高可用性アーキテクチャ
ONTAP Select

NetApp
February 26, 2026

目次

高可用性アーキテクチャ

ONTAP Selectの高可用性構成

高可用性オプションを確認して、環境に最適な HA 構成を選択します。

お客様は、アプリケーションワークロードをエンタープライズクラスのストレージアプライアンスからコモディティハードウェア上で稼働するソフトウェアベースのソリューションに移行し始めていますが、レジリエンス（回復力）とフォールトトレランスに対する期待とニーズは変わっていません。ゼロ復旧ポイント目標（RPO）を提供するHAソリューションは、インフラストラクチャスタック内のいずれかのコンポーネントの障害によるデータ損失からお客様を保護します。

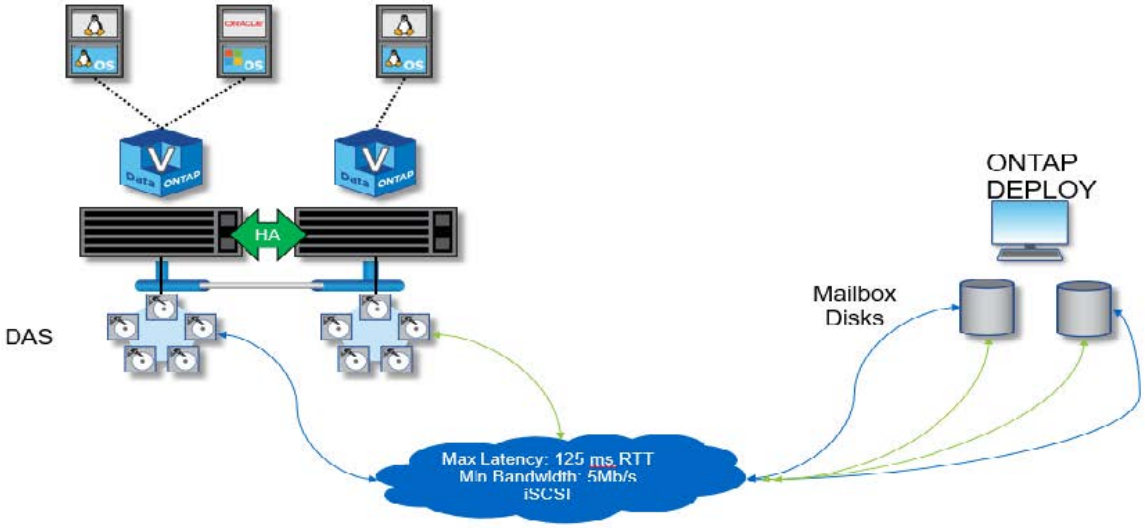
SDS 市場の大部分は、シェアードナッシング ストレージの概念に基づいて構築されており、ソフトウェア レプリケーションによって、異なるストレージ サイロに複数のユーザー データの複数のコピーを保存することでデータの復元力を実現しています。ONTAP Select はこの前提に基づいて構築されており、 ONTAPが提供する同期レプリケーション機能 (RAID SyncMirror) を使用して、クラスタ内にユーザー データの追加コピーを保存します。これは、HA ペアのコンテキスト内で行われます。すべての HA ペアは、ユーザー データの 2 つのコピーを保存します。1 つはローカル ノードによって提供されるストレージに、もう 1 つは HA パートナーによって提供されるストレージにです。ONTAP Selectクラスタ内では、HA と同期レプリケーションが結び付けられており、2 つの機能を切り離したり、独立して使用したりすることはできません。そのため、同期レプリケーション機能はマルチノード オファリングでのみ使用できます。

ONTAP Selectクラスタでは、同期レプリケーション機能はHA実装の機能であり、非同期SnapMirrorまたはSnapVaultレプリケーションエンジンの代替ではありません。同期レプリケーションは、HAとは独立して使用することはできません。

ONTAP Select HA導入モデルには、マルチノードクラスタ（4ノード、6ノード、または8ノード）と2ノードクラスタの2種類があります。2ノードONTAP Selectクラスタの顕著な特徴は、スプリットブレインシナリオを解決するために外部メディエーターサービスを使用することです。ONTAPDeploy VMは、構成ONTAPすべての2ノードHAペアのデフォルトのメディエーターとして機能します。

