Cloud computing is a combination of many resources interacting together by sharing and pooling resources efficiently. Resources: Storage, procesing, network bankwith, etc

<u>Services offered by the cloud</u>

- Compute (Compute power):
  * Virtual machines can be deployed rapidly
  * Hosted and bare metal hypervisors allow VM deployment as needed
  * Dynamic and static hard drives are used for VM deployements

- Storage

- Platform for running App and App development

- Containers

<u>The five main principles of cloud computing and benefits</u>

1 Pools of computing resources (Resource pooling): available to any subscribing user (Capabilities are available over the network)

  * The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.
  * It utilize pooled computing asset that can be externally purchased
  * The shift from Capital Expenses (CAPEX) to Operating expenses (OPeX)

There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.

2 Virtualization: High utilization of hardware

  * Virtualization is vital to the cloud because each server takes physical space and uses significant power and cooling. So getting high utilization of each server is vital to be cost effective

3 Elasticity: Dynamic scale without CAPEX

  * Synonym for dynamic scaling, refers to the ability to dynamically change how much resource is consumed in response to how much in needed.
  * **Resources can be requested as required:**
  * **Ability to scale up and scale down again as needed:**
    ~ Vertical scaling: An IT resource is scaled up by replacing it with a more powerful IT resource
  * **Ability to scale out and scale back again as needed:**
    ~ Horizontal Scaling): by adding more of the same IT resources
  * **Ability to scale applications: scaling up and scaling out**

4 Automation: Build, deploy, configure, provision, and move, all without manual intervention

The ability to automatically (via an API) provision and deploy a new virtual instance and, equivalently, to be able to free or de-provision an instance. A cloud-deployed App can provision new instances on an as-needed basis.

After the peak demand ebbs, and you don't need the additional resources, these virtual instances can be taken offline and you won't longer be billed

  * **Ability to automate** how our cloud platform responds to changes in demands

5 Metered billing: pay only for what you use

  * Metered billing model (like the electricity)
  * In the case of managed hosting there typically is an initial fee and an annual contract fee
  * The cloud model is pay as you go. There is no annual contract and no commitment for a specific level of consumption.

- On-demand self-service:A consumer can unilaterally provision computing capabilities as needed automatically without requiring human interaction with each service provider.

and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

- Monitoring and Load testing in the cloud

  * We can monitor our compute resources through various methods such as work metrics and resources metrics
  * The monitoring can help us determine whether we have sufficient compute resources available
  * The monitoring can help our application respond to change in demand to provide more compute resources or less compute resources as needed
  * **We can prepare for large demand by load testing our cloud-based compute resources**
  * **Measured services:** Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

- **Available via the Internet (Broad network access)** Capabilities are available over the network.

- **Private clouds can be available via private WAN**

<u>Cloud storage</u>

Advantages of cloud storage (online storage)

- Information is available from anywhere (from any network device)
- Files are backed up in case of disaster
- Online storage can supplement our local and off-site backups to help ensure redundancy of our important App and data.
- Your data will have additional fault tolerance
- Basic services are free
- You can have as much storage that you want if pay for it

Disadvantages of cloud storage:

Limited storage for free accounts (5Gb to 15Gb on avarage)
Accessible by others
Some cloud providers have gone out of business in the past
A network connection is required

<u>Utility model of resource usage</u>

- You pay-as-you-go and only pay for what you use
- You are billed for you usage
- You will pay more if you need more compute resources
- You will pay less if you need less compute resources
- The cloud bring us 'Utility Computing'

<u>Essential Characteristics</u>

On-Premise Computing:

- Requires hardware, space, electricity, cooling
- Requires managing OS, applications and updates
- Software Licensing
- Difficult to scale:
- Too much or too little capacity
- High upfront capital costs
- <u>You have complete control</u>

<u>Deployment models</u>

Private cloud: The cloud infrastructure is provisioned for exclusive use by a single organization...

