Fsas Technologies

C140-0068-02Z3

ETERNUS AX series オールフラッシュアレイ

All SAN Array (ASA) のデータ可用性と整合性

目次

図目次

表目次

はじめに

本書は、ETERNUS AX series All SAN Array (ASA) システムのさまざまなデータ保護機能およびデータ整合性機能について説明します。また、最大限の信頼性を実現する SAN ネットワークの設計、実装、および管理に関するベストプラクティスについても説明します。

第 2 版
2025 年 3 月

登録商標

本製品に関連する他社商標については、以下のサイトを参照してください。
https://www.fujitsu.com/jp/products/computing/storage/trademark/
本書では、本文中の™、® などの記号は省略しています。

本書の読み方

対象読者

本書は、ETERNUS AX の設定、運用管理を行うシステム管理者、または保守を行うフィールドエンジニアを対象としています。必要に応じてお読みください。

関連マニュアル

ETERNUS AX に関連する最新の情報は、以下のサイトで公開されています。
https://www.fujitsu.com/jp/products/computing/storage/manual/

本書の表記について

本文中の記号

本文中では、以下の記号を使用しています。

注意	お使いになるときに注意していただきたいことを記述しています。必ずお読みください。
備考	本文を補足する内容や、参考情報を記述しています。

1. ONTAP について

ONTAP は、インライン圧縮、無停止のハードウェアアップグレード、外部ストレージシステムから LUN をインポートする機能などをデフォルトで備えた、強力なデータ管理プラットフォームです。最大 12 のノードをクラスタ化して、iSCSI、ファイバチャネル (FC) 、および Nonvolatile Memory Express (NVMe) プロトコル経由で SAN にデータを同時に提供できます。さらに、スナップショットテクノロジーは ONTAP に不可欠な要素であり、重要なデータセットの数万件に及ぶバックアップの作成と、ほぼ即時のデータセットのクローン作成を可能にします。また、包括的なディザスタリカバリ機能も提供します。

2. All SAN Array

All SAN Array (ASA) システムは、ONTAP を実行するオールフラッシュシステム上に構築され、複数のワークロード用ストレージリソースを統合および共有したいお客様にエンタープライズクラスの SAN ソリューションを提供します。

ASA システムは、以下の機能を提供します。

- 99.9999% を超える、業界トップクラスの可用性
- スケールアップとスケールアウトの両方が可能な大規模クラスタ
- 業界トップクラスの企業向け性能
- 業界トップクラスのストレージ効率
- クラウドに対応した、最も完成度の高い接続性
- コストパフォーマンスに優れたシームレスなデータ保護

ASA は、オールフラッシュシステムのプラットフォーム上に構築され、SAN の継続的な可用性を実現します。ASA は、ストレージの計画的フェイルオーバーまたは計画外フェイルオーバーの際にデータへのアクセスを中断することなく提供し、SAN ワークロードの実行のみに特化したソリューションを通じて、実装、構成、管理を合理化します。以下の要件が含まれる場合、ASA の設定を推奨します。

- ホストからストレージへの対称アクティブ / アクティブパスを必要とするデータベースなどの、ミッションクリティカルなワークロード
- SAN ワークロードを分離するための専用システムを優先する

ONTAP のオールフラッシュシステムは、以下のお客様にも最適です。

- SAN クラスタを最大 12 ノードまで拡張する必要がある。
- アクティブ / アクティブ SAN パス管理に関して特に要件がない。
- NAS と SAN の混在したワークロードをサポートするために、統合プロトコルをサポートするクラスタを希望している。

3. ASA のアーキテクチャ：データの可用性と整合性

ストレージシステムには、データの保護とデータの可用性の確保という 2 つの基本的な要件があります。

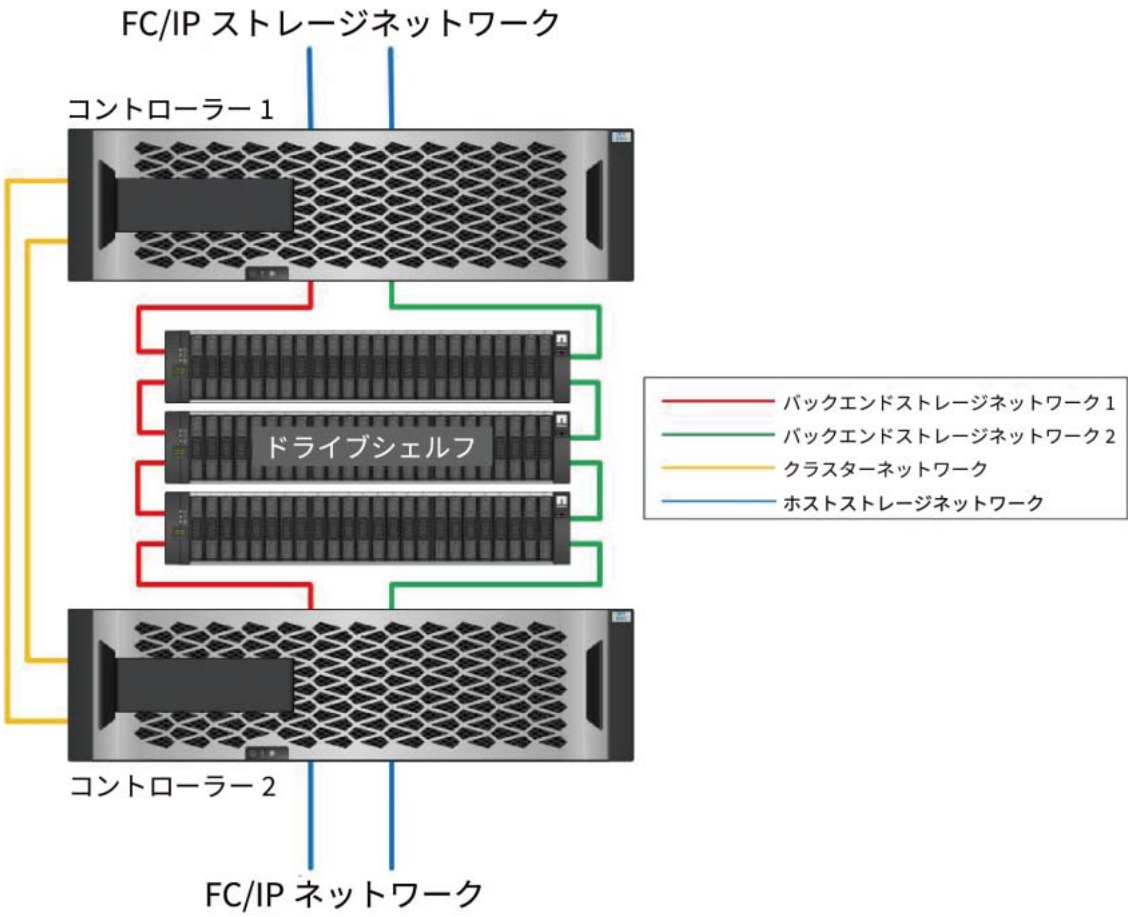
3.1 高可用性

ONTAP の高可用性機能の詳細については、本書では説明しません。ただし、データ保護と同様に、データベースインフラストラクチャを設計する場合は、この機能の基本を理解することが重要です。

