# On-Demand Capacity Reservation Private Preview

Welcome to the On-Demand Capacity Reservation Private Preview!

This document is to briefly describe the feature, provide onboarding steps for the preview program, and to share Azure Resource Manager (ARM) templates that can be used for testing purposes.

**Note:**

* The feature is currently in Private Preview phase. That means, customers should only use it for testing purposes. It is not recommended to use this feature at this point for any mission-critical workloads.
* The feature is currently supported via ARM templates only. Other client tools support such as Azure Portal, CLI, PS etc. will be coming later.
* During Private Preview, production SLAs will not be enforced.

**Billing**

During Private Preview, you will be charged for the virtual machines that are attached to a capacity reservation. You will also be charged for the unused capacity reservation spots. For more information on billing, please read “**VM Capacity Reservations: Usage and Billing”** section described later in the document.

## Provide Feedback

We would love to know what we can do to make Capacity Reservation better meet your needs. For any feedback, please fill [this](https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR6sD4u5Jlg1JlaZfjElD4-9UOFM5NDc5UUFRQjlBS1hOU1JSWUU3Tks2Si4u) form. Alternatively, you can reach out to the people below:

PM contacts:

* Varun Gupta <[gupta.varun@microsoft.com](mailto:gupta.varun@microsoft.com)>, Program Manager, Compute
* Bill DeForeest <bill.deforeest@microsoft.com>, Principal Program Manager, Compute

## On-Demand Capacity Reservation

Capacity Reservation allows customers to reserve compute capacity for the supported VM sizes in an Azure Region or an Availability Zone. It is a customer facing object that can be deployed using ARM APIs.

A Capacity Reservation has some basic properties always defined at creation:

* Each reservation is for one VM size (for example, Standard\_D2s\_v3).
* Each reservation is for one location (region). If that location has availability zones, then the reservation also may specify exactly one of those zones.
* Each reservation has a quantity of instances to be reserved.

To create a Capacity Reservation, these parameters are passed to Azure as a capacity request. If the subscription lacks the required quota or Azure does not have capacity available that meets the specification, the Capacity Reservation will fail to deploy. Either request more quota or try a different VM size/location/zone combination.

Once Azure accepts a reservation request, capacity reservation receives the applicable Azure availability SLA for use with virtual machines. The SLA will be defined later and will not be enforced during private preview.

Virtual Machines can now be allocated to the reservation.

* Up to the limit of the reserved quantity, no quota checks are required.
* For each instance of the reserved quantity that is consumed, Azure will emit usage incurred by the allocated VM.
* For each instance of the reserved quantity that is unused, Azure will emit VM usage for the reserved size. This is because capacity is being consumed by the reservation itself.

**Working with Capacity Reservations**

Capacity Reservations should be used to create the minimum required capacity for an application or workload to meet runtime requirements such as throughput or response time. Overprovisioning capacity reservations will increase the monthly bill without providing any real benefit.

Capacity Reservations do not create limits on VM deployments. Azure supports associating as many VMs as desired against a Capacity Reservation. Up to the reserved quantity, quota checks are omitted. Allocating more VMs against the reservation is subject to quota checks and Azure fulfilling the additional capacity. Once deployed, these additional VM instances can cause the quantity of VMs allocated against the reservation to exceed the reserved quantity. This is called “overallocation”.

One benefit of this design is that VMs do not have to be managed separately across reserved and unreserved capacity. When an application is scaled down from temporary peak usage, VMs can be removed by whichever rules are desired. Manage only the total instances reserved and the total allocated. When the two values are equal, your deployment is at the minimum usage.

The quantity reserved can be adjusted after initial deployment by changing the “*capacity”* property of the Capacity Reservation. If preferred, first add additional VMs and then adjust the reserved quantity to match. Azure also supports adding VMs that match the reservation and were allocated before the reservation was created.

In cases where the goal is to reserve capacity for just a specific set of VMs and no others, then only associate the specific set of critical VMs with the Capacity Reservation. If you add a new critical VM, adjust the quantity reserved to gain assurance. Deploy the other VMs normally without the Capacity Reservation property.

Other changes to a Capacity Reservation, such as a different VM size or location, are not permitted. The recommended approach is to delete the existing reservation and create a new reservation with the new requirements.

Capacity Reservation is always created as part of a Capacity Reservation group. This is to group several Capacity Reservations together in a single group for a multi-tier application/workload. As we add multi-VM deployment support for VMSS in the coming weeks, reservations across zones can also be grouped together

**Capacity Reservation Lifecycle**

When a Capacity Reservation is created, Azure will set aside the requested number of capacity instances in the specified location:

Graphical user interface

Description automatically generated

The capacity instances are not directly visible or addressable. Track the state of the overall Capacity Reservation through its “*capacity”* and “*virutalMachinesAllocated”* properties.

