ONTAPを使用したvSphere Metroストレージクラスタ

Enterprise applications

NetApp
May 03, 2024

目次

ONTAPを使用したvSphere Metroストレージクラスタ

ONTAPを使用したvSphere Metroストレージクラスタ

VMwareの業界をリードするvSphereハイパーバイザーは、vSphere Metro Storage Cluster（vMSC）と呼ばれるストレッチクラスタとして導入できます。

vMSCソリューションは、NetApp®MetroCluster™とSnapMirrorアクティブ同期（旧称SnapMirrorビジネス継続性（SMBC））の両方でサポートされており、1つ以上の障害ドメインで全体的な停止が発生した場合に高度なビジネス継続性を提供します。さまざまな障害モードへの耐障害性は、どの設定オプションを選択するかによって異なります。

vSphere環境向けの継続的可用性ソリューション

ONTAPのアーキテクチャは、柔軟性と拡張性に優れたストレージプラットフォームであり、データストアにSAN（FCP、iSCSI、NVMe-oF）サービスとNAS（NFS v3およびv4.1）サービスを提供します。NetApp AFF、ASA、FASの各ストレージシステムは、ONTAPオペレーティングシステムを使用して、ゲストストレージアクセス用にS3、SMB / CIFSなどの追加プロトコルを提供します。

NetApp MetroClusterは、ネットアップのHA（コントローラフェイルオーバーまたはCFO）機能を使用してコントローラ障害から保護します。また、ローカルSyncMirrorテクノロジ、災害時のクラスタフェイルオーバー（オンデマンドのコントローラフェイルオーバーまたはCFOD）、ハードウェアの冗長性、地理的な分離によって高レベルの可用性を実現します。SyncMirrorは、アクティブにデータを提供しているローカルプレックス（ローカルシェルフ上）と、通常はデータを提供していないリモートプレックス（リモートシェルフ上）の2つのプレックスにデータを書き込むことで、MetroCluster構成の2つの部分にわたってデータを同期的にミラーリングします。ハードウェアの冗長性は、コントローラ、ストレージ、ケーブル、スイッチ（ファブリックMetroClusterで使用）、アダプタなど、MetroClusterのすべてのコンポーネントで確保されています。

NetApp SnapMirrorアクティブ同期は、FCPおよびiSCSI SANプロトコルを使用してデータストアをきめ細かく保護するため、優先度の高いワークロードのみを選択的に保護できます。アクティブ/スタンバイ解決策であるNetApp MetroClusterとは異なり、ローカルサイトとリモートサイトの両方にアクティブ/アクティブアクセスを提供します。現時点では、アクティブ同期は非対称解決策であり、一方が他方よりも優先されるため、パフォーマンスが向上します。これにはAsymmetric Logical Unit Access（ALUA；非対称論理ユニットアクセス）機能が使用され、どのコントローラを優先するかがESXiホストに自動的に通知されます。ただし、NetAppでは、アクティブな同期によって完全対称アクセスがまもなく有効になることが発表されています。

2つのサイトにVMware HA / DRSクラスタを作成するために、ESXiホストをvCenter Server Appliance（vCSA）で使用および管理します。vSphere管理ネットワーク、vMotion®ネットワーク、および仮想マシンネットワークは、2つのサイト間の冗長ネットワークを介して接続されます。HA / DRSクラスタを管理するvCenter Serverは両方のサイトのESXiホストに接続でき、vCenter HAを使用して設定する必要があります。

を参照してください "vSphere Clientでクラスタを作成および構成する方法" をクリックしてvCenter HAを設定します。

また、 "VMware vSphere Metro Storage Cluster Recommended Practices"。

vSphere Metro Storage Clusterとは

vSphere Metro Storage Cluster（vMSC）は、仮想マシン（VM）とコンテナを障害から保護する認定済みの構成です。これは、ストレッチストレージの概念とESXiホストのクラスタを使用して実現されます。ESXiホストは、ラック、建物、キャンパス、さらには都市など、さまざまな障害ドメインに分散されます。NetApp MetroClusterとSnapMirrorのアクティブな同期ストレージテクノロジは、それぞれホストクラスタに対してRPO=0またはNearRPO=0の保護を提供するために使用されます。vMSCの構成は、物理的または論理的な「サイト」全体に障害が発生した場合でも、データを常に利用できるように設計されています。vMSC構成に含まれるストレージデバイスは、vMSC認定プロセスを完了したあとに認定されている必要があります。サポートされているすべてのストレージデバイスは、 "VMwareストレージ互換性ガイド"。

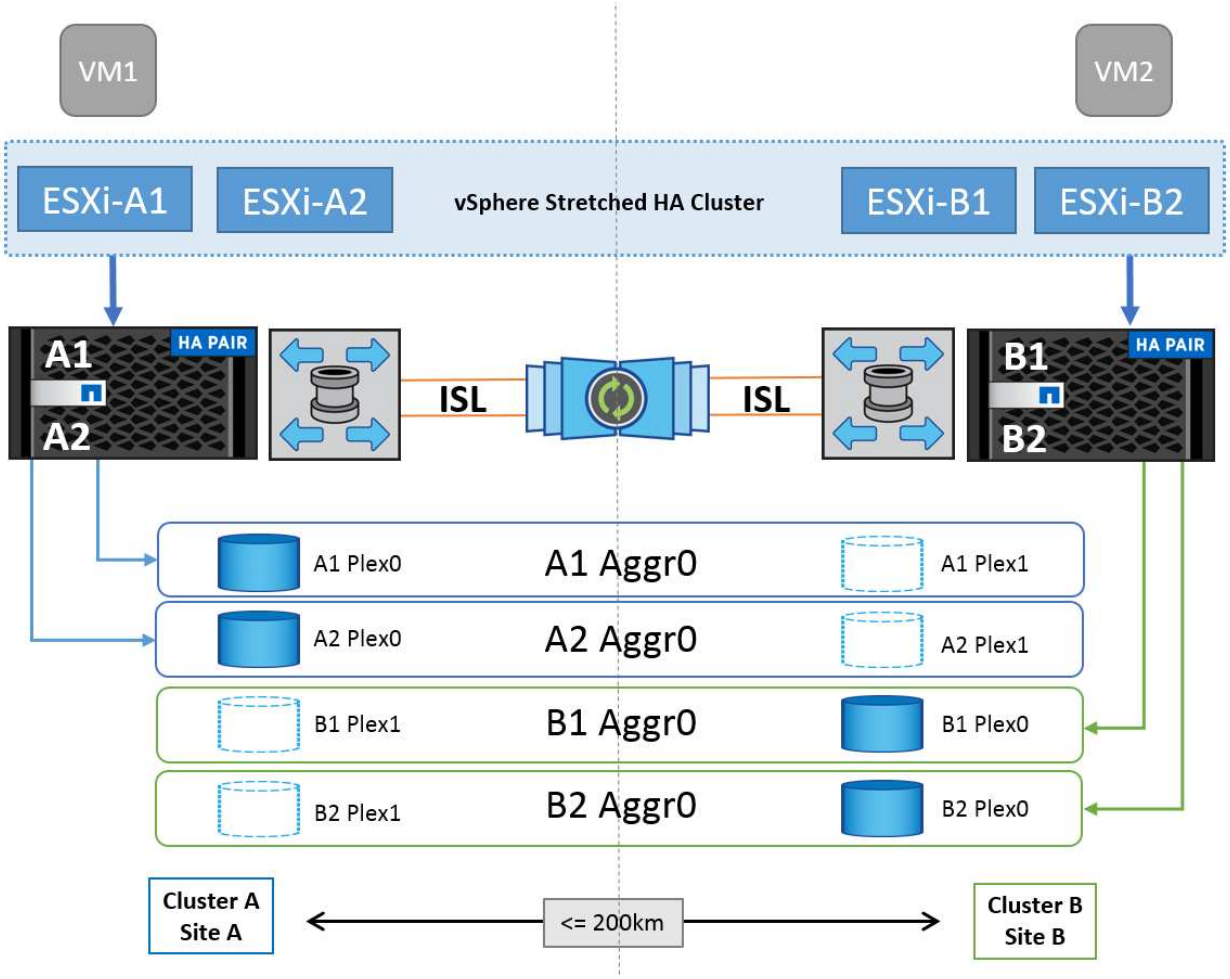
vSphere Metro Storage Clusterの設計ガイドラインの詳細については、次のドキュメントを参照してください。

- "NetApp MetroClusterによるVMware vSphereのサポート"
- "NetApp SnapMirrorビジネス継続性によるVMware vSphereのサポート" （SnapMirrorアクティブ同期）

レイテンシの考慮事項に応じて、NetApp MetroClusterを導入してvSphereで使用できます。

- ストレッチMetroCluster
- ファブリックMetroCluster

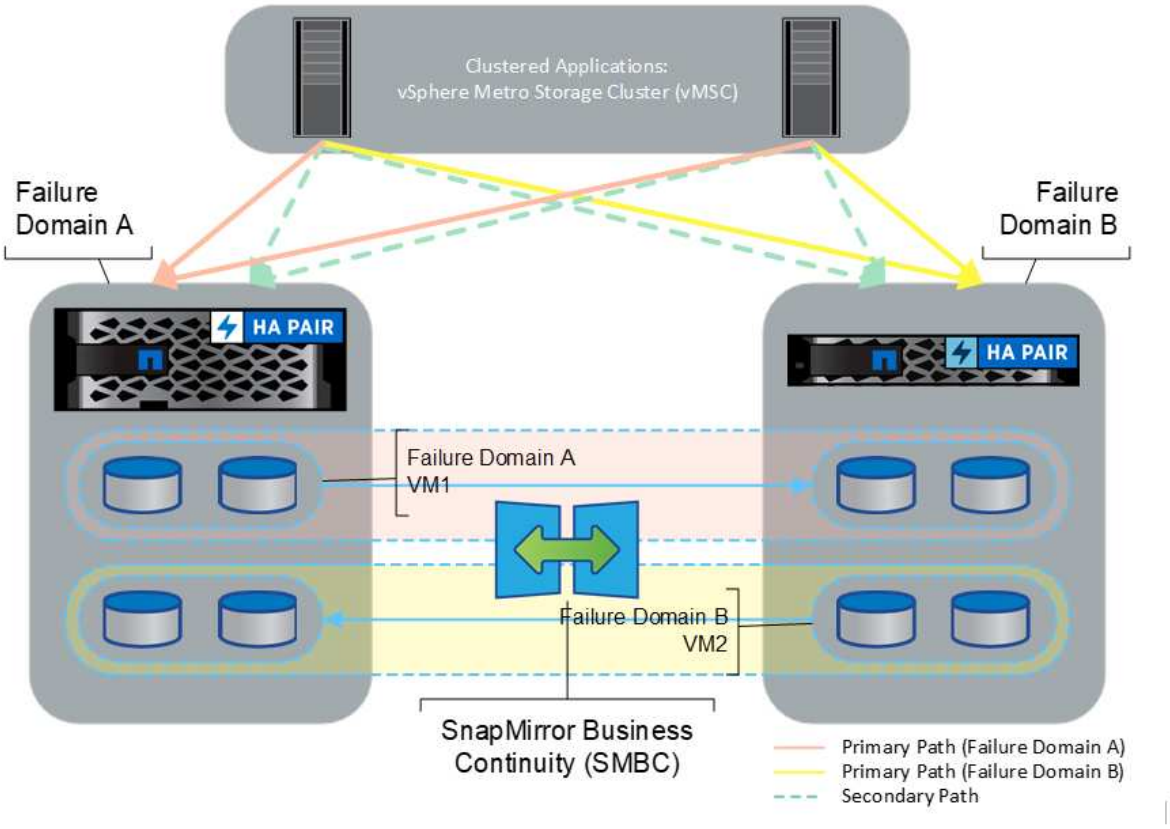
次の図は、ストレッチMetroClusterのトポロジ図の概要を示しています。



を参照してください "MetroCluster のドキュメント" を参照してくださいMetroCluster。

SnapMirror Active Syncは、2つの方法で導入することもできます。

- 非対称
- 対称（ONTAP 9.14.1でのプライベートプレビュー）



を参照してください "ネットアップのドキュメント" を参照し、SnapMirror Active Syncの設計と導入に関する情報を確認してください。

VMware vSphere解決策の概要

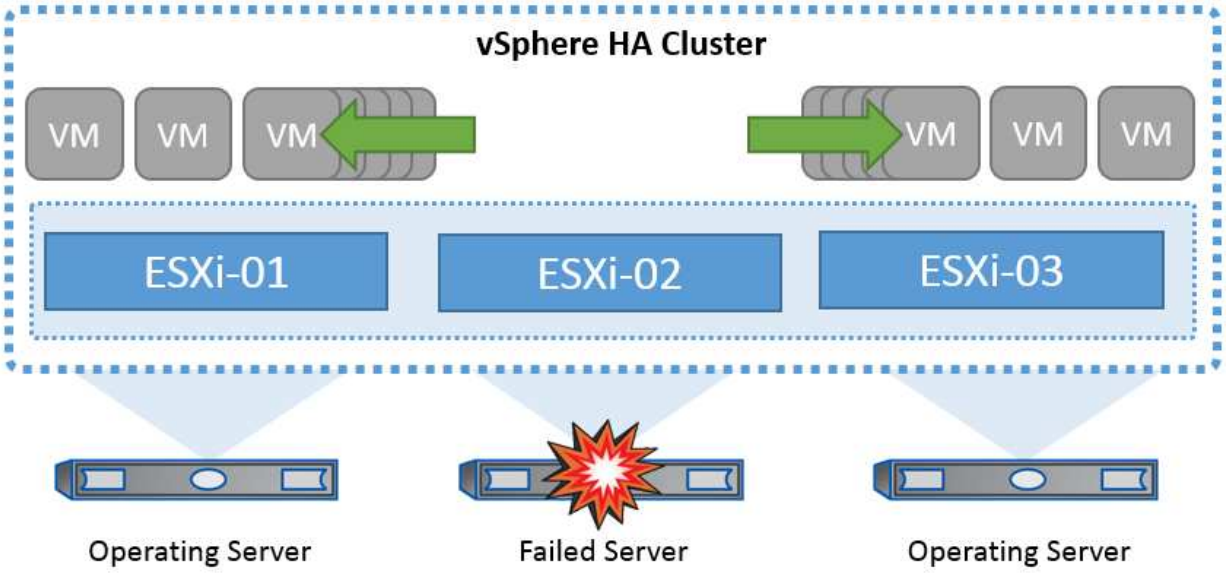
vCenter Server Appliance（vCSA）は、管理者がESXiクラスタを効果的に運用できるようにする、強力な一元管理システムであり、vSphere用の単一コンソールです。VMプロビジョニング、vMotion処理、High Availability（HA；高可用性）、Distributed Resource Scheduler（DRS；分散リソーススケジューラ）、Tanzu Kubernetes Gridなどの主要な機能を簡易化します。VMwareクラウド環境に欠かせないコンポーネントであり、サービスの可用性を考慮して設計する必要があります。

vSphereの高可用性

VMwareのクラスタテクノロジは、ESXiサーバを仮想マシンの共有リソースプールにグループ化し、vSphere High Availability（HA；高可用性）を提供します。vSphere HAは、仮想マシンで実行されるアプリケーションに対して、使いやすく高可用性を提供します。クラスタでHA機能を有効にすると、いずれかのESXiホストが

