Department of Computer Science and Engineering

23CSX507- Cloud Computing and Virtualization

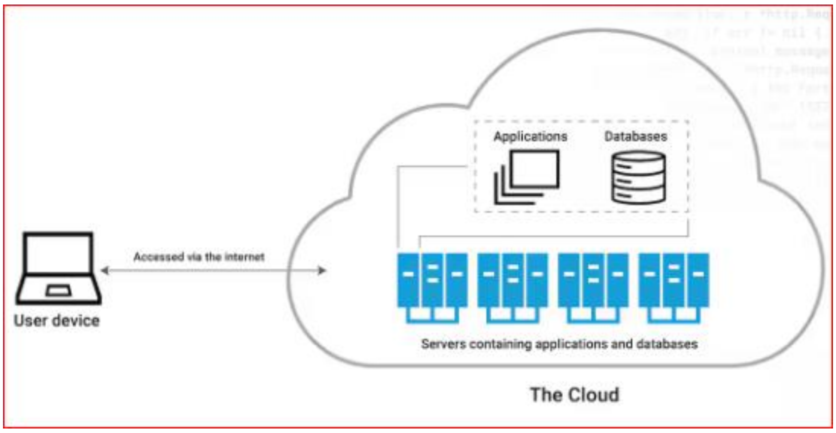
UNIT-I INTRODUCTION TO CLOUD COMPUTING

MODULE

Introduction to Cloud Computing: Overview, Roots of Cloud Computing, Layers and Types of Cloud, Desired Features of a Cloud, Benefits and Disadvantages of Cloud Computing, Cloud Infrastructure Management, Infrastructure as a Service Providers, Platform as a Service Providers, Challenges and Risks, Assessing the role of Open Standards.

OVERVIEW OF CLOUD COMPUTING

"The cloud" refers to servers that are accessed over the Internet, and the software and databases that run on those servers. Cloud servers are in data centres all over the world. By using cloud computing, users and companies do not have to manage physical servers themselves or run software applications on their own machines.



The cloud enables users to access the same files and applications from almost any device, because the computing and storage takes place on servers in a data centre, instead of locally on the user device. Therefore, a user can log into their Instagram account on a new phone after their old phone breaks and still find their old account in place, with all their photos, videos, and conversation history. It works the same way with cloud email providers like Gmail or Microsoft Office 365, and with cloud storage providers like Dropbox or Google Drive.

For businesses, switching to cloud computing it removes some IT costs and overhead: for instance, they no longer need to update and maintain their own servers, as the cloud vendor they are using will do that. This especially makes an impact for small businesses that may not have been able to afford their own internal infrastructure but can outsource their infrastructure needs affordably via the cloud. The cloud can also make it easier for companies to operate internationally, because employees and customers can access the same files and applications from any location.

ROOTS OF CLOUD COMPUTING

The roots of clouds computing can track by observing the advancement of several technologies, especially in hardware (virtualization, multi-core chips), Internet technologies (Web services, service-oriented architectures, Web 2.0), distributed computing (clusters, grids), and systems management (autonomic computing, data center automation). The convergence of technology fields that significantly advanced and contributed to the advent of cloud computing. The emergence of cloud computing itself is closely linked to the maturity of such technologies. We present a closer look at the technologies that form the base of cloud computing, with the aim of providing a clearer picture of the cloud ecosystem.

1. From Mainframes to Clouds

We are currently experiencing a switch in the IT world, from in-house generated computing power into utility-supplied computing resources delivered over the Internet as Web services. This trend is like what occurred about a century ago when factories, which used to generate their own electric power, realized that it is was cheaper just plugging their machines into the newly formed electric power grid. Computing delivered as a utility can be defined as "on demand delivery of infrastructure, applications, and business processes in a security-rich, shared, scalable, and based computer environment over the Internet for a fee".

2. SOA, Web Services, Web 2.0, and Mashups

The emergence of Web services (WS) open standards has significantly contributed to advances in the domain of software integration. Web services can glue together applications running on different messaging product plat- forms, enabling information from one application to be made available to others, and enabling internal applications to be made available over the Internet. WS standards have been created on top of existing ubiquitous technologies such as HTTP and XML, thus providing a common mechanism for delivering services, making them ideal for implementing a service-oriented architecture (SOA). The purpose of a SOA is to address requirements of loosely coupled, standards-based, and protocol-independent distributed computing. In a SOA, software resources are packaged as "services," which are well-defined, self- contained modules that provide standard business functionality and are independent of the state or context of other services. Services are described in a standard definition language and have a published interface .Many service providers, such as Amazon, delicious, Facebook, and Google, make their service APIs publicly accessible using standard protocols such as SOAP and REST. Consequently, one can put an idea of a fully functional Web application into practice just by gluing pieces with few lines of code.

3. Grid Computing

Grid computing enables aggregation of distributed resources and transparently access to them. Most production grids such as TeraGrid and EGEE seek to share compute and storage resources distributed across different administrative domains, with their focus being speeding up a broad range of scientific applications, such as climate modeling, drug design, and protein analysis. A key aspect of the grid vision realization has been building standard Web services-based protocols that allow distributed resources to be "discovered, accessed, allocated, monitored, accounted for, and billed for, etc., and in general managed as a single virtual system." The Open Grid Services Architecture (OGSA) addresses this need for standardization by defining a set of core capabilities and behaviors that address key concerns in grid systems. Virtualization technology has been identified as the perfect fit to issues that have caused frustration when using grids, such as hosting many dissimilar software applications on a single physical platform. In this direction, some research projects (e.g., Globus Virtual Workspaces) aimed at evolving grids to support an additional layer to virtualize computation, storage, and network resources.

