

ETERNUS AX series オールフラッシュアレイ , ETERNUS HX series ハイブリッドアレイ

MetroCluster ソリューションアーキテクチャと設計

目次

1.	MetroCluster の概要	7
1.1	SyncMirror テクノロジーによるデータ保護	8
1.2	MetroCluster を使用した HA データセンター	8
1.3	キャンパス、都市、地域レベルの保護	9
1.4	選択可能な保護レベル	9
1.5	WAN ベースの DR	9
1.6	管理の簡略化：初回設定のみ	9
1.7	アプリケーションの透過性	9
2.	アーキテクチャ	10
2.1	MetroCluster の物理アーキテクチャ	10
2.2	MetroCluster IP の機能	10
2.3	ローカルフェイルオーバー (HA) およびリモートスイッチオーバー (DR)	11
2.4	MetroCluster レプリケーション	13
2.4.1	構成レプリケーション	13
2.4.2	NVRAM レプリケーション	14
2.4.3	ストレージレプリケーション	15
2.4.4	SyncMirror ストレージレプリケーション	16
2.5	アグリゲートのスナップショットコピー	17
2.6	アクティブ／アクティブおよびアクティブ／パッシブ構成	18
2.7	Advanced Drive Partitioning (ADP)	18
2.7.1	ルート - データ - データ (RD2) パーティショニング	19
2.8	ミラーリングされていないアグリゲート	20
3.	展開オプション	21
3.1	拡張構成および拡張ブリッジ構成	21
3.2	IP 構成	21
4.	計画的なイベントと計画外イベントの復元性	22
4.1	シングルノード障害	22
4.2	サイト全体のコントローラー障害	22
4.3	ISL 障害	22
4.4	複数の連続障害	23
4.5	4 ノードおよび 8 ノードの無停止運用	23
4.6	スイッチオーバー後のローカルフェイルオーバーの影響	24
4.7	スイッチオーバープロセスの概要	24

4.8	MetroCluster Tiebreaker	24
4.8.1	MetroCluster Tiebreaker による障害の検出	25
4.8.2	サイト間接続障害の検出	25
4.8.3	サイト間接続の監視	26
4.8.4	Tiebreaker によって監視されるコンポーネント	26
4.8.5	Tiebreaker 障害のシナリオ	26
4.9	ONTAP Mediator	27
5.	技術要件	28
5.1	ハードウェアとソフトウェアの要件	28
6.	まとめ	29

目次

図 1.1	MetroCluster	7
図 2.1	HA および DR グループ	11
図 2.2	8 ノードの DR グループ	12
図 2.3	NVRAM の割り当て	14
図 2.4	書き込みデータブロックのミラーリング	15
図 2.5	ミラーリングされていないアグリゲート Plex0	16
図 2.6	MetroCluster でミラーリングされたアグリゲート	16
図 2.7	ルートアグリゲートおよびデータアグリゲート	17
図 2.8	ADP の論理的側面	19
図 2.9	48 ドライブの MetroCluster IP 構成の ADP の例	19
図 2.10	MetroCluster のミラーリングされていないアグリゲート	20
図 4.1	MetroCluster Tiebreaker の確認	25

表目次

表 2.1	MetroCluster IP の機能	11
表 3.1	ハードウェア要件	21
表 4.1	障害のタイプと復旧方法	23

はじめに

本書では、ONTAP 9.12.1 ストレージ管理ソフトウェアに搭載された MetroCluster 機能の、高度なアーキテクチャと設計概念について説明しています。

第 2 版
2025 年 3 月

登録商標

本製品に関連する他社商標については、以下のサイトを参照してください。

<https://www.fujitsu.com/jp/products/computing/storage/trademark/>

本書では、本文中の™、®などの記号は省略しています。

本書の読み方

対象読者

本書は、ETERNUS AX/HX の設定、運用管理を行うシステム管理者、または保守を行うフィールドエンジニアを対象としています。必要に応じてお読みください。

関連マニュアル

ETERNUS AX/HX に関連する最新の情報は、以下のサイトで公開されています。

<https://www.fujitsu.com/jp/products/computing/storage/manual/>

本書の表記について

■ 本文中の記号

本文中では、以下の記号を使用しています。

注意

お使いになるときに注意していただきたいことを記述しています。必ずお読みください。

備考

本文を補足する内容や、参考情報を記述しています。

1. MetroCluster の概要

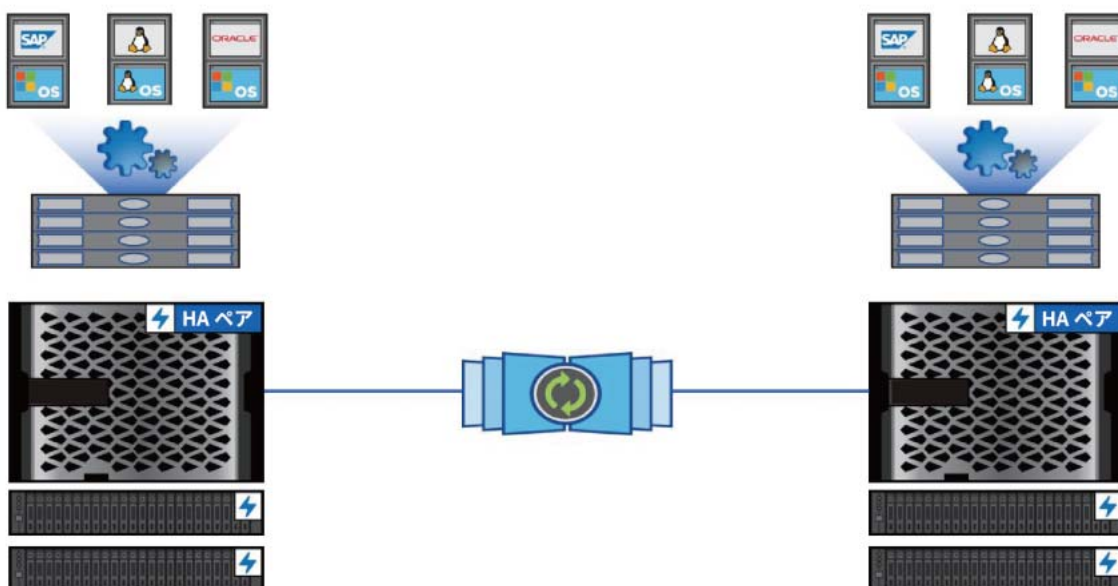
企業クラスのお客様は、コストと運用効率を維持しながら、増大するサービスレベル要件に対応する必要があります。データ量が急増し、より多くのアプリケーションが共有仮想インフラストラクチャに移行するにつれて、ミッションクリティカルなアプリケーションとその他のビジネスアプリケーションの両方の継続的な可用性に対するニーズが大幅に高まっています。

高度に仮想化されたインフラストラクチャ環境では、数百ものビジネスクリティカルなアプリケーションを実行するため、これらのアプリケーションが使用できなくなると企業は深刻な影響を受けます。このような重要なインフラストラクチャでは、データ消失ゼロで、システムリカバリを（数時間ではなく）数分で完了する必要があります。この要件は、プライベートクラウドインフラストラクチャ、パブリッククラウドインフラストラクチャ、および両者を組み合わせるハイブリッドクラウドインフラストラクチャに当てはまります。

MetroCluster ソフトウェアは、アレイベースのクラスタリングと同期レプリケーションを組み合わせて、継続的な可用性とデータ消失ゼロを実現するソリューションです。ホストベースのクラスタリングにつきものの依存性と複雑さが排除されるため、アレイベースのクラスタの管理が容易になります。MetroCluster は、すべてのミッションクリティカルなデータをトランザクション単位で即座に複製し、アプリケーションやデータへのアクセスを中断させません。また、従来のデータレプリケーションソリューションとは異なり、MetroCluster はホスト環境とシームレスに連携して、複雑なフェイルオーバースクリプトを作成および維持する手間を省きつつ、継続的なデータ可用性を提供します。MetroCluster には、以下の機能があります。

- 透過的なスイッチオーバーにより、ハードウェア、ネットワーク、またはサイトの障害から保護します。
- 計画的および計画外のダウンタイムと、変更管理の必要性を排除します。
- 運用を中断することなく、ハードウェアとソフトウェアをアップグレードします。
- 複雑なスクリプト、アプリケーション、オペレーティングシステムに依存することなく導入が可能です。
- VMware、Microsoft、Oracle、SAP、その他の重要なアプリケーションの継続的な可用性を実現します。

図 1.1 MetroCluster



MetroCluster は、組み込み型の高可用性 (HA) と、ハードウェアおよび ONTAP ストレージソフトウェアの無停止運用を強化し、ストレージおよびホスト環境全体の保護を追加します。スタンドアロンサーバー、HA サーバークラスタ、または仮想サーバーのうち、どのような環境であっても、MetroCluster はサイトのストレージが停止した場合のアプリケーションの可用性をシームレスに維持します。このようなシステム停止は、電力、冷却、またはネットワーク接続の損失、ストレージアレイのシャットダウン、または操作エラーによって発生することがあります。

1. MetroCluster の概要

1.1 SyncMirror テクノロジーによるデータ保護

MetroCluster は、アレイベースのアクティブ／アクティブなクラスタソリューションであり、複雑なフェイルオーバースクリプト、サーバーの再起動、アプリケーションのリスタートを必要としません。MetroCluster は、障害が発生した場合でも機能を維持するため、スイッチオーバーおよびスイッチバックの際もアプリケーションを中断させません。実際、MetroCluster を使用しているお客様のほとんどは、クラスタリカバリが行われてもアプリケーションの中断は発生しないと報告しています。MetroCluster は最高レベルの柔軟性を備えており、異なるプロトコルの混在をサポートすることで、あらゆる環境にシームレスに統合できます。

MetroCluster の導入には、以下のメリットがあります。

- 同期レプリケーションとアプリケーションへのシームレスなストレージプロモーションにより、厳格な SLA (ゼロ RPO [目標復旧時点]、および 2 分未満の RTO [目標復旧時間]) を達成します。
- 広範な SAN および NAS クライアント / ホストプロトコルに対するマルチプロトコルをサポートします。
- IP ネットワーク経由で同期レプリケーションをサポートします。
- MetroCluster 機能の使用料は不要です。
- 重要なデータのみをミラーリングします。具体的には、ミラーリングされたアグリゲートとミラーリングされていないアグリゲートをサポートしています。
- 外部 LUN インポート (FLI) によって、サードパーティストレージからのインポートが簡単に行えます。
- 重複除外、圧縮、コンパクションにより、ストレージとネットワークの効率性を実現します。
- SnapMirror テクノロジーとの統合により、非同期レプリケーション、距離、SLA の要件を満たします。

