DELL Technologies

ホワイト ペーパー

DELL EMC POWERSCALE ONEFS：技術概要

要約

このホワイト ペーパーでは、Dell EMC PowerScale スケールアウト NAS ストレージ ソリューションに搭載された OneFS オペレーティング システムの主な特長と機能に関する技術的な詳細情報を提供します。

2021 年 9 月

リビジョン

Version	日付	コメント
1.0	2013年11月	OneFS 7.1のイニシャル リリース
2.0	2014年6月	OneFS 7.1.1に伴うアップデート
3.0	2014年11月	OneFS 7.2に伴うアップデート
4.0	2015年6月	OneFS 7.2.1に伴うアップデート
5.0	2015年11月	OneFS 8.0に伴うアップデート
6.0	2016年9月	OneFS 8.0.1に伴うアップデート
7.0	2017年4月	OneFS 8.1に伴うアップデート
8.0	2017年11月	OneFS 8.1.1に伴うアップデート
9.0	2019年2月	OneFS 8.1.3に伴うアップデート
10.0	2019年4月	OneFS 8.2に伴うアップデート
11.0	2019年8月	OneFS 8.2.1に伴うアップデート
12.0	2019年12月	OneFS 8.2.2に伴うアップデート
13.0	2020年6月	OneFS 9.0に伴うアップデート
14.0	2020年9月	OneFS 9.1に伴うアップデート
15.0	2021年4月	OneFS 9.2に伴うアップデート
16.0	2021年9月	OneFS 9.3に伴うアップデート

謝辞

このホワイトペーパーは、次の作成者によって作成されました。

著者： Nick Trimbee

目次

はじめに

従来のストレージ モデルの3つのレイヤー（ファイル システム、ボリューム マネージャー、データ保護）は、小規模ストレージ アーキテクチャのニーズに応えるように進化してきた一方で、ペタバイト規模の非常に複雑なシステムには適応しません。OneFSオペレーティング システムはこれらに代わるストレージ モデルとして、拡張性の高いデータ保護が組み込まれたクラスター化ファイル システムを一元化し、ボリューム管理の必要性を排除しました。OneFSはスケールアウト インフラストラクチャの基盤となる構成単位で、非常に高い拡張性と効率性を実現し、すべてのDell EMC PowerScale NASストレージ ソリューションに搭載されています。

重要なこととして、OneFSは機械だけでなく、人的コストの部分も改善するように設計されています。OneFSにより、従来のストレージ システムで必要であった人的コストのほんの一部で大規模システムを管理できます。OneFSは複雑さを排除し、自動修復と自動管理機能を組み込むことで、ストレージ管理の負担を大幅に減らしています。また、OneFSはOSの非常に深いレベルで並列化を実現し、ほぼすべての主要なシステム サービスがハードウェアの複数のユニットに分散されています。これにより、OneFSはインフラストラクチャの拡大に合わせてほぼすべての領域を拡張し、現在から将来にわたるデータセットの増加にも対応します。

OneFSは完全に対称化されたファイル システムで、単一障害点がありません。クラスタリングを活用して、パフォーマンスと容量の拡張だけでなく、RAIDをはるかに超える「any-to-any」のフェイルオーバーと複数レベルの冗長性を実現します。ディスク サブシステムの傾向として、パフォーマンスの向上がゆっくりであるのに対し、ストレージ密度の増加は速いペースで進んでいます。OneFSはこの現状に対応し、冗長性の量を拡張する一方で、障害の修復速度を向上させました。これにより、OneFSは数ペタバイトの拡張性を提供する一方で、従来の小規模なストレージ システムより高い信頼性を実現します。

PowerScaleハードウェアは、OneFSが実行されるアプライアンスを提供します。ハードウェア コンポーネントは最適な組み合わせのものである一方でコモディティー ベースであることから、日々改善を続けるコモディティー ハードウェアのコスト曲線と効率曲線のメリットを得られます。OneFSはハードウェアをクラスターに自由に組み込むことも取り外すこともできるため、ハードウェアからデータやアプリケーションを分離することができます。データの寿命に限りはなく、進化するハードウェアの世代交代の影響を受けません。このため、データ移行やハードウェアのリフレッシュに伴うコストや苦痛を取り除きます。

OneFSは、大規模ホーム ディレクトリー、ファイル共有、アーカイブ、仮想化、ビジネス分析などの、エンタープライズ環境の非構造化「Big Data」ファイル ベース アプリケーションに最適です。このため、OneFSはエネルギー、金融サービス、インターネットおよびホスティング サービス、ビジネス インテリジェンス、エンジニアリング、製造、メディアおよびエンターテインメント、バイオインフォマティクス、科学研究、その他の高いパフォーマンスの処理が要求される分野など、大量のデータを扱う業界で広く採用されています。

対象読者

このホワイト ペーパーでは、Dell EMC PowerScaleクラスターの導入と管理に関する情報と、OneFSアーキテクチャの包括的な背景情報を示します。

このホワイト ペーパーの対象読者は、PowerScaleのクラスター化されたストレージ環境を構成および管理するすべてのユーザーです。ストレージ、ネットワーキング、オペレーティング システム、データ管理についての基本的な知識が読者にあることが前提となっています。

📖 OneFSのコマンドおよび機能の構成に関する詳細については、OneFS管理ガイドを参照してください。

OneFSの概要

OneFSは、従来のストレージ アーキテクチャの3つのレイヤー（ファイル システム、ボリューム マネージャー、データ保護）を1つの統合されたソフトウェア レイヤーに集約することで、OneFS搭載ストレージ クラスターで動作する単一のインテリジェントな分散型ファイル システムを構築します。

図 1：OneFS は、ファイル システム、ボリューム マネージャー、データ保護を単一のインテリジェントな分散システムに集約します。

これは、企業において現在の環境でスケールアウトNASを直接利用するうえで重要な、革新的な機能です。スケールアウトの原則であるインテリジェント ソフトウェア、コモディティ ハードウェア、分散アーキテクチャに忠実に従った形となっています。OneFSはオペレーティング システムであるだけでなく、クラスターにデータを処理して格納する基本ファイル システムでもあります。

PowerScaleノード

OneFSは、「クラスター」と呼ばれる専用のプラットフォーム ノード上のみで動作します。1つのクラスターは複数のノードで構成されます。各ノードは、メモリー、CPU、ネットワーク、Ethernetまたは低レイテンシーInfiniBandインターコネクト、ディスク コントローラー、ストレージ メディアを搭載したラックマウント可能なエンタープライズ アプライアンスです。このため、分散クラスター内の各ノードはコンピューティング機能のほか、ストレージまたは容量の機能も持ち合わせています。

Gen6アーキテクチャでは、OneFS 8.2以降で最大252ノードにスケールアップするクラスターを作成するには、4RU（ラック ユニット）フォーム ファクターの4ノードの単一シャーシが必要です。個々のノード プラットフォームでは、クラスターを形成するために少なくとも3ノードと3RUのラック スペースが必要です。種類の異なるノードもすべて1つのクラスターに組み込むことができるため、各種ノードによりさまざまな容量比率のスループット（IOPS：1秒あたりのI/O動作数）を実現できます。従来のGen6シャーシとPowerScaleオールフラッシュF900、F600、F200スタンドアロン ノードの両方が、同じクラスター内で問題なく共存します。

クラスターに追加される各ノードまたはシャーシにより、ディスク、キャッシュ、CPU、ネットワーク容量の総量が増加します。OneFSは各ハードウェアの構成単位を活用することで、全体で部分の合計よりも高い能力を実現しています。RAMが1つの一貫性のあるキャッシュに集約されているため、クラスターのどの部分に対してもI/Oが可能で、どこからでもキャッシュされたデータを利用できます。ファイル システム ジャーナルは、電源障害が発生した場合でも書き込みが安全であることを保証します。スピンドルとCPUを結合することで、クラスターの拡大に応じて、1ファイルまたは複数ファイルのアクセスにおけるスループット、容量、IOPSを高めています。クラスターのストレージ容量は数十TBから数十PBの範囲です。ストレージ メディアとノード シャーシの高密度化が続くにつれ、最大容量の増加は今後も続きます。

OneFS搭載のプラットフォーム ノードは、機能に応じていくつかのクラス、つまり階層に分類されます。

Tier	I/O Profile	Drive Media	Nodes
Performance	High Perf, Low Latency	Flash NVMe/SAS	F900 F810 F600 F800 F200
Hybrid / Utility	Concurrency & Streaming Throughput	SATA/SAS & SSD	H700 H600 H7000 H5600 H500 H400
Archive	Nearline & Deep Archive	SATA	A300 A200 A3000 A2000

表 1：ハードウェア階層とノード タイプ

ネットワーク

クラスターに関連づけられているネットワークには、内部ネットワークと外部ネットワークの2つがあります。

バックエンド ネットワーク

クラスター内のノード間通信はすべて、10、40、または100 GbのいずれかのEthernet、つまり低レイテンシーQDR InfiniBand (IB)で構成される専用のバックエンド ネットワークを介して実行されます。このバックエンド ネットワークは、高可用性を実現するために冗長スイッチで構成されており、クラスターのバックプレーンとして機能します。これにより、各ノードはクラスター内のコントリビューターとして機能し、ノード間通信は高速で低レイテンシーの専用ネットワークに分離されます。このバックエンド ネットワークでは、ノード間通信にインターネット プロトコル（IP）を使用します。

フロントエンド ネットワーク

クライアントはすべてのノードで利用可能なEthernet接続（10GbE、25GbE、40GbE、100GbEのいずれか）を使用してクラスターに接続します。各ノードではそれぞれのEthernetポートを使用できるため、クラスターで使用可能なネットワーク帯域幅の量はパフォーマンスと容量に比例して変化します。クラスターは、NFS、SMB、HTTP、FTP、HDFS、S3など、お客様ネットワークへの標準ネットワーク通信プロトコルをサポートしています。さらにOneFSは、IPv4環境とIPv6環境の両方と完全に統合します。

完全なクラスターの図

完全なクラスターは、次の図のようにハードウェア、ソフトウェア、ネットワークと組み合わされます。

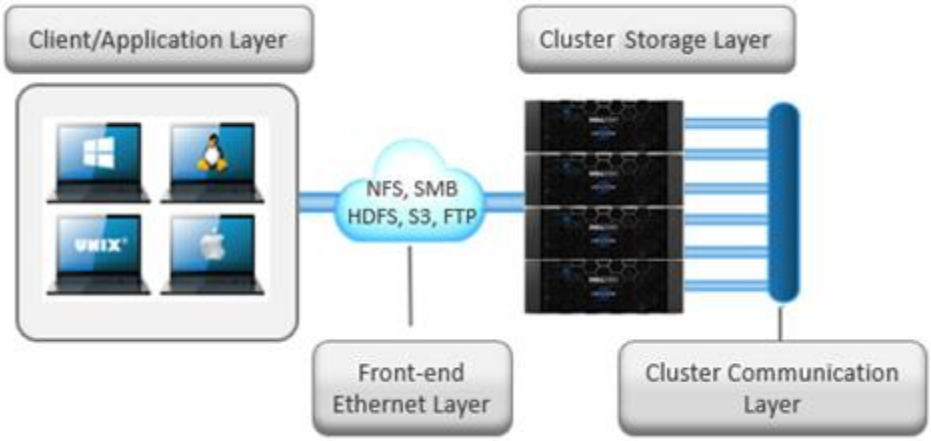

図 2：OneFS の稼働中の全コンポーネント

上の図は、ソフトウェア、ハードウェア、ネットワークのすべてが環境内で連携している完全なアーキテクチャを示しています。サーバーが提供する完全に分散化された単一のファイル システムは、スケールアウト環境でのワークロードや容量のニーズまたはスループットのニーズの変化に応じて動的に拡張可能です。

OneFS SmartConnectは、フロントエンドEthernetレイヤーで動作するロード バランサーであり、クラスター全体でクライアント接続を均等に分散させます。SmartConnectは、LinuxおよびUNIXクライアントの場合は動的なNFSフェールオーバーとフェールバックをサポートし、Windowsクライアントの場合はSMB3の継続的な可用性をサポートします。これにより、ノードに障害が発生した場合または予防メンテナンスが実行された場合に、処理中のすべての読み取り/書き込みをクラスターの別のノードに引き渡し、ユーザーやアプリケーションを中断することなく操作を完了できます。

フェイルオーバー中、クライアントはパフォーマンス インパクトを最小限に抑えて、クラスター内の残りのすべてのノードに均等に再分配されます。ノードが、障害を含めて何らかの理由でシャットダウンされると、そのノードの仮想IPが、クラスター内にある別のノードにシームレスに移行されます。オフライン ノードがオンラインに戻ると、SmartConnectは、ストレージとパフォーマンスの使用率を最大限に確保するために、クラスター全体のNFSクライアントとSMB3クライアントを自動的に再調整します。定期的なシステム保守やソフトウェア アップデートの場合、この機能を使用するとノードごとのローリング アップグレードが可能になるため、保守期間全体で可用性が完全に確保されます。

📖 詳細については、OneFS SmartConnectに関するホワイト ペーパーを参照してください。

OneFSソフトウェアの概要

オペレーティング システム

OneFSは、BSDベースのUNIX OS（オペレーティング システム）基盤に基づいています。ハード リンク、delete-on-close、アトミックなリネーム、ACL、拡張属性など、Linux/UNIXとWindowsの両方のセマンティクスがネイティブでサポートされています。ベースOSとしてBSDを使用した理由は、これが成熟した実証済みのオペレーティング システムであり、技術革新に向けてオープン ソース コミュニティを活用できるためです。OneFS 8.2以降では、基盤となるOSのバージョンはFreeBSD 11です。

クライアント サービス

クライアントがOneFSとの通信に使用できるフロントエンド プロトコルは、クライアント サービスと呼ばれます。サポートされているプロトコルの詳細な一覧については、サポートされているプロトコルに関するセクションを参照してください。OneFSがクライアントと通信する仕組みについて理解するために、I/Oサブシステムを2つに分けます。つまり、上半分の「イニシエーター」と下半分の「参加者」です。クラスター内の各ノードは、特定のI/O動作の参加者です。クライアントが接続するノードはイニシエーターであり、このノードはI/O動作全体の「キャプテン」として機能します。読み取りおよび書き込み処理については、後で詳しく説明します。

クラスター操作

クラスター化されたアーキテクチャには、クラスター自身の稼働状態とメンテナンスを処理するクラスター ジョブがあり、これらのジョブはすべてOneFSジョブ エンジンによって管理されます。このジョブ エンジンはクラスター全体にわたって動作し、大きなストレージ管理タスクや保護タスクを分割して制御下に置きます。これを実現するため、ジョブ エンジンはタスクを小さいワーク アイテムに分割し、ジョブ全体のこれらの部分を各ノード上の複数のワーカー スレッドに割り当てます（マップします）。ジョブ実行全体を通して進行状況が追跡および報告され、完了または終了時に詳細なレポートとステータスが提示されます。

ジョブ エンジンには包括的なチェック ポインティング システムが組み込まれており、これによってジョブの停止/開始だけでなく一時停止/再開も行うことができます。OneFSジョブ エンジン フレームワークには、拡張性に優れたインパクト管理システムも含まれています。

ジョブ エンジンは通常、予備または特に予約された容量とリソースを使用して、ジョブをクラスター全体のバックグラウンド タスクとして実行します。ジョブ自体は次の3つに大別できます。

ファイル システムのメンテナンス ジョブ

ファイル システムをバックグラウンドでメンテナンスするジョブで、通常はすべてのノードへのアクセスを必要とします。これらのジョブはデフォルト構成で実行される必要があり、多くの場合、クラスターが縮退した状態にあるときに実行します。ファイル システム保護やドライブの再構築がその例です。

機能サポート ジョブ

機能サポート ジョブはいくつかの拡張ストレージ管理機能を補助する処理を実行するもので、通常は特定の機能が構成されているときにのみ実行されます。重複排除やアンチ ウイルス スキャンがその例です。

ユーザー アクション ジョブ

これらのジョブは、何らかのデータ管理目標を達成するためにストレージ管理者が直接実行します。並列ツリーの削除や権限のメンテナンスがその例です。

次の表に、公開されているジョブ エンジン ジョブ、ジョブ エンジンが実行する操作、ファイル システムへのアクセス方法をまとめたリストを示します。

ジョブ名	ジョブの説明	アクセス方法
AutoBalance	クラスター内の空き領域のバランスを保つ。	**ドライブ+LIN**
AutoBalanceLin	クラスター内の空き領域のバランスを保つ。	**LIN**
AVScan	アンチウイルス サーバーによって実行されるウイルス スキャン ジョブ。	**ツリー**
ChangelistCreate	2つの連続したSyncIQスナップショット間の変更に関するリストを作成する	**変更リスト**
CloudPoolsLin	ファイル プール ポリシーに従い、クラウド プロバイダーにデータをアーカイブする。	**LIN**
CloudPoolsTreewalk	ファイル プール ポリシーに従い、クラウド プロバイダーにデータをアーカイブする。	**ツリー**
Collect	さまざまな障害状態により使用できなくなっているノードやドライブが原因で解放できなかったディスク領域を回収する。	**ドライブ+LIN**
ComplianceStoreDelete	SmartLockコンプライアンス モードのガベージ コレクション ジョブ。	**ツリー**
Dedupe	ファイル システム内の同一ブロックを重複排除する。	**ツリー**
DedupeAssessment	ドライ ランを実行して重複排除の利点を評価します。	**ツリー**
DomainMark	パスおよびその内容をドメインに関連づける。	**ツリー**
DomainTag	パスおよびその内容をドメインに関連づける。	**ツリー**
EsrsMftDownload	ライセンス ファイルのESRS管理ファイル転送ジョブ。	
FilePolicy	効率的なSmartPoolsファイル プール ポリシー ジョブ。	**変更リスト**
FlexProtect	障害のシナリオから復旧するためにファイル システムを再構築および再保護する。	**ドライブ+LIN**
FlexProtectLin	ファイル システムを再保護する。	**LIN**
FSAnalyze	InsightIQと組み合わせて使用されるファイル システムの分析データを収集する。	**変更リスト**
IndexUpdate	FilePolicyジョブとFSAnalyzeジョブ用の効率的なファイル システムの索引を作成してアップデートする。	**変更リスト**
IntegrityScan	ファイル システムのすべての不整合について、オンラインでの検証と修正を行う。	**LIN**
LinCount	ファイル システムの論理inode（LIN）をスキャンし、カウントする。	**LIN**

ジョブ名	ジョブの説明	アクセス方法
MediaScan	ドライブをスキャンしてメディア レベルのエラーを探す。	**ドライブ+LIN**
MultiScan	CollectジョブとAutoBalanceジョブを同時に実行する。	**LIN**
PermissionRepair	ファイルおよびディレクトリーの権限を修正する。	**ツリー**
QuotaScan	既存のディレクトリ パス上に作成されたドメインのクォータ アカウントを更新する。	**ツリー**
SetProtectPlus	デフォルトのファイル ポリシーを適用する。クラスター上でSmartPoolsがアクティブになっている場合、このジョブは無効になります。	**LIN**
ShadowStoreDelete	シャドウ ストアに関連づけられている領域を解放する。	**LIN**
ShadowStoreProtect	要求度の高い保護レベルのLINで参照されるシャドウ ストアを保護する。	**LIN**
ShadowStoreRepair	シャドウ ストアを修復する。	**LIN**
SmartPools	同一クラスター内のノード階層間で実行され、データを移動するジョブ。ライセンスが供与され、構成されている場合は、CloudPools機能も実行する。	**LIN**
SmartPoolsTree	サブツリー上にSmartPoolsファイル ポリシーを適用する。	**ツリー**
SnapRevert	スナップショット全体を先頭に戻します。	**LIN**
SnapshotDelete	削除されたスナップショットに関連づけられているディスク領域を解放する。	**LIN**
TreeDelete	クラスター自身から直接、ファイル システム内のパスを削除する。	**ツリー**
Undedupe	ファイル システム内での同一ブロックの重複排除を削除する。	**ツリー**
アップグレード	OneFSの今後のリリースでクラスターをアップグレードする。	**ツリー**
WormQueue	SmartLock LINのキューをスキャンする	**LIN**