4. Utility Computing

With increasing popularity and usage, large grid installations have faced new problems, such as excessive spikes in demand for resources coupled with strategic and adversarial behavior by users. Initially, grid resource management techniques did not ensure fair and equitable access to resources in many systems. Traditional metrics (throughput, waiting time, and slowdown) failed to capture the more subtle requirements of users. There were no real incentives for users to be flexible about resource requirements or job deadlines, nor provisions to accommodate users with urgent work.

In utility computing environments, users assign a "utility" value to their jobs, where utility is a fixed or time-varying valuation that captures various QoS constraints (deadline, importance, satisfaction). The valuation is the amount they are willing to pay a service provider to satisfy their demands. The service providers then attempt to maximize their own utility, where said utility may directly correlate with their profit. Providers can choose to prioritize high yield (i.e., profit per unit of resource) user jobs, leading to a scenario where shared systems are viewed as a marketplace, where users compete for resources based on the perceived utility or value of their jobs. Further information and comparison of these utility computing environments are available in an extensive survey of these platforms .

5. Hardware Virtualization

Cloud computing services are usually backed by large-scale data centers composed of thousands of computers. Such data centers are built to serve many users and host many disparate applications. For this purpose, hardware virtualization can be considered as a perfect fit to overcome most operational issues of data center building and maintenance. The idea of virtualizing a computer system resources, including processors, memory, and I/O devices, has been well established for decades, aiming at improving sharing and utilization of computer systems. Hardware virtualization allows running multiple operating systems and software stacks on a single physical platform. A software layer, the virtual machine monitor (VMM), also called a hypervisor, mediates access to the physical hardware presenting to each guest operating system a virtual machine (VM), which is a set of virtual platform interfaces Email Server Facebook App. A hardware virtualized server hosting three virtual machines, each one running distinct operating system and user level software stack. Several VMM platforms exist that are the basis of many utilities or cloud computing environments. The most notable ones, VMWare, Xen, and KVM, are outlined in the following sections.

VMWare ESXi – VMware is a pioneer in the virtualization market. Its ecosys- tem of tools ranges from server and desktop virtualization to high-level management tools. ESXi is a VMM (Virtual Machine Manager) from VMWare. It is a bare-metal hypervisor, meaning that it installs directly on the physical server, whereas others may require a host operating system. It provides advanced virtualization techniques of processor, memory, and I/O. Especially, through memory ballooning and page sharing, it can overcommit memory, thus increasing the density of VMs inside a single physical server.

Xen—The Xen hypervisor started as an open-source project and has served as a base to other virtualization products, both commercial and open-source. It has pioneered the paravirtualization concept, on which the guest operating system, by means of a specialized kernel, can interact with the hypervisor, thus significantly improving performance. In addition to an 10 open-source distribution, Xen currently forms the base of commercial hypervisors of several vendors, most notably Citrix XenServer and Oracle VM.

KVM—The kernel-based virtual machine (KVM) is a Linux virtualization subsystem. Is has been part of the mainline Linux kernel since version 2.6.20, thus being natively supported by several distributions. In addition, activities such as memory management and scheduling are carried out by existing kernel features, thus making KVM simpler and smaller than hypervisors that take control of the entire machine. KVM leverages hardware-assisted virtualization, which improves performance and allows it to support unmodified guest operating systems. currently, it supports several versions of Windows, Linux, and UNIX.

6. Virtual Appliances and the Open Virtualization Format

An application combined with the environment needed to run it (operating system, libraries, compilers, databases, application containers, and so forth) is referred to as a "virtual appliance. Packaging application environments in the shape of virtual appliances eases software customization, configuration, and patching and improves portability. Most commonly, an appliance is shaped as a VM disk image associated with hardware requirements, and it can be readily deployed in a hypervisor. OVF's extensibility has encouraged additions relevant to management of data centers and clouds. Mathews et al. have devised virtual machine contracts (VMC) as an extension to OVF. A VMC aids in communicating and managing the complex expectations that VMs have of their runtime environment and vice versa. A simple example of a VMC is when a cloud consumer wants to specify minimum and maximum amounts of a resource that a VM needs to function. similarly, the cloud provider could express resource limits to bound resource consumption and costs.

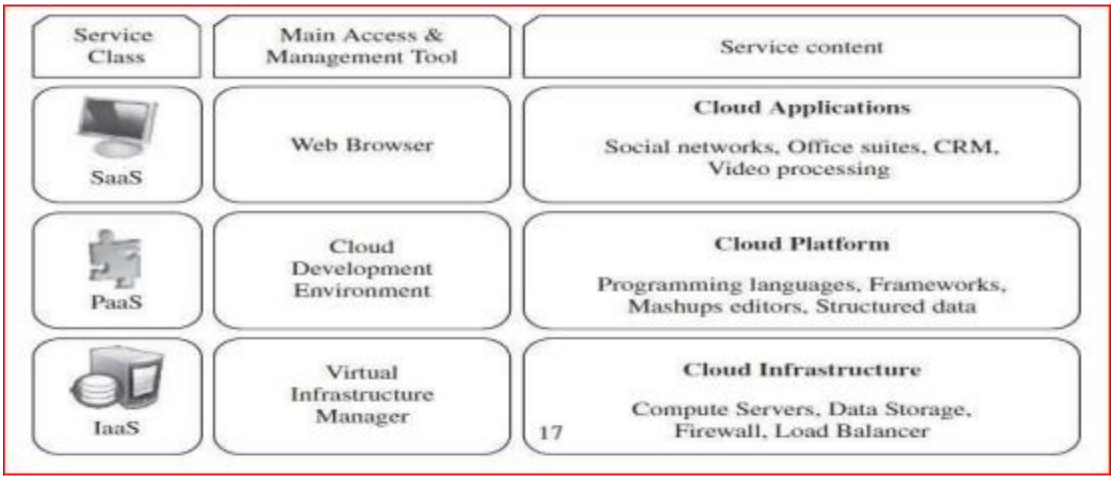
7. Autonomic Computing

Autonomic or self-managing, systems rely on monitoring probes and gauges (sensors), on an adaptation engine (autonomic manager) for computing optimizations based on monitoring data, and on effectors to carry out changes on the system. IBM's Autonomic Computing Initiative has contributed to define the four properties of autonomic systems: self-configuration, self- optimization, self-healing, and self-protection. IBM has also suggested a reference model for autonomic control loops of autonomic managers, called MAPE-K (Monitor Analyze Plan Execute—Knowledge).

