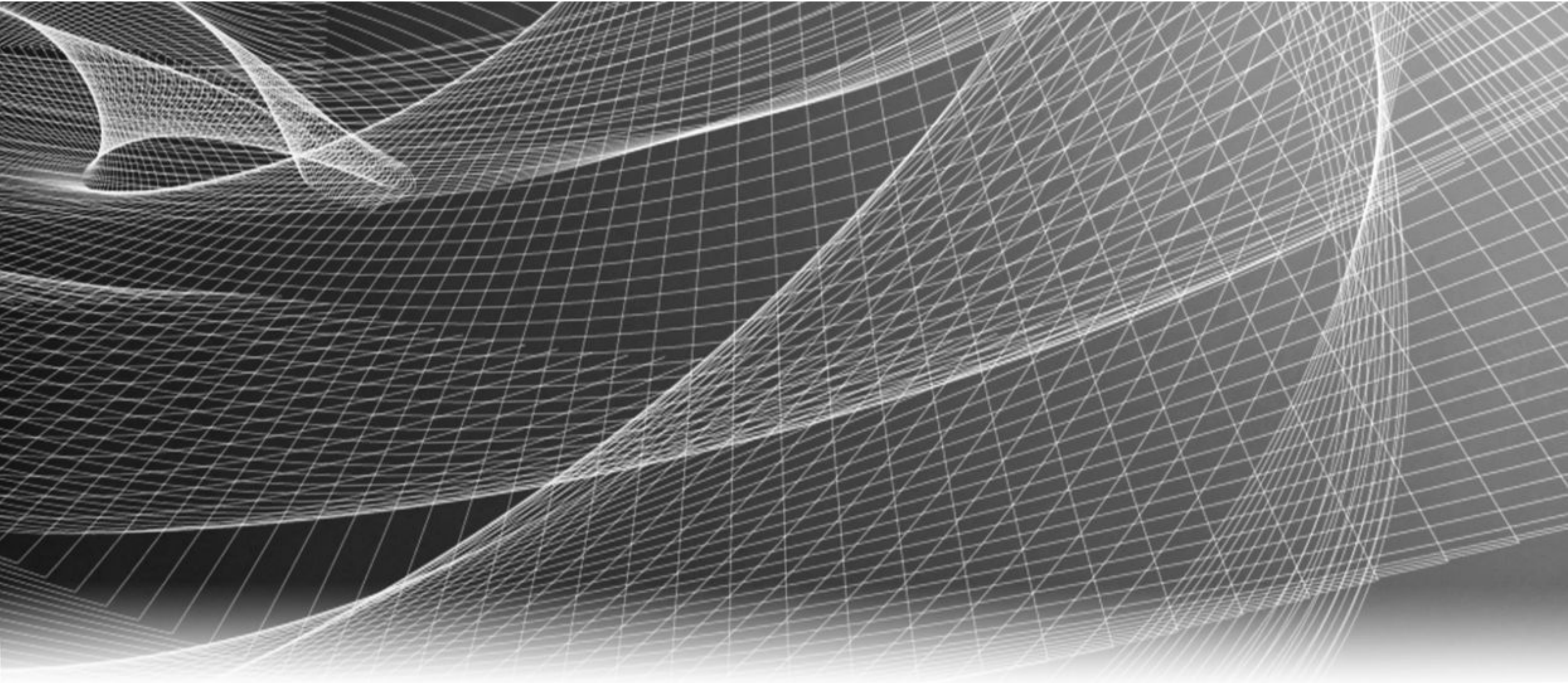


Elastic Cloud Storage (ECS)

Version 2.2

Planning Guide

302-002-524

02

Published February, 2016

EMC Corporation
Hopkinton, Massachusetts 01748-9103
1-508-435-1000 In North America 1-866-464-7381
www.EMC.com

CONTENTS

FIGURES

TABLES

CHAPTER 1

What is ECS?

Overview

ECS is a complete software-defined cloud storage platform that supports the storage, manipulation, and analysis of unstructured data on a massive scale on commodity hardware. ECS is specifically designed to support mobile, cloud, big data, and social networking applications. It can be deployed as a turnkey storage appliance or as a software product that can be installed on a set of qualified commodity servers and disks.

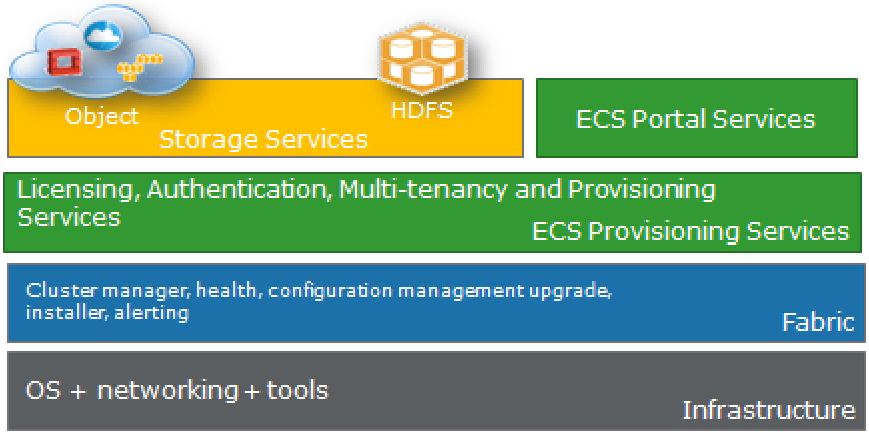
The ECS scale-out, geo-distributed architecture is a cloud platform that provides:

