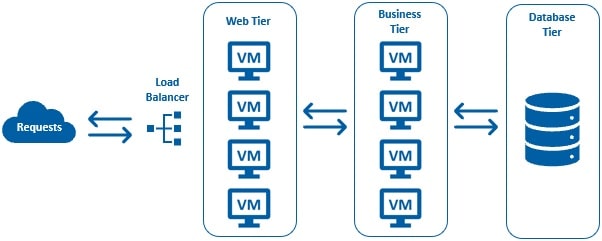
**Virtual machine scale sets**

## **What is a Virtual Machine Scale Set (VMSS)**

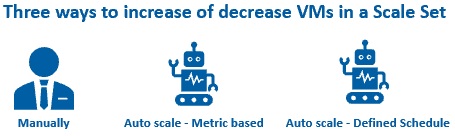
A Virtual Machine Scale Set lets us easily create and manage multiple virtual machines.

## **Why do we need a Virtual Machine Scale Set**

Let's understand this with a simple example. Consider a typical 3-tier application.



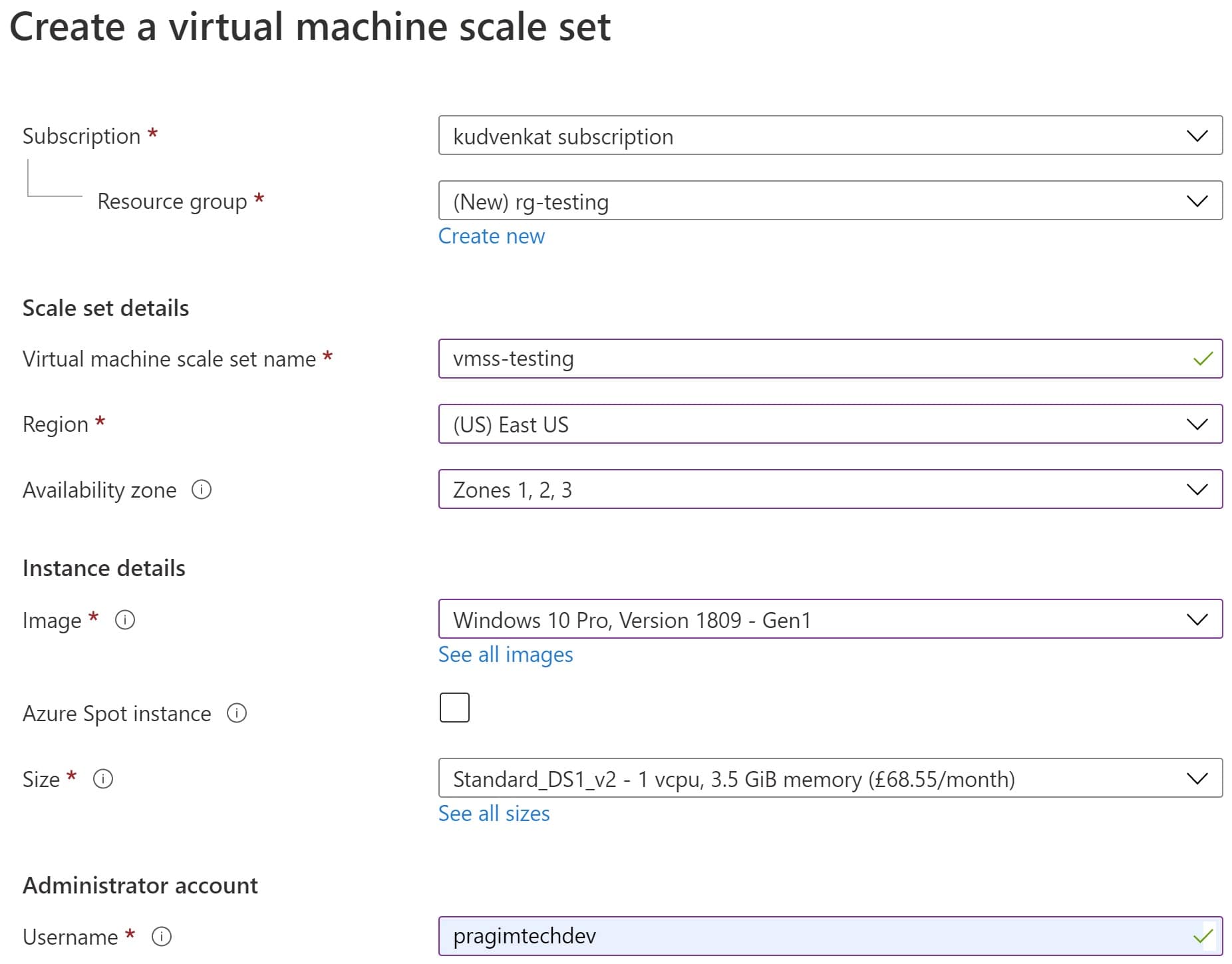
## **Virtual Machine Scale Set (VMSS) - Important points**

* A **Virtual Machine Scale Set** lets us easily create and manage multiple virtual machines.
* All VMs in a scale set are identical, meaning they are created from the same base operating system image and configuration. Obviously, this approach lets us easily manage hundreds of VMs without additional configuration.
* For traffic distribution, an **Azure Load Balancer** is also deployed along with the virtual machines in a scale set. There is no need to manually create the virtual machines or the load balancer. Just imagine the amount of time it saves.
* Even better, with the auto-scale feature, the number of virtual machines can automatically increase or decrease in response to demand. You can use metric based auto-scaling. There are many metrics from many sources in azure. For example, if the CPU utilization is over 70%, increase the VM instance count by 1. Not just scaling out, we can also auto-scale-in, when the demand subsides. For example, if the CPU utilization falls below 20%, decrease the VM instance count by 1.
* We can also increase or decrease the number of VMs in a scale-set based on a defined schedule. For example, let's say, next week, we are launching a new product or a business lines. So starting next week, we are going to have a heavy peak for 3 days. For this we can define a set schedule. For example, at 9:00 AM on Janury 1 increase the VM instance count to 50 and at 9:00 PM January 3, when our peak ends, bring the instance count back to our baseline configuration. How cool is that.
* 
* So, there are 3 ways to increase or decrease VMs in a scale set. Manually through the Azure portal or many of the azure APIs that are available. Auto scale based on metrics or a defined schedule.

# Create azure virtual machine scale set

**Step 1:**

As usual in the azure portal search for virtual machine scale set or it's acronym VMSS.