LAYERS AND TYPES OF CLOUD

Cloud computing services are divided into three classes, according to the abstraction level of the capability provided and the service model of providers, namely:

- Infrastructure as a service (IaaS)
- Platform as a service (PaaS)
- Software as a service (SaaS)



1. Infrastructure as a service (IaaS)

Infrastructure as a Service (IaaS) helps in delivering computer infrastructure on an external basis for supporting operations. Generally, IaaS provides services to networking equipment, devices, databases, and web servers.

Infrastructure as a Service (IaaS) helps large organizations, and large enterprises in managing and building their IT platforms. This infrastructure is flexible according to the needs of the client.

Advantages of IaaS

- IaaS is cost-effective as it eliminates capital expenses.
- IaaS cloud provider provides better security than any other software.
- IaaS provides remote access.

Disadvantages of IaaS

- In IaaS, users have to secure their own data and applications.
- Cloud computing is not accessible in some regions of the World.

2. Platform as a service (PaaS)

Platform as a Service (PaaS) is a type of cloud computing that helps developers to build applications and services over the Internet by providing them with a platform.

PaaS helps in maintaining control over their business applications.

Advantages of PaaS

- PaaS is simple and very much convenient for the user as it can be accessed via a web browser.
- PaaS has the capabilities to efficiently manage the lifecycle.

Disadvantages of PaaS

- PaaS has limited control over infrastructure as they have less control over the environment and are not able to make some customizations.
- PaaS has a high dependence on the provider.

3. Software as a service (SaaS)

Software as a Service (SaaS) is a type of cloud computing model that is the work of delivering services and applications over the Internet. The SaaS applications are called Web-Based Software or Hosted Software.

SaaS has around 60 percent of cloud solutions and due to this, it is mostly preferred by companies.

Advantages of SaaS

- SaaS can access app data from anywhere on the Internet.
- SaaS provides easy access to features and services.

Disadvantages of SaaS

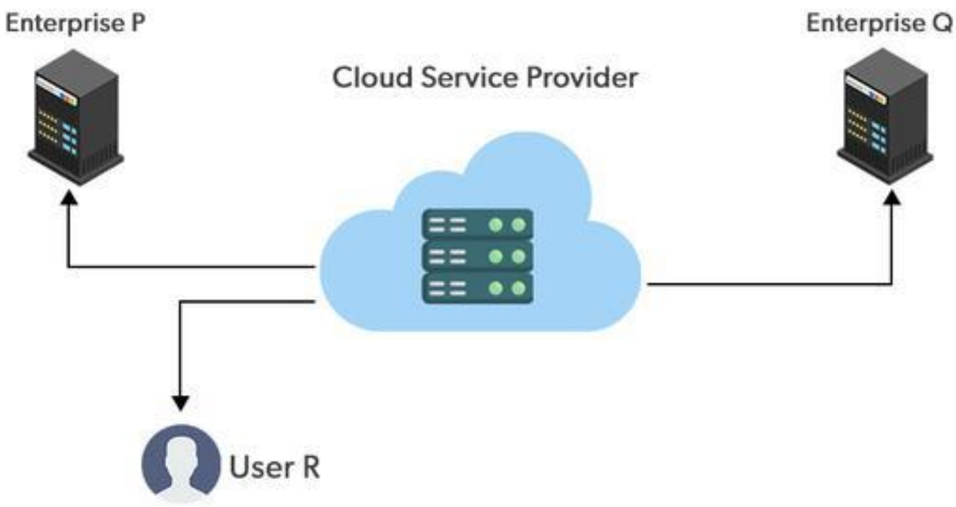
- SaaS solutions have limited customization, which means they have some restrictions within the platform.
- SaaS has little control over the data of the user.
- SaaS are generally cloud-based, they require a stable internet connection for proper working.

Different types of cloud computing deployment models are described below.

- Public Cloud
- Private Cloud
- Hybrid Cloud
- Community Cloud
- Multi-Cloud

Public Cloud

The public cloud makes it possible for anybody to access systems and services. The public cloud may be less secure as it is open to everyone. The public cloud is one in which cloud infrastructure services are provided over the internet to the general people or major industry groups. The infrastructure in this cloud model is owned by the entity that delivers the cloud services, not by the consumer. It is a type of cloud hosting that allows customers and users to easily access systems and services. This form of cloud computing is an excellent example of cloud hosting, in which service providers supply services to a variety of customers. In this arrangement, storage backup and retrieval services are given for free, as a subscription, or on a per-user basis. For example, Google App Engine etc.