図 1：OneFS ジョブ エンジンのジョブの説明

ファイル システム メンテナンス ジョブは、スケジュールに従って、または特定のファイル システム イベントに反応して、デフォルトで実行されますが、いずれのジョブ エンジン ジョブも、その（他のジョブに対する）優先度レベルおよびインパクト ポリシーを設定することによって管理できます。

インパクト ポリシーは、ある週のうちの一定時間に相当する1つ以上のインパクト インターバルで構成されます。各インパクト インターバルには、そのインターバルで使用するインパクト レベルを定義済みの選択肢の中から1つ選択できます。これにより、特定のクラスター処理に使用するクラスター リソースの量が決まります。選択可能なジョブ エンジンのインパクト レベルは次のとおりです。

- Paused
- 低
- 中
- 高

クラスターでの動作を円滑にするために、この粒度を使用してインパクト インターバルとインパクト レベルをジョブごとに構成できます。これらによって指定されたインパクト ポリシーにより、ジョブがいつ実行されてどの程度のリソースを消費できるかが決まります。

さらに、1～10の段階でジョブ エンジン ジョブの優先度が設定されます（数値が低いほど高い優先度を示します）。これは概念的にはUNIXのスケジューリング ユーティリティである「nice」に似ています。

ジョブ エンジンでは、最大3つのジョブを同時に実行することができます。このジョブの同時実行は次の条件によって制御されます。

- ジョブの優先度
- 除外セット：同時に実行できないジョブ（例：FlexProtectとAutoBalance）
- クラスターの稼働状態：ほとんどのジョブは、クラスターが劣化状態のときには実行できません。

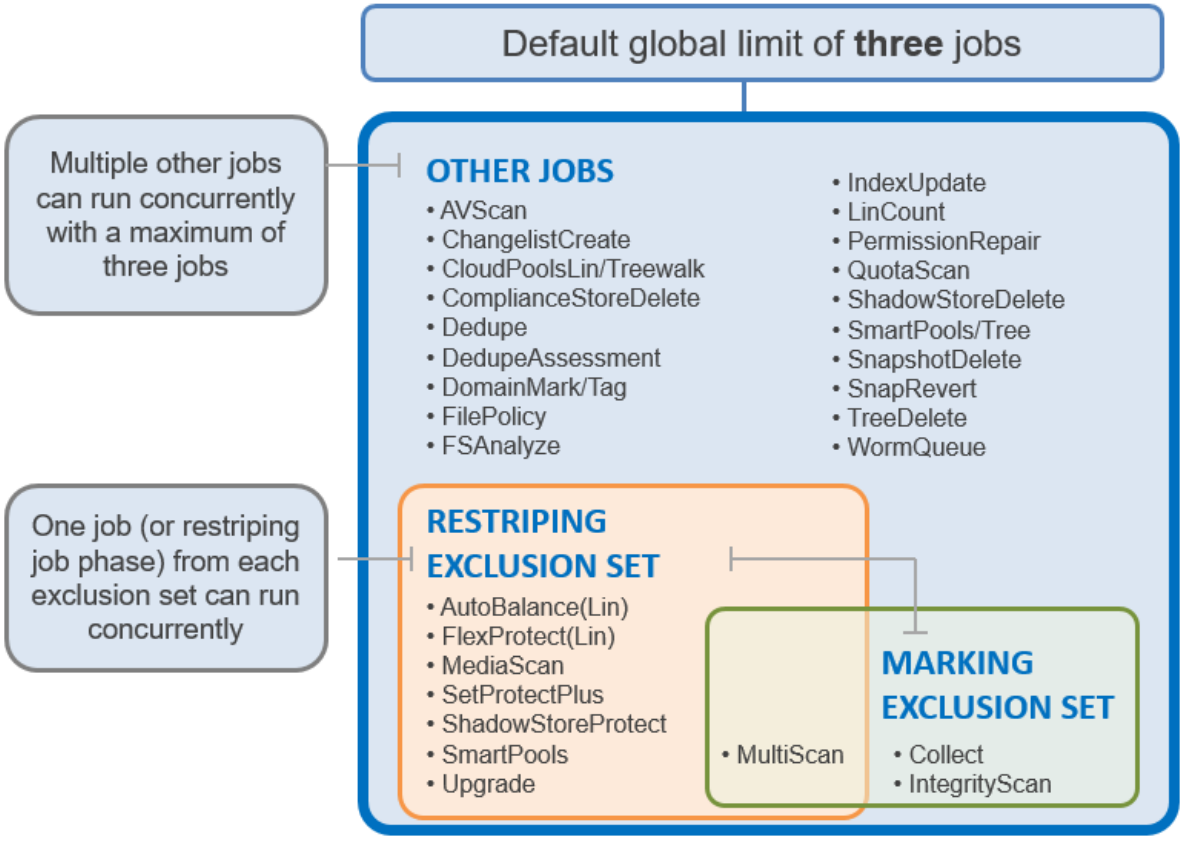

図 4：OneFS ジョブ エンジンの除外セット

📖 詳細については、OneFSジョブ エンジンに関するホワイト ペーパーを参照してください。

ファイル システム構造

OneFSファイル システムはUFS（UNIXファイル システム）に基づいているため、非常に高速な分散ファイル システムです。各クラスターで1つのネームスペースとファイル システムが作成されます。このため、ファイル システムはクラスター内のすべてのノードにわたって分散され、クラスター内の任意のノードに接続するクライアントからアクセスできます。パーティション設定はなく、ボリュームを作成する必要もありません。空き領域や権限のないファイルへのアクセスを物理ボリューム レベルで制限する代わりに、OneFSは同じ機能を共有やファイル権限を通じて、またSmartQuotasサービス（ディレクトリー レベルのクォータの管理を提供）を通じてソフトウェアで提供します。

📖 詳細については、OneFS SmartQuotas に関するホワイト ペーパーを参照してください。

すべての情報は内部ネットワークの各ノード間で共有されるため、データの書き込みと読み取りは任意のノードを対象に実行できます。これにより、複数のユーザーが同じデータセットに対して同時に読み取りと書き込みを行う際のパフォーマンスが最適化されます。

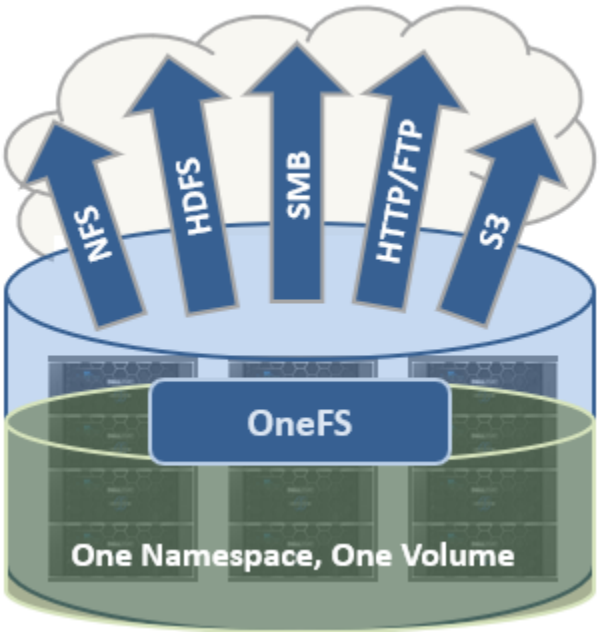

図 5：複数のアクセス プロトコルに対応する単一のファイル システム

OneFSは、1つのネームスペースを使用した真に単一のファイル システムです。データとメタデータは、冗長性と可用性を確保するために全ノードにわたってストライピングされます。ストレージは、ユーザーと管理者に対して完全に仮想化されています。ファイル ツリーは組織的に成長します。ツリーがどのように成長するか、またはユーザーがツリーをどのように使用するかについて計画を立てたり監視したりする必要はありません。ファイルを適切なディスクに階層化することについて、管理者が特別に注意を配る必要はありません。これはOneFS SmartPoolsによって自動的に処理され、単一ツリーに悪影響が及ぶこともありません。このような大きなツリーをどのようにレプリケートするかについて、特別に考慮する必要もありません。OneFS SyncIQサービスが、ファイル ツリーの形や深さに関係なく、ファイル ツリーの転送を1つ以上の代替クラスターへと自動的に並列化します。

このデザインはネームスペース統合（従来のNASに単一のネームスペースがあるように「見せる」ために広く使用されているテクノロジー）と比較されます。ネームスペース統合では、シンプルな「見た目上の」レイヤーを使用して、ボリューム内の個々のディレクトリがシンボリック リンク経由で最上位のツリーに結び付けられますが、ファイルは個別のボリュームで管理する必要があります。このモデルでは、LUNとボリューム（およびボリューム リミット）が存在しています。負荷を分散するには、ファイルをボリュームからボリュームへと手動で移動する必要があります。また管理者は、ツリーのレイアウトについて注意を払う必要があります。階層化はシームレスと呼ぶには程遠く、さまざまな手作業を継続的に行う必要があります。フェイルオーバーを実装するにはボリューム間のミラーリング ファイルが必要なため、効率性が低下するだけでなく、購入コスト、消費電力、冷却コストも増大します。総合的にみて、ネームスペース統合を使用した場合の管理者の負担は、シンプルなNASデバイスを使用した場合よりも大きくなります。このため、この種のインフラストラクチャを大規模に成長させることは困難なのが実状です。

データ レイアウト

OneFSでは、メタデータ用の物理ポインターとエクステントを使用して、ファイルとディレクトリのメタデータをinodeに格納します。通常、OneFS論理inode（LIN）のサイズは512バイトです。そのため、ほとんどのハード ドライブのフォーマットに使用されているネイティブ セクターに収まります。4KBセクターでフォーマットされるようになった高密度クラスのハード ドライブをサポートするため、8KB inodeもサポートされています。

ファイル システムではB-Treeが重点的に使用され、膨大な数のオブジェクトの拡張性と、データやメタデータの瞬時に近い検索が実現されています。OneFSは完全に対称的な高度に分散化されたファイル システムです。データとメタデータは、複数のハードウェアとデバイス間で常に冗長化されます。データはクラスター内のノード間で消去コーディングを使用して保護され、それによりクラスターが高効率化されるとともに、5ノード以上のクラスターで80%以上のraw容量に対する有効容量が実現します。メタデータ（通常はシステムの1%未満）は、パフォーマンスと可用性のためにクラスター内でミラーリングされます。OneFSはRAIDに依存しないため、冗長性の量はクラスターのデフォルトを超えてファイル レベルまたはディレクトリ レベルで管理者が選択できます。メタデータのアクセスとロックのタスクは、ピア ツー ピア アーキテクチャで見られるのと同様に、すべてのノードによって集合的に管理されます。この対称性が、アーキテクチャのシンプルさと耐障害性を高める鍵となっています。単一のメタデータ サーバー、ロック マネージャー、ゲートウェイ ノードは存在しません。

OneFSでは複数のデバイスから同時にブロックにアクセスする必要があるため、データとメタデータに使用されるアドレス指定スキームは{node, drive, offset}のタプルによって物理レベルでインデックスされます。例えば、ノード3のディスク2にあるブロックのブロック アドレスが12345の場合は{3,2,12345}となります。クラスター内のすべてのメタデータは、データ保護のために多重にミラーリングされます（少なくとも、関連づけられたファイルの冗長性にまで）。たとえば、あるファイルが「+2n」の消失訂正符号で保護されている場合、そのファイルは2件の同時障害にまで耐えることができ、そのファイルへのアクセスに必要なすべてのメタデータは3重にミラーリングされるため、やはり2件の障害に耐えることができます。このファイル システムは、その構造にかかわらず、クラスター内のすべてのノードのすべてのブロックを使用できるように設計されています。

他のストレージ システムでは、RAIDとボリューム管理レイヤーを通じてデータが送信されるため、データ レイアウトの効率性が失われ、ブロック アクセスが最適化されません。OneFSでは、ファイル ディレクトリーの配置はクラスター内のすべてのドライブのセクター レベルにいたるまで制御されます。これにより、データ配置とI/Oパターンが最適化され、不要な読み取り/変更/書き込み処理が回避されます。データをファイル単位でディスクに配置できるため、OneFSではストライピングのタイプを柔軟に制御できるだけでなく、ストレージの冗長性も、システム レベル、ディレクトリ レベル、さらにはファイル レベルで制御することが可能です。従来型のストレージ システムでは、RAIDボリューム全体を特定のパフォーマンス タイプと保護設定専用に使用する必要がありました。たとえば、一連のディスクを、あるデータベース用にRAID 1+0でアレンジするなどといったことが必要でした。これでは、（アイドル スピンドルを活用できないため）スピンドルの使用をストレージ全体にわたって最適化することが難しく、設計の柔軟性も奪われ、ビジネス要件に適応できなくなります。OneFSでは、個別の調整や柔軟な変更をいつでも、完全にオンラインで実行できます。

ファイル書き込み

OneFSソフトウェアはすべてのノードで同様に実行されます。つまり、すべてのノードにわたって実行される単一のファイル システムが作成されます。1つのノードがクラスターを制御したり「支配」したりすることはなく、すべてのノードが真のピアとなります。

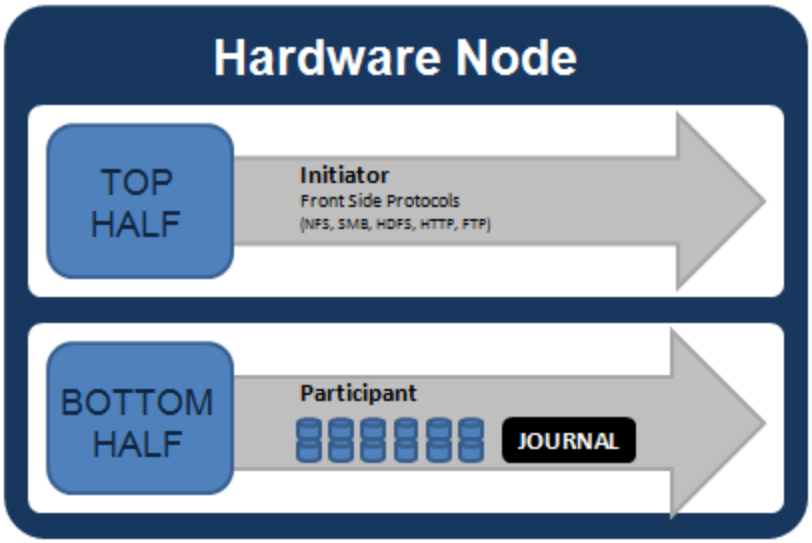

図 6：I/O に関わるノード コンポーネントのモデル

クラスターの全ノード内の、I/Oに関わるすべてのコンポーネントを大きく見ると、図6のように表すことができます。スタックが、上位レイヤー（イニシエーター）と下位レイヤー（参加者）に分割されています。この分割は、特定の読み取りや書き込みを分析するための「論理モデル」として使用されます。物理レベルでは、各ノード内のCPUおよびRAMキャッシュがイニシエーターと参加者のタスクを同時に処理し、I/Oはクラスター全体にわたって発生します。上の図では省略されていますが、実際には、キャッシュと分散型ロック マネージャーも存在します。これらについては、このホワイト ペーパーの後のセクションで説明します。

クライアントは、ファイルを書き込むためにノードに接続する際、そのノードの上半分、つまりイニシエーターに接続します。ファイルは、ノード（ディスク）の下半分（参加者）に書き込まれる前に、ストライプと呼ばれるより小さな論理チャンクへと分割されます。書き込みコアレッサーを使用したフェイルセーフ バッファリングは、書き込みを効率化し、読み取り/変更/書き込み処理を回避するために使用されます。各ファイル チャンクのサイズは、ストライプ ユニット サイズと呼ばれます。

OneFSでは、単にディスク全体ではなく、すべてのノードにわたってデータがストライピングされます。また、ソフトウェア消去コードやミラーリング テクノロジーを通じて、ファイル、ディレクトリ、関連づけられたメタデータが保護されます。OneFSでは、データ保護にリード ソロモン消去コード システムとミラーリングのいずれを使用するかを管理者の判断で選択できます（前者の方がより一般的に使用されます）。ミラーリングは、ユーザー データに適用される場合、高トランザクション パフォーマンスのケースでより多く使用される傾向があります。ユーザー データの大部分には通常、消去コーディングが使用されます。これは、ディスク上の効率性を犠牲にすることなく、極めて高いパフォーマンスを提供できるためです。消去コーディングでは、5ノード以上のRawディスクで80%以上の効率性を提供できます。これは大規模なクラスターでも可能で、4倍程度の冗長性を提供できます。ファイルのストライプ幅とは、ファイルが書き込まれるノード数です（ドライブ数ではありません）。これはクラスター内のノード数、ファイルのサイズ、保護設定（例：+2n）によって決定されます。

OneFSは、高度なアルゴリズムを使用して、効率性とパフォーマンスを最大限に高めるデータ レイアウトを決定します。クライアントがノードに接続する際には、そのノードのイニシエーターがそのファイルの書き込みデータ レイアウトの「キャプテン」として動作します。データ、消失訂正符号(ECC)保護、メタデータ、inodeがすべて、クラスター内の複数のノードに分散されるだけでなく、ノード内の複数のドライブにまで分散されます。

下の図7は、3ノードのクラスターの全ノードで発生するファイル書き込みを示したものです。

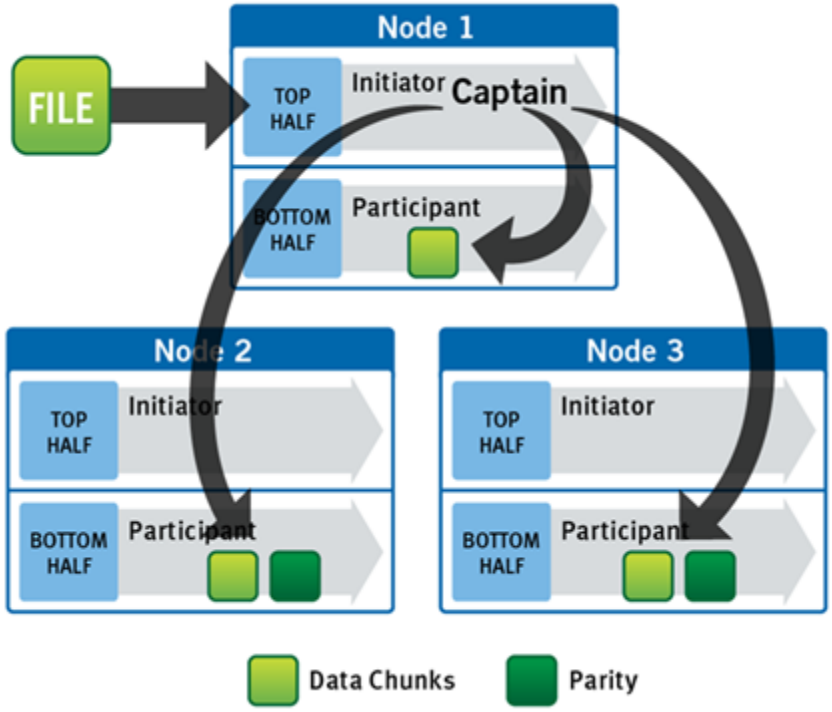

図 7：3 ノード クラスターでのファイル書き込み操作

OneFSでは、バックエンド ネットワークを使用してクラスター内の全ノードにデータが自動的に割り当てられてストライピングされるため、追加の処理は必要ありません。データが書き込まれる際、データは指定されたレベルで保護されます。書き込みが行われる際、OneFSは保護グループと呼ばれる極小単位にデータを分割します。保護グループには冗長性が組み込まれており、すべての保護グループが安全であれば、ファイル全体が安全ということになります。消去コードで保護されたファイルについては、保護グループは一連のデータ ブロックと、それらのデータ ブロックに対する消去コード セットで構成されます。ミラー ファイルについては、保護グループはブロック セットのすべてのミラーで構成されます。OneFSでは、書き込みの際、ファイルで使用されている保護グループのタイプを動的に切り替えることができます。これにより、多数の追加機能が使用可能になります。たとえば、クラスター内で一時的なノード障害が発生して必要な数の消去コードが使用できないような場合に、システムをブロックせずに継続することが可能です。このような場合には、ミラーリングを一時的に使用することで書き込みを継続することができます。ノードがクラスターにリストアされると、ミラーリングされたそれらの保護グループがシームレスかつ自動的に逆変換され、管理者の介入なく消去コード保護されます。