## **Disks tab**

Leave the default settings on the Disks tab

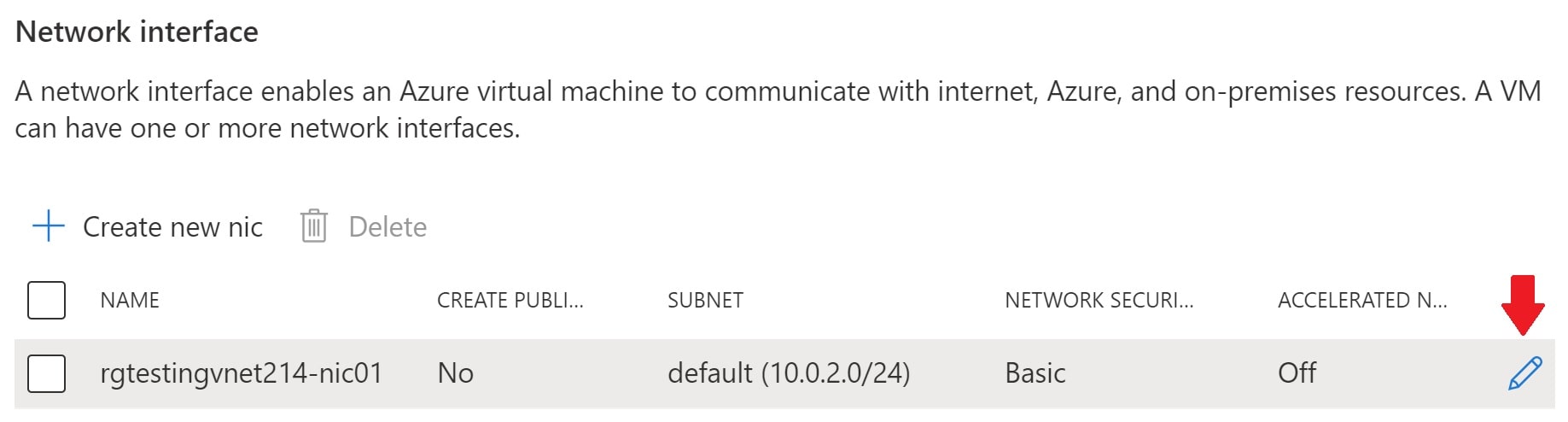
## **Networking tab**

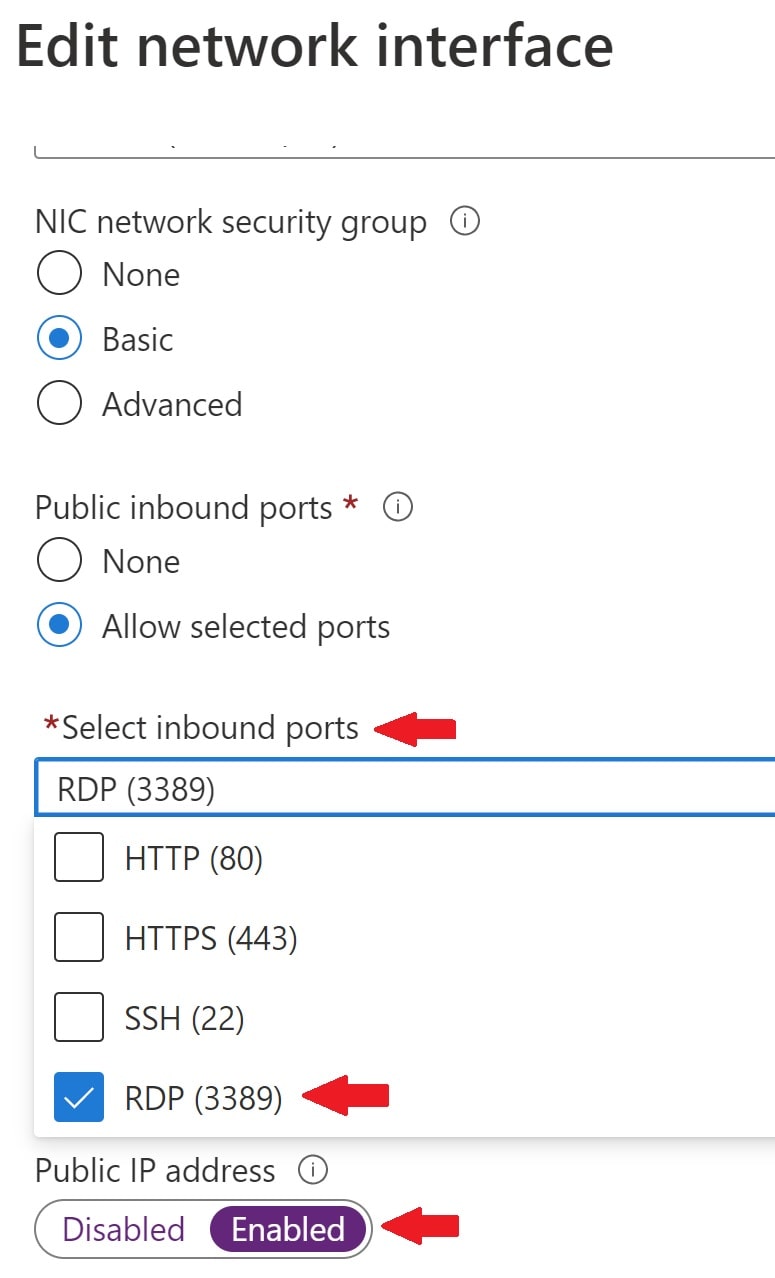
On the Networking tab we configure three things for our VMSS

1. **Virtual network**
2. **Network interface**
3. **Load balancing**

**Virtual network** -  Virtual network or VNet as it is commonly called allows Virtual Machines to securely communicate with each other, the internet, and on-premises networks. It is similar to a traditional network that you see in your own on-premise data center.

