White Paper

EMC ISILON ONEFS: TECHNISCHER ÜBERBLICK

Zusammenfassung

Dieses White Paper enthält eine ausführliche Beschreibung der wichtigsten Komponenten des EMC Isilon OneFS-Betriebs- und Dateisystems. Es umfasst Informationen zu Software, Hardware, verteilter Architektur und den verschiedenen Datensicherheitsmechanismen von OneFS, einer außerordentlich flexiblen Speicherlösung der Enterprise-Klasse mit einfachem Management und im Hinblick auf Kapazität und Performance extrem hoher Skalierbarkeit.

Juli 2014

EMC²

Art.-Nr.: H10719.4

Inhaltsverzeichnis

Einführung

Angesichts der Herausforderungen bei herkömmlichen Speicherarchitekturen und dem rasanten Anstieg von File-basierten Daten begannen die Gründer von Isilon Systems im Jahr 2000 mit der Entwicklung einer revolutionären neuen Speicherarchitektur: dem OneFS®-Betriebssystem. Der wesentliche Unterschied beim EMC® Isilon®-Speicher liegt in der Verwendung intelligenter Software, um Daten über eine Unmenge an handelsüblicher Hardware zu skalieren und so ein explosionsartiges Wachstum bei Performance und Kapazität zu ermöglichen. Die drei Ebenen des herkömmlichen Speichermodells (Dateisystem, Volume Manager und Datensicherheit) wurden im Laufe der Zeit auf die Anforderungen kleinerer Speicherarchitekturen hin weiterentwickelt, sind heute aber hochgradig komplex und nicht gut an Systeme mit einer Skalierung im Petabytebereich angepasst. OneFS ersetzt alle diese Ebenen und bietet ein vereinheitlichtes Clusterdateisystem mit integrierter skalierbarer Datensicherheit, für das kein Volume-Management erforderlich ist. OneFS ist ein grundlegender Baustein für Scale-out-Infrastrukturen und ermöglicht eine umfassende Skalierbarkeit und enorme Effizienz.

Entscheidend ist, dass OneFS nicht nur bei Computern, sondern auch im Hinblick auf den menschlichen Faktor skalierbar ist. So können große Systeme von einem Bruchteil der für herkömmliche Speichersysteme erforderlichen Mitarbeiter gemanagt werden. OneFS sorgt für weniger Komplexität und umfasst automatische Fehlerkorrektur und automatische Managementfunktionen, durch die der Arbeitsaufwand für das Speichermanagement drastisch reduziert wird. OneFS bietet außerdem Parallelität auf einer sehr tiefgehenden Betriebssystemebene, sodass nahezu jeder wichtige Systemservice auf mehrere Hardwareeinheiten verteilt wird. Dadurch kann OneFS bei einer Erweiterung der Infrastruktur in nahezu jede Richtung skaliert werden. Was heute funktioniert, wird also auch bei einem zukünftigen Anwachsen der Datasets weiterhin funktionsfähig sein.

OneFS ist ein vollständig symmetrisches Dateisystem ohne Single-Point-of-Failure, wobei Clustering nicht nur für die Skalierung von Performance und Kapazität eingesetzt wird, sondern auch für das n:n-Failover und für mehrere Redundanzlevel, die weit über die Möglichkeiten von RAID hinausgehen. Der Trend zu Festplattensubsystemen hat sich nach und nach positiv auf die Performance ausgewirkt, wobei zugleich die Speicherdichten zunahmen. OneFS reagiert auf diese Gegebenheiten, indem die Menge an Redundanz sowie die Geschwindigkeit der Reparaturen bei Ausfällen skaliert werden. Somit kann OneFS auf Systeme von mehreren Petabyte skaliert werden und ist gleichzeitig zuverlässiger als kleine herkömmliche Speichersysteme.

Die NAS-Scale-out-Hardware von Isilon stellt die Appliance für die Ausführung von OneFS dar. Dabei handelt es sich um Best-of-Breed-, aber gleichzeitig handelsübliche Hardwarekomponenten, was bedeutet, dass die Isilon-Hardware von den fortlaufenden Verbesserungen bei den Kosten- und Effizienzkurven handelsüblicher Hardware profitiert. Mit OneFS kann Hardware dem Cluster jederzeit hinzugefügt oder daraus entfernt werden, sodass eine Abstraktion von Daten und Anwendungen von der Hardware besteht. Die Daten haben uneingeschränkte Langlebigkeit und sind vor den Unbeständigkeiten zukünftiger Hardwaregenerationen geschützt. Kosten und Probleme im Zusammenhang mit Datenmigration und Hardwareaktualisierung werden vermieden.

OneFS ist die ideale Lösung für File-basierte und unstrukturierte „Big Data"-Anwendungen in Unternehmensumgebungen, darunter umfangreiche Stammverzeichnisse, Dateifreigaben, Archive, Virtualisierung und Geschäftsanalysen. Daher ist OneFS heute in vielen datenintensiven Branchen weit verbreitet, zum Beispiel Energie, Finanzdienstleistungen, Internet- und Hostingservices, Business Intelligence, Engineering, Fertigung, Medien und Unterhaltung, Bioinformatik, wissenschaftliche Forschung und andere leistungsfähige Computing-Umgebungen.

Überblick über EMC Isilon OneFS

OneFS führt die drei Schichten herkömmlicher Speicherarchitekturen – Dateisystem, Volume Manager und Datensicherheit – in einer einzigen Softwareschicht zusammen und schafft auf diese Weise ein einziges intelligentes, verteiltes Dateisystem, das auf einem Isilon-Speichercluster ausgeführt wird.

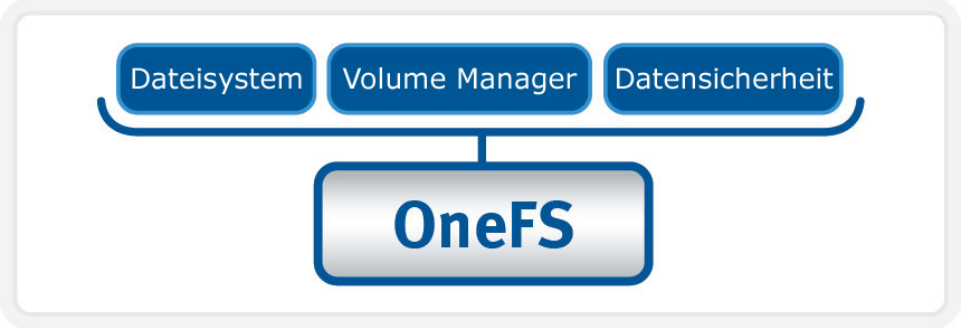

Abbildung 1: OneFS führt Dateisystem, Volume Manager und Datensicherheit in einem einzigen intelligenten, verteilten System zusammen.

Dies ist die Kerninnovation, mit der Unternehmen Scale-out-NAS in ihren heutigen Umgebungen direkt erfolgreich nutzen können. Die Software entspricht den Grundprinzipien von Scale-out: intelligente Software, handelsübliche Hardware und verteilte Architektur. OneFS ist nicht nur das Betriebssystem, sondern auch das zugrunde liegende Dateisystem, über das Daten im Scale-out-NAS-Cluster von Isilon gesteuert und gespeichert werden.

Isilon-Nodes

OneFS arbeitet ausschließlich mit den Scale-out-NAS-Nodes von Isilon, die als „Cluster" bezeichnet werden. Ein einziges Isilon-Cluster besteht aus mehreren Nodes, die als rackmontierbare Enterprise-Appliances aufgebaut werden und Folgendes umfassen: Arbeitsspeicher, CPU, Netzwerk, nicht flüchtiger Random Access Memory (NVRAM), InfiniBand-Verbindungen mit niedriger Latenz, Festplattencontroller und Speichermedien. Jeder Node im verteilten Cluster hat somit nicht nur Rechner- oder Verarbeitungsfunktionen, sondern auch Speicherfunktionen oder Kapazitätsmöglichkeiten.

Ein Isilon-Cluster kann lediglich aus drei Nodes bestehen und derzeit auf 144 Nodes skaliert werden (unterliegt dem größten 144-Port-InfiniBand-Switch, für den Isilon qualifiziert ist). Es gibt viele verschiedene Arten von Nodes, die alle in ein einziges Cluster integriert werden können, in dem verschiedene Nodes unterschiedliche Kapazitätsraten für Durchsatz oder Eingabe-/Ausgabevorgänge pro Sekunde (IOPS) bereitstellen.

Bei OneFS gib es keine integrierten Einschränkungen für die Anzahl der Nodes, die in ein einziges System einbezogen werden können. Jeder dem Cluster hinzugefügte Node erhöht den aggregierten Speicher, den Cache, die CPU und die Netzwerkkapazität. OneFS nutzt alle Hardwarebausteine auf eine Weise, dass das Endergebnis größer als die Summe seiner Teile ist. Der RAM wird in einem einzigen kohärenten Cache gruppiert, sodass für die I/O-Vorgänge in einem beliebigen Teil des Clusters Daten genutzt werden können, die an beliebigen Speicherorten zwischengespeichert sind. NVRAM wird zusammengefasst, um Schreibvorgänge mit hohem Durchsatz zu ermöglichen, die auch bei einem Stromausfall sicher sind. Spindeln und CPU werden kombiniert, um Durchsatz, Kapazität und IOPS bei wachsenden Clustern zu steigern, um auf eine oder mehrere Dateien zuzugreifen. Die Speicherkapazität eines Clusters beträgt mindestens 18 Terabyte (TB) und maximal 20 Petabyte (PB). Die maximale Kapazität erhöht sich mit steigender Dichte der Festplattenlaufwerke.

Die heute verfügbaren Isilon-Nodes sind entsprechend ihrer Funktion in eine Reihe von Klassen unterteilt:

- S-Serie: IOPS-intensive Anwendungen
- X-Serie: Viele gleichzeitige Workflows mit hohem Durchsatz
- NL-Serie: Nahezu Primärspeichern entsprechende Zugriffszeiten zu vergleichbaren Kosten wie beim Einsatz von Bandsystemen
- Performance-Accelerator: Unabhängige Skalierung für maximale Performance
- Backup-Accelerator: Skalierbare Hochgeschwindigkeitslösung für das Backup und die Wiederherstellung von Daten

Netzwerk

Es gibt zwei Typen von Netzwerken, die mit einem Cluster verbunden sind: interne und externe Netzwerke.

Back-end-Netzwerk

Die gesamte Kommunikation zwischen den Nodes in einem Cluster wird mithilfe eines proprietären Unicast-Protokolls (Node zu Node) durchgeführt. Bei der Kommunikation wird ein extrem schnelles IB-Netzwerk (InfiniBand) mit niedriger Latenz verwendet. Dieses Back-end-Netzwerk ist im Hinblick auf hohe Verfügbarkeit mit redundanten Switchen konfiguriert und fungiert als Grundlage für das Cluster. So kann jeder Node einen Beitrag im Cluster leisten und die Node-zu-Node-Kommunikation wird in einem privaten, hochleistungsfähigen Netzwerk mit niedriger Latenz isoliert. Für dieses Back-end-Netzwerk wird zur Kommunikation zwischen den einzelnen Nodes IP (Internet Protocol) über IB verwendet.

Front-end-Netzwerk

Clients stellen über Ethernetverbindungen (1 GbE oder 10 GbE), die auf allen Nodes verfügbar sind, eine Verbindung mit dem Cluster her. Da jeder Node eigene Ethernetports bietet, kann die im Cluster verfügbare Netzwerkbandbreite je nach Performance und Kapazität linear skaliert werden. Das Isilon-Cluster unterstützt standardmäßige Netzwerkkommunikationsprotokolle zu einem Kundennetzwerk, einschließlich NFS, CIFS, HTTP, FTP und HDFS.

Ansicht des gesamten Clusters

Das gesamte Cluster mit Hardware, Software und Netzwerken wird in der folgenden Ansicht zusammengestellt:

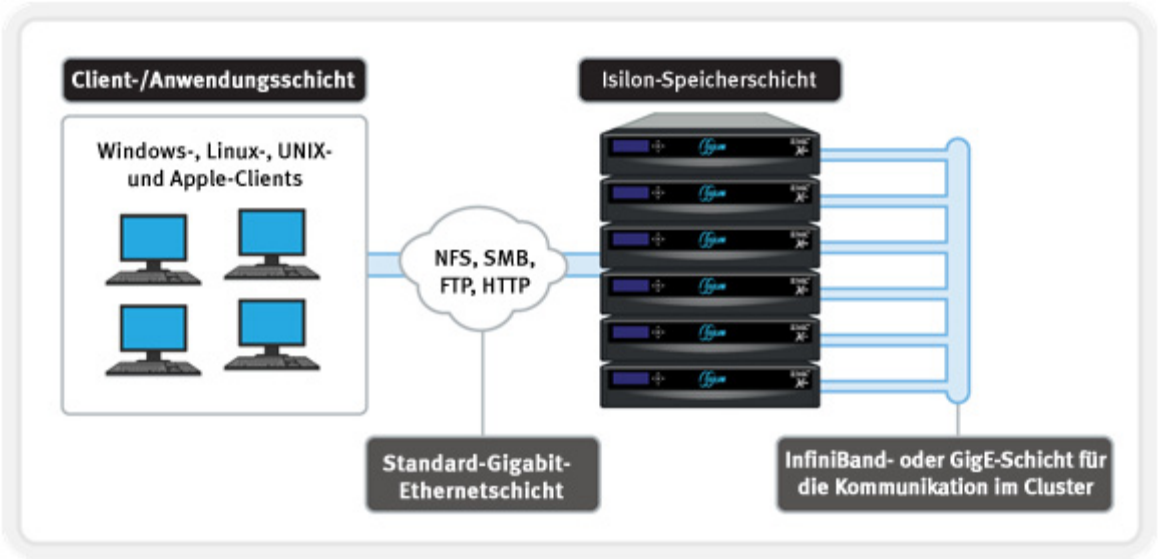

Abbildung 2: Alle zusammenarbeitenden Komponenten von OneFS

Abbildung 2 stellt die gesamte Architektur dar: Software, Hardware und Netzwerk arbeiten in Ihrer Umgebung mit Servern zusammen, sodass ein einziges, vollständig verteiltes Dateisystem entsteht, das je nach Workloads und Änderungen bei Kapazität oder Durchsatz in einer Scale-out-Umgebung dynamisch skaliert werden kann.

Überblick über die OneFS-Software

Betriebssystem

OneFS ist auf einem BSD-basierten UNIX-Betriebssystem aufgebaut. Es bietet nativen Support für die Semantik von Linux bzw. UNIX sowie Windows, einschließlich Hardlinks, Löschvorgang beim Schließen, atomares Umbenennen, ACLs und erweiterte Attribute. Als grundlegendes Betriebssystem wird BSD verwendet, da es sich um ein ausgereiftes und bewährtes Betriebssystem handelt, das von den Innovationen der Open-Source-Community profitiert.

Clientservices

Die von Clients zur Interaktion mit OneFS verwendeten Front-end-Protokolle werden als Clientservices bezeichnet. Im Abschnitt „Unterstützte Protokolle" finden Sie eine detaillierte Liste der unterstützten Protokolle. Um zu verstehen, wie OneFS mit Clients kommuniziert, wird das I/O-Subsystem in zwei Hälften aufgeteilt: die obere Hälfte bzw. der Initiator und die untere Hälfte bzw. der Teilnehmer. Jeder Node im Cluster ist ein Teilnehmer für einen bestimmten I/O-Vorgang. Der Node, mit dem der Client verbunden ist, ist der Initiator. Dieser Node fungiert als „Kapitän" für den gesamten I/O-Vorgang. Die Lese- und Schreibvorgänge werden weiter unten genauer beschrieben

Clustervorgänge

In einer Clusterarchitektur gibt es Clusterjobs, die für die Integrität und Verwaltung des Clusters selbst verantwortlich sind. Diese Jobs werden durch die OneFS-Job-Engine gesteuert. Die Job-Engine wird im gesamten Cluster ausgeführt und dient zum Aufteilen und Durchführen umfangreicher Speichermanagement- und Sicherungsaufgaben. Hierbei werden die Aufgaben in kleinere Arbeitselemente aufgeteilt und diese Teile des Gesamtjobs dann mehreren Worker Threads auf den einzelnen Nodes zugewiesen. Der Fortschritt wird während der Jobausführung überwacht und gemeldet. Nach Abschluss oder Abbruch wird ein detaillierter Bericht mit Statusangabe bereitgestellt.

