

ホワイトペーパー

Infinidat

ストレージアーキテクチャ 競争上の優位性をペタバ イト規模で実現する



要約

Infinidat® のエンタープライズ向けストレージソリューションは、独自の特許取得済みInfinidatストレージアーキテクチャをベースにしています。そのアーキテクチャは、ソフトウェアにより定義されたストレージ(SDS)機能を完全に抽象化したものに、既製ハードウェアを最適なかたちで統合したものです。Infinidatのソフトウェアを重視したアーキテクチャは、構築に30年以上の歳月を要しており、大容量、高速、コスト削減という矛盾する要件を解決します。まさにデータ管理設計面での進化と革新の証といえるでしょう。このホワイトペーパーでは、Infinidatを唯一無二のエンタープライズストレージへと導いたテクノロジーについて考察します。そのテクノロジーによってInfinidatは、オールフラッシュのパフォーマンスを超える速度(マイクロ秒の低レイテンシで、200万 IOPS超)でマルチペタバイト級の容量を提供しつつも、先例のない100%の可用性を備え、かつそのすべてを最低の総保有コストで実現しています。

設計原理

ストレージアーキテクチャの設計において、昨今のデータセンターが持つニーズを満たすには、複数の要件を満たす必要があります。

カテゴリー	要件
信頼性	企業は24時間365日事業を展開しており、ダウンタイムは許容されません
容量	デジタルトランスフォーメーション、細分化されたビッグデータアーキテクチャ、人工知能(AI)や機械学習(ML)により、データ量の増加は指数関数的に加速しています
パフォーマンス	データ規模の拡大に合わせて、パフォーマンスも拡大する必要があります。同一の(またはより短い)時間枠内で、同一の(またはより優れた)成果を提供しなければなりません
簡素化	管理者はDevOpsモデルへの移行を見据えて、簡易な操作、幅広いエコシステムの統合、ビルトインツールなどを期待しています。ストレージ管理にかかる時間を削減し、より多くの時間をアプリケーションやビジネスプロセスに費やすためです
統合	限られたユースケースにしか対応できないテクノロジーは過去の遺物。現代のストレージは最大限の効率、簡素化、コスト削減ができるよう、あらゆるユースケースに対応できなければなりません
費用	容量やパフォーマンス改善要件といっしょに、コストも増加するわけではありません。そのため、アーキテクチャの破壊的な変化が求められます

一方で、Amazon、Google、Azureのようなパブリッククラウドのインフラプロバイダーは、IT全体にかかる費用を削減できると主張します。彼らインフラプロバイダーは、小規模企業に対しても同じようにアピールしますが、小規模企業は大人数のITスタッフを集める余裕がなく、1人か2人の「よろず屋」に頼ってIT運用を維持するのが精一杯です。しかしながら、大企業や、リージョンクラウドおよびマネージドサービスプロバイダーにとってみれば、より効率的なIT運用、つまりビジネス、テクノロジー、財務それぞれの要件を満たすIT運用をすることで、自分たちのインフラ内でクラウドの利点をすべて享受することができます。さらにコストを削減し、データの主権も維持できるのです。

