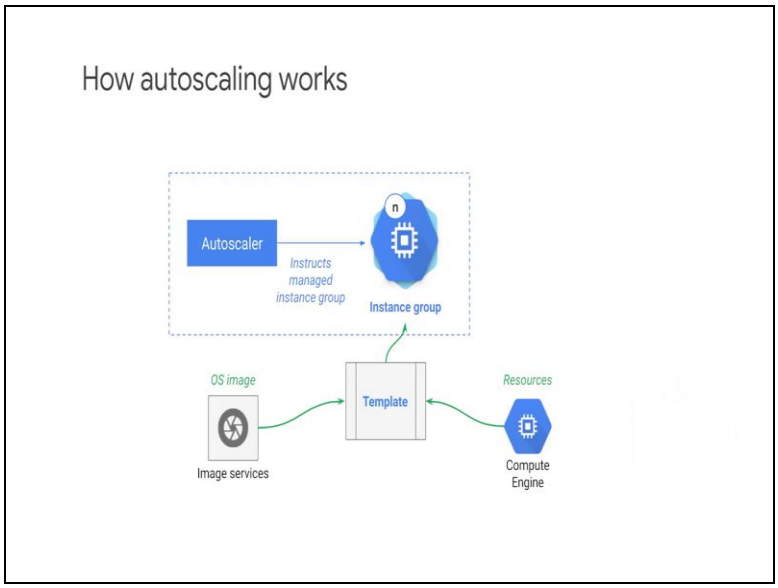
Google Cloud Computing Foundation Course
Sowmya Kannan
Google Cloud

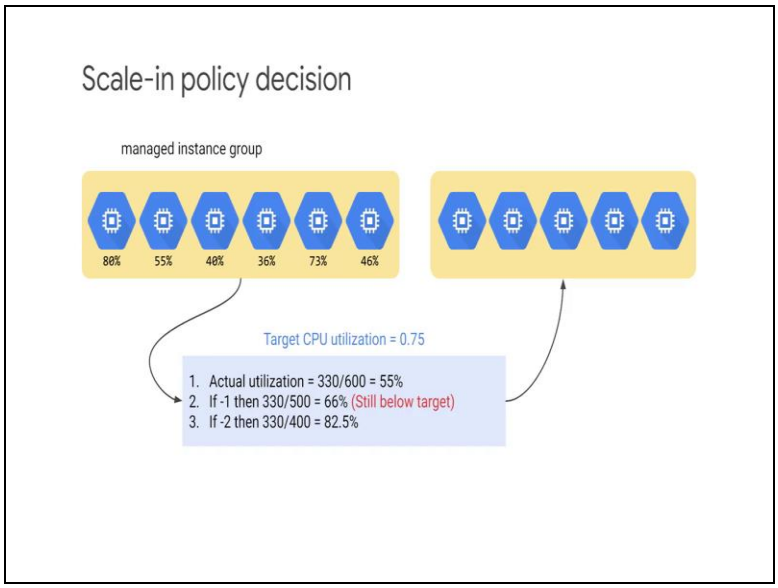
Lecture-20
Configuring Elastic Apps with Autoscaling

(Refer Slide Time: 00:07)



Your next topic looks at building elastic applications with auto scaling. Let us look at how auto scaling works. Auto scalar controls managed instance groups adding and removing instances using policies. A policy includes the minimum and maximum number of replicas. In this diagram n is any number of instance replicas based on a template. The template requisitions resource from compute engine identifies an OS image to boot and starts new VMs.

(Refer Slide Time: 00:42)



The percentage utilization that an additional VM contributes depends on the size of the group. The fourth VM added to a group offers a 25% increase in capacity to the group. The tenth VM added to a group only offers 10% more capacity even though the VMS are the same size. In this example auto scalar is conservative and rounds up. In other words it would prefer to start an extra VM that is not really needed than to possibly run out of capacity.

In this example removing one who VM does not get close enough to the target of 75% removing a second VM would exceed the target. Auto scalar behaves conservatively. So, it will shut down one VM rather than two VMs. It would prefer under utilization over running out of resources when they are needed.