Die Job-Engine enthält ein umfassendes Kontrollpunktsystem, sodass Jobs nicht nur gestartet und beendet, sondern auch angehalten und fortgesetzt werden können. Das Framework für die Job-Engine umfasst außerdem ein adaptives Auswirkungsmanagementsystem.

Die Job-Engine führt Jobs im gesamten Cluster in der Regel im Hintergrund aus, wobei freie oder speziell reservierte Kapazitäten und Ressourcen eingesetzt werden. Die Jobs selbst lassen sich in drei Hauptklassen einteilen:

- **Dateisystemverwaltung**

  Bei diesen Jobs wird eine Dateisystemverwaltung im Hintergrund durchgeführt. Sie benötigen in der Regel Zugriff auf alle Nodes. Diese Jobs müssen in den Standardkonfigurationen und oft unter heruntergestuften Clusterbedingungen ausgeführt werden. Beispiele: Dateisystemsicherung und Wiederherstellung von Laufwerken.

- **Funktionssupport**

  Die Supportjobs für Funktionen umfassen gewisse erweiterte Funktionen für das Speichermanagement und werden in der Regel nur dann ausgeführt, wenn die Funktion konfiguriert wurde. Beispiele hierfür sind Deduplizierung und Virenüberprüfungen.

- **Benutzeraktionen**

  Diese Jobs werden direkt vom Speicheradministrator ausgeführt, um gewisse Ziele beim Datenmanagement zu erfüllen. Beispiele hierfür sind parallele Strukturlöschvorgänge und die Berechtigungsverwaltung.

Die folgende Tabelle enthält eine umfassende Liste der verfügbaren Job-Engine-Jobs, Informationen zu den durchgeführten Vorgängen sowie den jeweiligen Methoden für den Dateisystemzugriff:

Jobname	Jobbeschreibung	Zugriffsmethode
AutoBalance	Ausgleich von freiem Speicherplatz im Cluster	Laufwerks- und LIN-Scans
AutoBalanceLin	Ausgleich von freiem Speicherplatz im Cluster	LIN-Scan
AVScan	Von ICAP-Servern ausgeführte Jobs zur Virenüberprüfung	Treewalk
Collect	Wiedergewinnung von Speicherplatz, der aufgrund eines nicht verfügbaren Node oder Laufwerks nicht freigegeben werden konnte, da diese verschiedene Ausfallbedingungen aufweisen	Laufwerks- und LIN-Scans
Dedupe	Deduplizierung identischer Blöcke im Dateisystem	Treewalk
DedupeAssessment	Probebewertung der Vorteile der Deduplizierung	Treewalk
DomainMark	Verknüpfung eines Pfads und seiner Inhalte mit einer Domain	Treewalk
FlexProtect	Wiederherstellung und erneute Sicherung des Dateisystems nach einem Ausfallszenario	Laufwerks- und LIN-Scans
FlexProtectLin	Erneute Sicherung des Dateisystems	LIN-Scan
FSAnalyze	Erfassen von Dateisystem-Analysedaten, die mit der Isilon InsightIQ™-Software verwendet werden	LIN-Scan
IntegrityScan	Durchführen einer Onlineüberprüfung und -korrektur von Dateisysteminkonsistenzen	LIN-Scan
MediaScan	Scannen von Laufwerken auf Fehler auf Medienlevel	Laufwerks- und LIN-Scans
MultiScan	Gleichzeitige Ausführung von Collect- und AutoBalance-Jobs	LIN-Scan
PermissionRepair	Korrektur von Berechtigungen für Dateien und Verzeichnisse	Treewalk
QuotaScan	Aktualisieren von Quota für auf einem vorhandenen Verzeichnispfad erstellte Domains	Treewalk
SetProtectPlus	Anwenden der standardmäßigen Datei-Policy. Dieser Job ist deaktiviert, wenn SmartPools auf dem Cluster aktiviert ist.	LIN-Scan
ShadowStoreDelete	Freigeben von Speicherplatz für einen Schattenspeicher	LIN-Scan

Jobname	Jobbeschreibung	Zugriffsmethode
SmartPools	Job zum Ausführen und Verschieben von Daten zwischen den Tiers von Nodes innerhalb desselben Clusters	LIN-Scan
SnapRevert	Umkehren eines gesamten Snapshot	
SnapshotDelete	Schaffen von freiem Speicherplatz durch gelöschte Snapshots	LIN-Scan
TreeDelete	Löschen eines Pfads im Dateisystem direkt aus dem Cluster	Treewalk

Abbildung 3: Jobbeschreibungen für die OneFS-Job-Engine

Die Verwaltungsjobs für das Dateisystem werden zwar nach einem festen Zeitplan oder als Reaktion auf ein bestimmtes Dateisystemereignis standardmäßig ausgeführt, aber für jeden Job-Engine-Job können sowohl das Prioritätslevel (im Verhältnis zu anderen Jobs) als auch die Auswirkungs-Policy konfiguriert werden.

Eine Auswirkungs-Policy umfasst ein oder mehrere Auswirkungsintervalle, die Zeitblöcke in einer bestimmten Woche definieren. Jedes Auswirkungsintervall kann auf ein vordefiniertes Auswirkungslevel konfiguriert werden, durch das die Menge der für einen bestimmten Clustervorgang zu verwendenden Clusterressourcen angegeben wird. Für die Job-Engine verfügbare Auswirkungslevel:

- **Ausgesetzt**
- **Niedrig**
- **Mittel**
- **Hoch**

Mit diesem Grad an Granularität können die Auswirkungsintervalle und -level für einzelne Jobs konfiguriert werden, um reibungslosere Clustervorgänge zu gewährleisten. Die resultierenden Auswirkungs-Policies geben vor, wann ein Job ausgeführt wird und welche Ressourcen dafür genutzt werden können.

Zusätzlich werden die Jobs der Job-Engine auf einer Skala von eins bis zehn priorisiert, wobei ein niedrigerer Wert eine höhere Priorität bedeutet. Dieses Konzept ähnelt dem der UNIX Scheduling Utility „nice".

Ab OneFS 7.1 können mit der Job-Engine bis zu drei Jobs gleichzeitig ausgeführt werden. Diese gleichzeitige Jobausführung beruht auf den folgenden Kriterien:

- Jobpriorität
- Ausschlusssätze: Jobs, die nicht gemeinsam ausgeführt werden können (d. h. FlexProtect und AutoBalance)
- Clusterintegrität: Die meisten Jobs können nicht ausgeführt werden, wenn sich das Cluster in einem heruntergestuften Status befindet.

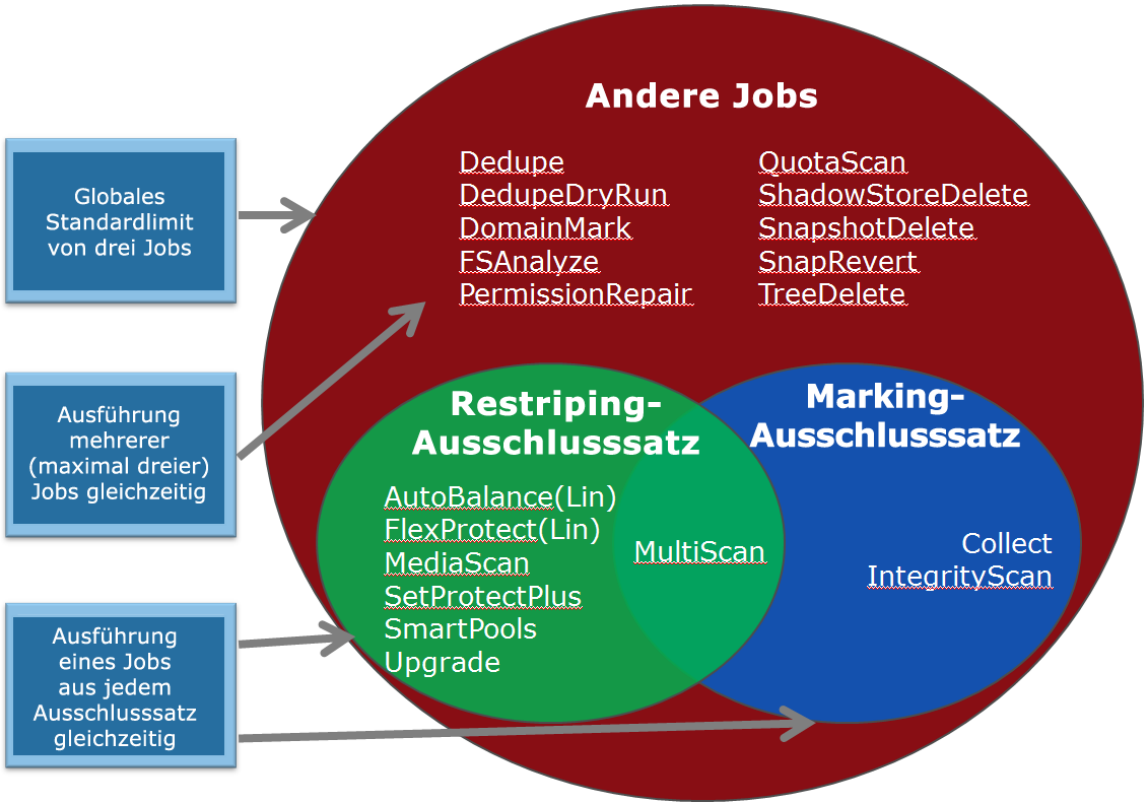

Abbildung 4: OneFS-Job-Engine – Ausschlusssätze

Struktur des Dateisystems

Das OneFS-Dateisystem basiert auf dem UNIX-Dateisystem (UFS). Es handelt sich daher um ein äußerst schnelles, verteiltes Dateisystem. Jedes Cluster erstellt einen einzigen Namespace und ein Dateisystem. Somit wird das Dateisystem über alle Nodes im Cluster verteilt und ist für Clients über eine Verbindung zu einem beliebigen Node im Cluster zugänglich. Es gibt keine Partitionierung und es ist keine Volume-Erstellung erforderlich. Statt den Zugriff auf freien Speicherplatz und nicht autorisierte Dateien auf Ebene des physischen Volume zu beschränken, bietet OneFS dieselbe Funktion in der Software über Freigabe- und Dateiberechtigungen und über den Isilon SmartQuotas™-Service, der ein Quotamanagement auf Verzeichnisebene bereitstellt.

Da alle Informationen über das interne Netzwerk von den Nodes gemeinsam verwendet werden, können Daten auf einen beliebigen Node geschrieben bzw. von einem beliebigen Node gelesen werden. Dies führt zu einer optimierten Performance, wenn mehrere Benutzer zur gleichen Zeit denselben Datensatz lesen bzw. darin schreiben.

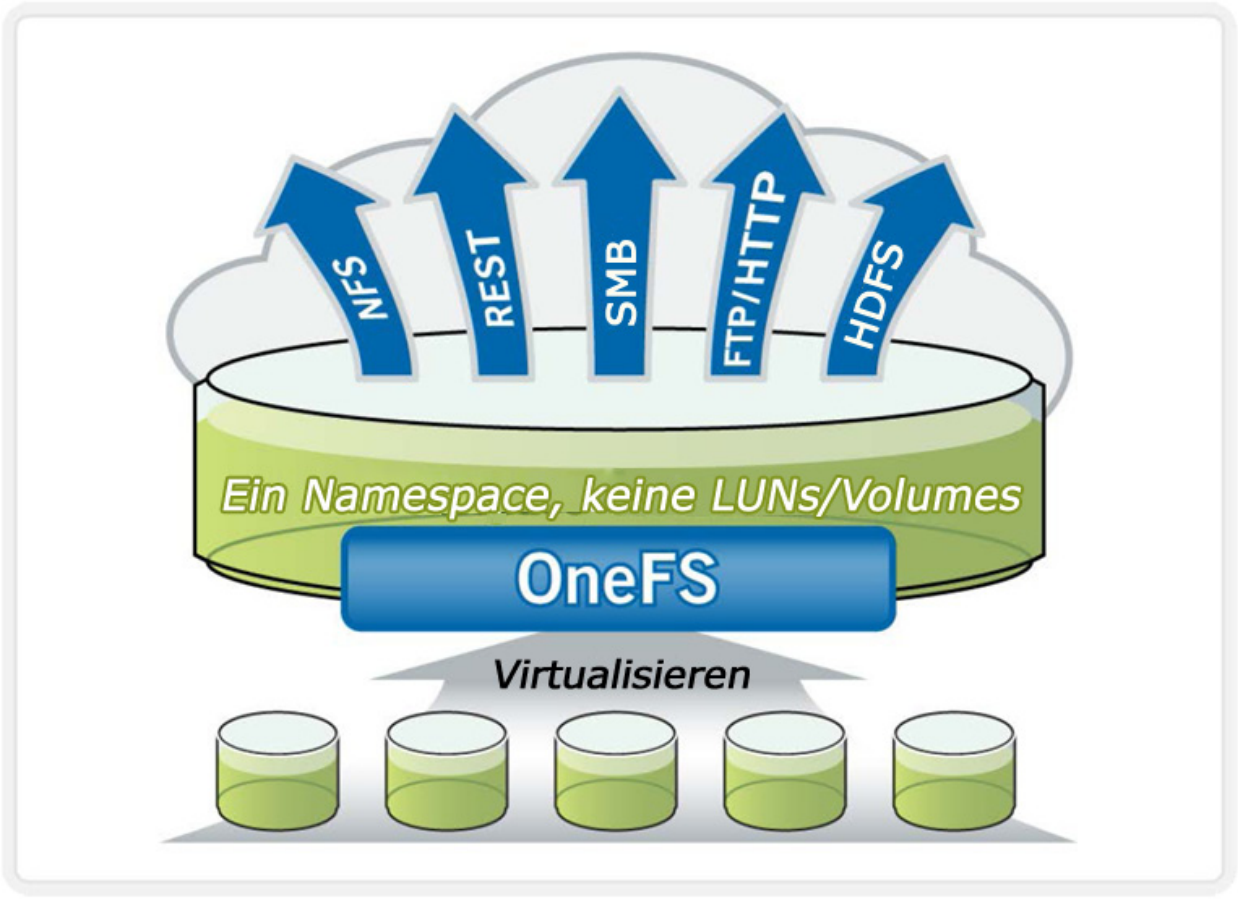

Abbildung 5: Ein einziges Dateisystem mit mehreren Zugriffsprotokollen

OneFS ist tatsächlich ein einziges Dateisystem mit einem globalen Namespace. Daten und Metadaten werden im Hinblick auf Redundanz und Verfügbarkeit auf mehrere Nodes in Stripes verteilt. Der Speicher wurde für die Benutzer und den Administrator komplett virtualisiert. Die Dateistruktur kann ohne Planung oder Überwachung der Benutzerverwendung organisch wachsen. Der Administrator braucht sich keine Gedanken über das Tiering von Dateien auf der entsprechenden Festplatte zu machen, da dies automatisch von Isilon SmartPools gemanagt wird, ohne die einzige Struktur zu unterbrechen. Es muss ebenfalls nicht berücksichtigt werden, wie solch eine große Struktur repliziert werden kann, denn der Isilon SyncIQ™-Service sorgt automatisch und unabhängig von der Form oder Tiefe der Struktur für eine parallele Übertragung der Dateistruktur auf ein oder mehrere alternative Cluster.