- Lower cost than public clouds
- Unmatched combination of storage efficiency and data access
- Anywhere read/write access with strong consistency that simplifies application development
- No single point of failure to increase availability and performance
- Universal accessibility that eliminates storage silos and inefficient ETL/data movement processes

ECS platform

The ECS platform includes the following software layers and services:

Figure 1 ECS platform services


Object
HDFS
Storage Services
ECS Portal Services
Licensing, Authentication, Multi-tenancy and Provisioning Services
ECS Provisioning Services
Cluster manager, health, configuration management upgrade, installer, alerting
Fabric
OS + networking + tools
Infrastructure

Portal services

Portal services include interfaces for provisioning, managing, and monitoring storage resources. The interfaces are:

- GUI: A built-in browser-based graphical user interface called the ECS Portal.
- REST: A RESTful API that you can use to develop your own ECS Portal.
- CLI: A command-line interface that enables you to perform the same tasks as the browser-based interface.

Storage services

Storage services are provided by the unstructured storage engine (USE) which ensures data availability and protection against data corruption, hardware failures, and data center disasters. It enables global namespace management across geographically dispersed data centers and geo-replication. The USE enables the following storage services:

- Object service: Provides the ability to store, access, and manipulate unstructured data. The object service is compatible with existing Amazon S3, OpenStack Swift APIs, EMC CAS, and EMC Atmos APIs.
- HDFS: Enables you to use your portal storage infrastructure as a Big Data repository that you can run Hadoop analytic applications against (in-place).

Provisioning service

The provisioning service manages the provisioning of storage resources and user access. Specifically, it handles:

- User management: Keeps track of which users have rights to administer the system, provision storage resources, and access objects using REST requests. ECS supports both local and domain users.
- Authorization and authentication for all provisioning requests: Queries the authentication domain to determine if users are authorized to perform management, provisioning, and access operations.
- Resource management: Enables authorized users to create storage pools, Virtual Data Centers, and replication groups.
- Multi-tenancy: Manages the namespace that represents a tenant, and their associated buckets and objects.

Fabric service

The fabric service is a distributed cluster manager that is responsible for:

- Cluster health: Aggregates node-specific hardware faults and reports on the overall health of the cluster.
- Node health: Monitors the physical state of the nodes, and detects and reports faults.
- Disk health: Monitors the health of the disks and file systems. It provides raw, fast, lock-free read/write operations to the storage engine, exposes information about the individual disk drives and their status so the storage engine can place data across the disk drives according to the storage engine's built-in data protection algorithms.
- Software management: Provides command line tools for installing and running services, and for installing and upgrading the fabric software on nodes in the cluster.

Infrastructure service

This layer provides the Linux OS running on the commodity nodes and it implements network interfaces and other hardware-related tools.

The ECS Appliance

The ECS appliance is available in the following models:

- U-Series: Unstructured storage servers with separate disk array enclosures engineered to maximize storage capacity. Available in Gen1 and upgraded Gen2 hardware configurations
- C-Series: High-density compute servers with integrated disks engineered with greater compute capacity.

The tables below describe the appliance components by series:

Table 1 U-Series components

Component	Description
40U rack	EMC Titan D racks that include: • Four power drops: two per side • Single phase PDUs with three phase wye and delta available • Front and rear doors • Racked by EMC manufacturing
Private switch	One 1 GbE switch
Public switch	Two 10 GbE switches
Nodes	Intel-based unstructured server in 4 and 8 node configurations
Disk array enclosure (DAE)	60 disk DAE drawers with 3.5 inch drives. Gen1 hardware uses 6TB disks and Gen2 hardware uses 8TB disks.

Table 2 C-Series components

Component	Description
40U rack	EMC Titan D Compute racks that include: • Four horizontal power drops • Single phase PDUs with three phase wye and delta available • Front and rear doors • Racked by EMC manufacturing
Private switch	One or two 1 GbE switches. The second switch is required for configurations with more than six server chassis (24 nodes).
Public switch	Two or four 10 GbE switches. The third and four switches are required for configurations with more than six server chassis (24 nodes).
Nodes	Intel-based unstructured server in 8 through 48 node configurations.
Disks	12 6TB 3.5 inch drives integrated with each node.

Network interfaces

Nodes have the following network interfaces:

- Public: A 10GbE interface that handles all network traffic. The interface is connected to the rack's 10GbE switches in a bonded configuration. The 10GbE switches are uplinked to the customer network through 1-4 10GbE uplinks.
- Private: A 1GbE interface used for internal administrative operations. All of the interfaces are private and are reserved for use by ECS traffic. Each node is automatically assigned two private IP addresses using the following scheme:

- 192.168.219.*port_number*: This network is used for installation and maintenance activities. It supports only rack-local traffic.
- 169.254.*Rack_ID.port_number*: This network handles the distributed configuration service for nodes in the cluster. It supports only data-center local traffic.

Provisioning storage resources

After you deploy ECS, you can use one of the ECS Portal services interfaces to provision the storage resources so that they can be used by S3, Swift, CAS, or Atmos applications.

Virtual data center (VDC)

VDCs are the top-level ECS resources. They are logical constructs that represent the collection of ECS infrastructure you want to manage as a cohesive unit. You can create a VDC to manage the resources of one or more physical racks, but the ECS resources in a single VDC must be part of the same Nile Area Network (NAN). A VDC is also referred to as a site or a zone.

You can deploy ECS software in multiple data centers to create a geo-federation. In a geo-federation, ECS behaves as a loosely coupled federation of autonomous virtual data centers in which you provision each VDC separately.

Storage pools

Storage pools allow you to logically partition the available storage resources (nodes) in a VDC. Storage pools provide the means for physically separating data based on application or multi-tenancy requirements. Storage pools require a minimum of four nodes. Data protection levels are defined by assigning storage pools to replication groups.