InfiniBox® アーキテクチャ

Infinidat の主力製品であるInfiniBox®は、主要な原則を念頭に置きながら、以下のすべての課題に対応できるように設計されました。

原則	根拠	課題
革新的なソフトウェア設計	ハードウェアとは異なり、ソフトウェアは時間の経過とともに最適化され、デグレードすることなくパフォーマンスを改善してきました。InfiniBoxは140を超えるソフトウェア特許をベースにした、本当の意味での「ソフトウェアにより定義されたストレージ(SDS)」です。Infinidatの革新的なソフトウェアには、特許取得済みのNeural Cacheアルゴリズム、メタデータ管理構造、次世代ストレージ機能などが含まれています。	パフォーマンス 簡素化 信頼性 費用
レジリエンシーを設計する	規模拡大に向けて設計する場合、レジリエンシーは必須です。InfiniBoxは、99.99999%の信頼性を実現できるよう設計され、三重に冗長化したアーキテクチャを使用しています。あらゆる必須コンポーネント(ソフトウェアとハードウェア)が、ダウンタイムやデータ喪失から少なくとも2つの冗長化(N+2)によって保護されています。Infinidatは、何万時間もの利用実績に基づき、100%のシステム可用性を保証します。これは、中立的な立場をとる業界最大手のガートナー社によるPeer Insights Reviewsによって、裏付けられています。	レジリエンシー コスト 簡素化 統合
規模拡大のアーキテクチャ	破壊的な料金での容量とパフォーマンスを達成するには、規模拡大が求められます。InfiniBoxは、10PBクラスを求める大口顧客に向けて設計されました。また、42Uラックでより優れた実効容量を実現しています。	統合 コスト 簡素化 容量
ハードウェアとソフトウェアの復元性が高い統合	Infinidatは、業界を牽引するハードウェアコンポーネントベンダーすべてを評価し、最も信頼性の高いベンダーを選び出して、InfiniBoxソリューションに組み込んでいます。このベストオブブリード式のアプローチは、完全に統合され試験済みのソリューションを顧客が利用できることを意味します。複雑なハードウェアの統合や管理作業を必要とする典型的なSDSとは対照的です。	信頼性 簡素化 統合
既製(COTS)ハードウェア	既製のハードウェアを使用し長期的な開発サイクルを回避することで、新たなテクノロジーを容易に導入することができます。これにはCPU、メモリータイプ、ストレージメディアが含まれます。既製のハードウェアとそれに関連したソフトウェアを使用することで、安定性も増します。同じハードウェアが世界中のシステムで利用されているためです。	コスト 信頼性 容量 簡素化 パフォーマンス

パフォーマンス高速化

InfiniBoxはフラッシュ最適化アレイで、DRAM、フラッシュメディア(SSD)、大容量NL-SASディスクを組み合わせることで、データを書き込み、読み取り、保存します。以下に、最小のレイテンシで最大のパフォーマンスを実現するのに、書き込みと読み取りをどのように高速化しているのかを説明します。データ配置の最適化に使用するアルゴリズムをNeural Cacheと呼びます。Neural Cacheは、スマートソフトウェアアルゴリズムを利用することで、オールフラッシュアレイ(AFA)よりも低いレイテンシを提供します。このセクションではその仕組みを説明します。大抵のトランザクションアプリケーションには2つ以上の個別のI/O(1つはログにトランザクションを書き込み、もう1つはデータベースにデータを書き込む)が必要であり、レイテンシがユーザーエクスペリエンスとアプリケーションの最大パフォーマンスの双方を左右する主要コンポーネントとなる点に留意してください。

メタデータ層

メタデータ層のレスポンスタイムはI/Oレイテンシに即座に影響を与えます。InfiniBoxは以下を用いてメタデータの操作を加速させます。

- ▶ **メタデータはすべてDRAMに存在する** — メタデータをDRAMに維持し、読み込みと書き込みの双方を加速させます
- ▶ **メタデータ構造** — InfiniBoxに書き込まれたすべてのデータは、その完全な履歴が「Trie」というメタデータ構造で管理されます。この特許取得済みの実装により、データ配置、関連する構成情報や仮想アドレス情報、データ保護の複数層を取り込みます。
- ▶ **Trieの効率** — Trieからの挿入、変更、削除はすべて同一のレイテンシで動作し、データの最初のバイトからマルチペタバイトにいたるまで、一貫したパフォーマンスを提供します。

書き込み高速化

InfiniBoxでは、前処理(パターン除去、圧縮、暗号化など)なしでDRAMに書き込むことができます。ホストに確認を送信する前に、低レイテンシのInfiniBandを経由して、別ノードのDRAMに第2コピーを作ります。外部フラッシュデバイスの代わりに、(CPUに直接取り付けられている)DRAMからの書き込みを承認すると、InfiniBoxが可能な限り低いレイテンシで書き込みを完了させることができます。

書き込みキャッシュが小さなバケットに細分化される(マトリックスアーキテクチャやデュアルコントローラーアーキテクチャのような)多くのアーキテクチャとは異なり、InfiniBoxは単一の巨大なメモリープールを使用して書き込みを受け付けます。これにより、大きな書き込みバーストを維持させることができ、頻繁に変更するデータをDRAMレイテンシで上書きできます。また、Neural Cache時間が、どのデータブロックがDRAMの速度からメリットを得て、どれをSSDおよびHDDにデステージすべきか、スマートに判断できるようにします。データを書き込みキャッシュに長く維持することで、Neural CacheはCPUやバックエンドの永続層における不要なワークロードを回避します。

