Keystone STaaS services

Keystone

NetApp
January 20, 2026

Table of Contents

Keystone STaaS services . . . 1
- Metrics and definitions used in Keystone . . . 1
  - Metrics measurement . . . 1
- Storage QoS in Keystone . . . 2
  - Adaptive QoS . . . 2
  - Adaptive QoS settings . . . 2
  - Configuration of adaptive QoS policy group . . . 3
  - Settings of adaptive QoS policies . . . 3
  - Block size calculation . . . 5
- Supported storage in Keystone . . . 6
  - Services for file, block, and object storage . . . 6
  - Services for cloud storage . . . 8
- Supported storage capacities in Keystone . . . 8
  - Logical capacity . . . 9
  - Committed capacity . . . 9
  - Consumed capacity . . . 9
  - Burst capacity . . . 9
  - Billed capacity . . . 10
- Performance service levels in Keystone . . . 10
  - Performance service levels for file and block storage . . . 10
  - Object storage . . . 11
  - Cloud storage . . . 12
- Capacity requirements for Keystone performance service levels . . . 12
  - Minimum capacity requirements for file and block services . . . 12
  - Minimum capacity requirements for object storage . . . 13
  - Capacity adjustments . . . 13

Keystone STaaS services

Metrics and definitions used in Keystone

The NetApp Keystone STaaS service uses several terms to measure metrics. You might want to learn more about these terms as you use Keystone.

The following terms and definitions are used within the Keystone STaaS service to measure metrics:

- Capacity: Measured in GiB, TiB, and PiB.
- IO density: IOPS/TiB: Number of input/output operations processed per second based on the total space that is being consumed by the workload, in tebibytes.
- Service availability
- Durability in accurate data access
- Latency and speed

Metrics measurement

- **Capacity measurement in GiB, TiB, and PiB**: Measurements of data storage capacity using base of 1024 (1 GiB = 1024^3 bytes, 1 TiB = 1024^4 bytes, and 1 PiB = 1024^5 bytes).
- **Operations counter chart in IOPS/TiB**: The protocol operations per second, requested by the application, divided by the size of the volume used by workloads.
- **Availability**: Measured as a percentage of the number of I/O requests successfully responded to by the service, divided by total number of I/O requests made to the service. This is measured at the service demarcation in a month and does not include the scheduled service downtime or unavailability of the facilities, network, or other services provided by the customer.
- **Durability**: Percentage of data accessed without loss of fidelity, excluding customer-caused deletion or corruption.
- **Latency**: Time to service an I/O request received from a client, measured at the service demarcation (storage controller I/O port).

Throughput performance metrics

Throughput performance metrics are applicable only for file and block services based on:

- 32 KB block sizes
- 70% read/30% write I/O mix

Variations in IO density

IO density calculated in IOPS/TiB and/or MBps/TiB varies based on the following factors:

- Workload characteristics
- Latency, excluding the following:
  - Application latency
  - Host latency

    - Latency in the customer network while transferring data to and from the controller ports
    - Overhead latency associated with data transfer to the object store in the case of FabricPool
    - The latency automatically applied by the QoS to keep IO within service level maximums
- The user and Snapshot copy data that is counted as part of the used capacity
- The allocated absolute minimum IOPS on each ONTAP volume, regardless of the amount of data in the volume:
    - Extreme: 1,000 IOPS
    - Premium: 500 IOPS
    - Performance, Standard, and Value: 75 IOPS
- While using the Advanced Data Protection add-on services, the target latency applies only to servicing IO requests from the local storage.

Volume AQoS

Each ONTAP volume should have the applicable adaptive quality of service (AQoS) policy applied. Otherwise, the capacity within each volume that does not have an AQoS policy applied is billed at the rate of the highest Service Level.

Storage QoS in Keystone

Keystone uses storage quality of service (QoS) to ensure that applications obtain consistent and predictable performance. Without QoS, certain workloads, such as those for booting of multiple systems, might consume most or all of the resources for a period of time and affect other workloads.

For information about QoS, see Guarantee throughput with QoS overview.

Adaptive QoS

Adaptive QoS (AQoS) is used by Keystone services to dynamically maintain the IOPS/TiB ratio based on the volume size. For information about AQoS policies, see About adaptive QoS.

Keystone provides you with AQoS polices that you can set up once your cluster is in production. You should ensure that all your volumes are associated with the correct AQoS policies that are already created and available in your system.

