Fsas Technologies

P3AG-5222-03Z0

ETERNUS AX series オールフラッシュアレイ，ETERNUS AC 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 IP ソリューションのアーキテクチャと設計

目次

図目次

表目次

はじめに

MetroCluster は、ETERNUS AX/AC/HX series で動作する ONTAP 用に開発された、ストレージの継続使用を可能にするソリューションです。MetroCluster IP はイーサネットベースのバックエンドストレージファブリックを用いた最新技術です。MetroCluster IP は高度な冗長構造でクリティカルなビジネスアプリケーションのニーズを満たします。MetroCluster IP は ONTAP に実装されているため個別のライセンスは不要で、クライアントまたはサーバは、ONTAP ストレージに NAS と SAN のどちらでも接続できるようになります。

第 3 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/AC/HX series 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/AC/HX series 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

MetroCluster 構成は、データセンター内外での高可用性 (HA：ハイアベイラビリティ)、データ損失ゼロ、無停止運用のために世界中の企業で数多く使われています。MetroCluster は ONTAP ソフトウェアが無料で提供する機能で、離れた場所と障害ドメインにある 2 つの ONTAP クラスタ間のデータと構成を同期的にミラーリングします。

MetroCluster は 2 つの目的を自動で管理することで、アプリケーションが継続的にストレージを利用できるようにします。

- データ損失ゼロ (目標復旧時点 [RPO] が 0)
  クラスタにデータが書き込まれると同時にミラーリングすることで実現します。
- ほぼゼロの停止時間 (目標復旧時間 [RTO] がほぼ 0)
  構成をセカンダリサイトへミラーリングし、データへのアクセスを自動化することで実現します。

MetroCluster は、2 つのサイトに独立して存在するクラスタ間でデータと構成を自動的にミラーリングすることで、使いやすさを提供します。一方のクラスタ内でストレージが割り当てられると、セカンダリサイトのもう一方のクラスタにも自動でミラーリングされます。SyncMirror はデータ損失 (RPO) ゼロですべてのデータの完全なコピーを作成します。これにより、一方のサイトの作業負荷をいつでももう一方のサイトに切り替えることができ、損失なしでデータを提供できます。

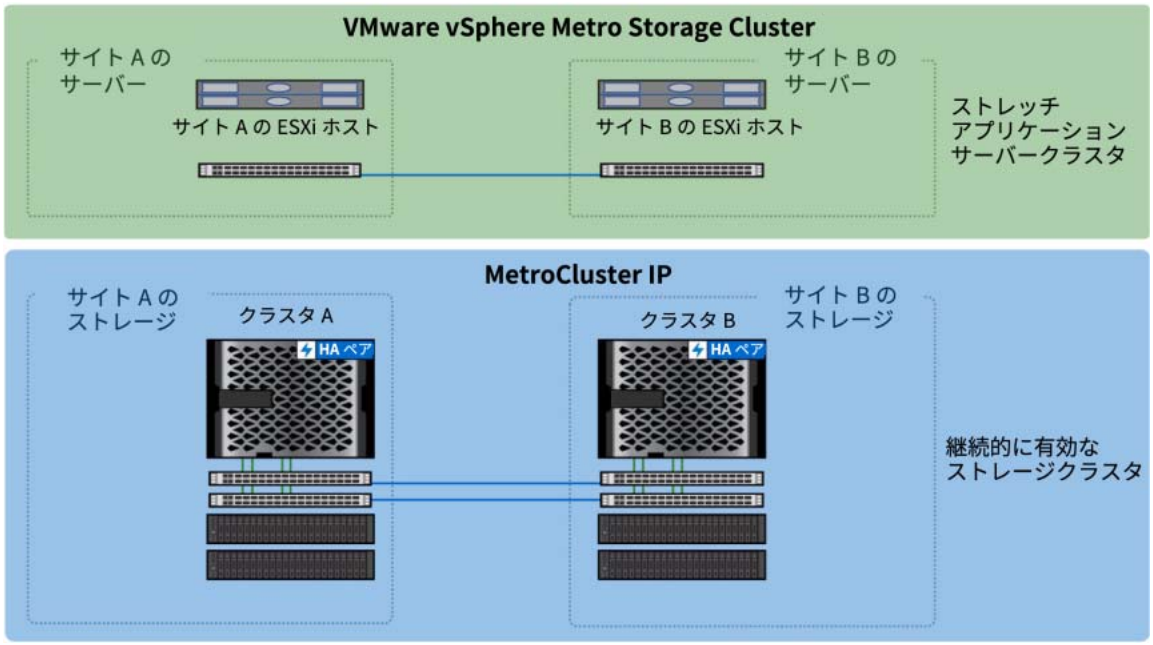
MetroCluster は、セカンダリサイトにある NAS 用データと SAN 用データへのアクセスを可能にするスイッチオーバー処理を管理します。MetroCluster を有効なソリューションとして設計するには、プロトコルのタイムアウト時間内またはそれよりも早い時間内 (通常は 120 秒未満) にスイッチオーバーを実行できるように構成とサイジングを設定します。これにより、ストレージプロトコルがタイムアウトする前にストレージが復旧し、ほぼ 0 秒の RTO を実現します。アプリケーションは障害の影響を受けることなく、継続してデータにアクセスできます。

MetroCluster には、バックエンドストレージファブリックによって定義されるいくつかのバリエーションがあります。MetroCluster の主なストレージファブリックは、FC と Ethernet の 2 つです。Ethernet のストレージファブリックを MetroCluster IP と呼びます。

1.1 継続的な可用性ソリューションの概要

MetroCluster はストレージの継続使用のニーズに応えます。この包括的なソリューションを、可用性を実現する他のアプリケーションと組み合わせることで、サイト全体で災害が発生した場合でも運用を継続できる、耐障害性に優れたアーキテクチャを実現できます。
例として、MetroCluster IP と VMware vSphere Metro Storage Cluster (vMSC) の組み合わせがあります。2 つの製品を組み合わせて回復性の高い仮想化インフラストラクチャを構築し、ビジネスに不可欠なアプリケーションへのニーズに応えます。MetroCluster IP はストレージの可用性を提供し、vMSC はサイト全体が停止した場合でも運用可能な、サイト間にまたがるコンピュータクラスタを提供します。

図 1.1 MetroCluster IP および VMware vSphere Metro Storage Cluster

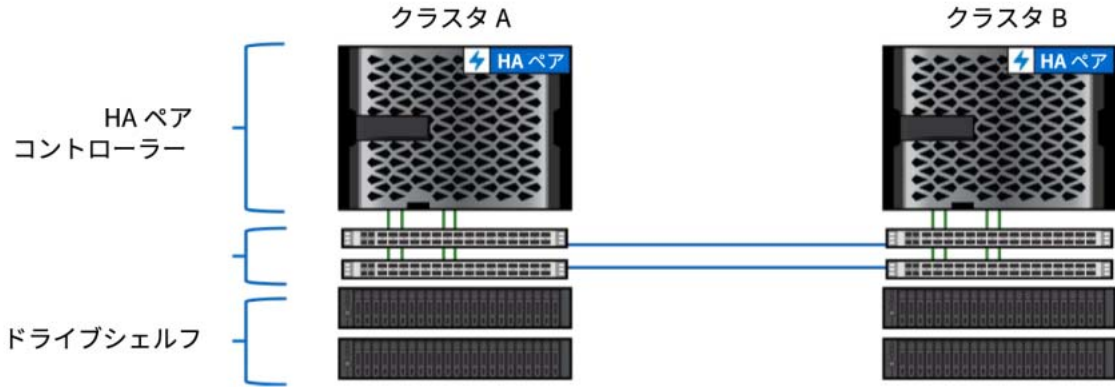


他のマルチサイトアプリケーションソリューションも、MetroCluster と連携して動作するアプリケーションやデータベースで使用できます。

2. MetroCluster IP アーキテクチャ

MetroCluster IP は、Ethernet ストレージファブリックを使用します。MetroCluster ストレージファブリックは、バックエンドストレージファブリックとも呼ばれ、ONTAP だけが使用します。ONTAP クラスタインターコネクト、MetroCluster SyncMirror、MetroCluster NVRAM ミラー通信に使用される独立した専用ネットワークです。

図 2.1　MetroCluster IP アーキテクチャ



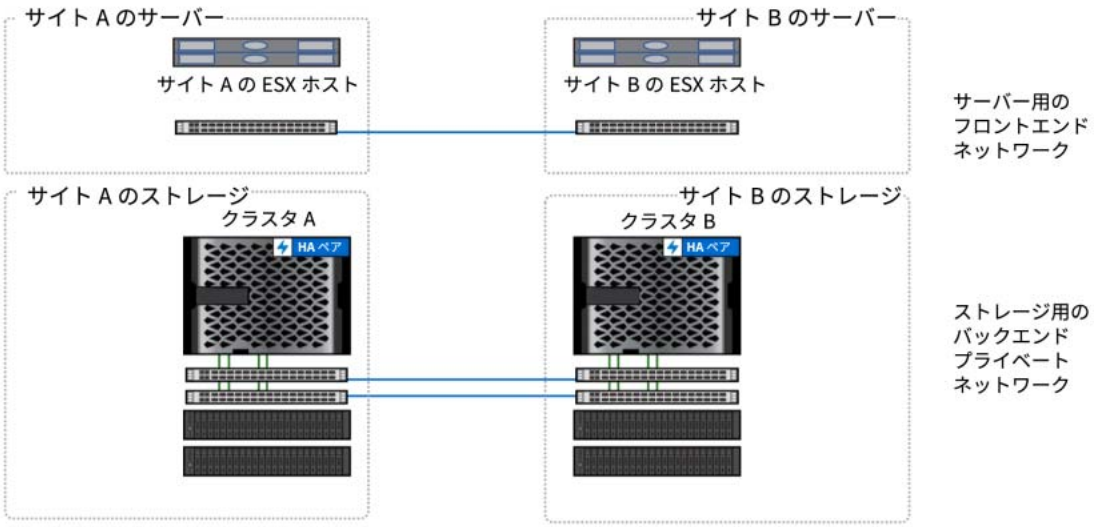
MetroCluster IP ハードウェアの概要

- 各サイトに 1 つの HA ペアコントローラ
- 各サイトに 2 台の高速イーサネットスイッチ
  - ローカルおよびリモートのレプリケーションで使用するクラスタ内／クラスタ間のスイッチを集約
- ドライブシェルフまたは内部ストレージ