Public Cloud

Advantages of the Public Cloud Model

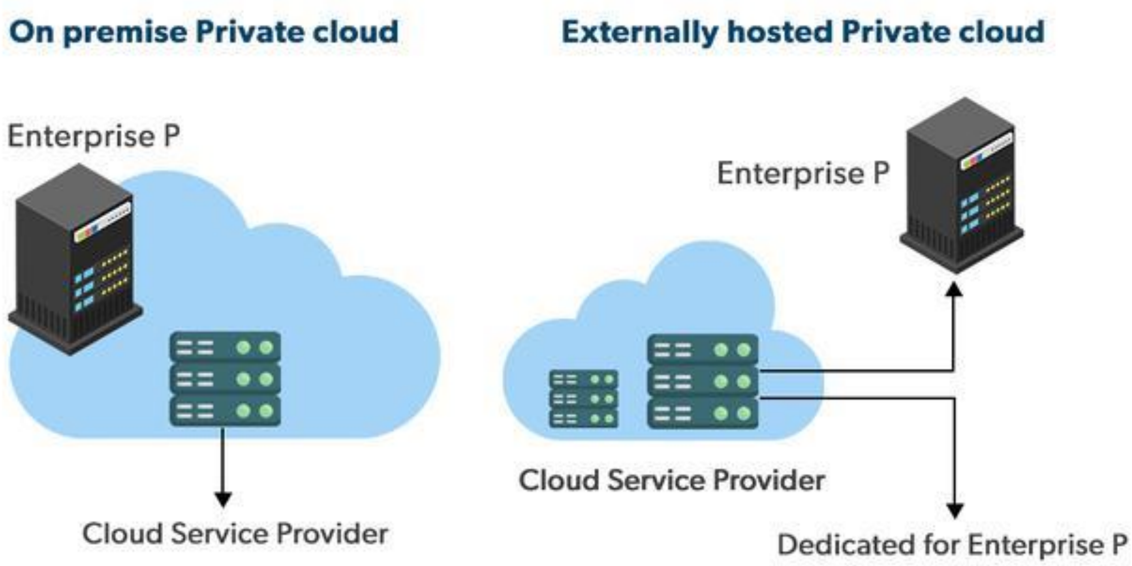
- Minimal Investment: Because it is a pay-per-use service, there is no substantial upfront fee, making it excellent for enterprises that require immediate access to resources.
- No setup cost: The entire infrastructure is fully subsidized by the cloud service providers, thus there is no need to set up any hardware.
- Infrastructure Management is not required: Using the public cloud does not necessitate infrastructure management.
- No maintenance: The maintenance work is done by the service provider (not users).
- Dynamic Scalability: To fulfill your company's needs, on-demand resources are accessible.

Disadvantages of the Public Cloud Model

- Less secure: Public cloud is less secure as resources are public so there is no guarantee of high-level security.
- Low customization: It is accessed by many public so it can't be customized according to personal requirements.

Private Cloud

The private cloud deployment model is the exact opposite of the public cloud deployment model. It's a one-on-one environment for a single user (customer). There is no need to share your hardware with anyone else. The distinction between private and public clouds is in how you handle all of the hardware. It is also called the "internal cloud" & it refers to the ability to access systems and services within a given border or organization. The cloud platform is implemented in a cloud-based secure environment that is protected by powerful firewalls and under the supervision of an organization's IT department. The private cloud gives greater flexibility of control over cloud resources.



Private Cloud

Advantages of the Private Cloud Model

- Better Control: You are the sole owner of the property. You gain complete command over service integration, IT operations, policies, and user behavior.
- Data Security and Privacy: It's suitable for storing corporate information to which only authorized staff have access. By segmenting resources within the same infrastructure, improved access and security can be achieved.

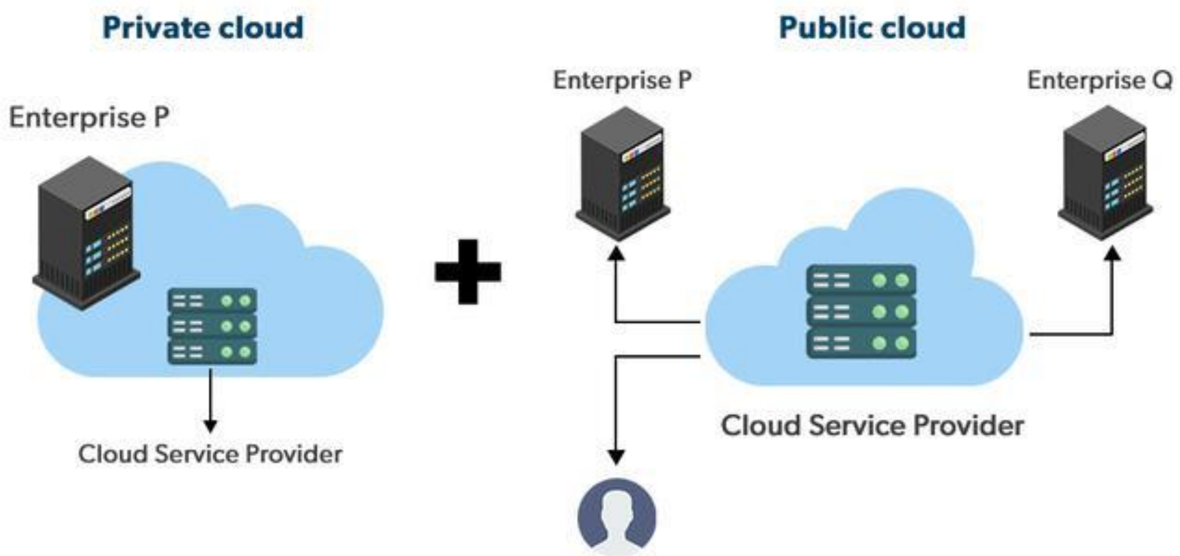
- Supports Legacy Systems: This approach is designed to work with legacy systems that are unable to access the public cloud.
- Customization: Unlike a public cloud deployment, a private cloud allows a company to tailor its solution to meet its specific needs.

