

Configure virtual machine availability

1. Introduction

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/1-introduction>

Introduction

Completed

Managing virtual machines at scale can be challenging, especially when usage patterns vary and demands on applications fluctuate. Azure Administrators need to be able to adjust their virtual machine resources to match changing demands. At the same time, they need to keep their virtual machine configuration consistent to ensure application stability. Achieving these goals means maintaining throughput and responsiveness while minimizing the costs of continually running a large collection of virtual machines.

Your company website uses virtual machines and manages large workloads. The IT department wants to ensure the virtual machines can dynamically adjust to increases and decreases in workloads. They also want to ensure there's a business continuity plan to provide for highly available machines. You're responsible for deploying highly available virtual machines. You decide to use Azure Virtual Machine Scale Sets and the autoscale feature.

In this module, you learn about scaling virtual machines. You learn about availability zones, availability sets, and update and fault domains. You also learn about scale sets and autoscale.

The goal of this module is to learn how to successfully respond to changing virtual machine workloads.

Learning objectives

In this module, you learn how to:

- Implement availability sets and availability zones.
- Implement update and fault domains.
- Implement Azure Virtual Machine Scale Sets.
- Autoscale virtual machines.

Skills measured

The content in the module helps you prepare for [Exam AZ-104: Microsoft Azure Administrator](#).

Prerequisites

- Familiarity with creating and managing Azure virtual machines.
- General knowledge of scaling infrastructure resources in fluctuating workloads.

2. Plan for maintenance and downtime

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/2-plan-for-maintenance-downtime>

Plan for maintenance and downtime

Completed

Azure Administrators prepare for planned and unplanned failures.

Things to know about maintenance planning

An availability plan for Azure virtual machines needs to include strategies for unplanned hardware maintenance, unexpected downtime, and planned maintenance.

- An **unplanned hardware maintenance** event occurs when the Azure platform predicts that the hardware or any platform component associated to a physical machine is about to fail. When the platform predicts a failure, it issues an unplanned hardware maintenance event. Azure uses Live Migration technology to migrate your virtual machines from the failing hardware to a healthy physical machine. Live Migration is a virtual machine preserving operation that only pauses the virtual machine for a short time, but performance might be reduced before or after the event.
- **Unexpected downtime** occurs when the hardware or the physical infrastructure for your virtual machine fails unexpectedly. Unexpected downtime can include local network failures, local disk failures, or other rack level failures. When detected, the Azure platform automatically migrates (heals) your virtual machine to a healthy physical machine in the same datacenter. During the healing procedure, virtual machines experience downtime (reboot) and in some cases loss of the temporary drive.
- **Planned maintenance** events are periodic updates made by Microsoft to the underlying Azure platform to improve overall reliability, performance, and security of the platform infrastructure that your virtual machines run on.

Note

Microsoft doesn't automatically update your virtual machine operating system or other software. You have complete control and responsibility for those updates. However, the underlying software host and hardware are periodically patched to ensure reliability and high performance.

3. Create availability sets

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/3-setup-availability-sets>

Create availability sets

Completed

An availability set is a logical feature you can use to ensure a group of related virtual machines are deployed together. The grouping helps to prevent a single point of failure from affecting all of your machines. The grouping ensures that not all of the machines are upgraded at the same time during a host operating system upgrade in the datacenter.

Things to know about availability sets

Let's review some characteristics of availability sets.

- All virtual machines in an availability set should perform the identical set of functionalities.
- All virtual machines in an availability set should have the same software installed.
- Azure ensures that virtual machines in an availability set run across multiple physical servers, compute racks, storage units, and network switches.

If a hardware or Azure software failure occurs, only a subset of the virtual machines in the availability set are affected. Your application stays up and continues to be available to your customers.

- You can create a virtual machine and an availability set at the same time.

A virtual machine can only be added to an availability set when the virtual machine is created. To change the availability set for a virtual machine, you need to delete and then recreate the virtual machine.

