application runs its own instance on the host's servers. Migrating a traditional non-networked or client-server application to this level of SaaS maturity typically requires the least development effort and reduces operating costs by consolidating server hardware and administration

SaaS Architectural Maturity Level 2—Configurability. The second level of SaaS maturity provides greater program flexibility through configuration metadata. At this level, many customers can use separate instances of the same application. This allows a vendor to meet the varying needs of each customer by using detailed configuration options. It also allows the vendor to ease the maintenance burden by being able to update a common code base.

SaaS Architectural Maturity Level 3—Multitenant Efficiency. The third maturity level adds multitenancy to the second level. This results in a single program instance that has the capability to serve all of the vendor's customers. This approach enables more efficient use of server resources without any apparent difference to the end user, but ultimately this level is limited in its ability to scale massively.

SaaS Architectural Maturity Level 4—Scalable. At the fourth SaaS maturity level, scalability is added by using a multitiered architecture. This architecture is capable of supporting a load-balanced farm of identical application instances running on a variable number of servers, sometimes in the hundreds or even thousands. System capacity can be dynamically increased or decreased to match load demand by adding or removing servers, with no need for further alteration of application software architecture.

The key characteristics of SaaS software are the following:

- Network-based management and access to commercially available software from central locations rather than at each customer's site, enabling customers to access applications remotely via the Internet.

- Application delivery from a one-to-many model (single-instance, multitenant architecture), as opposed to a traditional one-to-one model.

- Centralized enhancement and patch updating that obviates any need for downloading and installing by a user. SaaS is often used in conjunction with a larger network of communications and collaboration software, sometimes as a plug-in to a PaaS architecture.

Benefits of SaaS Model:

The SaaS model helps enterprises ensure that all locations are using the correct application version and, therefore, that theformat of the data being recorded and conveyed is consistent, compatible, and accurate. By placing the responsibility for an application onto the doorstep of a SaaS provider, enterprises can reduce administration and management burdens they would otherwise have for their own corporate applications. SaaS also helps to increase the availability of applications to global locations. SaaS also ensures that all application transactions are logged for compliance purposes. The benefits of SaaS to the customer are very clear;

- Streamlined administration

- Automated update and patch management services

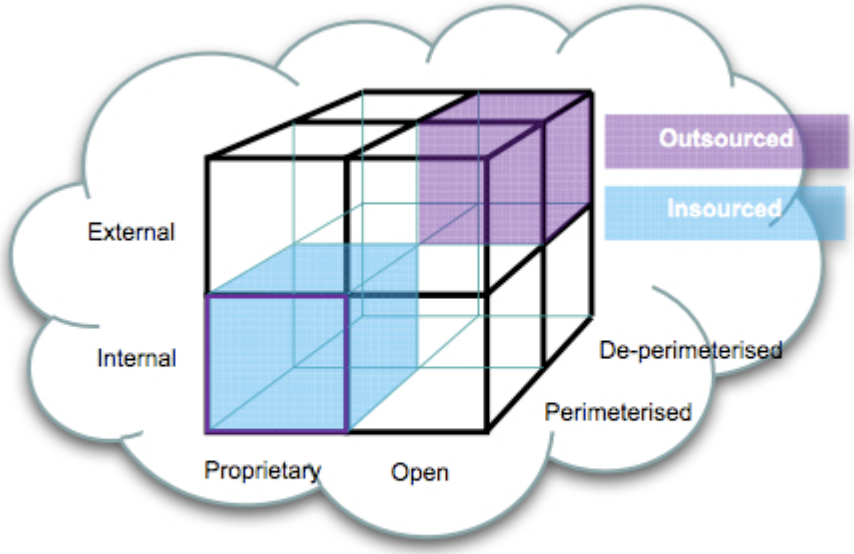
- Data compatibility across the enterprise (all users have the same version of software)

- Facilitated, enterprise-wide collaboration

- Global accessibility

Jericho Cube Model:

Refer the article ***"Cloud Cube Model: Selecting Cloud Formations for Secure Collaboration",*** *By Jericho Forum. (Provided and discussed in class)*



The Cloud Cube Model

XaaS:

Finally XaaS or 'anything as a service' is the delivery of IT as a Service through hybrid Cloud computing and is a reference to either one or a combination of Software as a Service (SaaS), Infrastructure as a Service (IaaS), Platform as a Service (PaaS). communications as a service (CaaS) or monitoring as a service (Maas). XaaS is quickly emerging as a term that is being readily recognized as services that were previously separated on either private or public Clouds are becoming transparent and integrated.

