ホワイト ペーパー
2016 年 3 月

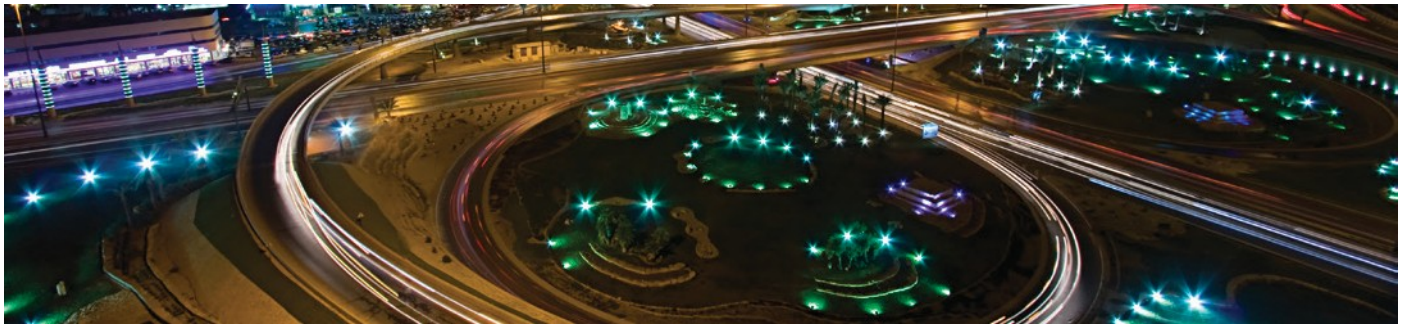
ハイパーコンバージェンス用の次世代データプラットフォーム

Intel® Xeon® プロセッサ搭載
Cisco HyperFlex™ Systems

目次

ハイパーコンバージェンス用の次世代データ プラットフォーム

ホワイト ペーパー
2016 年 3 月



概要
本書では、ハイパーコンバージド インフラストラクチャ環境のデータ ストレージを変革する Cisco HyperFlex™ HX Data Platform ソフトウェアについて解説します。このプラットフォームのアーキテクチャおよびソフトウェア定義型ストレージ アプローチを紹介し、IT インフラストラクチャの複雑化につながるストレージ サイロの解消方法を説明します。

新世代のアプリケーションやデータ用のプラットフォーム

IT アーキテクチャはアプリケーションによって規定されます。また、要件の進化に伴い、サーバ、ストレージ システム、ネットワーク ファブリックの関係は絶えず変化しています。仮想化環境や第 1 世代のハイパーコンバージド システムは一部の問題を解決できますが、新たなインフラストラクチャのサイロ化が生まれ、大規模なスケーラビリティには限界があり、ライフサイクル管理機能や強力なデータ セキュリティも備わっていません。Cisco HyperFlex™ Systems は柔軟性と拡張性に優れた新世代のソリューションを提供し、ハイパーコンバージド ソリューションの潜在能力を幅広いアプリケーション、ワークロード、ユース ケースで最大限に活用できるようにします。

Cisco HyperFlex Systems は、エンドツーエンドのソフトウェア定義型インフラストラクチャを目指して設計され、第 1 世代の製品では妥協していた点を克服しています。Cisco HyperFlex Systems では、Cisco Unified Computing System™（Cisco UCS®）サーバで構成されたソフトウェア定義型コンピューティング、新しい強力な Cisco HyperFlex HX Data Platform ソフトウェアを利用したソフトウェア定義型ストレージ、そして Cisco Application Centric Infrastructure（Cisco ACI™）とスムーズに統合する Cisco® ユニファイド ファブリックによるソフトウェア定義型ネットワーキング（SDN）が 1 つになっています。このように事前統合されたクラスタは、1 時間以内に稼働を開始でき、アプリケーションのリソース ニーズと一致するようにリソースを個別にスケーリングできます（図 1）。

Cisco HyperFlex HX Data Platform：ストレージ サイロの解消

IT 組織がサーバ仮想化を利用して物理サーバを統合している場合、アプリケーション固有のデータ需要が原因となり、多数のストレージ サイロが発生します。Cisco HyperFlex Systems の基盤である Cisco HyperFlex HX Data Platform は、専用の高性能な分散ファイル システムで、エンタープライズクラスのデータ管理サービスを幅広く備えています。この革新的なデータ プラットフォームは、分散ストレージ テクノロジーを再定義し、第 1 世代のハイパーコンバージド インフラストラクチャが抱えていた制約を克服します。

