テクニカル レポート

Oracle on MetroCluster

統合データプロテクション、ディザスタリカバリ、高可用性

NetApp
Jeffrey Steiner
2021年3月| TR-4592

概要

このドキュメントでは、NetApp® MetroCluster™ 同期レプリケーション機能でOracleデータベースを操作する場合のベストプラクティス、テスト手順、およびその他の考慮事項について説明します。

目次

表一覧

図一覧

OracleデータベースとNetApp MetroCluster

NetApp MetroClusterは、ミッションクリティカルなOracleワークロードに可用性とデータ損失ゼロの解決策を提供します。さらに、MetroClusterなどの統合ソリューションにより、今日の複雑なスケールアウト型Oracleデータベース、アプリケーション、仮想化インフラが簡素化されます。MetroClusterは、複数の外部データ保護製品や戦略を、1つのシンプルな中央集中型ストレージアレイに置き換えます。単一のクラスタストレージシステム内で、統合されたバックアップ、リカバリ、ディザスタリカバリ、高可用性（HA）を実現します。

MetroClusterテクノロジ

SyncMirrorによるデータ保護

最も単純な同期レプリケーションとは、変更があった場合、ミラーされたストレージの両側に確認応答を行う前に変更を加える必要があることを意味します。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しているときに、同期ミラーリングされたストレージのRedoログにデータが書き込まれるとします。ストレージ システムは、両方のサイトの不揮発性メディアに書き込みがコミットされるまで、書き込みを認識できません。これが終わると、データ損失のリスクなく安全に処理を続けることができます。

同期レプリケーションテクノロジの使用は、同期レプリケーション解決策を設計および管理するための最初のステップです。最も重要な考慮事項は、計画的および計画外のさまざまな障害シナリオで何が発生するかを理解することです。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。お客様から、Recovery Point Objective（RPO；目標復旧時点）がゼロ（データ損失ゼロ）の解決策を求められた場合は、すべての障害シナリオを検討する必要があります。具体的には、サイト間の接続が失われたためにレプリケーションが不可能になった場合、どのような結果が予想されますか。

MetroClusterヲシヨウシタHA

MetroClusterレプリケーションは、同期モードに効率的に切り替えられるように設計されたNetApp SyncMirror®テクノロジに基づいています。この機能は、同期レプリケーションを必要とする一方で、データサービスに高可用性も必要とするお客様の要件を満たします。たとえば、リモートサイトへの接続が切断されている場合は、ストレージシステムを非レプリケート状態で運用し続けることを推奨します。

多くの同期レプリケーションソリューションは、同期モードでしか動作できません。このタイプのall-or-nothingレプリケーションは、Dominoモードと呼ばれることがあります。このようなストレージシステムは、データのローカルコピーとリモートコピーが非同期になるのではなく、データの提供を停止します。レプリケーションが強制的に解除されると、再同期に非常に時間がかかることがあります。また、ミラーリングの再確立中にデータが完全に失われる可能性もあります。

SyncMirrorは、リモートサイトに到達できない場合はシームレスに同期モードを終了したり、接続がリストアされたときにRPO=0状態に迅速に再同期したりできます。再同期中にリモートサイトにある古いデータコピーを使用可能な状態で保持することもできるため、データのローカルコピーとリモートコピーが常に存在します。

Dominoモードが必要な場合は、Oracle DataGuardやホスト側のディスクミラーリングのタイムアウトの延長など、MetroCluster以外のオプションもあります。追加情報とオプションについては、NetApp®またはパートナーアカウントチームにお問い合わせください。

データベースのデータ保護

データベースのデータ保護アーキテクチャは、ビジネス要件によって定義する必要があります。これらの要件には、リカバリの速度、許容される最大データ損失、バックアップの保持のニーズが含まれます。データ保護計画では、データの保持とリストアに関するさまざまな規制要件も考慮する必要があります。最後に、さまざまなデータリカバリシナリオも考慮する必要があります。これらのシナリオには、ユーザやアプリケーションのエラー

に起因する一般的で予測可能なリカバリから、サイトの完全な停止を含むディザスタリカバリのシナリオまで、さまざまなシナリオがあります。

データ保護ポリシーとリカバリポリシーのわずかな変更は、ストレージ、バックアップ、リカバリのアーキテクチャ全体に大きな影響を与える可能性があります。データ保護アーキテクチャが複雑にならないように、設計作業を開始する前に標準を定義して文書化することが重要です。不要な機能や保護レベルは、不要なコストや管理オーバーヘッドにつながります。また、最初に見落とされた要件は、プロジェクトを間違った方向に進めたり、直前の設計変更を必要としたりする可能性があります。

目標復旧時間

Recovery Time Objective（RTO；目標復旧時間）は、サービスのリカバリに許容される最大時間を定義します。たとえば、人事データベースのRTOが24時間であるとします。これは、業務時間中にこのデータにアクセスできなくなるのは不便であっても、業務を継続できるためです。一方、銀行の総勘定元帳をサポートするデータベースでは、数分または数秒でRTOを測定できます。RTOをゼロにすることはできません。これは、実際のサービス停止と、ネットワークパケットの損失などの日常的なイベントを区別する方法が必要であるためです。ただし、一般的な要件はRTOがほぼゼロです。

目標復旧時点

RPOは、最大許容データ損失を定義します。データベースのコンテキストでは、通常、RPOは、特定の状況で失われる可能性のあるログデータの量です。製品のバグやユーザエラーによってデータベースが破損した一般的なリカバリシナリオでは、RPOはゼロ、つまりデータ損失がないことを意味します。リカバリ手順では、データベースファイルの以前のコピーをリストアし、ログファイルを再生して、データベースを希望する時点の状態にします。この処理に必要なログファイルは元の場所にすでに存在している必要があります。

通常とは異なる状況では、ログデータが失われる可能性があります。たとえば、偶発的または悪意のある `rm -rf *` データベースファイルがあると、すべてのデータが削除される可能性があります。唯一のオプションは、ログファイルを含むバックアップからリストアすることです。必然的に、一部のデータが失われます。従来のバックアップ環境でRPOを向上させる唯一の方法は、ログデータのバックアップを繰り返し実行することです。しかし、このアプローチには、データが絶えず移動し、バックアップシステムを継続的に実行されるサービスとして維持することが困難であるため、制限があります。高度なストレージシステムのメリットの1つは、偶発的または悪意のあるファイルの破損からデータを保護し、データを移動せずにRPOを向上できることです。

ディザスタ リカバリ

ディザスタリカバリには、物理的な災害が発生した場合にリカバリするために必要なITアーキテクチャ、ポリシー、および手順が含まれます。そのような災害には、洪水、火災、または悪意、過失、または単純なエラーによって引き起こされた製造された災害が含まれます。

ディザスタリカバリは、単なるリカバリ手順ではありません。これは、さまざまなリスクを特定し、データリカバリとサービス継続性の要件を定義し、適切なアーキテクチャと関連手順を提供する完全なプロセスです。

データ保護の要件を確立するには、一般的なRPOとRTOの要件と、ディザスタリカバリに必要なRPOとRTOの要件を区別することが重要です。一部のデータベース環境では、比較的正常なユーザエラーからデータセンターの破壊に至るまで、データ損失の状況に対して、RPOゼロとRTOほぼゼロを達成する必要があります。ただし、これらの高レベルの保護にはコストと管理上の影響があります。

一般に、ディザスタ以外のデータリカバリ要件は、次の2つの理由で厳しくする必要があります。まず、データベースに損害を与えるアプリケーションのバグやユーザエラーは、ほとんど避けられないほど予測可能です。2つ目は、ストレージシステムが破損していない場合でも、RPOをゼロにし、RTOを短縮できるバックアップ戦略を設計することです。容易に修復できる重大なリスクに対処しない理由はありません。そのため、ローカルリカバリのRPOとRTOの目標を積極的に設定する必要があります。

ディザスタリカバリのRTOとRPOの要件はさまざまです。これは、災害の可能性と、関連するデータの損

失やビジネスの中断の結果に基づいて判断されます。RPOとRTOの要件は、一般的な原則ではなく、実際のビジネスニーズに基づいている必要があります。論理的および物理的な複数の災害シナリオを考慮する必要があります。

論理的災害

論理的災害には、ユーザによるデータ破損、アプリケーションやOSのバグ、ソフトウェアの誤動作などがあります。論理的災害には、ウイルスやワームによる悪意のある攻撃や、外部によるアプリケーションの脆弱性の悪用も含まれます。この場合、物理インフラは破損していませんが、基盤となるデータは無効になります。

ランサムウェアと呼ばれる論理災害のタイプはますます一般的になりつつあり、攻撃ベクトルを使用してデータを暗号化します。暗号化はデータを損傷することはありませんが、サードパーティに支払いが行われるまで使用できなくなります。ランサムウェアのハッキングの標的にされる企業はますます増えています。

物理的災害

物理的災害には、インフラストラクチャのコンポーネントの障害がその冗長性機能を超え、データの損失やサービスの長期的な損失につながることが含まれます。たとえば、RAID保護はドライブレベルの冗長性を提供し、Host Bus Adapter（HBA；ホストバスアダプタ）を使用するとFibre Channel（FC；ファイバチャネル）ポートとFCケーブルの冗長性が提供されます。このようなコンポーネントのハードウェア障害は予測可能であり、可用性には影響しません。

データベース環境では、サイト全体のインフラを冗長コンポーネントで保護し、予測可能な唯一の物理的災害シナリオでサイトが完全に失われた時点まで保護できます。ディザスタリカバリ計画は、サイト間レプリケーションによって異なります。

同期および非同期のデータ保護

理想的な環境では、地理的に分散したサイト間ですべてのデータを同期的にレプリケートできます。このアプローチは、次のようないくつかの理由により、必ずしも実現可能ではありません。

- 同期レプリケーションでは、アプリケーションまたはデータベースを処理する前にすべての変更を両方の場所にレプリケートする必要があるため、書き込みレイテンシが避けられません。このようなパフォーマンスへの影響は許容できず、同期ミラーリングの使用が除外されることがよくあります。
- 100% SSDストレージの採用が増加しているため、書き込みレイテンシの増加に気付く可能性が高くなります。これは、想定されるパフォーマンスには数十万のIOPSと1ミリ秒未満のレイテンシが含まれているためです。100% SSDを使用するメリットを最大限に引き出すには、ディザスタリカバリ戦略を見直す必要があります。
- データセットはバイト単位で増え続けているため、同期レプリケーションを維持するのに十分な帯域幅を保証するという課題が生じています。
- データセットも複雑化し、大規模な同期レプリケーションの管理が困難になっています。
- クラウドベースの戦略では、多くの場合、レプリケーションの距離とレイテンシが長くなり、同期ミラーリングの使用がさらに困難になります。

NetAppは、最も厳しいデータリカバリ要件に対応する同期レプリケーションと、データベースのパフォーマンスと柔軟性を向上させる非同期ソリューションの両方を備えたソリューションを提供します。さらに、NetAppテクノロジは、Oracle DataGuardやSQL Server AlwaysOnなど、多くのサードパーティ製レプリケーションソリューションとシームレスに統合されます。

保持時間

データ保護戦略の最後の側面はデータ保持期間です。データ保持期間は次のように大きく異なります。

- 一般的な要件は、プライマリサイトに夜間バックアップを14日間、セカンダリサイトにバックアップを90日間保存することです。
- 多くのお客様が、さまざまなメディアに保存されたスタンドアロンの四半期ごとのアーカイブを作成します。
- 定期的に更新されるデータベースでは、履歴データは不要であり、バックアップは数日間だけ保持する必要があります。

規制要件では、365日の期間内に任意のトランザクションのポイントまでリカバリできることが規定されている場合があります。

NetApp ONTAPデータ保護の基礎

NetApp Snapshotコピーによるデータ保護

NetApp ONTAP® データ保護ソフトウェアの基盤は、NetApp Snapshot® テクノロジです。主な値は次のとおりです。

- **シンプル**：Snapshotコピーは、特定の時点のデータコンテナの内容の読み取り専用コピーです。
- **効率性**：Snapshotコピーの作成時にスペースは必要ありません。スペースが消費されるのは、データが変更されたときだけです。
- **管理性**：Snapshotコピーはストレージ OSに標準搭載されているため、Snapshotコピーに基づくバックアップ戦略を簡単に設定および管理できます。ストレージシステムの電源がオンになっていれば、バックアップを作成できます。
- **拡張性**ファイルとLUNの単一コンテナのバックアップを最大255個保持できます。複雑なデータセットの場合、データの複数のコンテナを、整合性のある単一のSnapshotコピーセットで保護できます。
- ボリュームに250個のSnapshotコピーが含まれているかどうかに関係なく、パフォーマンスには影響しません。

