テクニカル レポート

NetApp MetroCluster
解決策のアーキテクチャと設計

NetApp
Stephen B. Galla
2023年4月| TR-4705

概要

このドキュメントでは、NetApp ONTAP® 9.12.1ストレージ管理ソフトウェアのNetApp® MetroCluster機能のアーキテクチャの概要と設計の概念について説明します。

<<本レポートは機械翻訳による参考訳です。公式な内容はオリジナルである英語版をご確認ください。>>

目次

© 2024 NetApp, Inc. All rights reserved.

表一覧

図一覧

© 2024 NetApp, Inc. All rights reserved.

MetroClusterの概要

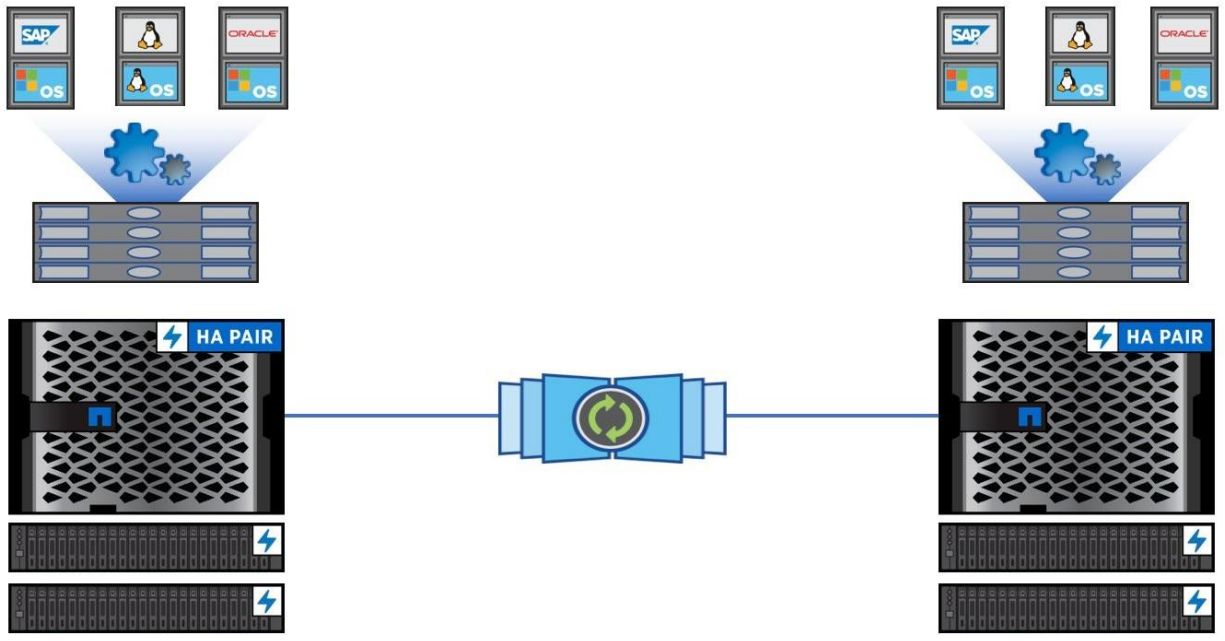
エンタープライズクラスのお客様は、コストと運用効率を維持しながら、増大するサービスレベルのニーズに対応する必要があります。データ量が急増し、共有仮想インフラに移行するアプリケーションが増えるにつれて、ミッションクリティカルなアプリケーションをはじめとするビジネスアプリケーションの可用性を継続的に維持する必要性は飛躍的に高まっています。

何百ものビジネス クリティカルなアプリケーションを実行する、高度に仮想化されたインフラを持つ環境で、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では、データ損失をゼロに抑え、システムのリカバリを数時間ではなく数分で完了することが求められます。この要件は、プライベート クラウドとパブリック クラウドの両方のインフラのほか、両者を橋渡しするハイブリッド クラウド インフラにも当てはまります。

NetApp 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、継続的な可用性を実現し、データ損失をゼロにする解決策です。アレイベースのクラスタは、通常ホストベースのクラスタリングに関連する依存関係や複雑さが排除されるため、管理がシンプルになります。MetroClusterは、トランザクションごとにすべてのミッションクリティカルなデータを即座に複製し、アプリケーションやデータへの中断のないアクセスを提供します。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、継続的なデータ可用性を実現します。また、複雑なフェイルオーバースクリプトを作成して管理する必要もありません。MetroClusterにより、次のことが可能になります。

- 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。
- 計画的停止と計画外停止、変更管理を排除
- 運用を中断せずにハードウェアとソフトウェアをアップグレード
- 複雑なスクリプト作成、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できます。
- VMware、Microsoft、Oracle、SAPなどの重要なアプリケーションの継続的可用性を実現します。

図1）MetroCluster



© 2024 NetApp, Inc. All rights reserved.

NetApp MetroClusterは、NetAppハードウェアとONTAPストレージ ソフトウェアに対する組み込みのHigh Availability（HA;高可用性）とノンストップ オペレーションを強化することで、ストレージとホストの環境全体を保護するさらなるレイヤとして機能します。スタンドアロンサーバ、HAサーバクラスタ、仮想サーバのいずれで構成されている環境でも、MetroClusterは、サイトストレージが停止してもアプリケーションの可用性をシームレスに維持します。停止は、電源、冷却機能、またはネットワーク接続の喪失のほか、ストレージアレイのシャットダウンや動作エラーが原因で発生する可能性があります。

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、サーバのリブート、またはアプリケーションの再起動は必要ありません。MetroClusterは、障害が発生しても自らのIDを維持するため、スイッチオーバーやスイッチバック時にアプリケーションを透過的に処理できます。実際、MetroClusterを使用するほとんどのお客様は、クラスタのリカバリ時にユーザがアプリケーションの中断を経験することはないと報告しています。MetroClusterは優れた柔軟性を発揮し、異種プロトコルをサポートして、あらゆる環境でシームレスな統合を実現します。

MetroClusterには、次のような利点があります。

- 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昇格により、きわめて厳しいSLA（RPO（目標復旧時点）= 0、RTO（目標復旧時間）< 2分）を達成
- SANおよびNASの幅広いクライアントプロトコルとホストプロトコルに対応するマルチプロトコルサポート。
- FCまたはIPネットワーク経由での同期レプリケーションがサポートされます。
- MetroCluster機能は無料です。
- 重要なデータのみをミラーする：ミラーされたアグリゲートとミラーされていないアグリゲートのサポート。
- Foreign LUN Import（FLI）により、サードパーティ製ストレージを簡単にインポートできます。
- 重複排除、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が向上します。
- NetApp SnapMirror® テクノロジとの統合により、非同期レプリケーション、距離、SLAの要件をサポート

対象

このドキュメントは、セールス エンジニア、フィールド コンサルタント、プロフェッショナル サービス担当者、ITマネージャー、パートナー エンジニアのほか、耐障害性と簡易性に優れたインフラを活用したいとお考えのお客様を対象としていますが、これに限定されません。

NetApp SyncMirrorテクノロジによるデータ保護

同期レプリケーションとは、簡単に言うと、変更をミラー元とミラー先の両方のストレージに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すると、同期的にミラーされるストレージのRedoログにデータが書き込まれます。ストレージシステムは、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確認応答しないでください。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にすぎません。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。お客様からRPO（データ損失ゼロ）をゼロにする解決策を求められた場合は、障害のシナリオを検討する必要があります。サイト間の接続が失われてレプリケーションが不可能な場合に予想される結果を特定する必要があります。

MetroClusterを使用した真のHAデータセンター

MetroClusterレプリケーションは、RAIDレベルで実装されるデータの同期ミラーリングを提供するNetApp SyncMirror®テクノロジに基づいています。SyncMirrorを使用して、同じNetApp 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。この2つのコピーはプレックスと呼ばれ、同時に更新され、常に同一になります。このテクノロジは、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ています。

© 2024 NetApp, Inc. All rights reserved.

注： サイト間のすべての接続が切断される部分的な障害が発生した場合、ストレージシステムはレプリケートされていない状態で動作を継続できます。

MetroClusterは、重要なビジネスアプリケーションを24時間体制で運用する必要がある組織に最適です。同じデータセンター内、建物間、キャンパス内、または地域間に配置されたNetApp AFFシステムとFASハイブリッドシステムの間でデータを同期的に複製することで、MetroClusterは、Disaster Recovery（DR;ディザスタ リカバリ）とビジネス継続性の戦略に透過的に適合します。

キャンパス、メトロ、地域の保護

また、NetApp MetroClusterは、サイト間の距離が最大700kmであることが検証されているため、キャンパス全体またはメトロポリタン全体のHAソリューションの設計、導入、およびメンテナンスが大幅に簡素化されます。サイト全体が停止している間、データ サービスは、自動化された1つコマンドを使用してわずか数秒でセカンダリ サイトにリストアされます。複雑なフェイルオーバー スクリプトや再起動の操作は必要ありません。

お客様が選択可能な保護機能

新たなレベルの柔軟性を実感し、幅広い選択肢から選んでビジネス継続性を実現できます。ONTAP 9ソフトウェアとともに導入すると、AFFコントローラとFASコントローラが混在している場合でも、MetroClusterを2ノードから4ノード、8ノードクラスタ（レプリケーションの両端に4ノード）に拡張できます。4ノード構成から8ノード構成へのスケールアップは、システム停止を伴わないプロセスです。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。

FCネットワークを介した同期レプリケーションは、2ノード構成、4ノード構成、および8ノード構成でサポートされます。IPネットワーク経由の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。

WANベースのDR

業務がメトロポリタン エリアを超えて地理的に分散している場合は、NetApp SnapMirrorソフトウェアを追加して、グローバル ネットワーク全体にデータを簡単かつ確実に複製できます。NetApp SnapMirrorソフトウェアは、MetroClusterソリューションと連携し、WAN接続を介してデータを高速で複製し、重要なアプリケーションを地域のシステム停止から保護します。

管理の簡易化：一度設定

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個別に管理されるため、ストレージ システムの管理、構成、メンテナンスで、重複する作業が求められます。この重複により管理オーバーヘッドが増えるうえ、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ことになります。MetroClusterは真のクラスタストレージ解決策であるため、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るため、重複する管理作業が不要になり、構成の一貫性が維持されます。

アプリケーション透過性

MetroClusterは、透過的で、あらゆるフロントエンド アプリケーション環境に対応できるように設計されており、アプリケーション、ホスト、クライアントに変更が必要になることはほとんどありません。スイッチオーバーの前後で接続パスが同じであり、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）はストレージの再接続や再検出を必要とせず、代わりに自動的に再開されます。SMBアプリケーション（共有の継続的可用性を備えたSMB 3を含む）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プロトコルの制限事項です。

© 2024 NetApp, Inc. All rights reserved.

