INFINIDAT

WHITE PAPER

INFINIDATのストレージアーキテクチャ

概要

INFINIDAT® のエンタープライズストレージソリューションは、特許取得済みのINFINIDAT独自のストレージアーキテクチャをベースに開発されており、完全に抽象化されたソフトウェア定義型ストレージ（SDS）をベスト・オブ・ブリードのコモディティハードウェアに統合しています。十分に検証されたハードウェアレファレンスプラットフォームと組み合わせてソフトウェアを出荷することによって、INFINIDATは業界で初めて、真にエンタープライズクラスのSDSソリューションを提供します。

このホワイトペーパーでは、セブン・ナイン（99.99999%）の比類なき可用性、ミリ秒以下の低遅延でフラッシュを上回る1M IOPS以上の高速性能、42Uシングルラックでマルチペタバイト規模の大容量、すべてを低総所有コスト（TCO）で提供し、卓越した信頼性を実現する唯一のプロバイダーであることを可能にしているINFINIDATのテクノロジーについて解説します。

設計原理

ストレージアーキテクチャの設計にあたっては、最新のデータセンターのニーズに対応するために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	24x7の運用、ダウンタイムは許容されない
容量	デジタル変革の加速に伴って飛躍的に増加するデータ、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)
性能	データ規模が拡大しても、同様(もしくは より短時間)のタイムフレームで同様(もしくはそれ以上)のパフォーマンスを実現 できなければならない
シンプルさ	シンプルな運用、広範なエコシステム統合、DevOpsモデルへ移行するためのビルトインツール、ストレージ管理に要する時間の短縮によって、管理者はよりアプリケーションやビジネスプロセスに集中できることを期待している
連携	ポイントテクノロジーはもはや過去のものであり、最新のストレージはあらゆるユースケースに対応し、効率や簡素化、コスト効率を最大限に高めることができなければならない
コスト	容量やパフォーマンスの増大に合わせて予算枠が拡大するわけではない。アーキテクチャレベルで破壊的な変革が求められる

同時に、Amazon、Google、AzureはITスタック全体のコスト削減をうたい、多数のITスタッフを揃えることのできない小規模な組織では、1人か2人のスタッフですべてをまかなってITオペレーション全体を維持するケースもよくあります。しかし、大企業やリージョナルクラウドプロバイダーの場合はビジネスやテクノロジー、財務上の要件に合った、より効率性の高いITスタックを導入することで、自社のインフラストラクチャ内でクラウドのあらゆる利点を提供しながらコストを削減し、データ主権を維持することができます。

INFINIBOXのアーキテクチャ

INFINIDATのフラッグシップ製品であるInfiniBox®は基本原理を念頭に、以下のすべての課題を解消できるよう設計されました。

主な目的	論拠	課題
革新的なソフトウェア	ハードウェアと違い、ソフトウェアは時間とともに最適化され、性能は劣化するのではなく向上していく。InfiniBoxは80以上の取得済み特許をベースとする真の意味のソフトウェア定義型(SDS)ストレージ	**性能** **シンプルさ** **信頼性** **コスト**
回復力に優れた設計	スケール設計では回復性が重要。InfiniBoxはすべての主要コンポーネント(ソフトウェアおよびハードウェア)に少なくとも2冗長構成(N+2)にしてダウンタイムやデータロスを防止する3重冗長構成でセブン・ナイン(99.99999%)のアップタイムを実現する設計	**回復性** **コスト** **シンプルさ** **連携**
大規模アーキテクチャ	破壊的な価格で容量と性能を実現するにはスケールが要求される。InfiniBoxは大規模顧客向けに設計され、42Uシングルラックで実効容量230TB以上、最大8.3PBまで拡張可能	**連携** **コスト** **シンプルさ** **容量**
ハードウェアとソフトウェアのタイトな統合	InfiniBoxが卓越した回復性を実現するためのさまざまな方法の1つとして、複数のハードウェアコンポーネントの検証を行った後、最も信頼性の高いもののみ市場に投入(リファレンスアーキテクチャ)。このアプローチによって、顧客は現場でのハードウェア統合にかかるコストや複雑な手間をかけることなく、管理オーバーヘッドなしに完全に統合したソリューションを得ることができる	**信頼性** **シンプルさ** **連携**
オフザシェルフのコモディティハードウェア(COTS)	コモディティハードウェアを使用して長期に及ぶ開発サイクルを回避することで、新規テクノロジーのより迅速な導入が可能に。対象はCPU、メモリタイプ、新種のストレージメディアが含まれる。すでに世界中で数千のシステムで使用されているコモディティハードウェアとそれに付随したソフトウェアを利用することでより高い確実性が得られる	**コスト** **信頼性** **容量** **シンプルさ** **性能**

