Cloud Computing Introduction and Foundations Concepts

Technology foundation concepts

References:

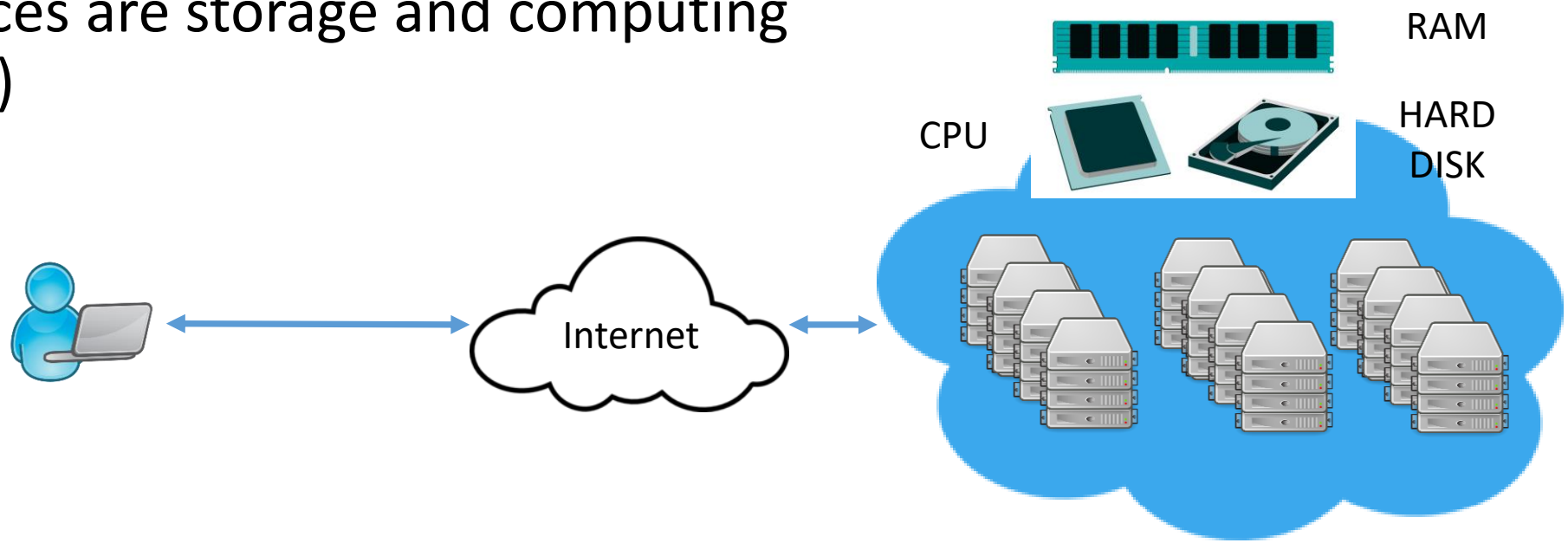
- Material provided by instructor

What is Cloud Computing?

- Cloud computing is now a ***buzzword***
- Historically the term ‘Cloud’ has been used as an abstraction of the network in system diagrams
- Today it is widely used to refer different concepts (sometimes even unproperly):
  - Cloud services
  - Cloud infrastructure
- In general, Cloud Computing refers to both the applications delivered as *services* over the internet, the *infrastructure* (hardware and software) in the datacenter that provides those services

Cloud Definition

- Cloud Computing original definition: ***cloud computing is a distinct IT environment designed for the purpose of remotely providing scalable and measured IT resources that are accessible via the Internet***
- Such resources are storage and computing (RAM + CPU)