アーキテクチャ

NetApp MetroClusterは、ストレージ インフラおよび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してい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、管理不要の継続的な可用性を実現し、アレイ内外の障害から保護します。

MetroCluster物理アーキテクチャ

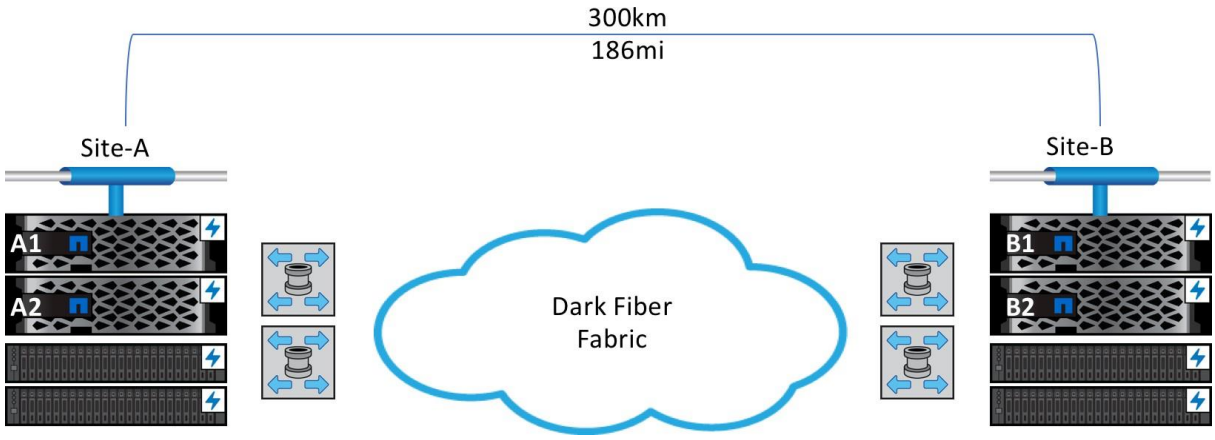
MetroCluster構成では、最大700km離れた2つのクラスタを使用してデータを保護します。それぞれのクラスタが、もう一方のデータおよび構成情報を同期的にミラーします。具体的には、すべてのStorage Virtual Machine（SVM）および関連する構成がレプリケートされます。クラスタを独立させることにより、データの分離と論理エラーに対する耐障害性が実現します。

一方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。これによってミラーされたSVMがアクティブになり、稼働しているサバイバ サイトから、ミラーされたデータの提供が再開されます。clustered Data ONTAP® 8.3以降では、MetroClusterの4ノード構成は、各サイトに2ノードのHAペアで構成されます。この構成では、ほとんどの計画的イベントと計画外イベントを、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で処理できます。もう一方のサイトへの完全なスイッチオーバーが必要となるのは、災害時またはテスト目的の場合のみです。スイッチオーバーとそれに続くスイッチバック処理では、クラスタ ワークロード全体がサイト間で転送されます。

MetroClusterの2ノード構成では、各サイトに1ノード クラスタがあります。計画的 / 計画外イベントは、スイッチオーバーとスイッチバックを使用して処理されます。スイッチオーバーとそれに続くスイッチバック処理では、クラスタ ワークロード全体がサイト間で転送されます。

次の図は、基本的な4ノードMetroCluster構成を示しています。データセンターAとBは、FCの場合は最大300km、IPの場合は最大700kmの距離でスイッチ間リンク（ISL）で接続されます。各サイトのクラスタは、HAペアの2つのノードで構成されています。このレポートでは、この構成と命名規則を使用します。

図2）4ノードのMetroCluster FC環境



2つのクラスタおよびサイトは2つの独立したネットワークによって接続され、レプリケーションが転送されます。クラスタ ピアリング ネットワークはIPネットワークで、サイト間のクラスタ構成情報のレプリケートに使用されます。共有ストレージファブリックはFC接続またはIP接続で、2つのクラスタ間のストレージおよびNVRAMの同期レプリケーションに使用されます。MetroCluster IPの場合、レプリケーションではNVRAMにiWARP、ディスクレプリケーションにiSCSIの両方が使用されます。すべてのストレージは、共有ストレージ ファブリックを介してすべてのコントローラで認識されます。

© 2024 NetApp, Inc. All rights reserved.

注意： iWARP (Internet Wide Area RDMA Protocol)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)を可能にするネットワークプロトコルです。サーバ、ストレージシステム、およびその他のネットワークデバイス間で高速かつ低遅延のデータ転送を可能にし、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。

MetroCluster FCとMetroCluster IPの比較

MetroCluster IPはONTAP 9.3で導入され、FC ISLを使用するMetroCluster FCとは異なり、ファブリックにはEthernet/IP ISLを使用します。さらに、MetroCluster IPクラスタでは、NVRAMレプリケーションとSyncMirrorレプリケーションの両方に高速イーサネットが使用されます。

MetroCluster IPには、他のMetroCluster以外のトラフィック（レイヤ2共有VLAN、レイヤ3-VIP/BGP）と共有されるサイト間リンクの使用など、運用コストを削減する機能がいくつかあります。ONTAP 9.7以降では、専用のスイッチがなくてもMetroCluster IPが提供されるため、MetroCluster IPの要件に準拠していれば既存のスイッチを使用できます。詳細については、『MetroCluster IPインストールおよび設定ガイド』を参照してください。

表1は、これら2つの構成の違いをまとめたもので、2つのMetroClusterサイト間でデータをレプリケートする方法を示しています。

表1）MetroCluster FCとMetroCluster IPの比較

機能	MetroCluster FC	MetroCluster IP
MetroClusterファブリック	FC ISL	Ethernet/IP ISL
ファブリック ファイバ スイッチ	1サイトに2つ	なし
SASブリッジ	1サイトに2つ	なし
ファブリック イーサネット スイッチ	なし	1サイトに2つ
FC-VIアダプタ	はい アダプタの数はコントローラに依存	なし
ファブリック25G / 40G / 100G イーサネット アダプタ	なし	プラットフォームに応じてノードごとに1つ。 アダプタはiWARPとiSCSIの両方のレプリケーションに使用される
クラスタ間	スイッチレスおよびスイッチ使用	スイッチレスおよびスイッチ使用
シェルフ	両方のサイトから物理的に見える	リモート クラスタには見えない
NVRAMレプリケーション	FCプロトコル	IP/iWARP
SyncMirrorレプリケーション	FCプロトコル	IP/iSCSI
構成レプリケーション サービス	変更なし	変更なし
MetroClusterサイズ	2ノード、4ノード、8ノード	4ノードと8ノード
MetroClusterストレッチ	はい	いいえ
アドバンスト ディスク パーティショニング	いいえ	○（AFFのみ）

ローカルフェイルオーバー（HA）とリモートスイッチオーバー（DR）

2ノード アーキテクチャのHAフェイルオーバーとリモートDRは、どちらもMetroClusterのスイッチオーバーとスイッチバックの機能を使用して実施されます。各ノードは、そのピアのHAパートナーおよびDRパートナーの両方として機能します。NVRAMは、4ノード構成のようにリモートパートナーにレプリケートされます。

4ノードアーキテクチャと8ノードアーキテクチャは、ローカルのHAフェイルオーバーとリモートのDRスイッチオーバーの両方を提供します。各ノードには、同じローカルクラスタにHAパートナーがあり、リモートクラスタにDRパートナーがあります（図3を参照）。A1とA2、およびB1とB2は、HAパートナーです。A1とB1、およびA2とB2は、DRパートナーです。NVRAMはHAとDRパートナーの両方にレプリケートされます。

© 2024 NetApp, Inc. All rights reserved.

詳細については、「キャンパス、メトロ、および地域の保護」の項を参照してください。ノードのDRパートナーはMetroClusterの設定時に自動的に選択され、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するようにクラスタを設定する前に、システムIDを書き留めておく必要があります。

図3）HAグループとDRグループ

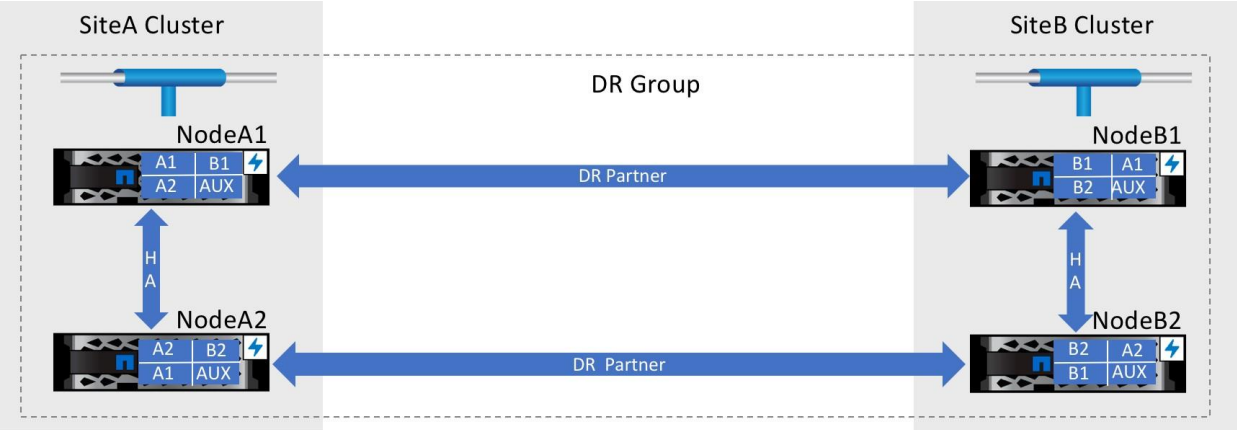
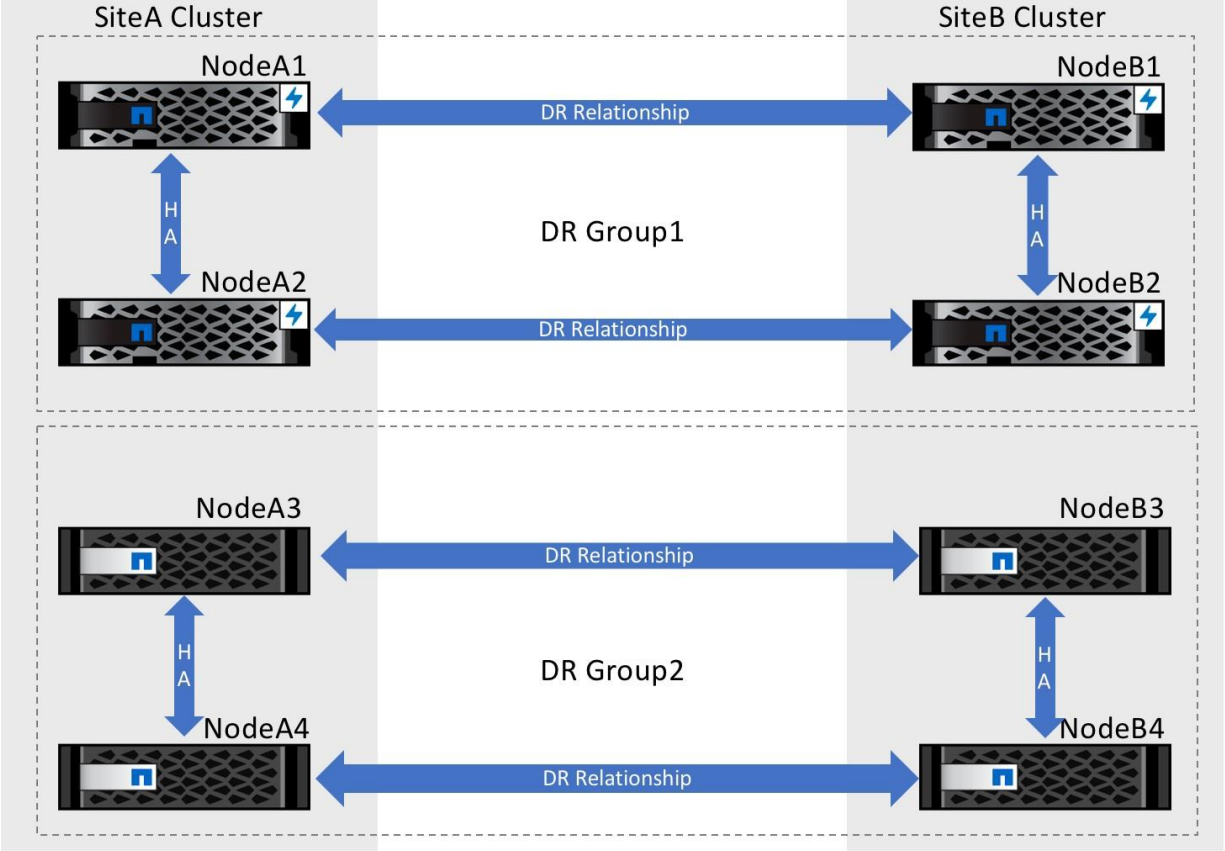