Disadvantages of the Private Cloud Model

- Less scalable: Private clouds are scaled within a certain range as there is less number of clients.
- Costly: Private clouds are more costly as they provide personalized facilities.

Hybrid Cloud

By bridging the public and private worlds with a layer of proprietary software, hybrid cloud computing gives the best of both worlds. With a hybrid solution, you may host the app in a safe environment while taking advantage of the public cloud's cost savings. Organizations can move data and applications between different clouds using a combination of two or more cloud deployment methods, depending on their needs.



Hybrid Cloud

Advantages of the Hybrid Cloud Model

- Flexibility and control: Businesses with more flexibility can design personalized solutions that meet their particular needs.

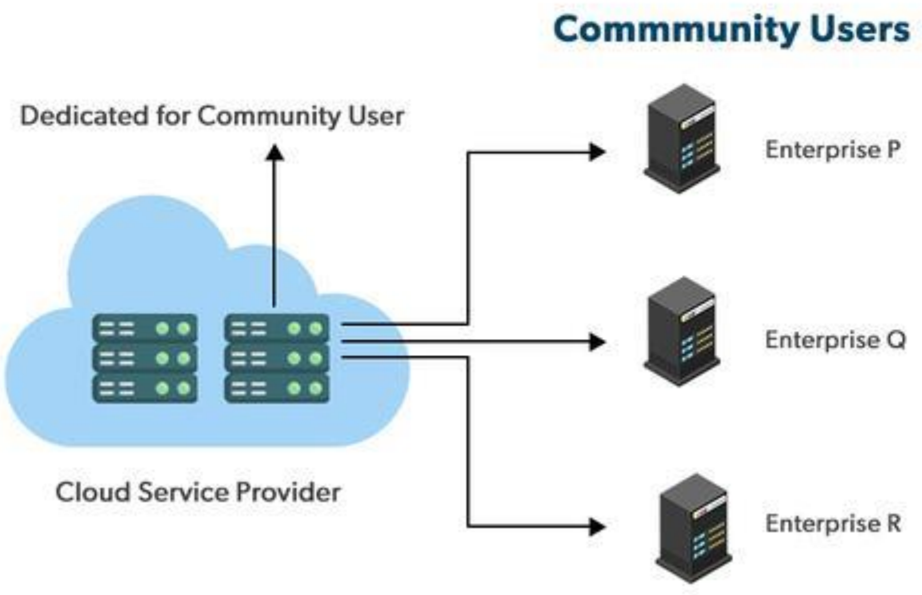
- Cost: Because public clouds provide scalability, you'll only be responsible for paying for the extra capacity if you require it.
- Security: Because data is properly separated, the chances of data theft by attackers are considerably reduced.

Disadvantages of the Hybrid Cloud Model

- Difficult to manage: Hybrid clouds are difficult to manage as it is a combination of both public and private cloud. So, it is complex.
- Slow data transmission: Data transmission in the hybrid cloud takes place through the public cloud so latency occurs.

Community Cloud

It allows systems and services to be accessible by a group of organizations. It is a distributed system that is created by integrating the services of different clouds to address the specific needs of a community, industry, or business. The infrastructure of the community could be shared between the organization which has shared concerns or tasks. It is generally managed by a third party or by the combination of one or more organizations in the community.



Advantages of the Community Cloud Model

- Cost Effective: It is cost-effective because the cloud is shared by multiple organizations or communities.
- Security: Community cloud provides better security.

- Shared resources: It allows you to share resources, infrastructure, etc. with multiple organizations.
- Collaboration and data sharing: It is suitable for both collaboration and data sharing.

Disadvantages of the Community Cloud Model

- Limited Scalability: Community cloud is relatively less scalable as many organizations share the same resources according to their collaborative interests.
- Rigid in customization: As the data and resources are shared among different organizations according to their mutual interests if an organization wants some changes according to their needs they cannot do so because it will have an impact on other organizations.

DESIRED FEATURES OF A CLOUD

The features of a cloud are essential to enable services that truly represent the cloud-computing model and satisfy expectations of consumers, and cloud offerings must be having following features:

1. Self-service
2. Per-usage metered and billed
3. Elastic,
4. Customizable.

1. Self-Service

Consumers of cloud computing services expect on-demand, nearly instant access to resources. To support this expectation, clouds must allow self-service access so that customers can request, customize, pay, and use services without intervention of human operators.

2. Per-Usage Metering and Billing

Cloud computing eliminates up-front commitment by users, allowing them to request and use only the necessary amount. Services must be priced on a shortterm basis (e.g., by the hour), allowing users to release (and not pay for) resources as soon as they are not needed. For these reasons, clouds must implement features to allow efficient trading of service such as pricing, accounting, and billing. Metering should be done accordingly for different types of service (e.g., storage, processing, and bandwidth) and usage promptly reported, thus providing greater transparency.

3. Elasticity

Cloud computing gives the illusion of infinite computing resources available on demand. Therefore, users expect clouds to rapidly provide resources in any quantity at any time. In particular, it is expected that the additional resources can be (a) provisioned, possibly automatically, when an application load increases and (b) released when load decreases (scale up and down).

4. Customization

In a multi-tenant cloud a great disparity between user needs is often the case. Thus, resources rented from the cloud must be highly customizable. In the case of infrastructure services, customization means allowing users to deploy specialized virtual appliances and to be given privileged (root) access to the virtual servers. Other service classes (PaaS and SaaS) offer less flexibility and are not suitable for general-purpose computing, but still are expected to provide a certain level of customization.