OneFSファイル システムのブロック長は8 KBです。8 KB未満のファイルには8 KBのブロック全体が使用されます。データの保護レベルによっては、この8KBのファイルに8KB以上のデータ容量が使用される場合もあります。ただし、データ保護設定の詳細については、このホワイト ペーパーの後のセクションで説明します。OneFSでは、ファイル システム内の膨大な数の小さなファイルを極めて高いパフォーマンスでサポートできます。ディスク上のすべての構造がこれらのサイズに適応できるよう設計されているため、オブジェクトの総数に関係なく、任意のオブジェクトにほぼ一瞬でアクセスできます。大きなファイルについては、OneFSでは連続する複数の8KBブロックを使用できます。その場合、最大16個の連続するブロックを単一ノードのディスク上にストライピングできます。ファイルのサイズが32KBの場合は、連続する8KBブロックが4つ使用されます。

さらに大きなファイルの場合、OneFSでは、16個の連続するブロックで構成されたストライプ ユニット（ストライプ ユニットあたり128 KB）を使用することで、シーケンシャル パフォーマンスを最大化できます。書き込みの際、データはストライプ ユニットに分割され、保護グループとして複数のノードに分散されます。データがクラスター全体にレイアウトされる際には、ファイルが常に保護されるよう、必要に応じて、消去コードまたはミラーが各保護グループ内で分散されます。

OneFSのAutoBalance機能の主な働きの1つは、データの再割り当てとリバランスを行うことで、ストレージ スペースの使用効率を高めることです（可能な場合）。多くの場合、大きなファイルのストライプ幅を増やすことで、（ノードの追加時に）新しい空き容量を利用し、ディスク上のストライピング効率を高めることができます。AutoBalanceは、ディスク上の効率性を維持し、ディスク上の「ホット スポット」を自動的に排除します。

「キャプテン」ノードの上半分であるイニシエーターは、修正された2フェーズ コミット トランザクションを使用して、クラスター内の複数のNVRAMに対する書き込みを安全に分散します（下の図8を参照）。

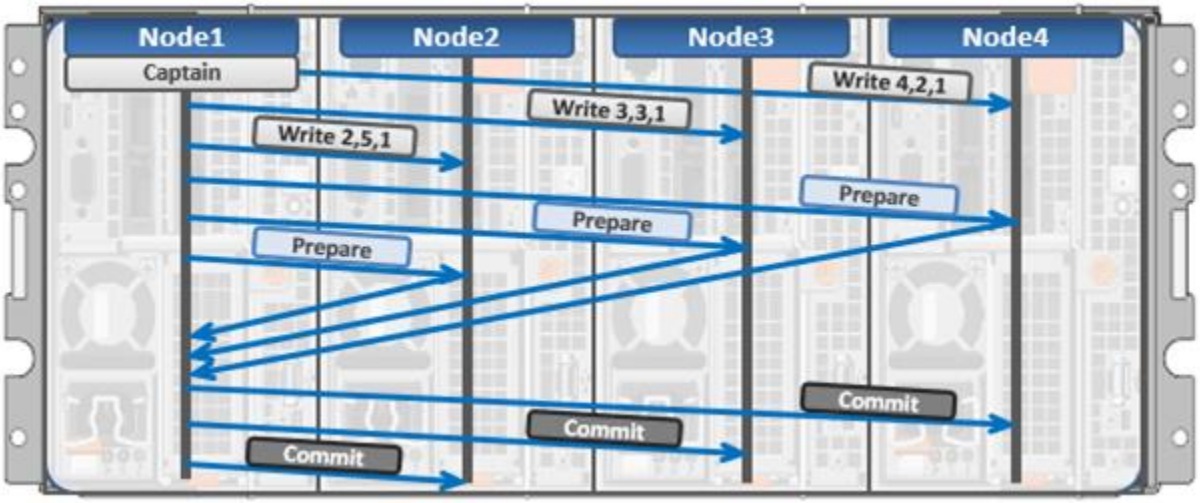

図 8：分散トランザクションと 2 フェーズ コミット

2フェーズ コミットでは、特定の書き込み内のブロックを所有するすべてのノードが処理に関連します。このメカニズムは、ストレージ クラスター内のすべてのノードで発生するすべてのトランザクションのジャーナル処理において、NVRAMに依存します。複数のNVRAMを並行して使用すると、高スループットの書き込みが可能になるだけでなく、電源障害を含むすべてのタイプの障害に対してデータの安全性が確保されます。ノードがトランザクション中に障害を起こした場合、そのトランザクションはそのノードを使用せずにただちに再開されます。ノードの復旧後、そのノードについて必要なアクションは、NVRAMからジャーナルを再生すること（数秒から数分で完了）と、場合によって、トランザクションに関連したファイルをAutoBalanceによって再バランシングすることだけです。リソース コストの高い「fsck」や「ディスク チェック」処理は必要ありません。時間のかかる再同期化も一切不要です。障害によって書き込みがブロックされることはありません。この特許取得済みのトランザクション システムが、OneFSで単一（および複数）の障害点を解消するための手段の1つとなっています。

書き込み処理では、イニシエーターがデータとメタデータのレイアウト、消去コードの作成、ロック管理および権限管理の通常のオペレーションを制御します。管理者は、Web管理またはCLIインターフェイスから、OneFSによって決定されたレイアウトをワークフローに応じていつでも最適化できます。管理者は、次のアクセス パターンをファイル単位またはディレクトリ単位で選択できます。

- コンカレンシー：多数の同時クライアントを想定した最適化により、クラスター上の同時ロードを強化します。この設定は、混在ワークロードに最善の動作を提供します。
- ストリーミング：最適化により、単一ファイルの高速ストリーミングを強化します。例えば、単一クライアントでの非常に高速な読み取りを可能にします。
- ランダム：最適化によってストライピングを調整し、プリフェッチ キャッシュの使用を無効にすることで予測不能なファイル アクセスの処理を強化します。

OneFSには、リアルタイムのアダプティブ プリフェッチも含まれており、管理者の介入なしに、認識可能なアクセス パターンを持つファイルに対して最適な読み取りパフォーマンスを提供します。

ⓘ OneFSが現在サポートしている最大ファイル サイズは、OneFS 8.2.2以降では16TBに増加しており、以前のリリースの最大4TBから増加しました。

OneFSのキャッシュ

OneFSキャッシュ インフラストラクチャの設計は、クラスター内の各ノードに存在するキャッシュをグローバルにアクセス可能な単一のメモリのプールに集約する仕組みに基づいています。この仕組みを実現するため、OneFSではNUMA (Non-Uniform Memory Access)と同様の効率的なメッセージング システムを使用しています。これにより、すべてのノードのメモリ キャッシュが、クラスター内の個々のノードのすべてで使用できるようになります。リモート メモリーへのアクセスは内部的な相互接続を介して行われるため、HDDにアクセスするよりもレイテンシーは大幅に低くなります。

リモート メモリー アクセスの場合、OneFSは、基本的に分散システム バスとして、冗長性がありサブスクライブ不足のフラットEthernetネットワークを利用します。ローカル メモリーほど高速ではありませんが、40Gb Ethernetのレイテンシーが低いことから、リモート メモリー アクセスは非常に高速と言えます。

OneFSキャッシュ サブシステムはクラスター全体にわたって一貫性があります。つまり、複数のノードのプライベート キャッシュに同じコンテンツが存在する場合、そのキャッシュ データはすべてのインスタンスで整合性が取れています。OneFSは、MESIプロトコルを使用してキャッシュの一貫性を維持します。このプロトコルは、共有キャッシュ全体ですべてのデータの整合性を取るために、「書き込み時無効化」ポリシーを実装しています。

OneFSは最大3レベルの読み取りキャッシュに加え、NVRAMベースのライト キャッシュまたはコアレッサーを使用します。これらの相互関係を次の図に示します。

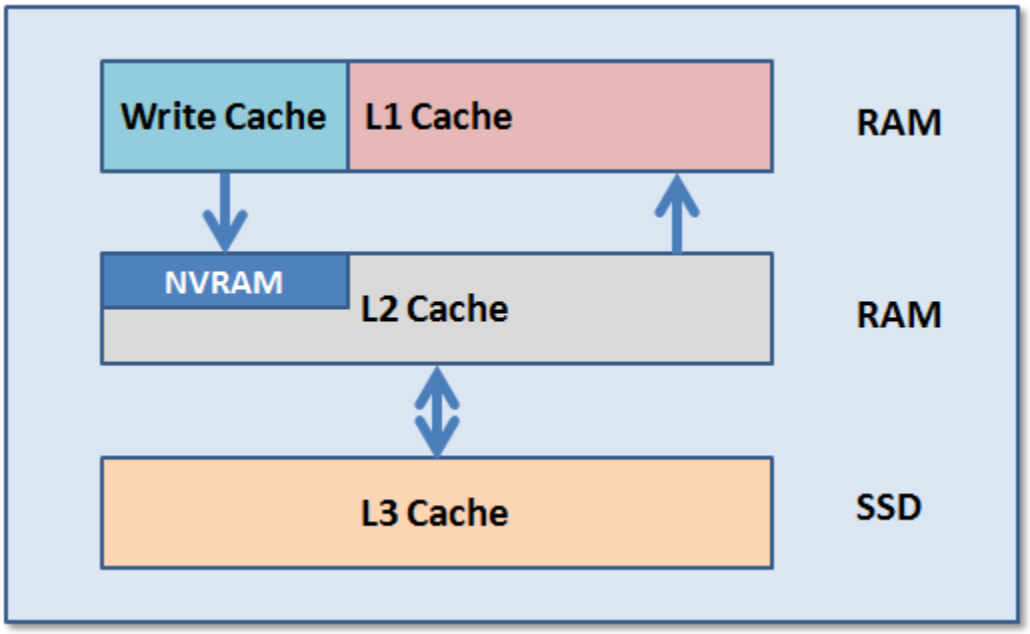

図 9：OneFS のキャッシュ階層

最初の2種類の読み取りキャッシュ、つまりレベル1（L1）およびレベル2（L2）はメモリ（RAM）ベースであり、プロセッサ（CPU）で使用されているキャッシュと同様です。この2つのキャッシュ階層は、すべてのプラットフォーム ストレージ ノードに存在します。

Name	Type	永続性	説明
L1キャッシュ	RAM	揮発性	フロントエンド キャッシュとも呼ばれ、クリーンなクラスター間で一貫したファイル システム データおよびメタデータ ブロック（クライアントからフロントエンド ネットワーク経由で要求）のコピーを保持します。
L2キャッシュ	RAM	揮発性	ファイル システム データとローカル ノードのメタデータのクリーン コピーを格納するバックエンド キャッシュです。
SmartCache / 書き込みコアレッサー	NVRAM	不揮発性	永続性のあるバッテリ バックアップされたNVRAMジャーナル キャッシュです。ディスクにコミットされていない保留中のフロントエンドファイルへの書き込みをバッファーします。
SmartFlash L3キャッシュ	SSD	不揮発性	L2キャッシュからエビクションされたファイルデータとメタデータ ブロックを格納し、L2キャッシュの容量を効率的に向上させます。

OneFSキャッシュの一貫性

OneFSキャッシュ サブシステムはクラスター全体にわたって一貫性があります。つまり、複数のノードのプライベート キャッシュに同じコンテンツが存在する場合、そのキャッシュ データはすべてのインスタンスで整合性が取れています。たとえば、以下の初期状態とイベントの順序について考えます。

1. ノード1とノード5がそれぞれ、共有キャッシュ内のあるアドレスに存在するデータのコピーを持っています。
2. ノード5が書き込み要求を受けて、ノード1のコピーを無効化します。
3. ノード5が値を更新します（下記を参照）。
4. ノード1は、共有キャッシュからデータを再度読み取り、アップデートされた値を取得する必要があります。

OneFSは、MESIプロトコルを使用してキャッシュの一貫性を維持します。このプロトコルは、共有キャッシュ全体ですべてのデータの整合性を取るために、「書き込み時無効化」ポリシーを実装しています。次の図は、キャッシュ内データがとり得るさまざまな状態とそれらの間の遷移を示します。図に示す状態は以下のとおりです。

- M – 変更：データはローカル キャッシュにのみ存在し、共有キャッシュの値から変更されています。変更されたデータのことを通常、「ダーティー」と呼びます。
- E – 排他：データはローカル キャッシュにのみ存在しますが、共有キャッシュの内容と一致します。このデータのことを通常、「クリーン」と呼びます。
- S – 共有：ローカル キャッシュのデータはクラスター内の他のローカル キャッシュにも存在する可能性があります。
- I – 無効：データのロック（排他または共有）が失われています。

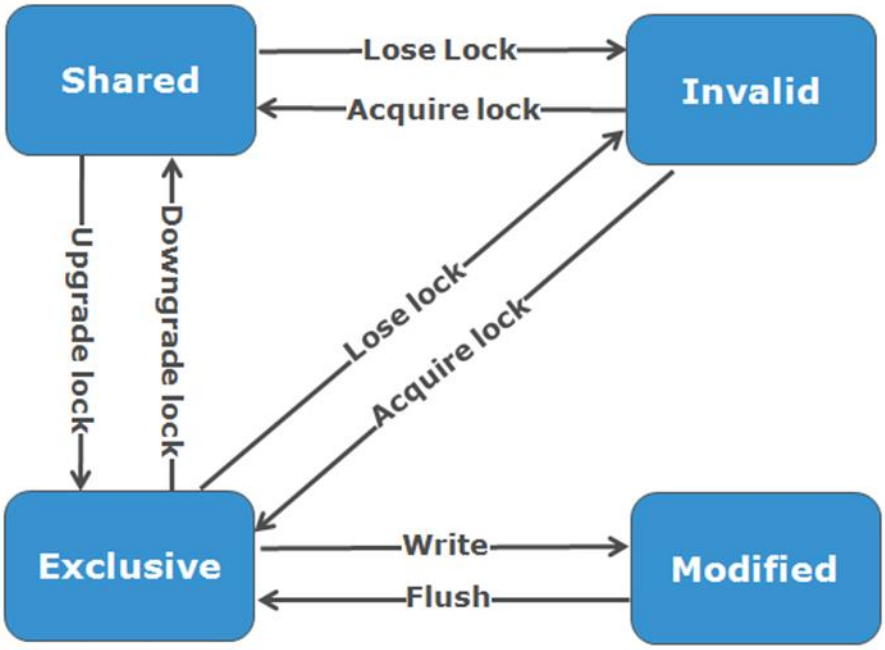

図 10：OneFS キャッシュの一貫性状態図

レベル1キャッシュ

レベル1キャッシュ（L1）、別名フロントエンド キャッシュは、そのノードに接続されたクライアント（イニシエーター）が使用するプロトコル層（NFSやSMBなど）に最も近いメモリーです。L1キャッシュの第一の目的は、リモート ノードからデータをプリフェッチすることです。データはファイル単位でプリフェッチされます。このキャッシュは、ノードのバックエンド ネットワークに伴うレイテンシーを軽減するために最適化されています。バックエンドの相互接続レイテンシーは比較的小さいため、L1キャッシュのサイズ、およびリクエストあたりの標準的なデータ保存量は、L2キャッシュよりも少なくなっています。

L1にはクラスター内の他のノードから取得したデータが格納されるため、L1はリモート キャッシュとも呼ばれます。クラスター全体にわたって一貫性がありますが、それが存在するノードのみが使用可能で、他のノードからアクセスすることはできません。ストレージ ノードのL1キャッシュのデータは、使用後すぐに破棄されます。L1キャッシュはファイル ベースのアドレス方式を使用しており、データのアクセスにはファイル オブジェクトへのオフセットが使用されます。

L1キャッシュはイニシエーターと同じノードのメモリを意味します。ローカル ノードのみがアクセス可能であり、通常このキャッシュはデータのマスター コピーではありません。これはCPUコアのL1キャッシュに似ています。CPUコアのL1キャッシュは、他のコアがメイン メモリに書き込むときに無効化される場合があります。

L1キャッシュの一貫性は、MESIに類似したプロトコルによって、前述の分散ロックを使用して管理されます。

OneFSは、最近要求されたinodeを保持する専用のinodeキャッシュも使用します。inodeキャッシュは一般に、パフォーマンスに大きな影響を与えます。クライアントはたいていの場合データをキャッシュし、ネットワークI/Oアクティビティはファイル属性やメタデータのリクエストが多数を占めます。これらはキャッシュされたinodeからすばやく返すことができるので、inodeキャッシュはパフォーマンスの向上に寄与します。

ⓘ ディスク ドライブを持たないクラスター アクセラレーター ノードでは、L1キャッシュの使い方が異なります。すべてのデータが他のストレージ ノードからフェッチされるので、読み取りキャッシュ全体がL1キャッシュになっています。また、キャッシュ エージングは、ストレージ ノードのL1キャッシュで通常使用されるドロップ ビハインド アルゴリズムではなく、LRU（Least Recently Used）エビクション ポリシーに基づきます。アクセラレータのL1キャッシュはサイズが大きく、格納されたデータは再び要求される可能性が高いため、データ ブロックは使用時にすぐにキャッシュから除去されることはありません。ただし、これはメタデータや更新頻度が高いワークロードにはそれほど効果的ではありません。アクセラレータのキャッシュは、ノードに直接接続されたクライアントにとってのみ有効です。

レベル2キャッシュ

レベル2キャッシュ（L2）、別名バックエンド キャッシュは、特定のデータ ブロックが保存されているノードのローカル メモリを意味します。L2キャッシュはクラスター内のどのノードからもグローバルにアクセス可能で、ディスク ドライブに直接データを要求しないことによって読み取り処理のレーテンシーを軽減するために使用されます。したがって、リモート ノードで使用するためにL2キャッシュにプリフェッチされるデータの量はL1キャッシュよりもはるかに多くなります。

L2キャッシュには、そのノードに存在するディスク ドライブから取得された、リモート ノードからのリクエストに応じて提供できるデータが格納されます。そのため、L2キャッシュはローカル キャッシュとも呼ばれます。L2キャッシュ データのエビクションは、LRU（Least Recently Used）アルゴリズムに従って行われます。