- You can build availability sets by using the Azure portal, Azure Resource Manager (ARM) templates, scripting, or API tools.

Note

Adding your virtual machines to an availability set doesn't protect your applications from operating system or application-specific failures. You need to explore other disaster recovery and backup techniques to provide application-level protection.

Things to consider when using availability sets

Availability sets are an essential capability when you want to build reliable cloud solutions. In your planning for availability sets, keep the following general principles in mind:

- **Consider redundancy.** To achieve redundancy in your configuration, place multiple virtual machines in an availability set.
- **Consider separation of application tiers.** Each application tier exercised in your configuration should be located in a separate availability set. The separation helps to mitigate single point of failure on all machines.
- **Consider load balancing.** For high availability and network performance, create a load-balanced availability set by using Azure Load Balancer. Load Balancer distributes incoming traffic across working instances of services that are defined in your load-balanced availability set.
- **Consider managed disks.** You can use Azure managed disks with your Azure virtual machines in availability sets for block-level storage.

4. Review update domains and fault domains

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/4-review-update-fault-domains>

Review update domains and fault domains

Completed

Azure Virtual Machine Availability Sets implements two node concepts to help Azure maintain high availability and fault tolerance when deploying and upgrading applications: *update domains* and *fault domains*. Each virtual machine in an availability set is placed in one update domain and one fault domain.

Things to know about update domains

An update domain is a group of nodes that are upgraded together during the process of a service upgrade (or *roll out*). An update domain allows Azure to perform incremental or rolling upgrades across a deployment. Here are some other characteristics of update domains.

- Each update domain contains a set of virtual machines and associated physical hardware that can be updated and rebooted at the same time.
- During planned maintenance, only one update domain is rebooted at a time.
- You can specify between 1 and 20 update domains when creating an availability set. If you don't specify a value, Azure defaults to five update domains.
- The update domain count is immutable after creation; to change it, you must delete and recreate the availability set.

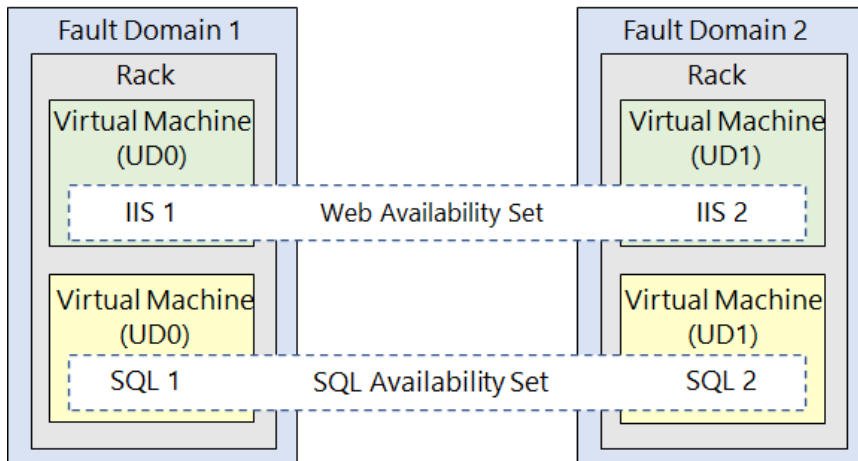
Things to know about fault domains

A fault domain is a group of nodes that represent a physical unit of failure. Think of a fault domain as nodes that belong to the same physical rack.

- A fault domain defines a group of virtual machines that share a common set of hardware (or *switches*) that share a single point of failure. An example is a server rack serviced by a set of power or networking switches.

- Two fault domains work together to mitigate against hardware failures, network outages, power interruptions, or software updates.

Let's look at a scenario with two fault domains that have two virtual machines each. The virtual machines in each fault domain are contained in different availability sets. The web availability set contains two virtual machines with one machine from each fault domain. The SQL availability set contains two different virtual machines with one from each fault domain.



