DELL Technologies

Technisches Whitepaper

ECS – Übersicht und Architektur

Zusammenfassung

Dieses Dokument enthält eine technische Übersicht und das Design von ECS™, der softwarebasierten Objektspeicherplattform auf Cloud-Ebene von Dell EMC™.

Februar 2021

H14071.18

Überarbeitungen

Datum	Beschreibung
Dezember 2015	Erstausgabe
Mai 2016	Aktualisiert für 2.2.1
September 2016	Aktualisiert für 3.0
August 2017	Aktualisiert für 3.1
März 2018	Aktualisiert für 3.2
September 2018	Aktualisiert für Gen3-Hardware
Februar 2019	Aktualisiert für 3.3
September 2019	Aktualisiert für 3.4
Februar 2020	Aktualisiert mit ECSDOC-628-Änderungen
Mai 2020	Aktualisiert für 3.5
November 2020	Aktualisiert für 3.6
Februar 2021	Aktualisiert für 3.6.1

Mitwirkung

Dieses Whitepaper wurde erstellt von:

Autor: Zhu, Jarvis

Inhaltsverzeichnis

Zusammenfassung

Unternehmen benötigen Optionen für die Nutzung von Public-Cloud-Services mit der Zuverlässigkeit und Kontrolle einer Private-Cloud-Infrastruktur. Dell EMC ECS ist eine softwarebasierte, IPv6-unterstützte Objektspeicherplattform auf Cloud-Ebene, die S3-, Atmos-, CAS-, Swift-, NFSv3- und HDFS-Storage-Services auf einer einzigen, modernen Plattform bietet.

Mit ECS können Unternehmen auf einfache Weise eine global verteilte Infrastruktur unter einem einzigen globalen Namespace managen, der den Zugriff auf Inhalte von überall aus erlaubt. ECS-Kernkomponenten sind für Flexibilität und Ausfallsicherheit in Schichten angelegt. Jede Schicht ist abstrahiert und kann unabhängig von den anderen skaliert werden, was für hohe Verfügbarkeit sorgt.

Der einfache RESTful-API-Zugriff für Speicherservices wird von Entwicklern zunehmend eingesetzt. Die Verwendung von HTTP-Semantik wie GET und PUT vereinfacht die erforderliche Anwendungslogik im Vergleich zu herkömmlichen, aber vertrauten pfadbasierten Dateivorgängen. Darüber hinaus ist das zugrunde liegende Speichersystem von ECS in hohem Maße konsistent, was bedeutet, dass es eine zuverlässige Reaktion gewährleistet. Anwendungen, die eine zuverlässige Datenbereitstellung gewährleisten müssen, erreichen dies ohne komplexe Codelogik durch den Einsatz von ECS.

1 Einleitung

Dieses Dokument enthält eine Übersicht über die Dell EMC ECS-Objektspeicherplattform. Es beschreibt die ECS-Designarchitektur und Kernkomponenten wie Speicherservices und Data-Protection-Mechanismen.

1.1 Zielgruppe

Dieses Papier richtet sich an alle, die den Nutzen und die Architektur von ECS verstehen möchten. Ziel ist es, Kontext mit Links zu weiterführenden Informationen bereitzustellen.

1.2 Geltungsbereich

Im Mittelpunkt dieses Dokuments steht in erster Linie die ECS-Architektur. Es enthält keine Installations-, Administrations- und Upgradeverfahren für ECS-Software oder -Hardware. Außerdem werden keine Details zur Verwendung und Erstellung von Anwendungen mit den ECS APIs behandelt.

Updates dieses Dokuments werden in regelmäßigen Abständen, im Allgemeinen bei der Veröffentlichung von Hauptversionen oder neuen Funktionen, durchgeführt.

2 Nutzen von ECS

ECS bietet Unternehmen und Serviceanbietern, die eine für schnelles Datenwachstum konzipierte Plattform suchen, einen erheblichen Mehrwert. Zu den Hauptvorteilen und Merkmalen von ECS, die es Unternehmen ermöglichen, verteilte Inhalte global zu managen und zu speichern, gehören die Folgenden:

- **Cloud-Ebene:** ECS ist eine Objektspeicherplattform für herkömmliche Workloads und Workloads der nächsten Generation. Die mehrschichtige softwarebasierte Architektur von ECS ermöglicht unbegrenzte Skalierbarkeit. Highlights der Funktionen umfassen:
  - Global verteilte Objektinfrastruktur
  - Skalierung auf mindestens Exabytegröße ohne Einschränkung der Kapazität auf Speicherpool, Cluster oder Verbundumgebung
  - Keine Begrenzungen für die Anzahl der Objekte in einem System, Namespace oder Bucket
  - Effizient bei kleinen und großen Datei-Workloads ohne Begrenzung der Objektgröße
- **Flexible Bereitstellung:** ECS bietet eine bisher unerreichte Flexibilität mit Funktionen wie:
  - Appliance-Bereitstellung
  - Reine Softwarebereitstellung mit Unterstützung für zertifizierte oder angepasste Hardware nach Branchenstandard
  - Unterstützung mehrerer Protokolle: Objekt (S3, Swift, Atmos, CAS) und Datei (HDFS, NFSv3)
  - Mehrere Workloads: Moderne Apps und langfristige Archivierung
  - Sekundärer Speicher für Data Domain Cloud Tier und Isilon mit CloudPools
  - Unterbrechungsfreie Upgradepfade auf ECS-Modelle der aktuellen Generation
- **Enterprise-Klasse:** ECS bietet Kunden mehr Kontrolle über ihre Datenressourcen mit Speicher der Enterprise-Klasse in einem sicheren und konformen System mit Funktionen wie:
  - Data-at-Rest (D@RE) mit Schlüsselrotation und externem Key-Management
  - Verschlüsselte standortübergreifende Kommunikation
  - Deaktiviert standardmäßig die Anschlüsse 9101/9206, um Unternehmen die Einhaltung von Compliance-Policy zu ermöglichen
  - Reporting, Policy- und ereignisbasierte Datensatzaufbewahrung sowie Plattformverstärkung für Compliance mit SEC Rule 17a-4 (f) einschließlich erweitertem Aufbewahrungsmanagement, z. B. für die gesetzliche Aufbewahrungsfrist und Min-/Max-Governance
  - Compliance mit Guidelines des DISA (Defense Information System Agency) STIG (Security Technical Implementation Guide) zur Sicherheitsverstärkung
  - Authentifizierung, Autorisierung und Zugriffskontrollen mit Active Directory und LDAP
  - Integration in Monitoring- und Warnmeldungsinfrastruktur (SNMP-Traps und SYSLOG)
  - Erweiterte Funktionen für Unternehmen (Mehrmandantenfähigkeit, Kapazitätsmonitoring und Warnmeldungen)
- **TCO-Reduktion:** Mit ECS können die Gesamtbetriebskosten (TCO) im Vergleich zu herkömmlichem Speicher und Public-Cloud-Speicher drastisch reduziert werden. Es bietet für langfristige Aufbewahrung sogar eine niedrigere TCO als Bandlaufwerke. Die Funktionen umfassen u. a.:
  - Globaler Namespace
  - Hohe Performance bei kleinen und großen Dateien
  - Nahtlose Centera-Migration
  - Vollständig konform mit Atmos REST
  - Geringer Managementoverhead
  - Kleine Stellfläche im Rechenzentrum
  - Hohe Speicherauslastung

Das Design von ECS ist für die folgenden primären Anwendungsbeispiele optimiert:

- **Moderne Anwendungen:** ECS wurde für moderne Entwicklungen wie Web-, Mobil- und Cloud-Anwendungen der nächsten Generation konzipiert. Die Anwendungsentwicklung wird durch stark konsistenten Speicher vereinfacht. Zusammen mit gleichzeitigem Lese-/Schreibzugriff von mehreren Standorten und Nutzern müssen Entwickler ihre Apps mit sich ändernder und wachsender ECS-Kapazität nicht neu codieren.
- **Sekundärer Speicher:** ECS wird als sekundärer Speicher verwendet, um primären Speicher für selten genutzte Daten freizugeben, während gleichzeitig ein angemessener Zugriff ermöglicht wird. Beispiele hierfür sind Policy-basierte Tiering-Produkte wie Data Domain Cloud Tier und Isilon CloudPools. GeoDrive, eine Windows-basierte Anwendung, bietet Windows-Systemen direkten Zugriff auf ECS, um Daten zu speichern.
- **Archiv mit Geoschutz:** ECS dient als sichere und kostengünstige lokale Cloud für Archivierung und langfristige Aufbewahrung. Die Verwendung von ECS als Archiv-Tier kann die primäre Speicherkapazität deutlich reduzieren. Um eine bessere Speichereffizienz bei Anwendungen als Archiv für inaktive Daten zu ermöglichen, ist zusätzlich zum Erasure-Coding-Standard (EC) 12+4 das Schema 10+2 verfügbar.
- **Globales Content Repository:** Repositories mit unstrukturierten Inhalten, die Daten wie Bilder und Videos enthalten, werden oft in teuren Speichersystemen gespeichert, sodass Unternehmen das enorme Datenwachstum nicht kosteneffizient handhaben können. ECS ermöglicht die Zusammenfassung mehrerer Speichersysteme in einem einzigen, global zugänglichen und effizienten Content Repository.
- **Speicher für das Internet der Dinge:** Das Internet der Dinge (Internet of Things, IoT) bietet neue Umsatzmöglichkeiten für Unternehmen, die Wert aus Kundendaten extrahieren können. ECS bietet eine effiziente IoT-Architektur für die Erfassung unstrukturierter Daten in einem extrem großen Maßstab. ECS ist die ideale Plattform für die Speicherung von IoT-Daten, unabhängig von Anzahl und Größe der Objekte oder angepassten Metadaten. ECS kann auch einige analytische Arbeitsabläufe rationalisieren, indem es die Analyse von Daten ohne zeitaufwendige Extraktions-, Transformations- und Ladeprozesse (ETL) direkt auf der ECS-Plattform ermöglicht. Hadoop-Cluster können Abfragen mithilfe von Daten ausführen, die von einer anderen Protokoll-API, z. B. S3 oder NFS, auf ECS gespeichert werden.
- **Videoüberwachungsnachweis-Repository:** Im Gegensatz zu IoT-Daten werden Videoüberwachungsdaten mit viel weniger Objekten, aber wesentlich höherer Kapazitätsbelastung pro Datei gespeichert. Dabei ist die Authentizität der Daten zwar wichtig, die Datenaufbewahrung hingegen nicht so sehr. ECS kann als kostengünstiger Erstspeicher oder als sekundärer Speicherort für diese Daten dienen. Videomanagementsoftware kann die umfangreichen angepassten Metadatenfunktionen für das Tagging von Dateien mit wichtigen Details wie Kamerastandort, Aufbewahrungsvorgaben und Data-Protection-Anforderungen nutzen. Außerdem können Metadaten verwendet werden, um die Datei in einen schreibgeschützten Status zu versetzen und eine Beweismittelkette für die Datei sicherzustellen.
- **Data Lakes und Analysen:** Daten-Storage und Analysen sind zu einem Alleinstellungsmerkmal im Wettbewerb und zu einer primären Quelle für die Wertschöpfung für Unternehmen geworden. Die Transformation von Daten in eine wertvolle Unternehmensressource ist jedoch ein komplexes Thema, das schnell den Einsatz von Dutzenden von Technologien, Tools und Umgebungen mit sich bringen kann. ECS bietet eine Reihe von Services, die Kunden bei der Erfassung, Speicherung, Steuerung und Analyse von Daten in jeder Größenordnung unterstützen.

3 Architektur

ECS ist mit einigen Kerndesignprinzipien wie globalem Namespace mit starker Konsistenz, Scale-out-Funktion, sicherer Mehrmandantenfähigkeit; und hoher Performance für kleine und große Objekte konzipiert. ECS ist als vollständig verteiltes System nach dem Prinzip von Cloud-Anwendungen aufgebaut, bei denen jede Funktion im System als unabhängige Schicht konzipiert ist. Bei diesem Design kann jede Schicht horizontal über alle Nodes im System skaliert werden. Ressourcen werden über alle Nodes verteilt, um die Verfügbarkeit zu erhöhen und die Last zu verteilen.

In diesem Abschnitt wird ausführlich auf die ECS-Architektur und das Design der Software und Hardware eingegangen.

3.1 Überblick

ECS wird auf einem Satz qualifizierter Hardware nach Branchenstandard oder als sofort einsetzbare Speicher-Appliance bereitgestellt. Die Hauptkomponenten von ECS umfassen:

- **ECS-Portal- und Bereitstellungsservices:** API-basierte Web-UI und CLI für Selfservice, Automatisierung, Reporting und Management von ECS-Nodes. Diese Schicht verarbeitet auch Lizenzierungs-, Authentifizierungs-, Mehrmandantenfähigkeits- und Bereitstellungsservices wie die Namespace-Erstellung.
- **Datendienste:** Services, Tools und APIs für den Objekt- und Dateizugriff auf das System.
- **Speicher-Engine:** Kernservice, der für das Speichern und Abrufen von Daten, das Managen von Transaktionen und das Schützen und Replizieren von Daten lokal und zwischen Standorten verantwortlich ist.
- **Fabric:** Clusteringservice für Integritäts-, Konfigurations- und Upgrademanagement sowie Warnmeldungen.
- **Infrastruktur:** SUSE Linux Enterprise Server 12 als Basisbetriebssystem für die gebrauchsfertige Appliance oder qualifizierte Linux-Betriebssysteme für eine Hardwarekonfiguration nach Branchenstandard.
- **Hardware:** Eine sofort einsetzbare Appliance oder qualifizierte Hardware nach Branchenstandard.

Abbildung 1 zeigt eine grafische Darstellung dieser Schichten, die in den folgenden Abschnitten ausführlich beschrieben werden.

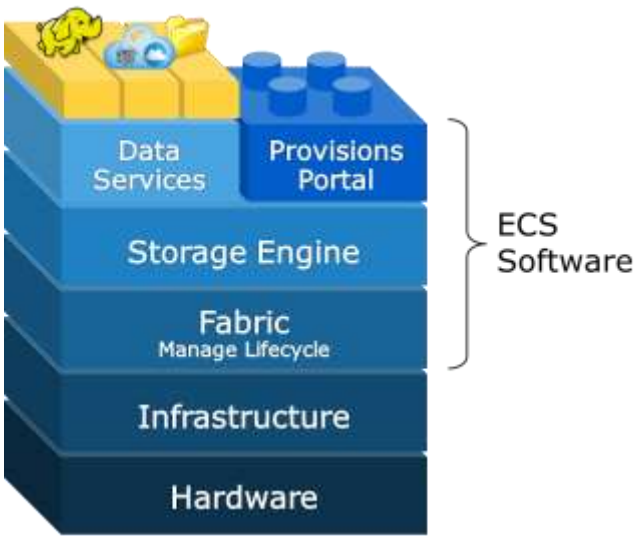

Abbildung 1 Schichten der ECS-Architektur

3.2 ECS-Portal- und Bereitstellungsservices

Speicheradministratoren managen ECS über die ECS-Portal- und Bereitstellungsservices. ECS bietet eine webbasierte GUI (Web-UI) für das Managen, Lizenzieren und Bereitstellen von ECS-Nodes. Das Portal hat unter anderem die folgenden umfassenden Reportingfunktionen:

- Kapazitätsauslastung pro Standort, Speicherpool, Node und Festplatte
- Performancemonitoring für Latenz, Durchsatz, Transaktionen pro Sekunde sowie Replikationsfortschritt
- Diagnoseinformationen, z. B. Recovery-Status für Nodes und Laufwerke

Das ECS-Dashboard bietet allgemeine Integritäts- und Performanceinformationen auf Systemlevel. Diese einheitliche Ansicht verbessert die Systemtransparenz insgesamt. Warnmeldungen benachrichtigen Nutzer über kritische Ereignisse, z. B. Kapazitätslimits, Quotenlimits, Festplatten- oder Node-Ausfälle sowie Softwarefehler. ECS bietet außerdem eine Befehlszeilenschnittstelle für das Installieren, Durchführen von Upgrades und Überwachen von ECS. Der Zugriff auf Nodes für die Verwendung der Befehlszeile erfolgt über SSH. Ein Screenshot des ECS-Dashboards wird in Abbildung 2 unten gezeigt.

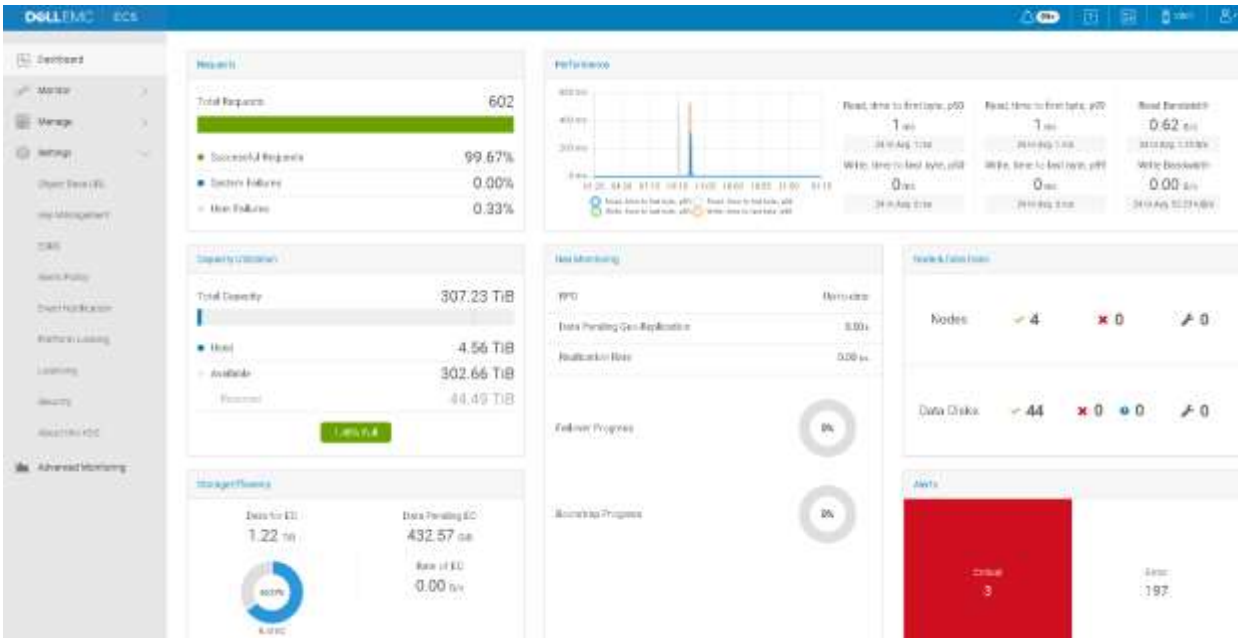

Abbildung 2 Web-UI-Dashboard von ECS

Detaillierte Performanceberichte sind auf der Benutzeroberfläche unter dem Ordner „Erweitertes Monitoring“ verfügbar. Die Berichte werden in einem Grafana-Dashboard angezeigt. Es sind Filter verfügbar, um nach bestimmten Namespaces, Protokollen oder Nodes zu suchen. Ein Beispiel für einen S3-Protokollperformancebericht finden Sie unten unter Abbildung 3.

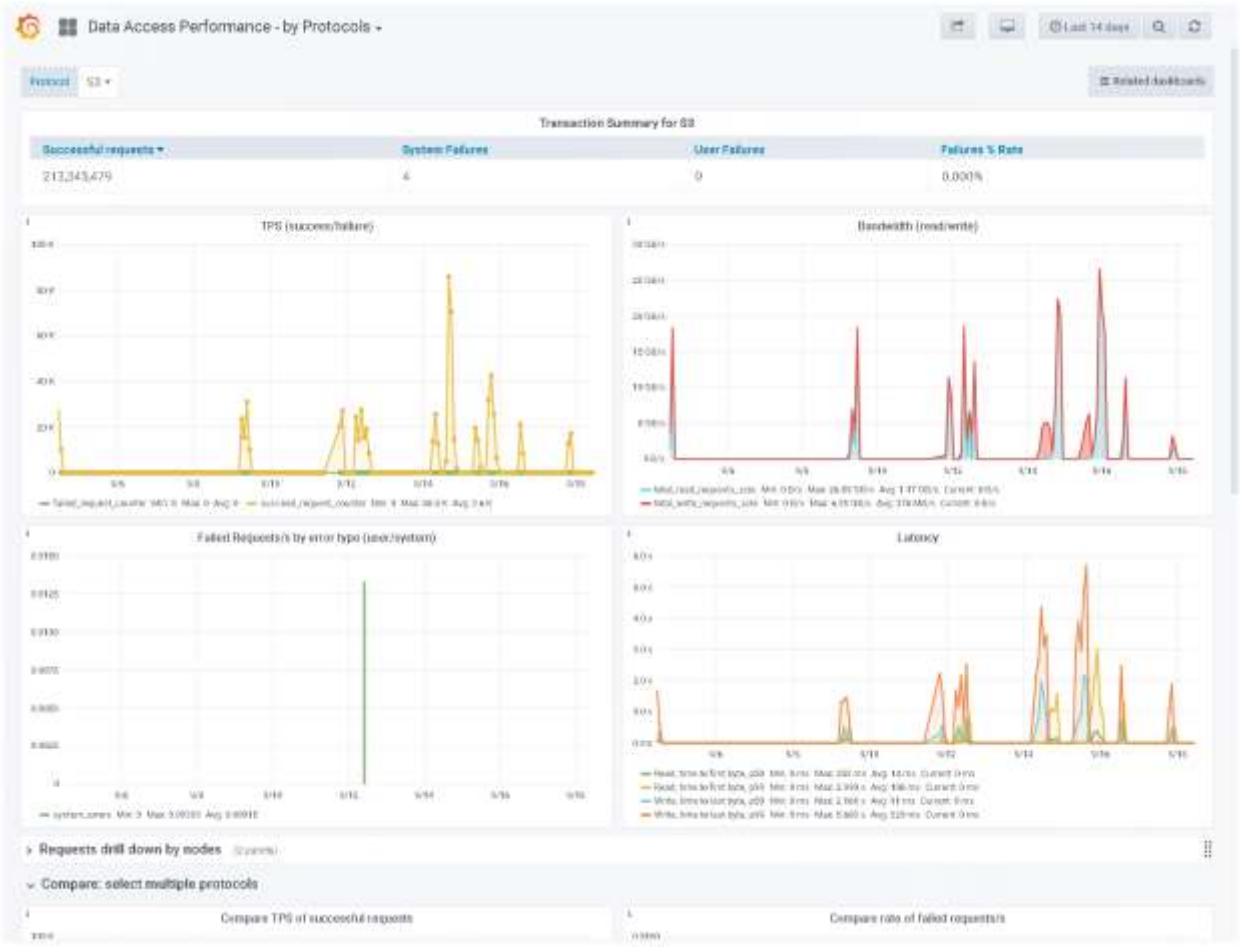

Abbildung 3 Visualisierung des erweiterten Monitorings mit Grafana

ECS kann auch über RESTful APIs gemanagt werden. Die Management-API ermöglicht es Nutzern, ECS innerhalb ihrer eigenen Tools, Skripte und neuen oder vorhandenen Anwendungen zu verwalten. Die ECS-Web-UI und die Befehlszeilentools basieren auf den ECS REST Management APIs.

