Dell EMC ECS: Technical FAQ

Abstract

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

March 2019

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

Acknowledgements

This paper was produced by the Dell EMC Unstructured Technical Marketing Engineering and Solution Architects team. Please send comments, suggestions, or feedback to unstructured.tme.sa@emc.com.

Table of contents

Executive summary

This document addresses technical frequently asked questions (FAQs) for the Dell EMC™ ECS™ platform.

This document will be reviewed, and may be updated, with each general availability (GA) release of ECS. It is not intended to be release-specific but may detail new features as they become available in ECS.

Audience

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

To make comments or suggestions, contact the ECS Technical Marketing team at ecstme@emc.com.

1 General FAQ

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

Answer: 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 99.9 (three 9s) based on request-error-rate estimates.

Question: How does ECS protect data?

Answer: At the heart of ECS software is a Storage Engine which 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 ~128MB of data. All chunks are triple-mirrored or erasure coded.

Metadata is written to a chunk, of which ECS creates three replica copies. Each replica copy is written to a single disk on different nodes.

User data is written using either triple mirror + in place erasure coding or inline erasure coding.

Triple mirror 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 utilizing 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 VDC. This process does not create replica copies which optimizes large write performance and saves disk I/O.

In the case of 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 copies 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, an XOR mechanism can be used which serves to reduce overhead significantly. See the architecture guide referenced at the end of this document for more details.

2 Hardware models and configurations

Question: What ECS appliance configurations are available?

Answer: Current Gen3 ECS hardware includes two classes of nodes including the EX300 and the EX3000. The EX3000 is further divided into EX3000S and EX3000D node types often referred to as EX3000S/D. The 'S/D' node types refers to the optional single-node chassis or dual-node chassis configuration.

The EX3000S single-node chassis configuration connects with all the node HDDs, which allows for dense storage configurations with either 45, 60, or 90 disks of 12TB HDD. The EX3000D dual-node chassis increases the number of CPU compute resources in the rack, but reduces the number of disks per node within the chassis to either 30 or 45 disks of 12 TB HDD.

The EX300 is a smaller class of node designed to lower the entry point for object-based storage and enable replacement of smaller Dell EMC Centera™ deployments. Each EX300 node houses 12 HDD. HDD sizes can be 1TB, 2TB, 4TB, or 8TB. All drive sizes must be consistently the same within the node. Node sizes may be mixed within a cluster as long as each new node size is introduced by a minimum size storage group.

In early 2018, product management did away with previous Gen2 U-Series model numbers. The Gen2 U-Series are referred to as Gen2 U-Series, x node, with x 8 or 12 TB drives. For historic reference here are the previous model definitions: 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).

Question: What switches are used?

Answer: With the release of ECS 3.2.2, 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.

Two optional Dell EMC Networking S5148F 25 GbE 1U Ethernet switches can be obtained for network connection or the customer may provide their own 10 GbE or 25 GbE HA pair for the FE network.

For the back-end (BE), two Dell EMC Networking S5148F 25 GbE 1U Ethernet switches with 48 x 25 GbE SFP ports and 6 x 100 GbE uplink ports must be included in the configuration.

As of ECS 3.2.2, the Arista® and Cisco® switch options have been discontinued.

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

Answer: ECS is a completely distributed system, with no Master-Slave or centralized architecture. Hence, one could add infinite number of nodes within a single site. There are deployments where we have 80+ racks and growing. For both the EX300 and the EX3000S/D node clusters, a customer can add nodes in 1 node increments after the initial storage pool minimum of 5 nodes. (Other rules for cluster expansion apply and may be referenced in the ECS 3.2.2 release notes.

3 Networking

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

Answer: Out of the box, each rack's two 10 GbE top-of-rack (TOR) switches are uplinked to the customer's network using between one to eight uplink ports per switch. The 10 GbE TOR switches carry all traffic with the exception of out-of-band (OOB) management. Best practice is to have a minimum of 2 uplinks per rack (4 minimum). Management, data (read/write) and geo-replication traffic can now be separated which allows for enhanced security and performance separation.

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

Answer: A minimum of one customer-provided IP address for each ECS node is required. That is, four or eight IPs per rack depending upon appliance type.