1.1 SyncMirror テクノロジーによるデータ保護

端的に言うと、同期レプリケーションでは、ミラーリングされたストレージの両側に変更が加えられます。たとえば、Oracle データベースがトランザクションをコミットすると、同期ミラー化されたストレージの REDO ログにデータが書き込まれます。ストレージシステムには、書き込み処理が両方のサイトの不揮発性メディアにコミットされるまで、書き込み処理の完了を認識させません。認識させないことによって、データ消失のリスクなしで安全に作業を進めることができます。

同期レプリケーションテクノロジーの使用は、同期レプリケーションソリューションの設計と管理における最初のステップにすぎません。最も重要なことは、さまざまな計画的および計画外の障害シナリオで何が発生するかを正確に把握することです。すべての同期レプリケーションソリューションが同じ機能を提供するわけではありません。お客様が RPO=0 (データ消失ゼロ) のソリューションを求めている場合は、障害シナリオを考慮する必要があります。サイト間の接続が切断されてレプリケーションを実行できない場合に予想される結果を、必ず確認する必要があります。

1.2 MetroCluster を使用した HA データセンター

RAID レベルで実施される MetroCluster レプリケーションは、データを同期ミラーリングする SyncMirror テクノロジーをベースにしています。SyncMirror を使用して、同じ WAFL ファイルシステムの 2 つのコピーで構成されるアグリゲートを作成できます。2 つのコピーはブックスと呼ばれ、同時に更新され、常に同一です。このテクノロジーは、通常の条件下で同期レプリケーションを必要とするほとんどのお客様の要件を満たします。

備考

部分故障でサイト間のすべての接続が切断された場合、ストレージシステムは動作を継続できますが、レプリケーションは行われません。

MetroCluster は、重要なビジネスアプリケーションに対して 24 時間 365 日の運用が必要な組織に最適なソリューションです。同一データセンター内の ETERNUS AX/HX series システム間、または建屋、キャンパス内、あるいは地域に存在する ETERNUS AX/HX series システム間でデータを同期レプリケーションすることにより、MetroCluster は通常の業務に影響を与えることなく、災害復旧 (DR) およびビジネスの継続を可能にします。

1.3 キャンパス、都市、地域レベルの保護

MetroCluster は、最大 700km までのサイト間距離を検証済です。このため、キャンパス全体や大都市全体で HA ソリューションの設計、実装、保守を大幅に簡略化することができます。サイト全体が停止している場合も、複雑なフェイルオーバースクリプトや再開手順は必要なく、自動化された単一のコマンドを使用して、データサービスを数秒でセカンダリサイトにリストアできます。

1.4 選択可能な保護レベル

ビジネスの継続を可能にする、新たなレベルの柔軟性を実現し、選択の幅を広げます。ONTAP 9 ソフトウェアを導入すると、ETERNUS AX/HX series システムが混在している場合でも、MetroCluster は 4 ノードから 8 ノード (レプリケーションの各エンドに 4 ノード) まで拡張できます。4 ノード構成から 8 ノード構成への拡張は、無停止のプロセスです。レプリケートするストレージプール (またはアグリゲート) を選択することもできるため、データセット全体を同期 DR 関係にコミットする必要はありません。

IP ネットワークを介した同期レプリケーションは、4 ノード構成と 8 ノード構成でサポートされます。

1.5 WAN ベースの DR

ビジネスが大都市圏の距離を超えて地理的に分散している場合は、SnapMirror ソフトウェアを追加して、グローバルネットワーク全体のデータを簡単かつ確実にレプリケートできます。SnapMirror ソフトウェアは、MetroCluster ソリューションと連携して WAN 接続経由で高速にデータをレプリケートし、地域レベルの障害が発生した場合も重要なアプリケーションを保護します。

1.6 管理の簡略化：初回設定のみ

ほとんどのアレイベースのデータレプリケーションソリューションでは、プライマリストレージアレイとセカンダリストレージアレイの間のレプリケーション関係がアレイごとに管理されるため、ストレージシステムの管理、構成、メンテナンスに重複した作業が必要になります。この重複によって管理オーバーヘッドが増加し、プライマリとセカンダリのストレージアレイ間で構成の不整合が発生した場合には、より大きなリスクにさらされるおそれがあります。MetroCluster はクラスタ化されたストレージソリューションであるため、アクティブ/アクティブストレージペアは単一のエンティティとして管理され、重複する管理作業を排除し、構成の整合性を維持します。

1.7 アプリケーションの透過性

MetroCluster は、フロントエンドアプリケーション環境に依存せずに透過的に動作するように設計されており、アプリケーション、ホスト、およびクライアントに対する変更が必要になることはほとんどありません。接続パスはスイッチオーバーの前後で同じであり、ほとんどのアプリケーション、ホスト、およびクライアント (NAS および SAN) は、ストレージを再接続または再検出する必要はなく、自動的に再開します。継続的可用性共有を使用する SMB 3 などの SMB アプリケーションは、スイッチオーバーまたはスイッチバック後に再接続する必要があります。これは SMB プロトコルの仕様です。

2. アーキテクチャ

MetroCluster は、ストレージインフラストラクチャとミッションクリティカルなビジネスアプリケーションの継続的な保護を必要とする組織向けに設計されています。地理的に離れたクラスタ間でデータを同期的にレプリケートすることにより、MetroCluster は、アレイの内部および外部の障害から保護する、ゼロタッチで継続的に利用可能なソリューションを提供します。

2.1 MetroCluster の物理アーキテクチャ

MetroCluster 構成では、最大 700 km 離れた位置にある 2 つの異なるクラスタを使用してデータを保護します。各クラスタは、他方のクラスタのデータと構成情報を同期的にミラーリングします。

事実上、すべてのストレージ仮想マシン (SVM) とそれに関連する構成がレプリケートされます。クラスタはそれぞれが独立しており、論理エラーからの切り離しが可能で、論理エラーからの復元力を備えています。

1 つのサイトで災害が発生した場合、管理者はスイッチオーバーを実行できます。これにより、ミラーリングされた SVM がアクティブ化され、障害の発生していないサイトからミラーリングされたデータの処理が再開されます。ONTAP でクラスタ化されている場合は、MetroCluster の 4 ノード構成は各サイトの 2 ノード HA ペアで構成されます。この構成では、計画イベントおよび計画外イベントのほとんどを、ローカルクラスタでの単純なフェイルオーバーとギブバックによって処理できます。他のサイトへのフルスイッチオーバーは、災害が発生した場合、またはテスト目的でのみ必要です。スイッチオーバーおよび対応するスイッチバック操作によって、クラスタ化されたワークロード全体がサイト間で転送されます。

2 つのクラスタとサイトは、レプリケーショントランスポートを提供する 2 つの異なるネットワークによって接続されます。クラスタピアリングネットワークは、サイト間でクラスタ構成情報をレプリケートするために使用される IP ネットワークです。共有ストレージファブリックは、2 つのクラスタ間のストレージおよび NVRAM 同期レプリケーションに使用される IP 接続です。MetroCluster IP の場合、NVRAM のレプリケーションに使用される iWARP とドライブレプリケーションに使用される iSCSI の両方のプロトコルが使用されます。すべてのストレージは、共有ストレージファブリックを通じてすべてのコントローラーから認識できます。

備考

iWARP (Internet Wide Area RDMA Protocol) は、イーサネットネットワーク上で Remote Direct Memory Access (RDMA) を有効にするネットワークプロトコルです。サーバー、ストレージシステム、その他のネットワークデバイス間の、高速かつ低レイテンシのデータ転送を可能にするとともに、従来のネットワーク通信プロトコルに関連するオーバーヘッドを削減します。

2.2 MetroCluster IP の機能

MetroCluster IP は、ファブリックに Ethernet/IP ISL を使用します。さらに、MetroCluster IP クラスタは、NVRAM と SyncMirror レプリケーションの両方に高速イーサネットを使用します。

MetroCluster IP には、他の非 MetroCluster トラフィックと共有するサイト間リンク (Layer2 で共有 VLAN、Layer3 で VIP/BGP) を使用するなど、運用コストを削減するいくつかの機能があります。ONTAP 9.7 から、MetroCluster IP は専用スイッチなしで提供され、MetroCluster IP の要件に準拠している既存のスイッチを使用できるようになりました。詳細については、「[MetroCluster IP 構成の設置](#)」を参照してください。

2. アーキテクチャ

2.3 ローカルフェイルオーバー (HA) およびリモートスイッチオーバー (DR)

表 2.1 は、2 つの MetroCluster サイト間でデータがどのようにレプリケートされるかを示しています。

表 2.1 MetroCluster IP の機能

機能	詳細
MetroCluster ファブリック	イーサネット /IP ISL
ファブリックファイバスイッチ	なし
SASブリッジ	なし
ファブリックイーサネットスイッチ	サイトごとに 2 つ
ファブリック 25G/40G/100G イーサネットアダプタ	プラットフォームに応じて、ノードごとに 1 つ。 iWARP と iSCSI の両方のレプリケーションにアダプタを使用
クラスタ間	スイッチ使用
シェルフ	リモートクラスタからは確認不可
NVRAM レプリケーション	IP/iWARP
SyncMirror レプリケーション	IP/iSCSI
構成レプリケーションサービス	変更なし
MetroCluster のサイズ	4 ノードおよび 8 ノード
MetroCluster 拡張構成	なし
Advanced Drive Partitioning	あり (ETRENUS AX series のみ)

2.3 ローカルフェイルオーバー (HA) およびリモートスイッチオーバー (DR)

2 ノードのアーキテクチャでは、MetroCluster のスイッチオーバー機能とスイッチバック機能を使用して、HA フェイルオーバーとリモート DR の両方を実行します。各ノードは、ピア関係における HA パートナーおよび DR パートナーとして機能します。NVRAM は、4 ノード構成のようにリモートパートナーに複製されます。