5. Review availability zones

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/5-review-availability-zones>

Review availability zones

Completed

Availability zones are a high-availability offering that protects your applications and data from datacenter failures. You can use availability zones to build high-availability into your application architecture by colocating your compute, storage, networking, and data resources within a zone and replicating in other zones.

Consider a scenario where you create three or more virtual machines across three zones in an Azure region. Your virtual machines are effectively distributed across three fault domains and three update domains. The Azure platform recognizes this distribution across update domains to make sure that virtual machines in different zones aren't updated at the same time.

Things to know about availability zones

Review the following characteristics of availability zones.

- Availability zones are unique physical locations within an Azure region.
- Each zone is made up of one or more datacenters that are equipped with independent power, cooling, and networking.
- To ensure resiliency, there's a minimum of three separate zones in all enabled regions.
- The physical separation of availability zones within a region protects applications and data from datacenter failures.
- Zone-redundant services replicate your applications and data across availability zones to protect against single-points-of-failure.

Things to consider when using availability zones

Azure services that support availability zones are divided into two categories.

Category	Description	Examples
Zonal services	Azure <i>zonal</i> services pin each resource to a specific zone.	- Azure virtual machines - Azure managed disks
Zone-redundant services	For Azure services that are zone-redundant, the platform replicates automatically across all zones.	- Azure Storage that's zone-redundant - Azure SQL Database

Note

Note: Standard IP addresses can be configured as zone-redundant (recommended for high availability), zonal (pinned to a specific zone), or non-zonal (regional) depending on your deployment choice.

Tip

To achieve comprehensive business continuity on Azure, build your application architecture with a combination of availability zones and Azure [regional pairs](#).

6. Compare vertical and horizontal scaling

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/6-compare-vertical-horizontal-scaling>

Compare vertical and horizontal scaling

Completed

A robust virtual machine configuration includes support for scalability. Scalability allows throughput for a virtual machine in proportion to the availability of the associated hardware resources. A scalable virtual machine can handle increases in requests without adversely affecting response time and throughput. For most scaling operations, there are two implementation options: *vertical* and *horizontal*.

Things to know about vertical scaling

Vertical scaling, also known as *scale up and scale down*, involves increasing or decreasing the virtual machine **size** in response to a workload. Vertical scaling makes a virtual machine more (scale up) or less (scale down) powerful.

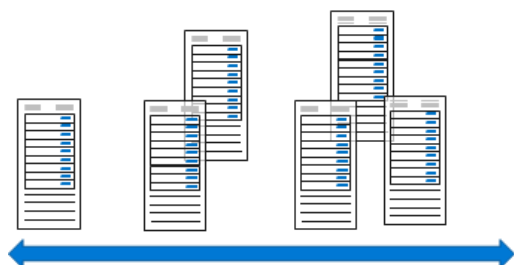


Here are some scenarios where using vertical scaling can be advantageous:

- If you have a service built on a virtual machine that's under-utilized such as on the weekend, you can use vertical scaling to decrease the virtual machine size and reduce your monthly costs.
- You can implement vertical scaling to increase your virtual machine size to support larger demand without having to create extra virtual machines.

Things to know about horizontal scaling

Horizontal scaling, also referred to as *scale out and scale in*, is used to adjust the **number** of virtual machines in your configuration to support the changing workload. When you implement horizontal scaling, there's an increase (scale out) or decrease (scale in) in the number of virtual machine instances.



Things to consider when using vertical and horizontal scaling

Review the following considerations regarding vertical and horizontal scaling. Think about which implementation might be required to support your company website.

- **Consider limitations.** Generally speaking, horizontal scaling has fewer limitations than vertical scaling. A vertical scaling implementation depends on the availability of larger hardware, which quickly hits an upper limit and can vary by region. Vertical scaling also usually requires a virtual machine to stop and restart, which can temporarily limit access to applications or data.
- **Consider flexibility.** When operating in the cloud, horizontal scaling is more flexible. A horizontal scaling implementation allows you to run potentially thousands of virtual machines to manage changes in workload and throughput.
- **Consider reprovisioning.** *Reprovisioning* is the process of removing an existing virtual machine and replacing it with a new machine. A robust availability plan considers where reprovisioning might be required and plans for interruptions to service. If reprovisioning might be required, determine if any data needs to be maintained and migrated to the new machine.