パフォーマンスを加速

InfiniBox®はDRAM、Flash media（SSD）および高容量のNL-SASディスクを使ってデータの書き込み、読み出し、格納を行うフラッシュ最適化アレイです。以下、InfiniBoxがどのように読み出し/書き込みを高速化し、最小限の遅延でパフォーマンスを最大化しているのかを説明します。

データの最適配置のために用いられているのが、ニューラルキャッシュ（Neural Cache）と呼ばれるアルゴリズムです。このセクションではスマートなソフトウェアアルゴリズムを活用することで、Neural Cacheがどのようにオールフラッシュアレイよりも低遅延を実現するのかを解説します。

ほとんどのトランザクションアプリケーションには、少なくとも2つのI/O（1つはトランザクションのログ書き込み用、もう1つはデータベースへのデータ書き込み用）が必要であり、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両方を決定する上で、これが主要コンポーネントの遅延を生じさせていることを覚えておくことが重要です。

メタデータレイヤー

メタデータレイヤーの応答時間はI/Oの遅延に直接影響を及ぼします。InfiniBoxは以下のようにメタデータの運用を高速化します。

- **メタデータはすべてDRAM内**：メタデータはDRAMに格納され、読み取り/書き込みを高速化
- **効率的なメタデータ構造**：挿入、修正、削除をすべてメタデータ構造（トライ木）から行い、遅延を同じにすることで一貫したパフォーマンスを提供

書き込みの高速化

InfiniBox は事前処理（パターン削除、圧縮、暗号化など）なしですべてをDRAMに書き込むことができ、低遅延のInfiniBandでホストに受け取りを送信する前に別のノードのDRAMでセカンドコピーを作ります。外付けのフラッシュデバイスの代わりにDRAMから書き込み（CPUに直付け）できるようにすることで、InfiniBoxは遅延を最低に抑えて書き込みを完了できます。

書き込みキャッシュをバケットに小さく分解する多くのアーキテクチャ（マトリックス構造やデュアルコントローラ構造など）とは違い、InfiniBoxは大容量の単一のメモリプールを使って書き込みを行います。これによってより大量の書き込みバーストも持続でき、頻繁に変更されるデータもDRAMレイテンシーで上書きし、どのデータをブロックすればDRAM速度を最も高められるのか、どのデータをSSDやHDDにデステージするべきか、Neural Cacheが高速でスマートな判断を下すことができます。書き込みキャッシュにデータを長く格納しておくことで、Neural CacheはCPUやバックエンドの永続性レイヤーに不要な負荷をかけずに済みます。

読み出しの高速化

最もアクティブなデータ、いわゆるホットデータのほとんどをフラッシュキャッシュに置いて、オールフラッシュアレイ（AFA）と同等のパフォーマンスを実現しようとする従来のストレージアレイとは異なり、InfiniBoxは革新的なNeural Cacheを使ってホットデータをすべてDRAMに格納しようとします。InfiniBox のNeural Cacheによって、読み取りのほとんどはDRAM速度で完了するため、従来のフラッシュに比べて1000倍高速になります。

