Elasticity and Scalability in Cloud Computing

Elasticity -

Elasticity refers to the ability of a cloud to automatically expand or compress the infrastructural resources on a sudden up and down in the requirement so that the workload can be managed efficiently.

This elasticity helps to minimize infrastructural costs and is helpful to address only those scenarios where the resource requirements fluctuate up and down suddenly for a specific time interval.

The main **purpose** of **Elasticity** is to **handle temporary, sudden** workload changes.

Scalability –

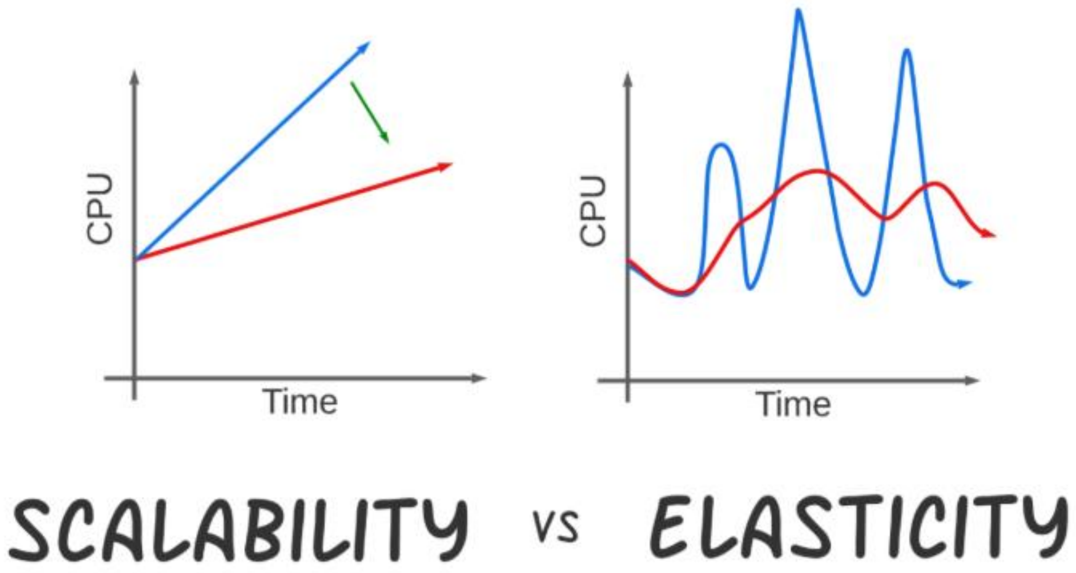
Cloud scalability is used to handle the growing workload where good performance is also needed to work efficiently with software or applications.

Scalability is the ability of a cloud system to **grow or shrink its resources** to handle **increasing or decreasing workloads over time**, based on **business or user growth**.

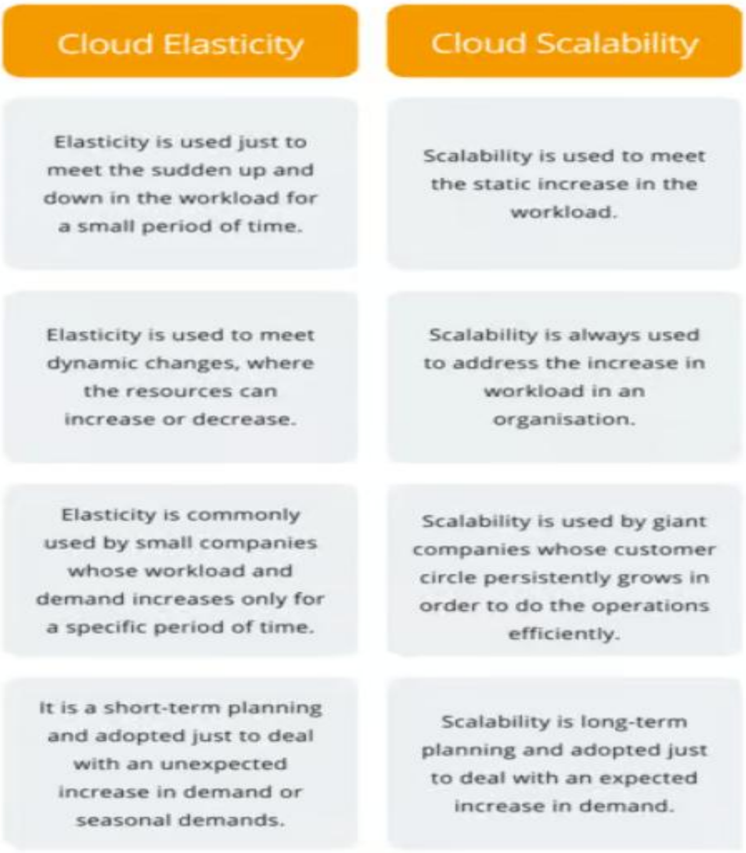
The main **purpose** of **Scalability** is to **handle steady, long-term growth**. Scalability is used to meet the static increase in the workload.

There are 3 types of Scalability :

1. Vertical Scalability (Scale-up): Increase power of existing resources (machines).
2. Horizontal Scalability (Scale-out): Resources are added in a horizontal row to handle load.
3. Diagonal Scalability: Combines both Vertical and Horizontal Scaling.



Though Scalability and Elasticity may seem similar, there is difference between both.



Cloud Elasticity	Cloud Scalability
Elasticity is used just to meet the sudden up and down in the workload for a small period of time.	Scalability is used to meet the static increase in the workload.
Elasticity is used to meet dynamic changes, where the resources can increase or decrease.	Scalability is always used to address the increase in workload in an organisation.
Elasticity is commonly used by small companies whose workload and demand increases only for a specific period of time.	Scalability is used by giant companies whose customer circle persistently grows in order to do the operations efficiently.
It is a short-term planning and adopted just to deal with an unexpected increase in demand or seasonal demands.	Scalability is long-term planning and adopted just to deal with an expected increase in demand.