応答しなくなったり分離されたりした場合に、各ESXiサーバが他のホストとの通信を維持します。 HAクラスタは、そのESXiホストで実行されていた仮想マシンのリカバリを、クラスタ内の残りのホスト間でネゴシエートできます。ゲストオペレーティングシステムに障害が発生すると、vSphere HAは影響を受ける仮想マシンを同じ物理サーバ上で再起動します。vSphere HAを使用すると、計画的停止の削減、計画外停止の防止、システム停止からの迅速なリカバリが可能になります。

vSphere HAクラスタ：障害が発生したサーバからVMをリカバリします。



VMware vSphereはNetApp MetroClusterまたはSnapMirrorのアクティブ同期を認識しないため、vSphereクラスタ内のすべてのESXiホストが、ホストおよびVMグループのアフィニティ構成に応じてHAクラスタ処理の対象となるホストとして認識されることを理解しておくことが重要です。

ホスト障害の検出

HAクラスタが作成されるとすぐに、クラスタ内のすべてのホストが選択対象になり、いずれかのホストがマスターになります。各スレーブはマスターに対してネットワークハートビートを実行し、マスターはすべてのスレーブホストに対してネットワークハートビートを実行します。vSphere HAクラスタのマスターホストは、スレーブホストの障害を検出する役割を果たします。

検出された障害のタイプによっては、ホストで実行されている仮想マシンのフェイルオーバーが必要になる場合があります。

vSphere HAクラスタでは、次の3種類のホスト障害が検出されます。

- 障害-ホストが機能を停止しました。
- 分離-ホストがネットワークから分離されます。
- パーティション-ホストとマスターホストとのネットワーク接続が失われます。

マスターホストは、クラスタ内のスレーブホストを監視します。この通信は、1秒ごとにネットワークハートビートを交換して行われます。マスターホストは、スレーブホストからのハートビートの受信を停止すると、ホストの稼働状況を確認してから、ホストに障害が発生したことを宣言します。マスターホストが実行する活性チェックでは、スレーブホストがいずれかのデータストアとハートビートを交換しているかどうかを確認します。また、マスターホストは、管理IPアドレスに送信されたICMP pingにホストが応答するかどうかをチェックして、単にマスターノードから隔離されているか、ネットワークから完全に隔離されているかを検出しま

す。これは、デフォルトゲートウェイに対してpingを実行することによって行われます。隔離アドレスを手動で指定することで、隔離検証の信頼性を高めることができます。

ベストプラクティス

NetAppでは、隔離アドレスを少なくとも2つ追加し、各アドレスをサイトローカルにすることを推奨しています。これにより、隔離検証の信頼性が向上します。

ホスト隔離時の対応

[Isolation Response]はvSphere HAの設定で、vSphere HAクラスタ内のホストが管理ネットワーク接続を失い、実行は継続した場合に仮想マシンでトリガーされる処理を決定します。この設定には、[Disabled]、[Shut Down and Restart VMs]、[Power Off and Restart VMs]の3つのオプションがあります。

[Shut Down]は、[Power Off]よりも優れています。[Power Off]では、最新の変更がディスクにフラッシュされたり、トランザクションがコミットされたりしません。仮想マシンが300秒以内にシャットダウンされない場合は、電源がオフになります。待機時間を変更するには、詳細オプションdas.isolationshutdowntimeoutを使用します。

HAは隔離時の対応を開始する前に、vSphere HAマスターエージェントがVM構成ファイルが格納されたデータストアを所有しているかどうかを確認します。そうでない場合、VMを再起動するマスターがないため、ホストは隔離時の対応をトリガーしません。ホストはデータストアの状態を定期的にチェックして、マスターロールを持つvSphere HAエージェントがデータストアを要求しているかどうかを判断します。

ベストプラクティス

NetAppでは、[Host Isolation Response]を[Disabled]に設定することを推奨しています。

ホストがvSphere HAマスターホストから分離またはパーティショニングされ、ハートビートデータストアまたはpingを介してマスターと通信できなくなると、スプリットブレイン状態が発生することがあります。マスターは、隔離されたホストの停止を宣言し、クラスタ内の他のホスト上のVMを再起動します。仮想マシンのインスタンスが2つ実行され、そのうちの1つだけが仮想ディスクの読み取りまたは書き込みを実行できるため、スプリットブレイン状態が発生します。VM Component Protection（VMCP）を設定することで、スプリットブレイン状態を回避できるようになりました。

VMコンポーネント保護（VMCP）

vSphere 6で強化されたHA関連機能の1つにVMCPがあります。VMCPは、ブロック（FC、iSCSI、FCoE）とファイルストレージ（NFS）のAll Paths Down（APD）状態とPermanent Device Loss（PDL）状態からの保護を強化します。

Permanent Device Loss（PDL）

PDLとは、ストレージデバイスに永続的に障害が発生した場合、または管理上削除されて元に戻ることがない場合に発生する状態です。NetAppストレージアレイは、デバイスが永続的に失われたことを宣言するSCSIセンスコードをESXiに発行します。vSphere HAの[Failure Conditions and VM Response]セクションで、PDL状態が検出されたあとの応答を設定できます。

ベストプラクティス

NetAppでは、[Response for Datastore with PDL]を[* Power off and restart VMs]に設定することを推奨しています。この状態が検出されると、vSphere HAクラスタ内の正常なホストでVMが即座に再起動されます。

すべてのパスがダウン（**APD**）

APDは、ストレージデバイスがホストからアクセスできなくなり、アレイへのパスが使用できなくなった場合に発生する状態です。ESXiは、これをデバイスの一時的な問題とみなし、再び使用可能になることを想定しています。

APD状態が検出されると、タイマーが開始されます。140秒後、APD状態が正式に宣言され、デバイスはAPDタイムアウトとしてマークされます。140秒が経過すると、[Delay for VM Failover APD]で指定された分数がカウントされます。指定した時間が経過すると、影響を受ける仮想マシンが再起動されます。必要に応じて異なる方法（[Disabled]、問題Events]、[Power Off and Restart VMs]）で応答するようにVMCPを設定できます。

ベストプラクティス

NetAppでは、[Response for Datastore with APD]を「* Power off and restart VMs（conservative）*」に設定することを推奨しています。

保守的とは、HAがVMを再起動できる可能性を示します。[Conservative]に設定すると、APDの影響を受けるVMは、別のホストで再起動できることがわかっている場合にのみ再起動されます。アグレッシブの場合、HAは他のホストの状態を認識していなくてもVMの再起動を試行します。その結果、VMが配置されているデータストアにアクセスできるホストがないと、VMが再起動されない可能性があります。

タイムアウトになる前にAPDステータスが解決され、ストレージへのアクセスが回復した場合は、明示的に設定していないかぎり、仮想マシンが不要に再起動されることはありません。環境がAPD状態から回復した場合でも応答が必要な場合は、[Response for APD Recovery After APD Timeout]を[Reset VMs]に設定する必要があります。

ベストプラクティス

NetAppでは、[Response for APD Recovery After APD Timeout]を[Disabled]に設定することを推奨します。

NetApp MetroCluster向け**VMware DRS**の実装

VMware DRSは、クラスタ内のホストリソースを集約する機能で、主に仮想インフラストラクチャ内のクラスタ内での負荷分散に使用されます。VMware DRSは、クラスタ内でロードバランシングを実行するために、主にCPUリソースとメモリリソースを計算します。vSphereはストレッチクラスタリングを認識しないため、両方のサイトのすべてのホストをロードバランシングの対象とします。サイト間トラフィックを回避するために、NetAppでは、VMの論理的な分離を管理するDRSアフィニティルールを設定することを推奨しています。これにより、サイト全体に障害が発生しないかぎり、HAとDRSでローカルホストのみが使用されるようになります。

クラスタ用のDRSアフィニティルールを作成する場合は、仮想マシンのフェイルオーバー時にvSphereがそのルールを適用する方法を指定できます。

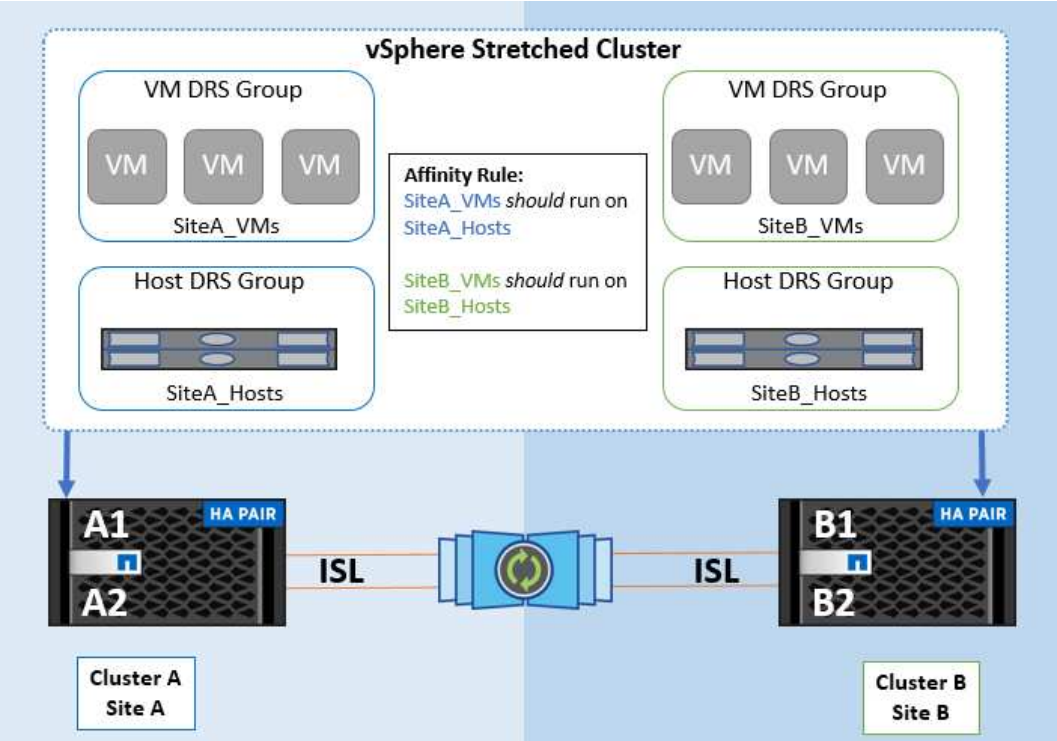
vSphere HAのフェイルオーバー動作を指定できるルールには、次の2種類があります。

- VMの非アフィニティルールでは、フェイルオーバー処理中に指定した仮想マシンが分離されたままになります。
- VMホストアフィニティルールは、フェイルオーバー処理中に、指定した仮想マシンを特定のホストまたは定義されたホストグループのメンバーに配置します。

VMware DRSのVMホストアフィニティルールを使用すると、サイトAとサイトBを論理的に分離して、特定のデータストアのプライマリ読み取り/書き込みコントローラとして設定されたアレイと同じサイトのホスト

でVMを実行できます。また、VMホストアフィニティルールを使用すると、仮想マシンはストレージに対してローカルなままになり、サイト間でネットワーク障害が発生した場合に仮想マシンの接続が確保されます。

次に、VMホストグループとアフィニティルールの例を示します。



ベストプラクティス

NetAppでは、障害が発生した場合にvSphere HAによって違反されるため、「must」ルールではなく「should」ルールを実装することを推奨しています。「must」ルールを使用すると、サービスが停止する可能性があります。

サービスの可用性は常にパフォーマンスより優先されるべきです。データセンター全体で障害が発生した場合、「must」ルールではVMホストアフィニティグループからホストを選択する必要があり、データセンターが使用できなくなっても仮想マシンは再起動されません。

NetApp MetroClusterでのVMware Storage DRSの実装

VMware Storage DRS機能を使用すると、データストアを1つのユニットに集約し、Storage I/O Controlのしきい値を超えた場合に仮想マシンディスクのバランスを調整できます。

Storage I/O Controlは、Storage DRS対応のDRSクラスタではデフォルトで有効になっています。Storage I/O Controlを使用すると、I/Oの輻輳時に仮想マシンに割り当てるストレージI/Oの量を管理者が制御できるため、重要度の高い仮想マシンを優先してI/Oリソースを割り当てることができます。

Storage DRSは、Storage vMotionを使用して、データストアクラスタ内の別のデータストアに仮想マシンを移行します。NetApp MetroCluster環境では、仮想マシンの移行をそのサイトのデータストア内で制御する必要があります。たとえば、サイトAのホストで実行されている仮想マシンAを移行する場合は、サイトAのSVMのデータストア内で移行するのが理想的です。そうしないと、仮想ディスクの読み取り/書き込みはサイト間リンクを介してサイトBから行われるため、仮想マシンは引き続き動作しますが、パフォーマンスは低下します。

ベストプラクティス

NetAppでは、ストレージサイトのアフィニティに従ってデータストアクラスタを作成することを推奨しています。つまり、サイトAに対するサイトアフィニティが設定されたデータストアクラスタと、サイトBに対するサイトアフィニティが設定されたデータストアを混在させないでください。

Storage vMotionを使用して仮想マシンを新規にプロビジョニングまたは移行するたびに、NetAppそれらの仮想マシンに固有のすべてのVMware DRSルールを手動で更新することを推奨します。これにより、ホストとデータストアの両方について、サイトレベルで仮想マシンのアフィニティが確保され、ネットワークとストレージのオーバーヘッドが削減されます。

vMSC設計および実装ガイドライン

本ドキュメントでは、ONTAPストレージシステムを使用するvMSCの設計と実装のガイドラインについて説明します。

NetAppストレージ構成

NetApp MetroCluster（MCC構成）のセットアップ手順については、次のWebサイトを参照してください。"MetroCluster のドキュメント"。SnapMirrorアクティブ同期の手順については、 "SnapMirror のビジネス継続性機能の概要"。

一度MetroClusterを設定すると、従来のONTAP環境を管理するようなものになります。Storage Virtual Machine（SVM）は、コマンドラインインターフェイス（CLI）、System Manager、Ansibleなどのさまざまなツールを使用してセットアップできます。SVMを設定したら、通常の運用に使用する論理インターフェイス（LIF）、ボリューム、論理ユニット番号（LUN）をクラスタに作成します。これらのオブジェクトは、クラスタピアリングネットワークを使用してもう一方のクラスタに自動的にレプリケートされます。

MetroClusterを使用していない場合は、SnapMirrorアクティブ同期を使用して、異なる障害ドメインにある複数のONTAPクラスタ間で、データストア単位でのきめ細かな保護とアクティブ/アクティブアクセスを実現できます。SnapMirrorアクティブ同期では、整合グループを使用して1つ以上のデータストア間で書き込み順序の整合性が確保されます。また、アプリケーションとデータストアの要件に応じて、複数の整合グループを作成することもできます。整合グループは、複数のデータストア間でのデータ同期が必要なアプリケーションに特に役立ちます。SnapMirror Active Syncでは、rawデバイスマッピング（RDM）とゲスト内iSCSIイニシエータを使用するゲスト接続ストレージもサポートされます。整合グループの詳細については、を参照してください。 "整合グループの概要"。

SnapMirrorアクティブ同期を使用するvMSC構成の管理は、MetroClusterとは多少異なります。まず、これはSANのみの構成であり、SnapMirrorのアクティブな同期でNFSデータストアを保護することはできません。次に、両方の障害ドメインのレプリケートされたデータストアにアクセスできるように、両方のLUNのコピーをESXiホストにマッピングする必要があります。