デステージする前に、各キャッシュサイクルはランダムに書き込みデータを収集します。そのデータを、書き込まれた際のデータの関連性を含め多数の要素に基づいて大きなシーケンシャルライトに変換します。これにより、それぞれの読み取り操作と協調してこれから必要となるだろうデータを判断する予測分析において、Neural Cacheをサポートします。

読み取り高速化

従来のストレージアレイでは、アクティブなデータ(「ホットデータ」とも言います)をフラッシュキャッシュに配置することで、AFAでパフォーマンスパリティを達成しようとします。それに対しInfiniBoxは、革新的なNeural Cacheを用い、ホットデータすべてをDRAMに配置します。InfiniBoxのNeural Cacheは、ほとんどの読み取りをDRAMが持つ速度で実行します。この速度は、フラッシュの1,000倍も高速です。

Infinidatのグローバルデータファブリックは、数エクサバイトのデータにまたがり、Neural CacheはDRAMからほぼすべての読み取りを実行できることが証明されています。顧客は、AFAよりも総保有コストを抑えて、「オールDRAMアレイ」体験を味わうことができます。

Neural Cacheは学習アルゴリズムであるため、時間の経過とともにパフォーマンスを最適化します。InfiniBoxは厚いSSDフラッシュ層を、DRAMミスの「クッション」として利用します。Neural Cacheは、I/Oパターンを学習してDRAMのデータ配置を最適化するため、フラッシュ層の機能はDRAMミスを処理する機能からI/Oパターンの変化を処理する機能へと変化します。これによりアルゴリズムは予想できなくなる可能性があります(例:DRAMに存在しないデータを必要とする定期監査)。

ソフトウェアアーキテクチャ

100%の可用性を維持するようInfiniBoxを設計する際に、Infinidatはソフトウェアを使用して予測ができないハードウェア障害に対処しています。InfiniBoxはアクティブ・アクティブ・アクティブのソフトウェアアーキテクチャとN+2設計を活用し、常時監視、自己回復、ハードウェア障害からの正常回復を、すべてのレベルで実現します。

RAIDからクラスター化されたサービスにいたるまで、すべてのコンポーネントがソフトウェアに実装されているため、新たなリリースのたびに継続的に最適化を図ることができます。GAバージョン(一般向けに提供開始されたバージョン)を発売してから5年にわたって、ソフトウェアを無停止でアップグレードするだけで、InfiniBoxの最大パフォーマンスは4倍以上も改善しました。これこそ、本当の意味でのSDSの力です。

クラスター化サービス

N+2アーキテクチャデザインに応じて、すべてのデータサービスがすべてのノードで実行され、かつすべてのノードでアクティブになります(クラスターにはパッシブノードはありません)。データサービスは、ファイバーチャネル(FC)ドライバーのような低レベルのコンポーネントも含め、ユーザー空間で実行するよう設計されています。カーネルではデータサービスが実行されないため、システムの他のサービスやノードの可用性に影響を及ぼすサービス障害は、1つとしてありません。さらに、それぞれのサービスは個別にわずか数秒で再起動できます。これらの設計原則が、データプロトコル(NFS、iSCSI、FC、NVMe-oF)のようなフロントエンドサービスや、Neural Cache、InfiniRaid®、InfiniSnap®といったバックエンドサービスに適用されます。

データサービスはクラスターマネージャー(CLM)により起動、監視されます。CLMは、サービスの問題を把握し必要に応じてサービスを再起動できます。障害が発生したサービスは、クラスターと再結合する前に再起動して自己診断が行われます。

正常に開始できなかったサービスはクラスターに結合されず、クラスター内での障害(ビザンチン故障)を回避します。クラスターマネージャーが特定のノードで複数回にわたって再起動に失敗したサービスを識別すると、再起動を停止し、Infinidatサポートに通知を行います。自動修復されたかどうかに関わらず、サービス障害はInfinidatのデータ解析プラットフォームに報告され、ソフトウェアの問題を検出し、コード品質を継続的に改善します。

ディスクレイアウト

InfiniBoxのディスクレイアウトは、InfiniRaidという特許取得済みのソフトウェアイノベーションにより管理されます。InfiniRaidはソフトウェアによって定義されたRAIDであり、すべてのデータ配置、データ保護、障害発生からの修復を制御します。InfiniRaidはデクラスター化したRAIDです。データレイアウトを物理層から分け、数千の仮想RAIDグループを使用するRAIDの一種であり、ドライブすべてにデータを分散し、リビルドによる負荷を分散します。InfiniRaidはRAIDグループを生成して、システムのドライブ2つごとにRAIDグループの最大2.5%を占めるようにします。このオーバーラップするRAIDグループの割合が低いことで、様々なメリットがあります。