2017年時点でINFINIDATのグローバルデータファブリックは数エクサバイトまで拡大し、Neural Cacheはほぼすべての読み出しをDRAMから実行できることを実証して、INFINIDATの導入企業は“オールDRAMアレイ”でAFA並みのエクスペリエンスが低コストで体験できるようになっています。

Neural Cacheは学習可能なアルゴリズムであり、時間の経過とともにパフォーマンスを最適化します。InfiniBoxはシックSSDフラッシュを活用し、DRAMのキャッシュミスのための“クッション”として機能させます。Neural CacheがI/Oパターンを学習してDRAMのデータ配置を最適化するにつれて、フラッシュレイヤーはDRAMミスの対応からアルゴリズムで予測不能な場合のI/Oパターンの変更へと機能を変えます（DRAM外のデータを要する定期監査など）。

ソフトウェアアーキテクチャ

InfiniBoxの設計にあたり、INFINIDATは99.99999%の信頼性を持続させるために、ハードウェアの不測の障害をソフトウェアを使って克服しました。InfiniBoxではトリプルアクティブのソフトウェアアーキテクチャとN+2設計で常時監視と自己修復機能を提供し、あらゆるレベルでハードウェア障害が発生しても正常にリカバリーすることができます。

RAIDからクラスタ化されたサービスまで、コンポーネントはすべてソフトウェアに実装され、新規リリースごとに常に最適化できます。InfiniBoxは一般提供の開始(GA)以来4年が経過し、ソフトウェアの無停止アップグレードによって性能は初回出荷時から4倍に向上しています。これこそ真のソフトウェア定義型ソリューションの威力と言えます。

クラスタ化されたサービス

すべてのデータサービスはあらゆるノードでN+2のアーキテクチャ設計に従って動作し、すべてのノードでアクティブ(クラスタ内のパッシブノードなし)になります。データサービスは、ファイバーチャネル(FC)ドライバのような低レベルコンポーネントを含め、ユーザースペースで動作するよう設計されています。カーネルで動作するデータサービスは存在しないため、障害がシステム内のほかのサービスに影響したり、ノードの可用性を低下させたりすることはありません。この原理はデータプロトコル(NFS、iSCSI、FC、FICON)のようなフロントエンドのサービスだけでなく、Neural Cache、InfiniRAID®、InfiniSnap®のバックエンドのデータサービスの設計にも適用されています。

データサービスはクラスタマネージャー(CLM)によって起動し、監視が行われて、問題を見つけると必要に応じてリスタートできます。何らかの障害が発生したサービスはリスタートし、自己診断を行ってから再びクラスタに戻ります。正常に起動できなかったサービスはクラスタ内での障害発生(ビザンチン障害)を避けるため、クラスタには加わりません。クラスタマネージャーは、特定のノードで起動に複数回失敗したサービスを特定した場合、再起動を停止してINFINIDATのサポートに通知します。

自動リカバリーの有無にかかわらず、あらゆるサービス障害はINFINIDATのデータアナリティクスプラットフォームに報告されて、ソフトウェアの問題を検出し継続的にコードの質を改善します。

ディスクレイアウト

InfiniBoxのディスクレイアウトは特許取得済みのINFINIDATの革新的なソフトウェアInfinRAIDによって管理されます。InfinRAIDはソフトウェア定義のRAID(Redundant Array of Independent Disks)で、すべてのデータの配置とデータ保護、障害発生シナリオからのリカバリーを制御します。

InfiniRAIDはデータレイアウトと物理レイヤーを切り離し、何千もの仮想RAIDグループを使用するRAIDの一種、非クラスタ化RAIDで、すべてのドライブを使用してデータを分散し、ホットスポットを防ぎます。InfiniRAIDはRAIDグループを生成するため、システム内のドライブ2個ごとにRAIDグループと共有するデータは最大2.5%以下に抑えられます。

