1 Introduction

The management of data this days can be difficult, it varies in volume, velocity, variety... and needs to be processed, stored, transfered... . Handling large scale distributed systems can involves high costs and latency. So it has been developed a new system to offer services: the Cloud.

A cloud service requires connectivity, reliability, efficiency, scalability, security and a pay-as-you-go system

By definition, cloud computing is a model for enabling on-demand network access to computing resources with minimal management effort or service provider interaction. The key characteristics are:

- shared resources
- broad network access
- On-demand automated reservation
- rapid elasticity
- Pay by use model

1.0.1 Obstacles

- Availability
- Data lock-in
- Data confidentiality
- Data transfer capability

Cloud services are moved into IoT devices (fog computing) to achieve various achievemt like lower latency, bandwidth costs and improve data privacy and resource ownership. This doesn't come without challenges. One of them is resource orchestration, or

2 Cloud ecosystem

Cloud computing is defined by deployment model, delivery models, infrastructure, resources and defining attributes.

Thanks to the NIST reference model, we identify, as actors of cloud computing services, the following:

- the service consumer,
- the broker, which manages relationships between providers and consumers
- the service provider
- the auditor, which makes independent assessment of the performance and security of a cloud service
- the carrier, which provides connectivity

2.0.1 Virtualization

Is the ability to hide the physical characteristics of the resources to the applications, systems and users using them

2.0.2 Tenancy

A single or multi -tenancy cloud service, is determined if the customer interacts with a personal instance of the software or a shared one. In the second case, the group of users is called a tenant. The two methods determine how resource management and costs are divided between users

2.0.3 Elasticity

This is the property to increase or decrease resource as needed, all within a short time. This allowes to fight the episodes of under or overprovisioning, or underutilisation and saturation of resources

Exercise:

peak demand (Pd)	average utilisation (Au)
pay as you go (Pg)	buying cost (Bc)
$Pd \times 24 =$ buying utilisation server hours (BuSh)	
$Au \times 24 =$ cloud utilisation server hours (CuSh)	

2.1 Transference and migration

2.2 Deployment models

they determine the ownership, size and access to cloud services

2.2.1 Public cloud

Developed for the general public or a lange industry group by an organisation sellid the service. The resources and infrastructure are managed by the provider. It uses a multi-tenancy model

2.2.2 Private cloud

Used only by an organisation, is developed by the consuming oranisation or another party. Resource are located on or off-premise based on the company preference

2.2.3 Community cloud

Used by a group of organisation with common concern (like policies or security considerations), is managed by the organisations or a third party

2.2.4 Hybrid cloud

It's a mashup of different type of clouds based on the various concerns

2.3 Delivery models

2.3.1 Software-as-a-Service (SaaS)

2.3.2 Platform-as-a-Service (PaaS)

2.3.3 Infrastructure-as-a-Service (IaaS)

2.4 The ecosystem

The main cloud providers are Amazon (Iaas), Google and Microsoft (SaaS and Paas), but we have also open source services like aucalyptus, OpenNebula and OpenStack

2.4.1 AWS

Composed of interconnected servers with high speed connection, it offers computing and storage services based on availability zones with different prices.

The user chooses an availability zone and instance type, which determines the hardware's specs. Then a Virtual Machine is installed on a located system and a IP address is provided (through DHCP). The user is able to interact through an AWS Management console, SDK libraries or raw REST requests.

Some examples of AWS services

- EC2 - Elastic Cloud Computing
  web service for launching applications under various OS
- S3 - Simple Storage System
  Service used to store large amount of data
- EBS - Elastic Block Store
  Provides block storage to EC2 instances, which sees them as disks
- SimpleDB
- SQS - Simple Queue Service
- CloudWatch

2.4.2 Google

Google offers both Saas like Gmail,docs, calendar... and PaaS: AppEngine, Google Drive, Google Base...

2.4.3 Microsoft

3 Virtualization

The first datacenters had most servers idle because of the inability of OSs to provide isolated ambients for multiple services (One application per server). To solve this, it has been implemented the virtualization of multiple servers inside the same physical machine

Virtualisation broadly describes the seperation of a service request from the underlying physical delivery of that service
(VMware definitio)

Virtualization offers various advantages and disadvantages:

Isolation	Additional overhead
Consolidation	more difficult handling of
Optimized energy consuption	heterogeneous hardware
Flexibility and agility	
easier disaster recovery	
rapid deployment of new servers	

3.1 Definitions

3.1.1 Layering

Used to simplify system complexity, in virtualization it separates hardware, software, OS, libraries and applications. The interfaces who manage the communication between layers are

- Application Program Interface (API)
- Application Binary Interface (ABI)
- Instruction Set Architecture (ISA)

3.1.2 Other definitions

Virtual Machine: Software emultaion fo a physical machine

Host OS: OS running on the physical machine

Guest OS: OS running on the VM

3.1.3 Hypervisor (VMM)

Is the software in charge of the virtualization, meaning assigning resources to each VM while guaranteering that different VM won't overlapt.

In practice are stripped-down OS, with a set of native drivers to manage hardware. The Virtual Machine Monitor must satisfy three characteristics:

- The environments virtualized must be identical to a real machine
- It must be efficient
- Should have the complete control of the physical resources

3.1.4 Other definitions

Virtual Hardware: HW provided by the VMM with the same characteristics of a given HW profile

3.2 CPU Virtualization

When a VMM assigns a CPU to the VM, it might use a different ISA from the physical architecture, so we need to emulate the CPU instead of virtualising it. This might result is lesser efficient VM.

3.2.1 X86

In the X86 world we have three levels of virtualization:

- Full virtualization
- Paravirtualization
- Hardware assisted virtualization

And defines 4 priviledge ring levels, ascending from 0 to 3 from most priviledged (level 0 is reserved for the kernel) to less priviledged. This is used to implement two model: 0/1/3 and 0/3/3 (VMM/guest OS/applications).
Using this system allowes the definition of different types of instructions:

- Privileged instruction: If run in the wrong context will generate a trap. Can't be executed by the guest OS
- Sensitive instruction: An instruction leaking information about the physical state of the CPU

  all sensitive instructions must be privileged instructions

3.2.2 Traps

A trap is when an instruction in user mode must be handled in kernel mode, by the hardware exception handler vector.
They occur with exceptions, system calls or hardware interrupts.

This allowed to develop the Trap&Emulate paradigm: the guest OS executes a privileged instruction, which launches a trap and is intercepted by the VMM, who emulates the privileged instruction if legitimate.
So, if the trap is caused by an application, it is passed to the guest OS. If it's caused by the guest OS, the VM state must be adjusted. All traps must be handled by the VMM

With the different cases:

- System call

  The CPU will trap the system call and send it to the interrupt handler vector, which is processed by the VMM who'll return it to the guest OS

- Privileged istruction

  They will be trapped to the VMM for emulation and then jumped back to the guest OS

- HW Interrupt

  The CPU traps to the interrupt handler of the VMM which jumps it to the guest OS interrupt handler

3.2.3 Dynamic Binary Translation (DBT)

x86 is difficult to virtualize because every privileged instruction is more timeconsuming and can lead to incorrect emulation of behaviors. The major solutions are changing the OS with paravirtualization or the dinamic detection of sensitive instructions, but the most common is hardware supported virtualization, which makes all sensitive instruction priviledged.

DBT is the fully virtualized approach which translates non virtualizable ISA at run-time