3.1.1 HA ペア

高可用性機能の基本単位は HA ペアです。

図 3.1 HA ペア



3.1.2 NVRAM

各ペアには、NVRAM データのレプリケーションをサポートするための冗長リンクが含まれています。NVRAM はライトキャッシュではありません。コントローラー内の RAM は、ライトキャッシュとして機能します。NVRAM の役割は、予期しないシステム障害に対する保護策として、データを一時的に保管することです。この点では、データベーストランザクションログに似ています。NVRAM とデータベーストランザクションログの両方を使用すると、データを迅速に保存し、データへの変更を可能な限り迅速に適用できます。ドライブ上の永続データへの更新は、チェックポイントと呼ばれるプロセスになるまで行われません。通常の操作では、NVRAM データもデータベース REDO ログも読み取られません。

コントローラーに突然障害が発生した場合、ドライブへの書き込みがまだ行われていない、保留中の変更が NVRAM に保存されている場合があります。パートナーコントローラーが障害を検出し、ドライブの制御を取って、NVRAM に保存されている必要な変更を適用します。

3.1.3 ケーブルの冗長性

上の図は HA ペアケーブルの例です。ただし、正確なレイアウトはコントローラーとドライブタイプによって異なります。いずれの場合も、冗長データパスが存在します。冗長データパスには、単一シャーシ内のバックプレーン上の物理ケーブルあるいは電気的な配線があります。コントローラーには、クラスタ内のもう一つのコントローラーと接続する最低 2 つのパスがあり、NVRAM を通じた変更の複製、I/O 操作を行うクラスタ内通信の補助、またはクラスタ内のデータを無停止で再配置します。

3.1.4 電源の冗長性

すべてのコントローラー、ドライブシェルフ、およびその他のコンポーネントには、冗長化された電源が搭載されています。システムは通常、デュアル PDU (配電ユニット) を搭載したサーバーラックに配置され、それぞれがデータセンター内の異なる UPS 保護回路に接続されます。

3.1.5 テイクオーバーとギブバック

テイクオーバーおよびギブバックとは、HA ペア内のノード間でストレージリソースの管理を移すプロセスのことです。テイクオーバーとギブバックには 2 つの側面があります。

- ホストがストレージシステムにアクセスするために使用するネットワーク接続の管理
- ストレージシステム内のドライブの管理

ASA システムでのテイクオーバーおよびギブバック中は、iSCSI や FC などの SAN ブロックプロトコルをサポートするネットワークインターフェイスの再配置が即座に実行されません。コントローラーに突然障害が発生した場合、パートナーコントローラーはそのインターフェイスを使用してデータを処理し続けます。プロセスの後半では、IP アドレスの移動 (iSCSI LIF フェイルオーバー) または HBA WWN の再配置 (FCP および NVMe/FC 永続ポート) によって、障害が発生したインターフェイスがパートナーコントローラーでオンラインになり、ホストはストレージとの間の障害が発生したパスを認識しなくなります。

備考

より大きなクラスタ内のノード間でのデータの再配置をサポートするために、追加のコントローラーへの追加パスを設定することもできますが、これは HA プロセスの一部ではありません。

テイクオーバーとギブバックの 2 つ目の側面は、ドライブの所有権の移動です。実際のプロセスは、テイクオーバーまたはギブバックの理由や発行されたコマンドラインオプションなど、複数の要因によって異なります。操作をできるだけ効率的に実行することが目的です。プロセス全体では数分かかるように見えますが、ドライブの所有権がノード間で移動するときに実際にかかる時間は、数秒と測定されています。

3.1.6 テイクオーバーのトリガー

テイクオーバーは、以下のようなさまざまな理由で実行されます。

- `storage failover takeover` コマンドを使用して手動でテイクオーバーを開始した場合。
- ソフトウェアまたはシステムの障害が発生し、コントローラーでパニックが発生した場合。パニックが完了し、コントローラーが再起動すると、ストレージリソースが返され、システムが正常に戻ります。
- コントローラーに電源喪失などの完全なシステム障害があり、再起動できない場合。
- パートナーコントローラーがハートビートメッセージを受信できない場合。この状況は、パートナーでハードウェア障害またはソフトウェア障害が起きた場合に発生する可能性があります。この障害はパニックを引き起こすことはありませんが、パートナーの正常な動作を阻害します。

3.1.7 NDO

「無停止運用」には、コントローラーの突然の障害に対処する機能だけでなく、コントローラーのオンラインアップグレードおよびメンテナンスを可能にする機能も含まれます。コントローラーがデータサービスの管理をパートナーに委譲すると、管理者は ONTAP OS のアップグレード、障害が発生したハードウェアの交換、新しいアダプタの追加、またはコントローラー自体のアップデートを行うことができます。

3.1.8 テイクオーバーにかかる時間

ASA システムでは、両方のコントローラーを経由するアクティブ - アクティブパスにより、ホスト OS はアクティブパスがダウンするのを待つことなく代替パスをアクティブにすることができます。これは、すでにホストがすべてのコントローラー上のすべてのパスを使用しており、ホストが安定状態にあるかどうかに関係なく常にアクティブパスが存在している場合、またはコントローラーのフェイルオーバー操作を実行している場合に可能です。

また、ASA には、SAN のフェイルオーバープロセスを大幅に高速化する独自の機能があります。各コントローラーは、主要な LUN メタデータをパートナーに継続的にレプリケートします。つまり、各コントローラは、障害が発生したコントローラーが故障する前に管理していたドライブの使用を開始するために必要なコア情報をすでに持っているため、フェイルオーバー処理が完了する前であっても、パートナーの突然の障害が発生した場合にただちにデータの提供を開始できるように準備されています。

表 3.1 テイクオーバーにかかる時間

テイクオーバータイプ	IO 再開時間
計画的テイクオーバー	2 ～ 3 秒
計画外テイクオーバー	2 ～ 3 秒

この時間は、オペレーティングシステムの完全な I/O 再開時間を反映しています。テイクオーバーにかかる時間は、ストレージシステムが IO に応答する能力を測定するだけであればこれより短くなりますが、より重要なのは、ホストから見た完全な IO 再開時間です。

3.2 データの整合性

ONTAP 内の論理データ保護には、以下の 3 つの主要な要件があります。

- データを破損から保護する必要があります。
- データをドライブ障害から保護する必要があります。
- データへの変更を損失から保護する必要があります。

この 3 つのニーズについては、以降のセクションで説明します。

3.2.1 ネットワークの破損：チェックサム

データ保護の最も基本的なレベルはチェックサムです。これは、データとともに格納される特殊なエラー検出コードです。ネットワーク転送中のデータの破損は、単一のチェックサム、場合によっては複数のチェックサムを使用して検出されます。

たとえば、FC フレームには、ペイロードが転送中に破損していないことを確認するために、Cyclic Redundancy Check (CRC：巡回冗長検査) と呼ばれるチェックサムの形式が含まれています。送信機は、データとデータの CRC の両方を送信します。FC フレームの受信側は、受信したデータの CRC を再計算して、送信した CRC と一致することを確認します。新しく計算された CRC がフレームに添付された CRC と一致しない場合はデータが破損していることになり、FC フレームは破棄または拒否されます。iSCSI の I/O オペレーションには、TCP/IP および Ethernet レイヤーでのチェックサムが含まれます。また、保護を強化するために、SCSI レイヤーでオプションの CRC 保護を追加することもできます。回線上のビット破損は TCP レイヤーまたは IP レイヤーによって検出され、パケットの再送信が発生します。FC の場合と同様に、SCSI CRC にエラーがあると、操作が破棄または拒否されます。