RAIDグループのオーバーラップの割合を低くしておくことには、いくつかの利点があります。

- **均等な分配**:あらゆるデータセットはサイズに関わらず、システム内のすべてのドライブに分散され、それぞれのアプリケーションのスループットを最大化します。
- **自己修復機能**:データレイアウトを最適化してホットスポットの可能性を自動的に解消します。
- **仮想スペア**:システム内で空き容量をすべてのディスクに均等に分散します。物理的なホットスペアがないことにより、再構築プロセスでデータの最適な再配分を行い、不要なコストを最少化します。F6000の場合、最大12ドライブまで拡張でき、システム内に十分な空き容量を確保します。
- **パフォーマンス保護**:1つのドライブに障害が生じても(データ保護は継続)RAID再構築の優先度は低いため("Rebuild-1")、アプリケーションのパフォーマンスを優先してシステムの予備リソースのみを使って実行されます。
- **迅速なリカバリー**:2つ目のドライブに障害が生じた場合、システムは中断した2つのドライブ間で共有するRAIDグループの2.5%の共有部分の再構築("Rebuild-2")を加速し、保護されないRAIDグループを解消した後に優先度の低いRebuild-1に戻ります。
- **InfiniSpares**:12のスペアに相当する容量を保証する以外に、InfiniBoxは必要に応じて空き容量をスペアとして活用することもできます。この革新的な手法によって、データ保護を損なうことなく最大100個のディスクの障害に対応できます。

データ保護サービス

InfiniBoxは多数のデータ保護サービスを提供して、顧客の資産を守ります。

- **スナップショット**：InfiniBoxにはノンロッキング、リダイレクトオンライト（redirect on write）の手法をベースにしたスナップショット機能InfiniSnapを備え、スナップショットの有無に関わらず一貫したパフォーマンスを生み出します。データセットごとに読み出し専用（データ保護用）または書き込み可能（テスト&開発環境用）のスナップショットを最大1000まで保存できます。InfiniSnapはDRAM内でスナップショットを実行するため、永続性レイヤーには一切書き込みを要求しません。
- **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秒間のRPO（Recovery Point Objective）を維持すると同時に、IPインフラストラクチャを利用することでコストと複雑さを低減できます。
- **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400マイクロ秒以下のストレージレイテンシーを維持しながらゼロRPOの同期データ保護を提供します。WANに問題が生じた場合（高遅延、接続消失など）、InfiniBoxの同期レプリケーションエンジンは自動的に非同期モードに戻ります。WANの回復と同時に自動的にレプリケーションできなかったすべてのデータを複製し、I/Oを中断することなく同期レプリケーションを再開します。

データ整理

InfiniBoxは複数のデータ整理手法を備え、ストレージコストをさらに削減します。

- **デフォルトのシンプロビジョニング**：すべてのボリュームはデフォルトでシンプロビジョニングされます。InfiniBoxはスマート容量プールも提供するため、割り当て超過/オーバープロビジョニングのリスクは、プールにアラートの閾地と緊急情報バッファを設定することで容易に回避でき、アプリケーションの可用性を保護します。
- **ゼロリクラメーション**：ホスト（物理または仮想）はディスク内のスペースを空け（LUN）、このスペースにwrite-sameオペレーションを通じてゼロを書き込む（より効率的）か、単純に1つ1つスペースにゼロを書き込みます。InfiniBoxは両方のケースを識別してこのスペースを削除し、元々書き込みが行われなかったようにすることでシンプロビジョニングを改善します。
- **圧縮**：InfiniBoxではデータの圧縮を書き込みキャッシュ（DRAM）からディスクへデステージする時に一回だけ行います。これによって書き込み速度を向上する（データ処理のための追加遅延なし）と同時に、数秒で上書きされる一時データを圧縮することのないようにします（CPUリソースの節約）。InfiniBoxはチャンクサイズ64KBのLZ4の圧縮機能を活用しており、従来のスモールブロック圧縮（一般的にオールフラッシュアレイで使用される）よりも高い圧縮率を実現します。
- **スナップショット**：InfiniBoxのスナップショットはシンで設計されており、フルコピーで容量にペナルティを与えないために役立ちます。