**Key properties of a Capacity Reservation**

|  |  |
| --- | --- |
| **Property** | **Description** |
| location | Azure region or zone in which the Capacity Reservation is created |
| sku | VM size |
| capacity | Total quantity of VMs reserved by the customer |
| virtualMachinesAllocated | List of VMs allocated against the Capacity Reservation. Length of this list gives the total quantity of capacity instances that are consumed. |
| virtualMachinesAssociated | List of VMs associated with the Capacity Reservation. The difference between virtualMachinesAssociatedand virutalMachinesAllocatedis that the latter doesn’t include the VMs that are in the deallocated state. However, the former does include such VMs. |

The above example will start with “*capacity”* as 2 and length of “*virutalMachinesAllocated” and “virutalMachinesAssociated”* as 0.

In order to consume Capacity Reservation, the VMs must be directly targeted to it. This gives customers the flexibility to only target their critical workloads against the reservation and non-critical workloads can go through regular allocation path.

When a VM is subsequently deployed against the Capacity Reservation, it will logically consume one of the reserved capacity instances:

Graphical user interface

Description automatically generated

The status of the Capacity Reservation will now show “*capacity”* as 2 and length of “*virutalMachinesAllocated” and “virutalMachinesAssociated”* as 1.

Allocations against the Capacity Reservation will succeed as along as the VMs have matching properties (location, size, etc.) and there is at least one empty capacity instance.

Using our example, when a third VM is allocated against the Capacity Reservation, the reservation enters the “overallocated” state. This third VM will require unused quota and additional capacity fulfillment from Azure. Once the third VM is allocated, the Capacity Reservation now looks like this:

Graphical user interface, application

Description automatically generated

The “*capacity”* is 2 and the length of “*virutalMachinesAllocated” and “virutalMachinesAssociated”* is 3. One advantage of this capability to over-allocate is that scale out will never be blocked just because the limit of reserved capacity is reached.

Another advantage is automatic minimum cost. To illustrate this, suppose the application scales down to the minimum two VMs. Since VM 0 needs an update, that VM is chosen for deallocation. The reservation automatically shifts to this state.

Graphical user interface

Description automatically generated

The “*capacity”* and the length of “*virutalMachinesAllocated”* are both 2. The length of *“virutalMachinesAssociated”* is 3 as VM 0, though deallocated, is still associated to the capacity reservation. In this state, the reservation has reached minimum VM usage as quantity reserved = length of *“virutalMachinesAllocated”*. In other words, Azure will always consume available reserved capacity first without any special management.

The Capacity Reservation will exist until explicitly deleted. To delete Capacity Reservation, first step is to dissociate all the VMs in the “*virtualMachinesAssociated”* property. Once this is done, the Capacity Reservation should look like this:

Graphical user interface

Description automatically generated

The status of the Capacity Reservation will now show “*capacity”* as 2 and length of “*virtualMachinesAssociated”* and “*virutalMachinesAllocated”* as 0.

From this state, the Capacity Reservation can be deleted. Once deleted, customers will not pay for the reservation anymore.

Graphical user interface

Description automatically generated

**VM Capacity Reservations: Usage and Billing**

When a capacity instance is empty, VM usage will be reported for the corresponding VM size and location. VM [Reserved Instances](https://azure.microsoft.com/en-us/pricing/reserved-vm-instances/) can cover some or all the Capacity Reservation usage even when VMs are not deployed. Depending on how long the Capacity Reservation is needed, pay-as-you-go might be a better choice for some situations instead of a 1 or 3-year commitment.

Continuing with our example, a Capacity Reservation with 2 Capacity Instances has been created. The subscription has access to one matching Reserved VM Instance of the same size. The result is two usage streams for the Capacity Reservation, one of which is covered by the Reserved Instance:

A screenshot of a computer

Description automatically generated with medium confidence

Since the customer has already paid for Reserved Instance, one of the usage streams for Capacity Reservation will be zeroed out. In the above example, the customer will pay for just one instance of capacity.

When a VM is allocated against the Capacity Reservation, the additional VM components (disks, network, extensions, and any other requested components) must also be allocated. Now the VM usage will reflect one allocated VM and one unused capacity instance. The Reserved VM Instance will zero out the cost of either the VM or the unused capacity instance. The other charges associated with the allocated VM will also appear on the bill. Here is an example:

A screenshot of a computer

Description automatically generated with medium confidence

**Important**

Capacity Reservations for virtual machines do not change Azure’s usage rules for other components. Disks or other assets associated with VMs must still be managed separately. In the above example, the managed OS disk for VM 0 will incur usage charges if it exists even when VM 0 is not allocated.

## Confirm that your subscription has access to creating Capacity Reservation

Confirm that subscription was successfully registered for Capacity Reservation by performing the following steps:

* 1. Login to the Azure CLI and run “az login”
  2. You can check the registration status by running following commands:
* az account set -s “SUBSCRIPTION\_NAME\_OR\_ID”
* az feature show --namespace microsoft.compute –name CapacityReservationPreview

*{*

*"id": "/subscriptions/<SubscriptionId>/providers/Microsoft.Features/providers/Microsoft.Compute/features/CapacityReservation",*

*"name": "Microsoft.Compute/CapacityReservationPreview",*

*"properties": {*

*"state": "Registered"*

*},*

*"type": "Microsoft.Features/providers/features"*

*}*

If the “state” says “NotRegistered”, please reach out to the PM contacts above.