MetroCluster は、2 つの独立した ONTAP クラスタ間でデータをミラーリングすることによって、ONTAP の可用性を強化します。各クラスタはサイトまたは障害ドメイン内にあり、ETERNUS AX/AC/HX series の標準 HA 機能を利用します。MetroCluster は、2 つの ONTAP クラスタ間でデータと構成の両方をミラーリングする機能を提供します。MetroCluster には、ストレージプロトコルの標準的なタイムアウト期間内に 1 つのサイトから別のサイトにフェイルオーバーできるように設計された、検証済みのシステムパラメータと制限が含まれています。

MetroCluster の機能とハードウェアは、標準的な ETERNUS AX/AC/HX series で認証されているサブセットです。

図 2.2　ストレージとサーバ



MetroCluster には、複数の機能領域またはコンポーネントに論理的に分割できるアーキテクチャがあります。レプリケーションなど、これらのコンポーネントがどのように動作するかを理解することは、適切に設計されたソリューションを構築し、ソリューションを管理するために重要です。

MetroCluster の主なコンポーネントは以下のとおりです。

- ディザスタリカバリグループ
- レプリケーション
- ネットワーク
- ストレージ

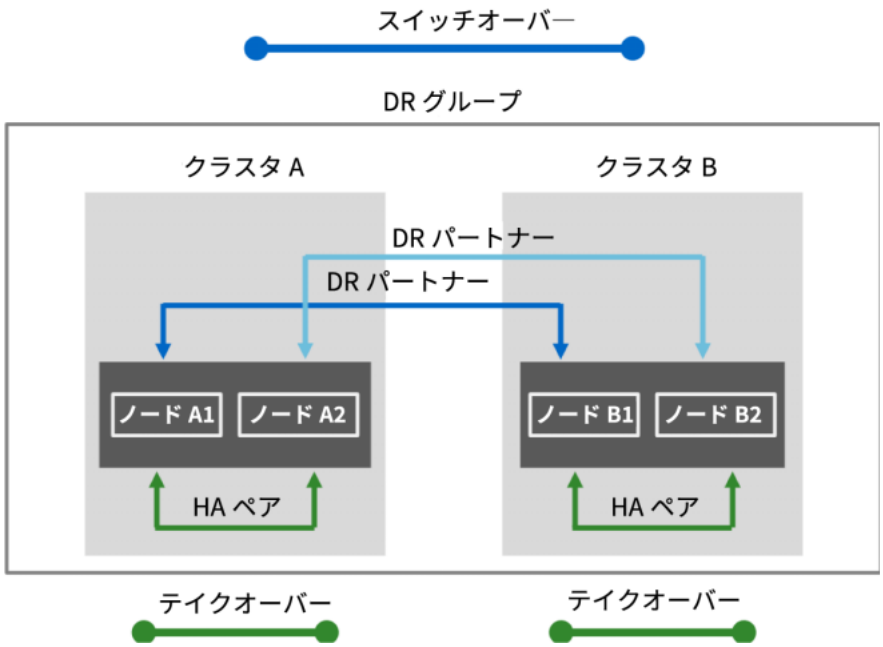
2.1 ディザスタリカバリグループ

MetroCluster IP は、グループとパートナーのディザスタリカバリの概念を使用して、フェイルオーバーとスイッチオーバーの関係を決定します。この 2 つのクラスタ (サイト A とサイト B) は、ディザスタリカバリグループとして一緒に構成されます。グループ内で、ノードはディザスタリカバリパートナーとして関連付けられます。

HA の関係は標準クラスタ内での関係と同じです。単独のコントローラ障害に備え、HA はローカルでフェイルオーバーを実行します。HA は、無停止での ONTAP 更新にも利用されます。サイト全体の障害の場合、ディザスタリカバリ関係を使用してサイト A からサイト B に切り替えます。これをスイッチオーバーと呼びます。

ディザスタリカバリパートナーは、MetroCluster の初期設定で構成し、変更されません。クラスタ A の 1 つのノードとクラスタ B の 1 つのノードをパートナーとして割り当てるコマンドがあります。残りのノードは、ディザスタリカバリグループの構成を完了するために自動的に割り当てられます。

図 2.3 MetroCluster HA とディザスタリカバリ



2.2 MetroCluster IP でのレプリケーション

MetroCluster IP は DAS (ダイレクトアタッチトストレージ) を利用するので、ドライブをストレージファブリックに接続するための外部 SAS ブリッジは必要ありません。この構成では、ディザスタリカバリグループ内の各ノードがストレージ iSCSI ターゲットプロキシとして機能し、ドライブをグループ内の他のノードにエクスポートします。IP ファブリック用のストレージ転送プロトコルは iSCSI (SCSI over TCP/IP) です。これにより、TCP/IP ファブリックを介した iSCSI イニシエータとターゲット間の通信が可能になります。

リモートストレージにアクセスするために、ディザスタリカバリグループの各ノードは iSCSI イニシエータを利用して、パートナーノードの iSCSI ターゲットと iSCSI セッションを確立します。iSCSI と、直接接続したストレージを使用することで、内蔵ドライブを備えたシステムも使用できます。この構成により、各ノードにディザスタリカバリのパートナーノードを提供し、内部ストレージおよび外部ドライブシェルフ内のストレージの両方にアクセスすることができます。

MetroCluster には 3 つのレプリケーションプランがあります。

- 構成レプリケーション
  MetroCluster (MC) は、2 つの ONTAP クラスタで構成されています。各クラスタには、独自のメタデータを含む固有の複製データベース (RDB) があります。スイッチオーバー中のクラスタ間のメタデータオブジェクトの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) 、メタデータボリューム (MDV) の 3 つのコンポーネントが含まれます。MC を使用すると、ピアリングネットワークを使用してクラスタ間で設定オブジェクトの複製が可能になり、変更が発生すると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 へ発生とほぼ同時に伝達されます。MDV は、クラスタピアリングネットワークが一時的に使用できなくなった場合のフォールバックメカニズムとして使用されます。
- NVRAM レプリケーション
  NVRAM レプリケーションには、フェイルオーバーまたはスイッチオーバーに伴うデータ損失から保護するために、ローカルノードの NVRAM をリモートのディザスタリカバリノードの NVRAM にコピーする作業が含まれます。NVRAM はローカルのハイアベイラビリティ (HA) パートナーにローカルでミラーリングされ、ディザスタリカバリパートナーにはリモートでミラーリングされます。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 補助パートナー用の 4 つのパーティションに分割されます。各ノードの NVRAM は、4 ノード構成で 2 回ミラーリングされます。これには、2 つの 4 ノードの DR グループを持つ 8 ノード構成が含まれます。ミラーリングは HA パートナーと DR パートナーの両方に対して実行され、アップデートは MetroCluster IP の ISL (Inter-Switch Link) 上で iWARP プロトコルを使用して送信されます。
- ストレージのレプリケーション
  MetroCluster IP では、RAID SyncMirror (RSM) を使用して、ローカルおよびリモートバックエンドドライブをミラーリングします。ディザスタリカバリグループの各ノードは、バックエンドストレージを論理的に共有するリモート iSCSI ターゲットとして機能します。リモートバックエンドドライブにアクセスする場合、そのノードはリモートのディザスタリカバリパートナーノードを経由して、iSCSI ターゲットを介して提供されるリモートドライブにアクセスします。ブロックは、NVRAM 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 による書き込みは RAID 層で行われるため、重複排除や圧縮などのストレージ効率が向上し、SyncMirror 操作によって書き込まれるデータが削減されます。

MetroCluster 内のレプリケーションの詳細については、マニュアルサイトに掲載の以下のリソースを参照してください。

- ETERNUS AX/HX series MetroCluster ソリューションアーキテクチャと設計

2.2.1 レプリケーションの暗号化

MetroCluster には、サイト間で送信されるデータを暗号化するメカニズムはありません。ネットワーク層やストレージ層でデータを暗号化するオプションがあります。

- ネットワーク層の暗号化は、波長分割多重 (WDM) デバイスおよびスイッチベースの暗号化を使用して実現できます。ラウンドトリップのレイテンシが要件の範囲内であれば、外部暗号化デバイスの使用がサポートされます。
- ホスト側でストレージ層のデータ暗号化機能を使用することもできます。ただし、この方法では ONTAP が通常提供するストレージ効率が得られません。

備考

ボリューム暗号化 (VE) を使用してボリュームに書き込まれたデータを暗号化できますが、すべての書き込みはホストによって書き込まれた暗号化されていないブロックデータを含めて NVRAM に複製されます。

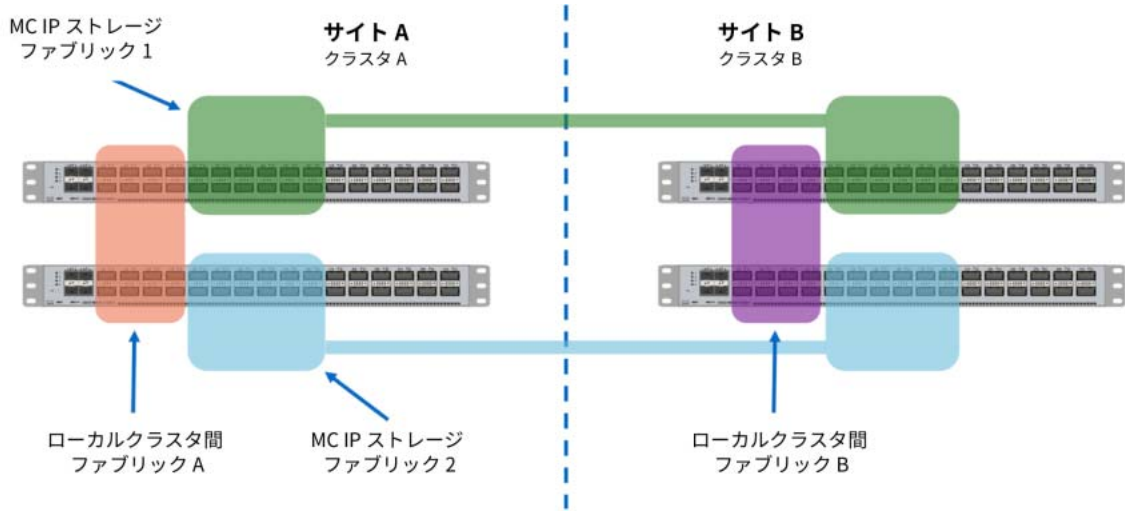
2.3 ネットワーク

MetroCluster には、2 つの独立したストレージファブリックがあります。

- クラスタインターコネクト
- MetroCluster IP ネットワーク

各ネットワークは特定の機能専用です。各ネットワークにマッピングする特定の仮想 LAN (VLAN) があり、個別のデータリンク層、つまり OSI 標準の第 2 層を作成します。

