高可用性アーキテクチャ
ONTAP Select

NetApp
April 12, 2024

目次

高可用性アーキテクチャ

ハイアベイラビリティ構成

高可用性オプションを確認して、環境に最も適した HA 構成を選択してください。

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で動作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、耐障害性とフォールトトレランスに対するニーズや期待は変わりません。Recovery Point Objective （ RPO ；目標復旧時点）がゼロの HA 解決策 は、インフラスタック内のコンポーネント障害によるデータ損失からお客様を保護します。

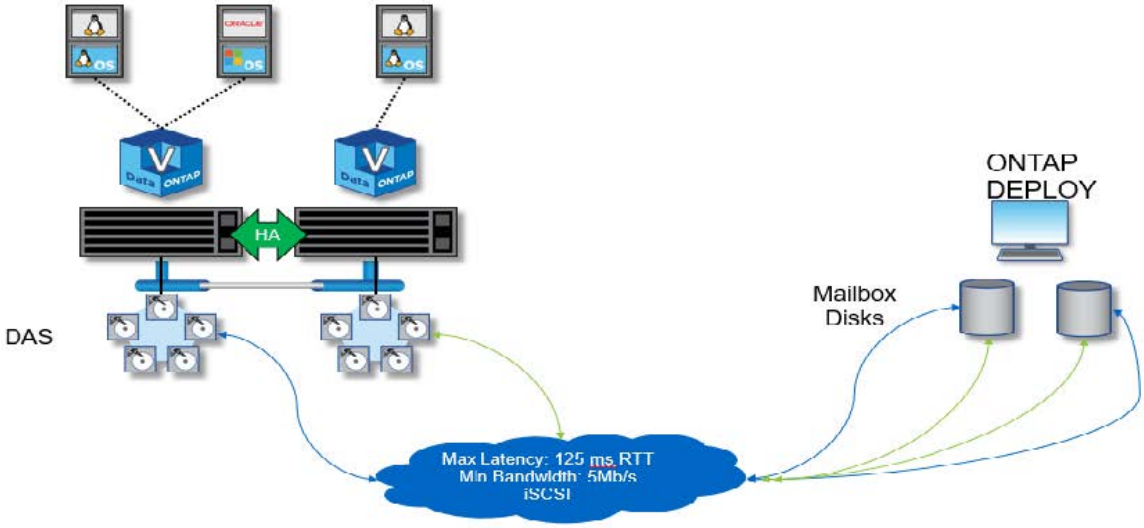
SDS 市場の大部分は、シェアードナッシングストレージという概念の上に成り立っています。これは、ソフトウェアレプリケーションでユーザデータの複数のコピーを複数のストレージサイロにまたがって格納することで、データの耐障害性を実現するというものです。ONTAP Select は、この前提の基に、 ONTAP 付属の同期レプリケーション機能（ RAID SyncMirror ）を使用してクラスタ内にユーザデータのコピーを追加で保存します。これは HA ペアのコンテキスト内で実行されます。HA ペアは、ユーザデータのコピーをローカルノードのストレージに 1 つ、 HA パートナーのストレージに 1 つ、合わせて 2 つ格納します。ONTAP Select クラスタ内では、 HA と同期レプリケーションが統合されており、 2 つの機能を切り離したり個別に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ソリューション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 の機能であり、非同期の SnapMirror または SnapVault レプリケーションエンジンに代わるものではありません。同期レプリケーションを HA から切り離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 、 6 、または 8 ノード）と 2 ノードクラスタの 2 つがあります。2 ノード ONTAP Select クラスタの注目すべき特徴は、スプリットブレインの状況を解決するために外部のメディエーターサービスを使用する点です。ONTAP Deploy VM は、設定するすべての 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。

