TCO (Total cost of ownership) = capEx (capital immobilisation) + opEx (pay for usage)

Essential characteristics cloud NIST :

On-demand self-service. A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

Broad network access. Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations). Resource pooling. 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. 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.

Rapid elasticity. Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

Measured service. Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

SaaS (Software) : cloud d'applications office 365, Google drive applications

PaaS (Plateform) : cloud lib et env, exemple kubernetes, gitlab, docker.

IaaS (Infra..) : cloud d'infrastructure (OS), AWS, openstack

(Hors NIST) FaaS (Function) : cloud de fonctions, AWS Lambda, Azure functions

Private cloud. The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

Community cloud. The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.

Public cloud. The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.

Hybrid cloud. The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

On-Demand Instances : Pay-by-the hour, Start and stop as you wish

Reserved Instances : Pay a yearly upfront fee and receive a discount on the hourly charge, Start and stop as you wish.

Spot Instances : Bid for unused EC2 capacity, Mention your Spot Price and if the market rate is less than your Bid, you get your instance, Instance automatically terminates if your Spot Price becomes less than the current market rate.

Chaque région est totalement indépendante. Chaque zone de disponibilité est isolée, mais les zones de disponibilité d'une région sont connectées par des liaisons à faible latence. Le diagramme suivant illustre la relation entre les régions et les zones de disponibilité.

Cluster — Set of machines where pods are deployed, managed and scaled.

- Nodes are connected via a “flat" network.
- Typical cluster sizes range from 1-200 nodes.

Pod — A pod consists of one or more containers that are guaranteed to be co-located on the same machine.

- Share storage volumes and a networking stack.
- A pod is the basic unit of scheduling.

Controller — A controller is a reconciliation loop that drives actual cluster state toward the desired cluster state.

- Replication Controller — Handles replication and scaling by running a specified number of copies of a pod across the cluster.

Service — Set of pods that work together, such as one tier of a multi-tier application. Kubernetes provides:

- Service discovery
- Request routing by assigning a stable IP address and DNS name
- Load balancing

Label — The user can assign key-value pairs (called labels) to any API object in the system (e.g., pods, nodes).

- Label selector — A query against a label that returns matching objects

Deployments add features and functionality compared to Replication Controllers to enable updating the deployed software without interrupting the service.

- Rolling updates
- Rollbacks