図 2.4 MetroCluster IP 複合ファブリック



標準のスイッチ設定として、RCF (Reference Configuration File) があります。スイッチベンダーとプラットフォームモデルに応じて、RCF ジェネレータツールを使用して RCF ファイルを作成します。入手方法については、サポート部門にお問い合わせください。スイッチの構成を変更するには、RCF ファイルを使用する必要があります。RCF ファイルは、4 つの個別構成ファイル (スイッチごとに 1 つ) を 1 つにまとめたものです。

RCF は、VLAN およびチャネルグループ識別子 (ID) を指定します。これらはバックエンドストレージスイッチ内でのみ使用されます。IP アドレスには特定の要件があります。

2.3.1 クラスタ相互接続

ONTAP クラスタ相互接続は、ETERNUS AX/AC/HX series ストレージシステムのノード通信に不可欠なローカル専用の高速ネットワークです。この接続では、すべてのクラスタトラフィックがノードを配置しているサイトに対してローカルであるため、サイト間は接続されません。効率的で信頼性の高い通信を保証し、拡張性をサポートします。

MetroCluster IP セットアップでは、専用スイッチポートがインターコネクトトラフィックを処理します。2 つの ISL ポートがサイトスイッチを接続し、冗長ローカル VLAN を形成します。ブレイクアウトケーブルは、ネイティブポートよりも低い速度で使用されます。

相互接続は、データネットワークや管理ネットワークとは別に、ストレージクラスタ内に専用のプライベートネットワークを形成します。スイッチ、ケーブル、NIC または HCA を使用して、冗長性、高速接続性、およびスケーラビリティを提供します。これらのコンポーネントは、クラスタ化されたストレージ環境で高いパフォーマンスと信頼性を保証します。

備考

- ETERNUS AX1200、ETERNUS AX2200、ETERNUS AC2100、および ETERNUS HX2200 は、VLAN を使用して分離し、MetroCluster IP ネットワークとクラスタポートを共有します。
- ONTAP 9.7 からは、ストレージコントローラ間でポートを相互接続することにより、スイッチレス相互接続のオプションを利用できるようになりました。これは、仕様に準拠したスイッチの構成で使用されます。

2.3.2 MetroCluster IP ネットワーク

MetroCluster IP 構成では、各サイトは 2 つの独立したストレージファブリックで構成されます。これらのローカルファブリックは、対応するリモートファブリックに接続されていますが、クラスタインターコネクトとは異なり相互には接続されていません。

各 MetroCluster IP スイッチには、ノード接続用に指定された複数のポートがあります。サイトごとに 2 つのノードを含む標準の 4 ノード MetroCluster IP 構成では、これらのポートのうち 2 つだけが使用されます。サイトの 2 つのノードは、それぞれ別の Ethernet インターフェイスを使用して両方のスイッチに接続します。通常、これらの接続には標準のクラスタ HA 設定に使用されるオンボードポートまたはインターフェイス標準の ONTAP インターコネクトポートが使用されます。

MetroCluster IP ネットワークは、インターネットを経由したリモートダイレクトメモリアクセス (RDMA) プロトコル (iWARP) に最適化された専用 Ethernet アダプタを採用しています。このアダプタは、TCP オフロードエンジン (TOE) と iSCSI オフロード機能の両方を備えており、高速 Ethernet 上での RDMA を可能にします。各ノードには 2 つの iWARP/iSCSI アダプタポートがあり、各スイッチに 1 つのポートが接続されています。スイッチは、図 2.4 に示すようにローカル接続されていない個別のファブリックを形成します。

ETERNUS AX1200、ETERNUS AX2200、ETERNUS AC2100、ETERNUS HX2200 などのプラットフォームでは、ソフトウェア iWARP が使用されます。これらのプラットフォームには固定数のネットワークポートがあり、その用途をフロントエンド、つまりホストサイドのデータアクセスへの使用に最適化するため、通常はクラスタインターフェースとしてリザーブされている 2 台のオンボードポート (e0a および e0b) の組み合わせとなっています。この構成により、クラスタトラフィックと MetroCluster IP トラフィックが同じポートを共有し、残りの 4 つのネットワークポートをホスト側のデータアクセスに使用できます。

備考

iWARP は、IETF RFC 5040 – A Remote Direct Memory Access Protocol Specification に記載されている、規格ベースのプロトコルです。

2.4 ストレージ

MetroCluster IP のストレージは、2 つのサイト間で直接共有されません。複数のサイトで一意のシェルフ ID を設定する必要はありませんが、各シェルフに一意の ID を割り当てることを推奨します。これにより、ストレージシステム内のシェルフの識別と管理が容易になります。各サイトのストレージには、ローカル HA ペアからのみ直接アクセスできます。「ストレージのレプリケーション」(P.11) で説明しているとおり、リモートストレージは iSCSI を使用したローカルノードで使用可能になります。

2.4.1 SyncMirror

SyncMirror または RSM (RAID SyncMirror) は、MetroCluster でアグリゲートのサイト間ミラーリングをするためのテクノロジーです。各アグリゲートで 2 つのプレックスを構築します。プレックスはそれぞれ「プール 0」および「プール 1」と呼びます。プール 0 にはノードのローカルストレージが含まれ、プール 1 にはリモートのミラーコピーが含まれます。

2.4.2 ADP

MetroCluster IP は、ETERNUS AX/AC series 上で、Root-Data-Data (RD2) パーティショニングを使用したアドバンストディスクパーティショニング (ADPv2) をサポートしています。アドバンストディスクパーティショニング (ADP) はストレージ効率を高める機能で、HA ペア内のコントローラとアグリゲートの間で物理ドライブの容量を共有できます。ADP を使用すると、ドライブ全体のパーティション化と比較して、使用可能な有効容量が増加し、ストレージ効率が最大 10~40% 向上します。RD2 パーティショニングでは、ドライブが 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されるため、1 台のドライブの容量と IOPS を ETERNUS AX/AC series の両方のコントローラで使用できます。ADP は、MC の初期化時にデフォルトで適用されます。

ADP の詳細については、マニュアルサイトに掲載の以下のリソースを参照してください。

- ETERNUS AX/AC/HX series MetroCluster IP インストールおよび設定ガイド

3. ソリューション設計

性能、容量、および回復性の要件に対応するには、ソリューションを適切に設計することが鍵となります。ソリューションを設計するための全体的な手順には、容量や性能の要件にあったサイジングの他、サポートされているホストとプラットフォーム構成の確認が含まれます。以下の点を考慮する必要があります。

- ホストとプロトコルのサポートの確認
- 性能要件を満たすソリューションのサイジング
- 容量要件を満たすソリューション (アクティブ／アクティブ、アクティブ／パッシブ構成) のサイジング
- システム制限の確認
- サイト間での ISL のサイジング
- ケーブル接続の要件

3.1 サポートの確認

ホスト側のプロトコルと OS のバージョンが ONTAP の設計と同様にサポートされていることを確認します。結果ページに表示されたアラートが MetroCluster に適用されるかどうかを確認します。

3.2 ハードウェアコンポーネント

MetroCluster IP セットアップに含まれる主なハードウェアコンポーネントの概要を以下に示します。特定のモデルと部品番号は、お客様の組織の構成と要件によって異なる場合があります。これらの詳細については、サポート部門にお問い合わせください。

3.2.1 プラットフォーム

MetroCluster IP 構成で一般的に使用されるモデルは、以下のプラットフォームで提供されています。

- ETERNUS AX series
  ETERNUS AX series は、フラッシュメモリ専用に設計されており、業界トップクラスの性能、密度、スケーラビリティ、セキュリティ、およびネットワーク接続性を提供します。エンタープライズ用のオールフラッシュアレイとして業界最小のレイテンシを提供し、最も要求の厳しいワークロードや AI/DL アプリケーションの実行に最適です。
- ETERNUS AC series
  ETERNUS AC series は、最新の QLC フラッシュテクノロジーを搭載したフラッシュメモリに、より多くのデータを移動できます。本システムは、大容量の導入に適した手頃な手段として、データセンターをオールフラッシュアレイに近代化し、クラウドへ接続します。
- ETERNUS AX/AC series ASA
  ETERNUS AX/AC series ASA モデルは、業界トップのパフォーマンスと信頼性を提供する ETERNUS AX/AC series をベースに構築されています。ETERNUS AX/AC series は、複数のワークロードに対応するストレージリソースの統合と共有を希望するお客様向けに、エンタープライズクラスの SAN ソリューションを提供しています。
- ETERNUS HX series
  ETERNUS HX series は、フラッシュドライブとハードディスクドライブを組み合わせて使用できるハイブリッドストレージシステムです。ETERNUS HX series は、容量とパフォーマンスの最適なバランスを提供することで、導入と運用を容易にし、将来の成長とクラウド統合に対応できる柔軟性を備えています。
  詳細については、マニュアルサイトを参照してください。

3.2.2 コントローラモデル

MetroCluster IP で使用可能なモデルは、さまざまなレベルのパフォーマンス、容量、拡張性を提供し、さまざまなビジネスニーズと予算要件に対応します。MetroCluster IP ソリューションを導入するときは、プライマリ (一次) サイトとセカンダリ (二次) サイトの両方に、同じタイプのストレージシステム (ETERNUS AX series、ETERNUS AC series、ETERNUS HX series のいずれか) と互換性のあるハードウェアおよびソフトウェアコンポーネントがあることを確認する必要があります。以下に、MetroCluster IP でサポートされるコントローラモデルを示します。

表 3.1 MetroCluster IP コントローラモデル

プラットフォーム	エントリ	ミッドレンジ
ETERNUS AX series	AX1200 AX1200 ASA AX2100 AX2100 ASA AX2200 AX2200 ASA	AX4100 AX4100 ASA
ETERNUS AC series	AC2100 AC2100 ASA	AC4100
ETERNUS HX series	HX2200	HX6100

特定のモデルと機能は変更される可能性があります。マニュアルサイトを参照してください。

3.2.3 ドライブシェルフ

ドライブシェルフは、ストレージシステムの容量を拡張するためのストレージデバイスとして機能します。MetroCluster IP のセットアップで使用すると、物理的に離れた 2 つのサイト間のデータレプリケーションを容易にし、データの高可用性を確保する上で重要な役割を果たします。ドライブシェルフには両サイトのノード (コントローラ) からアクセスできるため、2 つのサイト間でミラーボリュームを作成できます。したがって、ドライブシェルフは MetroCluster IP 構成の重要な要素であり、地理的に分散したサイト間でデータとアプリケーションの可用性を維持するために必要なストレージ容量とデータレプリケーション機能を提供します。ドライブにはさまざまなストレージ容量があり、物理的なサイズに応じて、通常 3.5 インチのラージフォームファクタ (LFF) 、または通常 2.5 インチのスモールフォームファクタ (SFF) のいずれかに分類されます。以下の表では、使用可能なシェルフと、サポートされている MetroCluster IP コントローラとの互換性を比較しています。