3.2.2 ドライブの破損：チェックサム

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

3.2.3 データの破損：書き込みの損失

検出が最も困難なタイプの破損の 1 つは、書き込みの消失または配置の誤りです。書き込みが確認されたら、正しい場所にあるメディアに書き込む必要があります。データとともに保存された単純なチェックサムを使用することで、インプレースでのデータ破損を比較的簡単に検出できます。しかし、単に書き込みが失われただけの場合は、以前のバージョンのデータがメディア上にまだ存在している可能性があり、そのチェックサムの根拠となるブロックに間違いがない場合があります。書き込みが物理的に誤った場所に配置された場合、書き込みによって他のデータが破損したとしても、保存されたデータに対して関連するチェックサムが再び有効になってしまいます。

この課題に対する解決策は以下のとおりです。

- 書き込み操作には、書き込みが検出されると予想される場所を示すメタデータを含める必要があります。
- 書き込み操作には、何らかのバージョン識別子を含める必要があります。

ONTAP は、ブロックを書き込むときに、そのブロックが属する場所のデータを含めます。例えば、後続の読み取りでブロックが検出されたが、456 番の位置で検出されたメタデータが示す位置が 123 番である場合、誤った場所に書き込まれたことが分かります。

完全に失われた書き込みを検出するのはさらに困難です。説明は非常に複雑ですが、基本的に ONTAP は、書き込み操作によってドライブ上の 2 つの異なる場所が更新されるようにメタデータを格納します。書き込みが失われた場合、データと関連するメタデータの後続の読み取りでは、2 つの異なるバージョン ID が表示されます。これは、ドライブによる書き込みが完了しなかったことを示します。

書き込みの消失や誤った位置への書き込み破損は非常にまれですが、ドライブの増加とデータセットのエクサバイト規模への増大に伴い、リスクは増大します。書き込み消失の検出は、重要なデータセットをサポートするすべてのストレージシステムに含める必要があります。

3.2.4 ドライブの故障：RAID4、RAID DP、および RAID-TEC

ドライブ上のデータブロックが破損していることが検出された場合、またはドライブ全体に障害が発生して完全に使用できない場合は、データを再構成する必要があります。これは、パリティドライブを備えた ONTAP で実行されます。データは複数のデータドライブにストライプ化され、パリティデータが生成されます。パリティデータは、元のデータとは別に保存されます。

ONTAP はもともと RAID 4 を使用していました。RAID 4 では、データドライブのグループごとに 1 つのパリティドライブが使用されます。そのため、グループ内でどれか 1 台のドライブで障害が発生しても、データが消失することはありません。パリティドライブに障害が発生したとしても、データは破損していないため、新しいパリティドライブを構築できます。1 台のデータドライブに障害が発生した場合は、残りのドライブをパリティドライブとともに使用して、失われたデータを再生成できます。

統計上、ドライブが小さい場合に 2 台のドライブが同時に障害を起こす確率はごくわずかでした。ドライブの容量が増加するにつれて、ドライブ障害後のデータの再構築に必要な時間も増加します。これにより、2 台目のドライブで障害が発生してデータが失われる確率が高くなりました。さらに、再構築プロセスでは、残りのドライブに大量の追加 I/O が作成されます。ドライブが古くなると、追加の負荷がかかることで 2 台目のドライブで障害が発生する確率が高くなります。最終的には、RAID 4 を継続的に使用することでデータ消失のリスクが増大しない場合でも、データ消失の影響はより深刻になります。RAID グループに障害が発生した場合に失われるデータが多いほど、データのリカバリに時間がかかり、業務の中断が長くなります。

これらの問題から、RAID 6 の一種である RAID DP テクノロジーが開発されました。このソリューションには 2 台のパリティドライブが含まれます。つまり、RAID グループ内で 2 台のドライブに障害が発生しても、データは失われません。成長し続けるドライブに対応して、3 台目のパリティドライブを導入した RAID-TEC テクノロジーが開発されました。

SAN の従来のベストプラクティスの中には、ストライプミラーリングとも呼ばれる RAID 1+0 の使用を推奨するものがあります。RAID 1+0 は、RAID DP に比べるとデータを保護する能力が高くありません。RAID 1+0 では 2 台のドライブが故障する状況が複数考えられるのに対し、RAID DP では考えられないからです。

また、従来の SAN ベストプラクティスを説明したマニュアルにも、パフォーマンス上の懸念から RAID 4/5/6 オプションよりも RAID 1+0 が優先されることを示すものがあります。これらの推奨事項では、RAID ペナルティについて言及している場合があります。これらの推奨事項は一般的に正しいものですが、ONTAP 内での RAID の実装には適用されません。パフォーマンスの問題は、パリティの再生成に関連しています。従来の RAID 実装では、書き込みを処理するときにパリティデータを再生成して書き込みを完了するため、複数のディスク読み取りが必要です。ペナルティは、書き込み操作の実行に必要な追加の読み取り IOPS として定義されます。

書き込みはパリティが生成されるメモリ内でステージングされ、その後単一の RAID ストライプとしてディスクに書き込まれるため、ONTAP では RAID ペナルティは発生しません。書き込み操作を完了するために読み取りは必要ありません。

要約すると、RAID 1+0 と比較した場合、RAID DP および RAID-TEC は、有効容量が大幅に増加し、ドライブ障害に対する保護が向上し、パフォーマンスを犠牲にすることはありません。

3.2.5 ハードウェア故障からの保護：NVRAM

遅延の影響を受けやすいワークロードにサービスを提供するストレージシステムでは、できるだけ早く書き込み操作を認識する必要があります。さらに、書き込み操作は、電源障害などの予期しない事象による損失から保護する必要があります。つまり、書き込み操作は少なくとも 2 つの場所に安全に保存する必要があります。

ASA システムは、これらの要件を満たすために NVRAM に依存しています。書き込みプロセスは、以下のように動作します。

1. インバウンド書き込みデータは RAM に格納されます。
2. ドライブ上のデータに対して変更を行う必要がある場合は、ローカルノードとパートナーノードの両方で NVRAM に記録されます。NVRAM は書き込みキャッシュではありませんが、データベース redo ログに似たジャーナルを保存する記録媒体です。通常の状態では、読み取りは行われません。これは、I/O 処理中の電源障害後などに行われるリカバリにのみ使用されます。
3. 次に、書き込みがホストに確認されます。

この段階で、書き込みプロセスはアプリケーションの観点からは完了しています。データは 2 つの異なる場所に保存されるため、損失から保護されます。最終的に変更はドライブに書き込まれますが、このプロセスは書き込みが確認された後に発生するためレーテンシには影響せず、アプリケーションの観点では帯域外となります。このプロセスもデータベースロギングに似ています。データベースへの変更はできるかぎり迅速に redo ログに記録され、変更はコミットされたものとして認識されます。データファイルの更新はかなり後に行われるため、処理速度に直接影響することはありません。

コントローラーに障害が発生すると、パートナーコントローラーが必要なドライブの所有権を取得し、NVRAM に記録されたデータを再生して、障害発生時に処理中だった I/O 操作を回復します。

3.2.6 冗長故障：NVFAIL

前述したように、少なくともう 1 台のコントローラー上のローカル NVRAM および NVRAM に記録されるまで、書き込みは確認されません。このアプローチでは、ハードウェアの障害や停電によって、処理中の I/O が失われないようにします。ローカル NVRAM に障害が発生したり、他のノードへの接続に障害が発生したりすると、データはミラーリングされなくなります。

