

Panasas ActiveStorソリ
ューション:アーキテクチャ
の概要



ホワイトペーパー



目次

エグゼクティブサマリー	3
Panasas ActiveStor	3
スケーラビリティ	3
パフォーマンス	3
信頼性	3
管理性	4
Panasas ActiveStorソリューションのコンポーネント	4
ディレクターノード	4
ストレージノード	4
Panasas PanFS	5
オブジェクトスケーラビリティに基づくパラレルファイルシステム	5
Object Storage Device File System	5
スケーラブルなメタデータサービス	5
クラスタ管理サービス	6
Panasas PanActive Manager	6
Panasas DirectFlow パラレルデータアクセスプロトコル	6
NFSおよびSMBプロトコル	6
メタデータアーキテクチャ	7
ファイルレベルのメタデータ	7
ブロックレベルのメタデータ	7
Panasas PanFSにおけるデータ保護	7
消去訂正符号 (Erasure-Coded) データ保護	7
スケールの拡大とともに信頼性が向上	8
データスクラビング	8
Extended File System Availability	8
結論	8

エグゼクティブサマリー

刻々と増え続ける情報を処理、保存、分析するニーズに押されて、ビジネスの成長やデータの増加にたやすく対応できるコンピューティングおよびストレージ環境の構築が緊急の課題となっています。増大するデータの大部分は、非構造化データです。わずか数キロバイトの小さいテキストファイルから、数テラバイトの画像およびビデオファイルまで、サイズはさまざまです。業界アナリストが異口同音に予測しているように、非構造化シーケンシャルデータが増加するペースは、構造化データのそれを大幅に上回っています。ネットワーク接続ストレージ(NAS)は、アプリケーションベースのデータ共有、容易な管理、Network File System(NFS)やServer Message Block(SMB)など複数のファイルアクセスプロトコルのサポートを前提に設計されているため、非構造化データの保存に最適です。

問題になるのは、従来のスケールアップNASシステムが成長していくと、いずれかの時点で物理的な限界に達することです。その限界を超えて成長した結果が、複数のストレージサイロであり、管理に余分なコストや労力がかかります。基本的なスケールアウトNASシステムは、この問題の解決に役立ったとはいえ、最新のワークフローで要求される高度なパフォーマンスが欠けています。それよりも優れたアプローチは、パフォーマンス重視のスケールアウトプロトコルを使用する、管理しやすく手頃な価格の、統合型のスケールアウトNASシステムです。

Panasas ActiveStor

高性能なスケールアウトNASであるPanasas® ActiveStor®は、他のNASアーキテクチャに見られる拡張性の問題やパフォーマンスの制約を克服する、先進的なテクノロジーソリューションです。Panasasのクラスタ化されたストレージアーキテクチャは、数十または数百テラバイトからスタートし、直線的にパフォーマンスを拡大しながら、シームレスかつ透過的に数ペタバイトまで拡張できる柔軟性を備えています。これにより、ストレージサイロが回避されます。クラスタにActiveStorユニットを追加するだけで、その都度、ストレージ容量、ダイナミックランダムアクセスメモリ(DRAM)キャッシュ、処理能力、ネットワーク帯域幅を直ちに拡大できます。最大規模のPanasas導入環境でも、すべてのデータが1つのネームスペースに存在し、1つの管理用グラフィカルユーザーインターフェイス(GUI)を使って操作することができ、非常に高い信頼性と可用性でデータが提供されます。

スケーラビリティ

Panasasのアーキテクチャ面における重要な成果の1つは、効率的なスケーラビリティの実現です。これは、1つのネームスペースを使用する共有ストレージプールに、1台以上のActiveStorエンクロージャーをクラスタ化することで達成されています。ActiveStorエンクロージャーを1台から10台、さらには100台に増やすことにより、パフォーマンスと容量が正比例して拡大され、ほぼ完璧なリニアスケーリングが可能です。

パフォーマンス

ストレージからデータが取り出されるのをじっと待つのは、誰にとっても苦痛です。しかし、データ

を集散的に使用する高性能なアプリケーションでは、膨大な量のデータがやり取りされ、待ち時間の条件も非常に厳しいのが普通です。

Panasas ActiveStorストレージなら、他のNASアーキテクチャよりずっと簡単に、こうした要件に対応することができます。Panasasは過去10年以上にわたって、これらの要件を前提に最適化されてきたからです。ActiveStorエンクロージャーを追加するたびに、ストレージ容量、DRAMキャッシュ、ネットワーク帯域幅が段階的かつ直線的に拡大される、スケールアウトアーキテクチャを採用しています。Panasasでは、アプリケーションと並行してストレージノードからデータが提供されます。これにより、集約帯域幅が増加するだけでなく、アプリケーションが1つのファイルに対して達成可能な帯域幅が増加します。さらに、ストレージノードからアプリケーションに直接、データが送信されます。中間サーバーを経由するホップがなく、余分なネットワークリンクも存在しません。

信頼性

NASシステムの総ストレージ容量が増加していくと、システムに存在するドライブ数が多いという単純な理由で、ドライブ障害の頻度が高くなる可能性があります。Panasasは、2レベルの保護機能を備えた、先進的なファイル単位/分散型の消去訂正符号(erasure-coding)ソフトウェア実装により、この問題を解決しています。第1レベルとして、各ファイルのデータがクラスタの全ストレージノードに分散される際、デュアルパリティの消去訂正符号による保護が適用されます。これにより、ドライブでもノード全体でも、最大2つの同時エラーが訂正されます。第2レベルの保護は、ドライブ内で一定の範囲のブロックがアクセス不可能になる、ドライブの部分的エラーを処理します。これに

Panasasアーキテクチャの主な概念

パラレルデータフロー

高速道路の車線が増えると交通の流れがスムーズになると同じように、Panasasアーキテクチャでは、ストレージ消費者とPanasasストレージの間のネットワークチャネルを追加することにより、I/Oスループットが向上します。従来のスケールアップアーキテクチャや、より伝統的なスケールアウトアーキテクチャでは、ストレージ消費者とネットワーク共有の間のネットワークパスが1つに制限されているため、データスループットが向上しません。