- ▶ 自己回復 — リビルドによる負荷は、データレイアウトの最適化により自動的に解消されます
- ▶ 仮想スペア — 空間容量はシステムのディスク全体に均一に広がっています。物理的なホットスペアはないため、再構築プロセスにより最適にデータを再分布し、不要なコストを最小化することができます。システムはF6000シリーズでの障害に対し、最大12ドライブまでの十分な予備容量を備えています。
- ▶ パフォーマンス保護 — 1回のドライブ障害(データは保護された状態)では、優先度の低いRAID再構築(「Rebuild-1」)を生成するのみで、アプリケーションのパフォーマンスを優先します。
- ▶ 迅速修復 — 2つ目のドライブに障害が発生すると、システムは障害のある2つのドライブ間で共有しているRAIDグループの共通する2.5%に対して、再構築(「Rebuild-2」)することを優先します。そのあとに優先度の低いRebuild-1に戻ります。これは、保護されていないRAIDグループがこれ以上ないためです。
- ▶ InfiniSpares — InfiniBoxは、12スペアと同等を保証する容量に加え、必要に応じて空き容量を予備容量として利用できます。このイノベーションにより、最大100本のディスクに障害が発生しても保護機能が喪失することはありません。

データ保護サービス

InfiniBox はデータ保護サービスを数多く提供し、顧客の資産を保護しています。

- ▶ **スナップショット** — InfiniBoxのスナップショット機構はInfiniSnapと呼ばれ、ノンロックングのリダイレクトオンライト方式をベースにしています。これは、スナップショットの有無にかかわらず安定したパフォーマンスを実現します。それぞれのデータセットは最大1000個のスナップショットを保持でき、それぞれが読み取りのみ(データ保護)または書き込み可能(テストや開発環境)のいずれかとなります。InfiniSnapは永続層への書き込みを必要とせず、DRAMでスナップショットを実行します。
- ▶ **不変のスナップショット** — InfiniSnapのスナップショットでは、「不変」というマークを付けることもできます。この不変のスナップショットを作成するときに定める要件およびタイマーに従って、一度書き込まれると変更や削除をすることができなくなります。この機能は、ランサムウェアや類似する脅威に対しての、優れた復元データ保護策となります。
- ▶ **低RPO非同期レプリケーション** — 非同期レプリケーションエンジンは、4秒のレプリケーション間隔で、業界で最も低い目標復旧時点(RPO)を実現、維持します。同時にIPインフラを使用して費用と複雑性を低減します。
- ▶ **同期レプリケーション** — 同期レプリケーションエンジンは、ゼロPROで同期データ保護機能を提供し、同時にストレージレイテンシを400 μ s(ミリ秒)未満に維持します。WANで(高レイテンシ、接続喪失といった)問題が生じた場合、InfiniBoxの同期レプリケーションエンジンは自動的に非同期レプリケーションモードにフェイルバックします。WANが復旧すると、レプリケーションは喪失したデータすべてを自動的に複製し、I/Oを妨げることなく同期レプリケーションを再開します。
- ▶ **Active-Activeレプリケーション** — InfiniBoxシステムのActive-Activeレプリケーションにより、メトロエリア にわたって、コンシステンシーグループに対して同時に読み取りと書き込みを実行できます。それらは、同じ容量への複数パスであるかのように現れる容量の外部イメージを維持し、同期レプリケーションを利用して容量が常に一定になるようにします。マスター・スレーブの関係がなくとも、任意の容量に更新書き込みを実行するのに外部ラウンドトリップは必要ありません。外部の、軽量の「witness」はスタンドアロンのノード、もしくはクラウドの仮想マシンに存在することができます。