ローカル NVRAM がエラーを報告すると、ノードはシャットダウンします。HA ペアが使用されている場合、このシャットダウンによってパートナーコントローラーへのフェイルオーバーが発生します。MetroCluster を使用すると、動作は選択した全体的な構成によって異なりますが、リモートコントローラーへの自動フェイルオーバーが行われる可能性があります。

いずれの場合も、障害が発生しているコントローラーが書き込みを確認していないため、データ喪失は発生しません。データ喪失とは、確認された書き込みが失われることを意味します。しかし、ファイルシステムとアプリケーションを使用したブロックストレージ管理では、原則として、確認応答が送信される前の未確認の書き込みが永続ストレージに存在するかしないかが分かりません。これは、ストレージで受信する前に書き込みが失われたのか、確認応答だけが失われたのかを OS から判断する方法がないためです。確認応答が受信されるまでは、書き込みの状態は不定となります。

4. データ保護

前章では、ストレージハードウェアに関するデータの可用性とデータの整合性について説明しました。データの可用性と整合性に関して同様に重要なのは、ユーザーやアプリケーションの回避できないエラーからリカバリする機能です。ストレージシステムから 99.9999% のアップタイムを必要とする企業は、増大するデータセットを迅速かつ確実にリカバリできるバックアップ / リカバリ戦略も計画する必要があります。

4.1 スナップショットコピーによるデータ保護

ONTAP データ保護ソフトウェアの基盤は、スナップショットテクノロジーです。主な有用性は以下のとおりです。

- **単純性**
  スナップショットコピーは、特定の時点におけるデータコンテナの内容の読み取り専用コピーです。
- **効率性**
  スナップショットコピーは、作成時に領域を必要としません。領域は、データが変更された場合にのみ消費されます。
- **管理性**
  スナップショットコピーはストレージ OS の基本機能であるため、スナップショットコピーに基づくバックアップ戦略の構成と管理が容易です。ストレージシステムの電源がオンになっていれば、バックアップを作成できます。
- **拡張性**
  単一 LUN で最大 1024 個のスナップショットをローカルに保持できます。複雑なデータセットの場合は、一貫性のある 1 組のスナップショットコピーによって、データの複数のコンテナを保護できます。
  ボリュームに 1024 個のスナップショットコピーが含まれているかどうかにかかわらず、パフォーマンスに影響はありません。

その結果、ONTAP 上で実行されているデータセットの保護は、シンプルで拡張性に優れています。バックアップでは、データを移動する必要はありません。したがって、バックアップ戦略をビジネスのニーズに合わせて調整でき、ネットワーク転送速度の制限、多数のテープドライブ、または高価なドライブのステージング領域の影響を受けることがありません。

4.2 ONTAP SnapRestore によるデータのリストア

スナップショットコピーからの ONTAP での高速データリストアは、SnapRestore テクノロジーによって実現されます。主な有用性は以下のとおりです。

- 個々のファイルまたは LUN は、16TB LUN または 4KB ファイルであっても数秒でリストアできます。
- LUN やファイルのコンテナ全体 (FlexVol ボリューム) は、10GB または 100TB のデータであっても数秒でリストアできます。

重要なアプリケーションが停止すると、重要なビジネスオペレーションが停止してしまいます。テープは破損する場合があり、ドライブベースのバックアップからのリストアでは、ネットワーク経由での転送に時間がかかることがあります。SnapRestore は、重要なデータセットをほぼ瞬時にリストアすることにより、これらの問題を回避します。ペタバイト規模のデータベースでも、わずか数分の作業で完全にリストアできます。

4.3 SnapMirror によるリモートデータ保護

SnapMirror は、管理が容易で、拡張性と効率性に優れたレプリケーションテクノロジーです。また、データだけでなくスナップショットもレプリケートできます。バックアップの一部またはすべてを、リモートサイトまたはクラウドに選択的に保存できます。これにより、どのような場合でもバックアップを利用できるようになります。また、スナップショットテクノロジーの効率性により、ネットワークインフラストラクチャおよびストレージ容量の要件が最小限に抑えられます。

4.4 ONTAP FlexClone によるデータのリストア

すべてのデータセットをそのままリストアできるわけではありません。場合によっては、データセットを復元するのではなく修復する必要があります。データのリストアと同じテクノロジーにより、現在のデータに影響を与えずにクローンを作成することもできます。

- 個々のファイルまたは LUN は、16TB LUN であっても 4KB ファイルであっても、数秒でクローンを作成できます。
- LUN やファイルのコンテナ全体 (FlexVol ボリューム) のクローンは、データが 10GB でも 100TB でも、数秒で作成できます。
- クローンは、ローカル、リモート、クラウドなど、データの任意のコピーから作成できます。

管理者は、クローンを確認し、必要に応じてデータを抽出し、データセットを修復できます。

5. ディザスタリカバリ

単一の ASA システムは、ハードウェアレベルでの最大の可用性と、ユーザーおよびアプリケーションのエラーに対処する高速リストア機能を提供しますが、災害についてはどうでしょうか。電源が完全に切断された場合、またはサイトが破壊された場合に、データの可用性を継続的に維持するにはどうすればよいのでしょうか。

ASA システムは、2 つのオプションをサポートします。MetroCluster および SnapMirror Business Continuity です。

備考

ASA システム上のスナップショットは、災害時のデータ損失が許容される場合、ディザスタリカバリのニーズに合わせて非同期にレプリケートすることもできます。

5.1 MetroCluster テクノロジー

MetroCluster は、ミッションクリティカルなワークロードのための高可用性、データ損失ゼロのソリューションを提供します。さらに、MetroCluster などの統合ソリューションは、今日の複雑なスケールアウト型エンタープライズアプリケーションと仮想化インフラストラクチャをシンプル化します。MetroCluster は、複数の外部データ保護製品および戦略を、中心となる 1 台のシンプルなストレージシステムに置き換えます。このストレージシステムは、単一クラスタのストレージシステム内で統合バックアップ、リカバリ、ディザスタリカバリ、および高可用性 (HA) を提供します。

5.1.1 MetroCluster による HA

MetroCluster レプリケーションは、SyncMirror テクノロジーに基づいています。このテクノロジーは、同期モードへの切り替えを効率的に行うように設計されています。この機能は、同期レプリケーションを必要としながら、データサービスの高可用性も必要とするお客様の要件を満たします。たとえば、リモートサイトへの接続が切断された場合、一般的には、レプリケーションされていない状態でストレージシステムを動作させることが望まれます。

多くの同期レプリケーションソリューションは、同期モードでのみ動作可能です。このタイプの全か無かの複製は、Domino モードと呼ばれることもあります。このようなストレージシステムでは、データのローカルコピーとリモートコピーが非同期になるのではなく、データの提供が停止します。レプリケーションが強制的に中断されると、再同期に非常に時間がかかり、ミラーリングの再確立時にデータが完全に失われるおそれがあります。

SyncMirror は、リモートサイトにアクセスできない場合に同期モードからシームレスに切り替えることができるだけでなく、接続がリストアされたときに RPO=0 の状態に迅速に再同期できます。リモートサイトのデータの古いコピーは、再同期中に使用可能な状態で保存することもできます。これにより、データのローカルコピーとリモートコピーが常に存在することが保証されます。