4 ノードアーキテクチャと 8 ノードアーキテクチャは、ローカル HA フェイルオーバーとリモート DR スwitchオーバーの両方を提供します。図 2.1 に示すとおり、各ノードには同一ローカルクラスタ上に HA パートナーを持ち、リモートクラスタ上に DR パートナーを持ちます。A1 と A2 は HA パートナーであり、B1 と B2 も HA パートナーです。ノード A1 と B1 は DR パートナーであり、A2 と B2 も DR パートナーです。[「2.4.2 NVRAM レプリケーション」\(P.14\)](#) で説明しているとおり、NVRAM は HA パートナーと DR パートナーの両方にレプリケートされます。MetroCluster を設定すると、ノードの DR パートナーが自動的に選択され、パートナーはシステム ID (NVRAM ID) の順序に従って選択されます。

システム ID はハードコードされており、変更できません。ローカルピアとリモートピアの間に適切なパートナーシップを作成できるように、クラスタを設定する前にシステム ID を記録しておいてください。

図 2.1 HA および DR グループ

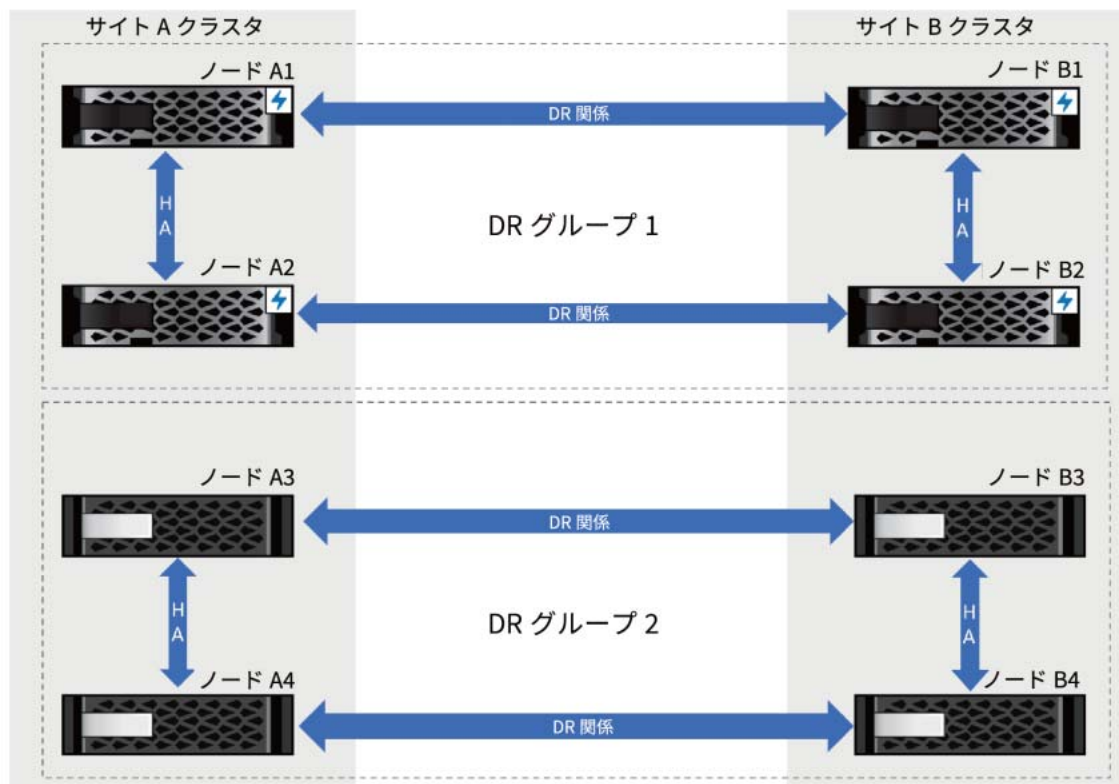


2. アーキテクチャ

2.3 ローカルフェイルオーバー (HA) およびリモートスイッチオーバー (DR)

図 2.2 は、8 ノードの MetroCluster 構成と DR グループの関係を示しています。8 ノード構成では、2 つの独立した DR グループがあります。DR グループ内のハードウェアは、サイト A とサイト B で同じである必要があります。ただし、DR グループ 1 のハードウェアと、DR グループ 2 のハードウェアを一致させる必要はありません。たとえば、DR グループごとに異なるハードウェアを使用することも可能です。

図 2.2 8 ノードの DR グループ



ローカル HA フェイルオーバーでは、HA ペアのノードの 1 つが、その HA パートナーの共有ストレージとサービスを一時的に引き継ぎます。たとえば、ノード A2 はノード A1 のリソースを引き継ぎます。テイクオーバーは、2 つのノード間のミラー NVRAM およびマルチパスストレージによって有効になります。たとえば、フェイルオーバーは無停止の ONTAP アップグレードを実行するために計画的に実施することも、パニックまたはハードウェア障害時に計画外に実施することもできます。ギブバックはフェイルオーバーの逆のプロセスです。障害が発生したノードは、テイクオーバーしたノードからリソースを取り戻します。ギブバックは常に計画された操作です。フェイルオーバーは常にローカル HA パートナーに対して行われ、どちらかのノードからも対となるノードにフェイルオーバーできます。

スイッチオーバー中は、自身のワークロードを実行しながら、ピアクラスタが他のクラスタのストレージおよびサービスを引き継ぎます。たとえば、サイト A がサイト B にスイッチオーバーすると、クラスタ B のノードは、クラスタ A が以前所有していたストレージとサービスの制御を一時的に引き継ぎます。スイッチオーバーが完了すると、クラスタ A の SVM はオンラインに戻り、クラスタ B で実行を継続できます。

スイッチオーバーは、たとえば、テストまたはサイトメンテナンスを実行するためにネゴシエート (計画) することも、サイトの 1 つを破壊する災害が発生した場合に強制 (計画外) することもできます。スイッチバックとは、正常に動作しているクラスタがスイッチオーバーされたリソースを取り戻して、安定した動作状態を復元するプロセスです。スイッチバックは 2 つのクラスタ間で調整され、常に計画された操作です。どちらのサイトも他のサイトに切り替えることができます。

また、サイトがスイッチオーバー中に後続の障害が発生するおそれがあります。たとえば、クラスタ B へのスイッチオーバー後に、ノード B1 で障害が発生したとします。B2 は自動的にすべてのワークロードを引き継ぎ、サービスを提供します。

2.4 MetroCluster レプリケーション

MetroCluster IP は直接接続したストレージを利用するため、ドライブをストレージファブリックに接続するために外部シリアル接続 SCSI (SAS) ブリッジを使用する必要がありません。災害復旧グループ内の各ノードは、グループ内の他のノードにドライブをエクスポートするストレージプロキシまたは iSCSI ターゲットとして機能します。iSCSI (SCSI over TCP/IP) は、IP ファブリック用のストレージ転送プロトコルです。iSCSI を使用すると、iSCSI イニシエータとターゲットが TCP/IP ファブリック経由で通信できます。災害復旧グループの各ノードは、リモートの災害復旧パートナーである iSCSI ターゲットとの iSCSI セッションを構成する iSCSI イニシエータを経由して、リモートストレージにアクセスします。

iSCSI および直接接続ストレージを使用すると、内蔵ドライブを備えたシステムを使用することもできます。iSCSI を使うと、そのノードの災害復旧パートナーノードが外部のドライブシェルフにあるストレージデバイスだけでなく内部ストレージにもアクセスできるようになります。

MetroCluster には、以下の 3 つのレプリケーションがあります。

- (1) 構成レプリケーション
- (2) NVRAM レプリケーション
- (3) ストレージレプリケーション

2.4.1 構成レプリケーション

MetroCluster は 2 つの ONTAP クラスタで構成され、各 ONTAP クラスタには、メタデータまたは構成情報を含む、クラスタごとのレプリケート済みデータベース (RDB) があります。スイッチオーバーが発生すると、停止したクラスタのメタデータオブジェクトが残りのクラスタでアクティブ化され、これらのオブジェクトを所有クラスタから他のクラスタに転送する必要があります。転送メカニズムには、クラスタピアリング、構成レプリケーションサービス (CRS)、およびメタデータを含むボリュームの 3 つのコンポーネントがあります。

- **クラスタピアリング**は、クラスタ間論理インタフェース (LIF) を使用して、2 つの ONTAP クラスタ間にお客様が用意した TCP/IP 接続を確立する方法です。これにより、クラスタ間で構成オブジェクトのレプリケーションが可能になり、MetroCluster および ONTAP SnapMirror ソフトウェアで使用されます。クラスタピアリングネットワークは、通常、フロントエンドまたはホスト側のネットワークであり、ストレージ仮想マシン (SVM)、LIF、ボリューム、アグリゲート、LUN などのオブジェクトを転送します。MetroCluster のピアリングネットワークは、通常の ONTAP クラスタと同じで、ホストがストレージにアクセスするために使用するフロントエンドネットワークと同じにすることもできます。レプリケーションは、お客様が用意したクラスタ間 LIF を備えた IP ネットワークであるピアリングネットワークを介して実行されます。
- **構成レプリケーションサービス (CRS)** は、MetroCluster 構成のコンポーネントであり、各クラスタで実行され、必要なメタデータオブジェクトを所有クラスタからピアクラスタのレプリケートされたデータベース (RDB) にレプリケートします。このサービスは、ピアリングネットワークを使用して、クラスタ間で構成オブジェクト (例: SVM、LIF、ボリューム、アグリゲート、LUN) とプロトコルオブジェクト (例: CIFS、NFS、SAN) をレプリケートします。クラスタピアリングネットワークで CRS に影響する中断が発生した場合、接続が再確立された後にレプリケーションが自動的に追いつきます。CRS は、メタデータボリュームと呼ばれるメタデータを格納するために、データアグリゲート上に小さなボリュームを必要とします。
- **メタデータを含むボリューム**は、MetroCluster 構成のクラスタメタデータ情報に使用されるステージングボリュームです。MetroCluster を構成すると、各クラスタに 10GB のボリュームが 2 つ作成されます。これらのボリュームは個別の非ルートアグリゲートに作成する必要があるため、MetroCluster を構成する前に、各クラスタに少なくとも 2 つのデータアグリゲートを作成することを推奨します。メタデータを含むボリュームは復元性を提供し、オブジェクトが作成または更新されるたびに更新が記録されます。ロギングが完了するまで、変更はローカル RDB にコミットされません。更新は、構成レプリケーションネットワークを介して他のクラスタの RDB に同期的に伝達されます。構成レプリケーションネットワークで一時的なエラーが発生したために変更を伝達できない場合、接続が復元された後に自動的に他のクラスタへ送信されます。