データ削減

InfiniBox では、以下のように複数のデータ整理手法を取り入れており、ストレージのコストをさらに削減することができます。

- ▶ **シンプロビジョニング** — すべての容量は、デフォルトでシンプロビジョニングされています。InfiniBoxはスマート容量プールも提供しているため、プールのアラート閾値や緊急のバッファを設定することで、割り当て超過／プロビジョニング超過のリスクを簡単に軽減することができます。結果として、アプリケーションの可用性を保護します。
- ▶ **ゼロリクレーション** — (物理または仮想)ホストがディスク(LUN)の領域をクリアすると、Write Same操作(より効率的)により、その領域にゼロを書き込みます。あるいは、シンプルにその領域に個別のゼロを書き込みます。InfiniBoxは、どちらの場合も識別し、書き込みがなかったかのようにこの領域を除去して、さらにシンプロビジョニングを改善します。
- ▶ **圧縮** — InfiniBox は、書き込みキャッシュ(DRAM)からディスクにデステージしてからのみ、データを圧縮します。これにより書き込みを高速化し(データ整理のためレイテンシは追加されない)、一方で数秒後に上書きされる一時データを圧縮しないようにします(CPUリソースを節約)。InfiniBoxの圧縮は、チャンクサイズ64KiB のLZ4を利用し、従来の小ブロック圧縮(一般的にフラッシュアレイすべてで使用される)よりも高い圧縮比を実現します。
- ▶ **スナップショット** — のスナップショットは仕様で組み込まれています。フルコピーにかかる容量やパフォーマンスのペナルティを回避するのに役立ちます。

ネットワークアーキテクチャ

ネットワークベースのサービスにおける可用性を確保するには、ネットワークの接続性が欠かせません。具体的に、インターネットプロトコル (IP) をベースとしたサービス (iSCSI、NFS、非同期レプリケーション、同期レプリケーション) では、IT 管理者は一般的に、ストレージシステムがフェイルオーバーに対応すること、またすばやく構成の問題を解決することを期待します。InfiniBox は、接続性の問題発生時にインスタント IP フェイルオーバーを使用してこの分野でイノベーションを起こし、IP アドレスを、関連サービスを提供できるネットワークインターフェイスに移します。

インスタント IP フェイルオーバーは、ハードウェア (ノード障害、イーサネットポート / ネットワークカード障害) とソフトウェア (特定ノードのサービス障害) の双方を含めたすべての障害シナリオに適用されます。他のサービスへの影響を最小化するために、当該ノードの異なるサービスの IP、または他のノードの IP が移動しないよう、InfiniBox は最小数の IP アドレスを移動します。

InfiniBox はさらに、仮想 MAC アドレス (VMAC) を利用し、それぞれの IP アドレスを VMAC に割り当てます。IP アドレスが移動すると、VMAC アドレスもそれに伴って移動します。これによりフェイルオーバー時間が削減され、それぞれのホストに変更を行わずに、スイッチの構成変更を行います。さらに、それにより Gratuitous ARP の問題を回避し、可用性を高めます。

InfiniBox はスマートネットワークモニタリング (IPv6 ping を使用) を採用して、構成ミスの可能性を識別します。構成ミスというのは、たとえば、ストレージネットワークインターフェイスがデータサービスに使用する VLAN にアクセスできなくなる、といったことです。InfiniBox で構成されたネットワークはそれぞれ常時監視されているため、仮にアプリケーションがストレージへのアクセスを失うといった事態になった場合でも、ストレージ管理者はすぐに問題解決にいたることができます。