5.1.2 MetroCluster と SyncMirror

ONTAP での同期レプリケーションは、SyncMirror によって提供されます。最も単純なレイヤーでは、SyncMirror は 2 つの異なる場所に RAID で保護された完全なデータセットを 2 つ作成します。データセットは、データセンター内の隣接する部屋に配置することも、何キロも離れた場所に配置することもできます。

SyncMirror は ONTAP と完全に統合されており、RAID レベルのすぐ上で動作します。したがって、スナップショットコピー、SnapRestore、FlexClone など、通常の ONTAP 機能はすべてシームレスに動作します。これはまだ ONTAP で行われるものです。同期データミラーリングの追加レイヤーが含まれているだけにすぎません。

SyncMirror データを管理する ONTAP コントローラーを総合して、MetroCluster と呼びます。多くの構成が利用可能であり、さまざまな一般的なディザスタリカバリ障害シナリオにおいて、同期ミラーリングされたデータへの高可用性アクセスを提供することが MetroCluster の主な目的となっています。

MetroCluster と SyncMirror を使用したデータ保護の主な有用性は、以下のとおりです。

- 通常のオペレーションでは、SyncMirror はロケーション間の同期ミラーリングを保証します。書き込み操作は、両方のサイトの不揮発性メディアに記録されるまで認識されません。
- サイト間の接続に障害が発生した場合、SyncMirror は自動的に非同期モードに切り替わり、接続が復元されるまで、データを提供するプライマリサイトを維持します。リストアされると、プライマリサイトに蓄積された変更を効率的に更新することによる再同期を速やかに実施します。完全な再初期化は必要ありません。

SnapMirror は SyncMirror ベースのシステムとも完全に互換性があります。たとえば、2 つの地理的なサイトにまたがる MetroCluster クラスタでプライマリデータベースが実行されているとします。このデータベースは、第 3 のサイトでの長期アーカイブ用、あるいは DevOps 環境でのクローン作成用として、バックアップをレプリケートすることもできます。

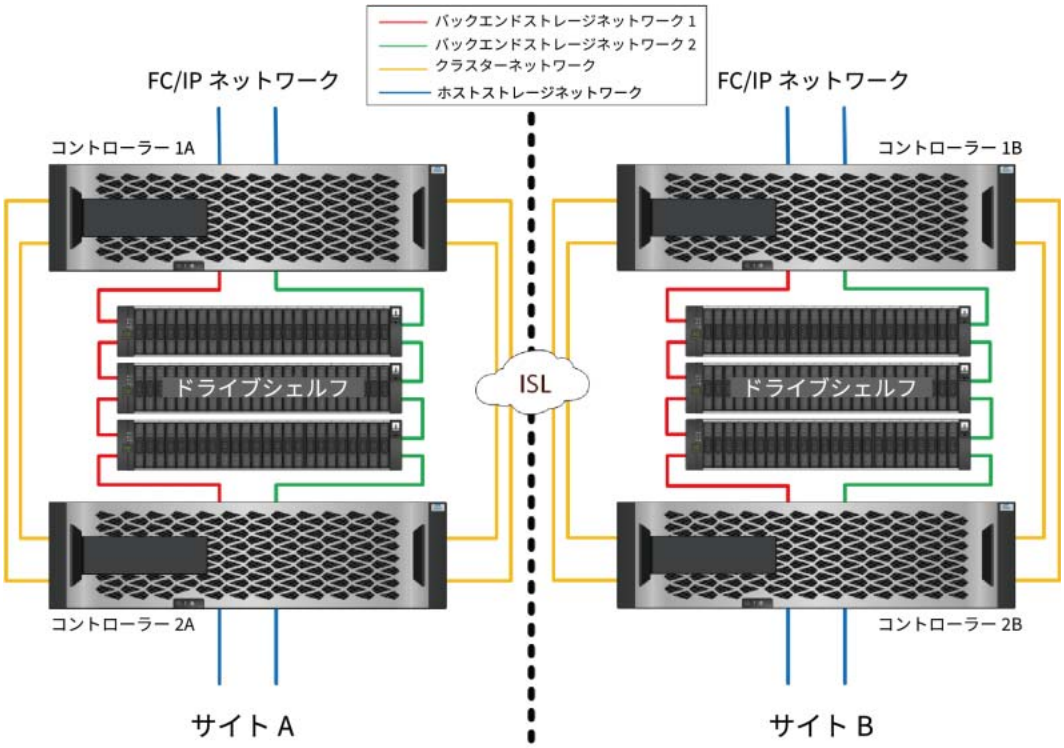
5.1.3 MetroCluster アーキテクチャ

本書では MetroCluster の完全な説明を記載していませんが、MetroCluster の主要な可用性機能を理解する必要があります。以下のセクションでは、IP ベースの MetroCluster を使用して説明します。今日では高速で低レイテンシの IP 回線がより容易に利用できるようになり、インフラストラクチャ要件がよりシンプルであるため、お客様のほとんどは IP 接続を選択します。

詳細については、ONTAP の公式マニュアルを参照してください。

IP 接続を使用する MetroCluster システムは、各サイトで HA ペアを使用して構成されています。

図 5.1 MetroCluster IP の基本アーキテクチャ



5.1.4 MetroCluster Resiliency の機能

上の図に示すように、MetroCluster ソリューションには単一点障害はありません。

- 各コントローラーには、ローカルサイトのドライブシェルフへの独立したパスが 2 つあります。
- 各コントローラーには、リモートサイトのドライブシェルフへの独立したパスが 2 つあります。
- 各コントローラーには、対となるサイトのコントローラーへの独立したパスが 2 つあります。
- HA ペア設定では、各コントローラーにローカルパートナーへのパスが 2 つあります。

つまり、MetroCluster のデータ処理機能を損なうことなく、構成内の任意の 1 つのコンポーネントを削除できます。この 2 つのオプションによる回復の唯一の違いは、サイト障害後も HA ペアのバージョンが全体的な HA ストレージシステムである点です。

5.1.5 サイト障害からの保護：NVRAM と MetroCluster

MetroCluster は、NVRAM データをローカルパートナーとリモートパートナーの両方にレプリケートすることにより、NVRAM データ保護を拡張します。書き込みは、すべてのパートナーにレプリケートされるまで確認されません。

このアーキテクチャは、サイト障害から処理中の I/O を保護します。このプロセスは、ドライブレベルのデータレプリケーションには関係しません。アグリゲートを担当するコントローラーは、アグリゲート内の両方のプレックスに書き込むことによってデータのレプリケーションを行いますが、サイトで障害が発生した場合には、処理中の I/O 損失に対する保護が必要です。レプリケートされた NVRAM データは、パートナーコントローラーが障害の発生したコントローラーを引き継ぐ必要がある場合にのみ使用されます。

5.1.6 サイトおよびシェルフ障害からの保護：SyncMirror とプレックス

SyncMirror は、RAID DP または RAID-TEC を強化するミラーリングテクノロジーですが、これらに取って代わるものではありません。SyncMirror は、2 つの独立した RAID グループの内容をミラーリングします。論理構成は以下のとおりです。