**Network interface** –

* A VM can have one or more Network interface. It is this Network interface that enables a VM to communicate with other VMs, internet and on-premises servers.
* By default, the virtual machines we create in azure are not accessible over the internet for obvious security reasons. However, you can override this by opening an inbound port.
* For example, let's say, we want to access the VM over the internet using RDP (Remote Desktop Protocol). For this we need to open the RDP inbound port. If you don't have this port open, you will not be able to RDP into the VM. Click on the pencil icon under Network interface category.
* 

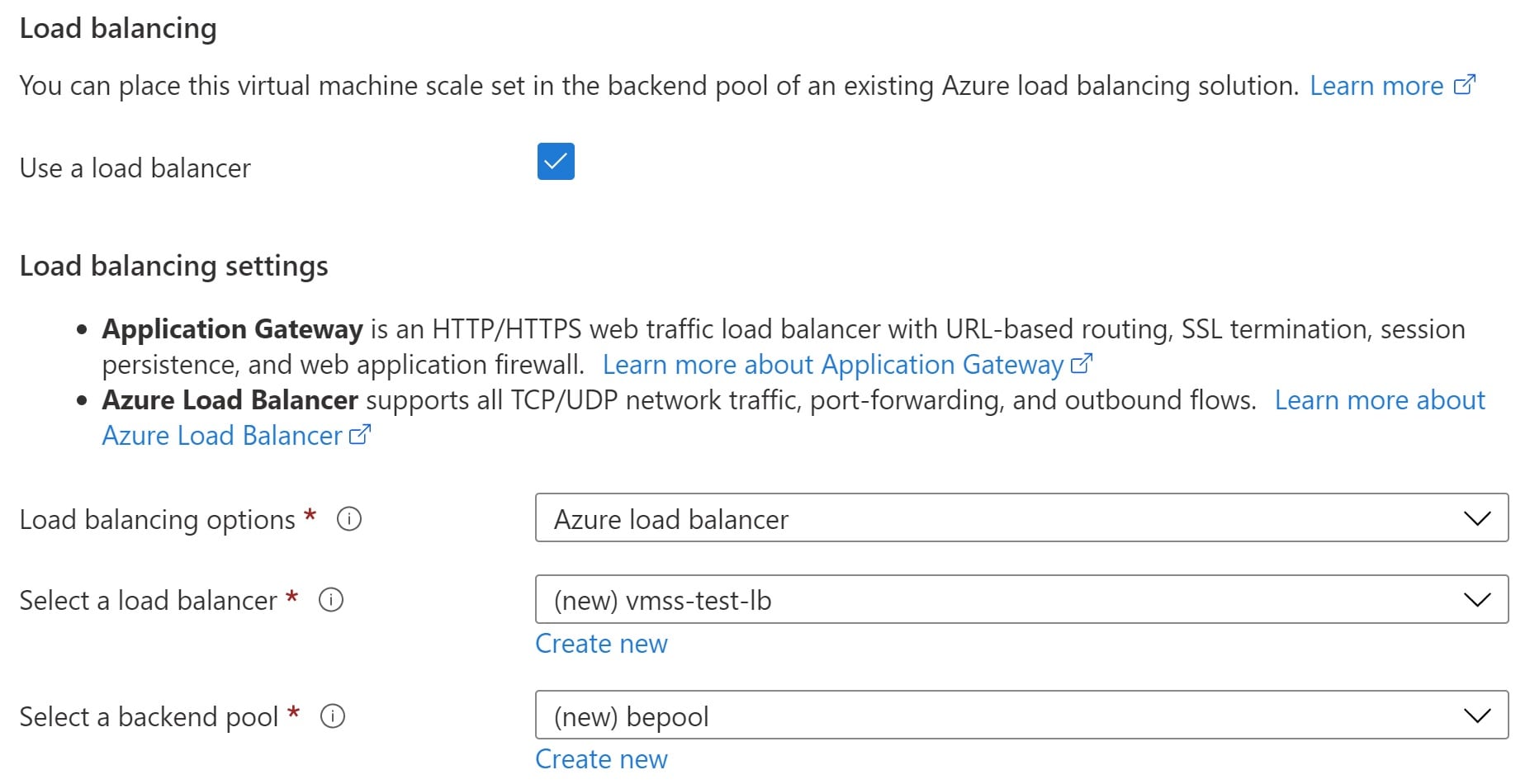


* Under *"Public inbound ports"*, select *"Allow selected ports"*
* Check *RDP (3389)* checkbox
* Enable Public IP address - Without this, we won't be able to communicate with the virtual machine from outside the virtual network.

With this configuration we are basically opening up port number 3389 for RDP communication.

### **Load balancer**

If you want a load balance to be created, check *"Use a load balancer"* checkbox.

