Cloud Computing

Dr. Nirmeen Abd-Elwahab

Introduction

- Technologies such *as cluster, grid, and now, cloud computing,* have all aimed to allowing access to large amounts of computing power in a fully virtualized manner, by aggregating resources (CPU , memory , disk space …etc) and offering a single system view.
- Utility computing describes a business model for on-demand delivery of computing power; consumers pay providers based on usage (“pay as- you-go”), similar to the way in which we currently obtain services from traditional public utility services such as water, electricity, gas, and telephony
- ***The main principle behind this model is offering computing, storage, and software “as a service.”***

What is “cloud computing”?

Buyya et al. [2] have defined it as follows:

“Cloud is a parallel and distributed computing system consisting of a collection of inter-connected **and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements (SLA)** established through negotiation between the service provider and consumers.”

Vaquero et al. [3] have stated

“clouds are a large pool of easily usable and accessible virtualized resources (such as hardware, development platforms and/or services).

These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing also for ***an optimum resource utilization.***

This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customized Service Level Agreements.”

there are countless other definitions……………….

The key characteristics of cloud computing

- The National Institute of Standards and Technology (NIST) characterizes cloud computing as

a pay-per-use model for enabling available , convenient, *on-demand* network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, services) that can be rapidly provisioned and released with *minimal management effort* or service provider interaction."

- Common characteristics between all definitions, that a cloud should have:

(i) pay-per-use (no ongoing commitment, utility prices);

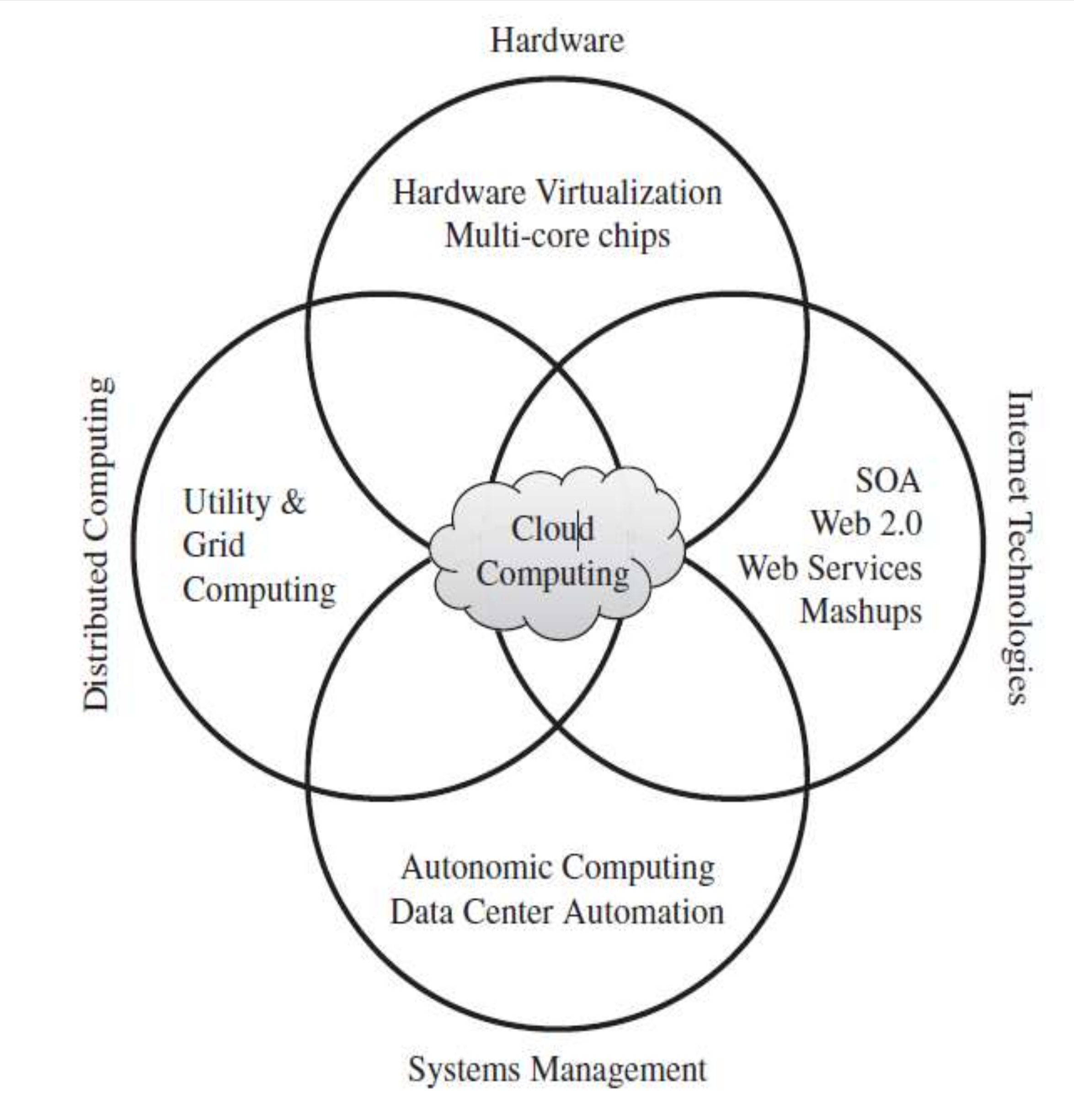
(ii) Elastic capacity and the illusion of infinite resources;

