9

Scaling Operations

Module Objectives

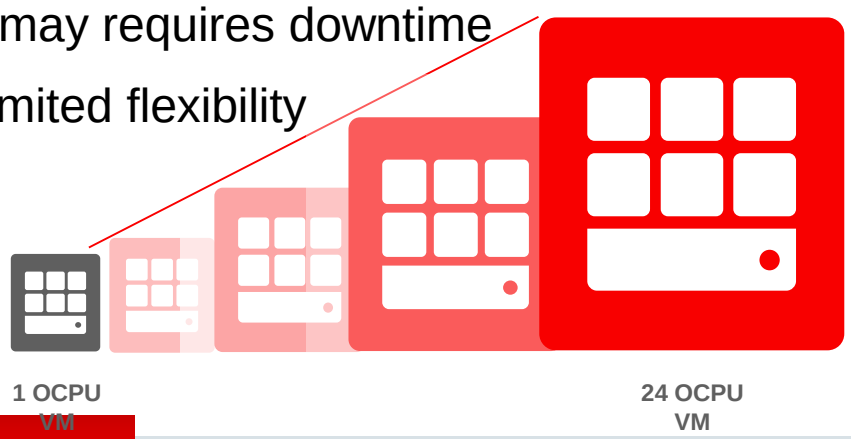
- Describe Scaling options
- Identify Horizontal and Vertical Scaling scenarios
- Describe Compute, Storage and Database Vertical Scale
- Describe Autoscaling and Metrics

Scaling Primer

Vertical Scaling

- scale-up/scale-down approach
- Increase/Decrees the capacity of a single instance (i.e CPU, RAM, Storage size)

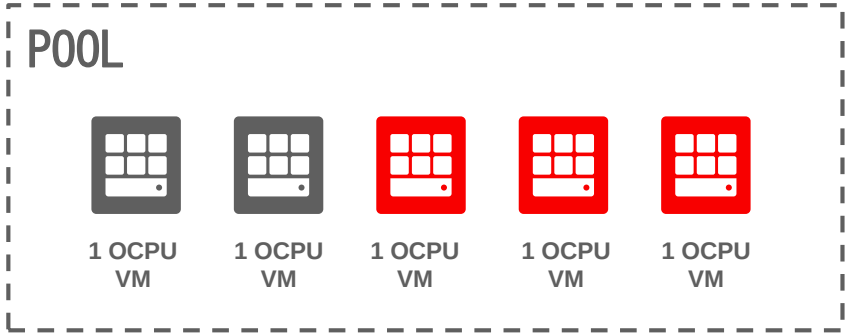
- **PRO**
  - Adapt to host monolithic application and workloads not compatible with distributed environment
- **CONS**
  - It may requires downtime
  - Limited flexibility



Horizontal Scaling

- scale-out/scale-in approach
- Increase / Decrease the number of nodes

- **PRO**
  - adapt to host clustered applications and distributed environment
  - Unlimited scaling
- **CONS**
  - It may requires to re-architect older applications and verticals workload solutions



Vertical Scaling

Vertical Scaling – Instance Offline Resize

The Oracle Cloud Infrastructure Compute service lets you change the instance shape

Resize Instance help cancel

Change the size of your instance to support changes in application workload.

Current Shape: VM.Standard2.1

This instance is running. You must stop the instance before you resize it. Learn more about resizing instances.

	Shape Name	OCPU	Memory (GB)	Local Disk (TB)	Network Bandwidth	Max Total VNICs
☑	VM.Standard2.1	1	15	Block Storage only	1 Gbps	2
☐	VM.Standard2.2	2	30	Block Storage only	2 Gbps	2
☐	VM.Standard2.4	4	60	Block Storage only	4.1 Gbps	2
☐	VM.Standard2.8	8	120	Block Storage only	8.2 Gbps	4
☐	VM.Standard2.16	16	240	Block Storage only	16.4 Gbps	8
☐	VM.Standard2.24	24	320	Block Storage only	24.6 Gbps	12