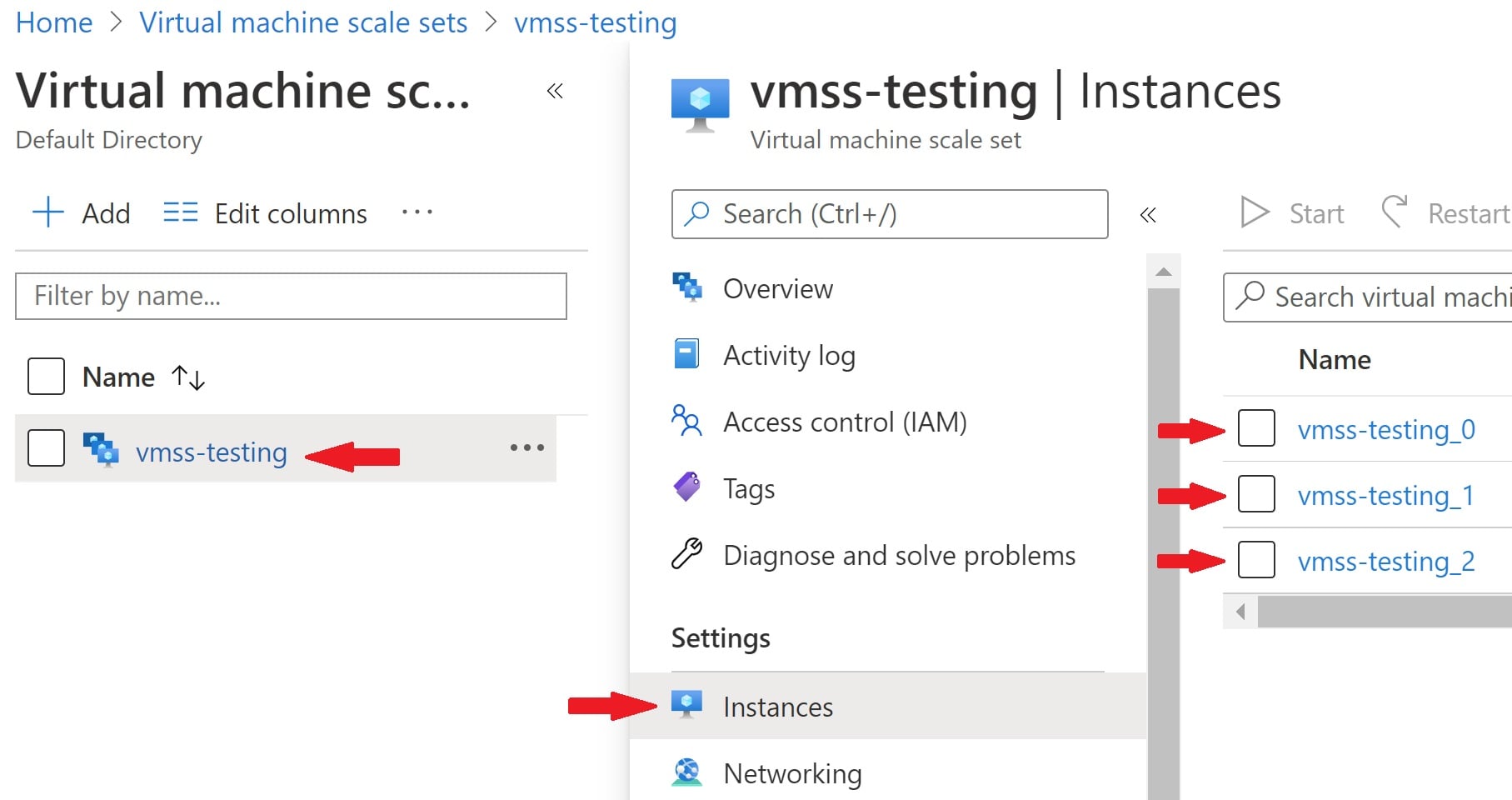


This automatically creates a new load balancer and distributes the traffic among all the VMs in our scale set.

## **Scaling tab**

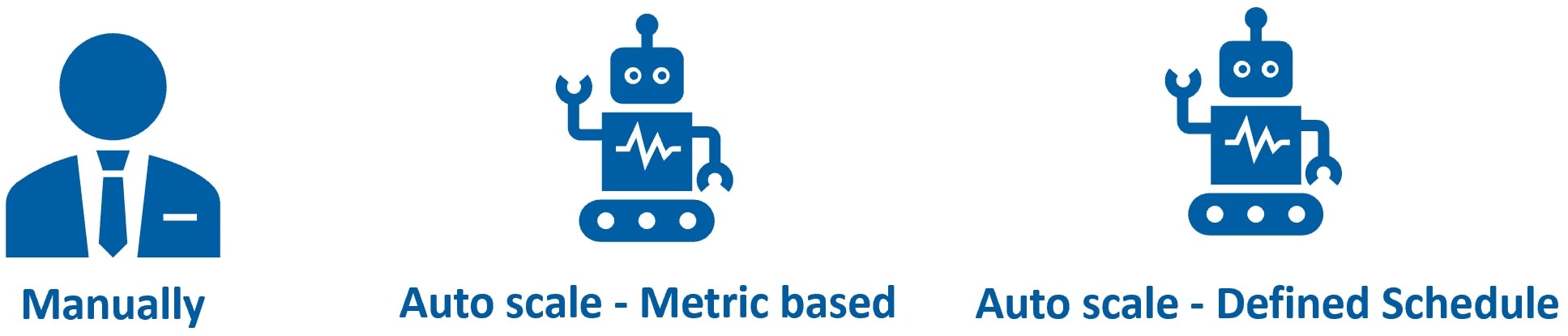
Initial instance count - This is where we specify the number of identical VMs we want in our scale-set. We can create upto 1000 VMs. Leave the rest of the defaults and click "Review + create"

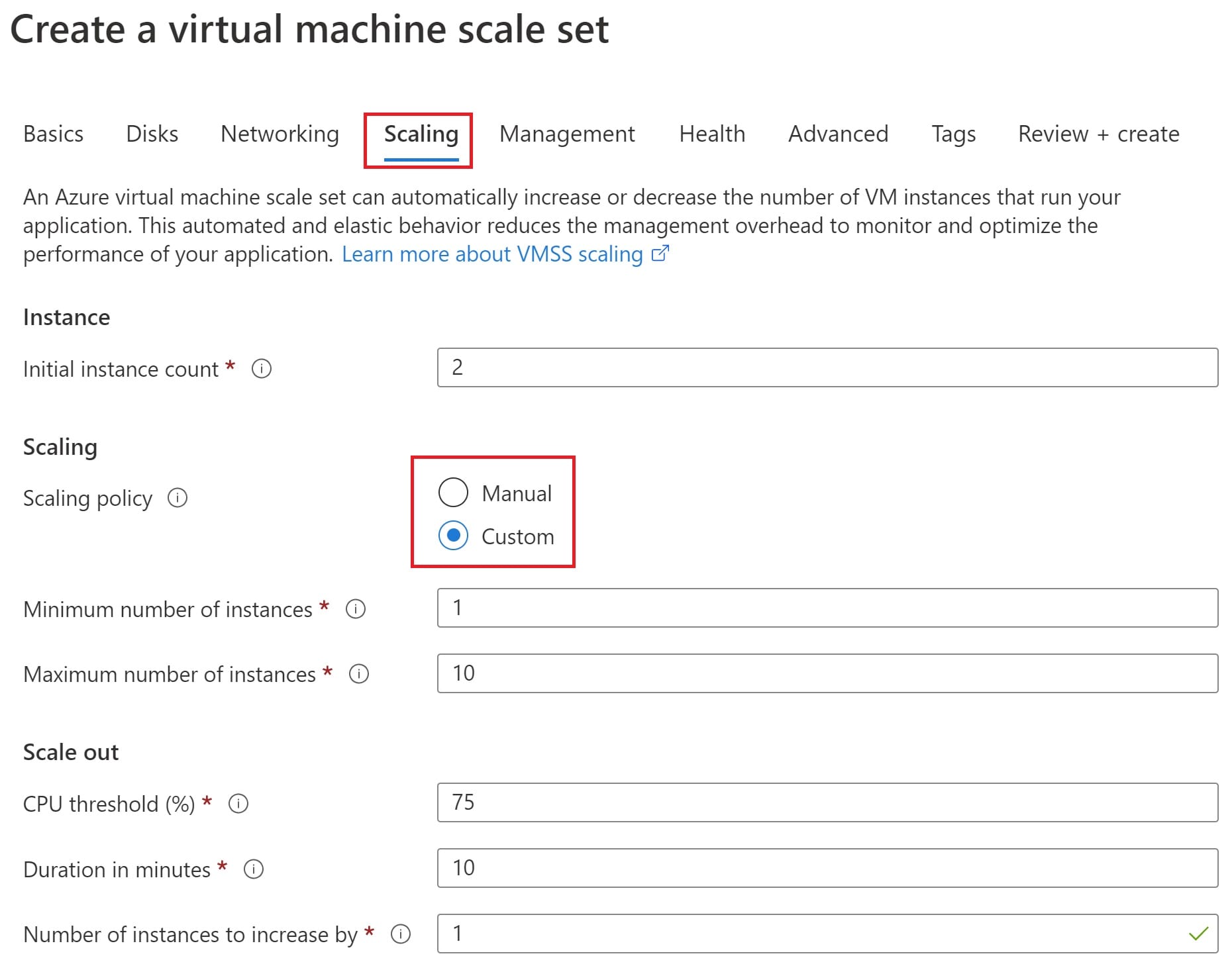
After a few minutes our virtual machine scale set must be created. After you see the message, your deployment is complete, search for VMSS and navigate to Virtual machine scale sets. You should see a Virtual machine scale set with the name vmss-testing