(iii) self-service interface;

(iv) resources that are abstracted or virtualized.

ROOTS OF CLOUD COMPUTING

Convergence of various advances leading to the advent of cloud computing.



From Mainframes to Clouds:

The mainframe era collapsed with the advent of fast and inexpensive microprocessors and IT data centers moved to collections of commodity servers.

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"

ROOTS OF CLOUD COMPUTING cont.

*** Internet technologies**

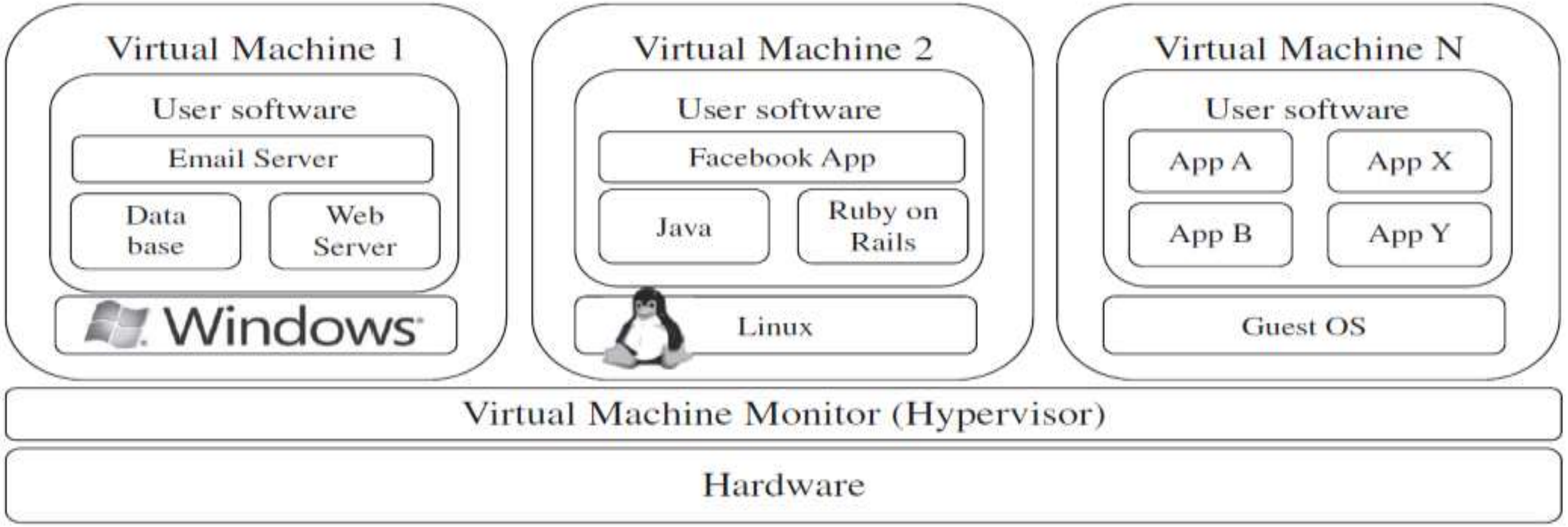
SOA, Web Services, Web 2.0, and Mash ups