そのため、ONTAPで実行されているデータベースの保護はシンプルで拡張性に優れています。データベースバックアップでは、データの移動は必要ありません。したがって、バックアップ戦略は、ネットワーク転送速度、多数のテープドライブ、ディスクステージング領域の制限ではなく、ビジネスのニーズに合わせて調整できます。

ONTAP SnapRestoreによるデータのリストア

ONTAP内のSnapshotコピーからのデータの高速リストアは、NetApp SnapRestore® テクノロジによって実現されます。主な値は次のとおりです。

- 個々のファイルやLUNは、2TBのLUNでも4KBのファイルでも、数秒でリストアできます。
- LUNやファイルのコンテナ（NetApp FlexVol® ボリューム）全体を、10GBまたは100TBのデータであれ、数秒でリストアできます。

重要なデータベースが停止すると、重要なビジネスの運用が停止します。テープが破損する可能性があり、ディスクベースのバックアップからリストアする場合でも、ネットワーク経由での転送に時間がかかることがあります。SnapRestoreでは、データベースをほぼ瞬時にリストアできるため、このような問題を回避できます。ペタバイト規模のデータベースでも、わずか数分で完全にリストアできます。

データレプリケーションとディザスタリカバリ

ほぼすべてのデータベースでデータレプリケーションが必要です。最も基本的なレベルでは、レプリカはオフサイトに保管されたテープ上のコピー、またはスタンバイデータベースへのデータベースレベルのレプリケーションです。ディザスタリカバリとは、サービスが壊滅的に失われた場合に、これらのレプリカコピーを使用してサービスをオンラインにすることです。

ONTAPは、ストレージアレイ内のさまざまな要件に対応するための複数のレプリケーションオプションを提供し、あらゆるニーズに対応します。これらのオプションは、リモートサイトへのバックアップの単純なレプリケーションから、同じプラットフォームでディザスタリカバリとHAの両方を実現する完全に自動化された同期解決策まで、さまざまです。

データベースに適用される主なONTAPレプリケーション機能は、NetApp SnapMirror® とNetApp SyncMirrorです。これらの機能はアドオン製品ではなく、ONTAPに完全に統合され、ライセンスキーを追加するだけでアクティブ化されます。また、ストレージレベルのレプリケーションだけが選択肢ではありません。Oracle DataGuardやMicrosoft SQL Server AlwaysOnなどのデータベースレベルのレプリケーションを、ONTAPに基づくデータ保護戦略に統合することもできます。

適切な選択は、特定のレプリケーション、リカバリ、保持の要件によって異なります。

ONTAP SnapMirror

SnapMirrorは、NetAppの非同期レプリケーション機能です。データベースやその関連アプリケーションなど、複雑で動的な大規模データセットの保護に最適です。主な値は次のとおりです。

- **管理性**：SnapMirrorは、ストレージソフトウェアに標準で組み込まれているため、設定と管理が容易です。アドオン製品は必要ありません。レプリケーション関係は数分で確立でき、ストレージシステム上で直接管理できます。
- **シンプル**：レプリケーションは、単一の整合グループとしてレプリケートされるLUNまたはファイルのコンテナであるFlexVolボリュームに基づいています。
- **効率性**：最初のレプリケーション関係が確立されると、変更内容のみがレプリケートされます。また、重複排除や圧縮などの効率化機能が維持されるため、リモートサイトに転送するデータ量がさらに削減されます。
- **柔軟性**：ミラーを一時的に解除してディザスタリカバリ手順をテストできます。その後、完全な再ミラーリングを必要とせずに、ミラーリングを簡単に再確立できます。ミラーを同期状態に戻すには、変更されたデータのみを適用する必要があります。ミラーリングを反転することで、災害が終了して元のサイトが稼働状態に戻ったあとに迅速に再同期することもできます。最後に、レプリケートされたデータの読み取り/書き込みクローンをテストと開発に使用できます。

MetroClusterとSyncMirror

ONTAPの同期レプリケーションはSyncMirrorによって提供されます。最も単純なレイヤでは、SyncMirrorは2つの異なる場所にRAID保護データの完全なセットを作成します。データセンター内の隣接する部屋に配置することも、数キロメートル離れた場所に配置することもできます。

SyncMirrorはONTAPと完全に統合されており、RAIDレベルのすぐ上で動作します。そのため、Snapshotコピー、SnapRestore、NetApp FlexClone®など、ONTAPの通常の機能はすべてシームレスに動作します。それはまだONTAPです。同期データミラーリングの追加レイヤが含まれているだけです。

SyncMirrorデータを管理する一連のONTAPコントローラをNetApp MetroClusterと呼び、多くの設定を使用できます。MetroClusterの主な目的は、さまざまな一般的な障害やディザスタリカバリのシナリオで、同期ミラーリングされたデータへのHAアクセスを提供することです。

MetroClusterとSyncMirrorを使用したデータ保護の主な価値は次のとおりです。

- 通常運用時には、SyncMirrorは複数のサイト間の同期ミラーリングを保証します。書き込み処理は、両方のサイトの不揮発性メディアに存在するまで確認応答されません。

- サイト間の接続に障害が発生すると、SyncMirrorは自動的に非同期モードに切り替わり、接続が回復するまでプライマリサイトがデータを提供し続けます。リストア時には、プライマリサイトに蓄積された変更を効率的に更新することで、迅速な再同期を実現します。完全な再初期化は必要ありません。

SnapMirrorは、SyncMirrorベースのシステムとも完全に互換性があります。たとえば、プライマリデータベースが2つの地理的なサイトに分散したMetroClusterクラスタで実行されているとします。このデータベースは、長期アーカイブやDevOps環境でのクローン作成のために、バックアップを第3のサイトにレプリケートすることもできます。

MetroCluster物理アーキテクチャ

MetroClusterクラスタには、次の3つの構成があります。

- IPセツソクノHAヘア
- FCセツソクノHAヘア
- FC接続のシングルコントローラ

注： 「接続」という用語は、クロスサイトレプリケーションに使用されるクラスタ接続を指します。ホストプロトコルを指しているわけではありません。MetroCluster構成では、クラスタ間通信に使用される接続のタイプに関係なく、ホスト側のプロトコルはすべて通常どおりサポートされます。

MetroCluster IP

IP接続を使用するMetroClusterテクノロジは、各サイトでHAペアを使用して構成されます。本書の執筆時点では、各サイトに1つのHAペアしか配置されていません。SAN接続MetroClusterにのみ、シングルコントローラオプションが含まれています。

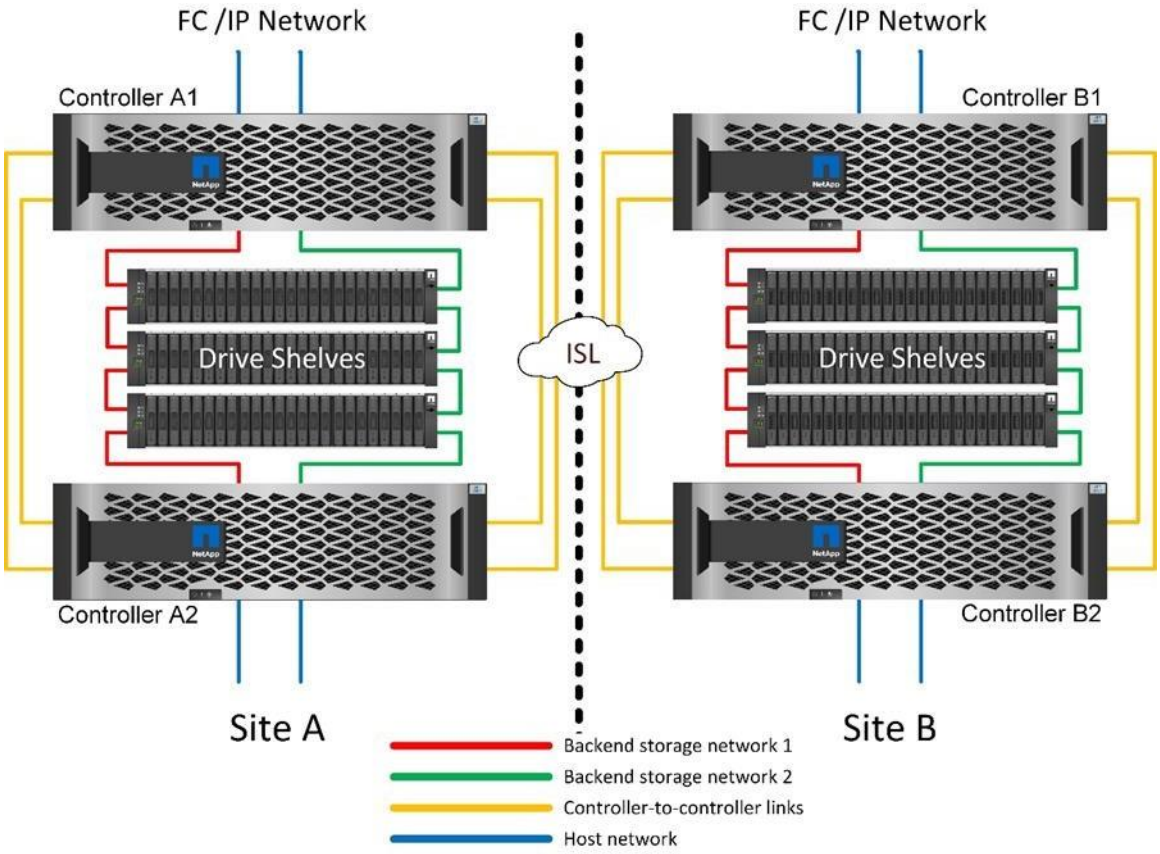
ほとんどのお客様は、インフラストラクチャの要件がシンプルであるため、IP接続を選択しています。これまでは、ダークファイバやFCスイッチを使用した、サイト間での高速接続はプロビジョニングが容易でしたが、今日では、高速で低レイテンシのIP回線がより容易に利用できるようになりました。

サイト間接続はコントローラのみであるため、アーキテクチャもシンプルです。FC SAN接続MetroClusterテクノロジでは、コントローラが反対側のサイトのドライブに直接書き込みを行うため、追加のSAN接続、スイッチ、およびブリッジが必要になります。一方、IP構成のコントローラは、コントローラを使用して反対側のドライブに書き込みます。