VMware vSphere HA の場合

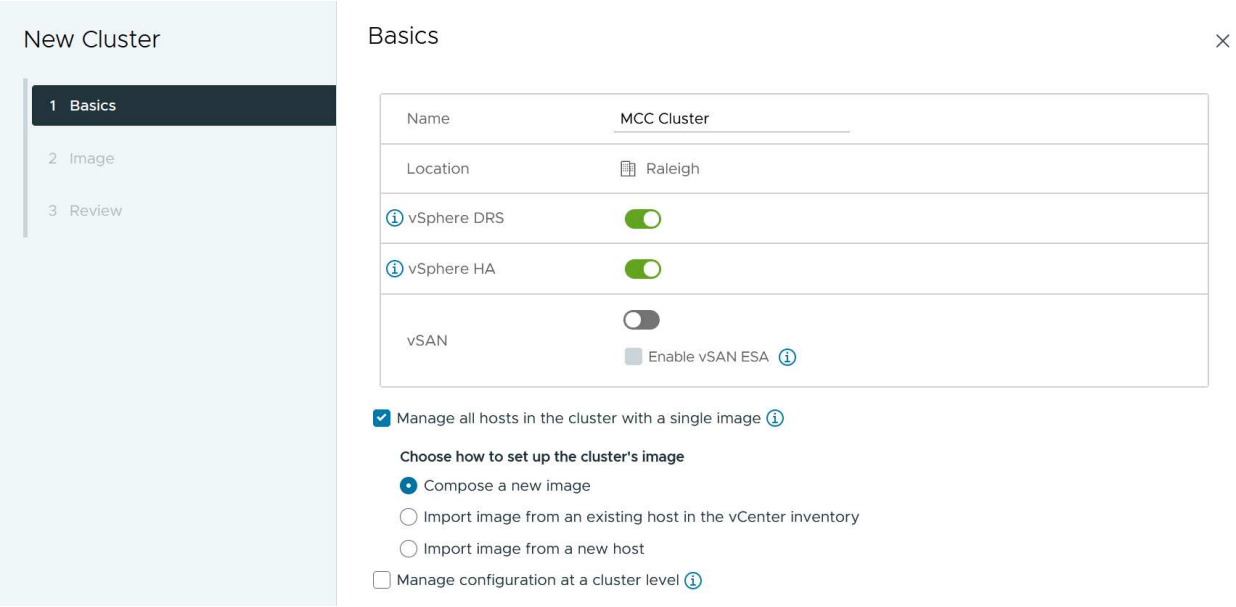
vSphere HAクラスタの作成

vSphere HAクラスタの作成は複数の手順で構成されます。詳細については、を参照してください。"docs.vmware.comのvSphere Clientでクラスタを作成および構成する方法"。つまり、最初に空のクラスタを作成してから、vCenterを使用してホストを追加し、クラスタのvSphere HAなどの設定を指定する必要があります。

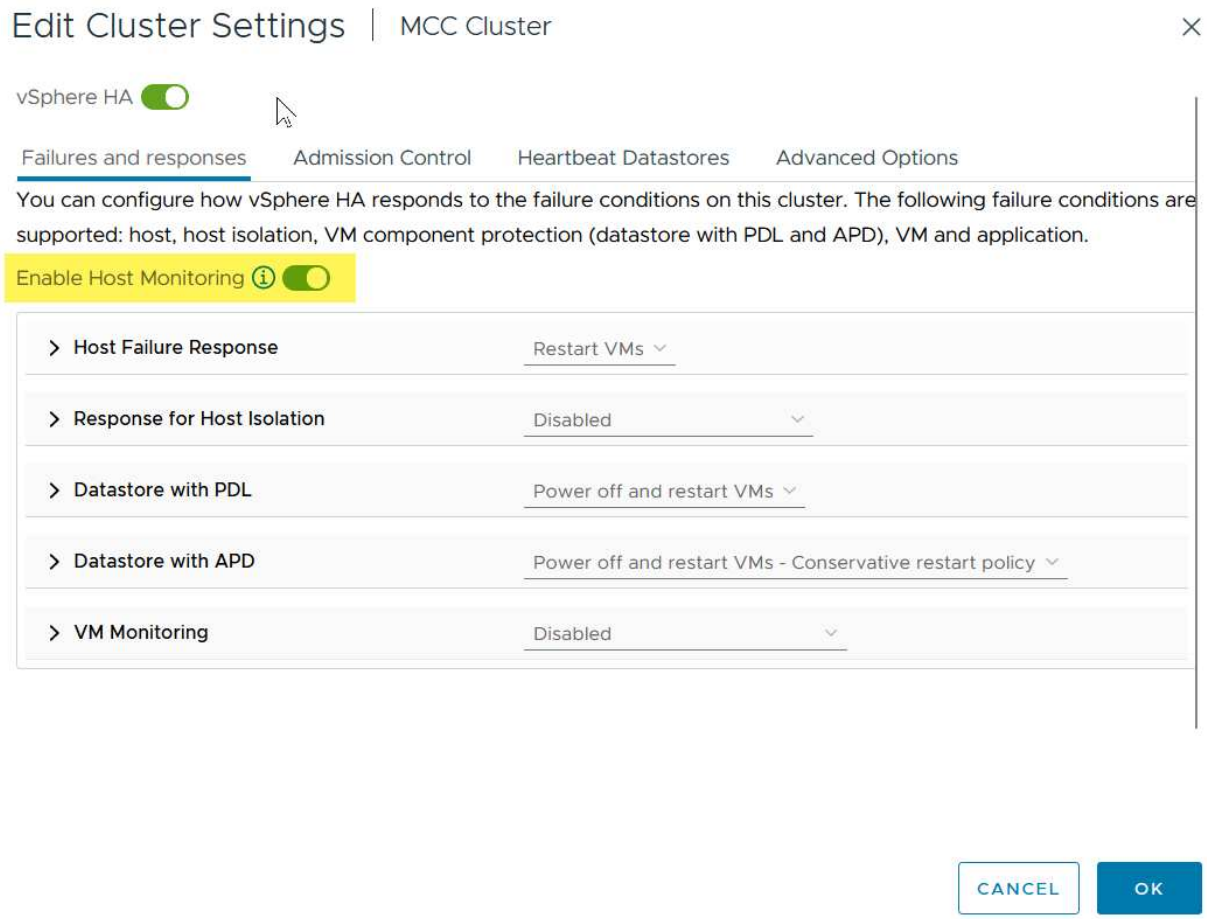
*注：*このドキュメントには、このドキュメントより優先されるものはありません。 "VMware vSphere Metro Storage Cluster Recommended Practices"

HAクラスタを設定するには、次の手順を実行します。

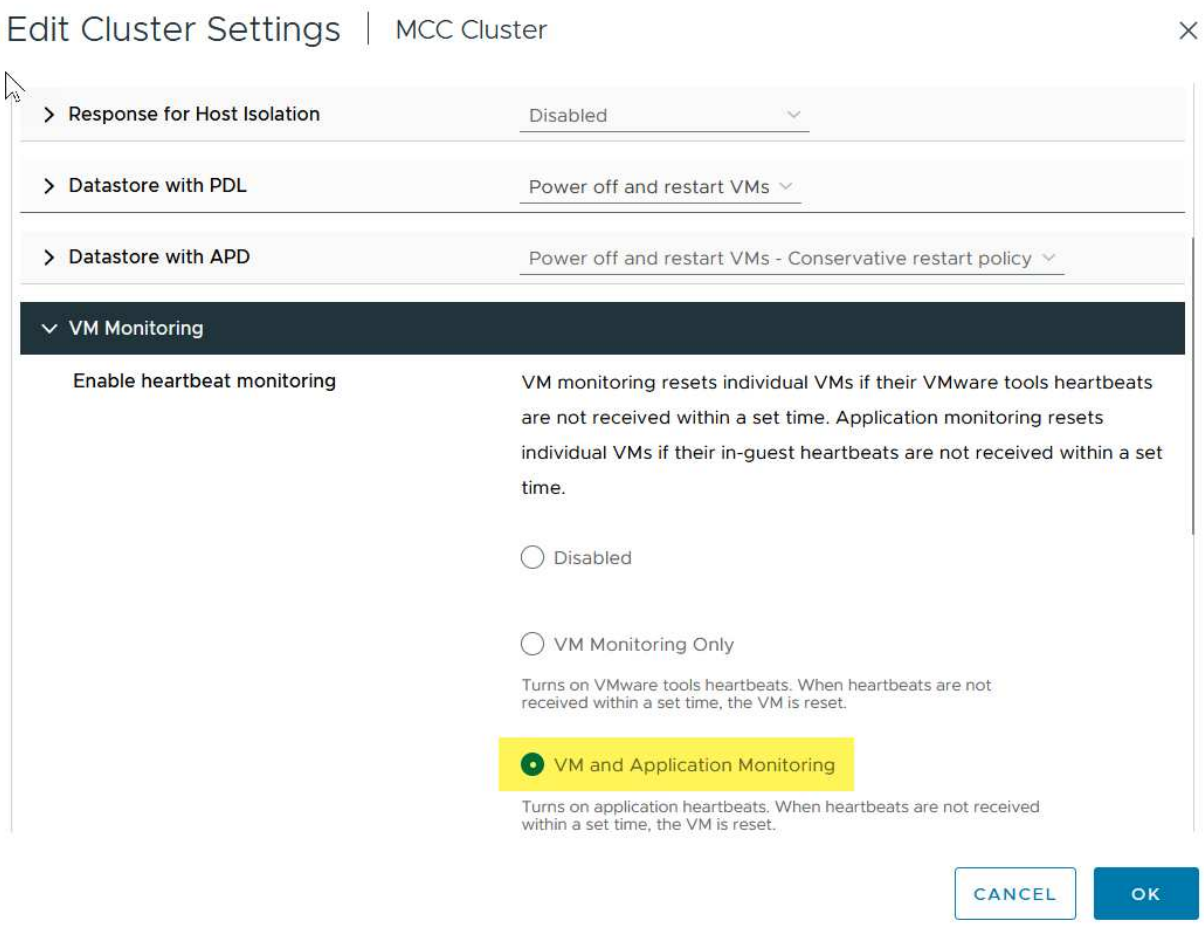
1. vCenter UIに接続します。
2. [Hosts and Clusters]で、HAクラスタを作成するデータセンターを選択します。
3. データセンターオブジェクトを右クリックし、[New Cluster]を選択します。[Basics]で、vSphere DRSとvSphere HAが有効になっていることを確認します。ウィザードの手順を実行します。



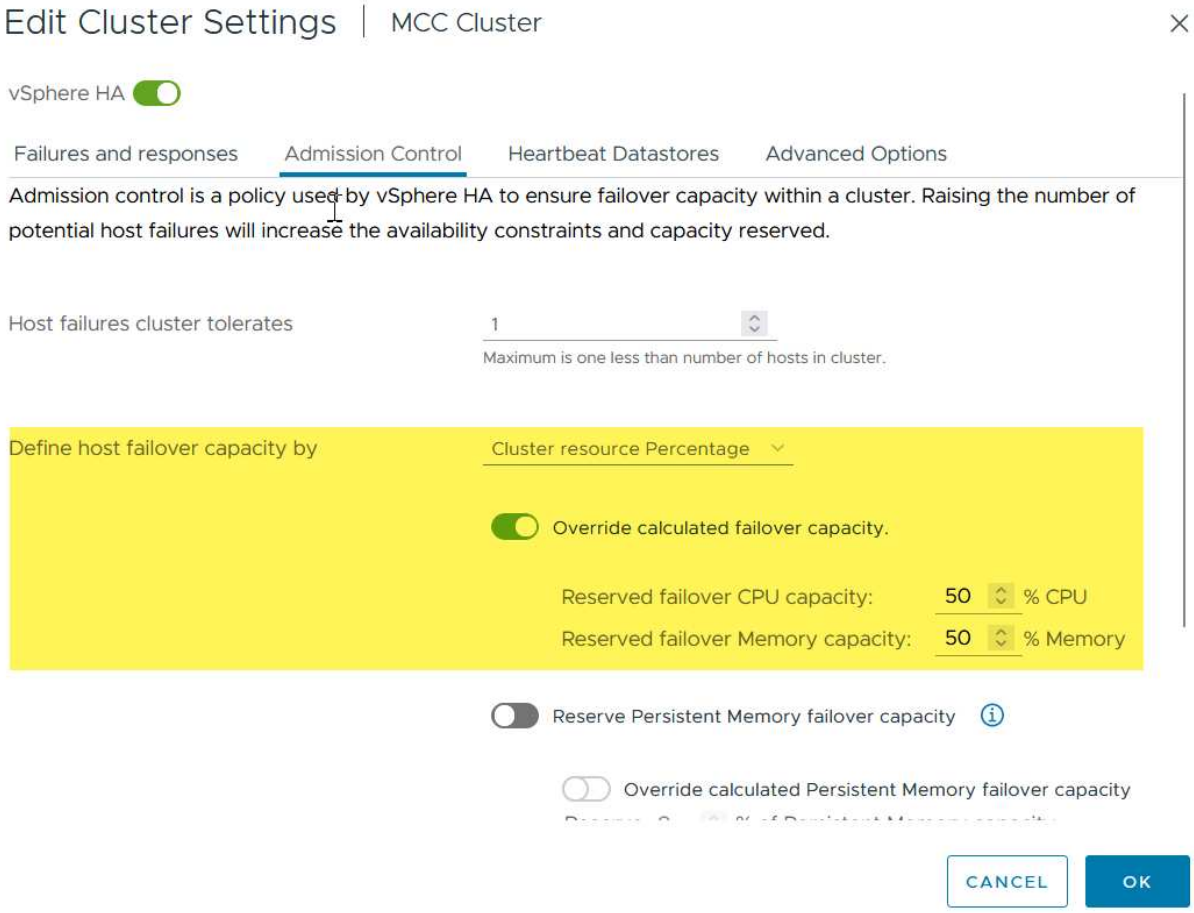
1. クラスタを選択し、[Configure]タブに移動します。[vSphere HA]を選択し、[edit]をクリック
2. [Host Monitoring]で、[Enable Host Monitoring]オプションを選択します。



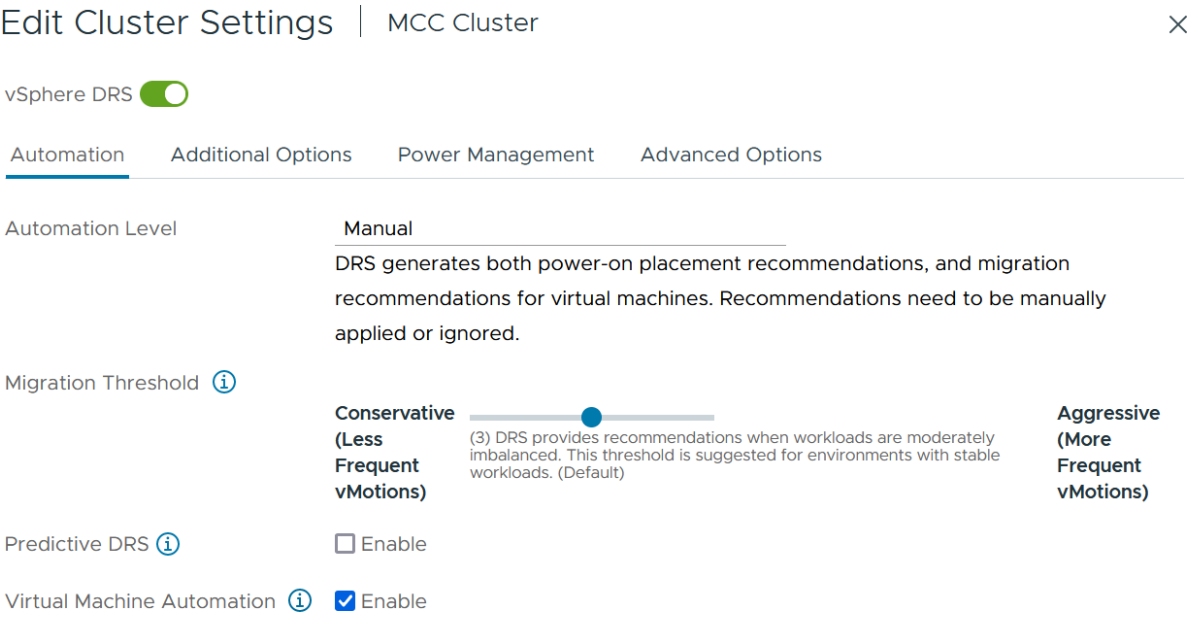
1. [Failures and Responses]タブの[VM Monitoring]で、[VM Monitoring Only]オプションまたは[VM and Application Monitoring]オプションを選択します。