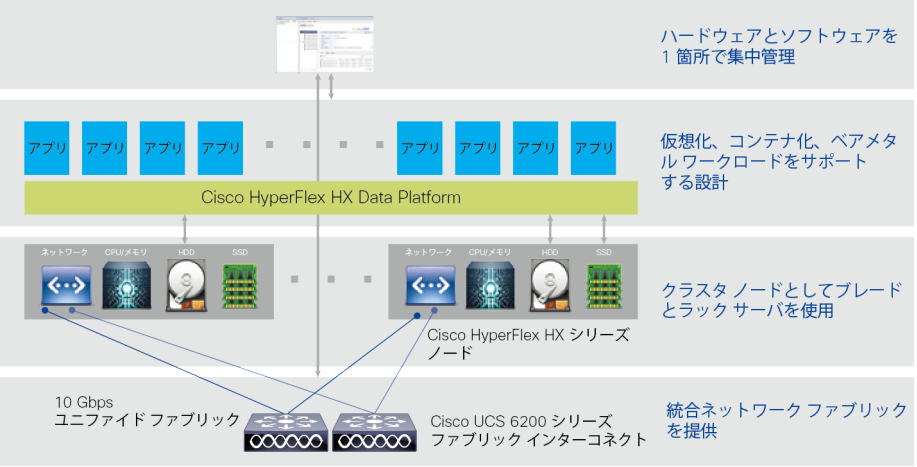


図 1. シスコ独自の機能を搭載した Cisco HyperFlex Systems は新世代のハイパーコンバージド ソリューションを実現

Cisco HyperFlex HX Data Platform には次のような特長があります。

- **エンタープライズクラスのデータ管理**:包括的なライフサイクル管理や高度なデータ保護を分散ストレージ環境で実現するために必要な機能です。複製、重複排除、圧縮、シン プロビジョニング、高速でスペース効率の高いクローン、スナップショットが含まれます。
- **データ管理のシンプル化**:ストレージ機能を既存の管理ツールに統合することで、アプリケーションの即時プロビジョニング、クローニング、スナップショットが可能になり、日常的な運用を大幅にシンプル化できます。
- **個別のスケーリング**:コンピューティング、キャッシング、キャパシティ層を個別にスケーリングできるため、ビジネス ニーズの変化に応じて柔軟に環境を拡張できます。
- **継続的なデータ最適化**:インライン データ重複排除および圧縮によってリソース使用率を向上させ、データ スケーリングの余裕を増やします。
- **動的なデータ配置**:ノード メモリ、エンタープライズクラスのフラッシュ メモリ（ソリッドステート ディスク（SSD）ドライブ上）、および永続ストレージ層（ハードディスク ドライブ（HDD）上）で動的にデータを配置し、パフォーマンスと耐障害性を最適化します。また、クラスタの拡張に応じて、データ配置を再調整します。
- **API ベースのデータ プラットフォーム アーキテクチャ**:データ仮想化を柔軟に実現し、既存のデータ タイプおよび新しいクラウドネイティブのデータ タイプをサポートします。

アーキテクチャ

Cisco HyperFlex Systems では、データ プラットフォームが 3 台以上の Cisco HyperFlex HX-Series ノードにわたって構築されているため、可用性の高いクラスタを実現できます。各ノードに搭載された Cisco HyperFlex HX Data Platform コントローラは、分散ファイル システムを実装し、フラッシュベースの内蔵 SSD ドライブと大容量 HDD を使用してデータを保存します。コントローラは 10 ギガビット イーサネットで相互に通信し、クラスタ内の複数のノードにわたる単一のストレージ プールを実現します（図 2）。ノードはファイル、ブロック、オブジェクト、API プラグインを使用し、データ レイヤを介してデータにアクセスします。ノードが追加されると、クラスタはそれに比例して拡張され、コンピューティング、ストレージ容量、I/O パフォーマンスを提供します。