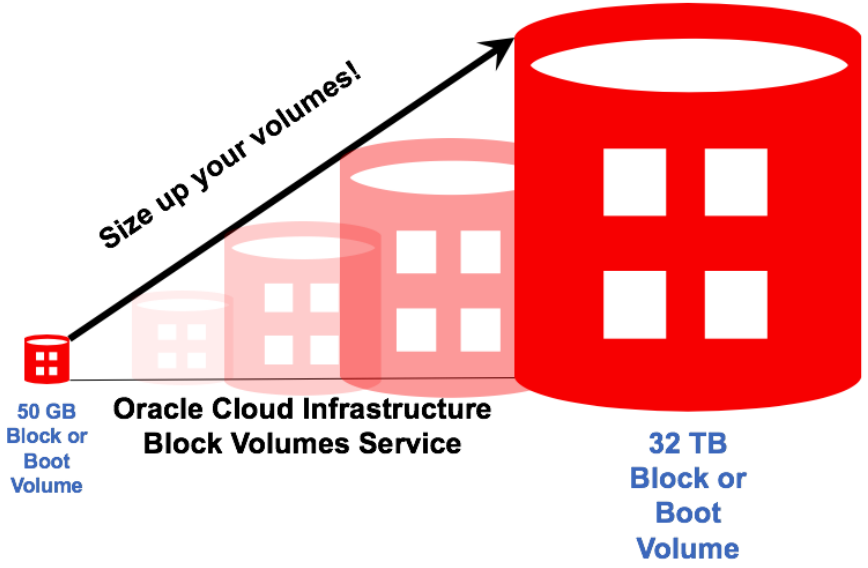
1 Selected Showing 6 Item(s)

Resize Cancel

- Scale-up and Scale-down supported
- New shape must have the some hardware architecture.
- Downtime is required. The instance must be stopped before resize it

Vertical Scaling – Block Volume Offline Resize

The Oracle Cloud Infrastructure Block Volume service lets you expand the size of block volumes and boot volumes. You have three options to increase the size of your volumes:



- Expand an existing volume in place with offline resizing.
- Restore from a volume backup to a larger volume.
- Clone an existing volume to a new, larger volume.

You can only increase the size of the volume, **you cannot decrease the size**.

Vertical Scaling – Boot Volume Linux Partition Resize

After resizing the instance Boot Volume, in order to take advantage of the larger size, you need to extend the partition for the boot volume.
Linux OS supports both Online and Offline partition resize.

Offline

1. Stop the instance
2. Detach the boot volume
3. Attach the boot volume to a second instance as a block volume
4. Run parted to edit the partition
5. Run ***xfs_growfs*** to grow the file system
6. Detach the volume from the second instance
7. Attach the volume to the original instance as a boot volume
8. Restart the instance