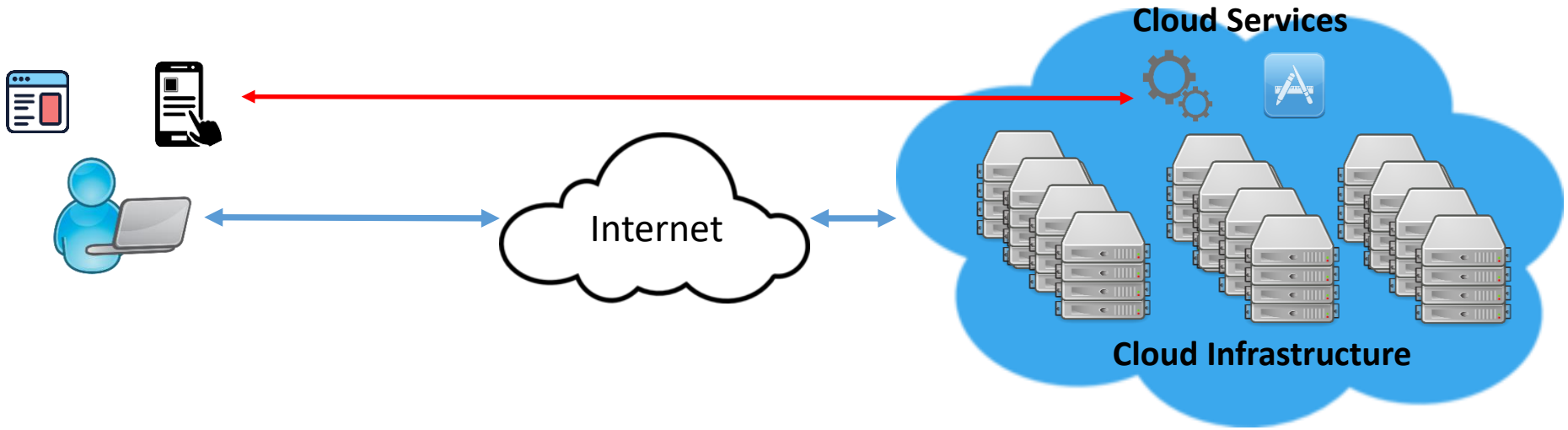


Cloud Services

- Such resources provided by the ***Cloud Infrastructure*** are used to implement ***Cloud Services***
- A Cloud Service can be itself an application or offer a pure service.
- When it implements an application, it usually exposes an interface that is accessed directly by users via **web pages (using a browser)**
- When it implements a service that is exploited by other applications (e.g. a local application running on a PC or a smartphone). In this case applications access the service through a set of programming interfaces API

Everything as a Service - XaaS

- Cloud services can implement a wide range of functions, everything can be provided ***as a service*** following this paradigm:
  Everything as a service – XaaS
  (https://en.wikipedia.org/wiki/As_a_service)