L2キャッシュのデータは、ローカル ノードによって、そのノードに存在するディスク ドライブへのオフセットを使用してアドレス指定されます。当該ノードはリモート ノードから要求されたデータがディスク上のどこに存在するかわかっているため、これはリモート ノード宛てのデータを取得する非常に高速な方法となります。リモート ノードは、特定のファイル オブジェクトに対してブロック アドレスのルックアップを実行することによってL2キャッシュにアクセスします。前述したように、ここではMESIの無効化は不要であり、キャッシュは書き込み中に自動的に更新され、トランザクション システムとNVRAMによって一貫性が維持されます。

レベル3キャッシュ

オプションの3番目の読み取りキャッシュ階層はSmartFlashまたはレベル3（L3）キャッシュと呼ばれ、ソリッド ステート ドライブ（SSD）を含むノードではこのキャッシュ階層も構成できます。SmartFlash（L3）はエビクション キャッシュであり、メモリからエージ アウトされたL2キャッシュ ブロックが入力されます。キャッシュに従来のファイル システム ストレージ デバイスではなくSSDを使用することにはいくつかの利点があります。たとえば、キャッシュ用に予約するとSSD全体が使用されるため、書き込みが非常に直線的かつ予測可能な形で発生します。これにより非常に優れた使用率が実現し、摩耗も著しく低下するため、特にランダム書き込みのワークロードを伴う通常のファイル システム用途と比較して耐久性が向上します。また、SSDをキャッシュ用に使用した場合、ストレージ階層として使用した場合と比較して、SSDのサイズ設定もはるかに単純になり、エラーの発生確率が低くなります。

次の図は、クライアントがOneFS読み取りキャッシュ インフラストラクチャおよび書き込みコアレッサーとどのようにやり取りするかを示します。L1キャッシュは必要な任意のノードのL2キャッシュとやり取りし、L2キャッシュはストレージ サブシステムとL3キャッシュの両方とやり取りします。L3キャッシュはノード内のSSDに存在し、同じノード プールの各ノードでL3キャッシュが有効になっています。

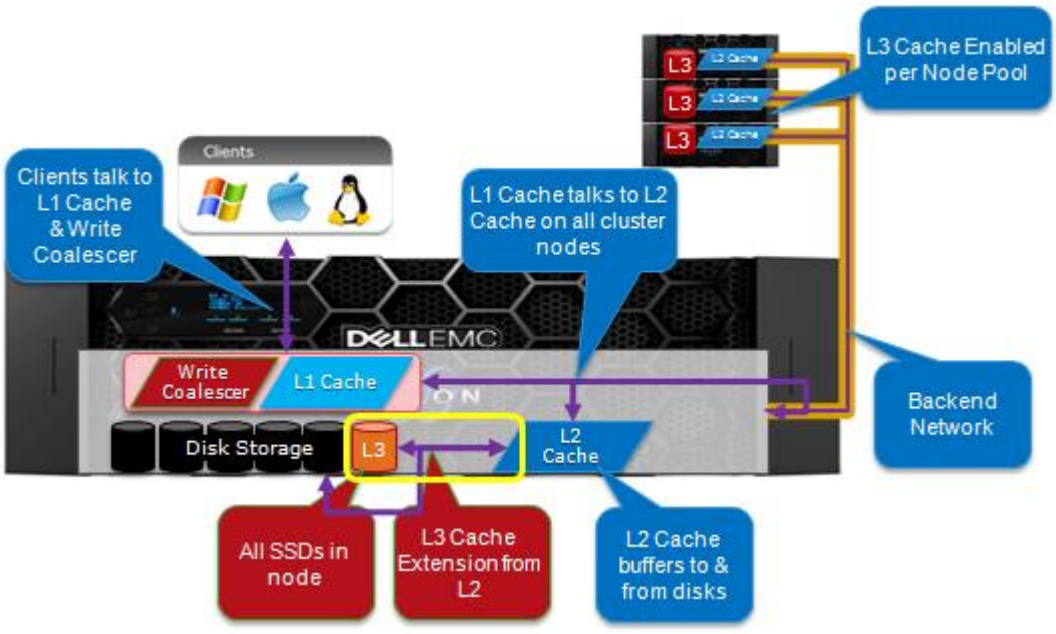

図 11：OneFS L1、L2、L3 のキャッシュ アーキテクチャ

OneFSでは、クラスター内の複数のノード間、場合によってはノード内の複数のドライブ間でファイルの書き込みが行われるように規定されているため、すべての読み取り要求にリモート（場合によってはローカル）データの読み取りが伴います。クライアントから読み取り要求が到着すると、OneFSは要求されたデータがローカル キャッシュにあるかどうかを判断します。ローカル キャッシュにあるデータは即座に読み取られます。要求されたデータがローカル キャッシュにない場合は、ディスクから読み取られます。ローカル ノードにないデータについては、そのデータが存在するリモート ノードから要求が行われます。他方の各ノードでは、キャッシュ ルックアップが新たに実行されます。キャッシュにあるデータは即座に返されますが、キャッシュにないデータはディスクから取得されます。

データがローカル キャッシュおよびリモート キャッシュ（場合によってはディスク）から取得されると、そのデータはクライアントに返されます。

ローカル ノードおよびリモート ノードの両方で読み取り要求を実行する手順の概要は、以下に示すとおりです。

ローカル ノード（要求を受け取るノード）：

1. 要求されたデータの一部がローカルのL1キャッシュに存在するかどうかを特定します。存在する場合は、クライアントに返します。
2. ローカル キャッシュに存在しない場合は、リモート ノードからデータを要求します。

リモート ノード：

1. 要求されたデータがローカルのL2またはL3キャッシュに存在するかどうかを特定します。存在する場合は、要求元のノードに返します。
2. ローカル キャッシュに存在しない場合は、ディスクから読み取って要求元のノードに返します。

書き込みキャッシュはクラスターへのデータの書き込み処理を高速化します。これは、小さな書き込み要求を1つにまとめ、それらをより大きなチャンクでディスクに送信することで、ディスク書き込みレイテンシーの量を大幅に減らす技術です。クライアントがクラスターに書き込みを行うと、OneFSはディスクへの書き込みをすぐには行わず、イニシエーター ノード上のNVRAMベースのジャーナル キャッシュにデータを一時的に書き込みます。OneFSはその後、キャッシュされたそれらの書き込みを、より都合の良いタイミングでディスクにフラッシュします。またこれらの書き込みは、ファイルの保護要件を満たすため、参加者ノードのNVRAMジャーナルにもミラーリングされます。そのため、クラスター分割や予期しないノード停止が起こった際にも、未コミットのライト キャッシュは完全に保護されます。

ライト キャッシュは次のように機能します。

- NFSクライアントが、+2nで保護されたファイルの書き込み要求をノード1に送信します。
- ノード1が、NVRAMライト キャッシュ（高速パス）への書き込みを受け入れ、その後、データ保護のために書き込みを特定のノードのログ ファイルにミラーリングします。
- 書き込み確認がNFSクライアントにただちに返され、ディスク書き込みレイテンシーが回避されます。
- ノード1の書き込みキャッシュがいっぱいになると、キャッシュが定期的にフラッシュされ、（前述の）2フェーズ コミット プロセスを通じて書き込みがコミットされて、適切な消失訂正符号(ECC)保護(+2n)が適用されます。
- ライト キャッシュと参加者ノード ログ ファイルがクリアされ、新しい書き込みを受け付け可能な状態になります。

📖 詳細については、OneFS SmartFlashに関するホワイト ペーパーを参照してください。

ファイル読み取り

データ、メタデータ、inodeがすべてクラスター内の複数のノードに分散されるだけでなく、ノード内の複数のドライブにまで分散されます。クラスターに対する読み取りや書き込みの際には、クライアントの接続しているノードが、操作の「キャプテン」として動作します。

読み取り処理では、「キャプテン」ノードがクラスター内のさまざまなノードからすべてのデータを収集し、それらをまとまった形でリクエスターに提供します。

コストが最適化された業界標準のハードウェアを使用しているため、クラスターではディスクに対するキャッシュの比率が高くなっています（ノードあたり数GB）。キャッシュは必要に応じて、読み取り用と書き込み用に動的に割り当てられます。このRAMベースのキャッシュはクラスター内のすべてのノードにわたって統一性と一貫性を持っているため、1つのノード上のクライアント読み取り要求は、別のノードですでに処理されたI/Oを活用することができます。これらのキャッシュされたブロックには、低レイテンシー バックプレーンに接続されたどのノードからも迅速にアクセスできるため、容量の大きい効率的なRAMキャッシュが可能となり、読み取りパフォーマンスが大幅に向上します。

クラスターが大きくなるほど、キャッシュのメリットも大きくなります。そのため、クラスター上でのディスクへのI/Oの量は通常、従来型のプラットフォームよりも大幅に低くなり、レイテンシーが軽減されてユーザー エクスペリエンスも向上します。

アクセス パターンとしてコンカレントまたはストリーミングが指定されたファイルの場合、OneFSでは、SmartReadコンポーネントで使用されるヒューリスティックに基づいてデータのプリフェッチを使用できます。SmartReadでは、L2キャッシュからデータの「パイプライン」を作成し、「キャプテン」ノード上のローカル「L1」キャッシュ内にデータをプリフェッチできます。これにより、すべてのプロトコルでシーケンシャルな読み取りパフォーマンスが大幅に向上し、データがRAMから直接ミリ秒単位で読み取られるようになります。高シーケンシャルの場合、SmartReadでは、非常に積極的なプリフェッチを行うことで、個々のファイルを極めて高いデータ転送レートで読み取りおよび書き込みできます。

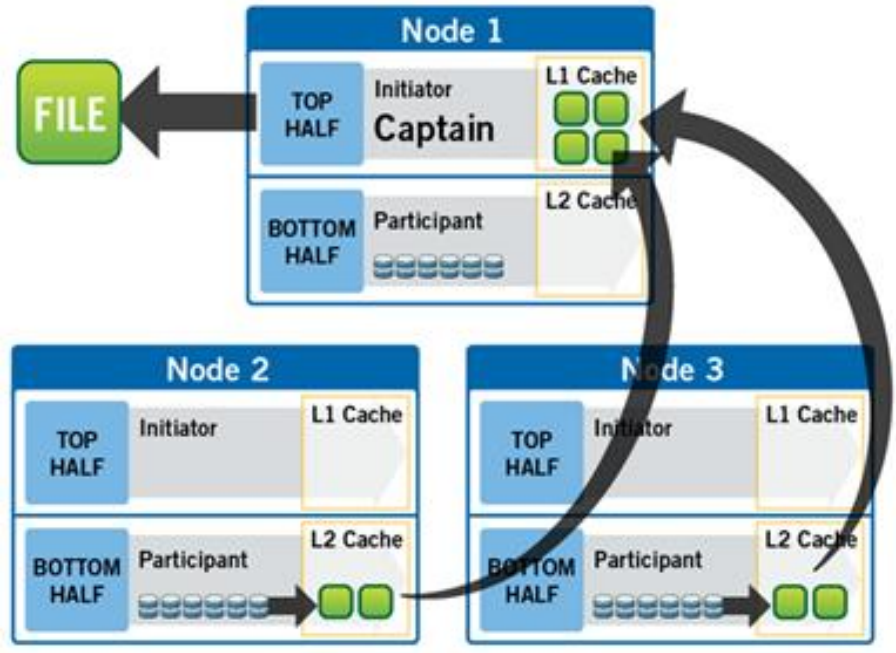

図 12：3 ノード クラスターでのファイル読み取り操作

図10は、3ノード クラスターのノード1に接続されたクライアントから要求されたシーケンシャル アクセスの非キャッシュ ファイルが、SmartReadでどのように読み取られるかを示したものです。

1. ノード1がメタデータを読み取り、ファイル データのすべてのブロックがどこにあるかを特定します。
2. ノード1はL1キャッシュもチェックし、要求されたファイル データがそこにあるかどうかも確認します。
3. ノード1は読み取りパイプラインを作成し、ファイル データの一部を保持するすべてのノードに対し、そのファイル データをディスクから取得するよう、同時要求を送信します。
4. 各ノードは、ファイル データのブロックをディスクからL2キャッシュ（使用可能な場合はL3 SmartFlashキャッシュ）にプルし、ファイル データをノード1に転送します。
5. ノード1は、受信したデータをL1キャッシュに記録しながら、ファイルをクライアントに提供します。その間、プリフェッチ プロセスは継続されます。
6. 高シーケンシャルの場合は、オプションで、L1キャッシュ内のデータを空いているRAMにドロップし、他のL1またはL2キャッシュのニーズに対応させることもできます。

SmartReadのインテリジェントなキャッシュにより、非常に高い読み取りパフォーマンスと、高レベルのコンカレント アクセスが実現します。注目すべき点は、ノード1が（低レーテンシーのクラスター相互接続を介して）ノード2のキャッシュからファイル データを取得する方が、自身のローカル ディスクにアクセスするよりも高速であるという点です。SmartReadのアルゴリズムは、プリフェッチのレベル（ランダム アクセスのケースについてはプリフェッチを無効化）とキャッシュ内のデータ保持期間を制御し、データのキャッシュ場所を最適化します。

ロックと並列性

OneFSには、ストレージ クラスター内の全ノードのデータのロックを管理する、完全分散型のロック マネージャーが備わっています。このロック マネージャーは拡張性に優れ、複数のロック「パーソナリティ」に対応しており、ファイル システム ロックだけでなく、クラスターでの一貫性を持ったプロトコル レベルのロック（SMB共有モード ロックやNFSアドバイザリー モード ロックなど）もサポートしています。OneFSではまた、CIFS OplockやNFSv4デリゲーションなどのデリゲート ロックもサポートされています。

クラスター内のすべてのノードはロック リソースのコーディネーターであり、各コーディネーターは、高度なハッシュ アルゴリズムに基づいてロック可能なリソースに割り当てられます。このアルゴリズムは、コーディネーターがほとんど常に要求のイニシエーターとは異なるノードになるように設計されています。ファイルについてロックが要求された場合、そのロックは共有ロック（複数のユーザーが同時にロックを共有できる（通常は読み取り用））または排他ロック（一度に1人のユーザーが許可される（通常は書き込み用））のどちらかになります。

下の図13は、さまざまなノードのスレッドがコーディネーターからロックを要求する方法の例を示しています。

1. ノード2を、これらのリソースのコーディネーターに指定します。
2. ノード4のスレッド1とノード3のスレッド2が、ノード2のファイルの共有ロックを同時に要求します。
3. ノード2は、要求されたファイルに排他ロックが存在するかどうかを確認します。
4. 排他ロックが存在しない場合は、ノード2が、要求されたファイル上で、ノード4のスレッド1とノード3のスレッド2に共有ロックを許可します。
5. 次に、ノード3とノード4が要求されたファイル上で読み取りを実行します。
6. ノード1のスレッド3が、ノード3とノード4の読み取りの実行中に、同じファイルについて排他ロックを要求します。
7. ノード2は、ノード3とノード4について、共有ロックを回収できるかどうかをチェックします。
8. ノード3とノード4は読み取り実行中であるため、ノード2はノード1のスレッド3にしばらく待機するよう求めます。
9. ノード1のスレッド3は、ノード2が排他ロックを許可するまでブロックしてから、書き込み処理を完了します。

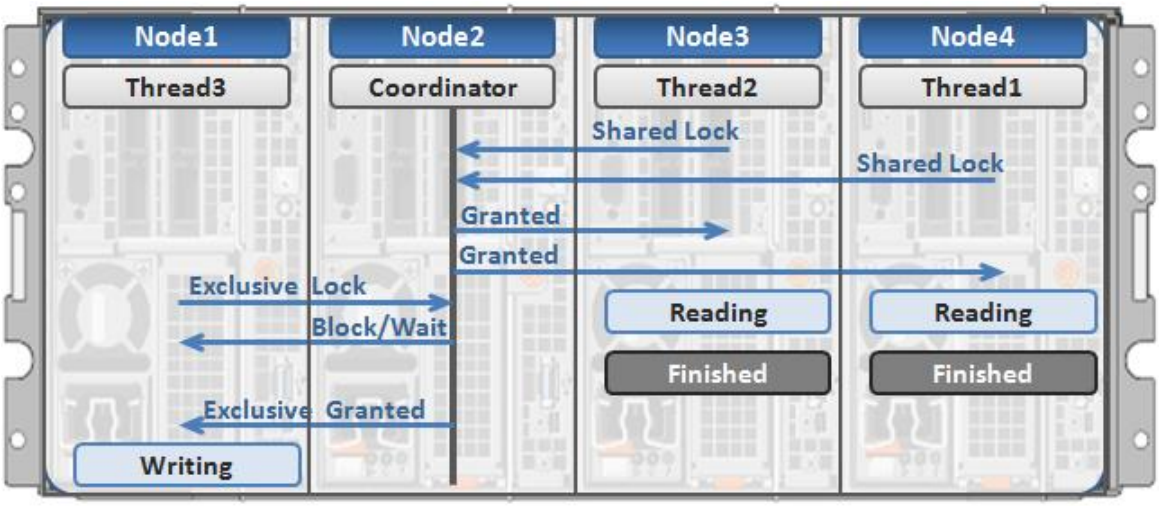

図 13：分散型ロック マネージャー

マルチスレッド化されたIO

サーバーの仮想化における大規模なNFSデータストアの使用が増加するにつれ、エンタープライズ アプリケーションでは、大容量ファイルに対する高スループットと低レーテンシーが求められるようになっています。これに対応するため、OneFS Multi-Writerでは、複数のスレッドの個々のファイルへのコンカレント書き込みがサポートされています。

前の例では、大容量ファイルへのコンカレント書き込みアクセスが、排他ロック機構によって制限され、ファイル レベル全体に適用される可能性があります。この潜在的なボトルネックを避けるために、OneFS Multi-Writerは、ファイルを個別の領域に分割し、ファイル全体ではなく個々の領域への排他書き込みロックを許可することによって、より細かい書き込みロックを提供します。このようにして、複数のクライアントによる、同じファイルの異なる部分へのコンカレント書き込みが可能になります。

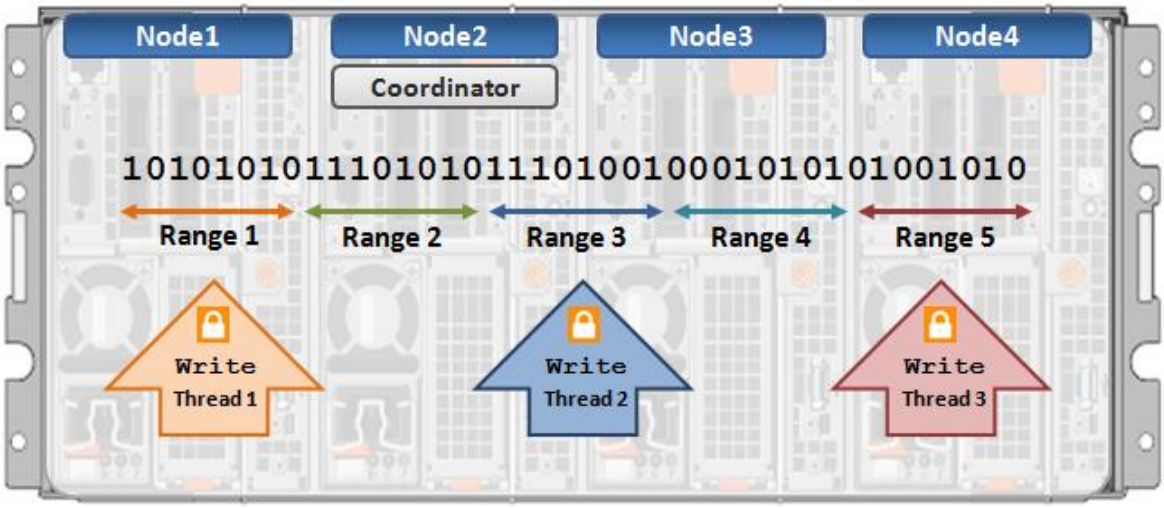

図 14：マルチスレッド化された IO ライター

データ保護

電源障害

ファイル システムへの変更に関する情報を格納するファイル システム ジャーナルは、電源障害などのシステム障害やクラッシュが発生した後、迅速かつ一貫したリカバリを実現できるように設計されています。ファイル システムは、ノードまたはクラスターが電源障害またはその他の障害から復旧した後、ジャーナル項目を再生します。ジャーナルがないと、ファイル システムは障害が発生した後に大容量ファイル システムで個別にすべての潜在的な変更を調査または確認する必要があり（「fsck」または「chkdsk」操作）、この操作には時間がかかる場合があります。