一方のクラスタの設定を変更すると、もう一方のクラスタに自動的に伝播されるため、データまたは構成情報が失われることなくスイッチオーバーが実現されます。更新は自動的に行われるため、MetroCluster 構成に固有の継続的な管理はほとんど必要ありません。構成レプリケーションネットワークで一時的なエラーが発生したために変更を伝達できない場合、接続が復元された後に自動的に他のクラスタへ送信されます。復元性を高めるために、クラスタ構成ネットワークには冗長ネットワークを使用することを推奨します。

以下の例では、MDV にシステムによって割り当てられた名前が割り当てられ、各クラスタから確認できるようになっています。最初の 2 つのボリュームはコマンドがクラスタ A から実行されたため、「オンライン」状態になったローカル MDV です。残る 2 つの MDV は、ホスティングアグリゲートで示す通りクラスタ B に属し、スイッチオーバーが実行されていないため現在は「オフライン」状態です。

tme-mcc-A::> volume show -volume MDV*							
Vserver	Volume	Aggregate	State	Type	Size	Available	Used%
tme-mcc-A	MDV_CRS_cd7628c7f1cc11e3840800a0985522b8_A	aggr1_tme_A1	online	RW	10GB	9.50GB	5%
tme-mcc-A	MDV_CRS_cd7628c7f1cc11e3840800a0985522b8_B	aggr1_tme_A2	online	RW	10GB	9.50GB	5%
tme-mcc-A	MDV_CRS_e8fef00df27311e387ad00a0985466e6_A	aggr1_tme_B1	-	RW	-	-	-
tme-mcc-A	MDV_CRS_e8fef00df27311e387ad00a0985466e6_B	aggr1_tme_B2	-	RW	-	-	-

2.4.2 NVRAM レプリケーション

NVRAM レプリケーションは、フェイルオーバーまたはスイッチオーバー時にデータを消失から保護するために、ローカルノードの NVRAM をリモートの災害復旧ノードにある NVRAM にコピーするプロセスです。ONTAP HA ペアでは、各ノードは HA インターコネクトを介して NVRAM を他のノードにミラーリングします。NVRAM は、各ノードの NVRAM ごとに 1 つずつ、合計で 2 つのセグメントに分割されます。[図 2.3](#) は、MetroCluster が一方のサイトに DR パートナーを持つことで追加のミラーリングを提供し、NVRAM が相互スイッチリンク (ISL) 接続を介して DR パートナーにミラーリングされる例を示します。4 ノード構成では、各ノードの NVRAM は HA パートナーと DR パートナーにそれぞれ 1 回ずつ (合計 2 回) ミラーリングされるため、各ノードの NVRAM は 4 つのセグメントに分割されます。

図 2.3 NVRAM の割り当て



書き込み操作は、ドライブに書き込まれる前にまず不揮発性メモリ (NVRAM) にステージングされ、すべての NVRAM セグメントが更新されると、発行元のホストまたはアプリケーションに確認応答が送信されます。各ストレージコントローラーの NVRAM は、ローカルの HA (高可用性) パートナーにローカルでミラーリングされ、パートナーサイトの災害復旧 (DR) パートナーにリモートでミラーリングされます。4 ノード構成では、不揮発性キャッシュは、ローカル、HA パートナー、DR パートナー、および DR 補助パートナーの 4 つのパーティションに分割されます。ローカル HA テイクオーバーの場合、DR 補助パートナーに自動的に切り替えることで、DR ミラーリングを続行できます。ギブバックが正常に完了すると、ミラーリングは自動的に DR パートナーに戻ります。たとえば、ノード B1 に障害が発生してノード B2 にテイクオーバーされた場合、ノード A1 のローカルキャッシュはノード B1 にミラーリングできないため、ミラーリングは DR の補助パートナーであるノード B2 に切り替わります。

DR パートナーの NVRAM に対する更新は、MetroCluster IP の iWARP プロトコルを使用して ISL 経由で送信されます。MetroCluster IP の場合、iWARP は、IP スタックの影響によるレイテンシを最小限に抑えるために RDMA 対応のネットワークアダプタを使用してハードウェアにオフロードされます。iWARP トラフィックをストレージレプリケーションよりも優先させるために、スイッチのサービス品質 (QoS) を使用します。

ただし、ISL レイテンシが増加すると、DR パートナーの NVRAM への書き込みの確認に時間がかかるため、書き込みパフォーマンスが影響を受け場合があります。一時的なサイト分離 (たとえば、すべての ISL がダウンし、リモートノードが応答しない) の場合に継続的なローカル操作を可能にするために、書き込みはシステムタイムアウト後に確認応答されます。リモート NVRAM ミラーリングは、少なくとも 1 つの ISL が使用可能になると自動的に再同期化されます。

データ消失を防ぐために、NVRAM トランザクションは、少なくとも 10 秒ごとに一度、整合性ポイントを介してドライブにコミットされます。コントローラーの起動時に、WAFL はドライブ上の最新の整合性ポイントを使用するため、停電やシステム障害後に長時間のファイルシステムチェックを行う必要がなくなります。ストレージシステムでは、最新の整合性ポイント以降に発生した可能性のあるデータ I/O 要求が失われないように、バッテリーバックアップされた NVRAM が使用されます。テイクオーバーまたはスイッチオーバーが発生すると、コミットされていないトランザクションがミラーリングされた NVRAM から再生され、データの損失が防止できます。

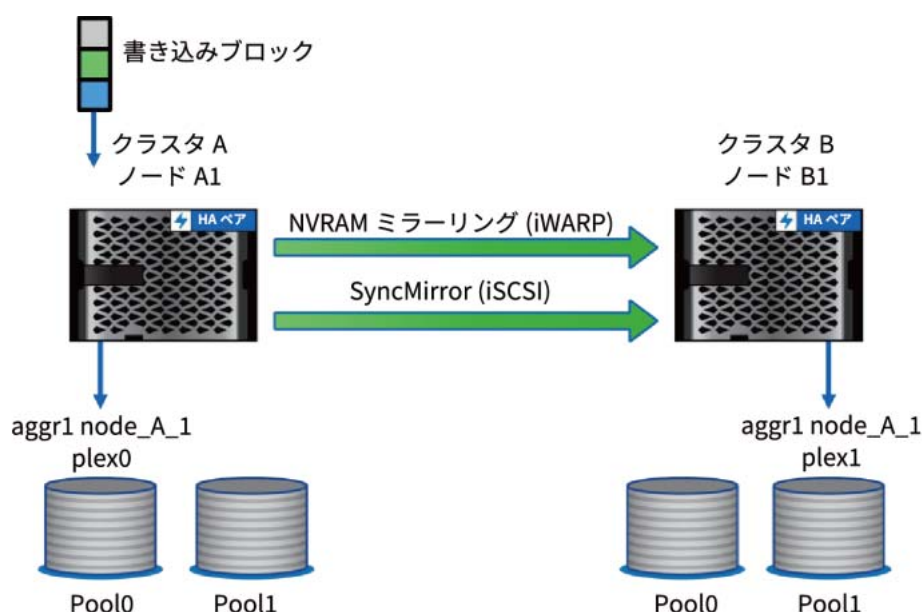
2.4.3 ストレージレプリケーション

ストレージレプリケーションは、RAID SyncMirror (RSM) を使用してローカルとリモートのバックエンドドライブをミラーリングします。MetroCluster IP では、災害復旧グループ内の各ノードをリモート iSCSI ターゲットとして機能させることによって、バックエンドストレージが論理的に共有されます。ノードがリモートのバックエンドドライブにアクセスするには、リモートの災害復旧パートナーノードを経由して、iSCSI ターゲットを介して処理されるリモートドライブにアクセスします。

図 2.4 に、NVRAM およびストレージの MetroCluster IP のレプリケーションを示します。ノード B1 は、ローカルに接続されたドライブを、iSCSI ターゲットを介して災害復旧グループ内のリモートパートナーノードにエクスポートします。

ノード A1 の pool0 ドライブはノード A1 にローカルで接続されていますが、pool1 リモートドライブはノード B1 がホストする iSCSI ターゲットを介してエクスポートされます。アグリゲートの aggr1 node_A_1 のローカル plex 0 は、pool0 からローカルで接続されたドライブで構成されています。アグリゲートの aggr1 node_A_1 のリモート plex 1 は、B1 に直接接続されたドライブから構成され、B1 にホストされた iSCSI ターゲットを経由して A1 にエクスポートされます。

図 2.4 書き込みデータブロックのミラーリング



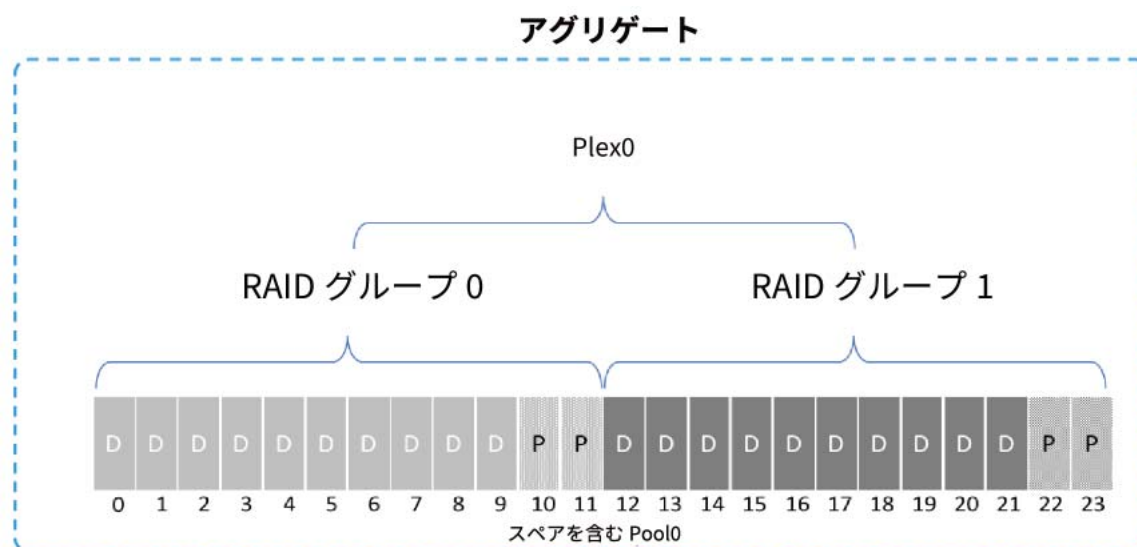
ブロックは、NVRAM (または NVMEM) と SyncMirror の両方を使用して、各サイトのペアになったノードに書き込まれます。SyncMirror は、ミラー化されたアグリゲートごとに、ローカルプレックスとリモートプレックスの 2 つにデータを書き込みます。SyncMirror の書き込みは RAID レイヤーで行われます。つまり、重複除外や圧縮などのストレージ効率化により、SyncMirror の操作で書き込まれるデータが削減されます。