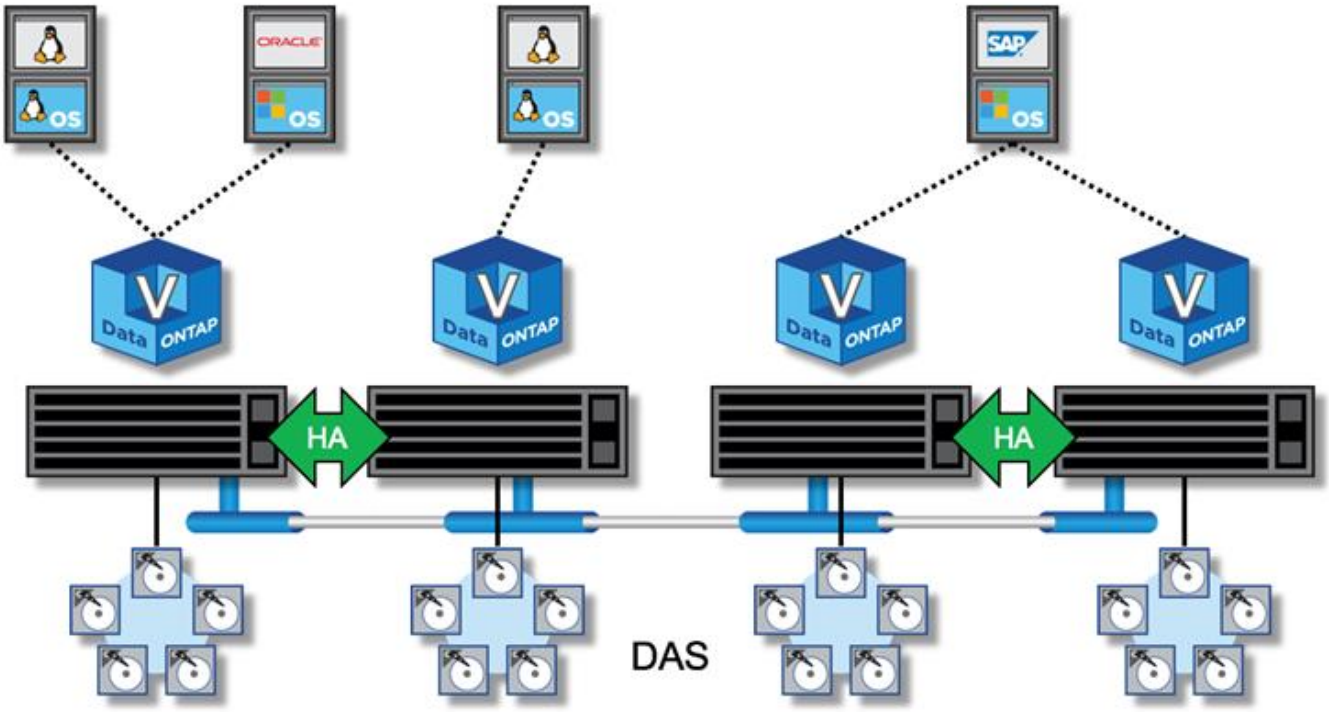
2 つのアーキテクチャは次の図に示されています。

リモートメディエーターとローカル接続ストレージを使用した**2**ノードの**ONTAP Select**クラスタ



2ノードのONTAP Selectクラスタは、1つのHAペアと1つのメディエータで構成されます。HAペア内では、各クラスタノードのデータアグリゲートが同期的にミラーリングされるため、フェイルオーバーが発生してもデータが失われることはありません。

ローカル接続ストレージを使用した**4**ノードの**ONTAP Select**クラスタ



- 4ノードのONTAP Selectクラスタは、2つのHAペアで構成されています。6ノードと8ノードのクラスタは、それぞれ3つと4つのHAペアで構成されています。各HAペア内では、各クラスタノードのデータアグリゲートが同期的にミラーリングされ、フェイルオーバーが発生してもデータが失われることはありません。
- DASストレージを使用する場合、物理サーバ上に存在できるONTAP Selectインスタンスは1つだけです。ONTAPONTAP Selectは、システムのローカルRAIDコントローラへの非共有アクセスを必要とし、ロ

ーカル接続されたディスクを管理するように設計されています。これは、ストレージへの物理的な接続がなければ不可能です。

2ノードHAとマルチノードHA

FASアレイとは異なり、HAペアのONTAP SelectノードはIPネットワーク経由でのみ通信します。つまり、IPネットワークは単一障害点（SPOF）となり、ネットワーク分割やスプリットブレインシナリオに対する保護が設計の重要な側面となります。マルチノードクラスタは、3つ以上の残存ノードによってクラスタクォーラムを確立できるため、単一ノード障害にも耐えることができます。2ノードクラスタでは、 ONTAP Deploy VMがホストするメディエーターサービスを利用して同じ結果を実現します。