1. [Admission Control]で、[HA Admission Control]オプションを[cluster resource reserve]に設定し、50%のCPU/MEMを使用します。



1. [OK]をクリックします。
2. [DRS]を選択し、[編集]をクリックします。
3. アプリケーションで必要な場合を除き、自動化レベルを手動に設定します。



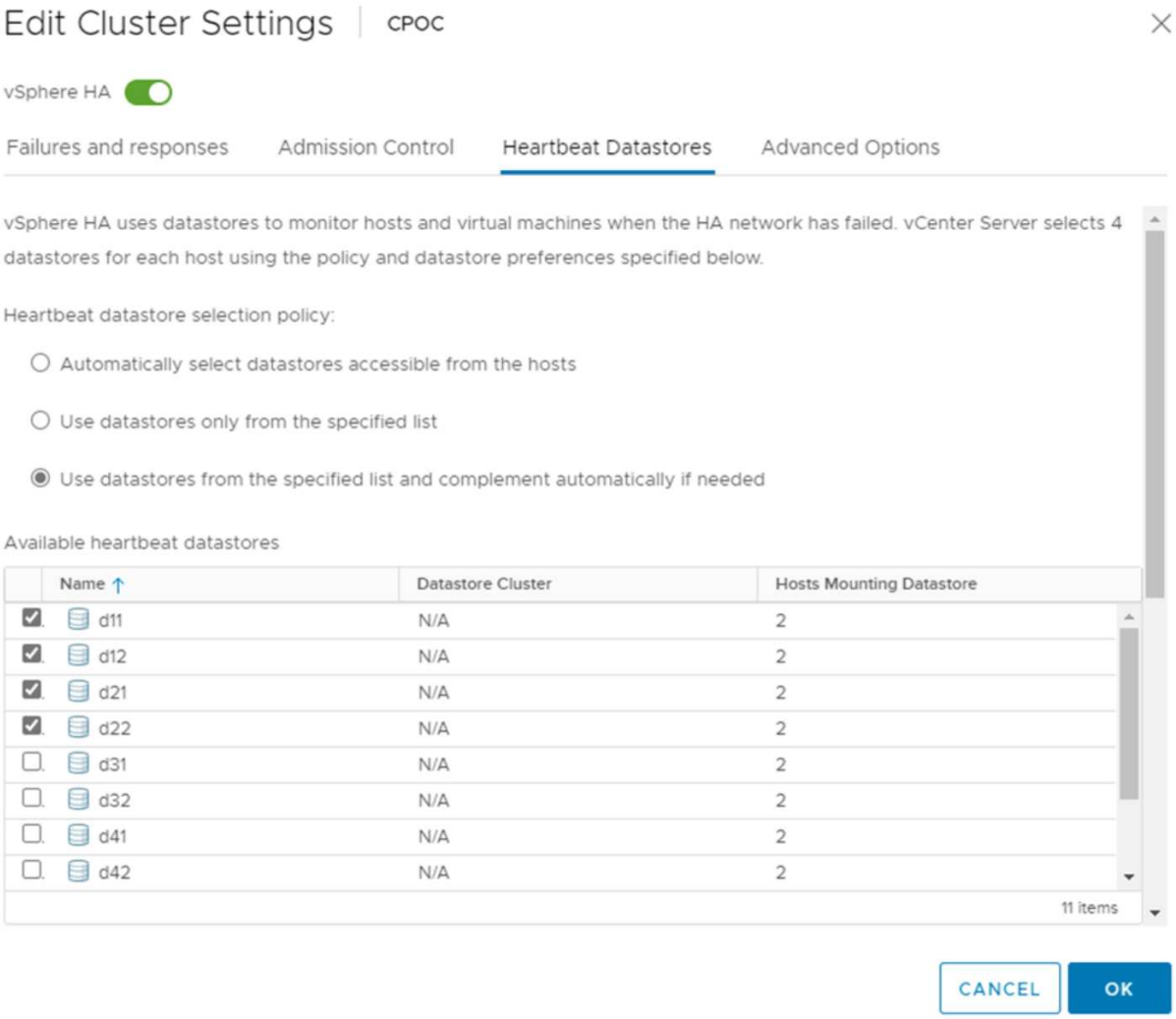
1. VMコンポーネント保護を有効にします。を参照してください。 "docs.vmware.com"。
2. MCCを使用するvMSCでは、次のvSphere HAの追加設定が推奨されます。

失敗	応答
ホスト障害です	VMの再起動
ホストの分離	無効
Permanent Device Loss（PDL；永続的デバイス損失）のあるデータストア	VMの電源をオフにして再起動する
すべてのパスがダウンしているデータストア（APD）	VMの電源をオフにして再起動する
ゲストが鼓動しない	VMのリセット
VM再起動ポリシー	VMの重要度に応じて決定
ホスト隔離時の応答	VMのシャットダウンと再起動
PDLを使用したデータストアの応答	VMの電源をオフにして再起動する
APDを使用するデータストアの応答	VMの電源をオフにして再起動する（控えめ）
APDのVMフェイルオーバーの遅延	3分
APDタイムアウトによるAPDリカバリの応答	無効
VM監視の感度	プリセット高

ハートビート用のデータストアの設定

vSphere HAでは、管理ネットワークに障害が発生した場合、データストアを使用してホストと仮想マシンを監視します。vCenterでのハートビートデータストアの選択方法を設定できます。ハートビート用のデータストアを設定するには、次の手順を実行します。

1. [Datastore Heartbeating]セクションで、[Use Datastores from the Specified List and Complement Automatically if Needed]を選択します。
2. vCenterで使用するデータストアを両方のサイトから選択し、[OK]を押します。

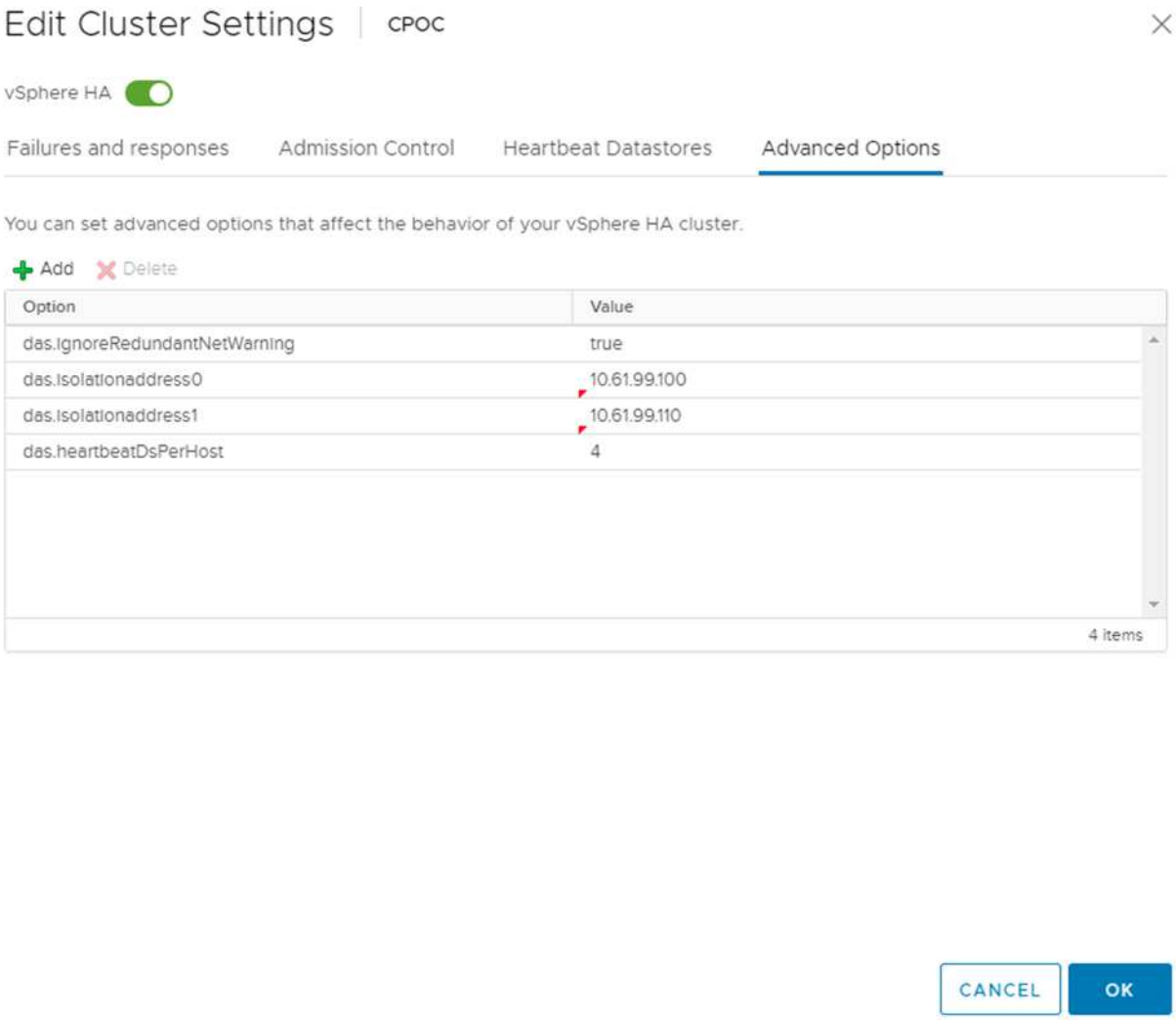


詳細オプションの設定

ホスト障害の検出

HAクラスタ内のホストがネットワークまたはクラスタ内の他のホストに接続できなくなると、分離イベントが発生します。デフォルトでは、vSphere HAは管理ネットワークのデフォルトゲートウェイをデフォルトの分離アドレスとして使用します。ただし、ホストがpingを実行するための追加の隔離アドレスを指定して、隔離応答をトリガーするかどうかを判断することができます。pingを実行できる隔離IPをサイトごとに1つずつ追加します。ゲートウェイIPは使用しないでください。使用するvSphere HAの詳細設定はdas.isolationaddressです。この目的には、ONTAPまたはメディエーターのIPアドレスを使用できます。

を参照してください "core.vmware.com" 詳細については、_を参照してください。



das.heartbeatDsPerHostという詳細設定を追加すると、ハートビートデータストアの数を増やすことができます。4つのハートビートデータストア（HB DSS）（サイトごとに2つ）を使用します。[Select from List but complent]オプションを使用します。これは、1つのサイトで障害が発生してもHB DSSが2つ必要になるためです。ただし、MCCやSnapMirrorのアクティブな同期で保護する必要はありません。

を参照してください "core.vmware.com" 詳細については、_を参照してください。

NetApp MetroCluster向けVMware DRSアフィニティ

このセクションでは、MetroCluster環境内のサイト/クラスタごとに、VMとホストのDRSグループを作成します。次に、VMホストアフィニティをローカルストレージリソースとアライメントするようにVM\Hostルールを設定します。たとえば、サイトAのVMがVMグループsitea_vmsに属し、サイトAのホストがホストグループsitea_hostsに属しているとします。次に、VM\Hostルールで、sitea_vmsをsitea_hostsのホストで実行するように記述します。

ベストプラクティス

- NetAppでは、「Must Run on Hosts in Group」という仕様ではなく、「Should Run on Hosts in Group」という仕様を使用することを強く推奨しています。サイトAのホストで障害が発生した場合、vSphere HAを使用してサイトAのVMをサイトBのホストで再起動する必要がありますが、後者の仕様では、HAがサイトBのVMを再起動することは難しいルールであるため許可されていません。前者の仕様はソフトルールで

あり、HAが発生した場合は違反となるため、パフォーマンスではなく可用性が確保されます。

*注：*仮想マシンがVMとホストのアフィニティルールに違反したときにトリガーされるイベントベースのアラームを作成できます。vSphere Clientで、仮想マシンの新しいアラームを追加し、イベントトリガーとして[VM is violating VM-Host Affinity Rule]を選択します。アラームの作成と編集の詳細については、を参照してください。 "vSphereの監視とパフォーマンス" ドキュメント

DRSホストグループの作成

サイトAとサイトBに固有のDRSホストグループを作成するには、次の手順を実行します。

1. vSphere Web Clientで、インベントリ内のクラスタを右クリックし、[Settings]を選択します。
2. [VM\Host Groups]をクリックします。
3. 追加をクリックします。
4. グループの名前を入力します（例：sitea_hosts）。
5. [Type]メニューから[Host Group]を選択します。
6. [Add]をクリックし、サイトAから目的のホストを選択して[OK]をクリックします。
7. 同じ手順を繰り返して、サイトBのホストグループをもう1つ追加します。
8. [OK] をクリックします。

DRS VMグループの作成

サイトAとサイトBに固有のDRS VMグループを作成するには、次の手順を実行します。

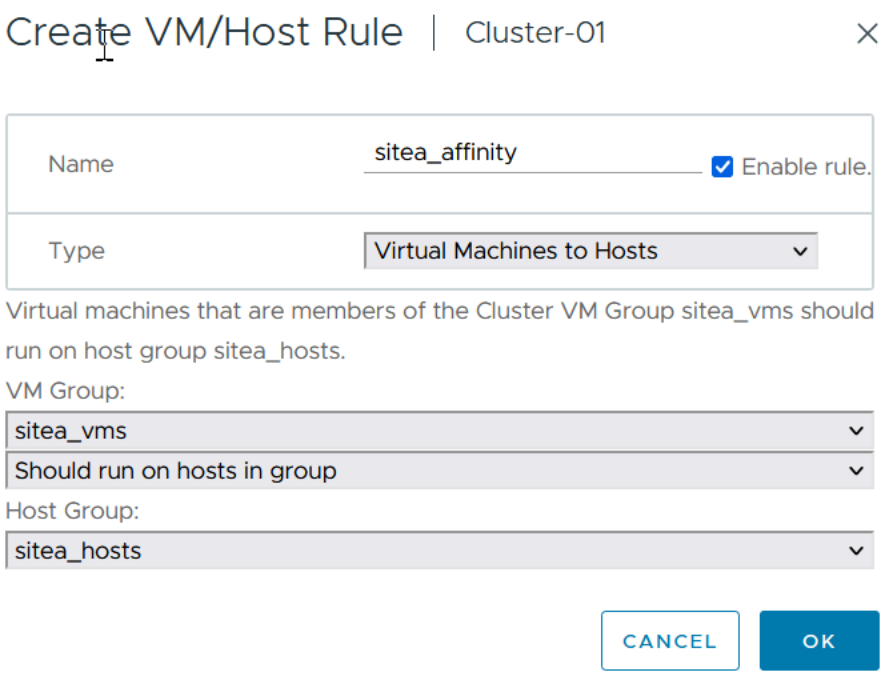
1. vSphere Web Clientで、インベントリ内のクラスタを右クリックし、[Settings]を選択します。
2. [VM\Host Groups]をクリックします。
3. 追加をクリックします。
4. グループの名前を入力します（例：sitea_vms）。
5. [Type]メニューから[VM Group]を選択します。
6. [Add]をクリックし、サイトAから目的のVMを選択して[OK]をクリックします。
7. 同じ手順を繰り返して、サイトBのホストグループをもう1つ追加します。
8. [OK] をクリックします。