## Scenarios supported in Private Preview

Below is the list of supported scenarios in Private preview:

1. Create a single zone or a multi-zone Capacity Reservation Group
2. Create Capacity Reservation(s) for the supported VM sizes inside a Capacity Reservation Group
3. Update reserved quantity for an existing Capacity Reservation
4. Delete Capacity Reservation and Capacity Reservation Group
5. Associate VM to a Capacity Reservation Group
6. Dissociate VM from a Capacity Reservation Group
7. Associate VMSS to a Capacity Reservation Group (ETA ~ 04/01)
8. Dissociate VMSS from a Capacity Reservation Group (ETA ~ 04/01)

**Important**

During the preview, new deployments that are targeted to the Capacity Reservation at the time of allocation will be supported. If customers want to attach an existing VM/VMSS to the Capacity Reservation, they have to deallocate and reallocate that VM/VMSS and target it to Capacity Reservation at the time of allocation. As we roll out more fixes, the need to deallocate and reallocate would disappear and old VM/VMSS should just be compatible with Capacity Reservation without having to deallocate.

**Supported Regions**

During the preview, the only regions that are supported for capacity reservation are **South Central US** and **West Central US**.

**Support for Virtual Machine Scale Sets**

Virtual Machines Scale Sets should be supported in Private Preview starting 04/01. Please expect a notification from Azure when VMSS is supported.

## Capacity Reservation APIs and ARM templates

* [Here](https://github.com/Azure/on-demand-capacity-reservation/blob/main/On-Demand%20Capacity%20Reservation-APIs.docx) is the link for Capacity Reservation APIs details.
* [Here](https://github.com/Azure/on-demand-capacity-reservation/blob/main/Sample%20ARM%20Template.zip) is the link for sample ARM templates to deploy Capacity Reservation.

## Capacity Reservation FAQ

1. **Does Capacity Reservation provide capacity guarantee?**

Yes, Capacity Reservation provides guarantee that capacity is exclusively reserved for the customer. However, in case of a big outage, it is possible that Azure is not able to make the capacity available. In that case, customer is eligible for refunds/credits as per our forthcoming SLA. Please note that SLA is not enforced during Private Preview.

1. **How much does Capacity Reservation cost?**

Capacity Reservation is charged at your negotiated PAYG rates. Any available Reserved Instances discount are automatically applied to the Capacity Reservation by our billing system.

1. **Do I need quota to create Capacity Reservation?**

Yes, you need quota to be able to create Capacity Reservation. To request more quota, please go to [this](https://docs.microsoft.com/en-us/azure/azure-portal/supportability/per-vm-quota-requests) link.

1. **Can I create a time bound Capacity Reservation?**

No, it is not possible. Capacity Reservation is created at a given point in time and will remain active until deleted by the customer.

1. **Is there any extra cost to use Capacity Reservation once it’s fully consumed by virtual machines?**

No, there is no extra cost to have Capacity Reservation once it’s fully consumed by virtual machines.

1. **What VM sizes are supported in Private Preview?**

Av2, B, D, E, and F series VM sizes are supported in Private Preview. Additional series will become available during public preview.

1. **What regions are supported during private preview?**

During the preview, only South-Central US and West Central US are supported. Other regions support will be rolled out in the future.

1. **What other deployment constructs are supported?**

Virtual Machine Scale Set will be supported very soon. VMSS Flex is not supported in Private preview but will likely be supported during public preview.

## Restrictions

Capacity reservation is fundamentally about setting aside VM capacity. Azure will do so in a way that supports many deployment scenarios. However, the capacity availability SLA does not apply to every possible VM placement option that could be requested. Azure will reject allocation against the Capacity Reservation for the following deployment scenarios:

1. Proximity Placement Groups
2. Availability Set
3. Update domains

At the time of VM allocation, customers can specify the number, size(s), and storage type of the VM managed disk(s). Storage capacity for disks is not typically an issue, but the exact desired configuration of disk capacity does fall outside the VM availability SLA. Including managed disk configuration in an on-demand capacity reservation is under investigation as a future enhancement.

In addition to the above, the following exclusions will also not deploy:

1. A series VMs do not support Capacity Reservations. Microsoft recommends migrating to Av2 or Dv3 series VMs to take advantage of Capacity Reservations and optionally Reserved Instance discounts.
2. Any VM size which uses InfiniBand networking is not supported at this time. Microsoft expects to add support for these sizes in the future.
3. “Fixed FD” count on multi-zone VMSS Scale Set mode (use “max spread” option instead).

### What’s coming up in Public Preview

Currently scheduled for Q2 CY2021, Public preview is expected to have the following enhancements:

* Support for specialized workloads such as HPC and GPU
* Create/Manage Capacity Reservation through Azure CLI /Power shell
* Create/Manage Capacity Reservation through Azure Portal
* Use Capacity Reservation for VMSS Flex