データへの直接アクセス

スケールアップNASや、より伝統的なスケールアウトNASアーキテクチャでは、データをアクセスする際、中間的なファイラーヘッドまたはNASクラスタの特定のノードを強制的に経由し、そこからデータが存在するストレージノードに到達します。Panasasソリューションでは、すべてのストレージ消費者が、スケールアウトNASクラスタのすべてのストレージに直接接続されます。その結果、パフォーマンスが向上するだけでなく、ホットスポットがなくなり、可用性が向上します。さらにPanasasソリューションでは、データバスからメタデータのアクセスが分離されています。これは、スケールアップおよびその他のスケールアウトストレージソリューションに見られる、輻輳のもう1つの側面を回避する上で重要です。

スケールアウトのネットワーク接続ストレージ

Panasasストレージという1つのプールを簡単に、段階的に拡大し、必要に応じて容量を増やし、パフォーマンスを上げることができます。個別に管理する複数のストレージサイロを作る必要がありません。このプロセス中、ストレージを1つのグローバルネームスペースとしてアクセス可能なオンライン状態が保たれます。対照的にスケールアップNASでは、1つのNASヘッドにドライブを随時追加して拡張が続けると、いつか限界に達します。さらに拡張しようとするれば、新たなストレージサイロを導入するか、大掛かりなフォークリフトアップグレードにより、NASファイラーヘッドを交換するしかありません。

インテリジェントな階層化が可能なハイブリッドストレージ

Panasasアーキテクチャでは、Flashおよびハードドライブという2種類のストレージメディアを活用します。小容量のファイルやファイルシステムメタデータは、フラッシュに保存されます。一方、大容量のファイルはハードドライブに保存されます。ハードドライブのストリーミング帯域幅は、大容量シーケンシャルファイルのデータ保存に理想的です。フラッシュの高速なランダムIOPSは、小容量ファイルやファイルシステムメタデータの保存に適しています。これらのアーキテクチャ概念を融合することで、ランダムおよびシーケンシャルI/Oのコストパフォーマンスが最大化され、魅力的なソリューション価格を実現しています。インテリジェントなハイブリッド階層化により、複雑性を隠匿しながら、ファイルシステムに関する単一のビューが維持されます。

消去訂正符号に基づくファイル単位のデータ保護

Panasasでは、ローストレージデバイス全体を保護する代わりに、ソフトウェアベースの消去訂正符号(erasure code)を使用して、個々のファイルを個別に保護します。ソフトウェアベースのデータ保護では、ハードウェアベースのRAIDで生じるスケーラビリティの問題が回避されると同時に、より柔軟性に富んだデータ保護レベルが実現されます。ファイルレベルで動作することにより、リビルドが迅速化され、従来のRAID技法のように可用性が低下せず、スケールの拡大に応じてデータ可用性が向上します。

より、セクターレベルのエラーなど、低レベルな問題をシステムが検知して訂正することができ、ノード全体のリビルドを実行する必要がありません。

従来のハードウェアまたはソフトウェアRAIDアーキテクチャでは、物理ドライブ全体が別の物理ドライブにリビルドされます。物理ドライブのリビルドは、代替ドライブへの書き込み帯域幅による制約を受けます。また、障害が発生したドライブが満杯でなくても、代替ドライブ全体に書き込みが行われます。Panabasソリューションでは、障害による影響のあったファイルだけがリビルドされます。また、専用のドライブ、すなわちアイドル状態の「ホットスベア」ドライブではなく、ストレージクラスタに分散している空きスペースを利用してリビルドが行われます。クラスタ全体を使ってファイルのリビルドするので、障害からの回復に必要な時間は、従来のRAIDアーキテクチャと比べてほんの数分の1に短縮されます。

管理性

Panabasスケールアウトストレージシステムは、1つのGUIを使って管理する1つのエンティティです。これは、ストレージシステムでサポートする単一のネームスペースに、何台のActiveStorエンクロージャーを統合していても同様です。ActiveStorエンクロージャーどうしの間で容量のアンバランスが生じている場合は、クラスタが自動的にリバランスを行います。障害発生時には、すべてのファイルに対し、消去訂正符号データ保護をフルレベルで自動的に再構築します。さらに、バックグラウンドで継続的にクラスタ内のすべてのファイルをスキャンし、潜在的なデータエラーがあればスクラビングします。クラスタの管理は、従来型NASで10のサイロを別々に管理するよりもはるかに簡単で、運用コストの削減につながります。

Panabas ActiveStorソリューションのコンポーネント

Panabas ActiveStorソリューションは、2種類のフォームファクターから選択できるActiveStorブレードハードウェアプラットフォームで動作する、Panabas PanFS®分散型クラスタファイルシステムで構成されています(図1参照)。

ActiveStorソリューションは、4RUのハードウェアエンクロージャーを使用して導入できます。このエンクロージャーに、11ノード(DBクラスのディレクターノードおよびストレージノード)を収容します。各エンクロージャーには、0~3個のDBクラスのディレクターノードと、8~11個のストレージノードを組み合わせて収容できます。

2RUのハードウェアエンクロージャーを使用して、ディレクター機能を導入することも可能です。より強力なASDクラスのディレクターノードを1~4個収容します。2種類のディレクターノード(DBクラスおよびASDクラス)の機能は同じです。PanFSクラスタ分散型ファイルシステムの中で同じ役割を果たし、同じ機能を担当します。この資料では以降、特に明記しない限り、この両方を単にディレクターノードと呼びます。