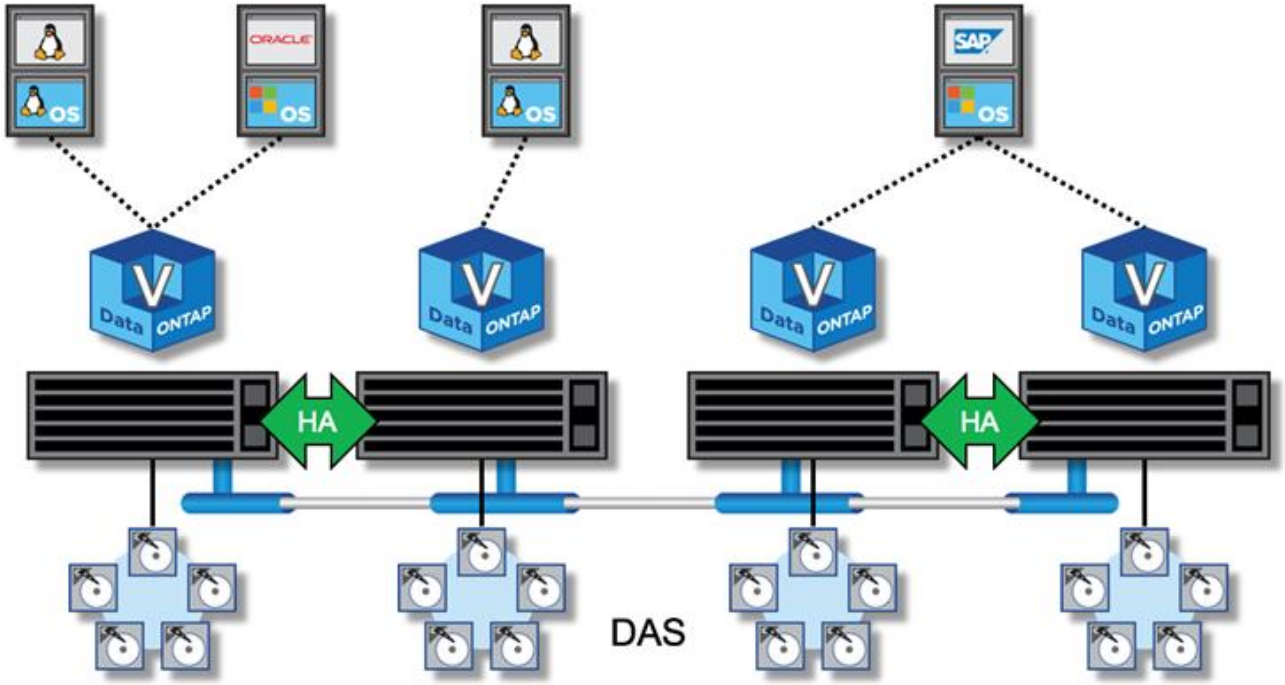
この 2 つのアーキテクチャを次の図に示します。

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



2 ノード ONTAP Select クラスタは、 1 つの HA ペアとメディエーターで構成されます。各 HA ペアでは、各クラスタノード上のデータアグリゲートが同期ミラーリングされ、フェイルオーバーが発生した場合に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する 4 ノード ONTAP Select クラスタ *



- 4 ノード ONTAP Select クラスタは、 2 つの HA ペアで構成されます。6 ノードクラスタと 8 ノードクラスタは、それぞれ 3 つと 4 つの HA ペアで構成されます。各 HA ペアでは、各クラスタノード上のデータアグリゲートが同期ミラーリングされ、フェイルオーバーが発生した場合にデータが失われることはありません。
- DAS ストレージを使用している場合、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。ONTAP Select は、システムのローカル RAID コントローラに排他的にアクセスする必要があり、かつローカル接続ディスクを管理するように設計されているため、ストレージとの物理的な接続が不

可欠です。

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、専用の IP ネットワーク経由で通信します。IP ネットワークが Single Point of Failure （ SPOF ；単一点障害）となるため、ネットワークパーティションやスプリットブレインの状況から保護することが、設計上の重要な要素となります。マルチノードクラスタでは、 1 つのノードで障害が発生しても残りの 3 つ以上のノードでクラスタクォーラムを確立可能なため、運用を継続できます。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使用して同様の保護が実装されます。

ONTAP Select ノードと ONTAP Deploy メディエーターサービスの間のハートビートネットワークトラフィックは最小限かつ耐障害性があるため、 ONTAP Deploy VM を 2 ノード ONTAP Select クラスタとは別のデータセンターでホストすることが可能です。

> 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、 ONTAP Deploy VM はそのクラスタに不可欠な要素です。メディエーターサービスを使用できない場合、 2 ノードクラスタはデータの提供を続けますが、 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。このため、 ONTAP Deploy のメディエーターサービスは、 HA ペアの各 ONTAP Select ノードとの安定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を適切に機能させるには、最小帯域幅が 5Mbps 、 RTT （最大ラウンドトリップ時間）が 125 ミリ秒でなければなりません。