Dieses Design sollte mit der Namespace-Aggregation verglichen werden. Dies ist eine allgemein verwendete Technologie, um einen herkömmlichen NAS so aussehen zu lassen, als hätte er einen einzigen Namespace. Mit der Namespace-Aggregation müssen Dateien nach wie vor in separaten Volumes verwaltet werden, aber eine einfache „Furnier"-Ebene ermöglicht das „Verkleben" einzelner Verzeichnisse in Volumes über symbolische Links mit einer „obersten" Struktur. In diesem Modell sind nach wie vor LUNs und Volume-Begrenzungen vorhanden. Im Hinblick auf den Lastenausgleich müssen die Dateien manuell von Volume zu Volume verschoben werden. Der Administrator muss die Dateistruktur umsichtig gestalten. Das Tiering ist bei Weitem nicht nahtlos, sodass erhebliche und kontinuierliche Eingriffe erforderlich sind. Beim Failover muss eine Spiegelung von Dateien zwischen Volumes erfolgen, was die Effizienz beeinträchtigt und die Kosten für Anschaffung, Strom und Kühlung erhöht. Insgesamt ist der Aufwand für den Administrator bei der Namespace-Aggregation höher als der für ein einfaches herkömmliches NAS-Gerät. Daher ist ein großes Wachstum für solche Infrastrukturen nicht möglich.

Datenlayout

In OneFS werden physische Pointer und Extents für Metadaten und Speicherdateien sowie Verzeichnismetadaten in Inodes verwendet. B-Strukturen werden häufig im Dateisystem eingesetzt und ermöglichen Skalierbarkeit auf Milliarden von Objekten und ein nahezu sofortiges Abfragen von Daten oder Metadaten. OneFS ist ein vollständig symmetrisches und hochgradig verteiltes Dateisystem. Daten und Metadaten sind immer über mehrere Hardwaregeräte redundant. Alle Daten sind anhand der Löschcodierung über die Nodes im Cluster geschützt. Dadurch entsteht ein hocheffizientes Cluster mit einem Verhältnis von mindestens 80 % Rohkapazität zur nutzbaren Kapazität auf Clustern von fünf Nodes oder mehr. Metadaten (normalerweise weniger als 1 % des Systems) werden im Cluster auf Performance und Verfügbarkeit gespiegelt. Da OneFS nicht von RAID abhängt, kann die Menge an Redundanz durch den Administrator auf Datei- oder Verzeichnisebene über die Standards des Clusters hinaus ausgewählt werden. Metadaten und Sperraufgaben werden von allen Nodes gemeinsam und gleichgestellt in einer Peer-to-Peer-Architektur gemanagt. Diese Symmetrie ist der Schlüssel für diese einfache und ausfallsichere Architektur. Es gibt keinen einzigen Metadatenserver, Sperrmanager oder Gateway-Node.

Da OneFS gleichzeitig auf Blöcke von verschiedenen Geräten zugreifen muss, wird das für Daten und Metadaten verwendete Adressierungsschema auf der physischen Ebene über einen Tupel von {Node, Laufwerk, Offset} indiziert. Beispiel: Wenn 12345 eine Blockadresse für einen Block auf Festplatte 2 von Node 3 ist, wird {3,2,12345} angezeigt. Alle Metadaten innerhalb des Clusters werden zur Datensicherung mehrfach gespiegelt, mindestens auf das Maß an Redundanz der zugehörigen Datei. Wenn beispielsweise eine Datei einen Löschcodeschutz von „N+2" hat, sie also zwei gleichzeitige Ausfälle überstehen könnte, werden alle für den Dateizugriff erforderlichen Metadaten dreimal gespiegelt, sodass diese ebenfalls zwei Ausfälle überleben könnten. Das Dateisystem ist so angelegt, dass jede Struktur grundsätzlich alle Blöcke auf allen Nodes im Cluster verwenden kann.

Andere Speichersysteme senden Daten über RAID und Volume-Managementebenen, was zu Ineffizienz beim Datenlayout und einem nicht optimierten Blockzugriff führt. Isilon OneFS steuert die Platzierung von Dateien direkt, bis hin zur Sektorebene jedes Laufwerks an einem beliebigen Ort im Cluster. Dies ermöglicht eine optimierte Datenplatzierung und optimierte I/O-Muster. Nicht erforderliche Lese-/Änderungs-/Schreibvorgänge werden vermieden. Da die Daten Datei für Datei auf Festplatten abgelegt werden, steuert OneFS die Art des Striping sowie die Redundanz des Speichersystems auf flexible Weise auf System-, Verzeichnis- und sogar Dateiebene. Bei herkömmlichen Speichersystemen müsste ein ganzes RAID-Volume einer bestimmten Performancekategorie- und Sicherungseinstellung zugewiesen werden. Beispielsweise kann eine Reihe von Festplatten für eine Datenbank in einer RAID 1+0-Struktur angeordnet werden. Dadurch lässt sich der Einsatz von Spindeln über den gesamten Speicherbestand nur schwer optimieren (da inaktive Spindeln nicht ausgeliehen werden können). Weiterhin führt dies zu unflexiblen Designs, die sich nicht auf die jeweiligen geschäftlichen Anforderungen anpassen lassen. OneFS ermöglicht jederzeit individuelles Tuning und flexible Änderungen, und zwar vollständig online.

Dateischreibvorgänge

Die OneFS-Software wird auf allen Nodes auf gleiche Weise ausgeführt. So entsteht ein einziges Dateisystem, das über alle Nodes hinweg ausgeführt wird. Das Cluster wird nicht von einem einzelnen Node oder „Master" gesteuert. Alle Nodes sind völlig gleichwertig.

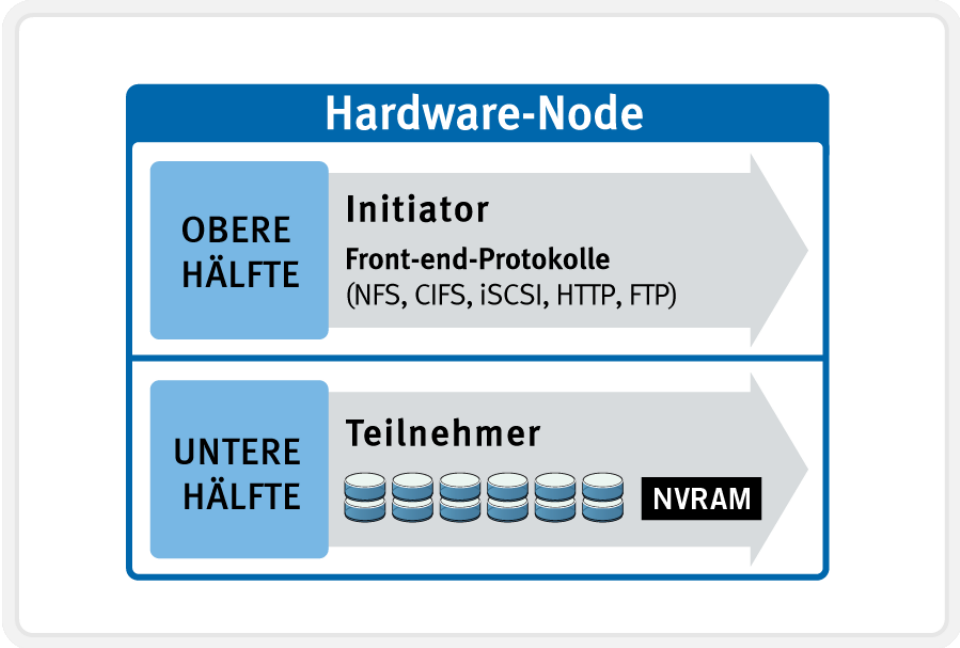

Abbildung 6: Modell von an I/O-Vorgängen beteiligten Node-Komponenten

Wenn Sie alle Komponenten innerhalb jedes Node eines Clusters in Betracht ziehen, die allgemein an I/O-Vorgängen beteiligt sind, sähe dies wie Abbildung 6 oben aus. Wir haben den Stapel in eine „oberste" Ebene, den Initiator, und eine „untere" Ebene, den Teilnehmer, aufgeteilt. Diese Aufteilung wird als „logisches Modell" für die Analyse aller Lese- oder Schreibvorgänge verwendet. Auf physischer Ebene bearbeiten CPUs und RAM-Cache in den Nodes gleichzeitig die Initiator- und Teilnehmeraufgaben für I/O-Vorgänge, die im Cluster stattfinden. Es gibt Caches und einen verteilten Sperrmanager, die aus Gründen der Einfachheit nicht im Diagramm oben aufgeführt sind. Sie werden in den späteren Abschnitten dieses Dokuments behandelt.

Wenn ein Client eine Verbindung mit einem Node herstellt, um in eine Datei zu schreiben, wird eine Verbindung mit der oberen Hälfte oder dem Initiator dieses Node hergestellt. Die Dateien werden in kleinere logische Segmente namens Stripes unterteilt, bevor sie in die untere Hälfte oder den Teilnehmer eines Node (Festplatte) geschrieben werden. Durch das ausfallsichere Puffern mithilfe eines Schreib-Coalescer wird dafür gesorgt, dass die Schreibvorgänge effizient sind und Lese-Änderungs-Schreibvorgänge vermieden werden. Die Größe der einzelnen Dateisegmente wird als Größe der Stripe-Einheiten bezeichnet.

OneFS verteilt die Daten in Stripes auf alle Nodes – nicht einfach auf Festplatten – und schützt die Dateien, Verzeichnisse und die damit verbundenen Metadaten über Softwarelöschcodes oder Spiegelungstechnologie. Für Daten verwendet OneFS (je nach Ermessen des Administrators) entweder das Reed-Solomon-Löschcodierungssystem für den Schutz der Daten oder (weniger häufig) die Spiegelung. Das Anwenden von Spiegelungen auf Benutzerdaten erfolgt eher in Fällen mit hoher Transaktionsperformance. Für die meisten Benutzerdaten wird in der Regel die Löschcodierung eingesetzt, da sie eine extrem hohe Performance ohne Beeinträchtigung der Festplatteneffizienz bietet. Die Löschcodierung kann mehr als 80 % Effizienz auf unformatierten Festplatten mit fünf Nodes oder mehr bereitstellen, auf großen Clustern sogar gleichzeitig mit einer Vierfachredundanz. Die Stripe-Breite einer bestimmten Datei ist die Anzahl der Nodes (nicht Festplatten), über die eine Datei geschrieben ist. Sie wird von der Anzahl der Nodes im Cluster, der Dateigröße und der Schutzeinstellung bestimmt (z. B. N+2).

OneFS verwendet erweiterte Algorithmen, um ein Datenlayout mit maximaler Effizienz und maximalem Schutz festzulegen. Wenn ein Client eine Verbindung mit einem Node herstellt, tritt der Initiator dieses Node als „Kapitän" für das Schreibdatenlayout dieser Datei auf. In einem Isilon-Cluster werden alle Daten, Parität, Metadaten und Inodes auf mehrere Nodes und sogar über mehrere Laufwerke innerhalb der Nodes verteilt. Abbildung 7 unten zeigt einen Dateischreibvorgang, der über alle Nodes in einem Cluster mit drei Nodes erfolgt.

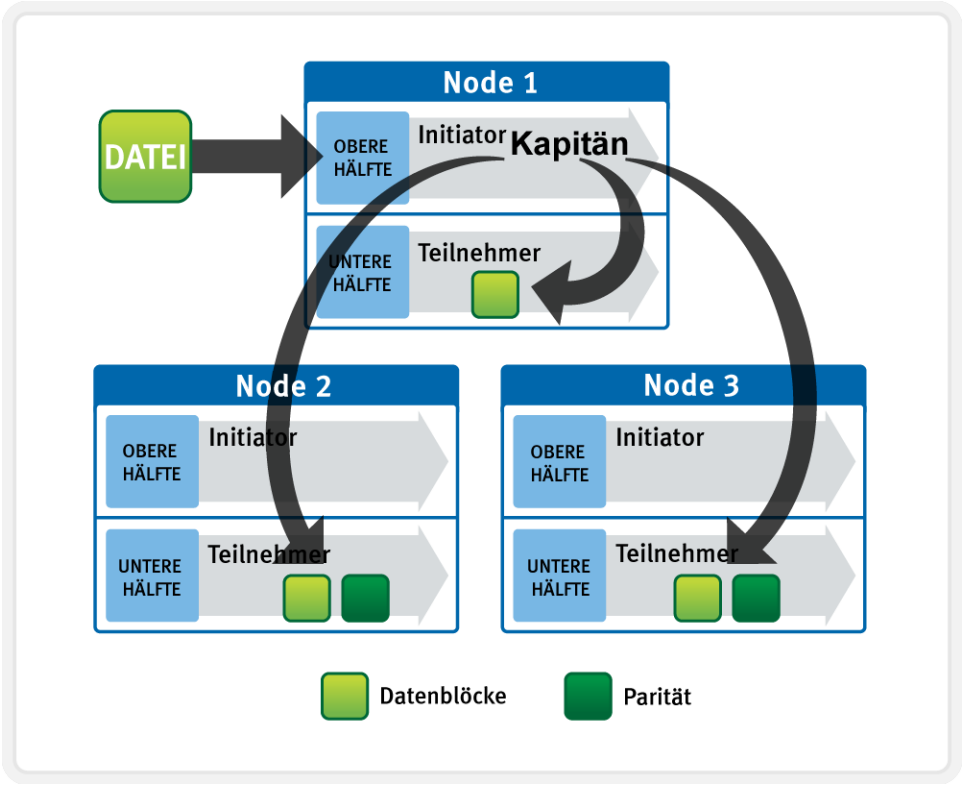

Abbildung 7: Ein Dateischreibvorgang auf einem Isilon-Cluster mit drei Nodes

OneFS verwendet das Back-end-Netzwerk von InfiniBand, um Daten automatisch auf alle Nodes im Cluster zu verteilen, sodass keine weitere Verarbeitung erforderlich ist. Wenn Daten geschrieben werden, werden sie auf der angegebenen Ebene geschützt.

Wenn Schreibvorgänge stattfinden, teilt OneFS die Daten in atomare Einheiten namens Schutzgruppen auf. Schutzgruppen verfügen über integrierte Redundanz. Wenn alle Schutzgruppen geschützt sind, ist auch die gesamte Datei geschützt. Für Dateien, die mit Löschcodes geschützt sind, besteht die Schutzgruppe aus einer Reihe von Datenblöcken sowie einer Reihe von Löschcodes für diese Datenblöcke. Bei gespiegelten Dateien besteht eine Schutzgruppe aus allen Spiegelungen eines Satzes an Blöcken. OneFS ist in der Lage, die in einer Datei verwendete Schutzgruppe dynamisch während des Schreibvorgangs zu ändern. Dies ermöglicht viele zusätzliche Funktionen, z. B. kann das System angewiesen werden, in Situationen ohne Blockierung weiterzuarbeiten, in denen temporäre Node-Ausfälle im Cluster verhindern würden, dass die gewünschte Anzahl von Löschcodes verwendet wird. In solchen Fällen kann vorübergehend die Spiegelung verwendet werden, damit die Schreibvorgänge fortgesetzt werden können. Wenn Nodes im Cluster wiederhergestellt werden, werden diese gespiegelten Schutzgruppen nahtlos und automatisch, also ohne Eingriff durch den Administrator, wieder in den Löschcodeschutz umgewandelt.