Panabasシステムのアーキテクチャ

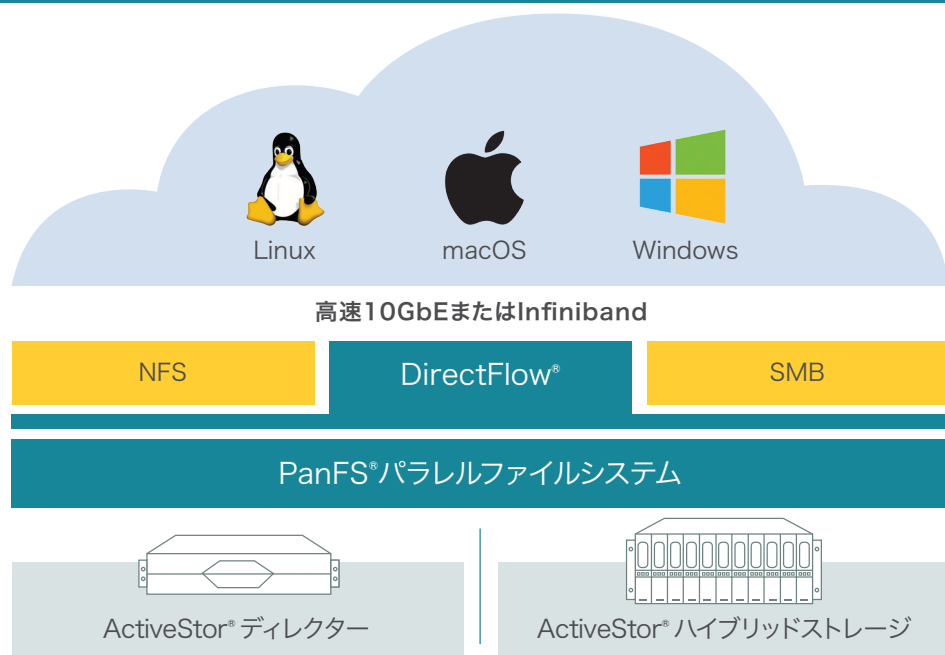


図1. 高度なスケラビリティと効率性

ディレクターノードとストレージノードの比率、およびディレクターノードのクラスは、固有のアプリケーションワークロードやファイルプロトコルの要件に応じて、ユーザー側で設定できます。

つのネームスペースに含まれるActiveStorエンクロージャーの数を増やすことにより、スケラビリティが達成されます。追加した数に比例して、ソリューションの容量、パフォーマンス、ネットワーク帯域幅が増加します。

ディレクターノード

ディレクターノードは、本アーキテクチャの「コントロールプレーン」を構成します。ユーザーデータの保存用ではなく、メタデータを管理します。各ディレクターノードには、プロセッサ、DRAM、および複数の高帯域幅イーサネットポートが搭載され、Panabasが開発したソフトウェアイメージを実行します。このソフトウェアイメージには、ストレージシステム全体のさまざまな側面に対応する制御プロセスおよびインターフェイスプロセスが含まれています。ディレクターノードの役割の1つは、すべてのディレクターノードおよびストレージノードの正常性および「クォーラムメンバーシップ」(ノードが正常に稼働し、Panabasクラスタに参加しているかどうか)を追跡することです。ディレクターノードは、ネームスペース(ファイル名およびディレクトリ階層構造)、ストレージノードにおけるユーザーデータの分散や一貫性、および障害回復動作(データスクラビングやリビルドなど)を管理します。また、ストレージクラスタ全体を1つのエンティティとして取り扱うGUIを提供します。

さらにディレクターノードは「ゲートウェイ」機能も提供し、PanabasアーキテクチャでネイティブのPanabas DirectFlow®プロトコルと、標準

で定義されたストレージプロトコルNFSおよびSMBの間における変換を行います。ディレクターノードは、データパス以外の場所で、これらの機能をすべて実行します。NFSまたはSMBとのプロトコル変換が行われる場合を除いて、ファイルのユーザーデータがディレクターノードを通過することはありません。新しい高性能なASDクラスのディレクターノードは、DBクラスのディレクターノードと比べて、ボリュームあたり約2倍のメタデータ処理を提供します。

ストレージノード

ストレージノードには、Panabasシステムで保存するすべてのアプリケーションおよびユーザーデータが含まれ、アーキテクチャの「データプレーン」を構成します。各ストレージノードは、フラッシュソリッドステートドライブ(SSD)×1、Serial ATA (SATA)ハードディスクドライブ(HDD)×2、プロセッサ×1、そしてクライアントシステムおよびディレクターノードとの接続用の冗長イーサネットポート1ペアを搭載した「ハイブリッド」デバイスです。4Uの各ActiveStorエンクロージャーには、冗長ファンおよび電源装置に加え、専用の内蔵UPS(無停電電源装置)が装備されています。そのためストレージノードソフトウェアは、新しく書き込まれたデータの電源保護されたキャッシュとして、すべてのDRAMを扱うことができます。フラッシュSSDは、小容量のユーザーデータファイルおよびPanFSファイルシステムメタデータへの高速アクセスに対応します。2つの大容量SATAドライブは、一緒にストライピングされており、大容量の非構造化シークエンシャルユーザーデータに適しています。ストレージノードは、Panabasが開発したObject Storage Device File System(OSDFS)と呼ばれるソフトウェアイメー

ジを実行します。OSDFSはPanasasアーキテクチャの他の部分と対話し(クライアントシステムとの直接インターフェイスを含む)、これらのシステムに代わって、SSDおよびHDDへのデータおよびメタデータの保存を管理します。

Panasas PanFS

Panasasは、分散型クラスタパラレルファイルシステムPanFSを開発し、これによって高性能なスケールアウトNASソリューションActiveStorを駆動しています。レガシーのブロックストレージレイとパラレルファイルシステムソフトウェアを緩やかにカップリングする他のストレージ製品とは違って、PanFSソリューションは、モジュール式で細かい構成が可能なActiveStorエンクロージャーを前提に、メンテナンスの容易な現場交換可能ユニット(FRU)として設計されています。その結果、高い信頼性と、容易な管理および成長力を兼ね備えた、エンタープライズクラスのソリューションとなっています。Panasasでは、ストレージクラスタを「レルム」と呼んでいます。

PanFSスケールアウトNAS/パラレルファイルシステムは、高性能、信頼性、管理の容易さを目標に設計されています。PanFSプラットフォームは、分散型のクラスタファイルシステム、ボリュームマネージャー、そして拡張性の高いソフトウェアベースの消去訂正符号データ保護エンジンの各機能を組み合わせた、高性能な1つのシステムとなっています。単一のネームスペースから、1秒あたり数百ギガバイトのデータを提供します。ユーザーデータ数十テラバイトから数十ペタバイトまで、どんな規模のPanFSオペレーティング環境でも、1人のシステム管理者が管理できます。

PanFSソリューションでは、使い勝手の良さを考慮し、必要に応じてネームスペースに何個のボリュームでも含めることが可能です。スループットとスケーラビリティを最大化するため、PanFSオペレーティング環境ではストレージノードのプール全体にデータを分散し、そのプールへの直接パラレルアクセスを実現しています。アクセスには複数のユーザー認証方式を使用することができ、Linux、macOS、Windows環境へのスムーズな統合のため、複数のアクセスプロトコルによってデータアクセスを制御できるようになっています。最高のアクセスパフォーマンスを必要とするお客様は、DirectFlowパラレルデータプロトコルを使って、PanFSアーキテクチャにアクセスできます。一方、従来のNFSおよびSMBプロトコルにより、同じネームスペースへのオープンアクセスが提供されます。