ブロックはローカルストレージから読み取られ、読み取り操作のパフォーマンスや ISL の使用には影響しません。

2.4.4 SyncMirror ストレージレプリケーション

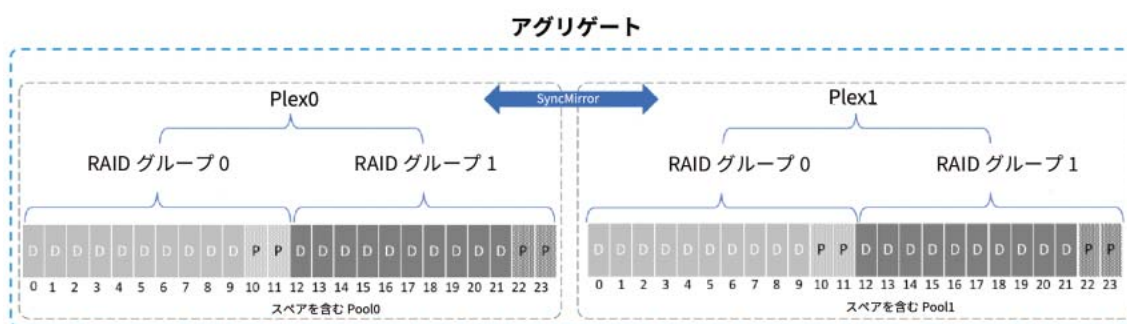
ONTAP システムは、アグリゲートからプロビジョニングされる FlexVol ボリュームにデータを保存します。各アグリゲートには WAFL ファイルシステムが含まれます。MetroCluster を使用しない構成では、各アグリゲートのドライブは、プレックスと呼ばれる単一または複数の RAID グループで構成されます (図 2.5)。プレックスは、コントローラーに接続されたローカルストレージに存在します。

図 2.5 ミラーリングされていないアグリゲート Plex0



MetroCluster 構成では、各アグリゲートは物理的に分離された 2 つのプレックス (ローカルプレックスとリモートプレックス) で構成されます (図 2.6)。すべてのストレージは共有され、MetroCluster 構成のすべてのコントローラーから認識できます。ローカルプレックスにはローカルプール (pool0) のドライブだけが含まれ、リモートプレックスにはリモートプールのドライブだけが含まれている必要があります。ローカルプレックスは、常に plex0 です。各リモートプレックスには、リモートであることを示す 0 以外の番号が割り当てられます (例えば、plex1 または plex2)。

図 2.6 MetroCluster でミラーリングされたアグリゲート



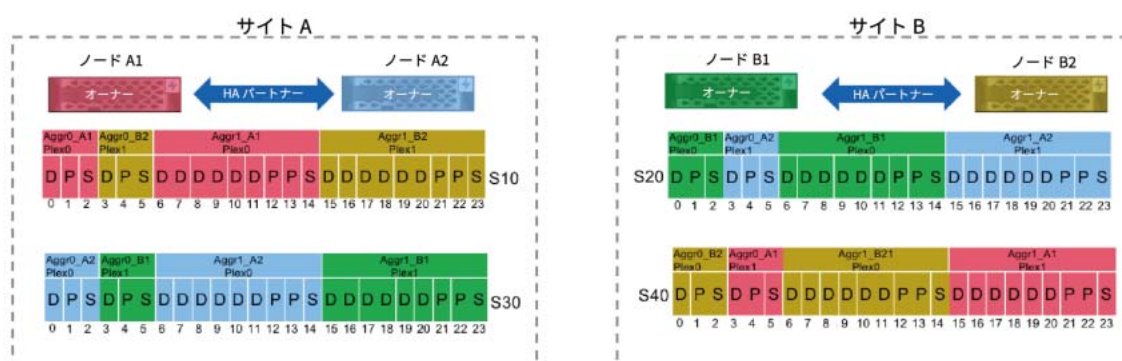
2. アーキテクチャ

2.5 アグリゲートのスナップショットコピー

MetroCluster では、ミラーリングされたアグリゲートとミラーリングされていないアグリゲートの両方がサポートされています。ミラーリングされていないアグリゲートは、ONTAP 9.8 以降の MetroCluster IP でサポートされます。MetroCluster を構成した後にアグリゲートを作成した場合は、`-mirror true` フラグを指定する必要があります。指定されていない場合、`create` コマンドが失敗します。`-diskcount` パラメータで指定されたドライブの数は、自動的に半分にになります。たとえば、使用可能なドライブが 6 つあるアグリゲートを作成するには、ドライブ数として 12 を指定する必要があります。これにより、ローカルプレックスにはローカルプールから 6 つのドライブが割り当てられ、リモートプレックスにはリモートプールから 6 つのドライブが割り当てられます。アグリゲートにドライブを追加する場合も、同じプロセスが適用されます。容量に必要なドライブ数の 2 倍を指定する必要があります。

図 2.7 の例は、ドライブがどのようにアグリゲートに割り当てられたかを示しています。各ノードには、ルートアグリゲートと 1 つのデータアグリゲートがあります。各ルートアグリゲートには、ノードごとに 6 台のドライブが含まれます。クラスタごとに最小 2 つのシェルフが使用され、6 台のうち 3 台がローカルプレックス上にあり、3 台がリモートプレックス上にあると仮定します。したがって、アグリゲートで使用可能な容量は 3 台のドライブです。同様に、各データアグリゲートには、ローカルドライブ 9 台とリモートドライブ 9 台の合計 18 台のドライブが含まれます。また、ETRENUS AX series で MetroCluster を使用する場合は、ルートアグリゲートは RAID4 を使用し、データアグリゲートは RAID DP または RAID-TEC を使用します。

図 2.7 ルートアグリゲートおよびデータアグリゲート



通常の MetroCluster 操作では、両方のプレックスが RAID レベルで同時に更新されます。クライアントとホストの I/O、またはクラスタメタデータのいずれからの書き込みでも、2 つのクラスタ間の ISL 接続を使用して、ローカルプレックスとリモートプレックスの 2 つの物理的な書き込み操作が生成されます。デフォルトでは、読み取りはローカルプレックスから実行されます。

2.5 アグリゲートのスナップショットコピー

自動的なアグリゲートのスナップショットコピーが行われると、デフォルトでアグリゲート容量の 5% がスナップショットコピーのためにリザーブされます。これらのスナップショットコピーは、必要に応じてアグリゲートを再同期するためのベースラインとして使用されます。

1 つのプレックスが使用できなくなった場合 (たとえば、シェルフやストレージアレイの障害など)、障害が発生したプレックスがリストアされるまで、影響を受けていないプレックスがデータを提供し続けます。障害が発生したプレックスが修復されると、プレックスは自動的に再同期され、両方のプレックスが整合します。再同期のタイプは自動的に決定され、実行されます。両方のプレックスが共通のアグリゲートのスナップショットコピーを共有する場合、このスナップショットコピーが部分的な再同期のベースとして使用されます。共通のスナップショットコピーがプレックス間で共有されていない場合は、完全な再同期が実行されます。

2.6 アクティブ／アクティブおよびアクティブ／パッシブ構成

MetroCluster では、対称的なスイッチオーバーおよびスイッチバックが自動的に有効になります。つまり、いずれかのサイトで災害が発生した場合に、そのサイトから別のサイトに切り替えることができます。したがって、両方のサイトがアクティブに独立したワークロードを処理するアクティブ／アクティブ構成は、製品の基盤となるものです。

もう 1 つの構成は、アクティブ／スタンバイまたはアクティブ／パッシブです。1 つのクラスタ (例えば、クラスタ A) のみが定常状態でアプリケーションワークロードをホストします。したがって、サイト A からサイト B への一方向のスイッチオーバーだけが必要です。クラスタ B の各ノードには、それぞれミラーリングされたルートアグリゲートとメタデータボリュームが必要です。後で要件が変更され、ワークロードがクラスタ B でプロビジョニングされる場合、アクティブ／パッシブからアクティブ／アクティブへの変更のために MetroCluster 構成を変更する必要はありません。いずれかのサイトで定義されたワークロード (SVM) は、他のサイトで自動的に複製および保護されます。

サポートされるもう 1 つのオプションは、HA ペアのアクティブ／パッシブで、2 つのノードのうち 1 つだけがワークロードをホストします。このオプションでは、クラスタごとに 1 つのデータアグリゲートのみが必要な小規模な構成が作成されます。

MetroCluster は、スイッチオーバー時にストレージアクセスパスの ID を保持します。スイッチオーバー後も LIF アドレスは維持され、NFS エクスポートと SMB 共有には同じ IP アドレスを使用してアクセスします。また、LUN は同じ LUN ID、ワールドワイドポート名 (WWPN)、または IP アドレスとターゲットポータルグループタグを持っています。このように ID が保持されるため、フロントエンドネットワークは両方のサイトにまたがる必要があり、フロントエンドクライアントとホストがパスと接続を認識できるようになります。ホストネットワークのための IP アドレス / サービスモビリティ要件を達成するために、MetroCluster は Layer2 (共有 VLAN) ネットワーキングと Layer3 (VIP/BGP) ネットワーキングの両方をサポートします。