メディエーターとして機能する ONTAP Deploy VM が一時的に使用できない場合、または永久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。その結果、新しい ONTAP Deploy VM は ONTAP Select ノードを管理できませんが、クラスタクォーラムのアルゴリズムには問題なく追加されます。ONTAP Select ノードと ONTAP Deploy VM の間の通信には、 IPv4 経由の iSCSI プロトコルが使用されます。ONTAP Select ノードの管理 IP アドレスがイニシエータで、 ONTAP Deploy VM の IP アドレスがターゲットです。したがって、 2 ノードクラスタを作成する場合、ノード管理 IP アドレスの IPv6 アドレスはサポートできません。2 ノードクラスタの作成時に、 ONTAP Deploy でホストされるメールボックスディスクが自動的に作成され、適切な ONTAP Select ノード管理 IP アドレスにマスクされます。設定はすべてセットアップ時に自動的に行われ、管理操作は不要です。クラスタを作成する ONTAP Deploy インスタンスが、そのクラスタのデフォルトのメディエーターとなります。

メディエーターの元の場所を変更する必要がある場合は、管理操作が必要です。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が、ネットアップでは、 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 ）の比較

2 ノードのアクティブ / アクティブ HA クラスタをより長い距離に伸長し、各ノードを異なるデータセンターに配置することが可能です。2 ノードクラスタと 2 ノードストレッチクラスタ（別名 MetroCluster SDS ）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、同じデータセンター内にある 2 つのノードが 300m 以内の範囲に配置されているクラスタです。一般に、両方のノードには、同じネットワークスイッチまたは一連の Interswitch Link （ ISL ；スイッチ間リンク）ネットワークスイッチへのアップリンクがあります。

2 ノード MetroCluster SDS とは、別の部屋、別の建物、別のデータセンターなど、物理的に 300m 以上離れたノードを持つクラスタです。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。MetroCluster SDS には専用ハードウェアは必要ありません。ただし、環境はレイテンシの要件（ RTT の最大 5 ミリ秒とジッタの 5 ミリ秒の合計 10 ミリ秒）と物理距離の要件（最大 10km ）に準拠している必要があります。

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

> 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ます。通常、 vNAS 構成では、 ONTAP Select VM と共有ストレージとの間のネットワークが原因で、本来のレイテンシが大きくなります。MetroCluster SDS 構成では、共有ストレージのレイテンシを含め、ノード間で最大 10 ミリ秒のレイテンシを実現する必要があります。つまり、これらの構成では共有ストレージのレイテンシを無視できないため、Select VM 間のレイテンシを測定するだけでは不十分です。

HA RSM とミラーされたアグリゲート

RAID SyncMirror （ RSM ）、ミラー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止します。

同期レプリケーション

ONTAP の HA モデルは、 HA パートナーの概念に基づいています。ONTAP Select は、 ONTAP の 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レプリケートして、 HA ペア内でユーザデータのコピーが 2 つ保持されるようにすることで、このアーキテクチャを非共有型コモディティサーバへと拡張します。

メディエーターを持つ 2 ノードクラスタは、 2 つのデータセンターにまたがることができます。詳細については、を参照してください "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に対して透過的であり、データアグリゲートのプロパティであり、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、ノードのフェイルオーバー時にデータ可用性を実現し、ハードウェア障害時の SPOF を回避するため、ミラーリングする必要があります。ONTAP Select クラスタ内のアグリゲートは、 HA ペアの各ノードが提供する仮想ディスクから作成され、次のディスクを使用します。

- 現在の ONTAP Select ノードが提供する 1 セットのローカルディスク
- 現在のノードの HA パートナーが提供する 1 セットのミラーディスク

> ミラーアグリゲートの作成に使用され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、プレックス 0 およびプレックス 1 と呼ばれます（ローカルミラーペアとリモートミラーペアを示す）。実際のプレックス番号は、環境によって異なる場合があります。

