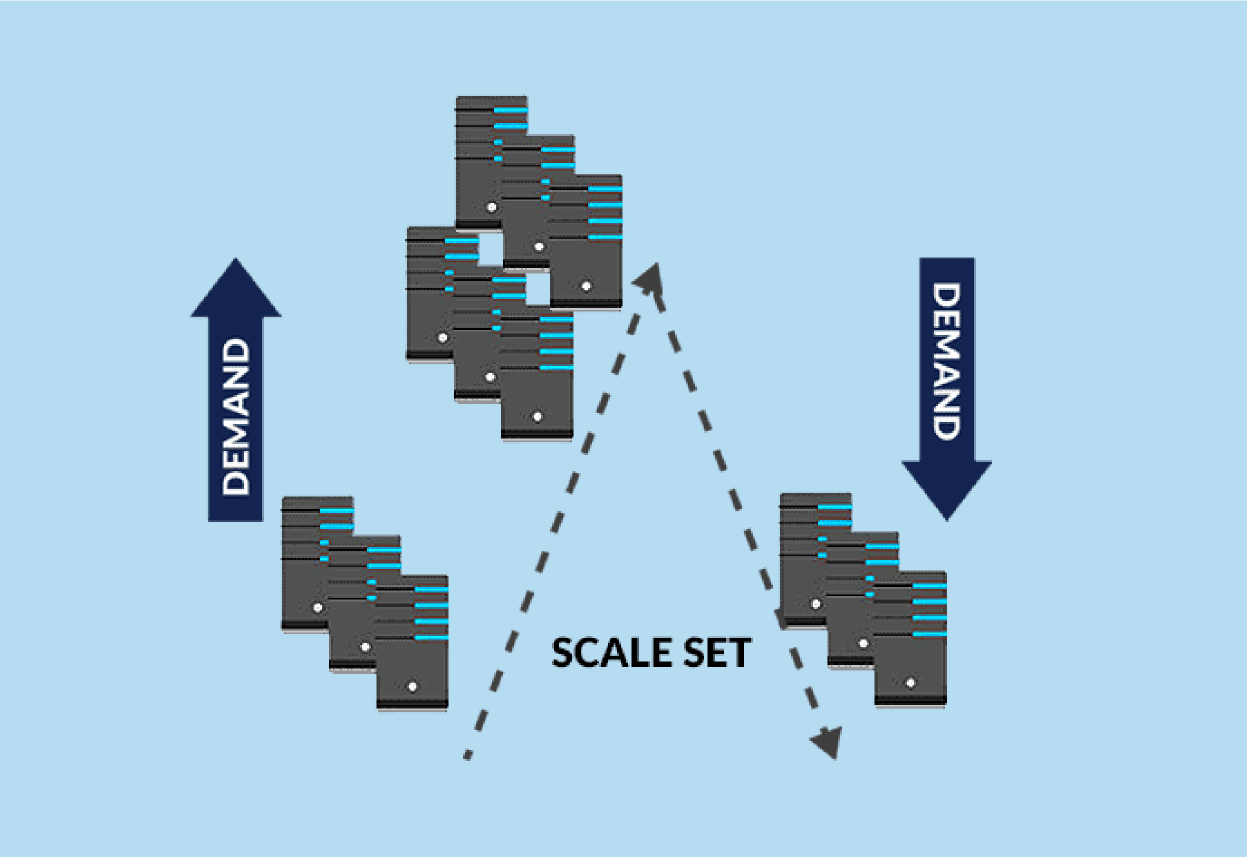
Virtual Machine Scale Set (VMSS) In Microsoft Azure

ne of the biggest benefits of cloud computing is that it can be *elastic*. Like a rubber band, the idea behind elastic computing is that you can stretch or shrink your cloud service usage to accommodate changes in workload.

Azure includes several services with elastic features. One of the newest, and least understood, is Azure Virtual Machine Scale Sets (VMSS).

What Are Virtual Machine Scale Sets (VMSS)?

Virtual Machine Scale Sets (VMSS), an interesting service offered by Microsoft Azure, helps to create and manage a set of identical, auto-scaling Virtual Machines (VMs). The number of VM instances can automatically increase or decrease based on scheduled conditions.



Auto-Scaling

Auto-scaling is a way to automatically scale up or down the number of compute resources that are being allocated to your application based on its needs at any given time.

The key point is that you can now design a scalable architecture that will automatically scale-up or scale-down to meet your needs over the lifetime of your setup regardless of how fast/slow or big/small your site grows over that time.

Here are the most popular ways of autoscaling:

- Horizontal Scaling
- Vertical Scaling

Vertical Scaling:

Vertical Scaling is an attempt to increase or decrease the capacity of a single machine, also called scaling up or down. Here the resources such as processing power, storage, memory, and more are added to an existing work unit.

It is done to increase the capacity of existing hardware or software by adding resources. It can enhance your server without manipulating your code. But it is limited by the fact that you can only get as big as the size of the server.

For example, you could move an application to a larger VM size.