OneFSはジャーナル ファイル システムであり、各ノードには、ファイル システムへのコミットされていない書き込みを保護するためのバッテリ バックアップNVRAMカードが搭載されています。NVRAMカードのバッテリは、再充電なしで数日持続します。ノードは起動時にジャーナルを確認し、ジャーナル処理システムで必要と判断された場合、ディスクに対する処理を選択的に再生します。

OneFSは、システム内にまだ存在しないすべてのトランザクションが記録されたことが確実な場合にのみマウントされます。たとえば、適切なシャットダウン手順に従わなかった場合、NVRAMバッテリが放電した場合、トランザクションが失われた可能性がある場合は、潜在的な問題を回避するために、ノードはファイル システムをマウントしません。

ハードウェア障害とクォーラム

クラスターを適切に機能させ、データ書き込みを許可させるには、ノードのクォーラムがアクティブになっていて、いつでも反応できる必要があります。クォーラムはマジョリティで定義します。個のノードを持つクラスターでは、書き込みを許可するために⌊/2⌋+1個のノードがオンラインになっている必要があります。たとえば、7ノードのクラスターでは、クォーラムに4つのノードが必要です。ノードまたはノードのグループが起動および応答しているが、クォーラムのメンバーではない場合は、読み取り専用で実行されます。

OneFSはクォーラムを使用して「スプリット ブレイン」状態を回避します。この状態は、クラスターが一時的に2つのクラスターに分割された場合に発生することがあります。クォーラム ルールに従うことにより、アーキテクチャでは、障害が発生したノードまたはオンラインに復帰するノードの数にかかわらず、書き込みが発生したときに、これまでに発生したすべての書き込みとの整合性が保たれます。クォーラムでは、特定のデータ保護レベルに移動するために必要なノード数も決まっています。消失訂正符号に基づく+の保護レベルの場合、クラスター内に少なくとも2+1個のノードが必要です。たとえば、+3nの構成には最小7つのノードが必要です。これにより、3個のノードで同時に障害が発生しても4つのノードのクォーラムを維持できるため、完全に機能し続けることができます。クラスターがクォーラムを下回った場合、ファイル システムは自動的に読み取り専用の保護モードに移行し、書き込みは拒否されます。ただしそれでも、使用可能なデータへの読み取りアクセスは可能です。

ハードウェア障害：ノードの追加と削除

グループ管理プロトコル(GMP)と呼ばれるシステムでは、クラスターの状態をいつでも全体的に把握でき、クラスター間のすべてのノードの状態を整合性のとれたビューに表示できます。クラスター相互接続で1つ以上のノードに接続できなくなった場合は、グループが「分割」されるか、クラスターから削除されます。すべてのノードが、そのクラスターの整合性のとれた新しいビューに表示されます（これは、クラスターが2つの異なるノード グループに分割されるようなものですが、1つのグループのみにクォーラムを設定できることに注意してください）。この分割された状態では、ファイル システム内のすべてのデータに接続でき、クォーラムを保持する側については、すべてのデータを変更できます。「障害が発生した」デバイスに格納されたデータは、クラスター内に保存された冗長性を使用して再構築されます。

ノードに再度接続できるようになると、「マージ」または追加が行われ、ノードがクラスター内に戻されます（2つのグループが1つにマージされます）。ノードを再構築または再構成することなく、クラスターに再度参加させることができます。これは、ドライブを再構築する必要があるハードウェアRAIDアレイとは異なります。分割中に一部の保護グループが上書きされてより狭いストライプに変換された場合、効率性を高めるため、AutoBalanceによって一部ファイルが再ストライプされる場合があります。

OneFSジョブ エンジンには、孤立ブロックのコレクターとして機能するCollectと呼ばれるプロセスもあります。書き込み処理中にクラスターが分割されると、ファイルに割り当てられているブロックを、クォーラム側で再割り当てする必要があることがあります。これにより、クォーラムではない側で、割り当てられたブロックが「孤立」することになります。クラスターが再度マージされると、Collectジョブはマーク アンド スイープ スキャンを並列実行してこれらの孤立したブロックを特定し、クラスターの空き領域として再利用します。

拡張性の高い再構築

OneFSは、データ割り当てまたは障害後のデータの再構成のどちらについてもハードウェアRAIDに依存しません。OneFSはファイル データの保護を直接管理するのではなく、エラーが発生したときに、並列化された方法でデータを再構築します。OneFSは、inodeデータをディスクからリニアに読み取ることによって、障害によってファイルのどの部分が影響を受けるかを特定します。影響を受ける一連のファイルは、ジョブ エンジンによって、クラスター間に分散する一連のワーカー スレッドに割り当てられます。ワーカー ノードは、ファイルを並行して修復します。これは、クラスター サイズが増えると、障害による再構築の時間が減ることを意味します。サイズの増加に伴うクラスターの復元性の維持において、効率性の面で非常に大きなメリットがあります。

仮想ホット スペア

RAIDに基づいた従来のストレージ システムでは、通常、障害が発生したドライブを個別にリカバリするために、1つ以上の「ホット スペア」ドライブのプロビジョニングが必要でした。ホット スペア ドライブは、RAIDセット内で障害が発生したドライブに置き換わります。これらのホット スペアが自動的に置き換えられずに別の障害が発生した場合は、システムに致命的なデータ消失が発生します。OneFSはホット スペア ドライブは使用せず、システム内の空き領域を借りて、障害からのリカバリを行います。この技術を仮想ホット スペアと呼びます。このとき、クラスターでは人が介入しない完全自動修復が可能です。管理者は仮想ホット スペアのリザーブを作成できるので、ユーザーが書き込み中であってもシステムは自己修復できます。

消失訂正符号を使用したファイル レベルのデータ保護

クラスターは、クラスターからのデータの供給を妨げることなく1つ以上の同時コンポーネント障害に耐えるように設計されています。これを実現するため、OneFSは、リード ソロモン エラー訂正 （N+M保護） による消失訂正符号ベースの保護またはミラーリング システムのどちらかの方法でファイルを保護します。データ保護はソフトウェアによってファイル レベルで適用されるため、システムは障害によって損なわれたファイルのリカバリのみに集中できます。ファイル セットまたはボリューム全体をチェックして修復する必要はありません。OneFSのメタデータとinodeは常に （リード ソロモン コーディングではなく） ミラーリングによって保護され、保護レベルはそれらが参照しているデータの保護レベルと同じかまたはそれ以上になります。

すべてのデータ、メタデータ、保護情報はクラスター内のすべてのノードにわたって分散されるため、クラスターは専用のパリティ ノードまたはドライブを必要とせず、メタデータの管理に専用のデバイスやデバイスのセットを使用する必要がありません。これにより、特定のノードが単一障害点になることを回避できます。すべてのノードが、実行されるタスクを等しく共有するため、ピア ツー ピア アーキテクチャにおける完全な対称性とロード バランシングが実現します。

OneFSでは、データ保護をいくつかのレベルで設定でき、これはクラスターまたはファイル システムをオフラインにすることなくいつでも修正できます。

消去コードで保護されたファイルの場合、ファイルの各保護グループは「N+M/bのレベルで保護されている」（N>MかつM>=b）と表現します。値NとMはそれぞれ、データに使用されるドライブの数と保護グループ内の消失訂正符号に使用されるドライブの数を表します。値bは、その保護グループのレイアウトに使用されるデータ ストライプ数を表します。このbについて以下に説明します。一般的でわかりやすいケースはb=1であり、保護グループが組み込まれていることを意味します。N台のドライブがデータに、M台のドライブが （消去コードに保存された） 冗長性に使用されていて、保護グループがノード セット全体にわたって正確に1つのストライプでレイアウトされていることを意味します。これにより、保護グループのMメンバーで同時に障害が発生しても、100%のデータ可用性を維持できます。消去コード メンバーMは、データ メンバーNから計算されます。下の図13は、通常の4+2の保護グループ （N=4、M=2、b=1） の場合を示しています。

OneFSではファイルは複数ノード全体に分散されるため、N+Mでストライプされたファイルは、個の同時ノード障害が発生した場合でも可用性を失うことなく障害に耐えることができます。したがって、OneFSは、障害がドライブ、ノード、またはノード内のコンポーネント （カードなど） のどこで発生しても、高可用性を提供できます。さらに、障害は、含まれるコンポーネントの数や種類にかかわらず、ノード単位でカウントされます。そのため、あるノードの5台のドライブに障害が発生しても、N+M保護の観点では1つの障害にしか数えられません。

OneFSではMのレベルが可変であり （最大4） 、最大で四重障害保護を提供できます。これは、今日一般に使用されているRAIDの最大レベルであるRAID-6の二重障害保護をはるかに超えています。ストレージの信頼性は、この冗長性の量とともに幾何学的に増加するため、+4nの保護は、従来のハードウェアRAIDよりも数段高い信頼性をもたらします。したがって、4 TBや6 TBなどの大容量SATAドライブを安心して増設できます。

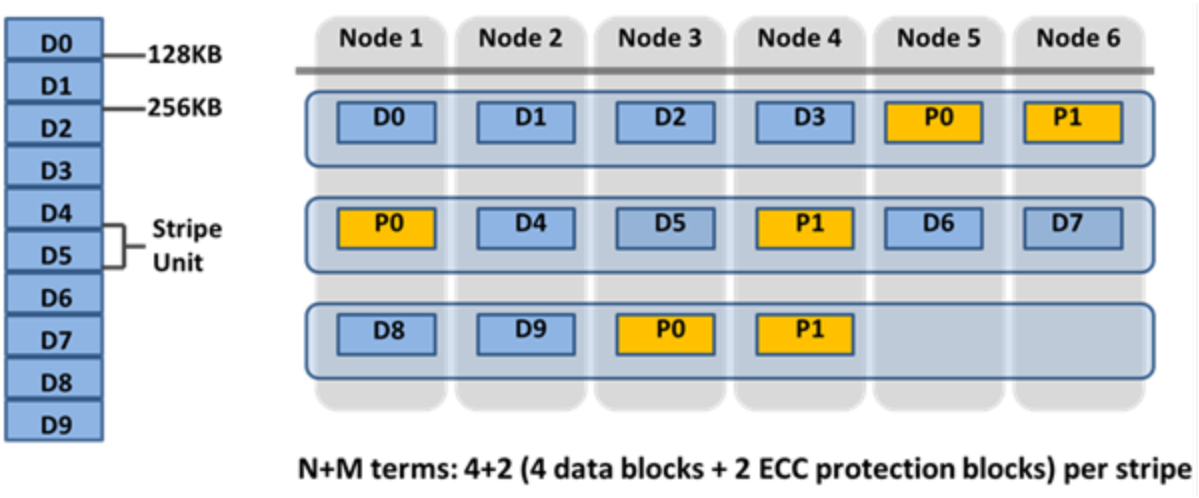

図 15：OneFS の冗長性 – N + M 消失訂正符号による保護

小さいクラスターは+1nで保護できますが、これは、1台のドライブまたはノードはリカバリーできても、2台の異なるノードの2台のドライブはリカバリーできないことを意味します。ドライブ障害が発生する可能性は、ノード障害と比べると桁違いに多くなります。大容量ドライブを持つクラスターの場合、単一ノードの復旧可能性でも許容されますが、複数のドライブ障害に備えた保護を提供することが望まれます。

二重ディスク冗長性と単一ノード冗長性が求められる状況に備えるため、2倍または3倍幅サイズの保護グループを構築できます。2倍または3倍幅の保護グループは、レイアウトするときに同じノード セット上で1回または2回「折り返されます」。各保護グループで冗長性に使用されるディスクの数は正確に2台になるため、このメカニズムにより、2台または3台のディスク障害または完全な単一ノード障害が発生した場合でもデータの可用性を失うことなく障害に耐えることができます。

小さいクラスターで最も重要なのは、このストライピング方法を採用するとディスク効率M/(N+M)が非常に高くなることです。たとえば、二重障害保護を備えた5ノードのクラスターでN=3、M=2を使用した場合、効率が1–2/5 = 60%の3+2保護グループが得られます。同じ5ノードのクラスターで各保護グループを2つのストライプでレイアウトした場合、M=2とするとNは8になるため、ディスク効率は1-2/(8+2) = 80%となります。この場合、二重ノード障害保護だけは犠牲となりますが、二重ドライブ障害保護は確保されます。

OneFSでは、保護スキームがいくつかサポートされています。これには、偏在する+2d:1nが含まれ、2つのドライブ障害または1つのノード障害から保護します。

ⓘ ベスト プラクティスは、特定のクラスター構成で推奨される保護レベルを使用することです。この推奨保護レベルは、OneFS WebUIストレージ プール設定ページで「推奨」として明示されており、通常はデフォルトで設定されます。現在のすべての第6世代ハードウェア構成では、推奨される保護レベルは「+2d:1n」です。

ハイブリッド保護スキームは、第6世代シャーシの高密度ノード構成に特に役立ちます。このような構成では、複数のドライブに障害が発生する確率がノード全体に障害が発生する確率をはるかに上回ります。複数のデバイスで同時に障害が発生するような可能性の低い事例において（ファイルが「保護レベルを超える」など）、OneFSはあらゆる可能性に再度保護を行い、影響を受ける各ファイルのエラーをクラスター ログに報告します。

また、OneFSでは、指定したコンテンツのミラーを2～8個作成できる、2～8倍のさまざまなミラーリング オプションをサポートしています。また、メタデータはデフォルトでFECより1レベル上でミラーリングされます。たとえば、ファイルが+2nで保護される場合、関連づけられたメタデータ オブジェクトは3倍でミラーリングされます。

OneFSのすべての保護レベルが次の表にまとめられています。

保護レベル	説明
+1n	1ドライブの障害または1ノードの障害に対応できる
+2d:1n	2ドライブの障害または1ノードの障害に対応できる
+2n	2ドライブの障害または2ノードの障害に対応できる
+3d:1n	3ドライブの障害または1ノードの障害に対応できる
+3d:1n1d	3ドライブの障害または1ノードと1ドライブの障害に対応できる
+3n	3ドライブの障害または3ノードの障害に対応できる
+4d:1n	4ドライブの障害または1ノードの障害に対応できる
+4d:2n	4ドライブの障害または2ノードの障害に対応できる
+4n	4ノードの障害に対応できる
2～8倍	構成に応じて、2～8ノードにミラーリング

OneFSを使用すると、管理者は、クライアントの追加中、またはデータの読み取りまたは書き込み中でも、リアルタイムで保護ポリシーを変更できます。

ⓘ クラスターの保護レベルを増やすと、クラスター上のデータ使用量が増えることがあります。

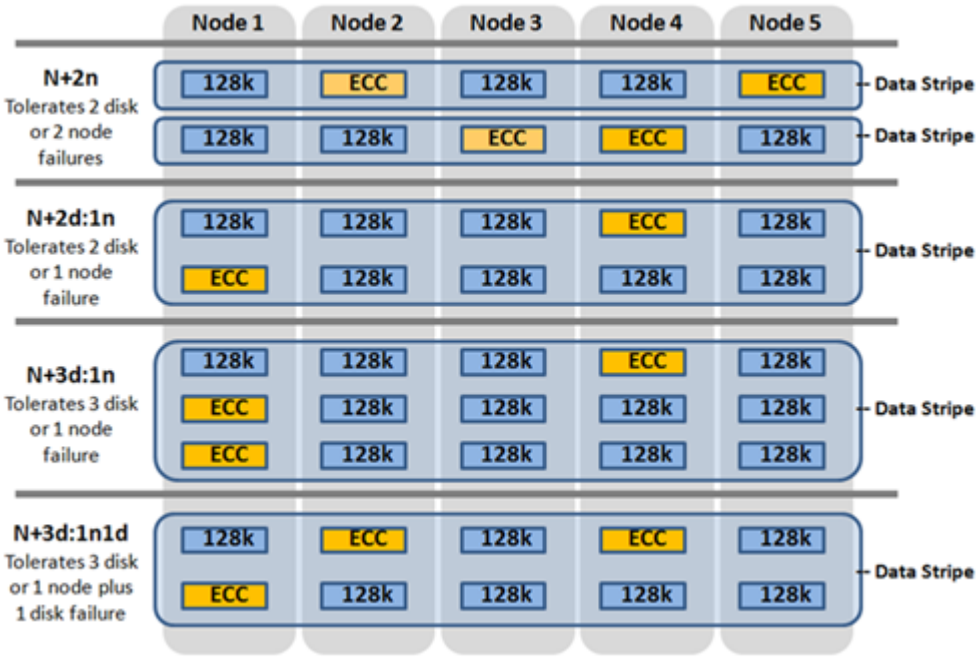

図 16：OneFS ハイブリッド消失訂正符号保護スキーム

ⓘ OneFSでは、新規クラスター インストールの過小保護に関するアラートも提供されます。クラスターが過小保護されている場合、クラスター イベント ログ システム（CELOG）は、アラートを生成して保護の不備を管理者に警告し、その特定のクラスター構成に適した保護レベルに変更するよう推奨します。

📖 詳細については、OneFSの高可用性とデータ保護に関するホワイト ペーパーを参照してください。

自動パーティション分割

OneFSのデータ階層化および管理は、SmartPoolsフレームワークによって処理されます。SmartPoolsでは、データ保護やレイアウト効率性の面から、数多くの大容量、同種ノードを小規模な、「平均データ消失時間」（MTTDL）により適したディスク プールに簡単に分割できます。たとえば、80ノードのH500クラスターは、一般的に+3d:1n1dの保護レベルで実行します。ただし、これを4つの12ノード ディスク プールにパーティション分割すると、各プールを+2d:1nで実行できます。これにより、保護のオーバーヘッドが軽減され、管理オーバーヘッドを増やさずにスペースの使用率を向上させることができます。

ストレージ管理の複雑化を避けるため、OneFSはクラスターを自動的に計算してディスクのプール（ノード プール）にパーティションし、MTTDLとスペース使用効率を最適化します。たとえば、前述のような80ノードのクラスターの場合でも、お客様が保護レベルの決定を行う必要はありません。

自動プロビジョニングでは、互換性のあるノード ハードウェアの全セットが、最大40ノード（1ノードあたり6ドライブ）からなるディスク プールに自動的に分割されます。デフォルトでは、これらのノード プールは+2d:1nの保護で実行されますが、複数のプールを論理階層に統合してSmartPoolsのファイル プール ポリシーで管理できます。ノードのディスクを複数の個別の保護プールに細分化することで、ノードの複数のディスクの同時障害への耐性は従来と比べ大幅に向上します。