Die Blockgröße im OneFS-Dateisystem beträgt 8 KB. Eine Datei, die kleiner als 8 KB ist, verwendet einen vollständigen Block von 8 KB. Je nach Datensicherheitslevel verwendet diese Datei von 8 KB möglicherweise mehr als 8 KB Datenspeicherplatz. Die Einstellungen für die Datensicherheit werden in einem späteren Abschnitt dieses Dokuments näher behandelt. OneFS kann Dateisysteme mit Milliarden kleiner Dateien bei höchster Performance unterstützen, da alle Festplattenstrukturen für eine Skalierung auf solche Größen entwickelt wurden. Darüber hinaus bietet OneFS unabhängig von der Gesamtanzahl der Objekte einen nahezu sofortigen Zugriff auf beliebige Objekte. Bei größeren Dateien nutzt OneFS mehrere zusammenhängende Blöcke von 8 KB. In diesen Fällen können bis zu 16 zusammenhängende Blöcke auf die Festplatte eines einzigen Node verteilt werden. Wenn eine Datei eine Größe von 32 KB aufweist, werden vier zusammenhängende Blöcke von 8 KB verwendet.

Bei noch größeren Dateien kann OneFS die sequenzielle Performance maximieren, indem eine Stripe-Einheit aus 16 zusammenhängenden Blöcken eingesetzt wird, sodass sich insgesamt 128 KB pro Stripe-Einheit ergeben. Während eines Schreibvorgangs werden die Daten in Stripe-Einheiten unterteilt und als Schutzgruppe über mehrere Nodes verteilt. Während die Daten über das gesamte Cluster hinweg abgelegt werden, werden je nach Bedarf Löschcodes oder Spiegelungen verteilt, um dafür zu sorgen, dass die Dateien jederzeit geschützt sind.

Eine der wichtigsten Funktionen der in OneFS integrierten AutoBalance-Funktion besteht darin, die Daten neu zuzuweisen und abzugleichen. So wird der Speicherplatz so optimal wie möglich genutzt. In den meisten Fällen kann die Stripe-Größe von großen Dateien erhöht werden, um neuen freien Speicherplatz zu nutzen (wenn Nodes hinzugefügt werden) und das Striping auf der Festplatte effizienter zu gestalten. AutoBalance sorgt für hohe Festplatteneffizienz, wodurch „Hotspots" auf der Festplatte automatisch eliminiert werden.

Die obere Initiatorhälfte des „Kapitän"-Node verwendet eine patentierte, modifizierte Zweiphasen-Commit-Transaktion, um die Schreibvorgänge sicher auf mehrere NVRAM im Cluster zu verteilen, wie in Abbildung 8 gezeigt.

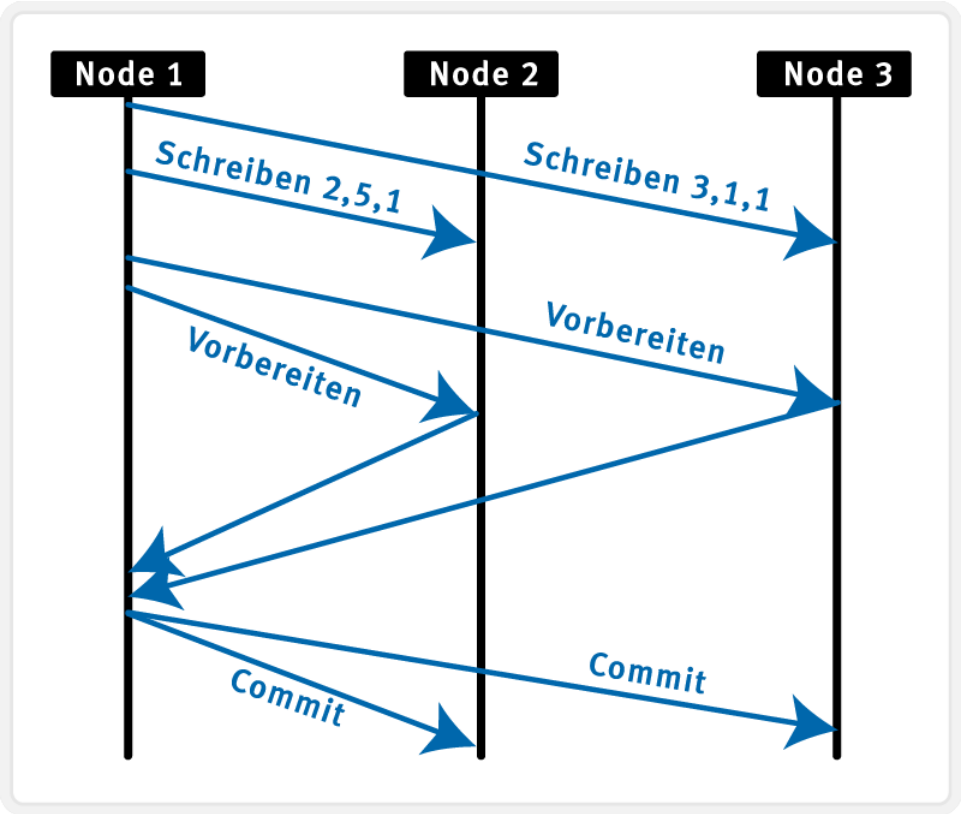

Abbildung 8: Verteilte Transaktionen und Zweiphasen-Commit-Transaktion

Jeder Node, der über Blöcke in einem bestimmten Schreibvorgang verfügt, ist an einer Zweiphasen-Commit-Transaktion beteiligt. Der Mechanismus beruht zur Protokollierung aller Transaktionen, die über jeden Node im Speichercluster erfolgen, auf NVRAM. Die Verwendung mehrerer NVRAM parallel ermöglicht Schreibvorgänge mit hohem Durchsatz bei gleichzeitiger Wahrung der Datensicherheit vor jeder Form von Ausfällen, einschließlich Stromausfällen. Wenn ein Node während einer Transaktion ausfällt, wird die Transaktion sofort ohne den betroffenen Node neu gestartet. Wenn der Node wieder einsatzbereit ist, muss er lediglich das entsprechende Journal von NVRAM wiedergeben – was Sekunden oder höchstens einige Minuten in Anspruch nimmt – und gelegentlich für die AutoBalance-Funktion Dateien neu abstimmen, die an der Transaktion beteiligt waren. Teure „fsck-Prozesse" oder „Festplattenprüfprozesse" sind in keinem Fall erforderlich. Zudem muss nie eine langwierige Neusynchronisierung durchgeführt werden. Schreibvorgänge werden nie aufgrund eines Ausfalls blockiert. Das patentierte Transaktionssystem ist eine der Methoden, mit denen OneFS Single-Points-of-Failure oder sogar mehrere Points-of-Failure ausschaltet.

Während eines Schreibvorgangs orchestriert der Initiator das Layout von Daten und Metadaten, die Erstellung von Löschcodes und den normalen Betrieb für Sperrmanagement und Berechtigungskontrolle. Ein Administrator kann über das Webmanagement oder die Befehlszeilenoberfläche jederzeit die von OneFS getroffenen Layoutentscheidungen für den jeweiligen Workflow optimieren. Der Administrator kann auf Datei- oder Verzeichnisebene eine Auswahl aus den unten aufgeführten Zugriffsmustern treffen:

- **Gleichzeitigkeit:** Optimierung auf die aktuelle Last auf dem Cluster, Möglichkeit für mehrere Clients gleichzeitig. Diese Einstellung bietet das beste Verhalten für gemischte Workloads.

- **Streaming:** Optimierung für das Hochgeschwindigkeitsstreaming einer einzigen Datei, sodass beispielsweise sehr schnelle Lesevorgänge mit einem einzigen Client ermöglicht werden
- **Zufällig:** Optimierung für unvorhersehbaren Zugriff auf die Datei durch Feinabstimmung des Striping und Deaktivieren aller Pre-Fetch-Caches

Caching in OneFS

Das Design der Cachinginfrastruktur in OneFS basiert auf einer Agggregierung des auf jedem Node in einem Cluster vorhandenen Caches in einen global zugänglichen Speicherpool. Dazu verwendet Isilon ein effizientes Messaging-System, ähnlich NUMA (Non-Uniform Memory Access). Dadurch wird der gesamte Speichercache des Node für jeden einzelnen Node im Cluster verfügbar. Der Zugriff auf den Remotespeicher erfolgt über einen internen Interconnect und weist eine erheblich niedrigere Latenz als beim Zugriff auf Festplattenlaufwerke auf.

Zum Zugriff auf den Remotespeicher nutzt OneFS das SDP (Sockets Direct Protocol) über ein IB (Infiniband) Back-end-Interconnect auf dem Cluster, im Grunde ein verteilter Systembus. Das SDP bietet eine effiziente, einem Socket ähnliche Schnittstelle zwischen Nodes, mit der bei Verwendung einer Switch-gestützten Sterntopologie dafür gesorgt wird, dass die Adressen von Remotespeichern immer nur einen IB-Hop entfernt sind. Der Zugriff auf den Remotespeicher ist zwar nicht so schnell wie der auf den lokalen Speicher, aufgrund der niedrigen Latenz von IB aber immer noch sehr schnell.

Das Cachingsubsystem von OneFS ist im gesamten Cluster kohärent. Wenn also der gleiche Inhalt in den privaten Caches mehrerer Nodes vorhanden ist, sind diese Cachedaten über alle Instanzen konsistent. OneFS verwendet das MESI-Protokoll, um Kohärenz im Cache aufrechtzuerhalten. Mit diesem Protokoll wird die Richtlinie „beim Schreibvorgang ungültig" implementiert, um zu ermöglichen, dass alle Daten im gesamten gemeinsamen Cache konsistent sind.

OneFS verwendet bis zu drei Ebenen für den Lesecache sowie einen NVRAM-gestützten Schreibcache oder Coalescer. Diese Caches und ihre allgemeine Interaktion sind im folgenden Diagramm dargestellt.

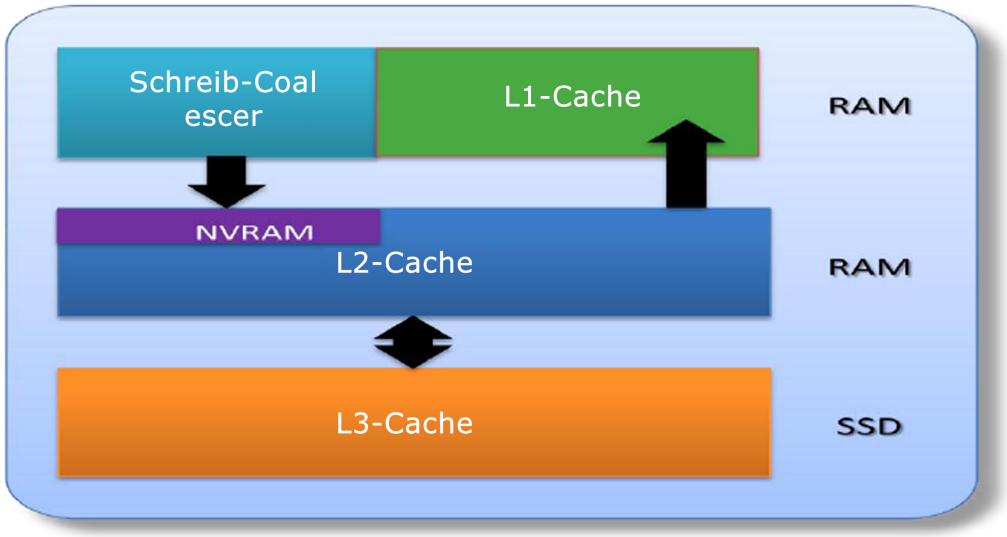

Abbildung 9: OneFS-Cachinghierarchie

Die ersten beiden Arten des Lesecaches, Level 1 (L1) und Level 2 (L2), sind speicherbasiert (RAM) und vergleichbar mit dem in den Prozessoren (CPUs) verwendeten Cache. Diese beiden Cacheebenen sind in allen Speicher-Nodes von Isilon vorhanden.

Ein optionaler dritter Tier des Lesecaches namens SmartFlash oder Level 3 (L3) kann ebenfalls für Nodes konfiguriert werden, die SSDs (Solid State Drives) enthalten. SmartFlash (L3) ist ein Leerungscache, der durch die vom Speicher gelöschten L2-Cacheblöcke aufgefüllt wird. Es gibt bei der Verwendung von SSDs für das Caching statt als herkömmliche Dateisystem-Speichergeräte mehrere Vorteile. Beispiel: Wenn SSDs für die Zwischenspeicherung reserviert sind, werden die gesamten SSDs verwendet und die Schreibvorgänge erfolgen auf äußerst lineare und vorhersehbare Art und Weise. Dies führt im Gegensatz zur normalen Dateisystemnutzung zu einer weitaus besseren Auslastung und darüber hinaus zu erheblich geringerer Abnutzung und verlängerter Lebensdauer, insbesondere bei zufälligen Schreibvorgängen. Im Vergleich zu einer Verwendung von SSDs als Storage Tier wird mit dem Einsatz von SSDs für den Cache auch die Dimensionierung der SSD-Kapazität sehr viel überschaubarer und weniger fehleranfällig.

Name	Typ	Persistenz	Beschreibung
L1-Cache	RAM	Flüchtig	Auch Front-end-Cache genannt, enthält bereinigte, im Cluster kohärente Kopien von Dateisystemdaten und Metadatenblöcken, die über NFS- und SMB-Clients usw. über das Front-end-Netzwerk angefordert werden
L2-Cache	RAM	Flüchtig	Back-end-Cache, enthält bereinigte Kopien von Dateisystemdaten und Metadaten auf einem lokalen Node
SmartCache/ Schreib-Coalescer	NVRAM	Nicht flüchtig	Persistenter, batteriegestützter NVRAM-Journalcache, der alle ausstehenden Schreibvorgänge an Front-end-Dateien puffert, die nicht auf die Festplatte geschrieben wurden
SmartFlash L3-Cache	SSD	Nicht flüchtig	Enthält die aus dem L2-Cache entfernten File-basierten Daten und Metadatenblöcke, wodurch sich die Kapazität des L2-Caches praktisch erhöht hat

OneFS schreibt vor, dass eine Datei über mehrere Nodes im Cluster und möglicherweise mehrere Laufwerke in einem Node geschrieben wird. Daher werden bei allen Leseanforderungen Remotedaten (und möglicherweise lokale Daten) gelesen. Wenn eine Leseanforderung von einem Client eingeht, überprüft OneFS, ob die angeforderten Daten im lokalen Cache vorhanden sind. Alle im lokalen Cache gespeicherten Daten werden sofort gelesen. Wenn die angeforderten Daten nicht im lokalen Cache vorhanden sind, werden sie von der Festplatte gelesen. Wenn Daten nicht auf dem lokalen Node gespeichert sind, erfolgt eine Anfrage vom Remote-Node, auf denen sich die Daten befinden. Auf jedem der anderen Nodes wird eine weitere Cacheabfrage durchgeführt. Alle Daten im Cache werden sofort zurückgegeben und alle nicht im Cache gespeicherten Daten werden von der Festplatte abgerufen.

Wenn die Daten vom lokalen und Remotecache (und ggf. der Festplatte) abgerufen wurden, werden sie an den Client zurückgegeben.

Die allgemeinen Schritte für die Umsetzung einer Leseanforderung auf einem lokalen und einem Remote-Node sind:

Auf dem lokalen Node (der Node, bei dem die Anfrage eingeht):

1. Ermitteln Sie, ob ein Teil der angeforderten Daten im lokalen L1-Cache vorhanden ist. Falls ja, kehren Sie zum Client zurück.
2. Wenn die Daten nicht im lokalen Cache vorhanden sind, fordern Sie sie von den Remote-Nodes an.

Auf den Remote-Nodes:

1. Ermitteln Sie, ob die angeforderten Daten im lokalen L2- oder L3-Cache vorhanden sind. Falls ja, kehren Sie zum anfordernden Node zurück.
2. Wenn die Daten nicht im lokalen Cache sind, lesen Sie sie von der Festplatte und kehren Sie zum anfordernden Node zurück.