表 3.2 ドライブシェルフの特長の比較

特長	NS224	DS224C	DS212C	DS460C
フォームファクタ	2U	2U	2U	4U
エンクロージャあたりのドライブ数	24 台 (SFF)	24 台 (SFF)	12 台 (LFF)	60 台 (LFF)
ドライブタイプ	SSD	SSD、HDD	SSD、HDD	SSD、HDD
I/O モジュール	デュアル NSM	デュアル IOM12	デュアル IOM12	デュアル IOM12
接続性	100Gb/s Eth	12Gb/s SAS	12Gb/s SAS	12Gb/s SAS

表 3.3 ETERNUS AX series および ETERNUS AX series ASA コントローラとドライブシェルフの互換性

特長	AX1200 AX1200 ASA	AX2100 AX2100 ASA	AX2200 AX2200 ASA	AX4100 AX4100 ASA
DS212C	–	–	–	–
DS224C	利用可能	利用可能	–	利用可能
DS460C	–	–	–	–
NS224	–	–	利用可能	利用可能

表 3.4 ETERNUS AC series および ETERNUS AC series ASA コントローラとドライブシェルフの互換性

特長	AC2100 AC2100 ASA	AC4100
DS212C	–	–
DS224C	–	–
DS460C	–	–
NS224	利用可能	利用可能

表 3.5 ETERNUS HX series コントローラとドライブシェルフの互換性

特長	HX2200	HX6100
DS212C	利用可能	利用可能
DS224C	利用可能	利用可能
DS460C	利用可能	利用可能
NS224	–	–

3.2.4 スイッチ

レイヤ 3 Ethernet スイッチは、データセンター間に堅牢な IP ベースの相互接続を確立し、データの整合性と無停止運用を確保するために不可欠です。10/25/40/50/100 ギガビットの Ethernet 接続をサポートしているため、迅速なデータ転送と最小限のレイテンシを実現します。

これらのスイッチは、ジャンボフレーム、VLAN タグ、リンクアグリゲーションなどの機能を提供し、MetroCluster 環境のパフォーマンスと信頼性を向上させます。ストレージコントローラ間のシームレスな通信を維持するために、スイッチは OSPF や BGP などのルーティングプロトコルをサポートし、最適なパス選択とネットワークロードバランシングを実現します。

MetroCluster IP の各環境には、冗長性を確保するためにサイトごとに 2 台ずつ、合計 4 台のスイッチが必要です。また、スイッチモデルを混在させることはできません。特定のプラットフォームおよび ONTAP バージョンでサポートされるスイッチモデルについては、サポート部門にお問い合わせください。

ブレークアウトケーブルは、10Gb ISL リンクで光モジュールを接続するなど、ネイティブポート速度よりも低いポート速度に使用されます。

3.2.4.1 適合スイッチ (仕様に準拠した既存スイッチを使用した MetroCluster IP)

ONTAP 9.7 では、動作検証済みのスイッチを必要としない特定のプラットフォームでの MetroCluster IP のサポートが導入されました。これにより、MetroCluster IP バックエンドストレージファブリック以外の目的で既存のスイッチを使用できます。このソリューションでは、各サイトでスイッチレスクラスタ構成を使用します (2 ノードのスイッチレスクラスタ)。このセットアップでは、クラスタインターコネクトのインターフェイスが相互接続され、MetroCluster IP インターフェイスが既存の MetroCluster 対応スイッチに接続されます。

詳細については、FTI マニュアルサイトに掲載の「ETERNUS AX/AC/HX series MetroCluster IP インストールおよび設定ガイド」の「MetroCluster 準拠スイッチの使用に関する考慮事項」を参照してください。

以下の表は、使用可能なスイッチモデルと、サポートされている MetroCluster コントローラモデルとの互換性を比較しています。

表 3.6 MetroCluster IP スイッチのモデル

モデル	PN	ネイティブポート速度
Cisco Nexus 9336C	X190200-CS-PE	40Gb/100Gb
Cisco Nexus 3232C	X190100	40Gb/100Gb

表 3.7 ETERNUS AX series および ETERNUS AX series ASA コントローラとスイッチの互換性

スイッチ	AX1200 AX1200 ASA	AX2100 AX2100 ASA	AX2200 AX2200 ASA	AX4100 AX4100 ASA
Cisco Nexus 9336C	対応	対応	対応	対応
Cisco Nexus 3232C	対応	対応	対応	対応

表 3.8 ETERNUS AC series および ETERNUS AC series ASA コントローラとスイッチの互換性

スイッチ	AC2100 AC2100 ASA	AC4100
Cisco Nexus 9336C	対応	対応
Cisco Nexus 3232C	対応	対応

表 3.9 ETERNUS HX series コントローラとスイッチの互換性

スイッチ	HX2200	HX6100
Cisco Nexus 9336C	対応	対応
Cisco Nexus 3232C	対応	対応

3.2.5 ネットワークアダプタ

MetroCluster の IP 構成では、専用のネットワークアダプタを使用して、データレプリケーションと IP ネットワーク経由の通信を効率的に行うことができます。これらのアダプタはプラットフォームに依存し、デュアルポートを備えた高速 Ethernet を提供し、2 つの異なるレイヤ 2 またはレイヤ 3 ネットワークへの接続を可能にします。さらに、このアダプタは、Ethernet ネットワーク経由でのリモートダイレクトメモリアクセス (RDMA) を有効にする iWARP オフロード機能を提供します。この機能により、サーバーとストレージシステムのメモリ同士でデータを直接転送できるため、レイテンシとオーバーヘッドが最小限に抑えられます。

これらのアダプタは、ストレージおよび NVRAM レプリケーションに使用されるノードとスイッチの接続を担当します。クラスタインターコネクトの場合は、個別のネットワークアダプターまたはネットワークポートが使用されます。

iWARP とネットワーク接続の実装方法はストレージのモデルによって異なり、ネットワークアダプタの要件と実装方法はプラットフォームのモデルによって異なります。例えば、

- ETERNUS AX1200、ETERNUS AX2200、ETERNUS AC2100、および ETERNUS HX2200 は、オンボードポートを使用します。
- ETERNUS AX4100、ETERNUS AC4100、ETERNUS HX6100 は、単一のネットワークアダプタを使用します。

ETERNUS AX1200、ETERNUS AX2200、ETERNUS AC2100、および ETERNUS HX2200 は、クラスタインターフェイスに組み合わされたソフトウェア iWARP を使用して、オンボードの e0a および e0b インターフェイス上でのトラフィック共有を実現します。これにより、バックエンドストレージに必要なポート数が削減され、ホスト側のデータインターフェイスで使用可能なポート数が最大化されます。

3.3 ソリューションのサイジング

特定のストレージ容量または性能要件を満たすために、ソリューションをサイジングできます。MetroCluster のサイジングは容量に関して、HA ペアのサイジングと似ています。MetroCluster では、ペアとなるサイトでデータをミラーコピーするため、ストレージデバイスの容量は HA ペアで使用する量の 2 倍になります。

パフォーマンスに関するサイジングでは、性能を決定する要素となる ISL を、ISL サイジングスプレッドシートで計算できます。

3.3.1 アクティブ／パッシブ構成

アクティブ／パッシブ構成では、2 つのストレージノードまたはクラスタを活用し、高可用性、ダウンタイムの最小化、フェイルオーバー / フェイルバック間のシームレスな移行を実現します。アクティブノードはクライアントの I/O 要求とストレージリソースを管理し、パッシブノードはアクティブノードの稼働状態を監視して、障害や保守の際に引き継ぐ準備を整えます。主なメリットとして、高可用性、無停止のフェイルオーバー / フェイルバック、継続的な監視が挙げられます。

アクティブ／パッシブ構成には以下の 2 種類があります。

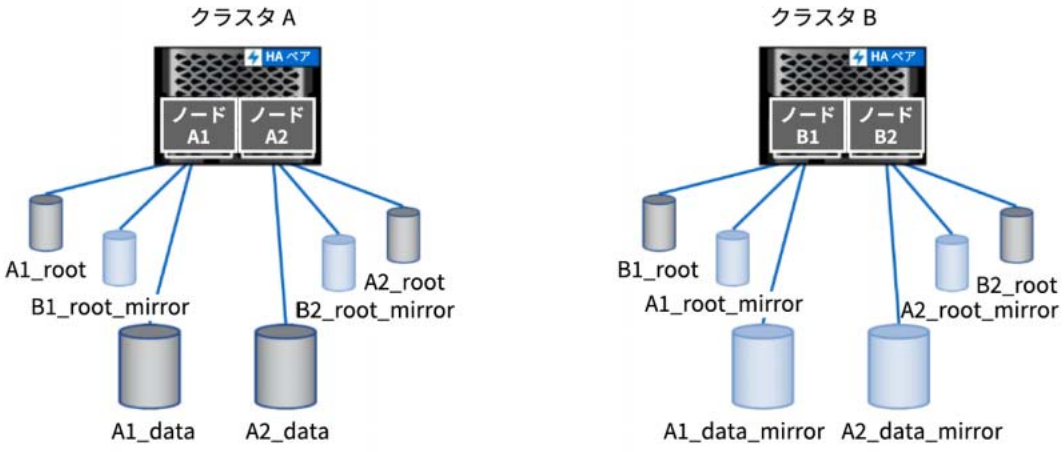
- アクティブ／パッシブクラスタ
- クラスタ内のアクティブ／パッシブ HA

3.3.1.1 アクティブ／パッシブクラスタ

アクティブ／パッシブクラスタ (またはサイト) では、1 つのサイト (アクティブサイト) が業務ワークロードに使用され、もう 1 つのサイトはフェイルオーバー用に最小容量で構成されます。これにより稼働ストレージとワークロードをアクティブサイトに持たせ、小規模なストレージを構築できます。セカンダリサイトがスイッチオーバーやサイト障害にのみ使われるため、ドライブ容量を必要としないことからコストパフォーマンスにも優れます。

アクティブ／パッシブクラスタ構成では、一方のクラスタにすべての pool0 ディスクがあり、もう一方のクラスタにすべての pool1 ディスクがあります。アクティブ／パッシブクラスタ (またはサイト) では、MetroCluster のメタデータを格納するボリュームをホストするため、小規模なデータアグリゲートを作成する必要があります。ルートボリュームと、メタデータを格納するボリューム用の小さいデータボリュームを除き、パッシブサイトにはデータのミラーコピーのみが含まれます。