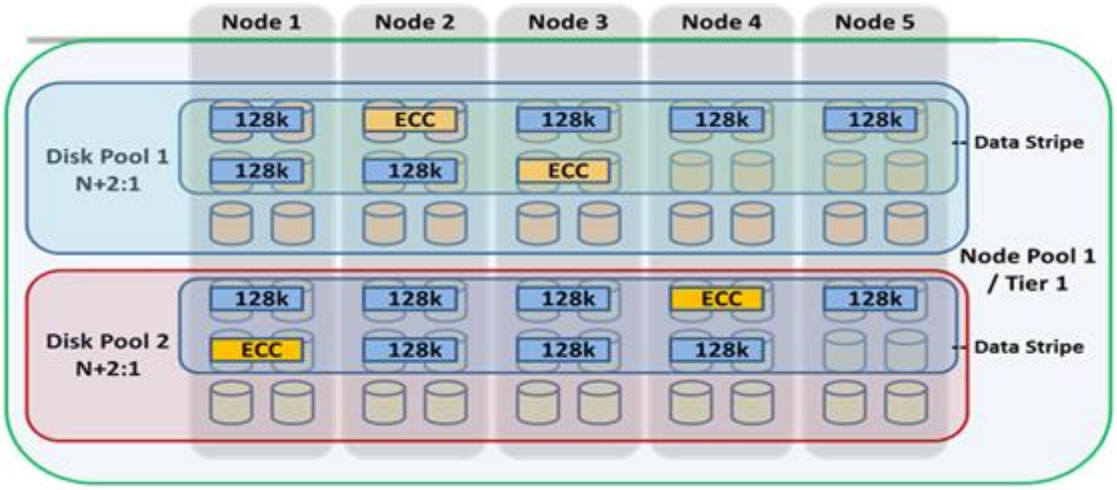

図 17：SmartPools を使用した自動パーティション分割

📖 詳細については、SmartPoolsに関するホワイト ペーパーを参照してください。

PowerScale Gen6モジュラー型ハードウェア プラットフォームは、単一の4RUシャーシにノードを4台搭載する高密度のモジュラー設計を採用しています。このアプローチにより、ディスク プール、ノード プール、「ネイバー」の概念が強化され、OneFS障害ドメインの概念にもう1つの耐障害性レベルが追加されます。各Gen6シャーシには、4台のコンピューティング モジュール（ノードごとに1つ）とノードあたり5台のドライブ コンテナ（スレッド）が搭載されます。

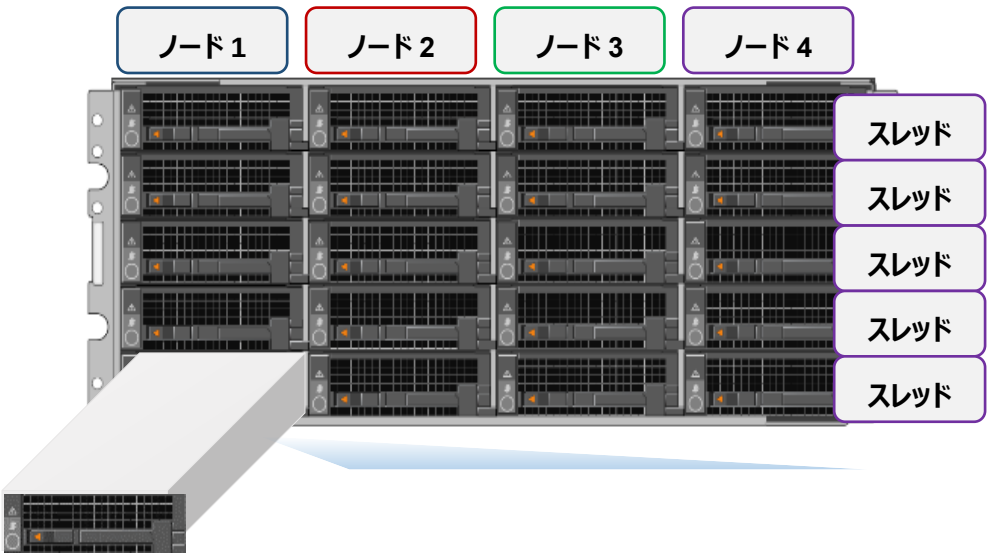

図 18：ドライブ スレッドを示した Gen6 プラットフォーム シャーシの前面図

各スレッドは、特定のシャーシの構成に応じて、シャーシの前面に向かってスライドする、3～6台のドライブを搭載したトレイです。ディスク プールは、ストレージ プール階層内の最小単位です。OneFSのプロビジョニングは、類似するノードのドライブをセット（ディスク プール）に分割し、各プールが別個の障害ドメインを表すことを前提として機能します。これらのディスク プールは、デフォルトでは+2d:1n（2台のドライブまたは1個のノード全体の耐障害性）で保護されます。

ディスク プールは、各Gen6ノードの5つのスレッドすべてに配置されます。たとえば、スレッドあたり3台のドライブを搭載したノードでは、次のディスク プール構成が使用されます。

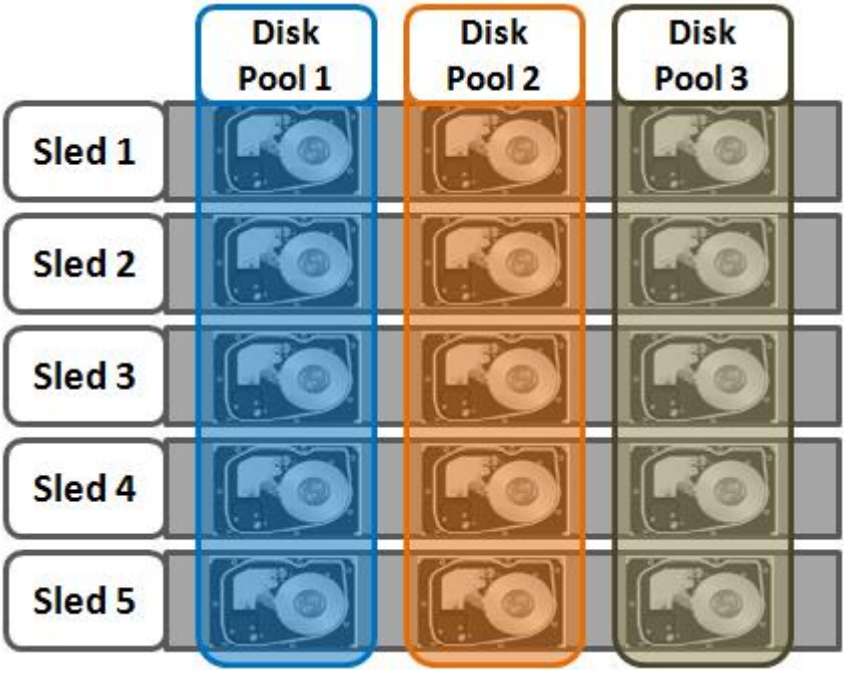

図 19：OneFS ディスク プール

ノード プールは、類似するストレージ ノード（互換性クラス）に及ぶディスク プールのグループです。これを次の図20に示します。ノード タイプが異なる複数のグループが、単一の異機種混在クラスターで連動できます。例えば、Fシリーズ ノード（IOPS集中型アプリケーションで使用）のノード プールが1つ、Hシリーズ ノード（主に高度な並列/シーケンシャル ワークロードで使用）のノード プールが1つ、Aシリーズ ノード（主にニアライン/ディープ アーカイブ ワークロード用）のノード プールが1つです。

これにより、OneFSを、SSD、高速SAS、大容量SATAなどで構成される1つのストレージ リソース プールとすることができ、さまざまなパフォーマンス、保護、容量特性を持たせることができます。つまり、このような異機種混在ストレージ プールは、統合された1つの管理ポイントにより、多様なアプリケーションとワークロードの要件に対応することができます。 また、新旧ハードウェアを混在させることもできるため、数世代の製品にわたっても投資の保護が簡単で、ハードウェアをシームレスに更新できます。

各ノード プールには、同じタイプのストレージ ノードのディスク プールのみが含まれ、ディスク プールが属するノード プールは1つだけとなります。例えば、1つのノード プールに1.6TB SSDドライブがあるFシリーズ ノードに対して、別のノード プールに10 TB SATAドライブがあるAシリーズのようになります。現在、PowerScale H700などのGen6ハードウェアの場合はノード プールごとに最低4ノード（1つのシャーシ）が必要であり、PowerScale F900などの自己完結型ノードの場合はプールごとに3ノードが必要です。

OneFSの「ネイバー」はノード プール内の障害ドメインであり、その目的は、全般的な信頼性を向上させて、ドライブ スレッドを間違って取り外すことによってデータが利用不可になるのを防ぐことにあります。PowerScale F200のような自己完結型ノードの場合、OneFSの最適サイズはノード プールあたり20ノードで、最大サイズは39ノードです。40番目のノードを追加すると、ノードは20ノードからなるネイバー2つに分割されます。

Gen6プラットフォームでは、ネイバーの最適サイズが20ノードから10ノードに変更されています。これにより、ノードペア ジャーナル障害とシャーシ全体の障害が同時に発生するのを防ぐことができます。

パートナー ノードとは、ジャーナルがミラーリングされているノードのことです。Gen6プラットフォームでは、以前のプラットフォームのように各ノードがNVRAMにジャーナルを格納するのではなく、ノードのジャーナルはSSDに格納され、すべてのジャーナルが別のノード上にミラー コピーを持ちます。ミラーリングされたジャーナルを含んでいるノードは、パートナー ノードと呼ばれます。ジャーナルに対する変更からは、いくつかの信頼性のメリットが得られます。たとえば、SSDは、状態を保持するために充電されたバッテリーを必要とするNVRAMよりも永続性と信頼性に優れています。また、ミラーリングされたジャーナルがある場合、両方のジャーナル ドライブが停止していないと、ジャーナルが失われたものとはみなされません。そのため、ミラーリングされたジャーナル ドライブの両方で障害が発生していない限り、どちらのパートナー ノードも正常に機能できます。

パートナー ノードによる保護を使用する場合、ノードは可能な限り複数の異なるネイバー（つまり、複数の異なる障害ドメイン）に配置されます。パートナー ノードによる保護は、クラスターがフル状態の5シャーシ（20ノード）に達すると可能になります。最初のネイバー分割の後、パートナー ノードは複数の異なるネイバーに配置されます。

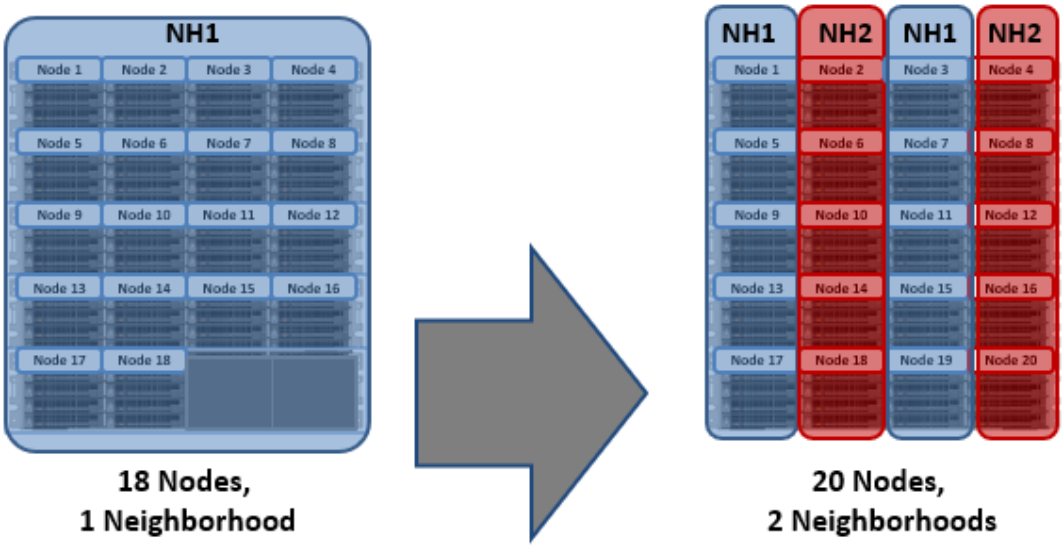

図20.20ノードで2つのネイバーに分割

パートナー ノードによる保護では、信頼性が高まります。なぜならば、両方のノードがダウンした場合に、ノードがそれぞれ異なる障害ドメインに存在するため、単一ノードの消失しか障害ドメインに影響しないからです。

シャーシ保護では、可能な場合、シャーシ内の4つの各ノードが別々のネイバーに配置されます。40ノード目でのネイバー分割によってシャーシ内のすべてのノードを別のネイバーに配置できるようになるため、シャーシ保護はノード数が40になると可能になります。そのため、38ノードのGen6クラスターが40ノードに拡張されると、2つの既存のネイバーが10ノードからなる4ネイバーに分割されます。

シャーシ保護により、シャーシ全体で障害が発生した場合にも、障害ドメインごとに1つのノードしか消失しません。

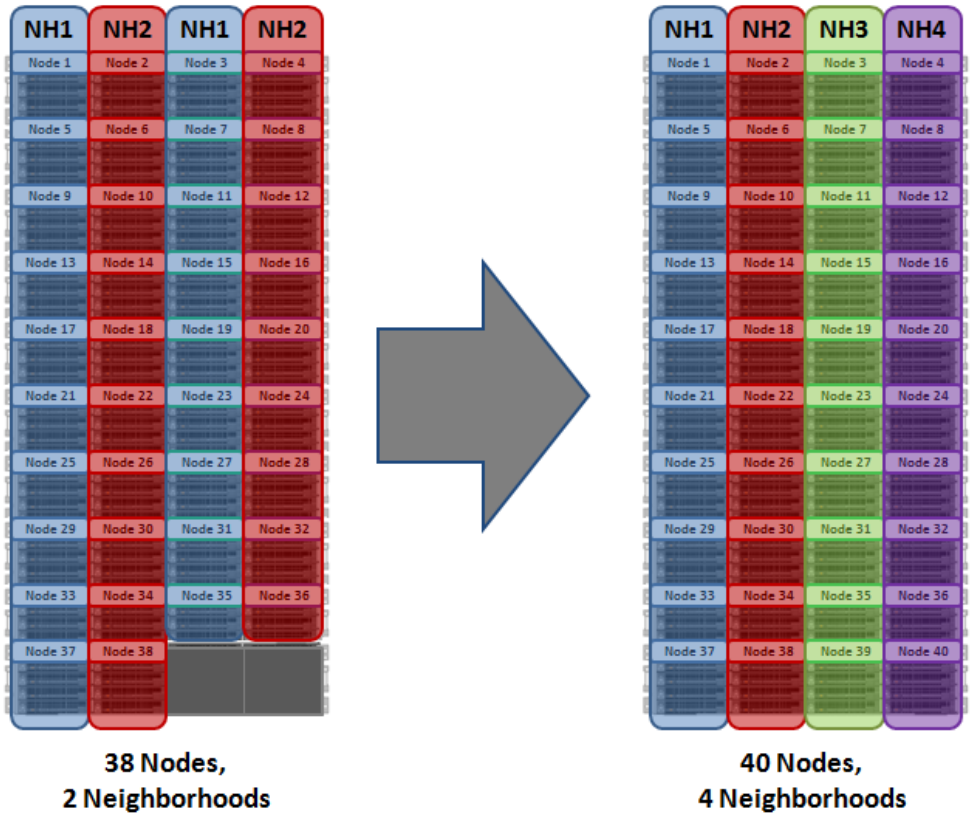

図 21：OneFS ネイバー – 4 つのネイバーに分割

ⓘ 40以上のノードからなるクラスターが4ネイバーを持ち、デフォルト レベルの+2d:1nで保護されている場合、このクラスターは、ネイバーごとに単一ノード障害に耐えることができます。これにより、単一Gen6シャーシの障害からクラスターを保護できます。

全体として、Gen6プラットフォーム クラスターでは、次のような機能拡張の直接的な結果として、同様の容量を持つ旧世代のクラスターよりも信頼性が少なくとも10倍向上します。

- ミラーリングされたジャーナル
- ネイバー サイズの縮小
- ミラーリングされたブート ドライブ

互換性

ノードの互換性がある場合は、類似しているが同一ではないノード タイプを既存のノード プールに対してプロビジョニングすることができます。OneFSでは、1つのノード プールに少なくとも3つのノードを含める必要があります。

ⓘ アーキテクチャに大きな違いがあるため、Gen6プラットフォーム、旧世代のハードウェア、PowerScaleノードの間にはノードの互換性がありません。

OneFSにはSSD互換性オプションも含まれており、SSD容量の異なるノードを1つのノード プールでプロビジョニングすることができます。

SSDとの互換性が作成されると「OneFS WebUI SmartPools Compatibilities」リストに表示され、「Tiers & Node Pools」リストにも表示されます。

ⓘ このSSDとの互換性を作成する場合、OneFSは、マージされる2つのプールに同じ数のSSD、階層、要求された保護、L3キャッシュの設定があることを自動的に確認します。これらの設定が異なる場合は、OneFS WebUIに、これらの設定の統合と調整を求めるプロンプトが表示されます。

📖 詳細については、SmartPools に関するホワイト ペーパーを参照してください。

サポートされるプロトコル

適切な認証情報や権限を持つクライアントは、クラスターと通信するために、次の標準の方法の1つを使用して、データの作成、変更、および読み取りを実行できます。

- NFS（ネットワーク ファイル システム）
- SMB/CIFS（Server Message Block/Common Internet File System）
- FTP（File Transfer Protocol）
- HTTP（Hypertext Transfer Protocol）
- HDFS（Hadoop Distributed File System）
- REST API（Representational State Transfer Application Programming Interface）
- S3（オブジェクト ストレージAPI）

NFSプロトコルの場合、OneFSはNFSv3とNFSv4の両方に加え、OneFS 9.3のNFSv4.1をサポートします。さらに、OneFS 9.2以降にはNFSv3overRDMAのサポートが含まれています。

Microsoft Windows側では、SMBプロトコルがバージョン3までサポートされます。SMB3ダイアレクトの一部として、OneFSでは次の機能がサポートされています。

- SMB3マルチパス
- SMB3継続的可用性と監視
- SMB3暗号化

SMB3暗号化は、共有単位、ゾーン単位、またはクラスター全体で構成できます。暗号化された共有に対応できるのは、SMB3暗号化をサポートしているオペレーティング システムのみです。これらのオペレーティング システムは、暗号化されていない接続を許可するようにクラスターが構成されている場合に、暗号化されていない共有にも対応できます。その他のオペレーティング システムは、暗号化されていない接続を許可するようにクラスターが構成されている場合に限り、暗号化されていない共有にアクセスできます。

クラスター内のすべてのデータのファイル システム ルートは/ifs（OneFSファイル システム）です。これはSMBプロトコルでは「ifs」共有(\\<cluster_name\ifs)、NFSプロトコルでは「/ifs」エクスポート(<cluster_name>:/ifs)として表すことができます。

ⓘ データはすべてのプロトコル間で共通であるため、ある1つのアクセス プロトコルによってファイルの内容に加えた変更は、他のどのプロトコルからもただちに確認できます。

OneFSは、フロントエンドEthernetネットワーク、SmartConnect、ストレージ プロトコルと管理ツールの完全なアレイにわたってIPv4とIPv6の両方の環境を完全にサポートします。