Durch das Caching von Schreibvorgängen wird das Schreiben von Daten auf einen Isilon-Cluster beschleunigt. Dies wird dadurch erreicht, dass kleinere Schreibanforderungen in einem Batch zusammengefasst und in größeren Segmenten an die Festplatte gesendet werden, was einen Großteil der Latenz beim Schreiben auf Festplatten vermeidet. Wenn Clients an das Cluster schreiben, schreibt OneFS die Daten vorübergehend in einen NVRAM-basierten Journalcache auf dem Initiator-Node statt sofort auf die Festplatte. OneFS kann diese zwischengespeicherten Schreibvorgänge dann zu einem späteren, passenderen Zeitpunkt an die Festplatte weiterleiten. Darüber hinaus werden diese Schreibvorgänge auch auf den NVRAM-Journalen der Teilnehmer-Nodes gespiegelt, um die Anforderungen für die Dateisicherung zu erfüllen. Im Falle einer Clusterteilung oder eines unerwarteten Node-Ausfalls sind daher selbst nicht übernommene zwischengespeicherte Schreibvorgänge vollständig geschützt.

Der Schreibcache arbeitet wie folgt:

- Ein NFS-Client sendet eine Schreibanforderung für eine Datei mit N+2-Schutz an Node 1.
- Node 1 nimmt die Schreibvorgänge in seinen NVRAM-Schreibcache auf (schneller Pfad) und spiegelt dann die Schreibvorgänge in die Protokolldateien der Teilnehmer-Nodes, um sie zu schützen.
- Bestätigungen für die Schreibvorgänge werden sofort an den NFS-Client zurückgegeben, sodass die Latenz beim Schreiben auf Festplatten vermieden wird.
- Der nach und nach anwachsende Schreibcache von Node 1 wird regelmäßig geleert und die Schreibvorgänge werden unter Anwendung des entsprechenden Paritätsschutzes (N+2) über den Zweiphasen-Commit-Prozess (siehe Beschreibung weiter oben) auf die Festplatte übertragen.
- Der Schreibcache und die Protokolldateien des Teilnehmer-Node werden geleert und können neue Schreibvorgänge aufnehmen.

Lesen von Dateien

In einem Isilon-Cluster werden alle Daten, Metadaten und Inodes auf mehrere Nodes und sogar über mehrere Laufwerke innerhalb der Nodes verteilt. Beim Lesen oder Schreiben in das Cluster fungiert der Node, an den ein Client angebunden ist, als „Kapitän" für den Vorgang.

Bei einem Lesevorgang erfasst der „Kapitän"-Node alle Daten aus den verschiedenen Nodes im Cluster und stellt sie der anfordernden Stelle auf zusammenhängende Weise bereit.

Dank kostenoptimierter Hardware nach Branchenstandard bietet das Isilon-Cluster bei der Übertragung vom Cache auf die Festplatte ein hohes Verhältnis (mehrere GB pro Node), das den Lese- und Schreibvorgängen je nach Bedarf dynamisch zugewiesen wird. Dieser RAM-basierte Cache ist über alle Nodes im Cluster einheitlich und kohärent, sodass eine Leseanforderung von einem Client an einen bestimmten Node die bereits an einen anderen Node weitergeleiteten I/O-Vorgänge nutzen kann. Diese zwischengespeicherten Blöcke sind über die InfiniBand-Backplane mit niedriger Latenz von beliebigen Nodes aus schnell zugänglich, was einen großen, effiziente RAM-Cache und somit eine erheblich schnellere Leseperformance ermöglicht. Je mehr das Cluster anwächst, desto größer werden die Vorteile beim Caching. Aus diesem Grund ist die Anzahl der I/O-Schreibvorgänge an die Festplatte in einem Isilon-Cluster in der Regel deutlich niedriger als auf herkömmlichen Plattformen, was zu niedrigerer Latenz und einem besseren Benutzererlebnis führt.

Bei Dateien mit einem gleichzeitigen oder Streamingzugriffsmuster nutzt OneFS die Pre-Fetch-Funktion von Daten auf Grundlage der von der SmartRead-Komponente in Isilon verwendeten Heuristik. SmartRead kann eine „Datenpipeline" aus dem L2-Cache lesen und über einen lokalen L1-Cache auf dem „Kapitän"-Node vorabrufen. Dadurch verbessert sich die sequenzielle Leseperformance über alle Protokolle hinweg erheblich und die Lesevorgänge werden innerhalb von Millisekunden direkt vom RAM bereitgestellt. In Fällen mit hohem sequenziellem Zugriff kann SmartRead eine äußerst große Datenmenge vorabrufen, sodass Lese- oder Schreibvorgänge für einzelne Dateien mit extrem hohen Datenraten erfolgen können.

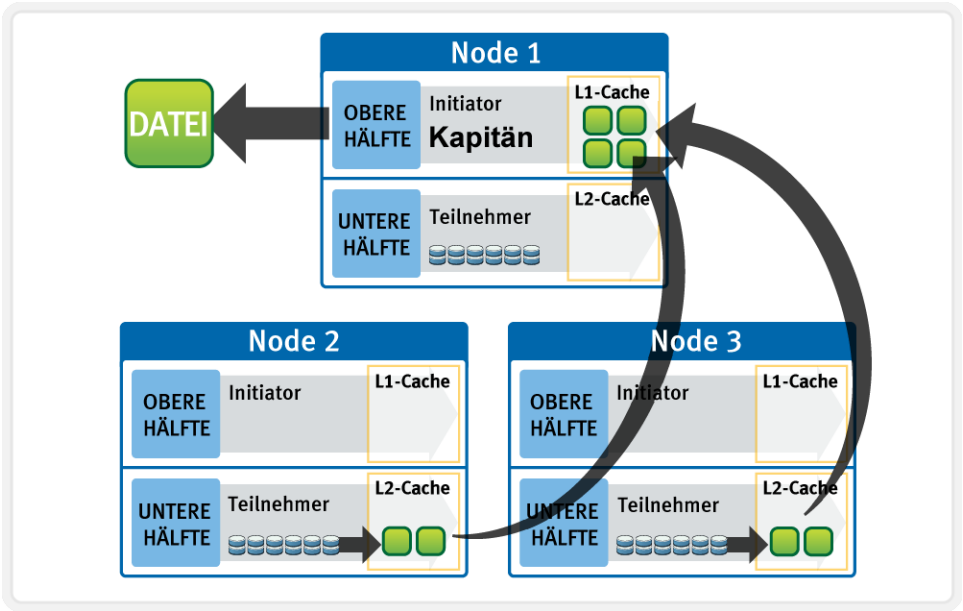

Abbildung 10: Lesevorgang für eine Datei auf einem Isilon-Cluster mit drei Nodes

Abbildung 10 zeigt, wie SmartRead eine nicht im Cache gespeicherte Datei mit sequenziellem Zugriff liest, wobei die Datei von einem Client angefordert wird, der in einem Cluster mit drei Nodes an Node 1 angebunden ist.

1. Node 1 liest Metadaten, um zu ermitteln, an welchem Ort sich die verschiedenen Blöcke von File-basierten Daten befinden.
2. Node 1 prüft außerdem seinen L1-Cache, um festzustellen, ob die angeforderten File-basierten Daten enthalten sind.
3. Node 1 erstellt eine Lesepipeline und sendet gleichzeitig Anforderungen an alle Nodes, die einen Teil der File-basierten Daten enthalten, um diese File-basierten Daten von der Festplatte abzurufen.
4. Die einzelnen Nodes rufen die Blöcke der File-basierten Daten von der Festplatte in ihren L2-Cache (oder, falls verfügbar, in den L3-SmartFlash-Cache) ab und übertragen die File-basierten Daten an Node 1.
5. Node 1 speichert die eingehenden Daten im L1-Cache und stellt die Datei gleichzeitig für den Client bereit. Dabei wird der Pre-Fetch-Prozess fortgesetzt.
6. In Fällen mit hohem sequenziellem Zugriff werden die Daten im L1-Cache optional „zurückgelassen", um RAM für andere L1- oder L2-Cacheanforderungen freizugeben.

Das intelligente Caching mit SmartReads ermöglicht eine sehr hohe Leseperformance mit einem hohen gleichzeitigen Zugriff. Noch wichtiger ist, dass Node 1 die File-basierten Daten vom Node 2-Cache (über den Cluster-Interconnect mit niedriger Latenz) schneller abrufen kann, als auf die eigene lokale Festplatte zuzugreifen. Die Algorithmen in SmartReads steuern, wie aggressiv der Pre-Fetch-Vorgang ausgeführt wird (Deaktivierung von Pre-Fetch für zufälligen Zugriff) und wie lange die Daten im Cache verbleiben. Außerdem optimieren sie das Caching von Daten, indem sie den entsprechenden Cache auswählen.

Sperren und gleichzeitiger Zugriff

OneFS verfügt über einen vollständig verteilten Sperrmanager, mit dem die Sperren für Daten in allen Nodes in einem Cluster gesteuert werden können. Der Sperrmanager ist hochgradig erweiterbar und ermöglicht unterschiedliche „Sperrarten", um sowohl Dateisystemsperren als auch clusterkohärente Sperren auf Protokollebene wie SMB-Freigabesperren- oder wahlfreie NFS-Sperren zu unterstützen. Darüber hinaus bietet OneFS Support für delegierte Sperren wie CIFS Oplocks und NFSv4-Delegierungen.

Jeder Node in einem Cluster ist ein Koordinator für das Sperren von Ressourcen. Ein Koordinator wird sperrbaren Ressourcen auf Grundlage eines erweiterten Hashing-Algorithmus zugewiesen. Der Algorithmus wurde so entwickelt, ist, dass der Koordinator sich am Ende fast immer auf einem anderen Node befindet, als der Initiator der Anforderung. Wenn eine Sperre für eine Datei angefordert wird, kann es sich um eine gemeinsam genutzte Sperre handeln (sodass mehrere Benutzer die Sperre gleichzeitig nutzen können, in der Regel für Lesevorgänge) oder eine exklusive Sperre (nur jeweils ein Benutzer, in der Regel für Schreibvorgänge).

Abbildung 11 zeigt ein Beispiel dafür, wie Threads von verschiedenen Nodes eine Sperre vom Koordinator anfordern können.

1. Node 2 wurde als Koordinator dieser Ressourcen festgelegt.
2. Thread 1 von Node 4 und Thread 2 von Node 3 fordern gleichzeitig eine gemeinsame Sperre für eine Datei von Node 2 an.

3. Node 2 überprüft, ob eine exklusive Sperre für die angeforderte Datei vorhanden ist.
4. Wenn keine exklusive Sperre besteht, gewährt Node 2 Thread 1 von Node 4 und Thread 2 von Node 3 gemeinsame Sperren für die angeforderte Datei.
5. Node 3 und Node 4 führen jetzt einen Lesevorgang für die angeforderte Datei durch.
6. Thread 3 von Node 1 fordert eine exklusive Sperre für dieselbe Datei an, die von Node 3 und Node 4 gelesen wird.
7. Node 2 überprüft bei Node 3 und Node 4, ob die gemeinsame Sperre aufgehoben werden kann.
8. Node 3 und Node 4 haben den Lesevorgang noch nicht abgeschlossen, sodass Node 2 Thread 3 von Node 1 auffordert, einen Augenblick zu warten.
9. Thread 3 auf Node 1 wird so lange blockiert, bis die exklusive Sperre durch Node 2 vergeben wird, und schließt den Schreibvorgang ab.

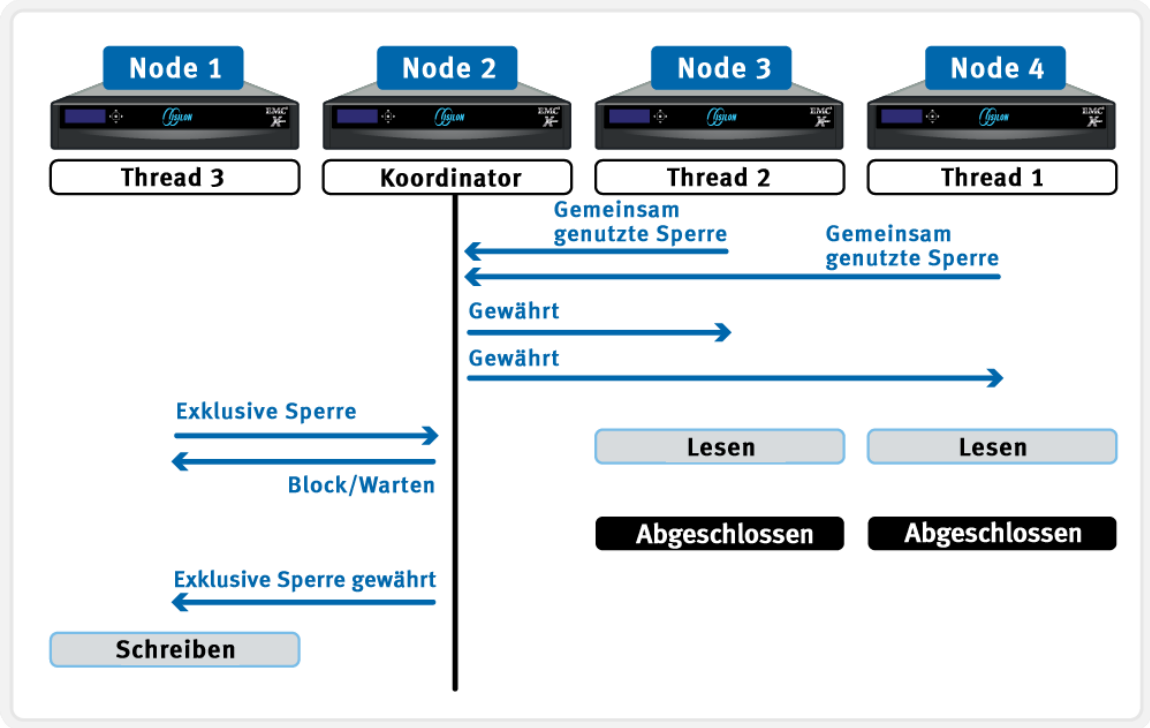

Abbildung 11: Verteilter Sperrmanager

I/O-Vorgänge mit Multithreading

Angesichts der zunehmenden Nutzung riesiger NFS-Datastores für die Servervirtualisierung und den Support von Anwendungen der Enterprise-Klasse sind für große Dateien ein hoher Durchsatz und geringe Latenz erforderlich. Aus diesem Grund unterstützt der OneFS Multi-Writer mehrere Threads gleichzeitig für Schreibvorgänge auf einzelne Dateien.

Im obigen Beispiel kann der gleichzeitige Schreibzugriff auf eine große Datei durch den exklusiven, auf der gesamten Dateiebene angewendeten Sperrmechanismus eingeschränkt werden. Um diesen potenziellen Engpass zu vermeiden, bietet der OneFS Multi-Writer eine feiner abgestimmte Schreibsperre, bei der nicht die gesamte Datei gesperrt, sondern die Datei in separate Bereiche unterteilt wird, denen dann exklusive Schreibsperren zugewiesen werden. Auf diese Weise können mehrere Clients gleichzeitig Schreibvorgänge für unterschiedliche Bereiche derselben Datei durchführen.