PanFSアーキテクチャ(図2参照)は、次の機能別ブロックで構成されています。

- ・ **パラレルファイルシステム** – オブジェクトストレージデバイスの動作とメタデータ処理の調整
- ・ **OSDFS** – クライアントシステムに代わって、ドライブへのデータの保存とドライブからのデータの取得を実行
- ・ **スケーラブルなメタデータサービス** – キャッシュ貫性とファイルメタデータの処理

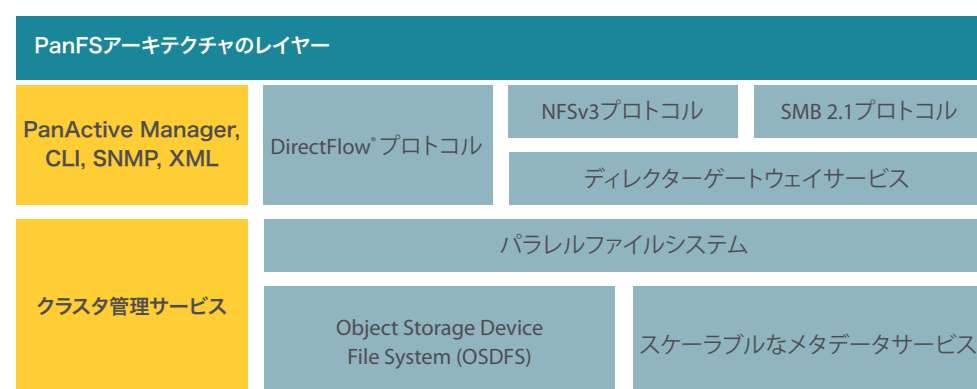


図2. 従来のスケールアウトNASソリューションの制約を克服。

- ・ **クラスタ管理サービス** – 現時点で正常かつ「クォラム」であるノードと、そうでないノード。ノードで障害が発生した場合の回復動作、設定変更など
- ・ **PanActive® Manager** – CLI、SNMP、XML管理インターフェイス
- ・ **DirectFlow プロトコル** – LinuxおよびmacOSに対応するネイティブの高性能なキャッシュ貫性のあるファイルアクセス
- ・ **NFSプロトコル** – Linux、macOS、その他に対応する標準のNFSv3プロトコル
- ・ **SMBプロトコル** – WindowsおよびmacOSに対応する標準のSMBv3.1プロトコル
- ・ **ディレクターゲートウェイサービス** – 上位レイヤーのプロトコル実装を介したPanFSアーキテクチャのエクスポートをサポート

オブジェクトスケーラビリティに基づくパラレルファイルシステム

Panasas PanFSクラスタパラレルファイルシステムは、基盤となる分散型オブジェクトストレージデバイスプールからの、スケールアウトNASファイルアクセスを提供します。オブジェクトベースのデータレイアウトは、Panasasアーキテクチャに備わった高度なスケーラビリティと効率性を支えている、重要な設計原則の1つです。PanFS環境内のファイルは、オブジェクトの内部に保存されています。ただし、Panasasではファイル単位の消去訂正符号(eraser coding)を使用して各ファイルを保護しています。そのため、ファイルは複数のオブジェクトに広がる形で「シャーディング」(すなわちストライピング)されます。これらのシャード(破片)を保存する多数のコンポーネントオブジェクトで構成された「仮想オブジェクト」に、ファイル全体が保存されます。PanFSオペレーティング環境には、コンポーネントオブジェクトだけが保存されます。仮想オブジェクトは、そのファイルを構成するコンポーネントオブジェクトおよびストライピングパラメータの集合を識別するマップに過ぎません。このマップが、コンポーネントオブジェクトの近くに保存されます。1つのストレージノードで発生した障害が、特定のファイルの複数のコンポーネントオブジェクトに影響しないようにするため、1つのファイル用のコンポーネントオブジェクトは、さまざまなストレージノードに分散されます。ファイルシステムメタデータ(ファイルに関する

情報)およびディレクトリも、やはりオブジェクトの内部に保存され、ファイルデータを含むオブジェクトと同じオブジェクトストアで維持されます。

DirectFlowプロトコルは、クライアントがオブジェクトストレージデバイスのデータを直接アクセスするための、高性能なネイティブ方式です。DirectFlowプロトコルを使用するクライアントは、メタデータマネージャーを迂回して、ストレージノードとの間でオブジェクトの読み書きをパラレルに実行します。DirectFlowクライアントは、メタデータマネージャーと帯域外でリモートプロシージャコール(RPC)により対話し、ファイルを保存しているオブジェクトのアクセス許可およびロケーション情報(例: マップ)を取得します。

Object Storage Device File System

PanFSデータサービスはすべて、ストレージノード機能の最上部にビルドされています。ストレージノードのソフトウェアは、Object Storage Device File System(OSDFS)と呼ばれる抽象化を実装しています。DirectFlowプロトコルは、DirectFlowクライアントソフトウェアがストレージノードに実行させる、オブジェクトレベルの読み取り/書き込みを指定します。

パフォーマンスを最適化するため、OSDFSには高度なキャッシング機能と、SSDおよびHDDを使ったインテリジェントなオブジェクトデータの配置が含まれています。OSDFSはキャッシュの読み取りによってデータの取得を改善するとともに、新しく書き込まれたデータを電源保護されたDRAMにキャッシングします。新しく書き込まれたデータを、より大容量のシーケンシャル領域に蓄積することで、データの断片化を減らします。したがって後の段階で同じデータを読み取るときも、シーケンシャルな読み取りが可能です。OSDFSは、ストレージノード内で各オブジェクトに最適な物理ストレージデバイスを、オブジェクトのサイズに基づいて判断します。60KB未満のPanFSメタデータおよびユーザーファイルは、ストレージノードのSSDに保存されます。60KBを超えるユーザーデータは、ストレージノードのハードドライブに保存されます。

