Dell ECS: Technical FAQ

December 2022

H16600.8

White Paper

Abstract

This document addresses technical frequently asked questions (FAQ) for the Dell ECS object storage platform.

Dell Technologies

Copyright

The information in this publication is provided as is. Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

Published in the USA December 2022 H16600.8.

Dell Inc. believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Contents

Executive summary

Overview

This document addresses high-level technical frequently asked questions (FAQs) for ECS.

Audience

This document is primarily intended for Dell personnel not familiar with ECS.

Revisions

Date	Description
October 2016	Initial release (ECS 3.0)
August 2017	Updated for ECS 3.1
March 2018	Updated for ECS 3.2
March 2019	Updated for ECS 3.3
September 2020	Update for ECS available number and ECS 3.5
May 2021	Updated for ECS 3.6.1
June 2021	Updated API connection limits information
March 2022	Updated for ECS 3.7
December 2022	Updated for ECS 3.8.0.1

Note: This document may contain language from third-party content that is not under Dell Technologies' control and is not consistent with current guidelines for Dell Technologies' own content. When such third-party content is updated by the relevant third parties, this document will be revised accordingly.

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email. You can also contact the ObjectScale and ECS Technical Marketing team at objectscale.tme@dell.com.

Author: ECS Technical Marketing

Note: For links to other documentation for this topic, see the ObjectScale and ECS Info Hub.

General FAQ

Question: What are the durability and availability numbers for ECS? How many nines does ECS guarantee?

Data durability of a storage system provides guarantees for data being stored in the system without loss or corruption. ECS supports local and multi-site protection of data, and regular systematic data integrity checks with self-healing capability. ECS durability is 99.999999999 (eleven 9s).

Data availability of a storage system provides guarantees for the system to successfully process data read/write requests and depends on factors outside of ECS control, including equipment/connectivity/power failures. ECS availability is 99.999 (five 9s) based on request-error-rate estimates.

Question: What is Chunk?

Data and system metadata are written in chunks on ECS. An ECS chunk is a 128MB logical container of contiguous space. Each Replication Group will have its own allocation of chunks for buckets and objects associated with the Replication Group. Each chunk can have data from different objects. ECS uses indexing to keep track of all parts of an object that might be spread across different chunks and nodes.

Chunks are written in an append-only pattern. The append-only behavior means that an application’s request to modify or update an existing object will not modify or delete the previously written data within a chunk, but rather the new modifications or updates will be written in a new chunk. For that reason, no locking is required for I/O and no cache invalidation is required.

Question: How does ECS protect data?

At the heart of ECS software is a Storage Engine that is responsible for laying out all data in 128 MB chunks across the system. User data, metadata, and system data are written to a different logical chunk that will contain 128 MB of data. Chunks are both triple-mirrored and erasure coded.

Metadata is written to a chunk, journal creates three copies. Each copy is written to a single disk on different nodes and btree is 12 + 4 in addition to another 12 data fragments.

User data less than 128 MB in size is initially written using triple mirror, with one copy being written in preparation for erasure coding.

Triple mirroring plus in place erasure coding is applicable to a chunk containing the data from any object that is less than 128 MB in size. ECS creates three replica copies of a chunk that contains user data. One copy is written in fragments that are spread across different nodes within the cluster. The remaining two copies are written in their entirety to different nodes. After a chunk is sealed parity is calculated and written to disk, after which the two replica copies written to individual nodes are removed. This process optimizes write performance for small objects, initially using triple mirroring for protection but ultimately leaves the chunk protected by erasure coding.

Inline erasure coding is used for objects that are 128 MB or larger. This process calculates parity as part of the initial write which is distributed across nodes in the Virtual Data Center (VDC). This process does not create replica copies which optimizes large write performance and saves disk I/O.