7. Implement Azure Virtual Machine Scale Sets

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/7-implement-scale-sets>

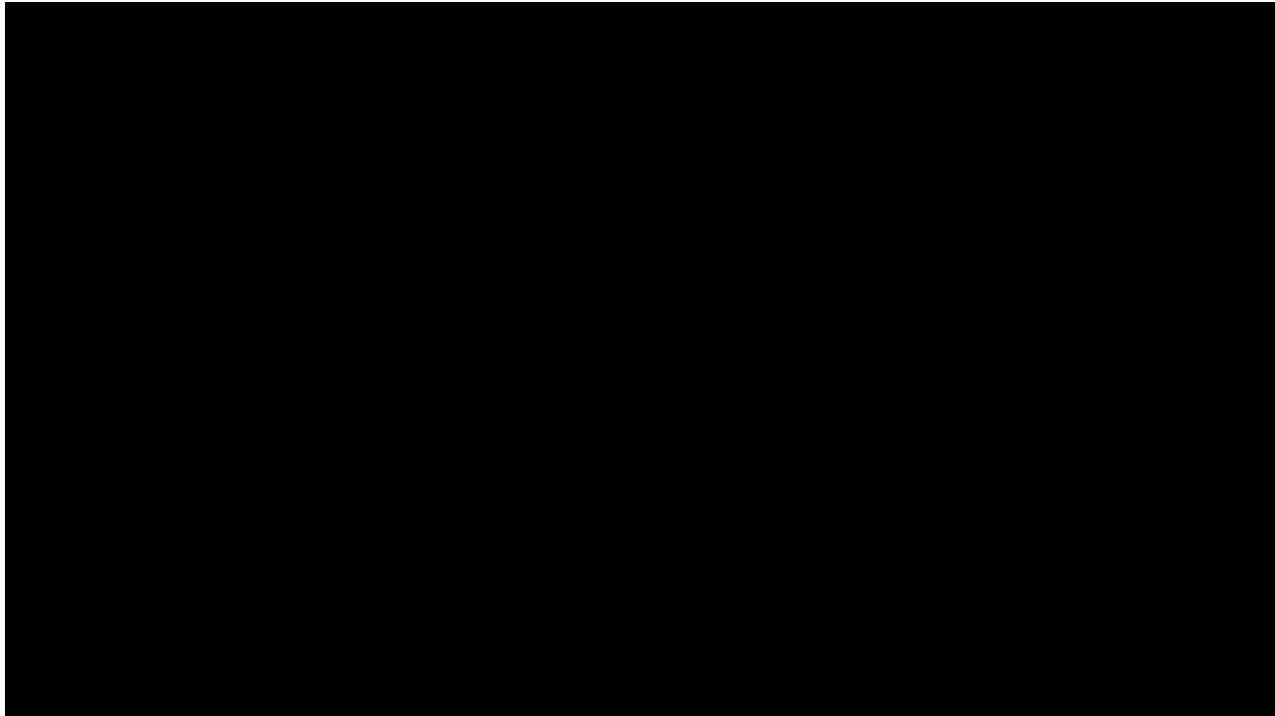
Implement Azure Virtual Machine Scale Sets

Completed

Azure Virtual Machine Scale Sets are an Azure Compute resource that you can use to deploy and manage a set of **identical** virtual machines. When you implement Virtual Machine Scale Sets and configure all your virtual machines in the same way, you gain true *autoscaling*. Virtual Machine Scale Sets automatically increases the number of your virtual machine instances as application demand increases, and reduces the number of machine instances as demand decreases.