A storage pool with eight nodes or more can be configured as Cold Storage for infrequently accessed files. A cold archive uses an erasure coding scheme that lowers the storage overhead.

Replication groups

Replication groups are logical constructs that define where storage pool content is protected, and the locations from which data can be read without WAN traffic. Replication groups can be local or global. Local replication groups protect objects within the same VDC against disk or node failures. Global replication groups span multiple VDCs and protect objects against disk, node, and site failures.

The strategy for defining replication groups depends on multiple factors including your requirements for data resiliency, the cost of storage, and physical versus logical separation of data.

Namespaces

Namespaces enable ECS to handle multi-tenant operations. They are assigned replication groups. Each tenant is defined by a namespace and a set of users who can store and access objects within that namespace. Namespaces can represent a department within an enterprise, or can be a different enterprise. Users of one namespace cannot access objects from another namespace. The SEC Compliance feature is enabled at the namespace level.

Buckets

Buckets are containers for object data. Buckets are created in a namespace so they are only available to namespace users that have the appropriate permissions. Namespace users with the appropriate privileges can create buckets and objects within buckets for each object protocol using its API. Buckets can be configured to support HDFS. Buckets configured for HDFS access can be read and written using its object protocol and also the HDFS protocol.

Within a namespace, it is possible to use buckets as a way of creating subtenants.

Users and roles

ECS supports the following types of users and roles:

- System Admin: Users in this role configure the VDC, storage pools, replication groups, LDAP namespaces, buckets, and users. The System Admin can also configure namespaces and perform namespace administration or they can assign a user who belongs to the namespace as the Namespace Admin. ECS has a root user account which is assigned to the System Admin role and can be used to perform initial configuration.
- System Monitor: Users in this role, can view all configuration data, but cannot make any changes. Local system monitors can change their passwords.ECS has a root user account which is assigned to the System Admin role and can be used to perform initial configuration.
- Namespace Admin: Users in this role configure namespace settings, such as quotas and retention periods, and can map domain users into the namespace and assign local users as object users for the namespace. Namespace Admin operations can also be performed from programmatic clients using the ECS REST API or the ECS Portal.
- Object user: Object users are end-users of ECS object storage. They access storage through object clients using the ECS supported access protocols (S3, Swift, CAS, or Atmos applications). Object users can be given privileges to read and write buckets and objects, within the namespace they are assigned to.

Monitoring and diagnostics

ECS provides monitoring, diagnostics, and event auditing through the ECS Portal. Monitoring pages allow overviews of storage, resources, services, and events. The Monitoring pages allow you to drill down to get the right view of diagnostic data. The Dashboard is the first page you see upon logging into the portal. It provides a quick summary of monitoring data.

CHAPTER 2

New Features

New features in ECS 2.2

New features and additions for ECS 2.2.

Cold Storage Archives

Cold storage archives store objects that do not change frequently and do not require the more robust default erasure coding (EC) scheme. The EC scheme used here is 10 data fragments and 2 coding fragments (10 + 2) compared to 10 + 4 for regular archives. This scheme yields less storage overhead. The efficiency is 1.2x.

You can specify a cold archive when creating a new storage pool with at least 6 nodes. After the storage pool is created, the EC scheme cannot be changed. This scheme can support the loss of a single node or two disks on two separate nodes. A single VDC with multiple storage pools can have both regular archives and cold archives.

Data at Rest Encryption (D@RE)

EMC 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. It is a required feature in financial and healthcare uses cases needing regulatory compliance.

ECS D@RE supports FIPS-140-2 Level 1 compliance using an AES 256-bit encryption algorithm.

D@RE can be applied at the bucket or namespace level in the ECS portal or with the ECS Management API. Support at the object level is also available using the Amazon S3 SSE constructs.

D@RE provides automated key management and encrypts inline and then stores the encrypted data on ECS storage media. Keys are segregated at the namespace level. User-supplied keys can be used with the S3 API.

D@RE requires an ECS license with D@RE specified.

HDFS

For ECS 2.2, the ECS HDFS is certified against Hadoop 2.7. Certified applications/components include HDFS, MapReduce, Pivotal HAWQ (requires ECS 2.2 HF1 or later), Yarn, Hbase, Hive, Pig and Zookeeper.

In addition, ECS 2.2 HDFS adds support for:

- ECS HDFS as the default filesystem in simple secure mode and Kerberos secure mode
- HDFS superuser and supergroup
- HDFS extended ACLs and extended attributes
- Hadoop proxy user.

Hortonworks support has been enhanced by integration with the Ambari stack to deploy and configure the Hadoop cluster with ECS as the the HDFS.

ECS 2.2 simplifies S3 cross-head access by allowing a default group (and permissions) to be assigned to objects created using the object protocol. This simplifies and enables immediate access to object data from HDFS.

ECS 2.2 simplifies the support for the migration from a Hadoop cluster using simple security to a Hadoop cluster secured by Kerberos. Files and directories, that have been created in the ECS HDFS in simple mode, can be accessed by Hadoop users and processes after migration to Kerberos.

ECS 2.2, adds support for LOCAL_USER mode. The concept of an anonymous user has been removed. In a simple, non-Kerberos cluster, LOCAL_USER mode is used where ownership (user and group) is assigned to a local Hadoop user, as is the case when using a regular Hadoop filesystem.

SEC Compliance

ECS supports SEC Rule 17a-4(f) standard for electronic record storage.