ONTAP SelectノードとONTAP Deploy メディエーター サービス間のハートビート ネットワーク トラフィックは最小限で、復元力に優れているため、 ONTAP Deploy VM をONTAP Select の2 ノード クラスタとは別のデータセンターでホストできます。

> ONTAP Deploy VMは、2ノードクラスタのメディエーターとして機能する場合、そのクラスタの不可欠な要素となります。メディエーターサービスが利用できない場合、2ノードクラスタは引き続きデータを提供しますが、 ONTAP Selectクラスタのストレージフェイルオーバー機能は無効になります。そのため、 ONTAP Deployメディエーターサービスは、HAペアの各ONTAP Selectノードとの継続的な通信を維持する必要があります。クラスタクォーラムが適切に機能するには、最小5Mbpsの帯域幅と最大125msのラウンドトリップ時間（RTT）レイテンシが必要です。

メディエータとして機能しているONTAP Deploy VM が一時的または永続的に使用できなくなる可能性がある場合は、セカンダリONTAP Deploy VM を使用して 2 ノード クラスタ クォーラムをリストアできます。これにより、新しいONTAP Deploy VM はONTAP Selectノードを管理できなくなりますが、クラスタ クォーラム アルゴリズムには正常に参加できるようになります。ONTAPONTAP SelectノードとONTAP Deploy VM 間の通信は、IPv4 経由の iSCSI プロトコルを使用して行われます。ONTAPONTAP Selectノードの管理 IP アドレスはイニシエータであり、 ONTAP Deploy VM の IP アドレスはターゲットです。したがって、2 ノード クラスタを作成するときに、ノード管理 IP アドレスとして IPv6 アドレスをサポートすることはできません。ONTAPONTAPがホストするメールボックス ディスクは、2 ノード クラスタの作成時に自動的に作成され、適切なONTAP Selectノードの管理 IP アドレスにマスクされます。設定全体はセットアップ時に自動的に実行されるため、それ以上の管理アクションは必要ありません。クラスタを作成するONTAP Deploy インスタンスは、そのクラスタのデフォルトのメディエータです。

元のメディエーターの場所を変更する必要がある場合は、管理アクションが必要です。元のONTAP Deploy VMが失われた場合でも、クラスタクォーラムをリカバリすることは可能です。NetAppでは、2ノード クラスタがインスタンス化されるたびにONTAP Deployデータベースをバックアップすることを推奨します。

2ノードHAと2ノードストレッチHA（MetroCluster SDS）

2ノードのアクティブ/アクティブHAクラスタをより長距離に拡張し、各ノードを異なるデータセンターに配置することも可能になります。2ノードクラスタと2ノードストレッチクラスタ（ MetroCluster SDSとも呼ばれます）の唯一の違いは、ノード間のネットワーク接続距離です。

2ノードクラスタとは、両ノードが同じデータセンター内に300m以内の距離にあるクラスタとして定義されます。通常、両ノードは同じネットワークスイッチまたはスイッチ間リンク（ISL）ネットワークスイッチセットへのアップリンクを備えています。

2ノードMetroCluster SDSは、ノードが物理的に300m以上（異なる部屋、異なる建物、異なるデータセンターなど）離れたクラスタとして定義されます。さらに、各ノードのアップリンク接続は別々のネットワークスイッチに接続されます。MetroClusterMetroClusterは専用のハードウェアを必要としません。ただし、環境は

レイテンシ（RTTは最大5ms、ジッターは最大5ms、合計10ms）と物理的距離（最大10km）の要件を満たす必要があります。

MetroCluster SDSはプレミアム機能であり、PremiumライセンスまたはPremium XLライセンスが必要です。Premiumライセンスは、小規模および中規模のVM、およびHDDとSSDメディアの作成をサポートします。PremiumXLライセンスは、NVMeドライブの作成もサポートします。

> MetroCluster SDSは、ローカル接続ストレージ（DAS）と共有ストレージ（vNAS）の両方でサポートされています。vNAS構成では、ONTAP Select VMと共有ストレージ間のネットワークが原因で、通常、固有のレイテンシが高くなります。MetroClusterMetroCluster構成では、共有ストレージのレイテンシを含め、ノード間のレイテンシを最大10ミリ秒に抑える必要があります。つまり、これらの構成では共有ストレージのレイテンシが無視できないため、Select VM間のレイテンシのみを測定するだけでは不十分です。