ECS unterstützt die folgenden Ereignisbenachrichtigungsserver, die über die Web-UI, API oder CLI festgelegt werden können:

- SNMP-Server (Simple Network Management Protocol)
- Syslog-Server

Im *ECS-Administratorhandbuch* finden Sie weitere Informationen und Details zur Konfiguration von Benachrichtigungsservices.

3.3 Datendienste

Für den Zugriff auf ECS-Speicherservices werden Standardobjekt- und -dateimethoden verwendet. Bei S3, Atmos und Swift werden RESTful APIs über HTTP für den Zugriff verwendet. Bei CAS (Content-Addressable Storage) wird eine proprietäre Zugriffsmethode/ein SDK verwendet. ECS unterstützt nativ alle NFSv3-Verfahren mit Ausnahme von LINK. Auf ECS-Buckets kann jetzt über S3a zugegriffen werden.

ECS bietet Multiprotokollzugriff, wobei auf Daten, die über ein Protokoll aufgenommen werden, über ein anderes zugegriffen werden kann. Das bedeutet, dass Daten über S3 aufgenommen und über NFSv3 oder Swift geändert werden können oder umgekehrt. Dieser Multiprotokollzugriff unterliegt aufgrund von Protokollsemantik und Darstellungen der Konzeption des Protokolls einigen Ausnahmen. In Tabelle 1 sind die Zugriffsmethoden und die Interoperabilität der Protokolle angegeben.

Tabelle 1 Von ECS unterstützte Datendienste und Protokollinteroperabilität

Protokolle		**Unterstützt**	**Interoperabilität**
Objekt	S3	Zusätzliche Funktionen wie Updates im Bytebereich und Rich ACLs	HDFS, NFS, Swift
	Atmos	Version 2.0	NFS (nur pfadbasierte Objekte, keine Objekte mit Objekt-ID-Stil)
	Swift	V2 APIs, Swift und Keystone-v3-Authentifizierung	HDFS, NFS, S3
	CAS	SDK v3.1.544 oder höher	–
File	HDFS	Kompatibilität mit Hadoop 2.7	S3, NFS, Swift
	NFS	NFSv3	S3, Swift, HDFS, Atmos (nur pfadbasierte Objekte, keine Objekte mit Objekt-ID-Stil)

Datendienste, die auch als Head-Services bezeichnet werden, sind für die Annahme von Clientanfragen, das Extrahieren erforderlicher Informationen und die Übergabe an die Speicher-Engine zur weiteren Verarbeitung verantwortlich. Alle Head-Services werden zu einem einzigen Prozess kombiniert: *dataheadsvc*. Dieser wird in der Infrastrukturschicht ausgeführt. Dieser Prozess wird außerdem in einem Docker-Container mit dem Namen *object-main* gekapselt, der auf jedem Node in ECS ausgeführt wird. Im Abschnitt *Infrastruktur* dieses Dokuments wird Docker ausführlicher behandelt. Portanforderungen des ECS-Protokollservice wie Port 9020 für die S3-Kommunikation sind im neuesten *ECS-Sicherheitskonfigurationsleitfaden* verfügbar.

3.3.1 Objekt

ECS unterstützt S3, Atmos, Swift und CAS APIs für den Objektzugriff. Mit Ausnahme von CAS werden Objekte oder Daten über die HTTP- oder HTTPS-Aufrufe GET, POST, PUT, DELETE und HEAD geschrieben, abgerufen, aktualisiert und gelöscht. Bei CAS werden standardmäßige TCP-Kommunikation und bestimmte Zugriffsmethoden und -Aufrufe verwendet.

ECS bietet eine Möglichkeit zur Suche nach Objekten mithilfe einer umfassenden Abfragesprache. Hierbei handelt es sich um eine mächtige Funktion von ECS, mit der S3-Objektclients mithilfe von System- und angepassten Metadaten nach Objekten in Buckets suchen können. Obwohl die Suche über beliebige Metadaten möglich ist, kann ECS durch die Suche nach Metadaten, die speziell für die Indexierung in einem Bereich konfiguriert wurden, Abfragen schneller zurückgeben, insbesondere für Buckets mit Milliarden von Objekten.

Pro Bucket können bis zu 30 nutzerdefinierte Metadatenfelder indexiert werden. Metadaten werden zum Zeitpunkt der Bucket-Erstellung angegeben. Die Metadatensuchfunktion kann auf Buckets mit aktivierter serverseitiger Verschlüsselung aktiviert werden. Allerdings werden indexierte Nutzermetadaten, die als Suchschlüssel verwendet werden, nicht verschlüsselt.

Hinweis: Das Schreiben von Daten in Buckets, die für die Indexierung von Metadaten konfiguriert sind, beeinträchtigt die Performance. Die Auswirkungen auf den Betrieb steigen mit zunehmender Anzahl der indexierten Felder. Die Auswirkungen auf die Performance müssen sorgfältig geprüft werden, um zu bestimmen, ob Metadaten in einem Bucket indexiert werden sollen, und, falls ja, wie viele Indexe verwaltet werden müssen.

Bei CAS-Objekten bietet die CAS-Abfrage-API eine ähnliche Möglichkeit, anhand von Metadaten nach Objekten zu suchen, die für CAS-Objekte verwaltet werden und nicht explizit aktiviert werden müssen.

Weitere Informationen zu ECS APIs und APIs für die Metadatensuche finden Sie im neuesten *ECS-Leitfaden für den Datenzugriff*. Informationen zu Atmos und S3 SDKs finden Sie auf der GitHub-Website „Dell EMC Data Services SDK“ oder „Dell EMC ECS“. Informationen zu CAS finden Sie auf der Website der Centera-Community. Zugriff auf zahlreiche Beispiele, Ressourcen und Unterstützung für Entwickler finden Sie in der ECS-Community.

Clientanwendungen wie S3 Browser und Cyberduck bieten eine Möglichkeit, in ECS gespeicherte Daten schnell zu testen oder darauf zuzugreifen. ECS Test Drive wird kostenlos von Dell EMC bereitgestellt und ermöglicht den Zugriff auf ein öffentlich zugängliches ECS-System zu Test- und Entwicklungszwecken. Nach der Registrierung für ECS Test Drive erhalten REST-Endpunkte Nutzeranmeldedaten für jedes der Objektprotokolle. ECS Test Drive kann von allen Nutzern verwendet werden, um ihre S3-API-Anwendung zu testen.

Hinweis: Nur die Anzahl der Metadaten, die pro Bucket indexiert werden können, ist in ECS auf 30 beschränkt. Es gibt keine Beschränkung für die Gesamtzahl der angepassten Metadaten, die pro Objekt gespeichert werden, sondern nur die für die schnelle Suche indexierte Zahl.

3.3.2 HDFS

ECS kann Hadoop-Dateisystemdaten speichern. Da ECS ein Hadoop-kompatibles Dateisystem ist, können Unternehmen Big Data-Repositories auf ECS erstellen, die Hadoop-Analysen nutzen und verarbeiten können. Der HDFS-Datendienst ist kompatibel mit Apache Hadoop 2.7 mit Unterstützung für fein abgestimmte ACLs und ein erweitertes Dateisystemattribut.

ECS wurde mit Hortonworks (HDP 2.7) validiert und getestet. ECS bietet auch Unterstützung für Services wie YARN, MapReduce, Pig, Hive/Hiveserver2, HBase, Zookeeper, Flume, Spark und Sqoop.

3.3.2.1 Hadoop S3A-Unterstützung

ECS unterstützt den Hadoop S3A-Client zur Speicherung von Hadoop-Daten. S3A ist ein Open-Source-Connector für Hadoop, basierend auf dem offiziellen Amazon Web Services (AWS)-SDK. Es wurde entwickelt, um Speicherskalierungs- und Kostenprobleme zu beheben, die viele Hadoop-Administratoren mit HDFS hatten. Hadoop S3A verbindet Hadoop-Cluster mit jedem S3-kompatiblen Objektspeicher, unabhängig davon, ob es sich um eine Public, Hybrid oder On-Premises Cloud handelt.

Hinweis: Die S3A-Unterstützung ist in Hadoop 2.7 oder höher verfügbar.

Abbildung: 4 Hadoop- und ECS-Architektur

Wie in Abbildung: 4 gezeigt, verweist die S3A-Konfiguration auf die ECS-Objektdaten, wenn der Kunde das Hadoop-Cluster auf herkömmlichem HDFS eingerichtet hat, um alle HDFS-Aktivitäten durchzuführen. Auf jedem Hadoop HDFS-Node würde jede herkömmliche Hadoop-Komponente den S3A-Client von Hadoop verwenden, um die HDFS-Aktivität durchzuführen.

Hadoop-Konfigurationsanalyse mithilfe der ECS-Servicekonsole

Die ECS-Servicekonsole (SC) kann Ihre Hadoop-Konfigurationsparameter in Bezug auf Verbindungen zu ECS für S3A lesen und interpretieren. Darüber hinaus bietet SC eine Funktion *Get_Hadoop_Config*, die die Hadoop-Clusterkonfiguration liest und die S3A-Einstellungen auf Tippfehler, Fehler und Werte prüft. Wenden Sie sich an das ECS-Supportteam, um Unterstützung bei der Installation von ECS SC zu erhalten.

Privacera-Implementierung mit Hadoop S3A

Privacera ist ein Drittanbieter, der einen clientseitigen Hadoop-Agent und eine Integration in Ambari für die granulare Sicherheit von S3 (AWS und ECS) implementiert hat. Obwohl Privacera Cloudera Distribution of Hadoop (CDH) unterstützt, unterstützt Cloudera (ein anderer Drittanbieter) Privacera nicht auf CDH.

Hinweis: CDH-Nutzer müssen ECS IAM-Sicherheitsservices verwenden. Wenn Sie sicheren Zugriff auf S3A ohne Verwendung von ECS IAM wünschen, wenden Sie sich an das Supportteam.

Weitere Informationen zur S3A-Unterstützung finden Sie im neuesten *ECS-Datenzugriffshandbuch*.

Hadoop S3A-Sicherheit

ECS IAM ermöglicht es dem Hadoop-Administrator, Zugriffs-Policies einzurichten, um den Zugriff auf S3A Hadoop-Daten zu steuern. Sobald die Zugriffs-Policies definiert sind, gibt es zwei Nutzerzugriffsoptionen für Hadoop-Administratoren, die konfiguriert werden müssen:

- IAM Nutzer/Gruppen
  - Erstellen von IAM-Gruppen, die mit Policies verknüpft sind
  - Erstellen von IAM-Nutzern, die Mitglieder einer IAM-Gruppe sind

- SAML-Assertionen (Verbundsnutzer)
  - Erstellen von IAM-Rollen, die mit Policies verknüpft sind
  - Konfigurieren von CrossTrustRelationship zwischen Identity Provider (AD FS) und ECS, die AD-Gruppen IAM-Rollen zuordnen

ECS-Administrator und Hadoop-Administrator müssen zusammenarbeiten, um geeignete Policies vorab zu definieren. In den folgenden fiktiven Beispielen werden drei Arten von Hadoop-Nutzern beschrieben, für die wir Policies erstellen werden. Sie lauten:

- **Hadoop Administrator –** alle Vorgänge, außer Bucket erstellen und Bucket löschen
- **Hadoop Power User –** alle Vorgänge, außer Bucket erstellen, Bucket löschen und Objekte löschen
- **Hadoop Read Only User –** nur Objekte auflisten und lesen

Weitere Informationen zu ECS IAM finden Sie auf der Seite ECS IAM 37.

3.3.2.2 ECS HDFS-Clientsupport

ECS wurde in Ambariintegriert, wodurch Sie die HDFS-Client-JAR-Datei von ECS mühelos bereitstellen und ECS HDFS als Standarddateisystem in einem Hadoop-Cluster angeben können. Die JAR-Datei wird auf jedem Node in einem teilnehmenden Hadoop-Cluster installiert. ECS bietet Dateisystem- und Speicherfunktionen, die den Namens- und Daten-Nodes in einer Hadoop-Bereitstellung entsprechen. ECS rationalisiert den Workflow von Hadoop, da keine Migration von Daten zu einem lokalen Hadoop DAS erforderlich ist und/oder mindestens 3 Kopien erstellt werden müssen. Abbildung 5 unten zeigt die HDFS-Client-JAR-Datei von ECS, die auf jedem Hadoop-Compute-Node installiert ist, sowie den allgemeinen Kommunikationsfluss.

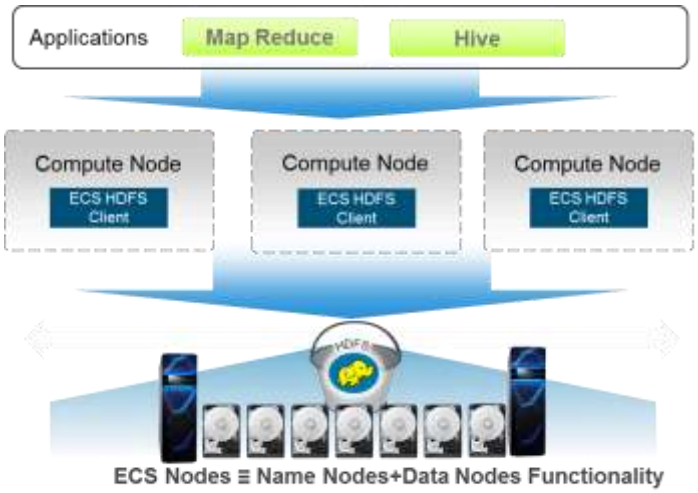

Abbildung 5 ECS als Namens- und Daten-Nodes für einen Hadoop-Cluster

Weitere Verbesserungen, die in ECS für HDFS hinzugefügt wurden, sind die Folgenden:

- **Proxynutzerauthentifizierung:** Identitätsbetrug für Hive, HBase und Oozie.
- **Sicherheit:** Serverseitige ACL-Durchsetzung und Hinzufügung von Hadoop-Super-User und Super-User-Gruppe sowie einer Standardgruppe für Buckets.

3.3.3 NFS

ECS bietet nativ Dateiunterstützung für NFSv3. Zu den Hauptfunktionen für den NFSv3-Dateidatendienst gehört Folgendes:

- **Globaler Namespace:** Dateizugriff von jedem beliebigen Node an jedem Standort.
- **Globale Sperre:** In NFSv3 wird das Sperren **nur empfohlen**. ECS unterstützt vorgabenkonforme Clientimplementierungen, die gemeinsames und ausschließendes bereichsbasiertes und obligatorisches Sperren ermöglichen.
- **Multiprotokollzugriff:** Zugriff auf Daten mit unterschiedlichen Protokollmethoden.

NFS-Exporte, -Berechtigungen und -Nutzergruppenzuweisungen werden mithilfe der Web-UI oder API erstellt. NFSv3-konforme Clients mounten Exporte mit Namespace- und Bucket-Namen. Hier ist ein Beispiel für den Befehl zum Mounten eines Bucket:

```
mount –t nfs –o vers=3 s3.dell.com:/namespace/bucket
```

Um Clienttransparenz während eines Node-Ausfalls zu erreichen, wird ein Load Balancer für diesen Workflow empfohlen.

ECS hat die anderen NFS-Serverimplementierungen wie *lockmgr*, *statd*, *nfsd* und *mountd* eng integriert, daher sind diese Services nicht von der zu managenden Infrastrukturschicht (dem Hostbetriebssystem) abhängig. Die NFSv3-Unterstützung bietet die folgenden Funktionen:

- Keine Designbeschränkungen für die Anzahl von Dateien oder Verzeichnissen
- Schreibgröße für Dateien von bis zu 16 TB.
- Möglichkeit zur Skalierung auf bis zu 8 Standorte mit einem einzigen globalen Namespace/Export
- Unterstützung für Kerberos- und AUTH_SYS-Authentifizierung

NFS-Dateiservices verarbeiten NFS-Anfragen von Clients; die Daten werden jedoch als Objekte in ECS gespeichert. Ein NFS-Datei-Handle wird einer Objekt-ID zugeordnet. Da die Datei im Wesentlichen einem Objekt zugeordnet ist, verfügt NFS über Funktionen wie den Objektdatendienst, einschließlich:

- Quotenmanagement auf Bucket-Level
- Verschlüsselung auf Objektlevel
- Write-Once-Read-Many (WORM) auf Bucket-Level
  - WORM wird während der Erstellung eines neuen Bucket mithilfe des Autocommit-Zeitraums implementiert.
  - WORM kann nur auf nicht-konforme Buckets angewendet werden.

3.3.4 Anschlüsse und Gateways

Mehrere Drittanbieter-Softwareprodukte haben die Möglichkeit, auf ECS Objektspeicher zuzugreifen. Unabhängige Softwareanbieter (ISVs) wie Panzura, Ctera und Syncplicity erstellen eine Schicht von Services, die Clientzugriff auf ECS-Objektspeicher über herkömmliche Protokolle wie SMB/CIFS, NFS und iSCSI bieten. Unternehmen können mit den folgenden Dell EMC Produkten auch auf Daten auf ECS-Speicher zugreifen oder sie hochladen:

- **Isilon CloudPools:** Policy-basiertes Tiering von Daten von Isilon auf ECS.
- **Data Domain Cloud Tier:** Automatisiertes, natives Tiering deduplizierter Daten von Data Domain zu ECS für die langfristige Aufbewahrung. Data Domain Cloud Tier bietet eine sichere und kostengünstige Lösung zur Verschlüsselung von Daten in der Cloud mit weniger Bedarf an Speicherplatz und Netzwerkbandbreite.
- **GeoDrive:** Stub-basierter ECS-Speicherservice für Microsoft® Windows®-Desktop-PCs und -Server.

3.4 Speicher-Engine

Im Zentrum von ECS steht die Speicher-Engine. Die Speicher-Engine-Schicht enthält die Hauptkomponenten, die für die Verarbeitung von Anfragen verantwortlich sind, sowie das Speichern, Abrufen, Schützen und Replizieren von Daten.

In diesem Abschnitt werden die Designprinzipien sowie die interne Darstellung und Handhabung von Daten beschrieben.

3.4.1 Speicherservices

Die Speicher-Engine von ECS umfasst die folgenden Services, wie in Abbildung 6 gezeigt.

Abbildung 6 Speicher-Engine-Services

Die Speicher-Engine-Services werden in einem Docker-Container gekapselt, der auf jedem ECS-Node ausgeführt wird, um einen verteilten und gemeinsam genutzten Service bereitzustellen.

3.4.2 Daten

Die primären Datentypen, die in ECS gespeichert sind, können wie folgt zusammengefasst werden:

- **Daten:** Gespeicherte Inhalte auf Anwendungs- oder Nutzerlevel, etwa ein Bild. „Daten“ wird synonym mit „Objekt“, „Datei“ oder „Inhalt“ verwendet. Anwendungen können für jedes Objekt eine unbegrenzte Menge an angepassten Metadaten speichern. Die Speicher-Engine schreibt Daten und zugehörige, von einer Anwendung bereitgestellte angepasste Metadaten gemeinsam in ein logisches Repository. Angepasste Metadaten sind eine robuste Funktion moderner Speichersysteme, die weitere Informationen oder Kategorisierungen der zu speichernden Daten liefern. Angepasste Metadaten werden als Schlüsselwertpaare formatiert und mit Schreibanfragen bereitgestellt.

- **Systemmetadaten:** Nutzerdaten- und systemressourcenbezogene Systeminformationen und -attribute. Systemmetadaten können wie folgt grob kategorisiert werden:
  - **Kennungen und Deskriptoren:** Eine Reihe von Attributen, die intern zur Identifizierung von Objekten und deren Versionen verwendet werden. Kennungen sind numerische IDs oder Hashwerte, die außerhalb des Softwarekontexts von ECS bedeutungslos sind. Deskriptoren definieren Informationen wie den Typ der Codierung.
  - **Verschlüsselungsschlüssel im verschlüsselten Format:** Datenverschlüsselungsschlüssel werden als Systemmetadaten betrachtet. Sie werden in verschlüsselter Form innerhalb der Tabellenstruktur des Kernverzeichnisses gespeichert.
  - **Interne Markierungen:** Eine Reihe von Indikatoren, die verwendet werden, um zu verfolgen, ob Updates oder Verschlüsselung im Bytebereich aktiviert sind, und um das Zwischenspeichern und Löschen zu koordinieren.
  - **Speicherortinformationen:** Attributsatz mit Index und Datenspeicherort, z. B. Byte-Offsets.
  - **Zeitstempel:** Attributsatz, der die Zeit nachverfolgt, z. B. beim Erstellen oder Aktualisieren von Objekten.
  - **Konfigurations-/Mandanteninformationen:** Namespace- und Objektzugriffssteuerung.

Daten- und Systemmetadaten werden in *Blöcken* auf ECS geschrieben. Ein ECS-Block ist ein logischer Container von 128 MB aus zusammenhängendem Speicher. Jeder Block kann Daten aus verschiedenen Objekten enthalten, wie unten in Abbildung 7 dargestellt. ECS verwendet die Indexierung, um alle Teile eines Objekts nachzuverfolgen, die möglicherweise über verschiedene Blöcke und Nodes verteilt sind.