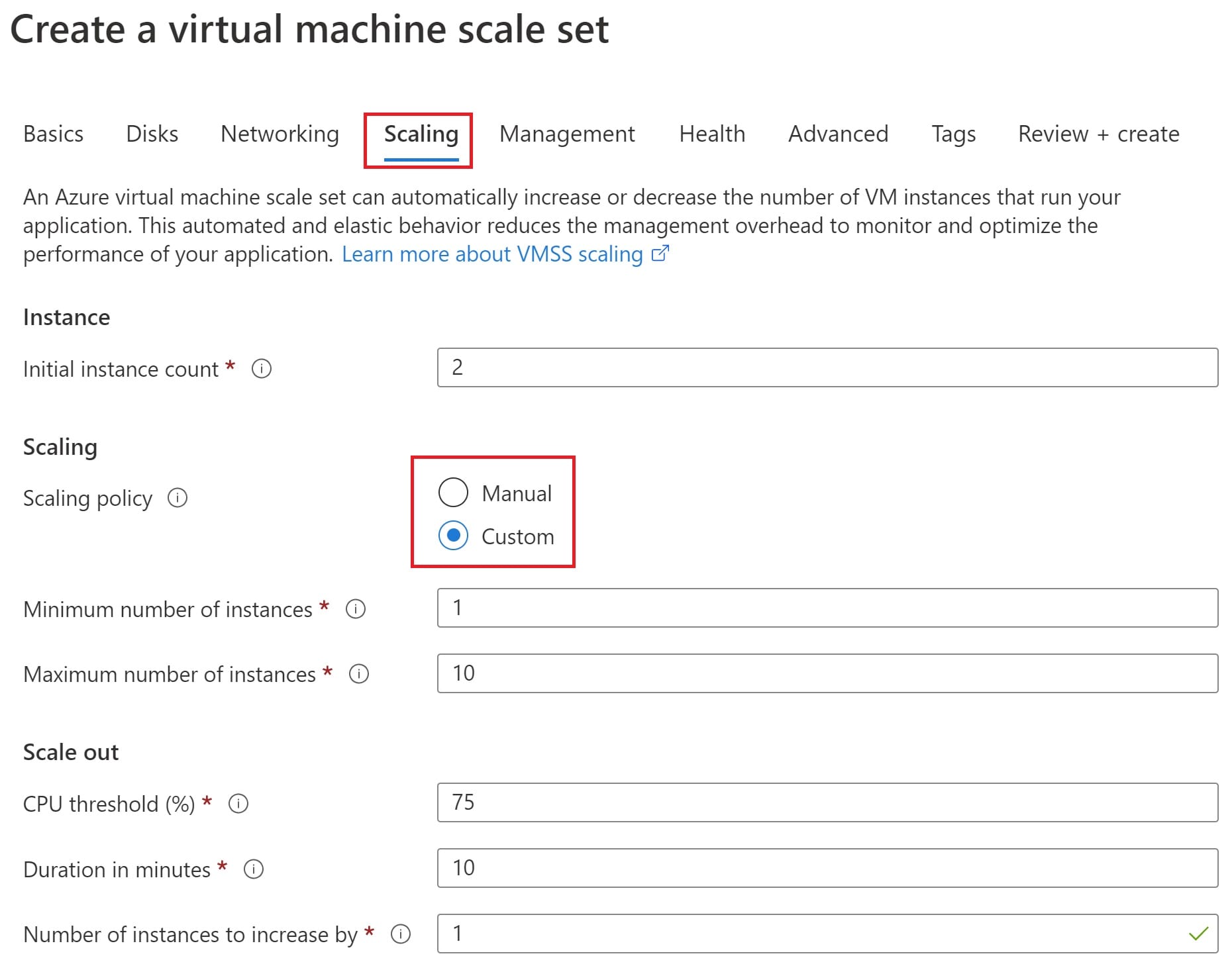
## **Virtual machine scale set scaling options**

In Azure, there are 3 options for scaling VM instances up and down.