図1 に、MetroCluster IPの基本アーキテクチャを示します。

追加情報については、ONTAPの公式ドキュメントおよび MetroCluster IP解決策のアーキテクチャと設計を参照してください。

図1）MetroCluster IPの基本アーキテクチャ

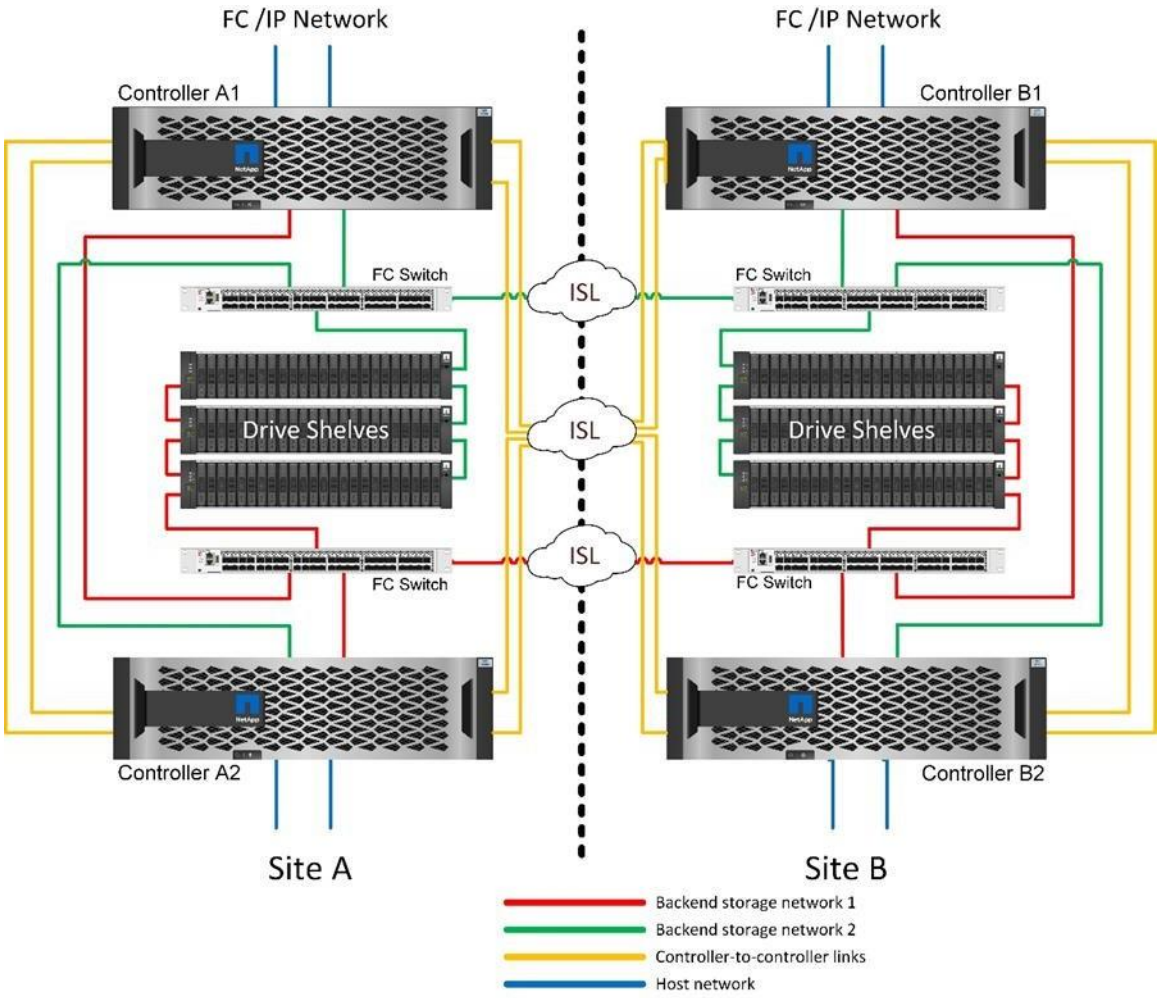


HAペアFC SAN接続MetroCluster

HAペアMetroCluster構成では、各サイトに2ノードを使用します。この設定オプションを使用すると、2ノードオプションに比べて複雑さとコストが増加しますが、サイト内の冗長性という重要なメリットがあります。単純なコントローラ障害では、WAN経由のデータアクセスは必要ありません。データアクセスは、代替ローカルコントローラを介してローカルのままです。

基本的な設計を図2に示します。

図2）HAペアFC SAN接続MetroClusterの基本アーキテクチャ



マルチサイトインフラの中には、HA構成向けに設計されたものではなく、プライマリサイトやディザスタリカバリサイトとして使用されるものもあります。この状況では、次の理由からHAペアのMetroClusterオプションが推奨されます。

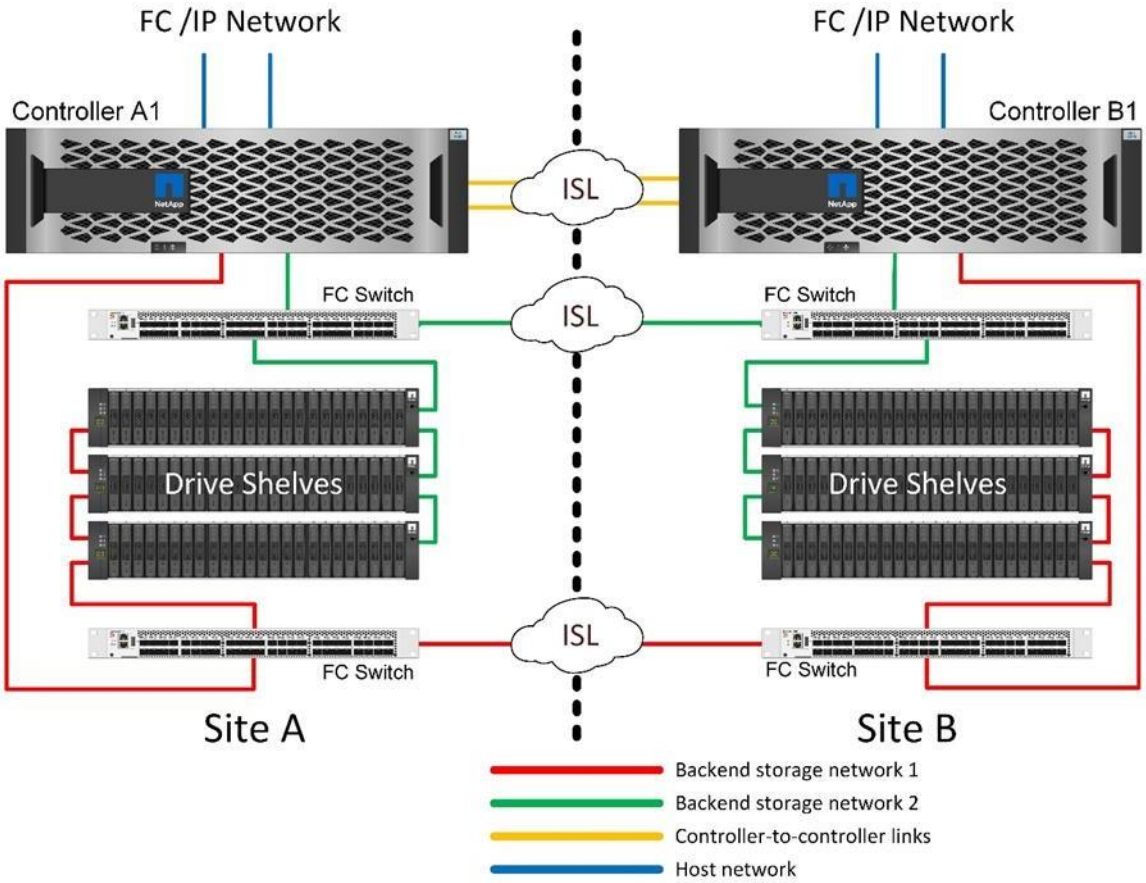
- 2ノードMetroClusterクラスタはHAシステムですが、コントローラに予期しない障害が発生した場合や計画的メンテナンスを行う場合は、反対側のサイトでデータサービスをオンラインにする必要があります。サイト間のネットワーク接続が必要な帯域幅をサポートできない場合は、パフォーマンスが低下します。唯一の選択肢は、さまざまなホストOSと関連サービスを代替サイトにフェイルオーバーすることです。HAペアMetroClusterクラスタでは、コントローラが停止すると同じサイト内で単純なフェイルオーバーが発生するため、この問題は解消されます。
- 一部のネットワークトポロジは、サイト間アクセス用に設計されていませんが、異なるサブネットまたは分離されたFC SANを使用します。この場合、代替コントローラが反対側のサイトのサーバにデータを提供できないため、2ノードMetroClusterクラスタはHAシステムとして機能しなくなります。完全な冗長性を実現するには、HAペアのMetroClusterオプションが必要です。
- 2サイトインフラを単一の高可用性インフラとみなす場合は、2ノードMetroCluster構成が適しています。ただし、サイト障害後もシステムが長期間機能しなければならない場合は、HAペアが推奨されます。HAペアは、単一サイト内でHAを提供し続けるためです。

2ノードFC SAN接続MetroCluster

2ノードMetroCluster構成では、サイトごとに1つのノードのみが使用されます。構成とメンテナンスが必要なコンポーネントが少ないため、HAペアオプションよりもシンプルです。また、ケーブル配線やFCスイッチの点でインフラストラクチャの必要性も軽減されています。最後に、コストを削減します。

基本的な設計を図3に示します。

図3）2ノードFC SAN接続MetroClusterの基本アーキテクチャ



この設計の明らかな影響は、1つのサイトでコントローラに障害が発生した場合、反対側のサイトからデータを利用できることです。この制限は必ずしも問題ではありません。多くの企業は、本質的に単一のインフラとして機能する、拡張された高速で低レイテンシのネットワークを使用したマルチサイトデータセンター運用を行っています。このような場合は、2ノードバージョンのMetroClusterが推奨されます。2ノードシステムは現在、複数のサービスプロバイダでペタバイト規模で使用されています。

MetroClusterの耐障害性機能

上の図に示すように、MetroCluster 解決策 には単一点障害（Single Point of Failure）はありません。

- 各コントローラに、ローカルサイトのドライブシェルフへの独立したパスが2つあります。
- 各コントローラに、リモートサイトのドライブシェルフへの独立したパスが2つあります。
- 各コントローラには、反対側のサイトのコントローラへの独立したパスが2つあります。
- HAペア構成では、各コントローラからローカルパートナーへのパスが2つあります。

つまり、構成内のコンポーネントを1つでも削除しても、MetroClusterのデータ提供機能を損なうことはありません。2つのオプションの耐障害性の違いは、サイト障害後もHAペアバージョンが全体的なHAストレージシステムになる点だけです。

MetroClusterの論理アーキテクチャ

ストレージシステムには、データを確実に保護し、データを利用できるようにするという2つの基本的な要件があります。ONTAPのデータ保護テクノロジについて詳しく説明することは本ドキュメントの範囲ではありませんが、さまざまな障害シナリオで何が起こるかを完全に理解するには、各レイヤを確認する必要があります。

データ保護

MetroClusterでの論理データ保護は、次の重要な要件で構成されます。

- ネットワーク上のデータ転送は、データ破損から保護する必要があります。
- ディスクに書き込まれたデータは、データ破損から保護する必要があります。
- ディスクに書き込まれたデータは、ドライブ障害から保護する必要があります。
- データへの変更は損失から保護する必要があります。
- 各サイトに1つずつ、データの独立した2つのコピーを同期しておく必要があります。

ネットワークの破損:チェックサム

最も基本的なデータ保護レベルはチェックサムです。チェックサムは、データと一緒に格納される特別なエラー検出コードです。ネットワーク転送中のデータの破損は、チェックサムを使用して検出されます。場合によっては、複数のチェックサムを使用します。

たとえば、FCフレームには巡回冗長検査（CRC）と呼ばれるチェックサム形式が含まれており、転送中にペイロードが破損していないことを確認できます。送信機は、データのデータとCRCの両方を送信します。FCフレームの受信側は、受信したデータのCRCを再計算して、送信されたCRCと一致することを確認します。新しく計算されたCRCがフレームに接続されたCRCと一致しない場合、データは破損し、FCフレームは破棄または拒否されます。iSCSI I/O処理には、TCP/IPおよびイーサネットレイヤでのチェックサムが含まれます。また、保護を強化するために、SCSIレイヤでオプションのCRC保護を含めることもできます。ワイヤ上のビットの破損はTCPレイヤまたはIPレイヤによって検出され、パケットが再送信されます。FCと同様に、SCSI CRCでエラーが発生すると、処理が破棄または拒否されます。

ドライブの破損：チェックサム

チェックサムは、ドライブに格納されているデータの整合性を検証するためにも使用されます。ドライブに書き込まれたデータブロックは、元のデータに関連付けられた予測不可能な数を生成するチェックサム機能で格納されます。ドライブからデータが読み取られると、チェックサムが再計算され、保存されているチェックサムと比較されます。一致しない場合は、データが破損しているため、RAIDレイヤでリカバリする必要があります。

データの不整合：失われた書き込み

検出するのが最も困難な種類の破損の1つは、書き込みの紛失または置き忘れです。書き込みが確認応答されたら、正しい場所にあるメディアに書き込む必要があります。インプレースデータの破損は、データとともに保存されたシンプルなチェックサムを使用することで、比較的簡単に検出できます。ただし、書き込みが失われても、以前のバージョンのデータが残っている可能性があり、チェックサムが正しいことになります。書き込みが間違った物理的な場所に配置された場合、書き込みによって他のデータが破壊されても、関連するチェックサムは保存データに対して再び有効になります。

この課題に対する解決策は次のとおりです。

- 書き込み処理には、書き込みが予想される場所を示すメタデータが含まれている必要があります。
- 書き込み処理には、何らかのバージョン識別子が含まれている必要があります。

ONTAPがブロックを書き込むときは、そのブロックが属する場所のデータも含まれます。後続の読み取りでブロックが識別されていても、メタデータにブロックが456の場所で見つかったときに123の場所に属していることが示されている場合、書き込みは誤って配置されています。

完全に失われた書き込みを検出することは、より困難です。ONTAPは、書き込み処理によってドライブ上の2つの場所が更新されるようにメタデータを格納しています。書き込みが失われると、その後のデータおよび関連するメタデータの読み取りで、2つの異なるバージョンIDが表示されます。この違いは、ドライブで書き込みが完了しなかったことを示しています。