2.7 Advanced Drive Partitioning (ADP)

備考

MetroCluster では、ADP は MetroCluster IP 構成の ETERNUS AX series システムでのみ使用できます。

ADP は、ETERNUSAX/HX series システムの HDD と SSD のストレージ効率を高める機能です。アグリゲートと HA ペア内のコントローラーとの間で物理ドライブの容量共有を可能にすることにより、ADP は、より少ない SSD で両方のノードの容量を拡張し、容量効率とコストを改善します。ADP を使用すると、ルートアグリゲートによって消費される容量が少なくなるため、データアグリゲートのプロビジョニングに使用できる容量が増えます。さらに、両方のコントローラー間でパリティドライブとスペアドライブを共有すると、HA ペア内で使用できる容量がさらに増加します。ADP は、パーティション化されていないドライブ全体を使用する場合に比べて、パーティション化したドライブからプロビジョニングすることでより効率的にストレージ容量を使用する方法です。

Advanced ADP の利点

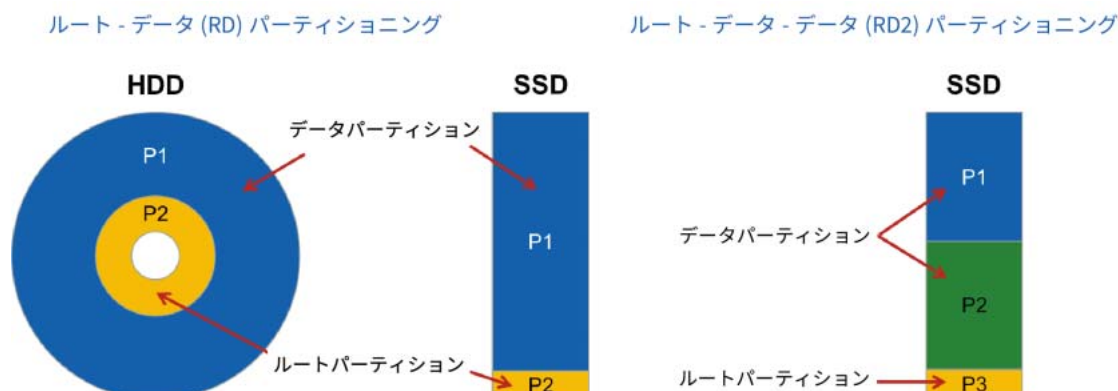
- ETERNUS AX/HX series システムの使用可能容量を拡大します。
- ストレージ効率が向上します (ドライブ全体のパーティション化と比較して、効率が 10 ~ 40% 向上します)。
- より少ない SSD で両方のノードに容量を拡張できます。
- コストを改善し、ストレージの競争力を向上させます。

パーティション化の方法

- ルート - データ (RD) パーティショニングドライブを 1 つのルートパーティションと 1 つのデータパーティションに分割します。

- ルート - データ - データ (RD2) パーティショニングドライブを 1 つのルートパーティションと 2 つのデータパーティションに分割します。

図 2.8 ADP の論理的側面



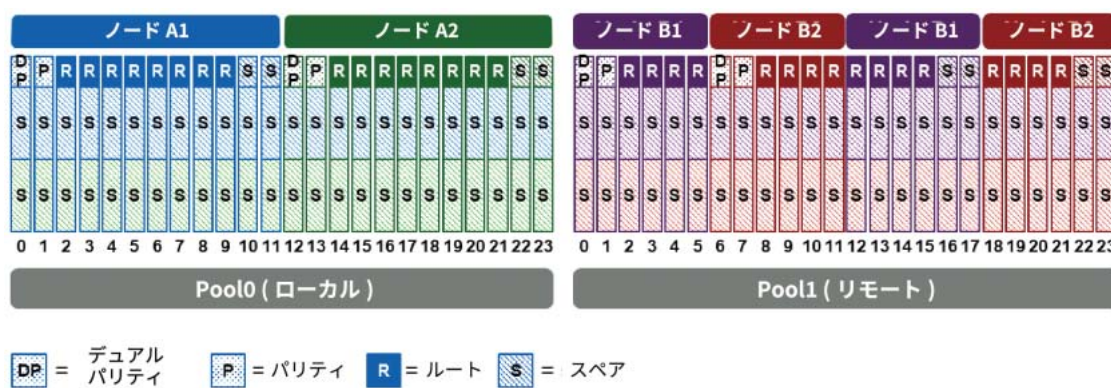
備考

ONTAP 9.7 以降を搭載した ETERNUS AX series システムの MetroCluster IP 構成では、ADP RD2 パーティショニングのみがサポートされます。ADP は、MetroCluster の初期化時にデフォルトで適用されます。

2.7.1 ルート - データ - データ (RD2) パーティショニング

ルート - データ - データ (RD2) パーティショニングは、ONTAP 9.0 以降のリリースで使用可能なストレージ効率化機能です。RD2 は、複数の SSD からパーティションを使用してルートアグリゲートを効率的にプロビジョニングするため、データアグリゲートで使用可能な容量が増えます。各 SSD は、小さい (シン) ルートパーティションと 2 つの大きな (シック) データパーティションの 3 つのパーティションに分割されます。SSD ごとに 2 つのデータパーティションがあるため、ETERNUS AX series または SSD のみで構成された ETERNUS HX series システムでは、1 台のドライブの容量と IOPS を両方のコントローラーで使用できます。RD2 でパーティションを分割できる SSD の最大数は 48 ですが、48 を超える SSD は HA ペアで RD2 パーティショニングを使用できます。RD2 パーティショニングを使用するために必要な SSD の最小数は 8 で、400GB 以上の SSD が RD2 パーティショニングをサポートしています。RD2 パーティショニングを使用するシステムのルートパーティションサイズは、コントローラーモデル、ONTAP リリース、およびシステムの初期設定中に接続された SSD の数によって異なります。スペアルートパーティションは、追加の領域が必要な場合にのみルートアグリゲートを拡張するために使用されますが、スペアデータパーティションは、障害が発生したデータパーティションを置き換えるホットスペアとして使用できます。[図 2.9](#) は、MetroCluster IP 構成の ETERNUS AX2200 の例を示します。48 台の NVMe SSD を搭載した NS224 が 1 台搭載された ETERNUS AX2200 を各サイトに 1 台ずつ構成しています。

図 2.9 48 ドライブの MetroCluster IP 構成の ADP の例

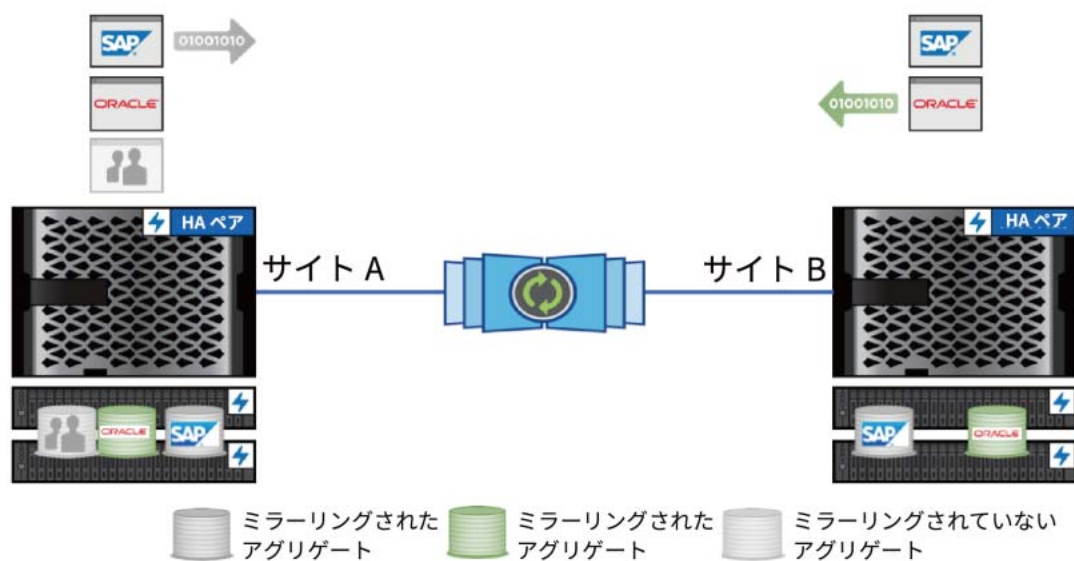


2.8 ミラーリングされていないアグリゲート

拡張およびファブリック接続構成の場合は ONTAP 9 以降で、IP 構成の場合は ONTAP 9.8 以降で、MetroCluster 構成で提供している冗長化したミラーリングを必要としないデータのための、ミラーリングされていないアグリゲートがサポートされました。ミラーリングされていないアグリゲートは、サイトが被災した場合に保護されません。これらのアグリゲートへの書き込み I/O は、ISL のサイジング時にカウントする必要があります。

図 2.10 は、ミラーリングアグリゲートの詳細な制御を示しています。SAP はサイト B クラスタにミラーリングされ、Oracle はサイト A クラスタにミラーリングされています。サイト A のホームユーザーディレクトリは重要なアグリゲートではなく、リモートクラスタにはミラーリングされません。サイト A で障害が発生した場合、このアグリゲートは使用できません。

図 2.10 MetroCluster のミラーリングされていないアグリゲート



備考

ADP で構成されたドライブを使用する場合、ノードによるパーティション所有権およびプール内のメンバーシップに関する固有ルールを理解することが重要です。また、ミラー化するドライブは、ピアリング関係の両側で対称である必要があります。ADP で構成されたパーティションをミラーリングされていないアグリゲートに使用すると、予期しない障害が発生する可能性があります。このような問題を回避するには、ミラーリングされていないアグリゲートをパーティション化されていないドライブに配置することを強くお勧めします。

3. 展開オプション

MetroCluster は、各サイトで同一のハードウェアが必要な、完全冗長構成です。さらに、MetroCluster は、拡張構成、ファブリック接続構成、および IP 構成に柔軟に対応します。[表 3.1](#) に、さまざまな展開オプションの概要と、サポートされるスイッチオーバー機能を示します。