Cloud Services Examples

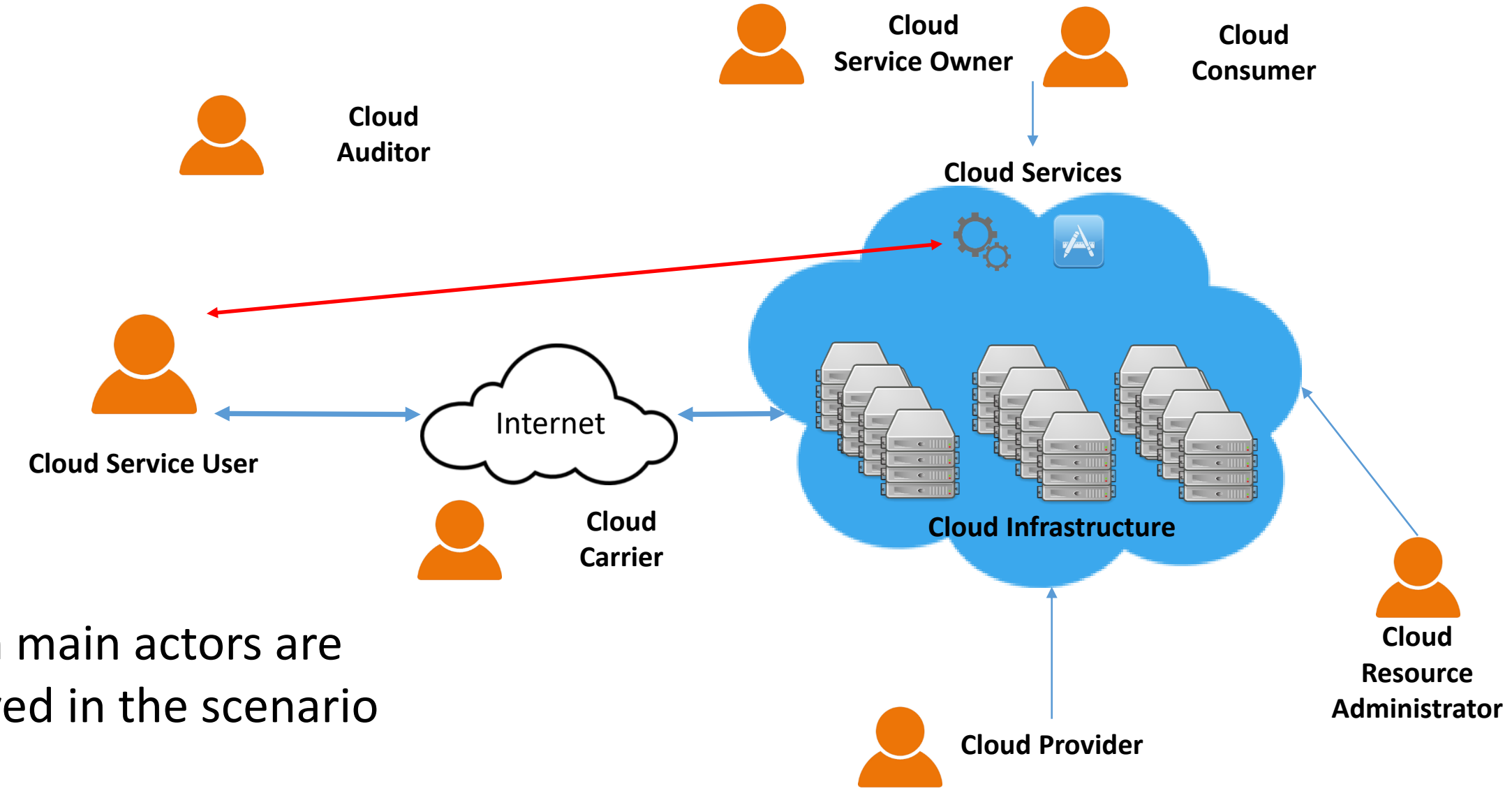
Cloud File Storage Service

Files are uploaded to the cloud, from there they can be downloaded on other devices or shared with other users. The service is accessed via web interface or via dedicated application

Cloud ERP Service

The ERP is deployed in the cloud. Employees access it through a web interface or a dedicated client (e.g. a smartphone application or a PC application) from anywhere

Roles



Seven main actors are involved in the scenario

Roles Definition

- **Cloud Provider**: The organization that provides cloud-based IT resources. It is usually responsible for creating and managing the Cloud Infrastructure. Normally IT resources are made available for lease by Cloud Consumers
- **Cloud Consumer**: An organization (or a human) that has a formal contract or arrangement with the Cloud Provider to use IT resources
- **Cloud Service Owner**: The person or organization that create a cloud service running on the resources provided by the cloud infrastructure

NOTE: a cloud consumer can own a cloud service, however, it does need to be the user of the service (it can sell it to other customers)

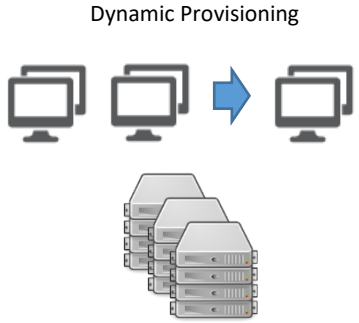
Roles Definition

- **Cloud Resource Administrator**: The person or the organization responsible for administering cloud-based IT resources. The administrator usually belongs to the cloud provider, but it can belong also to the cloud consumer
- **Cloud Auditor**: A third-party that conducts independent assessments of the cloud environment. Its role typically includes evaluating security and performance.
- **Cloud Carrier**: The party responsible for providing connectivity between users and the cloud provider
- **Cloud Service User**: The final user of a cloud service.

Cloud Model

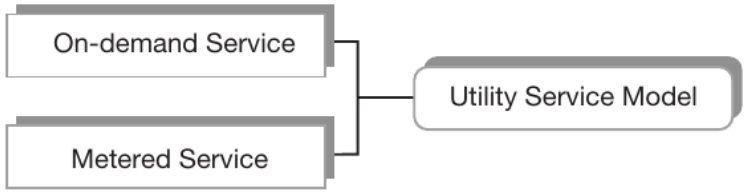
- Everything is based on the computing resources offered by the cloud provider
- As highlighted in its original definition, the cloud model for the delivery of IT resources is ***scalable and measured***
- This results in two main distinctive features:
  - It adopts a utility-based model

  - Resources can be provisioned dynamically



Utility-based Model

- IT resources are provided by **cloud providers**
- They adopt a **utility-based** approach: IT resources (computing, storage, connectivity) are delivered using a **‘pay-per-use’ price model**
- Usage of resources is metered, e.g. processing, storage or network
- Cloud consumers are billed as per their use, *they pay only what they get*
- A consumer for instance will be billed against his/her use of computing power (processor and memory), storage usage and network bandwidth consumption over time