Here the detailed step-by-step process

Vertical Scaling – Boot Volume Linux Partition Resize

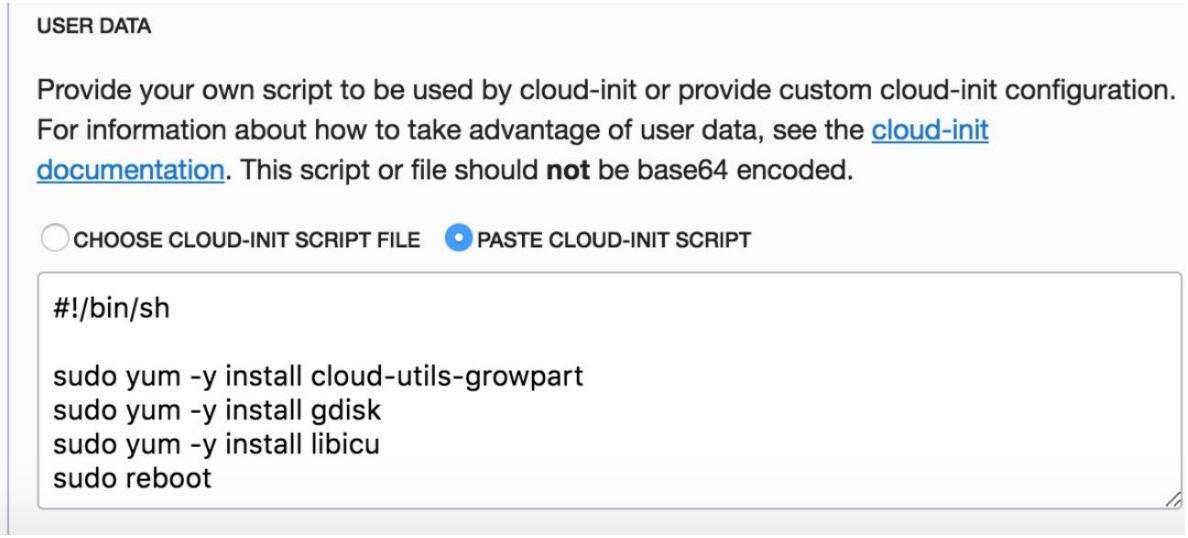
Online Manual Partition Resize

1. Use SSH to connect to your instance
2. Resize the partition using *growpart* and *gdisk*
3. Grow the file system using *xfs_growfs* or *resize2fs*

Online Automatic Partition Resize

On Oracle Linux and CentOS you can *cloud-init-growpart* along with *gdisk* and *cloud-init* to completely automate this process. .

You have to provide a cloud-init userdata script at provisioning time



Vertical Scaling – Boot Volume Windows Partition Resize

On Windows-based images, you can extend a partition using the Windows interface or from the command line using the *DISKPART* utility.

Windows Interface

1. Open Disk Management (in Windows 2008 it is in the Server Manager)
2. Use the Extend Volume wizard

Command Line

1. Open a command prompt as Administrator
2. Run *DISKPART*
3. Select and extend the volume

Vertical Scaling – DB-Systems

DB-Systems provides the ability to scale with no downtime

Virtual Machine (VM)	Bare Metal (BM)	Exadata
Storage Scale-up with no downtime	CPU Scale up and Scale Down with no downtime	CPU Scale-up and Scale-down with no downtime Rack-shape scale-up within a ¼ , ½ and Full rack.

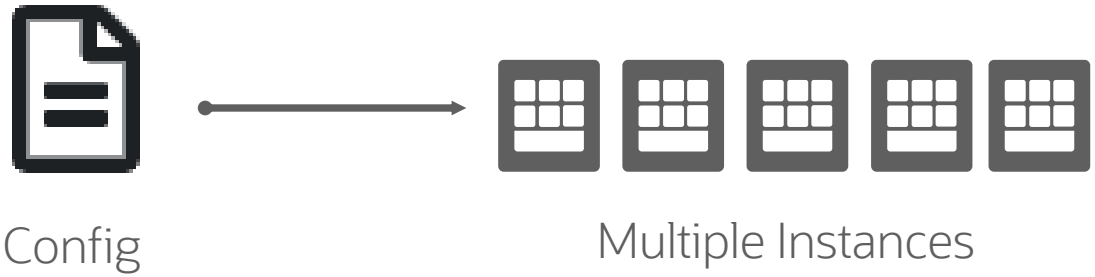
Horizontal Scaling & Autoscaling

Instance Configuration and Pool

Instance Configurations

- OS image, metadata, shape
- vNICs, Storage, subnets

Instance Configurations



- Different Availability Domains
- Manage all together (stop, start, terminate)
- Attach to a Load Balancer