VMホストルールの作成

サイトAとサイトBに固有のDRSアフィニティルールを作成するには、次の手順を実行します。

1. vSphere Web Clientで、インベントリ内のクラスタを右クリックし、[Settings]を選択します。
2. [VM\Host Rules]をクリックします。
3. 追加をクリックします。
4. ルールの名前を入力します（例：sitea_affinity）。
5. Enable Ruleオプションがオンになっていることを確認します。

6. [Type]メニューから[Virtual Machines to Hosts]を選択します。
7. VMグループを選択します（例：sitea_vms）。
8. ホストグループを選択します（例：sitea_hosts）。
9. 同じ手順を繰り返して、サイトBのVM\Hostルールをもう1つ追加します。
10. [OK] をクリックします。



NetApp MetroCluster向けVMware vSphere Storage DRS

データストアクラスタの作成

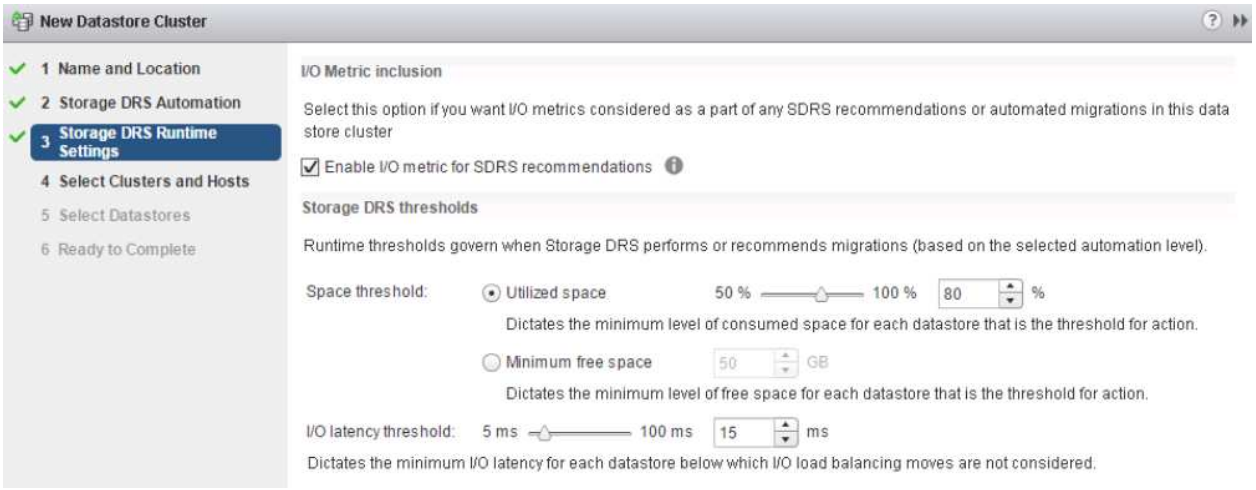
各サイトのデータストアクラスタを設定するには、次の手順を実行します。

1. vSphere Web Clientを使用して、[Storage]の下にあるHAクラスタが配置されているデータセンターに移動します。
2. データセンターオブジェクトを右クリックし、[Storage]>[New Datastore Cluster]を選択します。
3. [Turn on Storage DRS]オプションを選択し、[Next]をクリックします。
4. すべてのオプションを[No Automation（Manual Mode）]に設定し、[Next]をクリックします。

ベストプラクティス

- NetAppでは、移行が必要になるタイミングを管理者が判断して制御できるように、Storage DRSを手動モードで設定することを推奨しています。

1. [Enable I/O Metric for SDRS Recommendations]チェックボックスがオンになっていることを確認します。メトリック設定はデフォルト値のままにできます。



1. HAクラスタを選択し、[Next]をクリックします。

1. サイトAに属するデータストアを選択し、[Next]をクリックします。

1. オプションを確認し、[完了]をクリックします。
2. 同じ手順を繰り返してサイトBのデータストアクラスタを作成し、サイトBのデータストアのみが選択されていることを確認します。

vCenter Serverの可用性

vCenter Server Appliance（VCSA）はvCenter HAで保護する必要があります。vCenter HAでは、アクティブ/パッシブHAペアに2つのVCSAを導入できます。障害ドメインごとに1つ。vCenter HAの詳細については、"docs.vmware.com"。

計画的イベントと計画外イベントの耐障害性

NetApp MetroClusterとSnapMirrorのアクティブ同期は、NetAppハードウェアとONTAP®ソフトウェアの高可用性とノンストップオペレーションを強化する強力なツールです。

これらのツールは、ストレージ環境全体をサイト全体で保護し、データの可用性を確保します。スタンドアロンサーバ、高可用性サーバクラスタ、Dockerコンテナ、仮想サーバのいずれを使用している場合でも、NetAppテクノロジは、停電、冷却装置の障害、ネットワーク接続の障害、ストレージアレイのシャットダウン、または運用上のエラーが原因で全体が停止した場合でも、ストレージの可用性をシームレスに維持します。

MetroClusterとSnapMirrorのアクティブな同期では、計画的または計画外のイベントが発生した場合に、次の3つの基本的な方法でデータを継続できます。

- 冗長コンポーネントによる単一コンポーネント障害からの保護
- ローカルのHAテイクオーバー：1台のコントローラに影響するイベントに対応
- 完全なサイト保護–ストレージおよびクライアントのアクセスをソースクラスタからデスティネーションクラスタに移動することで、サービスを迅速に再開します。

つまり、1つのコンポーネントで障害が発生してもシームレスに運用が継続され、障害が発生したコンポーネントを交換すると自動的に冗長運用に戻ります。

シングルノードクラスタ（通常はONTAP Selectなどのソフトウェア定義バージョン）を除くすべてのONTAPクラスタには、テイクオーバーとギブバックと呼ばれるHA機能が組み込まれています。クラスタ内の各コントローラが別のコントローラとペアリングされ、HAペアが形成されます。これらのペアにより、各ノードはストレージにローカルで接続されます。

テイクオーバーは、データサービスを維持するために一方のノードがもう一方のノードのストレージをテイクオーバーする自動プロセスです。ギブバックは、通常動作に戻る逆のプロセスです。テイクオーバーは、ハードウェアのメンテナンス時やONTAPのアップグレード時などに計画的に行うことも、ノードのパニックやハードウェア障害による計画外で行うこともできます。

テイクオーバー時に、MetroCluster構成のネットワーク接続型ストレージ論理インターフェイス（NAS LIF）が自動的にフェイルオーバーされます。ただし、ストレージエリアネットワークLIF（SAN LIF）はフェイルオーバーせず、引き続き論理ユニット番号（LUN）への直接パスを使用します。

HAのテイクオーバーとギブバックの詳細については、 "HAペアの管理の概要"。この機能は、MetroClusterまたはSnapMirrorのアクティブな同期に固有ではないことに注意してください。

MetroClusterによるサイトのスイッチオーバーは、一方のサイトがオフラインになった場合、またはサイト全体のメンテナンスのために計画的に実行された場合に実行されます。オフラインになったクラスタのストレージリソース（ディスクおよびアグリゲート）の所有権がもう一方のサイトに引き継がれ、障害が発生したサイトのSVMがディザスタサイトでオンラインになって再起動されます。その際、クライアントとホストのアクセス用にIDは保持されます。

SnapMirrorのアクティブな同期では、両方のコピーが同時にアクティブに使用されるため、既存のホストは引き続き動作します。サイトのフェイルオーバーを正しく実行するには、NetAppメディエーターが必要です。

MCCを使用するvMSCの障害シナリオ

以降のセクションでは、vMSCおよびNetApp MetroClusterシステムで発生したさまざまな障害シナリオで想定される結果について説明します。

単一のストレージパス障害

このシナリオでは、HBAポート、ネットワークポート、フロントエンドデータスイッチポート、FCケーブル、イーサネットケーブルなどのコンポーネントで障害が発生すると、ストレージデバイスへの特定のパスがESXiホストによって停止とマークされます。HBA /ネットワーク/スイッチポートで耐障害性を提供してストレージデバイスに複数のパスが設定されている場合は、ESXiがパススイッチオーバーを実行するのが理想的です。この間、ストレージデバイスへの複数のパスを提供することでストレージの可用性が確保されるため、仮想マシンは影響を受けずに実行され続けます。

*注：*このシナリオではMetroClusterの動作に変更はなく、すべてのデータストアがそれぞれのサイトで引き続き実行されます。

ベストプラクティス

NFS / iSCSIボリュームを使用している環境ではNetApp、NFS vmkernelポート用に少なくとも2つのネットワークアップリンクを標準vSwitchに設定し、NFS vmkernelインターフェイスが分散vSwitchにマッピングされているポートグループに設定することを推奨します。NICチーミングは、アクティブ/アクティブまたはアクティブ/スタンバイのいずれかで設定できます。

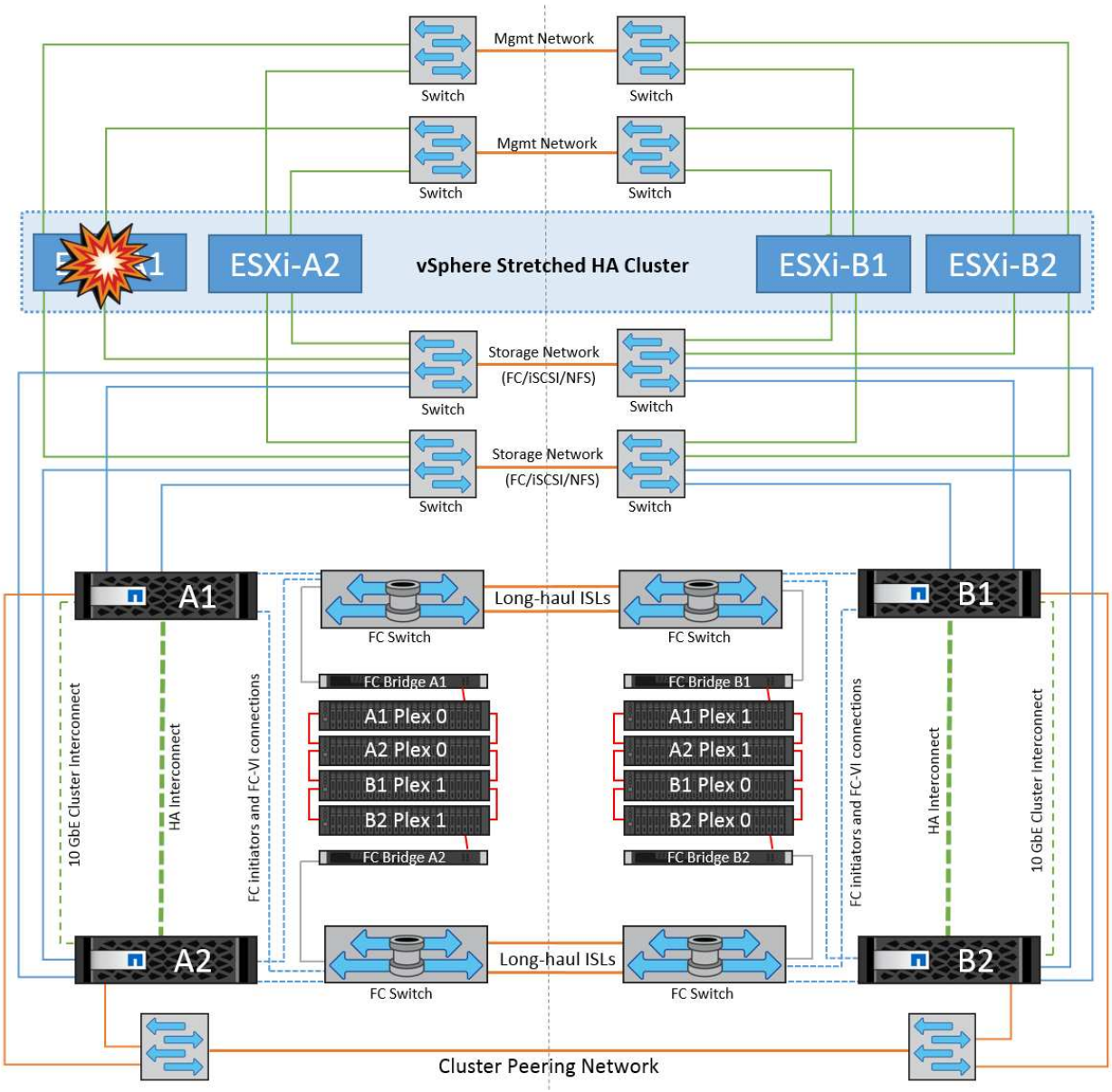
また、iSCSI LUNの場合は、vmkernelインターフェイスをiSCSIネットワークアダプタにバインドしてマルチパスを設定する必要があります。詳細については、vSphereストレージのドキュメントを参照してください。

ベストプラクティス

ファイバチャネルLUNを使用する環境でNetAppは、HBAを少なくとも2つ搭載し、HBA /ポートレベルでの耐障害性を保証することを推奨します。NetAppでは、ゾーニングを設定するためのベストプラクティスとして、単一のイニシエータから単一のターゲットへのゾーニングも推奨しています。

新規および既存のすべてのNetAppストレージデバイスにポリシーが設定されるため、Virtual Storage Console（VSC）を使用してマルチパスポリシーを設定する必要があります。

単一のESXiホスト障害



このシナリオでは、ESXiホストで障害が発生すると、VMware HAクラスタのマスターノードがネットワークハートビートを受信しなくなるため、ホスト障害を検出します。ホストが本当に停止しているのか、ネットワークパーティションだけなのかを判別するために、マスターノードはデータストアハートビートを監視し、ハートビートがない場合は、障害が発生したホストの管理IPアドレスに対してpingを実行して最終チェックを実行します。これらのチェックがすべて無効の場合、マスターノードはこのホストを障害が発生したホストであると宣言し、この障害が発生したホストで実行されていたすべての仮想マシンが、クラスタ内の残りのホストでリブートされます。