For geographically distributed systems, replication group policies determine how data is protected and where it can be accessed from. Data that is geo-replicated is protected by storing a primary copy of the data at the local site and a secondary copy of data at one or more remote sites. Each site is responsible for local data protection, meaning that both the local and secondary site will individually protect the data using erasure coding and/or triple mirroring. Replication is performed asynchronously, and data is added to a replication queue as it is written to the primary site. There are worker I/O threads continuously processing the queue. With more than two sites in a replication group, where “Replicate to All Sites" is off, an XOR mechanism can be used which serves to reduce overhead significantly. See the Dell ECS Overview and Architecture Guide for more details.

Question: What are the limitations regarding Storage Pool, VDCs, RGs, Namespaces, Buckets, Objects?

The storage that is associated with a VDC must be assigned to a storage pool and the storage pool must be assigned to one or more replication groups to allow the creation of buckets and objects. While the VDC can have a storage pool defined for each of the two Erasure Coding schemes (12+4 with minimum of five nodes or 10+2 with minimum of six nodes), the best practice is to have all nodes associated to one storage pool for a given VDC. A node cannot exist in more than one storage pool. The storage pool can span racks, but it is always within a site. When the storage pool reaches 90% of its total capacity, it does not accept write requests and it becomes a read-only system. A storage pool can be associated with more than one replication group.

The maximum number of VDC’s per ECS federation and/or replication group is eight.

Replication groups are limited to the possible variations in topology of federated VDCs. Best practice is to create replication groups that match the access and protection requirements for the data being written across the federated sites. Most customers create a local replication group for a VDC where data should be kept local to the VDC and one or more geo-replicated groups consisting of two or more VDCs for geo-protection of the data. Each replication group has chunks allocated for the buckets and objects associated to it.

There are no limitations to the number of namespaces that can be created. A namespace has a set of configured users who can store and access objects within the namespace. Users from one namespace cannot access the objects that belong to another namespace. An object in one namespace can have the same name as an object in another namespace. ECS can identify objects by the namespace qualifier.

There are no limitations to the number of buckets that can be created, nor are their limitations to the number of objects that can be created in a given bucket. A bucket is assigned to a namespace and object users are also assigned to a namespace. An object user can create buckets only in the namespace to which the object user is assigned.

Question: What are the detailed steps of ECS Tech Refresh?

Tech Refresh is a data-in-place service with automated procedures introduced in ECS v3.5.x. The technical steps for doing tech refresh include:

Rack Extend:

- Reimage new rack nodes using same OS version as the original rack nodes.
- Generate cluster.ini after upgrading to latest Service Console
- Edit cluster.ini and add new rack nodes
- Perform extend by adding new rack nodes to existing cluster
- Add newly extended nodes to storage pool and VDC from UI/API
- Rerun extend command for post extend checks
- Add extended nodes

Data Migration:

- Trigger data migration from original nodes(source) to newly extended nodes(target)
- Check the data migration progress and completion status using;
- Service Console Output
- ECS UI storage pool
- Grafana Migration Report

Node Evacuation:

- Run Node evacuation command on current installer node to setup new installer node.
- Run the Node evacuation command from new installer node
- Rerun the command on the same node to check the status
- Node evacuation complete

Checks after tech refresh:

- Service Console health check
- ECS UI without the evacuated nodes

For more detail, see the Professional Servers Procedures under SolVe.

Question: What are the types of Garbage Collection and how are they defined?

There are two garbage collection methods used by ECS to reclaim space from discarded chunks and depending on whether said chunks consist of entirely deleted objects, or a mixture of deleted and non-deleted objects. They are:

- Normal Garbage Collection - When an entire chunk is garbage, reclaim space.

- Partial Garbage Collection by Merge - When a chunk is 2/3 garbage, reclaim the chunk by merging the valid parts of with other partially filled chunks to a new chunk, reclaim space.

Question: What automation platforms are supported and what is included?

Although Ansible is not official supported with ECS today, it is however possible to use Ansible Playbooks to automate the creation of namespaces, object users, buckets, and so on.

Question: What kind of file types are supported by S3 select?

S3 select supports csv, json and parquet formats and it supports querying gzip/bzip2 compressed objects of the aforementioned file types.