Blöcke werden im Append-only-Muster geschrieben. Das Append-only-Verhalten bedeutet, dass die Anforderung einer Anwendung, ein vorhandenes Objekt zu ändern oder zu aktualisieren, die zuvor geschriebenen Daten in einem Block nicht ändert oder löscht, sondern dass die neuen Änderungen oder Updates in einen neuen Block geschrieben werden. Dadurch ist keine I/O-Sperre oder Cacheinvalidierung erforderlich. Das Append-only-Design vereinfacht außerdem die Datenversionserstellung. Alte Versionen der Daten bleiben in zuvor geschriebenen Blöcken gespeichert. Wenn die S3-Versionierung aktiviert ist und eine ältere Version der Daten erforderlich ist, kann sie mithilfe der S3 REST API abgerufen oder auf eine vorherige Version wiederhergestellt werden.

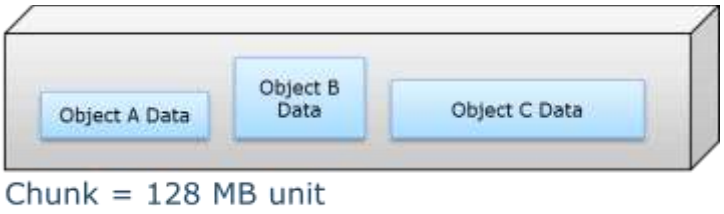

Abbildung 7 128-MB-Block, der Daten von 3 Objekten speichert

Im Abschnitt *Datenintegrität und Data Protection* unten wird erläutert, wie Daten auf Blocklevel geschützt werden.

3.4.3 Datenmanagement

ECS verwendet einen Satz logischer Tabellen, um Informationen zu den Objekten zu speichern. Schlüsselwertpaare werden schließlich auf der Festplatte in einer B+-Struktur gespeichert, damit Datenspeicherorte schnell indexiert werden können. Durch das Speichern des Schlüsselwertpaars in einer ausgeglichenen, durchsuchten Struktur wie einer B+-Struktur kann schnell auf den Speicherort der Daten und Metadaten zugegriffen werden. ECS nutzt eine protokollstrukturierte Zusammenführungsstruktur mit 2 Leveln, in der es 2 baumartigen Strukturen gibt: Eine kleinere Struktur befindet sich im Arbeitsspeicher (Arbeitsspeichertabelle) und die Haupt-B+-Struktur befindet sich auf der Festplatte. Die Suche nach Schlüsselwertpaaren erfolgt zunächst im Arbeitsspeicher und anschließend in der Haupt-B+-Struktur auf der Festplatte, falls erforderlich. Einträge in diesen logischen Tabellen werden zunächst in Journalprotokollen aufgezeichnet und diese Protokolle werden in 3-fach gespiegelten Blöcken auf Festplatten geschrieben. Die Journale werden verwendet, um Transaktionen nachzuverfolgen, die noch nicht in die B+-Struktur geschrieben wurden. Nach der Protokollierung jeder Transaktion in einem Journal wird die Tabelle im Arbeitsspeicher aktualisiert. Sobald die Arbeitsspeichertabelle voll ist oder ein bestimmter Zeitraum abgelaufen ist, wird sie mit Mergesort verarbeitet oder in der B+-Struktur auf der Festplatte abgelegt. Die Anzahl der vom System verwendeten Journalblöcke ist unbedeutend im Vergleich zu B+-Struktur-Blöcken. In Abbildung 8 wird dieser Prozess dargestellt.

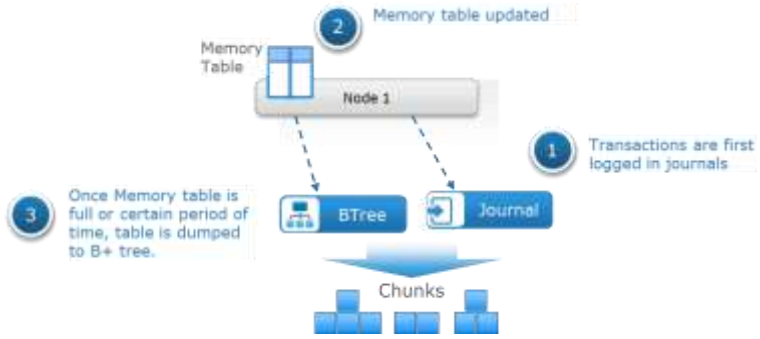

Abbildung 8 Arbeitsspeichertabelle in B+-Struktur abgelegt

Die in der Objekttabelle (OB) gespeicherten Informationen sind werden in Tabelle 2 unten gezeigt. Die OB-Tabelle enthält die Namen von Objekten und ihren Blockspeicherort mit einem bestimmten Offset und einer bestimmten Länge innerhalb dieses Blocks. In dieser Tabelle ist der Objektname der Schlüssel für den Index und der Wert ist der Speicherort des Blocks. Die Indexschicht in der Speicher-Engine ist für die Zuordnung von Objektnamen zu Blöcken verantwortlich.

Tabelle 2 Objekttabelleneinträge

Objektname	Blockspeicherort
ImgA	• C1:offset:length
FileB	• C2:offset:length • C3:offset:length

Die Blocktabelle (CT) zeichnet den Speicherort für jeden Block wie in Tabelle 3 beschrieben auf.

Tabelle 3 Blocktabelleneinträge

Block-ID	Position
C1	• Node1:Disk1:File1:Offset1:Length • Node2:Disk2:File1:Offset2:Length • Node3:Disk2:File6:Offset:Length

ECS wurde als verteiltes System entwickelt, sodass Speicher und Zugriff auf Daten über alle Nodes verteilt werden. Die für das Management von Objektdaten und Metadaten verwendeten Tabellen werden im Laufe der Zeit mit zunehmender Nutzung und Größe des Speichers immer größer. Die Tabellen werden in Partitionen unterteilt und verschiedenen Nodes zugewiesen, wobei jeder Node der Eigentümer der Partitionen ist, die er für jede der Tabellen hostet. Um den Speicherort eines Blocks zu erhalten, wird beispielsweise die Partitionsdatensatztabelle (PR) für den Eigentümer-Node abgefragt, der den Blockspeicherort kennt. Eine grundlegende PR-Tabelle wird in Tabelle 4 unten dargestellt.

Tabelle 4 Tabelleneinträge für Partitionsdatensätze

Partitions-ID	Owner
P1	Knoten 1
P2	Knoten 2
P3	Knoten 3

Wenn ein Node ausfällt, übernehmen andere Nodes die Eigentumsrechte an seinen Partitionen. Die Partitionen werden neu erstellt, indem der Stamm der B+-Struktur gelesen wird und die auf der Festplatte gespeicherten Journale wiedergegeben werden. Abbildung 9 zeigt das Failover der Partitionseigentumsrechte.

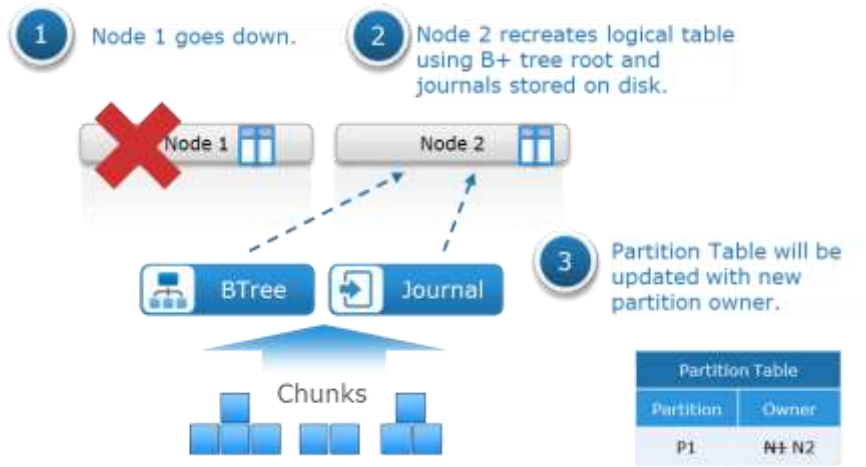

Abbildung 9 Failover der Partitionseigentumsrechte

3.4.4 Datenfluss

Speicherservices stehen auf jedem Node zur Verfügung. Daten werden durch über Laufwerke, Nodes und Racks verteilte EC-Segmente geschützt. ECS führt eine Prüfsummenfunktion aus und speichert das Ergebnis bei jedem Schreibvorgang. Wenn die ersten paar Byte der Daten komprimierbar sind, komprimiert ECS die Daten. Bei Lesevorgängen werden die Daten dekomprimiert und die gespeicherte Prüfsumme wird validiert. Dies ist ein Beispiel für den Datenfluss bei einem Schreibvorgang in 5 Schritten:

1. Der Client sendet Objekterstellungsanfrage an einen Node.
2. Der Node, der die Anfrage bearbeitet, schreibt die Daten des neuen Objekts in einen Repo-Block (kurz für Repository).
3. Beim erfolgreichen Schreiben auf die Festplatte findet eine PR-Transaktion für den Eintrag von Name und Blockspeicherort statt.
4. Der Partitionseigentümer zeichnet die Transaktion in Journalprotokollen auf.
5. Sobald die Transaktion in den Journalprotokollen aufgezeichnet wurde, wird eine Bestätigung an den Client gesendet.

Wie in Abbildung: 10 unten gezeigt, ist ein Beispiel für den Datenfluss für einen Lesevorgang für Festplattenlaufwerksarchitekturen wie Gen2 und EX300, EX500 und EX3000 dargestellt:

1. Eine Leseobjektanfrage wird vom Client an Node 1 gesendet.
2. Node 1 verwendet eine Hashfunktion mit dem Objektnamen, um zu bestimmen, welcher Node der Partitionseigentümer der logischen Tabelle ist, in der sich diese Objektinformationen befinden. In diesem Beispiel ist Node 2 Eigentümer und daher führt Node 2 eine Suche in den logischen Tabellen durch, um den Speicherort des Blocks zu ermitteln. In einigen Fällen kann die Suche auf 2 unterschiedlichen Nodes durchgeführt werden, z. B. wenn der Speicherort nicht in logischen Tabellen auf Node 2 zwischengespeichert wird.
3. Ausgehend vom vorherigen Schritt wird der Speicherort des Blocks für Node 1 bereitgestellt, der dann eine Byte-Offset-Leseanfrage für den Node ausgibt, auf dem sich die Daten befinden – in diesem Beispiel Node 3 –, und Daten an Node 1 sendet.
4. Node 1 sendet Daten an den anfordernden Client.

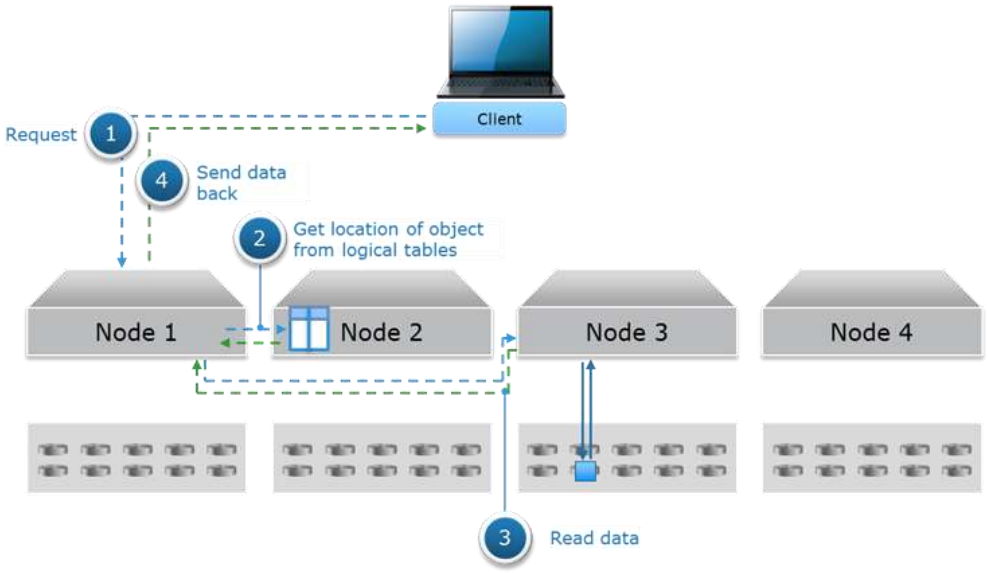

Abbildung: 10 Lesedatenfluss für Festplattenlaufwerksarchitektur

Wie in Abbildung: 11 unten gezeigt, ein Beispiel des Datenflusses für einen Lesevorgang für All-Flash-Architekturen wie EXF900:

1. Eine Leseobjektanfrage wird vom Client an Node 1 gesendet.
2. Node 1 verwendet eine Hashfunktion mit dem Objektnamen, um zu bestimmen, welcher Node der Partitionseigentümer der logischen Tabelle ist, in der sich diese Objektinformationen befinden. In diesem Beispiel ist Node 2 Eigentümer und daher führt Node 2 eine Suche in den logischen Tabellen durch, um den Speicherort des Blocks zu ermitteln. In einigen Fällen kann die Suche auf 2 unterschiedlichen Nodes durchgeführt werden, z. B. wenn der Speicherort nicht in logischen Tabellen auf Node 2 zwischengespeichert wird.
3. Im vorherigen Schritt wird der Speicherort des Blocks an Node 1 bereitgestellt, der dann die Daten direkt von Node 3 liest.
4. Node 1 sendet Daten an den anfordernden Client.

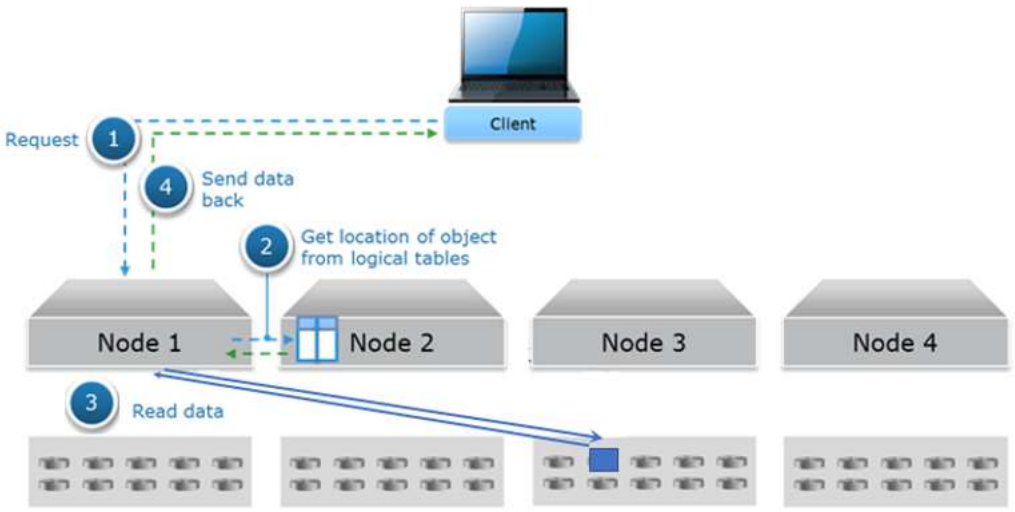

Abbildung: 11 Lesedatenfluss für All-Flash-Architektur

Hinweis: In der All-Flash-Architektur wie EXF900 kann jeder Node Daten direkt von einem anderen Node lesen, abgesehen von der Festplattenlaufwerksarchitektur, die jeder Node nur den Datenspeicher selbst lesen kann.

3.4.5 Schreiboptimierungen für Dateigröße

Für kleinere Schreibvorgänge im Speicher verwendet ECS eine *Box-Carting* genannte Methode, um die Auswirkungen auf die Performance zu minimieren. Beim Box-Carting werden mehrere kleinere Schreibvorgänge von 2 MB oder weniger im Arbeitsspeicher aggregiert und dann in einem einzigen Festplattenvorgang geschrieben. Box-Carting reduziert die Anzahl der Roundtrips auf die Festplatte, die für individuelle Schreibvorgänge erforderlich sind.

Bei Schreibvorgängen größerer Objekte können Nodes innerhalb von ECS Schreibanforderungen für dasselbe Objekt gleichzeitig verarbeiten und gleichzeitige Schreibvorgänge über mehrere Spindeln im ECS-Cluster hinweg nutzen. Daher kann ECS kleine und große Objekte effizient aufnehmen und speichern.

3.4.6 Speicherplatzrückgewinnung

Das Schreiben von Blöcken auf Append-only-Weise bedeutet, dass die Daten hinzugefügt oder aktualisiert werden, indem zunächst die ursprünglich geschriebenen Daten beibehalten werden und zum anderen neue Blocksegmente erstellt werden, die im Blockcontainer des ursprünglichen Objekts enthalten sein können, aber nicht müssen. Der Vorteil einer Append-only-Datenänderung ist ein Aktiv-Aktiv-Datenzugriffsmodell, das nicht durch Probleme mit Dateisperren herkömmlicher Dateisysteme behindert wird. Dies ist der Fall, wenn Objekte aktualisiert oder gelöscht werden und Daten in Blöcken nicht mehr referenziert oder benötigt werden. Die folgenden 2 Methoden zur automatischen Speicherbereinigung werden von ECS zur Rückgewinnung von Speicherplatz aus verworfenen vollständigen Blöcken oder Blöcken mit einer Mischung aus gelöschten und nicht gelöschten Objektfragmenten, die nicht mehr referenziert werden, verwendet:

- **Normale automatische Speicherbereinigung:** Wenn ein ganzer Block veraltet ist, wird Speicherplatz zurückgewonnen.
- **Partielle automatische Speicherbereinigung durch Zusammenführen:** Wenn ein Block zu 2/3 aus veralteten Objekten besteht, wird der Block zurückgewonnen, indem die gültigen Teile mit anderen teilweise gefüllten Blöcken zu einem neuen Block zusammengeführt werden, um Speicherplatz zurückzugewinnen.

Die automatische Speicherbereinigung wurde auch auf die ECS CAS-Datendienstzugriffs-API angewendet, um verwaiste BLOBs zu bereinigen. Verwaiste BLOBs, d. h. nicht referenzierte BLOBs, die in den in ECS gespeicherten CAS-Daten identifiziert wurden, werden für Speicherplatzrückgewinnung über normale Speicherbereinigungsmethoden ausgewählt.

3.4.7 SSD-Metadaten-Caching

ECS-Metadaten werden in B-Strukturen gespeichert. Jede B-Struktur kann Einträge im Arbeitsspeicher, journalbasierte Transaktionen und auf der Festplatte enthalten. Damit das System ein vollständiges Bild einer bestimmten B-Struktur erhalten kann, werden alle drei Speicherorte abgefragt, die häufig mehrere Suchen auf die Festplatte umfassen.

Um die Latenz für Metadatenabfragen zu minimieren, wurde ein optionaler SSD-basierter Cache-Mechanismus in ECS 3.5 implementiert. Der Cache enthält zuletzt aufgerufene B-Strukturseiten. Das bedeutet, dass Lesevorgänge auf den neuesten B-Strukturen immer auf den SSD-basierten Cache treffen und die Verwendung von rotierenden Festplatten vermieden wird.

Hier sind einige Highlights für die neue SSD-Metadaten-Caching-Funktion:

- Verbesserte systemweite Leselatenz und TPS (Transactions Per Second) für kleine Dateien
- Eine Flash-Festplatte mit 960 GB pro Node
- Neue Nodes aus der Fertigung enthalten die SSD-Festplatte als Option
- Vorhandene Vor-Ort-Nodes – Gen3 und Gen2 – können über Upgrade-Kits und Selfservice-Installation aktualisiert werden
- SSD-Festplatten können hinzugefügt werden, während ECS online ist
- Verbesserung für Analyse-Workloads mit kleinen Dateien, die schnelle Lesevorgänge großer Datenvolumen erfordern
- Alle Nodes in einem VDC müssen über SSDs verfügen, um diese Funktion zu aktivieren

Die ECS-Fabric erkennt, wenn ein SSD-Kit installiert wurde. Dadurch wird das System automatisch initialisiert und mit der Verwendung des neuen Laufwerks begonnen. Abbildung: 12 zeigt an, dass der SSD-Cache aktiviert ist.

Abbildung: 12 SSD-Cache aktiviert

SSD-Metadaten-Caching verbessert die Auflistung kleiner Lesevorgänge und Buckets. Wie wir in unserem Labor getestet haben, verbessert sich die Auflistungsleistung mit 10-MB-Objekten um 50 %. Die Leseleistung verbessert sich um 35 % bei 10-KB-Objekten und um 70 % bei 100-KB-Objekten.

3.4.8 Cloud DVR

ECS unterstützt die DVR-Funktion (Digital Video Recording), die zur Erfüllung einer gesetzlichen Copyright-Anforderung für Kabel- und Satellitenunternehmen dient. Die Anforderung besteht darin, dass für jede Aufzeichnungseinheit, die einem Objekt in ECS zugeordnet ist, eine festgelegte Anzahl von Kopien erstellt werden muss. Die festgelegte Anzahl von Kopien wird als „Fanout“ bezeichnet. Die vorab festgelegte Anzahl von Kopien (Fan-out) ist nicht wirklich eine Anforderung für Redundanz oder Performancesteigerung, sondern eher eine gesetzliche Urheberrechtsanforderung für Kabel- und Satellitenunternehmen. ECS unterstützt:

- Erstellen einer Fan-out-Anzahl von Kopien von Objekten, die in ECS erstellt wurden
- Lesevorgang einer bestimmten Kopie zulassen
- Löschen einer bestimmten Kopie zulassen
- Löschen aller Kopien zulassen
- Kopie einer bestimmten Kopie zulassen
- Auflisten von Kopien zulassen
- Bucket-Auflistung von Fan-out-Objekten zulassen

Die Cloud-DVR-Funktion kann über die Servicekonsole aktiviert werden. Zum ersten Mal müssen Sie die Cloud-DVR-Funktion über die Servicekonsole aktivieren. Nach der Aktivierung von Cloud DVR ist Cloud DVR standardmäßig für alle neuen Nodes aktiviert.