図4は、8ノードのMetroCluster構成とDRグループの関係を示しています。8ノード環境では、2つの独立したDRグループがあります。DRグループ内のハードウェアは、サイトAとサイトBで同じである必要があります。ただし、DRグループ1のハードウェアは、DRグループ2のハードウェアと一致する必要はありません。たとえば、グループ1でAFF A700を使用し、グループ2でFAS8200を使用するなど、ハードウェアはDRグループごとに異なる場合があります。

© 2024 NetApp, Inc. All rights reserved.

図4）8ノードのDRグループ



ローカルのHAフェイルオーバーでは、HAペアの一方のノードがHAパートナーの共有ストレージとサービスを一時的にテイクオーバーします。たとえば、ノードA2がノードA1のリソースをテイクオーバーします。テイクオーバーは、ミラーされたNVRAMおよび2つのノード間のマルチパス ストレージによって実現します。フェイルオーバーは、ONTAPのアップグレードを無停止で行うときなどに計画的に行う場合と、パニックやハードウェア障害のように計画外で発生する場合があります。ギブバックは逆方向のプロセスで、障害が発生したノードが、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的に実施されます。フェイルオーバーは常にローカルのHAパートナーに対して行われ、どちらのノードも相互にフェイルオーバーできます。

スイッチオーバー中、ピアクラスタは自身のワークロードを実行しながら、もう一方のクラスタのストレージとサービスをテイクオーバーします。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のノードが一時的にクラスタAが所有していたストレージとサービスの制御を引き継ぎます。スイッチオーバーが完了すると、クラスタAのSVMがオンラインに戻り、クラスタBで引き続き実行できます。

スイッチオーバーは、テストやサイトのメンテナンス目的で事前に決めて実施することも（計画的）、災害によって一方のサイトが使用できなくなった場合に強制的に実行（計画外）することもできます。スイッチバックは、スイッチオーバーされたリソースを稼働しているクラスタから元の場所へ戻し、安定した運用状態をリストアするプロセスです。スイッチバックは2つのクラスタ間で調整されて実施されるため、常に計画的処理です。どちらのサイトからももう一方のサイトにスイッチオーバーできます。

サイトがスイッチオーバー状態のときに障害が続けて発生する可能性もあります。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に障害が発生した場合は、B2が自動的にテイクオーバーし、すべてのワークロードに対応します。

© 2024 NetApp, Inc. All rights reserved.

MetroClusterレプリケーション

MetroCluster IPは直接接続型ストレージを利用するため、外付けのSerial-Attached SCSI（SAS；シリアル接続SCSI）ブリッジを使用してディスクをストレージファブリックに接続する必要がありません。ディザスタ リカバリ グループの各ノードは、グループ内の他のノードにディスク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、iSCSIイニシエータとターゲットがTCP/IPファブリックを介して通信できるようにするIPファブリックのストレージ転送プロトコルです。ディザスタ リカバリ グループの各ノードは、リモート ディザスタ リカバリ パートナーのiSCSI ターゲットとのiSCSIセッションを確立するiSCSIイニシエータを介してリモート 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、内蔵ディスクを備えたシステムを使用することもできます。iSCSIを使用すると、ノードは、外付けディスク シェルフに配置されたストレージ デバイスに加えて、内蔵ストレージへのディザスタ リカバリ パートナー ノードへのアクセスも可能です。

MetroClusterには、次の3つのレプリケーション プレーンがあります。

1. 設定レプリケーション
2. NVRAMレプリケーション
3. ストレージ レプリケーション

設定レプリケーション

MetroCluster構成は、2つのONTAPクラスタで構成されます。各クラスタには、独自のメタデータまたは設定情報を格納するレプリケートされたデータベース（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サバイバークラスタでアクティブ化されます。そのためには、これらのオブジェクトを所有権を持つクラスタからもう一方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、Configuration Replication Service（CRS；設定レプリケーションサービス）、メタデータを格納するボリュームの3つのコンポーネントがあります。

- **クラスタピアリングと** は、クラスタ間論理インターフェイス（LIF）を使用して、2つのONTAPクラスタ間にお客様指定のTCP/IP接続を確立する方法です。クラスタ間で構成オブジェクトをレプリケートでき、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、通常はフロントエンドまたはホスト側のネットワークであり、Storage Virtual Machine（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経由で実行されます。
- **Configuration Replication Service（CRS；設定レプリケーションサービス）** は、各クラスタで実行されるMetroCluster設定のコンポーネントで、必要なメタデータオブジェクトを所有者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、構成オブジェクト（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（CIFS、NFS、SANなど）をピアリングネットワークを使用してクラスタ間で共有します。CRSに影響するクラスタピアリングネットワークが中断された場合、接続が再確立されるとレプリケーションは自動的に停止します。CRSでは、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が必要です。
- 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を設定すると、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、別々のルート以外のアグリゲートに作成する必要があるため、MetroClusterを設定する前に、各クラスタで少なくとも2つのデータアグリゲートを使用することを推奨します。メタデータを含むボリュームは耐障害性を提供し、オブジェクトが作成または更新されるたびに更新がログに記録されます。変更はロギングが完了した時点でローカルのRDBにコミットされ、更新は、設定レプリケーションネットワーク経由でもう一方のクラスタのRDBに同期的に伝播されます。設定レプリケーション ネットワークの一時的エラーにより変更が伝播できない場合、接続のリストア後に、変更が自動的にもう一方のクラスタに送信されます。

一方のクラスタの構成に対する変更は、もう一方のクラスタに自動的に反映されるため、データや構成を失うことなくスイッチオーバーを実現できます。更新は自動で行われ、MetroCluster構成に固有の管理作業

© 2024 NetApp, Inc. All rights reserved.

はほとんど必要ありません。設定レプリケーションネットワークの一時的なエラーが原因で変更を伝播できない場合、接続のリストア後に変更が自動的にもう一方のクラスタに送信されます。耐障害性を高めるために、クラスタ構成ネットワークに冗長なネットワークを使用することを推奨します。

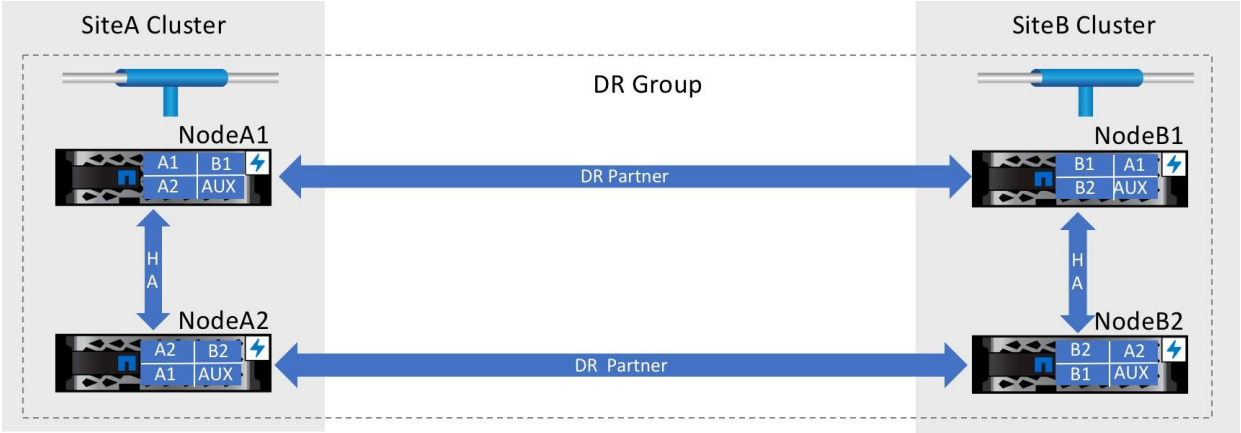
次の例では、MDVにシステムによって名前が割り当てられており、各クラスタで表示されていることがわかります。最初の2つのボリュームは、コマンドがクラスタAから実行されたため、状態が「online」のローカルMDVです。残りの2つのMDVは、ホストしているアグリゲートで示されるクラスタBに属し、スイッチオーバーが実行されないかぎり、現在オフラインになっています。

```
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            -          -     -
```

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のデータ損失を防ぐために、ローカルノードのNVRAMをリモートディザスタリカバリノードのNVRAMにコピーするプロセスです。ONTAP HAペアでは、各ノードのNVRAMがHAインターコネクトを介してもう一方のノードにミラーされます。NVRAMは、各ノードのNVRAM用に1つずつ、2つのセグメントに分割されます。図5は、MetroClusterがもう一方のサイトにDRパートナーノードを配置して追加のミラーリングを提供し、NVRAMがスイッチ間リンク（ISL）接続を介してDRパートナーにミラーリングされる様子を示しています。4ノード構成では、各ノードのNVRAMが2回（HAパートナーとDRパートナーに1回）ミラーされ、各ノードのNVRAMは4つのセグメントに分割されます。

図5）NVRAMの割り当て



書き込み処理はまず不揮発性メモリ（NVRAM）にステージングされてからディスクに書き込まれ、すべてのNVRAMセグメントが更新されたあとにのみ、実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の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され、パートナーサイトのディザスタリカバリ（DR）パートナーにリモートでミラーされます。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。ローカルでHAテイクオーバーが発生した場合、DR補助パートナーに自動的に切り替えて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で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をノードB1にミラーリングできないため、DR補助パートナーであるノードB2にミラーリングされます。

© 2024 NetApp, Inc. All rights reserved.