ONTAP Select HA RSMとミラーリングされたアグリゲート

RAID SyncMirror (RSM)、ミラー化されたアグリゲート、および書き込みパスを使用してデータ損失を防ぎます。

同期レプリケーション

ONTAP HAモデルは、HAパートナーの概念に基づいて構築されています。ONTAPONTAP Selectは、 ONTAPに搭載されているRAID SyncMirror （RSM）機能を使用してクラスタノード間でデータブロックを複製することで、このアーキテクチャを非共有型コモディティサーバの世界に拡張し、HAペアに分散されたユーザーデータの2つのコピーを提供します。

メディエータを備えた2ノードクラスタは、2つのデータセンターにまたがって構成できます。詳細については、セクションをご覧ください。 "2ノードストレッチHA（MetroCluster SDS）のベストプラクティス" 。

ミラーリングされたアグリゲート

ONTAP Selectクラスタは2～8ノードで構成されます。各HAペアにはユーザーデータのコピーが2つ含まれ、IPネットワークを介してノード間で同期的にミラーリングされます。このミラーリングはユーザーにとって透過的であり、データアグリゲートのプロパティとして、データアグリゲートの作成プロセス中に自動的に設定されます。

ONTAP Selectクラスタ内のすべてのアグリゲートは、ノードフェイルオーバー時のデータ可用性を確保し、ハードウェア障害発生時のSPOFを回避するためにミラーリングする必要があります。ONTAPONTAP Selectクラスタ内のアグリゲートは、HAペアの各ノードから提供される仮想ディスクから構築され、以下のディスクを使用します。

- ローカルのディスク セット（現在のONTAP Selectノードによって提供される）
- ミラーリングされたディスク セット (現在のノードの HA パートナーによって提供される)

> ミラーリングされたアグリゲートの構築に使用するローカルディスクとミラーディスクは、同じサイズである必要があります。これらのアグリゲートは、それぞれローカルミラーペアとリモートミラーペアを示すプレックス0とプレックス1と呼ばれます。実際のプレックス番号は、インストール環境によって異なる場合があります。

このアプローチは、標準的なONTAPクラスタの動作とは根本的に異なります。これは、 ONTAP Selectクラス

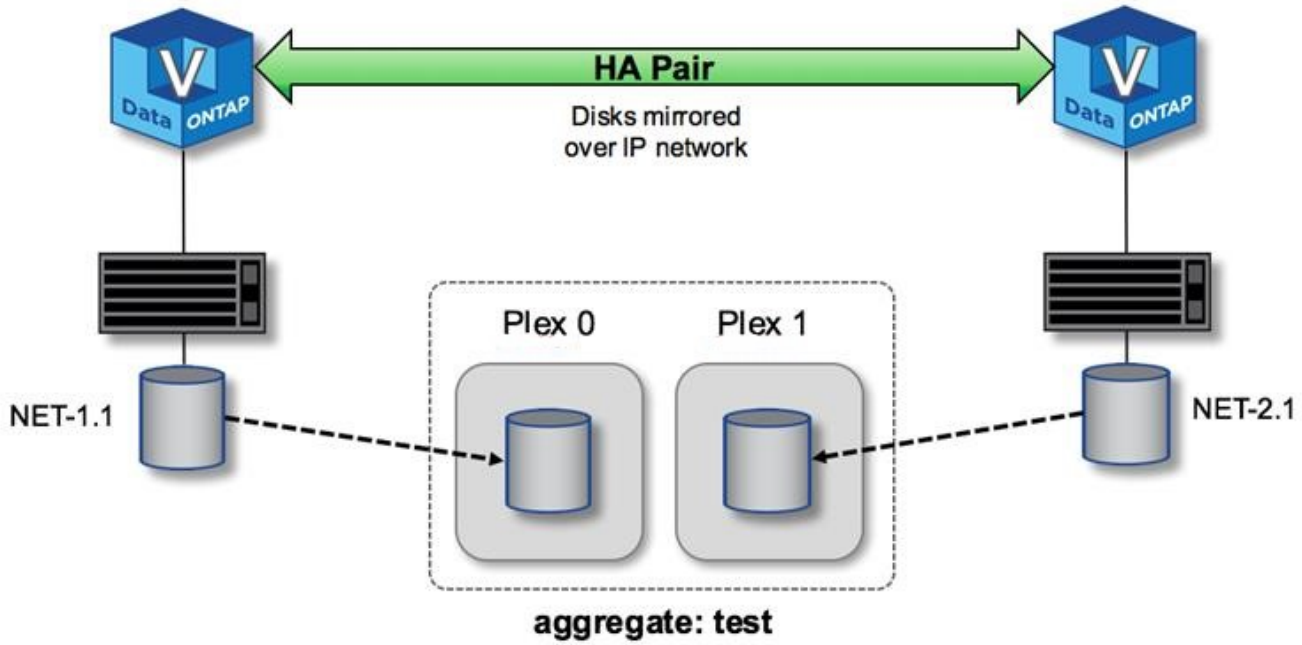
タ内のすべてのルートディスクとデータディスクに適用されます。アグリゲートには、データのローカルコピーとミラーコピーの両方が含まれます。したがって、N個の仮想ディスクを含むアグリゲートは、N/2ディスク分の固有のストレージを提供します。これは、データの2番目のコピーがそれぞれ固有のディスク上に存在するためです。