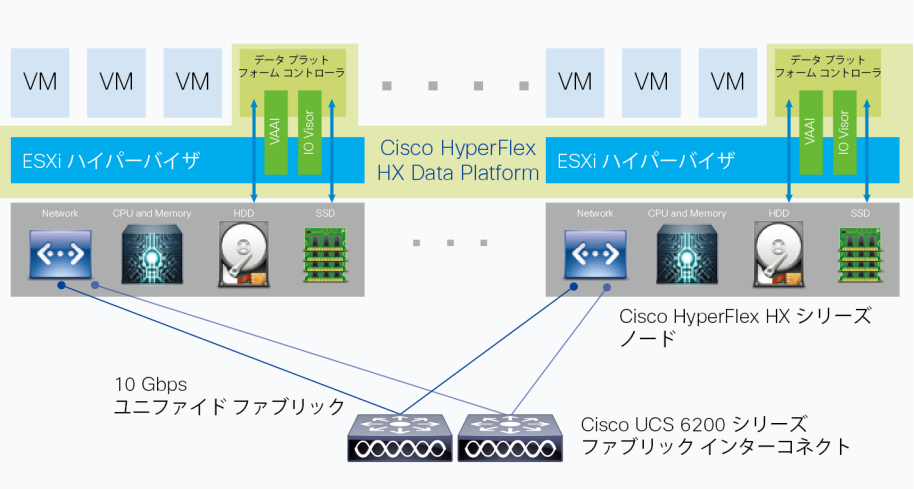


図 2. 分散型の Cisco HyperFlex System

VMware vSphere 環境では、コントローラは、プロセッサ コア数とメモリ量が専用に割り当てられた 1 台の仮想マシンを占有します。これにより、一貫したパフォーマンスを実現し、クラスタにある他の仮想マシンのパフォーマンスへの影響を防止できます。コントローラは、VMware VM_DIRECT_PATH 機能により、ハイパーバイザの介入なしですべてのストレージにアクセスできます。分散キャッシング レイヤの一部としてノードのメモリと SSD ドライブが使用され、分散キャパシティ ストレージ用にノードの HDD が使用されます。コントローラは、プリインストールされた以下の 2 つの VMware ESXi vSphere Installation Bundle（VIB）を使用して、データ プラットフォームを VMware ソフトウェアに統合します。

- **IO Visor**：この VIB はネットワーク ファイル システム（NFS）マウント ポイントを提供し、個別の仮想マシンに接続された仮想ディスク ドライブに対して ESXi ハイパーバイザがアクセスできるようにします。ハイパーバイザからは、単にネットワーク ファイル システムに接続されているように見えます。
- **VMware vStorage API for Array Integration（VAAI）**：このストレージ オフロード API は、vSphere が高度なファイル システム操作（スナップショット、クローニングなど）を要求できるようにします。コントローラは、実際のデータのコピーではなくメタデータの操作によって、これらの操作を発生させます。そのため、迅速な対応が可能で、新しいアプリケーション環境をすぐに導入できます。

機能の概要

Cisco HyperFlex HX Data Platform コントローラは、ハイパーバイザがアクセスするボリュームのすべての読み取り/書き込み要求を処理し、それにより仮想マシンからのすべての I/O を仲介します（ハイパーバイザには、このデータ プラットフォームから独立した専用のブート ディスクがあります）。このデータ プラットフォームが実装しているログ構造ファイル システムは、SSD ドライブのキャッシング レイヤを使用して、読み取り要求および書き込み応答を高速化します。また、HDD で実装されたパーシステンス レイヤも使用します。

データ分散化

新しいデータはクラスタ内のすべてのノードにわたって分散され、キャッシング層を使用してパフォーマンスが最適化されます（図 3）。すべてのノードに均等に保存されているストライプ ユニットに対して、新しいデータをマッピングすることで、効果的なデータ分散化を実現します。データ レプリカの数は、設定したポリシーによって決まります。アプリケーションが