ハードウェアアーキテクチャ

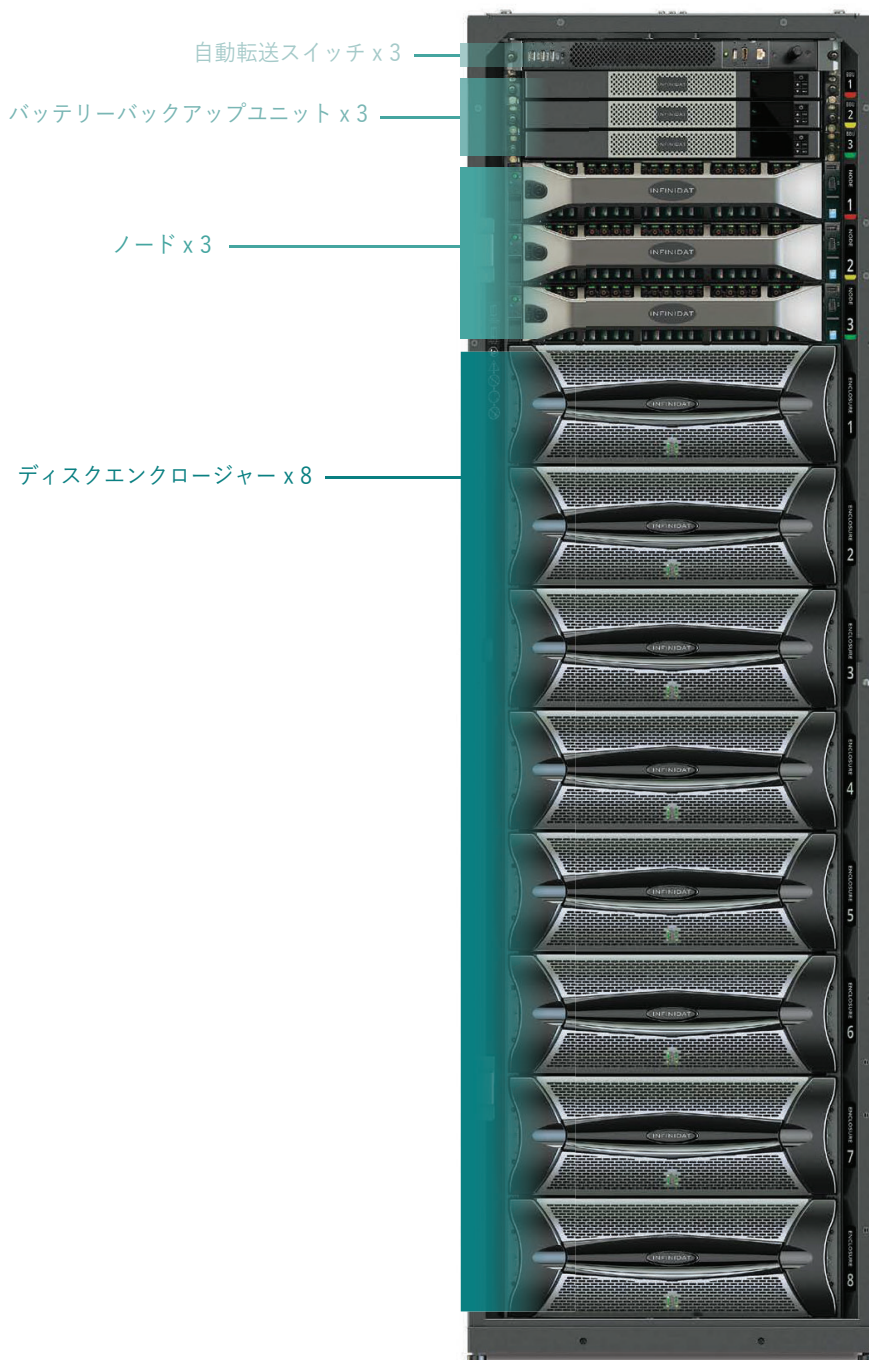
InfiniBoxはソフトウェアにより定義されたストレージシステム(SDS)であり、COTSハードウェアを活用しています。設計の一環として、InfiniBoxはソフトウェアに投資し、COTSハードウェアの信頼性や費用対効果を高め、管理やサポートを簡素化しています。最も欠かせない設計原則はN+2です。すべてのコンポーネントが少なくとも三重の冗長性を備え、99.99999%の信頼性を確保するための設計がなされています。InfiniBoxソフトウェアと組み合わせると100%のシステムの可用性が実現します。

InfiniBoxシステムはここに示す通り、ラックにあらかじめ組み立てられた状態で提供します。

ノード

ノードはInfiniBoxのストレージコントローラです。3つの完全に冗長化されたノードがアクティブ・アクティブ・アクティブクラスターで動作し、I/Oが3つのノードすべてを通じてシームレスに流れるようにします。ノードは高速InfiniBandと直接相互接続し、RDMAを使用してメモリーに直接アクセスします。これにより、可能な限り最低のレイテンシで、新たな書き込みをノード間ですばやく複製できるようにします。

1つのノードで障害が発生した場合、残りの2つのノードがその障害に対処します。複製されない書き込みキャッシュのいずれかの部分を再同期して完全なデータ保護を再開し、無停止で動作を維持します。また、N+2ノードアーキテクチャは、特定ノードの保守業務(例:コンポーネントの交換)を簡素化します。システムが2つのアクティブ・アクティブノードを起動し、データを保護しているためです。