[Unit 3: Building Cloud Networks]

Cloud Computing (CSC-458)

Jagdish Bhatta

Central Department of Computer Science & Information Technology

Tribhuvan University

Evolution from Managed service providers (MSP) to Cloud Computing:

Single Purpose architectures to multi-purpose architectures:

Data center virtualization:

Cloud data center

Service Oriented Architectures (SOA):

A cloud has some key characteristics: elasticity, self-service provisioning, standards based interfaces, and pay as you go. This type of functionality has to be engineered into the software. To accomplish this type of engineering requires that the foundation for the cloud be well designed and well architected. What about cloud architecture makes this approach possible? The fact is that the services and structure behind the cloud should be based on a modular architectural approach. A modular, component-based architecture enables flexibility and reuse. **A *Service Oriented Architecture (SOA)* is what lies beneath this flexibility.**

SOA is much more than a technological approach and methodology for creating IT systems. It's also a *business* approach and methodology. Companies have used the principles of SOA to deepen the understanding between the business and IT and to help business adapt to change.

One of the key benefits of a service oriented approach is that software is designed to reflect best practices and business processes instead of making the business operate according to the rigid structure of a technical environment. A service-oriented architecture is essentially a collection of services. A service is, in essence, a function that is well defined, self-contained, and does not depend on the context or state of other services. Services most often reflect logical business activities. Some means of connecting services to each other is needed, so services communicate with each other, have an interface, and are message-oriented. The communication between services may involve simple data passing or may require two or more services coordinating an activity. The services generally

communicate using standard protocols, which allows for broad interoperability. SOA encompasses legacy systems and processes, so the effectiveness of existing investments is preserved. New services can be added or created without affecting existing services. Service-oriented architectures are not new. **The first service-oriented architectures are usually considered to be the Distributed Component Object Model (DCOM) or Object Request Brokers (ORBs), which were based on the Common Object Requesting Broker Architecture (CORBA) specification.** The introduction of SOA provides a platform for technology and business units to meet business requirements of the modern enterprise. With SOA, your organization can use existing application systems to a greater extent and may respond faster to change requests. These benefits are attributed to several critical elements of SOA:

1. Free-standing, independent components
2. Combined by loose coupling
3. Message (XML)-based instead of API-based
4. Physical location, etc., not important

Combining Cloud and SOA:

Cloud services benefit the business by taking the best practices and business process focus of SOA to the next level. These benefits apply to both cloud service providers and cloud service users. Cloud service providers need to architect solutions by using a service-oriented approach to deliver services with the expected levels of elasticity and scalability. Companies that architect and govern business processes with reusable service-oriented components can more easily identify which components can be successfully moved to public and private clouds. A *service oriented architecture (SOA)* is a software architecture for building business applications that implement business processes or services through a set of loosely coupled, black-box components orchestrated to deliver a well defined level of service. This approach lets companies leverage existing assets and create new business services that are consistent, controlled, more easily changed, and more easily managed. SOA is a business approach to designing efficient IT systems that support reuse and give the businesses the flexibility to react quickly to opportunities and threats.

Characterizing SOA :

The principal characteristics of SOA are described in more detail here:

- **SOA is a black-box component architecture.** The *black box* lets you reuse existing business applications; it simply adds a fairly simple adapter to them. You don't need to know every detail of what's inside each component; SOA hides the complexity whenever possible.

- **SOA components are loosely coupled.** Software components are *loosely coupled* if they're designed to interact in a standardized way that minimizes dependencies. One loosely coupled component passes data to another component and makes a request; the second component carries out the request and, if necessary, passes data back to the first. Each component offers a small range of simple services to other components. A set of loosely coupled components does the same work that software components in tightly structured applications used to do, but with loose coupling you can combine and recombine the components in a bunch of ways. This makes a world of difference in the ability to make changes easily, accurately, and quickly.

- **SOA components are orchestrated to link through business processes to deliver a well-defined level of service.** SOA creates a simple arrangement of components that, together, deliver a very complex business service. Simultaneously, SOA must provide acceptable service levels. To that end, the components ensure a dependable service level. Service level is tied directly to the best practices of conducting business, commonly referred to as *business process management (BPM)* — BPM focuses on effective design of business process and SOA allows IT to align with business processes.

Open Source Software in Data Centers:

Refer Page number 75 to 100 of ***Cloud Computing: Implementation Management and Security****, John W. Rittinghouse and James F. Ransome*

[Unit 4: Security in Cloud Computing]

Cloud Computing (CSC-458)

Jagdish Bhatta

Central Department of Computer Science & Information Technology

Tribhuvan University

Cloud Security Challenges:

Software as a Service Security:

The seven security issues which one should discuss with a cloud-computing vendor:

1. **Privileged user access** —inquire about who has specialized access to data, and about the hiring and management of such administrators.
2. **Regulatory compliance**—make sure that the vendor is willing to undergo external audits and/or security certifications.
3. **Data location**—does the provider allow for any control over the location of data?
4. **Data segregation** —make sure that encryption is available at all stages, and that these encryption schemes were designed and tested by experienced professionals.
5. **Recovery** —Find out what will happen to data in the case of a disaster. Do they offer complete restoration? If so, how long would that take?
6. **Investigative support** —Does the vendor have the ability to investigate any inappropriate or illegal activity?