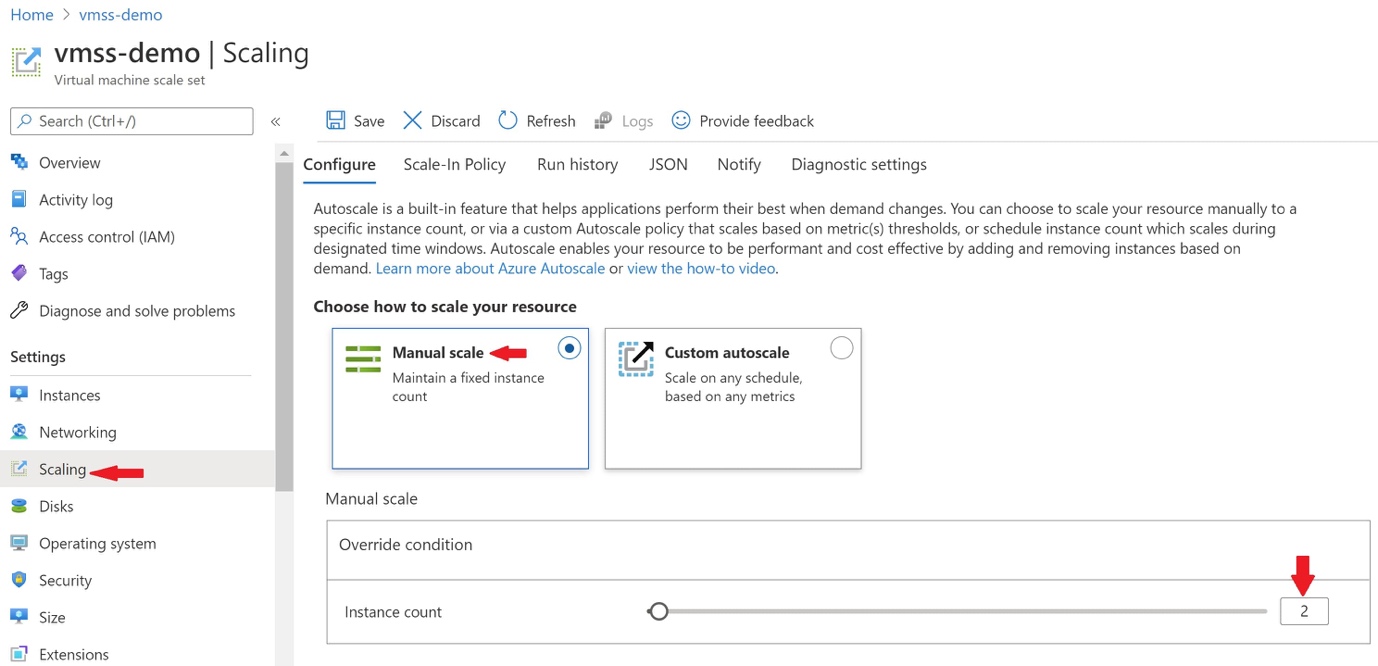
1. Manually through the Azure portal
2. Auto scale based on metrics or
3. Auto scale based on a defined schedule
4. 



You can configure scaling options either at the time of creating the virtual machine scale set or after it is created. At the time of creating the scale-set, you configure scaling on the Scaling tab as you can see from the image below.



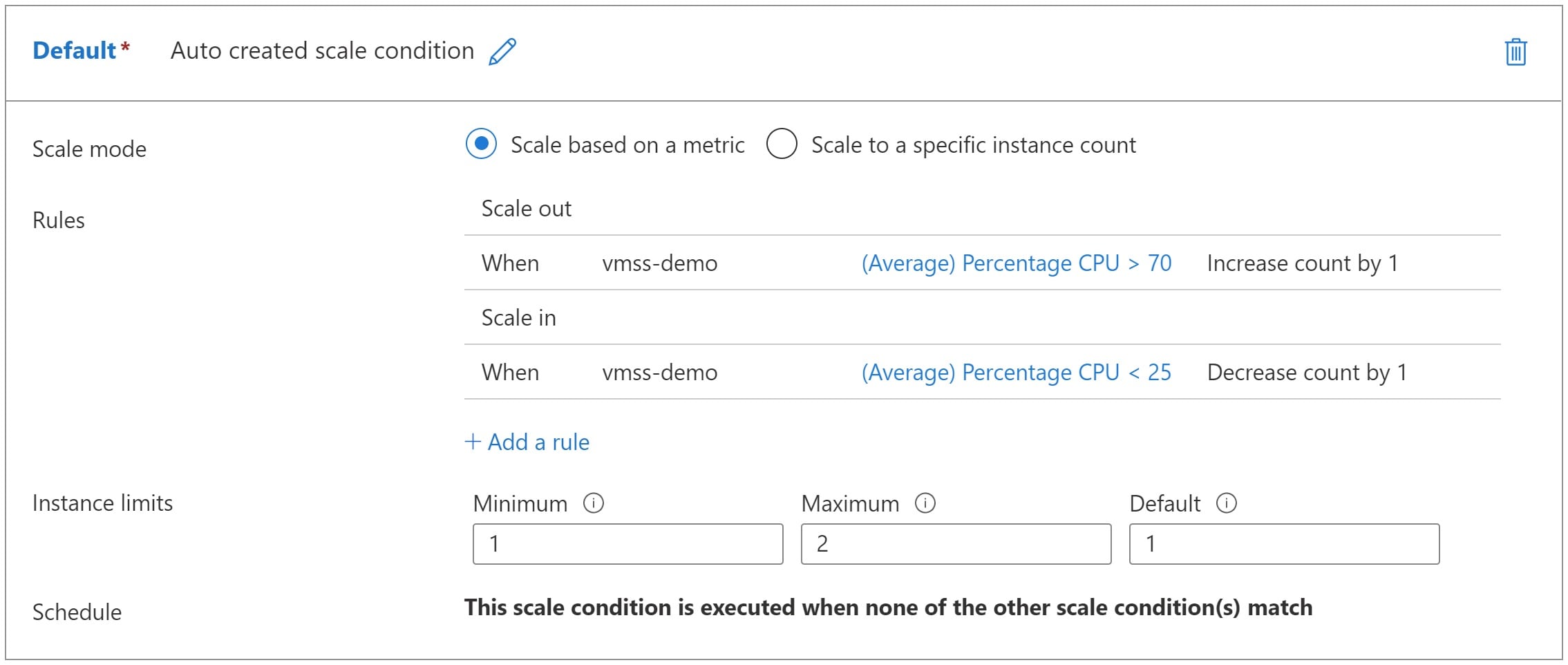
## **Virtual machine scale set - Manual Scale**