表 3.1 ハードウェア要件

特徴	IP 構成	ファブリック接続構成
		4 ノードまたは 8 ノード
コントローラー数	4 または 8	4 または 8
FC スイッチストレージファブリック	なし	あり
IP スイッチストレージファブリック	あり	なし
FC-SAS 間ブリッジ	なし	あり
直接接続ストレージ	あり (ローカル接続のみ)	なし
ローカル HA のサポート	あり	あり
自動スイッチオーバーのサポート	あり (Mediator あり)	あり
ミラーリングされていないアグリゲートのサポート	あり	あり
アレイ LUN のサポート	なし	あり

3.1 拡張構成および拡張ブリッジ構成

MetroCluster の拡張構成では、2 つのストレージノードの距離を通常 270 m まで拡張し、高いレベルの復元性と災害復旧機能を提供します。これは、高速リンクを使用して 2 つのクラスタを接続し、それらの間でデータを同期ミラーリングすることによって実現されます。一方のクラスタに障害が発生したり、使用できなくなった場合、もう一方のクラスタがシームレスに引き継ぎ、アプリケーションとユーザーに継続的なデータアクセスを提供します。MetroCluster の拡張ブリッジ構成では、2 つのプライマリクラスタ間に FC-SAS 間ブリッジを追加することで、拡張構成がさらに最大 500 m まで拡張されます。拡張ブリッジ構成を使用すると、より複雑なアーキテクチャをサポートしたり、データセンター間の距離を広げることができます。これらの構成はいずれも、データセンターの導入に最適であり、インフラストラクチャの要件 (例: ケーブル配線、FC スイッチ、ラックスペース) が縮小されています。

3.2 IP 構成

MetroCluster IP は、IP ネットワークを使用して、最大 700 km までの 2 つのサイト間で同期レプリケーションを実行します。ETERNUS AX/HX series のストレージでサポートされる MetroCluster IP は、4 ノードおよび 8 ノードのアーキテクチャに導入できます。

備考

MetroCluster の最新のサポート条件 (ハードウェア、ソフトウェア、サイジング、および制限事項) に関しては、Fusion を参照してください。

4. 計画的なイベントと計画外イベントの復元性

本章では、障害と災害の種類、および MetroCluster 構成がどのように可用性およびデータを維持し、障害から復旧させるかについて説明しています。

4.1 シングルノード障害

ローカル HA ペアの 1 つのコンポーネントで障害が発生した場合を想定しています。4 ノード構成の MetroCluster でこの障害が起きた場合、故障したコンポーネントの種類によっては、機能していないノードのストレージリソースに自動テイクオーバーもしくはネゴシエートテイクオーバーが実行される場合があります。データ復旧については、「[HA ペアの管理 - 概要](#)」を参照してください。

4.2 サイト全体のコントローラー障害

電源の喪失、機器の交換、または災害によって、サイトのすべてのコントローラーモジュールに障害が発生した場合を想定しています。通常、MetroCluster 構成では障害と災害を区別できません。ただし、MetroCluster Tiebreaker ソフトウェアなどの監視ソフトウェアを使用すれば、障害と災害を区別できます。ISL およびスイッチが機能していて、ストレージにアクセスできる場合、サイト全体のコントローラー障害によって自動スイッチオーバーが発生することがあります。

「[HA ペアの管理 - 概要](#)」に、コントローラー自体の故障がないにも関わらず発生するサイト全体のコントローラー障害からの復旧、および 1 つ以上のコントローラーの故障を含む障害から回復する方法についての詳細を示します。

4.3 ISL 障害

サイト間のリンクに障害が発生した場合を想定しています。この場合、MetroCluster 構成は何も行いません。各ノードは引き続きデータを正常に処理しますが、DR サイトへアクセスできないため、ミラーリングしたデータはそれぞれの DR サイトに書き込まれません。

4.4 複数の連続障害

複数のコンポーネントに連続して障害が発生した場合を想定しています。たとえば、コントローラーモジュール、スイッチファブリック、シェルフの順に障害が発生すると、ストレージのフェイルオーバー、ファブリックの冗長性、SyncMirror が発生し、ダウンタイムとデータ消失を防止します。

[表 4.1](#) に、障害のタイプと対応する DR メカニズム、および復旧方法を示します。

AUSO は、MetroCluster IP 構成で ONTAP Mediator および ONTAP 9.7 以降を使用している場合にのみサポートされます。

表 4.1 障害のタイプと復旧方法

障害タイプ	DR メカニズム	復旧方法の概要
シングルノード障害	ローカル HA 障害	自動フェイルオーバーおよびギブバックが有効の場合は対処不要です。
サイト障害	MetroCluster スイッチオーバー	ノードが復元されたら、metrocluster healing および metrocluster switchback コマンドを使用して、手動修復とスイッチバックを実行する必要があります。
サイト全体のコントローラー障害	AUSO 被災サイトのストレージにアクセスできる場合のみ可能です。	ノードが復元されたら、metrocluster healing および metrocluster switchback コマンドを使用して、手動修復とスイッチバックを実行する必要があります。
ISL 障害	MetroCluster スイッチオーバーは実行されません。2 つのクラスタは独立してデータを処理します。	このタイプの障害に対処は必要ありません。接続を復元すると、ストレージは自動的に再同期されます。
複数の連続障害	ローカル HA フェイルオーバーの後に、metrocluster switchover-forced-ondisaster コマンドを使用した MetroCluster の強制スイッチオーバーを実行します。 障害が発生したコンポーネントによっては、強制スイッチオーバーが不要な場合があります。	ノードが復元されたら、metrocluster healing および metrocluster switchback コマンドを使用して、手動修復とスイッチバックを実行する必要があります。

4.5 4 ノードおよび 8 ノードの無停止運用

単一のノードに限定された障害の場合、ローカル HA ペア内のフェイルオーバーおよびギブバックによって、無停止運用を継続できます。この場合、MetroCluster 構成はリモートサイトへのスイッチオーバーを必要としません。

これらの MetroCluster 構成は各サイトの 1 つ以上の HA ペアで構成されており、各サイトでローカル障害に耐えることができるため、パートナーサイトにスイッチオーバーすることなく無停止運用を実行できます。HA ペアの動作は、MetroCluster 以外の構成における HA ペアと同様です。

パニックまたは電力喪失によるノード障害によって、自動スイッチオーバーが発生する場合があります。

ローカルフェイルオーバー後に 2 つ目の障害が発生した場合は、MetroCluster のスイッチオーバーによって無停止運用が継続されます。同様に、2 つめの障害発生時のスイッチオーバー動作が故障していないノードの 1 つで完了した後は、ローカルフェイルオーバーによって無停止運用が継続されます。この場合、正常に動作している 1 つのノードが、DR グループ内の他の 3 つのノードのデータを処理します。

4.6 スイッチオーバー後のローカルフェイルオーバーの影響

MetroCluster のスイッチオーバーが発生し、正常に動作しているサイトで問題が発生した場合は、ローカルフェイルオーバーによって無停止運用が継続されます。しかし、システムが冗長構成ではなくになっているため、危険にさらされています。

もしもスイッチオーバーの後にローカルフェイルオーバーが発生した場合、MetroCluster 構成のすべてのストレージシステムのデータを 1 つのコントローラーで処理するため、リソース関連の問題が発生するおそれがあります。障害が発生しなかったコントローラーには、さらに別の障害が発生した場合の脆弱性があります。

4.7 スイッチオーバープロセスの概要

MetroCluster のスイッチオーバー操作では、ストレージとクライアントのアクセスをソースクラスタからリモートサイトのクラスタに移動することで、災害発生後すぐにサービスを再開できます。予期される変更と、スイッチオーバーが発生した場合に実行する必要があるアクションを認識しておく必要があります。

スイッチオーバーの動作中、システムは以下の処理を実行します。

- 被災サイトに属するドライブの所有権が、DR パートナーに変更されます。これは HA ペアのローカルフェイルオーバーの場合と同様で、障害が発生したパートナーに属するドライブの所有権が正常に動作しているパートナーに変更されます。
- 被災クラスタのノードに属しているものの、障害が発生していないサイトにあるため動作が存続しているブレックスは、障害が発生していないサイトでオンラインになります。
- 被災サイトに属する同期ソース SVM は、ネゴシエートスイッチオーバー中にだけダウンします。
 - このアプローチは、ネゴシエートスイッチオーバーにのみ適用されます。
- 被災サイトに属する同期先 SVM が起動します。

スイッチオーバーされている間は、DR パートナーのルートアグリゲートはオンラインになりません。

`metrocluster switchover` コマンドは、MetroCluster 構成のすべての DR グループのノードをスイッチオーバーします。たとえば、8 ノードの MetroCluster 構成では、このコマンドは両方の DR グループのノードをスイッチオーバーします。

サービスをサイトにスイッチオーバーするだけの場合は、サイトを保護せずにネゴシエートスイッチオーバーを実行する必要があります。ストレージまたは機器の信頼性が低い場合は、被災サイトを保護してから、計画外スイッチオーバーを実行する必要があります。保護することにより、ドライブの電源が段階的に投入されたときの RAID の再構築が防止されます。