Hardware models and configurations

Question: What ECS appliance configurations are available?

Current Gen3 ECS hardware includes classes of nodes for the EX500, EX5000, and EXF900:

- The EX500 appliance is a dual socket, multi-core CPU node type which aims to provide a balance of performance and density. With options for 12 or 24 drives, this series provides a versatile option for midsize enterprises looking to support modern application and/or deep archive use cases. HDD size can be 2TB, 4TB, 8TB, 12TB,16TB, and 20TB.
- EX5000 is the next generation of ECS dense appliance that refreshes EX3000. EX5000 has a maximum capacity of 11.2 PB of raw storage per rack using the 16TB disks, and can grow into exabytes across several sites, providing a deep and scalable solution. EX5000 is a 5U system that can hold up to 100 drives per system. These nodes are available in two different configurations: EX5000S and EX5000D:

  - The EX5000S is a single-node chassis with options for 25, 50, 75, and 100 drives.
  - The EX5000D is a dual-node chassis with 50 and 100 drives.

  EX5000 is designed with 192GB memory and Intel® Xeon® Gold 6230R 26 core CPUs, making it up to 3x faster than EX3000. EX5000 is supported to be used to expand an existing EX3000 VDC, however, it requires a new rack due to higher power needs. To add EX5000 to an existing rack, submit an RPQ to Dell Professional Services.
- The EXF900 is a dual socket, multi-core CPU node with an all-flash object storage solution of hyper-converged nodes for low latency and high Input or Transactions Per Second (TPs) ECS deployments. It has options for 12 or 24 disks and 3.84TB, 7.68 and 15.36 NVMe SSD per node.

All drive sizes must be consistently the same within the node. Node sizes may be mixed within a cluster if each new node size is introduced by a minimum size storage group.

Here are the previous model definitions: EX300 (3,072TB), EX3000 (11,520TB), U400 (320TB), U400E (400TB), U480E (480TB), U400T (640TB), U2000 (1,920TB), U2800 (2,880TB), U4000 (3,840TB), D4500 (4,480TB), D5600 (5600 TB), D6200 (6,272TB), and D7800 (7840TB).

An optional SSD drive for metadata caching is available as of ECS 3.5. Former deployments are upgradable using separate drive upgrade kits. If a caching SSD is added to a Virtual Data Center (VDC), every node in the VDC is required to have the caching SSD installed.

ECS version 3.8.0.1 supports memory upgrade expansion to 192 GB on EX300, EX500, EX3000, and Gen 2 platforms, with the support of Dell Professional Services.

Question: What switches are used?

ECS switching components and architecture were modified to include a back-end (BE) switch network for private admin connections, and a front-end (FE) switch network for customer public network connection. It should be noted that all node-to-node communication still travels over the FE switches for current releases.

For EX500 and EX5000 series:

- Two optional Dell Networking S5248F 25 GbE 1U Ethernet switches can be obtained for network connection or the customer can provide their own 25 GbE HA pair for the FE network. If S5248F 25 GbE front-end switches are used, then 2 x 200 GbE VLT cables are needed.
- For the backend (BE), two Dell Networking S5248F 25 GbE 1U Ethernet switches with 48 x 25 GbE SFP ports and 4 x 100 GbE uplink ports and 2 x 200 GbE VLT must be included in the configuration.

For EXF900 series:

- Two optional Dell Networking S5248F 25 GbE 1U Ethernet switches can be obtained for network connection, or the customer may provide their own 25 GbE HA pair for the FE network. If S5248F 25 GbE front-end switches are used, then 2 x 200 GbE VLT cables are needed.
- For the back-end (BE), two Dell Networking S5248F 25 GbE 1U Ethernet switches with 48 x 25 GbE SFP ports and 4 x 100 GbE uplink ports and 2 x 200 GbE VLT must be included in the configuration.
- For intra-rack back-end networks, two aggregation switches Dell S5232 with 28 x 100 GbE ports and 4 x 100 GbE VLT are needed between multiple racks.

Question: Is there a limit on number of nodes that ECS can scale up to?