Führen Sie den folgenden Befehl in der Servicekonsole aus, um die Cloud-DVR-Funktion zu aktivieren:

```
service-console run Enable_CloudDVR
```

Die Cloud-DVR-Funktion unterstützt APIs. Weitere Informationen finden Sie im *ECS-Datenzugriffshandbuch.*

3.5 Fabric

Die Fabric-Schicht bietet Clustering-, Systemintegritäts-, Softwaremanagement-, Konfigurationsmanagement- und Upgradefunktionen sowie Warnmeldungen. Sie ist für das stetige Ausführen von Services und das Managen von Ressourcen wie Festplatten, Containern und Netzwerk verantwortlich. Sie verfolgt Änderungen an der Umgebung wie etwa Fehlererkennung nach, reagiert auf diese und gibt Warnmeldungen im Zusammenhang mit der Systemintegrität aus. Die Fabric-Schicht umfasst die folgenden Komponenten:

- **Node Agent:** Managt Hostressourcen (Festplatten, Netzwerk, Container usw.) und Systemprozesse.
- **Lifecyle Manager:** Anwendungs-Lebenszyklusmanagement, das das Starten von Services, Recovery, Benachrichtigungen und Fehlererkennung umfasst.
- **Persistence Manager:** Koordiniert und synchronisiert die verteilte Umgebung in ECS.
- **Registrierung:** Docker-Image-Speicher für ECS-Software.
- **Ereignisbibliothek:** Enthält den Satz von Ereignissen, die auf dem System auftreten.
- **Hardware Manager:** Liefert Status- und Ereignisinformationen sowie die Bereitstellung der Hardwareschicht für Services höherer Level. Diese Services wurden zur Unterstützung handelsüblicher Hardware integriert.

3.5.1 Node Agent

Der Node Agent ist ein ressourcenschonender, in Java geschriebener Agent, der nativ auf allen ECS-Nodes ausgeführt wird. Zu den Hauptaufgaben gehören das Management und die Steuerung von Hostressourcen (Docker-Container, Festplatten, Firewall, Netzwerk) und das Monitoring von Systemprozessen. Zu den Beispielen für das Management gehören das Formatieren und Mounten von Festplatten, das Öffnen erforderlicher Ports, das Sicherstellen, dass alle Prozesse ausgeführt werden, und das Ermitteln von öffentlichen und privaten Netzwerkschnittstellen. Er verfügt über einen Ereignisstream, der geordnete Ereignisse für einen Lifecycle Manager bereitstellt, um Ereignisse anzuzeigen, die im System auftreten. Eine Fabric CLI ist hilfreich, um Probleme zu diagnostizieren und den allgemeinen Systemstatus zu überprüfen.

3.5.2 Lifecycle-Management

Der Lifecycle Manager wird auf einer Untergruppe von 3 oder 5 Nodes ausgeführt und managt den Lebenszyklus von Anwendungen, die auf Nodes ausgeführt werden. Jeder Lifecycle Manager ist für das Nachverfolgen mehrerer Nodes verantwortlich. Das Hauptziel ist das Management des gesamten Lebenszyklus der ECS-Anwendung vom Start bis zur Bereitstellung, einschließlich Fehlererkennung, Recovery, Benachrichtigung und Migration. Er überprüft die Streams vom Node Agent und regt den Agent zur Handhabung der Situation an. Wenn ein Node ausgefallen ist, reagiert er auf Ausfälle oder Unregelmäßigkeiten im Status des Node, indem er das System in einem als fehlerfrei bekannten Zustand wiederherstellt. Wenn eine Lifecycle Manager-Instanz ausgefallen ist, übernimmt eine andere ihre Funktion.

3.5.3 Registrierung

Die Registrierung enthält die ECS-Docker-Images, die während der Installation, beim Upgrade und beim Node-Austausch verwendet werden. Ein Docker-Container namens *fabric-registry* wird auf einem Node innerhalb des ECS-Racks ausgeführt und enthält das Repository mit ECS-Docker-Images sowie die benötigten Informationen für Installationen und Upgrades. Obwohl die Registrierung nur auf jeweils einem Node gleichzeitig verfügbar ist, werden alle Docker Images lokal auf jedem Node zwischengespeichert, sodass alle der Registrierung zur Verfügung stehen.

3.5.4 Ereignisbibliothek

Die Ereignisbibliothek wird in der Fabric-Schicht verwendet, um die Lebenszyklus- und Node-Agent-Ereignisstreams bereitzustellen. Vom System erzeugte Ereignisse werden in gemeinsam genutztem Speicher und auf der Festplatte gespeichert, um Verlaufsinformationen zum Status und zur Integrität des ECS-Systems zu liefern. Diese geordneten Ereignisstreams können verwendet werden, um das System in einem bestimmten Zustand wiederherzustellen, indem Sie die gespeicherten geordneten Ereignisse abspielen. Einige Beispiele für Ereignisse sind Node-Ereignisse wie gestartet, beendet oder heruntergestuft.

3.5.5 Hardware Manager

Der Hardware Manager ist in den Fabric Agent integriert, um Hardware nach Branchenstandard zu unterstützen. Seine Hauptaufgabe ist es, hardwarespezifische Status- und Ereignisinformationen sowie die Bereitstellung der Hardwareschicht für Services höherer Level zu liefern.

3.6 Infrastruktur

Auf ECS-Appliance-Nodes wird derzeit SUSE Linux Enterprise Server 12 für die Infrastruktur ausgeführt. Für ECS-Software, die auf angepasster Hardware nach Branchenstandard bereitgestellt wird, kann das Betriebssystem auch RedHat Enterprise Linux oder CoreOS sein. Angepasste Bereitstellungen werden über einen formellen Anforderungs- und Validierungsprozess durchgeführt. Docker ist in der Infrastruktur installiert, um die gekapselten ECS-Schichten bereitzustellen. Die ECS-Software ist in Java geschrieben, sodass die Java Virtual Machine als Teil der Infrastruktur installiert wird.

3.6.1 Docker

ECS wird als Java-Anwendung auf dem Betriebssystem ausgeführt und in mehreren Docker-Containern gekapselt. Die Container sind isoliert, teilen sich aber die zugrunde liegende(n) Betriebssystemressourcen und Hardware. Einige Teile der ECS-Software werden auf allen Nodes ausgeführt und einige werden auf einem oder mehreren Nodes ausgeführt. Die folgenden Komponenten werden u. a. in einem Docker-Container ausgeführt:

- **object-main:** Enthält die Ressourcen und Prozesse im Zusammenhang mit den Datendiensten, der Speicher-Engine sowie den Portal- und Bereitstellungsservices. Wird auf jedem Node in ECS ausgeführt.
- **fabric-lifecycle:** Enthält die Prozesse, Informationen und Ressourcen, die für das Monitoring auf Systemlevel, das Konfigurationsmanagement und das Integritätsmanagement erforderlich sind. Es wird immer eine ungerade Anzahl von fabric-lifecycle-Instanzen ausgeführt. So werden beispielsweise 3 Instanzen auf einem 4-Node-System und 5 Instanzen auf einem 8-Node-System ausgeführt.

- **fabric-zookeeper:** Zentraler Service zur Koordination und Synchronisation von verteilten Prozessen, Konfigurationsinformationen, Gruppen und Namensservices. Er wird als Persistence Manager bezeichnet und auf einer ungeraden Anzahl von Nodes ausgeführt, z. B. 5 in einem System mit 8 Nodes.
- **fabric-registry:** Registrierung der ECS-Docker-Images. Pro ECS-Rack wird nur eine Instanz ausgeführt.

Es gibt andere Prozesse und Tools, die außerhalb eines Docker-Containers ausgeführt werden, nämlich Tools für den Fabric Node Agent und die Hardwareabstraktionsschicht. In Abbildung 13 unten wird ein Beispiel dafür gezeigt, wie ECS-Container in einer Bereitstellung mit 8 Nodes ausgeführt werden können.

Abbildung 13 Beispiel für Docker-Container und -Agents in einer Bereitstellung mit 8 Nodes

Abbildung 14 zeigt die Befehlszeilenausgabe des Befehls *docker ps* auf einem Node, auf dem die 4 von ECS in Docker verwendeten Container angezeigt werden. Es wird eine Liste mit allen objektbezogenen Services angezeigt, die im System verfügbar sind.

```
admin@hop-u300-11-pub-01:~> sudo docker ps
CONTAINER ID        IMAGE                                                  COMMAND                  CREATED             STATUS
      PORTS               NAMES
7ba30ce42be2        ecs-monitoring/telegraf:3.5.0.0-825.b6b07cf9           "/entrypoint.sh "        5 weeks ago         Up 5 weeks
                          object-telegraf
e22513635eab        ecs-monitoring/grafana:3.5.0.0-825.b6b07cf9            "/entrypoint.sh "        5 weeks ago         Up 5 weeks
                          object-grafana
ee9db1ea40bc        emcvipr/object:3.5.0.0-120417.6a158e139f1              "/opt/vipr/boot/boot."   5 weeks ago         Up 5 weeks
                          object-main
d11a7acd55e5        ecs-monitoring/throttler:3.5.0.0-825.b6b07cf9          "/entrypoint.sh "        5 weeks ago         Up 5 weeks
                          object-throttler
f9402679f7bb3       ecs-monitoring/fluxd:3.5.0.0-825.b6b07cf9              "/entrypoint.sh "        5 weeks ago         Up 5 weeks
                          object-fluxd
c7b8530a8bb9        caspian/fabric:3.5.0.0-4076.7d40a97                    "./boot.sh lifecycle"    5 weeks ago         Up 5 weeks
                          fabric-lifecycle
bffdf8836853        caspian/fabric-zookeeper:3.5.6.0-99.0354df7            "./boot.sh 1 1*169.2."   5 weeks ago         Up 5 weeks
                          fabric-zookeeper
f4420f7f7d51        caspian/fabric-registry:2.3.1.0-68.10d1aca             "/opt/docker-registr."   5 weeks ago         Up 5 weeks
                          fabric-registry
admin@hop-u300-11-pub-01:~> sudo docker
hop-u300-11-pub-01:/ # cd /opt/storageos/
hop-u300-11-pub-01:/opt/storageos # ls bin/*svc
bin/blobsvc     bin/coordinatorsvc  bin/eventsvc  bin/objcontrolsvc  bin/storagemanagementsvc
bin/casavc      bin/dataheadsvc     bin/filesvc   bin/objheadsvc     bin/sysavc
bin/controlsvc  bin/ecsportalsvc    bin/hdfssvc   bin/resourcesvc    bin/transformsvc
```

Abbildung 14 Prozesse, Ressourcen, Tools und Binärdateien im object-main-Container

4 Appliance-Hardwaremodelle

Flexible Einstiegspunkte ermöglichen ECS eine schnelle Skalierung auf Petabyte und Exabyte an Daten. Eine ECS-Lösung kann bei minimaler Beeinträchtigung des Geschäfts durch Hinzufügen von Nodes und Festplatten sowohl in Kapazität als auch in Performance linear skaliert werden.

ECS-Appliance-Hardwaremodelle sind durch Hardwaregenerationen charakterisiert. Die Appliance-Serie der 3. Generation, die als Gen3 oder EX Serie bezeichnet wird, umfasst 3 Hardwaremodelle. Dieser Abschnitt bietet eine allgemeine Übersicht über die EX Serie. Vollständige Details finden Sie im *Hardwareleitfaden zur ECS EX Serie*.

Informationen zur ECS-Appliance-Hardware der 1. und 2. Generation finden Sie im *Hardwareleitfaden zu den Dell EMC ECS D und U Serien*.

4.1 EX Serie

Appliance-Modelle der EX Serie basieren auf standardmäßigen Dell Servern und Switchen. Die Serie umfasst die folgenden Produkte:

- **EX300:** Das EX300-System hat eine Einstiegsrohkapazität von 60 TB. Dies ist die ideale Speicherplattform für native Cloud-Anwendungen und digitale Transformationsinitiativen von Kunden. Das EX300-System ist ideal für die Modernisierung von Centera-Bereitstellungen geeignet. Vor allem aber kann das EX300-System kostengünstig auf größere Kapazitäten skaliert werden. Es bietet 12 Laufwerke pro Node und 1 TB, 2 TB, 4 TB, 8 TB, 16 TB Festplattenoptionen (alle gleich im Node)
- **EX500:** Das EX500-System ist die neueste Edition der Appliance und ist dafür konzipiert, Wirtschaftlichkeit bei hoher Dichte zu liefern. Mit Optionen für 12 oder 24 Laufwerke, 8 TB, 12 TB und 16 TB Festplattenoptionen (alle gleich im Node). Das Cluster reicht von 480 TB bis 6,1 PB pro Rack. Diese Serie ist eine vielseitige Option für mittelständische Unternehmen, die moderne Anwendungen und/oder umfassende Archivanwendungen unterstützen möchten.
- **EX3000:** Die EX3000 verfügt über eine maximale Kapazität von 11,5 PB Roh-Storage pro Rack, 30 bis 90 Laufwerke pro Node, 12-TB- oder 16-TB-Festplatten und kann über mehrere Standorte hinweg bis in den Exabyte-Bereich wachsen, was eine tiefgreifende und skalierbare Rechenzentrumslösung darstellt, die sich ideal für Workloads mit größeren Datenmengen eignet. Diese Nodes sind in 2 unterschiedlichen Konfigurationen verfügbar, die als EX3000S und EX3000D bezeichnet werden. Das EX3000S-System ist ein Single-Node- und das EX3000D-System ein Dual-Node-Gehäuse. Diese hochdichten Nodes sind Hot-Swap-fähig. Sie beginnen bei mindestens 30 Festplatten pro Node. 30 Laufwerke pro ECS-Node sind der Punkt, an dem der Performancegewinn durch das Hinzufügen von mehr Laufwerken abnimmt. Mit mindestens 30 oder mehr Laufwerken in jedem Node sind die Performanceerwartungen bei jedem EX3000-Node unabhängig von der Anzahl der Laufwerke ähnlich.
- **EXF900:** EXF900 ist eine All-Flash-Objektspeicherlösung mit hyperkonvergenten Nodes für ECS-Bereitstellungen mit niedriger Latenz und hohen IOPS. Mit Optionen für 12 oder 24 Laufwerke werden 3,84-TB-NVMe-SSD-Festplattenoptionen (7,68-TB-NVMe-SSD-Treiber werden unterstützt, wenn Hardware verfügbar ist). Diese Plattform beginnt mit einer Roh-Storage-Minimalkonfiguration von 230 TB und kann auf 1,4 PB Rohkapazität pro Rack skaliert werden. Abbildung: 15 zeigt einen Node von EXF900 an.

Abbildung: 15 EXF900-Node

Hinweis: Die SSD-Lesecachefunktion gilt nicht für EXF900; Cloud DVR wird auf EXF900 nicht unterstützt; Tech Refresh wird mit EXF900 nicht unterstützt; EXF900 kann nicht mit anderen Nicht-EXF900-Hardware in einem VDC koexistieren; EXF900 kann nicht mit anderen Nicht-EXF900-Hardwarekomponenten in GEO vorhanden sein (alle Standorte müssen EXF900 sein).

Die Optionen für die Einstiegskapazität der EX Serie ermöglichen es Kunden, eine ECS-Bereitstellung mit nur der benötigten Kapazität zu beginnen und das System bei sich ändernden Anforderungen in der Zukunft problemlos zu erweitern. Weitere Informationen zu den Appliances der EX Serie finden Sie im *Technischen Datenblatt für ECS-Appliances*, das auch die bisherigen Appliances der Gen2 U und D Serien beschreibt.

Updates nach der Bereitstellung werden für Nodes der EX Serie nicht unterstützt. Dazu gehören:

- Austauschen der CPU
- Ändern der Arbeitsspeicherkapazität
- Erweitern der Festplattengröße

4.2 Appliance-Networking

Beginnend mit der Veröffentlichung der Appliances der EX Serie wird ein redundantes Paar dedizierter Back-end-Managementswitche verwendet. Durch die Umstellung auf neue Appliance-Switche kann ECS jetzt einen Front- und einen Back-end-Switching-Modus für die Konfiguration einführen.

EX300, EX500 und EX3000 Appliances verwenden alle jeweils den Dell EMC S5148F für das Front-end-Switchpaar und das Back-end-Switchpaar. Die EXF900 Appliance verwendet den Dell EMC S5248F für das Front-end-Switchpaar und das Back-end-Switchpaar und S5232F für den Aggregations-Backend-Switch. Beachten Sie, dass Kunden die Option haben, anstelle der Dell EMC-Switches eigene Front-end-Switche zu verwenden.

4.2.1 S5148F – öffentliche Front-end-Switches

2 optionale Dell EMC S5148F-25-GbE-1-HE-Ethernetswitche können für die Netzwerkverbindung erworben werden oder der Kunde kann sein eigenes 10-GbE- oder 25-GbE-HA-Paar für die Front-end-Verbindung bereitstellen. Die öffentlichen Switche werden oft als *Hare* und *Rabbit* oder einfach als Front-end bezeichnet.

Achtung: Es müssen Verbindungen zwischen dem Netzwerk des Kunden und den Front-end-Switchen (Rabbit und Hare) bestehen, damit die Architektur mit hoher Verfügbarkeit der ECS Appliance aufrechterhalten wird. Wenn der Kunde entscheidet, sein Netzwerk nicht auf die erforderliche HA-Weise zu verbinden, gibt es keine Gewährleistung für eine hohe Datenverfügbarkeit bei Verwendung dieses Produkts.

Diese Switche bieten 48 25-GbE-SFP28-Ports und 6 100-GbE-QSFP28-Ports. Weitere Informationen zu diesen beiden Porttypen:

- SFP28 ist eine erweiterte Version von SFP+
  - SFP+ unterstützt bis zu 16 Gbit/s, SFP28 unterstützt bis zu 28 Gbit/s
  - Selber Formfaktor
  - Abwärtskompatibel mit SFP+-Modulen
- QSFP28 ist eine erweiterte Version von QSFP+
  - QSFP+ unterstützt bis zu 4 Lanes mit 16 Gbit/s, QSFP28 unterstützt bis zu 4 Lanes mit 28 Gbit/s
    - > Aggregierte QSFP+-Lanes können bis zu 40-Gbit/s-Ethernet erreichen
    - > Aggregierte QSFP28+-Lanes können bis zu 100-Gbit/s-Ethernet erreichen
  - Selber Formfaktor
  - Abwärtskompatibel zu QSFP+-Modulen
  - Kann in 4 einzelne Lanes mit SFP28 aufgeteilt werden

Hinweis: 2 100-GbE-LAG-Kabel werden mit öffentlichen Dell EMC S5148F-25-GbE-Switchen geliefert. Unternehmen, die ihre eigenen öffentlichen Switche bereitstellen, müssen die erforderlichen LAG-Kabel, SFPs oder externen Verbindungskabel bereitstellen.

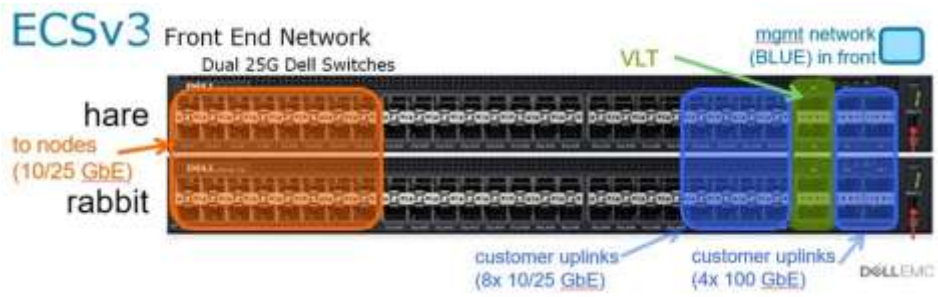

Abbildung 16 Portzuweisung und -nutzung des Front-end-Netzwerkswitchs

Abbildung 16 oben bietet eine visuelle Darstellung der Art und Weise, wie Ports verwendet werden, um ECS-Node-Datenverkehr sowie Kundenuplinkports zu ermöglichen. Dies ist der Standard für alle Implementierungen.

4.2.2 S5148F – private Back-end-Switches

Die beiden benötigten Dell EMC S5148F-25-GbE-1-HE-Ethernetswitche mit 48 25-GbE-SFP-Ports und 6 100-GbE-Uplinkports sind in jedem ECS-Rack enthalten. Diese werden häufig als *Fox*- und *Hound*- oder Back-end-Switche bezeichnet und sind für das Managementnetzwerk verantwortlich. In zukünftigen ECS-Versionen ermöglichen die Back-end-Switche auch eine Netzwerktrennung für den Replikationsdatenverkehr. Der Hauptzweck des privaten Netzwerks liegt in Remotemanagement und -konsole, PXE-Start für den Installationsmanager und Funktionen für rack- und clusterübergreifende Verwaltung und Bereitstellung. Abbildung 17 zeigt eine Vorderansicht von 2 Dell 25-GbE-Switchen.

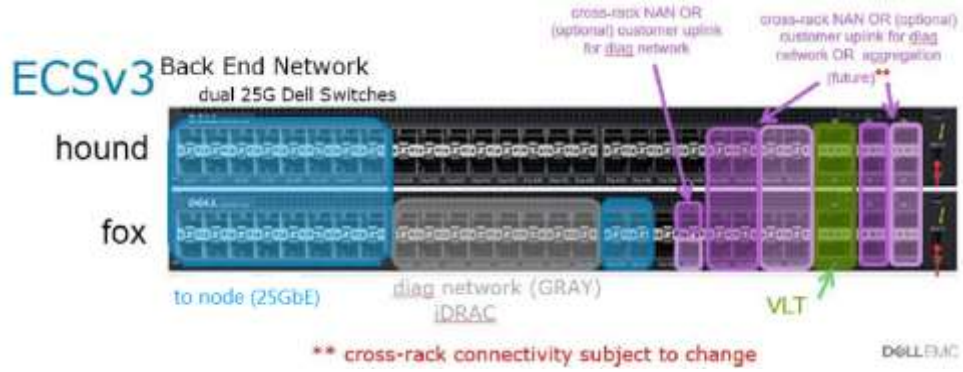

Abbildung 17 Portzuweisung und -nutzung des Back-end-Netzwerkswitchs

Das Diagramm oben bietet eine visuelle Darstellung, wie Ports verwendet werden sollen, um Ports für ECS-Managementverkehr und -Diagnose zu ermöglichen. Diese Portzuweisungen sind der Standard für alle Implementierungen. Mögliche Ports für zukünftige Nutzung sind in violett hervorgehoben, diese Nutzung kann sich jedoch in Zukunft ändern.