Dynamically Provisioned Resources

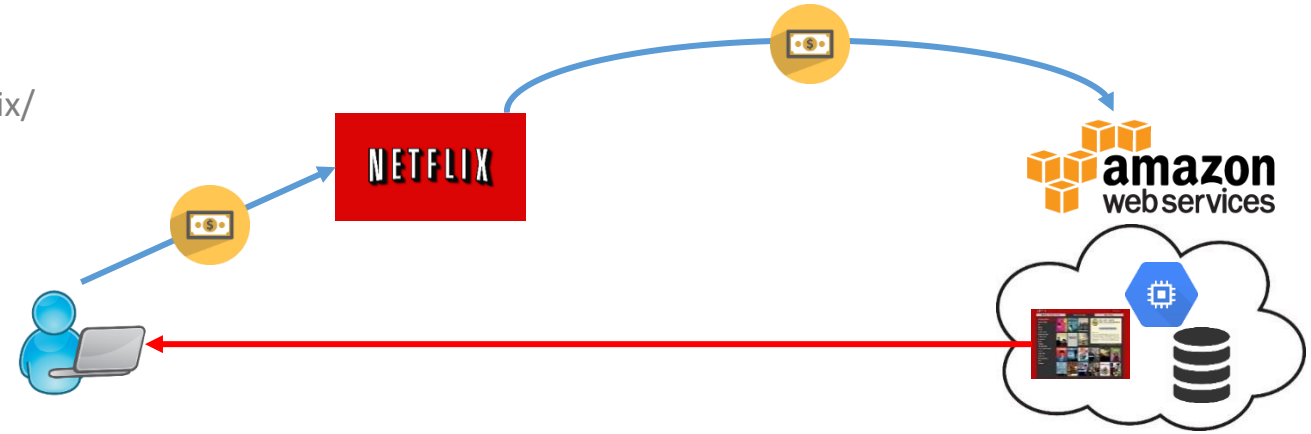
- The cloud infrastructure is designed to provision resources **dynamically** on demand
- Resources must be allocated/instantiated by the cloud provider in a short amount of time
- Resources are *managed with minimal effort* *without the need for human intervention*
- Resources can be provisioned not only as per request of a human but also under the orchestration of a software (the cloud infrastructure as we will see must provide an interface to provision resources to both humans and machines)

Business Model

- The cloud computing *business model* is simple:
- Cloud computing providers (e.g. Amazon, Rackspace, Google) sell IT resources (storage and computing in this case)
- Other companies (the Cloud consumer) buy such resources to create services and applications they can use internally or sell to their customers

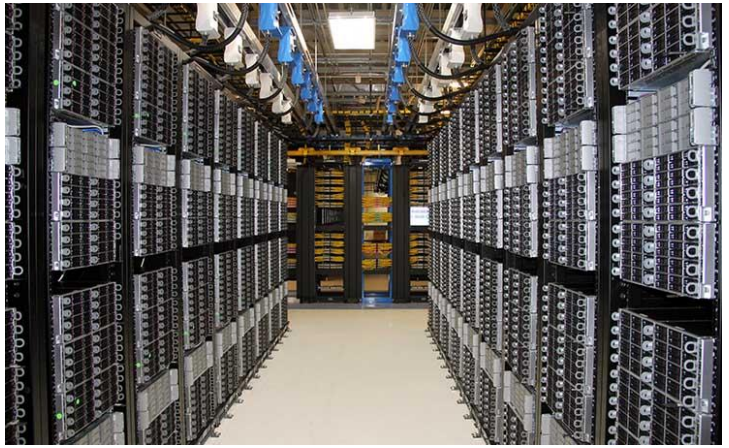
Example: **Netflix**
https://aws.amazon.com/solutions/case-studies/netflix/

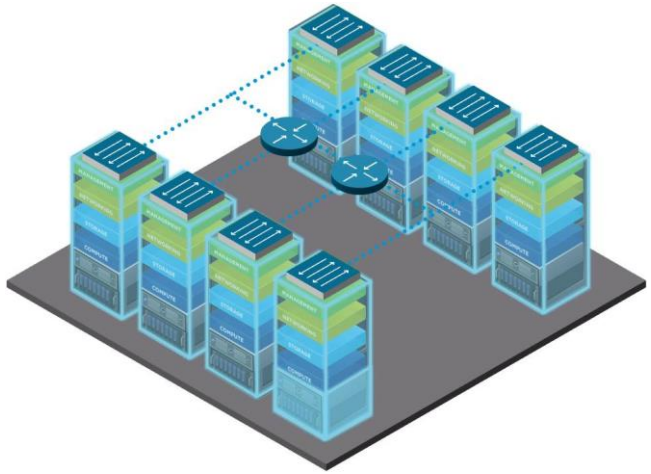
⚠ **We will see more examples soon**



Cloud Computing Infrastructure

- Cloud infrastructure is deployed in datacenters by the cloud provider
- Datacenter: a dedicated space to house servers and network equipment
- Such hardware provide IT resources such as:
  - Computing
  - Storage
  - Networking
- Eventually such resources are accessed by multiple cloud consumers at the same time via internet connection