この手順は、他のサイトが安定しており、オフラインにする予定がない場合にのみ使用してください。

4.8 MetroCluster Tiebreaker

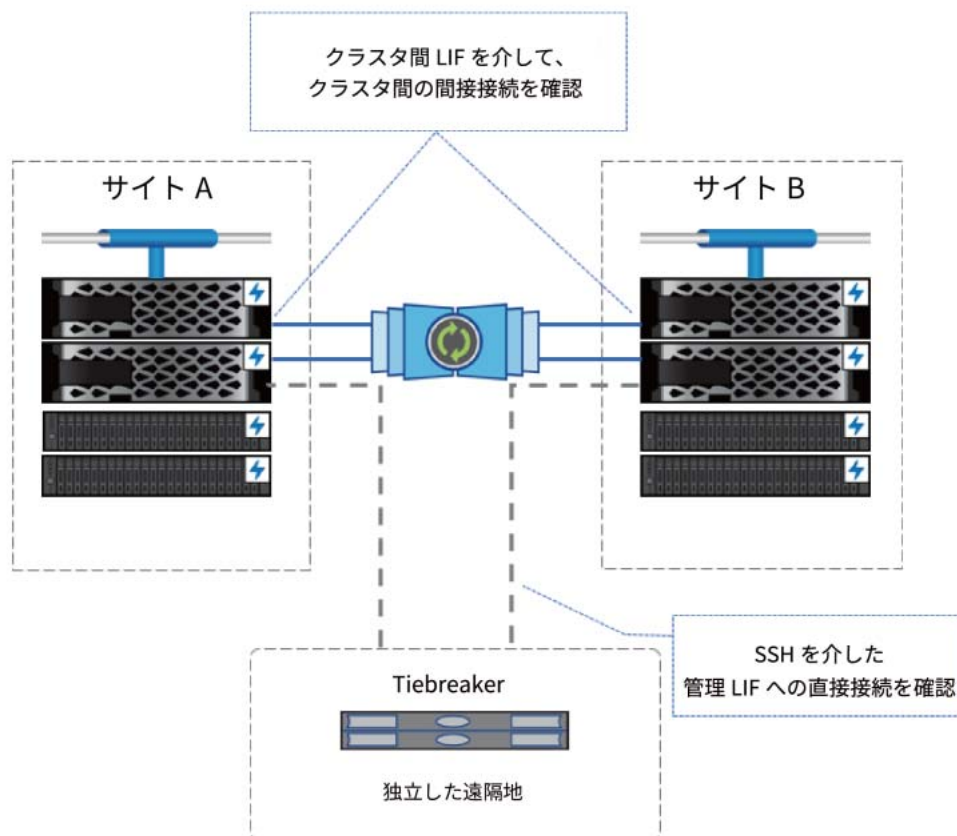
MetroCluster Tiebreaker ソフトウェアは、サイト間のすべての接続が失われた場合に警告を出します。MetroCluster Tiebreaker ソフトウェアは、ONTAP 9.7 以降でサポートされているすべての MetroCluster 構成をサポートします。

Tiebreaker ソフトウェアは Linux ホスト上にあります。別のサイトから 2 つのクラスタとそれらの間の接続状況を監視する場合にのみ、Tiebreaker ソフトウェアが必要です。これにより、クラスタ内の各パートナーは、サイト間リンクがダウンした場合に ISL 障害とサイト障害を区別することができます。

複数の Tiebreaker 監視が競合することを避けるため、MetroCluster 構成ごとに 1 つの MetroCluster Tiebreaker で監視を行ってください。

MetroCluster Tiebreaker ソフトウェアは、MetroCluster 構成のノードとクラスタに到達できるかどうかをチェックして、サイト障害が発生したかどうかを判断します。また、Tiebreaker ソフトウェアは、特定の条件下で警告を出します。図 4.1 に示すように、MetroCluster Tiebreaker は直接障害と間接障害を検出するため、ファブリックに障害がなければ、Tiebreaker はスイッチオーバーを開始しません。

図 4.1 MetroCluster Tiebreaker の確認



4.8.1 MetroCluster Tiebreaker による障害の検出

Tiebreaker ソフトウェアは Linux ホスト上にあります。Tiebreaker ソフトウェアが必要になるのは、2 つのクラスタとそれらの間の接続ステータスを別のサイトから監視する場合だけです。これにより、クラスタ内の各パートナーは、サイト間リンクがダウンした場合に ISL 障害とサイト障害を区別することができます。

Linux ホストに Tiebreaker ソフトウェアをインストールした後、MetroCluster 構成のクラスタを構成して、障害状態を監視できます。

4.8.2 サイト間接続障害の検出

MetroCluster Tiebreaker ソフトウェアは、サイト間のすべての接続が失われた場合に警告します。以下のタイプのネットワークパスが MetroCluster によって使用され、MetroCluster Tiebreaker によって監視されます。

- **クラスタ間ピアリングネットワーク**

このタイプのネットワークは、2 つのクラスタ間の冗長 IP ネットワークパスで構成されます。クラスタピアリングネットワークは、SVM 構成のミラーリングに必要な接続を提供します。1 つのクラスタ上のすべての SVM の構成は、パートナークラスタにミラーリングされます。

- **IP ネットワーク**

このタイプのネットワークは、2つの冗長 IP スイッチネットワークで構成されます。各ネットワークには2つの IP スイッチがあり、各スイッチファブリックの一方のスイッチがクラスタと同じ場所に配置されています。各クラスタには2つの IP スイッチがあり、各スイッチファブリックに1つずつあります。すべてのノードは、同じ場所にある各 FC スイッチに接続しています。データは ISL を介してクラスタ間でレプリケートされます。

4.8.3 サイト間接続の監視

Tiebreaker ソフトウェアは、定期的にノードからサイト間接続のステータスを取得します。NV 相互接続が失われ、クラスタ間ピアリングが ping に応答しない場合、クラスタはサイトが切り離されているものと見なし、Tiebreaker ソフトウェアが「AllLinksSevered」警告を出します。クラスタが「AllLinksSevered」ステータスを認識し、他方のクラスタにネットワーク経由で到達できない場合、Tiebreaker ソフトウェアが「disaster」警告を出します。

4.8.4 Tiebreaker によって監視されるコンポーネント

Tiebreaker ソフトウェアは、MetroCluster 構成の各コントローラーを監視します。監視は、IP ネットワーク上にホストされたノード管理 LIF とクラスタ管理 LIF への複数のパスを経由する冗長接続によって実施されます。

Tiebreaker ソフトウェアは、MetroCluster 構成の以下のコンポーネントを監視します。

- ローカルノードインターフェースを介したノード
- クラスタ指定のインターフェースを介したクラスタ
- 被災サイトへ接続できるかどうかを評価するための、障害が発生していないクラスタ (NV 相互接続、ストレージ、およびクラスタ間ピアリング)

Tiebreaker ソフトウェアとクラスタ内のすべてのノード、およびクラスタ自体との間の接続が失われると、クラスタは Tiebreaker ソフトウェアによって「到達不能」と判断されます。接続障害の検出には約 3～5 秒かかります。Tiebreaker ソフトウェアからクラスタに到達できない場合、障害が発生していないクラスタ (まだ到達可能なクラスター) は、Tiebreaker ソフトウェアが警告を出す前に、パートナークラスターへのすべてのリンクが切断されたことを示す必要があります。

障害が発生していないクラスタが、FC (NV 相互接続およびストレージ) およびクラスタ間ピアリングを介して被災サイトのクラスタと通信できなくなると、すべてのリンクが切断されます。

4.8.5 Tiebreaker 障害のシナリオ

Tiebreaker ソフトウェアは、被災サイトのクラスタ (すべてのノード) が停止しているか到達不能であり、障害の発生していないサイトのクラスタが「AllLinksSevered」ステータスを示している場合に警告を出します。

以下の場合では、Tiebreaker ソフトウェアは警告を出しません (または警告が拒否されます)。

- 8 ノードの MetroCluster 構成で、被災サイトの 1 つの HA ペアが停止している場合。
- 被災サイトのすべてのノードが停止しているクラスタで障害が発生していないサイトの 1 つの HA ペアが停止しており、障害が発生していないサイトのクラスタが「AllLinksSevered」ステータスを示している場合。Tiebreaker ソフトウェアは警告を出しますが、ONTAP はその警告を拒否します。この場合、手動スイッチオーバーも拒否されます。
- Tiebreaker ソフトウェアが被災サイトの少なくとも 1 つのノードまたはクラスターインターフェイスに到達できる場合、または障害が発生していないサイトが FC (NV 相互接続とストレージ) またはクラスタ間ピアリングを介して被災サイトのいずれかのノードに到達できる場合。

4.9 ONTAP Mediator

ONTAP 9.7 以降には、ONTAP Mediator と呼ばれる、障害処理用の新しい MetroCluster IP ソリューションが含まれています。ONTAP には、MetroCluster IP に AUSO 機能を提供する ONTAP Mediator の使用などの追加機能が搭載されています。ONTAP Mediator は、MetroCluster ノードとは別の (第 3 の) 障害ドメインにある Red Hat Enterprise Linux または CentOS Linux の物理サーバーまたは仮想サーバーにインストールされます。ONTAP Mediator の要件と障害の詳細については、「[MetroCluster IP 構成の設置](#)」を参照してください。

備考

Tiebreaker と ONTAP Mediator の両方で同じ MetroCluster 構成を管理することはサポートされていません。MetroCluster 構成の管理に使用できる製品はいずれか 1 つだけです。

5. 技術要件

5.1 ハードウェアとソフトウェアの要件

MetroCluster を設定するときは、サポートされているハードウェアおよびソフトウェアコンポーネントを慎重に検討することが重要です。特定のハードウェアコンポーネントは、お客様の環境によって異なる場合があります。MetroCluster 構成で使用する ONTAP システム、ストレージレイ、および FC スイッチは、必要な要件を満たす必要があります。MetroCluster の実装に必要なソフトウェアコンポーネントは ONTAP だけです。ONTAP は標準機能であり、個別のライセンスは必要ありません。標準の ONTAP ライセンスは、クライアント側とホスト側のプロトコルをサポートします。これに加えて、SnapMirror が非同期ミラーリングを使用してデータを保護し、バックアップデータ保護のためにデータを第 3 のクラスタにレプリケートするための追加機能を提供します。ハードウェアとソフトウェアの要件に関する最新情報については、入手可能な技術資料を参照してください。

MetroCluster IP のハードウェアおよびソフトウェア要件の詳細については、以下のドキュメントを参照してください。

- [ETERNUS AX/HX series MetroCluster IP ソリューションアーキテクチャと設計](#)

最新のハードウェア構成と相互運用性については、以下のツールを使用してください。

- [Fusion](#)

6. まとめ

IP ファブリックのサポートを含む MetroCluster のさまざまな展開オプションは、最も高い柔軟性と高レベルのデータ保護を提供し、すべてのプロトコル、アプリケーション、仮想化環境でシームレスなフロントエンド統合を実現します。

ETERNUS AX series オールフラッシュアレイ , ETERNUS HX series ハイブリッドアレイ
MetroCluster ソリューションアーキテクチャと設計

C140-0043-02Z3

発行年月 2025 年 3 月

発行責任 エフサステクノロジーズ株式会社

- 本書の内容は、改善のため事前連絡なしに変更することがあります。
- 本書の内容は、細心の注意を払って制作致しましたが、本書中の誤字、情報の抜け、本書情報の使用に起因する運用結果に関しましては、責任を負いかねますので予めご了承ください。
- 本書に記載されたデータの使用に起因する第三者の特許権およびその他の権利の侵害については、当社はその責を負いません。
- 無断転載を禁じます。