4.2.3 S5248F – öffentliche Front-end-Switches

Dell EMC bietet ein optionales HA-Paar an S5248F-Front-end-25-GbE-Switchen für die Kundennetzwerkverbindung an das Rack an. Es verfügt über zwei 200-GbE-VTL-Kabel (QSFP28-DD) pro HA-Paar. Diese Switche werden als Hare- und Rabbit-Switche bezeichnet. Abbildung 18 zeigt eine visuelle Darstellung der Art und Weise, wie Ports verwendet werden, um ECS-Node-Datenverkehr sowie Kundenuplinkports zu ermöglichen.

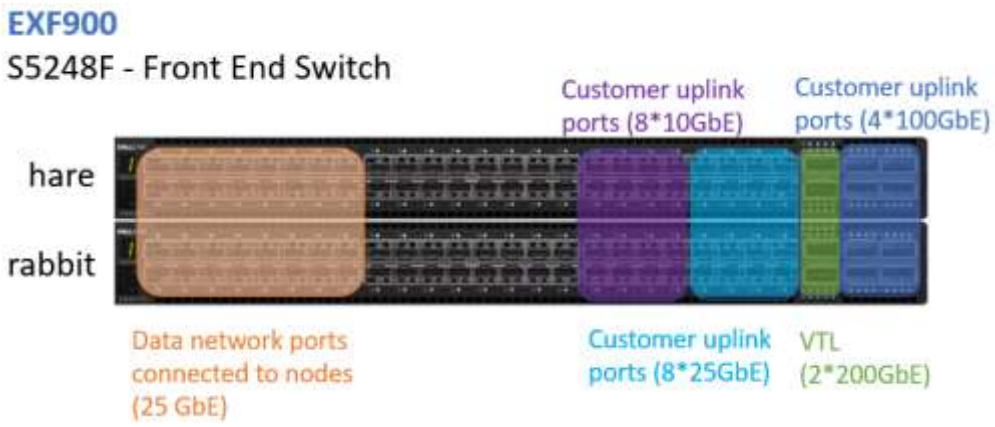

Abbildung 18 Portzuweisung und -nutzung des Front-end-Netzwerkswitchs

4.2.4 S5248F – private Back-end-Switches

Dell EMC bietet 2 S5248F-25-GbE-Back-end-Switche mit 2 200-GbE-VLT-Kabeln (QSFP28-DD). Diese Switche werden als Hound- und Fox-Switche bezeichnet. Alle iDRAC-Kabel von Nodes und alle Kabelverbindungen für das Front-end-Switchmanagement gehen zum Fox-Switch. Abbildung 19 bietet eine visuelle Darstellung, wie Ports verwendet werden sollen, um Ports für ECS-Managementverkehr und -Diagnose zu ermöglichen. Diese Portzuweisungen sind der Standard für alle Implementierungen.

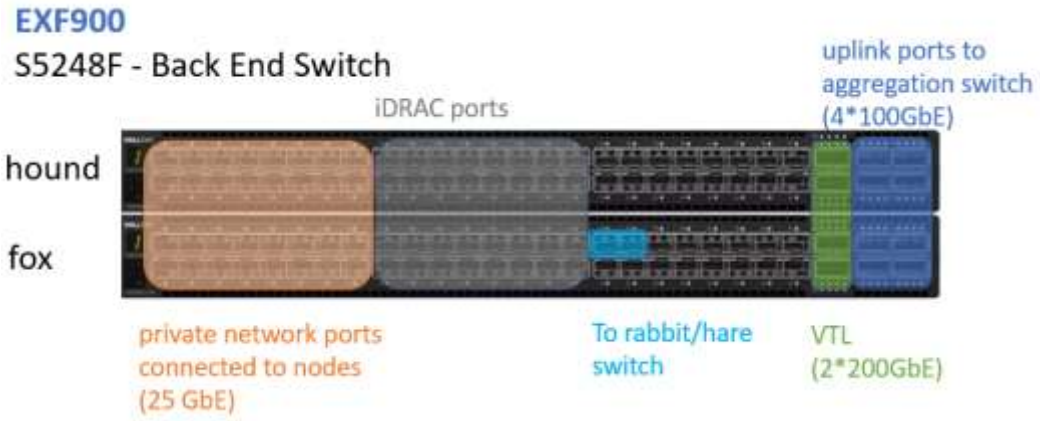

Abbildung 19 Portzuweisung und -nutzung des Back-end-Netzwerkswitchs

4.2.5 S5232 – Aggregationsswitch

Dell EMC bietet 2 S5232F-100-GbE-Back-end-Aggregationsswitche (AGG1 und AGG2) mit 4 100-GbE-VLT-Kabeln. Diese Switche werden als Falcon- und Eagle-Switche bezeichnet. Im folgenden Abbildung: 20 sind alle beschrifteten Ports mit Portzuweisungen angegeben. Diese Konfiguration ermöglicht das Anschließen von bis zu 7 Racks von EXF900-Nodes.

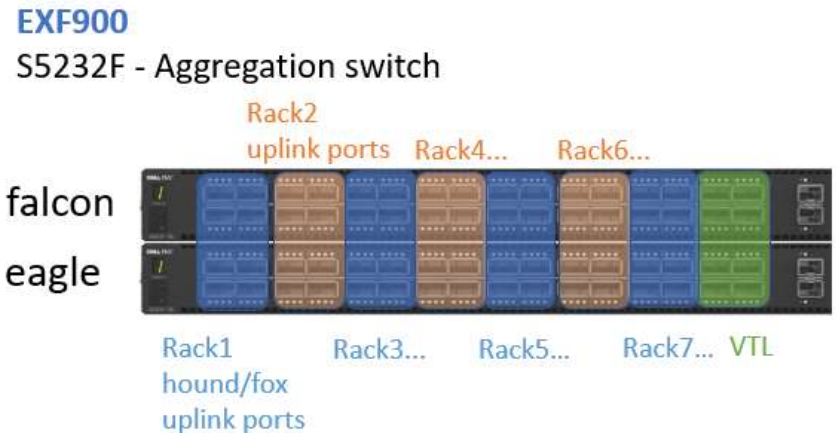

Abbildung: 20 Zuweisung und Nutzung des Aggregationsswitch-Ports

Weitere Informationen zum Netzwerk und zur Verkabelung finden Sie im Hardwarehandbuch für die *ECS EX Serie.*

5 Separate Netzwerke

ECS unterstützt die Trennung verschiedener Arten von Netzwerkverkehr zur Isolierung zu Sicherheits- und Performancezwecken. Folgende Arten von Datenverkehr können getrennt werden:

- Management
- Replikation
- Daten

Es gibt einen Betriebsmodus, der als *Netzwerktrennungsmodus* bezeichnet wird. In diesem Modus kann jeder Node auf Betriebssystemlevel mit bis zu 3 IP-Adressen oder logischen Netzwerken für jede der verschiedenen Arten von Datenverkehr konfiguriert werden. Diese Funktion wurde entwickelt, um die Flexibilität zu schaffen, entweder 3 separate logische Netzwerke für Management, Replikation und Daten zu erstellen oder sie zur Erstellung von 2 logischen Netzwerke zu kombinieren, sodass beispielsweise Management- und Replikationsdatenverkehr über ein logisches Netzwerk und Datenverkehr über ein anderes logisches Netzwerk läuft. Ein 2. logisches Datennetzwerk für reinen CAS-Datenverkehr kann konfiguriert werden, das die Trennung des CAS-Datenverkehrs von anderen Arten des Datenverkehrs wie S3 ermöglicht.

Die ECS-Implementierung separater Netzwerke erfordert, dass jeder logische Netzwerkverkehr mit Services und Ports verknüpft wird. Beispielsweise kommunizieren die ECS-Portalservices über die Ports 80 oder 443, sodass diese Ports und Services mit dem logischen Managementnetzwerk verbunden werden. Ein 2. Datennetzwerk kann konfiguriert werden, es dient jedoch nur für CAS-Datenverkehr. In Tabelle 5 unten sind die Services hervorgehoben, die auf einen logischen Netzwerktyp festgelegt sind. Eine vollständige Liste der den Ports zugeordneten Services finden Sie im aktuellen *ECS-Sicherheitskonfigurationsleitfaden*.

Tabelle 5 Zuordnung von Services zu logischen Netzwerken

Services	Logisches Netzwerk	Kennung
Web-UI und API, SSH, DNS, NTP, AD, SMTP	Management	public.mgmt
Clientdaten	Daten	public.data
	Nur CAS-Daten	public.data2
Replikationsdaten	Replikation	public.repl
SRS (Dell EMC Secure Remote Services)	Basierend auf dem Netzwerk ist SRS-Gateway verbunden	public.data oder public.mgmt

Hinweis: ECS 3.6 ermöglicht S3-Datenzugriff auf Daten (Standard) und Data2-Netzwerk (obwohl S3 auf Data2 nicht standardmäßig aktiviert ist). Um den S3-Datenzugriff im Data2-Netzwerk zu aktivieren, sind public.data erforderlich und wenden Sie sich an den ECS-Remotesupport.

Die Netzwerktrennung kann logisch mit unterschiedlichen IP-Adressen erreicht werden, wobei unterschiedliche VLANs virtuell oder unterschiedliche Kabel physisch verwendet werden. Der Befehl *setrackinfo* wird verwendet, um IP-Adressen und VLANs zu konfigurieren. Die VLAN-Konfiguration auf Switchlevel oder Clientseite liegt in der Verantwortung des Kunden. Für die physische Netzwerktrennung müssen Kunden eine Anfrage zur Produktqualifizierung (Request for Product Qualification, RPQ) einreichen, indem sie sich an Dell EMC Global Business Service wenden. Weitere Informationen zur Netzwerktrennung finden Sie im Whitepaper *ECS-Netzwerke und Best Practices*, das eine allgemeine Übersicht über die Netzwerktrennung bietet.

6 Sicherheit

ECS-Sicherheit wird ist Administrations-, Transport- und Datenlevel implementiert. Die Authentifizierung von Nutzern und Administratoren erfolgt über Active Directory, LDAP-Methoden, Keystone oder direkt im ECS-Portal. Sicherheit auf Datenlevel erfolgt über HTTPS für Daten in Bewegung und/oder serverseitige Verschlüsselung für gespeicherte Daten.

6.1 Authentifizierung

ECS unterstützt Active Directory-, LDAP- und Keystone- und IAM-Authentifizierungsmethoden, um Zugriff für das Management und die Konfiguration von ECS bereitzustellen. Es gibt jedoch Einschränkungen, wie in Tabelle 6 gezeigt. Weitere Informationen finden Sie im aktuellen *ECS-Sicherheitskonfigurationsleitfaden*.

Tabelle 6 Unterstützte Authentifizierungsmethoden

Authentifizierungsmethode	Unterstützt
Active Directory	• AD-Gruppenunterstützung für Managementnutzer • AD-Gruppenunterstützung für Objektnutzermethoden zur Selbstbereitstellung mithilfe von Selfservice-Schlüsseln über API • Unterstützung für mehrere Domains
LDAP	• Managementnutzer können sich über LDAP einzeln authentifizieren • LDAP-Gruppen werden für Managementnutzer nicht unterstützt • LDAP wird für Objektnutzer unterstützt (Selfservice-Schlüssel über API) • Unterstützung für mehrere Domains
Keystone	• RBAC-Policies werden noch nicht unterstützt • Keine Unterstützung für Token ohne Gültigkeitsbereich • Keine Unterstützung für mehrere Keystone-Server pro ECS-System
IAM	• Bereitstellung von Identitätsverbund und Single Sign-On (SSO) über SAML 2.0-Standards • Nur über das S3-Protokoll verfügbar

6.2 Datendienstauthentifizierung

Der Objektzugriff über RESTful APIs wird über HTTPS (TLS v1.2) gesichert. Eingehende Anforderungen werden mithilfe definierter Methoden wie Hash-based Message Authentication Code (HBAC), Kerberos- oder Tokenauthentifizierung authentifiziert. Tabelle 7 unten zeigt die verschiedenen Methoden, die für jedes Protokoll verwendet werden.

Tabelle 7 Datendienstauthentifizierung

Protokolle		**Authentifizierungsmethoden**
Objekt	S3	V2 (HMAC-SHA1), V4 (HMAC-SHA256)
	Swift	Token – Keystone v2 und v3 (Bereichsbeschränkung, UUID, PKI-Token), SWAuth v1
	Atmos	HMAC-SHA1
	CAS	PEA-Datei mit geheimem Schlüssel
File	HDFS	Kerberos
	NFS	Kerberos, AUTH_SYS

6.3 Data-at-Rest-Verschlüsselung (D@RE)

Complianceanforderungen verlangen oft die Verwendung von Verschlüsselung zum Schutz von Daten, die auf Festplatten geschrieben werden. In ECS kann die Verschlüsselung auf Namespace- und Bucket-Level aktiviert werden. Zu den Hauptfunktionen von ECS D@RE gehören:

- Native Verschlüsselung ruhender Daten ohne großen Aufwand: einfach zu aktivieren, einfache Konfiguration
- CIPHERs(AES-256 CTR) verwendet
- RSA-Verschlüsselung für öffentlichen Schlüssel mit einer Länge von 2048 Bit
- Unterstützung für externes Key-Management (EKM) auf Clusterlevel:
  - Gemalto SafeNet
  - Security Key Lifecycle Manager von IBM
- Schlüsselrotation
- Unterstützung für S3-Verschlüsselungssemantik durch HTTP-Header wie *x-amz-server-side-encryption*
- Compliance gemäß FIPS 140-2 mit kryptografischen Sicherheitsstandards der US-Regierung

Hinweis: Der Modus FIPS 140-2 bewirkt, dass nur genehmigte Algorithmen innerhalb von D@RE verwendet werden; die Compliance gemäß FIPS 140-2 gilt nur für das D@RE-Modul, nicht für das gesamte ECS-Produkt.

ECS verwendet eine Schlüsselhierarchie, um Daten zu ver- und entschlüsseln. Der native Key-Manager speichert einen privaten Schlüssel, den alle Nodes gemeinsam nutzen, um den Primärschlüssel zu entschlüsseln. Bei der EKM-Konfiguration wird der Primärschlüssel vom EKM bereitgestellt. Vom EKM bereitgestellte Schlüssel befinden sich auf ECS nur im Arbeitsspeicher. Sie werden nie in persistentem Speicher in ECS gespeichert.

Wenn ein neues ECS-System in einer geografisch replizierten Umgebung einem vorhandenen Verbund beitritt, wird der Primärschlüssel mithilfe des öffentlich-privaten Schlüssels des bestehenden Systems extrahiert und mit dem neuen öffentlich-privaten Schlüsselpaar verschlüsselt, das vom dem Verbund beigetretenen neuen System erzeugt wurde. Ab diesem Punkt ist der Primärschlüssel global und beiden Systemen innerhalb des Verbunds bekannt. Bei Verwendung von EKM rufen alle Systeme im Verbund den Primärschlüssel aus dem Key-Management-System ab.

6.3.1 Schlüsselrotation

ECS unterstützt das Ändern von Verschlüsselungsschlüsseln. Dies kann regelmäßig erfolgen, um die Menge der durch einen bestimmten Satz von Schlüsselverschlüsselungsschlüsseln (Key Encryption Keys, KEK) geschützten Daten zu begrenzen, oder als Reaktion auf ein mögliches Leck bzw. einen möglichen Angriff geschehen. Ein Rotations-KEK-Datensatz wird in Kombination mit anderen übergeordneten Schlüsseln verwendet, um virtuelle Wrapping-Schlüssel zum Schutz von Datenverschlüsselungsschlüsseln (Data Encryption Keys, DEK) und Namespace-KEKs zu erstellen.

Rotationsschlüssel werden nativ erzeugt oder von einem EKM bereitgestellt und verwaltet. ECS verwendet den aktuellen Rotationsschlüssel, um virtuelle Wrapping-Schlüssel zu erstellen und alle DEK oder KEK zu schützen, unabhängig davon, ob das Key-Management nativ oder extern erfolgt.

Während der Schreibvorgänge packt ECS den zufällig erzeugten DEK in einen virtuellen Wrapping-Schlüssel, der mit dem Bucket und dem aktiven Rotationsschlüssel erstellt wurde.

Im Rahmen der Schlüsselrotation packt ECS alle Namespace-KEK-Datensätze in einen neuen virtuellen Primär-KEK, der aus dem neuen Rotationsschlüssel, dem zugehörigen geheimen Kontext und dem aktiven Primärschlüssel erstellt wurde. Dies geschieht, um den Zugriff auf Daten zu schützen, die durch die vorherigen Rotationsschlüssel geschützt sind.

Die Verwendung eines EKM wirkt sich auf den Lese-/Schreibzugriff für verschlüsselte Objekte aus.
Die Rotation von Schlüsseln ermöglicht zusätzliche Data Protection durch die Verwendung von virtuellen Wrapping-Schlüsseln für DEKs und Namespace-KEKs. Die virtuellen Wrapping-Schlüssel werden nicht dauerhaft gespeichert und werden von 2 unabhängigen Hierarchien persistenter Schlüssel abgeleitet.
Durch die Verwendung von EKM wird der Rotationsschlüssel nicht in ECS gespeichert und trägt zusätzlich zur Datensicherheit bei. Wir fügen in erster Linie neue KEK-Datensätze hinzu und aktualisieren aktive IDs, löschen jedoch niemals etwas.

Diese weiteren Punkte sind in Bezug auf die Schlüsselrotation auf ECS zu berücksichtigen:

- Der Vorgang der Schlüsselrotation ändert nur den aktuellen Rotationsschlüssel. Die vorhandenen Primär-, Namespace- und Bucket-Schlüssel werden während des Schlüsselrotationsprozesses nicht geändert.
- Die Schlüsselrotation auf Namespace- oder Bucket-Level wird nicht unterstützt. Der Rotationsbereich ist jedoch auf Clusterlevel, sodass alle neuen im System verschlüsselten Objekte betroffen sind.
- Vorhandene Daten werden aufgrund von rotierenden Schlüsseln nicht erneut verschlüsselt.
- ECS unterstützt die Rotation von Schlüsseln während Ausfällen nicht.
  - TSO während der Rotation: Die Schlüsselrotationsaufgabe wird angehalten, bis das System nach einem TSO wiederhergestellt wird.
  - PSO ist aktuell im Gange. ECS muss nach einem PSO wiederhergestellt werden, bevor eine Schlüsselrotation möglich ist. Wenn ein PSO während der Rotation auftritt, schlägt die Rotation sofort fehl.

- Bucket-Verschlüsselung ist nicht erforderlich, um die Objektverschlüsselung über S3 durchzuführen.
- Indexierte Clientobjektmetadaten, die als Suchschlüssel verwendet werden, werden nicht verschlüsselt.

Weitere Informationen zu D@RE, zum EKM und zur Schlüsselrotation finden Sie im neuesten *ECS-Sicherheitskonfigurationsleitfaden*.

6.4 ECS IAM

Mit ECS Identify and Access Management (IAM) können Sie den Zugriff auf die ECS-S3-Ressourcen kontrollieren und sichern. Diese Funktion sorgt dafür, dass jede Zugriffsanfrage auf eine ECS-Ressource identifiziert, authentifiziert und autorisiert ist. Mit ECS IAM können Administratoren Nutzer, Rollen und Gruppen hinzufügen. Der Administrator kann den Zugriff auch einschränken, indem er Policies zu den ECS IAM-Entitäten hinzufügt.

Hinweis: ECS IAM ist nur für die Verwendung mit S3 vorgesehen. Es ist für CAS- oder dateisystemfähige Buckets nicht aktiviert.

ECS IAM setzt sich aus den folgenden Komponenten zusammen

- **Account-Management –** Mithilfe der Kontoverwaltung können Sie IAM-Identitäten in jedem Namespace managen, z. b. Nutzer, Gruppen und Rollen.
- **Zugriffsmanagement –** Der Zugriff wird durch die Erstellung von Policies und das Anhängen der IAM-Identitäten oder-Ressourcen verwaltet.
- **Identitätsverbund –** Identität wird von SAML (Security Assertion Markup Language) eingerichtet und authentifiziert. Nachdem die Identität hergestellt wurde, verwenden Sie den Secure Token Service, um temporäre Zugangsdaten zu erhalten, die für den Zugriff auf die Ressource verwendet werden.
- **Secure Token Service –** Ermöglicht Ihnen die Anforderung temporärer Zugangsdaten für kontoübergreifenden Zugriff auf Ressourcen und auch für Nutzer, die mithilfe der SAML-Authentifizierung von einem Unternehmensidentitätsanbieter oder Verzeichnisdienst authentifiziert werden

Mit IAM können Sie kontrollieren, wer authentifiziert und autorisiert ist, ECS-Ressourcen zu nutzen, indem Sie Folgendes erstellen und verwalten

- **Nutzer –** Ein IAM-Nutzer repräsentiert eine Person oder Anwendung im Namespace, die mit ECS-Ressourcen interagieren kann
- **Gruppen –** Eine IAM-Gruppe ist eine Sammlung von IAM-Nutzern. Mit Gruppen können Sie Berechtigungen für eine Sammlung von IAM-Nutzern festlegen
- **Rollen –** Die IAM-Rolle ist eine Identität, die von allen Personen übernommen werden kann, die die Rolle benötigen. Eine Rolle ähnelt einem Nutzer, einer Identität mit Berechtigungsrichtlinien, die festlegen, was die Identität tun kann und was nicht.
- **Policies –** Eine IAM-Policy ist ein Dokument im JSON-Format, das die Berechtigungen für eine Rolle definiert. Zuweisen und Anhängen von Policies zu IAM-Nutzern, IAM-Gruppen und IAM-Rollen.
- **SAML-Anbieter –** Security Assertion Markup Language (SAML) ist ein offener Standard für den Austausch von Authentifizierungs- und Autorisierungsdaten zwischen Parteien, insbesondere zwischen einem Identitätsanbieter und einem Serviceanbieter. Der SAML-Anbieter in ECS wird verwendet, um die Vertrauensbeziehung zwischen einem SAML-kompatiblen Identitätsanbieter (IdP) und ECS herzustellen.

Jedem ECS-System wird ein ECS-IAM-Konto zugewiesen. Dieses Konto unterstützt mehrere Namespaces und verfügt über zugehörige IAM-Entitäten, die in seinem Namespace definiert sind.