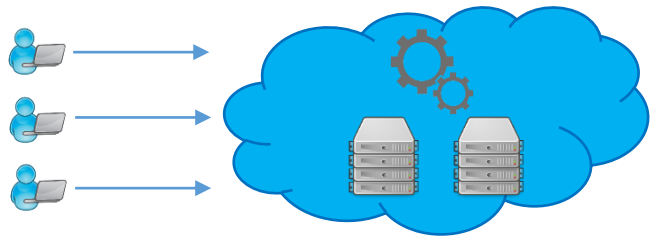
Multi-tenancy

- The traditional computing architectures are multi-users, thus allowing multiple users to access the resources of a system
- The system (e.g. a server) is installed and configured by an administrator
- When the number of users (each one with its own requirements) increases, it becomes impossible for the administrator to handle all the requirements
- Ideally, we would like to have a multi-tenant architecture in which multiple consumers are served at the same time
- Tenants are not user, they are the administrators of their system, they can control all its aspects
- The cloud computing infrastructure is multi-tenant, in order to be scalable we want cloud consumer to have the impression to be in control of the assigned resources

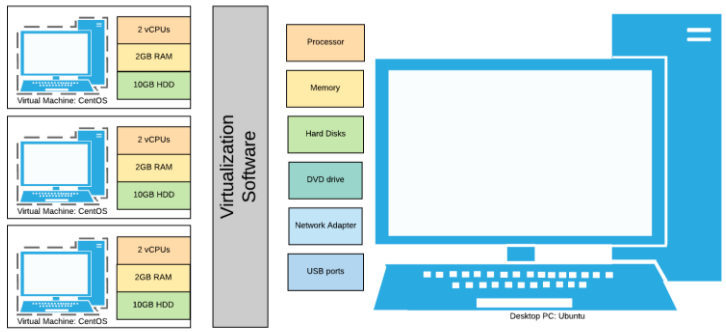
Virtualization

- Virtualization is a key enabler to create a multi-tenant infrastructure
- *Virtualization is a broad concept, it refers to every technique that allows to create a virtualized version of something*, i.e. IT resources such as a disk, a CPU, a Server, a program…
- By using virtualization multiple virtual copies of the real resources can be created
- Cloud consumers can have access not to the real versions but the virtualized ones, thus having the illusion that they are in complete control