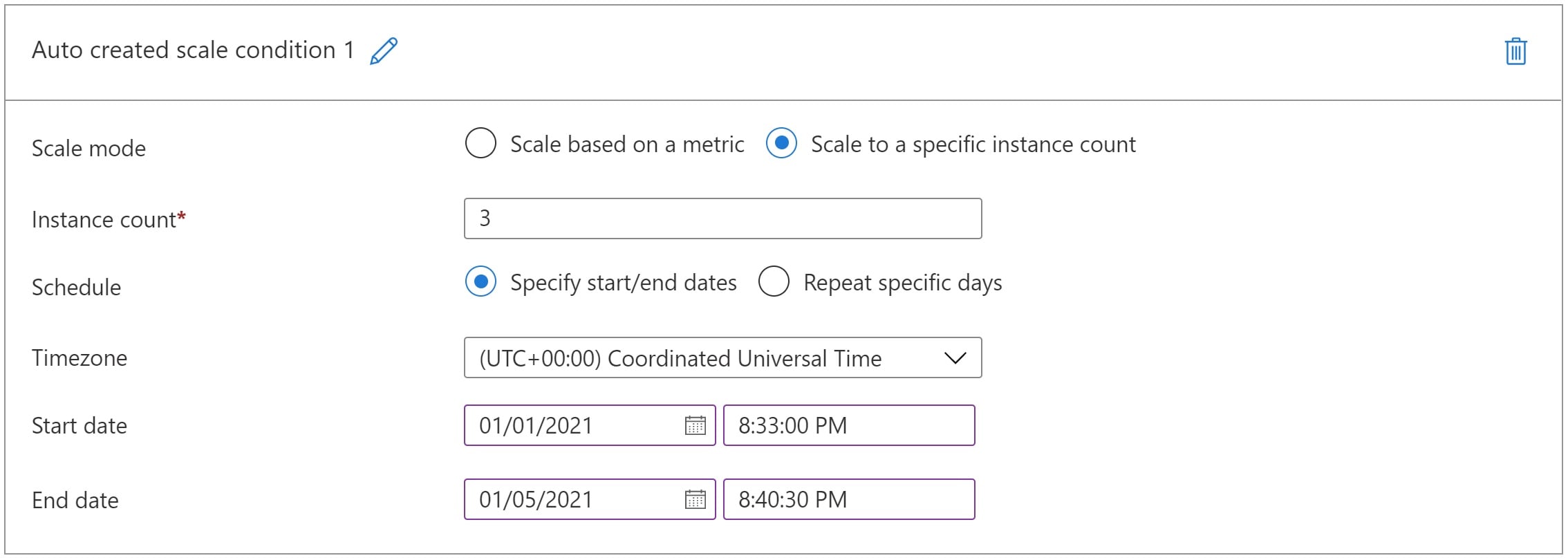
In the azure portal, navigate to your virtual machine scale set. Click on Scaling tab under Settings. To scale manually, specify the number of VM instances in the textbox provided. In a real world, it's a bit tedious and time consuming if we have to manually scale-out and scale-in everytime the demand for our application changes. However, it's an option, just in case, if you want to manually scale.

## **Virtual machine scale set - Custom Auto Scale**

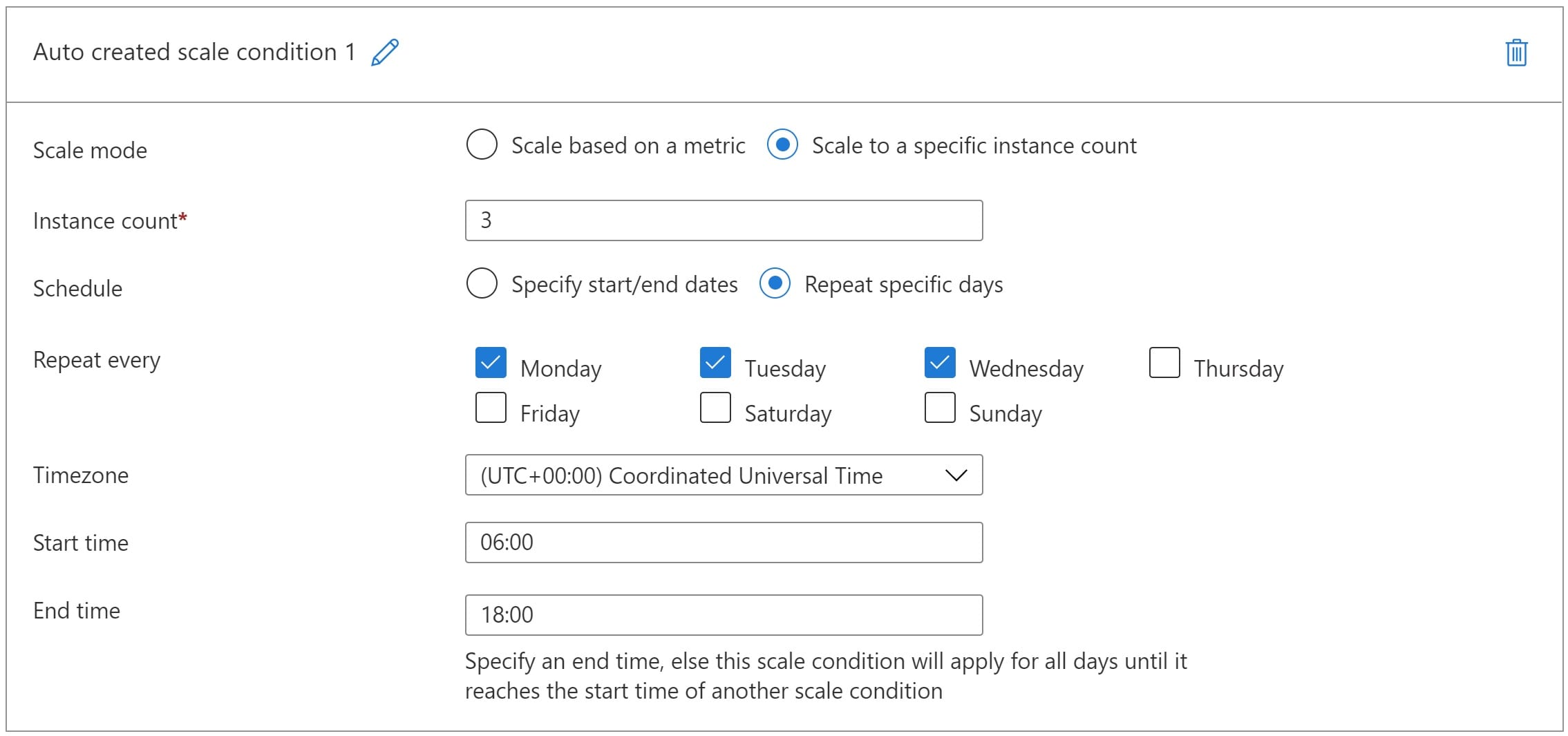
There are 2 ways to auto-scale. Auto scale based on metrics. For example, the following auto-scale condition increase the VM instance count by 1, when the average CPU utilization is greater than 70%. The instance count is decreased by 1, when the average CPU utilization falls below 25%.



Auto scale based on a schedule to a specific instance count. The following scale condition increases the total instance count to 3 on Jan 1, 2021 and on Jan 5, 2021 it falls back to the baseline configuration.