Web Services (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

with the advent of Web 2.0, information and services may be programmatically aggregated, acting as building blocks of complex compositions, called service mash ups. Many service providers, such as Amazon, del.icio.us, Facebook, and Google, make their service APIs publicly accessible using standard protocols such as SOAP and REST

ROOTS OF CLOUD COMPUTING cont.



*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's 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.

ROOTS OF CLOUD COMPUTING cont.

*distributed computing

Grid Computing Grid computing enables aggregation of distributed resources and transparently access to them. Most production grids seek to share compute and storage resources distributed across different administrative domains,

The development of standardized protocols for several grid computing activities has contributed—theoretically—to allow delivery of on-demand computing services over the Internet.

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

ROOTS OF CLOUD COMPUTING cont.

Autonomic Computing *The increasing complexity* of computing systems has motivated research on autonomic computing, which seeks to improve systems by decreasing human involvement in their operation also ,
The large data centers of cloud computing providers must be managed in an efficient way. In this sense, the concepts of autonomic computing inspire software technologies for data center automation, which may perform tasks such as: *management of service levels* of running applications; *management of data center capacity*; *proactive disaster recovery; and automation of VM provisioning*

IBM's Autonomic Computing Initiative has contributed to define <u>*the four properties of autonomic systems*</u>: <u>self-configuration</u>,
<u>self optimization</u>, <u>self-healing</u>, and <u>self-protection</u>.
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 CLOUDS

Cloud computing services are divided into *three classes* :

1. Infrastructure as a Service, Offering virtualized resources (computation, storage, and communication) on demand is known as Infrastructure as a Service (**IaaS**)

2. Platform as a Service (PaaS) offers an environment on which developers create and deploy applications and do not necessarily need to know how many processors or how much memory that applications will be using.

3. Software as a Service (SaaS) offers applications to users through Web portals ,such as , word processing and spreadsheet can now be accessed as a service in the Web.