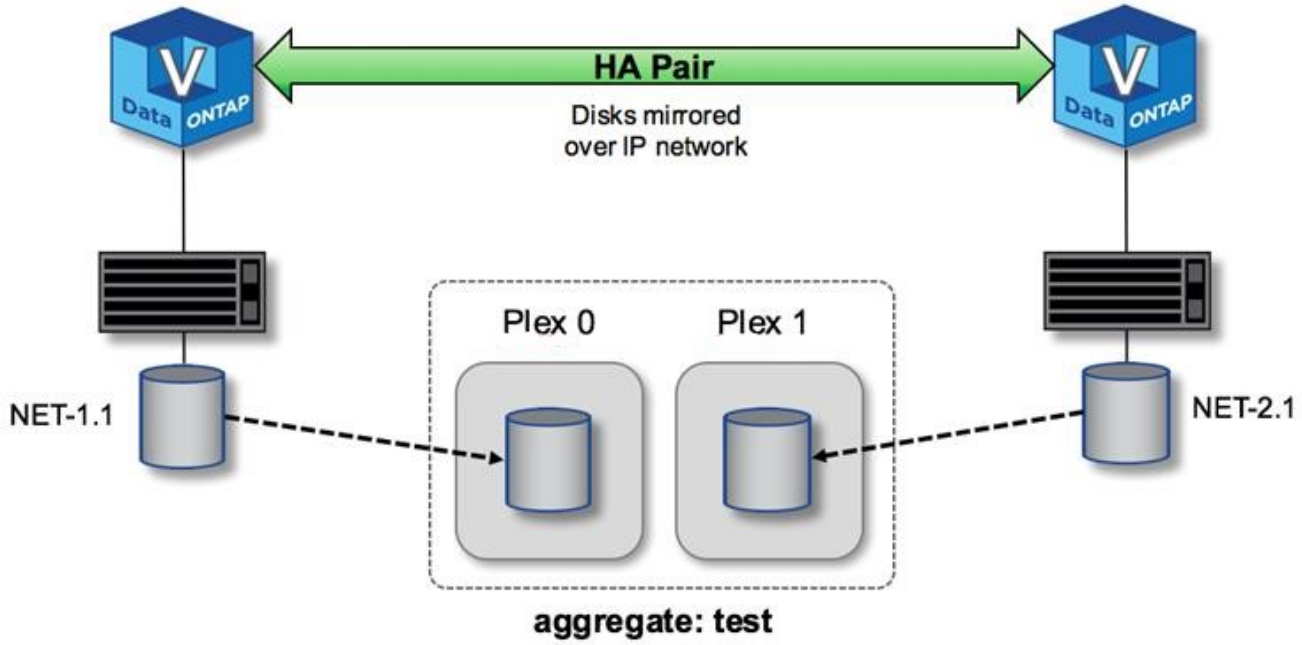
このアプローチは、標準的な ONTAP クラスタの動作とは根本的に異なります。この環境 は、 ONTAP Select クラスタ内のすべてのルートディスクとデータディスクを指します。アグリゲートには、データのローカルコピーとミラーコピーの両方が含まれます。したがって、 N 個の仮想ディスクを含むアグリゲートは、データの 2 番目のコピーが固有のディスクに保存されるため、 N /2 個分の一意のストレージを提供します。

次の図は、 4 ノード ONTAP Select クラスタ内の HA ペアを示しています。このクラスタには、両方の HA パートナーのストレージを使用する（テスト用の）アグリゲートが 1 つあります。このデータアグリゲートは、 2 セットの仮想ディスクで構成されます。 1 つは ONTAP Select が所有するクラスタノードが提供するローカルセット（プレックス 0 ）、もう 1 つはフェイルオーバーパートナーが提供するリモートセット（プレックス 1 ）です。

プレックス 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。プレックス 1 は、ミラーディスク、つまりユーザデータの 2 つ目のレプリケート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プレックス 0 にディスクを提供し、そのノードの HA パートナーはプレックス 1 にディスクを提供します。

次の図では、 2 本のディスクで構成されるミラーアグリゲートがあります。このアグリゲートの内容が 2 つのクラスタノード間でミラーされます。この場合、ローカルディスク NET-1.1 はプレックス 0 バケットに配置され、リモートディスク NET-2.1 はプレックス 1 バケットに配置されます。この例では、アグリゲート test は左側のクラスタノードによって所有され、ローカルディスク NET-1.1 と HA パートナーのミラーディスク NET-2.1 を使用します。

- ONTAP Select ミラーアグリゲート *



ONTAP Select クラスタの導入時にシステム上のすべての仮想ディスクが自動的に正しいプレックスに割り当てられるため、ディスクの割り当てに関してユーザが追加の作業を行う必要ありません。そのため、ディスクが間違ったプレックスに誤って割り当てられることがなく、最適なミラーディスク構成が確実に作成されます。