- ドライブは、場所に基づいて 2 つのプールに構成されます。1 つのプールはサイト A のすべてのドライブで構成され、もう 1 つのプールはサイト B のすべてのドライブで構成されます。
- 次に、アグリゲートと呼ばれる共通のストレージプールが、ミラーリングされた RAID グループのセットに基づいて作成されます。各サイトから同数のドライブが引き出されます。たとえば、20 台のドライブで構成された SyncMirror のアグリゲートは、サイト A のドライブ 10 台とサイト B のドライブ 10 台で構成されます。
- 特定のサイトの各ドライブセットは、ミラーリングの使用に関係なく、1 つ以上の完全な冗長化 RAID DP または RAID-TEC グループとして自動的に構成されます。このようにミラーリングの下に RAID を使用することで、サイトが消失した後もデータを保護できます。

図 5.2 SyncMirror

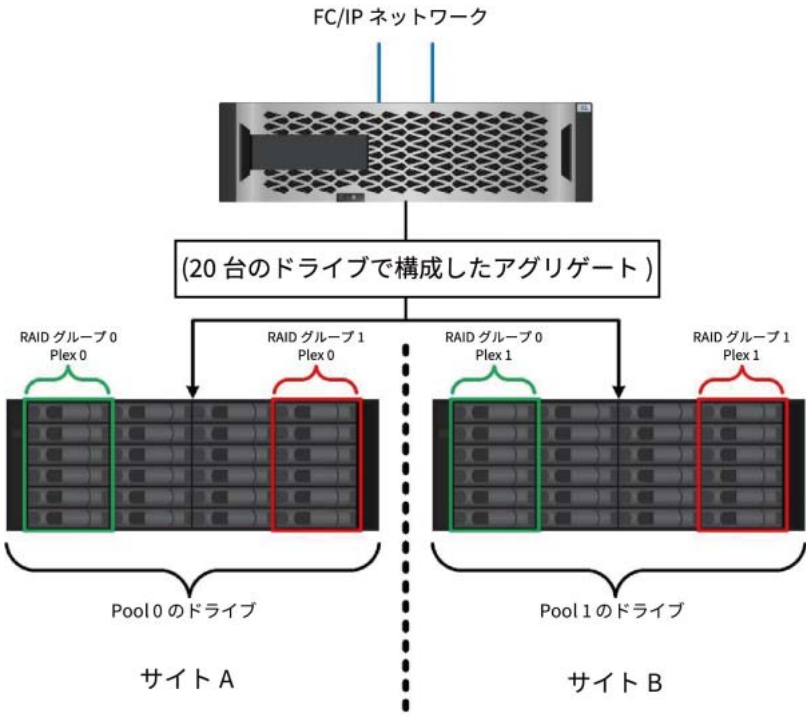


図 5.2 に、SyncMirror 構成の例を示します。24 台のドライブから成るアグリゲートは、サイト A のシェルフに属する 12 台のドライブと、サイト B のシェルフに属する 12 台のドライブで構成されています。各ドライブは、2 つのミラーリングされた RAID グループに分けられます。RAID グループ 0 には、サイト A に属する 6 台のドライブから成るプレックスをサイト B に属する 6 台のドライブから成るプレックスにミラーリングしたものが含まれます。同様に、RAID グループ 1 には、サイト A に属する 6 台のドライブから成るプレックスをサイト B に属する 6 台のドライブから成るプレックスにミラーリングしたものが含まれます。

通常、SyncMirror は MetroCluster システムでリモートミラーリングを行うために使用され、各サイトにデータのコピーが 1 つずつ作成されます。場合によっては、1 つのシステムに追加レベルの冗長性を提供するために使用されています。特に、シェルフレベルの冗長性を提供します。ドライブシェルフには、すでに冗長電源と冗長コントローラーが搭載されており、全体的にはシートメタルにすぎませんが、場合によっては追加の保護が保証されることがあります。たとえば、あるお客様は、自動車のテスト中に使用する携帯式のリアルタイム分析プラットフォームに SyncMirror を導入しました。このシステムは、独立電源供給と独立 UPS システムを備えた 2 つの物理ラックに分割されました。

5.1.7 ハードウェア支援型テイクオーバー

サービスプロセッサは、ONTAP システムに内蔵された帯域外管理装置です。サービスプロセッサ自体の IP アドレスによってアクセスされ、コントローラーが動作しているかどうかに関係なく、コンソールへの直接アクセスやその他の管理機能に使用されます。

ONTAP 自体は、パートナーノードからのハートビートを検出しなくなった後に障害が発生したノードのテイクオーバーをトリガーできますが、これにはタイムアウトが関係しています。ハードウェア支援型テイクオーバーでは、サービスプロセスを使用して、障害をより迅速に検出し、テイクオーバーをただちに開始することで、テイクオーバープロセスを高速化します。パートナーのハートビートが停止したことを ONTAP が認識するのを待たずに実行されます。

5.1.8 スイッチオーバーとスイッチバック

スイッチオーバーおよびスイッチバックという用語は、MetroCluster 構成のリモートコントローラー間でボリュームを移行するプロセスを指します。このプロセスは、リモートノードにのみ適用されます。MetroCluster が 4 つのボリュームからなる構成で使用されている場合、ローカルノードのフェイルオーバーは、前述のテイクオーバーおよびギブバックプロセスと同じです。

5.1.9 計画内スイッチオーバーと計画内スイッチバック

計画内スイッチオーバーまたは計画内スイッチバックは、ノード間のテイクオーバーまたはギブバックに似ています。このプロセスには複数のステップがあり、数分かかるように見える場合がありますが、実際には、ストレージとネットワークリソースの多段階の適切な移行が行われています。制御が転送される瞬間は、完全なコマンドが実行されるのに必要な時間よりもはるかに速くなります。

テイクオーバーギブバックとスイッチオーバースイッチバックの主な違いは、SAN 接続への影響です。ローカルテイクオーバーギブバックでは、ホストはローカルノードへのすべての SAN パスを失い、ネイティブ MPIO に依存して使用可能な代替パスに切り替えます。ポートは再配置されません。スイッチオーバーおよびスイッチバックにより、コントローラー上の仮想 FC ターゲットポートは他のサイトに移行します。この方法は、仮想 FC ポートが SAN から瞬間的に移行して、代替コントローラー上に再出現する際に効果的です。

5.1.10 MetroCluster IP を使用した ONTAP Mediator

ONTAP Mediator は、MetroCluster IP およびその他の特定の ONTAP ソリューションで使用されます。従来の Tiebreaker サービスとして機能します。

主な機能は、NVRAM と SyncMirror が同期されているかどうかを判断することです。MetroCluster では、NVRAM と基盤となるアグリゲートのプレックスが同期されているため、各サイトのデータが同一であることを確認でき、データ損失のリスクなしにスイッチオーバーを続行できます。

ONTAP では、フェイルオーバーまたはスイッチオーバーが強制実行されない限り、データが同期していない場合、フェイルオーバーまたはスイッチオーバーは許可されません。このように条件の変更を強制する場合、データが元のコントローラーに残されている可能性があり、データ損失を許容することを認識してください。

5.2 SnapMirror Business Continuity

MetroCluster は、サービスの停止を最小限に抑えつつ、災害発生時に環境全体のデータ損失をゼロに抑えることができる理想的なソリューションです。ただし、すべてのお客様がアレイ全体に対して RPO=0 のデータ保護を望んでいるわけではありません。選択したデータセットのみが RPO=0 の同期データ保護を必要とする場合があります。

このニーズに対応するため、ONTAP 9.9.1 では SM-BC (SnapMirror Business Continuity) が導入されました。SM-BC と SM-S (SnapMirror Synchronous) はレプリケーションエンジンを共有していますが、SM-BC には、透過的なアプリケーションフェイルオーバーやフェイルバックなどの追加機能が含まれています。