次の図は、4ノードのONTAP Selectクラスタ内のHAペアを示しています。このクラスタ内には、両方のHAパートナーのストレージを使用する単一のアグリゲート（テスト）があります。このデータアグリゲートは、2つの仮想ディスクセットで構成されています。1つはONTAP Selectを所有するクラスタノード（Plex 0）が提供するローカルセット、もう1つはフェイルオーバーパートナー（Plex 1）が提供するリモートセットです。

Plex 0はすべてのローカルディスクを保持するバケットです。Plex1はミラーディスク、つまりユーザーデータの2番目の複製コピーを格納するディスクを保持するバケットです。アグリゲートを所有するノードはPlex 0にディスクを提供し、そのノードのHAパートナーはPlex 1にディスクを提供します。

次の図には、2つのディスクを持つミラーリングされたアグリゲートがあります。このアグリゲートの内容は2つのクラスタノード間でミラーリングされており、ローカルディスクNET-1.1はPlex 0バケットに、リモートディスクNET-2.1はPlex 1バケットに配置されています。この例では、アグリゲートtestは左側のクラスタノードに所有されており、ローカルディスクNET-1.1とHAパートナーミラーディスクNET-2.1を使用しています。

- ONTAP Selectミラー アグリゲート*



ONTAP Selectクラスタを導入すると、システム上のすべての仮想ディスクが適切なプレックスに自動的に割り当てられるため、ユーザーがディスク割り当てに関して追加の手順を踏む必要はありません。これにより、ディスクが誤ったプレックスに誤って割り当てられるのを防ぎ、最適なミラーディスク構成を実現します。

書き込みパス

クラスタノード間のデータブロックの同期ミラーリングと、システム障害発生時のデータ損失ゼロという要件は、 ONTAP Selectクラスタ内での書き込みパスに大きな影響を与えます。このプロセスは2つの段階から構成されます。