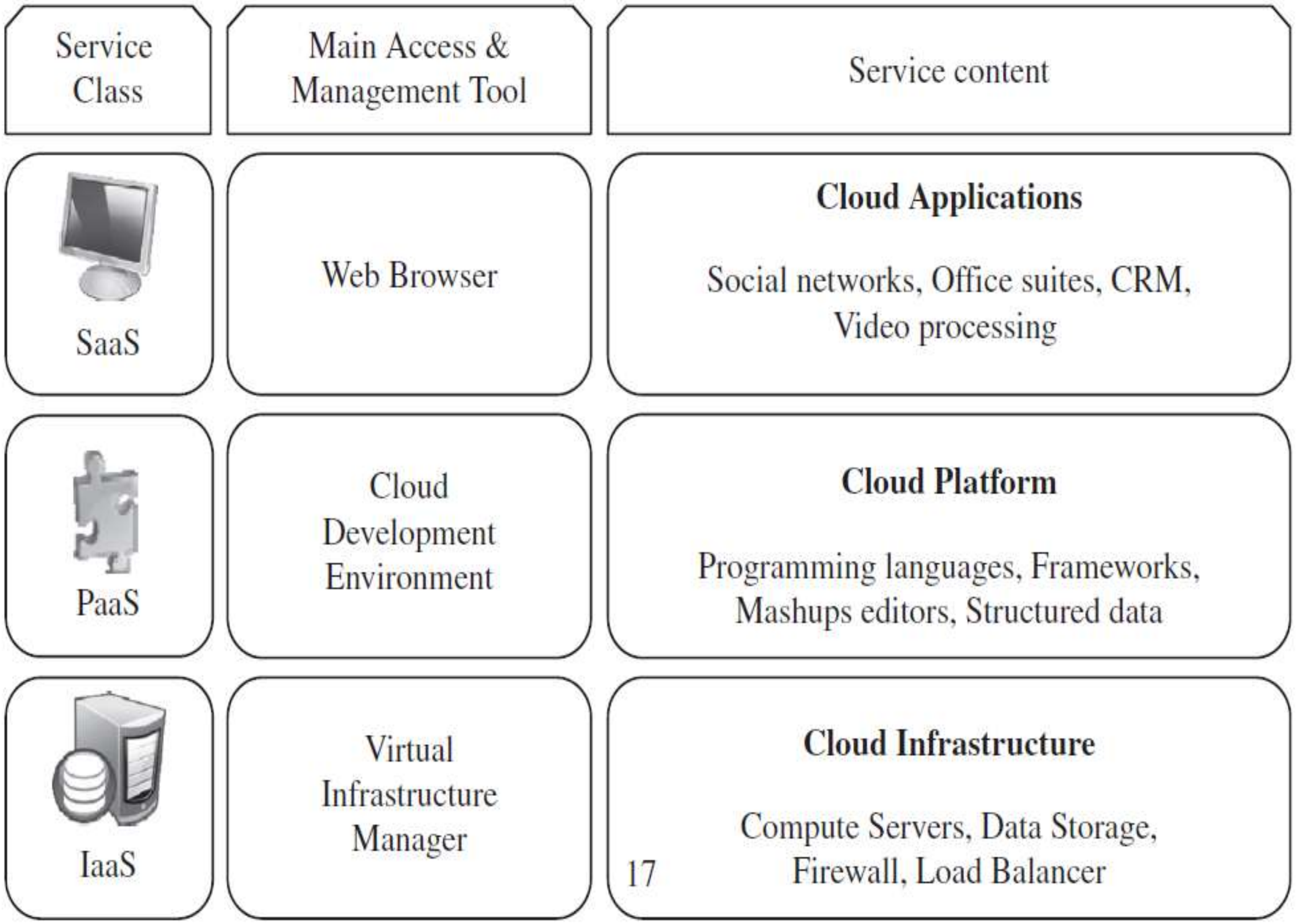


FIGURE 1.3. The cloud computing stack.

Deployment Models

A cloud can be classified as

- public,
- private,
- community,
- hybrid

community cloud
is “shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations).”