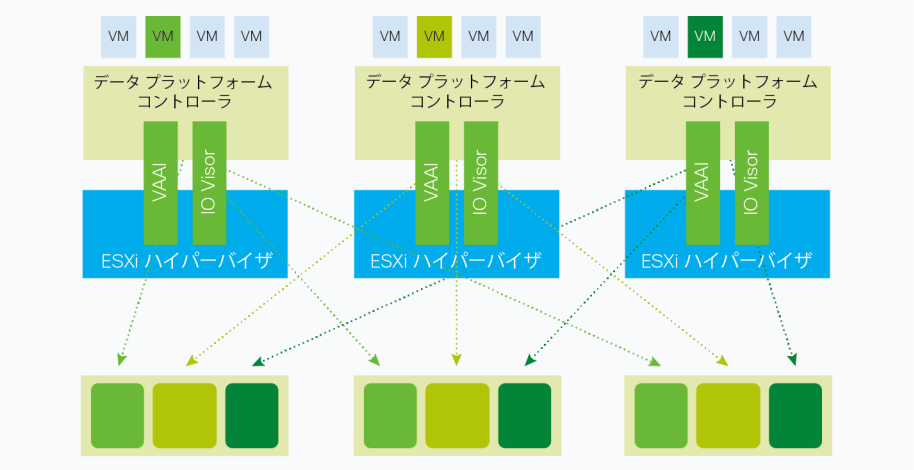


図 3. データはクラスタ内の複数のノードにわたってストライプ

データを書き込む際に、データはストライプ ユニットに基づいて適切なノードに送信されます。ストライプ ユニットには、関連する情報ブロックが含まれています。このデータ分散化アプローチを、同時に複数のストリームを書き込める機能と組み合わせることで、ネットワークとストレージのホット スポットを回避できます。また、仮想マシンの場所にかかわらず同じ I/O パフォーマンスを実現でき、ワークロード配置の柔軟性も向上します。局所的なアプローチを採用し、利用可能なネットワークや I/O リソースを十分に活用できていない他のアーキテクチャとは対照的な特徴です。

- **データ書き込み操作**：書き込み操作では、データはローカルの SSD キャッシュに書き込まれ、並行してレプリカがリモートの SSD に書き込まれます。その後で書き込み操作が認識されます。
- **データ読み取り操作**：読み取り操作では、ローカルのデータは通常、ローカルの SSD から直接読み取られます。ローカルではないデータは、リモート ノードの SSD から取得されます。これにより、プラットフォームがすべての SSD を読み取り操作に使用できるため、ボトルネックが削減され、優れたパフォーマンスを実現できます。

VMware Dynamic Resource Scheduling（DRS）などのツールを使用して仮想マシンを新しい場所に移動する場合に、Cisco HyperFlex HX Data Platform ではデータを移動する必要がありません。このアプローチにより、システム間での仮想マシンの移動に伴う影響やコストを大幅に削減できます。

データ操作

このデータ プラットフォームが実装しているログ構造ファイル システムは、SSD ドライブのキャッシング レイヤを使用して、読み取り要求および書き込み応答を高速化します。また、キャパシティ レイヤが HDD で実装されています。新しいデータは、可用性要件に対応できるように複数のノードにわたってストライプされます（通常は 2 台または 3 台のノードです）。設定したポリシーに基づき、新しい書き込み操作は、クラスタ内の別のノードにある SSD ドライブに複製されます。その後で、永続的な書き込み操作として認識されます。このアプローチにより、SSD やノードで障害が発生した場合のデータ損失の可能性を低減できます。その後、書き込み操作は長期保存用の安価で高密度な HDD にデステージされます。高性能な SSD ドライブを低コストで大容量な HDD と組み合わせることにより、最適なコストでアプリケーション データを保存し、高速に取得できるようになります。

ログ構造ファイル システムにより、書き込まれるブロックがキャッシュに集められます。構成可能な書き込みログが一杯になるまで、またはワークロード条件によって HDD ディスクへのデステージが必要になるまで集められます。既存データが（論理的に）上書きされた場合、ログ構造アプローチでは、新しいブロックを追加してメタデータを更新するだけで済みます。データをデステージする際の書き込み操作は、1 回のシーク操作と、大量のシーケンシャル データの書き込みで構成されています。このアプローチは、読み取り、変更、書き込みという従来のモデルと比較して、大幅にパフォーマンスを改善できます。従来のモデルでは、シーク操作が何回も実行され、少量のデータしか同時に書き込めませんでした。

各ノードのディスクにデステージされたデータは、重複排除されて圧縮されます。このプロセスは書き込み操作が認識された後に行われるため、これらの操作でパフォーマンスに悪影響が及ぶことはありません。小さな重複排除ブロック サイズにより、重複排除率が向上します。圧縮でデータ フットプリントがさらに削減されます。その後、データは HDD ストレージに移動され、書き込みキャッシュ セグメントは解放されて再利用可能になります（図 4）。

ホット データ セット（永続層から頻繁に読み取られるデータまたは最近読み取られたデータ）は、SSD ドライブとメモリの両方にキャッシュされます（図 5）。最も頻繁に使用されるデータをキャッシング レイヤに保持することで、仮想化アプリケーションに対する Cisco HyperFlex Systems のパフォーマンスが向上します。仮想マシンがデータを変更する際に、そのデータはキャッシュから読み取られる可能性が