Instance Configuration and Pool – Use Cases

Instance Configurations

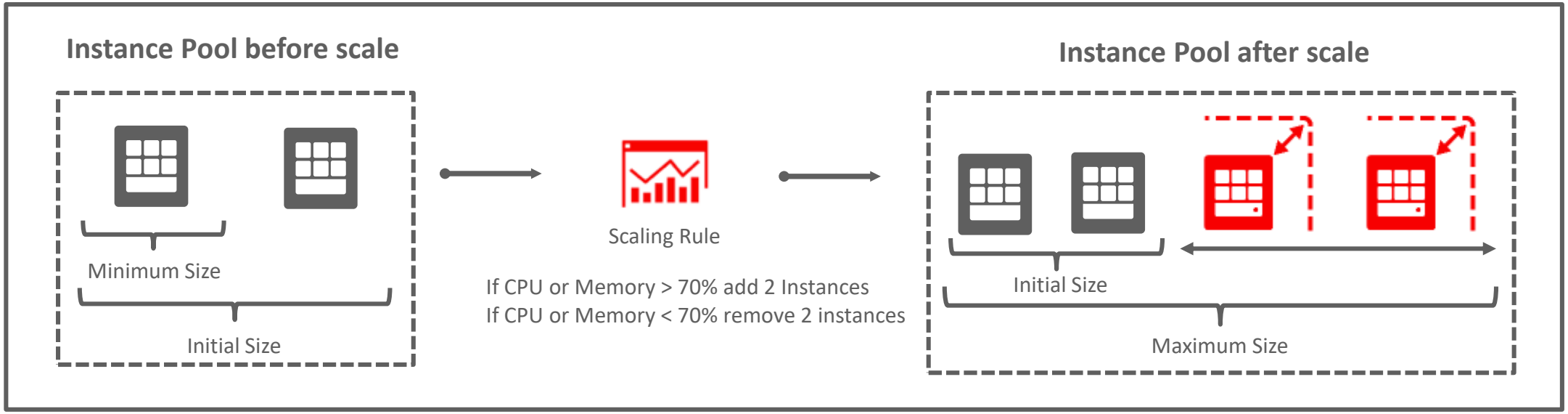
- Clone an instance and save to a configuration file
- Create standardized baseline instance templates
- Easily deploy instances from CLI with a single configuration file
- Automate the provisioning of many instances, its resources and handle the attachments

Instance Pools

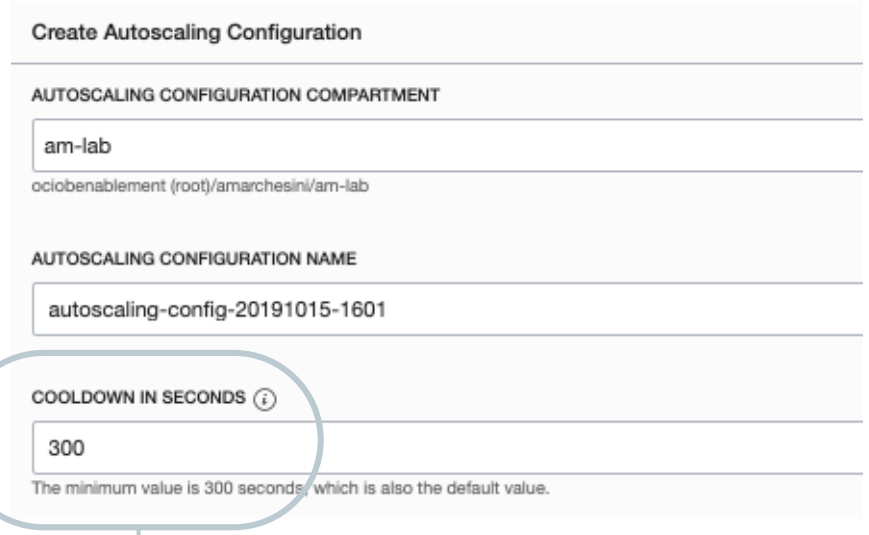
- Centrally manage a group of instance workloads that are all configured with a consistent configuration
- Update a large number of instances with a single instance configuration change
- Maintain high availability and distribute instances across availability domains within a region
- Scale out instances on-demand by increasing the instance size of the pool

Autoscaling Configurations

- Autoscaling enables you to automatically adjust the number of Compute instances in an instance pool based on performance metrics such as CPU or Memory utilization.
- When an instance pool scales in, instances are terminated in this order: the number of instances is balanced across availability domains, and then balanced across fault domains. Finally, within a fault domain, the oldest instance is terminated first.