Community cloud: The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns

Public cloud: The cloud infrastructure is provisioned for open use by the general public

Hybrid cloud: The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public)

Service Models

- Infrastructure as a Service (IaaS) :

The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where 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, and deployed applications

- Amazon Elastic Compute Cloud (EC2)

- Platform as a Service (PaaS)

The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

- Google App Engine , Microsoft Azure

- Software as a Service (SaaS)

The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

- Framework as a Service (FaaS): Allows developers to to exted the prebuilt functionality of the SaaS applications

- force.com exteds the selesforce.com SaaS

Load balancing

What is load balancing (load balancer)?

It's a technique (or we could say a device: **load balancer)** used to distribute the incoming traffic among available servers so that the request can be handled and the response given at a faster rate.

Why do we need load balancers?

For example: sometimes many systems can flood the resources of a target web server with many requests all at once. When a server is overloaded with connections, new connections cannot be handled.

- In order to minimize downtime and delays for the requests coming to a server, cloud providers also provide load balancer as a service (LBaaS) that distributes the requests across multiple instances of the application

- The load balancers increase performance with optimal utilization of the server instances

- Providing **high availability** is one of the driving factors of load balancing

Load balancing strategies:

Least-connection

Round-robin

What are blade servers

A blade server is a stripped-down server computer with a modular design optimized to minimize the use of physical space and energy.

Moving workloads to the cloud

- Real machines can be ritualized and moved to the cloud
- Virtual machines on earth can be migrated to the cloud
- Live migration allows the machine to keep running during the move
- To connect to the virtual machine:
  - RDP to Window VM's
  - SSH for linux

Practical consideration when choosing a cloud provider (7-1 Practical considerations when choosing a cloud provider.pdf)

* '''Service-level agreements (SLAs)'''
** Each vendor has their own specific SLAs that specify (entre otras cosas) their guaranteed level of availability

* '''Estabilidad de la compañía:''' Al escoger un proveedor, uno de los factores que se debe considerar es la estabilidad de la compañía. Si la compañía sale del mercado, se tendría que cambiar de proveedor y esto puede ser un dolor de cabeza. Es por ello que las compañías generalmente escogen una compañía grande y estable (Google, Amazon, Microsoft) a menos que otra compañía ofrezca servicios más adaptados a sus necesidades.
** Is the company financially stable
** How long has the vendor been around

'''Tehnical operation considerations:'''
* '''Availability:''' El servicio tiene que estar disponible sin interrupciones.
* '''Performance:''' how fast it is.

** "Availability:''' The availability of a system is often measured in 9s, which describe the percent valud of availability (is specified in the SLAs):
*** Three 9s: 99.9% availability
*** Five 9s: 99.999% availability

** The availability and performance are due to many factors:
*** How redundant and well provisioned the cloud vendor's data center are.
*** Cloud infrastructures are buil with a huge number of servers and are designed with the expectation that the individual components in the system might fail.

When disigning your cloud model, you can adopt tow possible strategies

* Desigh with the posibility of failure in mind
* Plan to fail fast but recover quickly

The fundamental economic benefit that cloud computing brigs to the table is related to the magical conversion of CAPEX to OPEX. The initial barrier of starting a project is drastically reduced

Changes in the overall business model that businesses traditionally followed

- Ability to convert CAPEX into OPEX
- Ability to offset of even avoid CAPEX and possibly reduce OPEX as well.

Disaster preparedness and recovery

DRaaS: Recovery as a service (RaaS), sometimes referred to as disaster recovery as a service (DRaaS), is a category of cloud computing used for protecting an application or data from a natural or human disaster or service disruption at one location by enabling a full recovery in the cloud. RaaS differs from cloud-based backup services by protecting data and providing standby computing capacity on demand to facilitate more rapid application recovery. RaaS capacity is delivered in a cloud-computing model so recovery resources are only paid for when they are used, making it more efficient than a traditional disaster recovery warm site or hot site where the recovery resources must be running at all times.

BaaS: Mobile backend as a service (MBaaS), also known as "backend as a service" (BaaS), is a model for providing web app and mobile app developers with a way to link their applications to backend cloud storage and APIs exposed by back end applications.