Availability set and Scale set

**Scale set-**

It’s nothing but vmss. it’s a future to set a scale (range) of set of VMs.

Here we do configuration on scaling (min number of instances, max and default number of instances)

Also we mention the limit of use to do scale in and scale out.

Scale in – cpu or memory usage %

Number of instance - Number of instance need to increase

Use time – if usage is more then given % for time (5 minutes, 10 minutes or any) increate the count by givennumber of instances.

Scale down – same configuration as scale up tile period to check % and no of instance to scale down.

Here the instances are coming under auto scaling which is deploying on vmss so, by default those instances will have high availability. From azure end fault domain and update domain will be created and the instances will be distributed.

We no need to handle it manually.

NOTE- vmss will run multiple instances, any instance can come to run and can be deleted at any time.

So how to run application on that and give access to end user. Because if multiple vms are running and having different ips where do end user hit request. How do it balance load.

To build this solution azure load balancer comes to picture.

The load balancer backend pool target we can add vmss. Hence load balancer will automatically pick any new instance comes to place and also will remove the deleted instance.

We can see the instances on load balancer backend pool instances.

Here the load balancer will have health check configured as well.

Also in advance setting we can mention auto replace instance if any instance runs into issue.

How to configure application on vmss instances? Do we need to configure it on all instance separately???

* Hence the ans is no, because it’s not possible to configure on individual vms. Also the new vm will not get the application.
* On vmss we do configuration of application on vmss that will be applied on all vms. Also the port applied on vmss will be inherited to all of the vms.

Note- here in vmss each server will have their own ip, we can configure RDP rule and can connect to any of the instance.  
  
we do see the vmss is also having public ip, what the use of that ? why do we pay for that ip?

* Well the ans is the vmss ip will be added to load balancer or any of the frontend service to add that vmss to backendpool.

**Availability Set-**

In the availability set we do create a availability set by defining the number of fault domain and the number of update domain we need.

Then we add servers to that availability set and from azure end it will be distributed among fault domain and update domain.

What if fault domain?

* It saves the servers from physical failures. The fault domain arrangement arrange the servers into multiple racks in same datacenter.
* Each rack is having its own power supply and network connection. If one rack goes down, the servers in that fault domain will be down hence other domain servers will keep running.

Note- each fault domain indicated to different rack for the server.

Q -> here unlike vmss we are adding the servers manually. So, in each servers we need to configure the required application. We need to create Load balancer and add those instances to the load balancer.

What is update domain?

* Update domain arrange the servers from any rack, update domain does not depends on rack or any of the physical infrastructure. It’s just logical grouping of servers on those servers update will apply accordingly.
* Update domain is totally separate from fault domain.

What is the required of update domain? We are installing the updated on servers manually?

* Yes on vm we have full access on os hence os we need to update manually. During this planed update from Microsoft . azure will update the underlying infrastructure. Such as hosted vmware update, network or any.

Those updates to make the server more reliable and faster.

Do you know?

* During unplanned activity or any hardware failure time if the physical server goes down. Azure will migrate the server to a good physical server, it may have some temporary outage and the sever will be rebooted.
* It may have data loss on temporary disk.
* The temporary disk (D:/) given on server to store logs data. It’s recommended to do not store important data that disk. That disk is a small size and that’s on physical server which is hosting the vm. So if the vm change to any new physical server it will have new temporary disk and the data stored on old temporary disk will be lost.

what is An unplanned hardware maintenance on azure vm?

* Unplanned hardware maintenance on an Azure VM occurs when the Azure platform predicts that the hardware or any platform component associated with a physical machine is about to fail. To minimize the impact on the virtual machines (VMs) hosted on that hardware, Azure initiates an unplanned hardware maintenance event. During this event, Azure uses Live Migration technology to move the VMs from the failing hardware to a healthy physical machine. [This process typically only pauses the VM for a short time, preserving memory, open files, and network connections1](https://learn.microsoft.com/en-us/azure/virtual-machines/understand-vm-reboots).
* If Live Migration cannot be used, the VM may experience unexpected downtime, during which it is rebooted and possibly loses temporary drive data. [However, the attached OS and data disks are always preserved1](https://learn.microsoft.com/en-us/azure/virtual-machines/understand-vm-reboots).

unexpected downtime from azure can migrate the server to a healthy rack in same datacenter during the time server reboot is fine but what may cause that the server may lost the temporary disk?

* During an unexpected downtime, Azure can indeed migrate the server to a healthy rack within the same datacenter. However, the temporary disk (also known as the ephemeral disk) might be lost during this process. This loss can occur because the temporary disk is stored on the local physical server that the VM is running on. When the VM is moved to a different physical server, the data on the temporary disk does not move with it.
* The temporary disk is intended for short-term storage and is not backed up. It is typically used for operations like page files or temporary data that can be recreated if lost. For critical data, it’s recommended to use persistent storage options like Azure Managed Disks, which are not affected by such migrations.

suppose i am creating a scale set of vm which have 5 servers, how it will manage my servers on rack to maintain fault domain and update domain? for example consider that datacenter have 3 racks?

* When you create a scale set of VMs in Azure, the platform automatically manages the distribution of your VMs across fault domains and update domains to enhance availability and reliability.
* **Fault Domains**

Fault domains represent a group of VMs that share a common power source and network switch. In your scenario with 3 racks in the datacenter, each rack would typically represent a fault domain. Azure will distribute your 5 VMs across these fault domains to ensure that a failure in one rack (fault domain) does not affect all your VMs. For example, with 3 fault domains, Azure might place 2 VMs in the first fault domain, 2 in the second, and 1 in the third.

* **Update Domains**

Update domains are used to ensure that not all VMs are updated or rebooted at the same time during planned maintenance. Azure assigns each VM in your scale set to an update domain. With 5 VMs, Azure will distribute them across multiple update domains to minimize downtime. For instance, if you have 5 update domains, each VM will be placed in a different update domain.

**Example Distribution**

Given your setup with 3 racks (fault domains) and 5 VMs, the distribution might look like this:

* **Fault Domain 1 (Rack 1)**: VM1, VM2
* **Fault Domain 2 (Rack 2)**: VM3, VM4
* **Fault Domain 3 (Rack 3)**: VM5

For update domains, assuming 5 update domains:

* **Update Domain 1**: VM1
* **Update Domain 2**: VM2
* **Update Domain 3**: VM3
* **Update Domain 4**: VM4
* **Update Domain 5**: VM5

okay so update domain does not depends on rack? any numbers of vms on any rack can be in any group of fault domain?

* That’s correct! Update domains are independent of fault domains (racks). This means that VMs in any rack can belong to any update domain. The primary goal of update domains is to ensure that not all VMs are updated or rebooted simultaneously during planned maintenance, thereby minimizing downtime.