**Title: Cloud Scalability**

**Sathish Reddy Thadamla**

**Computer Science Department**

**University of Central Missouri**

**Lee’s Summit, Missouri**

[**SXT58180@ucmo.edu**](mailto:SXT58180@ucmo.edu)

|  |
| --- |
| **Scalability:** |
| Scalability is a term used to describe how the application will handle increased loads of traffic volume. [1] |
| Cloud scalability came as a boon to enterprises and organizations who were constantly worried about data storage and protection resources and costs. Before the cloud, data storage and management required a lot of investments and infrastructure from companies. It is to such a scenario that cloud solutions came in with the signature advantages of cost efficiency and easy scalability. |
| One of the most influential characteristics of cloud computing is its scalability. A scalable cloud will scale to meet the industry's growing computing, networking, infrastructure, and software demands. |
| **Cloud Scalability:** |
| Cloud scalability in cloud computing refers to the ability to increase or decrease IT resources as needed to meet changing demand. It’s the ability of a system to accommodate larger or smaller loads.[2] |
| Businesses can scale up or down (vertically) and out or in (horizontally). For example, scaling up makes hardware stronger; scaling out adds additional nodes. More on these later. |
| **Why is Cloud Scalable?** |
| A scalable cloud architecture is made possible through virtualization. Unlike physical machines whose resources and performance are relatively set, virtual machines (VMs) are highly flexible and can be easily scaled up or down. They can be moved to a different server or hosted on multiple servers at once; workloads and applications can be shifted to larger VMs as needed. |
| Third-party cloud providers also have all the vast hardware and software resources already in place to allow for rapid scaling that an individual business could not achieve cost-effectively on its own. |
| **Types of Scalabilities** |
| There are mainly three types of cloud scalabilities   * Vertical Scalability * Horizontal Scalability * Diagonal Scalability |
| **Vertical Scalability:** |
| Vertical scaling is also known as scaling up (or down). It means resizing an existing resource with no change to your code. You’re simply running the same code on a higher- or lower-spec machine. [3] |
| When you scale vertically, you enlarge or diminish a resource to change the capacity of your existing infrastructure. That infrastructure becomes more or less powerful. For example, you could move a web application to a larger virtual machine or add more CPU to an existing server |
| Image title |
| **Horizontal Scalability:** |
| Horizontal scaling also goes by the name scaling out (or in) [3]. When your business scales horizontally, you add or remove instances of a resource or infrastructure. It involves breaking a sequential piece of logic into smaller pieces. That’s then executed in parallel across multiple resources or infrastructures. |
| With Amazon Web Services (AWS), an example of horizontal scaling is changing the number of nodes in a computing system. The size of each individual node remains the same. |
| Horizontal scaling is more labor-intensive than vertical scaling. It’s particularly important for organizations with high availability services requiring minimal downtime |
| High availability: The aim of a high availability system is being available 99.999% of the time, or as close to that as possible. Usually, that means configuring a failover system with the ability to handle the same workloads as the primary system. |
| Vertical Scaling and Horizontal Scaling in AWS | My Passion Behind Blogging |
| **Diagonal Scalability:** |
| Diagonal scaling refers to a combination of vertical and horizontal scaling. Your business grows vertically within existing infrastructure until it reaches a tipping point. At that stage, you add more resources to scale out horizontally. |
| How to Build Scalable Applications in AWS Cloud |
| **Difference between Vertical and Horizontal Scaling:** |
| [3] The main and fundamental difference between the two is horizontal Scaling Breaking a sequential piece of logic into smaller parts so that it can be performed in parallel through several machines. Whereas the logic does not need to adjust, vertical scaling is simpler. Instead, you're simply running the same code on more powerful machines. There are some of other differences are there. |