With Virtual Machine Scale Sets, you don't need to pre-provision your virtual machines. It's easier to build large-scale services that target large compute, big data, and containerized workloads. As workloads increase, more virtual machine instances can be added. As workloads decrease, virtual machines instances can be removed. The process of adding and removing machines can be manual or automated, or a combination of both.

Increase app availability and scalability with Azure Virtual Machine Scale Sets



Things to know about Azure Virtual Machine Scale Sets

Review the following characteristics of Azure Virtual Machine Scale Sets.

- Virtual Machine Scale Sets support the use of Azure Load Balancer for basic layer-4 traffic distribution, and Azure Application Gateway for more advanced layer-7 traffic distribution and TLS/SSL termination.
- You can use Virtual Machine Scale Sets to run multiple instances of your application. If one of the virtual machine instances has a problem, customers continue to access your application through another virtual machine instance with minimal interruption.
- There are two types of orchestration modes available for Azure virtual machine scale sets: Uniform and Flexible. In **Uniform orchestration mode**, all virtual machine instances are created from the same base operating system image and configuration. In **Flexible orchestration mode**, VMs can use different images, sizes, or configurations within the same scale set. The orchestration mode must be chosen when the scale set is created.
- Customer demand for your application might change throughout the day or week. To meet customer demand, Virtual Machine Scale Sets implements autoscaling to automatically increase and decrease the number of virtual machines.

8. Create Virtual Machine Scale Sets

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/8-create-scale-sets>

Create Virtual Machine Scale Sets

Completed

You can implement Azure Virtual Machine Scale Sets in the Azure portal. You specify the number of virtual machines and their sizes, and indicate preferences for using Azure Spot instances, Azure managed disks, and allocation policies.

In the Azure portal, there are several settings to configure to create an Azure Virtual Machine Scale Sets implementation.

Create a virtual machine scale set ...

[Basics](#) [Disks](#) [Networking](#) [Scaling](#) [Management](#) [Health](#) [Advanced](#) [Tags](#) [Review + create](#)

Azure virtual machine scale sets let you create and manage a group of load balanced VMs. The number of VM instances can automatically increase or decrease in response to demand or a defined schedule. Scale sets provide high availability to your applications, and allow you to centrally manage, configure, and update a large number of VMs.

Orchestration

A scale set has a "scale set model" that defines the attributes of virtual machine instances (size, number of data disks, etc). As the number of instances in the scale set changes, new instances are added based on the scale set model.

Orchestration mode * ⓘ ☒ **Uniform:** optimized for large scale stateless workloads with identical instances
☐ **Flexible:** achieve high availability at scale with identical or multiple virtual

Instance details

Image * ⓘ

Ubuntu Server 20.04 LTS - x64 Gen2

[See all images](#) | [Configure VM generation](#)

VM architecture ⓘ ☐ Arm64
☒ x64

Run with Azure Spot discount ⓘ ☐

Size * ⓘ

Standard_D2s_v3 - 2 vcpus, 8 GiB memory

[See all sizes](#)

- **Orchestration mode:** Choose how the scale set manages the virtual machines. Flexible orchestration is the default and recommended mode for new deployments. For most new workloads, accept the Flexible default unless you have a specific requirement for identical instances.
- **Image:** Choose the base operating system or application for the VM.
- **VM Architecture:** Azure provides a choice of x64 or Arm64-based virtual machines to run your applications. x64-based VMs provide the most software compatibility. Arm64-based VMs provide up to 50% better price-performance than comparable x64 VMs.
- **Size:** Select a VM size to support the workload that you want to run. The size that you choose then determines factors such as processing power, memory, and storage capacity. Azure offers a wide variety of sizes to support many types of uses. Azure charges an hourly price based on the VM's size and operating system.

Under the **Advanced** tab, you can also select the following:

- **Spreading algorithm:** The spreading algorithm determines how VMs in the scale set are balanced across fault domains. With max spreading, VMs are spread across as many fault domains as possible in each zone. With fixed spreading, VMs are always spread across exactly five fault domains. In the case where fewer than five fault domains are available, a scale set using "Max spreading" completes, while a scale set using "Fixed spreading" fails. For this reason, Microsoft recommends using **Max spreading** for your implementation.