7. **Long-term viability** —What will happen to data if the company goes out of business? How will data be returned, and in what format?

To address the security issues listed above, SaaS providers will need to incorporate and enhance security practices used by the managed service providers and develop new ones as the cloud computing environment evolves. The baseline security practices for the SaaS environment as currently formulated are discussed in the following sections.

- Security Management (People): One of the most important actions for a security team is to develop a formal charter for the security organization and program. This will foster a shared vision among the team of what security leadership is driving toward and expects, and will also foster "ownership" in the success of the collective team. The charter should be aligned with the strategic plan of the organization or company the security team works for. Lack of clearly defined roles and responsibilities, and agreement on expectations, can result in a general feeling of loss and confusion among the security team about what is expected of them, how their skills and experienced can be leveraged, and meeting their performance goals. Morale among the team and pride in the team is lowered, and security suffers as a result.

- **Security Governance:** A security steering committee should be developed whose objective is to focus on providing guidance about security initiatives and alignment with business and IT strategies. A charter for the security team is typically one of the first deliverables from the steering committee. This charter must clearly define the roles and responsibilities of the security team and other groups involved in performing information security functions. Lack of a formalized strategy can lead to an unsustainable operating model and security level as it evolves. In addition, lack of attention to security governance can result in key needs of the business not being met, including but not limited to, risk management, security monitoring, application security, and sales support. Lack of proper governance and management of duties can also result in potential security risks being left unaddressed and opportunities to improve the business being missed because the security team is not focused on the key security functions and activities that are critical to the business.

- **Risk Management:** Effective risk management entails identification of technology assets; identification of data and its links to business processes, applications, and data stores; and assignment of ownership and custodial responsibilities. Actions should also include maintaining a repository of information assets. Owners have authority and accountability for information assets including protection requirements, and custodians implement confidentiality, integrity, availability, and privacy controls. A formal risk assessment process should be created that allocates security resources linked to business continuity.

- **Risk Assessment:** Security risk assessment is critical to helping the information security organization make informed decisions when balancing the dueling priorities of business utility and protection of assets. Lack of attention to completing formalized risk assessments can contribute to an increase in information security audit findings, can jeopardize certification goals, and can lead to inefficient and ineffective selection of security controls that may not adequately mitigate information security risks to an acceptable level. A formal information security risk management process should proactively assess information security risks as well as plan and manage them on a periodic or as-needed basis. More detailed and technical security risk assessments in the form of threat modeling should also be applied to applications and infrastructure. Doing so can help the product management and engineering groups to be more proactive in designing and testing the security of applications and systems and to collaborate more closely with the internal security team. Threat modeling requires both IT and business process knowledge, as well as technical knowledge of how the applications or systems under review work.

- **Security Monitoring and Incident Response:** Centralized security information management systems should be used to provide notification of security vulnerabilities and to monitor systems continuously through automated technologies to identify potential issues. They should be integrated with network and other systems monitoring processes (e.g., security information management, security event management, security information and event management, and security operations centers that use these systems for

dedicated 24/7/365 monitoring). Management of periodic, independent third-party security testing should also be included. Many of the security threats and issues in SaaS center around application and data layers, so the types and sophistication of threats and attacks for a SaaS organization require a different approach to security monitoring than traditional infrastructure and perimeter monitoring. The organization may thus need to expand its security monitoring capabilities to include application- and data-level activities. This may also require subject-matter experts in applications security and the unique aspects of maintaining privacy in the cloud. Without this capability and expertise, a company may be unable to detect and prevent security threats and attacks to its customer data and service stability.

- **Third-Party Risk Management:** As SaaS moves into cloud computing for the storage and processing of customer data, there is a higher expectation that the SaaS will effectively manage the security risks with third parties. Lack of a third-party risk management program may result in damage to the provider's reputation, revenue losses, and legal actions should the provider be found not to have performed due diligence on its third-party vendors.

Security Architecture Design:

A security architecture framework should be established with consideration of processes (enterprise authentication and authorization, access control, confidentiality, integrity, non-repudiation, security management, etc.), operational procedures, technology specifications, people and organizational management, and security program compliance and reporting. A security architecture document should be developed that defines security and privacy principles to meet business objectives. Documentation is required for management controls and metrics specific to asset classification and control, physical security, system access controls, network and computer management, application development and maintenance, business continuity, and compliance. A design and implementation program should also be integrated with the formal system development life cycle to include a business case, requirements definition, design, and implementation plans. Technology and design

methods should be included, as well as the security processes necessary to provide the following services across all technology layers:

1. Authentication
2. Authorization
3. Availability
4. Confidentiality
5. Integrity
6. Accountability
7. Privacy

The creation of a secure architecture provides the engineers, data center operations personnel, and network operations personnel a common blueprint to design, build, and test the security of the applications and systems. Design reviews of new changes can be better assessed against this architecture to assure that they conform to the principles described in the architecture, allowing for more consistent and effective design reviews.