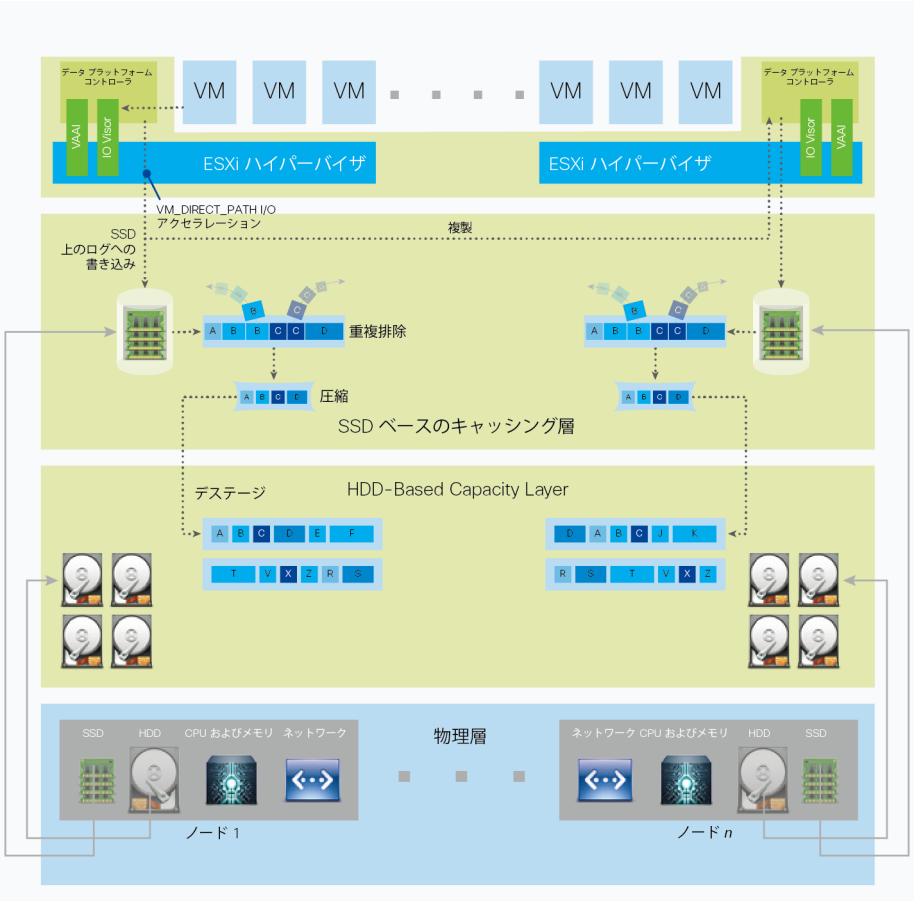


図 4. Cisco HyperFlex HX Data Platform 書き込みオペレーション フロー

高いため、HDD ディスク上のデータを読み取って展開する必要はほとんど生じません。Cisco HyperFlex HX Data Platform ではキャッシング層が永続層から分離されており、I/O パフォーマンスとストレージ容量を個別にスケーリングできます。

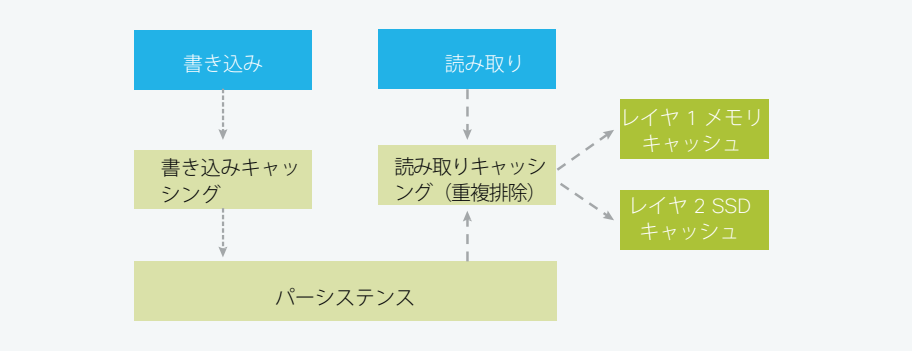


図 5. 分離されたデータ キャッシングとデータ パーシステンス

データ最適化

Cisco HyperFlex HX Data Platform には、精密なインライン重複排除や可変ブロック インライン圧縮が搭載されており、キャッシュ（SSD、メモリ）レイヤおよびキャパシティ（HDD）レイヤのオブジェクトに対して常に有効になっています。他のソリューションでは、パフォーマンス維持のためにこれらの機能を無効にする必要があります。一方、シスコのデータ プラットフォームの重複排除と圧縮機能は、パフォーマンスを維持、強化し、物理ストレージのキャパシティ要件を大幅に削減できるように設計されています。