Compliance has three components:

- Platform hardening: addressing common security vulnerabilities in software.
- Policy-based record retention: limiting the ability to change retention policies for records under retention.
- Compliance reporting: a system agent periodically reports the system's compliance status.
- Compliance is supported on ECS Appliances and not on ECS Software only installations.

Metadata Search

In ECS 2.2, the ECS S3-compatible protocol automatically associates system metadata with an object and allows users to associate custom metadata with an object. The metadata is in the form of name-value pairs.

The metadata search facility enables ECS to maintain an index of the objects in a bucket, based on their associated metadata, and allows S3 object clients to search for objects within buckets based on the indexed metadata, using a rich query language.

The metadata fields for which search indexes will be maintained (search keys) are configured for a bucket from the ECS Portal, the ECS Management REST API, or the S3 REST API.

Note

The metadata search feature cannot be enabled on a bucket when Data at Rest Encryption (D@RE) is enabled.

CAS Query Support

The CAS query API is now automatically available for all CAS buckets. See *Data Access Guide: Configure support for CAS SDK applications with the ECS Portal* for more information.

Bucket Tagging

In ECS 2.2, tags in the form of name-value pairs can be assigned to a bucket using the ECS Portal or the ECS Management REST API, enabling object data stored in the bucket to be categorized. For example, bucket data can be associated with a cost-center or project. Bucket tags and values are included in the metering reports for buckets and can be read by custom clients using the ECS Management REST API.

Bucket Default Quota

In ECS 2.2, it is possible to assign a default bucket quota to a namespace so that all buckets created in the namespace will have the specified quota unless a quota is explicitly assigned to individual buckets.

Erasure Coding Rebalancing

ECS object data is stored in chunks and chunks are broken into fragments based on an erasure coding (EC) scheme in order to improve storage efficiency.

ECS 2.2, provides improved EC rebalancing so that when new nodes are introduced into a site, the fragment distribution is recalculated to take into account the new nodes and the fragments are distributed accordingly. The redistribution applies to all chunks, including those that are XOR'ed.

Geo Copy to All Sites

A replication group can now be configured to copy data to all sites in the replication group. This feature increases performance at the cost of storage efficiency.

ECS Portal Features

The ECS portal has the following improvements:

- A Getting Started Checklist app that guides the initial root user through basic configuration.
- A monitoring dashboard that creates a quick view of the health and status of the VDC.
- A new About this VDC feature that allows users to see the software version on all nodes in the VDC.
- Improved filtering options on monitoring pages
- Online documentation for each portal page is available by clicking the global help icon in the global menu.
- Popup page- and field-level help is available by clicking the buttons next to page and field titles.

Auditing and Alerting

The Events page has been divided into two viewing panels, separating system alerts from auditing information. Now critical hardware and software errors along with operational events such as quota and licensing issues are now presented separately

Support for AD Groups as Management Users

Active Directory Groups can be assigned system administrator, system monitor, and namespace administrator roles.

ESRS

EMC Secure Remote Services (ESRS) is an IP-based, bi-directional remote connection between the customer's EMC environment and EMC that enables proactive remote monitoring, diagnosis, and repair - assuring availability and optimization of our customer's EMC products and solutions. ESRS enables EMC Customer Service to efficiently and quickly resolve many service issues without the need for onsite dispatch, resulting in significant cost savings and improved Total Customer Experience. ESRS Virtual Edition replaces ConnectEMC in ECS 2.2 and later.

ECS Software on Commodity Hardware (ECS DIY)

ECS has begun qualifying third-party commodity hardware to run ECS Software. EMC partners with a customer to qualify a particular third-party hardware platform. To find out more about the program, please contact your EMC representative.

CHAPTER 3

Data protection

Overview

Learn about how ECS protects unstructured data against node, disk, and site failures through replication and erasure coding.

ECS ensures durability, reliability, and availability of objects by creating and distributing three copies of objects and their metadata across the set of nodes in the local site. After the three copies are successfully written, ECS erasure-codes the object copies to reduce storage overhead. It handles failure and recovery operations automatically with no additional backup software or devices required.

Storage service

The storage service layer handles data availability and protection against data corruption, hardware failures, and data center disasters.

The unstructured storage engine (USE) is part of the storage services layer. It is a distributed shared service that runs on each node, and it manages transactions and persists data to nodes. The USE enables global namespace management across geographically dispersed data centers through geo-replication.

The USE writes all object-related data (such as, user data, metadata, object location data) to logical containers of contiguous disk space known as *chunks*. Chunks are open and accepting writes, or closed and not accepting writes. After chunks are closed, the storage engine erasure-codes them. The storage engine writes to chunks in an append-only pattern so that existing data is never overwritten or modified. This strategy improves performance because locking and cache validation is not required for I/O operations. All nodes can process write requests for the same object simultaneously while writing to different chunks.