Multi-tenant



Resource Virtualization

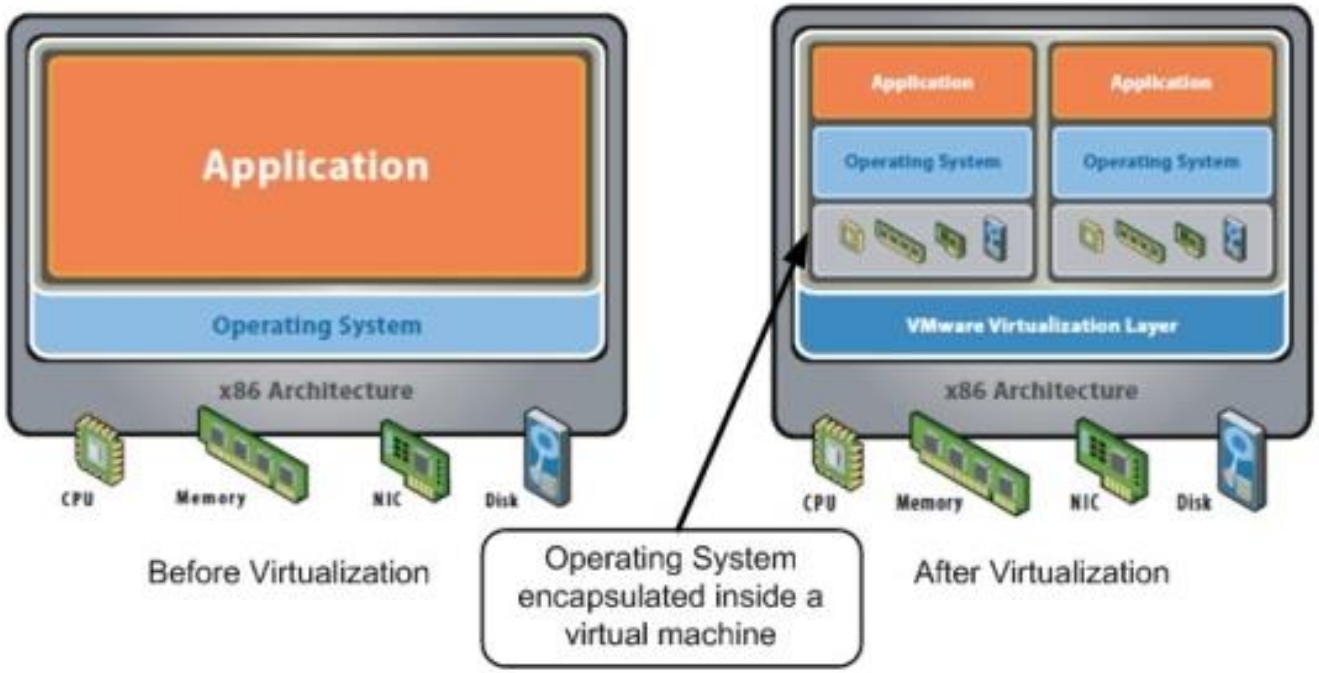


⚠ **We will see cover this in details later**

Hardware Virtualization

- The most popular virtualization technology used in cloud computing is *hardware virtualization* that allows to create a virtual representation of all hardware resources of a physical machine, a server creating a virtual server
- Hardware virtualization or *system level virtualization* provides an abstract execution environment on top of which a **full operating system** can be run
- It usually exploits different virtualization techniques to virtualize each component of the system
- Each virtual environment on which an operating system is run is called **Virtual Machine**

Hardware Virtualization

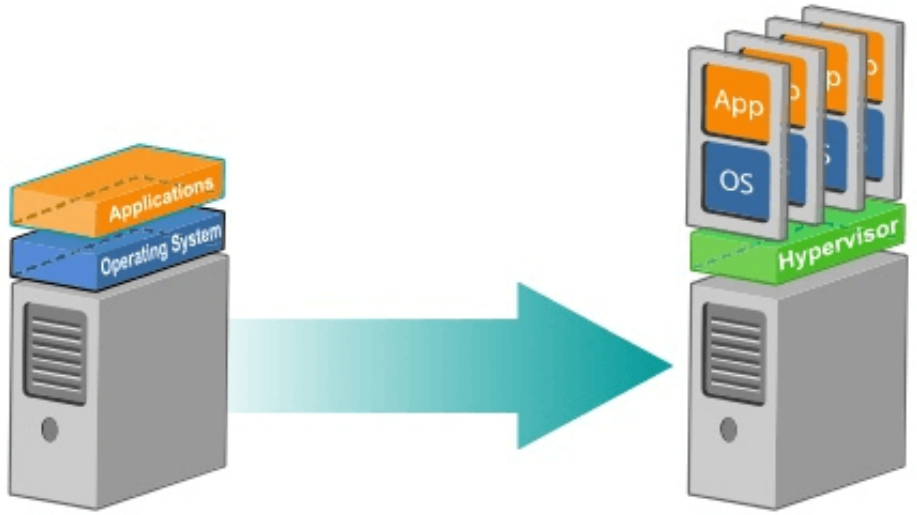


- Hardware Virtualization allows multiple Operating Systems to run on the same hardware.
- They access the virtualized version of each resource

Virtualization is often used in commercial PCs to run a different OS than the one the PC has (e.g. Linux on Windows)

Hypervisor

- The virtual hardware is managed and controlled by the **virtualization layer**
- Core element of such layer is the **Hypervisor** or Virtual Machine manager, a software that recreates the virtualized hardware environment in which the **guest operating system** runs
- The real machine on which the hypervisor runs is called **host**



Host

Virtualization Types

- Different virtualization types are available today, each one with its specific features to accommodate different application requirements:
  - Full virtualization
  - Para-virtualization
  - Operating system virtualization
- The main difference among them is the approach adopted, which results in different levels of abstraction and different overhead for its implementation, in terms of resources employed to virtualize the resources