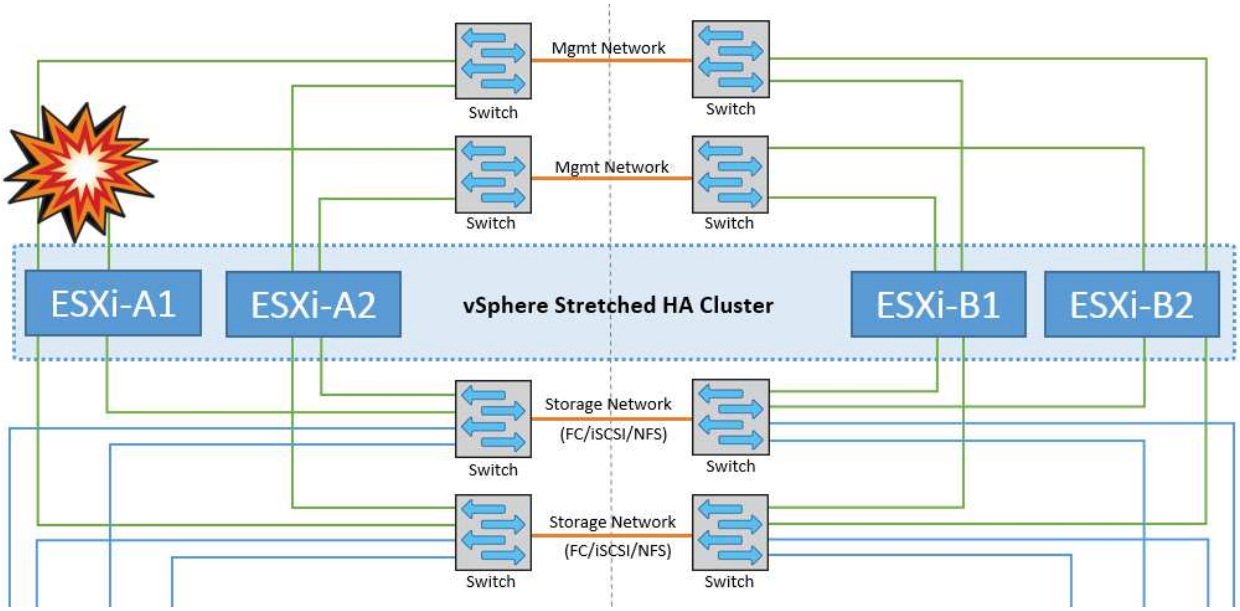
DRSのVMとホストのアフィニティルールが設定されている場合（VMグループsitea_vmsのVMはホストグループsitea_hostsのホストを実行する必要があります）、HAマスターは最初にサイトAで使用可能なリソースを確認します。サイトAに使用可能なホストがない場合、マスターはサイトBのホストでVMの再起動を試みます。

ローカルサイトのリソースに制約がある場合は、もう一方のサイトのESXiホストで仮想マシンが起動される可能性があります。ただし、DRSのVMとホストのアフィニティルールに違反した場合は、仮想マシンをローカルサイトの稼働しているESXiホストに移行することで修正されます。DRSが手動に設定されている場合、NetAppはDRSを起動し、推奨事項を適用して仮想マシンの配置を修正することを推奨します。

このシナリオではMetroClusterの動作に変更はなく、すべてのデータストアがそれぞれのサイトで引き続き実

行されます。

ESXiホストの分離

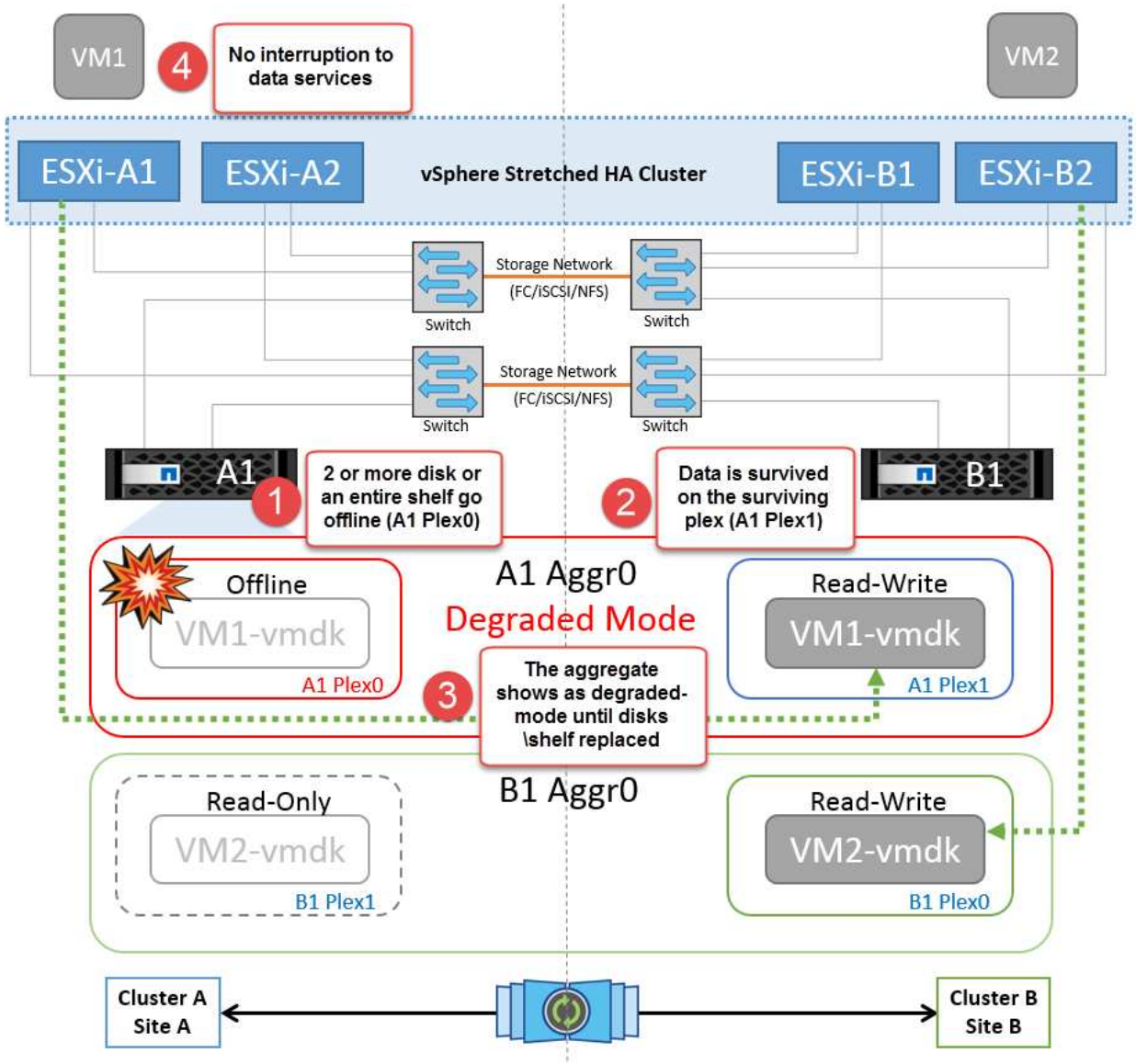


このシナリオでは、ESXiホストの管理ネットワークが停止すると、HAクラスタ内のマスターノードがハートビートを受信しなくなり、このホストがネットワークから分離された状態になります。障害が発生したか、隔離されているだけかを判別するために、マスターノードはデータストアハートビートの監視を開始します。ホストが存在する場合、ホストはマスターノードによって分離されていると宣言されます。構成されている隔離時の対応に応じて、ホストは仮想マシンの電源をオフにするか、シャットダウンするか、仮想マシンの電源をオンにしたままにするかを選択できます。分離応答のデフォルトの間隔は30秒です。

このシナリオではMetroClusterの動作に変更はなく、すべてのデータストアがそれぞれのサイトで引き続き実行されます。

ディスクシェルフの障害

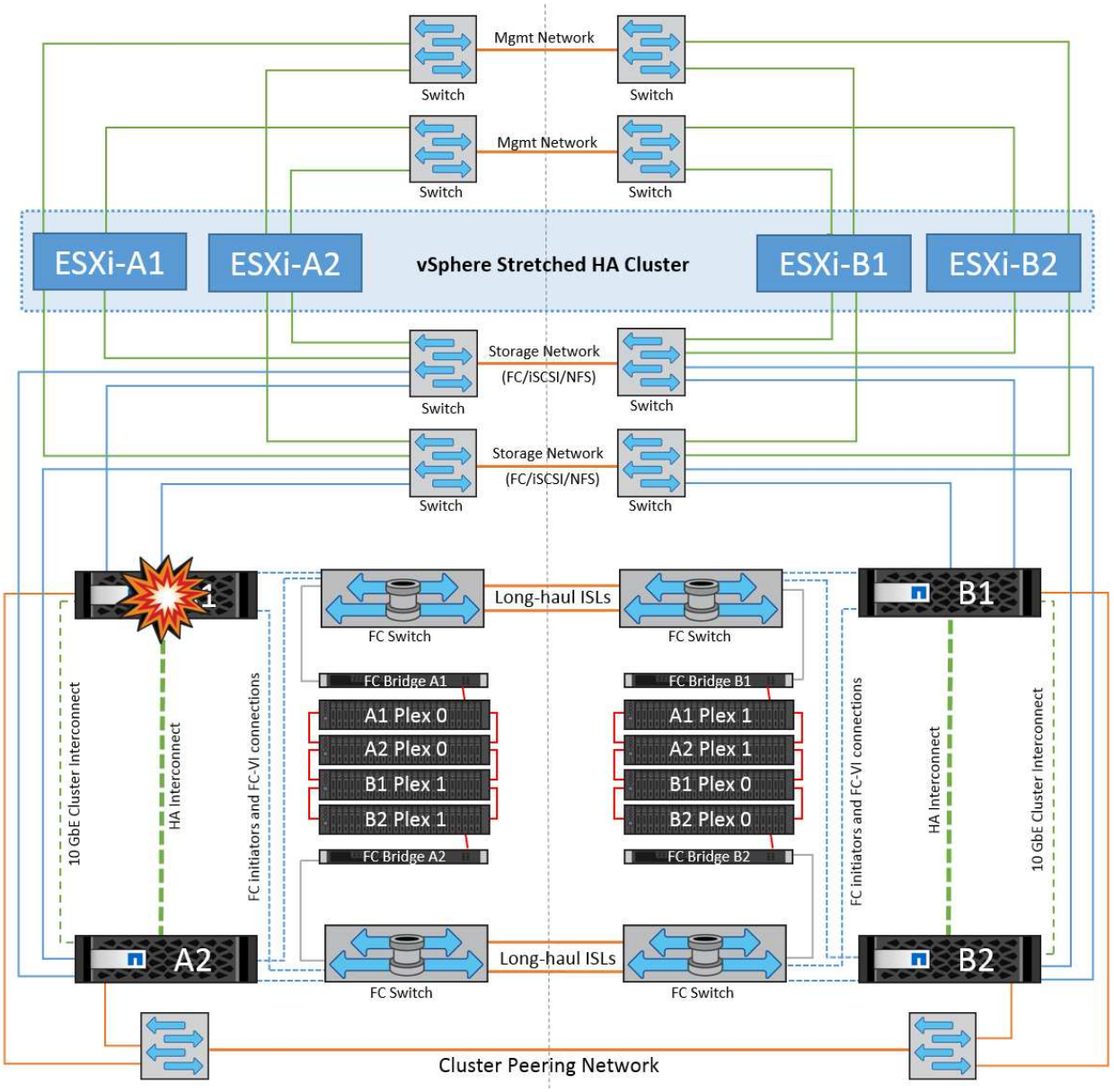
このシナリオでは、3本以上のディスクまたはシェルフ全体で障害が発生しています。データは、データサービスを中断することなく、稼働しているプレックスから提供されます。ディスク障害は、ローカルまたはリモートのプレックスに影響する可能性があります。アクティブなプレックスが1つしかないため、アグリゲートはデグレードモードになります。障害が発生したディスクを交換すると、影響を受けたアグリゲートが自動的に再同期されてデータが再構築されます。再同期後、アグリゲートは自動的に通常のミラーモードに戻ります。単一のRAIDグループ内の3本以上のディスクで障害が発生した場合は、プレックスを最初から再構築する必要があります。



*注：*この間、仮想マシンのI/O処理への影響はありませんが、データはISLリンクを介してリモートのディスクシェルフからアクセスされるため、パフォーマンスが低下します。

単一のストレージコントローラ障害

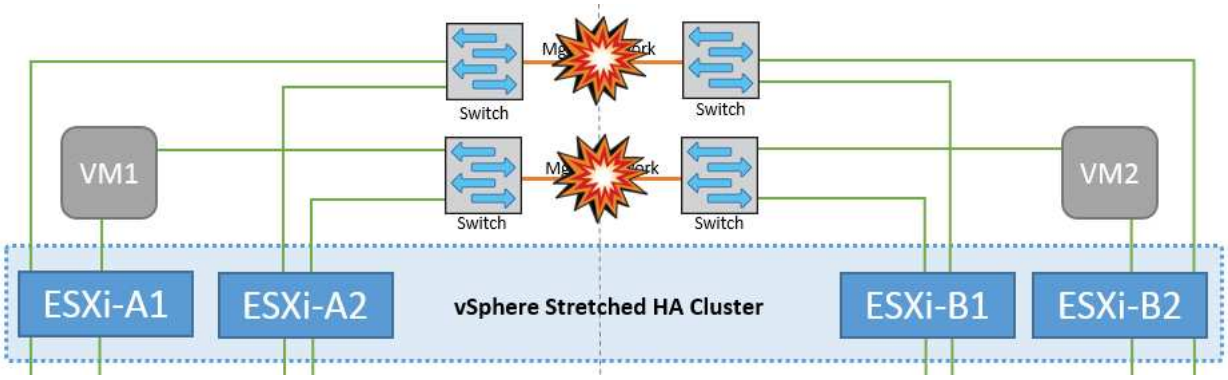
このシナリオでは、一方のサイトの2台のストレージコントローラのどちらかで障害が発生します。各サイトにHAペアがあるため、一方のノードで障害が発生すると、もう一方のノードへのフェイルオーバーが透過的かつ自動的にトリガーされます。たとえば、ノードA1に障害が発生した場合、そのストレージとワークロードは自動的にノードA2に転送されます。すべてのプレックスが引き続き使用可能なため、仮想マシンに影響はありません。2つ目のサイトのノード（B1とB2）は影響を受けません。また、クラスタ内のマスターノードは引き続きネットワークハートビートを受信するため、vSphere HAによる処理は行われません。



フェイルオーバーがローリングディザスタ（ノードA1からA2にフェイルオーバー）の一部である場合に、その後A2またはサイトA全体で障害が発生すると、災害後にサイトBでスイッチオーバーが発生する可能性があります。