ECS is a distributed system with no limit to the number of nodes.

Question: Which model of ECS node can intermix with EXF900

With ECS 3.7, you can create a VDC that consists of an All flash Array (such as EXF900 nodes) and any of the ECS Gen3 or Gen2 HDD-based nodes. EXF900 and the HDD nodes must be in their own Storage Pools.

It has the flexibility to add different media type nodes to your existing VDC or to create a new VDC.

- Fresh install with AFA/HDD intermixes mode
- Extend node and add storage pool to support AFA/HDD intermix mode
  - Extend HDD only VDC to co-existing VDC
  - Extend AFA only VDC to co-existing VDC

Networking

Question: What are the common network considerations customers make when deploying ECS?

Each rack has two top-of-rack (TOR) switches which are uplinked to the customer network. The TOR switches carry all traffic except for out-of-band (OOB) management. Best practice is to have a minimum of two uplinks per rack (four minimum). Management, data (read/write) and geo-replication traffic can be separated, virtually or physically, which allows for enhanced security and performance separation.

Question: How many IP addresses are required for an ECS deployment?

A minimum of one IP address for each ECS node is required.

For out of band management, one IP address is required for each ECS node to be managed.

Question: How many switch ports does the customer need to provide?

For each rack a minimum of two 10/25 GbE switch ports in the customer's infrastructure are required, one for each uplink from each TOR switch.

No switch ports are required for the BE switch(es) unless the customer wants to have RMM (Remote Management Module) for Gen2 and iDRAC (integrated Dell Remote Assistance Controller) for Gen3 access. RMM/iDRAC access is optional and would require one or two switch ports in the customer’s infrastructure depending on topology.

Question: How does the network separation work?

ECS supports separating different types of network traffic for security and performance isolation. There is a mode of operation called the network separation mode. In this mode, each node can be configured at the operating system level with up to three IP addresses, or logical networks, for each of the three different types of network traffic. This feature has been designed to provide the flexibility of either creating three separate logical networks for management, replication and data, or combining them to either create two logical networks, for instance management and replication traffic is in one logical network and data traffic in another logical network.

Software

Authentication

Question: Which authentication providers are available with ECS?

ECS supports Active Directory, LDAP, and Keystone v3, an OpenStack project that provides identity, token, catalog, and policy services. Keystone compatibility allows ECS to be a drop-in replacement for OpenStack Swift. Authentication providers enable control and monitoring of administration users, not object access users.

Note: Identity and Access Management (IAM) functionality was released in version 3.5 and IAM is only for data (bucket/object) accessed by the S3 protocol.

Monitoring

Question: What methods are available to monitor ECS?

The WebUI (ECS Portal), integrated Grafana, CLI, SNMP, REST API, and Dell SRM are available to provide information about the health of ECS. Also available is Dell Secure Remote Support (SRS) which provides a secure two-way communication between Dell storage systems and the support system for proactively identifying and responding to possible issues. SNMP queries for basic health metrics such as CPU and memory are available as are traps for critical events. Remote syslog support is available. SNMPv2 and SNMPv3 (USM mode) are supported.

Question: What metrics does the management system report (such as raw storage capacity, usable capacity, CPU utilization, and current IOPS) and how are they defined?

There is monitoring page in the ECS UI, it includes metering data for namespaces, or buckets within namespaces, capacity utilization, system Health, transactions, recovery status, disk bandwidth and geo replication. There is also an advanced monitoring section which uses Grafana which is integrated within the ECS UI, it includes data access performance metrics. For more details about these metrics, see the ECS monitoring guide.

Question: What types of Event Notification does ECS provide?

During the ECS installation process, your customer support representative can configure and start an snmpd server to support specific monitoring of ECS node-level metrics. A Network Management Station client can query these kernel level snmpd servers to gather information about memory and CPU usage from the ECS nodes, as defined by standard Management Information Bases (MIBs). It can query the snmpd servers that can run on each ECS node from Network Management Station clients for the following.

