Autoscaling

CPU threshold — Ortega, Ivan: Specify the CPU usage percentage threshold to trigger the scale-in autoscale rule

Number of VM's (scale in) — Ortega, Ivan: Specify the number of virtual machines to remove from your implementation when the scale-in autoscale rule is triggered.

Scale in

Duration — Ortega, Ivan: Duration in minutes is the amount of time that Autoscale engine will look back for metrics.

Number of VM's (scale out) — Ortega, Ivan: Specify the number of virtual machines to add to your Virtual Machine Scale Sets implementation when the scale-out autoscale rule is triggered

CPU threshold (scale out) — Ortega, Ivan: Specify the CPU usage percentage threshold to trigger the scale-out autoscale rule

Scale out

Schedule events — Ortega, Ivan: You can implement autoscaling and schedule events to automatically increase or decrease the capacity of your implementation at fixed times

Automatic adjusted capacity — Ortega, Ivan: Create autoscaling rules to define the acceptable performance for a positive customer experience

Scaling policy

Minimum number — Ortega, Ivan: Specify the minimum number of virtual machines that should be available when autoscaling is applied on your Virtual Machine Scale Sets implementation

Maximum number of VMs — Ortega, Ivan: Specify the maximum number of virtual machines that can be available when autoscaling is applied on your implementation.

Create a virtual machine scale set ...

Basics Disks Networking **Scaling** Management Health Advanced

Initial instance count * ⓘ [2]

Scaling

Scaling policy ⓘ   ◯ Manual scaling
                         ⦿ Autoscaling

Minimum number of instances * ⓘ [1]

Maximum number of instances * ⓘ [10]

Scale out

CPU threshold (%) * ⓘ [75]

Duration in minutes * ⓘ [10]

Number of instances to increase by * ⓘ [1 ✓]

Scale in

CPU threshold (%) * ⓘ [25]

Number of instances to decrease by * ⓘ [1 ✓]

Scale-In policy

Configure the order in which virtual machines are selected for deletion during a scale-in operation.

Scale-in policy
- Default - Balance across availability zones and fault domains, then delete VM with highest instance ID
- Newest VM - Balance across availability zones, then delete the newest created VM
- Oldest VM - Balance across availability zones, then delete the oldest created VM