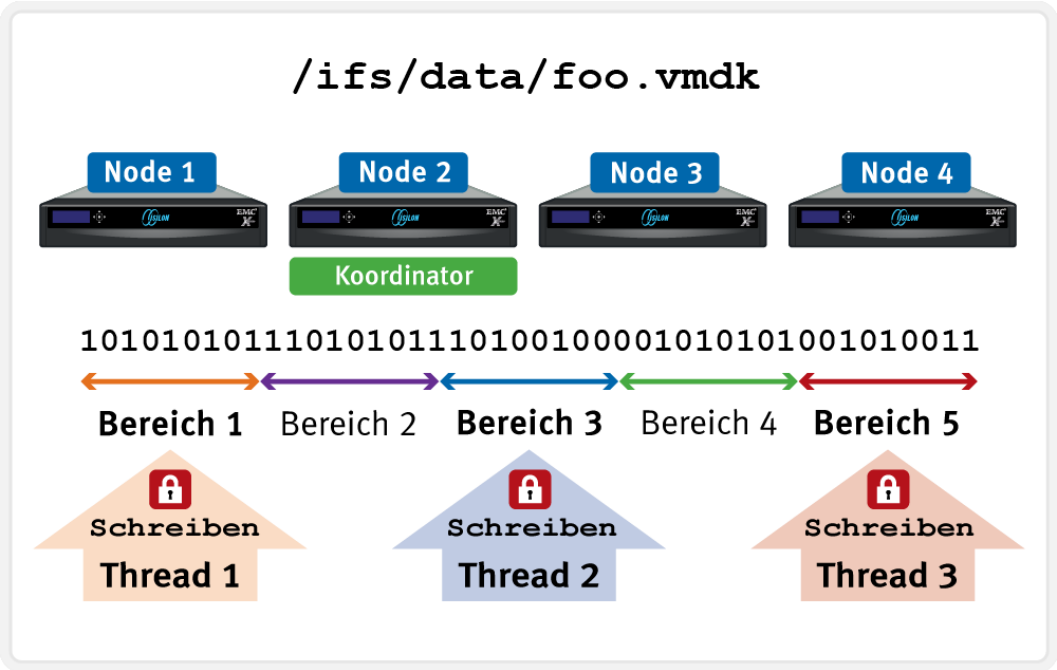

Abbildung 12: I/O-Writer mit Multithreading

Datensicherheit

Stromausfall

Ein Dateisystemjournal, in dem Informationen zu Änderungen am Dateisystem gespeichert werden, ist auf eine schnelle, konsistente Recovery nach Systemausfällen oder Abstürzen ausgelegt, z. B. bei einem Stromausfall. Das Dateisystem gibt die Journaleinträge erneut wieder, wenn ein Node oder Cluster nach einem Stromausfall oder einem anderen Ausfall wiederhergestellt wird. Ohne ein Journal müsste das Dateisystem nach einem Ausfall alle potenziellen Änderungen einzeln untersuchen und überprüfen („fsck"- oder „chkdsk"-Vorgang), was in einem großen Dateisystem viel Zeit in Anspruch nehmen kann.

OneFS ist ein Journaldateisystem, in dem jeder Node eine batteriegestützte NVRAM-Karte enthält, mit der nicht in das Dateisystem übernommene Schreibvorgänge geschützt werden. Die Batterieladung der NVRAM-Karte reicht für mehrere Tage, ohne dass ein Aufladen erforderlich ist. Wenn ein Node gestartet wird, prüft er sein Journal und gibt selektiv Transaktionen an die Festplatten wieder, wenn das Journalsystem dies für erforderlich hält.

OneFS wird nur dann gemountet, wenn sichergestellt werden kann, dass alle noch nicht im System gespeicherten Transaktionen aufgezeichnet wurden. Wenn beispielsweise nicht die richtigen Verfahren zum Herunterfahren eingehalten wurden und die NVRAM-Batterie entladen wurde, sind möglicherweise Transaktionen verloren gegangen. Um potenzielle Probleme zu vermeiden, wird das Dateisystem nicht durch den Node gemountet.

Hardwarefehler und Quorum

Damit das Cluster korrekt funktioniert und Datenschreibvorgänge akzeptieren kann, muss ein Quorum von Nodes aktiv sein und reagieren. Ein Quorum wird als einfache Mehrheit definiert: Ein Cluster mit x Nodes muss $\lfloor x/2 \rfloor+1$ Nodes online haben, um Schreibvorgänge zu ermöglichen. Beispiel: In einem Cluster mit sieben Nodes sind vier Nodes für ein Quorum erforderlich. Wenn ein Node oder eine Gruppe von Nodes aktiv und reaktionsfähig, aber nicht Mitglied eines Quorums ist, wird der Node oder die Gruppe von Nodes in einen schreibgeschützten Status versetzt.

Das System von Isilon verwendet ein Quorum, um „Split-Brain"-Bedingungen zu verhindern, die entstehen können, wenn das Cluster vorübergehend in zwei Cluster aufgeteilt wird. Durch Befolgen der Quorum-Regel sorgt die Architektur dafür, dass immer ein Schreibvorgang stattfindet, unabhängig davon, wie viele Nodes ausfallen oder wieder online sind, und dass der Schreibvorgang auf alle vorherigen Schreibvorgänge abgestimmt werden kann. Das Quorum gibt außerdem die Anzahl der Nodes vor, die zum Umstieg auf ein bestimmtes Datensicherheitslevel erforderlich sind. Für ein auf Löschcodes basierendes Sicherheitslevel $N+M$ muss das Cluster mindestens $2M+1$ Nodes umfassen. Beispielsweise sind für eine Konfiguration N+3 mindestens sieben Nodes erforderlich. So kann bei einem gleichzeitigen Verlust von drei Nodes auch weiterhin das Quorum von vier Nodes aufrechterhalten werden, damit das Cluster voll funktionsfähig bleibt. Wenn ein Cluster unter das Quorum fällt, wird das Dateisystem automatisch in einen gesicherten schreibgeschützten Status versetzt. Schreibvorgänge sind nicht mehr möglich, aber die verfügbaren Daten können nach wie vor gelesen werden.

Hardwarefehler – Hinzufügen/Entfernen von Nodes

Ein System namens Gruppenmanagementprotokoll ermöglicht jederzeit globale Kenntnisse des Clusterstatus und ermöglicht eine konsistente Ansicht des Status aller anderen Nodes über das gesamte Cluster. Wenn ein oder mehrere Nodes im gesamten Cluster-Interconnect nicht erreichbar sind, wird die Gruppe vom Cluster „geteilt" bzw. aus dem Cluster entfernt. Alle Nodes werden in einer neuen konsistenten Ansicht ihres Clusters aufgelöst. (Stellen Sie sich dies so vor, als ob das Cluster in zwei verschiedene Gruppen von Nodes aufgeteilt wird. Beachten Sie aber, dass nur eine Gruppe das Quorum haben kann.) In diesem Teilungsstatus sind alle Daten im Dateisystem zugänglich und für die Seite, die das Quorum aufrechterhält, modifizierbar. Alle auf dem „nachgeschalteten" Gerät gespeicherten Daten werden mithilfe der im Cluster integrierten Redundanz wiederhergestellt.

Wenn der Node wieder verfügbar ist, erfolgt ein „Zusammenführen" oder Hinzufügen, um die Nodes zurück in das Cluster zu bringen. (Die beiden Gruppen werden wieder zu einer Gruppe zusammengeführt.) Der Node kann dem Cluster wieder beitreten, ohne wiederhergestellt und neu konfiguriert zu werden. Dies steht im Gegensatz zu RAID-Arrays für Hardware, bei denen die Laufwerke wiederhergestellt werden müssen. Die AutoBalance-Funktion führt möglicherweise zur Steigerung der Effizienz für bestimmte Dateien ein erneutes Striping durch, wenn einige ihrer Schutzgruppen überschrieben und während der Teilung auf einen niedrigeren Stripe transformiert wurden.

Die OneFS-Job-Engine umfasst auch einen Prozess namens „Collect", der alle verwaisten Blöcke aufsammelt. Wenn ein Cluster während eines Schreibvorgangs geteilt wird, müssen möglicherweise einige der Datei zugewiesenen Blöcke auf der Quorum-Seite neu zugewiesen werden. Dadurch entstehen „verwaiste" zugeordnete Blöcke auf der Seite, die nicht das Quorum hält. Wenn das Cluster wieder zusammengeführt wird, sucht der Job „Collect" diese verwaisten Blöcke über einen parallelen Mark-and-Sweep-Scan, sodass sie als freier Speicherplatz für das Cluster zurückgewonnen werden können.

Skalierbare Wiederherstellung

OneFS nutzt für die Zuweisung oder Wiederherstellung von Daten nach einem Ausfall kein Hardware-RAID. Stattdessen wird der Schutz von File-basierten Daten in OneFS direkt gemanagt. Bei einem Ausfall stellt die Software die Daten auf parallele Weise wieder her. OneFS kann zeitlich konstant bestimmen, welche Dateien von einem Ausfall betroffen sind, indem die Inode-Daten auf lineare Weise direkt von der Festplatte gelesen werden. Der betroffene Dateisatz wird einem Satz Worker Threads zugewiesen, die von der Job-Engine auf die Cluster-Nodes verteilt werden. Die Worker Nodes reparieren die Dateien auf parallele Weise. Mit zunehmender Clustergröße verringert sich somit die Zeit für die Wiederherstellung nach Ausfällen. Dadurch steigert sich die Effizienz enorm, da die Ausfallsicherheit von Clustern auch dann beibehalten wird, wenn ihre Größe ansteigt.

Virtueller Hot Spare

Für die meisten herkömmlichen Speichersysteme, die auf RAID basieren, ist das Provisioning eines oder mehrerer Hot-Spare-Laufwerke erforderlich, um eine unabhängige Recovery fehlgeschlagener Laufwerke zu ermöglichen. Das Hot-Spare-Laufwerk ersetzt das fehlgeschlagene Laufwerk in einem RAID-Set. Wenn diese Hot Spares nicht ersetzt werden, bevor weitere Ausfälle stattfinden, besteht das Risiko eines katastrophalen Datenverlusts im System. OneFS vermeidet die Verwendung von Hot-Spare-Laufwerken und borgt für eine Recovery nach einem Ausfall einfach verfügbaren freien Speicherplatz vom System aus. Diese Methode wird als virtueller Hot Spare bezeichnet. Auf diese Weise wird eine automatische Fehlerkorrektur des Clusters ohne menschliches Eingreifen ermöglicht. Der Administrator kann eine virtuelle Hot-Spare-Reserve schaffen, sodass auch bei laufenden Schreibvorgängen durch die Anwender eine automatische Fehlerkorrektur durchgeführt werden kann.

N+M-Datensicherheit

Ein Isilon-Cluster ist so ausgelegt, dass ein oder mehrere gleichzeitige Komponentenausfälle möglich sind, ohne dass sich dies auf die Datenbereitstellung durch das Cluster auswirkt. Zu diesem Zweck setzt OneFS für den Schutz von Dateien einen Schutz mit Parität, die Reed-Solomon-Fehlerkorrektur (N+M-Schutz) oder ein Spiegelungssystem ein. Die Datensicherheit wird in der Software auf Dateiebene angewendet. So kann sich das System ganz auf die Wiederherstellung der Dateien konzentrieren, die durch den Ausfall betroffen sind, statt eine gesamte Dateigruppe oder ein gesamtes Volume zu prüfen und zu reparieren. OneFS-Metadaten und Inodes werden immer durch Spiegelung und nicht durch die Reed-Solomon-Codierung geschützt, und zwar mit mindestens dem Sicherheitslevel wie die Daten, auf die sie sich beziehen.

Da alle Daten, Metadaten und Paritätsinformationen über die Nodes des Clusters verteilt sind, benötigt das Isilon-Cluster zum Managen von Metadaten keinen dedizierten Paritäts-Node und kein dediziertes Laufwerk bzw. kein dediziertes Gerät oder keinen Satz von Geräten. Damit wird verhindert, dass ein einzelner Node einen Single-Point-of-Failure darstellt. Die auszuführenden Aufgaben werden gleichmäßig auf alle Nodes verteilt, sodass eine perfekte Symmetrie und ein perfekter Lastenausgleich in einer Peer-to-Peer Architektur entsteht.

Das Isilon-System bietet mehrere Level von konfigurierbaren Datensicherheitseinstellungen, die Sie jederzeit ändern können, ohne das Cluster oder Dateisystem offline nehmen zu müssen.

Für eine mit Löschcodes geschützte Datei gilt beispielsweise, dass jede ihrer Schutzgruppen auf Level *N+M/b* geschützt wird, wobei *N* > *M* und *M* >= *b*. Die Werte *N* und *M* stehen jeweils für die Anzahl der Laufwerke, die innerhalb der Schutzgruppe für Daten und für Löschcodes verwendet wird. Der Wert „b" bezieht sich auf die Anzahl der Daten-Stripes, auf die die Schutzgruppe abgelegt wurde, und wird nachfolgend erläutert. Ein häufiger und einfacher Fall besteht darin, das *b = 1* ist, was bedeutet, dass eine Schutzgruppe Folgendes umfasst: *N* Laufwerke an Daten, *M* Laufwerke an Redundanz, in Löschcodes gespeichert, und dass die Schutzgruppe genau über einen Stripe über eine Gruppe von Nodes abgelegt werden soll. Somit können *M* Mitglieder der Schutzgruppe gleichzeitig ausfallen und es wird nach wie vor eine Datenverfügbarkeit von 100 % zur Verfügung gestellt. Die *M* Löschcodemitglieder werden aus den N Datenmitgliedern berechnet. Abbildung 12 zeigt den Vorgang für eine reguläre Schutzgruppe mit 4+2 (*N = 4, M = 2, b = 1).*

Da OneFS-Stripes über Nodes verteilt werden, können Dateien, die an *N+M* verteilt sind, ohne Verlust von Verfügbarkeit 꾀 gleichzeitige Node-Ausfälle überstehen. Somit bietet OneFS Stabilität für beliebige Arten von Ausfällen, ob es sich um ein Laufwerk, einen Node oder eine Komponente innerhalb eines Node handelt (z. B. eine Karte). Darüber hinaus gilt ein Node unabhängig von der Anzahl oder Art von Komponenten, die darin fehlschlagen, als ein einziger Ausfall. Wenn also fünf Laufwerke in einem Node ausfallen, wird dies im Hinblick auf den *N+M*-Schutz lediglich als ein einziger Ausfall gezählt.

OneFS bietet sogar die einzigartige Möglichkeit, M variabel auf einen Wert bis zu vier festzulegen und somit für vierfachen Ausfallschutz zu sorgen. Das geht weit über die heute allgemein verwendete maximale RAID-Stufe hinaus, also der doppelte Ausfallschutz von RAID 6. Da sich die Zuverlässigkeit des Speichers geometrisch mit dieser Redundanzmenge erhöht, kann der N+4-Schutz um Klassen besser als herkömmliche RAID-Hardware sein. Dieser zusätzliche Schutz bedeutet, dass SATA-Laufwerke mit großer Kapazität, wie z. B. 3-TB- und 4-TB-Laufwerke, problemlos hinzugefügt werden können.