スケーラブルなメタデータサービス

PanFSメタデータサービスは、ディレクターノードで動作します。ファイルシステムのセマンティクをすべて実装し、全ストレージノードに広が

るデータのシャーディング(ストライピング)を管理します。これらのサービスが、分散型ファイルシステムの動作を管理します。具体的には、ファイルおよびオブジェクトレベルのメタデータ一貫性、クライアントキャッシュの一貫性、クライアントI/O中断からの回復能力、ストレージノード動作、ファイルへの安全なマルチユーザーアクセスなどです。Portable Operating System Interface (POSIX) とのコンプライアンス上、ディレクトリ階層構造やファイルメタデータの変更は、いずれもアトミックであることが要求されます。ディレクターノードはトランザクションログを使用して、このアトミック性を確保します。それぞれのローカルトランザクションログを、別のディレクターノードに同期複製することが、フォルトトレランスの基本となっています。そのディレクターノードが果たす役割の1つが、複製元のディレクターノードの指定「バックアップノード」です。バックアップノード関係の割り当てと再割り当ては、自動的に行われます。

クラスタ管理サービス

ActiveStorシステムはどれも、ストレージノードとディレクターノードで構成される1つのクラスタです。クラスタ内の各ノードが、共通のPanFSクラスタ管理サービスを実行します。該当ノードのハードウェアモニタリング、構成管理、ノード上で動作するサービスの制御を提供する追加のサービスにより、クラスタ管理サービスが強化されます。レルム内の3つ以上のディレクターノードからなる任意のサブセットを顧客側で選択し、「レップセット」の一部分にします。これらのノードは、グローバルな構成データベースの複製をホスティングします。レップセットに含まれるディレクターノードは、Paxosアルゴリズムを使った投票により、自分たちの中から1つを「レルムプレジデント」に選出します。プレジデントは、そのレルムのマスタークラスタ管理サービスを実行します。このプロセスにより、他のスケールアップストレージアーキテクチャによく見られる、いわゆる「分離脳」条件が回避されます。レルムが正常に動作するには、レルム内に少なくとも3つのディレクターノードが必要である理由がこれです。レルムプレジデントは、グローバルなシステム構成の変更、クラスタノード障害の検出と対処、ソフトウェアアップグレードやシステム再起動の処理を担当します。さらに、レルム内のどのノードでどのサービスを起動するかのも決定も、レルムプレジデントが担当します。

Panasas PanActive Manager

ストレージ管理者は、直感的なWebベースGUIであるPanasas PanActive Managerを通じて、PanFSクラスタ管理サービスと対話します。このGUIを使って、ActiveStorエンクロージャーのクラスタを、規模に関わらず1つのアプライアンスとして管理することができます。ネームスペース全体を対象とするストレージ管理のためのシングルポイントが提供されます。PanActive Managerは、新しいストレージの検出、負荷分散によるパフォーマンスの効率化といった主要なワークフローに加え、レポート作成、スナップショット、ユーザークォータの強制などのエンタープライズデータサービスを、自動的に実行します。管理者はPanActive Managerを介したPanFSオペレー

ティング環境のアクセスに加え、コマンドラインインターフェイス(CLI)、スクリプティング可能なXMLインターフェイス、標準ベースのSNMPインターフェイスを使用できます。

Panasas DirectFlowパラレルデータアクセスプロトコル

Panasas DirectFlowパラレルデータアクセスプロトコルの使用により、クライアントシステムはActiveStorストレージクラスタ内のすべてのストレージノードを直接、パラレルにアクセスすることができます。そのため、従来のプロトコルのようなI/Oボトルネックが回避されます。結果的に、NFSやSMBなどの業界標準プロトコルよりも高いパフォーマンスが得られます。それと同時に、これらのプロトコルではサポートされない、クライアントシステム全体でのキャッシュ一貫性もサポートされます。

PanFSオペレーティング環境へのアクセスを必要とするシステム上のDirectFlowクライアントは、クライアントのOSのカーネル内部で動作するファイルシステムモジュールとして実装されています。ディレクターノードとストレージノードには、これに対応するDirectFlowプロトコルサーバー側コンポーネントがあります。クライアント側のカーネルモジュールは、標準の仮想ファイルシステム(VFS)インターフェイスを実装しています。クライアントはPanFSプラットフォームを、標準に準拠したPOSIXファイルシステムとしてアクセス可能です。

DirectFlowクライアントは、シャーディングを使用してファイルを複数のフラグメントに分割しま

Panasas DirectFlow® プロトコル

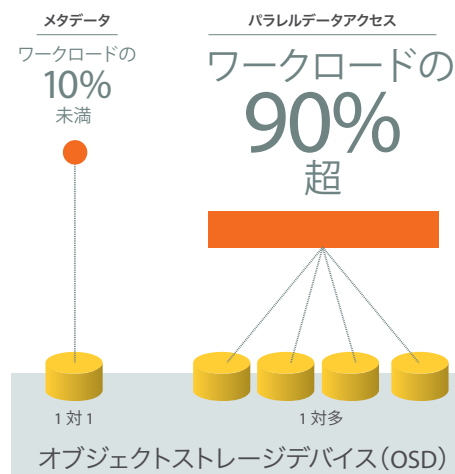


図3. クライアントがオブジェクトストレージデバイスのデータを直接アクセスするための、高性能なネイティブ方式

す。書き込まれたデータは、最初にシャーディングされます。次に、これらのシャード(破片)に対して消去訂正符号アルゴリズムが実行され、ファイルのその部分に関するデータ保護情報が生成されます。この保護情報も、シャードとして扱われます。その後、DirectFlowプロトコルによってすべてのシャードが独立したコンポーネントオブジェクトとして、それぞれ別のストレージノードに、パラレルに書き込まれます。同様に、クライアントは該当するストレージノードから直接、シャードコ