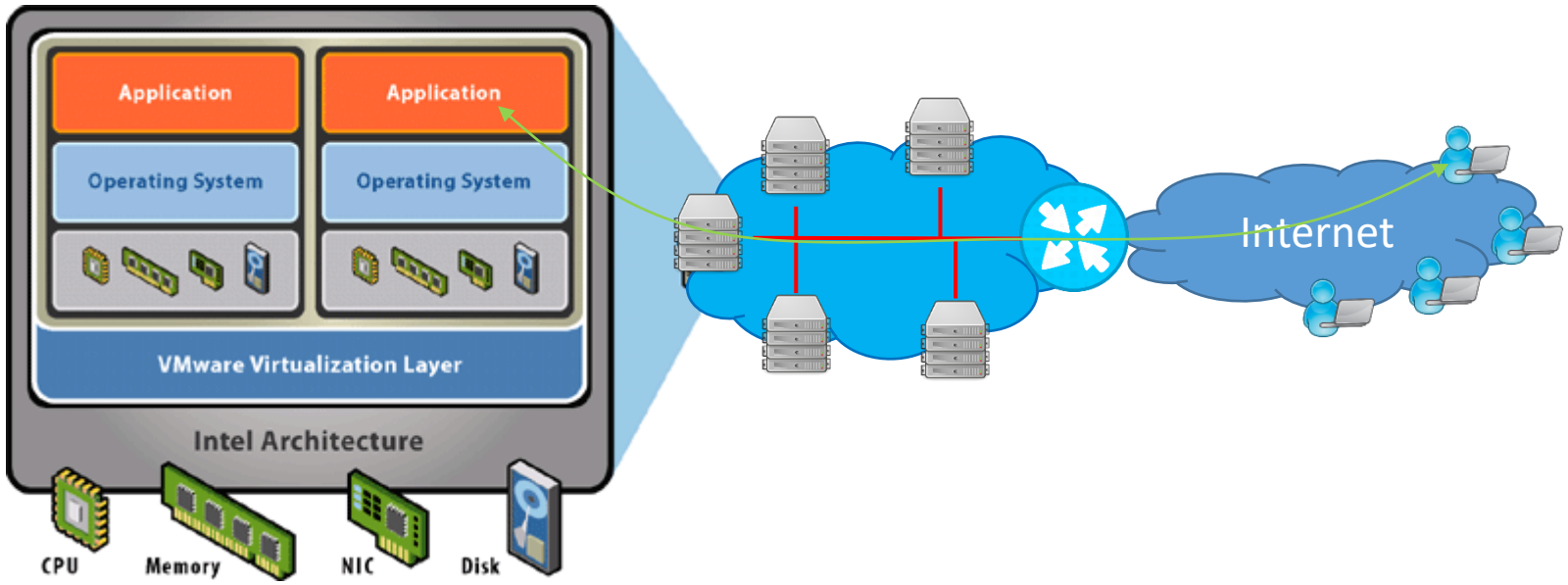
⚠ **We will see cover this in details later**

Cloud Virtualized Infrastructure

- In a datacenter, multiple servers (bare metal hardware) are installed and interconnected
- Each server runs a hypervisor for the execution of VMs
- VMs can be created dynamically as cloud consumers require new resources to host new cloud services, e.g. a web server hosting a site or a service collecting and analyzing data
- All physical servers are interconnected though a high-speed local connection (LANs)

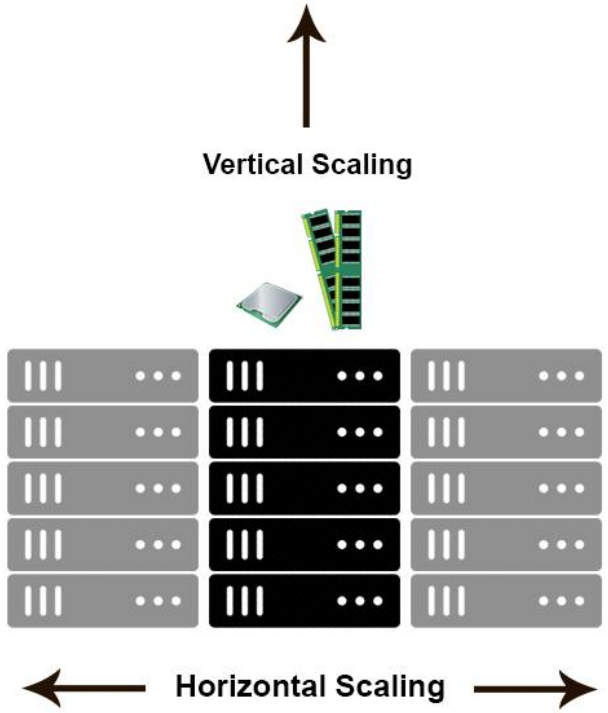
Cloud Virtualized Infrastructure

- Through this network infrastructure each VM can communicate with external hosts (to expose its cloud services) or can communicate with other VMs (to implement complex services, e.g. a web service that contacts a database)