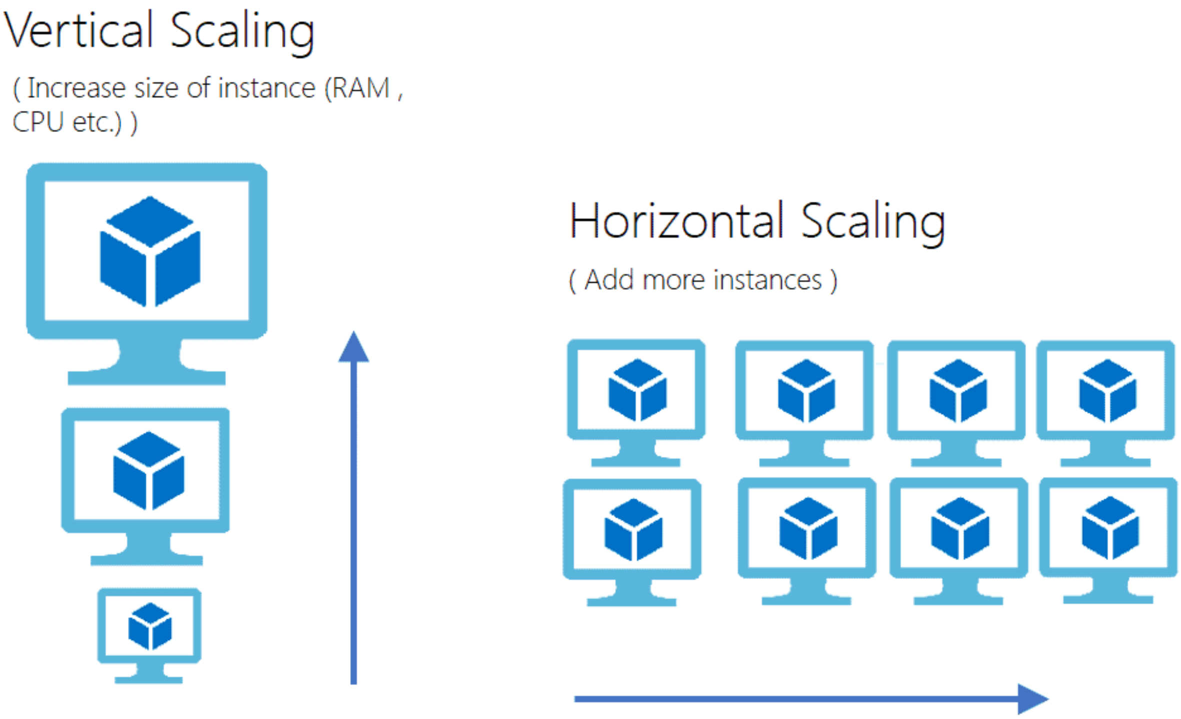
Also Read: Our blog post on az 104 Microsoft azure administrator Exam.

Horizontal Scaling:

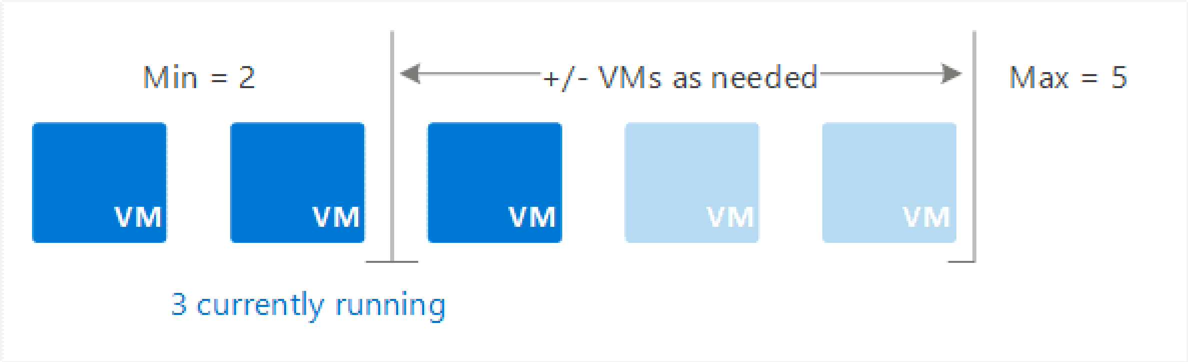
Horizontal Scaling is a must use technology – whenever a high availability of (server) services are required, also called scaling out and in, means adding or removing instances of a resource. **Scaling horizontally** involves adding more processing units or physical machines to your server or database.

It involves growing the number of nodes in the cluster, the application continues running without interruption as new resources are provisioned.

If demand drops, the additional resources can be shut down cleanly and deallocated.



In Autoscaling you have to specify a minimum and the maximum number of instances to run and add or remove VMs automatically based on a set of rules.



When rule conditions are met, one or more autoscale actions are triggered. You can add and remove VMs, or perform other actions.

Also checkout: Azure Availability zones to learn in detail on Availability Sets, Fault domains, Update domains, and Availability Zone.

Autoscale Settings

An **autoscale setting** is read by the autoscale engine to determine whether to scale up or down.

You can create custom autoscaling rules as needed for your situation. Rule types include:

- **Minimum Instance:** The minimum number of instances you want to deploy in your scale set.
- **Maximum Instances:** The maximum number of instances you want to deploy while scaling out. **(Note:** In Azure, you can have maximum, 1000 instances)
- **Metric-based** – It measures application load and add or remove VMs based on that load. For example, add instance in scale set when CPU usage is above 50%.
- **Time-based** – For example, trigger an instance every 8 am on Saturday in a given time zone.

Here are a **few points** which are important when we think about going with **Horizontal scaling** or **Vertical Scaling.**

- **Scaling up requires downtime**, in this case, you need to upgrade server's configuration like RAM, memory, CPU, etc. so while upgrading this configuration your server requires downtime. Once you are done with the update, the restart of the server is required.
- **Scaling up will Increase performance but not available** because its only one instance and it can go down anytime when it reaches out the scaling rules.
- **Scale-Out doesn't require downtime,** in Scale-out its creates new instances of server it doesn't touch the existing instance, so no downtime required.
- **Scale-Out, Increase performance and availability as well.** When autoscaling increase no. of instances and handle load/request using load balancer it increases the performance of the server and also increases availability as well.