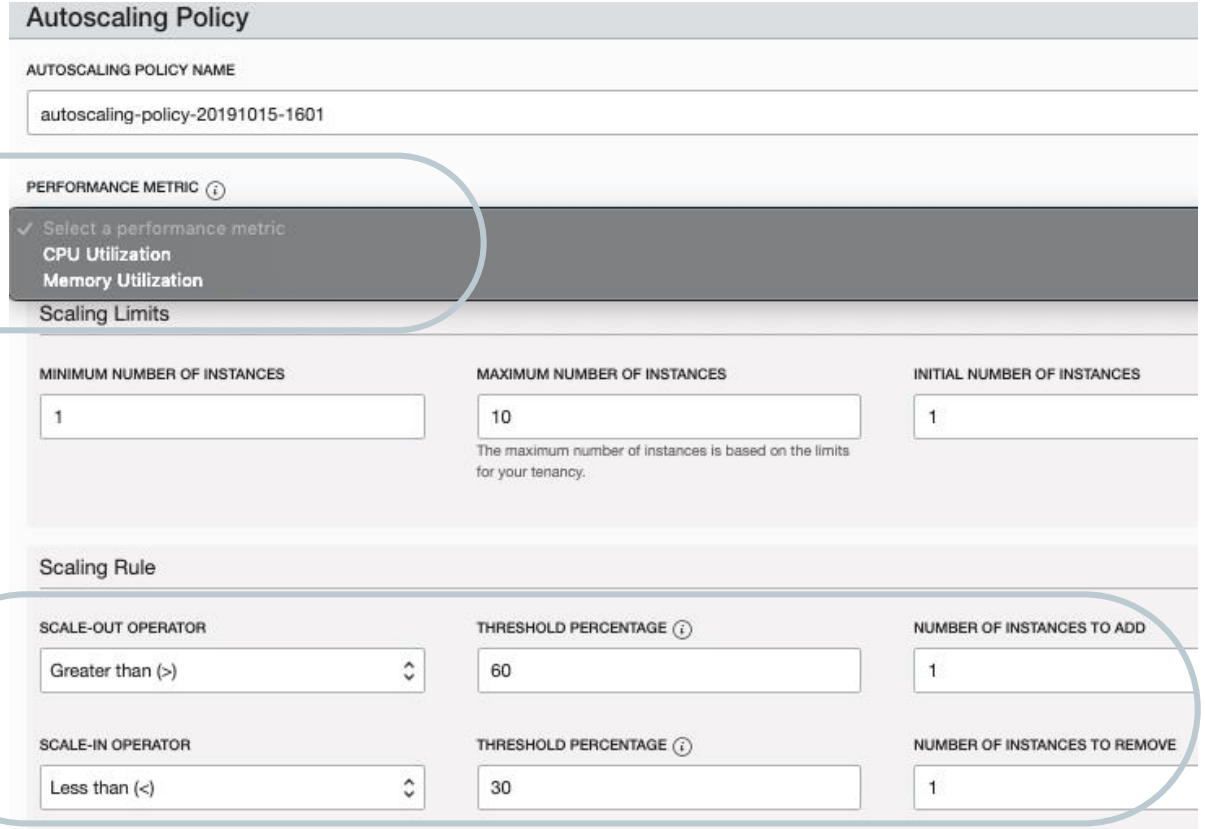


Autoscaling Configurations – Scaling Rules



The cooldown period gives the system time to stabilize before rescaling

The metric that triggers an increase or decrease in the number of instances in the pool can depend on either CPU or memory utilization