- Einzelne Namespaces unterstützen das Kontomanagement mithilfe der ECS-IAM-Entitäten wie Nutzer, Rollen und Gruppen.
- Policies, Berechtigungen, Zugriffskontrolllisten (Access Control List, ACL), die mit den ECS IAM-Entitäten verknüpft sind, und die ECS S3-Ressourcen unterstützen das Management des Zugriffs auf die ECS-IAM-Funktionen.
- ECS IAM unterstützt kontoübergreifenden Zugriff mithilfe von SAML (Security Assertion Markup Language) und Rollen.
- ECS IAM unterstützt den Zugriffsschlüssel für Amazon Web Services (AWS) für den Zugriff auf IAM und S3 in ECS.

Weitere Informationen zu ECS IAM finden Sie im neuesten *ECS-Sicherheitsleitfaden.*

6.5 Objekttagging

Das Objekttagging ermöglicht die Kategorisierung von Objekten durch Zuweisung von Tags zu den einzelnen Objekten. Ein einzelnes Objekt kann mehrere Tags haben, die mit ihm verknüpft sind, wodurch die mehrdimensionale Kategorisierung möglich wird.

Ein Tag könnte eine Art von vertraulicher Information beschreiben, wie z. B. eine Gesundheitsakte, oder Sie können ein Objekt einem bestimmten Produkt zuordnen, das als vertraulich eingestuft werden kann. Tagging ist eine Teilressource eines Objekts mit einem Lebenszyklus, integriert in Objektvorgänge. Sie können neuen Objekten Tags hinzufügen, wenn Sie diese hochladen, oder vorhandene Objekte mit Tags versehen. Es ist akzeptabel, Tags zur Etikettierung von Objekten mit vertraulichen Daten, wie personenbezogene Informationen (Personally Identifiable Information, PII) oder geschützte Integritätsinformationen (Protected Health Information, PHI), zu verwenden. Die Tags dürfen keine vertraulichen Informationen enthalten, da Tags ohne die eigentliche Leseberechtigung für ein Objekt angezeigt werden können.

6.5.1 Zusätzliche Informationen zum Objekttagging

In diesem Abschnitt finden Sie Informationen zum Objekttagging in IAM, Objekttagging mit Bucket-Policies, Umgang von Objekttagging während des TSO/PSO und Objekttagging während des Objektlebenszyklusmanagements. Hier sind zusätzliche Hinweise:

- Objekttagging in IAM
  - Die Schlüsselfunktion des Objekttagging als Kategorisierungssystem kommt zum Tragen, wenn es in IAM-Policies integriert wird. Auf diese Weise können Administratoren bestimmte Nutzerberechtigungen konfigurieren. Administratoren können z. B. eine Policy hinzufügen, mit der jeder auf Objekte mit einem bestimmten Tag zugreifen kann, oder sie können Berechtigungen für Nutzer, die Tags an bestimmten Objekten managen können, konfigurieren und diesen erteilen. Der andere wichtige Aspekt beim Objekttagging ist, wie und wo die Tags dauerhaft aufbewahrt werden. Dies ist wichtig, da es direkte Auswirkungen auf verschiedene Aspekte des Systems hat.

- Objekttagging mit Bucket-Policies
  - Mithilfe von Objekttagging können Sie Objekte kategorisieren. Außerdem wird Tagging mit verschiedenen Policies integriert. Die Lebenszyklusmanagement-Policy ermöglicht die Konfiguration auf Bucket-Ebene. Frühere Versionen von ECS unterstützen Verfall, Abbruch unvollständiger Uploads und Löschen der Markierung für abgelaufene Objekte (Delete Marker). Der Filter kann mehrere Bedingungen enthalten, einschließlich einer Tag-basierten Bedingung. Jedes Tag in der Filterbedingung muss mit dem Schlüssel und dem Wert übereinstimmen.
- Objekttagging während TSO/PSO
  - Das Objekttagging ist ein weiterer Eintrag, der in den Systemmetadaten festgelegt ist, während TSO/PSO ist keinen speziellen Umgang erfordert. Es gibt eine festgelegte Begrenzung für die Anzahl an Tags, die mit jedem Objekt verknüpft werden können, die Größe der Systemmetadaten im Zusammenhang mit dem Objekttagging liegt innerhalb der Speicherbeschränkungen.
- Objekttagging während des Objektlebenszyklusmanagements
  - Das Objekttagging ist Teil der Systemmetadaten und wird während des Lebenszyklusmanagements gleichzeitig mit den Systemmetadaten verarbeitet.
    Die Ablauflogik und der Lifecycle Delete Scanner müssen Tag-basierte Policies verstehen. Objekttags ermöglichen ein differenziertes Lebenszyklusmanagement für das Objekt, bei dem Sie zusätzlich zu einem Schlüsselnamenpräfix einen Tag-basierten Filter in einer Lebenszyklusregel angeben können.

Weitere Informationen zum ECS Objekttagging finden Sie im neuesten *ECS-Sicherheitskonfigurationsleitfaden*.

7 Datenintegrität und Data Protection

Für die Datenintegrität verwendet ECS Prüfsummen. Prüfsummen werden während Schreibvorgängen erstellt und mit den Daten gespeichert. Bei Lesevorgängen werden die Prüfsummen berechnet und mit der gespeicherten Version verglichen. Eine Hintergrundaufgabe scannt und überprüft die Prüfsummeninformationen proaktiv.

Für Data Protection verwendet ECS eine 3-Fachspiegelung von Journalblöcken und separate EC-Schemata für *repo*- und *btree*-Blöcke (Nutzer-Repository-Daten bzw. B+-Struktur).

Erasure Coding bietet im Vergleich zu herkömmlichen Schutzsystemen verbesserte Data Protection vor einem Festplatten-, Node- und Rackausfall bei guter Speichereffizienz. Die ECS-Speicher-Engine implementiert die Reed-Solomon-Fehlerkorrektur mithilfe von 2 Schemata:

- 12+4 (Standard): Der Block wird in 12 Datensegmente unterteilt. 4 Codierungssegmente (Parität) werden erstellt.
- 10+2 (Archiv für inaktive Daten): Der Block wird in 10 Datensegmente unterteilt. 2 Codierungssegmente werden erstellt.

Mit dem Standardwert von 12+4 werden die daraus resultierenden 16 Segmente über Nodes am lokalen Standort verteilt. Die Daten- und Coding-Segmente der einzelnen Blöcke werden gleichmäßig über Nodes im Cluster verteilt. Beispielsweise hat bei 8 Nodes jeder Node 2 Segmente (von insgesamt 16). Die Speicher-Engine kann einen Block aus 12 beliebigen der 16 Segmente rekonstruieren.

ECS benötigt mindestens 6 Nodes für die Option des Archivs für inaktive Daten, bei der das Schema 10+2 anstatt 12+4 verwendet wird. EC wird beendet, wenn die Anzahl der Nodes unter das für das EC-Schema erforderliche Minimum fällt.

Wenn ein Block voll ist oder ein festgelegter Zeitraum abgeschlossen wird, wird die Parität berechnet und die Coding-Segmente werden in der Fehlerdomain auf die Festplatten geschrieben. Blockdaten bleiben als eine einzige Kopie erhalten, die aus 16 Segmenten (12 Daten-, 4 Codesegmenten) bestehen, die über den gesamten Cluster verteilt sind. ECS verwendet Codefragmente nur dann für die Blockrekonstruktion, wenn ein Fehler auftritt.

Wenn die zugrunde liegende Infrastruktur eines VDC auf Node- oder Racklevel geändert wird, erkennt die Fabric-Schicht die Änderung und löst einen Scanner zur Abstimmung als Hintergrundaufgabe aus. Der Scanner berechnet mithilfe der neuen Topologie für jeden Block das beste Layout für EC-Segmente in Fehlerdomains. Wenn das neue Layout einen besseren Schutz als das vorhandene Layout bietet, verteilt ECS die EC-Segmente in einer Hintergrundaufgabe neu. Diese Aufgabe hat minimale Auswirkungen auf die Systemperformance. Es gibt jedoch während der Abstimmung einen Anstieg des Datenverkehrs zwischen den Nodes. Gleichzeitig werden die logischen Tabellenpartitionen auf die neuen Nodes verteilt und neu erstellte Journal- und B+-Strukturblöcke werden in weiterer Folge gleichmäßig alten und neuen Nodes zugewiesen. Die Neuverteilung verbessert den lokalen Schutz durch Nutzung aller Ressourcen in der Infrastruktur.

Hinweis: Es wird empfohlen, vor dem Hinzufügen von Laufwerken oder Nodes nicht zu warten, bis die Storage-Plattform vollständig ausgelastet ist. Eine vernünftige Speicherauslastungsschwelle liegt bei 70 % unter Berücksichtigung der täglichen Aufnahmerate und der erwarteten Bestell-, Liefer- und Integrationszeit der hinzugefügten Laufwerke/Nodes.

7.1 Compliance

Um die Complianceanforderungen von Unternehmen und der Branche (SEC Rule 17a–4(f)) für die Speicherung von Daten zu erfüllen, hat ECS Folgendes implementiert:

- **Plattformverstärkung:** Verstärkung behebt Sicherheitsschwachstellen in ECS, etwa Plattformsperren, um den Zugriff auf Nodes oder Cluster zu deaktivieren, alle nicht wesentlichen Ports (z. B. *ftpd, sshd*) werden geschlossen, vollständige Auditprotokollierung für sudo-Befehle und Unterstützung für SRS (Dell EMC Secure Remote Services), um den Remotezugriff auf Nodes zu deaktivieren.
- **Compliancereporting:** Ein System-Agent meldet den Compliancestatus des Systems, z. B. *Good*, was Einhaltung der Compliance anzeigt, oder *Bad*, was die Nichteinhaltung anzeigt.
- **Policy-basierte Datensatzaufbewahrung und Regeln:** Möglichkeit zur Begrenzung von Änderungen an aufbewahrten Datensätzen oder Daten mithilfe von Policies, Zeiträumen und Regeln.
- **Erweitertes Aufbewahrungsmanagement (Advanced Retention Management, ARM):** Zur Erfüllung der Centera-Complianceanforderungen wurde nur für CAS eine Reihe von Aufbewahrungsregeln definiert.
  - **Ereignisbasierte Aufbewahrung:** Ermöglicht Aufbewahrungsfristen, die beim Auftreten eines bestimmten Ereignisses beginnen.
  - **Gesetzliche Aufbewahrungsfrist:** Ermöglicht das vorübergehende Blockieren der Löschung von Daten, die für rechtliche Maßnahmen benötigt werden.
  - **Min-/Max-Kontrolle:** Bucket-spezifische Einstellung für die minimale und maximale Standardaufbewahrungsfrist.

Compliance wird auf Namespace-Level aktiviert. Aufbewahrungsfristen werden auf Bucket-Level konfiguriert. Complianceanforderungen zertifizieren die Plattform und deshalb ist die Compliancefunktion nur für ECS verfügbar, das auf Appliance-Hardware ausgeführt wird. Informationen zum Aktivieren und Konfigurieren der Compliance in ECS finden Sie im aktuellen *ECS-Leitfaden für den Datenzugriff* und im aktuellen *ECS-Administratorhandbuch.*

8 Bereitstellung

ECS kann als einzelne oder auf mehrere Standorte verteilte Instanz bereitgestellt werden. Die Komponenten einer ECS-Bereitstellung umfassen Folgendes:

- **Virtuelles Rechenzentrum (VDC):** Ein Cluster, der in der Regel auch als Standort oder geografische Region bezeichnet wird und aus einem Satz von ECS-Infrastruktur besteht, der von einer einzigen Fabric-Instanz gemanagt wird.
- **Speicherpool (SP):** SPs können als Untergruppe der Nodes und ihres zugehörigen Speichers betrachtet werden, die zu einem VDC gehört. Ein Node kann nur einem SP angehören. EC wird auf SP-Level mit einem Schema von 12+4 oder 10+2 festgelegt. Ein SP kann als Tool zur physischen Trennung von Daten zwischen Clients oder Gruppen von Clients verwendet werden, die auf ECS auf Speicher zugreifen.
- **Replikationsgruppe (RG):** RGs legen fest, wo der SP-Inhalt geschützt ist und an welchen Speicherorten auf Daten zugegriffen werden kann. Eine RG mit einem einzigen Mitgliedsstandort wird manchmal als lokale RG bezeichnet. Daten werden immer dort lokal vor Festplatten-, Node- und Rackausfällen geschützt, wo sie geschrieben werden. RGs mit 2 oder mehr Standorten werden oft als globale RGs bezeichnet. Globale RGs umfassen bis zu 8 VDCs und schützen vor Festplatten-, Node-, Rack- und Standortausfällen. Ein VDC kann zu mehreren RGs gehören.
- **Namespace:** Ein Namespace ist im Prinzip identisch mit einem Mandanten in ECS. Ein wichtiges Merkmal eines Namespace besteht darin, dass Nutzer von einem Namespace nicht auf Objekte zugreifen können, die einem anderen Namespace angehören.
- **Buckets:** Buckets sind Container für Objekte, die in einem Namespace erstellt und manchmal als logischer Container für Untermandanten angesehen werden. In S3 werden Objektcontainer Buckets genannt. Dieser Ausdruck wurde in ECS übernommen. In Atmos ist das Äquivalent eines Bucket ein Untermandant, in Swift ist das Äquivalent eines Bucket ein Container und bei CAS entspricht ein Bucket einem CAS-Speicherpool. Buckets sind globale Ressourcen in ECS. Jeder Bucket wird in einem Namespace erstellt und jeder Namespace wird in einer RG erstellt.

ECS nutzt die folgenden Infrastruktursysteme:

- **DNS** (erforderlich): Forward und Reverse Lookups, die für jeden ECS-Node erforderlich sind.
- **NTP** (erforderlich): Network Time Protocol-Server.
- **SMTP** (optional): Simple Mail Transfer Protocol-Server für das Senden von Warnmeldungen und Berichten.
- **DHCP** (optional): Erforderlich, wenn IP-Adressen über DHCP zugewiesen werden.
- **Authentifizierungsanbieter** (optional): ECS-Administratoren können mithilfe von Active Directory- und LDAP-Gruppen authentifiziert werden. Objektnutzer können mithilfe von Keystone authentifiziert werden. Authentifizierungsanbieter sind für ECS nicht erforderlich. ECS verfügt über integrierte Funktionen für das lokale Nutzermanagement. Beachten Sie jedoch, dass Nutzer, die lokal erstellt werden, nicht zwischen VDCs repliziert werden.

- **Load Balancer** (erforderlich, wenn der Workflow es vorgibt, andernfalls optional): Die Clientlast sollte auf alle Nodes verteilt werden, um alle im System verfügbaren Ressourcen effektiv zu nutzen. Wenn eine dedizierte Load Balancer Appliance oder ein dedizierter Load-Balancer-Service erforderlich ist, um die Last über ECS-Nodes zu managen, sollte dies als erforderlich betrachtet werden. Entwickler, die Anwendungen mithilfe der ECS S3 SDK schreiben, können die integrierte Load-Balancer-Funktion nutzen. Anspruchsvolle Load Balancers können zusätzliche Faktoren berücksichtigen, etwa vom Server gemeldete Last, Antwortzeiten, aktiver/nicht aktiver Status, Anzahl der aktiven Verbindungen und geografischer Standort. Der Kunde ist dafür verantwortlich, den Clientdatenverkehr zu managen und Zugriffsanforderungen festzulegen. Unabhängig von der Methode gibt es einige grundlegende Optionen, die in der Regel berücksichtigt werden, darunter manuelle IP-Zuweisung, DNS-Rundlaufverfahren, clientseitiger Lastenausgleich, Load Balancer Appliances und geografische Load Balancers. Im Folgenden finden Sie eine kurze Beschreibung der einzelnen Methoden:

  - **Manuelle IP-Zuweisung:** IP-Adressen werden manuell an Anwendungen verteilt. Dies wird in der Regel nicht empfohlen, da es möglicherweise keine Last verteilt oder Fehlertoleranz bietet.
  - **DNS-Rundlaufverfahren:** Ein DNS-Eintrag wird erstellt, der alle Node-IP-Adressen enthält. Clients fragen den DNS ab, um vollständig qualifizierte Domainnamen für ECS-Services aufzulösen, und erhalten die IP-Adressen eines zufälligen Node als Antwort. Dies kann einen „Pseudo-Lastenausgleich“ ermöglichen. Diese Methode bietet möglicherweise keine Fehlertoleranz, da häufig manuell eingegriffen wird, um IP-Adressen ausgefallener Nodes aus dem DNS zu entfernen. Bei dieser Methode können TTL-Probleme (Time to Live) auftreten. Einige DNS-Serverimplementierungen können DNS Lookups für einen bestimmten Zeitraum zwischenspeichern, sodass Clients, die innerhalb eines kurzen Zeitrahmens eine Verbindung herstellen, möglicherweise an dieselbe IP-Adresse gebunden werden, wodurch die Lastverteilung auf die Daten-Nodes reduziert wird. Die Verwendung von DNS für die Verteilung des Datenverkehrs in einem DNS-Rundlaufverfahren wird nicht empfohlen.
  - **Lastenausgleich:** Load Balancers sind der häufigste Ansatz für die Verteilung der Clientlast. Clients können Datenverkehr zum Load Balancer senden, wo er empfangen und an einen fehlerfreien ECS-Node weitergeleitet wird. Proaktive Integritätsprüfungen oder der Verbindungsstatus werden verwendet, um die Verfügbarkeit der einzelnen Nodes für Service-Requests zu überprüfen. Nicht verfügbare Nodes werden für die Verwendung entfernt, bis sie eine Integritätsprüfung bestehen. Die Auslagerung der CPU-intensiven SSL-Verarbeitung kann genutzt werden, um diese Ressourcen auf ECS freizugeben.
  - **Geografischer Lastenausgleich:** Der geografische Lastenausgleich nutzt DNS, um Suchvorgänge an eine Appliance, z. B. Riverbed SteelApp, weiterzuleiten, die Geo-IP oder einen anderen Mechanismus nutzt, um den besten Standort für die Weiterleitung des Clients zu ermitteln.

8.1 Bereitstellung am Einzelstandort

Während der anfänglichen Bereitstellung eines einzelnen Standorts oder eines einzelnen Clusters werden Nodes zuerst zu einem SP hinzugefügt. SPs sind logische Container für physische Nodes. Die SP-Konfiguration umfasst das Auswählen der erforderlichen Mindestanzahl verfügbarer Nodes und das Auswählen des standardmäßigen EC-Schemas 12+4 oder des EC-Schemas 10+2 für das Archiv für inaktive Daten. Kritische Warnmeldungslevel können während der SP-Konfiguration anfänglich und später festgelegt werden. Ein EC-Schema kann jedoch nach der SP-Initialisierung nicht mehr geändert werden. Der erste erstellte SP wird als System-SP festgelegt und zum Speichern von Systemmetadaten verwendet. Der System-SP kann nicht gelöscht werden.

Cluster enthalten in der Regel wie in Abbildung 21 gezeigt 1 oder 2 SPs, einen für jedes EC-Schema. Wenn ein Unternehmen jedoch eine physische Trennung von Daten benötigt, werden zusätzliche SPs zur Implementierung von Grenzen verwendet.

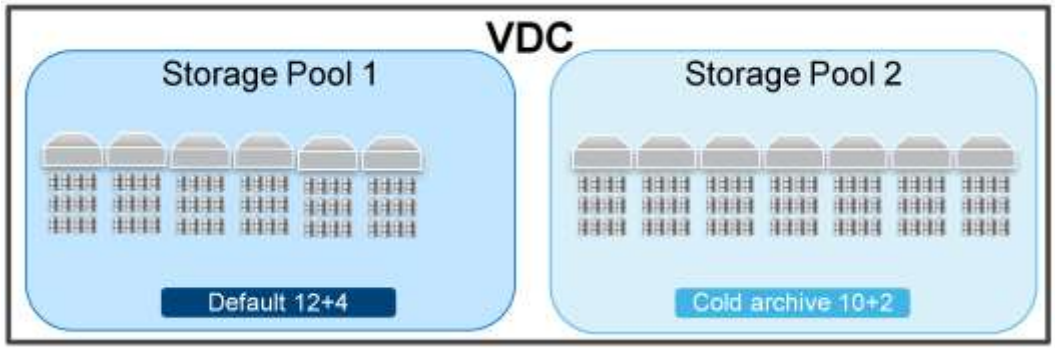

Abbildung 21 VDC mit 2 Storage-Pools, die mit unterschiedlichen EC-Schemata konfiguriert sind

Nach der Initialisierung des ersten SP kann ein VDC erstellt werden. Die VDC-Konfiguration umfasst die Festlegung von Replikations- und Managementendpunkten. Beachten Sie, dass zwar eine System-SP-Initialisierung vor der VDC-Erstellung erforderlich ist, die VDC-Konfiguration jedoch keine SPs, sondern die IP-Adressen von Nodes zuweist.

Nachdem ein VDC erstellt wurde, werden RGs konfiguriert. RGs sind globale Ressourcen mit einer Konfiguration, bei der mindestens ein VDC selbst bei der Einzel- oder Ersteinrichtung zusammen mit einem der SPs des VDC festgelegt wird. Eine RG mit einem einzigen VDC-Mitglied schützt Daten lokal auf Festplatten-, Node- und Racklevel. Im nächsten Abschnitt werden RGs mit Bereitstellungen an mehreren Standorten behandelt.

Namespaces sind globale Ressourcen, die erstellt und einer RG zugewiesen werden. Auf Namespace-Level werden Aufbewahrungs-Policies, Quoten, Compliance- und Namespace-Administratoren definiert. Der Zugriff während eines Ausfalls (Access During Outage, ADO) kann auf dem Namespace-Level konfiguriert werden, die im nächsten Abschnitt behandelt wird. In der Regel ist es der Namespace-Level, auf dem Mandanten organisiert werden. Mandanten können eine Anwendungsinstanz oder ein Team, ein Nutzer, eine Geschäftsgruppe oder eine andere Gruppierung sein, die für das Unternehmen Sinn ergibt.

Buckets sind globale Ressourcen, die sich über mehrere Standorte erstrecken können. Bei der Bucket-Erstellung muss dieser einem Namespace und einer RG zugewiesen werden. Auf Bucket-Level werden Eigentumsrechte und Datei- oder CAS-Zugriff aktiviert. Abbildung 22 unten zeigt einen SP in einem VDC mit einem Namespace, der 2 Buckets enthält.