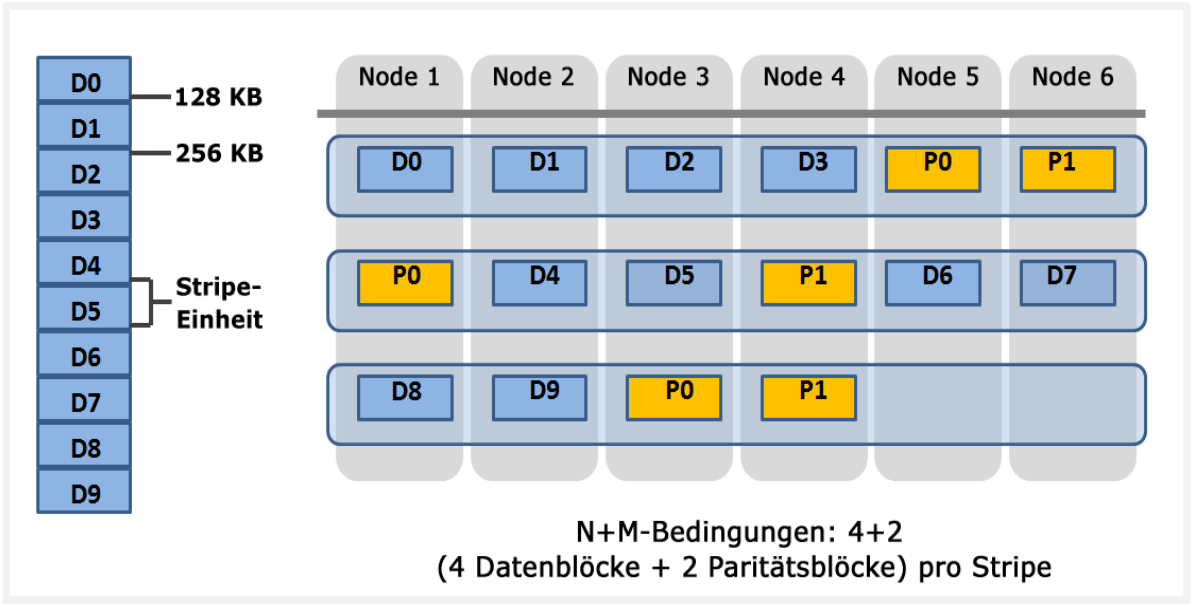

Abbildung 13: OneFS-Redundanz – N+M-Paritätsschutz

Kleinere Cluster können mit N+1-Schutz gesichert werden, was aber Folgendes bedeutet: Ein einziges Laufwerk oder ein Node kann zwar wiederhergestellt werden, zwei Laufwerke in zwei unterschiedlichen Nodes aber nicht. Laufwerksausfälle sind sehr viel wahrscheinlicher als Node-Ausfälle. Für Cluster mit großen Laufwerken ist es empfehlenswert, Schutz für mehrere Laufwerksausfälle einzurichten, obwohl Single-Node-Wiederherstellbarkeit akzeptabel ist.

Um für eine Situation vorzusorgen, bei der doppelte Festplattenredundanz und Single-Node-Redundanz gefordert wird, können Sie Schutzgruppen von „doppelter Länge" erstellen. Diese Schutzgruppen mit doppelter Länge „umhüllen" sozusagen dieselbe Node-Gruppe, wenn sie ausgelegt werden. Da jede Schutzgruppe genau zwei Festplatten für Redundanz enthält, ermöglicht dieser Mechanismus einem Cluster entweder, einen doppelten Laufwerkausfall oder einen vollständigen Node-Ausfall ohne Beeinträchtigung der Datenverfügbarkeit zu überstehen.

Ein wichtiger Punkt bei kleinen Clustern besteht darin, dass diese Striping-Methode äußerst effizient ist, da die Effizienz auf dem Laufwerk $M/(N+M)$ beträgt. Beispiel: Wenn auf einem Cluster mit fünf Nodes und doppeltem Ausfallschutz N = 3, M = 2 verwendet wird, ergibt sich eine Schutzgruppe von 3+2 mit einer Effizienz von *1−2/5* oder 60 %. Wenn aber im selben Cluster mit fünf Nodes jede Schutzgruppe über zwei Stripes abgelegt wird, ist N jetzt 8 und M = 2, sodass eine Effizienz auf der Festplatte von *1-2/(8+2)* oder 80 % erreicht und gleichzeitig der doppelte Ausfallschutz für das Laufwerk erhalten wird und nur eine Schutzfunktion des doppelten Node-Ausfallschutzes geopfert werden muss.

OneFS unterstützt Löschcode-Sicherheitslevel von N+1, N+2, N+3 und N+4 (bei einem Ausfall von bis zu vier Laufwerken oder einem Ausfall aller Nodes pro Cluster) sowie N+2:1 (doppelter Laufwerks- und Single-Node-Ausfallschutz) und N+3:1 (dreifacher Laufwerks- und Single-Node-Ausfallschutz). Darüber hinaus werden für Dateien und Metadaten gespiegelte Sicherheitslevel zwischen 2x und 8x (bis zu achtmal gespiegelt) bereitgestellt. Die Datensicherheit in OneFS ist dazu sehr flexibel, da der Schutz auf einzelne Dateien, Verzeichnisse und deren Inhalte oder das gesamte Dateisystem angewendet werden kann.

OneFS ermöglicht dem Administrator, die Schutz-Policys in Echtzeit zu ändern, während Clients angebunden sind und Daten lesen und schreiben. Hinweis: Durch Erhöhen des Sicherheitslevels eines Clusters erhöht sich möglicherweise die Speichermenge, die durch die Daten auf dem Cluster belegt wird.

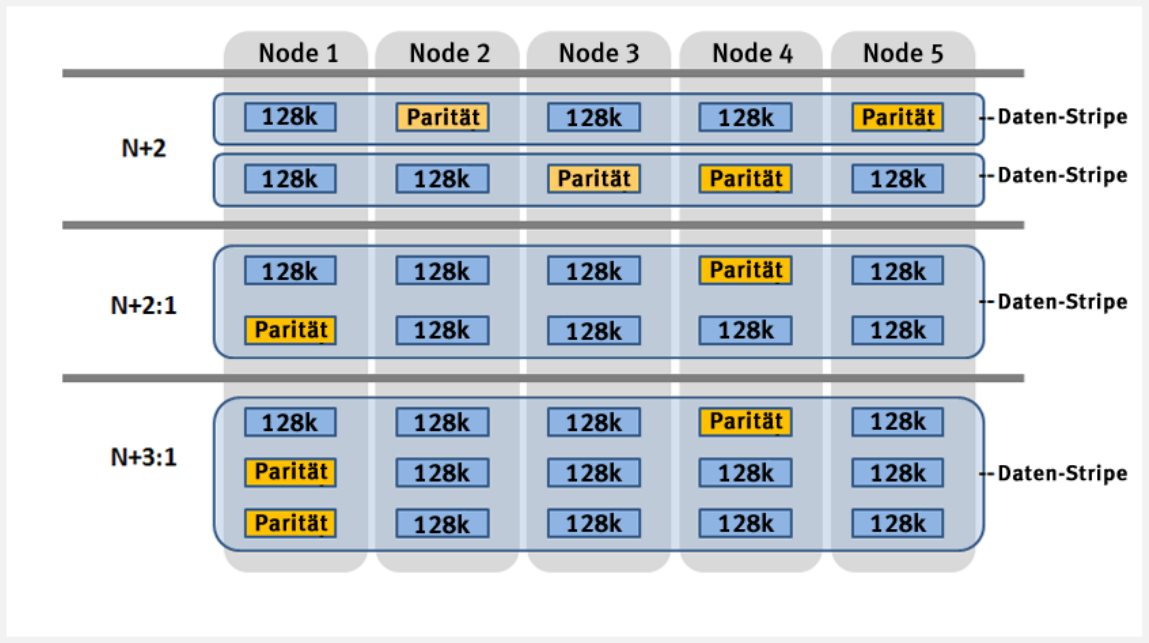

Abbildung 14: Hybride Paritätsschutzmodelle in OneFS (N+M:x)

Automatische Partitionierung

Daten-Tiering und -management in OneFS erfolgen durch das SmartPools-Framework. Im Hinblick auf die Effizienz von Datensicherheit und -layout erleichtert SmartPools die Unterteilung vieler homogener Nodes mit hoher Kapazität in kleinere Laufwerkspools mit einem geringeren MTTDL (Mean Time to Data Loss). Zum Beispiel wird ein NL-Cluster (Nearline) mit 80 Nodes normalerweise mit einem Sicherheitslevel von +4 ausgeführt. Durch eine Partitionierung in vier Laufwerkspools mit jeweils 20 Nodes kann jeder Pool mit +2:1 ausgeführt werden. Dadurch wird der Schutz-Overhead verringert und die Speicherauslastung wird verbessert, ohne dass eine Nettozunahme beim Management-Overhead zu verzeichnen ist.

Um das Ziel einer Vereinfachung des Speichermanagements umzusetzen, berechnet OneFS das Cluster automatisch und partitioniert es in Pools von Laufwerken oder „Node-Pools", die für MTTDL und eine effiziente Speicherauslastung optimiert sind. Daher werden Entscheidungen zum Sicherheitslevel, wie im obigen Beispiel mit dem Cluster mit 80 Nodes angeführt, nicht dem Kunden überlassen, es sei denn, dies wird gewünscht.

Dank automatischem Provisioning wird jede Gruppe entsprechender Node-Hardware automatisch in Node-Pools unterteilt, die bis zu 40 Nodes und sechs Laufwerke pro Node enthalten. Diese Node-Pools werden standardmäßig mit +2:1 geschützt und mehrere Pools können dann in logischen Tiers kombiniert und anhand der Dateipool-Policies von SmartPools verwaltet werden. Durch die Unterteilung der Laufwerke eines Node in mehrere, separat geschützte Pools sind die Nodes bei mehreren Festplattenausfällen sehr viel robuster, als dies bisher möglich war.

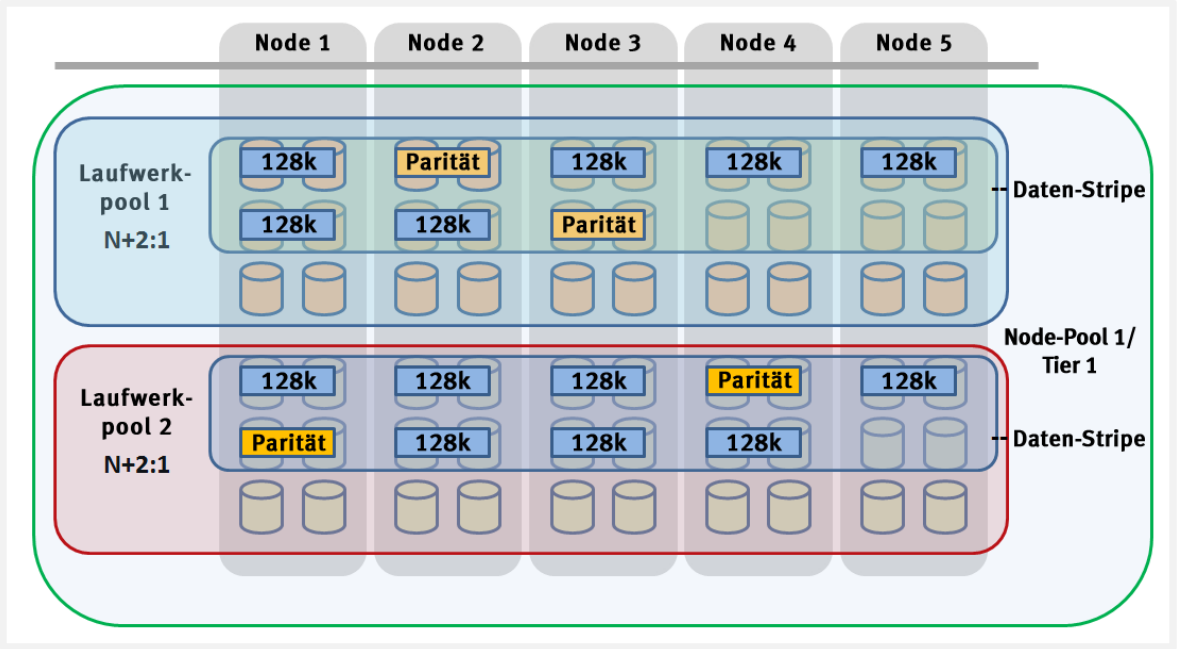

Abbildung 15: Automatische Partitionierung mit SmartPools

Unterstützte Protokolle

Clients mit den richtigen Anmeldedaten und Berechtigungen können Daten mithilfe einer der unterstützten Standardmethoden für die Kommunikation mit dem Cluster erstellen, ändern und lesen:

- NFS (Network File System)
- SMB/CIFS (Server Message Block/Common Internet File System) einschließlich Support für SMB3 Multi-Channel
- FTP (File Transfer Protocol)
- HTTP (Hypertext Transfer Protocol)
- HDFS (Hadoop Distributed File System)
- REST API (Representational State Transfer Application Programming Interface)

Standardmäßig sind nur die SMB-/CIFS- und NFS-Protokolle im Isilon-Cluster aktiviert. Das Stammverzeichnis des Dateisystems für alle Daten im Cluster ist **/ifs** (das Isilon OneFS-Dateisystem). Dies wird über das SMB-/CIFS-Protokoll als eine „ifs"-Share (\\<*cluster_name**ifs*) und über das NFS-Protokoll als ein „/ifs"-Export (*<cluster_name>:/ifs)* dargestellt.

Hinweis: Die Daten sind allen Protokollen gemein, sodass Änderungen am Dateiinhalt über ein Zugriffsprotokoll sofort von allen anderen Protokollen sichtbar sind.

Dynamische Skalierung/Skalierung nach Bedarf

Performance und Kapazität

Im Gegensatz zu herkömmlichen Speichersystemen, die „hochskaliert" werden müssen, wenn zusätzliche Performance oder Kapazität erforderlich ist, ermöglicht OneFS einem Isilon-Speichersystem ein „Scale-out" und damit die nahtlose Erweiterung des vorhandenen Dateisystems oder Volume auf eine Kapazität mit mehreren Petabyte. Gleichzeitig wird die Performance linear erhöht.

Das Hinzufügen von Kapazität und Performance zu einem Isilon-Cluster ist erheblich einfacher als bei anderen Speichersystemen. Der Speicheradministrator braucht nur drei einfache Schritte durchzuführen: einen weiteren Node im Rack hinzufügen, den Node an das InfiniBand-Netzwerk anbinden und das Cluster anweisen, den zusätzlichen Node hinzuzufügen. Der neue Node bietet zusätzliche Kapazität und Performance, da jeder Node CPU, Speicher, Cache, Netzwerk, NVRAM und I/O-Kontrollpfade umfasst.

Die in OneFS integrierte AutoBalance-Funktion verschiebt Daten automatisch und auf kohärente Weise über das InfiniBand-Netzwerk, sodass auf dem Cluster vorhandene Daten in diesen neuen Speicher-Node verschoben werden. Dank des automatischen Ausgleichs wird dafür gesorgt, dass der neue Node kein Hot Spot für neue Daten wird und dass vorhandene Daten von den Vorteilen eines leistungsfähigeren Speichersystems profitieren können. Die AutoBalance-Funktion in OneFS ist auch für den Anwender vollständig transparent und kann so eingestellt werden, dass die Auswirkung auf leistungsfähige Workloads minimiert wird. Diese Funktion allein ermöglicht eine transparente Skalierung in OneFS von 18 TB auf 20 PB bei laufendem Betrieb, ohne zusätzliche Managementzeit für den Administrator oder erhöhte Komplexität innerhalb des Speichersystems.

Ein umfangreiches Speichersystem muss die für verschiedene Workflows erforderliche Performance bereitstellen, unabhängig davon, ob diese Workflows sequenziell, gleichzeitig oder zufällig sind. Zwischen und innerhalb einzelner Anwendungen existieren verschiedene Workflows. OneFS erfüllt dank intelligenter Software all diese Anforderungen gleichzeitig. Noch wichtiger ist aber, dass in OneFS Durchsatz und IOPS linear mit der Anzahl der in einem einzigen System vorhandenen Nodes skaliert werden. Aufgrund der ausgeglichenen Datenverteilung, des automatischen Ausgleichs und der dezentralen Verarbeitung können Sie mit OneFS weitere CPUs, Netzwerkports und Arbeitsspeicher bei der Skalierung des Systems nutzen.