1. Cost: It reduces the huge capital costs of buying hardware and software.
2. Speed: Resources can be accessed in minutes, typically within a few clicks.
3. Scalability: We can increase or decrease the requirement of resources according to the business requirements.
4. Productivity: While using cloud computing, we put less operational effort. We do not need to apply patching, as well as no need to maintain hardware and software. So, in this way, the IT team can be more productive and focus on achieving business goals.
5. Reliability: Backup and recovery of data are less expensive and extremely fast for business continuity.
6. Security: Many cloud vendors offer a broad set of policies, technologies, and controls that strengthen our data security.

CLOUD INFRASTRUCTURE MANAGEMENT

A key challenge IaaS providers face when building a cloud infrastructure is managing physical and virtual resources, namely servers, storage, and net- works, in a holistic fashion.

The software toolkit responsible for this orchestration is called a virtual infrastructure manager (VIM). This type of software resembles a traditional operating system—but instead of dealing with a single computer, it aggregates resources from multiple computers, presenting a uniform view to user and applications. Other terms include infrastructure sharing software and virtual infra- structure engine.

There are two categories of tools used to manage cloud they are

1. Cloud toolkits - includes those that “expose a remote and secure interface for creating, controlling and monitoring virtualize resources,” but do not specialize in VI management.
2. The virtual infrastructure managers - provide advanced features such as automatic load balancing and server consolidation, but do not expose remote cloud-like interfaces.

The availability of a remote cloud-like interface and the ability of managing many users and their permissions are the primary features that would distinguish “cloud toolkits” from “VIMs.” However, here we place both categories of tools under the same group (of the VIMs) and, when applicable, we highlight the availability of a remote interface as a feature.

Virtually all VIMs we investigated present a set of basic features related to managing the life cycle of VMs, including networking groups of VMs together and setting up virtual disks for VMs. These basic features pretty much define whether a tool can be used in practical cloud deployments or not. On the other hand, only a handful of software present advanced features (e.g., high availability) which allow them to be used in large-scale production clouds.

Features:

Virtualization Support - The multi-tenancy aspect of clouds requires multiple customers with disparate requirements to be served by a single hardware infrastructure. Virtualized resources (CPUs, memory, etc.) can be sized and resized with certain flexibility. These features make hardware virtualization, the ideal technology to create a virtual infrastructure that partitions a data center among multiple tenants.

Self-Service, On-Demand Resource Provisioning- Self-service access to resources has been perceived as one the most attractive features of clouds. This feature enables users to directly obtain services from clouds, such as spawning the creation of a server and tailoring its software, configurations, and security policies, without interacting with a human system administrator. This capability eliminates the need for more time-consuming, labor-intensive, human- driven procurement processes familiar to many in IT. Therefore, exposing a self-service interface, through which users can easily interact with the system, is a highly desirable feature of a Vi manager.

Multiple Backend Hypervisors- Different virtualization models and tools offer different benefits, drawbacks, and limitations. Thus, some Vi managers provide a uniform management layer regardless of the virtualization technology used.

This characteristic is more visible in open-source Vi managers, which usually provide 15 pluggable drivers to interact with multiple hypervisors. In this direction, the aim of libvirt is to provide a uniform API that Vi managers can use to manage domains (a VM or container running an instance of an operating system) in virtualized nodes using standard operations that abstract hypervisor specific calls.

Storage Virtualization-Virtualizing storage means abstracting logical storage from physical storage. By consolidating all available storage devices in a data center, it allows creating virtual disks independent from device and location. Storage devices are commonly organized in a storage area network (SAN) and attached to servers via protocols such as Fiber Channel, iSCSI, and NFS; a storage controller provides the layer of abstraction between virtual and physical storage.

In the VI management sphere, storage virtualization support is often restricted to commercial products of companies such as VMWare and Citrix.

Interface to Public Clouds— Researchers have perceived that extending the capacity of a local in-house computing infrastructure by borrowing resources from public clouds is advantageous. In this fashion, institutions can make good use of their available resources and, in case of spikes in demand, extra load can be offloaded to rented resources. A VI manager can be used in a hybrid cloud setup if it offers a driver to manage the life cycle of virtualized resources obtained from external cloud providers. To the applications, the use of leased resources must ideally be transparent.

Virtual Networking— Virtual networks allow creating an isolated network on top of a physical infrastructure independently from physical topology and locations. A virtual LAN (VLAN) allows isolating traffic that shares a switched network, allowing VMs to be grouped into the same broadcast domain. Additionally, a VLAN can be configured to block traffic originated from VMs from other networks. Similarly, the VPN (virtual private network) concept is used to describe a secure and private overlay network on top of a public network (most commonly the public Internet) Support for creating and configuring virtual networks to group VMs placed throughout a data center is provided by most VI managers. Additionally, VI managers that interface with public clouds often support secure VPNs connecting local and remote VMs.

Dynamic Resource Allocation— In cloud infrastructures, where applications have variable and dynamic needs, capacity management and demand prediction are especially complicated. This fact triggers the need for dynamic resource allocation aiming at obtaining a timely match of supply and demand. Energy consumption reduction and better management of SLAs can be achieved by dynamically remapping VMs to physical machines at regular intervals. Machines that are not assigned any VM can be turned off or put on a low power state. A number of VI managers include a dynamic resource allocation feature that continuously monitors utilization across resource pools and reallocates avail- able resources among VMs according to application needs.

INFRASTRUCTURE AS A SERVICE PROVIDERS

Public Infrastructure as a Service providers commonly offer virtual servers containing one or more CPUs, running several choices of operating systems and a customized software stack. In addition, storage space and communication facilities are often provided.