図 3.1 アクティブ／パッシブクラスタまたはサイト



3.3.1.2 アクティブ／パッシブ HA 構成

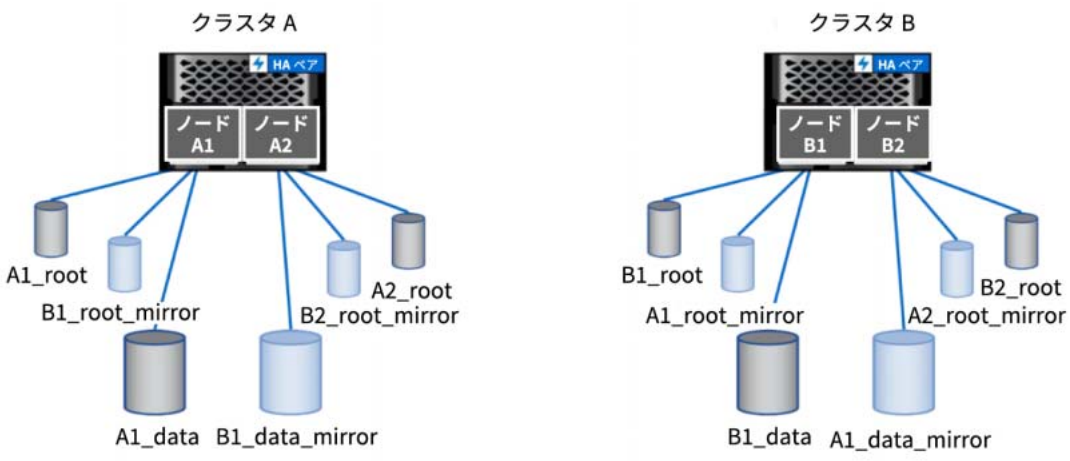
アクティブ／パッシブ HA 構成では、HA ペアのノードの 1 つにストレージが割り当てられます。この構成は通常、小規模な構成で容量を最大化する場合に使用されます。

備考

ETERNUS AX/AC series の場合は、ストレージがノード間で均等に分散されている必要があります。

この例では、各ノードがルートボリュームを持つドライブを所有しています。ローカルのアクティブデータボリュームはノード 1 にホストされ、リモートミラーコピーはノード 2 にホストされます。

図 3.2 アクティブ／パッシブ HA



3.4 ネットワーク構成

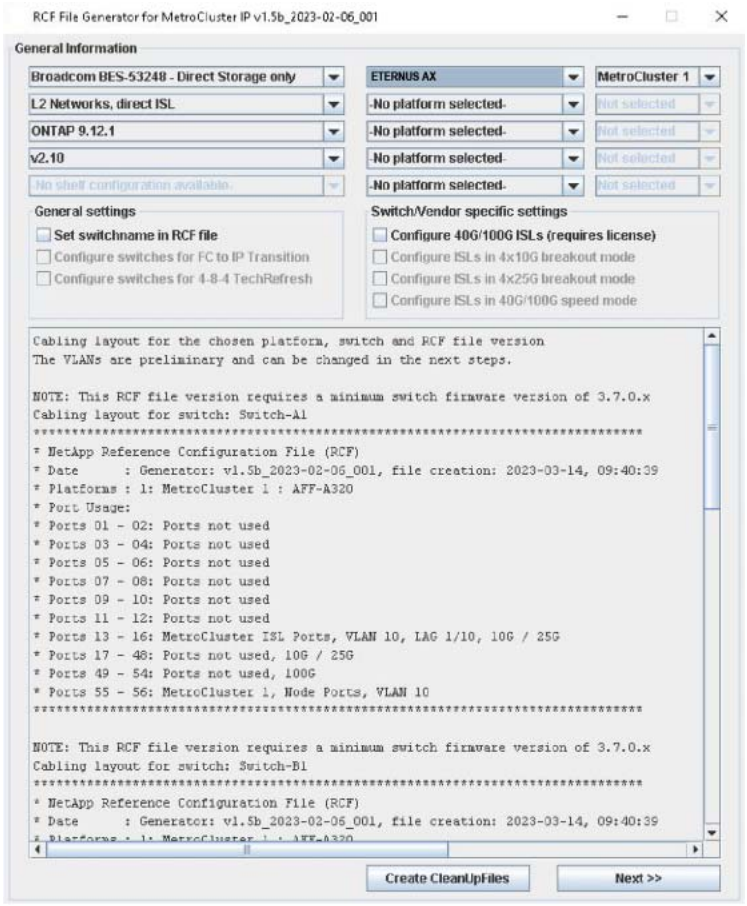
ネットワーク構成では、他のネットワークと重複しない VLAN および IP アドレスが必要です。VLAN は、プライベートバックエンドファブリック用で、スイッチの RCF ファイルによって自動的に割り当てられます。

RCF ファイルジェネレーターユーティリティを使用すると、MetroCluster IP スイッチのすべてのモデルおよびサポートされているプラットフォーム用の RCF ファイルを作成できます。RCF ファイルジェネレーターユーティリティでは、お客様が用意した VLAN ID を使って、共有レイヤ 2 サイト間ネットワークで使用する RCF ファイルを作成できます。また、RCF ファイルジェネレーターは、レイヤ 3 ネットワーク構成をサポートしています。RCF ファイルジェネレーターからの出力には、選択したプラットフォーム、スイッチ、および RCF ファイルバージョンのケーブル接続レイアウトが記述されます。

備考

ONTAP 9.7 では、ETERNUS AX2100 および ETERNUS HX2200 コントローラの VLAN ID の変更が有効になっていません。ただし、ONTAP 9.8 以降では、ETERNUS AX2100/AX2200 および ETERNUS HX2200 の VLAN を指定できます。デフォルトは 10 と 20 で、ユーザ指定の VLAN は 100 より大きく 4096 より小さい値にする必要があります。

図 3.3 RCF ファイルジェネレーター



通常のスイッチ構成では、将来の拡張に備えてほとんどのポートが未使用です。スイッチは 2 つの冗長ファブリックを作成し、各ノードにはクラスタ接続と MetroCluster IP ノード接続があります。ノードからスイッチまでの距離に応じて、ケーブル接続にはいくつかの方法があります。可能であれば、ノードとスイッチを同じラックに配置して、光ケーブルを必要とせずに Twinax Copper ケーブルを使用できるようにすることが最適です。

スイッチはネイティブポート速度 (モデルによって 10/25Gbps、40Gbps、100Gbps) に対応しています。ネイティブポートは、ブレークアウトモードで動作することもできます。ブレークアウトモードでは、ポートは個別のインターフェイスとして使用される 4 つの個別のレーンに分割されます。ブレークアウトモードで動作している場合、40Gb スイッチのポートはケーブルまたは光モジュールに依存する 4 つの 10Gbps インターフェースとして動作することが可能です。ただし、すべての光モジュールがブレークアウトモードをサポートしているわけではありません。

100Gb スイッチもブレークアウトモードをサポートしています。ネイティブポート速度で動作する場合、1 つの物理インターフェースは 4 つの 25Gbp インターフェースとして動作します。特定のケーブルおよび光モジュールは、ブレークアウトモードでの動作をサポートしています。各プラットフォームで速度を指定する必要がある場合は、RCF ファイルにブレークアウトポートを事前に設定しておきます。

備考

コマンドラインから RCF ファイルジェネレーターを起動できます。JAVA_HOME 環境変数を設定する必要があります。RCF ファイルジェネレーターを実行するには、以下のコマンドを入力します。

```
java -jar RcfFileGenerator.jar
```

3.5 クラスタファブリックスイッチ間リンク (ISL)

クラスタ相互接続ファブリックは、MetroCluster IP 構成のノード間の通信に使用される専用ネットワークです。このネットワークでは、信頼性の高い最適なパフォーマンスを確保するために、特定のスイッチ要件とケーブルタイプが必要です。

ISL ネットワークに使用するスイッチにはネイティブポートが必要であり、高速データ転送をサポートするためにジャンボフレームを設定する必要があります。ファブリックには最低 2 台のスイッチが必要です。

Twinax Copper ケーブルまたは光ファイバーケーブルを使用することを推奨します。通常、7m までの距離には Twinax Copper ケーブルが使用され、それよりも長い距離には光ファイバーケーブルが使用されます。干渉を減らし、信頼性の高いデータ転送を保証するために、ケーブルは高品質で適切にシールドされている必要があります。

3.6 MetroCluster IP スイッチ間リンク設計

MetroCluster IP 構成で高可用性と耐障害性を確保するには、2 つのスイッチを高速かつ低レイテンシのリンクで接続する ISL 設計が必要です。ISL リンク設計には、いくつかの重要な要素があります。レプリケーショントラフィックをサポートするための十分な帯域幅、迅速なデータ転送のための低いレイテンシ、高可用性のための冗長性、レプリケーショントラフィックに優先順位を付けるための適切な QoS 設定、不正アクセスを防止するためのセキュアな認証などです。

MetroCluster IP の ISL リンク設計のプロセスは、サイト間の距離やネットワーク機能など、サイト固有の要因により複雑になる場合があります。ストレージ性能の要件、ダイレクトファイバーが使用可能であるかどうか、多重化デバイス、および既存のファイバーインストラクチャなど、さまざまな要件を考慮して各リンクに必要なコンポーネントや対応可能な距離を決定する必要があります。

現在の MetroCluster IP の ISL ネットワーキングの Round Trip Time (RTT) 要件は、7 ミリ秒以下のレイテンシ、3 ミリ秒以下のジッタ、0.01% 未満のパケット損失です。これらの要件により、レイテンシは最大で 10 ミリ秒となります。これは、データ転送に厳密な時間制限がある同期レプリケーションには不可欠です。

MetroCluster IP の ISL 構成を設計する場合、サポートされる最大距離を提供する適切な光モジュールおよびそれに対応する光ケーブル構成を選択することが重要です。短距離の場合は、マルチモードの光モジュールとケーブルが適しており、コストパフォーマンスにも優れています。長距離の場合は、ロングレンジの光モジュールとシングルモードのファイバーケーブルが必要です。最大距離については、通信プロバイダに問い合わせてください。

場合によっては、個別のアベイラビリティゾーンにあるラックで分離された単一のデータセンターに通常の Ethernet ケーブル接続を仕様可能です。ネイティブスイッチポートの ISL の速度は、スイッチやモジュールによって 25Gb、40Gb、または 100Gb になります。

つまり、サイト固有の要因を考慮した MetroCluster IP の ISL リンク設計の慎重な計画と実装により、スイッチ間の高速で信頼性の高い安全なレプリケーショントラフィックを確保できます。