DRパートナーのNVRAMへの更新は、MetroCluster FCの場合はFC-VI接続を経由し、MetroCluster IPに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、IPスタックの影響によるレイテンシを最小限に抑えます。スイッチのQuality of Service（QoS；サービス品質）を使用すると、FC-VIおよびiWARPトラフィックがストレージレプリケーションよりも優先されます。ただし、ISLのレイテンシが上昇すると、DRパートナーのNVRAMへの書き込み確認に時間がかかるため、書き込みパフォーマンスに影響する可能性があります。一時的にサイトが分離された場合（すべてのISLが停止し、リモートノードが応答しない場合など）でもローカルの処理を継続できるようにするために、システムタイムアウト後に書き込みが確認されます。少なくとも1つのISLが使用可能になると、リモートのNVRAMミラーは自動的に再同期されます。

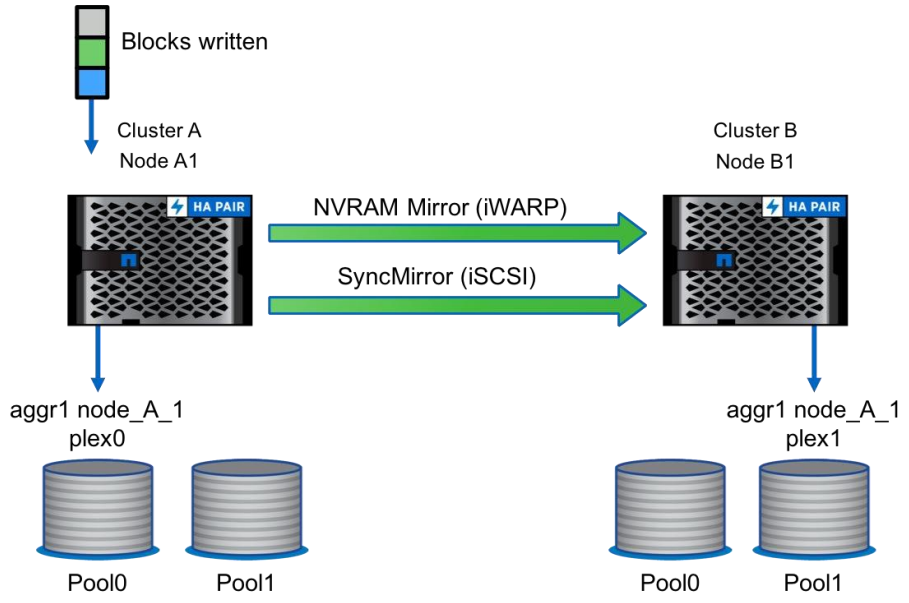
データ損失を防ぐために、NVRAMトランザクションは少なくとも10秒に1回、整合ポイントを使用してディスクにコミットされます。WAFLは、コントローラのブート時にディスク上の最新の整合ポイントを使用するため、停電やシステム障害の発生後に長時間をかけてファイルシステムをチェックする必要がありません。最新の整合ポイントのあとに発生したデータI/O要求が失われないように、ストレージ システムではバッテリでバックアップされたNVRAMを使用します。テイクオーバーまたはスイッチオーバーが発生すると、コミットされていないトランザクションがミラーNVRAMから再生され、データ損失が回避されます。

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、ローカルおよびリモートのバックエンド ディスクをミラーリングします。MetroCluster IPは、ディザスタ リカバリ 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で、バックエンド ストレージを論理的に共有します。ノードがリモート バックエンド ディスクにアクセスするには、リモート 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ディスクにアクセスします。

図6 は、NVRAMとストレージのレプリケーションのMetroCluster IPプレーンを示しています。ノードB1 は、ローカルに接続されたディスクを、iSCSIターゲットを介してディザスタ リカバリ 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。NodeA1 pool0ディスクはNodeA1にローカルに接続され、pool1リモートディスクはB1でホストされる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 `aggr1 node_A_1` ローカルは `plex 0`、pool0からローカルに接続されたディスクで構成されます。アグリゲート `aggr1 node_A_1` リモートは `plex 1`、B1に直接接続されたディスクで構成され、B1でホストされているiSCSIターゲットを介してA1にエクスポートされます。

図6）書き込みデータブロックのミラーリング



© 2024 NetApp, Inc. All rights reserved.

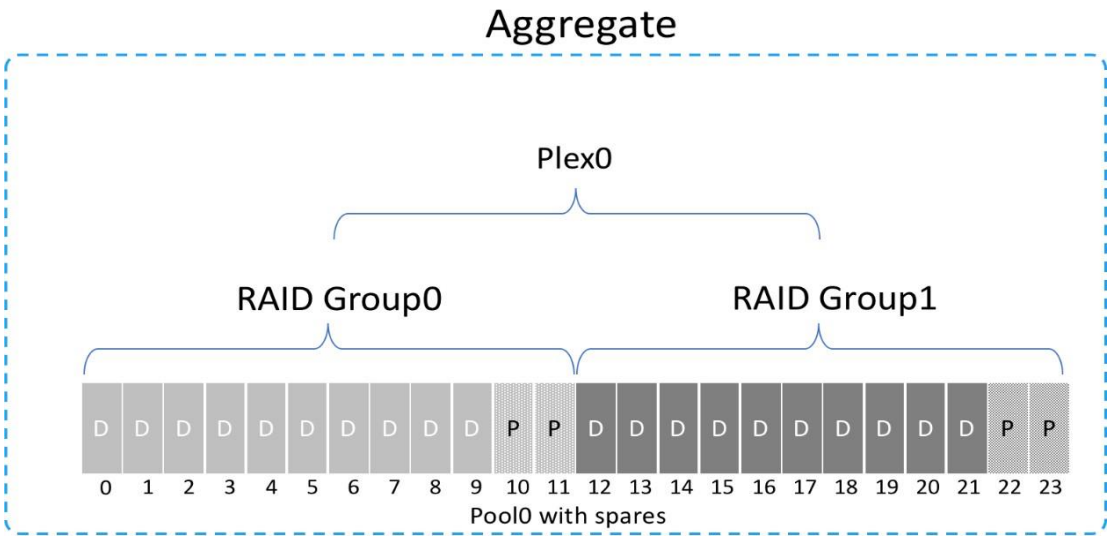
NVRAM（NVMEM）とSyncMirrorの両方を使用して、各サイトのペアノードにブロックが書き込まれます。SyncMirrorは、ミラーリングされたアグリゲートごとに2つのプレックス（ローカルプレックス1つとリモートプレックス1つ）にデータを書き込みます。SyncMirrorの書き込みはRAIDレイヤで行われます。つまり、重複排除や圧縮などのストレージの効率性によって、SyncMirror処理により書き込まれるデータが削減されます。

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

SyncMirrorストレージレプリケーション

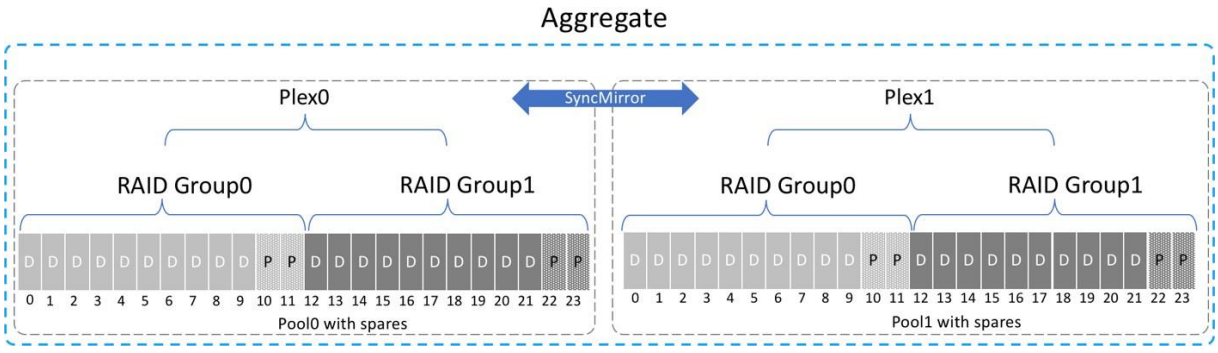
ONTAPシステムは、アグリゲートからプロビジョニングされたNetApp FlexVol® ボリュームにデータを格納します。各アグリゲートにはWAFLファイルシステムが含まれています。MetroClusterを使用しない構成では、各アグリゲート内のディスクは、プレックスと呼ばれる単一または複数のRAIDグループで構成されます（図7）。プレックスは、コントローラに接続されているローカル ストレージに存在します。

図7）ミラー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8）。MetroCluster構成では、すべてのストレージが共有され、すべてのコントローラから認識できます。ローカル プレックスにはローカル プール（pool0）のディスクのみが含まれ、リモート プレックスにはリモート プールのディスクのみ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。リモートプレックスには 0 、リモートであることを示す以外の番号が付けられます（、など plex1 plex2）。

図8）MetroClusterのミラーされたアグリゲート

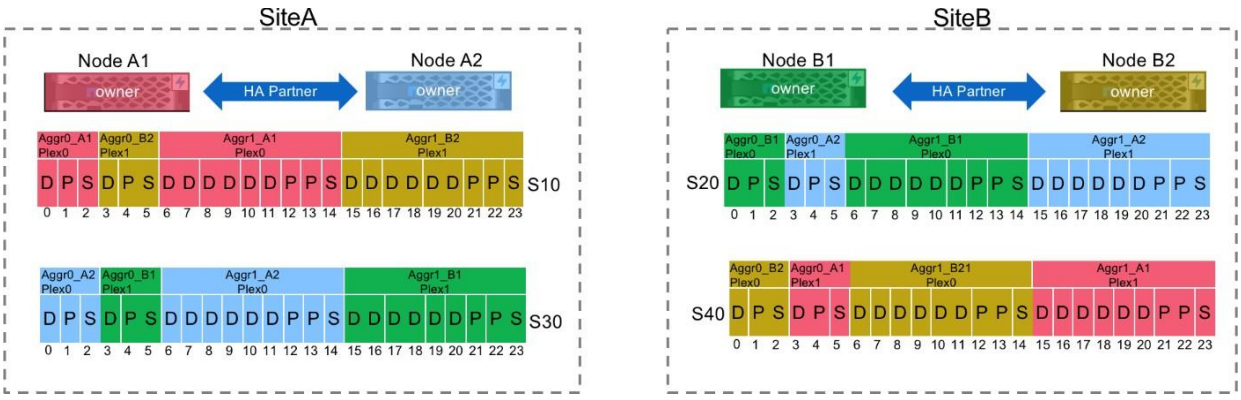


© 2024 NetApp, Inc. All rights reserved.