書き込みパス

クラスタノード間でデータブロックが同期ミラーリングされること、またシステム障害発生時にデータ損失ゼロが求められることは、受信した書き込みが ONTAP Select クラスタ全体に伝播される際のパスに大きく影響します。このプロセスは、次の 2 つの段階で構成されます

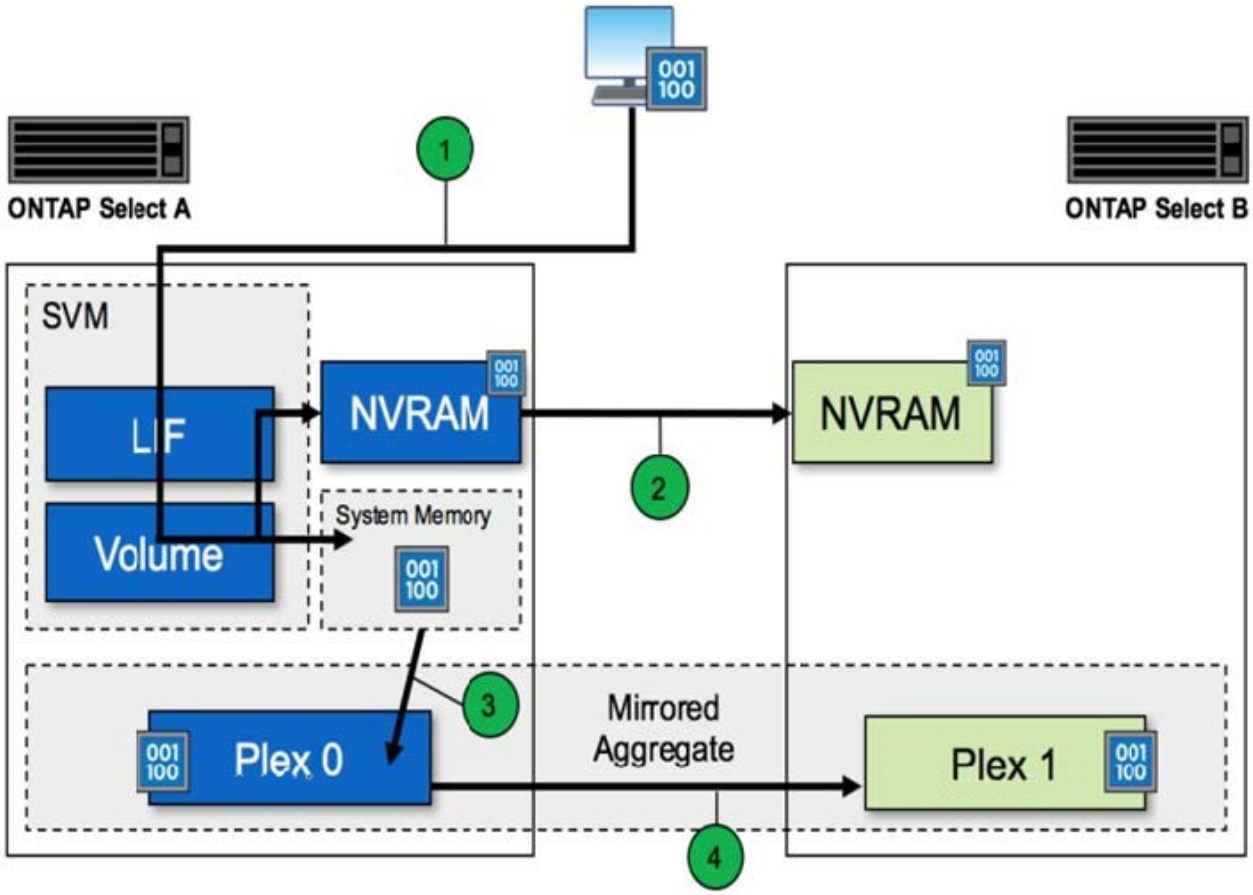
- 確認応答
- デステージ

ターゲットボリュームへの書き込みはデータ LIF 経由で行われ、 ONTAP Select ノードのシステムディスク上の仮想 NVRAM パーティションにコミットされてから、クライアントに確認応答が返されます。HA 構成の場合は、確認応答の前に、 NVRAM への書き込みがターゲットボリュームを所有するノードの HA パートナーにただちにミラーされます。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 パートナーノードで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 にコミットされると、 ONTAP はこのパーティションの内容を適切な仮想ディスクに定期的に移動します。これがデステージと呼ばれるプロセスです。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 1 回だけ実行され、 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 書き込みパスのワークフロー *