また、OneFS CloudPoolsは次のクラウド プロバイダーのストレージAPIをサポートしており、以下をはじめとした多数のストレージ ターゲットにファイルをスタブ化できます。

- Amazon Web Services S3
- Microsoft Azure
- Google Cloud Service
- Alibaba Cloud
- Dell EMC ECS
- OneFS RAN (RESTful Access to Namespace)

📖 詳細については、CloudPools管理ガイド**を参照してください。**

無停止の操作 - プロトコル サポート

OneFSは、LinuxおよびUNIXクライアントの場合は動的なNFSv3およびNFSv4フェールオーバーとフェールバックをサポートし、Windowsクライアントの場合はSMB3の継続的な可用性をサポートすることによって、データの可用性に貢献しています。これにより、ノードに障害が発生した場合または予防メンテナンスが実行された場合に、処理中のすべての読み取り/書き込みをクラスターの別のノードに引き渡し、ユーザーやアプリケーションを中断することなく操作を完了できます。

フェイルオーバー中、クライアントはパフォーマンス インパクトを最小限に抑えて、クラスター内の残りのすべてのノードに均等に再分配されます。ノードが、障害を含めて何らかの理由でシャットダウンされると、そのノードの仮想IPが、クラスター内にある別のノードにシームレスに移行されます。

オフライン ノードがオンラインに戻ると、SmartConnectは、ストレージとパフォーマンスの使用率を最大限に確保するために、クラスター全体のNFSクライアントとSMB3クライアントを自動的に再調整します。定期的なシステム保守やソフトウェア アップデートの場合、この機能を使用するとノードごとのローリング アップグレードが可能になるため、保守期間全体で可用性が完全に確保されます。

ファイル フィルタリング機能

OneFSのファイル フィルタリング機能をNFSおよびSMBクライアントで使用して、エクスポート、共有、アクセス ゾーンへの書き込みを許可または禁止することができます。この機能では、セキュリティの問題、生産性の低下、スループットの問題、ストレージのクラッターを引き起こす可能性のあるファイルについて、特定タイプのファイル拡張子をブロックできます。設定には、明示的なファイル拡張子をブロックする除外リストを使用するか、特定のファイル タイプのみの書き込みを明示的に許可する包含リストを使用します。

データ重複排除 - SmartDedupe

SmartDedupe製品は、組織のデータを格納するために必要な物理ストレージの量を減らすことによってクラスターのストレージ効率を最大限に高めます。ディスク上のデータに同一のブロックがないかをスキャンし、重複を排除することで効率性を確保します。この方法は一般にポスト プロセス重複排除または非同期重複排除と呼ばれています。

SmartDedupeは重複するブロックを検出すると、これらブロックの片方のみをシャドウ ストアと呼ばれる特殊なファイル セットに移動します。このプロセスで、重複するブロックは実際のファイルから削除され、シャドウ ストアを参照するポインターに置き換えられます。

ポスト プロセス重複排除では、新しいデータは最初にストレージ デバイスに保存され、それに続くプロセスでデータの共通性が分析されます。つまり、書き込みパスで追加の計算が必要ないため、ファイルの最初の書き込みまたはファイル変更のパフォーマンスは影響を受けません。

SmartDedupeアーキテクチャ

OneFS SmartDedupeのアーキテクチャは次の主要モジュールで構成されます。

- 重複排除制御パス
- 重複排除ジョブ
- 重複排除エンジン
- シャドウ ストア
- 重複排除インフラストラクチャ

SmartDedupeの制御パスは、OneFS Web管理インターフェイス（WebUI）、コマンド ライン インターフェイス（CLI）、RESTfulプラットフォームAPIで構成され、重複排除ジョブの構成管理、スケジュール設定、制御を行います。ジョブそのものは高度に分散されたバックグラウンド処理で、クラスター内のすべてのノード間で重複排除の調整を管理します。ジョブの制御にはファイル システムで一致するデータ ブロックのスキャン、検出、および共有が含まれ、これは重複排除エンジンとあわせて実行されます。重複排除インフラストラクチャ レイヤーは、共有データ ブロックをシャドウ ストアに統合するカーネル モジュールです。シャドウ ストアはファイル システム コンテナであり、物理データ ブロックとともに、共有ブロックへの参照（ポインタ）を保持します。各要素について、以下で詳しく説明します。

図 22：OneFS SmartDedupe のモジュラー型アーキテクチャ

📖 詳細については、OneFS SmartDedupeに関するホワイト ペーパーを参照してください。

シャドウ ストア

OneFSのシャドウ ストアは、データを共有可能な形で保存できるファイル システム コンテナーです。このため、OneFS上のファイルには、物理データ ファイルまたは共有ブロックへのポインタ（参照）をシャドウ ストアに含めることができます。

シャドウ ストアは通常のファイルと似ていますが、一般に、通常のファイルのinodeに関連するメタデータをすべて含むことはありません。特に、時間ベースの属性（作成時刻、変更時刻など）は明示的に除外されます。各シャドウ ストアは最大256ブロックを含むことができ、各ブロックは32,000のファイルからの参照が可能です。参照の上限である32,000個を超えると、新しいシャドウ ストアが作成されます。さらに、シャドウ ストアは他のシャドウ ストアを参照しません。また、シャドウ ストアにはハード リンクがないため、シャドウ ストアのスナップショットは許可されません。

ⓘ シャドウ ストアは重複排除に加えて、OneFSファイル クローンとSmall File Storage Efficiency (SFSE)にも使用されます。

Small File Storage Efficiency

シャドウ ストアを使用するもう1つの主要機能は、OneFS Small File Storage Efficiencyです。この機能は、一般的な医療用のPACSワークフローなど、アーカイブ データセットを構成することが多い小容量のファイルを格納するために必要な物理ストレージの量を削減することで、クラスターのスペース使用率を最大化します。

効率性を実現するには、フル コピーのミラーによって保護されている小容量ファイルのディスク上のデータをスキャンし、シャドウ ストアに格納します。これらのシャドウ ストアは、ミラーリングするのではなくパリティ保護され、通常は80%以上のストレージ効率を実現します。

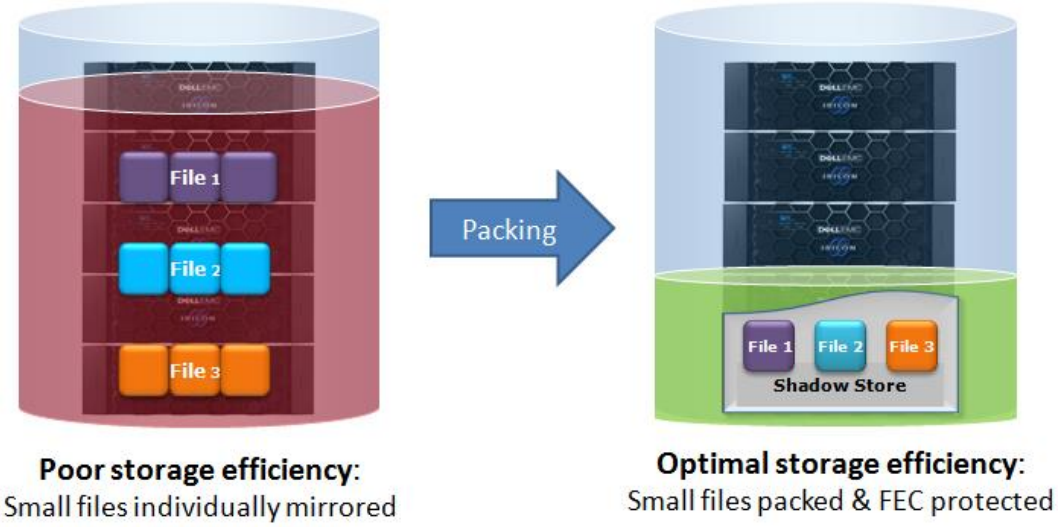

図 23：小容量ファイルのコンテナ化

Small File Storage Efficiencyでは、ストレージ使用率を向上させるために、読み取りレイテンシーのパフォーマンスがわずかに低下します。アーカイブされたファイルは当然書き込み可能なままですが、シャドウ参照のあるコンテナ化されたファイルが削除、トランケート、上書きされると、未参照のブロックがシャドウ ストアに残ることがあります。これらのブロックは後で解放されるため、穴が開いてストレージ効率が低下する可能性があります。

実際の効率の低下は、シャドウ ストアで使用される保護レベルのレイアウトによって異なります。 コンテナ内のすべてのブロックには、最大1つの参照ファイルがあり、パックされたサイズ（ファイル サイズ）が小さいため、コンテナ化されたファイルと同様に、保護グループのサイズが小さいほど、より影響を受けやすくなります。

上書きや削除の結果として生じるファイルのフラグメンテーションを軽減するために、デフラグツールが用意されています。このシャドウ ストア デフラグツールは、ShadowStoreDeleteジョブに統合されています。デフラグメンテーション処理は、各コンテナ ファイルを論理チャンク（それぞれ約32 MBずつ）に分割し、各チャンクのフラグメンテーションを評価することで機能します。

フラグメント化されたチャンクのストレージ効率がターゲットを下回ると、そのチャンクはデータを別の場所に退避することによって処理されます。デフォルトのターゲットの効率は、シャドウ ストアによって使用される保護レベルで使用可能な最大ストレージ効率である90%です。保護グループのサイズを大きくすると、ストレージの効率がこの閾値を下回る前に、より高いレベルのフラグメント化を許容できます。

インライン データ削減

OneFSインライン データ削減は、F900、F810、F600、F200オールフラッシュ ノード、H700/7000およびH5600ハイブリッド シャーシ、A300/3000アーカイブ プラットフォームで使用できます。OneFSアーキテクチャは、次の主要なコンポーネントで構成されます。

- データ削減プラットフォーム
- 圧縮エンジンとチャンク マップ

- ゼロ ブロックの削除フェーズ
- 重複排除インメモリー インデックスとシャドウ ストア インフラストラクチャ
- データ削減に関するアラートとレポート作成フレームワーク
- データ削減制御パス

インライン データ削減の書き込みパスは、次の 3 つの主要なフェーズで構成されます。

- ゼロ ブロックの削除
- インライン重複排除
- インライン圧縮

インライン圧縮と重複排除の両方がクラスターで有効になっている場合は、最初にゼロ ブロック削除の実行、続いて重複排除、その後に圧縮が行われます。この順序によって、各フェーズで各後続フェーズの作業範囲を低減することができます。

図 24：インライン データ削減ワークフロー

F810 にはハードウェア圧縮オフロード機能が含まれており、F810 シャーシ内の各ノードに Mellanox Innova-2 Flex アダプターが搭載されています。つまり、圧縮と解凍は最小限のレイテンシーで Mellanox アダプターによって透過的に実行されるため、ノードの高価な CPU とメモリー リソースを消費する必要がありません。

PowerScale F900、F810、F600、F200、H700/7000、H5600、A300/3000 ノードの場合、igzip のソフトウェア実装とともに、OneFS ハードウェア圧縮エンジンは zlib を使用します。ソフトウェア圧縮は、圧縮ハードウェア障害が発生した場合のフォールバックとしても使用されます。ハードウェア圧縮機能のない F810 以外のノードの混在クラスターでは、圧縮ハードウェア障害が発生した場合のフォールバックとして使用されます。OneFS では 128KB の圧縮チャンク サイズが採用され、各チャンクは 8KB のデータ ブロック 16 個で構成されます。これは、OneFS がデータ保護ストライプ ユニットに使用するのと同じサイズであり、追加のチャンク パッキングのオーバーヘッドを回避することでシンプルさと効率性を実現するため、最適です。

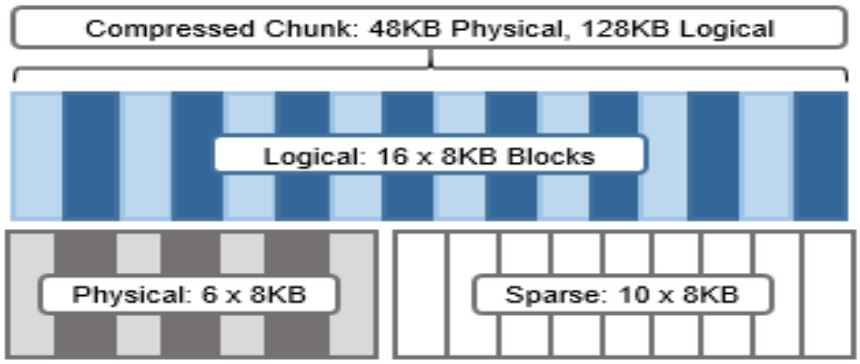

図 25：圧縮チャンクと OneFS の透過的なオーバーレイ

前の図をご覧ください。圧縮後、このチャンクのサイズは、16 個から 6 個の 8KB のブロックに削減されます。これは、このチャンクのサイズが現在、物理的に 48KB であることを意味します。OneFS は、物理属性に対して透過的な論理オーバーレイを提供します。このオーバーレイは、ファイル システムのコンシューマーが圧縮の影響を受けないように、バッキング データが圧縮されているかどうか、およびチャンク内のどのブロックが物理またはスパースであるかを示します。このように、圧縮されたチャンクは、実際の物理サイズに関係なく、論理的に 128KB として表されます。

効率化による節約には、圧縮を実行するために少なくとも 8KB（1 つのブロック）が必要です。そうでない場合、そのチャンクまたはファイルが引き継がれ、元の圧縮不可能な状態のままになります。たとえば、8KB のファイル（1 つのブロック）の削減になる 16KB のファイルは圧縮されます。ファイルは圧縮された後、FEC によって保護されます。

圧縮チャンクが複数のノード プールにまたがることはありません。これにより、データを解凍または再圧縮して保護レベルを変更したり、リカバリー済みの書き込みを実行したり、保護グループの境界をシフトさせたりする必要がなくなります。

ダイナミックな拡張/オン デマンドの拡張

パフォーマンスと容量

パフォーマンスや容量の追加が必要なときには「スケールアップ」する以外に方法がない従来のストレージ システムとは異なり、ストレージ システムでは、OneFSによって既存のファイル システムまたはボリュームをペタバイト規模の容量にシームレスに拡張すると同時に、直線的にパフォーマンスを向上させる「スケールアウト」が可能です。

クラスターに容量やパフォーマンス能力を追加することは他のストレージ システムに比べてはるかに簡単です。ストレージ管理者が従う必要があるステップは、ラックに別のノードを追加する、バックエンド ネットワークにノードを追加する、クラスターに別のノードを追加するように指示するの3つだけです。各ノードにはCPU、メモリー、キャッシュ、ネットワーク、NVRAMおよびI/O制御パスウェイが組み込まれているため、ノードを追加するだけで容量を増やし、同時にパフォーマンスを向上させることができます。

OneFSのAutoBalance機能は、バックエンド ネットワーク全体にわたって、一貫性のある方法でデータを自動的に移動し、クラスター内の既存データを新しいストレージ ノードに移動させます。この自動再バランシングにより、新しいノードが新しいデータのホット スポット（負荷の高い部分）になることを回避し、既存のデータはより強力なストレージ システムのメリットを享受できます。また、OneFSのAutoBalance機能は、エンド ユーザーにとって完全に透過的であり、高パフォーマンスのワークロードへの影響を最小限に抑えるように調整できます。OneFSではこの機能だけで、管理者の管理時間が長くなったりストレージ システム内の複雑さが増すことなく、TBからPBまでのスケールアウトを透過的にオンザフライで実行できます。

大規模なストレージ システムでは、ワークフローがシーケンシャル、コンカレント、ランダムのいずれであっても、さまざまなワークフローで求められるパフォーマンスを提供できる必要があります。アプリケーション間や個別のアプリケーション内に、さまざまなワークフローが存在します。OneFSは、インテリジェント ソフトウェアによってそれらすべてのニーズに対応します。さらに重要なのは、OneFSでは、スループットおよびIOPSは、単一システム内に存在するノードの数に合わせて直線的に拡張されます。バランシングされたデータ分散、自動リバランシング、分散処理により、OneFSでは、システムの拡張に合わせて、追加のCPU、ネットワーク ポート、メモリを利用できます。

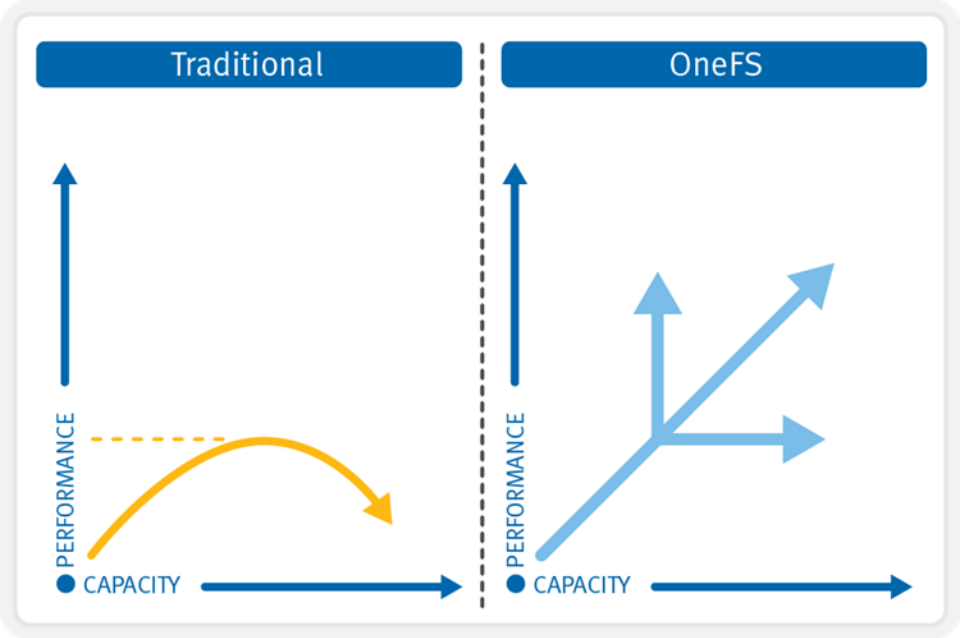

図 26：OneFS の直線的な拡張性の向上

Interfaces

管理者は環境内でストレージ クラスターを管理するために、複数のインターフェイスを使用できます。

- Web管理ユーザー インターフェイス（「WebUI」）
- SSHネットワーク アクセスまたはRS232シリアル接続を使用したコマンド ライン インターフェイス
- ノード自体のLCDパネル（機能の追加/削除を簡単に実行可能）
- RESTfulプラットフォームAPI（クラスターの構成と管理をプログラムによって制御および自動化することが可能）