An ONTAP volume is non-compliant if it does not have an AQoS policy applied. A volume without a QoS policy is the last on the list of priority for the system to provide any available input-output operations. However, if any input-output operations are available, then the volume could consume all available IOs.

If you have not applied AQoS policies to your volumes, those volumes will be measured and charged at the highest service level as per your subscription. This may result in unintended burst charges.

Adaptive QoS settings

The Adaptive QoS (AQoS) settings vary with service levels.

Policy name	Extreme	Premium	Performance	Standard	Value
Expected IOPS/TiB	6,144	2,048	1,024	256	64
Expected IOPS Allocation	Allocated space				
Peak IOPS/TiB	12,288	4,096	2,048	512	128
Peak IOPS Allocation	Used space				
Block Size	32K				

Configuration of adaptive QoS policy group

You can configure adaptive QoS (AQoS) policies to automatically scale a throughput ceiling or floor to volume size. Not all Keystone service levels are aligned with the default ONTAP QoS policies. You can create custom QoS policies for them. For configuring a policy, you should be aware of the following:

- **Policy group name**: The name of the AQoS policy group. For example, `Keystone_extreme`.
- **VServer**: The name of the VServer or storage VM (storage virtual machine).
- **Expected IOPS/TiB**: The minimum number of IOPS, per allocated TiB per volume, that the system attempts to provide when enough system IOPS are available.
- **Peak IOPS/TiB**: The maximum number of IOPS, per used TiB per volume, that the system allows the volume to reach before it throttles the IOPS through injection of latency.
- **Expected IOPS allocation**: This parameter controls whether the expected IOPS available to the volume is based on the allocated or used size of the volume. In Keystone, this is based on the allocated space.
- **Peak IOPS allocation**: This parameter controls whether the peak IOPS available to the volume is based on the allocated or used size of the volume. In Keystone, this is based on the used space.
- **Absolute minimum IOPS**: The lowest number of expected IOPS that will be applied to a volume if the volume size is very small and would otherwise result in an unacceptable number of IOPS. This value defaults to 1,000 for `Extreme`, 500 for `Premium`, and 250 for `Performance`, and 75 for `Standard` and `Value` service levels.

> This is not IOPS density (for example, 75 IOPS/TiB), but an absolute minimum number of IOPS.

For information about IO density, see Metrics and definitions used in Keystone Services. For more information about AQoS policy groups, see Use adaptive QoS policy groups.

Settings of adaptive QoS policies

The settings for adaptive QoS (AQoS) policies for each service level are described in the following sections. The minimum and maximum volume sizes for each service level provided here allow for optimal IOPs and latency values for a volume. Creating too many volumes outside of these guidelines may negatively impact performance in those volumes.

Settings for Extreme service level

Settings and commands for the Extreme service level:

- Sample command:

```
qos adaptive-policy-group create -policy-group <Keystone_extreme> -vserver
<SVM_name> -expected-iops 6144 -peak-iops 12288 -expected-iops-allocation
allocated-space -peak-iops-allocation used-space -block-size 32K -absolute
-min-iops 1000
```

- Recommended minimum volume size: 100GiB, 0.1TiB
- Recommended maximum volume size: 10TiB

Settings for Premium service level

Settings and commands for the Premium service level:

- Sample command:

```
qos adaptive-policy-group create -policy-group <Keystone_premium> -vserver
<SVM_name> -expected-iops 2048 -peak-iops 4096 -expected-iops-allocation
allocated-space -peak-iops-allocation used-space -block-size 32K -absolute
-min-iops 500
```

- Recommended minimum volume size: 500GiB, 0.5TiB
- Recommended maximum volume size: 50TiB

Settings for Performance service level

Settings and commands for the Performance service level:

- Sample command:

```
qos adaptive-policy-group create -policy-group <Keystone_performance>
-vserver <SVM_name> -expected-iops 1024 -peak-iops 2048 -expected-iops
-allocation allocated-space -peak-iops-allocation used-space -block-size
32K -absolute-min-iops 250
```

- Recommended minimum volume size: 500GiB, 0.5TiB
- Recommended maximum volume size: 80TiB

Settings for Standard service level

Settings and commands for the Standard service level:

- Sample command:

```
qos adaptive-policy-group create -policy-group <Keystone_standard>
-vserver <SVM_name> -expected-iops 256 -peak-iops 512 -expected-iops
-allocation allocated-space -peak-iops-allocation used-space -block-size
32K -absolute-min-iops 75
```

- Recommended minimum volume size: 1TiB
- Recommended maximum volume size: 100TiB

Settings for Value service level

Settings and commands for the Value service level:

- Sample command:

```
qos adaptive-policy-group create -policy-group <Keystone_value> -vserver
<SVM_name> -expected-iops 64 -peak-iops 128 -expected-iops-allocation
allocated-space -peak-iops-allocation used-space -block-size 32K -absolute
-min-iops 75
```

- Recommended minimum volume size: 1TiB
- Recommended maximum volume size: 100TiB

Block size calculation

Note these points before you calculate the block size by using these settings:

- IOPS/TiB = MBps/TiB divided by (block size * 1024)
- Block size is in KB/IO
- TiB = 1024GiB; GiB = 1024MiB; MiB = 1024KiB; KiB = 1024Bytes; as per base 2
- TB = 1000GB; GB = 1000MB; MB = 1000KB; KB = 1000Bytes; as per base 10

Sample block size calculation

To calculate the throughput for a service level, for example `Extreme` service level:

- Maximum IOPS: 12,288
- Block size per I/O: 32KB
- Maximum throughput = (12288 * 32 * 1024) / (1024*1024) = 384MBps/TiB

If a volume has 700GiB of logical used data, the available throughput will be:

```
Maximum throughput = 384 * 0.7 = 268.8MBps
```

Supported storage in Keystone

Keystone STaaS services support file and block storage with ONTAP, object storage with StorageGRID, and Cloud Volumes ONTAP.

Keystone STaaS provides standard and optional services for your storage.

Keystone STaaS standard services: Standard services are included within the base subscription and are not charged separately.

Keystone STaaS add-on services: These are optional, chargeable services that provide additional utilities and benefits on top of standard Keystone STaaS subscription services.

Keystone STaaS services can be used at the same time. For example, a cloud storage subscription can have the same term as with file, block, and object storage subscriptions. A cloud service can be included at any point during the service term of an existing storage subscription. However, if you do not plan to renew an existing file, block, and object subscription, a cloud storage subscription cannot be added during the last 90 days of the subscription.

Services for file, block, and object storage

Keystone STaaS services for ONTAP file and block storage, and StorageGRID object storage, support multiple features and protocols, and described in the following table:

Storage	Platform	Protocols	Supported features
File storage	ONTAP	NFS and CIFS	Supported ONTAP features: • FlexVol • FlexGroup • Snapshot copies • SnapMirror (Asynchronous) • SnapVault • SnapLock Enterprise • FabricPool/Cloud tiering • SnapRestore • FlexClone • SnapCenter (license is included but is not a part of Keystone services, and management is not guaranteed) • Autonomous ransomware protection[1]

Storage	Platform	Protocols	Supported features
Block storage	ONTAP	FC and iSCSI	Supported ONTAP features: • FlexVol • FlexGroup • Snapshot copies • SnapMirror (Asynchronous) • SnapVault • SnapLock Enterprise • FabricPool/Cloud tiering • SnapRestore • FlexClone • SnapCenter (license is included but is not a part of Keystone services, and management is not guaranteed)
Object storage	StorageGRID	S3	Supports multiple information lifecycle management (ILM) policies across multiple sites[2]

[1] For information about ransomware protection in ONTAP, see Autonomous Ransomware Protection.
[2] Each site requires a separate subscription.

Services for cloud storage

Keystone STaaS provides cloud storage services. Keystone STaaS supports Cloud Volumes ONTAP data management capabilities on Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform.

Hyperscalar-based compute, storage, and network services required by Cloud Volumes ONTAP are not provided by NetApp as a part of Keystone STaaS subscriptions; these subscriptions need to be procured directly from hyperscalar cloud service providers.

Supported storage capacities in Keystone

The NetApp Keystone STaaS service supports several types of storage capacities. Understanding these different capacity terms can help as you use Keystone.

Logical capacity

This is the data placed on the Keystone infrastructure by a customer. All Keystone capacities refer to a logical capacity.
For example, if a 1 TiB file is stored on the Keystone infrastructure then a minimum of 1 TiB of capacity should be purchased.