データ重複排除

データ重複排除は、クラスタ内のすべてのストレージ（メモリ、SSD ドライブ、HDD など）で使用されます。このプラットフォームは、特許出願中の Top-K Majority アルゴリズムを基盤とし、実証的研究の成果を活用しています。実証的研究では、小さなデータ ブロックに分割し、少数のデータ ブロックに基づけば、ほとんどのデータは大幅に重複排除できる可能性があることが示されています。このような頻繁に使用されるブロックのみに対してフィンガープリントとインデックスを作成することにより、少量のメモリで高い重複排除率を実現でき、クラスタ ノード内の価値の高いリソースであるメモリを節約できます（図 6）。データは、スペースを節約するためにパーシステンス層で重複排除されるだけではなく、キャッシング層に読み込まれる際にも重複排除された状態が維持されます。このアプローチにより、キャッシング層に格納できるワーキング セットが増加するため、読み取りパフォーマンスが向上します。

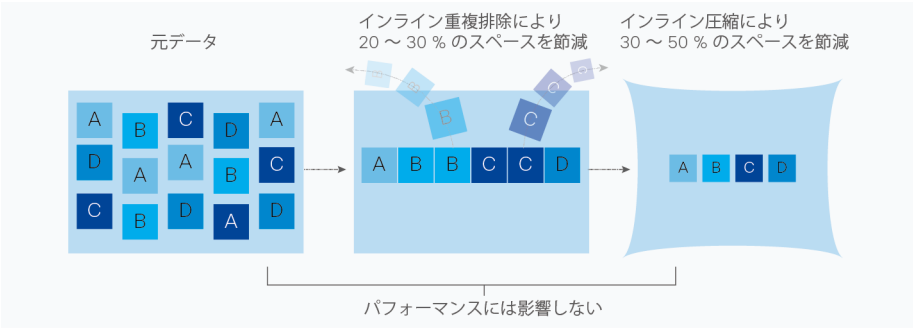


図 6. Cisco HyperFlex HX Data Platform はパフォーマンスに影響を与えずにデータ ストレージを最適化

インライン圧縮

Cisco HyperFlex HX Data Platform は、データ セットに対して高性能なインライン圧縮を使用し、ディスク領域を節約します。他の製品も圧縮機能を提供していますが、パフォーマンスに悪影響を及ぼす場合が少なくありません。一方、シスコのデータ プラットフォームは CPU オフロード指示を使用して、圧縮操作によるパフォーマンスへの影響を軽減します。さらに、ログ構造の分散オブジェクト レイヤは、以前に圧縮されたデータの変更（書き込み操作）に影響を及ぼしません。新規の変更は圧縮されて新しい場所に書き込まれ、既存の（古い）データは削除用にマークされます（そのデータをスナップショット内で保持する必要がある場合を除きます）。書き込み操作の前に、変更対象のデータを読み取る必要はありません。一般的には読み取り、変更、書き込みという手順が必要ですが、この機能はその手順に伴う悪影響を回避し、書き込みパフォーマンスを大幅に向上させます。

ログ構造の分散オブジェクト

Cisco HyperFlex HX Data Platform では、ログ構造の分散オブジェクト ストア レイヤにより、データをグループ化して圧縮します。データは、重複排除エンジンでフィルタされ、自己アドレス可能なオブジェクトになります。これらのオブジェクトは、ログ構造でディスクに順次書き込まれます。すべての新しい I/O（ランダム I/O を含む）は、キャッシング層（SSD、メモリ）および永続層（HDD）の両方に順次書き込まれます。オブジェクトはクラスタ内のすべてのノードにわたって分散されるため、ストレージ容量を均等に使用できます。