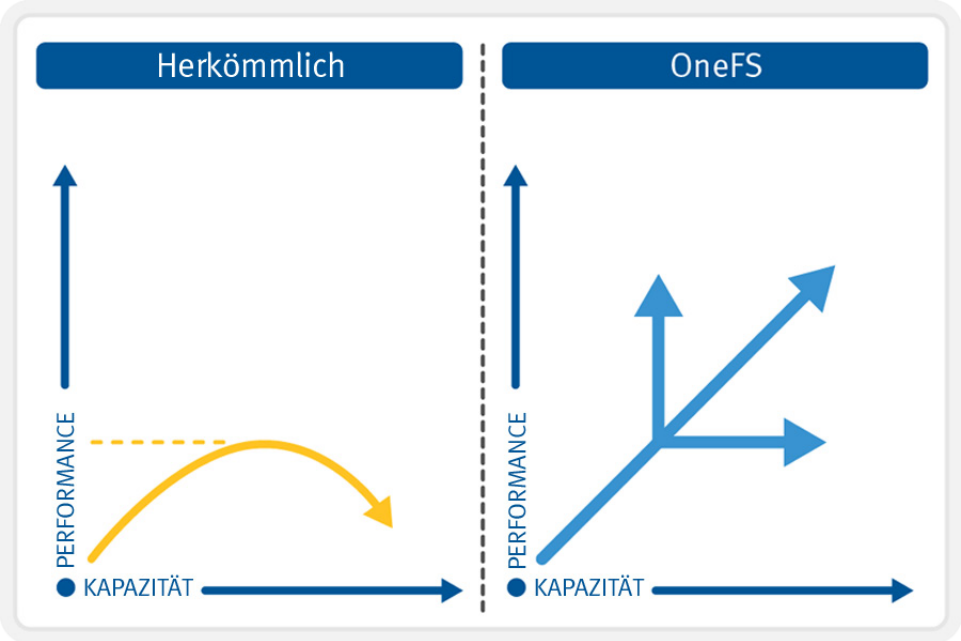

Abbildung 16: Lineare Skalierbarkeit in OneFS

Schnittstellen

Administratoren können zur Verwaltung eines Isilon-Speicherclusters in ihrer Umgebung mehrere Schnittstellen verwenden:

- Webverwaltungsschnittstelle (WebUI)
- Befehlszeilenoberfläche über SSH-Netzwerkzugriff oder einen seriellen RS232-Anschluss
- LCD-Bereich auf den Nodes selbst für einfache Funktionen zum Hinzufügen/Entfernen
- RESTful-Plattform-API zur programmatischen Steuerung und Automatisierung von Clusterkonfiguration und -management

Authentifizierung und Zugriffskontrolle

Authentifizierungsservices stellen eine Sicherheitsmaßnahme dar, da sie die Anmeldedaten von Benutzern überprüfen, bevor diese auf Dateien zugreifen und Dateien verändern können. OneFS unterstützt vier Methoden für die Authentifizierung von Benutzern:

- Active Directory (AD)
- LDAP (Lightweight Directory Access Protocol)
- NIS (Netzwerkinformationsservice)
- Lokale Benutzer und Gruppen

OneFS unterstützt die Verwendung mehrerer Authentifizierungstypen. Es wird jedoch empfohlen, dass Sie sich mit den Interaktionen zwischen Authentifizierungstypen gründlich vertraut machen, bevor Sie mehrere Methoden für das Cluster aktivieren. Detaillierte Informationen zur richtigen Konfiguration mehrerer Authentifizierungsmodi finden Sie in der Produktdokumentation.

Active Directory

Active Directory, eine LDAP-Implementierung von Microsoft, ist ein Verzeichnisservice, der Informationen über die Netzwerkressourcen speichern kann. Active Directory bietet viele Funktionen, aber der Hauptgrund für die Verbindung des Clusters mit der Domain besteht in der Authentifizierung von Benutzern und Gruppen.

Sie können die Active Directory-Einstellungen eines Clusters über die Webverwaltungsschnittstelle oder die Befehlszeilenoberfläche konfigurieren und verwalten. Es wird jedoch empfohlen, nach Möglichkeit die Webverwaltungsschnittstelle zu verwenden.

Jeder Node im Cluster verwendet dasselbe Active Directory-Rechnerkonto, wodurch die Verwaltung sehr vereinfacht wird.

LDAP

Das LDAP (Lightweight Directory Access Protocol) ist ein Netzwerkprotokoll zum Definieren, Abfragen und Ändern von Services und Ressourcen. Der Hauptvorteil von LDAP besteht in der offenen Gestaltung der Verzeichnisservices und der Fähigkeit, LDAP plattformübergreifend zu verwenden. Mithilfe von LDAP kann das Clusterspeichersystem von Isilon Benutzer und Gruppen authentifizieren, bevor diesen Zugriff auf das Cluster gewährt wird.

NIS

NIS (Network Information Service) von Sun Microsystems ist ein Protokoll für Verzeichnisservices, mit dem das Isilon-System Benutzer und Gruppen, die auf das Cluster zugreifen, authentifizieren kann. NIS, auch als „Gelbe Seiten" bezeichnet, unterscheidet sich vom Protokoll NIS+, das vom Isilon-Cluster nicht unterstützt wird.

Lokale Benutzer

Das Clusterspeichersystem von Isilon unterstützt die Authentifizierung lokaler Benutzer und Gruppen. Sie können Konten für lokale Benutzer und Gruppen direkt mithilfe der WebUI auf dem Cluster erstellen. Die lokale Authentifizierung kann nützlich sein, wenn Verzeichnisservices – Active Directory, LDAP oder NIS – nicht verwendet werden, oder wenn ein bestimmter Benutzer bzw. eine bestimmte Anwendung Zugriff auf das Cluster benötigt.

Zugriffszonen

Zugriffszonen bieten eine Methode zur logischen Partitionierung des Clusterzugriffs und zum Zuweisen von Ressourcen zu eigenständigen Einheiten, wodurch eine gemeinsam genutzte Mandantenumgebung bereitgestellt wird. Zur Vereinfachung dieses Vorgangs werden die drei wichtigsten externen Zugriffskomponenten in Zugriffszonen zusammengefasst:

- Clusternetzwerkkonfiguration
- Dateiprotokollzugriff
- Authentifizierung

Daher sind Isilon SmartConnect™-Zonen mit einer Reihe von SMB-/CIFS-Shares und einem oder mehreren Authentifizierungsanbietern zur Zugriffskontrolle verknüpft. Überlappende Shares sind zulässig und ermöglichen das zentrale Management eines einzelnen Stammverzeichnis-Namespace, werden aber für mehrere Mandanten bereitgestellt und gesichert. Dies ist besonders in Unternehmensumgebungen nützlich, in denen mehrere separate Geschäftsbereiche über eine zentrale IT-Abteilung bedient werden. Ein weiteres Beispiel: Während eines Projekts zur Serverkonsolidierung werden mehrere Windows-Dateiserver zusammengeführt, die mit separaten, nicht vertrauenswürdigen Active Directory-Strukturen verbunden sind.

Beim Einsatz von Zugriffszonen umfasst die integrierte Systemzugriffszone standardmäßig eine Instanz jedes unterstützten Authentifizierungsanbieters, alle verfügbaren SMB-Shares und alle verfügbaren NFS-Exporte.

Diese Authentifizierungsanbieter können mehrere Instanzen von Microsoft Active Directory, LDAP, NIS und lokale Benutzer- oder Gruppendatenbanken umfassen.

Rollenbasierte Administration

Bei der rollenbasierten Administration handelt es sich um ein Zugriffskontrollsystem auf Grundlage von Clustermanagementrollen (Roles Based Access Control, RBAC), das die Berechtigungen von „Root"- und „Administratorbenutzern" in detailliertere Berechtigungen aufteilt und die Zuweisung dieser Berechtigungen zu bestimmten Rollen ermöglicht. Diese Rollen können dann anderen, nicht privilegierten Benutzern zugeteilt werden. So kann beispielsweise Rechenzentrumsmitarbeitern Lesezugriff für das gesamte Cluster zugewiesen werden, wodurch sie umfassenden Monitoringzugriff erhalten, aber keine Konfigurationsänderungen vornehmen können. OneFS bietet eine integrierte Sammlung von Rollen, einschließlich Audit, System- und Sicherheitsadministratoren sowie die Möglichkeit, benutzerdefinierte Rollen zu erstellen. Die rollenbasierte Administration ist in die OneFS-Befehlszeilenoberfläche, die WebUI und die Plattform-API integriert.

Softwareupgrade

Durch die Durchführung eines Upgrades auf die neueste Version von OneFS können Sie alle neuen Funktionen, Fehlerkorrekturen und Merkmale auf dem Isilon-Cluster nutzen. Cluster können auf zweierlei Weise aktualisiert werden: Simultanes oder fortlaufendes Upgrade

Simultanes Upgrade

Bei einem simultanen Upgrade wird das neue Betriebssystem installiert. Ferner werden alle Nodes im Cluster gleichzeitig neu gestartet. Für ein simultanes Upgrade muss der Service bei Neustart der Nodes vorübergehend für einen Zeitraum von weniger als zwei Minuten unterbrochen werden.

Fortlaufendes Upgrade

Bei einem fortlaufenden Upgrade wird das Upgrade einzeln durchgeführt und die Nodes im Cluster werden nacheinander neu gestartet. Während eines fortlaufenden Upgrades bleibt das Cluster online und stellt Clients weiterhin ohne Serviceunterbrechungen Daten bereit. Ein fortlaufendes Upgrade kann nur innerhalb einer OneFS-Produktreihe mit einer Codeversion und nicht zwischen unterschiedlichen Hauptrevisionen der Codeversion von OneFS durchgeführt werden.

Durchführen des Upgrades

Isilon empfiehlt dringend, vor dem Upgrade ein installationsvorbereitendes Verifizierungsskript auszuführen. Mit diesem Skript wird überprüft, ob die Konfiguration in der aktuellen Installation von OneFS mit der Version von OneFS kompatibel ist, für die das Upgrade erfolgt. Ein Patch, der das Skript enthält, ist beim Isilon Customer Service erhältlich.

Während eines Upgrades wird außerdem eine installationsvorbereitende Upgradeprüfung durchgeführt, um dafür zu sorgen, dass nur eine unterstützte Konfiguration für das Upgrade zugelassen wird. Wenn eine nicht unterstützte Konfiguration gefunden wird, wird das Upgrade angehalten und Troubleshooting-Anweisungen werden angezeigt. Eine installationsvorbereitende Upgradeprüfung vor dem Start des Upgrades trägt dazu bei, Unterbrechungen aufgrund von nicht kompatiblen Konfigurationen zu vermeiden.

Wenn der Administrator das Upgradepaket vom Isilon Customer Service heruntergeladen hat, kann das Upgrade entweder über die Befehlszeilenoberfläche oder die Webverwaltungsschnittstelle durchgeführt werden. Eine Überprüfung des Upgrades kann über eine der beiden Schnittstellen erfolgen, indem Sie nach erfolgreicher Durchführung des Upgrades den Integritätsstatus des gesamten Clusters prüfen.

EMC Isilon-Software für Datensicherheit und -management

Isilon bietet ein umfassendes, auf Ihre Anforderungen abgestimmtes Softwareportfolio für Datensicherheit- und management an:

InsightIQ™	Performancemanagement	Maximiert die Performance Ihres Isilon-Scale-out-Speichersystems mit innovativen Tools für Performancemonitoring und -reporting
SmartPools™	Ressourcenmanagement	Implementiert eine hocheffiziente, automatisierte Tiered-Storage-Strategie zur Optimierung der Speicherperformance und -kosten
SmartQuotas™	Datenmanagement	Managt und weist Quotas zu für eine nahtlose Partitionierung und Thin Provisioning von Speicher in einfach zu managende Segmente auf Cluster-, Verzeichnis-, Unterverzeichnis-, Benutzer- und Gruppenebene
SmartConnect™	Datenzugriff	Ermöglicht einem Lastenausgleich für Clientverbindungen und ein dynamisches NFS-Failover und -Failback von Clientverbindungen zwischen Speicher-Nodes zur Optimierung der Nutzung von Clusterressourcen
SnapshotIQ™	Datensicherheit	Effizienter und zuverlässiger Schutz Ihrer Daten mit sicheren, nahezu sofortigen Snapshots bei geringem bis gar keinem Performance-Overhead Beschleunigte Recovery wichtiger Daten mit nahezu sofortigen Snapshot-Wiederherstellungen nach Bedarf
Isilon für vCenter	Datenmanagement	Management von Isilon-Funktionen im vCenter
SyncIQ™	Datenreplikation	Asynchrone Replikation und Verteilung großer, geschäftskritischer Datasets an mehrere freigegebene Speichersysteme an mehreren Standorten für eine zuverlässige Disaster-Recovery-Fähigkeit Einfaches Failover und Failback mit einem Tastendruck zur erhöhten Verfügbarkeit geschäftskritischer Daten
SmartLock™	Datenaufbewahrung	Schutz Ihrer wichtigen Daten vor versehentlichem, vorzeitigem oder böswilligem Ändern oder Löschen mit unserem softwarebasierten WORM-Ansatz (Write Once Read Many) und Einhaltung strenger Compliance- und Governance-Anforderungen wie SEC 17a-4
SmartDedupe™	Datendeduplizierung	Maximierte Speichereffizienz durch Scannen des Clusters auf identische Blöcke und nachfolgendes Eliminieren von Duplikaten, wodurch die Menge an erforderlichen physischen Speichermedien verringert wird

Weitere Informationen zu allen oben genannten Isilon-Softwareprodukten erhalten Sie in der Produktdokumentation.

Fazit

Mit OneFS können Unternehmen und Administratoren in einem einzigen Dateisystem und einem Volume über eine einzige zentrale Administration eine Skalierung von 18 TB bis auf 20 PB vornehmen. OneFS bietet hohe Performance und hohen Durchsatz oder beides, und das ohne zusätzliche Komplexität beim Management.

Rechenzentren der nächsten Generation müssen auf nachhaltige Skalierbarkeit ausgelegt werden. Sie müssen die Vorteile der Automatisierung und Kommerzialisierung von Hardware nutzen, die vollständige Auslastung der Netzwerk-Fabric ermöglichen und maximale Flexibilität für Unternehmen bieten, die darauf bedacht sind, sich ständig ändernde Anforderungen erfüllen zu können.

OneFS ist das Dateisystem der Zukunft, mit dem sich alle diese Herausforderungen meistern lassen. OneFS bietet die folgenden Vorteile:

- Vollständig verteiltes, einziges Dateisystem
- Leistungsfähiges, vollständig symmetrisches Cluster
- Datei-Striping über alle Nodes in einem Cluster hinweg
- Weniger Komplexität dank automatisierter Software
- Dynamische Inhaltsverteilung
- Flexibler Schutz von Daten
- Hohe Verfügbarkeit
- Webbasierte und Befehlszeilenadministration

OneFS ist ideal für File-basierte und unstrukturierte Big Data-Anwendungen in Unternehmensumgebungen – wie umfangreiche Stammverzeichnisse, Dateifreigaben, Archive, Virtualisierung und Geschäftsanalysen – sowie für vielfältige datenintensive und leistungsfähige Computing-Umgebungen geeignet, z. B. Energieversorgung, Finanzdienstleistungen, Internet- und Hostingservices, Business Intelligence, Technik, Fertigung, Medien und Unterhaltung, Bioinformatik und wissenschaftliche Forschung.

Über EMC

Die EMC Corporation ermöglicht Unternehmen und Serviceprovidern, ihre Geschäftsprozesse zu verändern und IT as a Service bereitzustellen. Cloud-Computing ist ein entscheidender Faktor für diese Transformation. Durch innovative Produkte und Services beschleunigt EMC die Nutzung von Cloud-Computing und hilft IT-Abteilungen, ihre wertvollsten Ressourcen, nämlich Daten, auf dynamischere, zuverlässigere und kostengünstigere Weise zu speichern, zu verwalten, zu schützen und zu analysieren. Zusätzliche Informationen zu EMC finden Sie unter http://germany.emc.com.