ネットワークアーキテクチャ

ネットワークベースのあらゆるサービスにとって、可用性を高めるためにはネットワークのアクセシビリティが重要です。特にインターネットプロトコル(IP)ベースのサービス(iSCSI、NFS、同期/非同期レプリケーションなど)に対し管理者は通常、フェールオーバーに対応し、コンフィギュレーションの問題を素早く克服できるストレージシステムを期待します。InfiniBoxはインスタントIPフェールオーバーを使ってこの分野で革新を起こし、接続に問題が生じた場合、IPアドレスをネットワークインターフェースに移して必要なサービスを提供できるようにしました。

インスタントIPフェールオーバーはハードウェア(ノード障害、Ethernetポート/ネットワークカード障害等)とソフトウェア(特定ノードでのサービス停止)の両方を含めて、あらゆるケースの障害に適用されます。他のサービスに与える影響を最少化するため、InfiniBoxは移動するIPアドレスは最低限に留め、同じノードの異なるサービスや別のノードのIPは移動しないようにします。

InfiniBoxは仮想MACアドレス(VMAC)も活用し、各IPアドレスにVMACを割り当てます。IPアドレスを移動する際にはVMACアドレスも合わせて移動します。これによって、フェールオーバーの時間をなくし、コンフィギュレーションの変更は各ホストに伝達することなく、スイッチ上で行うようになります。さらに、ARPに余分な問題を生じさせることを避けることで可用性を高めます。

InfiniBoxはスマートネットワーク監視(IPv6ピンを使用)を採用して潜在的なコンフィギュレーションエラーを特定し、ストレージネットワークのインターフェースがデータサービスで使用するVLANへのアクセスからブロックされる事態などが生じないようにします。InfiniBox内の各ネットワークの設定は常に監視され、「このアプリケーションのストレージへのアクセスが失われたのはなぜか」とストレージ管理者が疑問に感じるよりもずっと早く、しばしばこの段階で解消されます。

ハードウェアアーキテクチャ

InfiniBoxはコモディティ(COTS)ハードウェアを活用したソフトウェア定義型ストレージシステムです。設計の段階でINFINDATはCOTSハードウェアの信頼性やコスト効率を高め、管理やサポートを簡素化するためのソフトウェアに投資を行いました。セブン・ナイン(99.99999%)の信頼性を実現するために最も重要なのがN+2の設計原理で、すべてのコンポーネントは3重以上の冗長構成になっています。InfiniBoxのシステムはあらかじめ以下の構成で1ラックに格納されています。

ノード

ノードはInfiniBoxの中のストレージコントローラです。完全冗長化された3つのノードがトリプルアクティブクラスタ内で動作し、I/O をシームレスに3つのノード間で動かすことができます。ノードは高速なInfiniBandで直接に相互接続し、RDMAを使用してメモリに直接アクセスするため、新規の書き込みをノード間で素早く複製でき、遅延を最低限に抑えることができます。

1つのノードで障害が生じると残りの2つのノードが役割を引き継ぎ、複製されなくなった書き込みキャッシュのどの部分でもリサイクルして完全なデータ保護を再開し、中断なしのオペレーションを維持します。N+2のノードアーキテクチャーも特定のノードのメンテナンスオペレーション(コンポーネントの入れ替えなど)を簡素化し、システムには引き続きダブルアクティブノードが動作してデータを保護します。

