1.5 ELASTICITY IN CLOUD COMPUTING

- Elasticity is defined as the ability of a system to add and remove resources (such as CPU cores, memory, VM and container instances) to adapt to the load variation in real time.
- Elasticity is a dynamic property for cloud computing.
- Elasticity is the degree to which a system is able to adapt to workload changes by provisioning and deprovisioning resources in an autonomic manner, such that at each point in time the available resources match the current demand as closely as possible.

Elasticity = Scalability + Automation + Optimization

- Elasticity is built on top of scalability.
- It can be considered as an automation of the concept of scalability and aims to optimize at best and as quickly as possible the resources at a given time.
- Another term associated with elasticity is the efficiency, which characterizes how cloud resource can be efficiently utilized as it scales up or down.
- It is the amount of resources consumed for processing a given amount of work, the lower this amount is, the higher the efficiency of a system.
- Elasticity also introduces a new important factor, which is the speed.
- Rapid provisioning and deprovisioning are key to maintaining an acceptable performance in the context of cloud computing
- Quality of service is subjected to a service level agreement

Classification

Elasticity solutions can be arranged in different classes based on

- Scope
- Policy
- Purpose
- Method

a.Scope

- Elasticity can be implemented on any of the cloud layers.
- Most commonly, elasticity is achieved on the IaaS level, where the resources to be

provisioned are virtual machine instances.

- Other infrastructure services can also be scaled
- On the PaaS level, elasticity consists in scaling containers or databases for instance.
- Finally, both PaaS and IaaS elasticity can be used to implement elastic applications, be it for private use or in order to be provided as a SaaS
- The elasticity actions can be applied either at the infrastructure or application/platform level.
- The elasticity actions perform the decisions made by the elasticity strategy or management system to scale the resources.
- Google App Engine and Azure elastic pool are examples of elastic Platform as a Service (PaaS).
- Elasticity actions can be performed at the infrastructure level where the elasticity controller monitors the system and takes decisions.
- The cloud infrastructures are based on the virtualization technology, which can be VMs or containers.
- In the embedded elasticity, elastic applications are able to adjust their own resources according to runtime requirements or due to changes in the execution flow.
- There must be a knowledge of the source code of the applications.
- Application Map: The elasticity controller must have a complete map of the application components and instances.
- Code embedded: The elasticity controller is embedded in the application source code.
- The elasticity actions are performed by the application itself.
- While moving the elasticity controller to the application source code eliminates the use of monitoring systems
- There must be a specialized controller for each application.

b.Policy

- Elastic solutions can be either manual or automatic.
- A manual elastic solution would provide their users with tools to monitor their systems and add or remove resources but leaves the scaling decision to them.

Automatic mode: All the actions are done automatically, and this could be classified into reactive and proactive modes.

Elastic solutions can be either reactive or predictive

Reactive mode: The elasticity actions are triggered based on certain thresholds or rules, the system

reacts to the load (workload or resource utilization) and triggers actions to adapt changes accordingly.

- An elastic solution is reactive when it scales a posteriori, based on a monitored change in the system.
- These are generally implemented by a set of Event-Condition-Action rules.

Proactive mode: This approach implements forecasting techniques, anticipates the future needs and triggers actions based on this anticipation.

- A predictive or proactive elasticity solution uses its knowledge of either recent history or load patterns inferred from longer periods of time in order to predict the upcoming load of the system and scale according to it.

c.Purpose

- An elastic solution can have many purposes.
- The first one to come to mind is naturally performance, in which case the focus should be put on their speed.
- Another purpose for elasticity can also be energy efficiency, where using the minimum amount of resources is the dominating factor.
- Other solutions intend to reduce the cost by multiplexing either resource providers or elasticity methods
- Elasticity has different purposes such as improving performance, increasing resource capacity, saving energy, reducing cost and ensuring availability.
- Once we look to the elasticity objectives, there are different perspectives.
- Cloud IaaS providers try to maximize the profit by minimizing the resources while offering a good Quality of Service (QoS),
- PaaS providers seek to minimize the cost they pay to the

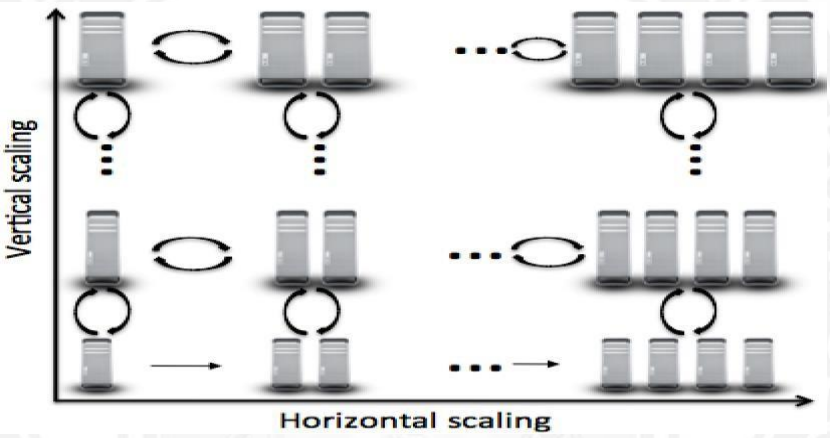
Cloud.

- The customers (end-users) search to increase their Quality of Experience (QoE) and to minimize their payments.
- QoE is the degree of delight or annoyance of the user of an application or service

d.Method

- Vertical elasticity, changes the amount of resources linked to existing instances on-the-fly.
- This can be done in two manners.

- The first method consists in explicitly redimensioning a virtual machine instance, i.e., changing the quota of physical resources allocated to it.
- This is however poorly supported by common operating systems as they fail to take into account changes in CPU or memory without rebooting, thus resulting in service interruption.
- The second vertical scaling method involves VM migration: moving a virtual machine instance to another physical machine with a different overall load changes its available resources



- Horizontal scaling is the process of adding/removing instances, which may be located at different locations.
- Load balancers are used to distribute the load among the different instances.
- Vertical scaling **is** the process of modifying resources (CPU, memory, storage or both) size for an instance at run time.
- It gives more flexibility for the cloud systems to cope with the varying workloads

Migration

- Migration can be also considered as a needed action to further allow the vertical scaling when there is no enough resources on the host machine.
- It is also used for other purposes such as migrating a VM to a less loaded physical machine just to guarantee its performance.
- Several types of migration are deployed such as live migration and no-live migration.
- Live migration has two main approaches
    - post-copy
    - pre-copy
- Post-copy migration suspends the migrating VM, copies minimal processor state to the

target host, resumes the VM and then begins fetching memory pages from the source.

- In pre-copy approach, the memory pages are copied while the VM is running on the source.
- If some pages are changed (called dirty pages) during the memory copy process, they will be recopied until the number of recopied pages is greater than dirty pages, or the source VM will be stopped.
- The remaining dirty pages will be copied to the destination VM.

Architecture

- The architecture of the elasticity management solutions can be either centralized or decentralized.
- Centralized architecture has only one elasticity controller, i.e., the auto scaling system that provisions and deprovisions resources.
- In decentralized solutions, the architecture is composed of many elasticity controllers or application managers, which are responsible for provisioning resources for differentcloud-hosted platforms

Provider

- Elastic solutions can be applied to a single or multiple cloud providers.
- A single cloud provider can be either public or private with one or multiple regions or datacenters.
- Multiple clouds in this context means more than one cloud provider.
- It includes hybrid clouds that can be private or public, in addition to the federated clouds and cloud bursting.
- Most of the elasticity solutions support only a single cloud provider.