スイッチ間リンクの障害

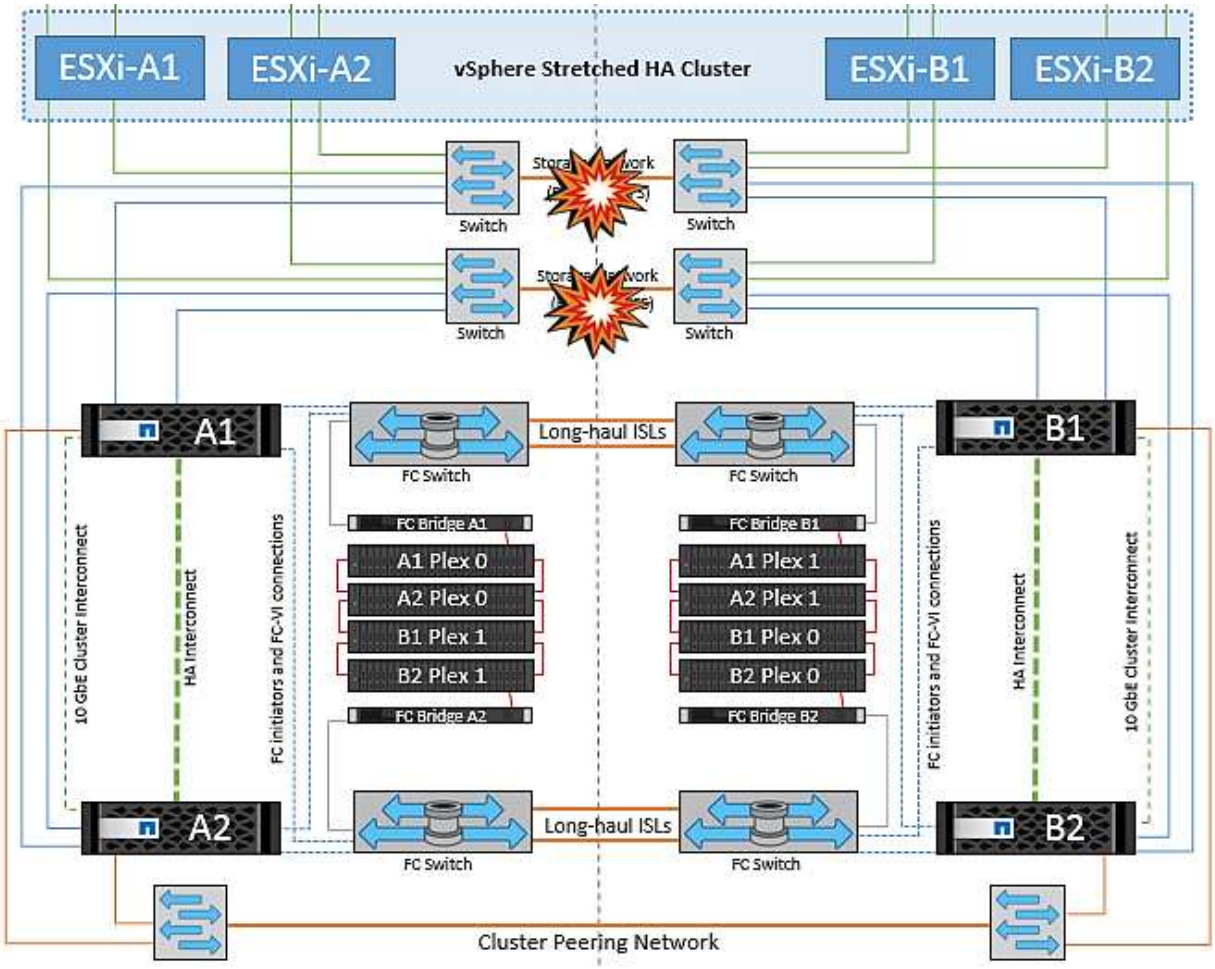
管理ネットワークでのスイッチ間リンク障害



このシナリオでは、フロントエンドホスト管理ネットワークのISLリンクで障害が発生し、サイトAのESXiホストがサイトBのESXiホストと通信できなくなります。これにより、特定のサイトのESXiホストからHAクラスタ内のマスターノードにネットワークハートビートを送信できなくなるため、ネットワークが分割されます。そのため、パーティションのために2つのネットワークセグメントがあり、各セグメントにマスターノードがあり、特定のサイト内でVMがホスト障害から保護されます。

*注：*この間、仮想マシンは実行されたままであり、このシナリオではMetroClusterの動作に変更はありません。すべてのデータストアがそれぞれのサイトで引き続き実行されます。

ストレージネットワークのスイッチ間リンク障害

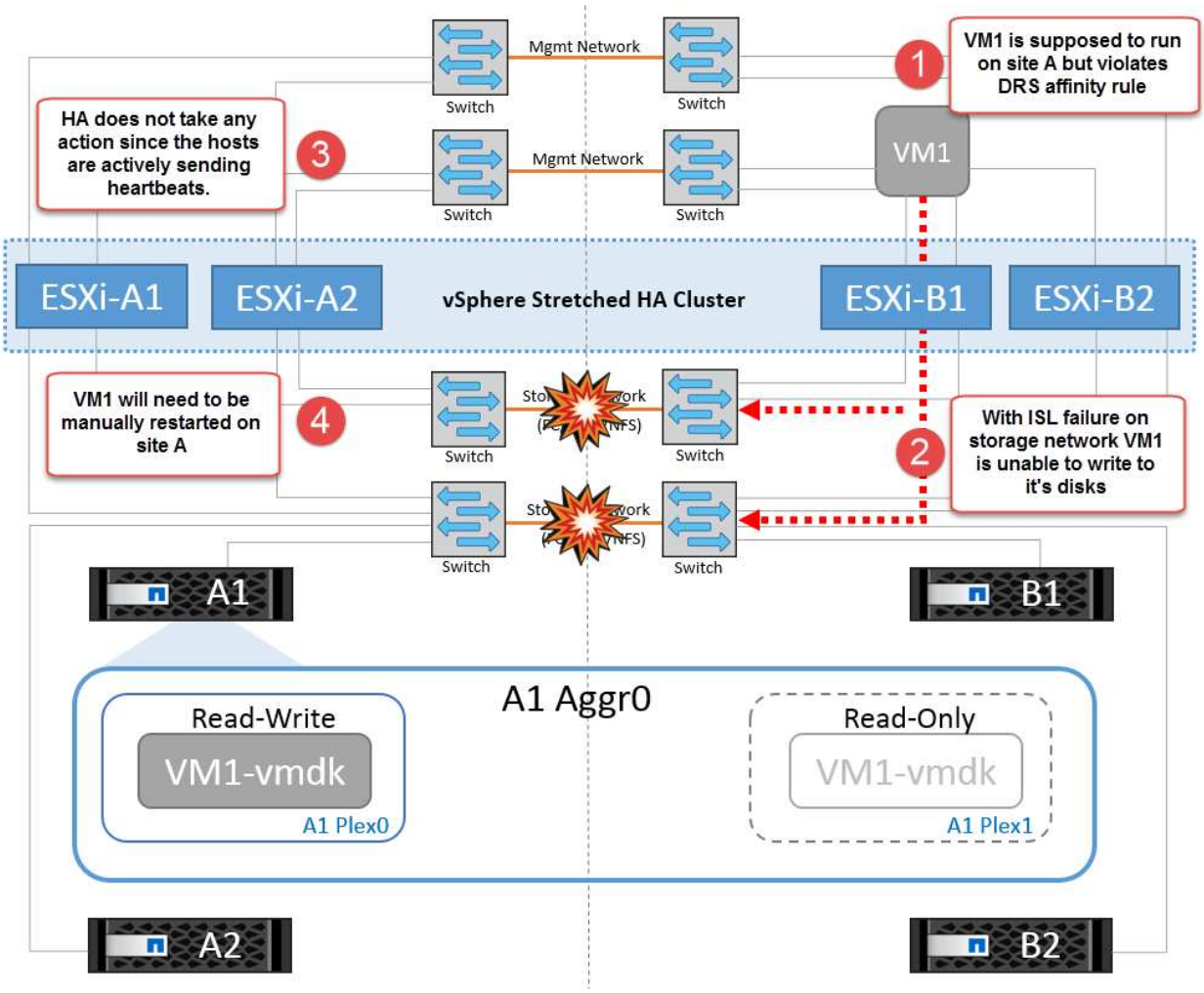


このシナリオでは、バックエンドストレージネットワークのISLリンクで障害が発生すると、サイトAのホストはサイトBのクラスタBのストレージボリュームまたはLUNにアクセスできなくなります。その逆も同様で

す。VMware DRSルールは、ホストとストレージサイトのアフィニティによって、サイト内で影響を与えることなく仮想マシンを実行できるように定義されています。

この間、仮想マシンはそれぞれのサイトで実行されたままになり、このシナリオではMetroClusterの動作に変更はありません。すべてのデータストアがそれぞれのサイトで引き続き実行されます。

何らかの理由でアフィニティルールに違反した場合（ローカルクラスタAのノードにディスクが配置されているサイトAから実行されていたVM1がサイトBのホストで実行されている場合など）、仮想マシンのディスクにISLリンクを介してリモートからアクセスされます。ISLリンクで障害が発生すると、ストレージボリュームへのパスが停止し、その仮想マシンが停止するため、サイトBで実行されているVM1はディスクに書き込むことができなくなります。この場合、ホストからハートビートがアクティブに送信されるため、VMware HAによる処理は行われません。これらの仮想マシンは、それぞれのサイトで手動で電源をオフにしてオンにする必要があります。次の図は、VMがDRSアフィニティルールに違反していることを示しています。

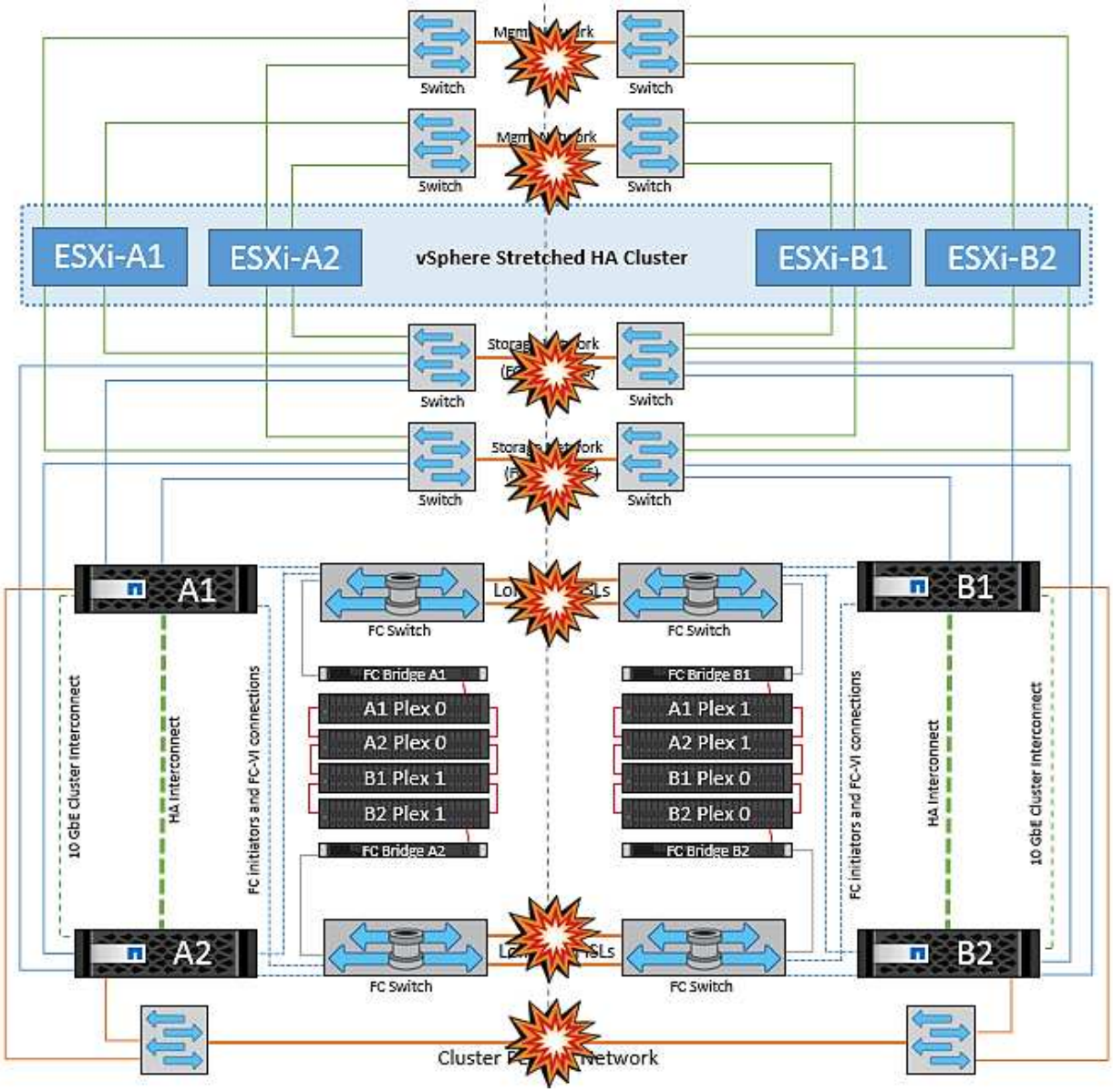


すべてのスイッチ間障害またはデータセンターの完全なパーティション

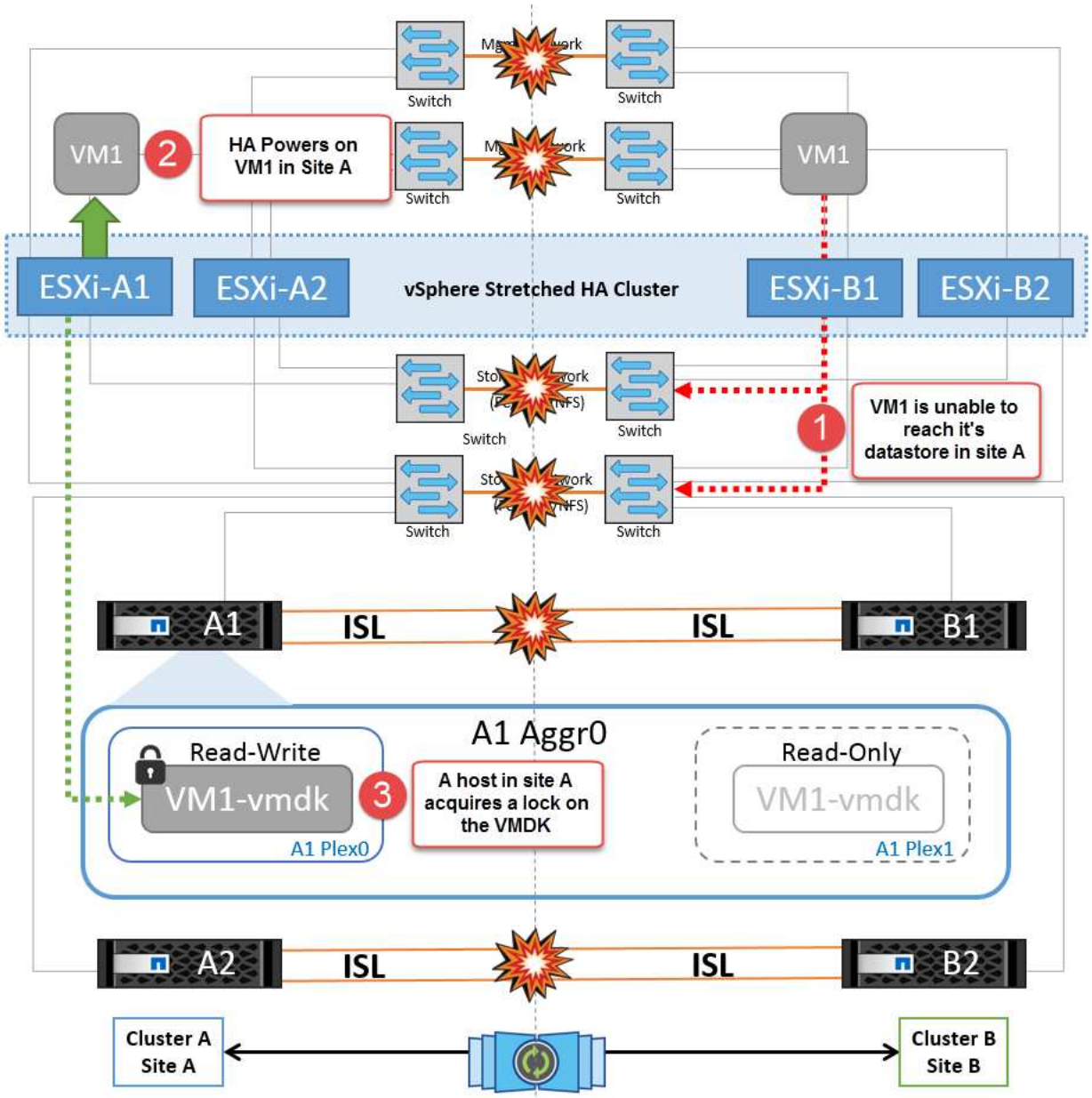
このシナリオでは、サイト間のすべてのISLリンクが停止し、両方のサイトが相互に分離されます。管理ネットワークやストレージネットワークでのISL障害などのシナリオで説明したように、ISL全体で障害が発生しても仮想マシンは影響を受けません。