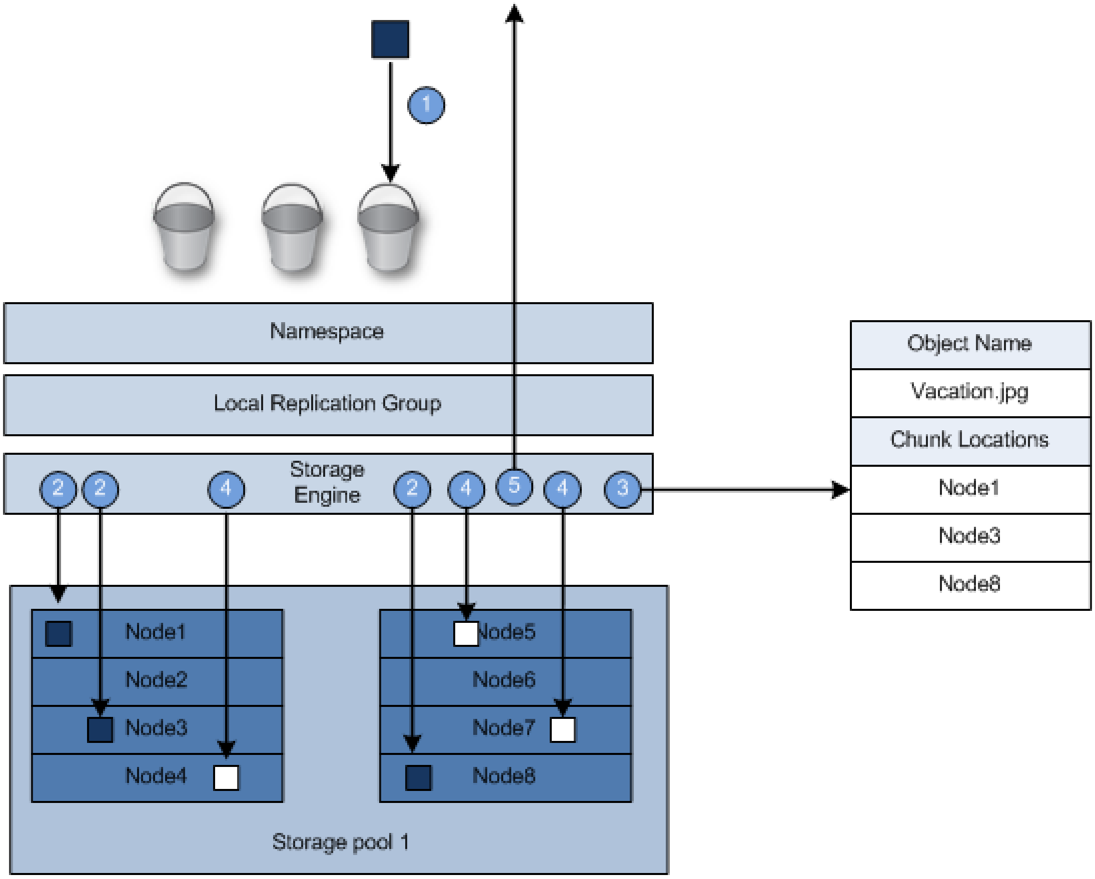
The storage engine tracks object location through an index that records object name, chunk id, and offset. Chunk location is separately tracked through an index that records chunk id and a set of disk locations. The chunk location index contains three disk location pointers before erasure coding, and multiple location pointers after erasure coding. The storage engine performs all of the storage operations (such as, erasure coding and object recovery) on chunks.

Object creates

Object creates: one VDC

The following figure shows how the storage engine writes object data when there is a single VDC. In this example, there is a single appliance deployed at the site, but the same principles apply when more appliances are deployed. The eight nodes are in a single storage pool within a single replication group.

Figure 2 Single site: object creates



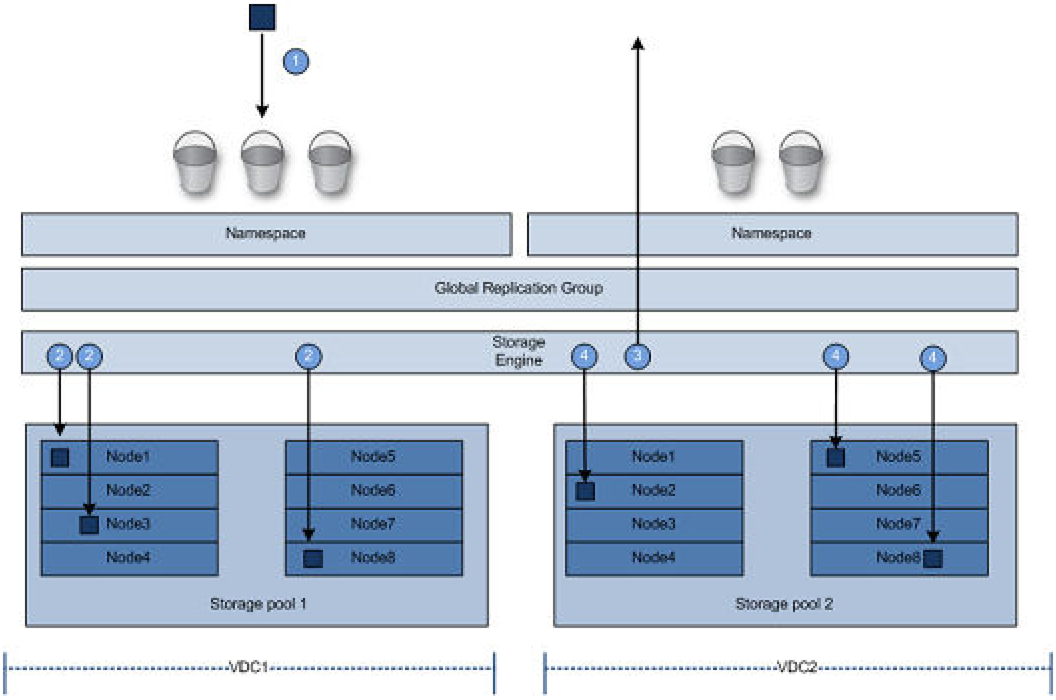
1. An application creates an object in a bucket.
2. The storage engine writes the object to one chunk. The disk locations corresponding to this chunk are on three different disks/nodes, so the writes go to three different disks/nodes in parallel. The storage engine can write the object to any of the nodes that belong to the bucket's replication group. The VDC where the object is created is the object's owner.
3. The storage engine records the disk locations of the chunk in the chunk location index, and the chunk id and offset in the object location index.
4. The storage engine writes the object location index to one chunk and the disk locations corresponding to the chunk to three different disks/nodes, so the writes go to three different disks/nodes in parallel. The index locations are chosen independently from the object chunk locations.
5. After all of the disk locations are written successfully, the storage engine acknowledges the write to the application.