MetroClusterでは、ミラーアグリゲートとミラーされていないアグリゲートの両方がサポートされますミラーアグリゲートは、MetroCluster FC（ONTAP 9.0以降）およびMetroCluster IP（ONTAP 9.8以降）でサポートされます。`-mirror true` MetroClusterの設定後にアグリゲートを作成する場合は、フラグを使用する必要があります。このフラグを指定しないと `create` コマンドは失敗します。`-diskcount` パラメータで指定したディスク数は、自動的に半分になります。たとえば、使用可能なディスクが6本あるアグリゲートを作成するには、ディスク数として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本のディスクが割り当てられ、リモート プレックスにはリモート プールから6本のディスクが割り当てられます。アグリゲートにディスクを追加する際も同じです。容量として必要なディスク数の倍の数を指定する必要があります。

図9の例は、アグリゲートへのディスクの割り当て状況を示しています。各ノードには、ルート アグリゲートとデータ アグリゲートが1つずつあります。各ルート アグリゲートには、ノードごとに6本のドライブがあります。ここでは、クラスタごとに最低2つのシェルフが使用され、そのうち3つがローカル プレックス上に、3つがリモート プレックス上にあることを前提としています。したがって、アグリゲートの使用可能容量は3本のドライブです。同様に、データアグリゲートにはそれぞれ18本のドライブ（ローカルドライブ9本とリモートドライブ9本）が含まれています。MetroCluster、特にAFFでは、ルートアグリゲートはRAID 4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9）ルートアグリゲートとデータアグリゲート



MetroClusterの通常運用時は、両方のプレックスがRAIDレベルで同時に更新されます。クライアントおよびホストI/Oからかクラスタ メタデータからかに関係なく、すべての書き込みで物理的書き込み処理が2つ生成されます。1つはローカル プレックスに対して、もう1つはリモート プレックスに対してで、書き込みには2つのクラスタ間のISL接続が使用されます。デフォルトでは、読み取りはすべてローカル プレックスから実行されます。

アグリゲートのSnapshotコピー

自動アグリゲートNetApp Snapshot™ コピーが作成され、デフォルトでは、アグリゲート容量の5%がこれらのSnapshotコピー用に予約されます。これらのSnapshotコピーは、必要なときに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。

一方のプレックスが使用できなくなった場合（シェルフやストレージ アレイの障害などのため）、障害が発生したプレックスがリストアされるまでは、影響を受けていないプレックスがデータの提供を継続します。障害が発生したプレックスが修復されると、プレックスが自動的に再同期され、両方のプレックスの整合性が確保されます。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Snapshotコピーを共有している場合、このSnapshotコピーが部分的な再同期の基準として使用されます。プレックス間でSnapshotコピーが共有されていない場合は、完全な再同期が実行されます。

© 2024 NetApp, Inc. All rights reserved.

アクティブ/アクティブ構成およびアクティブ/パッシブ構成

MetroClusterは、対称型のスイッチオーバーとスイッチバックに対して自動的に有効になります。つまり、どちらのサイトで災害が発生した場合にも、そのサイトからもう一方のサイトにスイッチオーバーできるということです。したがって、両方のサイトが独立したワークロードをアクティブに処理するアクティブ / アクティブ構成が基本的な構成です。

代替構成として、アクティブ / スタンバイまたはアクティブ / パッシブがあります。この構成では、1つのクラスタ（クラスタAなど）だけがアプリケーション ワークロードを安定した状態でホストします。そのため、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のノードには、独自のミラーされたルートアグリゲートとメタデータボリュームも必要です。あとで要件が変更されてクラスタBでワークロードがプロビジョニングされ、その結果アクティブ / パッシブからアクティブ / アクティブに変更されても、MetroCluster構成の変更が必要になることはありません。いずれかのサイトで定義されたワークロード（SVM）はすべて自動的にレプリケートされ、もう一方のサイトで保護されます。

サポートされているもう1つの選択肢はHAペア内のアクティブ / パッシブ構成で、2つのうち1つのノードだけがワークロードをホストします。このオプションを使用すると、クラスタごとに必要なデータアグリゲートが1つだけの小規模構成が作成されます。

MetroClusterでは、スイッチオーバー時にストレージ アクセス パスのIDが保持されます。LIFのアドレスはスイッチオーバー後も維持され、NFSエクスポートとSMB共有は同じIPアドレスを使用してアクセスされます。また、LUNのLUN ID、WWPN（Worldwide Port Name）、またはIPアドレスとターゲット ポータル グループ タグも同じです。保持されているこのIDのため、フロントエンドのクライアントおよびホストがパスと接続を認識できるよう、フロントエンド ネットワークを両方のサイトにまたがって設定する必要があります。ホストネットワークの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はレイヤ2（共有VLAN）とレイヤ3（VIP/BGP）の両方のネットワークをサポートしています。

アドバンストドライブパーティショニング（ADP）

注： MetroClusterの場合、ADPはMetroCluster IP構成のAFFシステムでのみ使用できます。

ADPは、AFFシステムとFASシステムでHDDとSSDの両方のストレージ効率を高める機能です。ADPでは、HAペア内のアグリゲートとコントローラ間で物理ドライブの容量を共有できるため、少ないSSDで両方のノードの容量を拡張できるため、効率とコストが向上します。ADPを使用すると、ルートアグリゲートで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可能な容量が多くなります。さらに、両方のコントローラでパリティドライブとスペアドライブを共有することで、HAペアで使用可能な容量がさらに増加します。ADPは、パーティショニングされていないドライブ全体を使用するよりも効率的にストレージ容量をプロビジョニングする方法です。

Advanced ADPの利点：

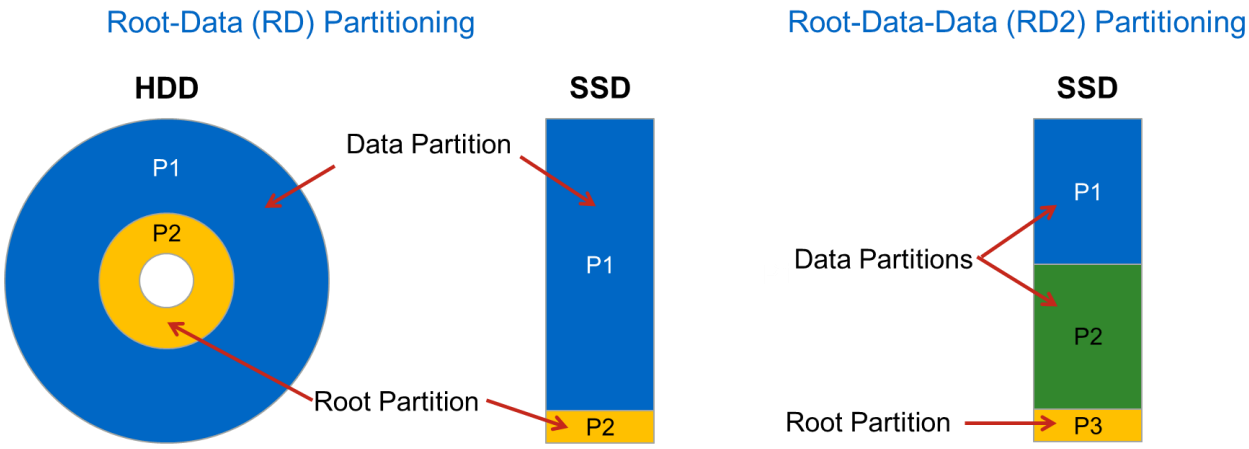
- AFFシステムとFASシステムの使用可能容量と実効容量が増加します。
- ストレージ効率を向上（ディスクパーティショニング全体と比較して10~40%の効率向上）
- 少ないSSDで両方のノードに容量を拡張できます。
- コストとストレージの効果的な競争力が向上します。パ

ーティション分割の方法：

- 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。

© 2024 NetApp, Inc. All rights reserved.

図10）ADPメソッドの論理ビュー

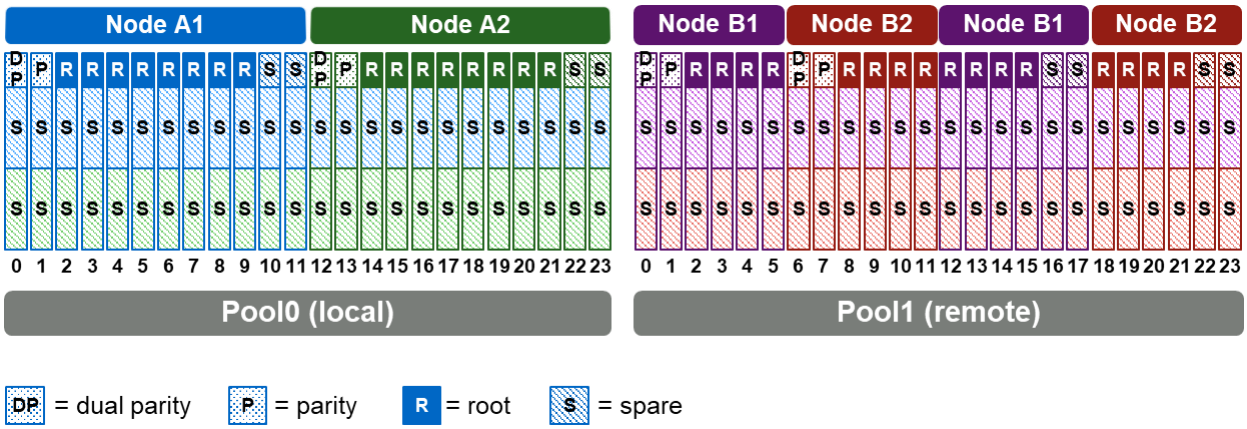


注： ONTAP 9.0以降を使用するAFFシステムのMetroCluster IP構成では、ADP RD2パーティショニングのみがサポートされます。ADPは、MetroClusterの初期化時にデフォルトで適用されます。ADPは、MetroCluster FC構成のAFFシステムではサポートされません。

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、NetApp ONTAP 9.0以降のリリースで使用できるStorage Efficiency機能です。RD2は、複数のSSDの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し、データアグリゲートに使用できる容量を増やします。各SSDは3つのパーティション（小容量（シン）ルートパーティションと2つの大容量（シック）データパーティション）に分割されます。SSDごとにデータパーティションが2つあるため、AFFまたはオールSSD FASシステムの両方のコントローラで、1本のドライブの容量とIOPSを使用できます。RD2パーティショニングで使用できるSSDの最大数は48本ですが、HAペアでは48本を超えるSSD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本で、400GB SSD以上が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、コントローラモデル、ONTAPリリース、およびシステムの初期化時に接続された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、追加のスペースが必要な場合にルートアグリゲートを拡張するためにのみ使用できます。一方、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の代わりにホットスペアとして使用できます。図12は、AFF A250×1、各サイトにNS224×1、NVMe SSDx48を搭載したMetroCluster IP構成のADPの例を示しています。

図11）48ドライブMetroCluster IP構成のADPの例



ADPの詳細については、次のドキュメントを参照してください。

© 2024 NetApp, Inc. All rights reserved.