Panasasアーキテクチャにおけるマルチプロトコルアクセス

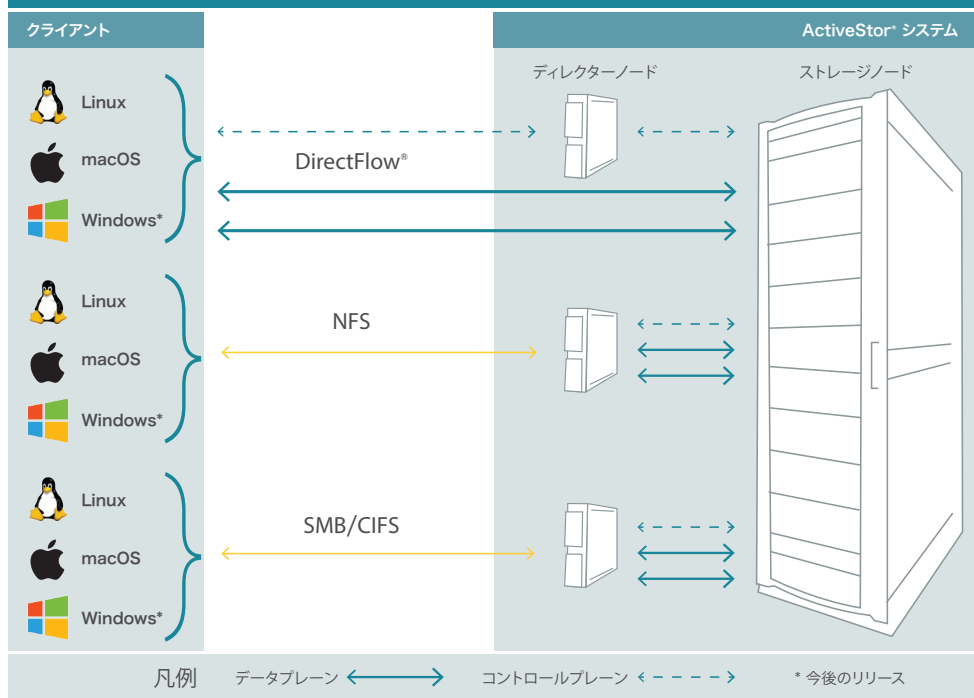


図4. Panasas PanFSオペレーティング環境では、NFSまたはSMBプロトコルによるスケーラブルなアクセスが、クライアントシステムに提供されます。

ンポーネントオブジェクトをパラレルに取得することで、ファイルを読み取り、これらのシャードを組み合わせ、目的とするファイルにします(図3参照)。クライアント側で消去訂正符号(eraser coding)を実行することにより、従来のスケールアウトNASソリューションのような、ストレージ側での消去訂正符号によるパフォーマンスボトルネックが排除されています。

NFSおよびSMBプロトコル

PanFSオペレーティング環境では、ネイティブのDirectFlowアクセスに加えて、標準のNFSまたはSMBプロトコルによるスケラブルなアクセスも、クライアントシステムに提供されます。これがディレクターノード上で「ゲートウェイ」として動作することで(図4参照)、Linux、macOS、Windowsクライアントからなる異種混在型のIT環境に、PanFSプラットフォームを統合することが可能です。

NFSサービスは、標準のFreeBSD NFSサーバーのチューニング済みバージョンであり、NFSv3プロトコルをサポートします。PanFS環境へのSMBアクセスには、2つの選択肢があります。プロトコル間のキャッシュ一貫性をサポートするPanasonicの従来のSMB実装と、パフォーマンスが大幅に向上しながらも、引き続きプロトコル間のキャッシュ一貫性をサポートする、Panasonicの新しいSMB実装です。NFSとSMBの両方から同時にファイルをアクセスする必要性が予測される場合、プロトコル間のキャッシュ一貫性が重要です。PanFS SMBサービスの旧バージョンは、DBクラスおよびASDクラスのディレクターノードのどちらでも実行可能であり、Microsoft SMBプロトコル仕様に準拠した、商用ライセンス取得済みのSMB 2.1実装から派生しています(PanasonicはMicrosoftプロトコルライセンスです)。新しいPanFS SMBサービスは、より強力なASDクラスのディレクターノードだけで動作し、Sambaバージョン4.6.6をベースとしています。ユーザーはいずれかのSMBゲートウェイソリューションを使用して、Microsoft WindowsまたはmacOS環境で作成したファイルを容易に管理することができます。ユーザー認証の管理には、Active DirectoryやLightweight Directory Access Protocol(LDAP)など、さまざまな選択肢があります。

PanFSプラットフォームでは、管理者がWindowsのセキュリティ識別子(SID)をLinuxのユーザーIDにマッピングすることが可能です。したがって、特定のユーザーのWindows、macOS、Linuxアカウントに対し、同時にストレージクォータを適用することができます。

メタデータアーキテクチャ

ストレージ管理者は、PanFSグローバルネームスペース内でボリュームを簡単に作成できます。

ボリュームは、ディレクトリとファイルの一般的な階層構造であり、PanFSオペレーティング環境内で共通のストレージ容量のプールを共有します。ただし、ボリュームによって全体的なネームスペースを各エリアに分割し、それぞれ異なる管理統制を行うことが可能です。たとえば、ユーザー1人あたりの容量クォータを、ボリュームレベルで定義することができます。ボリュームごとに独自のクォータを設定できます。また、スナップショットをボリュームレベルで取得します。ボリュームは、ファイルシステムネームスペースでトップレベルのディレクトリとして表示されます。ボリュームごとに管理サービスのセットが関連付けられ、別のディレクターノード上で動作する各ボリュームマネージャーにより、そのボリュームの動作(例:クォータやスナップショットの実装)が制御されます。システムがスケールアウトされると、ディレクターノードの数が増加するので、このパーティショニング方式により、全体的なファイルシステムメタデータのパフォーマンスを、ネームスペース全体の成長とともにリニアに拡大できます。

「メタデータ」という用語は、いくつかの異なるものを同時に意味する場合があります。通常、ベンダーによって意味が異なります。一般論として、メタデータはデータ以外のほとんど全部を意味します。PanFSオペレーティング環境では、次のものがメタデータとして扱われます。

- ・ レルムプレジデントは、グローバルな構成データベースをメタデータとして扱います。
- ・ ディレクターノードは、ディレクトリ構造およびファイル名の情報をメタデータとして扱います。
- ・ ディレクターノードは、アクセス制御リスト(ACL)、アクセス許可ビット、ファイル所有者などのファイル属性を、メタデータとして扱います。
- ・ ディレクターノードは、ファイル(仮想オブジェクト)と、そのファイルを含むシャード済みコンポーネントオブジェクトのID集合とのマッピングを、メタデータとして扱います。
- ・ ストレージノードは、コンポーネントオブジェクトIDと、そのストレージノード内にあるそのコンポーネントオブジェクトのブロックアドレスの集合とのマッピングを、メタデータとして扱います。