Committed capacity

The minimum logical capacity billed each month during the subscription:

- Capacity is committed to each performance service level.
- Committed capacity and additional service levels can be added during the term.

Changes to committed capacity

During the tenure of a subscription, you can change the committed capacities. However, there are certain preconditions:

- The committed capacity can be decreased based on certain conditions. For information, see Capacity reduction.
- The committed capacity cannot be increased 90 days prior to the expiry of your subscription, unless the subscription is to be renewed for an additional 12-month term.
- You can request changes to committed capacity through the Console or from your Keystone Success Manager (KSM).
  For information about requesting changes, see NetApp Keystone support.

Consumed capacity

Consumed capacity refers to the capacity (in TiB of storage) currently being consumed on the service. It is the sum of:

- The logical capacity used to store all instances and types of user data (like copies, mirrored copies, and versions).
- The logical capacity used to store clone volumes that are more than 10% of the size of the original volume.
- The physical capacity used to store the differential data from Snapshot copies.
- The allocated physical capacity.

Burst capacity

The NetApp Keystone service enables you to use additional capacity on top of the committed capacity for a service level. This is referred to as the burst capacity usage.

Note these points:

- Burst capacity is agreed upon in the Keystone agreement. It is usually set up to 20% above the committed capacity, and is charged at the same rate as the committed capacity.
- Burst capacity can be consumed on an elastic basis and is charged on a daily basis for the consumed average.

Billed capacity

Monthly bill = (committed capacity [TiB] * committed rate [$/TiB]) + (daily average provisioned burst capacity [TiB] * burst rate [$/TiB]). The monthly bill contains a minimum charge based on the committed capacity.

The monthly bill varies beyond the minimum charge based on daily average burst capacity consumption.

Performance service levels in Keystone

Keystone STaaS offers data storage capacity at pre-defined performance service levels. Each volume managed by Keystone services is associated with a performance service level.

A subscription can have multiple rate plans and each rate plan corresponds to a performance service level. Each rate plan has a committed capacity per performance service level.

Each performance service level is defined by its I/O density, that is IOPS/TiB/volume. This is the ratio of performance (input/output operations per second [IOPS]) and used storage capacity (TiB) which is IOPS/TiB at average latency per volume.

You select performance service levels based on your storage environment, and storage and consumption needs. The base performance service levels are available for you by default. Specific performance service levels are additionally available, if you have opted for add-on services. For example, for the advanced data protection add-on service, the *Advanced Data-Protect* performance service level is assigned to your subscription.

A detailed service description for NetApp Keystone STaaS performance service levels is available here.

The base performance service levels for the supported storage types, file, block, object, and cloud services are described in the following sections:

Performance service levels for file and block storage

Supported protocols: NFS, CIFS, iSCSI, and FC

Perfromance service level	Extreme	Premium	Performance	Standard	Value
Sample workload types	Analytics, databases, mission-critical apps	VDI, VSI, software development	OLTP, OLAP, containers, software development	File shares, web servers	Backup
Maximum IOPS/logical TiBs stored per volume	12,288	4,096	2,048	512	128
Maximum IOPS/logical TiBs allocated per volume	6,144	2,048	1,024	256	64

Maximum MBps/logical TiBs stored per volume @ 32K B/S	384	128	64	16	4
Target 90th percentile latency	<1 ms	<2 ms	<4 ms	<4 ms	<17 ms
Block size	32K				
Committed and metered capacity type	Logical				

More on performance service levels for file and block storage

The base performance service level metrics depend on the following conditions:

- The performance service levels for file and block storage support ONTAP 9.7 and later.
- IOPS/TiB/volume, MBps/TiB/volume, and latency values for performance service levels are based on the amount of data stored in the volume, 32KB block size, and a random combination of 70% read and 30% write IO operations.
- Actual IOPS/TiB/volume and MBps/TiB/volume may vary based on the actual or assumed block size, system workload concurrency, or input-output operations.
- Latency does not include the following:
  - application or host latency
  - customer network latency to or from the controller ports
  - overheads associated with the data transfer to the object store in case of FabricPool
  - latency automatically applied by QoS to keep IO within performance service level maximums
- Latency values are not applicable to MetroCluster write operations. These write operations are dependent on the distance of remote systems.
- If one or more volumes on a storage system do not have an AQoS policy assigned, then these volumes are considered as non-compliant volumes, and no target performance service levels are applicable for those systems.
- *Expected IOPS* is targeted for FabricPool only if the tiering policy is set to "none" and no blocks are in the cloud. *Expected IOPS* is targeted for volumes that are not in a SnapMirror synchronous relationship.
- Workload IO operations need to be balanced across all deployed controllers, as determined by the Keystone order.