ESXiホストがサイト間でパーティショニングされると、vSphere HAエージェントがデータストアハートビートをチェックし、各サイトでローカルのESXiホストがデータストアハートビートを対応する読み書き可能なボリューム/ LUNに更新できるようになります。サイトAのホストは、ネットワーク/データストアハートビートがないため、サイトBの他のESXiホストで障害が発生したと見なします。サイトAのvSphere HAはサイトB

の仮想マシンの再起動を試行しますが、ストレージISLの障害が原因でサイトBのデータストアにアクセスできなくなるため、再起動は失敗します。同様の状況がサイトBでも繰り返されます。



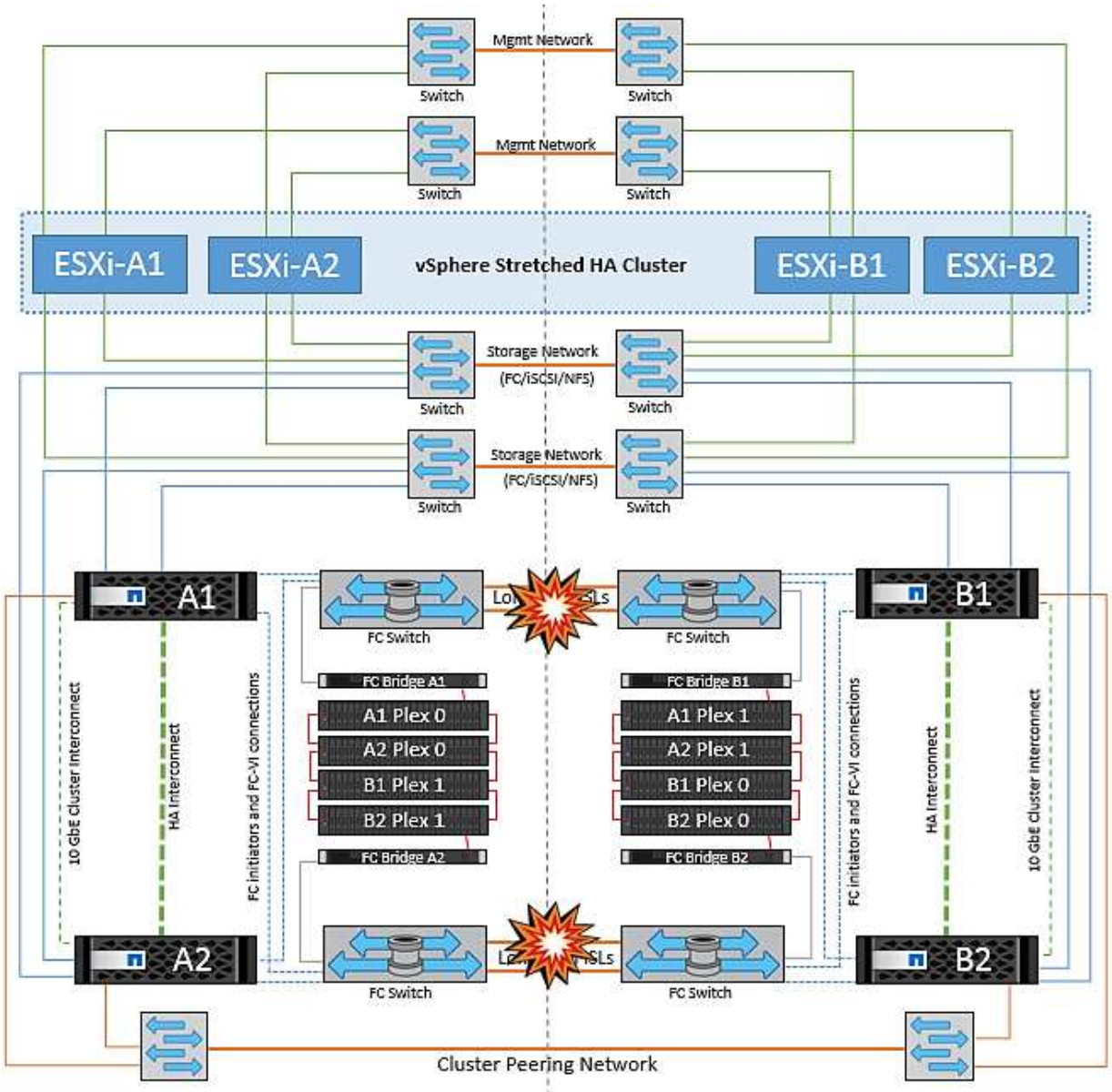
NetAppでは、DRSルールに違反した仮想マシンがないかどうかを確認することを推奨しています。リモートサイトから実行されている仮想マシンはデータストアにアクセスできないため停止し、vSphere HAはその仮想マシンをローカルサイトで再起動します。ISLリンクがオンラインに戻ると、同じMACアドレスで仮想マシンのインスタンスが2つ実行されることはないため、リモートサイトで実行されていた仮想マシンが強制終了されます。



NetApp MetroClusterの両方のファブリックのスイッチ間リンク障害

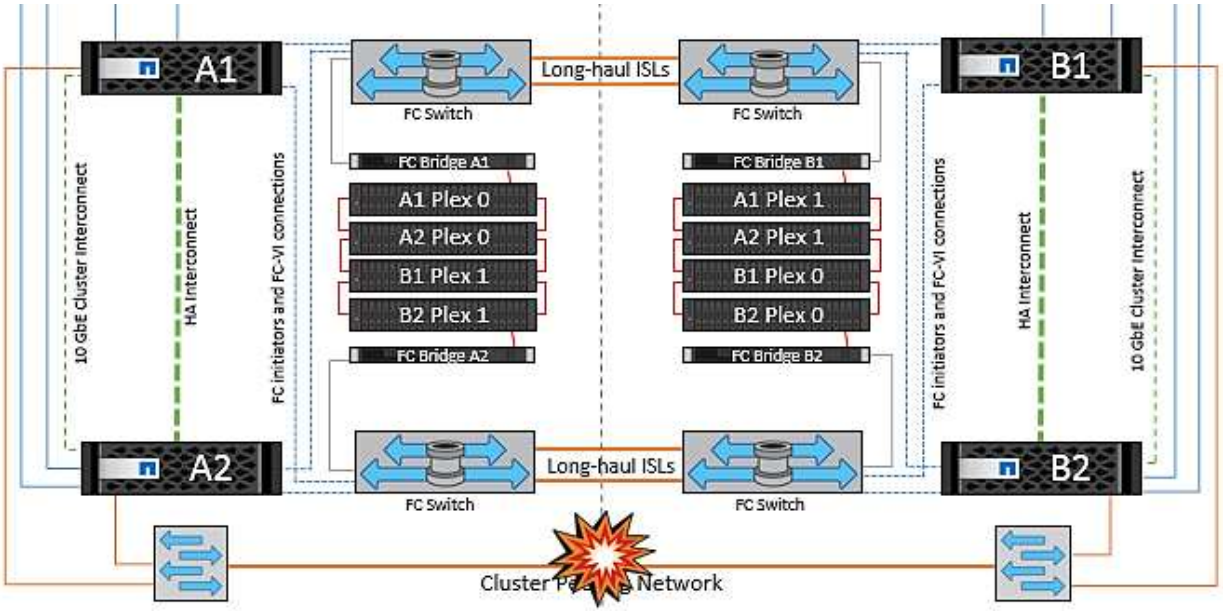
1つ以上のISLで障害が発生した場合、トラフィックは残りのリンクを経由して続行されます。両方のファブリックのすべてのISLで障害が発生し、ストレージとNVRAMのレプリケーション用のサイト間のリンクがなくなった場合、各コントローラはローカルデータの提供を継続します。少なくとも1つのISLをリストアすると、すべてのプレックスの再同期が自動的に実行されます。

すべてのISLが停止したあとに発生した書き込みは、もう一方のサイトにミラーリングされません。そのため、構成がこの状態のときに災害時にスイッチオーバーを実行すると、同期されていなかったデータが失われます。この場合、スイッチオーバー後のリカバリを手動で行う必要があります。ISLが長期間使用できなくなる可能性がある場合は、災害時のスイッチオーバーが必要な場合にデータ損失のリスクを回避するために、すべてのデータサービスをシャットダウンすることができます。この処理を実行するかどうかは、少なくとも1つのISLが使用可能になる前にスイッチオーバーが必要な災害が発生する可能性と比較して判断する必要があります。また、ISLで連鎖的に障害が発生した場合は、すべてのリンクで障害が発生する前に、いずれかのサイトへの計画的スイッチオーバーをトリガーすることもできます。



ピアクラスタのリンク障害

ピアクラスタのリンクで障害が発生した場合、ファブリックのISLはアクティブなままであるため、データサービス（読み取りと書き込み）は両方のサイトで両方のプレックスに対して継続されます。クラスタ設定の変更（新しいSVMの追加、既存のSVMでのボリュームやLUNのプロビジョニングなど）は、もう一方のサイトに伝播できません。これらはローカルのCRSメタデータボリュームに保持され、ピアクラスタリンクのリストア時にもう一方のクラスタに自動的に伝播されます。ピアクラスタのリンクがリストアされる前に強制スイッチオーバーが必要な場合は、スイッチオーバープロセスの一環として、サバイバーサイトにあるメタデータボリュームのリモートレプリケートコピーから、未処理のクラスタ構成変更が自動的に再生されます。



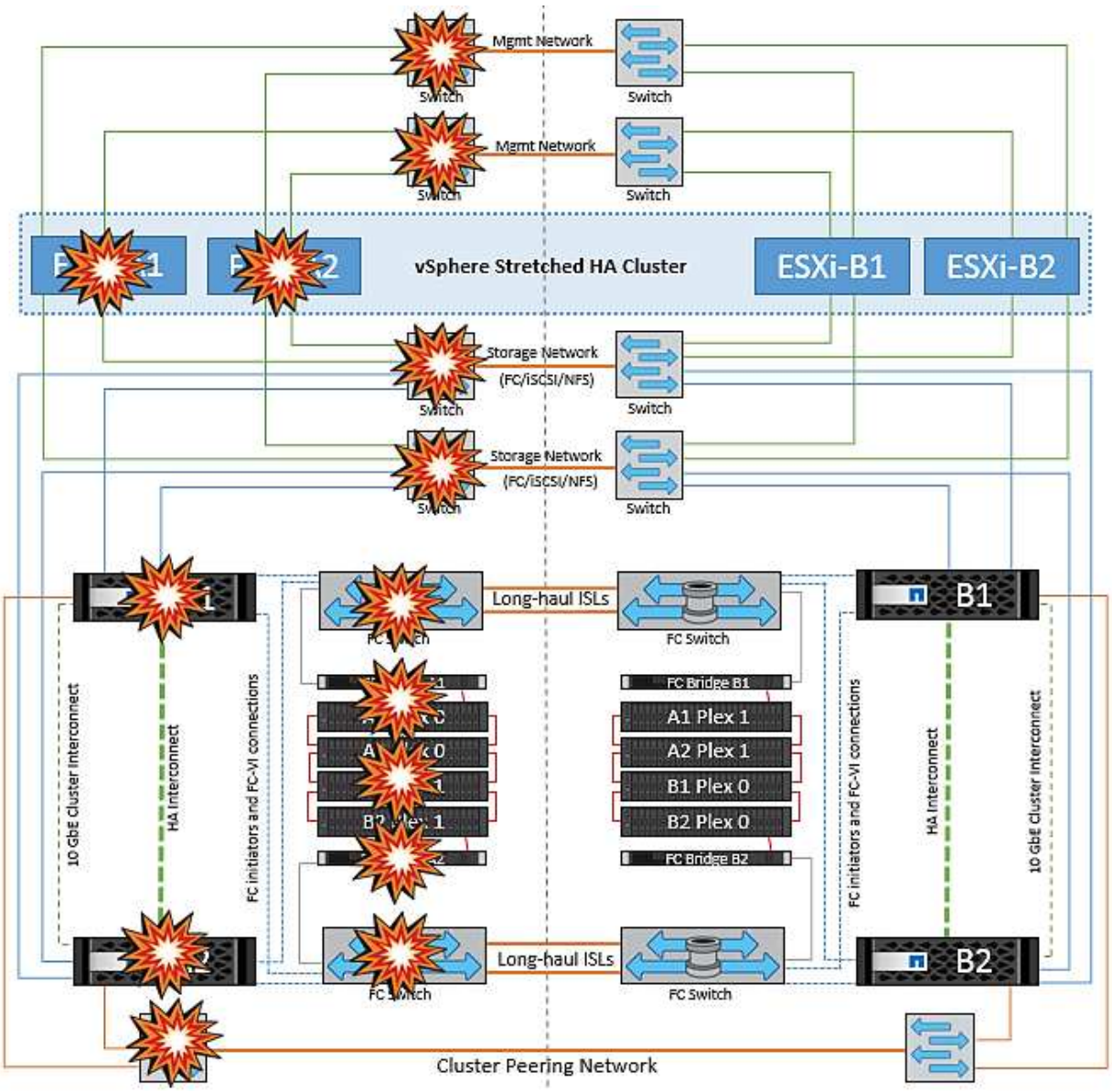
サイト全体の障害

サイトA全体で障害が発生した場合、サイトAのESXiホストが停止しているため、サイトBのESXiホストはサイトAのESXiホストからネットワークハートビートを受信しません。サイトBのHAマスターは、データストアハートビートが存在しないことを確認し、サイトAのホストで障害が発生したことを宣言して、サイトAの仮想マシンをサイトBで再起動しようとします。この間に、ストレージ管理者はスイッチオーバーを実行して障害が発生したノードのサービスをサバイバーサイトで再開し、サイトAのすべてのストレージサービスをサイトBでリストアします。サイトAのボリュームまたはLUNがサイトBで使用可能になると、HAマスターエージェントはサイトAの仮想マシンをサイトBで再起動しようとします。

vSphere HAマスターエージェントがVMの再起動（VMの登録と電源投入を含む）に失敗した場合、遅延後に再起動が再試行されます。再起動の間隔は、最大30分まで設定できます。vSphere HAは、再起動を最大試行回数（デフォルトでは6回）試行します。

注： HAマスターは、Placement Managerが適切なストレージを検出するまで再起動の試行を開始しません。したがって、サイト全体で障害が発生した場合は、スイッチオーバーの実行後に再起動が試行されます。

サイトAがスイッチオーバーされた場合は、サバイバーサイトBのいずれかのノードで障害が発生しても、サバイバーノードにフェイルオーバーすることでシームレスに対応できます。この場合、4つのノードの作業は1つのノードだけで実行されます。この場合のリカバリでは、ローカルノードへのギブバックを実行します。その後、サイトAがリストアされるとスイッチバック処理が実行され、構成の安定した運用が再開されます。


Mgmt Network
Switch
Mgmt Network
Switch
vSphere Stretched HA Cluster
ESXi-B1
ESXi-B2
Storage Network
(FC/iSCSI/NFS)
Switch
Storage Network
(FC/iSCSI/NFS)
Switch
Long-haul ISLs
FC Switch
B1
FC Bridge B1
A1 Plex 1
A2 Plex 1
B1 Plex 0
B2 Plex 0
FC Bridge B2
10 GbE Cluster Interconnect
HA Interconnect
FC initiators and FC-VI connections
FC initiators and FC-VI connections
HA Interconnect
10 GbE Cluster Interconnect
Long-haul ISLs
FC Switch
B2
Cluster Peering Network

著作権に関する情報

商標に関する情報