Features

IAAS offers a set of specialized features that can influence the cost benefit ratio to be experienced by user applications when moved to the cloud.

The most relevant features are:

1.Geographic distribution of data centers.

2. Variety of user interfaces and APIs to access the system.

3. Specialized components and services that aid Particular applications (e.g., load-balancers, firewalls).

4. Choice of virtualization platform and operating systems and

5. Different billing methods and period (e.g., prepaid vs. postpaid, hourly vs. monthly).

Geographic Presence - To improve availability and responsiveness, a provider of worldwide services would typically build several data centers distributed around the world. For example, Amazon Web Services presents the concept of availability zones and regions 20 for its EC2 service. Availability zones are distinct locations that are engineered to be insulated from failures in other availability zones and provide inexpensive, low-latency network connectivity to other availability zones in the same region. Regions, in turn, are geographically dispersed and will be in separate geographic areas or countries.

User Interfaces and Access to Servers - Ideally, a public IaaS provider must provide multiple access means to its cloud, thus catering for various users and their preferences. Different types of user interfaces (UI) provide different levels of abstraction, the most common being graphical user interfaces (GUI), command-line tools (CLI), and Web service (WS) APIs.

GUIs are preferred by end users who need to launch, customize, and monitor a few virtual servers and do not necessarily need to repeat the process several times. On the other hand, CLIs offer more flexibility and the possibility of automating repetitive tasks via scripts (e.g., start and shutdown a number of virtual servers at regular intervals).

Advance Reservation of Capacity - Advance reservations allow users to request for an IaaS provider to reserve resources for a specific time frame in the future, thus ensuring that cloud resources will be available at that time. However, most clouds only support besteffort requests that means users can request server whenever resources are available .

Amazon Reserved Instances is a form of advance reservation of capacity, allowing users to pay a fixed amount of money in advance to guarantee resource availability at anytime during an agreed period and then paying a discounted hourly rate when resources are in use. However, only long periods of 1 to 3 years are offered; therefore, users cannot express their reservations in finer granularities—for example, hours or days.

Automatic Scaling and Load Balancing - Automatic scaling is a highly desirable feature of IaaS clouds. It allows users to set conditions for when they want their applications to scale up and down, based on application-specific metrics such as transactions per second, number of simultaneous users, request latency, and so forth.

When the number of virtual servers is increased by automatic scaling, incoming traffic must be automatically distributed among the available servers. This activity enables applications to promptly respond to traffic increase while also achieving greater fault tolerance.

Service-Level Agreement - Service-level agreements (SLAs) are offered by IaaS providers to express their commitment to delivery of a certain QoS. To customers it serves as a warranty. An SLA usually include availability and performance guarantees. Additionally, metrics must be agreed upon by all parties as well as penalties for violating these expectations.

Most IaaS providers focus their SLA terms on availability guarantees, specifying the minimum percentage of time the system will be available during a certain period. For instance, Amazon EC2 states that “if the annual uptime Percentage for a customer drops below 99.95%

for the service year, that customer is eligible to receive a service credit equal to 10% of their bill.3”

Hypervisor and Operating System Choice - Traditionally, IaaS offerings have been based on heavily customized open-source Xen deployments. IaaS providers needed expertise in Linux, networking, virtualization, metering, resource management, and many 21 other low-level aspects to successfully deploy and maintain their cloud offerings.

More recently, there has been an emergence of turnkey IaaS platforms such as VMWare VCloud and Citrix Cloud Center (C3) which have lowered the barrier of entry for IaaS competitors, leading to a rapid expansion in the IaaS marketplace.

PLATFORM AS A SERVICE PROVIDERS

Public Platform as a Service providers commonly offer a development and deployment environment that allow users to create and run their applications with little or no concern to low-level details of the platform. In addition, specific programming languages and frameworks are made available in the platform, as well as other services such as persistent data storage and in memory caches.

Features

Programming Models, Languages, and Frameworks: Programming models made available by IaaS providers define how users can express their applications using higher levels of abstraction and efficiently run them on the cloud platform.

Each model aims at efficiently solving a particular problem. In the cloud computing domain, the most common activities that require specialized models are: processing of large dataset in clusters of computers (MapReduce model), development of request-based Web services and applications; definition and orchestration of business processes in the form of workflows (Workflow model); and high-performance distributed execution of various computational tasks.

For user convenience, PaaS providers usually support multiple programming languages. Most commonly used languages in platforms include Python and Java (e.g., Google AppEngine), .NET languages (e.g., Microsoft Azure), and Ruby (e.g., Heroku). Force.com has devised its own programming language (Apex) and an Excel-like query language, which provide higher levels of abstraction to key platform functionalities.

A variety of software frameworks are usually made available to PaaS developers, depending on application focus. Providers that focus on Web and enterprise application hosting offer popular frameworks such as Ruby on Rails, Spring, Java EE, and .NET.

Persistence Options:

A persistence layer is essential to allow applications to record their state and recover it in case of crashes, as well as to store user data. Web and enterprise application developers have chosen relational databases as the preferred persistence method. These databases offer fast and reliable structured data storage and transaction processing, but may lack scalability to handle several peta bytes of data stored in commodity computers. In the cloud computing domain, distributed storage technologies have emerged, which seek to be robust and highly scalable, at the expense of relational structure and convenient query languages.

CHALLENGES AND RISKS

Despite the initial success and popularity of the cloud computing paradigm and the extensive availability of providers and tools, a significant number of challenges and risks are inherent to this new model of computing. Providers, developers, and end users must consider these challenges and risks to take good advantage of cloud computing. Issues to be faced include user privacy, data security, data lock-in, availability of service, disaster recovery, performance, scalability, energy- efficiency, and programmability.