Vulnerability Assessment:

Vulnerability assessment classifies network assets to more efficiently prioritize vulnerability-mitigation programs, such as patching and system upgrading. It measures the effectiveness of risk mitigation by setting goals of reduced vulnerability exposure and faster mitigation. Vulnerability management should be integrated with discovery, patch management, and upgrade management processes to close vulnerabilities before they can be exploited.

Data Privacy:

A risk assessment and gap analysis of controls and procedures must be conducted. Based on this data, formal privacy processes and initiatives must be defined, managed, and sustained. As with security, privacy controls and protection must an element of the secure architecture design. Depending on the size of the organization and the scale of operations, either an individual or a team should be assigned and given responsibility for maintaining

privacy. A member of the security team who is responsible for privacy or a corporate security compliance team should collaborate with the company legal team to address data privacy issues and concerns. As with security, a privacy steering committee should also be created to help make decisions related to data privacy. Typically, the security compliance team, if one even exists, will not have formalized training on data privacy, which will limit the ability of the organization to address adequately the data privacy issues they currently face and will be continually challenged on in the future. The answer is to hire a consultant in this area, hire a privacy expert, or have one of your existing team members trained properly. This will ensure that your organization is prepared to meet the data privacy demands of its customers and regulators.

For example, customer contractual requirements/agreements for data privacy must be adhered to, accurate inventories of customer data, where it is stored, who can access it, and how it is used must be known, and, though often overlooked, Request for Interest/Request for Proposal questions regarding privacy must answered accurately. This requires special skills, training, and experience that do not typically exist within a security team. As companies move away from a service model under which they do not store customer data to one under which they do store customer data, the data privacy concerns of customers increase exponentially. This new service model pushes companies into the cloud computing space, where many companies do not have sufficient experience in dealing with customer privacy concerns, permanence of customer data throughout its globally distributed systems, cross-border data sharing, and compliance with regulatory or lawful intercept requirements.

Data Security:

The ultimate challenge in cloud computing is data-level security, and sensitive data is the domain of the enterprise, not the cloud computing provider. Security will need to move to the data level so that enterprises can be sure their data is protected wherever it goes. For example, with data-level security, the enterprise can specify that this data is not allowed to go outside of the United States. It can also force encryption of certain types of data, and permit only specified users to access the data. It can provide compliance with the Payment

Card Industry Data Security Standard (PCI DSS). True unified end-to-end security in the cloud will likely requires an ecosystem of partners.

Application Security:

Application security is one of the critical success factors for a world-class SaaS company. This is where the security features and requirements are defined and application security test results are reviewed. Application security processes, secure coding guidelines, training, and testing scripts and tools are typically a collaborative effort between the security and the development teams. Although product engineering will likely focus on the application layer, the security design of the application itself, and the infrastructure layers interacting with the application, the security team should provide the security requirements for the product development engineers to implement. This should be a collaborative effort between the security and product development team. External penetration testers are used for application source code reviews, and attack and penetration tests provide an objective review of the security of the application as well as assurance to customers that attack and penetration tests are performed regularly. Fragmented and undefined collaboration on application security can result in lower-quality design, coding efforts, and testing results.

Virtual Machine Security:

In the cloud environment, physical servers are consolidated to multiple virtual machine instances on virtualized servers. Not only can data center security teams replicate typical security controls for the data center at large to secure the virtual machines, they can also advise their customers on how to prepare these machines for migration to a cloud environment when appropriate.

Firewalls, intrusion detection and prevention, integrity monitoring, and log inspection can all be deployed as software on virtual machines to increase protection and maintain compliance integrity of servers and applications as virtual resources move from on-premises to public cloud environments. By deploying this traditional line of defense to the virtual machine itself, you can enable critical applications and data to be moved to the cloud securely. To facilitate the centralized management of a server firewall policy, the

security software loaded onto a virtual machine should include a bidirectional stateful firewall that enables virtual machine isolation and location awareness, thereby enabling a tightened policy and the flexibility to move the virtual machine from on-premises to cloud resources. Integrity monitoring and log inspection software must be applied at the virtual machine level.

This approach to virtual machine security, which connects the machine back to the mother ship, has some advantages in that the security software can be put into a single software agent that provides for consistent control and management throughout the cloud while integrating seamlessly back into existing security infrastructure investments, providing economies of scale, deployment, and cost savings for both the service provider and the enterprise.