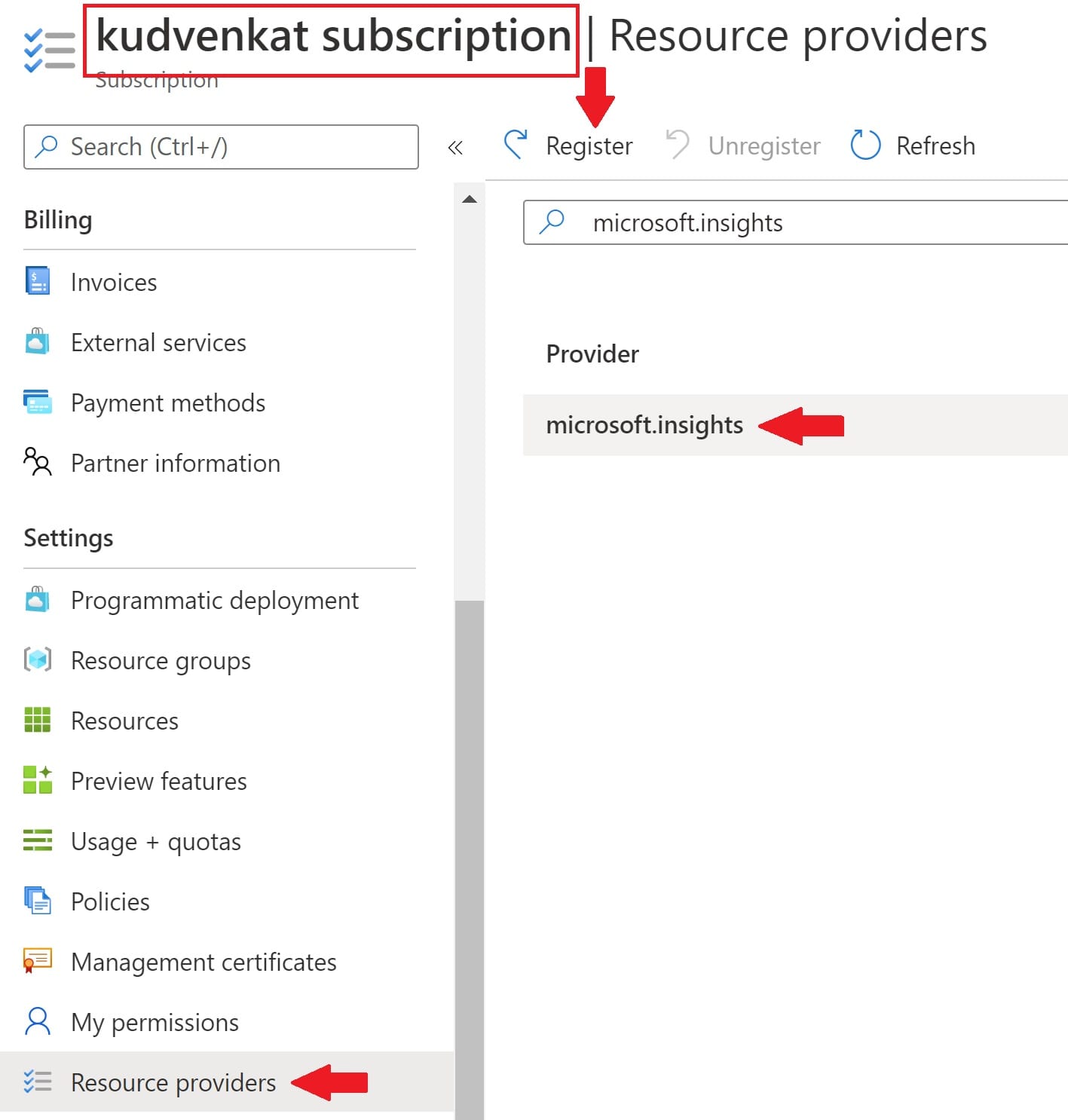


Scale to a specific instance count can be repeated on specific days of the week.



If you get the error - The subscription is not registered to use namespace 'microsoft.insights

1. Navigate to your Subscription blade in the azure portal
2. Click on Resource Providers
3. Search for Microsoft.Insights
4. Select it and click Register



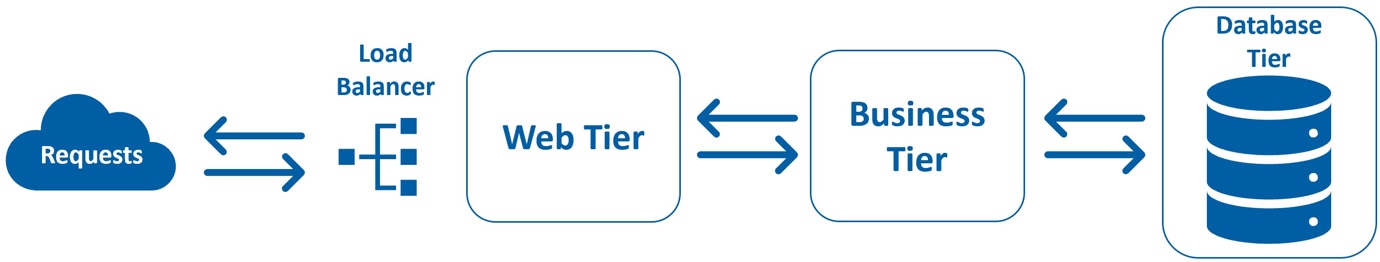
Failed to update configuration for 'vmss-demo'. {"error":{"code":"MissingSubscriptionRegistration","message":"The subscription is not registered to use namespace 'microsoft.insights'. See https://aka.ms/rps-not-found for how to register subscriptions.","details":[{"code":"MissingSubscriptionRegistration","target":"microsoft.insights","message":"The subscription is not registered to use namespace 'microsoft.insights'. See https://aka.ms/rps-not-found for how to register subscriptions."}]}}.

## **Summary**

If your application demand increases, the load on the VM instances in your scale set also may increase. If this increased load is consistent, rather than just a brief demand, you can configure autoscale rules, both to increase and reduce the number of VM instances in the scale set.

# When to use azure virtual machines ?

Something like a typical n-tier application, with a RESTful service accessing a database server.



It doesn't really matter what technology we have used to build this web application. ASP.NET Web forms, MVC, Java, PHP, or Python. If you want to host your web application, you may use an **Azure App Service.** There is no need for you to use a Virtual Machine. However, there are several use cases, why we may want to create and use a virtual machine.