書き込みの破損が失われたり置き忘れられたりすることは非常にまれですが、ドライブが増え続け、データセットがエクサバイト規模になると、リスクが増大します。データベースワークロードをサポートするストレージシステムには、Lost Write検出機能を含める必要があります。

ドライブ障害：RAID、RAID DP、RAID-TEC

ドライブ上のデータブロックが破損していることが判明した場合、またはドライブ全体で障害が発生して完全に使用できなくなった場合は、データを再構成する必要があります。この再構成は、ONTAPでパリティドライブを使用して行われます。データが複数のデータドライブにストライピングされ、パリティデータが生成されます。パリティデータは、「実際の」データとは別に格納されます。

ONTAPは元々 RAID 4を使用していました。RAID 4は、データドライブのグループごとにパリティドライブを1本使用します。その結果、グループ内のいずれかのドライブで障害が発生してもデータが失われることはありませんでした。パリティドライブで障害が発生してもデータは破損しておらず、新しいパリティドライブを構築できました。1本のデータドライブで障害が発生した場合は、残りのドライブをパリティドライブと一緒に使用して失われたデータを再生成します。

ドライブが小さい場合、2本のドライブで同時に障害が発生する可能性はほとんどありませんでした。ドライブ容量の増大に伴い、ドライブ障害発生後のデータの再構築に必要な時間も増加しています。そのため、2つ目のドライブ障害が発生してデータが失われる時間が長くなりました。さらに、再構築プロセスでは、稼働しているドライブに余分なI/Oが作成されます。ドライブが古くなると、負荷が増えて2つ目のドライブ障害が発生するリスクも高まります。最後に、RAID 4を継続して使用することでデータ損失のリスクが増加しなかったとしても、データ損失の影響はより深刻になります。RAIDグループで障害が発生した場合に失われるデータが増えるほど、データのリカバリにかかる時間が長くなり、業務の中断が長くなります。

これらの問題により、NetAppはRAID 6の一種であるRAID DP® テクノロジを開発しました。この解決策にはパリティドライブが2本含まれているため、RAIDグループ内の2本のドライブで障害が発生してもデータが失われることはありません。RAID DPに加えて、NetAppはRAID-TEC™ テクノロジも開発しました。これは、ドライブのサイズが拡大し続けているため、3つ目のパリティドライブを導入します。

一部の履歴データベースのベストプラクティスでは、ストライプミラーリングとも呼ばれるRAID 10の使用を推奨しています。このタイプのミラーリングでは2本のディスクで障害が発生するのに対し、RAID DPでは何も発生しないため、RAID DPよりもデータ保護が劣ります。

また、パフォーマンス上の懸念から、RAID 4 / 5 / 6よりもRAID 10が推奨されることを示す履歴データベースのベストプラクティスもいくつかあります。これらの推奨事項は、RAIDペナルティを意味する場合があります。これらの推奨事項は正しくありますが、ONTAP内でのRAIDの実装には適用されません。パフォーマンスの問題はパリティ再生に関連しています。従来のRAID実装では、データベースによって実行されるルーチンのランダムライトを処理するには、パリティデータを再生成して書き込みを完了するために、複数のディスク読み取りが必要です。ペナルティは、書き込み処理の実行に必要な追加の読み取りIOPSとして定義されます。

書き込みはメモリでステージングされ、パリティが生成されてから単一のRAIDストライプとしてディスクに書き込まれるため、ONTAPではRAIDペナルティは発生しません。書き込み処理を完了するための読み取りは必要ありません。

要約すると、RAID DPとRAID-TECは、RAID 10と比較して使用可能な容量がはるかに多く、ドライブ障害に対する保護が強化され、パフォーマンスが低下することはありません。

ハードウェア障害からの保護:NVRAM

データベースワークロードを処理するストレージアレイでは、書き込み処理をできるだけ迅速に処理する必要があります。さらに、書き込み処理は、電源やデバイスの障害などの予期しないイベントによって発生する損失から保護する必要があります。

AFFシステムとFASシステムは、これらの要件を満たすためにNVRAMを利用しています。書き込みプロセスは次のように機能します。

1. インバウンド書き込みデータはRAMに格納されます。
2. ディスク上のデータに加えなければならない変更は、ローカルノードとパートナーノードの両方のNVRAMに記録されます。NVRAMは書き込みキャッシュではなく、データベースのRedoログに似たジャーナルです。通常の条件下では、読み取りは行われません。I/O処理中に電源障害が発生した場合など、リカバリにのみ使用されます。
3. その後、書き込みがホストに確認応答されます。

この段階の書き込みプロセスはアプリケーションの観点からは完了しており、データは2つの異なる場所に格納されるため、損失から保護されます。変更は最終的にはディスクに書き込まれますが、このプロセスは書き込みが確認されたあとに実行されるため、レイテンシに影響しないため、アプリケーションの観点からは帯域外です。このプロセスは、再びデータベースロギングのようなものです。データベースに対する変更はできるだけ早くREDOログに記録され、変更がコミットされたことが確認されます。データファイルの更新はかなり遅れて行われ、処理速度に直接影響することはありません。

コントローラに障害が発生した場合は、必要なディスクの所有権がパートナーコントローラに引き継がれます。次に、パートナーコントローラがNVRAMに記録されたデータを再生して、障害発生時に転送中だったI/O処理をリカバリします。

サイト障害からの保護：NVRAMとMetroCluster

MetroClusterは、次の方法でNVRAMデータ保護を拡張します。

- 2ノード構成では、NVRAMデータがスイッチ間リンク（ISL）を使用してリモートパートナーにレプリケートされます。
- HAペア構成では、NVRAMデータがローカルパートナーとリモートパートナーの両方にレプリケートされます。
- 書き込みは、すべてのパートナーにレプリケートされるまで確認応答されません。

このアーキテクチャは、NVRAMデータをリモートパートナーにレプリケートすることで、転送中のI/Oをサイト障害から保護します。このプロセスは、ドライブレベルのデータレプリケーションには関係ありません。アグリゲートを所有するコントローラは、アグリゲート内の両方のプレックスに書き込み、データレプリケーションを実行します。ただし、サイト障害が発生した場合でも、転送中のI/O損失から保護する必要があります。レプリケートされたNVRAMデータは、障害が発生したコントローラをパートナーコントローラがテイクオーバーする必要がある場合にのみ使用されます。

サイトおよびシェルフ障害からの保護：SyncMirrorとプレックス

SyncMirrorは、RAID DPやRAID-TECを強化するミラーリングテクノロジですが、これに代わるものではありません。2つの独立したRAIDグループの内容をミラーリングします。論理構成は次のとおりです。

1. ドライブは、場所に基づいて2つのプールに構成されます。1つのプールはサイトAのすべてのドライブで構成され、2つ目のプールはサイトBのすべてのドライブで構成されます。
2. 次に、アグリゲートと呼ばれる共通のストレージプールが、RAIDグループのミラーセットに基づいて作成されます。各サイトから同じ数のドライブが引き出されます。たとえば、20ドライブのSyncMirrorアグリゲートは、サイトAの10本のドライブとサイトBの10本のドライブで構成されます。

3. サイト上の各ドライブセットは、ミラーリングを使用せずに、完全に冗長化された1つ以上のRAID DPグループまたはRAID-TECグループとして自動的に構成されます。ミラーリングの下でRAIDを使用することで、サイトが失われた場合でもデータを保護できます。

図4）SyncMirror

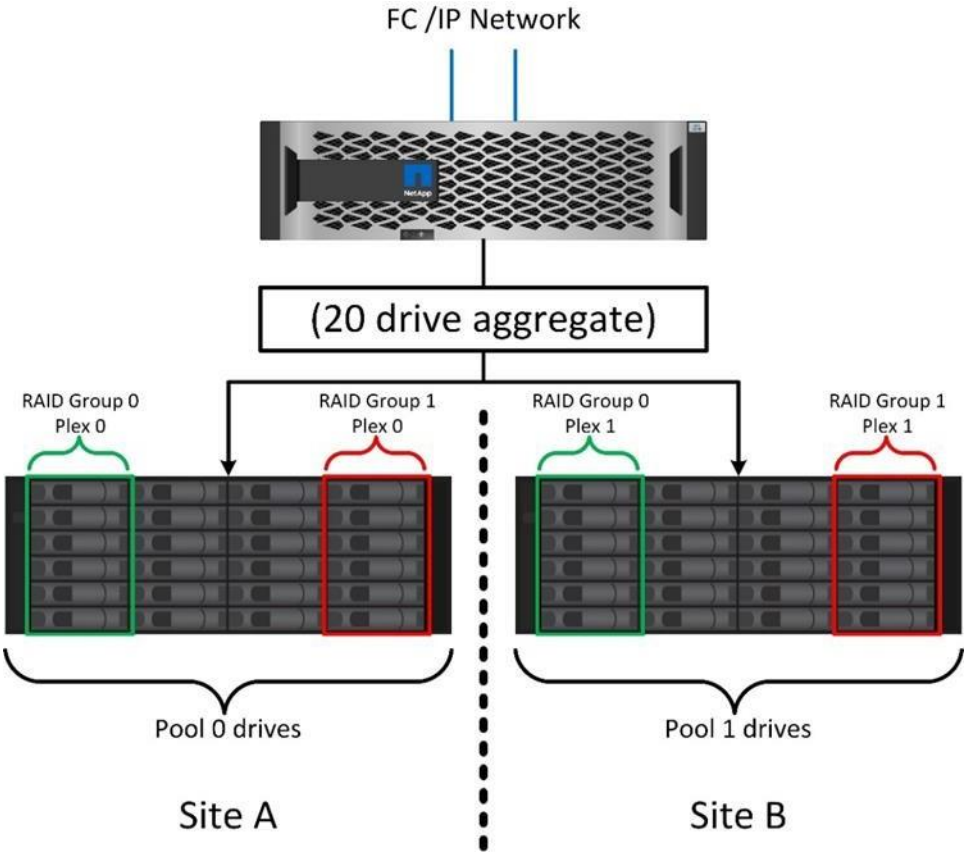


図4 は、SyncMirror構成の例を示しています。コントローラ上の24ドライブのアグリゲートは、サイトAで割り当てられたシェルフの12本のドライブと、サイトBで割り当てられたシェルフの12本のドライブで構成されています。ドライブは2つのミラーRAIDグループにグループ化されます。RAIDグループ0には、サイトAの6ドライブのプレックスが含まれており、サイトBの6ドライブのプレックスにミラーリングされています。同様に、RAIDグループ1にはサイトAの6ドライブのプレックスが含まれており、サイトBの6ドライブのプレックスにミラーリングされています。

SyncMirrorは通常、MetroClusterシステムに非同期のリモートミラーリングを提供するために使用され、各サイトにデータのコピーが1つずつ配置されます。場合によっては、1つのシステムで追加レベルの冗長性を提供するために使用されます。シェルフレベルの冗長性を提供します。ドライブシェルフにはすでにデュアル構成の電源装置とコントローラが搭載されており、全体的にはシートメタル程度ですが、追加の保護が保証される場合があります。たとえば、あるNetAppのお客様は、自動車テストで使用するモバイルリアルタイム分析プラットフォームにSyncMirrorを導入しています。システムは、独立した電源供給と独立したUPSシステムを備えた2つの物理ラックに分かれていました。

冗長性エラー：NVFAIL

前述したように、書き込みの確認応答は、少なくとも1台の他のコントローラでローカルのNVRAMとNVRAMに記録されるまで返されません。このアプローチにより、ハードウェア障害や停電が発生しても、転送中のI/Oが失われることはありません。ローカルのNVRAMに障害が発生したり、他のノードへの接続に障害が発生したりすると、データはミラーリングされなくなります。

ローカルNVRAMからエラーが報告されると、ノードはシャットダウンします。このシャットダウンにより、HAペアが使用されている場合はパートナーコントローラにフェイルオーバーされます。MetroClusterでは、動作は選択した全体的な設定によって異なりますが、リモートノートに自動的にフェイルオーバーされる場合があります。いずれの場合も、障害が発生したコントローラが書き込み処理を認識していないため、データは失われません。

リモートノードへのNVRAMレプリケーションがブロックされるサイト間接続障害は、より複雑な状況です。書き込みがリモートノードにレプリケートされなくなるため、コントローラで重大なエラーが発生した場合にデータが失われる可能性があります。さらに重要なことは、このような状況で別のノードにフェイルオーバーしようとするとデータが失われることです。

