Cloud Elasticity

Cloud elasticity is commonly used to refer to the degree to which public cloud providers can adapt dynamically to grow or shrink in response to changing resource demands.

Table of contents

What is cloud elasticity?

Cloud elasticity is one of the fundamental properties of public cloud computing services. The more elastic cloud services are, the quicker they are able to expand or contract to varying resource demands – ideally without impacting end-user performance. The goal of cloud elasticity is to automatically solve a common age-old system design problem: the trade-off between forward capacity planning and cost-efficiency.

Traditionally, when designing a system, engineers and architects would need to plan for and provision sufficient computing capacity in order to handle the maximum possible peaks in demand. For a retailer or bank, for example, this could be the annual Black Friday sales when the number of users visiting a website and making purchases is likely to be at their absolute peak.

Of course, the problem with this approach is that Black Friday occurs just once a year, and there are 364 other days in the year where this level of capacity may not be required. Having this be the existing infrastructure and relying on resource over-provisioning isn't cost-effective and often means that companies have provisioned computing resources far in excess of those required to serve the number of users at any given time.

How cloud elasticity helps with a system's ability to quickly expand

As mentioned, there is sometimes a demand for more resources, but oftentimes, the number of resources allocated can be much lower. Cloud elasticity solves this problem by allowing users to dynamically adapt the number of cloud resources – for example, the number of virtual machines – provisioned at any given time.

This means that companies relying on cloud computing solutions don't have to worry about anticipating rapid and unpredictable changes or unexpected demand spikes; cloud elasticity ensures that changing demands will be matched by makeshift resource allocation.

Typically, this will be controlled by system monitoring tools – tracking the utilization of a target resource such as CPU load or memory consumption, and matching the amount of resources deployed in order to keep utilization at a performant and cost-effective level. In this way, available resources can be conserved for peak usage or a traffic surge, removing resources and adding resources when it makes sense.

With cloud elasticity, users avoid paying for unused capacity or idle resources while maintaining the ability to scale up and respond to peaks in demand for their systems.

In the context of public cloud environments, users are able to purchase capacity on-demand, and on a pay-as-you-go basis. This means that during peaks in demand, such as Black Friday, when system monitoring detects increased utilization above a usual baseline, it can respond by purchasing additional virtual machines in order to handle these spikes in traffic. As the traffic then falls away, these additional virtual machines can be automatically shut down. This feature is often referred to as auto-scaling.

What is cloud scalability?

Cloud scalability is an important enabler of cloud elasticity – it's the ability to increase the capacity of a given system without impacting performance. Usually, cloud scalability is referred to across two distinct dimensions, vertical scalability or horizontal scalability:

- **Vertical scalability** or *scaling up*, refers to adding more compute power to an existing system or virtual machine. For example, if we increased the size of our MongoDB instance from two cores to four cores, we would have scaled it *vertically*
- **Horizontal scalability** or *scaling down* refers to adding additional compute *instances*, often of the same size, to the system. For example, if we added additional replicas to our existing MongoDB cluster, we would have scaled it *horizontally*

With most modern public clouds, you can use managed cloud services, such as MongoDB Atlas, to make it easily scale a cloud-based application both horizontally and vertically.

It is worth noting, however, that there is an inherent limit to systems that rely on vertical scaling – since there is usually a maximum server size available on all public clouds. The same is usually not true for horizontal scaling – where it's possible to scale solutions out from a single server to tens of thousands of servers.

Cloud elasticity vs. cloud scalability

Scalability and elasticity have similarities, but important distinctions exist. Cloud scalability is a feature of cloud computing, particularly in the context of public clouds, that enables them to be elastic. If a cloud resource is scalable, then it enables stable system growth without impacting performance.

This could mean adding additional virtual machines to an application, increasing the size of an existing database server, or increasing the number of available compute functions in a system with a serverless architecture. All of these features enable users to increase the number of resources available to a system in order to meet increasing demand.

Cloud elasticity is a feature that enables a system to *scale automatically* in response to demand for resources. An important concept of cloud elasticity is the ability of a system to be able to rapidly add more resources in order to meet peaks in demand, but also enables allocating fewer resources when they are no longer required in order to reduce cloud costs and be cost-effective.

Cloud elasticity is usually enabled by closely integrated system monitoring tools that are able to interact with cloud APIs in real-time to both request new resources, as well as retire unused ones.

Cloud elasticity is also enabled by a number of other recent improvements to the way applications are designed for the cloud, such as the increasing popularity of NoSQL

databases, stateless computing, and a shift towards microservice architectures.

Elasticity in cloud computing

The ability to develop new applications in an elastic fashion is one of the key drivers behind the adoption of a public cloud: the ability for systems to grow or shrink automatically in response to demand results in solutions that are both highly performant but also extremely cost-effective.

In response to this, cloud platforms are investing significant effort in new products which make it easy for users to take advantage of the pay-as-you-go nature of their engagement model.

Historically, cloud elasticity referred only to the ability to auto-scale a fleet of virtual machines. However, elasticity in cloud computing has grown. A broad range of products offer capabilities which allow them to leverage cloud elasticity dynamically, and in real-time respond to demand:

- Database services like MongoDB Atlas, which offer both automatic vertical (cluster auto-scaling) and horizontal (cluster auto-sharding) scaling options
- Object storage services
- Managed machine learning services – for example, performing image recognition
- Message bus and queue style systems

Services such as these are available from your preferred major cloud provider, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP).

Related Resources

- What is a cloud database?
- What is multi-cloud?
- What are cloud-native applications?

- Public cloud vs. private cloud vs. hybrid cloud