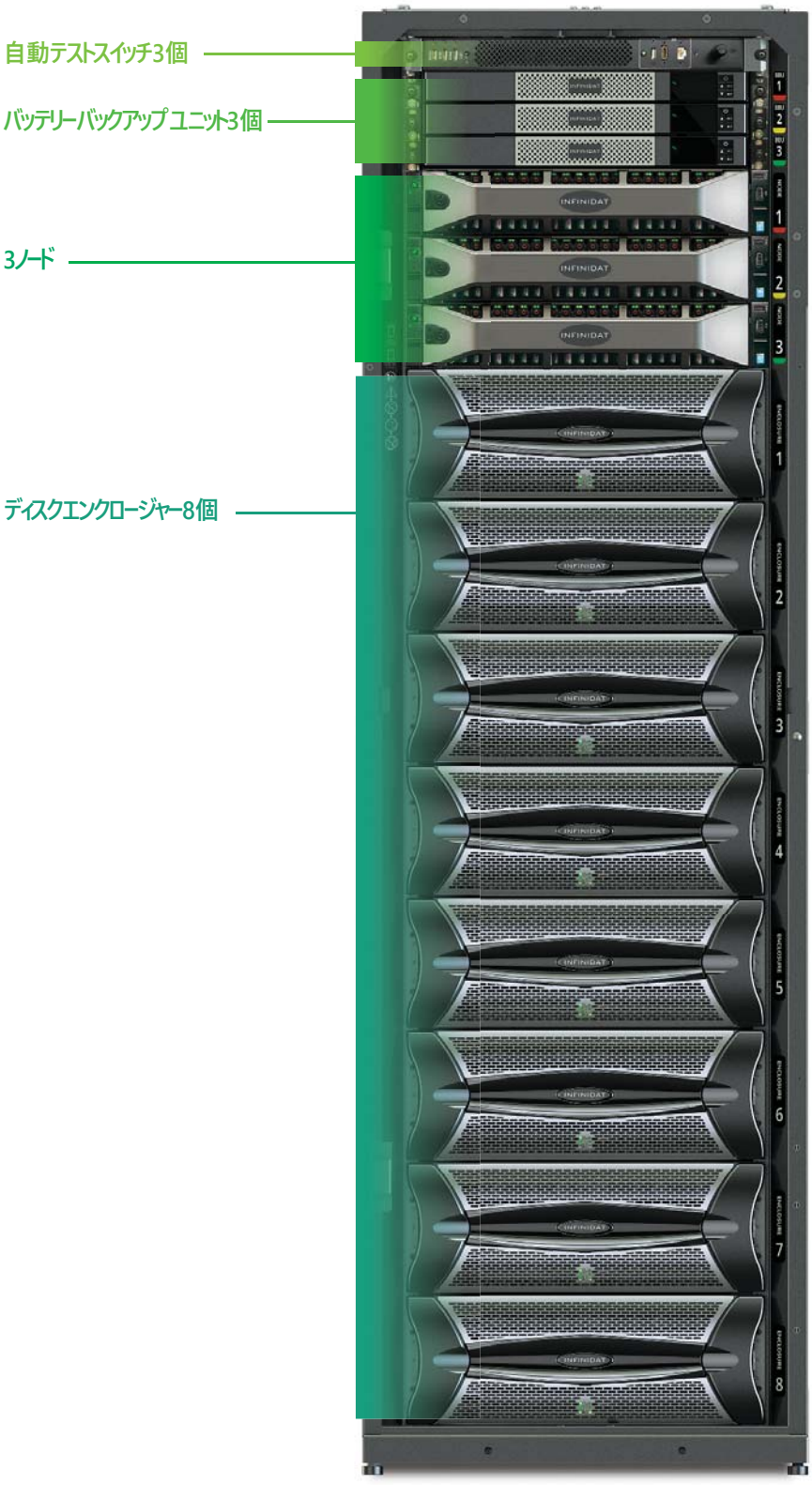

図 1 InfiniBoxのフロントエンド/バックエンドの接続性

物理的接続

ノードから顧客の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8ポート、合計24個のポート。ポートはすべてアクティブで、それぞれのホストには複数のパスが存在（1ノードにつき1個以上、1ノードに2個を推奨）。マルチパスによってポートやHBAに障害が発生しても影響はそれぞれのパスに限定され、アプリケーションには影響しません。
- **Ethernet（Eth）ポート** – ノードごと4ポート、合計12個のポート。カッパーもしくはオプティカル接続を提供し、iSCSI、NFS、同期/非同期レプリケーションプロトコルのサポートおよびInfiniSync（INFINIDAT独自の距離無制限のゼロRPOソリューション）との統合をサポート。スマートIPフェールオーバーをサポートし、あらゆる物理的障害によるシステムの接続性への影響を防止。