制御要素は、NVRAMが同期されているかどうかです。NVRAMが同期されていれば、ノード間のフェイルオーバーを安全に実行でき、データ損失のリスクはありません。MetroCluster構成では、NVRAMと基盤となるアグリゲートのプレックスが同期されていれば、データ損失のリスクなしにスイッチオーバーを実行できます。

データが同期されていない場合、ONTAPは、フェイルオーバーまたはスイッチオーバーを強制的に実行しないかぎり、フェイルオーバーまたはスイッチオーバーを許可しません。この方法で条件を変更すると、元のコントローラにデータが残っている可能性があり、データ損失が許容されることが確認されます。

データベースは、ディスク上のデータのより大きな内部キャッシュを保持するため、フェイルオーバーやスイッチオーバーを強制的に実行した場合、データベースが破損する可能性が特に高くなります。強制的なフェイルオーバーまたはスイッチオーバーが発生した場合、以前に確認済みの変更は事実上破棄されます。ストレージ アレイの内容は事実上時間を逆方向に進め、データベース キャッシュの状態はディスク上のデータの状態を反映しなくなりました。

この状況からデータベースを保護するために、ONTAPでは、NVRAM障害に対する特別な保護をボリュームに設定できます。この保護メカニズムがトリガーされると、ボリュームがNVFAILという状態になります。この状態になると、原因Aデータベースがクラッシュする I/Oエラーが発生します。このクラッシュにより、データベースはシャットダウンされ、古いデータが使用されなくなります。コミットされたトランザクションデータがログに含まれている必要があるため、データが失われないようにしてください。次の手順では、管理者がホストを完全にシャットダウンしてから、LUNとボリュームを手動で再度オンラインに戻します。これらの手順にはいくつかの作業が含まれる可能性がありますが、このアプローチはデータの整合性を確保するための最も安全な方法です。すべてのデータがこの保護を必要とするわけではありません。そのため、NVFAILの動作はボリューム単位で設定できます。

高可用性

完全な概要of ONTAP HA機能については、本ドキュメントでは説明しません。ただし、データ保護と同様に、データベースインフラを設計する際には、この機能の基本的な理解が重要です。

HAペア

HAの基本単位はHAペアです。各ペアには、NVRAMデータのレプリケーションをサポートするための冗長リンクが含まれています。NVRAMは書き込みキャッシュではありません。コントローラ内部のRAMは書き込みキャッシュとして機能します。NVRAMの目的は、予期しないシステム障害から保護するためにデータを一時的にジャーナルすることです。この点では、データベースのREDOログに似ています。

NVRAMとデータベースのRedoログはどちらもデータを迅速に格納するために使用されるため、データに対する変更をできるだけ迅速にコミットできます。ドライブ（またはデータファイル）上の永続的データの更新は、ONTAPとほとんどのデータベースプラットフォームの両方でチェックポイントと呼ばれるプロセスを実行するまで行われません。NVRAMデータまたはデータベースREDOログは、通常運用時に読み取られます。

コントローラで突然障害が発生した場合、ドライブにまだ書き込まれていない保留中の変更がNVRAMに保存されている可能性があります。パートナーコントローラが障害を検出し、ドライブを制御して、NVRAMに保存されている必要な変更を適用します。

HAペアとMetroCluster

MetroClusterには、2ノードとHAペアの2つの構成があります。2ノード構成の動作は、NVRAMに関してHAペアと同じです。突然の障害が発生した場合、パートナーノードはNVRAMデータを再生してドライブの整合性を確保し、確認済みの書き込みが失われていないことを確認できます。

HAペア構成では、ローカルパートナーノードにもNVRAMがレプリケートされます。MetroClusterを使用しないスタンドアロンHAペアの場合と同様に、単純なコントローラ障害ではパートナーノードでNVRAMが再生されます。サイト全体が突然失われた場合、リモートサイトには、ドライブの整合性を確保してデータの提供を開始するために必要なNVRAMもあります。

MetroClusterの重要な側面の1つは、通常の運用状態ではリモートノードがパートナーデータにアクセスできないことです。各サイトは本質的に、反対のサイトのパーソナリティを想定できる独立したシステムとして機能します。このプロセスはスイッチオーバーと呼ばれ、計画的スイッチオーバーでは、サイトの処理が無停止で反対側のサイトに移行されます。また、サイトが失われ、ディザスタリカバリの一環として手動または自動のスイッチオーバーが必要になる計画外の状況も含まれます。

テイクオーバーとギブバック

テイクオーバーとギブバックは、HAペアのノード間でストレージリソースの責任を移すプロセスです。テイクオーバーとギブバックには次の2つの側面があります。

- ドライブへのアクセスを許可するネットワーク接続の管理
- ドライブ自体の管理

CIFSおよびNFSトラフィックをサポートするネットワークインターフェイスには、ホームロケーションとフェイルオーバーロケーションの両方が設定されます。テイクオーバーでは、元の場所と同じサブネットが検出された物理インターフェイス上の一時的なホームにネットワークインターフェイスを移動します。ギブバックでは、ネットワークインターフェイスを元の場所に戻します。必要に応じて、正確な動作を調整できます。

iSCSIやFCなどのSANブロックプロトコルをサポートしているネットワークインターフェイスは、テイクオーバーやギブバックの実行時に再配置されません。代わりに、完全なHAペアを含むパスを使用してLUNをプロビジョニングする必要があります。これにより、プライマリパスとセカンダリパスが作成されます。

注： 大規模なクラスタ内のノード間でのデータの再配置をサポートするように追加のコントローラへのパスを設定することもできますが、この再配置はHAプロセスの一部ではありません。

テイクオーバーとギブバックの2つ目の側面は、ディスク所有権の移行です。具体的なプロセスは、テイクオーバー/ギブバックの理由やコマンドラインオプションを実行した理由など、複数の要因によって異なります。目標は、できるだけ効率的に操作を実行することです。全体的なプロセスには数分かかるように見えるかもしれませんが、ドライブの所有権がノードからノードに移行される実際の瞬間は、通常数秒で測定できます。

テイクオーバー時間

テイクオーバー処理やギブバック処理の実行中にホストI/Oが短時間中断されますが、正しく設定された環境ではアプリケーションが停止することはありません。I/Oが遅延する実際の移行プロセスは、秒単位で測定されます。ただし、ホストがデータパスの変更を認識してI/O処理を再送信するまでに、さらに時間がかかる場合があります。

中断の内容はプロトコルによって異なります。

- NFSおよびCIFSトラフィックをサポートするネットワークインターフェイスは、新しい物理的な場所への移行後に、ネットワークに対してAddress Resolution Protocol（ARP；アドレス解決プロトコル）要求を発行します。この要求により、ネットワークスイッチはMACアドレステーブルを更新し、I/Oの処理を再開します。計画的なテイクオーバーとギブバックが実行される際のシステム停止は秒単位で測定され、検出されないことがよくあります。ネットワークによっては、ネットワークパスの変更を完全に認識するのに時間がかかる場合があり、OSによっては短時間に大量のI/Oがキューイングされて再試行が必要になる場合があります。このキューイングにより、I/Oの再開に必要な時間が長くなる可能性があります。

- SANプロトコルをサポートするネットワークインターフェイスが新しい場所に移行されない。ホストOSが使用中のパスを変更する必要があります。ホストで検出されるI/Oの一時停止は、複数の要因によって異なります。ストレージシステムの観点から見ると、I/Oを処理できない時間はわずか数秒です。ただし、ホストOSによっては、I/Oがタイムアウトしてから再試行するまでに時間がかかることがあります。新しいOSではパスの変更をより迅速に認識できますが、古いOSでは通常、変更を認識するのに最大30秒かかります。

表1に、ストレージシステムがデータベース環境にデータを提供できない場合の想定テイクオーバー時間を示します。

表1）想定されるテイクオーバー時間

	NAS	SAN向けに最適化されたOS	SAN
計画的なテイクオーバー	15秒	2～10秒	2～10秒
計画外のテイクオーバー	30 秒	2～15秒	30 秒

テイクオーバーのトリガー

テイクオーバーが発生する状況は次のとおりです。

- `storage failover takeover` コマンドを使用してテイクオーバーを手動で開始します。
- ソフトウェアまたはシステムの障害が発生し、コントローラがパニック状態になる。パニックが完了してシステムがリブートすると、ストレージリソースがギブバックされ、システムは通常の状態に戻ります。この動作はデフォルトで、必要に応じて変更できます。
- コントローラに電源の喪失など、システム全体の障害が発生し、リブートできない。
- パートナーコントローラがハートビートメッセージを受信できません。この状況は、パートナーでハードウェアまたはソフトウェアの障害が発生してもパニックにはならないものの、正常な動作が妨げられている場合に発生する可能性があります。
- ノードを手動で停止すると、`- inhibit_takeover` パラメータをに指定してコマンドを実行しないかぎり、テイクオーバーがトリガーされる可能性があります `true`。
- ノードを手動でリブートすると、`- inhibit_takeover` パラメータをに指定してコマンドを実行しないかぎり、テイク `true` オーバーがトリガーされることがあります。
- ハードウェアアシストテイクオーバーが有効になっている場合は、サービスプロセッサがパートナーノードの障害を検出したときにテイクオーバーをトリガーできます。

ハードウェア アシスト テイクオーバー

サービスプロセッサは、AFFシステムおよびFASシステムに組み込まれているアウトオブバンド管理デバイスです。独自のIPアドレスでアクセスされ、コントローラが動作しているかどうかに関係なく、コンソールへの直接アクセスやその他の管理機能に使用されます。

ONTAPだけでは、パートナーノードからのハートビートが検出されなくなり、タイムアウトが発生した場合に、障害が発生したノードのテイクオーバーがトリガーされます。ハードウェアアシストテイクオーバーでは、サービスプロセスを使用して障害をより迅速に検出し、テイクオーバーをすぐに開始することで、テイクオーバープロセスにかかる時間を短縮します。パートナーのハートビートが停止したことをONTAPが認識するまで待機しません。

スイッチオーバーとスイッチバック

スイッチオーバーとスイッチバックという用語は、MetroCluster構成のリモートコントローラ間でボリュームを移行するプロセスを指します。このプロセスでは、リモートノードのみが環境されます。4ボリューム構成でMetroClusterを使用する場合のローカルノードのフェイルオーバーは、前述したテイクオーバーとギブバックのプロセスと同じです。

計画的スイッチオーバーとスイッチバック

計画的スイッチオーバーまたはスイッチバックは、ノード間のテイクオーバーやギブバックのような機能です。このプロセスには複数の手順があり、数分かかるように見える場合もありますが、ストレージリソースとネットワークリソースを複数の段階で適切に移行します。制御が転送されると、完全なコマンドの実行に必要な時間よりもはるかに短時間でモーメントが発生します。

テイクオーバー/ギブバックとスイッチオーバー/スイッチバックの主な違いは、FC SAN接続への影響です。ローカルのテイクオーバー/ギブバックでは、ローカルノードへのFCパスがすべて失われ、ホストのネイティブのMicrosoft Multipath I/O（MPIO；マルチパスI/O）を使用して使用可能な代替パスに切り替えます。ポートは再配置されません。スイッチオーバーとスイッチバックでは、コントローラの仮想FCターゲットポートがもう一方のサイトに移行します。一時的にSAN上に存在しなくなり、代わりのコントローラに再表示されます。

SyncMirrorタイムアウト

SyncMirrorは、シェルフ障害から保護するONTAPのミラーリングテクノロジです。シェルフが離れた場所に配置されている場合は、リモートデータ保護が実現します。

SyncMirrorは汎用同期ミラーリングを提供しません。その結果、可用性が向上します。一部のストレージシステムでは、一定のオールオアナッシングミラーリング（Dominoモードと呼ばれることもあります）を使用します。リモートサイトへの接続が失われるとすべての書き込みアクティビティが停止する必要があるため、この形式のミラーリングはアプリケーションで制限されます。そうしないと、書き込みは一方のサイトに存在し、もう一方のサイトには存在しません。通常、このような環境では、サイト間の接続が短時間（30秒など）以上切断された場合にLUNがオフラインになるように構成されます。

この動作は、データベース環境の一部に適しています。しかし、ほとんどのデータベースには、通常の動作条件下で保証された同期レプリケーションを提供しながら、レプリケーションを一時停止できる解決策が必要です。サイト間の接続が完全に失われると、多くの場合、災害が近い状況とみなされます。通常、このようなデータベース環境は、接続が修復されるか、データを保護するためにデータベースをシャットダウンする正式な決定が下されるまで、オンラインのままでデータを提供します。リモート・レプリケーションの障害のみが原因で'データベースの自動シャットダウンが必要になることは'通常ではありません