3.6.1 近距離のラック間接続

ラック内または隣接ラック間など、近接した個別の MetroCluster IP 構成では、ISL リンクに銅線ケーブルを使用できます。これは、研究所の構成またはテスト構成でよくみられます。表 3.10 および表 3.11 では、40Gb および 100Gb の両方のスイッチのスイッチ間を接続する Copper ケーブルの概要を示します。

表 3.10 40Gb スイッチ間の距離 (ケーブルのおおよその長さ) が 3m ～ 5m の場合

距離	利用するケーブルの概要
1m 未満	ケーブル、Copper、QSFP+-QSFP+、40GbE、1m
3m 未満	ケーブル、Copper、QSFP+-QSFP+、40GbE、3m
5m 未満	ケーブル、Copper、QSFP+-QSFP+、40GbE、5m

表 3.11 100Gb スイッチ間の距離 (ケーブルのおおよその長さ) が 3m ～ 5m の場合

距離	利用するケーブルの概要
1m 未満	ケーブル、Copper、QSFP28-QSFP28、100GbE、1m
2m 未満	ケーブル、Copper、QSFP28-QSFP28、100GbE、2m
5m 未満	ケーブル、Copper、QSFP28-QSFP28、100GbE、5m

同様に、ラック間の光ケーブルの使用も可能です。これにより、距離が光モジュールの仕様内であれば、シンプルな ISL 構成が可能になります。

表 3.12 40GbE スイッチ用の短距離光モジュール

距離	モジュール PN	利用するケーブルの概要
OM4 で最大 400m	X65401	XCVR、QSFP+、光、40GbE、短波

表 3.13 100GbE スイッチ用の短距離光モジュール

距離	モジュール PN	利用するケーブルの概要
OM4 で最大 100m	X65405	XCVR、QSFP28、光、100GbE、短波

表 3.14 40Gb および 100Gb 光ケーブル

ケーブル長	モジュール PN	利用するケーブルの概要
2m	X66200-2	ケーブル、光、OM4、MPO/MPO タイプ B
5m	X66200-5	ケーブル、光、OM4、MPO/MPO タイプ B
15m	X66200-15	ケーブル、光、OM4、MPO/MPO タイプ B
30m	X66200-30	ケーブル、光、OM4、MPO/MPO タイプ B

3.6.2 キャンパスリンク

ショートレンジ間で直接ファイバー接続を使用するキャンパスリンクは、ラック間 ISL 接続と類似しています。相違点は、マルチモードケーブルとショートレンジ光モジュールの使用に比べ、ロングレンジ光モジュールとシングルモードケーブルを使用することで、より長い距離を実現できることです。

ロングレンジモジュールを必要とするリンクの設計については、Cisco サポートマトリックスで指定されているスイッチモデルを確認し、Cisco 光モジュールのデータシートを参照して、距離と接続の仕様を決定します。

3.6.3 専用ファイバーリンク

専用ファイバーリンクは、近接する建屋間を接続するキャンパスネットワークでの使用が一般的です。専用ファイバーリンクを使用すると、多数のファイバー接続から少数のファイバーリンクに信号を多重化することもできます。これにより、使用率を最大化し、サイト間に必要なファイバー数を削減できます。光信号の多重化は波長分割多重方式 (WDM) と呼ばれ、粗波長分割多重方式 (CWDM) と高密度波長分割多重方式 (DWDM) の 2 種類があります。

CWDM は、DWDM よりも少数の波長を多重化できます。

CWDM は通常パッシブデバイスであり、光モジュールから受け取った光を多重化および分離して一つの信号に変換し、1 つのファイバーペアで送信できるようにします。光モジュールは波長が固定されており、この波長はチャネルとも呼ばれます。2 つの異なるファイバー信号を多重化するためには、各源信号は、異なる波長を使用する光モジュールから生成する必要があります。CWDM 光モジュールは Cisco から入手でき、8 つの異なる波長をサポートします。これにより、8 つのファイバー信号リンクを 1 つのファイバーリンクに多重化できます。CWDM マルチプレクサは受動側であり、信号を多重化または分離した光だけを格納します。DWDM と比較すると、この方式では多重化デバイスと関連する光モジュールのコストを削減できます。

DWDM は、CWDM デバイスと同様の方法で信号を統合します。CWDM との主な違いは、光モジュールが生成する信号がより精密であり、狭い信号幅でより細い分光幅の信号を送ることができ、信号間のスペースも CWDM より小さいことです。これにより、より多くの信号を組み合わせてサイト間ファイバーリンクで伝送することが可能になります。DWDM デバイスには、アクティブとパッシブの二種類があります。パッシブデバイスの場合は CWDM と同じように、光モジュールが送信した特定の波長またはチャネルを DWDM 装置が 1 つの信号にします。この信号は、サイト間のロングレンジファイバーケーブルで送信されます。

DWDM 装置はアクティブデバイスとしても使用できます。アクティブデバイスの場合、スイッチと DWDM 装置間の信号は標準光を使用し、DWDM 装置でサイト間通信のファイバーリンクに使用可能な波長の信号に変換します。

距離に関しては、可能な転送距離や、あわせるべきリンク特性の仕様は光モジュールに依存します。長距離の転送には、信号増幅器が必要になる場合があります。DWDM に適用する増幅器には、光増幅器などいくつかの種類があります。最適な構成設計については電気通信の専門家に相談することをお勧めします。

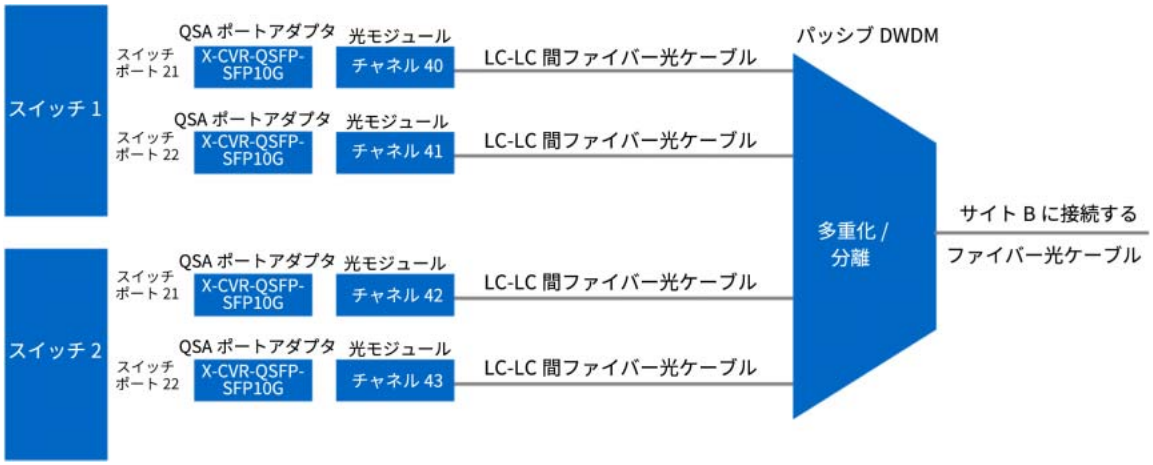
シスコは粗波長マルチプレクサおよび高密度マルチプレクサのいずれにも使用可能な 10Gb SFP+ モジュールを提供しています。いずれの WDM 方式でも、サイト間の単一のファイバーペアで複数の光信号を統合または多重化して、スイッチまたは装置へ転送する前に分離します。

図 3.4 は、パッシブ DWDM を使用した、特定のチャネルにおける光モジュールのマッピング例です。各サイトのすべての光モジュールは、固有のチャネルを使用します。対になるサイト B の、光モジュールの同じチャネルと一致させてください。

DWDM モジュールの使用例

- サイト A、スイッチ 1 のポート 21 と、サイト B、スイッチ 1 のポート 21 (チャネル 40 上の 2 つの光モジュールを使用)
- サイト A、スイッチ 2 のポート 21 と、サイト B、スイッチ 2 のポート 21 (チャネル 41 上の 2 つの光モジュールを使用)
- サイト A、スイッチ 1 のポート 22 と、サイト B、スイッチ 1 のポート 22 (チャネル 42 上の 2 つの光モジュールを使用)
- サイト A、スイッチ 2 のポート 22 と、サイト B、スイッチ 2 のポート 22 (チャネル 43 上の 2 つの光モジュールを使用)

図 3.4 サイト A で、10Gb 光モジュールと Quad Small Form-factor Pluggable Adapter(QSA) を使用したパッシブ DWDM の例



この例の 1 つ目の光モジュールはシスコの 10GBASE-DWDM SFP+ (部品番号 DWDM-SFP10G-45.32) で、ITU40 チャネルが 1545.32nm の波長 (100-GHz ITU grid) で動作するモジュールです。このサイトでの設定を完了するには、41、42、および 43 それぞれのチャネルに対応するモジュールがあと 3 つ必要です。サイト B には、まったく同じ構成の光モジュール、ポートアダプタ、パッシブ DWDM が搭載されます。

3.6.4 都市間リンク

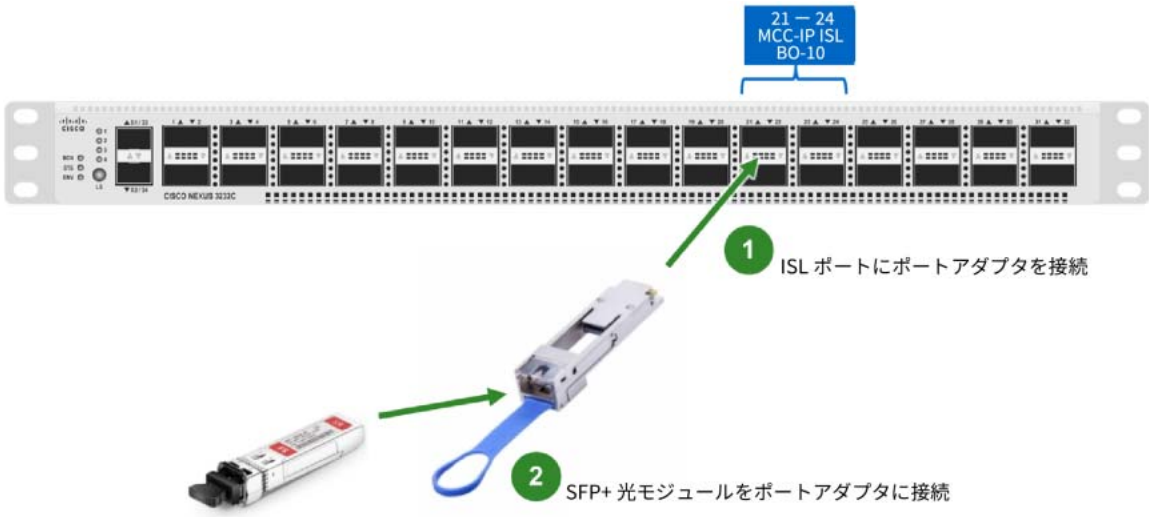
最長距離リンクには、主にアクティブ DWDM または Telco 回線が使用されます。スイッチから大部分の通信機器またはアクティブ DWDM 装置への接続は、データセンターまたはラック間構成で使用されるのと同じ光モジュールで行います。「3.6.1 近距離のラック間接続」(P.23) を参照してください。