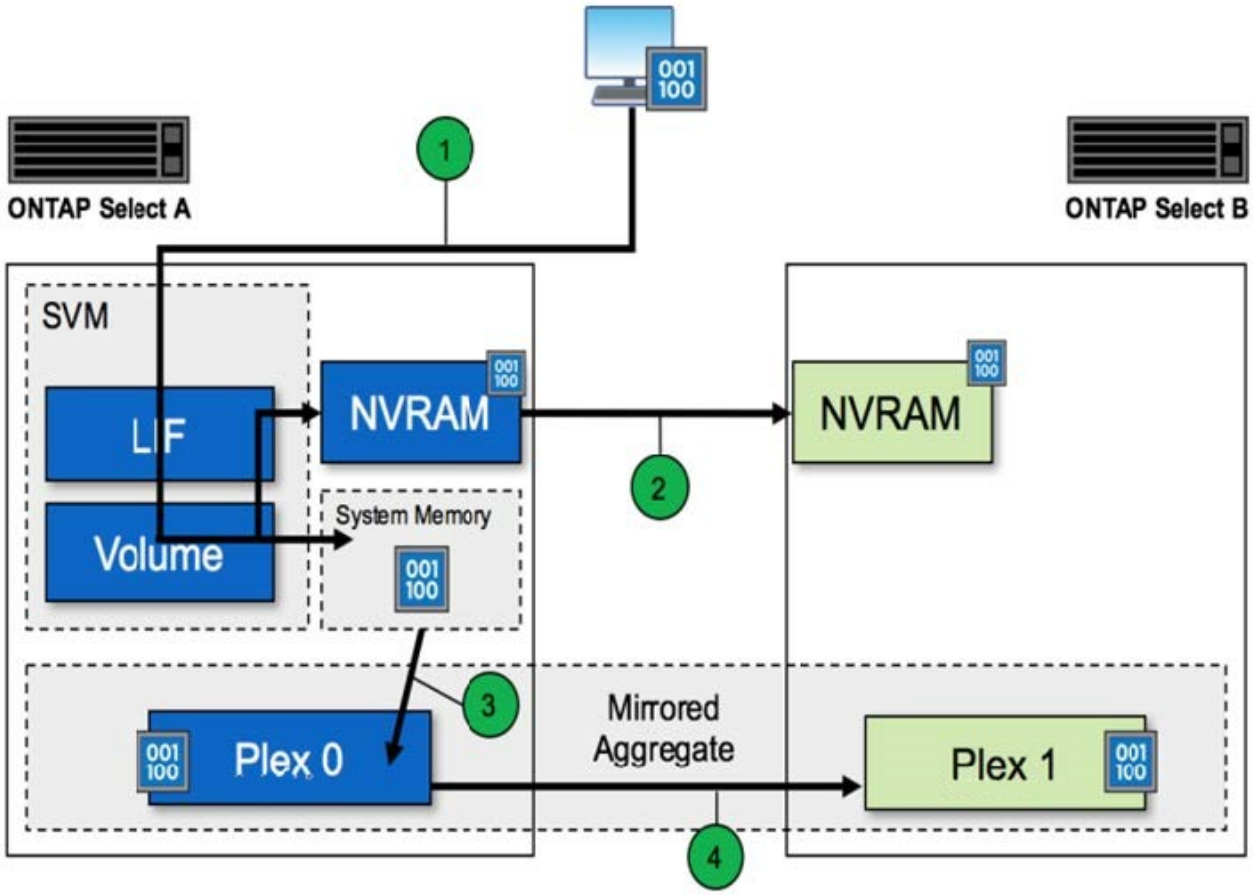
- 了承
- デステージング

ターゲットボリュームへの書き込みはデータLIFを介して行われ、 ONTAP Selectノードのシステムディスク上にある仮想NVRAMパーティションにコミットされた後、クライアントに確認応答が返されます。HA構成では、これらのNVRAMへの書き込みは確認応答される前に、ターゲットボリュームの所有者のHAパートナーに即座にミラーリングされるため、追加のステップが発生します。このプロセスにより、元のノードでハードウェア障害が発生した場合でも、HAパートナーノード上のファイルシステムの整合性が確保されます。

書き込みがNVRAMにコミットされた後、 ONTAPは定期的にこのパーティションの内容を適切な仮想ディスクに移動します。このプロセスはデステージングと呼ばれます。このプロセスは、ターゲットボリュームを所有するクラスタノードで一度だけ実行され、HAパートナーでは実行されません。

次の図は、ONTAP Selectノードへの着信書き込み要求の書き込みパスを示しています。

- ONTAP Select書き込みパスワークフロー*



着信書き込み確認には次の手順が含まれます。

- 書き込みは、ONTAP Selectノード A が所有する論理インターフェイスを介してシステムに入ります。
- 書き込みはノード A のNVRAMにコミットされ、HA パートナーであるノード B にミラーリングされます。
- I/O 要求が両方の HA ノードに存在すると、その要求はクライアントに確認応答されます。

NVRAMからデータ アグリゲート (ONTAP CP) へのONTAP Selectデステージングには、次の手順が含まれま

す。

- 書き込みは仮想NVRAMから仮想データ アグリゲートにデステージされます。
- ミラー エンジンは、ブロックを両方のプレックスに同期的に複製します。

ONTAP Select HAはデータ保護を強化します

高可用性 (HA) ディスク ハートビート、HA メールボックス、HA ハートビート、HA フェイルオーバー、およびギブバックは、データ保護を強化するために機能します。

ディスクハートビート

ONTAP Select HAアーキテクチャは、従来のFASアレイで使用されているコードパスの多くを活用していますが、いくつか例外があります。その一つが、ディスクベースのハートビートの実装です。これは、クラスタノードがネットワークの分離によるスプリットブレインの発生を防ぐために使用する、ネットワークベースではない通信方法です。スプリットブレインとは、通常はネットワーク障害によって引き起こされるクラスタのパーティショニングによって発生する現象で、各ノードが他方のノードがダウンしていると認識し、クラスタリソースの乗っ取りを試みます。