5.2.1 モード

SM-BC は、2 つのモードのいずれかで動作します。同期モードは MetroCluster に似ています。通常の操作では、RPO=0 が維持され、すべての書き込みがローカルシステムとリモートシステムにコミットされます。ただし、書き込みをレプリケートできない場合、同期モードはタイムアウトし、処理を続行できます。サイト障害が発生すると、リモートサイトが元のデータと同期しなくなるため、データが失われます。

これはほとんどのお客様に推奨されるモードですが、変更を両方のレプリカにコミットする必要があるか、逆にまったくコミットしないワークロードの場合、SM-BC には StrictSync モードが含まれます。この場合、変更をレプリケートできない状態が長く続くと、I/O を実行しているオペレーティングシステムにエラーが報告されます。これにより、通常はアプリケーションがシャットダウンされます。

5.2.2 パスアクセス

SM-BC は、プライマリストレージシステムとリモートストレージシステムの両方からホストオペレーティングシステムがストレージデバイスを認識できるようにします。

ローカルコントローラーへのパスはアクティブ / 最適化パスとして指定され、リモートコントローラーへのパスはアクティブ / 非最適化パスとして指定されます。通常の操作では、すべての I/O はアクティブ / 最適化パスをホストするローカルコントローラーによって処理されます。サイト障害やリモートサイトへのストレージフェイルオーバーが発生した場合、アクティブ / 非最適化パスは最適化パスに移行します。

5.2.3 フェイルオーバー

SM-BC は、計画的および計画外の 2 種類のストレージフェイルオーバー操作をサポートします。これらの操作は、わずかに異なる方法で機能します。

計画的なフェイルオーバーは、リモートサイトへの迅速なスイッチオーバーのために管理者が手動で開始します。一方の計画外のフェイルオーバーは、3 番目のサイトの Mediator によって自動的に開始されます。計画的なフェイルオーバーの主な目的は、段階的なパッチ適用とアップグレードを実行すること、ディザスタリカバリテストを実行すること、完全なビジネス継続性機能を証明するために年間を通じてサイト間でオペレーションを切り替える正式なポリシーを採用することです。

5.2.4 ストレージハードウェア

他のディザスタリカバリストレージソリューションとは異なり、SM-BC は非対称プラットフォームの柔軟性を提供します。つまり、各サイトのハードウェアが同一である必要はありません。この機能により、SM-BC のサポートに使用するハードウェアのサイズを適切に設定できます。本番ワークロード全体をサポートする必要がある場合は、リモートストレージシステムをプライマリサイトと同一にすることができますが、災害によって I/O が減少した場合は、リモートサイトのシステムのサイズを小さくする方がコストパフォーマンスがよくなります。

5.2.5 ONTAP Mediator

ONTAP Mediator は、当社のファームウェアダウンロードサイトからダウンロードできるソフトウェアアプリケーションです。Mediator は、プライマリサイトとリモートサイトの両方のストレージクラスタのフェイルオーバー操作を自動化します。オンプレミスまたはクラウドにホストされた小規模な仮想マシン (VM) に適用することができます。設定完了後は、両方のサイトのフェイルオーバーを監視する第 3 のサイトとして機能します。

Mediator はこのスプリットブレーンを認識し、マスターコピーを保持するノードで I/O を再開します。サイト間の接続がオンラインに戻ると、代替サイトは自動的に再同期を実行します。

6. SAN 構成のベストプラクティス

SAN の可用性を最大化するには、以下のベストプラクティスが重要です。大半はホストおよび FC ネットワーク構成に適用され、SAN の実装、オペレーティングシステム、マルチパスソフトウェアのさまざまな側面と制限が原因となっています。これらのベストプラクティスからの逸脱は有用性がある場合もありますが、管理者は考えられる結果とリスクを慎重に検討する必要があります。

6.1 独立した FC ファブリック

FC SAN ホストでは、ホストまたはネットワークスイッチの 1 つのポートで障害が発生してもシステムが停止しないようにするため、冗長ネットワーク接続が必要です。これらの 2 つのネットワーク接続では、独立した FC ファブリックも使用する必要があります。フルメッシュファブリックを使用すると、過剰な数のパスがホストに公開され、SAN 全体に影響するユーザーエラーのリスクが高まります。

6.2 独立した IP サブネット

最大の可用性を必要とする iSCSI ホストおよび NVMe/TCP ホストは、独立したサブネット上で少なくとも 2 つのネットワークアダプタ (NIC) を使用する必要があります。すべての TCP/IP 通信に共通のサブネットを使用すると、その 1 つのサブネット全体が中断され、停止するリスクが高まります。さらに、多くの OS には内部ルーティングテーブルがあり、その結果、使用可能な NIC のうち 1 つだけがネットワーク通信に使用されます。追加の NIC が存在する場合がありますが、共通のサブネットを共有している場合は、OS で使用できません。

LACP トランキングなどのホストボンディングもサブネットごとに使用できます。

たとえば、HA iSCSI または NVMe/TCP は以下のように設定できます。

- アドレスが 192.168.1.10/24 のホスト上の NIC#1
- アドレスが 192.168.2.10/24 のホスト上の NIC#2
- アドレスが 192.168.1.1/24 の ONTAP コントローラー #1 上の 2 ポート LACP トランク
- アドレスが 192.168.2.1/24 の ONTAP コントローラー #2 上の 2 ポート LACP トランク

その結果、トランクインターフェイス上の ONTAP コントローラーからの SAN リソースの負荷を分散した可用性と、ホスト上の負荷を分散した冗長性が得られます。2 つのサブネットを使用すると、ネットワークの中断によって SAN 接続が完全に中断されることがなくなります。

6.3 LUN パスの制限

近年のネットワークの SAN ホストは、通常、LUN またはネームスペースへのパスを 4 つ以上必要としません。パスが 8 つを超える構成にしないでください。

パスの数が多すぎると、OS の起動とパスのフェイルオーバーが遅延します。パスの数が多すぎると、パスの検出と管理に関するホスト OS のバグが露呈する場合があります。最終的に、ホスト上の SAN デバイスを管理する際にユーザーエラーが発生するリスクは、公開されるパスの数が増えるほど高くなります。

6.4 LUN ／ネームスペース (NS) のサイジング

最大規模の ONTAP コントローラーでも、わずか 8 つの LUN またはネームスペースで 100% のパフォーマンス容量を実現できます。1 つのアプリケーションが理論上の最大パフォーマンス容量を消費することが予想される場合は、追加が必要になることがありますが、8 つの LUN またはネームスペースを超えてパフォーマンスが徐々に向上することはほとんどありません。

LUN またはネームスペースの数が多いということは、パスの数が増えていることを意味します。そのため、前述と同じ問題が発生します。問題を回避するには、使用する LUN の数を少なくし、サイズを大きくします。たとえば、サイズが 8TB の通常のデータベースは、4 つの 2TB LUN またはネームスペースに配置する必要があります。I/O が特に高い場合は、8 つの 1TB LUN ／ネームスペースが有効です。

単一のデータセットをサポートする LUN またはネームスペースの推奨最大数は、コントローラーあたり 64 の LUN またはコントローラーあたり 16 のネームスペースです。これは、単一のコントローラーから単一のホストにアドバタイズできるストレージリソースの最大数ではありません。ただし、1 つのデータセットのワークロードに使用するリソースの最大数です。たとえば、10 のデータベースは 10 の異なるワークロードを表し、それぞれが 8 つの LUN を持ち、合計で 80 になります。