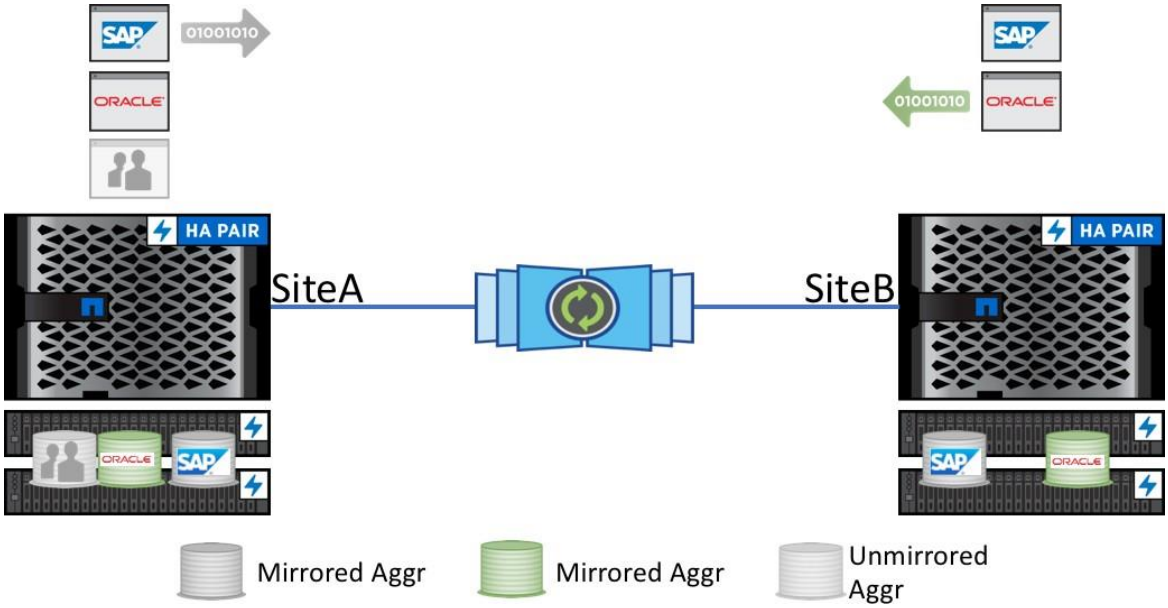
- テクニカルFAQ：アドバンストドライブパーティショニング（ADP）
- 技術プレゼンテーション：アドバンストドライブパーティショニング（ADP）

ミラーされてないアグリゲート

ストレッチおよびファブリック接続構成ではONTAP MetroCluster 9、IP構成ではONTAP 9.8以降では、MetroCluster構成で提供される冗長ミラーリングを必要としないデータについては、ミラーされていないアグリゲートがサポートされます。ミラーされていないアグリゲートは、サイト障害の際には保護されません。また、ISLのサイジング時に、これらのアグリゲートへの書き込みI/Oを考慮する必要があります。

図10は、アグリゲートのミラーリングの詳細な制御を示しています。SAPはサイトBのクラスタにミラーリングされ、OracleはサイトAのクラスタにミラーリングされます。サイトAのホーム ユーザ 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12）MetroClusterでのミラーされていないアグリゲート



MetroCluster FCでミラーされていないアグリゲートを検討する場合は、次の点に注意してください。

- MetroCluster FC構成でのミラーされていないアグリゲートは、アグリゲート内のリモート ディスクにアクセスできる場合、スイッチオーバー後にのみオンラインになります。ISLで障害が発生すると、ローカル ノードはミラーされていないリモート ディスクのデータにアクセスできなくなる可能性があります。アグリゲートで障害が発生すると、ローカルノードが原因でリブートされる可能性があります。
- ドライブとアレイLUNは、特定のノードによって所有されます。アグリゲートを作成する場合は、そのアグリゲート内のすべてのドライブが同じノード（アグリゲートのホーム ノードになるノード）によって所有されている必要があります。
- アグリゲート名は、MetroCluster構成を計画する際に決定した命名基準を満たしている必要があります。

注： ADPが設定されたディスクを使用する場合は、ノード別のパーティション所有権とプール内のメンバーシップに関する特定のルールを理解することが重要です。また、ミラーリングするドライブは、ピア関係の両側で対称にする必要があります。ミラーされていないアグリゲートにADPが設定されたパーティションを使用すると、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ために、パーティショニングされていないドライブにミラーされていないアグリゲートを導入することを強く推奨します。

© 2024 NetApp, Inc. All rights reserved.

導入オプション

MetroClusterは、各サイトに同一のハードウェアを必要とする、完全冗長構成です。さらに、MetroClusterでは、ストレッチ、ファブリック接続、IPを柔軟に構成できます。表2は、さまざまな導入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ています。

表2）ハードウェア要件

項目	IP設定	ファブリック接続構成		ストレッチファブリック構成	
		4ノードまたは8ノード	2ノード	2ノードブリッジ接続	2ノード直接接続型
コントローラ数	4または8	4または8	2	2	2
FCスイッチストレージファブリック	いいえ	はい	はい	いいえ	いいえ
IPスイッチストレージファブリック	はい	いいえ	いいえ	いいえ	いいえ
FC-to-SASブリッジ	いいえ	はい	はい	はい	はい
直接接続型ストレージ	○（ローカル接続のみ）	いいえ	いいえ	いいえ	はい
ローカルHAのサポート	はい	はい	いいえ	いいえ	いいえ
自動スイッチオーバーのサポート	○（メディエーターあり）	はい	はい	はい	はい
ミラーされていないアグリゲートのサポート	はい	はい	はい	はい	はい
アレイLUNのサポート	いいえ	はい	はい	はい	はい

ストレツチオヨヒストレツチブリッジコウセイ

MetroClusterストレッチ構成は、2つのストレージノードをより長い距離（通常は最大270m）に拡張し、ワンランク上の耐障害性とディザスタリカバリ機能を提供します。そのためには、高速リンクを使用して2つのクラスタを接続し、クラスタ間でデータを同期的にミラーリングします。一方のクラスタに障害が発生したり使用不能になったりした場合は、もう一方のクラスタがシームレスにテイクオーバーし、アプリケーションやユーザへの継続的なデータアクセスを提供します。MetroClusterストレッチブリッジ構成では、2つのプライマリクラスタ間にFC-to-SASブリッジを追加することで、ストレッチ構成を最大500mまで拡張できます。ストレッチブリッジ構成は、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より長い距離にまたがる場合に使用できます。どちらの構成もデータセンター環境に最適で、インフラ（ケーブル配線、FCスイッチ、ラックスペースなど）の必要性が軽減されています。

© 2024 NetApp, Inc. All rights reserved.

フアフリツクセツソクノFCコウセイ

MetroCluster FCでは、ファイバチャネル（FC）テクノロジを使用して、2つのサイト間（最長300km）の同期レプリケーションを実行します。MetroCluster FCは、AFFとFASの両方のプラットフォームでサポートされており、2ノード、4ノード、8ノードのアーキテクチャで導入できます。

IP設定

MetroCluster IPは、IPネットワークを使用して2つのサイト間（最大700km）の同期レプリケーションを行います。AFF、FAS、Cシリーズのプラットフォームでサポートされ ているMetroCluster IPは、4ノードおよび8ノードのアーキテクチャで導入できます。

注： MetroClusterのストレッチ構成とFC構成の詳細については、TR-4375：『NetApp MetroCluster FC for ONTAP 9.8』を参照してください。

注：MetroCluster IP構成の詳細については、TR-4689：『MetroCluster IP -解決策Architecture and Design』を参照してください。

注：MetroClusterでサポートされる最新のハードウェア、ソフトウェア、サイジング、制限については、Interoperability Matrix Tool、Hardware Universe、Fusionを参照してください。

計画的イベントと計画外イベントの耐障害性

このセクションでは、障害や災害の種類と、MetroCluster構成で可用性、データ保護、修復を維持する方法について説明します。

単一ノード障害

ローカルのHAペアの1コンポーネントで障害が発生した場合を考えてみてください。4ノードMetroCluster構成では、障害が発生したコンポーネントによっては、障害ノードのストレージリソースの自動テイクオーバーまたはネゴシエートテイクオーバーが実行される可能性があります。データリカバリについては、『ONTAP 9：ハイアベイラビリティ構成ガイド』を参照してください。2ノードMetroCluster構成の場合は、Automatic Unplanned Switchover（AUSO;自動計画外スイッチオーバー）が行われます。

サイトレベルのコントローラ障害

電源の損失、機器の交換、または災害が原因で、サイトの全コントローラ モジュールに障害が発生した場合を考えてみてください。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では、これら2つを区別できます。ISLとスイッチが稼働していてストレージにアクセス可能な場合、サイト規模のコントローラ障害によって自動スイッチオーバーが実行される可能性があります。

『ONTAP 9：ハイアベイラビリティ構成ガイド』には、サイト全体のコントローラ障害（コントローラ障害を含まない）および1つ以上のコントローラを含む障害からリカバリする方法が詳しく記載されています。

ISL障害

サイト間のリンクに障害が発生した場合を考えてみてください。この場合、MetroCluster構成による処理は何も行われません。各ノードは通常どおりデータを提供しますが、対応するDRサイトにアクセスできないため、ミラー データの書き込みは行われません。

複数の連続的な障害

複数のコンポーネントで障害が連続して起こる場合を考えてみてください。たとえば、コントローラ モジュール、スイッチ ファブリック、およびシェルフで連続して障害が発生した結果、ダウンタイムやデータ損失から保護するために、ストレージのフェイルオーバー、ファブリックの冗長処理、およびSyncMirrorが順次行われる場合です。

© 2024 NetApp, Inc. All rights reserved.