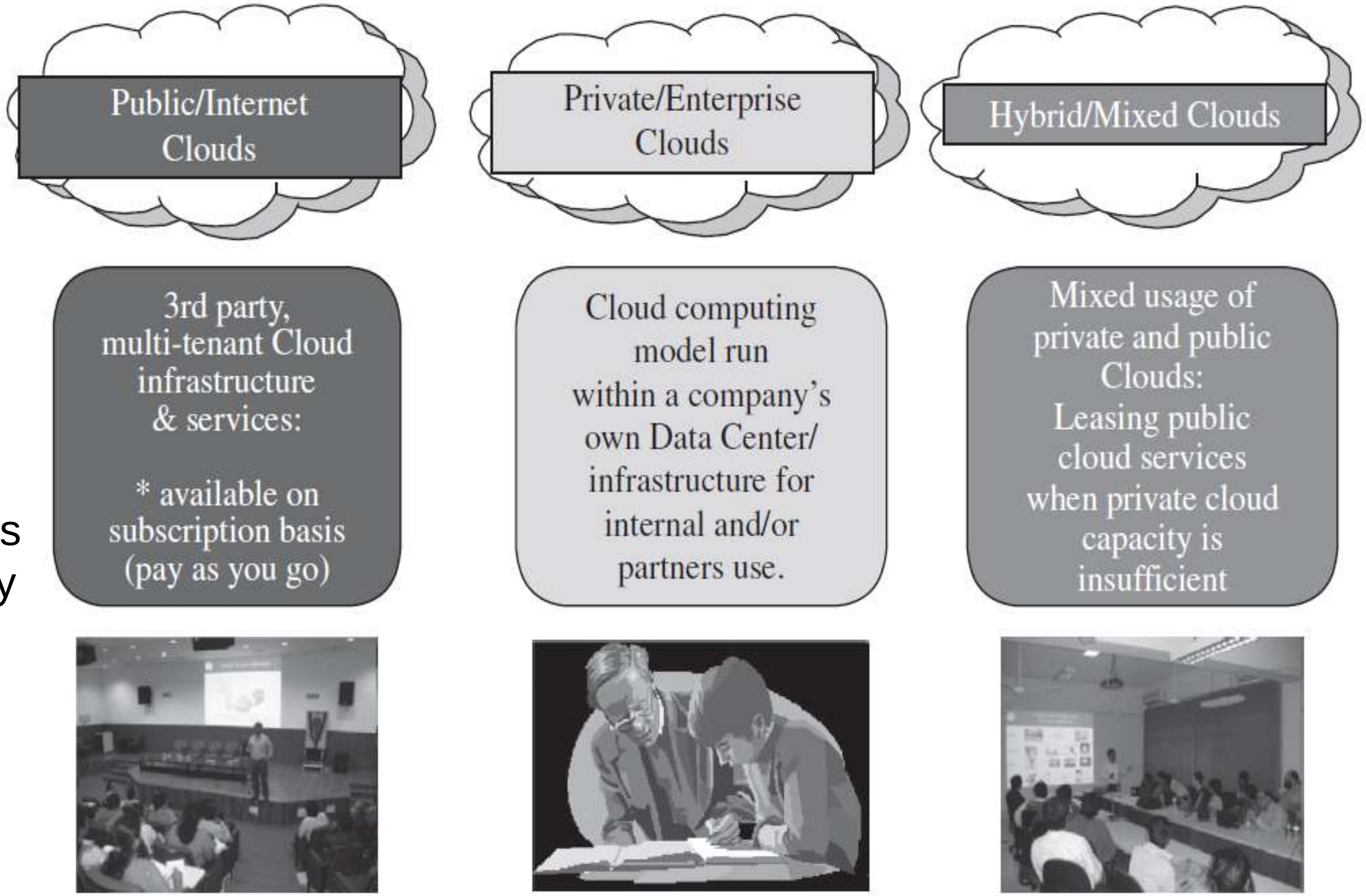


FIGURE 1.4. Types of clouds based on deployment models.

DESIRED FEATURES OF A CLOUD

Certain features of a cloud are essential to enable services

(i) self-service, clouds must allow self-service access so that customers can request, customize, pay, and use services without intervention of human operators

(ii) per-usage metered and billed, 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

(iii) 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)*

(iv) 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.

CLOUD INFRASTRUCTURE MANAGEMENT

A key challenge **IaaS** providers face when building a cloud infrastructure *is managing physical and virtual resources, namely servers, storage, and networks , in a holistic fashion.* ***to rapidly and dynamically provision resources to applications***

Software tools:

1. ***cloud toolkits:*** creating, controlling and monitoring virtualize resources. can also manage virtual infrastructures.
2. ***Virtual infrastructure manager (VIM):*** provide advanced features, it aggregates resources from multiple computers, presenting a uniform view to user and applications.

Features of VIMs

1. ***Virtualization Support:*** 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.*
2. ***Self-Service, On-Demand Resource Provisioning:*** configurations, and security policies, without interacting with a human system administrator. This capability “eliminates the need for more time-consuming”.
3. ***Storage Virtualization.*** Storage devices are commonly organized in *a storage area network (**SAN**) and attached to servers via protocols to creating virtual disks* .
4. ***Interface to Public Clouds.*** In this fashion, institutions can make good use of their available resources and, *it can rented extra resources on demand*.
5. ***Virtual Networking:*** Virtual networks (**VLAN**) allow creating an isolated network on physical infrastructure independently. Additionally, *a VLAN can be configured to block traffic originated from VMs from other networks.*

Features of VIMs cont.