SyncMirrorは、タイムアウトの柔軟性を備えた同期ミラーリングの要件に対応しています。リモートコントローラやプレックスへの接続が失われると、30秒のタイマーがカウントダウンを開始します。カウンタが0に達すると、ローカルデータを使用して書き込みI/O処理が再開されます。データのリモートコピーは使用可能ですが、接続が回復するまで時間内に凍結されます。再同期では、アグリゲートレベルのSnapshotを使用してシステムをできるだけ迅速に同期モードに戻します。

特に、この種の汎用的なオールオアナッシングDominoモードレプリケーションは、アプリケーションレイヤでより適切に実装されていることがよくあります。たとえば、Oracle DataGuardには最大保護モードが用意されており、どのような状況でも長時間のインスタンスレプリケーションが保証されます。設定可能なタイムアウトを超えてレプリケーションリンクに障害が発生すると、データベースはシャットダウンします。

ファブリック接続MetroClusterによる自動無人スイッチオーバー

自動無人スイッチオーバー（AUSO）は、クロスサイトHAの形式を提供するファブリック接続MetroCluster機能です。前述したように、MetroClusterには2つのタイプ（各サイトに1台のコントローラを配置する場合と、各サイトに1台のHAペアを配置する場合）があります。HAオプションの主な利点は、コントローラの計画的シャットダウンと計画外シャットダウンのどちらでもすべてのI/Oをローカルで処理できることです。シングルノードオプションのメリットは、コスト、複雑さ、インフラの削減です。

AUSOの主な価値は、ファブリック接続MetroClusterシステムのHA機能を向上させることです。各サイトが反対側のサイトの健常性を監視し、データを提供するノードがなくなると、AUSOによって迅速なスイッチオーバーが実行されます。このアプローチは、可用性の点でHAペアに近い構成になるため、サイトごとにノードが1つだけのMetroCluster構成で特に役立ちます。

AUSOでは、HAペアレベルで包括的な監視を行うことはできません。HAペアには、ノード間の直接通信用の2本の冗長な物理ケーブルが含まれているため、きわめて高い可用性を実現できます。さらに、HAペアの両方のノードが冗長ループ上の同じディスクセットにアクセスできるため、1つのノードが別のノードの健常性を監視するための別のルートが提供されます。

MetroClusterクラスタは複数のサイトにまたがって存在し、ノード間の通信とディスクアクセスの両方がサイト間ネットワーク接続に依存します。クラスタの残りの部分のハートビートを監視する機能には制限があります。AUSOは、ネットワークの問題が原因で、もう一方のサイトが使用できない状況ではなく、実際にダウンしている状況を区別する必要があります。

その結果、HAペアのコントローラで、システムパニックなどの特定の理由で発生したコントローラ障害が検出された場合、テイクオーバーが要求されることがあります。また、接続が完全に失われた場合（ハートビートの損失とも呼ばれます）、テイクオーバーを促すこともあります。

MetroClusterシステムで自動スイッチオーバーを安全に実行できるのは、元のサイトで特定の障害が検出された場合のみです。また、ストレージシステムの所有権を取得するコントローラは、ディスクとNVRAMのデータが同期されていることを保証できる必要があります。コントローラは、ソースサイトとの通信が失われて稼働している可能性があるため、スイッチオーバーの安全性を保証できません。スイッチオーバーを自動化するためのその他のオプションについては、次のセクションのMetroCluster Tiebreaker（MCTB）解決策に関する情報を参照してください。

MetroCluster Tiebreakerとファブリック接続MetroCluster

NetApp MetroCluster Tiebreakerソフトウェアを第3のサイトで実行して、MetroCluster環境の健常性の監視、通知の送信を行うことができます。また、災害発生時にスイッチオーバーを強制的に実行することもできます。Tiebreakerの完全な概要は NetApp Support Siteで確認できます が、MetroCluster Tiebreakerの主な目的はサイトの損失を検出することです。また、サイトの損失と接続の損失を区別する必要があります。たとえば、Tiebreakerがプライマリサイトに到達できなかったためにスイッチオーバーが発生しないようにします。そのため、Tiebreakerはリモートサイトからプライマリサイトに接続する機能も監視します。

AUSOによる自動スイッチオーバーもMCTBと互換性があります。AUSOは、特定の障害イベントを検出し、NVRAMとSyncMirrorのプレックスが同期されている場合にのみスイッチオーバーを実行するように設計されているため、迅速に対応します。

一方、Tiebreakerはリモートに配置されているため、サイトの停止を宣言する前にタイマーが経過するのを待つ必要があります。Tiebreakerは最終的にAUSOの対象となるコントローラ障害を検出しますが、一般的にはAUSOがスイッチオーバーを開始しており、Tiebreakerが機能する前にスイッチオーバーを完了している可能性があります。Tiebreakerから送信される2つ目のswitchoverコマンドは拒否されます。

注意： MCTBソフトウェアでは、強制的にスイッチオーバーを実行する際に、NVRAM（不揮発性RAM）が同期されているかどうかは検証されません。メンテナンス作業中に自動スイッチオーバーが設定されている場合は無効にして、NVRAMまたはSyncMirrorプレックスの同期が失われるようにしてください。

また、MCTBは、次の一連のイベントにつながるローリングディザスタに対応できない場合があります。

1. サイト間の接続が30秒以上中断されます。
2. SyncMirrorレプリケーションがタイムアウトし、プライマリサイトで処理が続行されるため、リモートレプリカは古くなります。
3. プライマリサイトが失われた場合。

その結果、プライマリサイトにレプリケートされていない変更が存在することになります。その場合、次のようないくつかの理由でスイッチオーバーが望ましくない可能性があります。

- 重要なデータはプライマリサイトに存在し、最終的にリカバリ可能になる可能性があります。スイッチオーバーによってデータベースの動作が継続されると、重要なデータは実質的に破棄されます。
- サバイバーサイトのデータベースで、サイト障害時にプライマリサイトのストレージリソースを使用していた場合、データがキャッシュされている可能性があります。スイッチオーバーでは、キャッシュと一致しない古いバージョンのデータが生成されます。
- サバイバーサイトのオペレーティングシステムで、サイト障害時にプライマリサイトのストレージリソースを使用していた場合、キャッシュデータがある可能性があります。スイッチオーバーでは、キャッシュと一致しない古いバージョンのデータが生成されます。

最も安全な方法は、Tiebreakerがサイト障害を検出した場合にアラートを送信するように設定し、スイッチオーバーを強制的に実行するかどうかをユーザが決定できるようにすることです。キャッシュされたデータをクリアするには、まずデータベースやオペレーティングシステムをシャットダウンしなければならない場合があります。さらに、NVFAIL設定を使用して保護を強化し、フェイルオーバープロセスを合理化することもできます。

MetroCluster IPを使用したONTAPメディエーター

ONTAPメディエーターは、MetroCluster IPおよびその他の特定のONTAPソリューションで使用されます。これは、前述のMetroCluster Tiebreakerソフトウェアと同様に従来のTiebreakerサービスとして機能しますが、自動無人スイッチオーバーの実行という重要な機能も備えています。

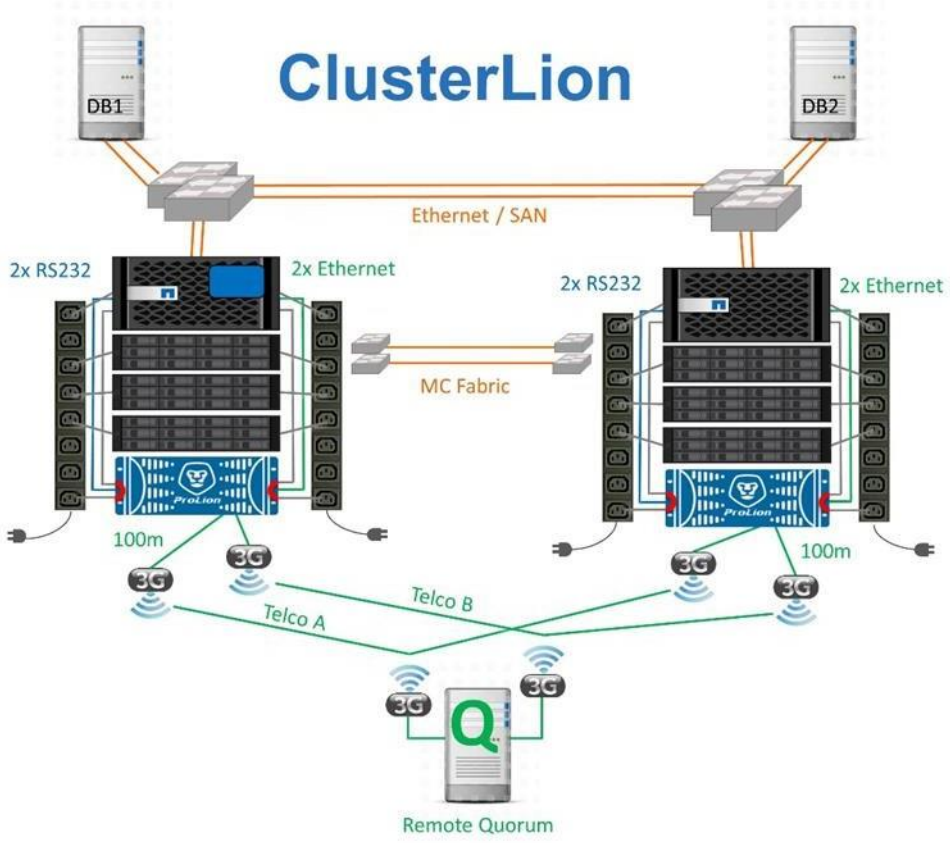
ファブリック接続MetroClusterは、反対側のサイトのストレージデバイスに直接アクセスできます。このアクセスにより、一方のMetroClusterコントローラがドライブからハートビートデータを読み取って他のコントローラの健常性を監視できるようになります。同じアクセスで、1台のコントローラが別のコントローラの障害を認識し、スイッチオーバーを実行できます。

一方、MetroCluster IPアーキテクチャでは、すべてのI/Oがコントローラとコントローラの接続を介して排他的にルーティングされるため、リモートサイトのストレージデバイスに直接アクセスすることはありません。直接アクセスによって障害を検出してスイッチオーバーを実行するコントローラの能力が制限されることはありません。そのため、サイトの損失を検出して自動的にスイッチオーバーを実行するためには、ONTAPメディエーターがTiebreakerデバイスとして必要になります。

ClusterLionを使用した3番目の仮想サイト

ClusterLionは、仮想の第3サイトとして機能する高度なMetroCluster監視アプライアンスです。このアプローチにより、完全に自動化されたスイッチオーバー機能により、MetroClusterを2サイト構成で安全に導入できます。さらに、ClusterLionでは、追加のネットワークレベル監視を実行し、スイッチオーバー後の処理を実行できます。完全なドキュメントはProLionから入手できます。図5 に、いくつかの特長を示します。

図5）ProLionのアーキテクチャ



- ClusterLionアプライアンスは、直接接続されたイーサネットケーブルとシリアルケーブルでコントローラの健常性を監視します。
- 2つのアプライアンスは、冗長3Gワイヤレス接続で相互に接続されています。
- ONTAPコントローラへの電源は、内部リレーを介して配線されます。サイト障害が発生すると、内部UPSシステムを搭載したClusterLionによって電源接続が切断されてからスイッチオーバーが実行されます。このプロセスにより、スプリットブレイン状態が発生しないようにします。
- ClusterLionは、30秒のSyncMirrorタイムアウト内にスイッチオーバーを実行するか、まったく実行しません。
- ClusterLionでは、NVRAMプレックスとSyncMirrorプレックスの状態が同期されていないかぎり、スイッチオーバーは実行されません。
- ClusterLionでは、MetroClusterが完全に同期されている場合にのみスイッチオーバーが実行されるため、NVFAILは必要ありません。この構成では、計画外スイッチオーバーが発生しても、拡張Oracle RACなどのサイトスパニング環境をオンラインのまま維持できます。
- ファブリック接続MetroClusterとMetroCluster IPの両方がサポートされます。

OracleとNVFAIL

フェイルオーバーまたはスイッチオーバーが強制的に実行されると、データベースは大規模な内部キャッシュを保持するため、破損の影響を受けやすくなります。強制フェイルオーバーまたは強制MetroClusterスイッチオーバーが発生した場合、以前に確認済みの変更は事実上破棄されます。ストレージアレイの内容が時間を遡るようになり、データベースキャッシュの状態がディスク上のデータの状態を反映しなくなります。この不整合により、データが破損します。