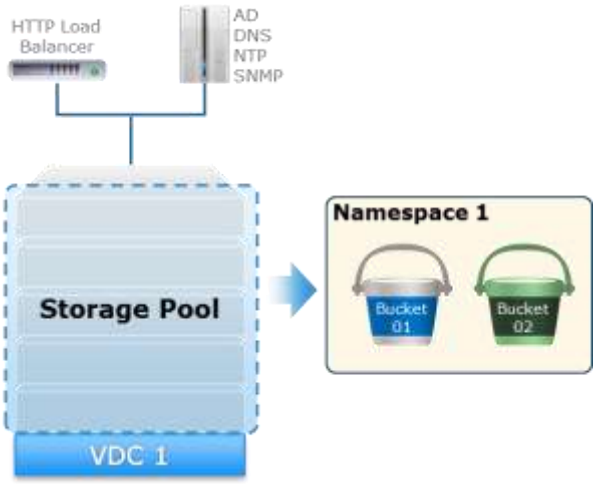

Abbildung 22 Beispiel für die Bereitstellung an einem einzelnen Standort

8.2 Bereitstellung an mehreren Standorten

Eine Bereitstellung an mehreren Standorten, die auch als Verbundumgebung oder Verbund-ECS bezeichnet wird, kann sich über bis zu 8 VDCs erstrecken. Daten werden in ECS auf Blocklevel repliziert. Nodes, die an einer RG teilnehmen, senden ihre lokalen Daten asynchron an einen oder alle anderen Standorte. Daten werden mithilfe von AES256 verschlüsselt, bevor sie mittels HTTP über das WAN gesendet werden. Die wichtigsten Vorteile, die bei der Zusammenführung mehrerer VDCs zu einem Verbund erkannt werden, sind die folgenden:

- Zusammenführung des Managements mehrerer VDCs in einer einzigen logischen Ressource
- Schutz auf Standortlevel zusätzlich zu lokal auf Node-, Festplatten- und Racklevel
- Geografisch verteilter Zugriff auf Speicher in einer überall aktiven, stark konsistenten Art und Weise

In diesem Abschnitt zur Bereitstellung an mehreren Standorten werden die für Verbund-ECS spezifischen Funktionen beschrieben, darunter:

- **Datenkonsistenz:** ECS bietet standardmäßig einen stark konsistenten Speicherservice.
- **Replikationsgruppen:** Globale Container, die verwendet werden, um Schutz und Zugriffsgrenzen festzulegen.
- **Geo-Caching:** Optimierung für Workflows für den Remotestandortzugriff in Bereitstellungen an mehreren Standorten.
- **ADO:** Clientzugriffsverhalten während eines vorübergehenden Standortausfalls (Temporary Site Outage, TSO).

8.2.1 Datenkonsistenz

ECS ist ein stark konsistentes System, das Eigentumsrechte verwendet, um eine maßgebliche Version der einzelnen Namespaces, Buckets und Objekte beizubehalten. Die Eigentumsrechte werden dem VDC zugewiesen, in dem der Namespace, der Bucket oder das Objekt erstellt wird. Beispiel: Wenn der Namespace NS1 im VDC1 erstellt wird, besitzt VDC1 NS1 und ist für die Verwaltung der maßgeblichen Version von Buckets in NS1 verantwortlich. Wenn ein Bucket B1 auf VDC2 in NS1 erstellt wird, besitzt VDC2 B1 und ist verantwortlich für die Aufrechterhaltung der maßgeblichen Version des Bucket-Inhalts sowie des Eigentümer-VDC jedes Objekts. Ebenso gilt: Wenn etwa ein Objekt O1 in B1 in VDC3 erstellt wird, besitzt VDC3 O1 und ist für die Verwaltung der maßgeblichen Version von O1 und der zugehörigen Metadaten verantwortlich.

Die Ausfallsicherheit der Data Protection an mehreren Standorten geht auf Kosten des erhöhten Speicherschutzoverheads und der WAN-Bandbreitennutzung. Indexabfragen sind erforderlich, wenn auf ein Objekt über einen Standort zugegriffen oder es über ihn aktualisiert wird, der das Objekt nicht besitzt. Auf ähnliche Weise sind Indexabfragen über WAN auch zum Abrufen von Informationen wie einer maßgeblichen Liste von Buckets in einem Namespace oder Objekten in einem Bucket erforderlich, die einem Remotestandort gehören.

Das Wissen, wie ECS Eigentumsrechte nutzt, um Daten auf Namespace-, Bucket- und Objektlevel autorisierend zu verfolgen, hilft Administratoren und Anwendungseigentümern, Entscheidungen beim Konfigurieren ihrer Umgebung bezüglich des Zugriffs zu treffen.

8.2.2 Aktive Replikationsgruppe

Während der RG-Erstellung ist die Einstellung *Replicate to All Sites* verfügbar, die entweder standardmäßig deaktiviert ist oder durch Umschalten aktiviert werden kann. Das Replizieren von Daten an allen Standorten bedeutet, dass die einzeln auf die einzelnen VDCs geschriebenen Daten auf allen anderen RG-Mitglieds-VDCs repliziert werden. Beispielsweise bedeutet eine Verbund-ECS-Instanz mit x Standorten und einer aktiven RG, die zum Replizieren von Daten an allen Standorten konfiguriert ist, einen x-fachen Schutzoverhead oder x * 1,33 (bzw. 1,2 bei EC für Archiv für inaktive Daten) an Gesamt-Datenschutzoverhead. Die Replikation an allen Standorten kann insbesondere bei kleineren Datenvolumen sinnvoll sein, bei denen lokaler Zugriff wichtig ist. Wenn diese Einstellung deaktiviert ist, werden alle auf die einzelnen VDCs geschriebenen Daten in einem anderen VDC repliziert. Der primäre Standort, an dem das Objekt erstellt wird, und der Standort, der die replizierte Kopie speichert, schützen die Daten lokal mit dem EC-Schema, das dem lokalen SP zugewiesen ist. Das bedeutet, dass nur die ursprünglichen Daten über das WAN repliziert werden und keine zugehörigen EC-Coding-Segmente.

Daten, die in einer aktiven RG gespeichert sind, sind für Clients über alle verfügbaren RG-Mitglieds-VDCs zugänglich. In Abbildung 23 unten sehen Sie ein Beispiel für ein Verbund-ECS-System, das mit VDC1, VDC2 und VDC3 erstellt wurde. 2 RGs werden angezeigt. RG1 hat ein einzelnes Mitglied, VDC1, und RG2 hat alle 3 VDCs als Mitglieder. 3 Buckets werden angezeigt: B1, B2 und B3.

In diesem Beispiel haben Clients mit Zugriff auf:

- VDC1 Zugriff auf alle Buckets.
- VDC2 und VDC3 nur Zugriff auf Buckets B2 und B3.

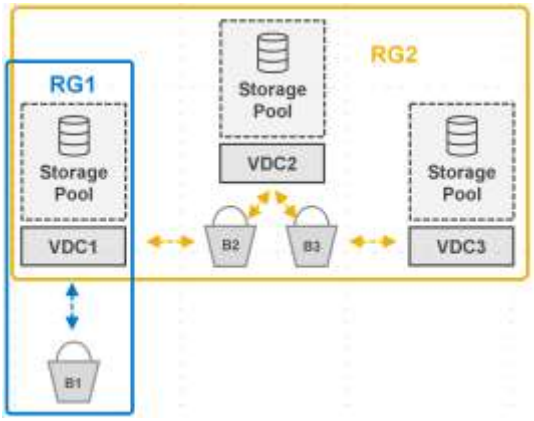

Abbildung 23 Zugriff auf Bucket-Level nach Standort mit Replikationsgruppen mit einem oder mehreren Standorten

8.2.3 Passive Replikationsgruppe

Eine passive RG hat 3 Mitglieds-VDCs. 2 der VDCs werden als aktiv gekennzeichnet und sind für Clients zugänglich. Das 3. VDC ist als passiv festgelegt und wird nur als Replikationsziel verwendet. Der passive Standort wird nur für Recovery-Zwecke verwendet und erlaubt keinen direkten Clientzugriff. Vorteile der Geo-Passive-Replikation:

- Reduzierter Speicherschutzoverhead durch höheres Potenzial für XOR-Vorgänge
- Kontrolle auf Administratorlevel des für reinen Replikationsspeicher verwendeten Standorts

Abbildung 24 zeigt ein Beispiel für eine Geo-Passive-Konfiguration, bei der VDC 1 und VDC 2 primäre Standorte (Quellstandorte) sind, die ihre Daten (Blöcke) auf dem Replikationsziel VDC 3 replizieren.

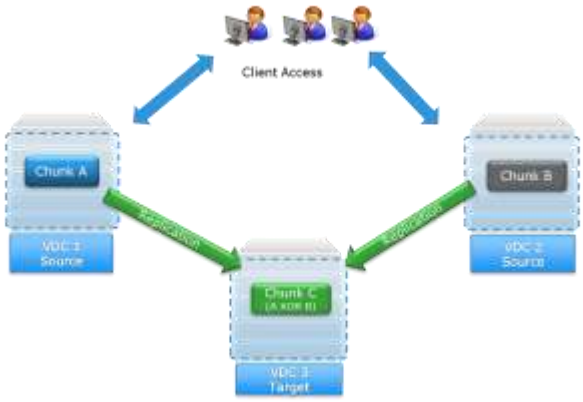

Abbildung 24 Client-Zugriff und Replikationspfade für Geo-Passive-Replikationsgruppe

Der Zugriff auf stark konsistente Daten an mehreren Standorten erfolgt mithilfe von Namespace-, Bucket- und Objekteigentumsrechten an RG-Mitgliedsstandorten. Standortübergreifende WAN-Indexabfragen sind erforderlich, wenn der API-Zugriff von einem VDC stammt, das nicht Eigentümer der erforderlichen logischen Konstrukte ist. WAN Lookups werden verwendet, um die maßgebliche Version der Daten zu bestimmen. Wenn also ein an Standort 1 erstelltes Objekt von Standort 2 gelesen wird, ist ein WAN Lookup erforderlich, um das Eigentümer-VDC des Objekts, Standort 1, abzufragen, um zu überprüfen, ob die Daten des Objekts, die auf Standort 2 repliziert wurden, die aktuelle Version der Daten sind. Wenn Standort 2 nicht über die neueste Version verfügt, werden die erforderlichen Daten von Standort 1 abgerufen. Andernfalls werden die zuvor darauf replizierten Daten verwendet. Dies ist unten in Abbildung 25 dargestellt.

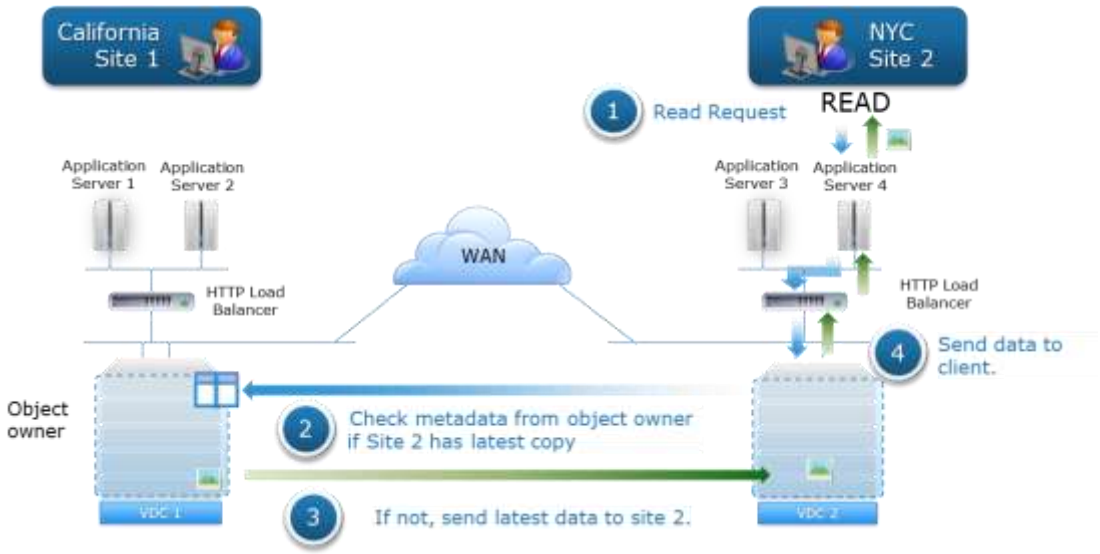

Abbildung 25 Leseanfrage an Nicht-Eigentümer-VDC löst WAN Lookup für Objekteigentümer-VDC aus

Der Datenfluss von Schreibvorgängen in einer geografisch replizierten Umgebung, in der 2 Standorte dasselbe Objekt aktualisieren, wird in Abbildung 26 gezeigt. In diesem Beispiel hat Standort 1 das Objekt ursprünglich angelegt und besitzt es. Das Objekt wurde gelöscht und die zugehörigen Journaltransaktionen wurden an Standort 1 auf die Festplatte geschrieben. Der Datenfluss für ein Update auf das an Standort 2 empfangene Objekt sieht folgendermaßen aus:

1. Standort 2 schreibt die Daten zunächst lokal.
2. Standort 2 aktualisiert die Metadaten (Journalschreibvorgänge) synchron mit dem Eigentümer des Objekts, Standort 1, und wartet auf die Bestätigung des Metadatenupdates von Standort 1.
3. Standort 1 bestätigt den Schreibvorgang für Metadaten auf Standort 2.
4. Standort 2 bestätigt den Schreibvorgang auf dem Client.

Hinweis: Standort 2 repliziert den Datenstandort 1, den Standort des Objekteigentümers, wie gewohnt asynchron. Wenn die Daten von Standort 1 bereitgestellt werden müssen, bevor sie von Standort 2 auf ihm repliziert werden, ruft Standort 1 die Daten direkt von Standort 2 ab.

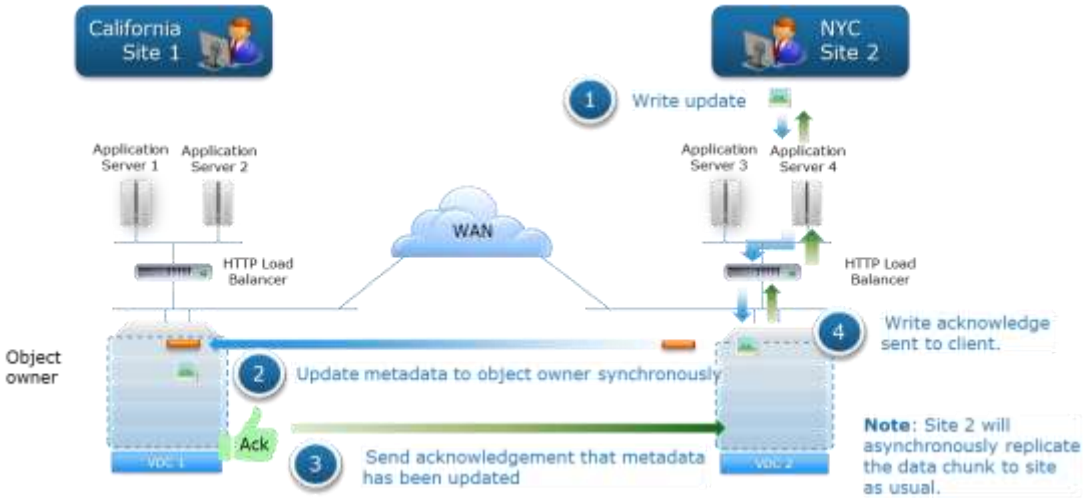

Abbildung 26 Datenfluss bei Update desselben Objekts in einer geografisch replizierten Umgebung

In Lese- und Schreibszenarien in einer geografisch replizierten Umgebung gibt es eine Latenz beim Lesen und Aktualisieren der Metadaten und beim Abrufen von Daten vom Objekteigentümerstandort.

Hinweis: Ab ECS 3.4 können Sie ein VDC aus einer Replikationsgruppe (RG) in einem Verbund mit mehreren VDCs entfernen, ohne dass dies Auswirkungen auf das VDC oder andere mit dem VDC verbundene RGs hat. Durch das Entfernen eines VDC aus einer RG wird kein PSO mehr initiiert (permanenter Standortausfall). Durch das Entfernen eines VDC aus einer RG wird die Recovery initiiert.

Weitere Informationen zur Replikationsgruppe finden Sie im neuesten *ECS-Administratorhandbuch.*

8.2.4 Geo-Caching von Remotedaten

ECS optimiert die Antwortzeiten für den Zugriff auf Daten, die an Remotestandorten gespeichert sind, indem Objekte, die über die WAN gelesen werden, lokal zwischengespeichert werden. Dies kann für Zugriffsmuster mit mehreren Standorten nützlich sein, bei denen Daten oft von einem Remotestandort oder einem Nicht-Eigentümer-Standort abgerufen werden. Nehmen wir beispielsweise eine geografisch replizierte Umgebung mit 3 Standorten – VDC1, VDC2 und VDC3 – an, wobei ein Objekt in VDC1 geschrieben wird und die replizierte Kopie des Objekts in VDC2 gespeichert wird. In diesem Szenario müssen die Objektdaten für eine Leseanforderung, die in VDC3 empfangen wurde, für das in VDC1 erstellte und in VDC2 replizierte Objekt entweder von VDC1 oder VDC2 an VDC3 gesendet werden. Das Geo-Caching von häufig abgerufenen Remotedaten trägt dazu bei, die Reaktionszeiten zu verkürzen. Ein LRU-Algorithmus (Least Recently Used) wird für das Caching verwendet. Die Geo-Cache-Größe wird angepasst, wenn Hardwareinfrastruktur wie Festplatten, Nodes und Racks zu einem geografisch replizierten SP hinzugefügt wird.

8.2.5 Verhalten beim Standortausfall

Ein vorübergehender Standortausfall (Temporary Site Outage, TSO) bezieht sich in der Regel entweder auf einen Ausfall der WAN-Konnektivität oder eines gesamten Standorts, z. B. während einer Naturkatastrophe. ECS verwendet Heartbeat-Mechanismen, um vorübergehende Standortausfälle zu erkennen und darauf zu reagieren. Der Clientzugriff und die Verfügbarkeit des API-Vorgangs auf Namespace-, Bucket- und Objektlevel während eines TSO wird durch die folgenden ADO-Optionen geregelt, die auf Namespace- und Bucket-Level festgelegt sind:

- **Deaktiviert (Standard):** Bei einem vorübergehenden Ausfall bleibt die starke Konsistenz erhalten.
- **Aktiviert:** Zugriff mit letztendlicher Konsistenz wird während eines vorübergehenden Standortausfalls erlaubt.

Datenkonsistenz während eines TSO wird auf Bucket-Level implementiert. Die Konfiguration wird auf Namespace-Ebene festgelegt, wodurch die standardmäßige ADO-Einstellung für ADO während der Erstellung neuer Buckets festgelegt wird. Es kann bei der Erstellung neuer Buckets außer Kraft gesetzt werden. Das bedeutet, dass TSO für einige Buckets und nicht für andere konfiguriert werden kann.

8.2.5.1 Zugriff während des Ausfalls (ADO) nicht aktiviert

Standardmäßig ist ADO nicht aktiviert und starke Konsistenz wird beibehalten. Alle Client-API-Anfragen, bei denen die maßgeblichen Namespace-, Bucket- oder Objektdaten erforderlich, aber vorübergehend nicht verfügbar sind, schlagen fehl. Die Objektvorgänge „Lesen“, „Erstellen“, „Aktualisieren“ und „Löschen“ sowie die Auflistung von Buckets, die nicht im Besitz eines verfügbaren Standorts sind, schlagen fehl. Außerdem schlagen Vorgänge zum Erstellen und Bearbeiten von Bucket, Nutzer und Namespace fehl.

Wie zuvor erwähnt ist der erste Standorteigentümer von Bucket, Namespace und Objekt der Standort, an dem die Ressource ursprünglich erstellt wurde. Während eines TSO können bestimmte Vorgänge fehlschlagen, wenn auf den Standorteigentümer der Ressource nicht zugegriffen werden kann. Die folgenden wichtigen Vorgänge sind während eines vorübergehenden Standortausfalls zulässig oder nicht zulässig:

- Das Erstellen, Löschen und Aktualisieren von Buckets, Namespaces, Objektnutzern, Authentifizierungsanbietern, RGs und NFS-Nutzer- bzw. -Gruppenzuordnungen sind von keinem Standort aus zulässig.
- Das Auflisten von Buckets innerhalb eines Namespace ist zulässig, wenn der Namespace-Eigentümerstandort verfügbar ist.

HDFS-/NFS-fähige Buckets, die dem nicht zugänglichen Standort angehören, sind schreibgeschützt.

8.2.5.2 ADO aktiviert

In einem Bucket mit aktiviertem ADO ermöglicht der Speicherservice während eines TSO letztendliche Konsistenz. In diesem Szenario werden Lesevorgänge und optional Schreibvorgänge von einem sekundären Standort (Nicht-Eigentümer) angenommen und berücksichtigt. Darüber hinaus führt ein Schreibvorgang auf einem sekundären Standort während eines TSO dazu, dass der sekundäre Standort die Eigentumsrechte an dem Objekt übernimmt. Diese Funktion ermöglicht es jedem VDC, weiterhin Lese- und Schreibvorgänge für Objekte aus Buckets in einem gemeinsam genutzten Namespace auszuführen. Schließlich wird die neue Version des Objekts zur maßgeblichen Version des Objekts während des Abgleichs nach dem TSO, selbst wenn das Objekt auf dem Eigentümer-VDC von einer anderen Anwendung aktualisiert wird.

Obwohl viele Objektvorgänge während eines Netzwerkausfalls fortgesetzt werden, sind bestimmte Vorgänge nicht zulässig, z. B. das Erstellen neuer Buckets, Namespaces oder Nutzer. Wenn die Netzwerkverbindung zwischen 2 VDCs wiederhergestellt wird, erkennt der Heartbeat-Mechanismus automatisch die Konnektivität, stellt den Betrieb wieder her und stimmt Objekte von den beiden VDCs ab. Wenn dasselbe Objekt sowohl auf VDC A als auch auf VDC B aktualisiert wird, ist die Kopie auf dem Nicht-Eigentümer-VDC die maßgebliche Kopie. Wenn also ein Objekt, das Eigentum von VDC B ist, während der Synchronisation sowohl auf VDC A als auch auf VDC B aktualisiert wird, ist die Kopie auf VDC A die maßgebliche Kopie, die aufbewahrt wird. Die andere Kopie wird verworfen und kann zur Rückgewinnung von Speicherkapazität verwendet werden.