When object chunks are full, the storage engine erasure-codes them. It does not erasure code the object location index chunks.

Object creates: federated VDCs (2 sites)

In a federated deployment of two VDCs, the storage engine writes object chunks to the local VDC and also to the remote VDC.

Figure 3 Two site: object creates



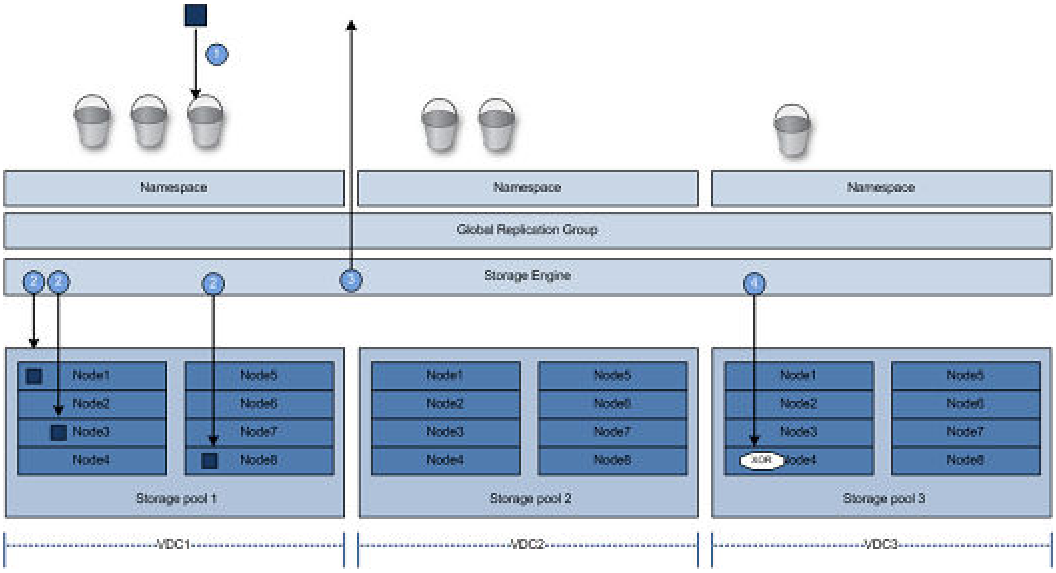
1. An application creates an object in a bucket.
2. The storage engine writes the object to one chunk at the site where it is ingested. The disk locations corresponding to this chunk are on three different disks/nodes, so the writes go to three different disks/nodes in parallel. The storage engine can write the object to any of the nodes that belong to the bucket's replication group. The storage engine records the disk locations of the chunk in the chunk location index, and the chunk id and offset in the object location index. The site where the object is originally ingested is the object's owner.
3. After all of the disk locations are written successfully, the storage engine acknowledges the write to the application.
4. The storage engine replicates the chunk to three nodes in the federated site. It records the chunk locations in the object location index (not shown in this diagram) also on three different nodes at the federated site.

When the chunks are full, the storage engine erasure-codes the object chunks. It does not erasure code the object location index chunks.

When two VDCs are in a replication group, both VDCs have a readable copy of the object.

Three sites: object creates

Figure 4 Object creates: federated VDCs (3 or more sites)



1. An application creates an object in a bucket.
2. The storage engine writes the object to one chunk at the site where it is ingested. The disk locations corresponding to this chunk are on three different disks/nodes, so the writes go to three different disks/nodes in parallel. It can write the object to any of the nodes that belong to the bucket's replication group. The storage engine records the disk locations of the chunk in the chunk location index, and the chunk id and offset in the object location index (not shown in this diagram). The VDC where the write received is the object's owner, and it contains a readable copy of the object.
3. After all of the disk locations are written successfully, the storage engine acknowledges the write to the application.
4. The storage engines replicates the chunks to nodes in another VDC within the replication group. To improve storage efficiency, the storage engine XOR's the chunks with other chunks from other objects also stored on the node.

When the chunks are full, the storage engine erasure-codes the XOR'd chunks. When possible, it writes XOR chunks directly in erasure-coded format without going through the replication phase. It does not erasure code the object location index chunks.

Replicate to All Sites Option

The Replicate to All Sites is an option available to an admin when creating a replication group.

A replication group with this feature enabled copies all data to all sites (VDCs) within the replication group. Having all data on all VDCs in the replication group provides data durability and improves local performance at all sites at the cost of storage efficiency.

This option can only be enabled at the time of creation and cannot be disabled later.

Object updates

When an application fully updates an object, the storage engine writes a new object (following the principles described earlier). The storage engine then updates the object location index to point to the new location. Because the old location is no longer referenced by an index, the original object is available for garbage collection.

Object reads

Object reads: single VDC

In a single site deployment, when a client submits a read request, the storage engine uses the object location index to find which chunks are storing the object, it retrieves the chunks or erasure-coded fragments, reconstructs and returns the object to the client.

Object reads: federated VDCs (2 sites)

In a two-site federation, the storage engine reads the object chunk or erasure coded fragments from the nodes on the VDC where the application is connected. In a two-site federation, object chunks exist on both sites.