表3に、障害のタイプと、対応するDRメカニズムとリカバリ方法を示します。

AUSOは、ONTAPメディエーターとONTAP 9.7以降を使用している場合にMetroCluster IP構成でのみサポートされます。

表3）障害のタイプとリカバリ方法

障害の種類	DRメカニズム		リカバリ方法の概要	
	4ノード構成	2ノード構成	4ノード構成	2ノード構成
単一ノード障害	ローカルのHA障害	AUSO	自動フェイルオーバーとギブバックが有効になっている場合は必要なし。	ノードのリストア後、`metrocluster heal -phase aggregates`、`metrocluster heal -phase root-aggregates`、および`Metrocluster switchback` コマンドを使用した手動の修復とスイッチバックが必要です。
サイト障害	MetroClusterスイッチオーバー時		ノードがリストアされたら、`metrocluster healing` コマンドと`metrocluster switchback` コマンドを使用して手動で修復とスイッチバックを行う必要があります。	
サイトレベルのコントローラ障害	AUSO ディザスタ サイトのストレージにアクセスできる場合のみ。	AUSO 単一ノード障害と同じ。	ノードがリストアされたら、`metrocluster healing` コマンドと`metrocluster switchback` コマンドを使用して手動で修復とスイッチバックを行う必要があります。	
ISL障害	MetroClusterのスイッチオーバーなし。2つのクラスタはそれぞれ独立してデータを提供。		必要なし。接続が回復すると、ストレージは自動的に再同期される。	
複数の連続的な障害	ローカルHAフェイルオーバーのあとに、`metrocluster switchover -forced-ondisaster` コマンドを使用してMetroCluster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は不要。	`metrocluster switchover -forced-ondisaster` コマンドを使用してMetroClusterの強制スイッチオーバーを実行します。	ノードがリストアされたら、`metrocluster healing` コマンドと`metrocluster switchback` コマンドを使用して手動で修復とスイッチバックを行う必要があります。	

© 2024 NetApp, Inc. All rights reserved.

4ノードおよび8ノードのノンストップオペレーション

1つのノードに限った問題が発生した場合、ローカルのHAペア内でのフェイルオーバーとギブバックにより、中断のないノンストップ オペレーションが実現します。リモート サイトへのスイッチオーバーは必要ありません。

このMetroCluster構成は各サイトに1つ以上のHAペアで構成されるため、各サイトでのローカルな障害に耐えることができ、パートナーサイトにスイッチオーバーすることなくノンストップオペレーションを実行できます。HAペアの動作は、MetroCluster以外の構成でのHAペアと同じです。
パニックまたは電源喪失によるノード障害は、自動スイッチオーバーを原因で実行できます。

ローカル フェイルオーバー後に2回目の障害が発生した場合、MetroClusterのスイッチオーバー イベントによって、中断のないノンストップ オペレーションが実現します。同様に、稼働しているノードのいずれかで2つ目の障害が発生し、スイッチオーバー処理が行われた後は、ローカル フェイルオーバー イベントによって、中断のないノンストップ オペレーションが実現します。この場合、正常稼働している1つのノードが、DRグループ内の他の3ノードにデータを提供します。

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後にサバイバ 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中断のないノンストップ オペレーションが実現します。ただし、冗長な構成ではなくなるため、システムはリスクにさらされます。

スイッチオーバー後にローカルフェイルオーバーが発生すると、MetroCluster環境内のすべてのストレージシステムのデータを1台のコントローラで処理できるため、リソース関連の問題が発生するリスクがあります。稼働しているコントローラは、追加の障害に対して脆弱です。

2ノードのノンストップオペレーション

注： 2ノード構成はMetroCluster FCでのみサポートされます。

2つのサイトのどちらかでシステムパニックが発生した場合、MetroClusterスイッチオーバーによって中断のないノンストップオペレーションが実現します。ノードとストレージの両方に影響する電源喪失がイベントの原因である場合、スイッチオーバーは自動では行われず、 `metrocluster switchover` コマンドが実行されるまでシステムが停止します。

すべてのストレージがミラーされているため、ノード障害時のHAペアのストレージフェイルオーバーのようなサイト障害が発生した場合に、スイッチオーバー処理を使用して無停止の耐障害性を実現できます。

2ノード構成では、HAペアで自動ストレージフェイルオーバーをトリガーするイベントと同じイベントによってAUSOもトリガーされます。つまり、2ノードMetroCluster構成では、HAペアと同じ保護レベルが設定されます。

スイッチオーバー プロセスの概要

MetroClusterスイッチオーバー処理を実行すると、ストレージおよびクライアントのアクセスがソースクラスタからリモートサイトクラスタに移されるため、災害発生後にサービスを即座に再開できます。想定される変更と、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操作を把握しておく必要があります。

スイッチオーバー処理中に、システムは次の処理を実行します。

- ディザスタ サイトに属するディスクの所有権をDRパートナーに変更します。これは、停止しているパートナーに属するディスクの所有権が正常な状態のパートナーに変更される、HAペアのローカルフェイルオーバーのような状況です。
- サバイバ サイトにあるサバイバ プレックスがディザスタ クラスタ内のノードに属する場合、そのサバイバ プレックスは、サバイバ サイトのクラスタでオンラインになります。
- ディザスタ サイトに属する同期元のSVMは、ネゴシエート スイッチオーバーの間のみ停止します。
  - この手法は、ネゴシエート スイッチオーバーにのみ適用されます。

© 2024 NetApp, Inc. All rights reserved.

- ディザスタ サイトに属する同期先のSVMを起動します。

スイッチオーバー中は、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構成内のすべての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切り替えます。

サービスをリモート サイトにスイッチオーバーするだけの場合は、サイトをフェンシングせずにネゴシエートスイッチオーバーを実行します。ストレージまたは機器を信頼できない場合は、ディザスタ サイトをフェンシングしてから、計画外スイッチオーバーを実行する必要があります。フェンシングにより、ディスクに電源が順次投入されたときのRAIDの再構築が回避されます。

この手順は、もう一方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。

MetroCluster FCスイッチオーバーとIPスイッチオーバーの違い

MetroCluster IP構成では、リモート ディスクへのアクセスは、iSCSIターゲットとして機能するリモートDRパートナー ノードを介して行われます。したがって、スイッチオーバー操作でリモート ノードが停止された場合、リモート ディスクにはアクセスできません。この手法により、MetroCluster FCの構成に次のような違いが生じます。

- ローカル クラスタが所有する、ミラーされたアグリゲートがデグレードされます。
- リモート クラスタからスイッチオーバーされた、ミラーされたアグリゲートがデグレードされます。

MetroCluster 9.5では、MetroCluster IPの自動修復という新機能が採用されています。この機能は、DRテストなどの計画的なスイッチオーバーとスイッチバックを実行する際に、ルート アグリゲートとデータ アグリゲートの修復を組み合わせて、簡易化した1つのプロセスに統合します。

NetApp 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ると警告を発行します。MetroCluster Tiebreakerソフトウェアは、ONTAP 8.3および9.0~9.7でサポートされるすべてのMetroCluster構成をサポートします。

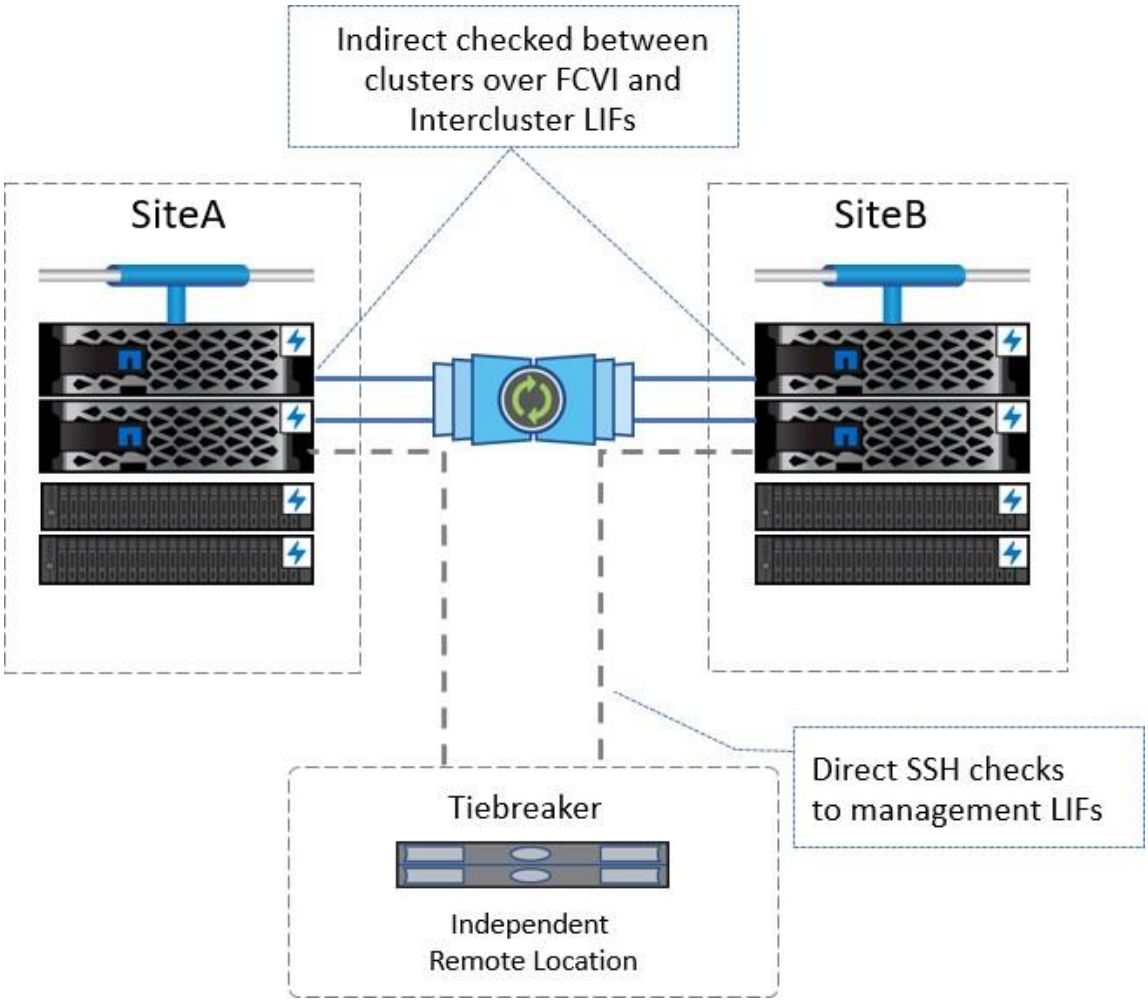
TiebreakerソフトウェアはLinuxホストにインストールされます。Tiebreakerソフトウェアは、2つのクラスタおよびクラスタ間の接続ステータスを第3のサイトから監視する場合にのみ使用します。これにより、クラスタ内の各パートナーでISL障害（サイト間リンクが停止した場合）とサイト障害を区別することができます。

複数のTiebreakerモニタ間の競合を回避するため、MetroCluster構成ごとにMetroCluster Tiebreakerモニタは1つにする必要があります。

NetApp MetroCluster Tiebreakerソフトウェアは、MetroCluster構成のノードおよびクラスタに到達できるかどうかをチェックして、サイト障害の有無を判断します。また、Tiebreakerソフトウェアは、特定の状況でアラートをトリガーします。図11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、ファブリックに損傷がない場合はスイッチオーバーを開始しません。

© 2024 NetApp, Inc. All rights reserved.