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

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

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

No switch ports are required for the private 1 GbE switch (Out-of-Band management (OOB)) unless customer desires to have RMM (Remote Management Module) access. RMM access is optional and would require 1 or 2 switch ports in the customer’s infrastructure depending on topology. The management switch directly connects to each rack's 10 GbE switches and in a multi-rack environment would connect to the other rack’s 1 GbE management switch.

4 Software

4.1 Authentication

Question: Which authentication providers are available with ECS?

Answer: 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 only enable control and monitoring of ECS administration users, not object access users.

4.2 Monitoring

Question: What methods are available to monitor ECS?

Answer: The WebUI (ECS Portal), CLI, SNMP, and REST API are available to provide information on the health of ECS. Also available is EMC Secure Remote Support (ESRS) which provides a secure two-way communication between EMC storage systems and the EMC 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 as of 3.0. SNMPv2 and SNMPv3 (USM mode) are supported.

4.3 ECS internals

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

Answer: ECS implements a box-carting method to perform all writes to disk. All client writes go through a buffer and threads on the backend of the buffer grab up to 2 MB of data (which may contain data from one or more clients) at a time. With several small writes from the buffer, the writes to disk are completed in bulk as opposed to one-by-one as the clients provide them. With fewer trips to write to disk, small write efficiency can be more in line with large writes.

5 APIs and protocols

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

Answer: Yes and no. CAS objects are only accessible via their own API. NFS and Dell EMC Atmos™ can access each other's created objects using path-based style. S3, NFS, Swift (as of ECS 3.1), 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?

Answer: ECS HDFS provides a primary or secondary HDFS-compatible file system. An ECS client (a .jar file) installs on the data nodes of an existing Hadoop cluster and registers ECS as a file system that is available for MapReduce jobs as well as Pig, Hive queries etc. The compute runs in the Hadoop nodes. HDFS data on ECS is protected like all other data on ECS, triple-mirrored on ingest and subsequently erasure-coded, or erasure-coded on ingest when possible during writes of 128 MB or larger.

5.1 Centera

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

Answer: 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?

Answer: 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 via the Web UI.

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

Answer: 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.

5.2 NFS

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

Answer: 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?

Answer: 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?

Answer: 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 good 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.

Note: Access checks for write and updates to a file or directory are always accurate as they happen in a deeper layer.

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

Answer: 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?

Answer: For NFS only, the maximum file size allowed is 4 TB.

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

Answer: 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 or/or 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 currently 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?

Answer: ECS NFS supports sys, krb5, krb5i, krb5p.

Question: Does ECS support all NFS v3 operations?

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

ECS NFS will support the LINK procedure in the future. LINK requires additional structures which aren't currently in place - more time is required. SYMLINK is supported along with all of the other NFS v3 operations.

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

Answer: Read access via NFS is available, just like with all other protocol access, during a site outage. Write access via NFS depends on the zone ownership of the effected 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 writeable.

6 Fault tolerance

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

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 HA Design white paper listed in the useful links section of this paper.

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

Answer: Any request for system metadata owned by a node that isn't responding, will trigger the requested metadata ownership to be redistributed across the remaining nodes in the site. Once 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.

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.
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 1–3 nodes: reads and writes are successful, erasure coding stops and new writes will be triple mirrored.

- Concurrent failure: When nodes fail concurrently it means nodes fail almost at the same time, 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 4 sites → 3 sites → 2 sites → 1 site. This requires that the remaining nodes have sufficient space to complete recovery operations.

For more details on node failures, see the ECS HA design white paper listed in appendix A.1.

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

Answer: 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 ECS HA design white paper listed in appendix A.1.

A Technical support and resources

Dell.com/support is focused on meeting customer needs with proven services and support.

Storage technical documents and videos provide expertise that helps to ensure customer success on Dell EMC storage platforms.

A.1 Related resources

Note: Links in this section may require login access to Dell EMC Support site or Dell EMC internal site.

- Dell EMC ECS: High Availability Design
- UDS/ECS Technical Marketing Collateral
- ECS Sizing Tool (Web based presales solution sizing tool)
- ECS FastPass
- ECS Appliance in EMC Community
- Enablement Center for ECS
- Presales Index
- ISV Partner Validation Registry