Object storage

Supported protocol: S3

Performance service level	Object
Workload type	Media repository, archiving
Maximum IOPS/logical TiB stored per volume	N/A

Maximum MBps/logical TiB stored per volume	N/A
Average Latency	N/A
Committed and metered capacity type	Physical

Latency does not include overheads associated with data transfer to the object store in case of FabricPool storage.

Cloud storage

Supported protocol: NFS, CIFS, iSCSI, and S3 (AWS and Azure only)

Performance service level	Cloud Volumes ONTAP
Workload type	Disaster Recovery, software development/testing, business apps
Maximum IOPS/logical TiB stored per volume	N/A
Maximum MBps/logical TiB stored per volume	N/A
Average Latency	N/A

- Cloud native services, such as compute, storage, networking, are invoiced by cloud providers.
- These services are dependent on cloud storage and compute characteristics.

Related information

- Supported storage capacities
- Metrics and definitions used in Keystone services
- Quality of Service (QoS) in Keystone
- Keystone pricing

Capacity requirements for Keystone performance service levels

The capacity requirements for Keystone STaaS performance service levels differ among the file, block, object, and cloud storage offerings supported by the Keystone STaaS subscription.

Minimum capacity requirements for file and block services

The minimum capacity and incremental capacity allowed per subscription is described in the following table. The minimum capacity per performance service level is defined to be the same across Keystone sales motions. The capacity above the minimum capacity either at the beginning of the subscription, or as an add-on service to the subscription, or after reallocation during the subscription is also structured in the table.

Capacity	Extreme	Premium	Performance	Standard	Value

Minimum capacity [in TiB]	25	100
Incremental capacity (and in multiples) allowed at start of subscription [in TiB]	25	25
Incremental capacity (and in multiples) allowed as add-on during subscription [in TiB]	25	25

Minimum capacity requirements for object storage

You can see the minimum capacity requirements for object storage in the following table:

Capacity	Data tiering	Object	Cloud Volumes ONTAP	Cloud Backup service
Minimum capacity [in TiB]	Not applicable	500	4	4
Incremental capacity (and in multiples) allowed at start of subscription [in TiB]	Not applicable	100	1	1
Incremental capacity (and in multiples) allowed as add-on during subscription [in TiB]	Not applicable	100	1	1

Capacity adjustments

Learn more about capacity adjustments:

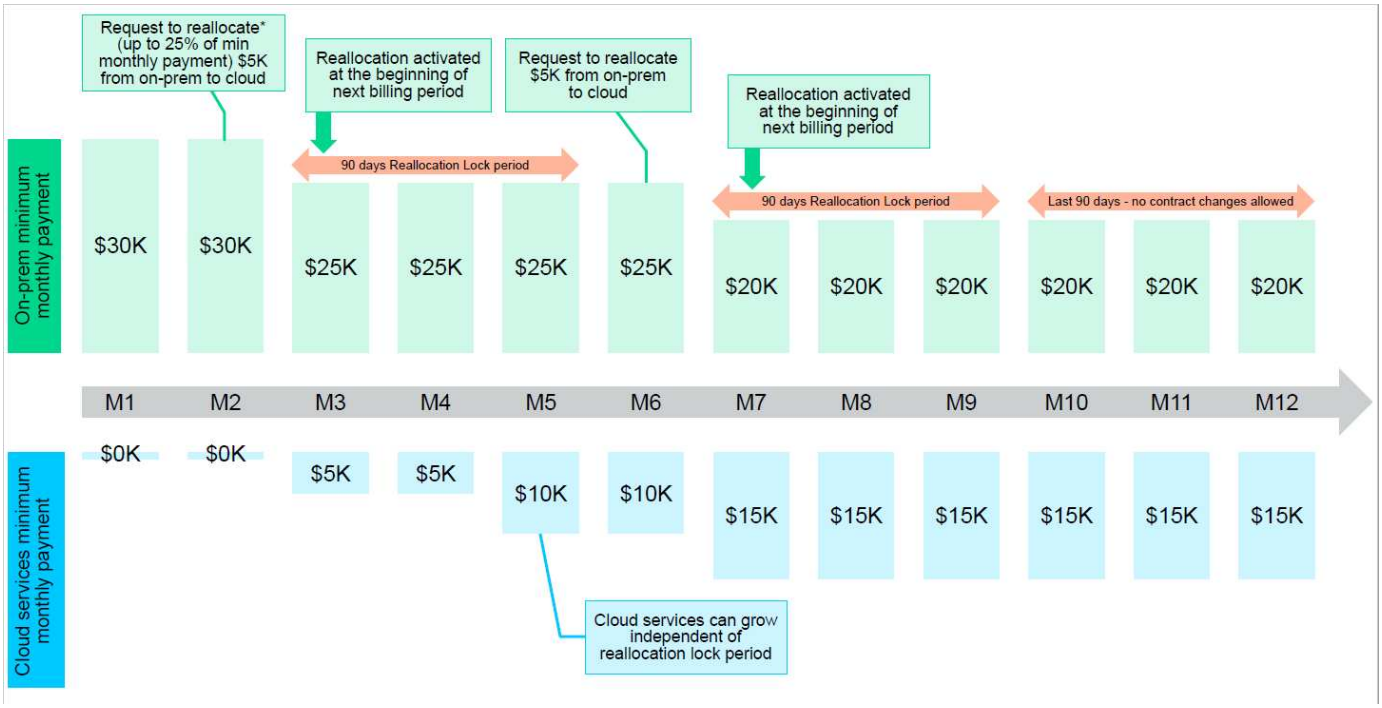
- Capacity can be added anytime during the term, except for the last 90 days of the contract term, in the increments per performance service level as described in the tables in the previous section. Addition of capacity or services is allowed within the last 90 days of the contract term as long as there is a consent of service renewal. Any addition in capacity, new service on-prem or cloud can co-term with the existing term. The invoice sent to you following the activation of the new services reflects the revised billing. Committed capacity of cloud services cannot be reduced at any point during the subscription term. Meanwhile, committed capacity and committed spend on the on-premises services during the term of the contract can

be reduced based on certain criteria as defined in the following section *Capacity reduction*.

- A burst capacity is available at each site, based on the Keystone agreement. Usually, it is 20% above the committed capacity for a performance service level. Any burst usage is billed only for that billing period. If you have additional burst requirement greater than the capacity you agreed upon, contact support.
- Committed capacity can be altered during a contract term, only under certain conditions, as described in the following section *Capacity reduction*.
- Increasing capacity or changing to higher performance service level during a subscription term is allowed. However, moving from a higher performance service level to a lower performance service level is not permitted.
- Any change request in the last 90 days of the service term requires a renewal of the service for a minimum of one year.

Capacity reduction

Capacity reduction (annual) is applicable to the *Annual in Advance* payment model and on-premises only deployments. It is not available for cloud services or hybrid cloud services. It provides provision for on-premises capacity, which can be reduced by up to 25% per service level per subscription. This reduction is allowed once every year to be made effective at the beginning of the next annual billing period. On-premises service-based annual payments should be >=$200K anytime during the term in order to take advantage of capacity reduction. Because it is supported only for on-premises deployments, this billing model does not provide reallocation in spending from on-premises to cloud services. An example of annual capacity reduction is illustrated in the following image.