備考

機器によっては、アクティブ DWDM で ISL トラフィックの暗号化が可能な場合があります。

都市間リンクの最適な構成設計については、電気通信の専門家に相談することをお勧めします。

図 3.5 10Gb ポートアダプタを使用した ISL



Cisco DWDM 光モジュールには、以下の 3 種類の構成があります。

- Cisco DWDM-SFP10G-XX .XX モジュール
  - DWDM 専用モジュールは、ITU 100GHz 波長 (チューニング不可能) を 40 種類サポート
- Cisco DWDM-SFP10G-C モジュール
  - DWDM チューナブルモジュールは、最大 80km までの ITU 50GHz 波長 (チューニング可能) を 96 種類サポート
- Cisco DWDM-SFP10G-C-S モジュール
  - チューニング可能なトランシーバモジュールは Ethernet のみ
  - DWDM チューナブルモジュールは、最大 70km までの ITU 50GHz 波長 (チューニング可能) を 96 種類サポート

3.6.5 ISL ネットワーキング : 専用リンク、共有レイヤ 2、またはレイヤ 3

MetroCluster IP 構成では、サイト間のレプリケーションに専用または共有 ISL を使用できます。専用 ISL 構成では、ISL スイッチに、MetroCluster IP スイッチの ISL ポートの速度と一致するネイティブ速度のポートが必要です。MetroCluster IP スイッチとお客様のネットワークスイッチ間の ISL の数と速度が一致している必要はありませんが、各 MetroCluster スイッチを中間スイッチに接続する ISL の数と速度は、両方の MetroCluster サイトで同じである必要があります。

共有レイヤ 2 ネットワーク構成では、複数の MetroCluster IP 構成で同一の ISL 用中間ネットワークを共有できます。ただし、適切な容量の確保と ISL の適切なサイジングを行い、データレプリケーションのために低レイテンシと高スループットを確保する必要があります。MetroCluster トラフィックが必要なサービスレベルを満たすように、MetroCluster IP スイッチ間のパスに沿ったクラスマップ、ポリシーマップ、QoS アクセスマップを構成する必要があります。

パケット損失は 0.01% 以下である必要があり、サポートされるジッタ値はラウンドトリップで 3 ミリ秒です。共有 ISL の理論上の最大スループットは、スイッチモデルとポートタイプによって異なります。MetroCluster IP スイッチを共有ネットワークに接続する ISL の数は、スイッチモデルとポートタイプによって異なります。

ONTAP 9.9.1 以降では、MetroCluster IP 構成に IP ルーティング (レイヤ 3) バックエンド接続を実装することもできます。MetroCluster バックエンドスイッチは、ルーターに直接、または間に他のスイッチを介して、ルーテッド IP ネットワークに接続されます。4 ノードの MetroCluster 構成のみがサポートされ、MetroCluster トラフィックの動的ルーティングはサポートされません。MetroCluster サイトごとに 2 つのサブネット (各ネットワークに 1 つずつ) が必要です。

つまり、MetroCluster IP の導入に適したネットワークを確保するためには、特定の構成に関する要件と推奨事項を慎重に確認する必要があります。

4. 運用と管理

MetroCluster の運用と管理では、MetroCluster の稼働状態のチェックまたは検証と監視を行います。ほとんどの操作については、ONTAP のマニュアルに MetroCluster を含むストレージの管理手順が記載されています。

4.1 ハイアベイラビリティ (HA) およびディザスタリカバリ (DR)

MetroCluster には、ローカルまたはクラスタ (サイト) レベルの障害発生時に継続的なデータの可用性と効率的なシステム機能を確保するために、2 つの主要な復旧メカニズムがあります。これらのメカニズムには、ローカルレベルの復旧 (テイクオーバーおよびギブバック) とクラスタレベルの復旧 (スイッチオーバーおよびスイッチバック) があります。

ローカルで発生した障害や、ONTAP アップグレードなどの無停止運用は、HA パートナーによって処理されます。MetroCluster では、標準の ONTAP 用語を使用して、HA の運用を説明します。

4.1.1 ローカルレベルの復旧 (HA)

MetroCluster IP のハイアベイラビリティ (HA) ペアに関連するテイクオーバープロセスとギブバックプロセスにより、フォールトトレランスと無停止運用を実現します。

- テイクオーバーとは、HA ペアのノード (ローカルノード) が、データ処理のエラーまたは停止中にパートナー (パートナーノード) のストレージの制御を引き継ぐプロセスです。ソフトウェアまたはシステムの障害、停電、ハートビートメッセージの中断、手動による開始など、さまざまな状況でテイクオーバーが開始されることがあります。ローカルノードは、パートナーの障害ステータスを検出し、データ処理を引き継ぐことによってデータの可用性とシステム機能を維持します。
- ギブバックとは、問題から回復した後、またはメンテナンスを完了した後に、パートナーノードがストレージの制御を再開するプロセスのことです。ローカルノードは、ストレージをパートナーノードに返し、通常の操作を復元します。テイクオーバーは、さまざまな状況で自動的に発生することもありますが、storage failover takeover コマンドを使用して手動で開始することもできます。

4.1.2 サイトレベルの復旧 (DR)

スイッチオーバーおよびスイッチバックは、MetroCluster IP のディザスタリカバリ (DR) プロセスに関連して、あるクラスタサイトが別のクラスタサイトのタスクを引き継ぐことを可能にし、メンテナンスや災害からの復旧を容易にします。これらの操作は、ONTAP System Manager でサポートされています。

- スイッチオーバーは、クラスタサイト (サイト A) が別のクラスタサイト (サイト B) からストレージ制御とクライアントアクセスを引き継ぐプロセスです。このプロセスにより、テスト、メンテナンス、またはサイト障害時に無停止での運用が可能になります。計画的スイッチオーバー (NSO) は、ディザスタリカバリのテストまたは計画されたメンテナンスに使用できます。計画外スイッチオーバー (USO) は、いずれかのサイトに影響を与える災害に対応して発生します。System Manager はスイッチオーバーの実行可能性を判断し、それに応じてワークロードを調整します。

スイッチオーバーの後、System Manager は MetroCluster IP 構成の修復プロセスを 2 つのフェーズで完了します。最初のフェーズでは、ミラーリングされたプレックスの再同期と、ルートアグリゲートのディザスタサイトへのスイッチバックを行います。次のフェーズでは、スイッチバックプロセスのためにサイトを準備します。

- スイッチバックとは、サイト B でメンテナンスと修復が完了した後、サイト A からサイト B にストレージとクライアントアクセスの制御を戻す操作です。スイッチバックを成功させるには、サイト B でホームノードとストレージシェルフに電源が入ること、修復フェーズが問題なく完了していること、サイト A のすべてのアグリゲートがミラーステータスになっていること、スイッチバック操作を実行する前に実施したすべての設定変更が完了していることなど、特定の条件を満たす必要があります。

修復は、MetroCluster IP 構成での計画的スイッチオーバーの操作中に自動化されます。計画外スイッチオーバー発生後の自動修復がサポートされています。これにより、metrocluster heal コマンドを発行する必要がなくなります。

4.2 クォーラムウィットネス

ストレージクラスタでは、高可用性を維持し、データの破損を防ぎ、フェイルオーバーとリカバリを容易にするために、クォーラムウィットネスを確認することが不可欠です。クォーラムウィットネスはタイブレーカーとして機能し、大多数のノードが相互に通信できるようにすることで、クラスタノードが独立して動作したことでデータの不整合が生じるスプリットブレインを回避することができます。MetroCluster IP は、クォーラムウィットネスとして MetroCluster Tiebreaker または ONTAP Mediator のいずれかをサポートします。

4.2.1 ONTAP Mediator ソフトウェア

ONTAP には、MetroCluster IP 用の新しい ONTAP Mediator ソフトウェアソリューションのリリースが含まれます。このソフトウェアは第 3 の障害ゾーンに常駐し、MetroCluster IP が自動計画外スイッチオーバー (AUSO) を実行できるようにします。さらに、2 つのサイト間でデータのミラーリングに障害が発生した場合に、AUSO を無効化する機能があります。この機能はサイト間でリンクダウンが発生した場合に自動スイッチオーバーを抑制するため、管理者は手動でのスイッチオーバーが適切かどうかを判断できます。

新しい ONTAP Mediator サービスは、一方のサイトにある MetroCluster IP のノードの 1 つで構成されます。ONTAP は自動ですべてのノードと 2 つ目のクラスタを構成します。初期リリースでは、ONTAP 9.7 以降で動作する MetroCluster IP と Mediator ソフトウェア (バージョン 1.0 以降) が必要です。

ONTAP では、ONTAP Mediator の構成用に新しいコマンドを提供しています。

```
metrocluster configuration-settings mediator add -mediator-address <mediator-ip>
metrocluster configuration-settings mediator remove
```

新しく Mediator を追加する場合、Mediator の有効な認証情報の入力が要求されます。認証情報は導入時に設定されます。この設定は Mediator システムにログインし、以下のコマンドを実行することで変更できます。

- Mediator のアカウント名を変更する場合

```
/opt/netapp/lib/ontap_mediator/tools/mediator_change_user
```

- Mediator のパスワードを変更する場合

```
/opt/netapp/lib/ontap_mediator/tools/mediator_change_password
```

- Mediator の状態を確認する場合

```
systemctl status ontap_mediator
```

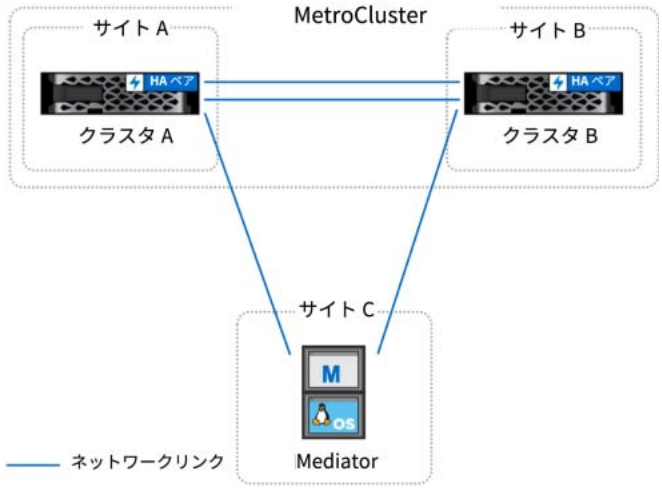
- ログ収集ツールを実行する場合

```
/opt/netapp/lib/ontap_mediator/tools/mediator_generate_support_bundle
```