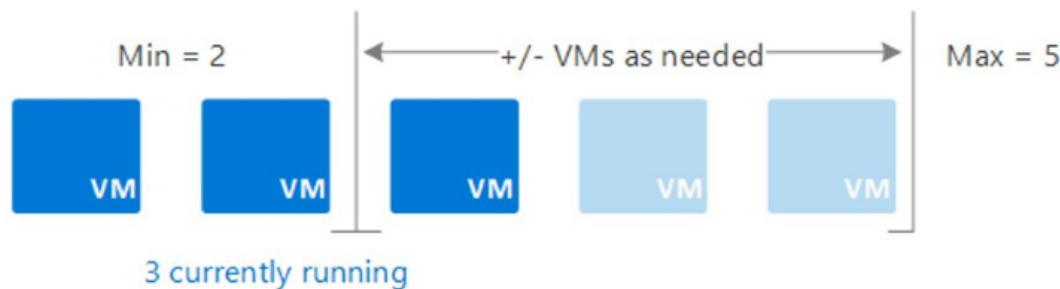
9. Implement autoscale

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/9-implement-autoscale>

Implement autoscale

Completed

An Azure Virtual Machine Scale Sets implementation can automatically increase or decrease the number of virtual machine instances that run your application. This process is known as *autoscaling*. Autoscaling allows you to dynamically scale your configuration to meet changing workload demands.



Autoscaling minimizes the number of unnecessary virtual machine instances that run your application when demand is low. Your customers continue to receive an acceptable level of performance as demand grows and more virtual machine instances are automatically added.

Things to consider when using autoscaling

Review the following considerations about autoscaling. Think about how this process can benefit your company website implementation.

- **Consider automatic adjusted capacity.** You can create autoscaling rules to define the acceptable performance for a positive customer experience. When the defined thresholds are met, the autoscale rules act to adjust the capacity of your Virtual Machine Scale Sets implementation.
- **Consider scale out.** If your application demand increases, the load on the virtual machine instances in your implementation increases. If the increased load is consistent, rather than a brief demand, you can configure autoscale rules to increase the number of virtual machine instances in your implementation.
- **Consider scale in.** On an evening or weekend, your application demand might decrease. If the decreased load is consistent over a period of time, you can configure autoscale rules to decrease the number of virtual machine instances in your implementation. The scale-in action reduces the cost to run your Virtual Machine Scale Sets implementation as you only run the number of instances required to meet the current demand.
- **Consider scheduled events.** You can implement autoscaling and schedule events to automatically increase or decrease the capacity of your implementation at fixed times.
- **Consider overhead.** Using Azure Virtual Machine Scale Sets with autoscaling reduces your management overhead to monitor and optimize the performance of your application.

10. Configure autoscale

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/10-configure-autoscale>

Configure autoscale

Completed

When you create an Azure Virtual Machine Scale Sets implementation in the Azure portal, you can enable manual or autoscaling. For optimal performance, you should define a minimum, maximum, and default number of virtual machine instances to use.

In the Azure portal, you can select the scaling mode.

Create a virtual machine scale set ...

Scaling

Scaling mode ⓘ

- ☒ **Manually update the capacity:** Maintain a fixed amount of instances.
- ☐ **Autoscaling:** Scaling based on a CPU metric, on any schedule.
- ☐ **No scaling profile:** manual attach virtual machines after deployment
-  Improve your availability by selecting multiple zones

Instance count * ⓘ

[Configure scaling options](#)

Scaling mode

- **Manually update the capacity:** Maintain a fixed instances count. Set the **Instance count** to the number of virtual machines in the scale set (0 - 1000). Configure the **Scale-in policy** which is the order virtual machines are selected for deletion. For example, you could balance across zones and then delete the virtual machine with the highest instance ID.
- **Autoscaling:** Scaling based on a CPU metric, or any schedule.