SNMP MIBs:

- MIB-2
- DISMAN-EVENT-MIB
- HOST-RESOURCES-MIB
- UCD-SNMP-MIB

It can query ECS nodes for the following basic information by using an SNMP Management Station or equivalent software:

- CPU usage
- Memory usage
- Number of processes running

Question: What are the refresh times for data showed in the ECS UI?

On monitoring displays, you can force a table to refresh with the latest data by clicking the Refresh icon.

Question: What are the definitions for total, used, available, and reserved capacity?

- Total: Total capacity of the system that is online. This is the total of the capacity that is already used and the capacity still free for allocation.
- Used: Used online capacity in the system.
- Available/reserved: Online capacity available for use, including the approximately 10% of the total capacity that is reserved for failure handling and for performing erasure encoding or XOR operations.

ECS internals

Question: How does ECS efficiently write both small and large unstructured data?

For smaller writes to storage, ECS uses a method called box carting to minimize impact to performance. Box carting aggregates multiple smaller writes of 2MB or less in memory and writes them in a single disk operation. Box carting limits the number of roundtrips to disk required process individual writes.

For writes of larger objects, nodes within ECS can process write requests for the same object simultaneously and take advantage of simultaneous writes across multiple spindles in the ECS cluster. Writes of objects less than 128MB in size are tripled mirrored across three nodes, with one of the three being distributed in its erasure coded format, waiting for the chunk to completely fill before calculating the EC bits. Objects larger than 128MB in size skip the triple mirroring step, with 128MB portions of the object distributed in its erasure coded format across the nodes in the storage pool. Thus, ECS can ingest and store small and large objects efficiently.

APIs and protocols

Question: Can the same object in ECS be accessed by different protocols?

CAS objects are only accessible by their own API. NFS and Dell Atmos can access each other’s created objects using path-based style. S3, NFS, Swift, and HDFS can interoperate such that objects created in any of these protocols can be accessed by any one of them as well.

Question: How does ECS HDFS work?

ECS supports the Hadoop S3A client for storing Hadoop data. S3A is an open source connector for Hadoop, based on the official Amazon Web Services (AWS) SDK. It was created to address storage scaling and cost problems that many Hadoop admin were having with HDFS. Hadoop S3A connects Hadoop clusters to any S3 compatible object store.

Question: What is the API connection limit?

All the REST APIs (S3, Atmos, Swift) go through an http server component called netty. The effective limit for REST protocols has been 1000 active connections per node.

For the CAS API, it has a maximum of 600 active connections per node.

Centera

Question: What is ECS CAS missing that Centera customers should be aware of?

Data shredding is not available currently in ECS CAS. As of 3.0 all Advanced Retention Management (ARM) features such as event based retention, litigation holds, and the Min/Max governor are available. NOTE: ARM features are available for CAS only.

Question: What does the Transformation engine do?

The Centera Transformation and Migration feature allows organizations to natively (within ECS software) and seamlessly migrate ECS-compatible applications to ECS from Centera. The practically non-disruptive application cutover allows data to be moved as a background process. As of 3.0, migrations can be administered using the Web UI.

Question: Are there any caveats to migrating data from Centera to ECS?

Yes. ECS Sync and Datadobi migration tools cannot migrate Centera legacy data to ECS without disruption to existing EBR and/or LH information. ECS's built-in transformation engine does support migrating Centera legacy data with EBR and/or LH.

NFS

Question: Which ports are required for clients to access over NFS?

2049 (nfsd TCP, mountd TCP/UDP, statd TCP), 10000 (nlockmgr TCP/UDP) and 111 (portmapper TCP/UDP)

Note: Services are part of ECS software and not exposed by the underlying operating system.

Question: What are the primary use cases for NFS on ECS?

- Archiving or primarily sequential writes.
- Applications currently using file that will be modified later to use object.
- Multi-protocol access to object data, as with data loading for HDFS analytics, for example.

Question: What are the implications of the "server-side" metadata caching the ECS NFS implementation uses to increase performance by reducing related disk operations?