Security, Privacy, and Trust: Security and privacy affect the entire cloud computing stack, since there is a massive use of third-party services and infrastructures that are used to host important data or to perform critical operations. In this scenario, the trust toward providers is fundamental to ensure the desired level of privacy for applications hosted in the cloud. Legal and regulatory issues also need attention. When data are moved into the Cloud, providers may choose to locate them anywhere on the planet. The physical location of data centers determines the set of laws that can be applied to the management of data. For example, specific cryptography techniques could not be used because they are not allowed in some countries. Similarly, country laws can impose that sensitive data, such as patient health records, are to be stored within national borders.

Data Lock-In and Standardization: A major concern of cloud computing users is about having their data locked-in by a certain provider. Users may want to move data and applications out from a provider that does not meet their requirements. However, in their current form, cloud computing infrastructures and platforms do not employ standard methods of storing user data and applications. Consequently, they do not interoperate and user data are not portable.

The answer to this concern is standardization. In this direction, there are efforts to create open standards for cloud computing. The Cloud Computing Interoperability Forum (CCIF) was formed by organizations such as Intel, Sun, and Cisco in order to "enable a global cloud computing ecosystem whereby organizations are able to seamlessly work together for the purposes for wider industry adoption of cloud computing technology." The development of the Unified Cloud Interface (UCI) by CCIF aims at creating a standard programmatic point of access to an entire cloud infrastructure. In the hardware virtualization sphere, the Open Virtual Format (OVF) aims at facilitating packing and distribution of software to be run on VMs so that virtual appliances can be made portable— that is, seamlessly run on hypervisor of different vendors.

Availability, Fault-Tolerance, and Disaster Recovery: It is expected that users will have certain expectations about the service level to be provided once their applications are moved to the cloud. These expectations include availability of the service, its overall performance, and what measures are to be taken when something goes wrong in the system or its components. In summary, users seek for a warranty before they can comfortably move their business to the cloud. SLAs, which include QoS requirements, must be ideally set up between customers and cloud computing providers to act as warranty. An SLA specifies the details of the service to be provided, including availability and performance 25 guarantees.

Additionally, metrics must be agreed upon by all parties, and penalties for violating the expectations must also be approved.

Resource Management and Energy-Efficiency: One important challenge faced by providers of cloud computing services is the efficient management of virtualized resource pools. Physical resources such as CPU cores, disk space, and network bandwidth must be sliced and shared among virtual machines running potentially heterogeneous workloads. The multi-dimensional nature of virtual machines complicates the activity of finding a good mapping of VMs onto available physical hosts while maximizing user utility. Dimensions to be considered include: number of CPUs, amount of memory, size of virtual disks, and network bandwidth. Dynamic VM mapping policies may leverage the ability to suspend, migrate, and resume VMs as an easy way of preempting low-priority allocations in favor of higher-priority ones. Migration of VMs also brings additional challenges such as detecting when to initiate a migration, which VM to migrate, and where to migrate. In addition, policies may take advantage of live migration of virtual machines to relocate data center load without significantly disrupting running services. In this case, an additional concern is the trade-off between the negative impact of a live migration on the performance and stability of a service and the benefits to be achieved with that migration. Another challenge concerns the outstanding amount of data to be managed in various VM management activities. Such data amount is a result of particular abilities of virtual machines, including the ability of traveling through space (i.e., migration) and time (i.e., check pointing and rewinding), operations that may be required in load balancing, backup, and recovery scenarios. In addition, dynamic provisioning of new VMs and replicating existing VMs require efficient mechanisms to make VM block storage devices (e.g., image files) quickly available at selected hosts. Data centers consumer large amounts of electricity. According to a data published by HP, 100 server racks can consume 1.3MWof power and another 1.3 MW are required by the cooling system, thus costing USD 2.6 million per year. Besides the monetary cost, data centers significantly impact the environment in terms of CO_2 emissions from the cooling systems.

ASSESSING THE ROLE OF OPEN STANDARDS:

Open Standards" are standards made available to the general public and are developed and maintained via a collaborative and mutually agreed process.

It facilitate interoperability and data exchange among different products or services and are intended for widespread adoption.

The cloud computing technology is the result of the convergence of many different standards.

Cloud computing changes the manner in which services and applications are deployed. Hence, the role of open standards becomes crucial.

The cloud computing industry is working with the following architectural standards:

- Platform virtualization of resources
- Service-oriented architecture
- Web-application frameworks
- Deployment of open-source software
- Standardized Web services
- Autonomic systems
- Grid computing

Without standards, the industry creates proprietary systems with vendor lock-in.

- Since, clients do not want to be locked into any single system, there is a strong industry push to create standards based clouds.
- These standards help to enable different business models that cloud computing vendors can support, such as SaaS, Web 2.0 applications, and utility computing.
- These businesses require open standards so that data is both portable and universally accessible.

Open Cloud platform technologies

OpenStack

An open source project by the company Rackspace.com(one of the largest Iaas cloud service provider).

OpenStack Compute software will automatically create large groups of virtual private servers on industry-standard systems.

OpenStack Storage is the software that will create redundant object-based storage using clusters of commodity servers and storage systems.

EUCALYPTUS

- The company Eucalyptus Systems was formed in 2009 to support the commercialization of the Eucalyptus Cloud Computing Platform.
- A Linux-based software platform for IaaS systems based on computer clusters. Most of the major Linux vendors support this project. It works with a number of technologies for system virtualization.
- It has an interface that can connect to Amazon's compute and storage cloud systems (EC2 and S3) .
- EUCALYPTUS: "Elastic Utility Computing Architecture for Linking Your Programs to Useful Systems".