図13）MetroCluster Tiebreakerのチェック



MetroCluster Tiebreakerによる障害の検出

TiebreakerソフトウェアはLinuxホストにインストールされます。Tiebreakerソフトウェアは、2つのクラスタおよびクラスタ間の接続ステータスを第3のサイトから監視する場合にのみ使用します。これにより、クラスタ内の各パートナーでISL障害（サイト間リンクが停止した場合）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ら、災害状況を監視するMetroCluster構成内のクラスタを設定できます。

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ると警告を発行します。MetroClusterでは次のタイプのネットワーク パスが使用され、MetroCluster Tiebreakerによって監視されます。

- FCネットワーク：この種類のネットワークは、2つの冗長FCスイッチ ファブリックで構成されます。各スイッチ ファブリックには2つのFCスイッチがあり、各スイッチ ファブリックの1つのスイッチはクラスタと同じ場所に配置されます。各クラスタには、各スイッチ ファブリックから1つずつ、2つのFCスイッチがあります。すべてのノードは、同じ場所に配置されている各FCスイッチにFC（NVインターコネクトおよびFCPイニシエータ）接続されています。データは、クラスタからクラスタへ、ISL経由でレプリケートされます。

© 2024 NetApp, Inc. All rights reserved.

- クラスタ間ピアリング ネットワーク： この種類のネットワークは、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設定をミラーリングするために必要な接続を提供します。一方のクラスタのすべてのSVMの設定が、パートナー クラスタにミラーされます。
- IPネットワーク： この種類のネットワークは、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1つのスイッチはクラスタと同じ場所に配置されます。各クラスタには、各スイッチ ファブリックから1つずつ、2つのIPスイッチがあります。すべてのノードは、同じ場所に配置されている各FCスイッチに接続されています。データは、クラスタからクラスタへ、ISL経由でレプリケートされます。

サイト間接続の監視

Tiebreakerソフトウェアは、サイト間接続のステータスをノードから定期的に取得します。NVインターコネクト接続が失われ、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分離されているとみなし、Tiebreakerソフトウェアは「AllLinksSevered」アラートをトリガーします。クラスタが「AllLinksSevered」ステータスを識別し、もう一方のクラスタがネットワーク経由で到達できない場合、Tiebreakerソフトウェアは「disaster」アラートをトリガーします。

Tiebreakerで監視されるコンポーネント

Tiebreakerソフトウェアは、IPネットワークでホストされるノード管理LIFとクラスタ管理LIFへの複数のパスを介して冗長接続を確立することで、MetroCluster構成内の各コントローラを監視します。

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

- ローカル ノード インターフェイスを介したノード
- クラスタ指定インターフェイスを介したクラスタ
- 障害が発生していないクラスタ（ディザスタ サイトへの接続が確立されているかどうか（NVインターコネクト、ストレージ、クラスタ間ピアリング）を評価するため）

Tiebreakerソフトウェアとクラスタ内のすべてのノードとの間の接続、およびクラスタ自体へ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」と宣言されます。接続障害を検出するには、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、サバイバークラスタ（到達可能なクラスタ）がパートナークラスタへのすべてのリンクが切断されたことを示すまで、Tiebreakerソフトウェアはアラートをトリガーしません。

障害が発生していないクラスタが、FC（NVインターコネクトとストレージ）およびクラスタ間ピアリングを介してディザスタ サイトのクラスタと通信できなくなった場合、すべてのリンクは切断されます。

Tiebreakerの障害シナリオ

Tiebreakerソフトウェアは、ディザスタ 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、サバイバ サイトのクラスタが「AllLinksSevered」ステータスを示している場合、アラートをトリガーします。

Tiebreakerソフトウェアは、次のいずれかの場合はアラートをトリガーしません（またはアラートは拒否されます）。

- 8ノードのMetroCluster構成で、ディザスタ サイトの1つのHAペアが停止している場合。
- ディザスタ サイトのすべてのノードが停止しているクラスタで、サバイバ サイトの1つのHAペアが停止しており、サバイバ サイトのクラスタが「AllLinksSevered」ステータスを示している場合。Tiebreakerソフトウェアはアラートをトリガーしますが、ONTAPはアラートを拒否します。この場合、手動によるスイッチオーバーも拒否されます。
- Tiebreakerソフトウェアがディザスタ サイトの少なくとも1つのノードまたはクラスタ インターフェイスに到達できる、またはサバイバ サイトがFC（NVインターコネクトとストレージ）かクラスタ間ピアリングのいずれかを使用してディザスタ サイトのいずれかのノードに到達できる場合。

© 2024 NetApp, Inc. All rights reserved.

ONTAP Mediator

ONTAP 9.7には、ONTAPメディエーターと呼ばれる障害を処理するための新しいMetroCluster IP解決策が含まれています。ONTAPには、ONTAPメディエーターを使用してMetroCluster IPにAUSO機能を提供するなどの機能が追加されています。ONTAPメディエーターは、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の物理サーバまたは仮想サーバにインストールします。

ONTAPメディエーターの要件および障害の詳細については、MetroCluster IPインストールおよび設定ガイドを参照してください。

注： TiebreakerとONTAPメディエーターの両方を使用した同じMetroCluster構成の管理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1つだけです。

テクノロジ要件

ハードウェアとソフトウェアの要件

MCCを構成する際は、サポートされるハードウェアコンポーネントと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。使用するハードウェアコンポーネントは、お客様の導入環境や、MetroCluster FCとIPのどちらを選択したかによって異なります。MCC構成で使用するONTAPシステム、ストレージアレイ、およびFCスイッチが必要な要件を満たしていることが不可欠です。MCCの実装に必要なソフトウェアコンポーネントは、ONTAPのみです。これは標準機能であり、別途ライセンスは必要ありません。標準のONTAPライセンスには、クライアント側とホスト側のプロトコルに加え、非同期ミラーを使用してデータを保護するSnapMirrorや、バックアップデータを保護するために第3のクラスタにデータをレプリケートするXDPの追加機能が含まれています。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、利用可能な技術リソースを参照してください。

MetroCluster FCのハードウェアとソフトウェアの要件の詳細については、次のドキュメントを参照してください。

- TR-4375：『NetApp MetroCluster FC』

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

- TR-4689：『MetroCluster IP -解決策Architecture and Design』

ハードウェア構成と相互運用性に関する最新情報については、次のツールを使用してください。

- NetApp Hardware Universe
- Interoperability Matrix Tool
- Fusion

まとめ

FCファブリックとIPファブリックの両方のサポートなど、MetroClusterにはさまざまな導入オプションが用意されており、きわめて高い柔軟性、高度なデータ保護、あらゆるプロトコル、アプリケーション、仮想環境でのシームレスなフロントエンド統合を実現できます。

© 2024 NetApp, Inc. All rights reserved.

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。

- TR-4375：NetApp MetroCluster FC
  https://www.netapp.com/pdf.html?item=/media/13482-tr4375.pdf
- TR-4689：NetApp MetroCluster IP
  https://www.netapp.com/pdf.html?item=/media/13481-tr4689pdf.pdf
- MetroCluster IP 40Gbスイッチ技術プレゼンテーション（NetApp Field Portal、ログインが必要）
  https://fieldportal.netapp.com/content/729700
- MetroCluster IP 100Gbスイッチ技術プレゼンテーション（NetApp Field Portal、ログインが必要）
  https://fieldportal.netapp.com/content/757495
- MetroCluster FC Technical FAQ（NetApp Field Portal、ログインが必要）
  https://fieldportal.netapp.com/content/617080
- MetroCluster IP Technical FAQ（NetApp Field Portal、ログインが必要）
  https://fieldportal.netapp.com/content/748972
- MetroCluster IP and FC ISL Sizing Spreadsheet（NetApp Field Portal、ログインが必要）
  https://fieldportal.netapp.com/content/699509
- NetApp Interoperability Matrix Tool
  https://imt.netapp.com/
- NetApp Hardware Universe
  https://hwu.netapp.com/
- MetroClusterのドキュメント
  https://docs.netapp.com/us-en/ontap-metrocluster/index.html
- MetroClusterリソース
  http://mysupport.netapp.com/metrocluster/resources
- NetAppサポート サイト
  https://mysupport.netapp.com/site/
- NetAppの製品ドキュメント
  https://docs.netapp.com

バージョン履歴

バージョン	日付	ドキュメントの改訂履歴
バージョン1.0	2019年11月	ONTAP 9.7
バージョン2.0	2023年4月	ONTAP 9.12.1

© 2024 NetApp, Inc. All rights reserved.

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

Copyright © 2024 NetApp, Inc. All Rights Reserved. Printed in the U.S.このドキュメントは著作権によって保護されています。著作権所有者の書面による事前承諾がある場合を除き、画像媒体、電子媒体、および写真複写、記録媒体、テープ媒体、電子検索システムへの組み込みを含む機械媒体など、いかなる形式および方法による複製も禁止します。

NetApp の著作物から派生したソフトウェアは、次に示す使用許諾条項および免責条項の対象となります。

このソフトウェアは、NetApp によって「現状のまま」提供されています。NetApp は明示的な保証、または商品性および特定目的に対する適合性の暗示的保証を含み、かつこれに限定されないいかなる暗示的な保証も行いません。NetApp は、代替品または代替サービスの調達、使用不能、データ損失、利益損失、業務中断を含み、かつこれに限定されない、このソフトウェアの使用により生じたすべての直接的損害、間接的損害、偶発的損害、特別損害、懲罰的損害、必然的損害の発生に対して、損失の発生の可能性が通知されていたとしても、その発生理由、根拠とする責任論、契約の有無、厳格責任、不法行為（過失またはそうでない場合を含む）にかかわらず、一切の責任を負いません。

NetApp は、ここに記載されているすべての製品に対する変更を随時、予告なく行う権利を保有します。NetApp による明示的な書面による合意がある場合を除き、ここに記載されている製品の使用により生じる責任および義務に対して、NetApp は責任を負いません。この製品の使用または購入は、NetApp の特許権、商標権、または他の知的所有権に基づくライセンスの供与とはみなされません。

このマニュアルに記載されている製品は、1 つ以上の米国特許、その他の国の特許、および出願中の特許により保護されている場合があります。

本書に含まれるデータは市販の製品および / またはサービス（FAR 2.101 の定義に基づく）に関係し、データの所有権は NetApp, Inc.にあります。米国政府は本データに対し、非独占的かつ移転およびサブライセンス不可で、全世界を対象とする取り消し不能の制限付き使用権を有し、本データの提供の根拠となった米国政府契約に関連し、当該契約の裏付けとする場合にのみ本データを使用できます。前述の場合を除き、NetApp, Inc.の書面による許可を事前に得ることなく、本データを使用、開示、転載、改変するほか、上演または展示することはできません。国防総省にかかる米国政府のデータ使用権については、DFARS 252.227-7015(b)項で定められた権利のみが認められます。

商標に関する情報

NetApp、NetAppのロゴ、https://www.netapp.com/company/legal/trademarks/に記載されているマークは、NetApp, Inc.の商標です。その他の会社名と製品名は、それを所有する各社の商標である場合があります。