Metadata is cached by nodes for the NFS operations they serve to clients. The cache allows for serving metadata quicker than is possible if disks access is required for each operation. Changes to metadata are not globally tracked across nodes and as such will not get reflected instantly across nodes. If client1 and client2 both connect to the same ECS node, both of them see the same information since it is either being served from the same cache or disk.

Metadata in cache is considered alive until it times out. This means if a change is made directly on disk for an object, and a client subsequently performs a listing operation on that object, older data from cache will be returned to the client until the time at which the cache expires. After expiration requests for the metadata will be served from disk and cache repopulated with the most recent information. If client1 and client2 connect to different ECS nodes, then there is a possibility that they see different information, if related metadata exists in cache, until the cache times out. Basically, metadata is cached locally to each ECS node and is not globally coherent.

Question: Are there no known limitations or hard-coded values for max number of directories or files per directory?

There are no known limitations, but the more files in the directory, the longer listing contents will take to complete.

Question: Is there a max file size in ECS?

For NFS only, the maximum file size allowed is 16 TB.

Question: How do storage administrators configure NFS access to a bucket?

Along with creating the exports, for a user to access a file over NFS, a mapping must be created between a bucket user and UNIX UID and GID. With this mapping ECS can translate the UID and GID received over the wire as part of the NFS operation to a bucket user to determine access. ECS does not retrieve user mapping from authentication sources, that is, all mappings must be created by a storage administrator.

Similarly, for access from a client configured with Kerberos, a mapping between principal names and UID/GID is required so that ECS can return a UID and GID over the wire to the client in its response.

Question: What authentication methods are supported for ECS NFS?

ECS NFS supports sys, krb5, krb5i, krb5p.

Question: Does ECS support all NFS v3 operations?

ECS NFS supports all NFSv3 procedures EXCEPT for LINK: Create hard link to an object.

Question: Can files be accessed over NFS during a site outage?

Read access over NFS is available, just like with all other protocol access, during a site outage. Write access over NFS depends on the zone ownership of the affected path during outage.

For example, a three-site replication group contains an Access During Outage (ADO) enabled namespace, ns1, and ns1 contains file-system-enabled bucket, b1, which is also configured with ADO enabled. A three-directory-deep path exists in b1 and each directory was created and is owned by a different site/zone. /ns1/b1/dir1/dir2/dir3.

If site 2, owner of dir2, is temporarily unavailable, contents in dir2 are read-only during outage but contents in dir1 (owned by site 1) and dir3 (owned by site 3) remain writable.

Fault tolerance

Question: What is the expected behavior during the loss of one or more disks?

Data that exists on failed disks will be reconstructed using either the remaining erasure coded data and parity fragments or the replica copies.

For more details on node failures, see the Dell ECS: High Availability Design white paper.

Question: What is the expected behavior during loss of node(s)?

Any request for system metadata owned by a node that is not responding, will trigger the requested metadata ownership to be redistributed across the remaining nodes in the site. When this completes, the request will complete successfully. Data that exists on disks from the unresponsive node will be reconstructed using either the remaining erasure coded data and parity fragments or the replica copies.

- Concurrent failure: When nodes fail concurrently it means nodes fail almost simultaneously, or a node fails before recovery from a previous failed node completes.
- One-by-one failure: When nodes fail one by one it means one node fails, all recovery operations complete and then a second node fails. This can occur multiple times and is analogous to a VDC going from something like four sites to three sites to two sites to one site. This requires that the remaining nodes have sufficient space to complete recovery operations.

For erasure-coded content for each single site, the following chart is provided.

EC scheme	Nodes in VDC	Concurrent failure	One-by-one failure
12+4	5 nodes	Loss of 1 node: reads and writes are successful, erasure coding continues. Loss of 2 or 3 nodes: some reads will fail, new writes will stop, erasure coding stops and new writes will be triple mirrored.	Loss of 1 node: reads and writes are successful, erasure coding continues. Loss of 2 or 3 nodes: all reads will succeed, new writes will stop, erasure coding stops and new writes will be triple mirrored.