PanFSプラットフォームでは、メタデータはファイルレベルおよびブロックレベルという、2つの基本的なカテゴリに分類されます。ファイルレベルのメタデータは、ディレクターによって操作されますが、ストレージノードに保存されます。ブロックレベルのメタデータは、ストレージノードの内部でしか使用・認識されません。

ファイルレベルのメタデータ

PanFSオペレーティング環境では、ファイルに関する数種類のメタデータが維持されます。メタデータとしては、所有者、サイズ、変更日時など、一

般的なユーザー可視情報も含まれます。加えて、各ファイルの仮想オブジェクトと、一連のコンポーネントオブジェクトIDとの関係を示す、ファイル単位のマップも維持されます。各ファイルのシャード済みデータの部分が、どのコンポーネントオブジェクトに保存されているか、これらのオブジェクト間でデータがどのようにストライピングされているかが、このマップによって示されます。Panasonicソリューションは基本的に、このファイルメタデータを、ファイルの3つのコンポーネントオブジェクトに関連付けられた「拡張属性」として保存します。これらのコンポーネントオブジェクトは、所定の「属性保存コンポーネント」です。それ以外のコンポーネントオブジェクトには、個々のオブジェクトの長さや変更日時など、基本的な属性しかありません。

ブロックレベルのメタデータ

ストレージノード内部のソフトウェア(OSDFS)は、ブロックレベルのメタデータの管理に「遅延割り当て」方式を使用します。この方式では、データ、ブロックポインター、オブジェクト記述子をバッチでまとめて、DRAMへの大容量の書き込みを行い、その後でSSDおよびHDDに送信します。DRAM内のOSDFS書き込みバッファは、ActiveStorプラットフォームに組み込まれたUPSで保護されており、電源異常などのシステム障害が発生した場合には、ディスクにフラッシュされます。

OSDFSは、典型的な最新の「ローカル」ファイルシステムです。従来のdirect/indirect/double-indirect方式を使用して、各コンポーネントオブジェクトのバイト群が保存されているドライブ上の場所を追跡します。コピーオンライト参照カウントリング用に最適化された、独自のbitmaplikeデータ構造により、空きスペースを追跡します。コピーオンライトのサポートは、PanFSのスペース効率に優れたスナップショット技法に不可欠です。

従来のNASシステムでは、ブロックレベルメタデータの管理によって、使用可能な処理能力の大半が消費されていました。低レベルのドライブ管理をストレージノードに委任することで、PanFSディレクターノードが実行すべき処理は、同等のSANまたはNASファイルシステムマネージャー(ファイルシステムの全ファイルに加え、全ブロックを追跡しなければならない)と比べて、桁違いに少なくなります。

Panasonic PanFSにおけるデータ保護

PanFSアーキテクチャには、複数のレイヤーからなる、スケラビリティの高い先進的なデータ保護機能が備わっています。その結果、高度な可用性と耐障害性を示しながら、高いパフォーマンスを提供するシステムとなっています。

消去訂正符号(Erasure-Coded)データ保護

PanFSプラットフォームには、2レベルの保護機能を備えた、先進的なファイル単位／分散型の消去訂正符号(erasure-coding)ソフトウェアが実装されています。第1レベルとして、各ファイルのデータがクラスタの全ストレージノードに分散される際、デュアルパリティの消去訂正符号による保護が適用されます。これにより、ドライブでもノード全体でも、最大2つの同時エラーが訂正されます。第2レベルの保護は、ドライブ内で一定の範囲のブロックがアクセス不可能になる、ドライブの部分的エラーを処理します。これにより、セクターレベルのエラーなど、低レベルな問題をシステムが検知して訂正することができ、ノード全体のリビルドを実行する必要がありません。

PanFSオペレーティング環境におけるデータ保護のトップレベルは、他のアーキテクチャのようにドライブ単位での計算またはRAIDグループ内での計算はなく、ファイル単位で計算されます。ファイルは全ストレージノードに広がる形でシャーディングされます。つまり、各ファイルがそれぞれ別のストレージノードのサブセットにストライピングされることを意味します。いずれかのストレージノードで障害が発生しても、その障害はファイルの1つのシャード(破片)にしか影響しません。消去訂正符号アルゴリズムにより、消失したデータを残りのシャードから簡単にリビルドすることが可能です。

PanFSオペレーティング環境におけるデータ保護の第2のレベルは、「垂直パリティ」と呼ばれます。垂直パリティは、各ストレージノードで発生するドライブレベルのセクター障害(すなわち、回復不能読み取りエラー[unrecoverable read errors, URE])からの保護を提供します。全面的なドライブ障害よりもはるかに発生頻度が高いUREは、場合によってはエラーにならなかったりする、過渡的な状況の可能性もあります。システムがファイルの集合を再構築している間に、ストレージノードでUREが発生した場合、システムから見れば追加の障害であり、「障害が多すぎる(too many failures)」のリスクが生じ、そのファイルのデータ損失が引き起こされます。垂直パリティは、このようなUREを検出して訂正することにより、「too many failures」のケースを回避します。

スケールの拡大とともに信頼性が向上

物理的RAIDグループをベースとする従来のストレージ製品は、ハードウェア実装がソフトウェア実装かを問わず、複数のドライブグループと、各グループで発生する1つないし2つのドライブ障害からの回復が可能なアルゴリズムを、データ保護戦略の基盤としています。これらのグループは通常、12〜24ドライブ程度で非常に小さく、サイズが固定されているのが一般的です。何らかの障害が発生すると、RAIDコントローラーが「ホットスเปア」プールから使われていないアイドル状態のドライブを引き出し、アルゴリズムを使用して、この新しいドライブに、故障したドライブと同じデータを保存します。

このアプローチには、パフォーマンスに悪影響を及ぼしかねない、いくつかの重要な因果関係があります。

- ・RAIDグループの「リビルド」を実行する速度は、新しいドライブがリビルド用のデータを受け付ける速さによって制約されます。
- ・ドライブの容量が増え続けた場合、そのドライブの書き込み帯域幅は、仮に増えるとしても、容量の増加率と同じ割合で拡大されるわけではありません。
- ・ドライブの故障率は、ドライブの寿命全体を通じてほぼ一定しています。そのため、RAIDグループが最初のドライブ障害から回復している間に、別のドライブで障害が発生する確率は、最初のドライブをリビルドするのに必要な時間の長さによって左右されます。リビルドに時間がかかるほど、「二重障害」およびデータ損失の可能性が高くなります。
- ・ドライブの「ホットスペアプール」のコストを支払っていることになりますが、このコストは、必要なパフォーマンスや容量には貢献しません。