画像1 ラック前面

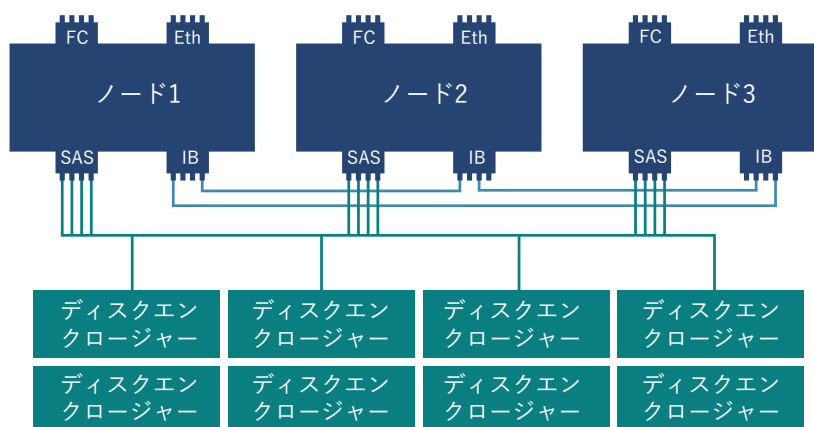
物理的接続性

ノードから顧客ファブリックへのフロントエンドの接続性

- ▶ **ファイバーチャネル(FC)** — ノードごとに8個のポート、合計24個のポート。すべてのポートがアクティブなため、それぞれのホストが複数のパスを確認します(ノードごとに少なくとも1つ。推奨は2つ)。マルチパスにより、ポート障害またはHBA障害は個々のパスにのみ影響を与え、アプリケーションには影響を与えません。
- ▶ **イーサネット(Eth)ポート** — ノードごとに最大6個のポート、合計18個のポートは、銅接続または光接続を提供し、iSCSI、NVMe/TCP、NFS、同期レプリケーション、非同期レプリケーションプロトコルをサポートします。これらのポートは、スマートIPフェイルオーバーをサポートし、物理的障害がシステムのアクセス性に影響を及ぼさないようにしています。

内部では、ノードは冗長的バックエンド接続性も提供します。

- ▶ **InfiniBand (IB)** ポート — クラスターの相互接続に使用します。別のノードからのノード切断を引き起こすInfiniBand障害は、これらの2つのノードが3つ目のノードで通信する原因となります。ノードが残りの両ノードから切断されると、切断が解決されるまでクラスターから正常に取り除かれます。
- ▶ **SAS** ポート — ノードをすべてのディスクエンクロージャーに接続します。特定のノードがディスクの一部へアクセスできなくなるSAS障害が発生した場合、InfiniBandを使用して別のノード経由で遠隔からそれらのディスクにアクセスします。



画像2 InfiniBox のフロントエンドとバックエンドの接続性

ノードには冗長電源があり、異なるバッテリーバックアップユニット(BBU)から給電されます。同様にBBUは、複数のパワーインレットを通じて給電されるため、電源の問題が生じても無停止で動作することができます。

自動電源切替スイッチ

自動電源切替スイッチ(ATS)が、BBUへの給電を制御し、電源の1つに停電が生じた場合でも、バッテリーに常時入力電流が流れるようにします。ATSは2つの電源の1つに障害が発生すると、瞬時に切り替えることができ、BBUへの電力が中断しないようにします。

バッテリーバックアップユニット

BBUは、短時間の停電が発生した場合、復旧するまでの間(例:発電機が完全にアクティブになるまでの間)、InfiniBoxのノードへの電力を維持し、システムを停止しなければならない状況を回避します。また、長時間の停電が発生した場合は、電力を供給してDRAMキャッシュから適切にデータをデステージし、InfiniBoxが常に適切なシャットダウン手順を踏めるようにします。BBUは監視され、それぞれが週に1度自動的に検査されます。バッテリーの状態が正常であるか確認し、停電が実際に発生した場合にシステムを確実に保護できるようにします。

結論

独自のInfiniBoxアーキテクチャは、信頼性、パフォーマンス、容量、総保有コストそれぞれの間にこれまで存在したトレードオフを解消しました。ストレージイノベーションの推進力としてソフトウェアを重視することにより、時間とともに改良できるソリューションを実装できるようになったのです。こうして、IT関連の組織は、予算は限られながらも、ビジネスの可能性を得ようと新たな取り組みを実施することができています。InfiniBoxを使用すれば、かつてないほどの確信を持って、デジタルトランスフォーメーションやビッグデータイニシアチブに取り組むことができるのです。