EC scheme	Nodes in VDC	Concurrent failure	One-by-one failure
10+2	6 nodes	Loss of 1 node: reads and writes are successful, erasure coding stops and new writes will be triple mirrored. Loss of 2 nodes: some reads will fail, new writes will be successful. Loss of 3 nodes: some reads and writes will fail.	Loss of 1 node: reads and writes are successful, erasure coding stops and new writes will be triple mirrored.

The basic rules for determining what operations fail in a single site with multiple node failures include:

- If you have three or more concurrent node failures, some reads and writes will fail due to the potential loss of all three replica copies of the associated triple mirrored metadata chunks.
- Writes require a minimum of three nodes.
- Erasure coding will stop, and erasure coded chunks will be converted to triple mirror protection if the number of nodes is less than the minimum required for each erasure coding scheme. Since the default erasure coding scheme, 12+4 requires four nodes, erasure coding will stop if there are fewer than four nodes. For cold storage erasure coding, 10+2, erasure coding will stop if there are less than six nodes.
- If the node count goes below the minimum required for the erasure coding scheme, erasure coded chunks will be converted to triple mirror protection. As an example, in a VDC with default erasure coding and four nodes, after a node failure the following would happen:
    - Node failure causes four fragments to be lost.
    - Missing fragments are rebuilt.
    - Chunk creates three replica copies, one on each node.
    - EC copy is deleted.

For more details on node failures, see the Dell ECS: High Availability Design white paper.

Question: What are the types of site outages and How ECS handle it?

Site outages can be classified as a temporary site outage (TSO) or a permanent site outage (PSO). A TSO is a failure of the WAN connection between two sites, or a temporary failure of an entire site (for example, a power failure). A site can be brought back online after a TSO. ECS can detect and automatically handle these types of temporary site failures. A PSO is when an entire site becomes permanently unrecoverable, such as when a disaster occurs. In this case, the System Administrator must permanently fail over the site from the federation to initiate failover processing. VDCs in a geo-replicated environment have a heartbeat mechanism. Sustained loss of heartbeats for a configurable duration (by default, 15 minutes) indicates a network or site outage and the system transitions to identify the TSO.

If a disaster occurs, an entire site can become unrecoverable; it is referred to in ECS as a permanent site outage (PSO). ECS treats the unrecoverable site as a temporary site failure, but only if the entire site is down or unreachable over the WAN. If the failure is permanent, the System Administrator must permanently fail over the site from the federation to initiate failover processing. This initiates resynchronization and reprotection of the objects that are stored on the failed site. The recovery tasks that are run as a background process.

Starting with version 3.7, ECS supports recovery from a multiple simultaneous site (N-1 site) failure. This shortens the data recovery time. The customer must contact Dell to support the operation. It only supports the replication group setting with replication to all sites.

For more details, see the Dell ECS: High Availability Design white paper.

Question: What is the expected behavior during loss of a site?

If a single site is temporarily unavailable, in a replication group containing more than one site, some operations will be limited such as:

- File systems within HDFS/NFS buckets that are owned by the unavailable site are read-only.
- Buckets, namespaces, object users, authentication providers, replication groups, and NFS user and group mappings cannot be created, deleted, or updated from any site (replication groups can be removed from a VDC during a permanent site failover).
- You cannot list buckets for a namespace when the namespace owner site is not reachable.
- OpenStack Swift users cannot log in to OpenStack during a TSO because ECS cannot authenticate Swift users during the TSO. After the TSO, Swift users must re-authenticate.
- Create, read, update objects and list object in a bucket may be interrupted depending upon replication group options configured on the bucket.

For more details on site failures, see the Dell ECS: High Availability Design white paper.

Security and compliance

Question: What type of data encryption is available, and where is it applied?

Data at Rest Encryption (D@RE) is simple, low-touch server-side encryption. It supports enterprises and service providers seeking to protect sensitive data on storage media. In ECS encryption can be enabled at the namespace and bucket levels.

