Understand billing

Keystone

NetApp
January 20, 2026

Table of Contents

Understand billing

Learn about Keystone pricing

The NetApp Keystone STaaS pay-as-you-go subscription service offers flexible and scalable consumption with predictable and upfront pricing for your storage requirements.

Keystone provides you with the following billing facilities:

- You can pay based on IOPS and latency committed capacity to meet various workload needs. The different performance service tiers - Extreme, Premium, Performance, Standard, Value, and Object enable you to manage your storage based on your purchased performance service level.
- It presents predictable billing for the committed capacity and pay-per-use for variable (burst) capacity usage.
- You can select a bundle price for hardware, core OS, and support for one $/TiB price. You have a single invoice for each storage type, file, block, object, or cloud storage services.
- You can select a flexible term for the services and payment options, such as monthly, quarterly, or annually.

Keystone billing is based on committed capacity and variable burst consumption.

For information about different capacities supported in Keystone, see Supported storage capacities in Keystone.

Related information

- Billing based on committed capacity
- Metering based on consumed capacity
- Billing based on burst consumption
- Billing based on miscellaneous volume types
- Billing schedules

Understand Keystone committed capacity billing

Committed capacity is the capacity committed for a particular performance service level while purchasing the subscription.

Committed capacity can be the total capacity for various performance service levels in a single subscription, as accepted by you and NetApp/partner. This capacity is stated on each Keystone order and is billed, regardless of the actual capacity consumption.

For information about different capacities supported in Keystone, see Supported storage capacities in Keystone.

Understand Keystone consumed capacity metering

Keystone STaaS has metering based on the capacity consumed by you during your service usage. Consumed capacity is the capacity that your workloads actually use.

As a part of the Keystone service deployment, NetApp continuously monitors and measures the consumption of the service. At least once in every five minutes, a consumption record is generated by the system, detailing the current consumed capacity for your subscription. These records are aggregated over the billing period to generate invoices and usage reports.

For information about different capacities supported in Keystone, see Supported storage capacities in Keystone.

Understand Keystone burst consumption billing

Keystone STaaS billing is based on *burst capacity*, which is the capacity consumed by you, on top of the committed capacity of your subscription.

Your burst limit is determined and specified in your Keystone agreement. It is 20% above the committed capacity.

Committed capacity is the capacity committed to you while purchasing the subscription. The committed capacity and burst capacity are measured per performance service level. Consumed capacity is the capacity that your workloads actually use.

When the consumed capacity is greater than the committed capacity for a performance service level, burst consumption is recorded and charged accordingly. Usually, it is 20% above the committed capacity. The usage above the burst capacity is indicated as "Above Burst Limit".

This process occurs for each consumption record generated. Burst consumption, therefore, is a reflection of both the amount and tenure of your over-consumed capacities on top of your committed capacities. To learn more, refer to View consumption trends of your Keystone subscriptions.

For information about different capacities supported in Keystone, see Supported storage capacities in Keystone.

Learn about Keystone billing for specific volume configurations

Understanding Keystone billing for specific configurations can help you optimize service usage and manage costs. The configurations include cloned volumes, advanced data protection, temporary volumes, QoS policies, SnapMirror destinations, LUNs, and system/root volumes.

Billing for cloned volumes

If volumes are cloned in ONTAP and you use them for backing up and restoring your data, you can continue using the clones without any additional payments. However, cloned volumes used for any other purpose in your business for an extensive duration are charged.

Note the following:

- Clone volumes are free from charging as long as their size is less than 10% of the parent volume (the physical capacity used in the clone volume compared to the physical capacity used in the parent volume).
- There is no 24-hour grace period for cloned volumes. Only the size of the clone is considered.

- Once the clone volume exceeds 10% of the physical size of the parent, the clone is billed as a standard volume (logical used capacity).

Billing for advanced data protection

Advanced data protection uses NetApp MetroCluster to mirror data between two physically separated clusters. For MetroCluster mirrored aggregates, data is written twice, once on each cluster. The Keystone service charges for consumption on each side independently, resulting in two identical consumption records. The add-on charges are applied on all the capacities in the subscription, irrespective of whether the data is at the source, or it is mirrored or unmirrored data.

Each MetroCluster site has its own subscription and billing. For MetroCluster configurations, the usage measurement accounts for how the storage is utilized across both sites. If you have a MetroCluster setup with 100 TiB per site, only 50 TiB is used actively on each site, with the remaining 50 TiB at each site serving as mirrored backup. The advanced data protection add-on charges are calculated based on this 100 TiB of total active usage and split across both sites through their respective subscriptions at 50 TiB each.

If you monitor your clusters through ONTAP System Manager (System Manager) or Active IQ Unified Manager (Unified Manager), you might see a discrepancy between the consumption reported on these tools and Keystone. System Manager and Unified Manager do not report volumes on the mirrored (remote) cluster, and in doing so, reports half the consumption metrics that the Keystone service reports.

Example:

Site A and Site B are set up in a MetroCluster configuration. When a user creates a volume of 10 TB in site A, an identical volume of 10 TB is created in site B. Keystone identifies 10 TB of consumption in each site, for a total increase of 20 TB. System Manager and Unified Manager report a 10 TB volume created in site A, but do not report a 10 TB volume in Site B.

Additionally, all volumes created on a Keystone system with advanced data protection will be counted towards the consumption of advanced data protection, regardless of whether those volumes are mirrored or not.

Billing for temporary volumes

Occasionally, temporary (TMP) volumes are created by ONTAP when moving volumes. These temporary volumes are short-lived, and the consumption on these volumes is not measured for billing.

Billing and adaptive QoS policies

Keystone measures consumption based on service levels. Each service level is associated with a specific adaptive quality of service (QoS) policy. During deployment, you will be informed of the details of each adaptive QoS policy for your subscribed Keystone services. During storage management operations, ensure that your volumes have the appropriate adaptive QoS policies assigned as per your subscribed service levels to avoid unexpected billing.
For more information about adaptive QoS policies in ONTAP, see Guarantee throughput with QoS overview.

Billing for SnapMirror destinations

The pricing for the SnapMirror destination volume is governed by the adaptive QoS policy for the service level assigned on the source. However, if the source does not have an associated adaptive QoS policy, the destination is billed based on the lowest available service level.

Billing for LUNs

For LUNs, the same billing pattern is followed as for the volumes that are governed by adaptive QoS policies. If separate adaptive QoS policies are set on LUNs, then:

- The size of the LUN is counted for consumption according to the associated service level of that LUN.
- The remainder of the space in the volume, if any, is charged according to the adaptive QoS policy of the service level set on the volume.

System and root volumes

System and root volumes are monitored as a part of the overall monitoring of the Keystone service but are not counted or billed. The consumption on these volumes is exempted for billing.

Learn about Keystone billing schedules

Keystone STaaS subscriptions are billed on a monthly or annual basis.

Monthly billing

Invoices are sent monthly. For the month in which the services are availed, an invoice is sent in the next month. For example, the invoice for the services you have used in January is delivered at the beginning of February. This invoice includes the charges for the committed capacity and if applicable, any burst usage.

Annual billing

An invoice is generated at the beginning of each subscription year for the minimum payment of the committed capacity. It is generated on the start date of the subscription.
Another invoice is sent at the end of a subscription quarter, summing up the applicable charges of any burst usage accrued in that quarter.
If the committed capacity is changed during a subscription, an invoice is sent on the same day for the prorated minimum payments for the rest of that subscription year. The billing is calculated from the day when the change in the committed capacity is effective.

Copyright information

Trademark information