キャッシングは、アプリケーション層またはサーバ層で行われます。たとえば、プライマリサイトとリモートサイトの両方でアクティブなサーバを使用するOracle Real Application Cluster（RAC）構成では、Oracle SGA内のデータがキャッシュされます。強制スイッチオーバー処理によってデータが失われると、SGAに格納されているブロックがディスク上のブロックと一致しない可能性があるため、データベースが破損するリスクがあります。

キャッシュの使用は、OSファイル システム レイヤではあまり明らかではありません。マウントされたNFSファイルシステムのブロックは、OSにキャッシュされる場合があります。または、見つかったLUNに基づいてクラスタ化されたファイルシステムで、プライマリサイトをリモートサイトのサーバにマウントして、データをキャッシュすることもできます。このような状況でNVRAMの障害、強制テイクオーバー、強制スイッチオーバーが発生すると、ファイルシステムが破損する可能性があります。

ONTAPシステムでは、NVFAILとそれに関連するパラメータを使用して、このシナリオからデータベースとオペレーティング システムを保護します。

NVFAIL

ONTAPストレージ上のすべてのデータベースボリュームで、 `nvfail` パラメータをに設定する必要があります `on`。

この設定は、データの整合性を損なうNVRAMジャーナリングの壊滅的な障害からボリュームを保護します。 `nvfail` パラメータは起動時に有効になります。NVRAMエラーが検出された場合は、コミットされていない変更が失われている可能性があり、ドライブの状態がデータベースキャッシュと一致していない可能性があります。ONTAPは `nvfail`、パラメータをに設定してボリュームを `on` に設定します `in-nvfailed-state`。その結果、データにアクセスしようとするすべてのデータベースプロセスにI/Oエラーが発生し、データベースの保護クラッシュまたはシャットダウンが発生します。

dr-force-nvfail

`dr-force-nvfail` パラメータは、特定の計画外MetroClusterスイッチオーバーイベントからデータを保護します。「ハイアベイラビリティ」の項で説明したように、SyncMirrorではリモート接続が切断された場合に30秒のタイムアウトが発生します。ローリングディザスタによって最初にレプリケーションが中断され、数分後にサイトの残りの部分が破壊される可能性があります。また、データのプライマリコピーとリモートコピーの状態が同期されていない場合、メンテナンス中に災害が発生する可能性があります。いずれかのアプリケーションがリモートサイトのデータにアクセスしていた場合、スイッチオーバーによってデータの古いコピーがキャッシュの状態と一致しなくなる可能性があります。

`nvfail` パラメータは、主に単一のHAペア内のテイクオーバーとギブバックの手順に対応することを目的としています。NVRAMの不整合が検出されたMetroClusterスイッチオーバーイベントには適用されますが、強制スイッチオーバー中のデータ損失は保護されません。強制スイッチオーバーを実行する管理者は、基本的に、ディザスタサイトのコントローラにレプリケートされていないデータがある可能性があることを確認します。いずれにしても、データの残りのコピーをアクティブ化する必要があります。この状況では、という2つ目のパラメータを使用してデータベースを保護する必要もあります `dr-force-nvfail`。`dr-force-nvfail in-nvfailed-state` 強制スイッチオーバー中にと表示されたボリュームがになります。データが実際に同期されていない可能性があるため、このNVFAILは強制的に実行されます。リモートデータは元のデータと整合性がある可能性がありますが、プライマリサイトにアクセスできない場合、整合性を保証する方法はありません。

プライマリ `dr-force-nvfail` サイトとリモートサイトの両方のサーバがストレージシステム上のデータにアクセスするデータベースクラスタが必要になります。その結果、リモートサイトのサーバにキャッシュされたデータが存在する可能性があります。このデータはプライマリサイトにコミットされていますが、メンテナンスや異常なローリングディザスタによってリモートサイトにレプリケートされていません。このような状況で強制的スイッチオーバーを実行すると、データのリモートコピーがプライマリサイトの状態と正確に一致しない場合にデータが破損するリスクがあります。

災害発生時にリモートサイトのデータにサーバがアクセスしていない場合は、通常、 `dr-force-nvfail` ボリュームに設定する理由はありません。継続的な災害によって、元のデータと一致しないデータのコピーが残っている可能性はありますが、 `dr-force-nvfail` 強制的にパージする必要があるキャッシュデータがない場合は保護されません。

最大限の保護を実現するには、データをキャッシュするリモートサーバからアクセスされる可能性のあるすべて

のボリュームを dr-force-nvfail に設定する必要があります on。パラメータをに設定しても問題ない off のは、アプリケーションやオペレーティングシステムによってキャッシュされていないデータ、またはスイッチオーバープロセスが30秒のSyncMirrorタイムアウト内に実行されることが保証されている場合だけです。

注意： ONTAP 9.0より前のバージョンでは、dr-force-nvfail このパラメータを使用すると nvfailed、正常な計画的スイッチオーバーの実行中にボリュームも状態になります。

force-nvfail-all

-force-nvfail-all パラメータは、switchoverコマンドで使用するオプションの引数です。in-nvfailed-state true スイッチオーバーするすべてのボリュームに対してパラメータがに設定され、-dr- force-nvfail true まだ有効になっていないボリュームについてはパラメータがに設定されます。

この引数の主な用途は、災害時にリモートサイトを強制的にオンラインにする必要があり、元のサイトのデータがディザスタリカバリサイトのどこかにキャッシュされていたかどうかに関する質問がある場合です。nvfail dr-force-nvfail パラメータとパラメータがすべてのボリュームで正しく設定されている場合は、を使用する必要はありませ force-nvfail-allん。ただし、-force- nvfail-all さらに確実性を高めるために使用することが望ましい場合もあります。欠点としては in-nvfailed-state、すべてのボリュームからフラグをクリアする必要があること、およびそれらのボリュームにアクセスしようとするホストで発生するI/Oエラーなどがあります。

MetroCluster上のOracle単一インスタンス

前述したように、MetroClusterシステムが存在しても、データベースの運用に関するベストプラクティスが必ずしも追加されたり変更されたりするわけではありません。お客様のMetroClusterシステムで現在実行されているデータベースのほとんどはシングルインスタンスであり、TR-3633：『Oracle Databases on ONTAP』の推奨事項に従っています。スタンバイサーバリソースはリモートサイトに存在しますが、特別な自動化や統合は行われていません。

事前設定されたOSを使用したフェイルオーバー

SyncMirrorはディザスタリカバリサイトにデータの同期コピーを提供しますが、そのデータを利用できるようにするには、オペレーティングシステムと関連するアプリケーションが必要です。基本的な自動化により、環境全体のフェイルオーバー時間を大幅に短縮できます。Veritas Cluster Server（VCS）などのClusterware製品は、サイト間でクラスタを作成するためによく使用され、多くの場合、フェイルオーバープロセスは単純なスクリプトで実行できます。

マスターノードが失われた場合、代替サイトでデータベースをオンラインにするようにクラスタウェア（またはスクリプト）が設定されます。1つは、データベースを構成するNFSリソースまたはSANリソース用に事前設定されたスタンバイサーバを作成する方法です。プライマリサイトに障害が発生すると、クラスタウェアまたはスクリプトによって次のような一連の処理が実行されます。

1. MetroClusterスイッチオーバーの強制実行
2. FC LUNの検出の実行（SANのみ）
3. ファイルシステムのマウントおよび/またはAutomatic Storage Management（ASM）ディスクグループのマウント
4. データベースの起動

このアプローチの主な要件は、リモートサイトでOSを実行することです。Oracleバイナリを使用して事前に設定する必要があります。つまり、Oracleのパッチ適用などのタスクをプライマリサイトとスタンバイサイトで実行する必要があります。また、災害が発生した場合は、Oracleバイナリをリモートサイトにミラーリングしてマウントすることもできます。

実際のアクティベーション手順は簡単です。LUN検出などのコマンドでは、FCポートあたりのコマンド数が少なくて済みます。ファイルシステムのマウントはマウントコマンドにすぎません。データベースとASMの両方を、1つのコマンドでCLIから開始および停止できます。スイッチオーバーの前にディザスタリカバリサイトでボリュームとファイルシステムが使用されていない場合は、 dr-force-nvfail ボリュームでを設定する必要はありません。

仮想OSによるフェイルオーバー

データベース環境のフェイルオーバーを拡張して、オペレーティングシステム自体を含めることができます。理論的には、このフェイルオーバーはブートLUNで実行できますが、ほとんどの場合、仮想OSで実行されます。手順の手順は次のようになります。

1. MetroClusterスイッチオーバーの強制実行
2. データベースサーバ仮想マシンをホストするデータストアのマウント
3. 仮想マシンの起動
4. データベースを手動で起動するか、またはデータベースを自動的に起動するように仮想マシンを設定する

たとえば、ESXクラスタが複数のサイトにまたがっているとします。災害が発生した場合は、スイッチオーバー後にディザスタリカバリサイトで仮想マシンをオンラインにすることができます。災害発生時に仮想データベースサーバをホストするデータストアが使用されていない場合は `dr-force-nvfail` 、関連するボリュームでを設定する必要はありません。

MetroCluster上の拡張Oracle RAC

多くのお客様が、Oracle RACクラスタを複数のサイトにまたがって構成し、完全なアクティブ/アクティブ構成を実現することで、RTOを最適化しています。Oracle RACのクォーラム管理を含める必要があるため、設計全体が複雑になります。また、データは両方のサイトからアクセスされるため、強制的スイッチオーバーによって古いデータコピーが使用される可能性があります。

データのコピーは両方のサイトに存在しますが、データを提供できるのはアグリゲートを現在所有しているコントローラだけです。そのため、拡張RACクラスタでは、リモートのノードがサイト間接続でI/Oを実行する必要があります。その結果、I/Oレイテンシが増加しますが、このレイテンシは一般的には問題になりません。RACインターコネクトネットワークは複数のサイトにまたがって拡張する必要があるため、とにかく高速で低レイテンシのネットワークが必要です。レイテンシが増加して原因に問題が発生した場合は、クラスタをアクティブ/パッシブで運用できます。I/O負荷の高い処理は、アグリゲートを所有するコントローラに対してローカルなRACノードに対して実行する必要があります。リモートノードは、より軽いI/O処理を実行するか、純粋にウォームスタンバイサーバとして使用されます。

高可用性の拡張RACが必要な場合は、MetroClusterの代わりにASMミラーリングを検討する必要があります。ASMミラーリングでは、データの特定のレプリカを優先することができます。したがって、すべての読み取りがローカルに行われる拡張RACクラスタを構築できます。読み取りI/Oがサイトを経由することはないため、レイテンシは最小限に抑えられます。すべての書き込みアクティビティは引き続きサイト間接続を転送する必要がありますが、このようなトラフィックは同期ミラーリング解決策では回避できません。

注： 仮想ブートディスクを含むブートLUNをOracle RACで使用する場合は、`miscount` パラメータの変更が必要になることがあります。RACタイムアウトパラメータの詳細については、TR-3633：『Oracle Databases on ONTAP』を参照してください。

2サイト構成

2サイトの拡張RAC構成では、すべてではないが多くの災害シナリオに無停止で対応できる高可用性データベースサービスを提供できます。

RAC投票ファイル

MetroClusterに拡張RACを導入する場合は、クォーラム管理を最初に検討する必要があります。Oracle RACには、クォーラムを管理するための2つのメカニズム（ディスクハートビートとネットワークハートビート）があります。ディスクハートビートは、投票ファイルを使用してストレージアクセスを監視します。単一サイトのRAC構成では、基盤となるストレージシステムがHA機能を提供していれば、単一の投票リソースで十分です。

以前のバージョンのOracleでは、投票ファイルは物理ストレージデバイスに配置されていましたが、現在のバージョンのOracleでは、投票ファイルはASMディスクグループに格納されていました。

注： Oracle RACはNFSでサポートされています。グリッドのインストールプロセスでは、一連のASMプロセスが作成され、グリッドファイルに使用されるNFSの場所がASMディスクグループとして提供されます。このプロセスはエンドユーザに対してほぼ透過的であり、インストール完了後にASMを継続的に管理する必要はありません。

2サイト構成の最初の要件は、無停止のディザスタリカバリプロセスを保証する方法で、各サイトが常に半数以上の投票ファイルにアクセスできるようにすることです。このタスクは、投票ファイルがASMディスクグループに格納される前は簡単でしたが、今日の管理者はASM冗長性の基本原則を理解する必要があります。