備考

同一 MetroCluster に Tiebreaker と ONTAP Mediator の両方を使用する構成管理は未サポートです。
MetroCluster の構成管理にはどちらか一方の製品を使用してください。

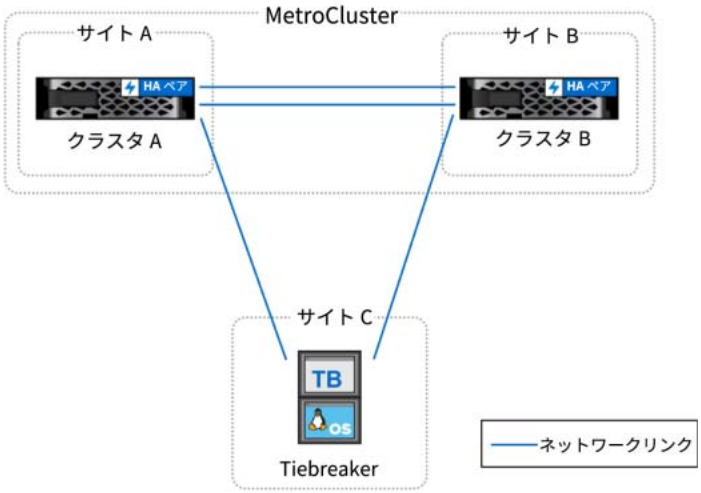
図 4.1 MetroCluster Mediator サイト



4.2.2 Tiebreaker ソフトウェア

MetroCluster Tiebreaker マネージメントパックは MetroCluster システムを監視し、サイト障害および ISL 故障を検出します。Tiebreaker ソフトウェアは Linux ホスト (通常は仮想マシン) にインストールされます。仮想マシンは、MetroCluster ソリューションの両クラスタの障害ドメインとは別の第 3 の障害ドメインに配置されます。

図 4.2 MetroCluster Tiebreaker サイト



Tiebreaker ソフトウェアは、ノード管理 LIF およびクラスタ管理 LIF を複数のパスで冗長接続し、各コントローラを監視します。

備考

同一 MetroCluster に Tiebreaker と ONTAP Mediator の両方を使用する構成管理は未サポートです。
MetroCluster の構成管理にはどちらか一方の製品を使用してください。

4.2.2.1 Tiebreaker サイト障害時の症状

サイトで障害が発生すると、Tiebreaker ソフトウェアから接続できるクラスタがどちらか一方だけの場合は、Tiebreaker ソフトウェアが警告を出していなくても、接続可能なクラスタからパートナークラスタへの通信もできなくなっているはずです。クラスタ同士の通信が可能な場合、Tiebreaker は、Tiebreaker ソフトウェアと接続できないクラスタ間のネットワークで接続障害が発生したと判断します。

図 4.3 Tiebreaker サイトのリンク障害

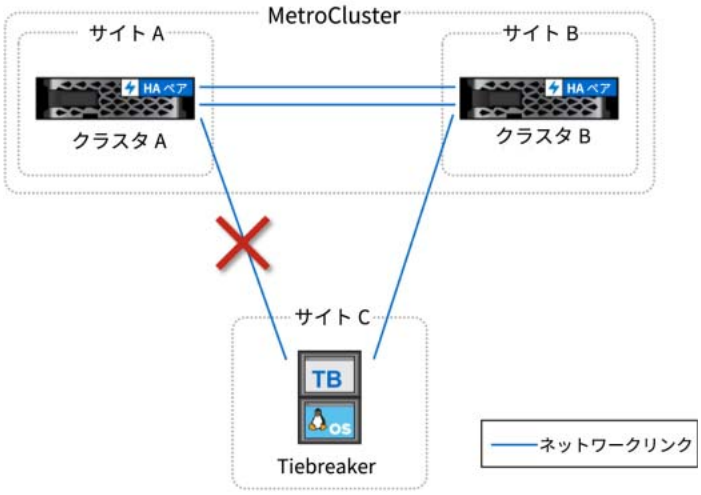
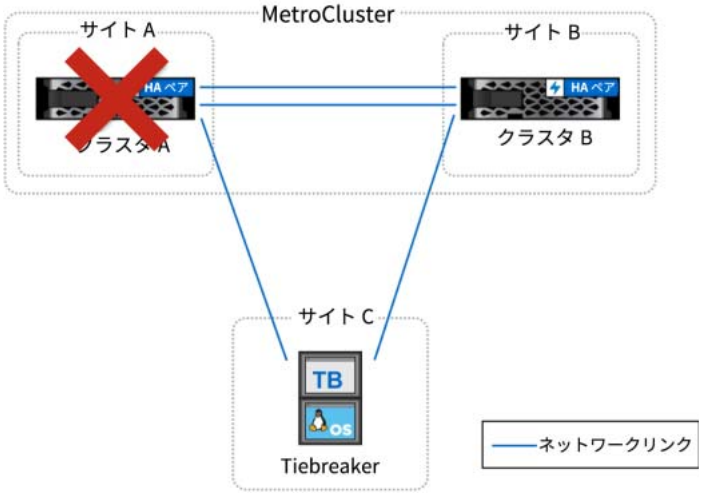


図 4.4 Tiebreaker サイトの障害



5. 相互運用

MetroCluster は、ONTAP の最も一般的な機能をサポートします。ただし、SnapMirror Synchronous などいくつかの ONTAP 機能は、現在サポートされていません。MetroCluster のサポートに関するガイダンスについては、ONTAP の機能に関するドキュメントを参照してください。

5.1 SnapMirror Asynchronous

SnapMirror Asynchronous は、MetroCluster でサポートされています。SnapMirror Asynchronous 保護は、ソースボリュームとデスティネーションボリューム間からターシャリ (三次) クラスタへの、計画的なブロックレプリケーションおよびミラー保護を提供します。MetroCluster システムは、SnapMirror レプリケーション関係のソースおよびデスティネーションとして構築することができます。詳細については、マニュアルサイトに掲載の以下のリソースを参照してください。

- ETERNUS AX/HX series Hybrid Arrays ONTAP 9 用 SnapMirror 構成およびベストプラクティスガイド

5.2 ONTAP FlexGroup ボリューム

FlexGroup ボリュームは、MetroCluster でサポートされています。FlexGroup ボリュームは自動負荷分散と拡張性に加えて高いパフォーマンスを提供するスケールアウトボリュームです。詳細については、マニュアルサイトに掲載の以下のリソースを参照してください。

- ETERNUS AX/HX series ONTAP FlexGroup ボリュームベストプラクティス / 実装ガイド

5.3 FlexCache

FlexCache テクノロジーは、MetroCluster IP でサポートされています。FlexCache は、ファイル分散の簡易化、WAN レイテンシの削減、および WAN 帯域幅コストの低減を実施するリモートキャッシュ機能です。この機能は、ファイルやボリューム全体ではなく、データセンターのローカル内または地理的に離れたリモートサイトにある、読み取り頻度の高いデータのみをキャッシュします。詳細については、マニュアルサイトに掲載の以下のリソースを参照してください。

- ETERNUS AX/HX series ONTAP での FlexCache

5.4 FabricPool

FabricPool は、MetroCluster IP でサポートされています。FabricPool は、オールフラッシュ (SSD) アグリゲートをパフォーマンス層として使用し、パブリッククラウドサービス内のオブジェクトストアをクラウド層として使用する、ONTAP のハイブリッドストレージソリューションです。詳細については、マニュアルサイトに掲載の以下のリソースを参照してください :

- ETERNUS AX/HX series FabricPool のベストプラクティス

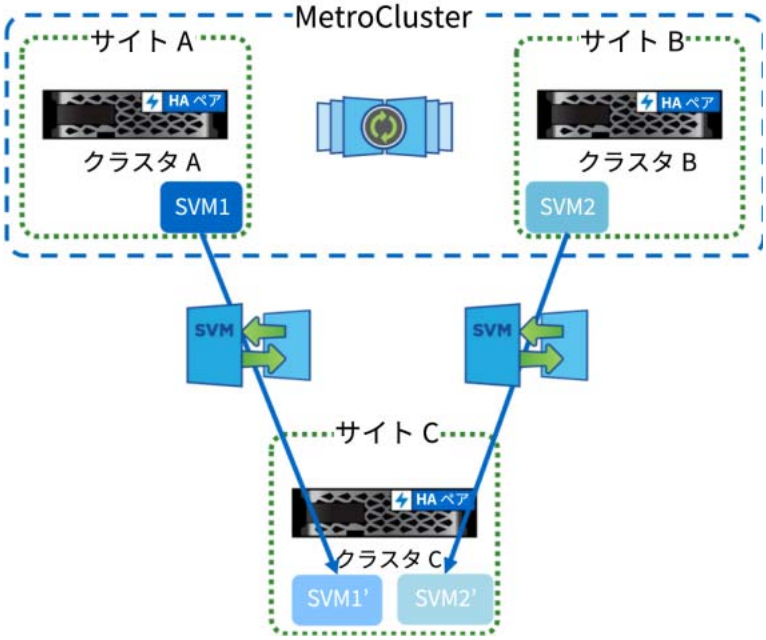
5.5 SVM ディザスタリカバリ (DR)

SVM DR は、MetroCluster IP と互換性があります。SVM DR は、SVM のデータ保護とディザスタリカバリ機能を提供するように設計されたソリューションです。この機能を使用すると、管理者はプライマリ SVM のレプリカを作成して、データと構成の最新のコピーをリモートサイトで使用できるようにし、災害の影響を軽減できます。これは、SnapMirror Asynchronous テクノロジーを利用して、プライマリサイト上のソース SVM と DR サイト上のデスティネーション SVM との間に関係を確立することによって実現されます。災害発生時には、セカンダリ SVM をアクティブにしてデータにアクセスできるため、ダウンタイムを最小限に抑えることができます。

備考

SVM は、MetroCluster 構成の 1 つのサイトからのみ保護できます。ONTAP 9.11.1 からは、MetroCluster の両方のサイトから SVM を保護する機能が導入されています。

図 5.1 SVM ディザスタリカバリ



ETERNUS AX series オールフラッシュアレイ , ETERNUS AC 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 IP ソリューションのアーキテクチャと設計

P3AG-5222-03Z0

発行日　2025 年 3 月
発行責任　エフサステクノロジーズ株式会社

Fsas Technologies