Question: Which EKMs are supported?

ECS supports External Key Management using external key managers that are Key Management Interoperability Protocol version 1.4 (KMIP v1.4) compliant. ECS delegates the storage and protection of top-level Key Encrypting Key (KEK), the Master Key to the external EKM. ECS 3.3 and later versions support Safenet KeySecure (Gemalto Safenet) and ECS 3.4 supports the IBM SKLM 3.0 (Security Key Lifecycle Manager). ECS 3.6 supports Safenet KeySecure 8.11 with client certificate authentication only. ECS 3.8.0.1 supports Thales CipherTrust because Gemalto SafeNet KeySecure will end of life on December 31, 2023. ECS customers who are using KeySecure can migrate to CipherTrust Manager.

Question: How does the encryption affect throughput?

In general, the performance when accessing objects in an encryption-enabled namespace can be approximated as large reads are up to half as fast when encrypted. This behavior is not seen with large creates. Small reads are performed at a lower rate as well but not nearly as much as large reads. Small creates, as with large create, are nominally impacted by encryption.

Question: How does Multi-tenancy work on ECS?

ECS supports access by multiple tenants, where each tenant is defined by a namespace and the namespace has a set of configured users who can store and access objects within the namespace. Users from one namespace cannot access the objects that belong to another namespace.

Question: Which administrative activities will be logged?

ECS monitor metering provides critical information about viewing and using the monitoring pages in the ECS portal dashboard. In the Events page, all activity by users working with the portal, the ECS REST APIs, and the ECS CLI. Other audit types include upgrade activities. For more details, see the ECS monitoring guide.

Question: How does the consistency checker process work?

ECS is a strongly consistent system that uses ownership to maintain an authoritative version of each namespace, bucket, and object. Ownership is assigned to the VDC where the namespace, bucket or object is created. For example, if a namespace, NS1, is created at VDC1, VDC1 owns NS1 and responsible for maintaining the authoritative version of buckets inside NS1. If a bucket, B1, is created at VDC2 inside NS1, VDC2 owns B1 and is responsible for maintaining the authoritative version of the bucket contents, as well as each object's owner VDC. Similarly, if an object, O1, is created inside B1 at VDC3, VDC3 owns O1 and is responsible for maintaining the authoritative version of O1 and associated metadata.

Question: What is the difference between a IAM user, and an object user created locally?

Local users include management users which and object users. Management users are for configuring, administering, and monitoring the logical components of the ECS architecture. Object users are users of the ECS object store. They access ECS through object clients that are using the object protocols that ECS supports (S3, Atmos, OpenStack Swift, and CAS). Object users can be assigned Unix-style permissions to access buckets exported as file systems for HDFS.

IAM is accessible only by S3 protocol and allows for fine grained access controls where IAM users are joined with policies that grant or restrict access to namespaces, buckets and objects. Management users in ECS have complete access to IAM capabilities.

Question: How does SAML work?

SAML is an open standard for exchanging authentication and authorization data between an identity provider and a service provider. SAML provider in ECS is used to establish trust between a SAML-compatible Identity Provider (IdP) and ECS. For more information, see the ECS data access guide.

Question: What are the ECS compliance features?

The ECS Appliance meets the storage requirements for the following standards, as verified by Cohasset Associates, Inc.

- Securities and Exchange Commission (SEC) in regulation 17 C.F.R. & 240.17a-4(f)
- Commodity Futures Trading Commission (CFTC) in regulation 17 C.F.R. & 1.31(b)-(c)
- DISA STIG security standards which address the CAT I and CAT II vulnerabilities
- FIPS 140-2 mode enforces the use of approved-only algorithms within D@RE; FIPS 140-2 compliance is only for the D@RE module, not the entire ECS product.

References

Dell Technologies documentation

The following Dell Technologies documentation provides other information related to this document. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell Technologies representative.

- *ObjectScale and ECS Info Hub*
- *Dell ECS Data Access Guide*
- *Dell ECS Monitoring Guide*
- *Dell ECS Overview and Architecture*