  内部的には、ノードはバックエンドの冗長接続も提供します。
- **InfiniBand（IB）ポート** – クラスタの相互接続に使用。InfiniBandの障害で他のノードからの接続が中断されると、この2つのノードは第3のノードを通じて通信を行います。2つのノードの両方との接続が中断した場合、このノードは中断が解消されるまでクラスタから適切に除外されます。
- **SASポート** – ノードをすべてのディスクエンクロージャーと接続します。SASの障害によって特定のノードから一部ディスクへのアクセスが失われた場合、InfiniBandを使って別のノードからリモートでこれらのディスクにアクセスします。
  ノードには冗長電源が提供されるとともに、別のバッテリバックアップユニット（BBU）から電力が供給されます。複数の電源インレットを通じて電力が供給されることで電力に問題が生じてもオペレーションを中断することなく運用できます。

自動テストスイッチ（ATS）

ATSはバッテリーバックアップユニット（BBU）への電力供給を制御し、電源の1つが停電していてもバッテリーに常に電流が流れるようにします。ATSは2つの電源を瞬時に切り替えることができ、1つが停止してもBBUへの電源を維持します。

バッテリーバックアップユニット（BBU）

BBUは短時間の停電の間（発電機が完全に稼働するまで）、InfiniBoxのノードの電源を維持し、システムをシャットダウンせずに済むようにします。停電が長引いた場合にはDRAMキャッシュからデータを適切に削除し（デステージ）、InfiniBoxがいつでも適切な手順を踏んでシャットダウンできるようにします。

BBUはモニタリングされ、ユニットごとに1週間に1度、テストを自動的に実施して電源が正常に機能することを確認し、実際に停電が発生した場合にシステムを保護できる状態を保ちます。

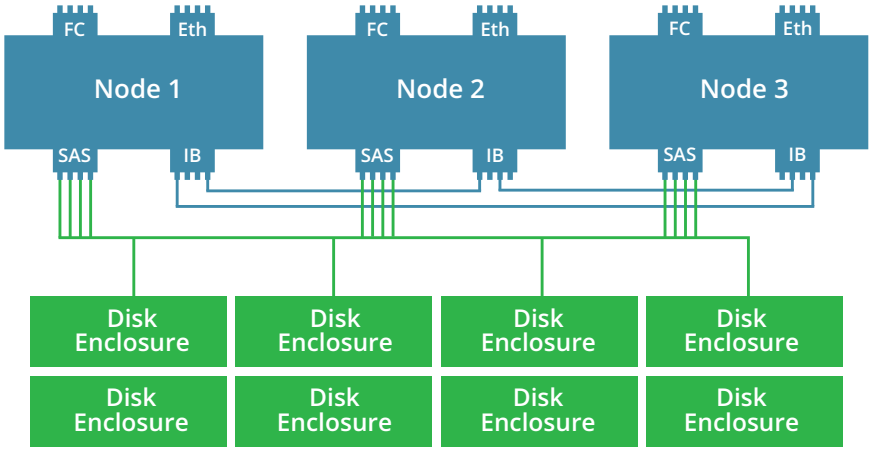

図 2 InfiniBoxのフロントエンド/バックエンドの接続性

まとめ

InfiniBoxの独自アーキテクチャは性能、耐障害性、容量、コストのいずれかに妥協を強いられる従来の考え方を打ち破ります。IT部門は初めて、IT予算を増大させたり自社の利益を損なったりすることなく、自社のクリティカルなデータをすべて蓄えられるようになります。InfiniBoxを導入することで、企業は自信を持ってデジタルトランスフォーメンションを進め、ビッグデータの活用に取り組むことができるでしょう。

jp_sales@INFINIDAT.com | 03-4243-6343