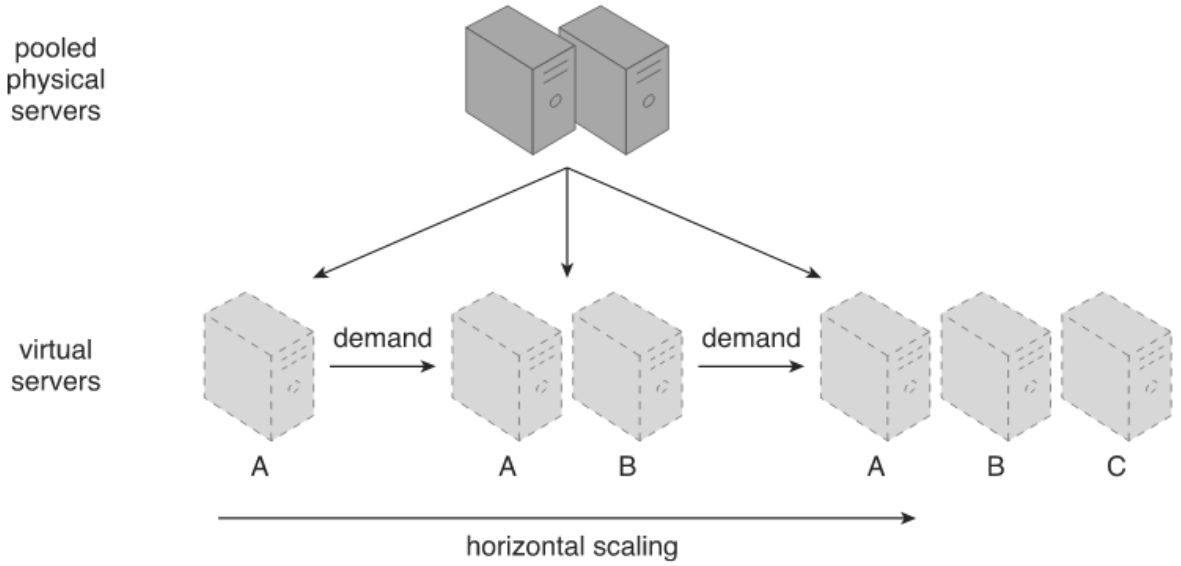
Dynamic provisioning

- Hypervisor allows to manage VMs and their configuration dynamically
- Cloud providers can instantiate VMs to cloud consumers **on demand**
- The set of resources allocated to each cloud consumer can scale dynamically, to handle changing conditions/loads
- Two scaling mechanisms are possible depending on the specific strategy:
  - Vertical scaling
  - Horizontal scaling



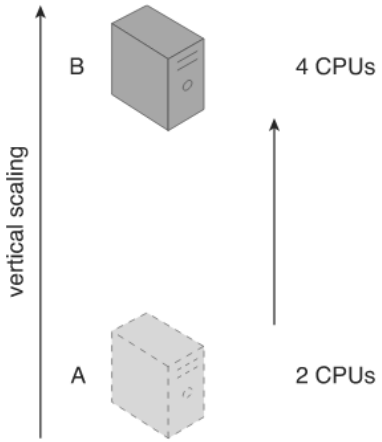
Horizontal Scaling

- Allocating or releasing the same type of IT resources, e.g. the same type of Virtual Servers
- Scaling out -> new resources are allocated
- Scaling down -> resources are deallocated
- For instance, new VMs are allocated running the same cloud service to handle more traffic



Vertical Scaling

- When an existing IT resource is replaced by another with higher or lower capacity, or the configuration of the same resource is modified to have higher or lower capacity
- Scaling up -> capacity is increased
- Scaling down -> capacity is decreased
- For instance, a VM is replaced/reconfigured to have 4 CPUs instead of 2 CPUs



Scaling Comparison

- Horizontal scaling is the most popular scaling mechanism in cloud computing
- Vertical scaling is less common in cloud environments due to downtime required while the replacement/reconfiguration is taking place

Horizontal Scaling	Vertical Scaling
less expensive (through commodity hardware components)	more expensive (specialized servers)
IT resources instantly available	IT resources normally instantly available
resource replication and automated scaling	additional setup is normally needed
additional IT resources needed	no additional IT resources needed
not limited by hardware capacity	limited by maximum hardware capacity

Metering

- Virtualization allows also the actual metering of computing resources
- Access and usage of virtualized resources can be easily monitored and measured
- In traditional computing, only a basic set of metering functions were available, they were not adequate for measuring the actual usage for metering
- Usage of different resources like, processing, storage or network bandwidth can be performed
- Cloud consumers are billed per their use proportionally