エンタープライズクラスのHA実装では、このようなシナリオを適切に処理する必要があります。ONTAPは、カスタマイズされたディスクベースのハートビート方式によってこれを実現します。これは、クラスタノードがハートビートメッセージを渡すために使用する物理ストレージ上の場所であるHAメールボックスの役割です。これにより、クラスタは接続状態を判断し、フェイルオーバー発生時にクォーラムを定義できます。

共有ストレージ HA アーキテクチャを使用するFASアレイでは、 ONTAP は次の方法でスプリット ブレインの問題を解決します。

- SCSIの永続的予約
- 永続的なHAメタデータ
- HA 状態は HA インターコネクト経由で送信されます

ただし、 ONTAP Selectクラスタのシェアードナッシングアーキテクチャでは、ノードは自身のローカルストレージのみを参照でき、HAパートナーのローカルストレージは参照できません。そのため、ネットワークパーティショニングによってHAペアの両側が分離されている場合、クラスタクォーラムとフェイルオーバーの動作を決定する前述の方法は利用できません。

スプリットブレイン検出および回避の既存の方法は使用できませんが、シェアードナッシング環境の制約に適合するメディエーションの方法が依然として必要です。ONTAPONTAP Selectは既存のメールボックスインフラストラクチャをさらに拡張し、ネットワークパーティショニング発生時のメディエーション手段として機能できるようにします。共有ストレージが利用できないため、メディエーションはNAS経由のメールボックスディスクへのアクセスを通じて実行されます。これらのディスクは、iSCSIプロトコルを使用して、2ノードクラスタ内のメディエータを含むクラスタ全体に分散されています。したがって、クラスタノードはこれらのディスクへのアクセスに基づいて、インテリジェントなフェイルオーバーの決定を行うことができます。ノードがHAパートナー以外の他のノードのメールボックスディスクにアクセスできる場合、そのノードは正常に稼働していると考えられます。

> メールボックス アーキテクチャと、クラスタ クォーラムおよびスプリット ブレインの問題を解決するためのディスクベースのハートビート方式のため、 ONTAP Selectのマルチノード バリアントでは、2 ノード クラスタに対して 4 つの個別のノードまたはメディエーターが必要になります。

HAメールボックス投稿

HAメールボックスアーキテクチャは、メッセージポストモデルを採用しています。クラスタノードは、一定の間隔で、クラスタ全体の他のすべてのメールボックスディスク（メディエータを含む）に、ノードが稼働中であることを示すメッセージをポストします。正常なクラスタ内では、どの時点でも、クラスタノード上の単一のメールボックスディスクに、他のすべてのクラスタノードからポストされたメッセージが保持されます。

各 Select クラスター ノードには、共有メールボックス アクセス専用の仮想ディスクが接続されています。このディスクは、ノード障害またはネットワーク パーティション分割の発生時にクラスター仲介の手段となることが主な機能であるため、メディエーター メールボックス ディスクと呼ばれます。このメールボックス ディスクには各クラスター ノードのパーティションが含まれており、他の Select クラスター ノードによって iSCSI ネットワーク経由でマウントされます。これらのノードは、メールボックス ディスクの適切なパーティションに定期的にヘルス ステータスを送信します。クラスター全体に広がるネットワーク アクセス可能なメールボックス ディスクを使用すると、到達可能性マトリックスを通じてノードのヘルスを推測できます。たとえば、クラスター ノード A および B はクラスター ノード D のメールボックスには送信できますが、ノード C のメールボックスには送信できません。また、クラスター ノード D はノード C のメールボックスには送信できないため、ノード C がダウンしているかネットワークから分離されており、テイクオーバーする必要がある可能性が高くなります。