Quarterly spend reallocation

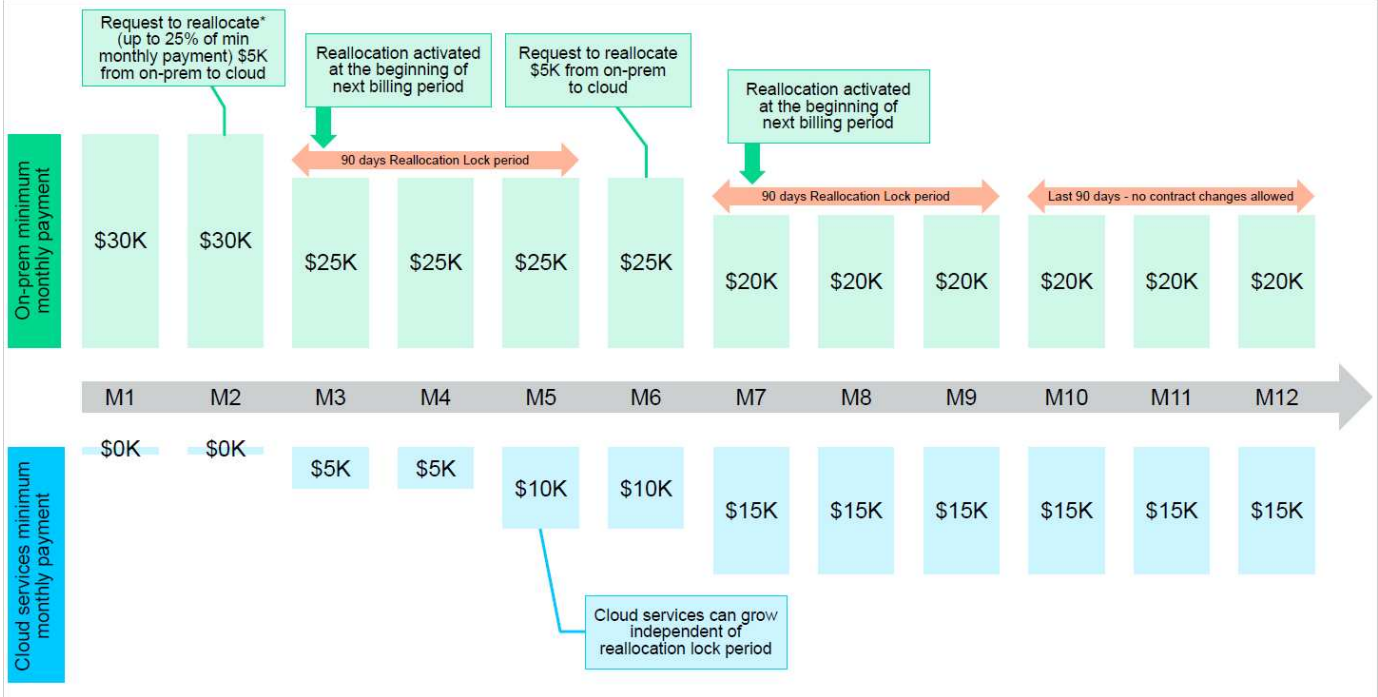
Keystone STaaS offers you the option to reallocate on-premises service spend to Cloud Volumes ONTAP spend.

Requirements and conditions at a subscription level:

- Applies only to monthly billing in arrear model.

- Applies only to subscriptions with 1, 2, or 3-year term commitments.
- Capacity for Cloud Volumes ONTAP and Cloud Backup service should be purchased through Keystone.
- Up to 25% of the existing on-premises, service-based monthly payments can be used for reallocation to cloud services.
- Reallocation requests are made effective only after 90 days from the previous activation date of the reallocation.
- Reallocation cannot be done from cloud services back to on-premises services.
- A request to reallocate should be formally submitted by the customer or partner to Keystone Success Manager (KSM) at least one week before the next billing cycle.
- New requests go into effect only from the consecutive billing cycle.

You can allocate a portion of your expenses towards your subscribed file, block, or object storage performance service levels to hybrid cloud storage services. Up to 25% of the Annual Contract Value (ACV) can be reallocated on a quarterly basis to Cloud Volumes ONTAP Primary and Cloud Volumes ONTAP Secondary services:



This table provides a set of sample values to demonstrate how the reallocation of expenses works. In this example, `$5000` from the monthly spend is reallocated to hybrid cloud storage service.

Before allocation	**Capacity (TiB)**	**Monthly designated expense**
Extreme	125	37,376
After reallocation	**Capacity (TiB)**	**Monthly designated expense**
Extreme	108	37,376
Cloud Volumes ONTAP	47	5,000
		37,376

The reduction is of (125-108) = 17 TiB of the capacity allocated for the Extreme performance service level. On

spend reallocation, the allotted hybrid cloud storage is not of 17 TiB but an equivalent capacity that $5000 can purchase. In this example, for $5000, you can get 17 TiB on-prem storage capacity for the Extreme performance service level and 47 TiB hybrid cloud capacity of Cloud Volumes ONTAP performance service level. Therefore, the reallocation is with respect to the spend, not capacity.

Contact your Keystone Success Manager (KSM) if you want to reallocate expenses from your on-premises services to cloud services.

Copyright information

Trademark information