6. ***Dynamic Resource Allocation.*** In cloud infrastructures, where applications have variable and dynamic needs, capacity management and demand prediction are especially complicated. *That's why **VMs** make a dynamically remapping to physical machines.*
7. ***Virtual Clusters.*** Several **VI managers** can manage groups of **VMs**.
8. ***Reservation and Negotiation Mechanism.*** To support complex requests. When users request computational resources to available at a specific time, *requests are termed advance reservations (**AR**).*
9. ***High Availability and Data Recovery.*** The high availability (**HA**) feature of VI managers *aims at minimizing application downtime and preventing business put off*. The **HA** solution monitors failures of system components such as servers, **VMs**, disks, and network.
   - **and ensures that a duplicate VM serves the application in case of failures.**
   - **Data backup in clouds should take into account the high data volume involved in VM management.**

TABLE 1.1. Feature Comparison of Virtual Infrastructure Managers

	License	Installation Platform of Controller	Client UI, API, Language Bindings	Backend Hypervisor(s)	Storage Virtualization	Interface to Public Cloud	Virtual Networks	Dynamic Resource Allocation	Advance Reservation of Capacity	High Availability	Data Protection
Apache VCL	Apache v2	Multi-platform (Apache/ PHP)	Portal, XML-RPC	VMware ESX, ESXi, Server	No	No	Yes	No	Yes	No	No
AppLogic	Proprietary	Linux	GUI, CLI	Xen	Global Volume Store (GVS)	No	Yes	Yes	No	Yes	Yes
Citrix Essentials	Proprietary	Windows	GUI, CLI, Portal, XML-RPC	XenServer, Hyper-V	Citrix Storage Link	No	Yes	Yes	No	Yes	Yes
Enomaly ECP	GPL v3	Linux	Portal, WS	Xen	No	Amazon EC2	Yes	No	No	No	No
Eucalyptus	BSD	Linux	EC2 WS, CLI	Xen, KVM	No	EC2	Yes	No	No	No	No
Nimbus	Apache v2	Linux	EC2 WS, WSRF, CLI	Xen, KVM	No	EC2	Yes	Via integration with OpenNebula	Yes (via integration with OpenNebula)	No	No
OpenNEbula	Apache v2	Linux	XML-RPC, CLI, Java	Xen, KVM	No	Amazon EC2, Elastic Hosts	Yes	Yes	Yes (via Haizea)	No	No
OpenPEX	GPL v2	Multiplatform (Java)	Portal, WS	XenServer	No	No	No	No	Yes	No	No
oVirt	GPL v2	Fedora Linux	Portal	KVM	No	No	No	No	No	No	No
Platform ISF	Proprietary	Linux	Portal	Hyper-V XenServer, VMWare ESX	No	EC2, IBM CoD, HP Enterprise Services	Yes	Yes	Yes	Unclear	Unclear
Platform VMO	Proprietary	Linux, Windows	Portal	XenServer	No	No	Yes	Yes	No	Yes	No
VMWare vSphere	Proprietary	Linux, Windows	CLI, GUI, Portal, WS	VMware ESX, ESXi	VMware vStorage VMFS	VMware vCloud partners	Yes	VMware DRM	No	Yes	Yes

Features of Infrastructure as a Service Providers

1. ***Geographic Presence.*** *To improve availability and responsiveness*, a provider of worldwide services *would typically build several data centers distributed around the world to be insulated from probable failures*. For example, **Amazon** Web Services.
2. ***User Interfaces and Access to Servers.*** GUIs, Different types of user interfaces (UI), are preferred by end users who need to launch, customize, and monitor a few virtual servers and do not necessary need to repeat the process several times.
3. ***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*.
4. ***Automatic Scaling and Load Balancing.*** Applications often need to scale **up** and **down** to meet varying load conditions. 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.*

Features of Infrastructure as a Service Providers cont.

5. ***Service-Level Agreement.*** That's meaning the reliability and availability of the system. *the maximum percentage of time the system will be available during a certain period according to Service-Level agreement.*
6. ***Hypervisor and Operating System Choice.*** **IaaS** providers needed expertise in Linux, networking, virtualization, metering, resource management, and many other low-level aspects to successfully deploy and maintain their cloud offerings.