6.5 単一イニシエーターのゾーニング

常に単一イニシエーターのゾーニングを使用します。複数イニシエーターのゾーニングは多くの場合無害ですが、一部の OS や HBA ／ファームウェアの組み合わせでは、イニシエータークロストークによって断続的な問題が発生します。発生する問題は、深刻な場合もあれば予期しない場合もあります。単一イニシエーターのゾーニングでは、1 つのイニシエーターを別のイニシエーターから切り離すことにより、この問題を回避します。複数ターゲットのゾーニングは使用可能です。

6.6 SAN Host Utilities のマニュアルに応じた SAN の設定

ほとんどのオペレーティングシステムはインストール時に正常に動作しますが、一部の構成では、正常に動作させるために追加の設定が必要になる場合があります。

6.7 sanlun ユーティリティを使用したパス状態の確認

ホストユーティリティは、サポートされているすべてのオペレーティングシステムにインストールする必要があります。重要なユーティリティは、`sanlun` コマンドです。ユーザーは、`sanlun lun show-p` を実行して、パスの状態を確認できます。これは、ONTAP または SAN インフラストラクチャのアップグレードを実行する前に特に重要になります。システム停止が報告された場合のサポート例の多くは、突き詰めればパスの喪失の結果です。ホストが最初にインストールされたときに単一のコントローラーだけが SAN にゾーニングされていた場合か、SAN 構成の一時的な変更が原因になっています。

パスの数が正しいことを確認し、HA ペアの両方のコントローラーを構成に含めると、このタイプの見落としを防ぐことができます。検証を実施すれば、SAN に変更が加えられてシステムが停止する前に、OS の構成ミスや誤動作が検出されます。

`sanlun` コマンド以外にも、関連する OS のマルチパス管理ツールを使用できます。

6.8 Linux LVM に関する注意事項

Linux LVM には、パスの変更中に I/O エラーやアプリケーションのクラッシュを引き起こす設計上の欠陥があります。ブート時に、マルチパスドライバと LVM ドライバがほぼ同時に起動するため、競合状態が発生します。ほとんどの場合、multipathd は LVM が起動する前にデバイスの作成を終了しますが、必ずそうなるとは限りません。

LVM がデバイスをスキャンしたときにマルチパスデバイスが存在しなかったため、LVM がシングルパスデバイスを使用して PV デバイスを作成する可能性があります。その PV を使用している LV がマウントされていてフェイルオーバーが発生すると、その PV の唯一のパスが使用できなくなるため、その PV は消失します。このバグを引き起こすほどに大量の LUN またはネームスペースを持つ構成はごく少数ですが、実例が確認されています。

pvs の出力には、安全でない状態であることを示す兆候があります。特定の PV がマルチパスデバイスを使用していないことを警告できます。

```
  WARNING: Not using device /dev/mapper/3600a0980383038616e3f4a53716a4c7a for PV 4ZZweF-tjt9-
wLxC-CdPU-oQmT-78Wy-My6st2.
  WARNING: Not using device /dev/mapper/3600a0980383038616e3f4a53716a4d32 for PV O3IihV-zEaH-
J82B-fF8B-NGvz-dlPe-uUgb1r.
  WARNING: Not using device /dev/mapper/3600a0980383038616e3f4a53716a4d31 for PV XvjZty-T1qx-
7aHc-nrtI-yh3N-CWAv-U5gwrX.
  WARNING: Not using device /dev/mapper/3600a0980383038616e3f4a53716a4d30 for PV tl9BmZ-3dCY-
Lfvs-s7xR-3jfN-NLLT-dFGLc0.
```

この問題は、`/etc/lvm/lvm.conf` を変更することで対処できます。デフォルト設定は以下のとおりで、lvmd はすべてのデバイスの物理ボリュームをスキャンします。

```
filter = [ "a|.*/|" ]
```

一般に、以下の設定が有効ですが、慎重にテストする必要があります。

```
filter = [ "a|^/dev/sda[1-9]$|", "a|^/dev/mapper/*|", "r|^/dev/*|" ]
global_filter = [ "a|^/dev/sda[1-9]$|", "a|^/dev/mapper/*|", "r|^/dev/*|" ]
```

このフィルタにより、lvmd は /dev/sda* および /dev/mapper/* 上の物理ボリュームのみをスキャンします。ブートデバイスが /dev/sda パーティションでない場合、この設定は再起動を妨げる可能性があります。たとえば、サーバーのローカルブートデバイスが /dev/xda デバイスとして表示される場合があります。詳細については、LVM の公式マニュアルを参照してください。

注意

このファイルを変更した場合は、サーバーを再起動して、再起動が成功したことを確認します。また、エラーを修正するためにコンソールでログオンする準備をしてください。

6.9 /etc/sysconfig/oracleasm エラーに関する注意事項

Oracle データベースを ASMlibm とともに使用する場合は、`/etc/sysconfig/oracleasm` がシングルパスデバイスを検出していないことを確認してください。Linux 上のシングルパスデバイスは、マルチパスデバイスと並行して動作します。ASMlib は、マルチパスデバイスのみを検出するように構成する必要があります。

例 :

```
# ORACLEASM_SCANORDER: Matching patterns to order disk scanning
ORACLEASM_SCANORDER="mpath dm" (OR ORACLEASM_SCANORDER="dm")

# ORACLEASM_SCANEXCLUDE: Matching patterns to exclude disks from scan
ORACLEASM_SCANEXCLUDE="sd"
```

6.10 Solaris で host_config スクリプトを使用する場合の注意事項

特に、Solaris では、ONTAP マルチパスデバイスが正しく認識されるようにするために、特定の構成手順が必要です。ホスト構成の手順にしたがわないと、回復力が低下し、ZFS のパフォーマンスに重大な問題が発生する可能性があります。

6.11 NVFAIL

重要なデータを含む ONTAP ストレージ上の SAN ボリュームでは、nvfail パラメーターを on に設定する必要があります。

データベースなどのアプリケーションがドライブ上のデータの大規模な内部キャッシュを保持しているため、フェイルオーバーまたはスイッチオーバーが強制された場合、SAN ワークロードは特に破損しやすくなります。強制的なフェイルオーバーまたはスイッチオーバーが発生すると、以前に確認された変更は事実上破棄されます。ストレージシステムの内容は時間的に逆行し、データベースキャッシュの状態はディスク上のデータの状態を反映しなくなります。

nvfail 設定は、データの整合性が問題になる NVRAM ジャーナル処理の致命的な障害からボリュームを保護します。nvfail パラメーターは、起動時に有効になります。NVRAM エラーが検出された場合は、コミットされていない変更が失われ、ドライブの状態がデータベースキャッシュと一致しない場合があります。次に、ONTAP は nvfail パラメーターが on になっているボリュームを in-nvfail-state に設定します。その結果、データにアクセスしようとするプロセスに I/O エラーが発生し、データベースの保護クラッシュまたはシャットダウンが発生します。

ETERNUS AX series オールフラッシュアレイ All SAN Array (ASA) のデータ可用性と整合性

C140-0068-02Z3

発行年月　2025 年 3 月
発行責任　エフサステクノロジーズ株式会社

Fsas Technologies