このようなアーキテクチャアプローチを使用する場合、スケールが拡大するにつれ、全体的なデータ信頼性は低下していくことになります。一方、PanFSプラットフォームは、スケール(およびパフォーマンス)が拡大するとともに、データ信頼性が向上するように設計されています。この事実の裏側にある最も重要な要因の1つは、パラレルな再構築です。PanFSは、ドライブではなくファイルをリビルドします。ファイルは全ストレージノードの中で、それぞれ異なるサブセットにストライピングされます。したがって、すべてのストレージノードの帯域幅を利用しながら、障害による影響のあったすべてのファイルを一度に、パラレルにリビルドすることが可能です。各ファイルの生き残ったフラグメントは、そのフラグメントのあるストレージノードから読み取られます。各ファイルの再構築された部分は、他のストレージノードに書き込まれます。

PanFSによるリビルドは、ドライブ1台分のパフォーマンスに制限されません。リビルドパフォーマンスも、データアクセスパフォーマンスと同様、直線的に拡大されます。リビルド時間が短いため、すべてのデータ保護レイヤーでファイルが完全にカバーされない、脆弱性の期間が短縮されます。各ファイルの再構築、すなわちファイルの生き残りのデータと消去訂正符号パリティ情報を読み取り、消失したデータを計算し、このリビルドされたデータを新しいストレージノードに書き込むという実際の処理は、システムのすべてのディレクターノードに分散されます。システムにディレクターノードを追加すればするほど、再構築パフォーマンスがリニアに向上します。

データスクラビング

PanFSオペレーティング環境では、バックグラウンドですべてのファイルのスクラビングが実行されます。ファイルごとに、そのファイルのすべてのレ

ベルのデータ保護情報が、そのファイルに含まれているデータと一致するかどうかチェックされます。たとえば消去訂正符号パリティ情報で、消去訂正符号のレベルがデータと一致しないことが判明した場合、残りの2レベルの保護を使ったファイルレベルの再構築により、そのエラーが訂正されます。データ保護システムの最下部にある垂直パリティレイヤーにより、ドライブメディアの潜在的な「ビットロット」エラーを検出し、訂正することが可能です。データスクラビングは、潜在的なエラーを早期にキャッチして問題を解決し、データ損失のリスクを未然に防ぐ上で重要です。

Extended File System Availability

PanFSソリューションでは、デュアルパリティ消去訂正符号および垂直パリティの2つのレベルのデータ保護に加え、ネームスペース、ディレクトリ階層構造、ファイル名に関して、Extended File System Availability (EFSA)と呼ばれる追加のデータ保護レイヤーが提供されます。たとえば3つ以上のストレージノードで同時に障害が起こるような極めて稀なイベントが発生し、消去訂正符号で回復することが不可能な場合、PanFSシステムは影響のあったファイルを認識し、それらを不良ファイルとして囲い込みます。このシステムにはディレクトリ階層構造とファイル名に関する追加の保護レイヤーがあるため、どのファイルが障害の影響を受けたかを、フルパス名で管理者に正確に知らせることが通常、可能です。そのため、障害による影響を受けなかったファイル(通常、大多数のユーザーファイル)については、中断のないユーザーアクセスが確保されます。一方、影響を受けた少数のファイルについては、ログに記録され、バックアップまたはその他のソースからの回復が行われます。

結論

Panasas ActiveStorアーキテクチャとPanFSオペレーティング環境は、従来のスケールアップNASアーキテクチャのパフォーマンス上の限界を取り除き、より伝統的なスケールアウトNASソリューションの制約を克服する設計となっています。包括的でタイトに統合されたPanasasソリューションは、数テラバイト、数ペタバイトものデータに対するパフォーマンスの高い直接/パラレルアクセスを実現すると同時に、従来のNASシステムを拡張したときに見られる、信頼性の問題を回避します。

大容量の非構造化データセットに対応する、真に直線的なスケールパリティ

Panasasソリューションを使用すると、ストレージのパフォーマンスと容量を簡単に、シームレスに拡大できます。ActiveStorエンクロージャーを10台から100台、さらには1,000台に増やすことにより、パフォーマンス、容量、クライアントアクセスが同じ倍率で拡大され、ほぼ完璧なリニアスケールリングが可能です。

超高速パラレルパフォーマンス

Panasonic DirectFlowプロトコルなら、最も高度なデータ転送要件に対応できます。複数のストレージノードへの直接パラレルデータI/Oを実行するクライアントを通じて、すばやく結果を取得できます。

混合ワークロード

多数のアプリケーションを同時に実行しても、Panasonic ActiveStorアーキテクチャのI/Oパフォーマンスは終始一定しています。これは大小さまざまな非構造化データセットが混在するストレージとの間で読み取り／書き込みを行う場合も同様です。

エンタープライズ級の信頼性と可用性

ActiveStorスケールアウトNASを拡張するにつれ、信頼性と可用性が向上します。Panasonicの先進的なファイル単位／分散型の消去訂正符号 (erasure- coding) ソフトウェア実装では、2レベルの保護が提供されます。1つのレベルでは、ストレージノードで障害が発生した時点で問題を検知し、訂正します。もう1つのレベルでは、セクターレベルのエラーなど、ストレージノード内部の低レベルな問題をシステムが検知して訂正することができ、ノード全体のリビルドを実行する必要がありません。さらにPanasonicソリューションでは、すべてのストレージ消費者が、スケールアウトNASクラスタのすべてのストレージに直接接続されます。その結果、パフォーマンスが向上するだけでなく、ホットスポットがなくなり、可用性が向上します。

容易な管理、コストの削減

Panasonic ActiveStorソリューションは、完全に統合されたスケールアウトNASシステムであり、規模がどれほど大きくても、シングルポイントでの管理が可能です。NASシステムに期待されるエンタープライズデータサービスを提供し、商業市場ワークロードの高度なパフォーマンスを前提に最適化されています。多数の異なる非構造化データワークロードを1つのスケールアウトActiveStorソリューションに集約して、高度なパフォーマンスと使いやすさを両立させることにより、データセンターにおけるストレージの複雑性を解消します。

Panasonic高性能スケールアウトNASについての詳細は、www.panasonic.comをご覧ください。