TABLE 1.2. Feature Comparison Public Cloud Offerings (Infrastructure as a Service)

	Geographic Presence	Client UI API Language Bindings	Primary Access to Server	Advance Reservation of Capacity	SLA Uptime	Smallest Billing Unit	Hypervisor	Guest Operating Systems	Automated Horizontal Scaling	Load Balancing	Runtime Server Resizing/ Vertical Scaling	Instance Hardware Capacity		
												Processor	Memory	Storage
Amazon EC2	US East, Europe	CLI, WS, Portal	SSH (Linux), Remote Desktop (Windows)	Amazon reserved instances (Available in 1 or 3 years terms, starting from reservation time)	99.95%	Hour	Xen	Linux, Windows	Available with Amazon CloudWatch	Elastic Load Balancing	No	1–20 EC2 compute units	1.7–15 GB	160–1690 GB 1 GB–1 TB (per EBS volume)
Flexiscale	UK	Web Console	SSH	No	100%	Hour	Xen	Linux, Windows	No	Zeus software loadbalancing	Processors, memory (requires reboot)	1–4 CPUs	0.5–16 GB	20–270 GB
GoGrid		REST, Java, PHP, Python, Ruby	SSH	No	100%	Hour	Xen	Linux, Windows	No	Hardware (F5)	No	1–6 CPUs	0.5–8 GB	30–480 GB
Joyent Cloud	US (Emery-ville, CA; San Diego, CA; Andover, MA; Dallas, TX)		SSH, VirtualMin (Web-based system administration)	No	100%	Month	OS Level (Solaris Containers)	OpenSolaris	No	Both hardware (F5 networks) and software (Zeus)	Automatic CPU bursting (up to 8 CPUs)	1/16–8 CPUs	0.25–32 GB	5–100 GB
Rackspace Cloud Servers	US (Dallas, TX)	Portal, REST, Python, PHP, Java, C#/.NET	SSH	No	100%	Hour	Xen	Linux	No	No	Memory, disk (requires reboot) Automatic CPU bursting (up to 100% of available CPU power of physical host)	Quad-core CPU (CPU power is weighed proportionally to memory size)	0.25–16 GB	10–620 GB

Features of Platform as a Service Providers

1. ***Programming Models, Languages, and Frameworks.*** *A variety of software frameworks are usually made available to **PaaS** developers*, depending on application focus. Programming models made available by **IaaS** providers (*how users can express their applications using higher efficiently run them on the cloud platform*).
2. ***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*. For example, Amazon and Google have a data store offer schema-less for that.

TABLE 1.3. Feature Comparison of Platform-as-a-Service Cloud Offerings

	Target Use	Programming Language, Frameworks	Developer Tools	Programming Models	Persistence Options	Automatic Scaling	Backend Infrastructure Providers
Aneka	.Net enterprise applications, HPC	.NET	Standalone SDK	Threads, Task, MapReduce	Flat files, RDBMS, HDFS	No	Amazon EC2
AppEngine	Web applications	Python, Java	Eclipse-based IDE	Request-based Web programming	BigTable	Yes	Own data centers
Force.com	Enterprise applications (esp. CRM)	Apex	Eclipse-based IDE, Web-based wizard	Workflow, Excel-like formula language, Request-based web programming	Own object database	Unclear	Own data centers
Microsoft Windows Azure	Enterprise and Web applications	.NET	Azure tools for Microsoft Visual Studio	Unrestricted	Table/BLOB/ queue storage, SQL services	Yes	Own data centers
Heroku	Web applications	Ruby on Rails	Command-line tools	Request-based web programming	PostgreSQL, Amazon RDS	Yes	Amazon EC2
Amazon Elastic MapReduce	Data processing	Hive and Pig, Cascading, Java, Ruby, Perl, Python, PHP, R, C++	Karmasphere Studio for Hadoop (Net-Beans-based)	MapReduce	Amazon S3	No	Amazon EC2

CHALLENGES AND RISKS

A significant number of challenges and risks are inherent to this new model of computing:

- **Data security,**
- **data lock-in,**
- **availability of service,**
- **disaster recovery,**
- **performance,**
- **scalability,**
- **energy-efficiency,**
- **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

Data Lock-In and Standardization

- *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

- The development of the Unified Cloud Interface (UCI) by CCIF aims to create a standard programmatic point of access to an entire cloud infrastructure.

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.

- An SLA specifies the details of the service to be provided, including availability and performance guarantees.

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.*

- *Data centers consumer large amounts of electricity*. 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.

- Overcome energy consuming by *judiciously consolidating workload onto smaller number of servers and turning off idle resources*.

Thanks

?