ASMディスクグループには `external`、 `normal`、、の3つの冗長性オプションがあります。 `high`つまり、ミラーリングされていない、ミラーリングされている、3方向ミラーリングされているということです。と呼ばれる新しいオプション `Flex` も使用できますが、めったに使用されません。冗長デバイスの冗長性レベルと配置によって、障害が発生した場合の動作が制御されます。例：

- 外部冗長リソースを持つディスクグループに投票ファイルを配置すると、サイト間接続が失われた場合に一方のサイトが削除されます。
- サイトごとにASMディスクが1本しかない通常の冗長性を備えたディスクグループに投票ファイルを配置すると、どちらの サイトもマジョリティクォーラムを持たないためにサイト間接続が失われた場合に、両方のサイトでノードが削除されます。
- 冗長性の高いディスクグループに投票ファイルを配置し、一方のサイトに2本のディスクを配置し、もう一方のサイトに1本のディスクを配置すると、両方のサイトが動作していて相互にアクセスできる場合に高可用性処理が可能になります。ただし、シングルディスクサイトがネットワークから分離されている場合、そのサイトは削除されます。

RACネットワークハートビート

Oracle RACネットワークハートビートは、クラスタインターコネクト経由でノードに到達できるかどうかを監視します。クラスタに残すには、あるノードが他のノードの半数以上にアクセスできる必要があります。この要件により、2サイトアーキテクチャのRACノード数は次のように選択されます。

- サイトごとに同じ数のノードを配置すると、ネットワーク接続が失われた場合に1つのサイトが削除されます。
- 1つのサイトにN個のノードを配置し、 もう一方のサイトにN+1個のノードを配置することで、ノード数の多いサイトがネットワーククォーラムを維持し、ノード数の少ないサイトがサイト間の損失によって削除されることが保証されます。 接続性：

Oracle 12cR2より前のバージョンでは、サイト障害時にどの側で削除するかを制御することは不可能でした。各サイトのノード数が同じ場合、削除はマスターノード（通常は最初にブートするRACノード）によって制御されます。

Oracle 12cR2では、ノードの重み付け機能が導入されています。この機能により、管理者はOracleによるスプリットブレイン状態の解決方法をより細かく制御できます。簡単な例として、次のコマンドはRAC内の特定のノードの優先順位を設定します。

```
[root@jfs3-a ~]# /grid/bin/crsctl set server css_critical yes
CRS-4416: Server attribute 'CSS_CRITICAL' successfully changed. Restart Oracle High Availability
Services for new value to take effect.
```

Oracle High-Availability Servicesを再起動すると、構成は次のようになります。

```
[root@jfs3-a lib]# /grid/bin/crsctl status server -f | egrep '^NAME|CSS_CRITICAL='
NAME=jfs3-a
CSS_CRITICAL=yes
NAME=jfs3-b
CSS_CRITICAL=no
```

`jfs3-a` これで、ノードがクリティカルサーバとして指定されました。2つのRACノードが分離されている場合、jfs3-aは存続し、`jfs3-b` 削除されます。

注： 詳細については、Oracleのホワイトペーパー『Oracle Clusterware 12c Release 2 Technical Overview』を参照してください。

12cR2より前のバージョンのOracle RACでは、CRSログを確認することでマスターノードを特定できます。

```
[root@jfs3-a ~]# /grid/bin/crsctl status server -f | egrep '^NAME|CSS_CRITICAL='
NAME=jfs3-a
CSS_CRITICAL=yes
NAME=jfs3-b
CSS_CRITICAL=no
 [root@jfs3-a ~]# grep -i 'master node' /grid/diag/crs/jfs3-a/crs/trace/crsd.trc
2017-05-04 04:46:12.261525 :   CRSSE:2130671360: {1:16377:2} Master Change Event; New Master Node
ID:1 This Node's ID:1
2017-05-04 05:01:24.979716 :   CRSSE:2031576832: {1:13237:2} Master Change Event; New Master Node
ID:2 This Node's ID:1
2017-05-04 05:11:22.995707 :   CRSSE:2031576832: {1:13237:221} Master Change Event; New Master
Node ID:1 This Node's ID:1
2017-05-04 05:28:25.797860 :   CRSSE:3336529664: {1:8557:2} Master Change Event; New Master Node
ID:2 This Node's ID:1
```

このログは、マスターノードがで 2 、ノード jfs3-a のIDがであることを示します 1。この事実は、それが jfs3-a マスターノードではないことを意味します。マスターノードのIDは、コマンドで確認できます olsnodes -n。

```
[root@jfs3-a ~]# /grid/bin/olsnodes -n
jfs3-a 1
jfs3-b 2
```

IDがのノード 2 はで jfs3-b、マスターノードです。各サイトに同じ数のノードがある構成では、jfs3-b 何らかの理由で2つのセットのネットワーク接続が失われた場合でも、を含むサイトが稼働します。

マスターノードを識別するログエントリがシステムから期限切れになる可能性があります。この場合、Oracle Cluster Registry（OCR）バックアップのタイムスタンプを使用できます。

```
[root@jfs3-a ~]# /grid/bin/ocrconfig -showbackup

jfs3-b     2017/05/05 05:39:53     /grid/cdata/jfs3-cluster/backup00.ocr     0

jfs3-b     2017/05/05 01:39:53     /grid/cdata/jfs3-cluster/backup01.ocr     0

jfs3-b     2017/05/04 21:39:52     /grid/cdata/jfs3-cluster/backup02.ocr     0

jfs3-a     2017/05/04 02:05:36     /grid/cdata/jfs3-cluster/day.ocr     0

jfs3-a     2017/04/22 02:05:17     /grid/cdata/jfs3-cluster/week.ocr     0
```

次の例は、マスターノードがであることを示してい jfs3-bます。また、jfs3-a jfs3-b 5月4日の2時05分から21時39分までの間にマスターノードがからに変更されたことを示しています。マスターノードを識別するこの方法は、前回のOCRバックアップ以降にマスターノードが変更されている可能性があるため、CRSログもチェックされている場合にのみ使用できます。この変更が発生した場合は、OCRログに表示されます。

ほとんどのお客様は、環境全体と各サイトで同数のRACノードにサービスを提供する単一の投票ディスクグループを選択しています。ディスクグループは、データベースが格納されているサイトに配置する必要があります。接続が失われると、リモートサイトが削除されます。リモートサイトにはクォーラムがなくなり、データベースファイルにもアクセスできなくなりますが、ローカルサイトは通常どおり稼働し続けます。接続が回復したら、リモートインスタンスを再びオンラインにすることができます。

災害が発生した場合は、サバイバーサイトでデータベースファイルと投票ディスクグループをオンラインにするためにスイッチオーバーが必要です。災害によってAUSOでスイッチオーバーがトリガーされた場合、クラスタが同期されていてストレージリソースが正常にオンラインになるため、NVFAILはトリガーされません。AUSOは高速処理であり、disk timeout 期限が切れる前に完了する必要があります。

サイトが2つしかないため、自動化された外部タイブレークソフトウェアを使用することは不可能であり、強制スイッチオーバーは手動で行う必要があります。

3サイト構成

3つのサイトで拡張RACクラスタを構築する方がはるかに簡単です。MetroClusterシステムの各半分をホストする2つのサイトもデータベースワークロードをサポートし、3つ目のサイトはデータベースとMetroClusterシステムの両方のTiebreakerとして機能します。MetroClusterタイブレークの詳細については、「MetroCluster Tiebreaker with Fabric-attached MetroCluster」のセクションを参照してください。Oracle Tiebreakerの構成は、投票に使用するASMディスクグループのメンバーを3つ目のサイトに配置するだけです。また、3番目のサイトに運用インスタンスを配置して、RACクラスタ内のノード数を奇数にすることもできます。

メモ： 拡張RAC構成でNFSを使用する場合の重要な情報については、「クォーラム障害グループ」に関するOracleのドキュメントを参照してください。要するに、softオプションを含めるようにNFSのマウントオプションを変更しなければならない場合があります。この変更は、（3番目のサイトでホストされている）クォーラムリソースへの接続が失われても、プライマリOracleサーバまたはOracle RACプロセスが停止しないようにするためのものです。

拡張RACおよびNVFAIL

手動で強制NVFAILを使用した拡張RAC

拡張RACクラスタでスイッチオーバーを強制する最も安全なオプションは、`-force-nvfail- all` コマンドラインでを指定することです。このオプションは、キャッシュされたすべてのデータが確実にフラッシュされるようにするための緊急措置として使用できます。災害が発生したサイトにあるストレージリソースをホストが使用している場合、I/Oエラーまたは古いファイルハンドル（`ESTALE`）エラーが発生します。Oracleデータベースがクラッシュし、ファイルシステムが完全にオフラインになるか、読み取り専用モードに切り替わります。

スイッチオーバーが完了したら、`in-nvfailed-state` フラグをクリアしてLUNをオンラインにする必要があります。このアクティビティが完了したら、データベースを再起動できます。これらのタスクを自動化してRTOを短縮できます。

dr-force-nvfailを使用した拡張RAC

最も安全な構成は `dr-force-nvfail`、リモートサイトからアクセスする可能性のあるすべてのボリュームにフラグを設定することです。そのため、`in-nvfailed-state` スイッチオーバー中はボリュームを使用できなくなります。スイッチオーバーが完了したら、`in-nvfailed-state` フラグをクリアしてLUNをオンラインにする必要があります。これらのアクティビティが完了したら、データベースを再起動できます。
これらのタスクを自動化してRTOを短縮できます。

結果は `-force-nvfail-all` フラグを使用する場合と似ています。ただし、影響を受けるボリュームの数は、古いキャッシュを使用するアプリケーションまたはオペレーティングシステムから保護する必要があるボリュームだけに制限される場合があります。

dr-force-nvfailを使用しない拡張RAC

`dr-force-nvfail` データベースボリュームでを使用しない環境には、次の2つの重要な要件があります。

- 強制スイッチオーバーは、プライマリサイトの障害から30秒以内に実行する必要があります。
- メンテナンスタスクの実行中や、SyncMirrorプレックスやNVRAMレプリケーションが同期されていないその他の状況では、スイッチオーバーを実行しないでください。

最初の要件を満たすには、Tiebreakerソフトウェアを使用します。Tiebreakerソフトウェアは、サイト障害から30秒以内にスイッチオーバーを実行するように設定されています。これは、サイト障害が検出されてから30秒以内にスイッチオーバーを実行する必要があるという意味ではありません。これは、サイトが動作し

ていることが確認されてから30秒が経過した場合に強制的にスイッチオーバーを実行しても安全ではないことを意味します。

2つ目の要件は、MetroCluster構成が同期されていないことが判明した場合に、自動スイッチオーバー機能をすべて無効にすることで部分的に満たすことができます。NVRAMレプリケーションとSyncMirrorプレックスの健常性を監視できるTiebreaker解決策を使用することを推奨します。クラスタが完全に同期されていない場合、Tiebreakerはスイッチオーバーをトリガーしません。

NetApp MCTBソフトウェアは同期ステータスを監視できないため、何らかの理由でMetroClusterが同期されていない場合は無効にする必要があります。ClusterLionにはNVRAM監視機能とプレックス監視機能が搭載されており、MetroClusterシステムが完全に同期されていることが確認されないかぎり、スイッチオーバーをトリガーしないように設定できます。

詳細情報の入手方法

このドキュメントに記載されている情報の詳細については、以下のドキュメントやWebサイトを確認してください。

- MetroClusterのドキュメント リソース
  https://www.netapp.com/support-and-training/documentation/metrocluster/
- ONTAP 9ドキュメント センター
  https://docs.netapp.com/ontap-9/index.jsp
- NetAppの製品ドキュメント
  https://www.netapp.com/support-and-training/documentation/

本ドキュメントに記載されている製品や機能のバージョンがお客様の環境でサポートされるかどうかについては、NetApp サポート サイトで Interoperability Matrix Tool (IMT)を参照してください。NetApp IMT には、NetApp がサポートする構成を構築するために使用できる製品コンポーネントやバージョンが定義されています。サポートの可否は、お客様の実際のインストール環境が公表されている仕様に従っているかどうかによって異なります。

機械翻訳に関する免責事項

原文は英語で作成されました。英語と日本語訳の間に不一致がある場合には、英語の内容が優先されます。公式な情報については、本資料の英語版を参照してください。翻訳によって生じた矛盾や不一致は、法令の順守や施行に対していかなる拘束力も法的な効力も持ちません。

著作権に関する情報

商標に関する情報

TR-4592-0421-JP

NetApp