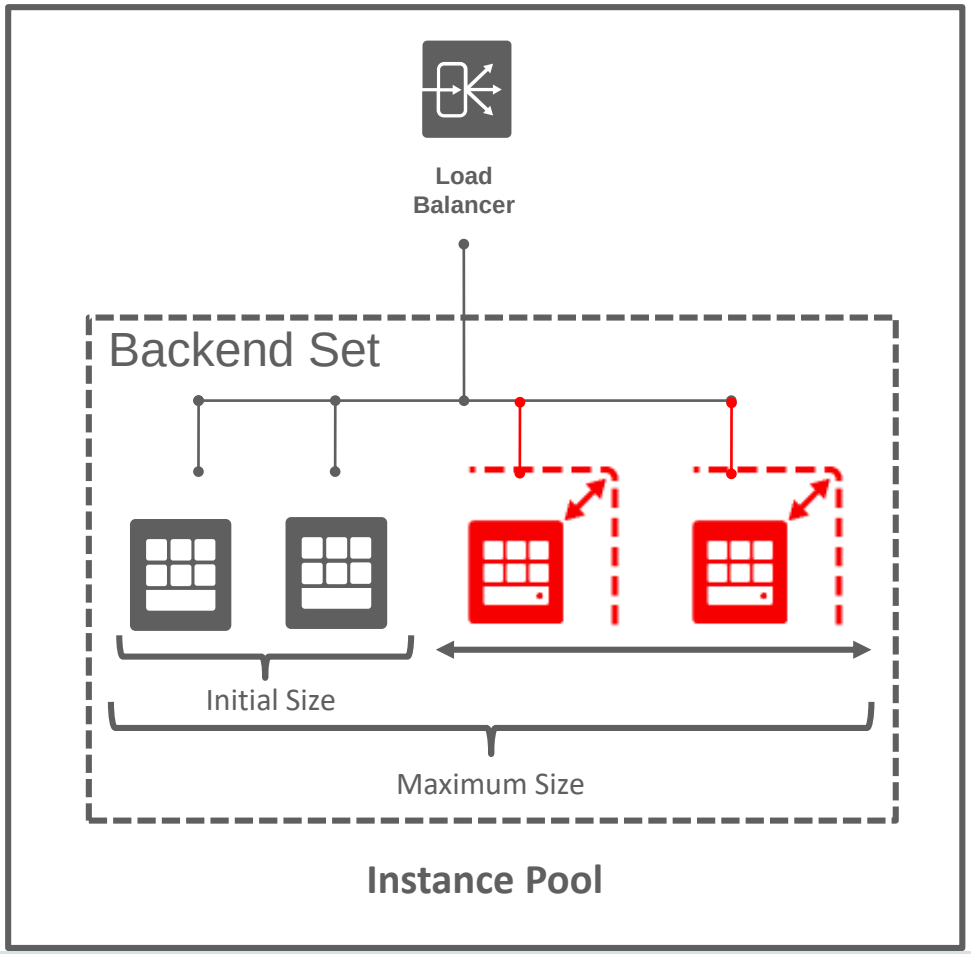


Scaling rules depend on thresholds that the performance metric must reach to trigger a scaling event.

Autoscaling and Load Balancer

Load Balancer instance can be attached to an instance pool configuration.

- On scale-out the new nodes are automatically added to the specified backend set.
- On scale-in the terminated nodes are automatically removed from the backend set



Autonomous DB Scaling options

On demand Scale

- Not constrained by fixed building blocks, no predefined shapes
- Independently scale compute or storage
- Resizing occurs instantly, fully online
- Memory, IO bandwidth, concurrency scales linearly with CPU
- No downtime

Autoscaling

- Automatically increase the number of CPU cores by up to three times the assigned CPU core count value, depending on demand for processing. The auto scaling feature reduces the number of CPU cores when additional cores are not needed
- No downtime

Summary

- Understand scale-up/down and scale-out/in options
- Understand Instance Configuration, Pools and Autoscaling
- Setup a Compute Autoscaling Policy
- Understand Autonomous DB scaling options