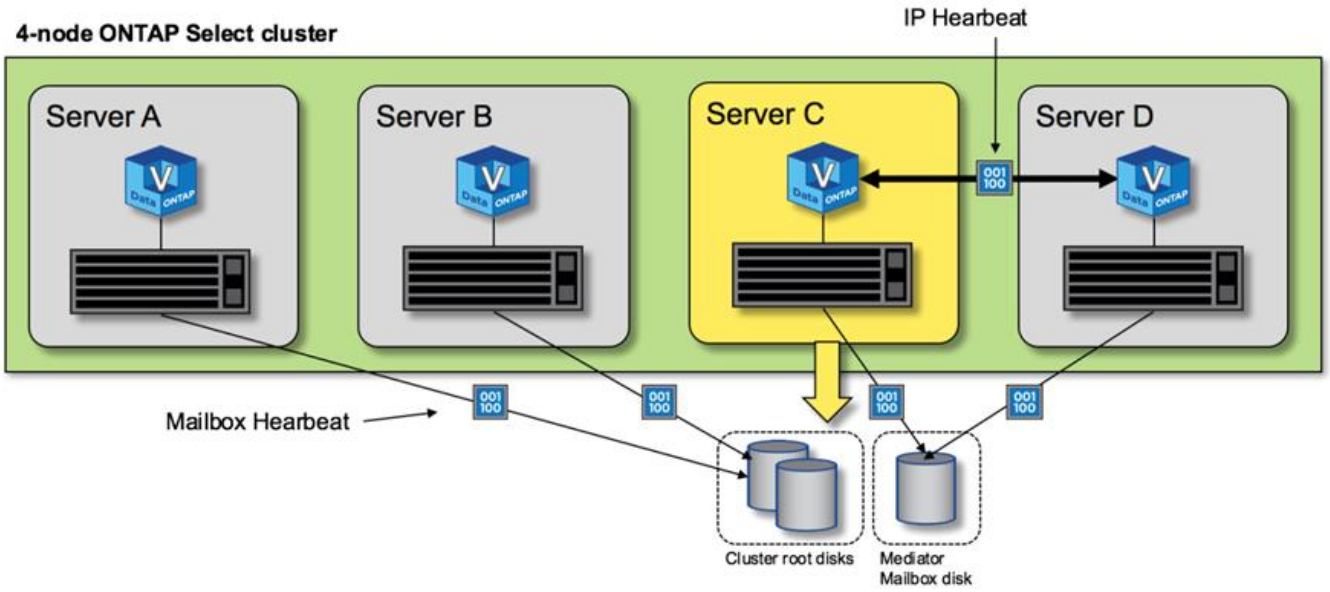
HAの鼓動

NetAppのFASプラットフォームと同様に、ONTAP SelectはHAインターコネクトを介して定期的にHAハートビート メッセージを送信します。ONTAP Selectクラスタ内では、これはHAパートナー間に存在するTCP/IPネットワーク接続を介して実行されます。さらに、ディスクベースのハートビートメッセージが、メディエータメールボックスディスクを含むすべてのHAメールボックスディスクに渡されます。これらのメッセージは数秒ごとに渡され、定期的に読み戻されます。これらの送受信頻度により、 ONTAP Selectクラスタは約15秒以内にHA障害イベントを検出できます。これはFASプラットフォームで利用可能な時間枠と同じです。ハートビートメッセージが読み取られなくなると、フェイルオーバーイベントがトリガーされます。

次の図は、単一のONTAP Selectクラスタ ノード（ノード C）の観点から、HA インターコネクトおよびメディエーター ディスクを介してハートビート メッセージを送受信するプロセスを示しています。

ネットワーク ハートビートは HA インターコネクトを介して HA パートナーであるノード D に送信され、ディスク ハートビートはすべてのクラスター ノード A、B、C、D のメールボックス ディスクを使用します。

4ノードクラスタにおける**HA**ハートビート：定常状態



HAのフェイルオーバーとギブバック

フェイルオーバー処理中、残存ノードはHAパートナーのデータのローカルコピーを使用して、ピアノードへのデータ提供を引き継ぎます。クライアントI/Oは中断することなく継続されますが、ギブバックを実行する前に、このデータへの変更をレプリケートする必要があります。ONTAPONTAP Selectは強制ギブバックをサポートしていません。強制ギブバックを実行すると、残存ノードに保存されている変更が失われるためです。

再起動されたノードがクラスタに再参加すると、同期戻し操作が自動的に開始されます。同期戻しに必要な時間は、レプリケートする必要がある変更の数、ノード間のネットワーク遅延、各ノードのディスクサブシステムの速度など、いくつかの要因によって同期戻しに必要な時間が、自動ギブバックウィンドウの10分を超える可能性があります。この場合、同期戻し後に手動でギブバックを行う必要があります。同期戻しの進行状況は、次のコマンドで監視できます。

```
storage aggregate status -r -aggregate <aggregate name>
```