Configure autoscaling

Autoscaling is based on a scaling condition.

Add a scaling condition



Scale mode

- ☒ **Autoscaling:** Scaling based on a CPU metric, on any schedule

Default instance count * ⓘ

Instance limit

Minimum * ⓘ

Maximum * ⓘ

Scale out

CPU threshold greater than * ⓘ

Increase instance count by * ⓘ

Scale in

CPU threshold less than * ⓘ

Decrease instance count by * ⓘ

Query duration

Minutes * ⓘ

The engine will query CPU usage for the past 60 minutes before executing the scaling to avoid reacting to transient spikes.

Schedule

Schedule type *

- ☒ Specify start/end dates
- ☐ Repeat specific days

- **Default instance count.** The initial number of virtual machines deployed in this scale set (0-1000).
- **Instance limit.** The minimum instance count you want this condition to scale down to. The maximum instance count you want this condition to scale up to.

- **Scale out.** The CPU usage percentage threshold for triggering the scale-out autoscale rule. The number of instances to scale out by.
- **Scale in.** The CPU usage percentage threshold for triggering the scale-in autoscale rule. The number of instances to scale in by.
- **Query duration:** This duration is the time the Autoscale engine looks back for the metric usage average. This look back is to allow your metric to stabilize.
- **Schedule:** Specify the start and end dates. You can also repeat the schedule on specific days.

11. Module assessment

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/11-knowledge-check>

Module assessment

Completed

12. Summary and resources

<https://learn.microsoft.com/en-us/training/modules/configure-virtual-machine-availability/12-summary-resources>

Summary and resources

Completed

Azure provides several high availability options for virtual machines. You can achieve high availability by using availability sets, availability zones, and Azure Virtual Machine Scale Sets.

In this module, you learned how to configure virtual machine availability by using availability sets and availability zones with update and fault domains. You discovered how to autoscale virtual machines and configure vertical and horizontal scaling. You reviewed how to implement Virtual Machine Scale Sets, including storage resiliency and scalability options.

The main takeaways from this module are:

- Azure Virtual Machine Scale Sets allow for the deployment and management of a group of identical virtual machines, making it easier to build large-scale services.
- Autoscaling with Virtual Machine Scale Sets helps optimize performance by automatically adjusting the number of instances based on workload demands.
- Availability sets and availability zones are important features in Azure for achieving high availability and fault tolerance for virtual machines.

Learn more with Copilot

Copilot can assist you in configuring Azure infrastructure solutions. Copilot can compare, recommend, explain, and research products and services where you need more information. Open a Microsoft Edge browser and choose Copilot (top right) or navigate to copilot.microsoft.com. Take a few minutes to try these prompts and extend your learning with Copilot.

- How do virtual machine scale sets work with Azure availability zones and sets?
- What is the difference between manual and autoscaling of Azure virtual machines?

Learn more with documentation

- [Availability options for Azure Virtual Machines](#). This article provides an overview of the availability options for Azure virtual machines (VMs).
- [Autoscale with Azure Virtual Machine Scale Sets](#). This article reviews when use Virtual Machine Scale Sets.
- [Create virtual machines in a scale set using Azure portal](#). This article steps through using Azure portal to create a Virtual Machine Scale Set.

Learn more with self-paced training

- [Introduction to Azure Virtual Machines \(sandbox\)](#). Learn about the decisions you make before creating a virtual machine, the options to create and manage the VM, and the extensions and services you use to manage your VM.
- [Implement scale and high availability with Windows Server VM](#). You learn how to implement scaling for virtual machine scale sets and load-balanced VMs. You also learn how to implement Azure Site Recovery.
- [Introduction to Azure Virtual Machine Scale Sets](#). Learn about what Azure Virtual Machine Scale Sets do, how they work, and when you should use Azure Virtual Machine Scale Sets as a solution for your organization.