このプラットフォームは順次レイアウトの使用により、フラッシュメモリの耐久性を高め、HDD の読み取り/書き込みパフォーマンスの特性を最大限に活用します（HDD はシーケンシャル I/O 操作に最適です）。読み取り、変更、書き込みという手順ではないため、圧縮やスナップショット操作のパフォーマンスには影響がほとんどないか、まったくありません。また、全体的なパフォーマンスに対しても同様です。

データ ブロックはオブジェクトに圧縮され、固定サイズのセグメント内に順次配置されます。次に、それらのセグメントはログ構造で順次配置されます（図 7）。ログ構造セグメント内の圧縮された各オブジェクトは、キーを使用して一意にアドレス可能です。各キーはフィンガープリントされ、チェックサムと保存されており、高度なデータ整合性を実現します。また、オブジェクトが時系列で書き込まれるため、このプラットフォームはメディアやノードの障害から迅速に回復できます。障害が発生してデータが削除された場合、その時点より後にシステムが受信したデータのみを再度書き込めば回復できるからです。

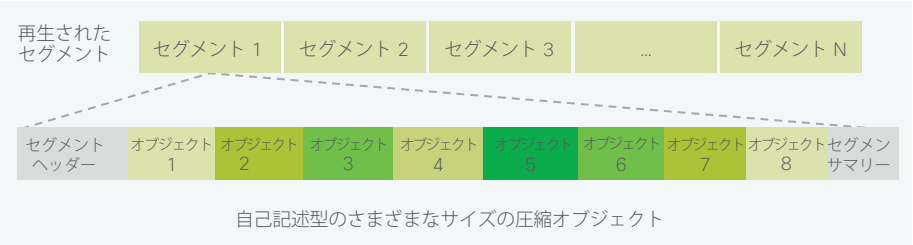


図 7. Cisco HyperFlex HX Data Platform のログ構造ファイル システムのデータ レイアウト

データ サービス

Cisco HyperFlex HX Data Platform では、スペース効率の高いデータ サービスがスケーラブルに実装されています。これらのサービスには、シン プロビジョニング、スペース再利用、ポインタベースのスナップショット、クローンが含まれ、パフォーマンスには影響が及びません。

シン プロビジョニング

このプラットフォームでは、予測に基づいてディスク容量を購入してインストールする必要がないため、長期間使用されないままになるリスクを避けられ、ストレージを効率的に使用できます。仮想データ コンテナは、任意の量の論理スペースをアプリケーションに示せます。一方、必要な物理記憶域の量は、書き込まれるデータによって決定されます。そのため、ビジネス要件に応じて、既存ノードのストレージを拡張したり、ストレージ集約型ノードを追加してクラスタを拡張したりできます。必要になる前の段階で容量確保のためにストレージを購入しておく必要性はなくなります。

スナップショット

Cisco HyperFlex HX Data Platform はメタデータベースのゼロコピー スナップショットを使用して、バックアップ操作やリモート複製を容易にします。これらの機能は、データを常時利用できる必要がある企業にとって極めて重要です。スペース効率の高いスナップショットにより、物理ストレージ容量の消費を心配せずに、データのオンライン バックアップを頻繁に実行できます。データはオフライン移動や、これらのスナップショットからの迅速な復元が可能です。

- **高速なスナップショットの更新**：スナップショットに変更されたデータが含まれている場合、新しい場所に書き込まれ、メタデータが更新されます。読み取り、変更、書き込みという手順を実行する必要はありません。
- **高速なスナップショットの削除**：スナップショットは迅速に削除できます。このプラットフォームでは、SSD 上の少量のメタデータを削除するだけで済みます。デルタ ディスク技術を使用するソリューションで必要となる、長時間の統合プロセスは不要です。
- **柔軟なスナップショット**：Cisco HyperFlex HX Data Platform を使用すると、個別のファイル ベースでスナップショットを作成できます。仮想環境内で、これらのファイルは仮想マシンのドライブにマッピングされます。柔軟に異なる仮想マシンに異なるスナップショット ポリシーを適用できます。

高速でスペース効率の高いクローン

Cisco HyperFlex HX Data Platform では、クローンは書き込み可能なスナップショットです。クローンを使用すると、テストや開発環境用に、仮想デスクトップやアプリケーションなどの項目を迅速にプロビジョニングできます。これらの高速でスペース効率の高いクローンにより、ストレージ ボリュームを迅速に複製可能です。そのため、仮想マシンをメタデータの操作だけで複製でき、実際のデータ コピーは書き込み操作に対してのみ実行されます。このアプローチでは、何百個ものクローンの作成や削除を数分で行えます。フルコピー方式と比較すると、このアプローチは時間を大幅に節約し、IT の俊敏性や生産性を高められます。