Wenn mehr als 2 VDCs Teil einer Replikationsgruppe sind und die Netzwerkverbindung zwischen einem VDC und den anderen beiden unterbrochen wird, werden Schreib-/Aktualisierungs-/Eigentümerschaftsvorgänge wie mit 2 VDCs fortgesetzt, der Prozess der Reaktion auf Leseanforderungen ist jedoch komplexer, wie unten beschrieben.

Wenn eine Anwendung ein Objekt anfordert, das einem VDC gehört, das nicht erreichbar ist, sendet ECS die Anforderung an das VDC mit der Sekundärkopie des Objekts. Die sekundäre Kopie wurde jedoch möglicherweise einem Datenkontraktionsvorgang unterzogen, bei dem es sich um einen XOR-Vorgang zwischen 2 unterschiedlichen Datasets handelt, aus dem ein neues Datenvolumen hervorgeht. Das VDC am sekundären Standort muss die Blöcke des Objekts aus dem ursprünglichen XOR-Vorgang abrufen und für diese Blöcke mit der Recovery-Kopie einen XOR-Vorgang ausführen. Dieses Verfahren gibt den Inhalt des Blocks zurück, der ursprünglich im ausgefallenen VDC gespeichert war. Die Blöcke aus dem wiederhergestellten Objekt können dann neu zusammengesetzt und zurückgegeben werden. Wenn die Blöcke rekonstruiert werden, werden sie auch zwischengespeichert, sodass das VDC schneller auf nachfolgende Anforderungen reagieren kann. Beachten Sie, dass die Rekonstruktion zeitaufwendig ist. Mehr VDCs in einer Replikationsgruppe implizieren mehr Blöcke, die von anderen VDCs abgerufen werden müssen, daher dauert die Rekonstruktion des Objekts länger.

Bei einem Notfall kann ein komplettes VDC nicht wiederherstellbar werden. ECS behandelt das nicht wiederherstellbare VDC als vorübergehenden Systemausfall am Standort. Wenn der Ausfall dauerhaft ist, muss der Systemadministrator ein dauerhaftes Failover für das VDC aus dem Verbund ausführen, um die Failover-Verarbeitung zu initiieren. Dadurch werden die Neusynchronisierung und der erneute Schutz der im ausgefallenen VDC gespeicherten Objekte initiiert. Die Recovery-Aufgaben werden als Hintergrundprozesse ausgeführt. Sie können den Recovery-Fortschritt im ECS-Portal überprüfen.

Eine zusätzliche Bucket-Option ist für *schreibgeschützte (read-only, RO)* ADO verfügbar. Dadurch wird sichergestellt, dass sich die Objekteigentumsrechte nie ändern, und die Möglichkeit von Konflikten, die durch Objektaktualisierungen auf den ausgefallenen und verfügbaren Standorten während eines vorübergehenden Standortausfalls verursacht werden, wird eliminiert. Der Nachteil von RO ADO besteht darin, dass während eines vorübergehenden Standortausfalls keine neuen Objekte erstellt werden können und keine vorhandenen Objekte im Bucket aktualisiert werden können, bis alle Standorte wieder online sind. Die RO-ADO-Option ist nur während der Erstellung von Buckets verfügbar und kann danach nicht mehr geändert werden. Diese Option ist standardmäßig deaktiviert.

Tabelle 8 Ausfalltoleranz bei mehreren Standorten

Fehlermodell	Toleranz
Geografisch replizierte Umgebung	Ausfall von bis zu 1 Standort

8.3 Ausfalltoleranz

ECS ist so konzipiert, dass es eine Reihe von Geräteausfallsituationen mit einer Reihe von Fehlerdomains toleriert. Die Ausfallbedingungen umfassen einen variierenden Bereich, einschließlich:

- Ausfall einer einzelnen Festplatte in einem einzigen Node
- Ausfall mehrerer Festplatten in einem einzigen Node
- Ausfall einer Festplatte in mehreren Nodes
- Ausfall mehrerer Festplatten in mehreren Nodes
- Ausfall eines einzelnen Node
- Ausfall mehrerer Nodes
- Verlust der Kommunikation mit einem replizierten VDC
- Verlust eines gesamten replizierten VDC

In einer Konfiguration mit 1 Standort, 2 Standorten oder einer geografisch replizierten Konfiguration hängen die Auswirkungen des Ausfalls von der Anzahl und Art der betroffenen Komponenten ab. Allerdings bietet ECS auf jedem Level Mechanismen zum Schutz vor Auswirkungen von Komponentenausfällen. Viele dieser Mechanismen wurden bereits in diesem Papier diskutiert, werden aber hier und in Abbildung 27 erneut behandelt, um zu zeigen, wie sie zur Lösung angewendet werden. Dazu gehören:

- Disk failure
  - Kein Speichern von EC-Segmenten oder Replikatkopien aus demselben Block auf derselben Festplatte
  - Prüfsummenberechnung für Schreib- und Lesevorgänge
  - Erneute Überprüfung von Prüfsumme durch Hintergrundkonsistenzprüfung

- Node failure
  - Gleichmäßige Verteilung von Segmenten oder Replikatkopien eines Blocks auf Nodes in einem VDC
  - ECS-Fabric sorgt dafür, dass die Services ausgeführt werden, und managt Ressourcen wie Festplatten und Netzwerke
  - Partitionsdatensätze und -tabellen, die durch ein Failover der Partitionseigentumsrechte von Node zu Node geschützt sind
- Ausfall des Racks innerhalb eines VDC
  - Gleichmäßige Verteilung von Segmenten der Replikatkopien eines Blocks auf Racks in einem VDC
  - Eine fabric-registry-Instanz wird in jedem Rack ausgeführt und kann auf jedem anderen Node im selben Rack neu gestartet werden, wenn der Node ausfällt

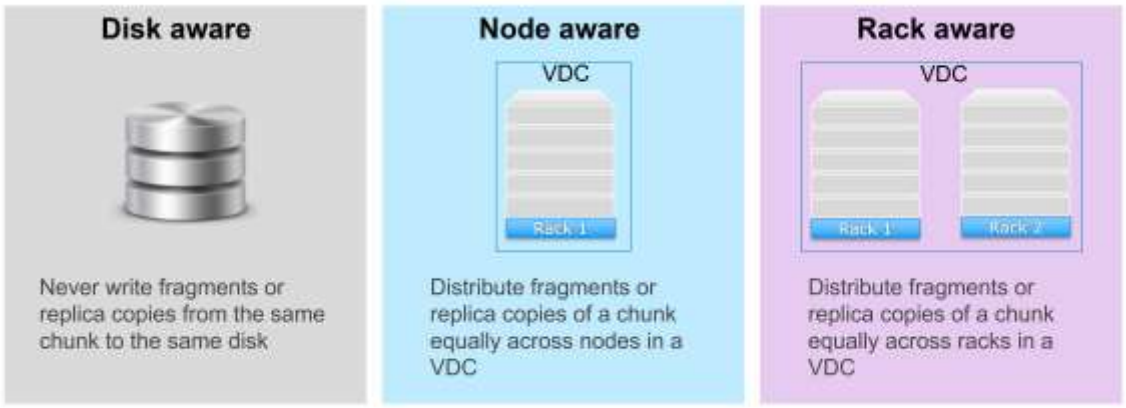

Abbildung 27 Schutzmechanismen auf Festplatten-, Node- und Racklevel

Das folgende Diagramm beschreibt den Typ und die Anzahl der Komponentenausfälle, gegen die jedes EG-Schema gemäß der Basisrackkonfiguration schützt. Tabelle 9 zeigt, wie wichtig es ist, die Auswirkungen von Schutzfehlerdomains auf die allgemeine Daten- und Serviceverfügbarkeit zu berücksichtigen, und gibt die Anzahl der Nodes an, die für jedes EC-Schema erforderlich sind.

Tabelle 9 Erasure-Coding-Schutz für Fehlerdomains

EC-Schema	Anzahl der Nodes im VDC	Anzahl der Blockfragmente pro Node	EC Data Protection vor …
12 + 4 Standard	5 oder weniger	4	• Ausfall von bis zu 4 Festplatten oder • Ausfall von 1 Node
	6 oder 7	3	• Ausfall von bis zu 4 Festplatten oder • Ausfall von 1 Node und 1 Festplatte aus einem 2. Node
	8 oder mehr	2	• Ausfall von bis zu 4 Festplatten oder • Ausfall von 2 Nodes oder • Ausfall von 1 Node und 2 Festplatten
	16 oder mehr	1	• Ausfall von 4 Nodes oder • Ausfall von 3 Nodes und Festplatten von 1 zusätzlichen Node oder • Ausfall von 2 Nodes und Festplatten von bis zu 2 anderen Nodes oder • Ausfall von 1 Node und Festplatten von bis zu 3 anderen Nodes oder • Ausfall von 4 Festplatten von 4 unterschiedlichen Nodes
10+2 Cold Storage	11 oder weniger	2	• Ausfall von bis zu 2 Festplatten oder • Ausfall von 1 Node
	12 oder mehr	1	• Verlust einer beliebigen Anzahl von Festplatten von 2 verschiedenen Nodes oder • Ausfall von 2 Nodes

8.4 Automatisierung des Festplattenaustauschs

Ab ECS 3.5 können Kunden ausgefallene Festplatten mithilfe eines intuitiven Workflows des ECS-Portals (Web UI) durch Dell EMC Services ersetzen. Die Funktion bietet Folgendes:

- Do-it-yourself-Lösung zur Behebung von Ausfällen
- Schnellere Problemlösung
- Betriebliche Flexibilität und Gesamtbetriebskosteneinsparungen

Die Wartungsseite im ECS-Portal bietet Administratortransparenz für alle Festplatten in jedem Node. Wenn ein Laufwerk ausfällt, initiiert das System automatisch die Wiederherstellung. Alle Arten von Ressourcen auf dem Laufwerk werden wiederhergestellt. Wenn das Laufwerk bereit ist, aus dem Node entfernt zu werden, zeigt das ECS-Portal die Schaltfläche „Replace“ an, wie in Abbildung: 28 gezeigt.

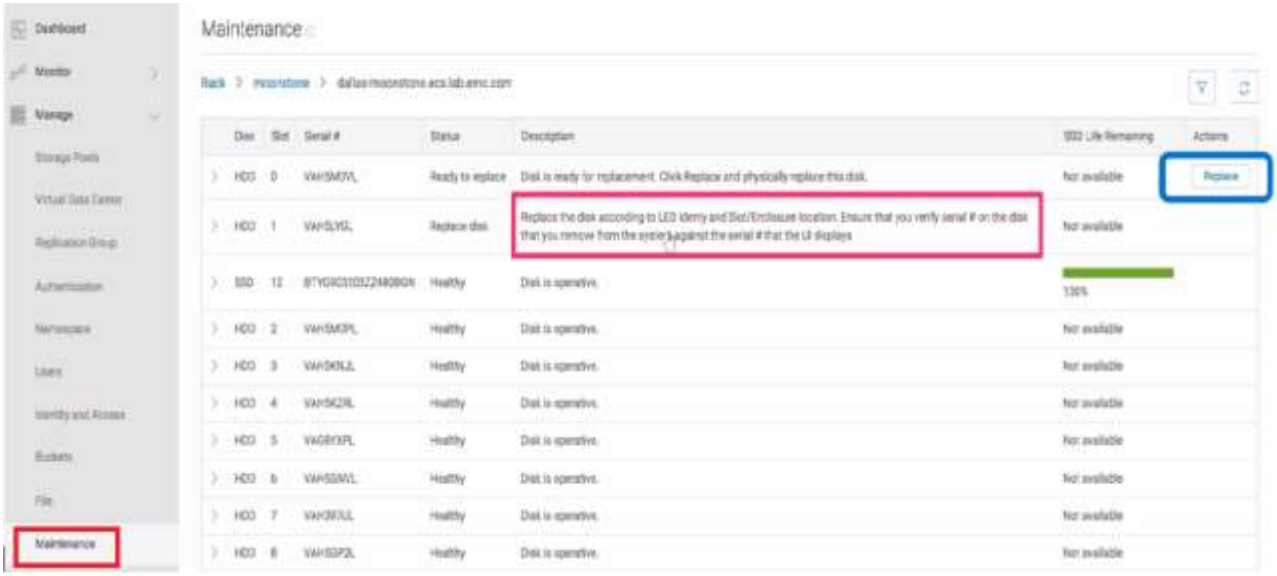

Abbildung: 28 Automatisierung des Festplattenaustauschs

Hinweis: Es kann jeweils nur ein Laufwerk ausgetauscht werden. Dies dient dazu, den Austausch des falschen Laufwerks zu vermeiden.

8.5 Tech Refresh

Tech Refresh ist ein von Dell EMC Professional Services angewiesenes Engagement, das ab ECS 3.5 verfügbar ist, um ältere Hardware-Nodes mithilfe der integrierten Softwarefunktion unterbrechungsfrei aus ECS-Clustern zu entfernen. Es handelt sich um einen effizienten und ressourcenintensiven Vorgang, der präzise gedrosselt werden kann. Diese Funktion reduziert den Overhead, der zuvor mit der Außerbetriebnahme von ECS-Hardware verbunden war.

Tech Refresh umfasst drei Teile:

- **Node-Erweiterung:** Hinzufügen von Gen3-Nodes zu einem vorhandenen Cluster
- **Ressourcenmigration:** Verschieben aller Ressourcen von vorhandenen Nodes auf Gen3-Nodes
- **Node-Evakuierung:** Bereinigen alter Nodes und Entfernen aus dem Cluster

Professional Services sollten bei der Tech-Refresh-Wartung einbezogen werden. Weitere Informationen zu Tech Refresh finden Sie im neuesten *ECS Tech Refresh Guide.*

9 Speicherschutzoverhead

Jedes VDC-Mitglied in einer RG ist für den eigenen EC-Schutz von Daten auf lokalem Level verantwortlich. Das heißt, Daten werden repliziert, nicht aber zugehörige Coding-Segmente. Obwohl EC mehr Speichereffizienz als andere Schutzarten aufweist, z. B. eine vollständige Kopie von Laufwerken, wird ein inhärenter Speicherverbrauchsoverhead auf lokalem Level verursacht. Wenn jedoch Sekundärkopien außerhalb des Standorts repliziert werden und alle Standorte Zugriff auf Daten haben müssen, wenn ein einzelner Standort nicht verfügbar ist, wird der Speicherbedarf höher als bei der Verwendung herkömmlicher Schutzmethoden mit Datenkopiervorgängen von Standort zu Standort. Dies gilt insbesondere, wenn eindeutige Daten über 3 oder mehr Standorte verteilt werden.

ECS bietet einen Mechanismus, mit dem die Effizienz des Speicherschutzoverheads steigen kann, wenn 3 oder mehr Standorte in einem Verbund sind. In einer replizierten Umgebung mit 2 VDCs repliziert ECS Blöcke vom primären oder Eigentümer-VDC an einem Remotestandort, um hohe Verfügbarkeit und Ausfallsicherheit bereitzustellen. Es gibt keine Möglichkeit, die 100 % Kosten des Schutzoverheads einer vollständigen Kopie der Daten in einer ECS-Verbundbereitstellung mit 2 Standorten zu umgehen.

Nehmen wir als Beispiel 3 VDCs in einer Umgebung mit mehreren Standorten an: VDC1, VDC2 und VDC3. Auf jedem VDC werden eindeutige Daten von jedem der anderen VDCs repliziert. VDC2 und VDC3 können zum Schutz eine Kopie ihrer Daten an VDC1 senden. VDC1 hätte daher seine eigenen Originaldaten sowie replizierte Daten von VDC2 und VDC3. Das bedeutet, dass VDC1 die 3-fache Menge der Daten speichert, die an einem eigenen Standort geschrieben werden.

In diesem Fall kann ECS einen XOR-Vorgang der Daten von VDC2 und VDC3, die lokal in VDC1 gespeichert sind, durchführen. Diese mathematische Operation vergleicht die gleiche Menge eindeutiger Daten und Blöcke und gibt ein Ergebnis in einem neuen Block aus, der genügend Merkmale der beiden ursprünglichen Datenblöcke enthält, um die Wiederherstellung jedes der beiden ursprünglichen Sätze zu ermöglichen. Wo früher also 3 einzigartige Sätze von Datenblöcken in VDC1 gespeichert waren, die das 3-Fache der verfügbaren Kapazität verbrauchten, gibt es jetzt nur noch 2 – den ursprünglichen lokalen Datensatz und die durch XOR reduzierten Schutzkopien.

Wenn VDC3 in diesem Szenario nicht verfügbar ist, kann ECS VDC3-Datenblöcke mithilfe von Blockkopien, die von VDC2 abgerufen werden, und den (C1 ⊕ C2-)Daten aus VDC3, die lokal in VDC1 gespeichert sind, rekonstruieren. Dieses Prinzip gilt für alle 3 an der RG beteiligten Standorte und hängt von den 3 VDCs mit eindeutigen Datenvolumen ab. Abbildung 29 zeigt eine XOR-Berechnung mit 2 Standorten, die auf einen 3. Standort repliziert werden.

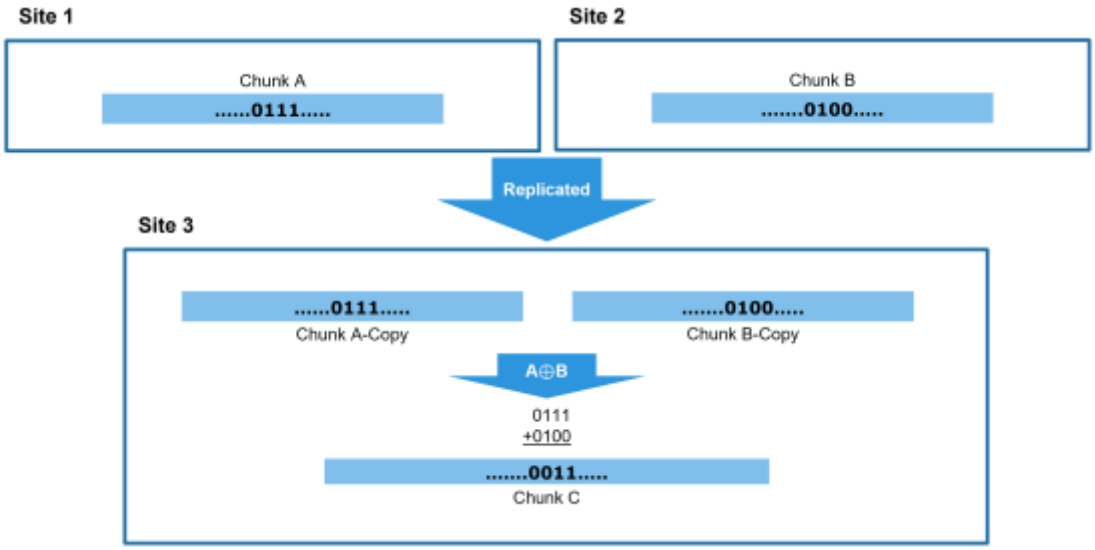

Abbildung 29 Effizienz von XOR Data Protection

Wenn für geschäftliche Service Level Agreements eine optimale Lesezugriffsgeschwindigkeit selbst bei einem vollständigen Standortausfall erforderlich ist, zwingt die Einstellung „Replicate to All Sites“ ECS, vollständige Kopien replizierter Daten zu erstellen, die an allen Standorten gespeichert werden. Dies führt erwartungsgemäß zu einem Anstieg der Speicherkosten im Verhältnis zur Anzahl der an der RG teilnehmenden VDCs. Daher würde eine Konfiguration mit 3 Standorten zu einem 3-fachen Speicherschutzoverhead führen. Die Einstellung „Replicate to All Sites“ ist während der RG-Erstellung verfügbar und kann nicht hin- und hergeschaltet werden.

Mit zunehmender Anzahl von Verbundstandorten wird die XOR-Optimierung dank Replikation effizienter bei der Reduktion von Speicherschutzoverhead. Tabelle 10 enthält Informationen zum Speicherschutzoverhead basierend auf der Anzahl der Standorte für das normale EC von 12+4 und EC für Archiv für inaktive Daten von 10+2. Es wird gezeigt, wie ECS umso speichereffizienter werden kann, je mehr Standorte miteinander verbunden sind.

Hinweis: Um den Overhead durch replizierte Daten bei 3 bis 8 Standorten zu reduzieren, müssen eindeutige Daten relativ gleichmäßig an jedem Standort geschrieben werden. Durch das gleichmäßig verteilte Schreiben von Daten auf Standorte verfügt jeder Standort über eine ähnliche Anzahl von Replikatblöcken. Eine ähnliche Anzahl von Replikatblöcken an jedem Standort führt zu einer ähnlichen Anzahl von XOR-Vorgängen, die an jedem Standort ausgeführt werden können. Die maximale Speichereffizienz an mehreren Standorten wird erreicht, indem die maximale Anzahl von Replikatblöcken, die mithilfe von XOR gespeichert werden, reduziert wird.

Tabelle 10 Speicherschutzoverhead

Anzahl der Standorte in der RG	12+4 EC	10+2 EC
1	1,33	1,2
2	2,67	2,4
3	2,00	1,8
4	1,77	1,6
5	1,67	1,5
6	1,60	1,44
7	1,55	1,40
8 (max. Anzahl der Standorte in RG)	1,52	1,37

10 Fazit

Unternehmen sind mit immer höheren Daten- und Speicherkosten konfrontiert, insbesondere im Public-Cloud-Bereich. Die Scale-out- und geografisch verteilte Architektur von ECS bietet eine Vor-Ort-Cloud-Plattform, die auf Exabyte an Daten skaliert werden kann und dabei deutlich geringere *Gesamtbetriebskosten* als der Public-Cloud-Speicher verursacht. ECS ist eine hervorragende Lösung aufgrund seiner Vielseitigkeit, Hyper-Skalierbarkeit, leistungsstarken Funktionen und der Verwendung handelsüblicher Hardware.

A Technischer Support und Ressourcen

Dell.com/support konzentriert sich auf die Erfüllung der Kundenanforderungen mit bewährtem Service und Support.

Technische Dokumentation und Videos zum Thema Speicher liefern das Know-how, das zum Kundenerfolg mit Dell EMC Speicherplattformen beiträgt.