書き込みの確認応答は、次の手順で行われます。

- ONTAP Select ノード A が所有する論理インターフェイス経由で書き込みがシステムに送信されます
- 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 にミラーされます
- I/O 要求が両方の HA ノードに到達した時点で、クライアントに要求の確認応答が返されます。

NVRAM からデータアグリゲートへの ONTAP Select のデステージ（ ONTAP CP ）は、次の手順で行われます。

- 仮想 NVRAM から仮想データアグリゲートに書き込みがデステージされます。
- ミラーエンジンが両方のプレックスにブロックを同期的にレプリケートします。

HA の詳細

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、ギブバックの機能を使用してデータ保護を強化します。

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たコードパスの多くを利用しますが、例外もあります。その 1 つが、ディスクベースのハートビートで採用されている非ネットワークベースの通信方法で、クラスタノードはこの通信方法を使用してネットワーク分離によって発生するスプリットブレインを回避します。スプリットブレインのシナリオはクラスタパーティショニングの結果であり、一般にネットワーク障害が原因で発生します。スプリットブレインが発生すると、それぞれのサイドが相手が停止したと判断してクラスタリソースをテイクオーバーしようとします。

エンタープライズクラスの HA 実装では、このタイプのシナリオを適切に処理する必要があります。ONTAP では、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を使用してこの処理を実行します。この処理に使用されるのが HA メールボックスで、物理ストレージ上にあって、クラスタノード間でのハートビートメッセージのやり取りに使用されます。これにより、クラスタはフェイルオーバー時に接続の有無を判断し、クォーラムを定義できます。

共有ストレージの 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 インターコネクト経由で送信された HA 状態です

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、ノードが自身のローカルストレージしか認識できず、 HA パートナーのローカルストレージは認識できません。このため、ネットワークパーティショニングによって HA ペアの両サイドが分離されると、前出の方法ではクラスタクォーラムとフェイルオーバー動作を判断できなくなります。

既存の方法でスプリットブレインの検出と回避を行うことはできませんが、シェアードナッシング環境の制約の範囲内で使用できるメディエーションの手段は依然として必要です。ONTAP Select を使用すると、既存のメールボックスインフラを拡張して、ネットワークパーティショニングが発生した場合にメディエーションの手段として機能させることができます。共有ストレージを使用できないため、メディエーションは、 NAS 経由でメールボックスディスクにアクセスすることで実施されます。これらのディスクは、 iSCSI プロトコルを使用して、 2 ノードクラスタのメディエーターを含むクラスタ全体に分散されます。そのため、これらのディスクへのアクセスに基づいて、クラスタノードからインテリジェントなフェイルオーバーの決定を下すことができます。ノードがその HA パートナーの外部にある他のノードのメールボックスディスクにアクセスできれば、そのノードは正常に稼働していると考えられます。

> このように、クラスタクォーラムとスプリットブレインの問題を解決するためにメールボックスアーキテクチャとディスクベースのハートビートを使用することが、マルチノードの ONTAP Select では 4 つの独立したノード、またはメディエーターを使用する 2 ノードクラスタのいずれかが必要とされる理由です。

HA メールボックスへの投稿

HA メールボックスアーキテクチャでは、メッセージの投かんモデルが使用されます。クラスタノードは、メ

ディエーターを含めてクラスタ内の他のすべてのメールボックスディスクにメッセージを繰り返し送信し、ノードが稼働していることを通知します。正常なクラスタでは、どの時点においても、あるクラスタノードの 1 つ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投かんされたメッセージが存在します。

各 Select クラスタノードには、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ニングが発生した場合にクラスタメディエーションの手段として機能するため、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ノード用のパーティションが含まれ、他の Select クラスタノードから iSCSI ネットワークを介してマウントされます。各 Select クラスタノードは、メールボックスディスクの該当するパーティションに定期的に健常性ステータスを投かんします。ネットワークにアクセス可能なメールボックスディスクをクラスタ全体に分散させることで、到達可能かどうかという観点からノードの健常性を推測できます。たとえば、クラスタノード A と B は、クラスタノード D のメールボックスには投かんできますが、ノード C のメールボックスにはポストできませんまた、クラスタノード D がノード C のメールボックスに投かんできないため、ノード C は停止しているかネットワークから分離されている可能性があり、テイクオーバーの必要があります。