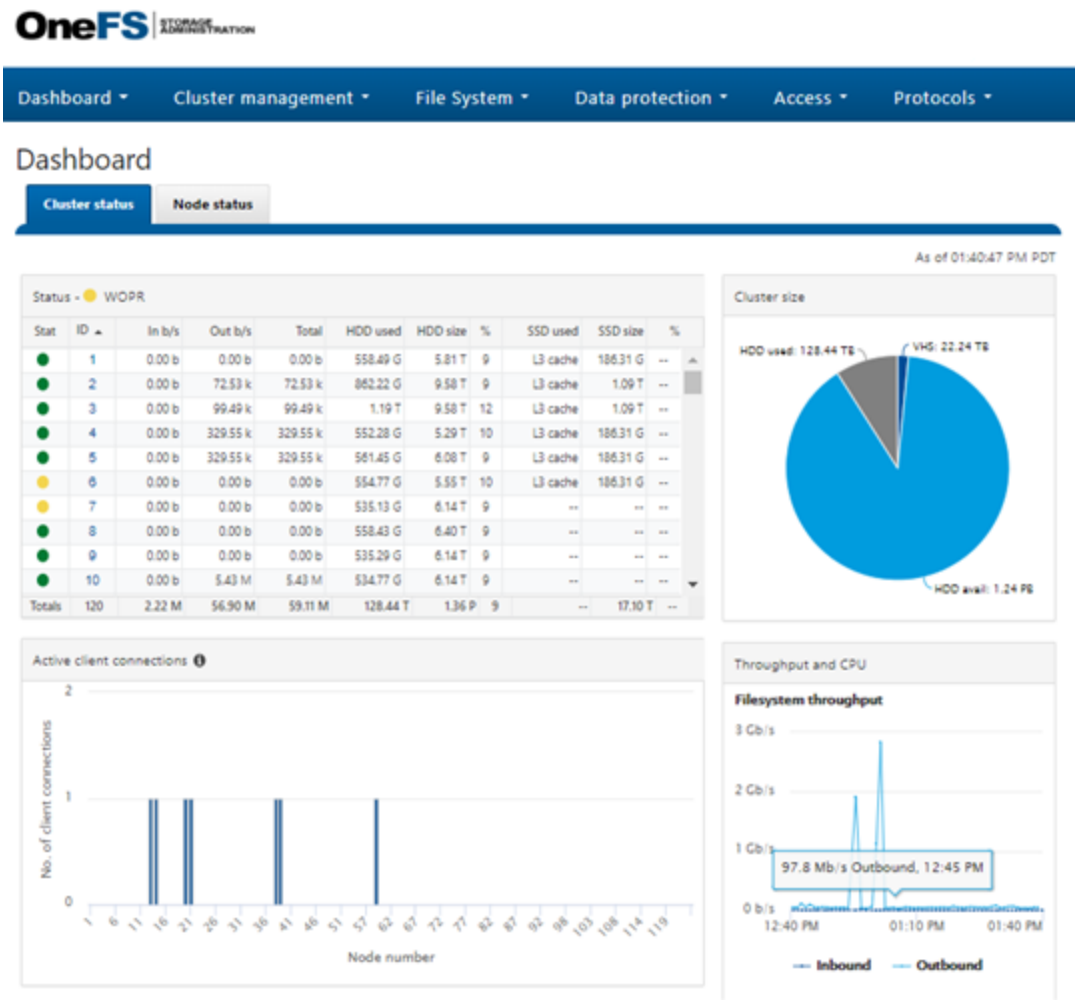

図 27：OneFS Web ユーザー インターフェイス

📖 OneFSのコマンドおよび機能の構成に関する詳細については、OneFS管理ガイドを参照してください。

認証とアクセス制御

認証サービスは、ユーザーがファイルにアクセスして変更を加える前にユーザーの認証情報を検証するセキュリティ レイヤーを提供します。OneFSは、次の4通りのユーザー認証方法をサポートしています。

- AD（Active Directory）
- LDAP（Lightweight Directory Access Protocol）
- NIS（Network Information Service）
- ローカル ユーザーおよびグループ

OneFSは、複数の認証タイプの使用をサポートしています。ただし、クラスターで複数の認証方法を有効にするときは、事前に各認証タイプの相互作用を十分に理解することをお勧めします。複数の認証モードを適切に構成する方法について詳しくは、製品ドキュメントを参照してください。

Active Directory

Active Directoryは、Microsoftが開発したLDAP実装であり、ネットワーク リソースの情報を格納できるディレクトリ サービスです。Active Directoryには多くの機能がありますが、クラスターをドメインに参加させる主な理由は、ユーザー認証およびグループ認証を行うためです。

クラスターのActive Directory設定は、Web管理インターフェイスまたはコマンド ライン インターフェイスを使用して構成および管理できますが、可能な限りWeb管理インターフェイスを使用することをお勧めします。

クラスターの各ノードは、同じActive Directoryマシン アカウントを共有しているため、管理が非常に簡単です。

LDAP

LDAP（Lightweight Directory Access Protocol）は、サービスおよびリソースの定義、クエリー、変更を行うために使用するネットワーキング プロトコルです。LDAPの主なメリットは、ディレクトリ サービスがオープンであり、LDAPが多くのプラットフォームでサポートされていることです。クラスター化されたストレージ システムでは、LDAPを使用してユーザーおよびグループを認証してクラスターへのアクセスを許可できます。

NIS

NIS (Network Information Service)は、Sun Microsystemsによって設計されたディレクトリー サービス プロトコルです。OneFSはNISを使って、クラスターにアクセスするユーザーおよびグループを認証できます。NISはYP（イエロー ページ）と呼ばれることもあり、OneFSがサポートしていないNIS+とは異なります。

ローカル ユーザー

OneFSは、ローカル ユーザーおよびグループ認証をサポートします。ローカル ユーザー アカウントやグループ アカウントは、Web管理インターフェイスを使用して、クラスターに直接作成できます。ローカル認証は、Active Directory、LDAP、NISといったディレクトリ サービスを使用しない場合、または特定のユーザーやアプリケーションがクラスターにアクセスする必要がある場合に役立ちます。

アクセス ゾーン

アクセス ゾーンは、クラスター アクセスを論理的に区分化してリソースを自己完結型ユニットに割り当てる方法を提供し、共有テナント（またはマルチテナント）環境を実現します。これを容易にするために、アクセス ゾーンはコアとなる次の3つの外部アクセス コンポーネントを結合します。

- クラスター ネットワーク構成
- ファイル プロトコル アクセス
- 認証

このようにして、SmartConnectゾーンはゾーンごとにSMB共有、NFSエクスポート、HDFSラック、アクセス制御を提供する1つ以上の認証プロバイダーのセットに関連づけられます。これにより、複数のテナントを対象にプロビジョニングとセキュリティ保護を行うことのできる、一元管理の単一ファイル システムのメリットがもたらされます。これは、複数の異なるビジネス ユニットが中央のIT部門によって管理されるエンタープライズ環境では、特に便利です。もう1つの例は、サーバー統合イニシアティブ中に、信頼されない別個のActive Directoryフォレストに参加している複数のWindowsファイル サーバーをマージするときです。

アクセス ゾーンでは、組み込みのシステム アクセス ゾーンに、サポート対象の各認証プロバイダーのインスタンス、利用可能なすべてのSMB共有、利用可能なすべてのNFSエクスポートがデフォルトで含まれています。

これらの認証プロバイダーには、Microsoft Active Directory、LDAP、NIS、ローカル ユーザー データベースまたはグループ データベースの複数のインスタンスを含めることができます。

役割ベースの管理

役割ベースの管理は、「ルート」および「管理者」ユーザーの権限をより細かい特権に分割し、これらの特権を特定の役割に割り当てることが可能なクラスター管理のRBAC（役割に基づいたアクセス制御）システムです。これらの役割は、他の権限を持たないユーザーに割り当てることができます。たとえば、データセンターのオペレーション スタッフに全クラスターに対する読み取り専用権限を割り当てた場合、これらのスタッフはフル モニタリング アクセスが可能ですが、構成は変更できません。OneFSには、監査、システムおよびセキュリティ管理者などの一連の組み込みの役割が用意されているほか、アクセス ゾーン単位またはクラスター全体でカスタムの役割を定義することもできます。役割ベースの管理はOneFSコマンド ライン インターフェイス、WebUI、プラットフォームAPIに統合されています。

📖 マルチプロトコル環境におけるID管理、認証、アクセス制御の詳細については、『OneFSマルチプロトコル セキュリティ ガイド』を参照してください。

OneFSの監査

OneFSには、クラスター上のシステム構成、NFS、SMB、HDFSプロトコル アクティビティを監査する機能が用意されています。これにより組織は、データ ガバナンスとコンプライアンスのさまざまな要件を満たすことができます。

すべての監査データはクラスターのファイル システム内に格納および保護され、監査トピックごとに整理されます。ここからは、Dell EMC Common Event Enabler（CEE）フレームワークを介して、Varonis DatAdvantageやSymantec Data Insightなどのサード パーティ アプリケーションに監査データをエクスポートできます。OneFSプロトコル監査はアクセス ゾーンごとに有効にすることができ、クラスター全体できめ細かく制御できます。

ノードあたり最大5つのCEEサーバーへのクラスターによる監査イベントの書き込みを、負荷分散された並列構成で行うことができます。これにより、OneFSでエンタープライズ グレードの監査ソリューションをエンドツーエンドで実現できます。

📖 詳細については、OneFS監査に関するホワイト ペーパーを参照してください。

ソフトウェアのアップグレード

OneFSの最新バージョンにアップグレードすると、新しい機能やバグ修正を利用できます。クラスターは、同時アップグレードまたはローリング アップグレードの2つの方法でアップグレードできます。

同時アップグレード

同時アップグレードは、新しいオペレーティング システムのインストールとクラスター内のすべてのノードの再起動を同時に実行します。同時アップグレードを実行すると、ノードが再起動されるまでの間、一時的なサービスの中断（2分程度）が発生します。

ローリング アップグレード

ローリング アップグレードでは、クラスター内の各ノードを個別にアップグレードし、順番に再起動します。ローリング アップグレード中もクラスターはオンラインのままで、サービスの中断が発生することなくデータをクライアントに提供します。OneFS 8.0より前では、ローリング アップグレードは、OneFSの単一のコード バージョン ファミリー内でのみ実行可能で、OneFSのメジャー コード バージョンのリビジョン間では実行できません。OneFS 8.0以降では、すべての新しいリリースが以前のバージョンからローリング アップグレードできるようになります。

無停止アップグレード

NDU（無停止アップグレード）を使用すると、クラスター管理者は、エンド ユーザーがエラーや中断なくデータにアクセスしている間に、ストレージOSをアップグレードできます。クラスター上のオペレーティング システムを更新すると、ローリング アップグレードを簡単に行うことができます。このプロセスでは、一度に1つのノードが新しいコードにアップグレードされ、そこに接続されているアクティブなNFSおよびSMB3クライアントがクラスター内の他のノードに自動的に移行されます。部分アップグレードも許可されており、これによってクラスター ノードのサブセットをアップグレードできます。また、アップグレード中にノードのサブセットを拡張することもできます。アップグレードを一時停止して再開できるため、お客様は複数の短い保守期間にわたってアップグレードを分散できます。さらに、OneFS 8.2.2以降では並列アップグレードが提供されます。これにより、クラスターはネイバー全体（障害ドメイン）を一度にアップグレードでき、大規模なクラスターのアップグレードにかかる期間が大幅に短縮されます。OneFS 9.2以降では、OSとファームウェアのアップグレードを組み合わせて、アップグレードを並行して実行できるようにすることでアップグレードの影響と期間が大幅に削減されます。9.2以降にはドレイン ベースのアップグレードも含まれます。これにより、すべてのSMBクライアントがノードから切断されるまで、ノードがプロトコル サービスをリブートまたは再起動するのを防ぎます。

ロールバック対応

OneFSは、アップグレード ロールバックをサポートしています。これにより、コミットされていないアップグレードを含んだクラスターを以前のバージョンのOneFSに戻すことができます。

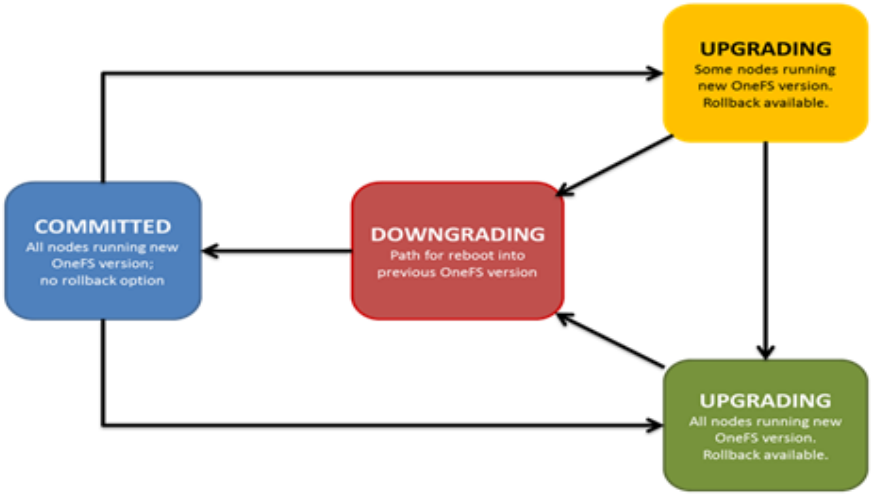

図 28：OneFS のスムーズなアップグレードの状態

自動ファームウェア アップデート

OneFS搭載のクラスターは、スムーズなファームウェア アップデート プロセスの一環として、新規および交換用ドライブの自動ドライブ ファームウェア アップデートをサポートしています。ファームウェア アップデートは、ドライブ サポート パッケージを介して提供されます。これにより、クラスター全体で既存のドライブと新規ドライブの管理がシンプルになると同時に、合理化されます。その結果、ドライブのファームウェアが最新の状態になり、既知のドライブの問題を原因とした障害が発生する可能性を低減できます。このように、自動ドライブ ファームウェア アップデートは、OneFSの高可用性および無停止操作戦略を構成する重要な要素です。ドライブとノードのファームウェアは、ローリング アップグレードとしても、クラスター全体を再起動することによっても適用できます。

OneFS 8.2以前のノード ファームウェア アップデートは一度に1つのノードにインストールする必要があり、特に大規模なクラスターでは時間のかかる操作でした。同時にアップデートするノードのリストを提供することで、ノード ファームウェア アップデートをクラスター全体で計画できるようになりました。アップグレード ヘルパー ツールを使用すると、同時にアップデートできるノードの任意の組み合わせと、同時にアップデートすべきでないノード（たとえば、ノードペアに含まれるノード）の明示的なリストを選択することができます。

アップグレードの実行

アップグレードの一環として、OneFSはインストール前検証チェックを自動的に実行します。このチェックは、現在のOneFSのインストール構成がアップグレードするOneFSのバージョンと互換性があるかどうかを検証します。サポート対象外の構成が見つかった場合は、アップグレードが停止し、問題を解決するための手順が表示されます。アップグレードを開始する前にインストール前アップグレード チェック スクリプトを実行しておくと、構成に互換性がないためにアップグレードが中断する事態を防ぐことができます。

OneFSデータ保護および管理ソフトウェア

OneFSは、ユーザーのニーズに応えるデータ保護および管理ソフトウェアの包括的なポートフォリオを提供します。

ソフトウェア モジュール	機能	説明
CloudIQ™	クラスターの正常稼働モニタリング	クラスターの正常稼働をプロアクティブに監視するためのインテリジェントな予測分析を実施します。
InsightIQ™	パフォーマンス管理	革新的なパフォーマンス監視/レポート作成ツールを使用して、クラスターのパフォーマンスを最大化します。
DataIQ™	データ分析と管理	データがファイルとオブジェクト ストレージ、オンプレミスまたはクラウド内のどこにあっても、数秒でデータを検索、アクセス、管理できます。一元化された管理ポイントで異機種混在ストレージ システム全体にわたって総合的に表示できるため、事実上、サイロ化したデータが解消されます。
SmartPools™	リソース管理	非常に効率的で自動化された階層型ストレージを導入して、ストレージのパフォーマンスとコストを最適化します。
SmartQuotas™	データ管理	ストレージを管理の容易なセグメントにシームレスに分割してシン プロビジョニングするために、クォータを割り当てて管理します。ストレージの分割はクラスター、ディレクトリ、サブディレクトリ、ユーザー、グループの各レベルで行うことができます。
SmartConnect™	データ アクセス	ストレージ ノード全体にわたりクライアント接続のロード バランシングと動的NFSフェイルオーバー/フェイルバックを可能にして、クラスター リソースの利用を最適化します。
SnapshotIQ™	Data Protection	安全でほぼ即時のスナップショットを、パフォーマンスのオーバーヘッドをほとんど発生させることなく作成し、データを効率的かつ確実に保護します。ほぼ即時のオン デマンドのスナップショット リストアによって重要なデータを迅速にリカバリします。OneFSの書き込み可能なスナップショットを使用して、読み取り専用スナップショットのスペースと時間効率に優れた変更可能なコピーを作成します。
SyncIQ™	データ レプリケーション	大規模でミッション クリティカルなデータセットを複数の場所にある複数の共有ストレージ システムにレプリケーションして非同期的に分散することで、信頼性の高い災害復旧機能を提供します。ミッション クリティカルなデータの可用性を高めるために、プッシュボタン式のシンプルなフェイルオーバーとフェイルバック機能を備えています。
SmartLock™	データ保持	IsilonのソフトウェアベースのWORM（Write Once Read Many）アプローチによって過失、または故意の変更/削除から重要なデータを保護し、SEC 17a-4要件などの厳格なコンプライアンスおよびガバナンス ニーズを満たします。
SmartDedupe™	データ重複排除	クラスターをスキャンして同一ブロックを検出してから重複を排除し、必要な物理ストレージの量を減らすことによって、ストレージ効率を最大限に高めます。
CloudPools™	クラウド階層化	CloudPoolsにより、クラスター上のどのデータをクラウド ストレージにアーカイブするかを定義できます。クラウド プロバイダーには、Microsoft Azure、Google Cloud、Amazon S3、Dell EMC ECS、ネイティブOneFSが含まれます。

表 3：Dell EMC PowerScale データ サービス ポートフォリオ

詳細については、製品マニュアルを参照してください。

まとめ

OneFSオペレーティング システムを搭載したDell EMCスケールアウトNASソリューションにより、組織は単一のファイル システム、単一ボリューム、単一の管理ポイントを使用して、TBからPBまで拡張できるようになります。OneFSでは、管理の複雑さが増すことなく、高パフォーマンス、高スループット、またはその両方が実現します。

持続可能な拡張性のためには、次世代データセンターを構築する必要があります。次世代データセンターは、自動化機能やハードウェアのコモディティ化を活用し、ネットワーク ファブリックを完全に消費します。また、変化し続ける需要を満たすという企業の目的を達成するために、最大レベルの柔軟性を提供します。

OneFSは、これらの課題に立ち向かうための次世代ファイル システムです。OneFSは以下を実現します。

- 完全分散型の単一ファイル システム
- 高パフォーマンス、完全対称のクラスター
- クラスター内のすべてのノードにわたるファイル ストライピング
- 自動化されたソフトウェアによる複雑さの排除
- 動的コンテンツ バランシング
- 柔軟なデータ保護
- 高可用性
- Webベースのコマンド ライン管理

OneFSは、エンタープライズ データ レイク環境のファイル ベースおよび非構造化「Big Data」アプリケーション（大規模なホーム ディレクトリ、ファイル共有、アーカイブ、仮想化、ビジネス分析など）に最適です。また、データ集約型のハイ パフォーマンス コンピューティング環境（エネルギー探査、金融サービス、インターネットおよびホスティング サービス、ビジネス インテリジェンス、エンジニアリング、製造、メディアおよびエンターテイメント、バイオインフォマティクス、科学研究など）でも幅広く活用できます。

次のステップ

PowerScale NASストレージ ソリューションがお客様の組織にどのようなメリットをもたらすかについては、Dell EMCのセールス担当者または認定販売店にお問い合わせください。

機能の比較や詳細情報を入手するには、Dell EMC PowerScaleにアクセスしてください。

Dell EMC PowerScale ソリューションの詳細はこちら

Dell EMC エキスパートに問い合わせる

他の リソースを表示

#DellEMCStorage で会話に参加