Object reads: federated VDCs (3 sites or more sites)

If the requesting application is connected to the VDC that owns the object, the storage engine reads the object chunk or erasure coded fragments from the nodes on the VDC. If the requesting application is not connected to the owning VDC, the storage engine retrieves the object chunk or erasure coded fragments from the VDC that owns the object, copies them to the VDC the application is connected to, and returns the object to the application. The storage engine keeps a copy of the object in its cache in case another request is made for the object. If another request is made, the storage engine compares the timestamp of the object in the cache with the timestamp of the object in the owning VDC. If they are the same, it returns the object to the application; if the timestamps are different, it retrieves and caches the object again.

Erasure coding

ECS uses erasure coding (EC) to provides better storage efficiency without compromising data protection.

The storage engine implements the Reed Solomon 12 + 4 erasure coding scheme in which an object is broken into 12 data fragments and 4 coding fragments. The resulting 16 fragments are dispersed across the nodes in the local site. The storage engine can reconstruct an object from any of the 12 fragments.

Table 3 Storage overhead when deploying multiple sites

Number of sites	Storage Overhead
1	1.33
2	2.67
3	2.00
4	1.77
5	1.67
6	1.60
7	1.55
8	1.52

ECS requires a minimum of four nodes running the object service in a single site. It tolerates failures based on the number of nodes.

Table 4 Node failure tolerance at a single site

Total nodes	Node failure tolerance for writes
4	1
8 - 24	2

When an object is erasure coded, the original chunk data is present as a single copy that consists of 16 fragments dispersed throughout the cluster. When an object has been erasure-coded, ECS can read objects directly without any decoding or reconstruction. ECS only uses the code fragments for object reconstruction when there is hardware failure.

This erasure scheme is resilient up to four drive failures.

Erasure coding for cold archives

Cold storage archives store objects that do not change frequently and do not require the more robust default EC scheme. The EC scheme used here is 10 data fragments and 2 coding fragments (10 + 2). The efficiency is 1.2x.

You can specify a cold archive (Cold Storage) when creating a new storage pool. After the storage pool is created, the EC scheme cannot be changed. This scheme can support the loss of a single node or one drive out of six or two drives out of 12 on two separate nodes.

EC requirements

Table 5 Requirements for regular and cold archives compared

Use case	How enabled	Minimum required nodes	Minimum required disks	Recommended disks	EC efficiency	EC scheme
Regular archive	Default	4	16*	32	1.33x	12 + 4
Cold archive	Configured by admin	6	12*	24	1.2x	10 + 2

Note

*Since the minimum deployable configuration for a C-Series appliance is two appliances with 12 disks, 24 disks is the effective minimum.

Recovery on disk and node failures

ECS continuously monitors the health of the nodes, their disks, and objects stored in the cluster. Since ECS disperses data protection responsibilities across the cluster, it is able to automatically re-protect at-risk objects when nodes or disks fail.

Disk health

ECS reports disk health as Good, Suspect, or Bad.

- Good — The disk's partitions can be read from and written to.
- Suspect — The disk has not yet met the threshold to be considered bad.
- Bad — A certain threshold of declining hardware performance has been met. Once met, no data can be read or written.

ECS writes only to disks in good health; it does not write to disks in suspect or bad health. ECS reads from good disks and from suspect disks. When two of an object's chunks are located on suspect disks, ECS writes the chunks to other nodes.

Node health

ECS reports node health as Good, Suspect, Degraded, or Bad.

- Good: The node is available and responding to I/O requests in a timely manner. Internal health monitoring indicates that it is in good health.
- Suspect: The node is available, but is reporting internal health information such as a fan failure (if there are multiple fans), a single power supply failure (if there are redundant power supplies). Or, the node is unreachable by the other nodes, but it is visible to BMC probes and is in an unknown state.
- Degraded: The node is available but is reporting bad or suspect disks.
- Bad: The node is reachable, but internal health monitoring indicates poor health. For example, the node's fans are offline, the CPU temperature is too high, there are too many memory errors, and so on. Bad health can also be reported when the node is offline, and BMC probes indicate the health is not acceptable.

ECS writes only to nodes in good health; it does not write to nodes in suspect, degraded, or bad health. ECS reads from good and suspect nodes. When two of an object's chunks are located on suspect nodes, ECS writes two new chunks of it to other nodes. When a node is reported as suspect or bad, all of the disks it manages are also considered suspect or bad.

Data recovery

When there is a failure of a node or drive in the site, the storage engine:

1. Identifies the chunks or EC fragments affected by the failure.
2. Writes copies of the affected chunks or EC fragments to good nodes and disks that do not currently have copies.

Data rebalancing after adding new nodes

When the number of nodes at a site is expanded due to the addition of new racks or storage nodes, new erasure coded chunks are allocated to the new storage and existing data chunks are redistributed (rebalanced) across the new nodes. Four or more nodes must exist for erasure coding of chunks to take place. Addition of new nodes over and above the required 4 nodes will result in EC rebalancing.

The redistribution of EC fragments is performed as a background task so that the chunk data continues to be accessible during the redistribution process. In addition, the new fragment data is distributed as a low priority to minimize network bandwidth consumption.

Fragments are redistributed according to the same EC scheme with which they were originally encoded. So, if a chunk was written using the cold storage EC scheme, the cold storage scheme will be used when creating the new fragments for redistribution.

Site fail over and recovery

ECS provides protection against a site failure due to a disaster or other problem that causes a site to go offline or to be disconnected from the other sites in a geo-federated deployment.

Temporary site failure

Temporary site failures occur when network connectivity is interrupted between federated VDCs or when a VDC goes down temporarily. When a VDC goes down, the **Replication Group** page displays the status `Temporarily unavailable` for the VDC that is unreachable.