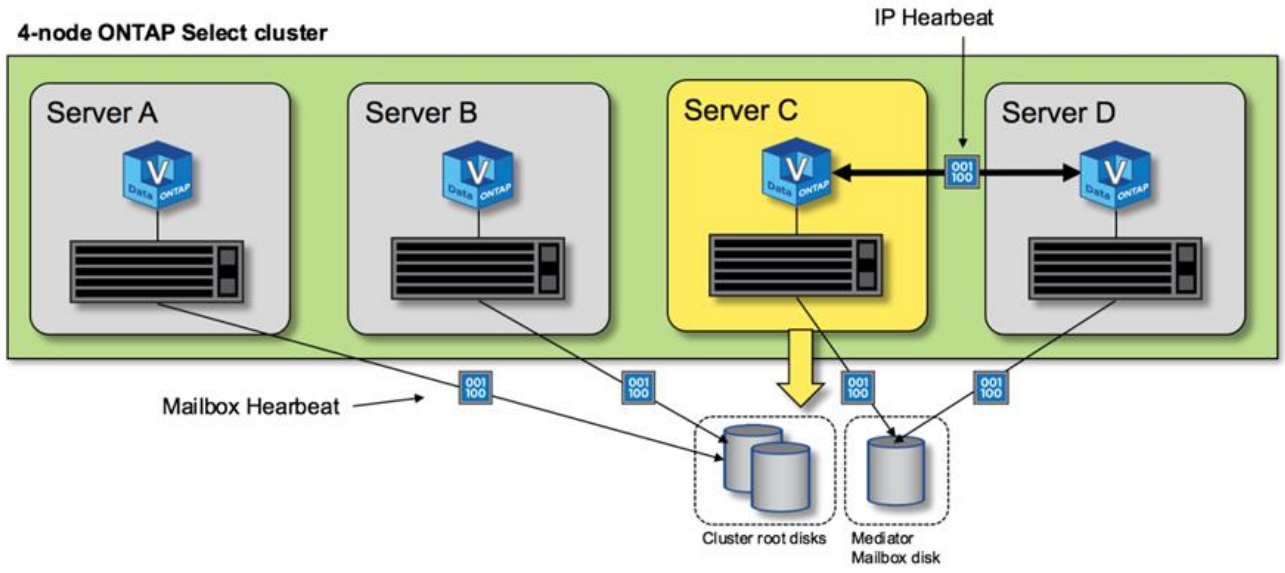
HA ハートビート

ネットアップの 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、この処理は HA パートナー間の TCP / IP ネットワークを介して行われます。また、ディスクベースのハートビートメッセージは、メディエーターの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送信されます。これらのメッセージは数秒ごとに送信され、定期的に読み取られます。メッセージが頻繁に送受信されることで、 ONTAP Select クラスタは HA 障害イベントを FAS プラットフォームと同じくほぼ 15 秒以内に検出できます。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 クラスタノードであるノード C から見た、 HA インターコネクトディスクとメディエーターディスクを介したハートビートメッセージの送受信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、一方、ディスクハートビートはクラスタノード A 、 B 、 C 、 D のすべてに存在するメールボックスディスクを使用して送信されます

- 4 ノードクラスタでの HA ハートビート：安定状態 *



HA フェイルオーバーおよびギブバック

フェイルオーバー処理中、稼働しているノードは、 HA パートナーのデータのローカルコピーを使用して、ピアノードのデータ提供を担当します。クライアント I/O は中断なく継続しますが、ギブバックが発生する前に、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。ONTAP Select では強制ギブバックはサポートされません。強制ギブバックにより、障害を免れたノードに保存されている変更が失われるためです。

リブートされたノードがクラスタに再び参加すると、再同期処理が自動的にトリガーされます。再同期に必要な時間は、いくつかの要因によって異なります。たとえば、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レイテンシ、各ノードのディスクサブシステムの速度などです。再同期に必要な時間が、自動ギブバック時間の 10 分を超える可能性があります。この場合、再同期後の手動ギブバックが必要です。再同期の進捗状況は、次のコマンドを使用して監視できます。

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

著作権に関する情報

商標に関する情報