| |  |  |  | | --- | --- | --- | |  | **Horizontal Scaling (Scaling out)** | **Vertical Scaling (Scaling up)** | | **Databases** | In a database world, horizontal scaling is usually based on the partitioning of data (each node only contains part of the data). | In vertical scaling, the data lives on a single node and scaling is done through multi-core, e.g. spreading the load between the CPU and RAM resources of the machine. | | **Downtime** | In theory, adding more machines to the existing pool means you are not limited to the capacity of a single unit, making it possible to scale with less downtime. | Vertical scaling is limited to the capacity of one machine, scaling beyond that capacity can involve downtime and has an upper hard limit, i.e. the scale of the hardware on which you are currently running. | | **Concurrency** | Also described as distributed programming, as it involves distributing jobs across machines over the network. Several patterns associated with this model: Master/Worker\*, Tuple Spaces, Blackboard, MapReduce. | Actor model: concurrent programming on multi-core machines is often performed via multi-threading and in-process message passing | | **Message passing** | In distributed computing, the lack of a shared address space makes data sharing more complex. It also makes the process of sharing, passing or updating data more costly since you have to pass copies of the data. | In a multi-threaded scenario, you can assume the existence of a shared address space, so data sharing and message passing can be done by passing a reference. | |
|  |
| **Major Benefits of Cloud Scalability:** |
| The major cloud scalability benefits are driving cloud adoption for businesses large and small [4]: |
| **Cost-efficient:** |
| With zero upfront costs on costly equipment, scaling via the cloud saves your business money. With none of these initial outlays, there won’t be expensive upgrades either. There are savings in terms of IT staff, power, and cooling, too. |
| Waste and risk both minimize because you only pay cloud providers for what you use. What’s more, many applications run more cost-effectively in the cloud. |
| **Convenient:** |
| Often with just a few clicks, IT administrators can easily add more VMs that are available without delay and customized to the exact needs of an organization. That saves precious time for IT staff. Instead of spending hours and days setting up physical hardware, teams can focus on other tasks. |
| **Fast and Flexible:** |
| The beauty of cloud scalability is you can respond and adapt to shifting business requirements with speed and efficiency. For example, you can update storage and systems as and when you need to. As your business faces new challenges, cloud scalability offers you versatility and freedom. |
| You can reshape your infrastructure easily and even redesign your model. For instance, you can move from a private cloud model to a hybrid cloud or multi-cloud system if that suits your changing needs better |
| **Greater Power:** |
| Whatever the size of your business, cloud computing offers access to high-powered resources. Small or large, cloud computing scalability allows access to powerful software and data tools. Historically, infrastructure limitations prevented organizations from scaling quickly. Buying equipment and implementing processes took time and money |
| **Disaster Recovery:** |
| Scalable cloud applications allow you to reduce disaster recovery costs. You no longer have to foot the bill for the expensive building and maintenance costs associated with a secondary data center. |
|  |
| **Storage:** |
| Scalable cloud storage is crucial for a company that’s growing. Businesses can use cloud computing to scale their data storage to fit their needs. That means physical infrastructure and the associated costs become a thing of the past. |
| **When to use Cloud Scalability** |
| Scalable business models are nothing new: they allow businesses to meet changing needs. Cloud scalability applies to the IT side of your business. By making the most of cloud scalability, your business stays competitive. |
| Scalability drives migration to the cloud. When a business grows and needs more resources, cloud scalability enables a rapid response. It also offers a cost-effective solution. |
| **Conclusion:** |
| Having said that, it’s important to think about how your business should scale, to get the optimal experience. Is horizontal scaling optimal or should you scale vertically. Is auto-scaling right for organization. Whatever, cloud computing scalability is important for a company. |
| **References:** |
| [1] <https://www.vmware.com/topics/glossary/content/cloud-scalability> |
| [2] <https://www.netguru.com/blog/cloud-computing-scalability> |
| [3] <https://www.section.io/blog/scaling-horizontally-vs-vertically/> |
| [4] <https://vexxhost.com/blog/cloud-scalability/> |
| <https://airccse.org/journal/ijcses/papers/5414ijcses03.pdf> |
| <https://journalofcloudcomputing.springeropen.com/track/pdf/10.1186/s13677-019-0134-y.pdf> |
| <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.695.8518&rep=rep1&type=pdf> |