When buckets are configured with the **Access During Outage** property set to **On**, applications can read objects while connected to any site. When applications are connected to a site that is not the bucket's owner, the application must explicitly access the bucket to write to it or to view the contents. If an application modifies an object or bucket while connected to a VDC that is not the owner, the storage engine transfers ownership to the site where the change is initiated.

The following operations cannot be completed at any site in the geo-federation until the temporary failure is resolved regardless of the **Access During Outage** setting:

- Bucket: create or rename bucket, modify bucket properties, list buckets for a namespace when the namespace owner site is not reachable
- Namespace: create
- User: create

After the sites are reconnected, the storage engine starts a resync operation in the background. Use the portal's **Monitor** › **Recovery Status** to monitor the progress of the resync operation.

Permanent site fail over

If a disaster occurs at a site and the VDC cannot be brought back online, you must delete it.

CHAPTER 4

Planning an ECS Installation

Overview

Learn about the physical environment, data center, and multi-site requirements as well as the private management network topologies.

Site preparation

Review the Site Preparation Guide to learn about the environmental requirements associated with the 40U-D cabinet used by the ECS Appliance.

ECS installation readiness checklist

Review this list for the infrastructure components required for a successful installation.

An ECS appliance deployment consists of the following components:

- One or more racks.
  - The rack must be up linked to the customer network for both data traffic and remote management.

    The rack and all nodes must be powered on.

    The nodes must have valid IP addresses assigned by DHCP or configured statically .
- Infrastructure requirements: The data center environment must include the following servers that are reachable from all nodes.
  - DHCP server (if you are assigning IP addresses via DHCP)
  - DNS server (or forwarder)
  - NTP server
  - SMTP server

SSH must be enabled on all nodes.

The following ports are opened and used by the installer:

- Docker registry: 5000
- Lifecycle agent: 9240
- Object: 9020-9025,9040,9091,9094-9098,8088,9898,1095,1096,1098,9100,9101,9111,3218
- ZooKeeper: 9277,9278,9279

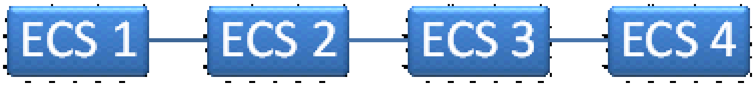
See the *Security Guide* for the list of ports that must be open.

Connecting ECS appliances in a single site

The ECS appliance management networks are connected together through the Nile Area Network. The NAN is created by connecting either port 51 or 52 to another turtle switch of another ECS appliance. Through these connections nodes from any segment can communicate to any other node in the NAN.

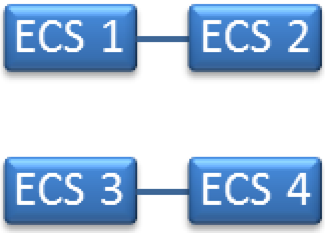
The simplest topology to connect the ECS appliances together does not require extra switch hardware. All the turtle switches can be connected together a linear or daisy chain fashion.

Figure 5 Linear or daisy-chain topology



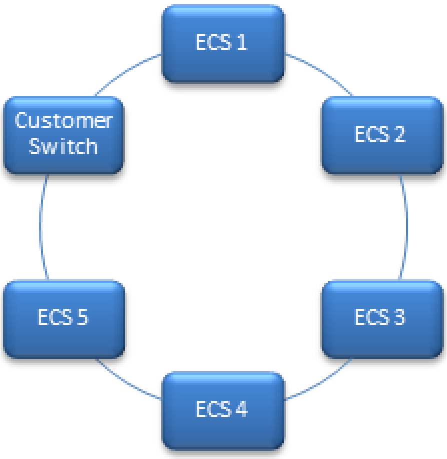
In this topology, if there is a loss of connectivity a split-brain can occur.

Figure 6 Linear or daisy-chain split-brain



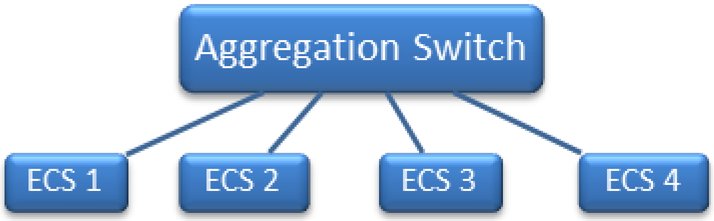
For a more reliable network, the ends of the daisy chain topology can be connected together to create a ring network. The ring topology is more stable because it would require two cable link breaks in the topology for a split-brain to occur. The primary drawback to the ring topology is that the RMM ports cannot be connected to the customer network unless an external customer or aggregation switch is added to ring.

Figure 7 Ring topology



The daisy-chain or ring topologies are not recommended for large installations. When there are four or more ECS appliances, an aggregation switch is recommended. The addition of an aggregation switch in a star topology can provide better fail over by reducing split-brain issues.

Figure 8 Star topology



Multi-site requirements

When planning for a multi-site ECS installation, ensure these requirements are met:

- A minimum of two VDCs is required.
- Each VDC in the multi-site configuration requires IP connectivity to the other VDCs.
  - Network latency: Ensure a maximum latency of 1000 ms between sites.
  - Free storage: If your disaster plan includes running for a period of time with one site permanently failed (instead of promptly recovering the site), each site needs enough free storage across all sites to accommodate data rebalancing. Across all sites, the amount of free space left should be, in total:

    ```
    free space across n sites =1.33*x/(n-1)/(n-2)
    ```

    where `x` is the total amount of user data across all n sites.

    This amount of free space is not required if you add a new site soon after the fail over, and do not continue to operate with (N-1) sites indefinitely.