クローンは作成時に重複排除されます。それぞれのクローンに差が生じ始めた場合は、共通するデータは共有され、固有のデータのみが新たな記憶域に保存されます。差が生じたクローン内にあるデータ重複は、重複排除エンジンが排除し、クローンのストレージ フットプリントをさらに削減します。そのため、ストレージ容量の使用状況を心配せずに、多数のアプリケーション環境を導入できます。

データの可用性

Cisco HyperFlex HX Data Platform では、ログ構造の分散オブジェクト レイヤが新しいデータを複製し、データの可用性を高めます。設定したポリシーに基づき、書き込みキャッシュに書き込まれたデータは、異なるノードにある 1 台以上の SSD ドライブに同時に複製されます。その後で、書き込み操作がアプリケーションに認識されます。このアプローチにより、新しい書き込みが迅速に認識されるとともに、SSD やノードの障害からデータを保護できます。SSD またはノードで障害が発生した場合、データの利用可能なコピーを使用して、別の SSD ドライブまたはノード上でレプリカが迅速に再作成されます。

ログ構造の分散オブジェクト レイヤは、書き込みキャッシュからキャパシティ レイヤに移動されたデータも複製します。この複製されたデータは、先ほどと同様に、HDD やノードの障害から保護されています。2 つのレプリカ（合計 3 つのデータ コピー）があるため、クラスタは 2 台の SSD ドライブ、2 台の HDD、または 2 台のノードで障害が発生しても対応でき、データ損失のリスクがありません。フォールトトレラントな構成および設定の一覧については、Cisco HyperFlex HX Data Platform のシステム管理者ガイドを参照してください。

Cisco HyperFlex HX コントローラ ソフトウェアで問題が発生した場合、そのノード内にあるアプリケーションからのデータ要求は、クラスタ内の別のコントローラに自動で転送されます。この機能を使用すると、クラスタやデータの可用性に影響を与えずに、コントローラ ソフトウェアのアップグレードやメンテナンスをローリング方式で実施できます。この自己修復機能は、Cisco HyperFlex HX Data Platform が実稼働アプリケーションに最適な理由の 1 つです。

データ リバランシング

分散ファイル システムには、堅牢なデータ リバランシング機能が必要です。Cisco HyperFlex HX Data Platform では、メタデータ アクセスでオーバーヘッドが生じず、リバランシングが極めて効率的です。リバランシングは、中断を伴わないオンライン プロセスであり、キャッシング レイヤと永続レイヤの両方で実行されます。データを詳細に特定して移動することで、ストレージ容量の使用を改善します。このプラットフォームでは、ノードやドライブが追加または削除された場合や、それらで障害が発生した場合に、既存データが自動でリバランスされます。新しいノードがクラスタに追加されると、その容量とパフォーマンスを新しいデータと既存データで利用可能になります。リバランシング エンジンは、既存データを新しいノードに分配し、クラスタ内にあるすべてのノードの容量とパフォーマンスが均等に使用されるようにします。ノードで障害が発生した場合や、ノードがクラスタから削除された場合は、リバランシング エンジンがデータのコピーを再構築し、それらのノードからクラスタ内にある利用可能なノードに分配します。

まとめ

Cisco HyperFlex HX Data Platform は、ハイパーコンバージド インフラストラクチャ環境のデータ ストレージを変革します。このプラットフォームのアーキテクチャおよびソフトウェア定義型ストレージ アプローチは、専用の高性能な分散ファイル システムを実現し、エンタープライズクラスの幅広いデータ管理サービスを提供します。革新的なデータ プラットフォームにより、分散ストレージ テクノロジーを再定義し、次世代のハイパーコンバージド インフラストラクチャを実現します。

関連情報

Cisco HyperFlex Systems の詳細については、
http://www.cisco.com/c/ja_jp/products/hyperconverged-infrastructure/index.html
を参照してください。

お問い合せ先

シスコシステムズ合同会社

〒107 - 6227 東京都港区赤坂 9-7-1 ミッドタウン・タワー
http://www.cisco.com/jp

LE-55406-00JA 16.04